3.4 Enter the following data into SPSS (time, in minutes, taken for subjects in a fitness trial to complete a certain exercise task):

31	39	45	26	23	56	45	80	35	37
25	42	32	58	80	71	19	16	56	21
34	36	10	38	12	48	38	37	39	42
27	39	17	31	56	28	40	82	27	37

Using SPSS select an appropriate graphing technique to illustrate the distribution. Justify your choice of technique against the other available options.

3.5 Consider the following list of prices, in whole dollars, for 20 used cars:

11,300	9200	8200	8600	10,600
11,100	7980	12,900	10,750	9200
13,630	9400	11,800	10,200	12,240
11,670	10,000	11,250	12,750	12,990

From these data construct a histogram using these class intervals:

7000–8499, 8500–9999, 10000–11499, 11500–12999, 13000–14499.

3.6 Construct a pie graph to describe the following data:

Migrants in local area, place of origin

Place	Number
Asia	900
Africa	1200
Europe	2100
South America	1500
Other	300
Total	6000

What feature of this distribution does your pie graph mainly illustrate?

3.7 From a recent newspaper or magazine find examples of the use of graphs. Do these examples follow the rules outlined in this chapter?

3.8 Use the **Employee data** file to answer the following problems with the aid of SPSS.

(a) I want to emphasize the high proportion of all cases that have clerical positions. Which graph should I generate and why? Generate this graph using SPSS, and add necessary titles and notes.

(b) Use a stacked bar graph to show the number of women and the number of men employed in each employment category. What does this indicate about the sexual division of labor in this company?

(c) Generate a histogram to display the distribution of scores for current salary. How would you describe this distribution in terms of skewness?

4

The tabular description of data

In the previous chapter we introduced the use of graphs as a means of displaying distributions. The power of graphs is their simplicity; the visual impact of a graph can sometimes convey a message better than the most advanced statistics. The simplicity of graphs can also be their weakness. We often do want to 'dig deeper' and extract more precise understandings of the data than can be gleaned from a chart. Obtaining a more detailed breakdown of a distribution usually begins with the construction of **frequency tables**.

Frequency (*f*) refers to the number of times a particular score appears in a set of data.

We will look at a variety of tables for presenting the frequency of scores in a distribution and the conclusions they allow us to reach about a variable. The tables we will cover are:

- listed data tables
- simple frequency tables
- relative frequency tables
- cumulative frequency tables
- tables with percentiles

Listed data tables

In Chapter 2 we worked with the results of a hypothetical survey of students for three separate variables: age, sex, and health level. A table such as the SPSS data table we created in Chapter 2 (Figure 2.20) is called a **listed data table**, since the score that each case has for each variable is *listed separately*. Such a table has as many rows as there are cases, and as many columns as there are variables for which observations have been taken. A listed data table, which presents the raw data for each case separately, allows us to calculate a variety of other descriptive statistics, which we will encounter later. This is why SPSS uses listed data as its format for data entry. However, listed data tables are not very informative as methods of *presenting* data, and, where we have a large number of cases, impractical. For example, with the hypothetical survey of 200 students space would prohibit the construction of a listed data table for these data.

Simple frequency tables

A more informative way of presenting the data is to construct a **simple frequency** table (or just **frequency table**), which presents the frequency distribution for a variable by tallying the number of times (*f*) each value of the variable appears in a distribution.

A simple frequency table reports, for each value of a variable, the number of cases that have that value.

A frequency table has in the first column the name of the variable displayed in the title row, followed by the categories or values of the variable down the subsequent rows. The second column then presents the frequencies for each category or value.

For example, the data we used to create graphs in Chapter 3 can alternatively be presented with a separate frequency table for each variable (Tables 4.1–4.3).

Table 4.1 Sex of students

Sex	Frequency (*f*)
Male	105
Female	93
Total	198

Table 4.2 Health rating of students

Health rating	Frequency (*f*)
Unhealthy	51
Healthy	56
Very healthy	71
Total	178

Table 4.3 Age of students

Age in years	Frequency (*f*)
17	6
18	28
19	34
20	41
21	30
22	25
23	12
24	9
25	8
Over 25	4
Total	197

Source: Hypothetical student survey
Notes: Totals do not equal 200 due to incomplete responses for individual items

These tables have the minimum structure that all frequency tables must display. They must:

- have a clear **title** indicating the variable and the cases for the distribution;
- have clearly **labelled categories** that are mutually exclusive;
- indicate the **total** number of cases;
- indicate the **source of data**, as in Table 4.3 (in most of the tables that follow in this book we will not follow this rule, since they are generally constructed from hypothetical data).
- indicate, where the total in the table is less than the number of survey respondents, why there is a difference, as in Table 4.3.

Notice also that in Table 4.1 we have placed males in the first row and females in the second. This may seem arbitrary given that, as this is a nominal scale, we can order the categories (the rows of the table) in any way we choose We have placed males first because it is commonplace with nominal variables to arrange the rows so that *the category with the highest frequency (what we will learn to call the mode) is the first row, the category with the second highest frequency is the second row, and so on.* The modal category is often of specific interest when analyzing the distribution of a nominal variable, and therefore it is convenient to present it first.

With Tables 4.2 and 4.3. however, the ordering of the categories is restricted by the fact that we are using ordinal and interval/ratio scales. For these levels of measurement, we generally *start with the lowest score in the distribution and then increase down the page.* Thus 17 is first row in Table 4.3, which is the lowest value for age, and then we gradually 'ascend' the scale as we move down the table row by row.

The other aspect to Table 4.3 that we should note is the use of the 'catch-all' category of Over 25. This is common with interval/ratio scales that have a long tail of values with only a small frequency of cases (usually less than five percent in total).

Example

The blood types of the 20 patients are recorded in the following listed data table (Table 4.4).

Table 4.4 Blood type of respondents

Case number	Blood type	Case number	Blood type
1	O	11	O
2	O	12	A
3	AB	13	B
4	A	14	O
5	A	15	O
6	O	16	A
7	A	17	O
8	AB	18	A
9	A	19	A
10	A	20	B

To describe these raw data in a simple frequency table we construct a table with the categories of the variable down the first column and the frequency with which each appears in the distribution down the second column (labelling each column appropriately) (Table 4.5). We then tally up the number of times each category appears in the distribution; we find that there are seven people with type O, two people with type AB, nine people with type A, and two people with type B.

Table 4.5 Blood type of respondents

Blood type	Frequency
A	9
O	7
B	2
AB	2
Total	20

Since blood type is a nominal variable, we have placed the category with the highest frequency (type A), which is called the **modal** category, in the first row.

Relative frequency tables: percentages and proportions

Some extra information can be calculated as part of a frequency table, if required. This is the **relative frequency distribution**.

Relative frequencies express the number of cases within each value of a variable as a percentage or proportion of the total number of cases.

In order to generate a relative frequency table, we need to acquaint ourselves with **percentages and proportions**.

According to Australia's census data (Australian Bureau of Statistics *Census of Population and Housing*, cat. no. 2720.0) in 1986 there were 324,167 one-parent families out of a total of 4,158,006 families. In 2001 there were 762,632 one-parent families out of a total of 4,936,828. What does this tell us about the changing nature of families? On the basis of this information we can say that there were more single-parent families in 2001 than in 1986. Absolute numbers, though, do not tell us much about the *relative* change in single-parent families. If, however, I said that such families accounted for 7.8 percent of all family types in 1986 and 15.4 percent in 2001, the pattern is immediately obvious: single-parent families have become a *relatively* larger group. By calculating these percentages we have in effect compensated for the different total number of families present in each year.

Percentages are statistics that standardize the total number of cases to a base value of 100.

The formula for calculating a percentage is:

$$\% = \frac{f}{n} \times 100$$

where f is the frequency of cases in a category, and n is the total number of cases in all categories.

We can see where the percentage figures came from in the example by putting ('substituting') the raw numbers into this formula:

$$1986: \frac{324,167}{4,158,006} \times 100 = 7.8\% \qquad 2001: \frac{762,632}{4,936,828} \times 100 = 15.4\%$$

It should be fairly clear that if I calculate the percentages for each family type in a given year and sum them, the total will be 100 percent. For example, if I add the percentage of single-parent families to the percentage of non-single-parent families in 2001, the total will be 100 percent; knowing that 15.4 percent of all families in 2001 were headed by a single parent allows me to calculate quickly the percentage of families *not* headed by a single parent:

$$100 - 15.4 = 84.6\%$$

Proportions are close cousins of percentages. A proportion (p) does exactly the same job as a percentage, except that it uses a base of 1 rather than 100. In fact, it is calculated in exactly the same way as a percentage, except for the fact that we do not multiply by 100:

$$p = \frac{f}{n}$$

The result is that we get a number expressed as a decimal. In the example above the results expressed as proportions are:

$$1986: \frac{324,167}{4,158,006} = 0.078 \qquad 2001: \frac{762,632}{4,936,828} = 0.154$$

Generally, percentages are easier to work with – for some reason people are more comfortable with whole numbers than with decimals. But in later chapters we will use proportions extensively, so it is important to learn the simple relationship between proportions and the more familiar percentages.

To convert a proportion into its corresponding percentage value, move the decimal point two places to the right (this is the same as multiplying by 100).

To convert a percentage into its corresponding proportion, move the decimal point two places to the left (this is the same as dividing by 100).

This may all seem straightforward. There are some words of caution that need to be borne in mind, though, when working with proportions and percentages, or when encountering them in other people's work. The first thing to look for when confronted with a percentage or proportion is the raw total from which they are calculated. This is because percentages and proportions are sometimes used to conceal dramatic differences in absolute size. An increase in the unemployment rate from 10 to 10.5 percent does not seem dramatic in statistical terms. But if this 0.5 percent represents 35,000 people it is, in socioeconomic terms, a large increase.

Conversely, a large change in percentage figures may be trivial when working with small absolute numbers. The number of people attending a pro-capital-punishment meeting may be 150 percent greater than the number that attended the last meeting, but if this is actually due to five people attending the recent meeting rather than the two who attended the previous one, it is hardly a dramatic rise. With small absolute numbers, small additions to either the total or the categories that make up the total will greatly affect the percentage figure calculated.

Now that we have familiarized ourselves with percentages and proportions we can use them to construct **relative frequency** tables for the data we introduced earlier. We can add to the table for each variable a column that expresses the percentage (or proportion) of cases that fall in each category. Table 4.6 shows the calculations involved in producing relative frequencies. Of course, when actually reporting results these calculations are not included, as in Table 4.7 and Table 4.8.

Table 4.6 Sex of students

Sex	Frequency	Percent (%)
Male	105	$\frac{105}{198} \times 100 = 53$
Female	93	$\frac{93}{198} \times 100 = 47$
Total	198	100

Table 4.7 Health rating of students

Health rating	Frequency	Percent (%)
Unhealthy	51	29
Healthy	56	31
Very healthy	71	40
Total	178	100

Table 4.8 Age of students

Age in years	Frequency	Percent (%)
17	6	3
18	28	14
19	34	17
20	41	21
21	30	15
22	25	13
23	12	6
24	9	5
25	8	4
Over 25	4	2
Total	197	100

Notice that the column of percentages must add up to 100 percent, since all cases must fall into one classification or another. Sometimes tables do not strictly follow this rule when numbers have been 'rounded off'. For example, for a particular table exact percentages to 1 decimal place may be 22.3%, 38.4%, and 39.3%. This may affect the readability of the table so the numbers are rounded off to the nearest whole number: 22%, 38%, 39%. These rounded numbers add up to only 99%. Where this occurs a footnote should be added to the table which states 'May not sum to 100 due to rounding', or words to that affect.

Cumulative frequency tables

With ordinal and interval/ratio data one further extension to the simple frequency table can be made. This is the addition of columns providing **cumulative frequencies** and **cumulative relative frequencies**. Since ordinal and interval/ratio data allow us to rank-order cases from lowest to highest, it is sometimes interesting to know the number, and/or percentage, of cases that fall *above or below a certain point on the scale*.

A **cumulative frequency table** shows, for each value in a distribution, the number of cases up to and including that value.

A **cumulative relative frequency table** shows, for each value in a distribution, the percentage or proportion of the total number of cases up to and including that value.

Sometimes all the absolute and relative frequencies and cumulative frequencies for a variable can be combined in the one table, as in Table 4.9, which shows the calculations for the first few rows of the table.

Table 4.9 Age (in years) of students

Age	Frequency	Cumulative frequency	Percent (%)	Cumulative percent %
17	6	6	3	3
18	28	28 + 6 = 34	14	$\frac{28+6}{197} = 17$
19	34	34 + 28 + 6 = 68	17	$\frac{34+28+6}{197} = 35$
20	41	41 + 34 + 28 + 6 = 109	21	$\frac{41+34+28+6}{197} = 55$
21	30	139	15	71
22	25	164	13	83
23	12	176	6	89
24	9	185	5	94
25	8	193	4	98
Over 25	4	197	2	100
Total	197		100	

With the distributions summarized in this way, I can now answer specific research questions that might be of interest. If I was interested in how many respondents are 19 years of age or younger I simply look at the sum of cases in the first three rows of Table 4.9. The cumulative frequency at this point is 68, which is 35% of all cases. Similarly, if I am interested in how many cases are over 19 years of age, I can see that since 35% are 19 or below, there must be 65% of cases (100 – 35 = 65%) above this age.

A common mistake in calculating cumulative percentages is to add the simple percentages for each row. The percentage for each row of the table contains a potential rounding error, so that adding these values up to get the cumulative percent may accumulate these rounding errors. For example, if we add the individual *percentages* for ages 17, 18, and 19 years we get a cumulative percent of 34%, rather than the correct figure of 35%, which is calculated directly from the *raw frequencies*.

Class intervals

One additional point needs to be made about working with interval/ratio data, as we have been with the age distribution of students in our example. With interval/ratio data we often use **class intervals** rather than individual values to construct a frequency distribution.

A **class interval** groups together a range of values for presentation and analysis.

The point of using class intervals is to collapse data into a few easy-to-work-with categories. But this increase in 'readability' comes at the cost of information, and therefore should not be undertaken if the data already come in a few, easily presented, values. In the example we have been working with, measuring age in whole years already provides a 'workable' number of values to organize the data into. It would not be useful to group these individual years into say 5 year class intervals, since this will only hide variation in the data that would otherwise help us answer our research question. *We only use class intervals if the range of values is so large that it makes presentation and analysis difficult.*

We will use the data represented in listed format in Table 4.10 for the income of 20 people to illustrate the usefulness of class intervals, and the general rules that apply to the construction of class intervals.

Table 4.10 Weekly income of 20 survey respondents: listed data

Case number	Income
1	$0
2	$0
3	$250
4	$300
5	$360
6	$375
7	$400
8	$400
9	$400
10	$420
11	$425
12	$450
13	$462
14	$470
15	$475
16	$502
17	$520
18	$560
19	$700
20	$1020

We can see that, even where we rank-order the cases from lowest to highest, a listed data table is not a useful summary of the data: a table with 20 rows of individual numbers does not get us far.

We can instead produce a simple frequency table by indicating the total number of cases that have each value of the variable contained in the data (Table 4.11). This frequency table has condensed the data slightly, but overall it has not simplified matters for us. We have so many individual values appearing in the distribution that when we use each one separately to group the cases, we still have a table with an impractical number of rows. To describe the data in a more meaningful way, we *cluster together ranges of values for people's income and indicate the total number of people that fall within each range* (Table 4.12).

Table 4.11 Weekly income of 20 survey respondents: simple frequency table

Weekly income	Frequency
$0	2
$250	1
$300	1
$360	1
$375	1
$400	1
$420	3
$425	1
$450	1
$462	1
$470	1
$475	1
$502	1
$520	1
$560	1
$700	1
$1020	1
Total	20

Table 4.12 Weekly income of 20 survey respondents: frequency table with class intervals

Weekly income	Frequency
$0	2
$1-100	0
$101-200	0
$201-300	1
$301-400	3
$401-500	9
$501-600	3
$601 or more	2
Total	20

We can see that the 'compact' version of the data distribution in Table 4.12 is easily interpreted. We can immediately observe the high frequency of cases within the $401–500 class interval. We can also see the spread of scores across the intervals.

Notice that in Table 4.12 the value of $0 is listed separately. It is common for readers of tables to be specifically interested in the number of cases that have a zero value for a particular variable. Such cases are often of special significance and therefore are usually separated from the rest of the distribution when constructing class intervals.

Notice also that we have not used the individual values that appear in the distribution to *label* each row. We have instead used **stated class limits**.

Stated class limits are the upper and lower bounds of a class interval that determine its width.

Generally, class intervals should have the same width, although at the lower and upper end of the data range we often have open-ended class intervals, such as the '$601 and more' interval in Table 4.12. The actual width of class intervals depends on the particular situation, especially the amount of information required. The wider the class intervals the easier it is to 'read' a distribution, but less information is communicated. For example, if we used class intervals that are $200 wide (i.e. $1–200, $201–400, etc.) a great deal of information will be lost. Cases that are very different in terms of the variable of interest, such as the person who earned $420 and the person who received $560 in weekly income, will now be considered to be the same. Generally, when collecting values into class intervals we lose information on the variation contained in the data, and *the wider the intervals the greater the loss of information.*

Conversely, if we have a very narrow width for the class intervals in a table we will be able to detect more variation in the data, but we will not simplify the data in a manageable and readable way. For example, if we used class intervals with a width of $50 for our income data (i.e. $1–50, $51–100, etc.) the number of rows in the table will not reduce down into the readable form we are after.

When constructing class intervals we need to ensure that the intervals are mutually exclusive. Thus in choosing $100 as the width of the class intervals in Table 4.12 the class intervals are $1–100, $101–200, $201–300, etc. Notice that the upper stated limit of each interval does not 'touch' the lower stated limit of the next interval: there appears to be a gap between 100 and 101, 200 and 201, 300 and 301, and so on. Won't some cases fall down this gap and not be included in any interval? *Provided that the unit with which the variable is measured is the same as that used to construct the class intervals, all cases will fall into one class or another.* We will be able to account for every case, in this example, because I have chosen to measure income in terms of whole dollars. Someone is in either the $1–100 group or the $101–200 group. A person cannot fall in between because of the units in which income is measured: no one can have an income of $100.63, simply because we have not measured income *at that* level of precision. If income is measured in a more precise unit, such as dollars and cents, the class intervals will then have to be expressed in dollars *and cents* as well to ensure the categories can 'capture' all possible scores.

Another concept that will be used when working with class intervals in later chapters is the **mid-point** (*m*) of the interval. The mid-point is the sum of the lower and upper limits divided by two:

$$\text{mid-point} = \frac{\text{lower limit} + \text{upper limit}}{2}$$

For example, the mid-point for the class interval $1–100 will be the sum of $1 and $100 divided by two:

$$m = \frac{1+100}{2} = \$50\ 50$$

Thus the frequency table for the data in Table 4.12, with stated limits and mid-points, is as shown in Table 4.13.

Table 4.13 Weekly income of 20 survey respondents

Weekly income	Mid-point	Frequency
$0	$0	2
$1-100	$50 50	0
$101–200	$150.50	0
$201–300	$250.50	1
$301–400	$350.50	3
$401–500	$450.50	9
$501–600	$550.50	3
$601 or more	$650.50	2
Total		20

The reason for laboring through this process of calculating limits and mid-points for tables using class intervals may not be immediately obvious. However, it does affect the types of calculations we might want to generate on the basis of such tables, as we will see when we come to Chapter 9. Some familiarity with their construction now will help us later.

Example

A drug is administered to a sample of 50 patients and the time elapsed (in seconds) before the drug has an effect is recorded for each patient. These times are:

78, 37, 99, 66, 90, 79, 80, 89, 68, 57, 71, 78, 53, 81, 77, 58, 93, 79, 98, 76, 60, 77, 49, 92, 83, 80, 74, 69, 90, 62, 84, 74, 73, 48, 75, 98, 32, 75, 84, 87, 65, 59, 63, 86, 95, 55, 70, 62, 85, 72

To construct class intervals for these data I have to define my intervals in the same unit of measurement as the raw data, which in this case is whole seconds. I also have to select interval widths that are neither too wide (which will conceal variation we are interested in) nor too small (which will not adequately condense the data into manageable groupings). This often takes a little trial and error; here I will choose ten second intervals, which, as you will hopefully agree after inspecting Table 4.14, provide an appropriate summary of the data.

Table 4.14 Drug response times

Time intervals (seconds)	Frequency
30–39	2
40–49	2
50–59	5
60–69	8
70–79	15
80–89	10
90–99	8
Total	50

The concentration of scores within a narrow range of times is now clearly evident, as well as the spread of scores around this range.

Percentiles

Another common way of grouping interval/ratio data into manageable and readable clusters is with the construction of **percentiles**. Instead of using the values of the variable to group cases, we use a particular **percentage of cases** to construct a table. The set of cases is rank-ordered, and 'split' into the number of groups of equal size defined by the chosen percentage. For example, **deciles** divide the cases into ten equal groups each containing 10 percent of cases; **quartiles** use four groups each containing 25 percent of cases, and **quintiles**, used in Table 4.15 to display the distribution of income among Australian households in 2002–03, into five groups each containing 20 percent of cases (ABS catalog no. 6523.0). This table rank-orders all households in terms of income, from the poorest to the richest, and then splits them into 5 equally sized quintiles. The first quintile comprises the 20 percent of households that are the poorest, the second decile comprises the next 20 percent of households, through to the fifth quintile, which comprises the richest 20 percent of households.

Table 4.15 Equivalized disposable household income, Australia 2002–03

Quintile	Share of disposable income %
Lowest	7.7
Second	12.8
Third	17.6
Fourth	23.7
Highest	38.3

The share of income held by each quintile gives a sense of income distribution at this point in time: income is not equally spread across households (according to this measure).

Frequency tables using SPSS

SPSS can generate the same tables that we created 'by hand' above (Table 4.16, Figure 4.1).

Table 4.16 The **Frequencies** command on SPSS (file: **Ch04.sav**)

SPSS command/action	Comments
1 From the menu select **Analyze/Descriptive Statistics/Frequencies**	This will bring up a dialog box headed **Frequencies**. This will contain an area with a list of the variables for which data have been entered
2 Select the variable(s) to generate a frequency table for by clicking on their name(s)	A number of frequency tables can be generated simultaneously by pasting more than one variable into the **Variable(s):** box. Here we want all three variables, so we will paste all of them
3 Click on ▸	This will paste the selected variable(s) into the area below **Variable(s):**, which is the list of variables for which a frequency table will be generated
4 Click on **OK**	

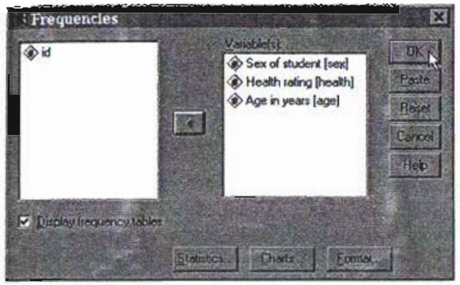


Figure 4.1 The **Frequencies** dialog box

Notice the appearance of the dialog box in Figure 4.1. It has some features in common with most of the dialog boxes we will encounter in later chapters so we will take a moment to note these.

- On the left of the box is an area with a list of the variables created in the **Data Editor**. This is called the **source variable list** that provides the list of variables we can analyze using the particular command we have chosen from the menu (in this instance we are using the **Analyze/Descriptive Statistics/Frequencies** command).
- On the right is another area which is initially blank, but which eventually contains the variable(s) we have actually chosen to analyze. This is called the **target variable list**.
- Variables can be moved back and forth from one list to the other, as we have done here, by clicking on them and then clicking on the ▸ button between the two lists.
- Many of the dialog boxes we will encounter have **default** settings. These are options that are preselected by SPSS; they will automatically be used when the **OK** button is clicked. For example, in the **Frequencies** dialog box you will notice a tick mark, ✓, in the tick-box next to **Display frequency tables**. This indicates that a frequency table will automatically be generated for each of the variables pasted into the target variable list. SPSS does not have to be specifically asked to generate the tables. If we did not want a frequency table to be generated for each of the target variables, we would click on this box to remove the tick mark. If at any point you want to return to the default setting for any given dialog box so you can begin a procedure from scratch, click the **Reset** button.
- There are a number of buttons available providing options that add to or refine the basic settings. Here we can also generate **Statistics** and **Charts**, and include **Format** options.

The instructions in Table 4.16 produce a table for each variable with raw, relative and cumulative frequencies (Figure 4.2).

Frequencies

Statistics

		Sex of student	Health rating	Age in years
N	Valid	198	178	197
	Missing	2	22	3

Frequency Table

Sex of student

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Female	93	46.5	47.0	47.0
	Male	105	52.5	53.0	100.0
	Total	198	99.0	100.0	
Missing	Did not answer	2	1.0		
Total		200	100.0		

Health rating

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Unhealthy	51	25.5	28.7	28.7
	Healthy	56	28.0	31.5	60.1
	Very healthy	71	35.5	39.9	100.0
	Total	178	89.0	100.0	
Missing	Don't know	18	9.0		
	Did not answer	4	2.0		
	Total	22	11.0		
Total		200	100.0		

Age in years

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	17	6	3.0	3.0	3.0
	18	28	14.0	14.2	17.3
	19	34	17.0	17.3	34.5
	20	41	20.5	20.8	55.3
	21	30	15.0	15.2	70.6
	22	25	12.5	12.7	83.2
	23	12	6.0	6.1	89.3
	24	9	4.5	4.6	93.9
	25	8	4.0	4.1	98.0
	26	2	1.0	1.0	99.0
	32	1	.5	.5	99.5
	39	1	.5	.5	100.0
	Total	197	98.5	100.0	
Missing	Did not answer	3	1.5		
Total		200	100.0		

Figure 4.2 SPSS Frequencies output

We can immediately compare these tables with the ones we generated 'by hand' above to confirm that all the figures are the same. The usefulness of the value labels that we specified in Chapter 2 should now be obvious. If we had not specified that 1=female and 2=male, for example, then the first table would not have these value labels printed along the left. Thus we might be left scratching our heads or hunting back through our notes to remember which category the value 1 represented and which category 2 represented. Here we have all the information printed with the output.

There are two limitations to SPSS tables to which attention needs to be drawn.

- If a category in the distribution has a zero frequency, even though it has been given a value label, it will not appear in a frequency table. For example, if there were no students in the survey who rated themselves as Healthy, SPSS will omit this category from the table, rather than print it with 0 in the frequency column.
- Cumulative frequencies are generated with a table even when they are not appropriate. Cumulative frequencies are not appropriate where we have a nominal scale, since the ordering of the categories is not fixed. Since the points on a nominal scale are not ordered it makes no sense to talk of the number or percentage of cases up to a certain point on the scale. Cumulative frequencies are also not appropriate where there are only two categories, since the simple frequencies and cumulative frequencies will be the same.

Valid cases and missing values

If you look closely at each table in the SPSS output you will see that there are columns headed **Percent** and another headed **Valid Percent**. The reason for printing these two columns in the frequency tables arises because data sometime include cases for which a variable has not been adequately measured. These are called, as we discussed in Chapter 2, missing cases, and the presence of missing cases will cause the values in the **Percent** and **Valid Percent** columns to diverge. The number of valid cases is equal to the total number of cases minus the number missing:

$$\text{valid cases} = \text{total cases} - \text{missing cases}$$

For example, we can see from the SPSS output that 2 students did not answer the question asking for their respective sex. In setting up the SPSS file, you will recall, these responses were coded as 3 for purposes of data entry, and the label 'Did not answer' attached to this code value, which was then defined as a missing value. The result is that the **Percent** column provides the percentage of cases in each category, including that for the missing value, *as a percentage of all 200 cases*. The **Valid Percent** column, on the other hand, calculates the percentage of cases in each category, excluding the missing values, based on *only the total number of valid cases* (the summary **Statistics** table at the top of the SPSS output summarizes, for each variable the number of total valid and total missing cases).

Improving the look of tables

You may regard, as I do, the format of basic SPSS tables not to be of 'report quality'. They obviously need some tidying up, such as the removal of unnecessary decimal places and changes to the layout of text and data. One option is to create a blank table using a word processor, into which we enter the results from the SPSS output. An alternative is to use the SPSS formatting options to improve the look of tables and then export them into a report. To explain how to do this would take us beyond the immediate needs of this text, but for those interested, a guide to formatting SPSS tables is included on the accompanying CD.

Exercises

4.1 How does a proportion differ from a percentage?

4.2 Why will a proportion always be smaller than its equivalent percentage value?

4.3 Convert the following proportions into percentages:

(a) 0.01 (b) 0.13 (c) 1.24 (d) 0.0045

4.4 Convert the following percentages into proportions:

(a) 12% (b) 14.4% (c) 167% (d) 4.5%

4.5 The following data represent time, in minutes, taken for subjects in a fitness trial to complete a certain exercise task.

31	39	45	26	23	56	45	80	35	37
25	42	32	58	80	71	19	16	56	21
34	36	10	38	12	48	38	37	39	42
27	39	17	31	56	28	40	82	27	37

The heart rate for each subject is also recorded in the same sequence as their respective time scores:

63	89	75	80	74	65	90	85	92	84
74	79	98	91	87	76	82	90	93	77
74	89	85	91	102	69	87	96	83	72
92	88	85	68	78	73	86	85	92	90

(a) Using the class intervals 1–9, 10–19, 20–29, and so on, organize the data for each of these variables into frequency tables, displaying both raw and cumulative frequencies and percentages. What are the mid-points of these class intervals?

(b) Open the file you created as part of Exercise 2.2 to store these data. Using the SPSS **Recode** command, generate a frequency table with these class intervals.

4.6 The following data indicate attendance at selected cultural venues across eight regions:

Region	People attending public libraries	People attending popular music concerts
A	1409	1166
B	1142	870
C	713	604
D	423	280
E	497	332
F	130	99
G	90	32
H	38	74
Total	4442	3456

For each of these variables calculate the relative frequencies for each region.

4.7 From a recent newspaper or magazine find examples that use the techniques outlined in this chapter. Do these examples follow the rules of description outlined here?

4.8 In Exercise 2.1 you created an SPSS file to store the data for the example we used in the text for the weekly income of 20 survey respondents:

$0, $0, $250, $300, $360, $375, $400, $400, $400, $420, $425, $450, $462, $470, $475, $502, $520, $560, $700, $1020

(a) Open this file in SPSS and generate a simple frequency table.

(b) Using the class intervals we employed in the text above, generate a frequency table.

4.9 In Exercise 2.3 you created an SPSS file for the following data:

Television watched per night (in minutes)	Main channel watched	Satisfaction with quality of programs
170	Commercial	Very satisfied
140	Public/government	Satisfied
280	Public/government	Satisfied
65	Commercial	Very satisfied
180	Commercial	Not satisfied
60	Commercial	Not satisfied
150	Public/government	Satisfied
160	Commercial	Not satisfied
200	Public/government	Satisfied
120	Commercial	Not satisfied

(a) Generate a frequency table for each of these variables.

(b) Recode minutes of TV watched into categories of less than 100 minutes, and 100 minutes or more. Generate a frequency table for this new variable. What is its level of measurement?

4.10 Using the **Employee data** file on the CD that comes with this book, generate frequency tables that will allow you to determine:

(a) The number of employees that are from minority groups.

(b) The percentage of employees that are from a minority group.

(c) The percentage of employees with 15 years of education or less.

(d) The percentage of employees whose starting salary was greater than $17,100.

4.11 Using the **Employee data** file collapse the beginning salary data into appropriate class intervals using the **Recode** command. Justify your choice of interval width, and determine the class mid-points.

5

Using tables to investigate the relationship between variables: Crosstabulations

The previous chapter discusses the way in which frequency tables can be used to describe the distribution of a *single* variable. This chapter extends the use of frequency tables to situations where we are interested in whether *two* variables are related. This is similar to the way we extended the use of graphs from the univariate context to the bivariate context in Chapter 3, thus allowing us to investigate whether a relationship exists between two variables.

Crosstabulations as descriptive statistics

We began Chapter 1 with the following research questions:

> 'What is the health status of the students in my statistics class?'
>
> 'Is there a relationship between the health status of the students in my statistics class and their sex?'
>
> 'Is any relationship between the health status and the sex of students in my statistics class affected by the age of the students?'

A simple frequency table allows us to describe the distribution of individual variables so that we can address questions such as the first of those listed above. Having dealt with this simple univariate question we are now ready to tackle the more complex bivariate problem presented in the second question. It may help at this point to return to the discussion of bivariate analysis in Chapter 1 before proceeding, where we emphasized the need to distinguish between the independent and dependent variables in the relationship.

We suspect that there is a relationship between health status and sex of students, and if there is such a relationship, it must be a one-way relationship running from the sex of the students to health status (it is not possible for the reverse to hold and the sex of students to be dependent on their respective health status). In other words, in our model, sex of students is the independent variable and health status is the dependent variable. Remember though that this is only a supposition for what we expect to find. The 'real world', or at least the data we gather from it, may not agree with this expectation. The two variables may not be related; instead they may be independent of each other. How can we organize the data we collect to inform us whether our model is correct or whether the two variables are in fact independent?

We introduced data for a hypothetical survey of students that included measurements for these two variables. Each student has a value assigned to them indicating their sex and another value indicating their health status. How can we organize these numbers in such a way as to reveal any relationship that may exist between the sex of a student and health status?

We could use the univariate methods we learnt in the previous chapter to construct separate frequency distributions for each variable (Tables 5.1 and 5.2). It is clear that these *separate* univariate frequency tables do not help us much. It is impossible to assess whether there is a relationship between the two variables, which is the aim of our research question.

Table 5.1 Frequency distribution for sex of students

Sex	Frequency
Male	105
Female	93
Total	198

Table 5.2 Frequency distribution for health status of students

Health rating	Frequency
Unhealthy	51
Healthy	56
Very healthy	71
Total	178

To capture any possible relationship that may exist between variables measured with scales that have only a few points we use a **bivariate table**, which is also known as a **contingency table** or **crosstabulation** (or 'crosstab' for short).

A bivariate table displays the joint frequency distribution for two variables.

The crosstabulation for the data we have (hypothetically) collected is presented in Table 5.3.

Table 5.3 Health rating by sex of students

Health rating	Sex		
	Female	Male	Total
Unhealthy	34	16	50
Healthy	29	27	56
Very healthy	17	54	71
Total	80	97	177

Source: Hypothetical student survey.
Note: 23 students did not provide responses for either or both variables

A crosstab shows the **joint frequency distribution** for two variables, since we can 'read off' the score any given case has for each of the variables simultaneously. Looking at Table 5.3, for example, we can see that there are 34 students who are both female *and* rate themselves as unhealthy. Since bivariate tables describe data in a way that reveals this joint distribution, it allows us to investigate whether two variables are related.

There are certain rules we follow in the construction of a bivariate table:

- *Give the table an appropriate title.* A crosstab should always have a title with clear labelling for both variables and the cases described by the table.
- *Clearly label the rows and columns with the variables described and the categories that make up each scale.*
- *Indicate the source of data.* This is usually done in the text immediately before or after the table, or as a footnote attached to the table (as in the example shown in Table 5.3).
- *Note any excluded data.* As with the source of data this can be done in the text immediately before or after the table, or as a footnote attached to the table.
- *Place the appropriate variables in the rows and columns.* If there is reason to believe that the two variables are not only related to each other, but that one of the variables is dependent on the other (a one-way relationship), the *independent variable should be arranged across the columns* and the *dependent variable down the rows*. In this example we have specified that sex is the independent (column) variable and health status is the dependent (row) variable.

- *For scales that can be ranked, ensure the scale increases down the rows/across the columns*. Notice that one of the variables, 'Health rating', is ordinal. Thus the categories that make up this scale can be ordered from lowest to highest. We therefore place the lowest point on the scale, the 'Unhealthy' category on the first row so that the scale increases down the page until we reach the highest point on the scale, which is the 'Very healthy' category.

In discussing the use of crosstabs as a means of describing data, we need to become familiar with some terminology:

- *The size and dimensions of the table*. The size of the table is defined as the number of categories for the row variable times the number of categories for the column variable. In this example there are three categories for health status and two categories for sex, producing a 3-by-2 table. If health status was measured on a four-point scale, on the other hand, the dimensions of the table will be 4-by-2.
- *The cells of the table*. Each square in the table that contains the number of cases that have a particular combination of values for the two variables is called a table cell.
- *The marginals of the table*. The entries in the Total column are called **column marginals** and the entries in the Total row are called **row marginals**. These provide the frequencies for the categories of each variable, much like the simple frequencies in Table 5.1 and 5.2.

Types of data suitable for crosstabulations

We use crosstabs to describe the relationship between two variables whose variation is expressed in only a few categories. Thus the most straightforward instance for using a crosstab is where both variables are measured on scales that respectively have only a small number of points. That is, regardless of the level of measurement, we use crosstabs if the data do not range over too many scores. As a rule of thumb, if each variable is measured on a scale with five or less points, the data will directly 'fit' into a crosstabulation. The data we used above for the relationship between sex and health rating is an example: sex has only two points in its scale of measurement (male and female), while health rating only has three (unhealthy, healthy, and very healthy).

The raw data we work with, however, do not always come neatly packaged into a small number of scores. A slightly more complicated situation we sometimes encounter is where one variable has only a few points of variation in the data, but the other has many points. Imagine, for example, that instead of the simple scale we used above we measured health status by asking students to rate themselves on a 10-point scale that ranges from 'Very unhealthy' at one extreme and 'Very healthy' at the other. If we try to present the raw data in a crosstab with the sex of students, the table will have ten rows (excluding the Total row) rather than the three that exist in Table 5.3. Such a table would be too big to easily interpret. Where one variable has only a few points of variation but the other has many, we cannot fit the original scale into a crosstab; we need to aggregate scores into broader groups (using the SPSS Recode command detailed in the supplementary chapter on the CD supplied with this text). Thus if I had a 10-point scale for health status and want to express its relationship with students' sex in a crosstab, I will collapse some of the scores together so I end up with fewer categories. Similarly, if I wish to see if there is a relationship between health rating and a student's age, I would need to group students together into broader age groups.

The only exception to this is where I have two variables that are *both* measured on interval/ratio scales and the data contain many values. I could conceivably collapse the values for each variable into broader groups, but a better option is to work with the original scales and use a scatterplot rather than a crosstab to display any possible relationship between the two variables (and calculate the measures of correlation specific to such data). We will explore this method of data description in Chapter 12.

Crosstabulations with relative frequencies

We constructed the crosstab to see if a student's health status was dependent on their sex. Looking at Table 5.3 we see that males do in fact tend to rate themselves as healthier than females, lending support to our theoretical model of the relationship. The problem with relying just on this table, though, is that the total number of females and total number of males are not equal. Thus there are more males rating themselves as 'Very healthy', partly because they do so at a higher rate than females, but also partly because there are simply more males in the sample. We can compensate for this and improve our ability to draw out any possible relationship contained in the data by calculating the **relative frequencies**, rather than just the absolute number of cases in each cell. The relative frequencies based on **column totals** (with calculations for the females) are as given in Table 5.4.

Table 5.4 Health rating by sex of students: Column percentages

Health rating	Sex		
	Female	Male	Total
Unhealthy	$\frac{34}{80}\times 100 = 43\%$	16%	28%
Healthy	$\frac{29}{80}\times 100 = 36\%$	28%	32%
Very healthy	$\frac{17}{80}\times 100 = 21\%$	56%	40%
Total	100% (80)	100% (97)	100% (177)

Thus 43 percent *of the total number of females* rate themselves as 'Unhealthy'. Note that in the cells of the table we only included the percentages. We therefore have also included in the marginals the total number of males and females from which these percentages are calculated.

The crosstab can, alternatively, provide the relative frequencies in terms of the **row totals**, as shown in Table 5.5, which includes the calculations for the 'Unhealthy' group. We can immediately see from this that 68 percent *of the total number of students rating themselves as unhealthy* are female.

Table 5.5 Health rating by sex of students: Row percentages

Health rating	Sex		
	Female	Male	Total
Unhealthy	$\frac{34}{50}\times 100 = 68\%$	$\frac{16}{50}\times 100 = 32\%$	100% (50)
Healthy	52%	48%	100% (56)
Very healthy	24%	76%	100% (71)
Total	45%	55%	100% (177)

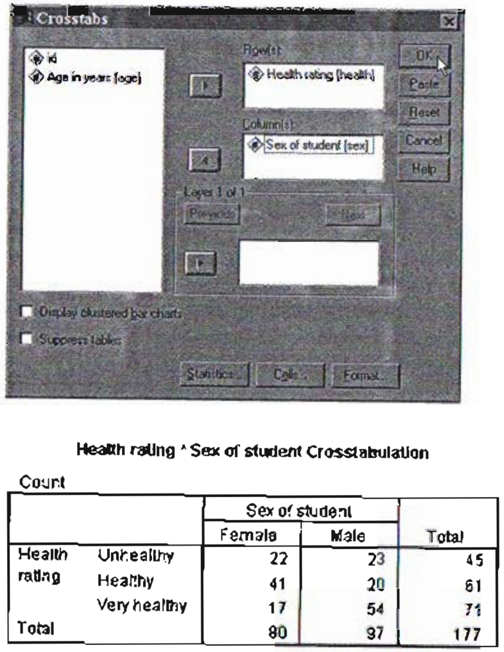
Sometimes we can combine in one table the raw data and the relevant percentages by adding extra columns or rows. The appropriate structure depends on the context in which the data are being used and the intended audience. As a general rule, where we have the independent variable across the columns *we are usually interested in generating the column percentages.* In bivariate analysis, we compare the groups formed by the independent variable (in this instance males and females), so the relevant percentages to calculate are based on the total number of cases in each of these groups, which should be arranged across the columns.

Crosstabulations using SPSS

The data from the previous example have been entered in SPSS, so we can see how to generate crosstabs (Table 5.6, Figure 5.1).

Table 5.6 Generating crosstabs on SPSS (file: Ch05.sav)

SPSS command/action	Comments
1 From the menu select: **Analyze/Descriptive Statistics/ Crosstabs**	This brings up the Crosstabs dialog box
2 Click on the variable in the source list that will form the rows of the table, which in this case is **Health rating**	This highlights **Health rating**
3 Click on ▸ that points to the target list headed **Row(s):**	This pastes **Health rating** into the **Row(s):** target list
4 Click on the variable in the source list that will form the columns of the table, which in this case is **Sex of student**	This highlights **Sex of student**
5 Click on ▸ that points to the target list headed **Column(s):**	This pastes **Sex of student** into the **Column(s):** target list
6 Click on **OK**	



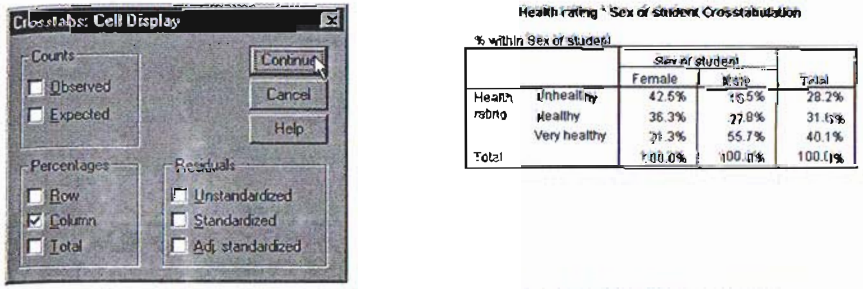
Health rating * Sex of student Crosstabulation

Count

		Sex of student		
		Female	Male	Total
Health rating	Unhealthy	22	23	45
	Healthy	41	20	61
	Very healthy	17	54	71
Total		80	97	177

Figure 5.1 SPSS Crosstabs dialog box and output

The crosstabs command can be extended to provide relative as well as absolute frequencies. This option is selected by clicking on the **Cells** button on the **Crosstabs** window. This will bring up another dialog box headed **Crosstabs: Cell Display** (Figure 5.2). This window provides the options for deciding how much information each cell will contain. The default setting, which we just used, is for the cells to contain the raw count only. If we want the row percentages in addition to the raw count we click on the small square next to **Row**. This will place a ✓ in the check-box to show that it is selected. Similarly, if we want column percentages we click on the check-box next to **Column**. Figure 5.2 also illustrates the output that results if we select column percentages only.



Health rating * Sex of student Crosstabulation

% within Sex of student

		Sex of student		
		Female	Male	Total
Health rating	Unhealthy	42.5%	16.5%	28.2%
	Healthy	36.3%	27.8%	31.6%
	Very healthy	21.3%	55.7%	40.1%
Total		100.0%	100.0%	100.0%

Figure 5.2 The Cell Display dialog box and SPSS crosstab output with only column percentages

Interpreting a crosstabulation: The pattern and strength of a relationship

We have introduced the construction of a very important descriptive tool in research: a crosstabulation. Its importance rests on the fact that so much data collected in research are data that only have a small number of categories or values. Having transformed a set of raw data into a crosstab, the task is then to interpret it – to assess whether it reveals that a relationship exists between the two variables. When interpreting a relationship evident in a crosstab we generally look for two features:

- pattern
- strength

These aspects of a relationship are clearer if we highlight the modal cell for each column (Table 5.7).

Table 5.7 Health rating by sex of students

Health rating	Sex		
	Female	Male	Total
Unhealthy	43%	16%	28%
Healthy	36%	28%	32%
Very healthy	21%	56%	40%
Total	100% (80)	100% (97)	100% (177)

By highlighting the modal cell for each column we can see that there is a relationship. Looking at the relative frequencies it is evident that nearly half of all females sampled rate themselves as unhealthy, whereas over half of all males rate themselves as very healthy. Thus we can interpret the table as suggesting a relationship exists between sex and health status, and the **pattern** of this relationship is that males perceive themselves healthier than females.

We can also assess the **strength** of this relationship by looking at the proportion of cases in each column 'captured' by the modal cell in each column. We can see that while the modal category for females is unhealthy, more than half fall into the other two categories. Similarly, while the majority of males rate themselves as very healthy, a large percentage (44%) of them do not. This result indicates to us that the relationship between health rating and sex of students is not very strong.

Interpreting a crosstabulation when both variables are at least ordinal

The previous sections discussed the construction of a crosstab, and how we go about interpreting any relationship revealed by such a crosstab. We looked at an example where we had one nominal variable (sex of students) and one ordinal variable (health status). The rules and procedures we learnt in this instance apply generally to the construction of a crosstab with variables measured at any combination of levels.

When *both* variables are measured at least at the ordinal level, however, the interpretation of the pattern of a relationship found in a crosstab can be taken one step further to incorporate a discussion of the **direction** and the **consistency** of the relationship.

Direction of the relationship

Notice that in the previous discussion of the relationship between the sex of students and their health rating we concluded that health is related to a student's sex such that females tend to rate their health lower than males. Because we are working with at least one variable that is measured on a nominal scale (sex of students) we can't talk about an increase or decrease in health being associated with an increase or decrease in a student's sex. It makes no sense to talk about students' sex increasing or decreasing.

When both variables are measured at least at the ordinal level, however, we can talk about the relationship having either a **positive** or **negative** direction.

A positive relationship exists where movement along the scale of one variable in one direction is associated with a movement in the same direction along the scale of the other variable.

A negative relationship exists where movement along the scale of one variable is associated with a movement in the opposite direction along the scale of the other variable.

For example, we might be interested in the relationship between income and the amount of TV someone watches. The amount of TV a person watches is measured by asking each person whether they watch TV 'never', 'some nights', or 'most nights'. Income is measured by grouping people according to whether they are low, medium, or high income earners. With *both* variables now measured at an ordinal level, if we do find that a relationship does indeed exist, we can talk about the direction of the relationship.

Assume that we have gathered data from 300 people measuring their respective incomes and the amount of TV they watch. We suspect that if there is a pattern of dependence between these two variables it will run from income to TV watching. Thus we will arrange the table with income (independent variable) across the columns and TV watching (dependent variable) down the rows.

All the rules we discussed earlier with respect to the construction of a crosstab still apply. But with both variables measured on an ordinal scale there is one important additional rule.

When crosstabulating two ordinal-level variables arrange the table so that the values of the independent variable increase across the page from left to right, and the values of the dependent variable increase down the page.

The application of this rule is illustrated in Table 5.8.

Table 5.8 Frequency of TV watching by income

TV watching	Income			
	Low	Medium	High	Total
Never	75 71%	15 15%	10 11%	100
Some nights	20 19%	70 70%	10 11%	100
Most nights	10 10%	15 15%	75 79%	100
Total	105	100	95	300

To help with the interpretation of the table, as in our earlier example, we highlight the modal cell for each column. We can immediately see that there is a relationship between these two variables, in that as income increases so too does the amount of TV watched. Thus we have a **positive relationship**. If the modal cells were all lined up along the other diagonal, from High/Never to Low/Most nights, the table will describe a **negative relationship**.

Consistency of the relationship

In addition to discussing the direction of the relationship, when working with two ordinal variables, we can also look at whether the relationship is **consistent**. Notice that *all* the modal cells in Table 5.8 are arranged along the positive diagonal, so that there is smooth progression in the relationship across the whole range of values. Such a pattern of dependence is called a consistent relationship. If on the other hand we observe the results contained in Table 5.9, we will still conclude that there is a relationship between the two variables, but we describe it as a **non-consistent relationship**.

Table 5.9 Frequency of TV watching by income: a non-consistent relationship

TV watching	Income			
	Low	Medium	High	Total
Never	75 71%	11 12%	15 15%	100
Some nights	20 19%	9 9%	70 70%	100
Most nights	10 10%	75 79%	15 15%	100
Total	105	95	100	300

We can see that at the low end of the income scale, as income rises TV watching also increases, but that the relationship reverses as we move further up the income scale.

Example

Research is conducted to see whether the English proficiency of migrants from non-English-speaking backgrounds improves over time. English proficiency is rated by a standard verbal assessment test as 'very poor', 'poor', 'average', or 'above average'. Length of time since migration is measured by classifying migrants as being resident for 'less than 1 year', 'between 1 and 2 years', '2–5 years', or 'over 5 years'. In total, 690 migrants of non-English-speaking background are surveyed

The raw data from this research are the 1380 numbers indicating for each person their English proficiency and their length of time since migration. These raw data are described in the contingency Table 5.10, which provides the relative frequencies as column percentages and also highlights the modal cell for each column.

The first point to note is the construction of Table 5.10. It is clear that if these two variables are related, the appropriate model for this relationship will be one-way dependence with time since migration as the independent variable and English proficiency as the dependent variable. There is no sense in which we could argue that the reverse is true: it is not reasonable to argue that English proficiency somehow determines how long someone has lived in a country. Thus we have placed time since migration along the columns, and English proficiency down the rows.

Table 5.10 English proficiency by time since migration

English proficiency	Time since migration				
	1 year or less	1–2 years	2–5 years	Over 5 years	Total
Very poor	70%	35%	5%	4%	184
Poor	20%	50%	10%	9%	154
Average	8%	11%	80%	81%	322
Above average	2%	4%	5%	6%	30
Total	150	180	160	200	690

The other aspect of the table's construction worth noting is that the quantity of each of these variables increases as we move across the columns or down the rows. We have two ordinal-level variables, so that we need to ensure the values of the variables move in the appropriate direction. That is, people with the least time since migration are in the first column, and time increases across the page. Similarly, the people with the lowest English proficiency are in the first row, and the strength of this variable increases as we move down the page.

The relationship can now be assessed in terms of its pattern and its strength. We can see that there is a general improvement in English proficiency reflecting a positive association between the two variables. The relationship is not perfectly consistent, as the effect of time since migration begins to peter out after 5 years of residency, and migrants' English skills reach the average level of the rest of the population. *After a point* there is clearly no association between these two variables.

In terms of the strength of the relationship we could argue that it is quite strong. For each column the modal cell seems to capture a very large proportion of cases in that column, indicating that for a majority of cases the pattern of association we have noted seems to hold.

Summary

We have investigated extensively the construction and interpretation of bivariate tables. We have seen that these tables are a useful way of describing categorical data in such a way as to reveal whether a relationship exists between two variables under investigation. We discussed the specific rules and procedures for transforming a collection of raw data into a compact crosstab, and the means for interpreting any relationship that a crosstab may reveal. With all tables we saw that this involved an assessment of the pattern and strength of the relationship. We have also seen that where both variables are measured at least at the ordinal level some additional aspects to a relationship can be gleaned from a crosstab, namely the direction of the relationship, and whether it is consistent.

Exercises

5.1 A study finds that the number of injured people at an accident is related to the number of ambulance officers attending the accident. Should ambulance officers stay away from accidents in order to reduce the injury rate?

5.2 For each of the following tables, calculate the column percentages.

(a)

Dependent	Independent		
	1	2	Total
1	30	60	90
2	45	50	95
Total	75	110	185

(b)

Dependent	Independent			
	1	2	3	Total
1	56	40	10	106
2	15	30	50	95
Total	71	70	60	201

5.3 For each of the following tables, calculate the row percentages.

(a)

Dependent	Independent		
	1	2	Total
1	30	60	90
2	45	50	95
Total	75	110	185

(b)

Dependent	Independent			
	1	2	3	Total
1	56	40	10	106
2	15	30	50	95
Total	71	70	60	201

5.4 Stratified samples of 30 people who voted for the Conservative Party at the last election and 30 people who voted for the Progressive Party at the last election are drawn to assess whether political preference is related to father's political preference:

Case	Voting preference	Father's voting preference	Case	Voting preference	Father's voting preference
1	Progressive	Progressive	31	Conservative	Conservative
2	Progressive	Progressive	32	Conservative	Other
3	Progressive	Progressive	33	Conservative	Conservative
4	Progressive	Conservative	34	Conservative	Conservative
5	Progressive	Progressive	35	Conservative	Conservative
6	Progressive	Progressive	36	Conservative	Progressive
7	Progressive	Progressive	37	Conservative	Conservative
8	Progressive	Progressive	38	Conservative	Conservative
9	Progressive	Conservative	39	Conservative	Progressive
10	Progressive	Conservative	40	Conservative	Other
11	Progressive	Progressive	41	Conservative	Conservative
12	Progressive	Progressive	42	Conservative	Conservative
13	Progressive	Other	43	Conservative	Conservative
14	Progressive	Progressive	44	Conservative	Conservative
15	Progressive	Progressive	45	Conservative	Conservative
16	Progressive	Progressive	46	Conservative	Other
17	Progressive	Other	47	Conservative	Conservative
18	Progressive	Progressive	48	Conservative	Other
19	Progressive	Progressive	49	Conservative	Progressive
20	Progressive	Progressive	50	Conservative	Conservative
21	Progressive	Progressive	51	Conservative	Conservative
22	Progressive	Progressive	52	Conservative	Conservative
23	Progressive	Other	53	Conservative	Progressive
24	Progressive	Other	54	Conservative	Progressive
25	Progressive	Progressive	55	Conservative	Conservative
26	Progressive	Progressive	56	Conservative	Conservative
27	Progressive	Conservative	57	Conservative	Other
28	Progressive	Progressive	58	Conservative	Conservative
29	Progressive	Progressive	59	Conservative	Conservative
30	Progressive	Progressive	60	Conservative	Other

(a) Which of these variables would you consider to be independent and which dependent? What are their respective levels of measurement?
(b) Construct a bivariate table to describe this result, either by hand or on SPSS.
(c) Looking at these raw figures, do you suspect a dependence between these variables? If so, how would you describe it in plain English?

5.5 Hypothetical samples of children from Australia, Canada, Singapore, and Britain are compared, in terms of the amount of TV they watch:

Amount of TV	Country				
	Canada	Australia	Britain	Singapore	Total
Low	23	25	28	28	104
Medium	32	34	39	33	138
High	28	30	40	35	133
Total	83	89	107	96	375

Can we say that the amount of TV watched is independent of country of residence?

5.6 A sample of 162 men between the ages of 40 and 65 years is taken and the state of health of each man is recorded. Each man is also asked whether he smokes cigarettes on a regular basis. The results are crosstabulated using SPSS:

Health Level * Smoking Habit Crosstabulation

Count

		Smoking Habit		
		Doesn't Smoke	Does Smoke	Total
Health Level	Poor	13	34	47
	Fair	22	19	41
	Good	35	9	44
	Very Good	27	3	30
Total		97	65	162

(a) What are the variables and what are their respective levels of measurement?
(b) Should we characterize any possible relationship in terms of one variable being dependent and the other independent? Justify your answer.
(c) From this table calculate the column percentages.

5.7 Use the **Employee data** file to answer the following questions:

(a) The total number of managers in the sample.
(b) The total number of males in the sample.
(c) The total number of male managers.
(d) The total number of male managers as a percentage of all managers.
(e) The percentage of female employees in custodial positions as a percentage of all females.

6

Measures of association for crosstabulations: Nominal data

The previous chapter looked at the construction of crosstabulations. Crosstabs are a means of organizing categorical data in such a way as to reveal whether a relationship exists between two variables. We used as an illustrative example the results contained in Table 6.1.

Table 6.1 Health rating by sex of students

Health rating	Sex		
	Female	Male	Total
Unhealthy	34	16	50
Healthy	29	27	56
Very healthy	17	54	71
Total	80	97	177

When analyzing a crosstab to see if a relationship exists we ask two related questions:

- What is the **pattern** of the relationship?
- How **strong** is the relationship?

We can see that in this crosstab the **pattern** of the relationship is such that females tend to rate their health lower than males. We can also describe, in verbal terms, the **strength** of the relationship. The variation in students' sex is related to a the variation in health rating. My impression from the table leads me to use words like 'mild' or 'moderate' to describe the strength of the relationship I observe. Notice, though, how subjective is this choice of words. You may read this and think that you would use words more like 'strong' or 'considerable' to describe the strength of the relationship in this crosstab.

It would be more objective to have a way of *measuring* the strength of the relationship evident in a crosstab. Rather than leave it to an eyeball judgment that might vary from person to person, it would be better to have a way of measuring the strength of a relationship that will give the same answer, regardless of the person making the judgment. This is precisely the function of **measures of association**. An analogy may aid this discussion. I may regard today as being a 'fairly warm' day, while another person may judge it to be 'very warm', while another person may feel that the temperature is 'pretty cool'. We are all experiencing – 'observing' – the same thing, which is today's temperature, but our subjective interpretations of this experience are different. If, however, we all refer to a standard thermometer and see that the temperature is 20 degrees, this is something we can all agree about. The thermometer shows the same number regardless of who looks at it. The thermometer is an objective quantification of temperature since it is based on a common standard. Similarly, while different people may look at a crosstab and verbally assess the strength of a relationship in different ways, measures of association can provide an unequivocal index of the strength of a relationship that will give the same answer for everyone.

Measures of association as descriptive statistics

Measures of association are descriptive statistics that quantify a relationship between two variables.

Measures of association indicate, in quantitative terms, the extent to which a change in the value of one variable is related to a change in the value of the other variable.

Association is another word for 'relationship' or 'dependence': when age increases does height also increase (or decrease)? Is a change in religious beliefs associated with a change in attitude to capital punishment? (The word 'correlation' is normally used when measuring the relationship between two continuous variables, but effectively means the same as association).

As we have discussed, graphs and tables are some ways of identifying a relationship that may be present between two variables. We can, in addition to these simple methods of description, calculate measures of association to actually quantify the impressions gained from these tools. The most important thing to remember about measures of association is that they are meant to help us *describe* data. Rather than just relying on a visual impression of a crosstab or graph, they can, *in the appropriate circumstances*, provide a single figure for the strength of association.

The problem with these measures is determining the appropriate circumstances in which they can provide this information. If the right circumstances do not apply then these numerical measures may be misleading. Thus while it is possible to generate these measures on their own, I would advise against presenting them independently of a crosstab. It is easier to 'see' a relationship embodied in a crosstab, which can indicate whether the conditions necessary for calculating measures of association are present.

Unfortunately, putting the concept of association into practice is a slippery problem. Working with measures of association can be a very frustrating experience because there are a large number to choose from, each with its own peculiarities and limitations, and often they do not lead to the same result. For example, many measures of association are sensitive to the decision as to which variable is designated as independent and which is dependent. Such measures are **asymmetric**. Asymmetric measures are useful where we believe that the relationship is such that one variable is dependent on the other. If, on the other hand, we suspect that the relationship is one of mutual dependence, or else we are simply not sure of which model is appropriate, we use **symmetric** measures that take on the same value regardless of the variable that is specified to be the independent variable and that which is specified to be the dependent variable.

Table 6.2 provides some guide for choosing between the more common measures detailed in the following chapters. The starting point for selecting a measure is the level at which each variable is measured, particularly whether the data allow ranking (ordinal and interval/ratio scales) or not (nominal scales). (Those wanting a more complete treatment of measures of association that covers the full range of measures available should consult either of the two following texts, which provide an excellent, although sometimes very technical, discussion: H.T. Reynolds, 1977, *The Analysis of Cross-Classifications*, New York: The Free Press; A.L. Liebetrau, 1983, *Measures of Association*, Beverly Hills, CA: Sage Publications.)

In constructing a measure of association it is desirable for it to have the following properties:

- It is ideal for measures of association to take on the value of 1 (or –1 where appropriate) in situations of perfect association. Unfortunately this is not always the case, and the cause of much of the frustration tied up with using measures of association. Some measures can take on values larger than 1, while others (such as gamma) can take on the value of 1 where perfect association does not exist.
- It is ideal for measures of association to take on the value of 0 in situations of no association. Unfortunately not all measures meet this ideal quality. Some measures such as lambda can take on a value of 0 even where an association is evident to the naked eye.
- Where both variables are measured at least at the ordinal level, a + or – sign should indicate the direction of association: whether an increase in the quantity of one variable is associated with an increase (positive association) or decrease (negative association) in the quantity of the other variable.

Table 6.2 Measures of association

Measure	Symmetry	Data consideration	Comment
Lambda	Asymmetric	At least one variable is nominal	May underestimate strength of a relationship where one variable is ordinal or interval/ratio. May equal 0 even where a relationship exists
Goodman and Kruskal tau	Asymmetric	At least one variable is nominal	
Eta	Asymmetric	Suitable where independent variable is nominal and dependent variable is interval/ratio	Similar in logic to Pearson's *r*
Somer's *d*	Asymmetric	Both variables at least are ordinal	
Gamma	Symmetric	Both variables at least are ordinal	Not suitable for non-consistent relationship
Kendall's tau-*b*	Symmetric	Both variables at least are ordinal	Suitable only for tables with the same number of rows and columns
Kendall's tau-*c*	Symmetric	Both variables at least are ordinal	
Spearman's rho	Symmetric	Both variables at least are ordinal with many points on the scale	Special case of Pearson's *r* applied to the ranks of the scores rather than raw scores
Pearson's *r*	Symmetric	Both variables are interval/ratio with many points on the scale	Suitable for linear relationships

The rest of this chapter discusses measures of association for two variables when one or both of the variables is measured at the nominal level. Before doing so, it is important to remember that all that these measures do is detect association. They do not necessarily show whether one variable *causes* a change in another. We may suspect theoretically that one variable causes a change in the other, but the statistics we will learn here cannot prove causation, only provide supporting evidence for a theoretical model. For example, a relationship between the number of storks in an area and the birth rate in that area has been observed (see T. Höfera, H. Przyrembelb and S. Verleger, 2004, New evidence for the theory of the stork, *Paediatric & Perinatal Epidemiology*, vol. 18, p. 88), and we may calculate a measure that quantifies this statistical relationship. However, we cannot go from this statistical regularity to the conclusion that the storks cause the birth rate!

Measures of association for nominal scales

A measure of association, as we discussed above, is a numerical index that indicates the strength of a relationship. Measures of association range between two extremes. One extreme is the case of **perfect association**. In the case of perfect association, all cases with a particular value for one variable have a certain value for the other variable. For example, Table 6.3 illustrates a crosstab where sex of students and health rating are perfectly associated.

Table 6.3 Perfect association

Health rating	Sex		
	Female	Male	Total
Unhealthy	80	0	80
Healthy	0	0	0
Very healthy	0	97	97
Total	80	97	177

We can see that knowing if a student is male or female allows us to state with perfect certainty what their respective health rating will be. Sex is, for this group of cases, a **perfect predictor** of health status. Put another way, a change from female to male will always be associated with a change in health from 'Unhealthy' to 'Very healthy'. With perfect association we can say that all the variation in the dependent variable (health) is *explained* by the variation in the independent variable (sex): *the difference between two cases in terms of their health can be explained just by referring to the difference in their sex.*

The opposite extreme, displayed in Table 6.4, is the case of **no association**: knowing how a case scores on one variable gives no indication as to how it scores on the other variable.

Table 6.4 No association

Health rating	Sex		
	Female	Male	Total
Unhealthy	22 28%	27 28%	50 28%
Healthy	26 32%	31 32%	56 32%
Very healthy	32 40%	39 40%	71 40%
Total	80	97	177

There is not a relationship in these data between students' sex and health rating. *For each of the categories of the independent variable, exactly the same pattern of responses exists for the dependent variable.*

The two cases of no association and perfect association form the two opposite ends of the scale. The case of no association is given a value of zero and perfect association a value of 1 (Figure 6.1).

No association
Perfect association
0
1

Figure 6.1 Scale for nominal measures of association

We never actually gather data that fit either of these two extremes. They simply act as reference points. Data normally fall somewhere in between, such as the example we have been working with (Table 6.5).

Table 6.5 Health rating by sex of students

Health rating	Sex		
	Female	Male	Total
Unhealthy	34 43%	16 16%	50 28%
Healthy	29 36%	27 28%	56 32%
Very healthy	17 21%	54 56%	71 40%
Total	80	97	177

A visual inspection of this crosstab tells us that there is some relationship between these variables, but it is also clear that this is not a case of perfect association. If you had to give the strength of the relationship in this table a number between 0 and 1, with 0 representing no association and 1 representing perfect association, what would you give it? Is it closer to the data in Table 6.3 or Table 6.4, or somewhere in the middle?

The calculation of **lambda** gives us this number. Lambda gives an exact numerical location for where our actual result falls along the continuum in Figure 6.1. It does this by measuring the 'statistical distance' between the table containing the actual data we observe and each of the two possible extreme situations of no association and perfect association.

Lambda is one of a class of measures called **proportional reduction in error** (PRE) measures. The logic behind PRE measures is that if two variables are associated, then we should be able to predict the score that a case has on one variable on the basis of the score it has for the other variable. If sex and health rating are indeed related, then we should be able to predict a student's health rating by knowing whether they are male or female, and the stronger the relationship the more accurate will be our prediction.

All PRE measures follow a similar procedure. We try to predict how the cases will be distributed in a bivariate table under two conditions:

- we predict the distribution of cases along the dependent variable *without* any knowledge of their scores for the independent variable;
- we predict the distribution of cases along the dependent variable *with* knowledge of their scores for the independent variable.

To see how we make these predictions assume that the 177 students in our survey are lined up outside a room, and they will walk in one by one. Before each person enters we have to guess – predict – their health rating (i.e. predict their scores on the dependent variable). In making these predictions you are given only one piece of information, which is that the majority of *all* 177 students rate themselves as 'Very healthy'.

What guess will you make before each person walks in the room? Knowing only that the *majority* of students rate themselves as 'Very healthy' the best guess is to predict that *all* 177 students rate themselves as 'Very healthy'. In other words, with no other information, guess the average! In effect this uses the no-association model in Table 6.4 as the prediction rule.

Now if there was not much of a relationship between these two variables this prediction rule will generate very few errors. The closer that the actual pattern of cases resembles the no-association model the fewer errors that will be made when using this prediction rule to guess a student's health. In our example, if we guess all 177 students rate themselves as 'Very healthy' we make a prediction error of 106. This is the number of students who actually rated themselves as either 'Unhealthy' or 'Healthy' that we have incorrectly guessed as being 'Very healthy'. We call this E_1:

$$E_1 = 50 + 56 = 106$$

Now let us assume that these 177 students are asked to re-enter the room randomly one by one. This time, though, before each one enters you are told whether they are female or male. Suspecting that there is an association between sex and health rating such that females tend to rate themselves as 'Unhealthy' and males rate themselves as 'Very healthy', you predict that *every* female rates herself as 'Unhealthy' and *every* male rates himself as 'Very healthy'. This is effectively using the perfect association model from Table 6.3 as the prediction rule.

Following this prediction rule we make 89 errors. This is made up of the (29+17=) 46 females who were incorrectly classified as 'Unhealthy' and the (16+27=) 43 males incorrectly classified as 'Very healthy'. We call this E_2:

$$E_2 = (29 + 17) + (16 + 27) = 89$$

The question is whether I have made fewer mistakes when given the extra information about each student's sex (the independent variable). Did my suspicion about a possible association between these two variables reduce the error rate when making these predictions? The reduction in errors is 106 – 89 = 17. We have made 17 fewer errors by using the perfect association prediction rule than when we used the no-association prediction rule, indicating that there is some relationship in the data.

Lambda calculates this reduction in errors as a proportion of E_1 where E_1 is the number of errors without information for the independent variable and E_2 is the number of errors with information for the independent variable

$$\lambda = \frac{E_1 - E_2}{E_1}$$

As a proportion, the error rate has been reduced by:

$$\lambda = \frac{E_1 - E_2}{E_1} = \frac{106 - 89}{106} = 0.16$$

Therefore, by having information about a student's sex (the independent variable) we are able to reduce errors when predicting their health ratings by 16 percent. This is the great advantage of PRE measures: they measure something meaningful, which is changes in prediction error rates, and thus have a specific interpretation.

We can see in Figure 6.2 that the result places the observed distribution of data much closer to the no-association extreme than to the perfect association extreme.

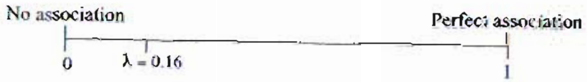


Figure 6.2

Lambda shows, in a clear-cut way, that although there is some relationship between these two variables, it is not very strong. Generally, we speak of association between variables as being weak, moderate, or strong (or some combination of these words, such as 'very weak' or 'moderately strong'). There is no sharp dividing line that determines when PRE values are to be called weak and when they are called strong, but to give a guide, one author suggests the terminology shown in Table 6.6. (See T.H. Black, 1993, *Evaluating Social Science Research*, London: Sage Publications, p. 137.)

Table 6.6 Interpreting values of lambda

Range	Relative strength
0.0	No relationship
0 > – 0.2	Very weak, negligible relationship
0.2 – 0.4	Weak, low association
0.4 – 0.7	Moderate association
0.7 – 0.9	Strong, high, marked association
0.9 – < 1.0	Very high, very strong relationship
1.0	Perfect association

We can see that for the data we are investigating the relationship is in the very weak range.

Properties of lambda

As a measure of association lambda has certain properties, some of which are desirable, but others (unfortunately) limit its applicability.

1. *Lambda will always equal 1 where data exhibit perfect association.* If we look at the way lambda is constructed it will have the desirable property that in the case of perfect association it will equal 1. If there is perfect association between two variables, the data will correspond exactly to the second of our prediction rules, producing no errors. If I make no errors with information about the independent variable ($E_2 = 0$) the value for lambda is:

$$\lambda = \frac{E_1}{E_1} = 1$$

2. *Lambda will always equal 0 where data exhibit no association.* If there is absolutely no association in the data, the observed results will conform exactly to the model of no association, and making predictions using the no-association model will yield no errors ($E_1 = 0$). This will generate a value for lambda of 0.

3. *Lambda will sometimes equal 0 where data exhibit some association.* Although lambda will always equal 0 when there is no association, the converse is not necessarily true: sometimes when lambda equals 0 there may indeed be association. This is a major limitation to the use of lambda and will be explored further at the end of the chapter.

4. *Lambda is an asymmetric measure of association.* This means that the value for lambda will be different depending on which of the two variables is considered to be independent and which is considered to be dependent. In other words, if in the example above we try to predict a person's sex based on their respective health rating, rather than the other way around, the value for lambda will change. Thus when using lambda we need to be explicit about the model of the relationship we think ties the two variables together. This makes lambda especially useful when we have strong reasons to believe that there is a one-way relationship between the two variables running in a certain direction.

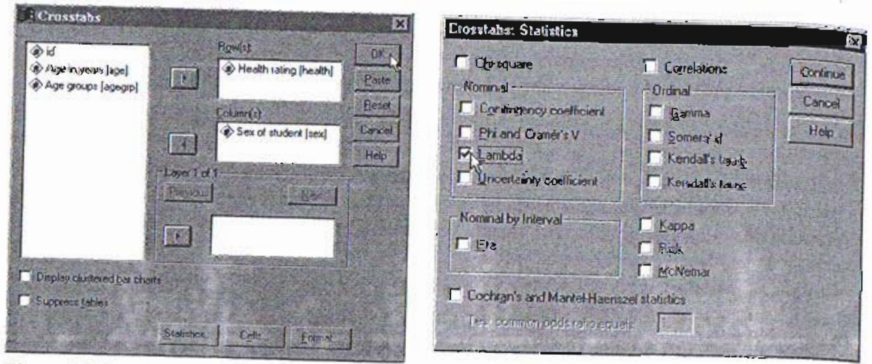
5. *Lambda ignores the ordering of categories for ordinal scales.* The value of 0.16 as a measure of the strength of the relationship for our crosstab above may strike you as lower than expected. To the naked eye, the relationship in Table 6.5 appears to be stronger than this value indicates. This has partially occurred because the calculation of lambda ignored the fact that the categories of health rating represent a quantitative increase in the variable; health is measured on an ordinal scale. Lambda does not take into account that in moving from female to male the modal response for health has increased from the lowest to the highest category. As far as lambda is concerned, the modal cell is in a *different* category, not a *higher* one. For example, assume that in Table 6.5, 54 males rates themselves as 'Healthy' and only 27 rated themselves as 'Very healthy'; the frequencies in these two categories are reversed. Common sense would suggest that the relationship is not as strong as it is in the original data distribution, since a switch from female to male causes the modal category for health to only jump one point on the scale. Yet lambda will still calculate the strength to be 0.16. Thus where we have one variable that is nominal and one that is ordinal, lambda may underestimate the strength of the relationship.

Lambda using SPSS

Lambda can be generated as an option of the **Crosstab** command in SPSS, which we introduced in Chapter 5 (Table 6.7 and Figure 6.3).

Table 6.7 Generating lambda on SPSS (file Ch06.sav)

SPSS command/action	Comments
1 From the menu select **Analyze/Descriptive Statistics/Crosstabs**	This brings up a window headed **Crosstabs**
2 Click on the variable that will form the rows of the table, which in this case is **Health rating**	This highlights **Health rating**
3 Click on ▸ that points to the area headed **Row(s):**	This pastes **Health rating** into the **Row(s):** target variable list
4 Click on **Sex of student**	This highlights **Sex of student**
5 Click on ▸ that points to the area headed **Column(s):**	This pastes **Sex of student** into the **Column(s):** target variable list
6 Click on the **Statistics** button	This brings up the **Crosstabs: Statistics** box. In the top-left corner you will see an area headed **Nominal Data**. These are the measures of association available when at least one variable is measured at the nominal level. In this instance **Sex of student** is measured at the nominal level
7 Select **Lambda** by clicking on the tick-box next to it	This places ✓ in the tick-box to show that lambda has been selected
8 Click on **Continue**	
9 Click on **OK**	



Row(s): Health rating [health]

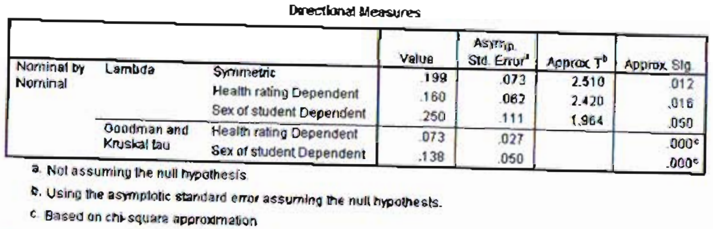
Column(s): Sex of student [sex]

☑ Lambda

Figure 6.3 The Crosstabs: Statistics dialog box

Notice that in the **Crosstabs: Statistics** dialog box we have the range of measures that we noted in Table 6.2 (plus others), broken down in a similar way by level of measurement. Thus we will be coming back to this dialog box frequently over the next two chapters as we work through the various measures of association.

If we follow the procedure in Table 6.7 we will obtain, along with the crosstab we generated in Chapter 5, the following table (Figure 6.4) labelled **Directional Measures** (which is SPSS's term for asymmetric measures of association). The table produces three versions of lambda: symmetric, asymmetric with Health rating as dependent, and asymmetric with Sex of student as dependent, from which we choose the one appropriate to our model of the relationship.


Directional Measures

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Lambda	Symmetric	.199	.073	2.510	.012
		Health rating Dependent	.160	.062	2.420	.016
		Sex of student Dependent	.250	.111	1.964	.050
	Goodman and Kruskal tau	Health rating Dependent	.073	.027		.000[c]
		Sex of student Dependent	.138	.050		.000[c]

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on chi-square approximation

Figure 6.4 SPSS Crosstabs: Statistics output

The symmetric version is used when there is no reason to suspect that one of the variables is dependent on the other, but rather that they are mutually dependent on each other. It is actually calculated as a weighted average of the two asymmetric versions: in this example the symmetric value of 0.199 falls somewhere in between the two asymmetric values of 0.160 and 0.250. The asymmetric version has two possible values, based on which of the two variables we believe is dependent. Here Health rating is dependent, and the value SPSS provides, 0.16, is the same as that we calculated by hand above, although it does so to 3 decimal places, rather than the 2 decimal places that we used in our hand calculations.

The table also produces the value of another nominal measure of association called Goodman and Kruskal *tau*, which has a much smaller value for the association than lambda. This indicates a 'problem' we will encounter a number of times in this and the next chapter, which is that different measures of association calculated on the *same* data will produce *different* values. This is because each measure conceptualizes the notion of association in slightly different ways and therefore will not always be in agreement.

The other columns in the **Directional Measures** table contain information that is not relevant at this point, but deals with issues that arise in later sections of this book. They deal with the problem of making an inference from a sample to a population.

Example

I suspect that there is a relationship between people's political orientation and their attitudes to equal rights legislation that has been proposed. I believe that political orientation is the independent variable and attitude to equal rights legislation is the dependent variable. One hundred people are selected and each person is asked to walk into a room. Before each person enters I have to guess whether that person favors or opposes equal rights legislation. However, I am given no information about each person's political orientation before they enter: I have to make a blind guess about each person's political beliefs (conservative or progressive). The only information I am given is that for the sample as a whole the modal response (the one with most cases) for the dependent variable is the 'support' category. Limited to this information, the best strategy is to guess that all 100 people support equal rights legislation. Since the modal category by definition is the category with the most observations we make the fewest errors, when we have no other information to inform our prediction, by predicting that all cases fall into it (Table 6.8).

Table 6.8 Prediction with no information about the independent variable

Attitude to equal rights	Political orientation
	Total
Oppose	0
Support	100
Total	100

I again have to predict the attitude of each person, but this time I am told the political orientation of each person. I use the perfect association model as the basis for prediction and guess that *all* conservative people oppose the legislation and *all* progressive people support it (Table 6.9).

Table 6.9 Prediction with information about the independent variable

Attitude to equal rights	Political orientation		
	Progressive	Conservative	Total
Oppose	0	50	50
Support	50	0	50
Total	50	50	100

The question is whether I have made fewer mistakes when given the extra information about each person's political leanings. To which one of these extremes does the actual data most closely conform? If there is little association between the two variables, the actual data will more closely resemble those in Table 6.8, whereas the stronger the association, the more closely the observed distribution will conform to that in Table 6.9. The extent to which the observed data are closer to one extreme or the other, or somewhere in between, will be expressed by the difference in error rates we make under each prediction rule.

Assume that the actual ('observed') data are those contained in Table 6.10.

Table 6.10 Observed frequency distribution

Attitude to equal rights	Political orientation		
	Progressive	Conservative	Total
Oppose	6	42	48
Support	44	8	52
Total	50	50	100

Even before we do any calculations an eyeball inspection of the crosstab will tell us that there is a strong association between these two variables, with a very high proportion of conservatives opposing and a high proportion of progressives supporting the legislation. We would place this table much closer to Table 6.9, which represents the case of perfect association. Thus the second prediction rule (perfect association) will dramatically reduce our errors when compared to the errors we make under the first prediction rule (no association). Before we proceed to actually calculate these error rates and lambda, can you guess what lambda will be for the actual survey data in Table 6.10 as a value between 0 (no association) and 1 (perfect association)?

Without any knowledge of the independent variable (i.e. whether a person is conservative or progressive), 42 conservatives who oppose the legislation are incorrectly classified as supporting it, and 6 progressives who oppose the legislation are incorrectly classified as supporting it. Therefore total errors made are:

$$E_1 = 42 + 6 = 48$$

With knowledge of a person's political orientation, however, 6 progressives are incorrectly classified as supporting the legislation, and 8 conservatives are incorrectly classified as opposing it. Therefore total errors made in this situation are:

$$E_2 = 8 + 6 = 14$$

Lambda calculates the difference in the two error rates as a proportion of the initial situation where *I* had no knowledge of the independent variable – hence the term 'proportional reduction in error':

$$\lambda = \frac{E_1 - E_2}{E_1} = \frac{48 - 14}{48} = 0.71$$

Therefore, by having information about political leaning we are able to minimize errors when predicting whether a person will support the proposed legislation by 71 percent. This indicates a strong relationship between the two variables. Did this figure correspond with the value you thought expressed the strength of the relationship based on just your visual inspection of the crosstab?

Limitations on the use of lambda

Despite its intuitive appeal and ease of calculation, a problem is all too frequently encountered when using lambda. The problem is one we have already noted above when discussing the properties of lambda. Lambda can have a value of 0 even though a relationship does exist between the two variables (which is evident just by looking at the crosstab). The cause of the problem is data that are highly skewed along the dependent variable.

Lambda will equal 0 when the modal category for the dependent variable is the same for all categories of the independent variable.

To see what this means in practice, we will analyze the following data (Table 6.11). In this hypothetical example respondents are asked whether the government is doing enough to alleviate poverty. Looking at the crosstab we can see that there is some relationship. A much higher percentage of under 45 year olds agree with the statement about government policy than people who are 45 or older. Clearly, there is some dependence between the two variables, and we might even describe it in verbal terms by saying that it appears to be fair to moderate in strength.

Table 6.11 Should the government do more to alleviate poverty?

Agree	Age group		
	Under 45	45 or over	Total
No	110 18%	168 42%	278
Yes	490 82%	232 58%	722
Total	600	400	1000

However, if we try to quantify this relationship with lambda we get a *measured* association of zero. Notice that the modal response for the dependent variable for all 1000 cases is 'yes', which is also the modal response for each of the two categories of the independent variable: the majority of people under 45 stated yes, and the majority of people aged 45 or over also stated yes. This skewed distribution in terms of the dependent variable will produce a lambda of zero, even when it is clear to the naked eye that some association does exist between the variables.

To see how, I need first to calculate the number of errors when predicting without knowledge of the independent variable (age group). I predict that all 1000 cases will fall in the 'yes' category, since this will minimize my error rate. I therefore make 278 mistakes:

$$E_1 = 278$$

With information about the independent variable, I will still make the same number of mistakes. Considering first the respondents aged under 45, I predict that all 600 respond 'yes' (110 mistakes). Second, I predict all 400 people aged 45 or over respond 'yes' (168 mistakes). This sums to 278 total errors, which is the same as predicting without knowledge of the respondents' sex.

$$E_2 = 278$$

The value for lambda will be:

$$\lambda = \frac{E_1 - E_2}{E_1} = \frac{278 - 278}{278} = 0$$

Lambda has failed to pick up the observable relationship, which is evident to the naked eye. This highlights one important rule:

Whenever lambda equals 0 inspect the relative frequencies to decide whether this actually reflects no association or whether it is due to a skewed distribution for the dependent variable.

If an inspection of the column percentages leads you to conclude that a value of 0 for lambda is due to a skewed distribution (as in this case), there are three options:

1. *Don't bother with measures of association*. Stick to the crosstab and the relative frequencies it contains, and base your conclusion regarding the relationship on this alone. This requires the researcher to make some subjective judgments, but as long as the crosstab is there for readers to assess for themselves, there is no problem with structuring an argument using only the relative frequencies as evidence. These frequencies sometimes 'speak for themselves': calculating more advanced statistics (and all the problems that sometimes come with them) may only serve to bury important information in an avalanche of suspect numbers. In the example above, we might say "Twenty-four percent more

people aged under 45 supported legislation to alleviate poverty than people 45 years and older. Slightly less than one-in-five people in the younger age group opposed the proposition whereas for the older people the level of opposition was nearly one-in-two."

2. *Calculate other measures of association.* There are other measures of association for nominal data that can be used if there are problems with lambda. Another PRE measure, which appeared in the SPSS output above, is the Goodman–Kruskal *tau*. Like lambda this is an asymmetric measure of association that ranges between 0 and 1. Unlike lambda it does not use the modal response for the independent variable in making predictions, but rather the frequency distribution of cases across all the categories of the independent variable. Since it is less sensitive to skewed marginal distributions than lambda it is a convenient alternative when skewness causes lambda to equal zero. Another measure of association is Cramer's *V*, which will always produce a value greater than 0 where an association exists between two variables. However, it does not have a simple interpretation in terms of PRE, and therefore cannot be used to assess the strength of a relationship for any given crosstab. It can be useful though when *comparing the strength of bivariate relationships across different tables*. The formula for Cramer's *V* is given presented in the Key Equations at the end of this book. Cramer's *V* is one of the options, along with lambda, for nominal data when choosing statistics in the SPSS Crosstabs command.
3. *Standardize the table so that the row totals are all equal* This is a slightly more complicated procedure, and one not often suggested by texts on statistics. (For a more complete discussion of standardization procedures and their use with measures of association see Y.M.M. Bishop, S.E. Feinberg, and P.W. Holland, 1975, *Discrete Multivariate Analysis: Theory and Practice*, Cambridge: MIT Press, pp. 392–3; and H.T. Reynolds, 1977, *The Analysis of Cross-Classifications*, London: Macmillan, pp. 31–3.)

Standardizing table frequencies (optional)

Standardizing a table involves trying to eliminate the variation brought about by the skewed distribution for the dependent variable, while still retaining the variation across the categories of the independent variable. When working with lambda we standardize the row marginals so that each row sums to 100. In a report that uses this procedure it should be made clear that lambda is not calculated on the raw data, by adding a comment or footnote such as: 'In calculating lambda, row marginals are standardized to sum to 100.' This involves the calculation of the row percentages, which are then treated as if they are real numbers of cases. That is, we calculate the percentage of the total 'yes' respondents that are under 45 and the percentage that are 45 or over. We do the same for the 'no' responses (Table 6.12).

Table 6.12 Should the government do more to alleviate poverty?

Agree	Age group		
	Under 45	45 or over	Total
No	$\frac{110}{278}\times 100 = 40\%$	$\frac{168}{278}\times 100 = 60\%$	100%
Yes	$\frac{490}{722}\times 100 = 68\%$	$\frac{232}{722}\times 100 = 32\%$	100%

We then use these percentage figures *as if* they are counts of actual cases, as in Table 6.13.

Table 6.13 Should the government do more to alleviate poverty?

Agree	Age group		
	Under 45	45 or over	Total
No	40	60	100
Yes	68	32	100

Remember that these are percentages: 40 represents 40 percent of 278 total no responses, and so on. But we treat them *as if* they are individual cases. This means that the total sample size 'is' 200 rather than 1000: the 100 yes 'respondents' and the *100* no 'respondents'.

Using the data from the standardized table (Table 6.13), I recalculate lambda. Without knowledge of the independent variable, I classify all 200 'respondents' in either yes or no, and therefore make 100 errors:

$$E_1 = 100$$

With knowledge of the independent variable I make the following predictions. Starting with the under 45s, I predict that all said yes, since this gives me the lowest error rate (40 mistakes). For 45 or over I predict that all said no, and therefore make 32 mistakes:

$$E_2 = 40 + 32 = 72$$

Lambda will therefore equal:

$$\lambda = \frac{E_1 - E_2}{E_1} = \frac{100 - 72}{100} = 0.28$$

After standardization, there turns out to be a weak to moderate association between these variables that lambda calculated on the original data could not extract.

Exercises

6.1 What is the difference between asymmetric and symmetric measures of association? Which is the appropriate measure to use in situations in which two variables are thought to be mutually dependent?

6.2 Why is it important, when calculating lambda, to decide whether one variable is likely to be dependent on the other, and if so to specify which is dependent and which is independent?

6.3 Calculate lambda for the following tables, and interpret the strength of any relationship:

(a)

Dependent	Independent		
	1	2	Total
1	30	60	90
2	45	50	95
Total	75	110	185

(b)

Dependent	Independent			
	1	2	3	Total
1	56	40	10	106
2	15	30	50	95
Total	71	70	60	201

(c)

Dependent	Independent			
	1	2	3	Total
1	70	40	10	106
2	50	45	38	133
3	43	30	14	87
Total	163	165	172	500

(d) (optional) If any of these tables produce a lambda equal to zero, standardize the distribution and recalculate lambda.

6.4 A researcher is interested in the relationship between gun ownership and attitude toward capital punishment. The researcher surveyed 3000 people and obtained the following results:

Capital punishment	Gun owners	Non-owners
For	849	367
Against	191	1593

Calculate lambda for these data and interpret the result.

6.5 A survey of 50 'blue-collar' and 50 'white-collar' workers asked respondents if they could sing their National Anthem from start to finish.

Blue-collar: Yes = 29, No = 21
White-collar: Yes = 22, No = 28

Arrange these data into a crosstabulation. What should be the dependent and independent variables? Calculate lambda for these data.

6.6 (optional) A study finds that the association between two variables, using Cramer's V as the measure, is 0.34. In the past, studies have measured association between the same variables using V as ranging from 0.15 to 0.21. How should the researchers report their result?

6.7 Open the **Employee data** file. Recode current salary into class intervals based on $10,000 income brackets. Use this recoded variable to:

(a) assess the strength of any association between current income and gender;
(b) assess the strength of any association between current income and employment category.

In your answers you should be careful to specify how you are modelling the relationships and choose the measure accordingly.

7

Measures of association for crosstabulations: Ranked data

Chapter 5 illustrated the use of crosstabulations as a means of summarizing data for two variables that we suspect are related, and which are measured on scales with only a few points. It is important to begin with a visual inspection of a crosstab – to 'eyeball' the table – in order to observe directly whether the two variables are independent or whether they exhibit some kind of relationship. A visual inspection of the table tries to identify the variation in the data, and based on this we interpret the nature of any relationship that the crosstab reveals. In Chapter 6 we also noted that measures of association can be calculated in conjunction with the table to give quantitative precision to any relationship we observe.

Data considerations

This chapter concentrates on measures of association for scales that can be ranked. In other words, both variables must be measured at either the ordinal or interval/ratio levels. We should note, though, that where both variables are measured on ranking scales with many points, the measures of correlation that we will discuss in Chapter 12 might be more appropriate to those discussed in this chapter. Thus the measures discussed here are generally used for ranked data with only a few points on the scale, and thus can also be effectively described in a crosstab (rather than a scatterplot).

The measures discussed in this chapter are PRE measures of association that are similar to lambda in their basic logic and how we interpret them. With lambda, we try to predict the value an *individual* case takes for the dependent variable. We do this by assuming first that there is no association between the variables, and then second by assuming that there is perfect association between the two variables. By comparing the error rates under each prediction rule we can assess the relationship contained in the set of data we actually collect.

We undertake a similar procedure with ranking scales, but we make use of the extra information about the variables given to us by the higher levels of measurement; with ordinal and interval/ratio scales, unlike nominal scales, we know how cases are *ranked* relative to each other. The following measures of association are based on our success in predicting these rankings.

For example, we introduced in Chapter 5 the data in Table 7.1, which display a positive relationship between Income and Frequency of TV watching, with each variable measured on an ordinal scale.

Table 7.1 Frequency of TV watching by income

TV watching	Income			
	Low	Medium	High	Total
Never	75 71%	15 15%	10 11%	100
Some nights	20 19%	70 70%	10 11%	100
Most nights	10 10%	15 15%	75 79%	100
Total	105	100	95	300

We can proceed to quantify the strong, positive relationship we observe in this crosstab by calculating the relevant ordinal measures of association. There are a number of PRE measures of association that can be calculated for such a table, varying slightly in their respective methods. All of the measures for ordinal data we will discuss have the common characteristic of being based on the distinction between **concordant** and **discordant pairs**.

Concordant pairs

Assume that one of the 75 high income people in Table 7.1 who watches TV most nights is named Alex, and one of the 70 medium income people who watch TV some nights is called Andrea. These two people can be ranked against each other for each of the two variables (Figure 7.1):

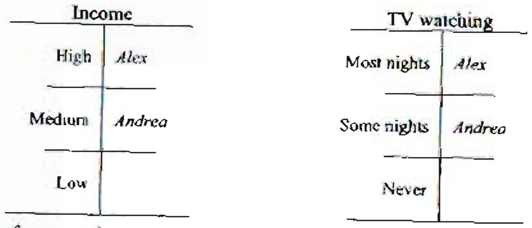


Figure 7.1 Ranking of a concordant pair

The ranking of this pair of cases is summarized Table 7.2.

Table 7.2 A concordant pair of cases

Independent variable: income	Dependent variable: TV watching
Alex ranked above Andrea (has a higher income)	Alex ranked above Andrea (watches more TV)

Therefore these two cases *are ranked the same for each variable*. This might sound like a strange use of language: how can they be the same if they have different values? The point is that they are *ranked* the same: Alex is ranked above Andrea for each variable. We describe such a pair of cases as a **concordant pair** (N_c).

A **concordant pair** is formed by two cases in a joint distribution that are ranked the same on both variables.

We have picked out two cases from the whole set of 300 cases that form a concordant pair. How do we calculate the *total* number of concordant pairs contained in the table? To see this look at the shaded cells in the crosstab from which we drew Andrea and Alex (Table 7.3).

Table 7.3 Frequency of TV watching by income

TV watching	Income		
	Low	Medium	High
Never	75	15	10
Some nights	20	70	10
Most nights	10	15	75

In the discussion above I formed a concordant pair by matching Alex, who is one of the 75 cases with a high income and also watches TV most nights, with Andrea, who is one of the 70 medium income earners who watches TV some nights. In fact I can pair Alex up with *each and every* one of the 70 people in the 'medium/some nights' cell, producing 70 concordant pairs: Alex plus each of the 70 people in the middle cell of the table (including Andrea). I can then do the same for each of the other 74 people with a high income and watch TV most nights. This will produce, in total, 75 lots of 70 concordant pairs:

$$75 \times 70 = 5250$$

Looking at Table 7.4, though, we see that there are still more pairs of cases that will form concordant pairs.

Table 7.4 Frequency of TV watching by income

TV watching	Income		
	Low	Medium	High
Never	75	15	10
Some nights	20	70	10
Most nights	10	15	75

Each of the 75 cases in the bottom-right cell is also ranked above each of the 15 cases in the 'never/medium' cell: they both have a higher income and watch TV more. So this will add the following number of concordant pairs:

$$75 \times 15 = 1125$$

In fact, any case will form a concordant pair with any other case in a cell *above and to the left or below and to the right* of it in the table (provided the table has been constructed with the values increasing down the rows and across the columns). The total number of concordant pairs, therefore, will be as shown in Table 7.5.

Table 7.5 Calculating concordant pairs

75	15	10	
20	70	10	(75×70)+(75×15)+(75×20)+(75×75) = 13,500
10	15	75	
			+
75	15	10	
20	70	10	(10×15)+(10×75) = 900
10	15	75	
			+
75	15	10	
20	70	10	(15×20)+(15×75) = 1425
10	15	75	
			+
75	15	10	
20	70	10	(70×75) = 5250
10	15	75	

$$N_c = 13{,}500 + 900 + 1425 + 5250 = 21{,}075$$

Discordant pairs

Now if I take one of the 10 people who have a low income and watch TV most nights, named Chris, and compare him with Andrea (one of the 70 people with medium income and watches TV some nights), the ranking will not be the same for both variables. Chris is ranked *below* Andrea in terms of income, but ranked *above* Andrea in terms of TV watching (Figure 7.2). Such cases are called **discordant pairs** (N_d).

A **discordant pair** is formed by two cases in a joint distribution whose ranking on one variable is different to their ranking for the other variable.

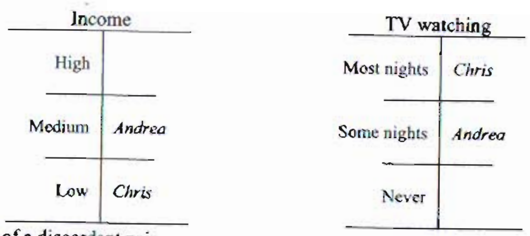


Figure 7.2 Ranking of a discordant pair

A case will form a discordant pair with any other case in the table that is in any cell *above and to the right/below and to the left*. To calculate the total number of discordant pairs we begin with the bottom-left cell in Table 7.6 and match it with all cells above and to the right or below and to the left.

Table 7.6 Calculating discordant pairs

75	15	10
20	70	10
10	15	75

$(10\times 70)+(10\times 15)+(10\times 10)+(10\times 10) = 1050$

+

75	15	10
20	70	10
10	15	75

$(20\times 15)+(20\times 10) = 500$

+

75	15	10
20	70	10
10	15	75

$(15\times 10)+(15\times 10) = 300$

+

75	15	10
20	70	10
10	15	75

$(70\times 10) = 700$

$$N_d = 1050 + 500 + 300 + 700 = 2550$$

Measures of association for ranked data

All PRE measures of association for ranked data use the difference between concordant and discordant pairs as the basis for assessing whether an association exists and determining its direction. The reason why we look at these concordant and discordant pairs is that they give us information that we can use in prediction. If two variables are positively associated then the crosstab will contain more concordant pairs than discordant pairs, and vice versa for negative association.

1. *Positive association between variables*. The data will contain a lot of concordant pairs and few discordant pairs. If this is this situation, and I am told a person ranks above another in terms of income, I will also predict that person ranks above the other in terms of frequency of TV watching as well.

$$\text{Positive association: } N_c - N_d > 0$$

2. *Negative association between variables*. The data will contain a lot of discordant pairs, so I will make the opposite prediction: knowing that a person ranks above another in terms of income will lead me to guess that that person ranks below the other in terms of frequency of TV watching.

$$\text{Negative association: } N_c - N_d < 0$$

3. *No association between variables*. The data will contain just as many concordant pairs as discordant pairs, and I will not increase my ability to predict the category of the dependent variable a case falls into by knowing its score for the independent variable.

$$\text{No association: } N_c - N_d = 0$$

There are four principal PRE measures of association for ordinal data: gamma, Somers' *d*, Kendal's tau-*b*, and Kendal's tau-*c*. These are all similar in that they have a PRE interpretation, and they all use the difference between N_c and N_d as the basis for assessing the strength of a relationship. The difference between them is in terms of how they standardize this difference. We will begin by exploring the simplest of these, which is gamma.

Gamma

Gamma is a common PRE measure of association for two variables measured at least at the ordinal level and arranged in a bivariate table. Gamma is a **symmetric measure of association** so that the value calculated will be the same regardless of which variable is specified as independent and which is specified as dependent. In other words, if we flipped the rows and columns around in our table, so that income is down the rows and TV watching is across the columns, the calculation of gamma will not be affected. Thus it is not sensitive to the particular model we believe characterizes the relationship between the two variables.

The formula for gamma expresses the difference between the number of concordant pairs and the number of discordant pairs *as a proportion of the total number of concordant and discordant pairs*. Using the data from our example, gamma will be:

$$G = \frac{N_c - N_d}{N_c + N_d} = \frac{21{,}075 - 2550}{21{,}075 + 2550} = 0.78$$

This indicates that we have a strong positive association between these two variables, which reinforces the conclusion we drew based just on the visual inspection of the crosstab.

The range of possible values for gamma is between –1 and 1. A gamma of –1 indicates perfect negative association: knowing that a case ranks above another along one variable indicates that it must rank below for the other variable. Such a result would be obtained if there were *only discordant pairs*, as in Table 7.7.

Table 7.7 Frequency of TV watching by income: perfect negative association

TV watching	Income		
	Low	Medium	High
Never	0	0	100%
Some nights	0	100%	0
Most nights	100%	0	0

If, on the other hand, there are *only concordant pairs* the value of gamma will be +1, indicating perfect positive association: knowing a case ranks above another for the independent variable indicates that it must also rank above for the dependent variable. Such a situation is reflected in Table 7.8.

Table 7.8 Frequency of TV watching by income: perfect positive association

TV watching	Income		
	Low	Medium	High
Never	100%	0	0
Some nights	0	100%	0
Most nights	0	0	100%

A gamma of zero indicates no association. If there are *just as many concordant pairs as there are discordant pairs*, then knowing the ranking along one variable gives no guide as to the ranking on the other variable. This situation is illustrated in Table 7.9.

Table 7.9 Frequency of TV watching by income: no association

TV watching	Income		
	Low	Medium	High
Never	50%	0	50%
Some nights	0	100%	0
Most nights	50%	0	50%

These three tables illustrate the three extreme points on a standardized scale measuring the strength of association between two ordinal variables (Figure 7.3).

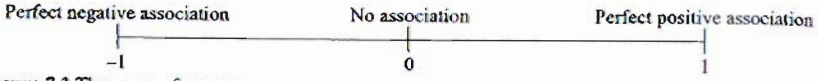


Figure 7.3 The range of gamma

Clearly the data for the example we are actually working with does not conform to any of these three extreme situations. It is a question of which prediction rule will be closest to the results we actually obtain. It is clear that the perfect positive association table is the one that the actual data most closely resemble, and gamma captures this quantitatively with a value of 0.78. It is not quite +1, but closer to it than to 0 or to −1.

Gamma is very popular in the literature because of its relative ease of calculation, although this advantage is now diminshed by the use of computer programs such as SPSS, which makes the calculation of all measures as easy as clicking buttons. I suspect that another element to its popularity is that compared to other ordinal measures of association, it generates the highest value for the strength of association for any given set of data.

Gamma does have some important limitations though, of which we need to be mindful. The first is that it is only a **symmetric** measure, and therefore does not take advantage of information provided by the data where we believe the most appropriate model for describing a relationship is one-way dependence (as we presumed in our example). The other main limitation is that, while perfect association will produce a value of +1 or −1, the converse is not always true: a gamma of +1 or −1 will not always indicate perfect association. It is possible to generate a value of −1 or +1 for a crosstab even where there is clearly less than perfect association. This occurs where the pattern of the relationship is not consistent. We should follow the rule, therefore, that *before using gamma a bivariate table should be inspected to assess whether the relationship is consistent*.

Both these limitations in fact stem from the same feature in the calculation of gamma. This is the failure of gamma to include **tied cases** in its formula. There are three types of tied cases.

1. *Cases tied on the independent variable* (T_x). These are pairs of cases that have the same score for the independent variable but have different scores for the dependent variable. These are usually any two cases in the same column of a crosstab but in different rows. In our example these are pairs of cases that have the same income but watch different amounts of TV.
2. *Cases tied on the dependent variable* (T_y). These are pairs of cases that have the same score for the dependent variable but which have different scores for the independent variable. In practical terms, these are any two cases in the same row of a crosstab but in different columns. In our example these are pairs of cases that watch the same amount of TV but have different income.
3. *Cases tied on both variables* (T_{yx}). These are cases that have the same score for both the variables. These are pairs of cases drawn from the same cell in the table. In our example these are pairs of cases that have the same income and watch the same amount of TV.

The other PRE measures of association for ordinal data seek to redress the limitations with gamma by including some or all of these tied cases in their calculation.

Somers' *d*

Somers' *d* is an **asymmetric measure of association**, in that it is sensitive to which variable is characterized as the independent variable and which is characterized as the dependent. Thus it is especially useful where we feel the relationship between two variables is best described by a one-way dependence model. The logic behind Somers' *d* is based on the idea that *two cases that vary in terms of the independent variable but do not vary in terms of the dependent variable* (they are *tied* on the dependent variable) reflect no association. In the example we have been working with, pairs tied on the dependent variable but not on the independent variable are those pairs of cases that are different in terms of income but watch exactly the same amount of TV. Somers' *d* calculates the association as a proportion of all concordant and discordant pairs *plus pairs tied on the dependent variable*:

$$d = \frac{N_c - N_d}{N_c + N_d + T_y}$$

To calculate the number of tied cases we take each cell, starting at top left, and multiply the number of cases it contains by the total number of cases in the cells to its right (Table 7.10). Substituting these calculations into the equation for Somers' *d* we get the following value:

$$d = \frac{N_c - N_d}{N_c + N_d + T_y} = \frac{21,075 - 2550}{21,075 + 2550 + 6350} = 0.62$$

A value of 0.62 indicates a moderate, positive association between these variables: an increase in income is associated with an increase in TV watching.

Table 7.10 Calculations for tied cases on the dependent variable

75	15	10	
20	70	10	(75× 15)+(75× 10) = 1875
10	15	75	
			+
75	15	10	
20	70	10	(15× 10) = 150
10	15	75	
			+
75	15	10	
20	70	10	(20× 70)+(20× 10) = 1600
10	15	75	
			+
75	15	10	
20	70	10	(70× 10) = 700
10	15	75	
			+
75	15	10	
20	70	10	(10× 15)+(10× 75) = 900
10	15	75	
			+
75	15	10	
20	70	10	(15× 75) = 1125
10	15	75	

$$T_y = 1875 + 150 + 1600 + 700 + 900 + 1125 = 6350$$

Notice that the equation for Somers' *d* is almost the identical equation to that for gamma, except for the term in the denominator for the number of dependent variable ties. As a result, whenever there are such tied cases, *d* will always have a lower value than gamma. In other words, by ignoring tied cases, gamma may overstate the strength of association between two variables in an asymmetric relationship when there are many tied cases.

Since Somers' *d* is an asymmetric measure of association we can actually calculate two alternative versions of it, for any given crosstab. We can calculate Somers' *d* with one variable as independent and the other as dependent, and we can then flip the variables around and calculate Somers' *d* again. In our example we have calculated *d* with income as the independent and TV watching as the dependent variable, since our theoretical model of this relationship depicts the causation as running in that direction. Someone with a different theory that postulated that somehow TV watching determines income would alternatively calculate *d* with income as dependent, and this will produce a different value.

Kendall's tau-*b*

Kendall's tau-*b* is a symmetric, PRE measure of association for ranked data arranged in a bivariate table. Its main feature is that it makes use of the information provided by cases tied on the dependent and on the independent variables:

$$\text{tau-}b = \frac{N_c - N_d}{\sqrt{(N_c + N_d + T_y)(N_c + N_d + T_x)}}$$

For the mathematically minded we note that tau-*b* is the geometric mean of the two alternative values for Somers' *d*. It is sometimes therefore referred to as the **symmetric version of Somers' *d***, even though this terminology is slightly confusing since *d* by definition is asymmetric. Since tau-*b* is the geometric mean of Somers' *d* it will have a value somewhere between the two values for *d* that can be calculated for any given crosstab.

Unfortunately, tau-*b* will only range between –1 and +1 where the number of rows in the crosstab is equal to the number of columns (a square table), and is therefore generally only used in this special case.

Kendall's tau-*c*

Kendall's tau-*c* is a symmetric, PRE measure of association much like tau-*b*. It is used in situations where a symmetric measure is desired for a table with an unequal number of rows and columns (which limits tau-*b*), and which has many tied cases (which limits gamma). The exact formula for tau-*c* is:

$$\text{tau-}c = \frac{2k(N_c - N_d)}{N^2(k-1)}$$

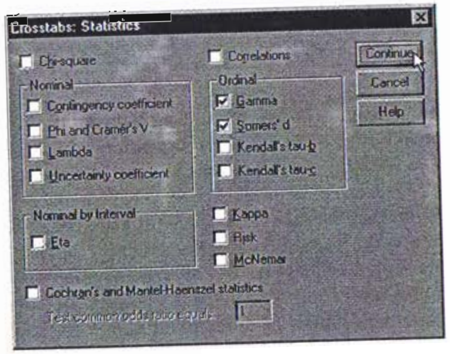
where *k* is the number of rows or the number of columns, whichever is smaller; and *N* is the total number of cases.

Measures of association using SPSS

Measures of association are available in SPSS as part of the **Crosstabs** command (Table 7.11, Figure 7.4). SPSS has produced the values for gamma and Somers' *d* in separate tables, since one is a symmetric measure and the other is asymmetric (which SPSS calls **Directional**). In either case the values generated by SPSS, in the column headed Value in each table, are the same as those we calculated by hand above. These reflect the moderate to strong association that exists between these two variables for these data. If there was a negative association, a negative sign will be printed in front of the value, provided that the data are arranged in the table in the correct format, with values increasing across the columns and down the rows.

Table 7.11 Ordinal measures of association on SPSS (file: **Ch07.sav**)

SPSS command/action	Comments
1 Select **Analyze/Descriptive Statistics/Crosstabs**	This brings up the **Crosstabs** dialog box
2 Click on **Frequency of TV watching** which is the variable in the source list that will appear down the rows of the table	This highlights **Frequency of TV watching**
3 Click on ▸ that points to the **Row(s):** target variables list	This pastes **Frequency of TV watching** into the **Row(s):** target variables list
4 Click on **Income level** which is the variable in the source list that will appear across the columns	This highlights **Income level**
5 Click on ▸ that points to the **Column(s):** target variables list	This pastes **Income level** into the **Column(s):** target variables list
6 Click on the **Statistics** button	This brings up the **Crosstabs: Statistics** box. You will see an area headed **Ordinal**, which provides a list of the measures of association available for this level of measurement
7 Select **Gamma** and **Somers' d** by clicking the boxes next to them	This places ✓ in the tick-boxes to indicate the statistics selected
8 Click on **Continue**	
9 Click on **OK**	



Directional Measures

			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Somers' d	Symmetric	.618	.043	14.192	.000
		Frequency of TV watching Dependent	.618	.043	14.192	.000
		Income level Dependent	.618	.043	14.192	.000

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Ordinal by Ordinal	Gamma	.784	.043	14.192	.000
N of Valid Cases		300			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Figure 7.4 SPSS **Crosstabs: Statistics** dialog box and output

Example of an asymmetric relationship

A public health researcher investigates whether a new drug improves rehabilitation for stroke victims. The researcher compares a group of 1013 stroke victims who do not take the drug with 588 stroke victims who do. Based on their ability to complete certain basic tasks the researcher classifies each person as showing 'no improvement', 'some improvement', 'moderate improvement', or 'strong improvement'. The researcher initially describes the data in the crosstab in Table 7.12.

It is very important to remember in constructing a bivariate table for ranked data that the values increase when going down the rows and across the columns. That is, the table begins with the lowest value for the row variable (which is normally the dependent variable) and moves down to the highest value. Similarly the first column should be the lowest value for the column variable (usually the independent variable) and increase across the page. This ensures that our procedures for calculating concordant and discordant pairs are appropriate.

Table 7.12 Effect of drug on health condition

Condition	Take drug?		
	No	Yes	Total
No improvement	42 4%	15 3%	57
Some improvement	86 9%	31 5%	117
Moderate improvement	316 31%	123 21%	439
Strong improvement	569 56%	419 71%	988
Total	1013	588	1601

Looking at the column percentages in this table it is evident that there is a relationship. For example, a higher percentage (71 percent) of people who have taken the drug showed strong improvement in their health condition than people who did not (56 percent). There is clearly a pattern of positive association: as drug taking increases (effectively from No to Yes) health condition also increases. We can also see, however, that the modal category for each group is 'strong improvement' indicating that there is not a very strong relationship evident in the data. In summary, our visual inspection of the table suggests a weak, positive association. By calculating the measures of association we get an exact quantitative measure of this impression. The calculations needed to determine the number of concordant pairs for these data are presented in Table 7.13.

Table 7.13 Calculating concordant pairs

42	15
86	31
316	123
569	419

(419×316)+(419×86)+(419×42) = 186,036

+

42	15
86	31
316	123
569	419

(123×86)+(123×42) = 15,744

+

42	15
86	31
316	123
569	419

(31×42) = 1302

$$N_c = 186{,}036 + 15{,}744 + 1302 = 203{,}082$$

The number of discordant pairs is calculated in Table 7.14.

Table 7.14 Calculating discordant pairs

42	15
86	31
316	123
569	419

(569×123)+(569×31)+(569×15) = 96,161

+

42	15
86	31
316	123
569	419

(316×31)+(316×15) = 14,536

+

42	15
86	31
316	123
569	419

(86×15) = 1290

$$N_c = 96{,}161 + 14{,}536 + 1290 = 11{,}1987$$

Putting all this into the formula for calculating gamma, we obtain:

$$G = \frac{N_c - N_d}{N_c + N_d} = \frac{203{,}082-111{,}987}{203{,}082+111{,}987} = 0.29$$

To calculate Somers' *d* we need to work out the number of pairs tied on the dependent variable, which is done in Table 7.15.

Table 7.15 Calculating tied cases on the dependent variable

42	15
86	31
316	123
569	419

(569×419) = 238,411

+

42	15
86	31
316	123
569	419

(316×123) = 38,868

+

42	15
86	31
316	123
569	419

(86×31) = 2666

+

42	15
86	31
316	123
569	419

(42×15) = 630

The total number of pairs tied on the dependent variable will be:

$$T_y = 238{,}411 + 38{,}868 + 2666 + 630 = 280{,}575$$

This will yield a value for Somers' *d* of:

$$d = \frac{N_c - N_d}{N_c + N_d + T_y} = \frac{203{,}082-111{,}987}{203{,}082+111{,}987+280{,}575} = 0.15$$

This is considerably weaker than the value for gamma, indicating the high number of tied cases. *Given that this is clearly a case of one-way dependence, Somers' d, as an asymmetric measure of association, is preferred.* In terms of the research question it would seem relevant to include in our calculations all those pairs of people who differed in terms of whether they took the drug yet showed no difference in health improvement.

Example of a symmetric relationship

A survey is conducted to assess whether the presence of union officials in the workplace is related to the accident rate for that workplace. The researcher thinks there is a relationship of mutual dependence between these variables: the level of unionization is affected by the accident rate, but also in turn affects the accident rate by raising consciousness and policing of safety regulations. The researcher will therefore use gamma, since it is a symmetric measure of association.

One hundred and seventy-seven workplaces are included in the survey and these are classified as having a low, moderate, or high level of union presence. These workplaces are also classified as having either a high or low accident rate. The results of the survey are presented in Table 7.16.

Table 7.16 Accident rates at the workplace by union presence

Accident rate	Union presence			
	Low	Moderate	High	Total
Low	17	32	35	84
High	43	27	23	93
Total	60	59	58	177

Can we detect an association between these variables?

To calculate gamma we begin with concordant pairs. For a 2-by-3 table such as this the combination of concordant pairs can be determined using the calculations in Table 7.17.

Table 7.17 Calculating concordant pairs

17	32	35	
43	27	23	(23×17)+(23×32) = 1127
			+
17	32	35	(27×17) = 459
43	27	23	

$$N_c = 1127 + 459 = 1586$$

To calculate the number of discordant pairs we work in the opposite direction (Table 7.18).

Table 7.18 Calculating discordant pairs

17	32	35	
43	27	23	(43×32)+(43×35) = 3311
			+
17	32	35	(27×35) = 945
43	27	23	

$$N_d = 3311 + 945 = 4256$$

Putting this information into the equation for gamma we get:

$$G = \frac{N_c - N_d}{N_c + N_d} = \frac{1586 - 4256}{1586 + 4256} = -0\ 46$$

This indicates that in predicting the order of pairs on one variable (accident rate), we will make 41 percent fewer errors if we take into account the way that the pairs are ordered on the other variable (level of unionization). There is a moderate, negative association between these two variables. Higher unionization is associated with a lower accident rate.

The other symmetric measure available to us is tau-*c* (tau-*b* is not appropriate since this is a table with a different number of rows and columns):

$$\text{tau-}c = \frac{2k(N_c - N_d)}{N^2(k-1)} = \frac{2(2)(1586-4256)}{177^2(2-1)} = -0.34$$

This is slightly lower than gamma, which is due to the presence of tied cases, but it still points to the existence of a moderate, negative relationship between these two variables.

Summary

We have investigated the calculation of a variety of PRE measures of association where both variables are measured at least at the ordinal level. Unfortunately, there is no easy rule for deciding which is the 'best' measure to use. Part of the problem lies with the notion of association itself, and the fact that this concept is operationalized in slightly different ways. For example, gamma, the tau measures, and rho are symmetric measures, whereas the Somers' *d* is asymmetric, so the choice should be guided by the model of the relationship we believe in. In practice, these measures usually 'point' in the same direction, in so far as they will generally give similar answers.

Exercises

7.1 If decreases in the value of a variable are associated with increases in the value of another variable, what is the direction of association?

7.2 Why do we not speak of association between two variables as being either positive or negative, when at least one variable is measured at the nominal level?

7.3 For the emboldened cells in each of the following tables, calculate the number of concordant pairs, assuming that the numbers on the edge of each table indicate the values of an ordinal scale:

(a)

	1	2	3
1	60	24	12
2	32	14	8

(b)

	1	2	3
1	60	24	12
2	32	14	8

(c)

	1	2	3	4
1	12	17	25	42
2	10	14	19	24
3	6	11	16	20

(d)

	1	2	3	4
1	12	17	25	42
2	10	14	19	24
3	6	11	16	20
4	3	9	14	22

7.4 For the **emboldened** cells in each of the following tables, calculate the number of discordant pairs, assuming that the numbers on the edge of each table indicate the values of an ordinal scale:

(a)

	1	2	3
1	60	24	12
2	32	**14**	8

(b)

	1	2	3
1	60	24	12
2	**32**	14	8

(c)

	1	2	3
1	60	**24**	12
2	32	14	8

(d)

	1	2	3	4
1	12	17	25	42
2	10	14	19	24
3	6	**11**	16	20

7.5 For the emboldened cells in each of the following tables, calculate the number of pairs of cases tied on the dependent variable but varying on the independent variable, assuming that the numbers on the edge of each table indicate the values of an ordinal variable:

(a)

	1	2	3
1	60	24	12
2	32	14	8

(b)

	1	2	3
1	60	24	12
2	**32**	14	8

(c)

	1	2	3
1	60	**24**	12
2	32	14	8

(d)

	1	2	3	4
1	12	17	25	42
2	10	14	19	24
3	6	**11**	16	20

7.6 For the example in Table 7.16, which looks at the relationship between accident rates and unionization in the workplace, calculate Somers' *d* and compare it to the value for gamma we calculated in the text.

7.7 Calculate gamma and Somers' *d* for the following table and interpret your result.

Mother working	Child achievement level			
	Poor	Good	High	Total
No	20	58	22	100
Part-time	15	62	23	100
Full-time	12	62	26	100
Total	47	182	71	300

7.8 Consider the following crosstabulation. The table displays the distribution of 162 patients whose health was assessed on a four-point scale, and who were also coded as smokers or non-smokers. This latter variable is considered ordinal for the purposes of this study since it indicates level of smoking.

	Smoking level		
Health level	Doesn't smoke	Does smoke	Total
Poor	13	34	47
Fair	22	19	41
Good	35	9	44
Very good	27	3	30
Total	97	65	162

(a) Looking at the raw distribution can you detect an association between these two variables? What is the direction of association? How will this direction manifest when calculating a measure of association?

(b) Calculate gamma and Somers' *d* and draw a conclusion about the relationship between health and smoking.

7.9 Open the **Employee data** file. Recode current salary into class intervals based on $10,000 income brackets. Use this recoded variable to assess the strength of the relationship between current income and employment category, treating the latter variable as ordinal variable indicating employment status. Why is tau-*b* not a useful measure in this instance?

8

Multivariate analysis of crosstabs: Elaboration

Chapters 5–7 analyzed the relationship between two variables. In those chapters it was assumed that any association observed in the data between two variables is due to a simple and direct relationship. A strong association in a bivariate table, however, does not necessarily mean that a simple direct relationship *in fact* exists; this is only how we have *interpreted* the data. There may be more complex relationships buried in the data, but we have not dug deep enough to find them.

The simplest way of extending – **elaborating** – the relationship discovered in a crosstab is to look at the possible impact that a third variable has on the original bivariate association. Depending on the outcome of this elaboration we may have to adjust our model of the relationship between the original two variables to take into account the influence of the third variable. There are three possible conclusions we can reach when we introduce a third variable into the analysis:

1. a **direct relationship** still exists (the third variable has no effect); or
2. either a **spurious** or **intervening relationship** exists; or
3. a **conditional relationship** exists.

We will investigate these possible outcomes by looking at examples of each in turn.

Direct relationship

We begin with an example where the original bivariate relationship does not change when we introduce a third variable. When the introduction of a third variable does not alter the original bivariate relationship, this will provide evidence that the simple direct model is the appropriate way of characterizing the relationship.

For example, we may have data on income and TV watching. Our theoretical model argues that income directly affects the amount of TV someone watches by affording them more or less leisure time. To express this we arrange the data in a crosstab and calculate a measure of association such as gamma (Table 8.1). These descriptive statistics tell us that there is a moderate to strong, positive relationship.

Table 8.1 TV watching by income level

TV watching	Income		
	Low	High	Total
Low	115 57%	95 32%	210
High	88 43%	204 68%	292
Total	203	299	502
Gamma=0.47			

When we argue that there is a direct relationship between two variables in this way we are effectively arguing that the relationship will be the same regardless of any other variable that may cause cases to vary from each other. In this example, we think income affects TV watching in the same way and to the same degree, regardless of any other variable that may cause cases to vary, such as sex, age, hair color, etc. This direct bivariate model, however, may appear to be overly simplistic. Surely there are other variables which impact on the amount of TV someone watches. Another researcher, for example, may feel that level of education also affects the amount of TV watched by individuals.

To assess the possible impact this new variable (level of education) has on the observed relationship between income and amount of TV watched, we divide the sample into two sub-groups: those who have no post-secondary education and those who have completed some post-secondary education. In technical terms education level is a **control variable**.

A control variable decomposes the data into sub-groups based on the categories of the control variable.

The effect of this control variable is to **generate a separate crosstab for each of the sub-groups defined by the control variable**. In this example, we first take *only those cases with no post-secondary education* and create a crosstab between their income and TV watching, ignoring those cases with some post-secondary education. We then take *only cases with some post-secondary education* and create a crosstab between their income and TV watching, ignoring people with no post-secondary education.

The resulting crosstabs are called **partial tables** and *we generate as many partial tables as there are categories for the control variable* (Table 8.2, Table 8.3). Here the control variable, 'Education level', only has two categories; we therefore generate two partial tables. (If we had three categories for the control variable, say 'no post-secondary', 'some post-secondary', 'a lot of post-secondary', we would generate three partial tables.)

Table 8.2 TV watching by income level: controlling for education level (no post-secondary education)

TV watching	Income		
	Low	High	Total
Low	78 57%	22 31%	100
High	58 43%	48 69%	106
Total	136	70	206
Gamma = 0.49			

Table 8.3 TV watching by income level: controlling for education level (post-secondary education)

TV watching	Income		
	Low	High	Total
Low	37 55%	73 32%	110
High	30 45%	156 68%	186
Total	67	229	296
Gamma = 0.45			

With this outcome we can see that the original relationship is reproduced almost exactly for each partial table. The value for gamma for each of the two partial tables is almost the same as that for the original table, before we controlled for education. In other words, regardless of the level of education, the relationship between income and TV watching still holds. The direct relationship we first observed is preserved even after controlling for the third variable. No matter how cases vary according to education level, the direct bivariate relationship remains basically the same, so we will not alter our initial model that characterized income and TV watching in a direct relationship.

Elaboration of crosstabs using SPSS

We can add control variables when generating a crosstab (Table 8.4, Figure 8.1) as part of the **Analyze/Descriptive Statistics/Crosstabs** command we introduced in Chapter 5. Note that Steps 8 and 9 are only optional when elaborating crosstabs, but the additional information they provide will help us interpret the results (Figure 8.2).

Table 8.4 Crosstabs with control variables on SPSS (file: **Ch8.sav**)

SPSS command/action	Comments
1 From the menu select **Analyze/Descriptive Statistics/Crosstabs**	This brings up the **Crosstabs** dialog box
2 Click on **TV watching**	This highlights **TV watching**
3 Click on ▸ pointing to the target list headed **Row(s):**	This pastes **TV watching** into the **Row(s):** target list
4 Click on **Income**	This highlights **Income**
5 Click on ▸ pointing to the target list headed **Column(s):**	This pastes **Income** into the **Column(s):** target list
6 Click on **Education level**	This highlights **Education level**
7 Click on ▸ pointing to the target list below **Layer 1 of 1**	This pastes **Education level** into the target list that contains the control variable. A crosstab will be generated for each value of the variable in this list
8 Click on the **Statistics** button and select **Gamma**	This will produce gamma for each partial table
9 Click on the **Cells** button and select **Column** percentages	This will generate the relative frequencies for each partial table based on the column totals
10 Click on **OK**	

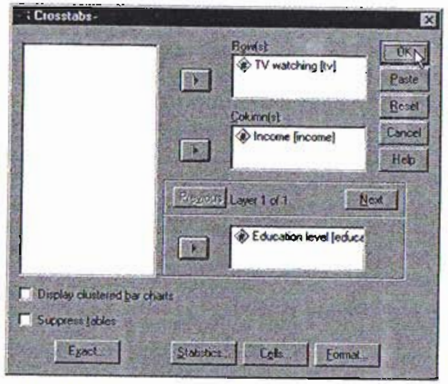
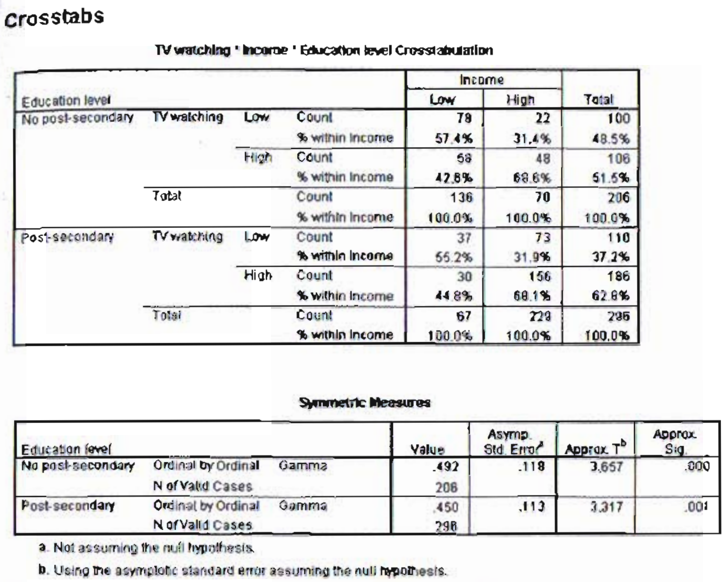


Figure 8.1 The **Crosstabs** dialog box

The table in Figure 8.2 is actually two crosstabs combined into one. The first half of the table is the crosstab of income and TV watching *for cases with no post-secondary education*, and immediately below it is the crosstab for those cases *with post-secondary education*. The percentage of cases watching a certain level of TV is the same for all income categories, *regardless of education level*.

This is reinforced by the values for gamma presented in the **Symmetric Measures** table. These gamma values are very similar to the value calculated on the unsegmented data in Table 8.1. The relationship between income and TV watching retains its strength and direction for each of the partial tables.



Crosstabs

TV watching * Income * Education level Crosstabulation

Education level				Income Low	Income High	Total
No post-secondary	TV watching	Low	Count	78	22	100
			% within Income	57.4%	31.4%	48.5%
		High	Count	58	48	106
			% within Income	42.6%	68.6%	51.5%
	Total		Count	136	70	206
			% within Income	100.0%	100.0%	100.0%
Post-secondary	TV watching	Low	Count	37	73	110
			% within Income	55.2%	31.9%	37.2%
		High	Count	30	156	186
			% within Income	44.8%	68.1%	62.8%
	Total		Count	67	229	296
			% within Income	100.0%	100.0%	100.0%

Symmetric Measures

Education level			Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
No post-secondary	Ordinal by Ordinal	Gamma	.492	.118	3.657	.000
	N of Valid Cases		206			
Post-secondary	Ordinal by Ordinal	Gamma	.450	.113	3.317	.001
	N of Valid Cases		298			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Figure 8.2 SPSS Crosstabs command output with a control variable

Partial gamma

Assume that when we introduce level of education into the analysis we instead obtain the following partial tables (Tables 8.5 and 8.6), rather than those in Tables 8.2 and 8.3.

Table 8.5 TV watching by income level: controlling for education level (no post-secondary education)

TV watching	Income		
	Low	High	Total
Low	102 75%	50 71%	152
High	34 25%	20 29%	54
Total	136	70	206
Gamma = 0.09			

Table 8.6 TV watching by income level: controlling for education level (post-secondary education)

TV watching	Income		
	Low	High	Total
Low	13 19%	45 20%	58
High	54 81%	184 80%	238
Total	67	229	296
Gamma = −0.007			

The relationship between income and TV watching that we observed in the original table has suddenly disappeared for each of the partial tables. It is clear to the naked eye that there is no association to speak of between income and TV watching, *once we have controlled for* *education level*. The original association we found has been 'washed out' by the introduction of the control variable. This impression is reinforced by the gamma values, which are now negligible in strength, unlike the combined gamma for the original bivariate table. In the original table, where the cases are not separated by level of education, gamma is 0.47. But the gamma values for each of the partial tables are very close to zero.

A more precise way of reaching this conclusion is to calculate the **partial gamma** for the data. The partial gamma is 'built-up' from the relationships embodied in the partial tables, rather than being calculated directly from the unsegmented data in Table 8.1. As we discussed in Chapter 7, gamma, is calculated on the basis of the number of concordant pairs and the number of discordant pairs. Concordant pairs, you remember, are pairs of cases that are ranked the same on each of the two variables, and thereby embody a positive relationship between the variables. Discordant pairs on the other hand are pairs of cases that are ranked differently on the two variables, reflecting a negative relationship between the variables.

If we add the concordant pairs across both partial tables and the discordant pairs across both partial tables we can calculate the **partial gamma**, which measures the direct relationship between the two variables we started with, *controlling for the third variable*. It is calculated by summing the concordant and discordant pairs across the partial tables. We still use all the cases in determining the partial gamma, but we are now doing it *after* separating the cases into two separate partial tables.

The process of calculating the partial gamma for these data is presented in Table 8.7.

Table 8.7 Calculating partial gamma

Table	Concordant pairs	Discordant pairs	Gamma
Original bivariate table	204×115 = 23,460	88×95 = 8360	0.47
Partial table 1	20×102 = 2040	34×50 = 1700	0.09
Partial table 2	13×184 = 2392	54×45 = 2430	–0.07
Total across partial tables	2040+2392 = 4432	1700+2430 = 4130	0.04

The partial gamma value for these data is only 0.04, indicating that there is very little *direct* relationship between income and TV watching, once we add level of education as a control.

Spurious or intervening relationship?

When the partial gamma is much lower than the original gamma calculated on the combined crosstab we should conclude that there is either a **spurious** relationship or **intervening** relationship between the first two variables. Before explaining each of these types of relationship, we need to point out that deciding which one explains the results of the elaboration is a *theoretical and not a statistical issue*. Having found that the original relationship disappears after elaborating a crosstab, it is up to us to decide how the three variables fit together, based on our understanding of how the world operates.

We might, for example, believe that the model represented in Figure 8.3 best explains the results we just analyzed.

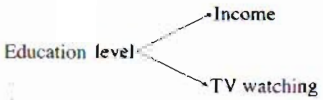


Figure 8.3 A spurious relationship

There is a spurious relationship between income and TV watching in that the relationship we originally observed between them (Table 8.1) does not exist; it is only a statistical outcome based on their respective relationships with the control variable. Education separately affects income and TV watching, but the latter two variables are not directly related to each other.

The classic example of a spurious relationship is the observed association between the presence of storks in an area and the birth rate (a reference to a study of this relationship appears in Chapter 6). Where there are many storks there is also a higher birth rate: the storks must be responsible for delivering babies! Of course this is a ridiculous argument and highlights the difference between a **statistical** relationship and a **causal** relationship. The observed relationship was explained by arguing that the same factors that caused the number of storks to vary across regions also caused the birth rate to vary. Specifically, rural areas attract storks, and they also attract people looking to start a family.

In other words, the relationship between the number of storks and the birth rate in a region is spurious. It does not really exist but is an artefact of two other relationships: the relationship between the type of region (rural, non-rural) and the number of storks, and the type of region and the birth rate.

Another researcher may look at the results of our elaboration of the crosstab between income and TV watching and instead characterize the relationship as in Figure 8.4.

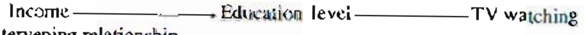


Figure 8.4 An intervening relationship

This researcher could make the argument that higher income earners can afford to undertake post-secondary education and then this affects how much TV they watch. Whether you think this argument is a good one or not is a matter for theoretical debate. Whether it is a more appropriate explanation of the results of the elaboration than the model of spurious relationship is open to discussion, but the statistical analysis itself cannot decide the issue. The statistical analysis merely indicate that *one* of these models best explains the results.

Conditional relationship

Assume that a researcher is interested in the extent to which patients respond to a program of exercise aimed at improving their cardiovascular system. The researcher organizes patients into low exercise and high exercise groups and observes whether there is any improvement in their cardiovascular systems (Table 8.8).

A visual inspection of Table 8.8, looking particularly at the (shaded) modal cells for each column, suggests that there is a strong, positive relationship between the variables. The exercise program does seem to work. To reinforce this impression the researcher calculates gamma, which produces a value of 0.68.

Table 8.8 Cardiovascular improvement by exercise level

Improvement	Exercise level		
	Low	High	Total
No	38 73%	11 34%	49
Yes	14 27%	21 66%	35
Total	52	32	84

The researcher could leave the results here, and conclude that a direct relationship has been observed between the independent variable (level of exercise) and the dependent variable (improvement level). However, the researcher believes that the actual relationship is more complex than this, and that there may be other factors left out of this analysis that may determine whether a patient's cardiovascular system improves. In particular, the researcher believes that whether a person has been a regular smoker will affect their chances of responding to the exercise program. The researcher therefore generates the crosstabulation, this time controlling for smoking level (Table 8.9 and Table 8.10).

Table 8.9 Cardiovascular improvement by exercise level: smokers only

Improvement	Exercise level		
	Low	High	Total
No	28 74%	7 70%	35
Yes	10 26%	3 30%	13
Total	38	10	48
Gamma = 0.09			

Table 8.10 Cardiovascular improvement by exercise level: non-smokers only

Improvement	Exercise level		
	Low	High	Total
No	10 71%	4 18%	14
Yes	4 29%	18 82%	22
Total	14	22	36
Gamma = 0.84			

When comparing these partial tables against the complete table we started with it is clear that the relationship works differently depending on smoking history. Regular smokers gained no improvement in their health levels as a result of the exercise program. But for non-smokers the relationship is even stronger than was evident in the complete table, a result that was 'diluted' by the inclusion of the smokers for whom the relationship does not seem to hold.

This is reinforced by the gamma values for each of these tables. For non-smokers, the value of gamma is 0.84, as opposed to 0.68 for the table as a whole. For smokers, though, there is practically no benefit from the exercise program. We can see that in gauging the effect of the control variable the measure of association is extremely useful, since it quantifies the changes that are brought about when the control variable is added.

As a result of this observation, the researcher changes the model which may tie the variables together. Instead of a simple one-way direct relationship, the researcher depicts the association in terms of a **conditional relationship**, as in Figure 8.5.

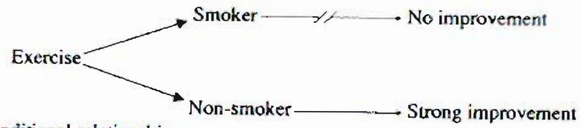


Figure 8.5 A conditional relationship

A conditional relationship is sometimes called **interaction**. Interaction exists where the relationship between two variables depends on the particular values of a third variable. Sometimes we might find that the relationship is reversed depending on the value of the control variable; for one sub-group the relationship might be positive, whereas for another sub-group the relationship might be negative.

Example

We want to investigate the relationship between intelligence and income. Intelligence is measured by a standard IQ test and respondents are divided into low and high IQ. Respondents are also divided into low or high income groups, depending on whether they earn below or above the median national income level.

The combined results for all 1000 people surveyed is presented in Table 8.11. This table illustrates a moderate association between intelligence, as measured by IQ, and income, and might lead to an interpretation that variation in intelligence causes the variation in income levels. People's earning capacity is to some extent predetermined by their respective IQs.

In order to avoid such a conclusion, we might argue that the IQ test as a measure of intelligence is biased. In particular we may feel that IQ scores are themselves a reflection of social class background, and this variable is a key determinant of income. To assess this we construct two partial tables, dividing the 1000 respondents into high social class and low social class sub-groups, producing the results in Tables 8.12 and 8.13.

Table 8.11 Income and intelligence

IQ	Income		
	Low	High	Total
Low	165 36%	95 18%	260
High	295 64%	445 82%	740
Total	460	540	1000
Gamma = 0.48			

Table 8.12 Income and intelligence: high social class only

IQ	Income		
	Low	High	Total
Low	20 18%	60 14%	80
High	90 82%	380 86%	470
Total	110	440	550
Gamma = 0.17			

Table 8.13 Income and intelligence: low social class only

IQ	Income		
	Low	High	Total
Low	145 41%	35 35%	180
High	205 59%	65 65%	270
Total	350	100	450
Gamma = 0.13			

We can see that the strength of the bivariate relationship is greatly diminished once we control for social class. There is little difference in the pattern of relative frequencies across the two partial tables. In fact, the partial gamma calculated on the basis of the partial tables is only 0.15. We have either a spurious relationship or an intervening relationship.

Summary

We have looked at the way in which the introduction of a third variable may alter a relationship we had previously observed between two variables. Indeed, the story can get even more complex when we allow for the impact of even more variables on the original bivariate relationship. Taking into account the possible effects of other variables involves **multivariate analysis**, and we have only just skimmed the surface in this chapter.

To help in drawing conclusions from the elaboration of crosstabs, Table 8.14 provides a useful guide to decision making (adapted from J. Healey, 1993, *Statistics: A Tool for Social Research*, Belmont, CA: Wadsworth, p. 428).

Table 8.14 Possible results when controlling for a third variable

Partial tables when compared with crosstab show:	Model	Implications for further analysis	Likely next step in statistical analysis	Theoretical implications
Same relationship between X and Y	Direct relationship	Disregard control variable	Select another control variable to test further the directness of the relationship	Model that X causes Y in a direct way is supported
Weaker or no relationship between X and Y	Spurious relationship	Incorporate control variable	Focus on the relationship between these three variables	Model that X causes Y is not supported
	or Intervening relationship	Incorporate control variable	Focus on the relationship between these three variables	Model that X causes Y is partially supported but must be revised to take control into account
Mixed relationships	Interaction/ conditional relationship	Incorporate control variable	Analyze sub-groups based on control variable separately	Model that X causes Y partially supported but must be revised to take control into account

Exercises

8.1 A study finds a strong positive relationship between a child's shoe size and the child's skills at mathematical problem solving. Explain.

8.2 What conclusion would you draw about the relationship between X and Y based on the following elaboration?

All cases

Y	X		
	1	2	Total
1	177	146	323
2	51	346	397
Total	228	492	720

Controlling for C(1)

Y	X		
	1	2	Total
1	153	52	205
2	44	123	167
Total	197	175	372

Controlling for C(2)

Y	X		
	1	2	Total
1	24	94	118
2	7	223	230
Total	31	317	348

8.3 An investigation of the relationship between age, concern for the environment, and political affiliation produces the following gamma values:

Gamma (age and concern for the environment): –0.57
Gamma (age and concern for the environment, liberals only): –0.22
Gamma (age and concern for the environment, conservatives only): –0.67
Partial gamma: –0.38

What conclusion should be drawn about the relationship, if any, between these three variables?

8.4 The following tables are based on a study of the likelihood of US courts to impose the death penalty, based on the racial characteristics of the victim and the defendant (M. Radelet, 1981, Racial characteristics and the imposition of the death penalty. *American Sociological Review*, 46, pp. 918–27).

All cases

Death penalty	Victim		
	White	Black	Total
No	184	106	290
Yes	30	6	36
Total	214	112	326

White defendant only

Death penalty	Victim		
	White	Black	Total
No	132	9	141
Yes	19	0	19
Total	151	9	160

Black defendant only

Death penalty	Victim		
	White	Black	Total
No	52	97	149
Yes	11	6	17
Total	63	103	163

What conclusions can you draw about the relationship between the race of the victim, the race of the defendant, and likelihood to impose the death penalty?

PART 3

Descriptive statistics: Numerical measures

9

Measures of central tendency

Part 2 looked at the description of data in graphical and tabular form. Tables and graphs as a form of describing data give some sense of the overall distribution of cases. For example, a quick glance at a frequency table or a histogram will identify the value that seems to be the 'center' of the distribution. However, we sometimes want to capture this feature of the data in more precise terms: what does the 'typical' or 'average' case look like?

Measures of central tendency

Measures of central tendency (also known as **measures of location**) are univariate descriptive statistics.

Measures of central tendency indicate the typical or average value for a distribution.

There are three common measures of central tendency: mode, median, and mean. Each measure embodies a different notion of average and, as Table 9.1 indicates, choosing which to calculate on a given set of data is restricted by the level at which a variable is measured.

Table 9.1 Measures of central tendency

Measure	Data considerations
Mode	Can be used with all levels of measurement, but not useful with scales that have many values
Median	Can be used with ranked data (ordinal and interval/ratio), but not useful for scales with few values
Mean	Can be used for interval/ratio data that are not skewed

In this table we can see one of the basic rules of statistics: *techniques that can be applied to a particular level of measurement can also be applied to a higher level.* For example, the measure of central tendency that can be calculated for nominal data (mode) can also be calculated for ordinal and interval/ratio data. This should be borne in mind as you read the rest of the book; when I refer to nominal-level statistical techniques I really mean 'nominal or above', and ordinal data techniques really refers to 'ordinal or above'. The converse, however, is not true: *measures that can be calculated for a particular level of measurement cannot always be calculated for lower levels.* The mean, for example, can only be calculated for the highest level of measurement (interval/ratio).

To see how each of these measures of central tendency is calculated we will use an extract of 20 cases from the hypothetical student survey we introduced in Chapter 2. The distributions for this sub-set of 20 students are presented in Tables 9.2, 9.3, and 9.4.

Table 9.2 Sex of respondents

Sex	Frequency
Male	12
Female	8
Total	20

Table 9.3 Health rating of respondents

Health rating	Frequency
Unhealthy	7
Healthy	5
Very healthy	8
Total	20

Table 9.4 Age of respondents

Age in years	Frequency
18	7
19	5
20	4
21	2
22	2
Total	20

The mode

We will start with the **mode** (M_o), which is the simplest measure of central tendency, and which can be calculated for all levels of measurement.

The mode is the value in a distribution with the highest frequency.

The mode is the only measure of central tendency that can be calculated for nominal data, and its great advantage over other choices is that it is very easy to calculate. A simple inspection of a frequency table is enough to determine the modal value or category.

For example, the category for sex that has the highest frequency in Table 9.2 is male, with 12 responses. For health rating in Table 9.3 the mode is 'very healthy', and for age in Table 9.4 the mode is 18 years.

Although it is exceptionally easy to determine the mode, occasionally people make the mistake of specifying as the mode the highest *frequency*, rather than the *score* with the highest frequency. That is, 12 might be reported as the mode for Table 9.2 since this is the highest frequency. This is incorrect – the important point to remember is that the mode is the *score* that occurs most frequently, not the number of times it appears in the distribution.

The mode has one feature that does not apply to the median or mean as measures of central tendency: there can be more than one mode for the same distribution. For example, assume we have the distribution for age shown in Table 9.5.

Table 9.5 Age of respondents

Age in years	Frequency
18	7
19	5
20	4
21	2
22	7
Total	25

We can see that two categories have the highest frequency: 18 years and 22 years. This is called a **bimodal** distribution. The median or the mean, on the other hand, will always produce only a single number as the average, regardless of the distribution.

The mode has one major limitation that arises especially when it is used to describe listed data for interval/ratio scales. Take, for example, the following scores that represent the time in seconds for a drug to take effect on a sample of patients, arranged in rank order:

33, 36, 36, 81, 82, 84, 86, 89, 91, 95, 97, 98

It is clear to the naked eye that the data are 'centered' somewhere in the 80–90 seconds range. Yet the mode for this distribution of listed data is 36 seconds since this appears twice in the distribution, whereas every other score only appears once. Clearly, the mode is not really reflecting the central tendency of this distribution. In such cases, we should either use other measures of central tendency, such as those we are about to discuss, or else organize the data into suitable class intervals, and *report the modal class interval*, rather than the individual modal score.

The median

With ordinal and interval/ratio data we can also calculate the **median** (M_d) score, along with the mode. We cannot calculate the median for nominal data since the determination of the median requires that the cases be rank-ordered from lowest to highest in terms of the quantity of the variable each case possesses. If all the cases in a distribution are ranked from lowest to highest, *the median is the value that divides the data in half.* Half of all the cases have a value for the variable greater than the median and half of all cases have a value less than the median. In other words, if I randomly select a case from a rank-ordered series, there is exactly a 50 percent chance that it will fall above the median and a 50 percent chance it will fall below the median.

For an **odd number** of rank-ordered cases, the median is the middle score.

For an **even number** of rank-ordered cases, the median is the mean of the two middle scores.

Thus if I lined up the 20 people in the survey (Figure 9.1), starting with the seven youngest that are 18 years old, followed by the 19 year olds, then the 20 and 21 year olds, and finally the two 22 year olds who are the oldest in the group, we can see that the mid-point of the distribution (between the 10th and 11th cases in line) is in the 19 years age group.

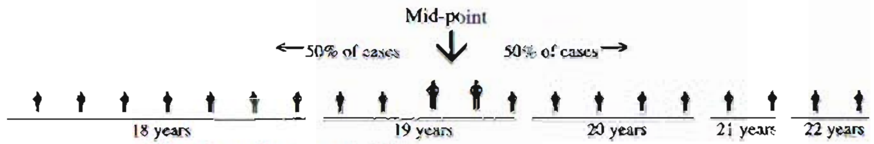


Figure 9.1 Calculating the median for ranked data

With an even number of cases, as we have here, the median is the average of the two middle scores, which are both 19 years, so the median will be 19 years:

$$\text{median} = \frac{19+19}{2} = 19 \text{ years}$$

However, if the 10th student was 19 years of age, and the 11th was 20 years of age, the median will then be 19.5 years:

$$\text{median} = \frac{19+20}{2} = 19.5 \text{ years}$$

If a cumulative relative frequency table has been generated (Table 9.6), an easier way to calculate the median is to identify the value at which the cumulative percent first passes 50.

Table 9.6 Age (in years) of respondents

Age	Frequency	Cumulative percentage
18	7	35%
19	5	60%
20	4	80%
21	2	90%
22	2	100%
Total	20	

The median has one limitation that is worth noting that arises especially with ordinal scales, such as that we have for the health rating of students. Although technically we can rank order cases from lowest to highest and find the middle score, it does not make much sense to do so when we *only have a small number of points in the scale* and therefore have a high proportion of cases in each of the categories. It does not really tell us much about the distribution to say that 50 percent of cases are 'Healthy or above' and 50 percent are 'Healthy or below'; in this circumstance the mode is a preferable measure of central tendency.

Example

Consider the following data:

93, 25, 87, 3, 56, 64, 12

To find the median of these data we first rank-order them from lowest to highest. Since there are seven cases (an odd number) the median value will be the 4th in line, i.e. 56:

Score:	3	12	25	56	64	87	93
Rank:	1st	2nd	3rd	4th	5th	6th	7th

If the same data set included one additional value of 98 the rank ordering will be:

Score:	3	12	25	56	64	87	93	98
Rank:	1st	2nd	3rd	4th	5th	6th	7th	8th

We now have eight cases (an even number). The median will therefore be the average of the 4th and 5th values:

$$\text{median} = \frac{56+64}{2} = 60 \text{ years}$$

The mean

With interval/ratio data the (arithmetic) **mean** can be calculated in addition to the mode and the median. The mean is the notion of average that is most commonly used, and in fact is often (incorrectly) synonymous with the term average.

The mean is the sum of all scores in a distribution divided by the total number of cases.

When calculating the mean for an entire population we use the Greek symbol μ (pronounced 'mu'). When calculating the mean for a sample, we use the Roman symbol $\bar{X}$ (pronounced 'X-bar'). The actual formula we use to calculate the mean depends on whether we have the data in listed form, or in a frequency table, or arranged into class intervals.

Listed data

If we have the raw data in listed form (with each individual datum listed separately) the equation for the mean of the population and the mean of a sample respectively are:

$$\mu = \frac{\Sigma X_i}{N}, \quad \bar{X} = \frac{\Sigma X_i}{n}$$

where N is the size of the population, n is the size of the sample, and X_i is each score in a distribution. The Σ (pronounced 'sigma') means 'the sum of' (or 'add up'), so we read these equations in the following way: 'the mean equals the sum of all scores divided by the number of cases'. Thus if I have the listed distribution of sample scores of 12, 15, 19, 27 the mean is:

$$\bar{X} = \frac{\Sigma X_i}{n} = \frac{12+15+19+22}{4} = 17$$

Frequency data

We sometimes do not have data presented in listed form, but instead have data grouped into a frequency table such as in Table 9.5. In this table we do not have the age for each person listed individually. Instead we have a frequency distribution of data grouped by years. In this case we use the following formula to calculate the mean for a sample:

$$\bar{X} = \frac{\Sigma fX_i}{n}$$

This formula instructs us to:

1. multiply each value in the distribution by the frequency (f) with which it occurs;
2. sum these products; and
3. divide the sum by the number of cases.

Here we have seven respondents aged 18, five aged 19, four aged 20, two aged 21, and another two aged 22. The mean is 19.35 years:

$$\bar{X} = \frac{(18\times7)+(19\times5)+(20\times4)+(21\times2)+(22\times2)}{20} = 19.35 \text{ years}$$

Frequency data using class intervals

Sometimes frequency tables only specify the class intervals in which data fall, rather than the specific values and the frequency with which each value occurs. A slightly more complicated procedure is involved when the data are grouped into class intervals, rather than by specific values. For example, we may be reading a report that includes Table 9.7 with the following information about children's ages.

Table 9.7 Children's ages grouped by class intervals

Intervals	Frequency
1–5 years	7
6–10 years	10
11–15 years	6
Total	23

The report, however, does not calculate the average age, so if we want this extra bit of description we need to calculate it for ourselves. With data grouped into class intervals we need to calculate the mid-points (m) and then multiply the frequencies by these mid-points:

$$\bar{X} = \frac{\Sigma fm}{n}$$

The procedure involved in using this equation is:

1. calculate the mid-point of each class interval;
2. multiply each mid-point by the number of cases in that interval;
3. sum these products; and
4. divide the total by the number of cases.

Thus for data in Table 9.7, the mid-points and the mid-points multiplied by the frequency of each class are as given in Table 9.8. Substituting these data into the formula we get (rounding to 1 decimal place):

$$\bar{X} = \frac{\Sigma fm}{n} = \frac{179}{23} = 7.8$$

Table 9.8 Calculations for the mean for class interval frequency data

Class intervals	Mid-point, (m)	Frequency, (f)	fm
1–5	3	7	3×7 = 21
6–10	8	10	8×10 = 80
11–15	13	6	13×6 = 78
Total		$n = 23$	$\Sigma fm = 179$

Choosing a measure of central tendency

The results we have generated for the age of these 20 students are summarized in Table 9.9.

Table 9.9 Age of respondents

Measure of central tendency	Value
Mode	18 years
Median	19 years
Mean	19.3 years

It is clear from this table that where more than one measure of average can be calculated we will not always get the same answer, even when calculated on the same raw data. This is because each measure defines 'average' in a slightly different way. In fact, unless the distribution is perfectly symmetrical, that is if the distribution is skewed, there will always be some difference in the various measures of central tendency. We can see examples of symmetrical and skewed distributions in Figure 9.2.

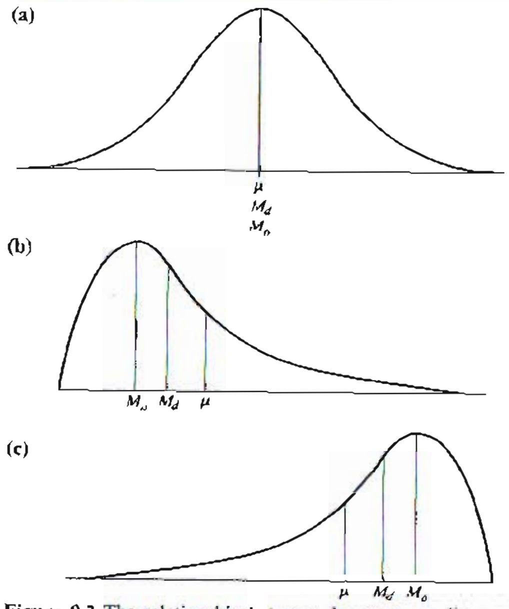


Figure 9.2 The relationship between the mean, median, and mode for a (a) symmetrical, (b) right-skewed, and (c) left-skewed distribution

The symmetrical curve has a nice bell-shape, and the measures of central tendency are all equal. With skewed distributions, though, the measures diverge. Notice also that in describing the direction to which a distribution is skewed we refer to the side of the curve that has the long tail, and not the side with the 'hump'.

Generally, when a distribution is heavily skewed the mean is a misleading notion of average. As the mean is calculated from *every* value in the distribution, it is influenced by extreme scores and outliers. For example, we may have the following exam scores:

$$X_1 = 60,\ X_2 = 62,\ X_3 = 66,\ X_4 = 67,\ X_5 = 69$$

With each datum listed separately, the mean for this distribution is 64.8:

$$\bar{X} = \frac{\Sigma X_i}{n} = \frac{60+62+66+67+69}{5} = 64.8$$

Consider the effect on the mean if the scores vary only slightly so that the fifth score is 95 instead of 69:

$$X_1 = 60,\ X_2 = 62,\ X_3 = 66,\ X_4 = 67,\ X_5 = 95$$

Even though only one score has changed, causing the distribution to skew to the right, the value of the mean has changed dramatically:

$$\bar{X} = \frac{\Sigma X_i}{n} = \frac{60+62+66+67+95}{5} = 70$$

The 'average' student suddenly looks a lot smarter, because of this one change. The median for both distributions, though, remains 66. This is the score that the student in the middle of the distribution receives. Since the median depends solely on the value of this one score at the mid-point, it is not 'pulled' in one direction or another by scores at the extreme ends of the range, and is, for interval/ratio data, therefore best used with a skewed distribution.

Measures of central tendency using SPSS: Univariate analysis

When we need descriptive statistics, such as those we discussed above, for only *one* group there are at least three different commands in SPSS that will provide them for us. Before discussing these, however, we note that the mode and median can be easily determined from frequency tables, and therefore for nominal and ordinal data we really do not need any special commands to assess central tendency. It is only with interval/ratio data upon which the mean can be calculated (along with other measures we will discuss in the next chapter) that the following commands are most relevant.

The various commands for generating measures of central tendency all appear under the **Analyze/Descriptive Statistics** option (Figure 9.3).

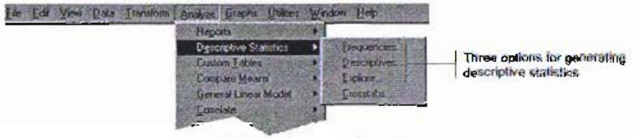
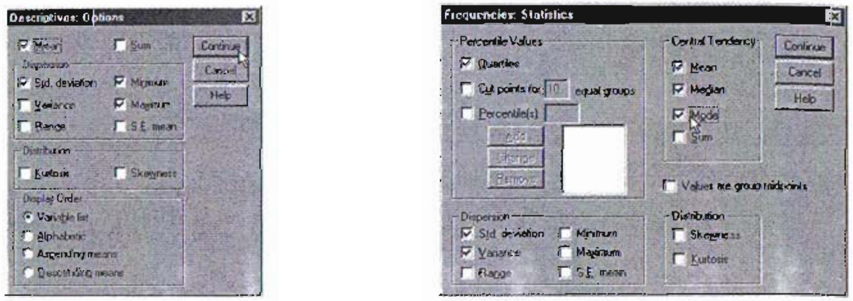


Figure 9.3 SPSS commands for univariate descriptive statistics

Ironically, the **Descriptives** command is the least useful of these three. If we compare the range of options available under this command with those available under the **Frequencies/Statistics** command, for example, we can see that the former only offers the mean as a measure of central tendency, and not the mode or median (Figure 9.4)


Descriptives: Options

☑ Mean ☐ Sum

Dispersion

☑ Std. deviation ☑ Minimum

☐ Variance ☑ Maximum

☐ Range ☐ S.E. mean

Distribution

☐ Kurtosis ☐ Skewness

Display Order

● Variable list

○ Alphabetic

○ Ascending means

○ Descending means

Continue Cancel Help

Frequencies: Statistics

Percentile Values

☑ Quartiles

☐ Cut points for: 10 equal groups

☐ Percentile(s):

Add Change Remove

Central Tendency

☑ Mean

☑ Median

☑ Mode

☐ Sum

☐ Values are group midpoints

Dispersion

☑ Std. deviation ☐ Minimum

☑ Variance ☐ Maximum

☐ Range ☐ S.E. mean

Distribution

☐ Skewness

☐ Kurtosis

Continue Cancel Help

Figure 9.4 *SPSS Descriptives and Frequencies/Statistics commands*

Notice that in either command we have many options from which to select when choosing descriptive statistics to be generated. SPSS does not discriminate between levels of measurement and will calculate anything we ask for. We need to be careful to select only the measures that are appropriate to the data we are analyzing and the question we want to answer. If we were analyzing the sex of respondents, for example, we would not select the mean or median option for measures of central tendency; SPSS will calculate them but the numbers are meaningless for nominal data. It is up to us to choose only the appropriate measures so that the output is not cluttered with unnecessary statistics.

The best option for generating descriptive statistics for interval/ratio data is the **Explore** command. To generate statistics using the **Explore** command we follow the procedure in Table 9.10, which will produce the output in Figure 9.5.

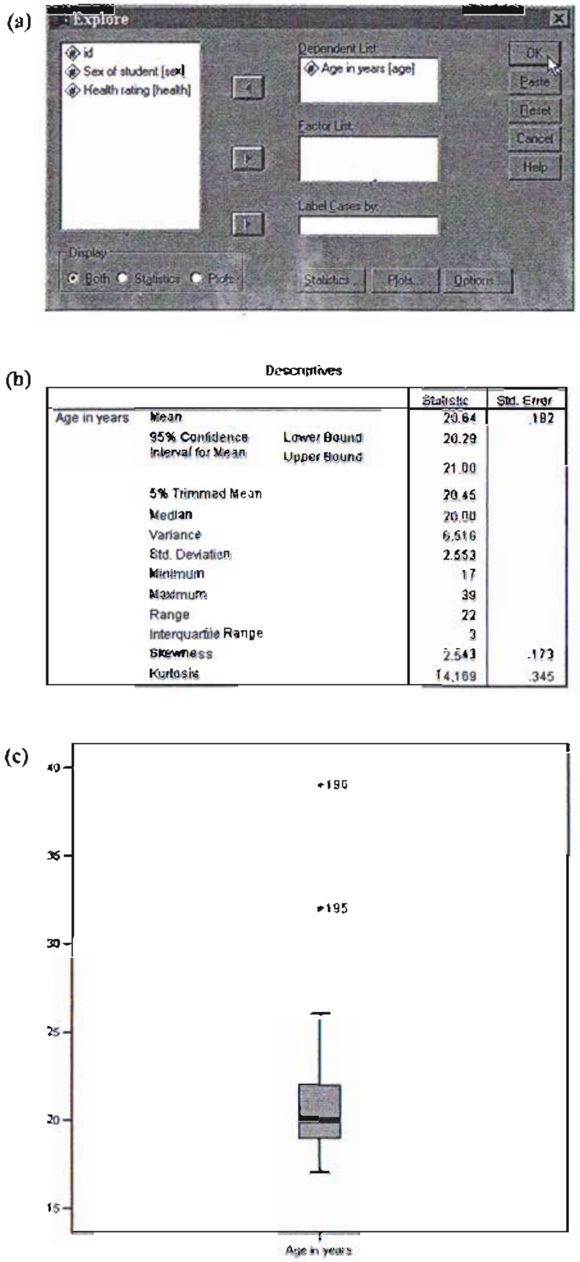
Table 9.10 The SPSS Explore command (*file*: Ch09.sav)

SPSS command/action	Comments
1 From the menu select **Analyze/Descriptive Statistics/Explore**	This brings up the **Explore** dialog box
2 Select **Age in years** from the source list of variables	
3 Click on ▸ pointing to the target list headed **Dependent** List:	This pastes **Age in years** into the target list headed **Dependent List:**
4 Click on **OK**	

The **Explore** command produces a **Descriptives** table that provides a number of statistics:

- the mean and the median (the mode is not usually a useful measure for interval/ratio scales as we discussed above);
- a more refined measures of central tendency called the **5% Trimmed Mean**. The trimmed mean is the arithmetic mean calculated when the largest 5% and the smallest 5% of the cases have been eliminated. Eliminating extreme cases from the computation of the mean results in a better estimate of central tendency when the data are skewed;
- other descriptive statistics (which we will cover in the next chapter) that measure dispersion;
- the 95% confidence interval around the mean so that we can make inferences, as we will discuss in Chapter 17;
- measures that help us assess the shape of the distribution, called measures of skewness and kurtosis, to which we will refer in Chapter 11.

The **Explore** command also creates a stem-and-leaf plot, which is not presented here; such plots were useful ways of tallying a distribution before personal computers, but with the advent of programs such as SPSS, stem-and-leaf plots are largely redundant.


(a)

Explore

id

Sex of student [sex]

Health rating [health]

Dependent List: Age in years [age]

Factor List:

Label Cases by:

Display: ● Both ○ Statistics ○ Plots

Statistics... Plots... Options...

OK Paste Reset Cancel Help

(b)

Descriptives

			Statistic	Std. Error
Age in years	Mean		20.64	.182
	95% Confidence Interval for Mean	Lower Bound	20.29	
		Upper Bound	21.00	
	5% Trimmed Mean		20.45	
	Median		20.00	
	Variance		6.516	
	Std. Deviation		2.553	
	Minimum		17	
	Maximum		39	
	Range		22	
	Interquartile Range		3	
	Skewness		2.543	.173
	Kurtosis		14.169	.345

(c)

Figure 9.5 (a) the SPSS Explore command (b) Descriptives output and (c) box-plot (stem-and-leaf plot omitted)

Following the stem-and-leaf plot is a box plot, Figure 9.5 (c), which graphically presents the statistics in the **Descriptives** box. The key elements of the box plot are:

- the heavy line in the middle of the box is the value for the median;
- the bottom edge of the box is the upper limit of the first quartile, and the upper edge of the box is the upper limit of the third quartile. The difference between the two, which is the height of the box, is thus the interquartile range;
- the 'whiskers' of the plot represent the range of values in which scores that do not represent extreme scores or *outliers* lie;
- two extreme scores are identified by * and by the row number in the data file in which they can be located. SPSS indicates any case that is more than 3 box lengths from the upper or lower quartiles as an extreme score. Any score that is between 1.5 and 3 box lengths from the upper and lower quartiles would be labelled by SPSS as 'outliers' and indicated in the plot by a O. Here there are no such 'outliers'. Note that the terminology that SPSS uses may be different from other definitions of what constitutes an 'outlier' you may come across, since there is no agreed upon criteria for designating scores as outliers. My preference is to call all scores that are disconnected from the main batch as outliers, rather than breaking them up into outliers and extreme scores.

Measures of central tendency using SPSS: Bivariate and multivariate analysis

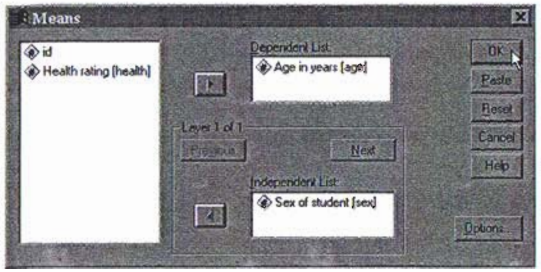
This chapter has discussed measures of central tendency largely in the univariate context. For example, we calculated mean age for *all* the students in the group for which we have data. However, this can be easily extended to the bivariate and multivariate contexts by simply calculating the relevant measures for each of the groups defined by the independent variable. Thus if I wanted to see whether male and female students were on average different in age (i.e. whether age of student was dependent on their sex), I would break the data up into male and female groups and then calculate measures of central tendency for each so that I can compare them.

In SPSS this can be done through a number of commands.

1. The **Analyze/Descriptive Statistics/Descriptives** and the **Analyze/Descriptive Statistics /Frequencies/Statistics** commands. These commands do not themselves provide the ability to break a data set up into comparison groups, but we can invoke the **Data/Split File** command prior to running these. The **Data/Split File** command is an especially useful function in SPSS, since it can extend any procedures that do not allow for the breaking up of data into comparison groups to the bivariate context. Once the **Data/Split File** command is used, we can specify an independent variable that will create the comparison groups, and then all subsequent commands will be performed on each of these groups. Thus by pasting Sex of students into the **Groups Based on** box, and then running the **Descriptives** command for age, we can get the mean age for males and for females separately so that we can compare them.

2. A better option, for the same reasons that we discussed in the previous section, is to use the **Analyze/Descriptive Statistics/Explore** command and paste the independent variable into the **Factor List:**. The summary statistics and plots that we generated in the previous section will now be produced for each of the groups defined by the variable(s) in this list.

3. The **Analyze/Compare Means/Means** command (Figure 9.6). If we paste **Age in years** into the **Dependent List:** and **Sex of student** into the **Independent List** the default setting is for SPSS to provide the mean age, the number of students, and the standard deviation for males, for females, and for the whole data set.



Report

Age in years

Sex of student	Mean	N	Std. Deviation
Female	21.28	90	3.065
Male	20.12	105	1.885
Total	20.66	195	2.559

Figure 9.6 The SPSS **Compare Means/Means** command with layer variable and output

There are two points to note about the **Compare Means/Means** command:

1. We can compare groups in terms of more than just their respective means. If we select **Options** for this command we can add to the list of statistics that can be generated, so that, contrary to the name of this command, we can compare more than just the means.

2. The **Compare Means** command can be extended to include more than two variables in the analysis (unlike the **Explore** command). The variables upon which we split the total data set are called **layer** variables, and we can have a number of layers. Thus I may want the data set to be first broken down by responses to **Health rating** and then by the **Sex of student** so that I can compare for Very healthy students whether there is a difference between females and males in terms of age. Here the first layer is **Health rating** (the highest level of division) and the second layer is **Sex of student** (which is the second order division which compares groups within each category of the first layer). This layering of variables is illustrated in Figure 9.7.

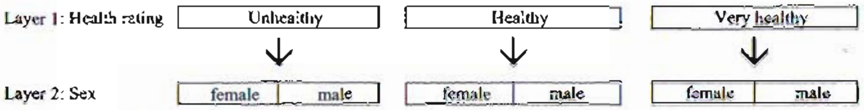


Figure 9.7 The logic of layer variables in SPSS

Summary

In this chapter we have worked through a number of ways of summarizing data so that we can identify the center of their distribution. Rather than rely on a simple visual inspection of a graph or frequency table to determine the central value for a set of scores, we can alternatively (or in addition) use an appropriate measure of central tendency. We have also seen, however, that each of these measures have their own peculiarities that affect their respective use, and that they **do not** always arrive at the same conclusion as to the where the center of a distribution lies.

Exercises

9.1 Can we calculate the mean for ordinal data? Why or why not?

9.2 What do the symbols μ and $\overline{X}$ represent?

9.3 In a set of eight scores the mean is 5. If seven of these scores are 9, 3, 4, 5, 6, 4, 7 what must the remaining score be?

9.4 Calculate the mean and median for each of the following distributions:

(a) 5	9	13	15	26	72			
(b) 121	134	145	212	289	306	367	380	453
(c) 1.2	1.4	1.9	2.0	2.4	3.5	3.9	4.3	5.2

9.5 A student switched from one class to another. This student's 'friends' commented that such a move raised the average IQ of each class. What does this comment suggest about the relationship of this student's IQ to the average in each class?

9.6 Consider the following data set:

43, 22, 56, 39, 59, 73, 60, 75, 80, 11, 36, 66, 45, 57, 20, 36, 68, 87, 50, 68, 9.

(a) Rank-order these values and determine the median.
(b) Calculate the mean.
(c) By comparing the value for the mean and the median, determine whether the distribution is symmetric, skewed to the left, or skewed to the right.
(d) If a score of 194 is added to this data set, how will it affect the median and the mean? Explain the changes to the previous calculation for these measures.

9.7 Calculate the mean, median, and mode for the following data regarding the annual income (in $'000) for people employed in a particular agency:

12	40	22	30	18	36	45	19	22	22	16	23
37	35	72	28	36	29	42	56	52	35	37	26
22	29	35	62								

9.8 The following data represent time, in minutes, taken for subjects in a fitness trial to complete a certain exercise task.

31	39	45	26	23	56	45	80	35	37	27	37
25	42	32	58	80	71	19	16	56	21	40	82
34	36	10	38	12	48	38	37	39	42	56	28
27	39	17	31								

In Exercise 5.5 you were asked to generate a frequency table by grouping these data into class intervals of 1–9, 10–19, 20–29, etc.

(a) Calculate the mean and median, using both the raw data and the grouped data. Are these values different from your calculations for the ungrouped data? Explain.
(b) If you created an SPSS data file for these data in Exercise 2.2 use SPSS to generate the relevant descriptive statistics for this variable.

9.9 Consider the following data sets:

Course of enrollment

Course	Frequency
Social science	32
Arts	45
Economics	21
Law	13
Other	8

Time spent studying for exams

Time	Frequency
1 hour	12
2 hours	25
3 hours	27
4 hours	30
5 hours	26

Satisfaction with employment

Satisfaction	Frequency
Very dissatisfied	12
Not satisfied	25
Satisfied	92
Very satisfied	38

For each of the data sets:

(a) Indicate the level of measurement.
(b) Calculate all possible measures of central tendency. Explain any differences between the measures and discuss which is most appropriate.

9.10 Consider the following data from a survey of employees of a factory:

School years completed

Years	Number of employees
1–4	127
5–8	500
9–12	784
13–16	59
17–20	8

(a) Calculate the mean, median, and mode of this distribution.
(b) If they differ, explain why.

9.11 Is 2100 the mode for the following distribution?

Migrants in local area, place of origin

Place	Number
Asia	900
Africa	1200
Europe	2100
South America	1500
Other	300
Total	6000

9.12 Using the **Employee data** file that comes with the SPSS program, calculate the appropriate descriptive statistics that will allow you to answer the following questions.

What is the difference between mean starting salary and mean current salary?

10

Measures of dispersion

We have seen that there are various ways by which the average of a distribution can be conceptualized and calculated. But how average is average? Consider the two distributions of cases according to annual income shown in Table 10.1.

Table 10.1 Annual incomes

Group A ($)	Group B ($)
5000	20,000
6500	28,500
8000	35,000
55,000	36,000
85,000	40,000

The mean income for each of these groups is the same:

$$\overline{X}_A = \frac{5000+6500+8000+55,000+85,000}{5} = \$31,900$$

$$\overline{X}_B = \frac{20,000+28,500+35,000+36,000+40,000}{5} = \$31,900$$

These distributions have the same mean, yet it is clear that there is also a major difference between the two. Although the mean is the same, the spread or **dispersion** of scores is very different.

Measures of dispersion are descriptive statistics that indicate the spread or variety of scores in a distribution.

We will begin with measures of dispersion for interval/ratio data: the range, interquartile range, standard deviation, and coefficient of relative variation. We will then explore a measure of dispersion for categorical data: the index of qualitative variation.

The range

The simplest measure of dispersion is the **range**.

The **range** is the difference between the lowest score and highest score in a distribution.

This is a quickly and easily calculated measure of dispersion, because it involves a straightforward subtraction of one score from another. Thus for the two distributions of income the ranges in Table 10.1 will be:

$$R_A = 85,000 - 5000 = \$80,000$$

$$R_B = 40,000 - 20,000 = \$20,000$$

We can immediately see that even though the two distributions have the same mean, there is considerable difference in the *spread* of scores around this average; Group A has much more variation.

The advantage of the range as a measure of dispersion is that it is very easily calculated, since it is simply the subtraction of one number from another. However, this advantage of the range is also its major limitation: it only uses the extreme scores, and therefore changes with the values of the two extreme scores. Consider the distribution of income for group B: all the cases fall in a $20,000 range between $20,000 and $40,000. If we add a sixth person to this group, whose annual income is $150,000, the range is suddenly stretched out by this one score. It is now $130,000. To compensate for the effect of such outliers, a slight variation on the range, called the **interquartile** range, can be generated.

The interquartile range

The **interquartile range** (IQR) overcomes the problems that can arise with the simple range by ignoring the extreme scores of a distribution. The IQR is the range for the middle 50 percent of cases in a rank-ordered series (Figure 10.1).

The **interquartile range** is the difference between the upper limits of the first quartile and the third quartile.

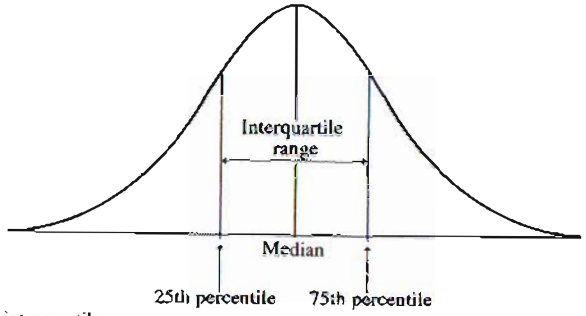


Figure 10.1 The interquartile range

To see how the IQR is calculated we will use the age data from the 20 survey respondents. There are 20 cases, so each quartile will consist of 20 ÷ 4 = 5 cases. The first quartile ends with a person who is 18 years of age. The third quartile ends with a person (the 15th) who is 20 years of age (Figure 10.2).

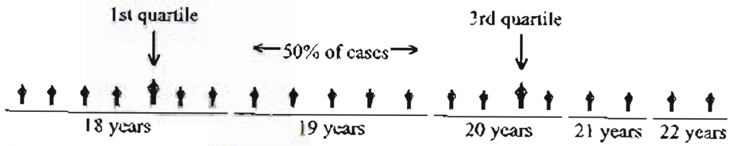


Figure 10.2 The interquartile range

The interquartile range is 2 years:

$$IQR = 20 - 18 = 2 \text{ years}$$

Unlike the simple range, the interquartile range will not change dramatically if we add one or two people who are much older or much younger to either end of the distribution.

The standard deviation

Many readers will have had the experience of dining out with a large group of people where one or two people proceed to order expensive meals and lots of drinks, and when the bill arrives these same people suggest dividing it up evenly to make the calculation of everyone's share easier! Everyone at the dinner table will be aware of the difference between the value of their own dinner and the cost of the 'average' meal so that they can gauge whether paying the average will put them ahead or behind. In this situation everyone is aware of the difference between average and spread, and how the mean may be a misleading representation of a distribution when taken just on its own.

In a similar manner the standard deviation tries to capture the average distance each score is from the average. The standard deviation assesses spread by employing in its calculation the difference between each score and the mean. As with the calculation for the mean, the formulas we use vary slightly depending on whether we have the data in listed form or in grouped form. In either case we use the Roman symbol, *s*, to symbolize the standard deviation for a sample, and the Greek letter, *σ*, for the standard deviation for a population. With listed data the standard deviation for the sample and population are respectively given by:

$$s = \sqrt{\frac{\Sigma\left(X_i - \bar{X}\right)^2}{n-1}} \quad \text{(sample)}, \quad \sigma = \sqrt{\frac{\Sigma\left(X_i - \mu\right)^2}{N}} \quad \text{(population)}$$

A close look at each of these formulas indicates how they capture the notion that the standard deviation is the average distance that each score is from the average. The numerator is simply the difference between each score and the mean, and the denominator adjusts those differences by the number of observations. The formulas are slightly more complicated, since the differences are squared and the square root of the whole lot taken (for reasons that are not necessary to the present discussion), but the basic idea is still evident.

To focus on the notion of the standard deviation more sharply, consider again the distribution of ages for our 20 survey respondents. We have already calculated the mean age to be 19.3 years. All the scores deviate from the mean, either above or below it, to a greater or lesser degree. This is illustrated in Figure 10.3.

The age of each person is plotted on a graph, with the line for the mean age running down the middle. The distance from the mean to each person's age is then drawn in. Respondents 7 and 13 are relatively a long way above the mean, while respondents 5, 8, 12, 14, and 18 are only slightly below the mean. What is the average of these distances?

Unfortunately, we cannot simply add all the **positive deviations** (scores above the mean) with all the **negative deviations** (scores below the mean), since by definition, these will sum to zero. This is why the equation for the standard deviation squares the differences: it thereby turns all the deviations into positive numbers, so that the larger the differences, the greater the value of the standard deviation.

Let us actually calculate the standard deviation for this distribution. We can use the equation above to do this, but we have only introduced it because it captures the idea that the standard deviation is the average distance from the mean. In actually calculating the standard deviation for listed data we work with a slightly different equation that is easier to compute, but which will always give us the same answer as the equation above:

$$s = \sqrt{\frac{\Sigma X_i^2 - \frac{\left(\Sigma X_i\right)^2}{n}}{n-1}}$$

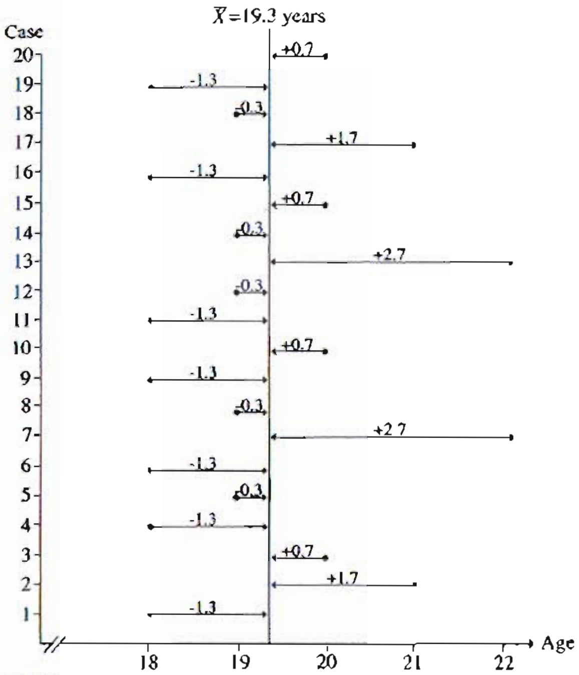


Figure 10.3 Deviations of scores around the mean

Table 10.2 Calculations for the standard deviation of age

Case	Age in years, X_i	X_i^2
1	18	324
2	21	441
3	20	400
4	18	324
5	19	361
6	18	324
7	22	484
8	19	361
9	18	324
10	20	400
11	18	324
12	19	361
13	22	484
14	19	361
15	20	400
16	18	324
17	21	441
18	19	361
19	18	324
20	20	400
Total	$\Sigma X_i = 387$	$\Sigma X_i^2 = 7523$

The term ΣX_i^2 reads 'the sum of all the squared scores', while the term $(\Sigma X_i)^2$ reads 'the sum of all the scores squared'. For the first term we square all the scores and then add them, while the second term reverses the procedure: we add up the scores and then square the sum. Table 10.2 goes through these steps.

Substituting the numbers from Table 10.2 into the equation for the standard deviation, we get 1.35 years:

$$s = \sqrt{\frac{\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}}{n-1}} = \sqrt{\frac{7523 - \frac{(387)^2}{20}}{20-1}} = 1.35 \text{ years}$$

In Table 10.2 we listed each respondent's age separately. However, we may not have the individual values for each case, but rather have data grouped in a frequency table. With data organized in a frequency table we use the following formula to compute the standard deviation for a sample:

$$s = \sqrt{\frac{\Sigma f X_i^2 - \frac{(\Sigma f X_i)^2}{n}}{n-1}}$$

In other words, we multiply each value by the frequency with which it appears in a distribution. Thus if we have the data for age arranged in a frequency table, rather than as a complete list of all ages, the computations will be as shown in Table 10.3. We obtain the same answer of 1.35 years as when we listed each case separately.

Table 10.3 Calculations for the standard deviation of frequency data

Age	Frequency (f)	X_i^2	fX_i^2	fX_i
18	7	18×18=324	7×324=2268	7×18=126
19	5	361	1805	95
20	4	400	1600	80
21	2	441	882	42
22	2	484	968	44
Total	n=20		$\Sigma f X_i^2$=7523	$(\Sigma f X_i)$=387

$$s = \sqrt{\frac{\Sigma f X_i^2 - \frac{(\Sigma f X_i)^2}{n}}{n-1}} = \sqrt{\frac{7523 - \frac{(387)^2}{20}}{20-1}} = 1.35 \text{ years}$$

Before moving on to other measures of dispersion, we should note that, as with the mean (which is part of the calculation), the standard deviation is not an appropriate measure of dispersion for data that are heavily skewed.

Coefficient of relative variation

The standard deviation has some limitations that are overcome by the **coefficient of relative variation** (CRV). The coefficient of relative variation is used:

- for comparing distributions measured in the same units but which have very different means, and
- for comparing distributions measured with different units.

There is no absolute way of saying, in the previous example, whether 1.35 years is a large or small amount of dispersion around the mean. Moreover, the standard deviation for one set of observations cannot be compared with that for another set of scores in order to decide which distribution is the more disperse. For example, two distributions may have standard deviations of 1.35 years, but if the means of each are 5 years and 50 years respectively, it is clear that a standard deviation of 1.35 represents *relatively* more variation for the distribution with the smaller mean.

In other instances we may wish to compare the variation for two separate variables each measured in different units. For example, we might be interested in whether the age of a group of respondents, which has a standard deviation of 1.35 years, displays more variation than their weekly income, which has a standard deviation of $65. We cannot compare these two standard deviations and say one variable is more disperse than the other, because each is measured with different units. We are effectively comparing apples with oranges.

To provide a *standardized* measure of dispersion, we calculate the coefficient of relative variation that expresses the standard deviation as a percentage of the mean:

$$CRV = \frac{s}{\bar{X}} \times 100$$

Using this formula with the distribution of ages for our 20 survey respondents we get:

$$CRV = \frac{s}{\bar{X}} \times 100 = \frac{1.35}{19.35} \times 100 = 7\%$$

If we had another group of people and their ages we can then calculate the CRV for that group and compare it with this one to see which has the greatest amount of dispersion. Thus if I found that a second group of respondents had a standard deviation for their ages of 5 years, and a mean of 21 years, the CRV will be:

$$CRV = \frac{s}{\bar{X}} \times 100 = \frac{5}{21} \times 100 = 24\%$$

This second set of people display more variation in their ages than the first. In fact we can actually say that they exhibit 17 percent more variation.

Index of qualitative variation

The measures of dispersion we have just considered all apply to the highest level of measurement of interval/ratio, since they require us to measure the distances between scores. Scales of measurement, as we know from Chapter 1, do not always permit these operations. How can we express variation in a distribution where the data are only categorical? A measure of dispersion is available for such data: the **index of qualitative variation** (IQV).

The index of qualitative variation is the number of differences between scores in a distribution expressed as a proportion of the total number of possible differences.

The IQV allows us to measure the amount of variation contained in a distribution, even where we only have nominal data. For example, in our earlier example we had a nominal variable, sex of respondents, whose variation cannot be captured by any of the measures of dispersion we have previously looked at (Table 10.4).

Table 10.4 Sex of respondents

Sex	Frequency
Male	12
Female	8
Total	20

The IQV locates the actual amount of variation contained in our data as falling somewhere between two possible extremes. One extreme possibility is if there is no variation in the data. This occurs when all the cases fall into the same category; in this example, if all the cases were either male or female. By definition, if all the cases have the same score for a variable, there is no variation. This then constitutes the *minimum* amount of variation that it is possible to observe.

The *maximum* amount of variation that we could possibly observe in a distribution is if the cases are evenly distributed across the categories of the variable, as would be the case if we obtained the results in Table 10.5.

Table 10.5 Sex of respondents: maximum possible variation

Sex	Frequency
Male	10
Female	10
Total	20

In this distribution we have 100 differences: each of the 10 females is different to each of the 10 males in terms of their sex.

There is a simple method for calculating the maximum possible number of differences that can be observed for any set of categorical data, using the following formula, where *k* is the number of categories.

$$\text{maximum possible differences} = \frac{n^2(k-1)}{2k}$$

We can use this formula to arrive at the maximum number of differences for the number of cases and categories in our example for respondents' sex:

$$\text{maximum possible differences} = \frac{n^2(k-1)}{2k} = \frac{20^2(2-1)}{2(2)} = 100$$

If we look at the actual distribution of responses in Table 10.4 it is evident that it more closely resembles the extreme of maximum variation (Table 10.5) than the situation of no variation. The IQV allows us to express this quantitatively. To do this we need to determine the number of observed differences in the distribution of scores we are analyzing. Take one of the 8 females. How many other people in the distribution are they different to in terms of sex? Clearly these are the 12 males in the distribution. For each of the 8 females there will be 12 other people in the distribution from whom they are different, producing a total of 96 observed differences.

The IQV for the sex of respondents will therefore be:

$$IQV = \frac{\text{observed differences}}{\text{maximum possible differences}} = \frac{96}{100} = 0.96$$

An IQV of 0.96 indicates that we have a very high amount of variation in the data for this variable. If, on the other hand, we did have all females or all males, so that there are no observed differences in the data, it is relatively easy to see that the IQV will equal 0, indicating no variation.

Let us now calculate the amount of variation, using this measure, evident in the distribution of responses according to health rating (Table 10.6).

Table 10.6 Health rating of respondents

Health rating	Frequency
Unhealthy	7
Healthy	5
Very healthy	8
Total	20

How many times do cases in this distribution differ from other cases? Starting with the 8 very healthy people, each of these are different in their health rating to the 5 healthy and 7 unhealthy people, producing 96 differences. To this can be added the 35 differences between the 5 healthy people and the 7 unhealthy people. The total number of observed differences is:

$$\text{observed differences} = (8 \times 5) + (8 \times 7) + (5 \times 7) = 131$$

The maximum number of differences we could observe (if the cases were evenly spread across the three categories) is:

$$\text{maximum possible differences} = \frac{n^2(k-1)}{2k} = \frac{20^2(3-1)}{2(3)} = 133.3$$

The IQV will therefore be:

$$IQV = \frac{\text{observed differences}}{\text{maximum possible differences}} = \frac{131}{133.3} = 0.98$$

This indicates that there is almost the maximum possible variation between these cases in terms of their health ratings. We can also say that there is about the same amount of variation among these cases in terms of their health rating as there is in terms of their sex.

Example

To see how all these measures, and those we discussed in the previous chapter for central tendency, apply in a given instance, let us go back to the data we introduced in previous chapters for the weekly income of 20 people in a sample:

$0, $0, $250, $300, $360, $375, $400, $400, $400, $420, $425, $450, $462, $470, $475, $502, $520, $560, $700, $1020

Notice that we have the data individually listed so that we will use the appropriate formulas, where relevant. Notice also that the data are interval/ratio, which opens up a wide choice in selecting measures of central tendency and measures of dispersion.

Starting with measures of central tendency, we begin with the mode. We can see without too much effort that the value that occurs the most is $400:

$$M_o = \$400$$

The data are also rank-ordered, from lowest to highest, so we can also calculate the median with relative ease. With 20 cases to work with (an even number) the median will be the average of the two middle scores; that is, the average of the incomes for the 10th and 11th people in line. These scores are $420 and $425, giving a median of $422.50:

$$M_d = \frac{420 + 425}{2} = \$422.50$$

The mean for this set of data is \$424:

$$\bar{X} = \frac{X_i}{n} = \frac{8489}{20} = \$424$$

We can see that the mean is only slightly higher than the median, which is higher than the mode, indicating that the data are skewed slightly to the right. This is obviously due to the one very high income earner who receives a weekly income of \$1020 (not an uncommon feature of income distribution data). However, the fact that these differences are not too large indicates that the distribution is only slightly skewed.

We will now calculate the measures of dispersion appropriate to this set of data to see the extent to which this average is a fair representation of the distribution. The range is the largest score (\$1020) minus the lowest score (\$0):

$$R = 1020 - 0 = \$1020$$

The one high score of \$1020, though, renders the simple range inaccurate as a measure of dispersion, so we will calculate the interquartile range as well. The first quartile ends with the income for the 5th person in the rank order (\$360), and the third quartile ends with the income for the 15th person in the rank order (\$475):

$$IQR = 475 - 360 = \$115$$

We can see that this is a 'truer' reflection of the spread of scores around the mean, which even an 'eyeball' inspection of the listed data tells us is not very large.

We will now calculate the standard deviation. To help calculate the relevant numbers to put into the equation I construct the following table (Table 10.7):

Table 10.7 Calculations for the standard deviation of income

Case	Income (\$), X_i	X_i^2
1	0	0
2	0	0
3	250	62,500
4	300	90,000
5	360	129,600
6	375	140,625
7	400	160,000
8	400	160,000
9	400	160,000
10	420	176,400
11	425	180,625
12	450	202,500
13	462	213,444
14	470	220,900
15	475	225,625
16	502	252,004
17	520	270,400
18	560	313,600
19	700	490,000
20	1020	1,040,400
Total	$\Sigma X_i = 8489$	$\Sigma X_i^2 = 4{,}488{,}623$

$$s = \sqrt{\frac{\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}}{n-1}} = \sqrt{\frac{4{,}488{,}623 - \frac{(8489)^2}{20}}{20-1}} = \$216$$

The standard deviation falls somewhere between the range and the interquartile range. It does not completely ignore the extreme cases, such as \$1020, which the IQR leaves aside, but it also does not give them as great a weight in the measurement of dispersion, as is the case with the simple range.

Assume that I am now presented with another set of cases that have a mean income of \$510 and a standard deviation of \$300. Which of these two distributions displays the greatest variation? It might be tempting to compare the standard deviations, but we know this is not an appropriate comparison given the differences in the means around which the scores deviate. Instead we need to calculate the CRV for each set of scores. For the first set of data the CRV will be:

$$CRV = \frac{s}{\bar{X}} \times 100 = \frac{216}{424.45} \times 100 = 51$$

For the second set of scores the CRV will be:

$$CRV = \frac{s}{\bar{X}} \times 100 = \frac{300}{510} \times 100 = 59$$

Thus I can say that the second distribution possesses 8 percent more variation in incomes than the first set of cases. It not only has a higher average, but is relatively more dispersed.

Measures of dispersion using SPSS

Measures of dispersion can be generated through the same commands that we discussed in the previous chapter for generating measures of central tendency (see pages 129–33). A summary of these commands is presented here:

1. The **Analyze/Descriptive Statistics/Descriptives** command, for univariate analysis, or for bivariate analysis using the **Data/Split File** command.
2. The **Analyze/Descriptive Statistics/Frequencies/Statistics** command, for univariate analysis, or for bivariate analysis using the **Data/Split File** command.
3. The **Analyze/Descriptive Statistics/Explore** command, for univariate analysis, or for bivariate analysis by placing the grouping variable into the **Factor List:**.
4. The **Analyze/Compare Means/Means** command, for bivariate analysis by using only one layer of variables, or for multivariate analysis by using more than one layer of variables. Despite its misleading name, this very useful command provides more descriptive statistics than just the mean under the **Options** sub-command.

None of these commands unfortunately provide the CRV or IQV, and only the **Analyze/Descriptive Statistics/Explore** command provides the interquartile range.

Summary

In this and the previous chapter we have worked through a number of ways of summarizing data and displaying a distribution. Many formulas and rules have been encountered and the options may seem a little overwhelming. Fortunately, computers have made life easy for us, and all the measures we have introduced, as we have seen, can be generated with the click of a button. However, life should not get too easy. There is a level of understanding that can only be obtained by working through the hand calculations, especially an understanding of the limits to many of the techniques we have introduced.

Exercises

10.1 What are the advantages and disadvantages of the range as a measure of dispersion?

10.2 What do the symbols s and σ represent?

10.3 Calculate the range and standard deviation for each of the following distributions:

(a) 5, 9, 13, 15, 26, 72
(b) 121, 134, 145, 212, 289, 306, 367, 380, 453
(c) 1.2, 1.4, 1.9, 2.0, 2.4, 3.5, 3.9, 4.3, 5.2

10.4 Consider the following data regarding the annual income (in $'000) for people employed in a particular agency:

12	40	22	30	18	36	45	19	22	22
37	35	72	28	36	29	42	56	52	35
16	23	37	26	22	29	35	62		

Calculate the range, interquartile range, and standard deviation for these data.

10.5 Using the **Employee data** file that comes with the SPSS program, calculate the appropriate descriptive statistics that will allow you to answer the following questions.

(a) Which of the two variables, mean starting salary and mean current salary, displays the most variation?

(b) What is the interquartile range for the amount of previous experience of employees, expressed in years?

11

The normal curve

In Chapter 4 we discussed the way that we can generate a frequency table to show the distribution of scores for a particular variable. In Chapters 9 and 10 we discussed how we can enter the frequency distribution of scores into specific equations that can give us precise numerical measures of the center and spread of that distribution. It is possible, though, to work 'backwards'; if we know the standard deviation and mean for a particular distribution, we can work out the frequency distribution from which these measures were calculated. To derive the frequency distribution of scores from these numerical measures, we need to assume that the distribution resembles a **normal curve**. The term 'normal' is not meant to signify 'usual' or 'common'. In fact, it might seem like a very artificial construct that is anything but normal. However, it does play a central role in statistical analysis and is the basis for many of the procedures that follow in later chapters. So while the reasons for studying this particular curve (among the multitude on which we can focus) may not be immediately apparent, hopefully they will become evident later.

The normal distribution

This chapter will try to 'circle in' on the nature of the normal distribution. We will begin with a very simple and approximate definition, gradually expanding on this definition as we become more familiar with it.

The **normal curve** is a smooth, unimodal curve that is perfectly symmetrical. It has 68.3 percent of the area under the curve within one standard deviation of the mean.

These features of the normal curve are illustrated in Figure 11.1. We can use the properties of the normal curve illustrated in Figure 11.1 to derive specific conclusions about a frequency distribution of scores that we think is normally distributed.

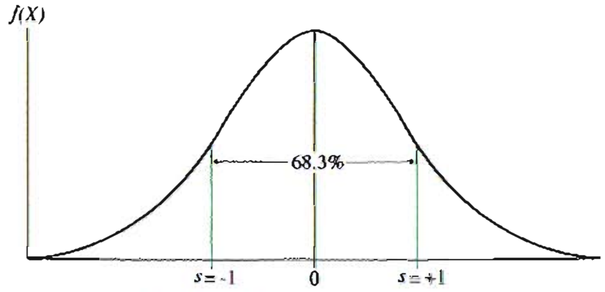


Figure 11.1 Areas under the standard normal curve

For example, we might be interested in people's ages and have the following descriptive statistics for the average and spread of ages for a population of 1200 people:

$$\mu = 35 \text{ years}, \sigma = 13 \text{ years}$$

If age is normally distributed, 68.3 percent of people in this population will fall within 1 standard deviation of the mean. In other words, 820 people (68.3 percent of 1200) will have ages somewhere between 22 years (35 – 13) and 48 years (35 + 13).

This property of the normal curve holds true regardless of the particular values for the standard deviation and mean for the cases we are dealing with. For example, we may have three different groups of 1200 people with ages described by the following statistics (Table 11.1, Figure 11.2).

Table 11.1 Average age and spread for three populations

Group	Mean age (years)	Standard deviation (years)	Age range of middle 68.3% of cases (years)
1	35	13	22 to 48
2	35	7	28 to 42
3	35	20	15 to 56

All three distributions have the same average age, but are different in terms of the *spread* of ages around the mean. If we can assume that they are each normal distributions we can derive the age ranges for the middle 68.3 percent of people in each group. If these are normal distributions, it follows that 820 people in group 1 will have ages between 22 and 48 years; 820 people in group 2 will have ages ranging from 28 to 42 years; and for group 3 the range is 15 to 56 years.

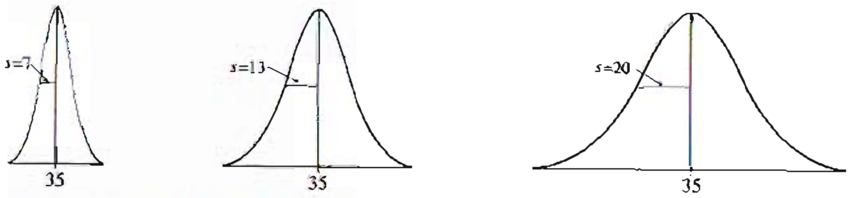


Figure 11.2 Three normal distributions with different standard deviations

This process of stating the spread of cases in terms of the number of standard deviations from the mean is called **standardizing the distribution**, and produces the **standard normal distribution** (Figure 11.3). The standard normal distribution has a mean of zero and a standard deviation of one (by definition, the mean is zero standard deviation units away from itself, and one standard deviation is one standard deviation unit away from the mean).

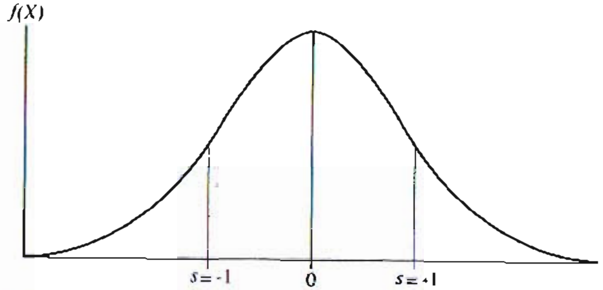


Figure 11.3 The standard normal distribution

This standardization procedure allows us to measure all normal distributions in terms of common units – standard deviations – regardless of the units in which the variable is initially measured. It is analogous to expressing the price of various goods from different countries in terms of a common currency. We may have a whole list of prices, some of which are expressed in US dollars, some in British pounds, others in Euros. But if we convert all the prices into a common unit such as the amount of gold each unit of currency will purchase, a comparison can be made. Similarly, a distribution may be expressed in terms of years, or crime rates, or births per thousand. But expressing these various distributions according to standard deviation units gives a common scale of measurement.

We noted that the normal curve is symmetrical. Since the curve is symmetrical the same percentage of cases that falls within a certain range *above* the mean also falls within the same range *below* the mean (Figure 11.4). In other words, if 68.3 percent of all cases fall within one standard deviation unit *either side* of the mean, half of this (34.15 percent) will fall *above* the mean, and the other half (34.15 percent) *below* the mean. For group 1 in Table 11.1, this will imply that 410 people will be between 22 and 35 years of age, and another 410 will be between 35 and 48 years of age.

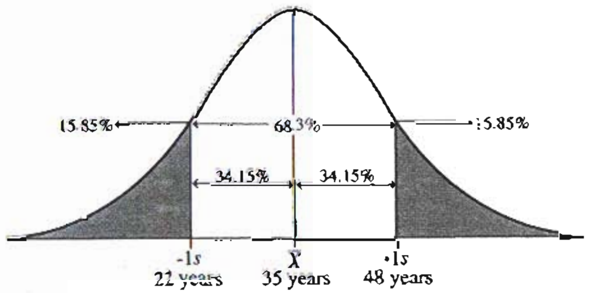


Figure 11.4 Distribution of age for group 1

The other thing to notice about the spread of cases under a normal curve in Figure 11.4 is that the percentage of cases falling *further* than one standard deviation from the mean is equal to the total number of cases (100%) minus the percentage that fall within the range (68.3 percent):

$$100 - 68.3 = 31.7\%$$

Again we can divide this region in two so that 15.85 percent of cases have ages above one standard deviation from the mean (i.e. for group 1 this is older than 48 years), and another 15.85 percent of cases are at the other end (or tail) of the curve. Thus if a woman from this group informs me that she is 52 years of age I will also know that she is in the oldest 16 percent of the population.

This simple exercise hopefully illustrates the usefulness of the normal curve. If we know, or can assume, that a distribution is normal, and we know its mean and standard deviation, we can then make a conclusion about the frequency distribution that underlies these measures. This makes the use of the normal curve important for two reasons, as follows.

1. *The normal curve as an aid to data description.* There are some *empirical* distributions (i.e. they exist in the 'real world') that are *fairly close to being normal*, which allows us to determine that a certain percentage of cases falls a specific distance above and/or below the mean. This is similar to the way in which we apply the equation for the area of the circle. A circle is defined as a shape where every point along the circumference is equidistant from

the center, or, to put it another way, the radius is constant. A figure defined in this way has an area equal to πr^2, but there are very few shapes that we encounter that *exactly* conform to this definition. This does not limit the applicability of the exact formula for a circle because there are many shapes in ordinary life that are *close enough* to a circle (they '**approximate**' a circle) such that using this formula to calculate their areas is not unreasonable. Just as with *figures* that are 'close enough' to being a circle, there are instances when it is not unrealistic to assume that a distribution is 'close enough' to being normal, even though, strictly speaking, it isn't. In other words, just as we never encounter perfect circles, yet still use the formula for the area of a circle in everyday life, we can make statements describing any empirical distribution that (we think) is approximately normal. Sometimes near enough is good enough. Many physical characteristics of people, such as height, are approximately normal. If we took a random sample of people and measured their height we would actually find that about 68 percent of cases fall within one standard deviation of the mean.

2. *The normal curve as a tool for inferential statistics.* The second reason for understanding the properties of a normal curve is that it forms the basis of the procedures that allow us to make inferences from a random sample to a population. The role of the normal curve in inferential statistics will be covered in Part 3, where the convenience of knowing the percentage of cases that fall above and below a certain distance from the mean will become very apparent.

Using normal curves to describe a distribution

The rest of this chapter will employ the normal curve as a descriptive tool, leaving its use as a tool for making an inference from a sample to a population for Part 3. We proceed by expanding slightly the definition of the normal curve, defining the percentage of the total area under the normal curve within two standard deviation units from the mean, and within three standard deviation units (Figure 11.5).

Between ±1 standard deviations from the mean of a normal distribution lies 68.3 percent of the area under the curve.

Between ±2 standard deviations from the mean of a normal distribution lies 95.4 percent of the area under the curve.

Between ±3 standard deviations from the mean of a normal distribution lies 99.7 percent of the area under the curve.

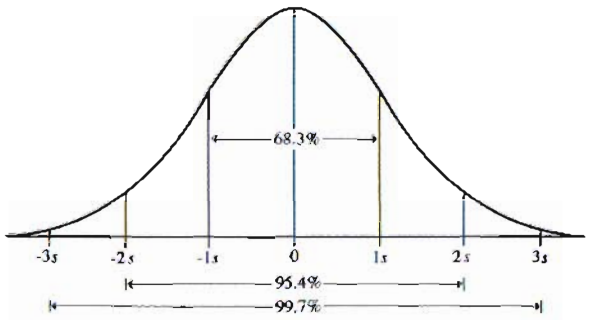
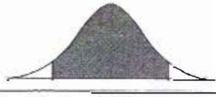
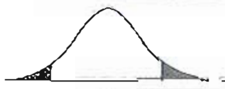
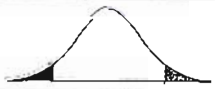


Figure 11.5 Areas under the standard normal curve

This information can be presented in a simple table (Table 11.2).

Table 11.2 Areas under the standard normal curve

Standard deviations from the mean	Area under curve between both points	Area under curve beyond both points (two tails)	Area under curve beyond one point (one tail)
			
±1	0.683	0.317	0.1585
±2	0.954	0.046	0.0230
±3	0.997	0.003	0.0015

There are two aspects to Table 11.2 worth noticing:

- Instead of expressing the area under the curve as a percentage, it is expressed as a proportion: 68.3 percent is converted to 0.683, and so on.
- The values in the first two columns will always sum to one (e.g. 0.683 + 0.317 = 1). This is because the two areas together must equal the total area under the curve.

The normal curve is a very specifically defined polygon, a type of graph we introduced in Chapter 3. This allows us to interpret the proportions in the table as **probabilities**. A probability in this context is simply the chance that any given case will have a certain value, or fall within a certain range of values. For example, assume that someone is chosen at random from group 1 and you have to guess their age. We can use the table to conclude that the probability that this person's age is somewhere between 22 years and 48 years (i.e. it is within one standard deviation either side of the mean) is 0.683, or around 68 in 100. The probability that the person has an age of less than 22 years is 0.1585, or around 16 in 100. This is common sense: there is usually a high probability that someone chosen at random from a set of cases will reflect the average. It is more likely that someone will be 'typical' rather than 'unusual'. This way of interpreting the area under the normal curve as a probability will be especially useful in the following chapters on inference.

z-scores

Instead of using the expression 'number of standard deviations from the mean' we will instead speak of *z*-scores. A *z*-score of +1 indicates one standard deviation above the mean. A *z*-score of −1.5 indicates 1.5 standard deviations below the mean. For a normal population or normal sample, we can work out the *z*-score associated with any actual value using the respective formulas:

$$Z = \frac{X_i - \mu}{\sigma} \text{ (population)}, \quad z = \frac{X_i - \bar{X}}{s} \text{ (sample)}$$

where:

X_i is the actual value measured in original units,
μ is the mean of the population,
σ is the standard deviation of the population,
$\bar{X}$ is the mean of the sample,
s is the standard deviation of the sample.

For example, consider the population of 1200 people in group 1 above, with a mean age of 35 years and standard deviation of 13 years. A member of this group tells me he is 61 years of age. Even before we complicate the matter with equations and *z*-scores, it is fairly clear that this person is much older than the average, so intuitively we can conclude that very few people will be this old or older. In fact, I can, at this point, use a verbal description and say that given the mean and standard deviation for this group only a 'handful' of people will be 61 years of age or more. We can, however, be more precise than this, and actually calculate what this 'handful' of people is as a proportion of the total. To do this I put the information into the formula for calculating *z*-scores for a population:

$$Z = \frac{X_i - \mu}{\sigma} = \frac{61 - 35}{13} = 2$$

This immediately tells me that 61 is two standard deviations above the mean. By referring to the last column in Table 11.2, we conclude that the proportion of people that are 61 years of age or more is only 0.023, or 2.3 percent of the total.

In fact, statisticians have worked out the area under the standard normal curve between the mean and *every* point along the horizontal axis of the normal curve. This information is summarized in a table that appears in the back of every statistics textbook (including this one, see Table A1 in the Appendix). Since we are going to work with the table for areas under the standard normal curve frequently throughout this chapter, and to familiarize ourselves with it, it is reproduced in Table 11.3.

Table 11.3 Areas under the standard normal curve

z	Area under curve between both points	Area under curve beyond both points	Area under curve beyond one point
±0.1	0.080	0.920	0.4600
±0.2	0.159	0.841	0.4205
±0.3	0.236	0.764	0.3820
±0.4	0.311	0.689	0.3445
±0.5	0.383	0.617	0.3085
±0.6	0.451	0.549	0.2745
±0.7	0.516	0.484	0.2420
±0.8	0.576	0.424	0.2120
±0.9	0.632	0.368	0.1840
±1	0.683	0.317	0.1585
±1.1	0.729	0.271	0.1355
±1.2	0.770	0.230	0.1150
±1.3	0.806	0.194	0.0970
±1.4	0.838	0.162	0.0810
±1.5	0.866	0.134	0.0670
±1.6	0.890	0.110	0.0550
±1.645	0.900	0.100	0.0500
±1.7	0.911	0.089	0.0445
±1.8	0.928	0.072	0.0360
±1.9	0.943	0.057	0.0290
±1.96	0.950	0.050	0.0250
±2	0.954	0.046	0.0230
±2.1	0.964	0.036	0.0180
±2.2	0.972	0.028	0.0140
±2.3	0.979	0.021	0.0105
±2.33	0.980	0.020	0.0100
±2.4	0.984	0.016	0.0080
±2.5	0.988	0.012	0.0060
±2.58	0.990	0.010	0.0050
±2.6	0.991	0.009	0.0045
±2.7	0.993	0.007	0.0035
±2.8	0.995	0.005	0.0025
±2.9	0.996	0.004	0.0020
±3	>0.996	<0.004	<0.0020

It may help at this point to reiterate why we bother defining the normal curve in such minute detail. Why have statisticians gone to such lengths as to actually work out and have printed a table that indicates the number of cases that fall within defined regions of a normal distribution? After all, there are an infinite number of possible frequency distributions we could come across – the distribution of families according to total income will be different from the distribution of cities according to crime rates, and neither will be remotely like a normal distribution. Why don't we construct tables that define the areas under these curves? First, there are many empirical distributions that are approximately normal so that this table will provide an aid in describing such distributions, and, second, because there is a distribution at the heart of inferential statistics that is normal, and which we will see in later chapters renders the normal curve exceptionally useful in making an inference from a sample to a population.

The rest of this chapter will work through a series of examples. The objective is to familiarize ourselves with the use of the normal curve as a descriptive tool. In the process, we will also familiarize ourselves with the procedures for looking up values in the area under the standard normal curve table, which will be useful for later chapters.

For example, assume that I have exam grades out of 100 for a sample of 100 students and obtain the following results:

$$\bar{X} = 60, \ s = 10$$

I graph these data on a frequency polygon and observe that the distribution looks approximately normal. Alternatively I can run the **Explore** command in SPSS that we introduced in Chapter 9 and refer to the **measures of skewness and kurtosis** that are printed in the **Descriptives** box:

1. The **measure of skewness** is a measure of the asymmetry of a distribution. The normal distribution is symmetric, and has a skewness value of zero. A distribution with a significant positive skewness has a long right tail. A distribution with a significant negative skewness has a long left tail. As a rough guide, a skewness value more than twice its standard error is taken to indicate a departure from symmetry.

2. The **measure of kurtosis** indicates the extent to which the scores are 'bunched' around the mean to form a 'tall peak' or else spread to form a 'flat hill'. For a normal distribution, the value of the kurtosis statistic is 0. A positive kurtosis value indicates that the observations cluster more and have longer tails than those in the normal distribution and a negative kurtosis value indicates the observations cluster less and have shorter tails.

If either method of assessing normality ('eyeball' inspection of the histogram or measuring skewness and kurtosis) indicates that this group of students is normally distributed (or close to it) according to exam scores I can then proceed to answer various questions about frequency distribution of this variable.

The area between the mean and a point on the distribution

I might want to know how many students are between the mean of 60 and a score of 65, which I consider to be a reasonable range of scores for students to achieve. The first thing to do is convert 65 into a *z*-score:

$$z = \frac{X_i - \bar{X}}{s} = \frac{65 - 60}{10} = 0.5$$

The next step is to refer to the table for the area under the standard normal curve and find the area between this point and the mean. A condensed version of the table is presented in Table 11.4 to show its use. For a z-score of 0.5 we get the result shown.

Table 11.4 Areas under the standard normal curve

z	Area under curve between both points	Area under curve beyond both points	Area under curve beyond one point
±0.1	0.080	0.920	0.4600
±0.2	0.159	0.841	0.4205
±0.3	0.236	0.764	0.3820
±0.4	0.311	0.689	0.3445
±0.5	0.383	0.617	0.3085
±0.6	0.451	0.549	0.2745
±0.7	0.516	0.484	0.2420
±0.8	0.576	0.424	0.2120
±0.9	0.632	0.368	0.1840
±1	0.683	0.317	0.1585
⋮	⋮	⋮	⋮
±3	>0.996	<0.004	<0.0020

In other words, 0.383 of all cases will have a grade of 5 marks *above or below* the mean. Since we are interested in only those students that are 5 marks *above* the mean, we divide 0.383 in half. This is illustrated in Figure 11.6

$$\text{proportion of students with grades between 60 and 65} = \frac{0.383}{2} = 0.1915$$

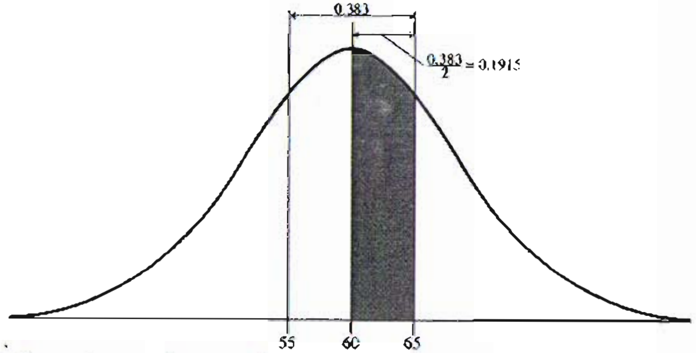


Figure 11.6 The area between the mean and one point

Thus, I can say that just over 0.19 (19 percent) of the students received grades between 60 and 65 (remember that a proportion can be transformed into a percentage by moving the decimal point two places to the right).

The area beyond a point on the distribution

A very similar logic applies to finding the percentage of cases that fall *beyond* a certain point on the distribution. For example, I might be interested in the percentage of students who did exceptionally well, which I regard to be a score over 65.

From the previous exercise we know that the z-score associated with a grade of 65 is:

$$z = \frac{X_i - \bar{X}}{s} = \frac{65-60}{10} = 0.5$$

This time, when referring to the table for the standard normal curve, we refer to the column for the area *beyond* the point defined by a z-score of 0.5. In other words, we are only interested in the area under one tail of the distribution (Table 11.5).

Table 11.5 Areas under the standard normal curve

z	Area under curve between both points	Area under curve beyond both points	Area under curve beyond one point
±0.1	0.080	0.920	0.4600
±0.2	0.159	0.841	0.4205
±0.3	0.236	0.764	0.3820
±0.4	0.311	0.689	0.3445
±0.5	0.383	0.617	0.3085
±0.6	0.451	0.549	0.2745
±0.7	0.516	0.484	0.2420
±0.8	0.576	0.424	0.2120
±0.9	0.632	0.368	0.1840
±1	0.683	0.317	0.1585
⋮	⋮	⋮	⋮
±3	>0.996	<0.004	<0.0020

This indicates that 0.3085 (30.85%) of students scored over 65.

If we look at the answers to these two problems we can see that the percentages sum to 50 (Table 11.6). This is because the two areas we have defined together make up exactly half the curve (Figure 11.7).

Table 11.6 Areas under the curve

Range of exam scores	Percentage of cases (%)
Between 60 and 65	19
65 or over	31
Total	50

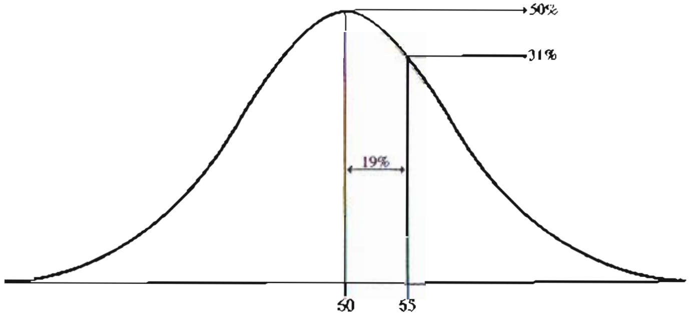


Figure 11.7 Areas under the normal curve

In a similar fashion I may be interested in the proportion of students that have failed the exam. I calculate the z-score associated with a grade of less than 50:

$$z = \frac{X_i - \bar{X}}{s} = \frac{50-60}{10} = -1$$

Looking at Table 11.5 I can see that there is 0.1586 of the curve beyond a z-score of –1, which indicates that nearly 16 percent of students failed.

The area between two points on a normal distribution

Another question in which I might be interested is the percentage of cases that fall within a range not bounded on one side by the mean. For example, I might be interested in the proportion of students that received a credit grade, which is a grade between 65 and 75.

The solution to this puzzle is apparent by looking at Figure 11.8. The area between 65 and 75 is the area left over if we subtract the area between 65 and the mean from the area between 75 and the mean. In other words we need to calculate two proportions, that bounded by the mean and 65 and that bounded by the mean and 75.

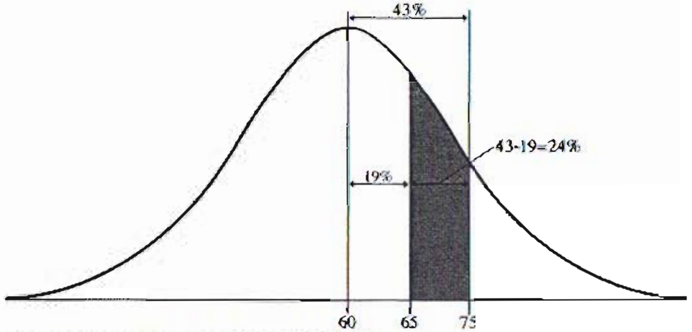


Figure 11.8 The area under the curve not bounded by the mean

We know from our earlier example that 19 percent of cases will have grades between 60 and 65. To determine the percentage of cases that will have a grade between 60 and 75, I first calculate the *z*-score for this range of scores:

$$z = \frac{X_i - \bar{X}}{s} = \frac{75-60}{10} = 1.5$$

From the table for the area under the standard normal curve (Table 11.3) 0.866 (86.6 percent) of cases will fall 1.5 *z*-scores above and below the mean, so that half of this (43.3 percent) will fall above the mean, with grades between 60 and 75. The result is 24 percent of students received a credit grade.

Calculating values from z-scores

In the above examples we wanted to identify the percentage of cases that have a certain range of grades. However, the problem we want to address might be slightly different. We might already have a predefined proportion of cases in which we are interested, and want to derive the grade range within which this percentage falls. For example, we might be interested in the range of scores that identify the middle 50 percent of students. Another way of posing this problem is to ask which scores mark the upper and lower bounds of the interquartile range.

We begin by looking at Table 11.7 to find the *z*-scores that will mark off the 0.5 (50 percent) region. We look down the column for the area under the curve between points and find the cell that has a probability of 0.5 (or the closest to it). The closest value to 0.5 is 0.516, which is associated with *z*-scores of +0.7 and –0.7. To convert these *z*-scores of –7 and +7 into the actual units (exam grades) in which we are measuring the variable, we rearrange the basic formula slightly:

$$z = \frac{X_i - \bar{X}}{s} \rightarrow X_i = \bar{X} \pm z(s)$$

Table 11.7 Areas under the standard normal curve

z	Area under curve between both points	Area under curve beyond both points	Area under curve beyond one point
±0.1	0.080	0.920	0.4630
±0.2	0.159	0.841	0.4205
±0.3	0.236	0.764	0.3820
±0.4	0.311	0.689	0.3445
±0.5	0.383	0.617	0.3085
±0.6	0.451	0.549	0.2745
±0.7	0.516	0.484	0.2420
±0.8	0.576	0.424	0.2120
±0.9	0.632	0.368	0.1840
±1	0.683	0.317	0.1585
⋮	⋮	⋮	
±3	>0.996	<0.004	<0.0020

If we put the two *z*-scores that define the region into this equation we obtain:

$$X_i = \bar{X} + z(s) = 60 + 0.7(10) = 67$$
$$X_i = \bar{X} - z(s) = 60 - 0.7(10) = 53$$

Therefore the 'middle' 50 percent of students scored between 53 and 67 in the exam. This also means that 25 percent of students are below 53 and 25 percent of students are above 67.

Normal curves on SPSS

We introduced the concept of the normal curve using a hypothetical survey of all 1200 people in a community, with mean age of 35 years and standard deviation of 13 years.

The data for this hypothetical survey have been entered into SPSS that will allow us to use SPSS to confirm the results we obtain from hand calculations. First, we can use SPSS to assess the extent to which the spread of scores can be described by a normal distribution. To do this we simply extend the procedure we learnt in Chapter 3 for generating a histogram. We can ask SPSS to generate a histogram, and also to 'fit' a normal curve onto this histogram. By looking at the results we can see the extent to which the distribution of data approximates a normal distribution. To generate a histogram, and a normal curve centered on the mean superimposed on the histogram, we use the procedure shown in Table 11.8. This procedure will generate the output shown in Figure 11.9.

Table 11.8 Generating a histogram with a normal curve on SPSS (file: **Ch11.sav**)

SPSS command/action		Comments
1	From the menu select **Graphs/Interactive/Histogram**	This brings up the **Histogram** dialog box
2	We drag **Age of respondent** into the blank box along the horizontal axis	
3	Click on the **Histogram** option	
4	Click on the small square next to **Normal curve**	A ✓ will appear in the check-box to indicate that a normal curve will be 'fitted' to the histogram
5	Click on **OK**	

Looking at the histogram with the normal curve superimposed on it, we can see that the histogram 'sort of' has the bell-shaped, symmetric features of the normal curve; it is **approximately** normal. A normal curve is a smooth continuous distribution – there are no 'jumps' from one value to another. We, on the other hand, are using a histogram with data arranged according to discrete units of measurement (age in whole years). Since histograms have jagged edges, brought about by the use of discrete units of measurement, they will not fit

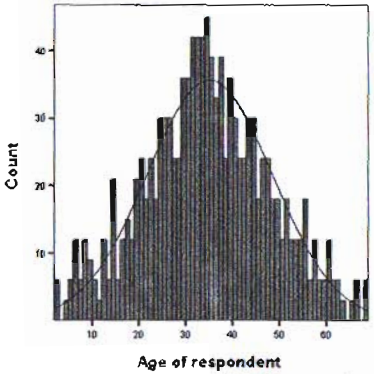


Figure 11.9 SPSS histogram with normal curve

the smoothly rising and falling normal curve. There are also some bars in which the normal curve does not pass exactly through the mid-point. For example, the bar for the middle group in the distribution, with a mid-point of 35 years, has more people in it than would be the case if the distribution was perfectly normal. Despite this variation, the distribution appears to the eye to be approximately normal.

Alternatively, or in addition to this, we can run the **Analyze/Descriptive Statistics/Explore** command and refer to the measures of skewness and kurtosis in the table. For these data we will obtain a measure of skewness of 0.026, which is very close to 0, indicating that the distribution is not very skewed. The measure of kurtosis is –0.265 indicates that the distribution has only slightly fatter tails (is less 'peaked') than a normal curve. These measures indicate that treating this distribution of scores as approximately normal is justified.

Since the distribution is approximately normal, we can use *z*-scores to analyze it. For example, we might be interested in the proportion of people who are not eligible to vote because of their age. This means finding the proportion of people who are less than 18 years of age. The first step is to determine how many *z*-scores 18 is from the mean of 35 years. Since we are working with a population distribution the appropriate formula is:

$$Z = \frac{X_i - \mu}{\sigma} = \frac{18-35}{13} = -1.3$$

Since we are only interested in the area in one tail of the distribution, we refer to the column for the area under the curve *beyond one point* when referring to Table 11.9.

Table 11.9 Areas under the standard normal curve

z	Area under curve between both points	Area under curve beyond both points	Area under curve beyond one point
±0.1	0.080	0.920	0.4600
±0 2	0.159	0.841	0.4205
±0 3	0.236	0.764	0.3820
±0.4	0.311	0.689	0.3445
±0 5	0.383	0.617	0.3085
⋮	⋮	⋮	
±1.1	0.729	0.271	0.1355
±1.2	0.770	0.230	0.1150
±1.3	0.806	0.194	0.0970
±1.4	0.838	0.162	0.0810
±1.5	0.866	0.134	0.0670
⋮	⋮	⋮	
±3	>0.996	<0.004	<0.0020

Only 0.097 (9.7 percent) of the curve lies beyond a *z*-score of –1.3 (Figure 11.10).

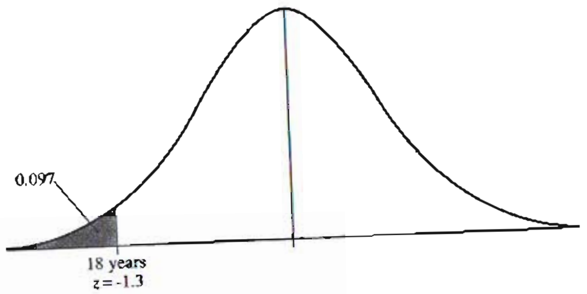


Figure 11.10 Area beyond $z = -1.3$

In the data file you will find that the *actual* percentage of cases whose age is less than 18 years is 10 percent. Thus by using the normal curve to describe the distribution we are very close to the actual results. We can confirm this by looking at the proportion of cases that fall within certain ranges around the mean. As we discussed above, for a normal curve we know that 68 percent of cases will fall within 1 *z*-score from the mean, and for this population this was bounded by 22 and 48 years of age. If we generate a frequency table and add the percentage of cases that have ages between 22 and 48 (inclusive) the sum will be 68 percent, which is consistent with the percentage of cases we expect to find in this range based on the normal curve.

This little exercise indicates that when a distribution is approximately normal, the calculation of *z*-scores can be a quick way of determining the frequency of cases within any range of values that may interest us. Of course, because any distribution is only approximately normal, the proportions obtained by using *z*-scores will not always be exactly equal to the actual proportion of cases within the range of values we are interested in.

Unfortunately SPSS does not provide the facility for calculating *z*-scores or associated areas under the normal curve, if we determine that a particular distribution is approximately normal. That is why we have so laboriously worked through so many hand calculations. An alternative to the hand calculation of *z*-scores and areas under the curve is to use the various web pages available that perform the calculations for you. A list of such pages is located at the web address members.aol.com/johnp71/javastat.html#Tables. One of these that is particularly well constructed is davidmlane.com/hyperstat/z_table.html which calculates *z*-scores based on desired areas and vice versa, and also illustrates the results with normal curves with clearly shaded areas.

Exercises

11.1 From the table for the area under the standard normal curve find the probability that a normally distributed variable will have a *z*-score:

(a) above 1.3
(b) below 1.3
(c) between 0.5 and 3.4
(d) between –2.3 and 2
(e) greater than 2.3 and less than –1.4
(f) less than –1.6 and greater than 1.6
(g) less than –1.96 and greater than 1.96.

For each of these regions draw a sketch of the normal curve and shade in the appropriate area.

11.2 If a set of cases is normally distributed, using the table for the area under the standard normal curve, find the *z*-score(s) that define the following proportions of cases:

(a) the middle 0.683 of cases
(b) the 0.018 cases with the highest scores
(c) the 0.05 cases with the lowest scores
(d) the 0.134 cases which together form the extremes of the distribution.

For each of these regions sketch the normal curve with the appropriate area shaded.

11.3 If *X* is a variable with a normal distribution, a mean of 60, and a standard deviation of 10, how many standard deviations from the mean are the following values for *X*?

(a) 60 (b) 52 (c) 85 (d) 43 (e) 73

11.4 A (hypothetical) study has discovered that the income of families headed by a single mother is normally distributed, with an average annual income of $17,500, and standard deviation of $3000. If the poverty line is considered to be $15,000, how many families headed by a single mother are living in poverty? Sketch the normal curve to illustrate your answer.

11.5 If the mean life of a certain brand of light bulb is 510 hours and the standard deviation is 30 hours, what percentage of bulbs lasts no more than 462 hours? (Assume a normal distribution.)

11.6 The average selling price of a new car is $19,800 and the standard deviation is $2300.

(a) What proportion of new cars will sell for less than $16,000?
(b) Within what limits will the middle 95 percent fall? (Assume a normal distribution.)

11.7 The reaction time of a motorist is such that when travelling at 60 km/h his average breaking distance is 40 meters with a standard deviation of 5 meters.

(a) If the motorist is travelling at 60 km/h and suddenly sees a dog crossing his path 47 meters away, what is the probability he will hit it?
(b) How far away will the dog have to be to have a 95 percent chance of not being hit? (Assume a normal distribution.)

11.8 (a) In the example used in this chapter for the distribution of the ages of 1200 community residents, calculate, using *z*-scores, the proportion of cases that are of working age, that is between 18 and 65 years old.
(b) Calculate the range of ages that determine the middle 50 percent of cases. Confirm your calculations by referring to the SPSS frequency table for this distribution.

11.9 Based on past results, a charity organization expects that donations for the forthcoming year can be modelled using a normal curve. It expects to receive donations of $1.5 million in the following year, with a standard deviation of $200,000. Its target is $1.7 million in donations.

(a) What is the expected probability of meeting this target?
(b) If the charity considers $1.2 million to be the minimum amount it requires to cover costs and meet the basic needs of the poor in its area, what is the expected probability that it will receive enough to meet this minimum?

11.10 A local energy-generating program is proposed using wind power. This form of energy generation is only viable if wind speed in a certain area is over 15 km/h for at least 25 percent of the time. The average wind speed is 12 km/h with a standard deviation of 6 km/h. Is there sufficient evidence to suggest that the project will be viable?

12

Correlation and regression

In Chapters 5–8 we explored methods for describing data that are grouped into categories. With only a few categories to express the range of variation, our initial means of describing such data is in the form of a crosstabulation. The crosstab shows the joint distribution for two variables and allows us visually to gauge whether there is an association between the two variables. If inspection of the relative frequency distribution in the table leads us to suspect that these two variables are related, the next step is to calculate measures of association that give a precise numerical value to any such suspicion.

However, if the data for the two variables under investigation have been collected at the interval/ratio level, and they have a large number of values, crosstabulations are not a convenient means of describing the distribution. The equivalent descriptive technique to a crosstabulation for interval/ratio data is a **scatter plot**.

Scatter plots

It is difficult to arrange interval/ratio data into a crosstabulation. Interval/ratio data do not usually fall into a small number of discrete categories such as large or small, old or young, etc. Since there are usually many values for variables measured at the interval/ratio level, a contingency table will have to have as many rows or columns as there are values in the data. If we are looking at the distribution of age in years for a country's population we will need over 100 rows of data to take account of the fact that age spreads out over a wide range. Such data can of course be collapsed into a few values, but this is at the cost of information. A scatter plot, which allows for the greater range of values that we usually have with interval/ratio scales, is the best way to organize such data to get an initial impression as to whether any correlation exists. A scatter plot (just like a crosstab) shows the combination of values that each case 'scores' on two variables simultaneously.

A scatter plot displays the joint distribution for two continuous variables. **Coordinates** on a scatter plot indicate the values each case takes for each of the two variables.

For example, we might be interested in the relationship between unemployment rates and the level of civil unrest across cities. From official statistics we obtain the information in Table 12.1, which presents the rate of unemployment (which we think is the independent variable, *X*) and the number of civil disturbances (which we think is the dependent variable, *Y*) for five cities.

Table 12.1 Unemployment and civil unrest in five cities

City	Unemployment rate, *X*	Civil disturbances, *Y*
A	25	17
B	13	15
C	5	10
D	10	5
E	2	4

Arranging this information in a scatter plot (Figure 12.1) makes these data easier to 'read' in order to determine whether an association exists.

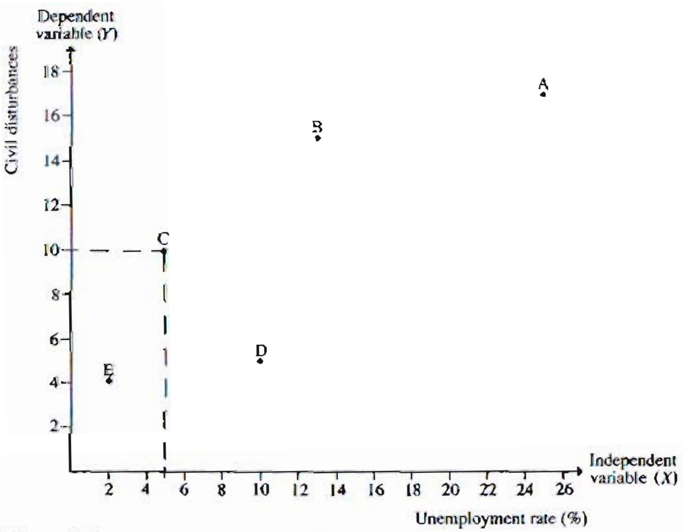


Figure 12.1 A scatter plot of data points

It is the convention to put the dependent variable, Y, on the vertical axis and the independent variable, X, on the horizontal axis when constructing a scatter plot. If we look at any one of these points and draw a straight line down to the horizontal axis, we can find the unemployment rate in that town. Similarly, by drawing a straight line across to the vertical axis we can 'read off' the number of civil disturbances. Grid lines for C have been drawn to illustrate this procedure. For this town the unemployment rate is 5 percent and there are also 10 incidents of civil unrest.

Looking at Figure 12.1, it can intuitively be seen that an association exists, because we can imagine a sloping line running through these five points. The direction of association is indicated by whether this imaginary line slopes up (positive) or down (negative). In this case the slope is positive, indicating that an increase in unemployment rate is associated with an increase in the number of civil disturbances. We can give quantitative expression to this imaginary line through the calculation of **linear regression** statistics. This extension of a scatter plot by calculating regression statistics for variables measured on interval/ratio scales with many points is directly analogous to the extension of crosstabulations by calculating measures of association when working with categorical data.

Linear regression

Each and every straight line that can be drawn in the area defined by the scatter plot has a unique equation that distinguishes it from every other line. Deriving this equation for any particular line is like giving a person a unique combination of first and last names so that this person can be differentiated from everybody else. The general formula for the equation of a line is much like a form with a space entitled Firstname and another space entitled Lastname.

$$Y = \text{Firstname} \pm \text{Lastname}$$

We write in the specific combination of names that identifies the relevant individual. If I try to identify somebody using just their first name, say Pablo, this will not differentiate that person from all the other people with the same first name. Similarly, if I identify someone just by their last name, say Picasso, this will not differentiate this person from all other people with the same last name. But writing both names *together* will identify a unique individual. Similarly with identifying a line. Thousands of straight lines can be drawn through the space marked out by the vertical and horizontal axes of a scatter plot. But to identify the individual line that we think best fits the scatter plot we need to provide it with a unique first and last name. The line's first name is its point of origin along the Y-axis. But obviously this is not enough to distinguish it from the multitude of lines that can start from the same point. This is illustrated in Figure 12.2, which shows only some of the lines that will share the same value for a in their equation.

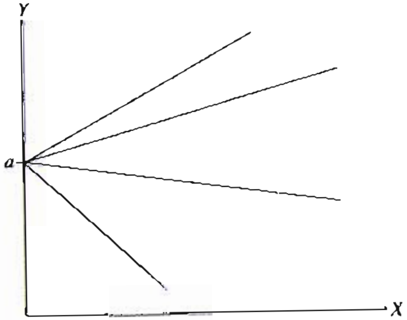


Figure 12.2 Straight lines with the same value for a but different values for b

Specifying the slope of a straight line on its own is also insufficient to distinguish it from all the others that could occupy the space. This is illustrated in Figure 12.3, which presents lines that will all have the same slope, but different origin.

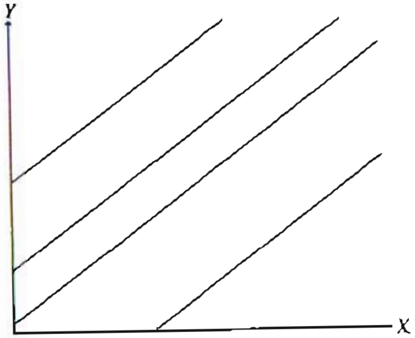


Figure 12.3 Straight lines with the same value for b but different values for a

However, if we specify *both* the point of origin on the Y-axis *and* the slope of the line from that point, then we are able to identify uniquely any line within the space. The trick is to come up with the unique combination of values for a and b that identify the **line of best fit**.

Regression analysis is simply the task of fitting a straight line through a scatter plot of cases that 'best fits' the data. Any straight line can be expressed in a mathematical formula. The general formula for a straight line is:

$$Y = a \pm bX$$

where:
Y is the dependent variable
X is the independent variable
a is the Y-intercept (the value of Y when X is zero)
b is the slope of the line
+ indicates positive association
– indicates negative association.

This formula says that a line is defined by two factors. One is its starting point along the vertical axis, a, and the second is the slope of the line from this point, ±b. It is the value of b that we are most interested in since any slope, either positive or negative, indicates some correlation between the two variables. In Figure 12.4 we see three different lines reflecting the value of b in the three alternative situations of positive, negative, and no correlation.

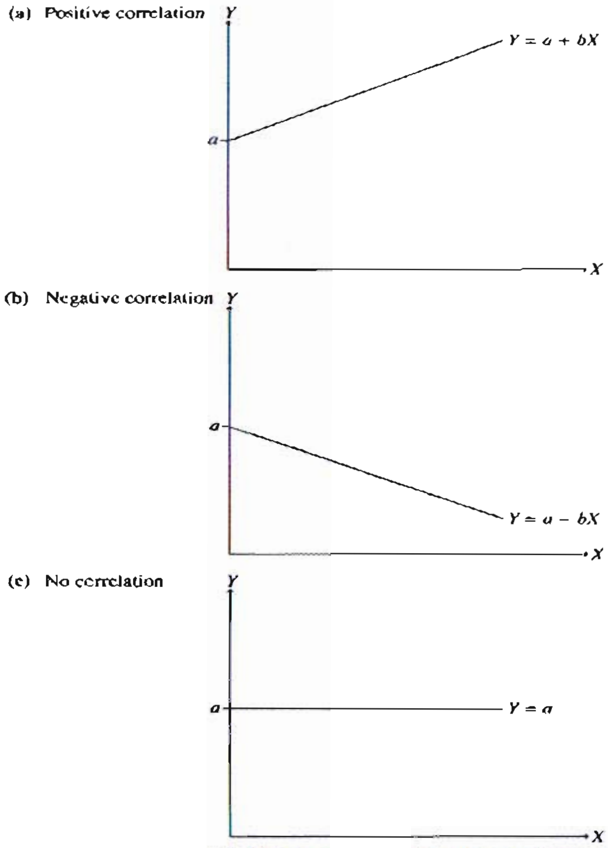


Figure 12.4 Three lines exhibiting (a) positive, (b) negative, and (c) no correlation

Looking at the data for the five cities, we can draw many straight lines through this scatter plot, and each of these lines will have its own unique formula. For example, in Figure 12.5 I have drawn a line that looks to me to fit the data pretty well. I could call this 'line 1' or 'line A' or 'my line'. Instead, I have called it by its mathematical name: $Y = 5 + 0.6X$.

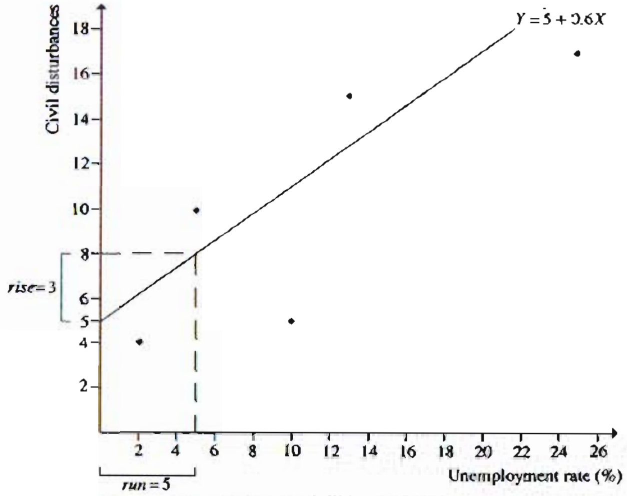


Figure 12.5 Determining the slope of a regression line

How did I arrive at the values in this equation?

- The value for a (5) is the point on the Y-axis where the line 'begins'. This is the number of civil disturbances we expect to find in a city with an unemployment rate of zero.
- The + sign means that the line has a positive slope, which indicates a positive correlation between these two variables.
- The value of 0.6 for b is the slope, or **coefficient**, of the regression line. The regression coefficient indicates by how much civil disturbances will increase if unemployment increases by 1 percent. Since the slope of any straight line is 'rise over run', to actually calculate this value I take any 'rise' in civil disturbances, such as the increase of 3 between 5 and 8. I then 'read off' the corresponding increase in the unemployment rate, which gives a 'run' of 5. Dividing rise over run, the slope will be:

$$b = \frac{\text{rise}}{\text{run}} = \frac{3}{5} = 0.6$$

The line we have just identified gives us a range of **expected values** for civil disturbance, depending on the value of the unemployment rate. The difference between the expected value and the actual value for civil disturbance at a particular unemployment rate is called the **residual** or **error term**.

The **residual** or **error term** is the difference between the observed value of the dependent variable and the value of the dependent variable predicted by a regression line.

Notice that no straight line will pass through all the points in a scatter plot. In fact, a 'good' line might not touch *any* of the points: there will usually be a gap between each plot and the regression line. Unless a point falls exactly on the line there will be a residual value. For example, my line predicts that, for city D with unemployment of 10 per cent, the number of civil disturbances will be 11:

$$Y = 5 + 0.6X = 5 + 0.6(10) = 11$$

Instead, there were five civil disturbances for city D with an unemployment rate of 10 percent. The error (*e*) *term* at this point is –6 (Figure 12.6):

$$e = Y_{actual} - Y_{expected} = 5 - 11 = -6$$

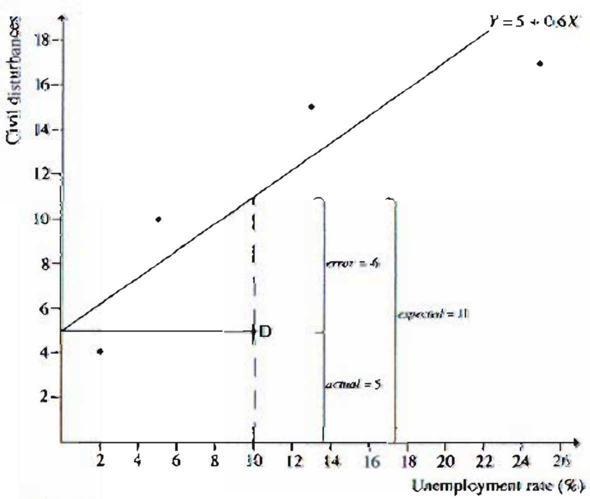


Figure 12.6 Observed and expected scores

I drew the particular line in Figure 12.6 on the basis of what looked to me, with the naked eye, to be the line that best fits these data. Someone else might think that they could draw a better line through these points, and this new line would have its own equation to define it, and the residuals between the expected values and actual values will be different to the ones I derived. It might be hard to determine which of these lines is the 'best' one just on the basis of our eyeball impression. We obviously need an objective principle for determining which line is the 'best'. Of all the possible lines that could run through the points, it seems plausible to suggest that the *best* line is the one that makes the residuals as small as possible: the one that *minimizes* the residuals.

Regression analysis uses this idea (although in a slightly more complicated form). The logic is called **ordinary least squares regression** (OLS): we want a line such that the gaps between the estimated values of *Y* and the actual values of *Y* (squared) are as small as possible. (We *square* the residuals, rather than just *sum* them, because the sum of residuals for *any* line that passes through the point that is the mean for both the dependent and independent variables will equal zero. To eliminate the effect of the positive and negative signs, the residuals are squared so that we are only dealing with positive numbers.)

Ordinary least squares regression is a rule that tells us to draw a line through a scatter plot that minimizes the sum of the squared residuals.

We could find the OLS regression line through a process of trial and error. We could keep drawing lines through the scatter plot, working out their respective equations and residuals, until we finally hit on the one that minimizes these residuals.

Fortunately there is an alternative. If we use the following two rules, we can derive the OLS regression line directly without having to go through an indeterminate process of trial and error.

1. *The OLS regression line must pass through a point whose coordinates are the averages of the dependent and independent variables* ($\overline{Y}$, $\overline{X}$). The average number of civil disturbances, $\overline{Y}$ (pronounced 'Y-bar'), is 11:

$$\overline{Y} = \frac{\Sigma Y_i}{n} = \frac{55}{5} = 11$$

The average unemployment rate, $\overline{X}$ (pronounced 'X-bar'), is 10.2:

$$\overline{X} = \frac{\Sigma X_i}{n} = \frac{51}{5} = 10.2$$

Thus the OLS regression line will pass through the coordinate point (10.2, 11).

2. *The slope of the regression line, b, is defined by the formula:*

$$b = \frac{\Sigma(X_i - \overline{X})(Y_i - \overline{Y})}{\Sigma(X_i - \overline{X})^2}$$

While this equation captures the essential idea that the line needs to minimize the squared differences between actual and expected values, the value of *b* is easier to calculate using the following formula:

$$b = \frac{n\Sigma(X_iY_i) - (\Sigma X_i)(\Sigma Y_i)}{n\Sigma X_i^2 - (\Sigma X_i)^2}$$

Although this formula still looks intimidating, if we work through it step by step we will see that it is a rather straightforward calculation. The calculations for city A are included in Table 12.2 to show how the numbers are derived.

Table 12.2 Calculations for the slope of the OLS regression line

City	Unemployment rate, X	Civil unrest, Y	X_i^2	Y_i^2	X_iY_i
A	25	17	25×25 - 625	17×17 - 289	25×17 - 425
B	13	15	169	225	195
C	5	10	25	100	50
D	10	5	100	25	50
E	2	4	4	16	8
	$\Sigma X_i = 55$	$\Sigma Y_i = 51$	$\Sigma X_i^2 = 923$	$\Sigma Y_i^2 = 655$	$\Sigma X_iY_i = 728$

Putting all these data into the equation for the slope of the regression line, we get 0.53:

$$b = \frac{n\Sigma(X_iY_i) - (\Sigma X_i)(\Sigma Y_i)}{n\Sigma X_i^2 - (\Sigma X_i)^2} = \frac{5(728) - (55)(51)}{5(923) - (55)^2}$$

$$= \frac{3640 - 2805}{4615 - 3025} = +0.53$$

The value of b, called the **regression coefficient**, is very important because it quantifies any correlation between two variables.

> **The regression coefficient** indicates by how many units the dependent variable will change, given a one-unit change in the independent variable.

Now that we have fixed the regression line through a specific point (the averages of X and Y) and also given it a 'last name' by calculating the slope of the line through this point, we can give it a complete label by deriving the value for a. We use the following formula, which uses both of the features of the regression line we have identified (it passes through the average of X and Y, and has a slope equal to b):

$$a = \bar{Y} - b\bar{X}$$

Therefore the value of a will be 4.4:

$$a = \bar{Y} - b\bar{X} = 10.2 - 0.53(11) = 4.4$$

Thus we can define the line of best fit, for this set of cases, with the following equation:

$$Y = 4.4 + 0.53X$$

In Figure 12.7 this regression line is drawn through the scatter plot.

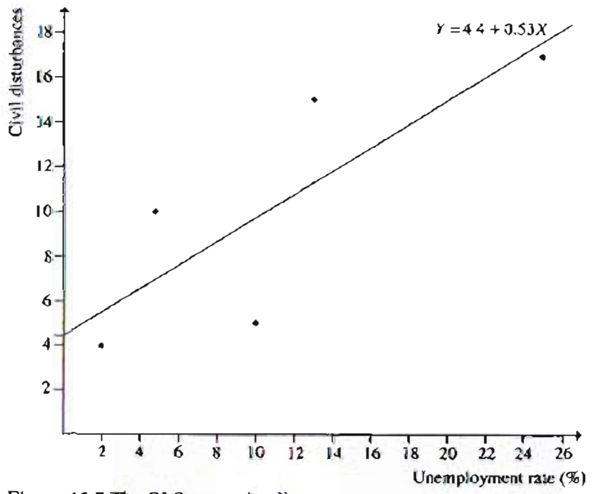


Figure 12.7 The OLS regression line

What does this tell us about the relationship between unemployment rates and civil disturbances, for this set of cases?

- There is a positive relationship between the two variables: an increase (decrease) in the unemployment rate is correlated with an increase (decrease) in the number of civil disturbances.
- We can quantify this positive correlation: an increase in the unemployment rate of 1 percent is correlated with an increase of 0.53 civil disturbances.

I can now use this formula for the purpose of prediction: I can predict the number of civil disturbances a city is likely to have, given a certain rate of unemployment. For example, if I was told that another city has an unemployment rate of 18 percent, my best guess will be to say that it experiences 13.9 civil disturbances:

$$Y = 4.4 + 0.53(18) = 13.9$$

If you are still a little confused about what this all means, imagine that you are arranging for a repairman to come and fix an appliance in your home. The charge is a flat fee of $50 for the visit plus $20 for each hour spent working in your home. We can summarize what we are required to pay the repairman using the following equation:

$$\$\text{payment} = 50 + 20(\text{number of hours})$$

For *any* given amount of time spent in the home we can calculate the total cost. For example, if the repairman comes and finds nothing wrong and therefore does not charge for time, we will still be obligated to pay $50 (the constant fixed amount) for the visit. If it takes 2 hours to fix a problem, on the other hand, we are up for $90. The regression line does essentially the same thing: it tells us what we predict will be the value of the dependent variable, given a certain value for the independent variable. The only difference is that we will never get the exact amount, since the data points do not all fall exactly on the regression line, so we have to allow for error.

Pearson's product moment correlation coefficient

We have seen that the value of b is an indicator of whether a correlation exists between two variables measured at the interval/ratio level, and also the direction of such correlation. But does it also indicate the *strength* of the relationship? Does a value of $b = 0.53$ indicate a strong, moderate, or weak association? Unfortunately it does not.

The problem is that the units of measurement vary from one situation to another. For example, if I use proportions rather than percentage points to measure unemployment rates, so that instead of using in my calculations 22, 20, 15, 10, 9, I use 0.22, 0.20, 0.15, 0.1, 0.09, the estimated value of b will be 53 rather than 0.53. The actual relationship I am looking at has not changed, only the units of measurement. In other words, the value of b is affected not only by the strength of the correlation, but also *by the units of measurement*. Therefore there is no way of knowing whether any particular value for b indicates a weak, moderate, or strong correlation.

To overcome this problem, we convert the value of b into a standardized measure of correlation called the **product moment correlation coefficient, Pearson's** ***r***. Pearson's r will always range between −1 and +1, regardless of the actual units in which the variables are measured. The formula for r is:

$$r = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left[(X_i - \bar{X})^2\right]\left[(Y_i - \bar{Y})^2\right]}}$$

or

$$r = \frac{n\Sigma(X_iY_i) - (\Sigma X_i)(\Sigma Y_i)}{\sqrt{\left[n\Sigma X_i^2 - (\Sigma X_i)^2\right]\left[n\Sigma Y_i^2 - (\Sigma Y_i)^2\right]}}$$

Fortunately, we have already calculated the elements of this equation in the table we used above for calculating b (Table 12.2). If we substitute the statistics from this table into this second formula we get 0.81:

$$r = \frac{n\Sigma(X_iY_i) - (\Sigma X_i)(\Sigma Y_i)}{\sqrt{\left[n\Sigma X_i^2 - (\Sigma X_i)^2\right]\left[n\Sigma Y_i^2 - (\Sigma Y_i)^2\right]}} = \frac{5(728) - 55(51)}{\sqrt{\left[5(923) - (55)^2\right]\left[5(655) - (51)^2\right]}}$$

$$= \frac{3640 - 2805}{\sqrt{[4615 - 3025][3275 - 2601]}} = 0.81$$

The value of r tells us the strength as well as the direction of association. A value of 0.81 indicates that the correlation between these two variables for this set of cases is a strong, positive one.

The problem with the correlation coefficient is that its relative values are not proportional to the relative strength of the relationship. In other words, an r of 0.6 is not twice as strong as an r of 0.3. This makes it difficult, and sometimes misleading, to compare the correlation coefficients for different pairs of variables. More generally, the correlation coefficient does not have any direct interpretation in PRE terms so it does not indicate the *confidence* we can place in our estimates, especially when making predictions. These problems are overcome by the square of the correlation coefficient, called the **coefficient of determination**, r^2.

Explaining variance: The coefficient of determination

We have already used the regression line to predict the number of civil disturbances in a city, given a particular rate of unemployment. But we also saw that there will usually be a margin of error in this prediction, depending on how closely the plots are clustered around the line. We can use the regression line to say that a certain increase in X will produce so much increase in Y, but if there are large error terms between the regression line and the actual data points to the likelihood that our predictions will be wrong and will be greater than in a situation where the scores are tightly packed around the regression line.

We can see in Figure 12.8 that even though the same regression line best fits both sets of plots, we will have a greater confidence in our predictive ability in (a) than in (b). This is because the regression line in (a) explains a greater proportion of the variance of Y than the line in (b). We therefore need some measure of how much of the variation in the dependent variable is explained by a regression line.

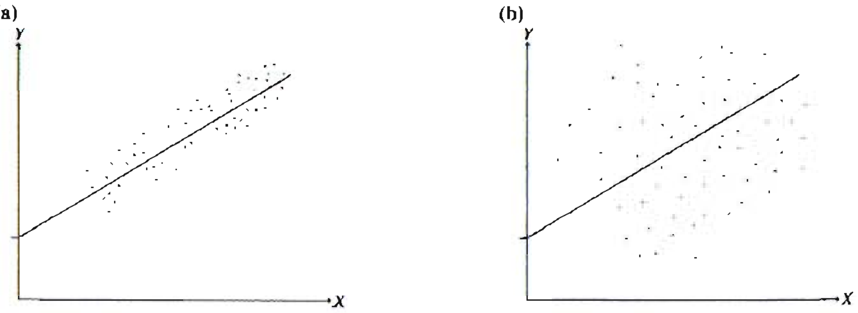


Figure 12.8 Regression lines that explain (a) a high amount and (b) a low amount of variation

Fortunately we can do this by simply squaring r and obtaining the **coefficient of determination**, r^2, the variance explained by the regression line relative to the variance explained in the case of no association:

$$r^2 = (0.81)^2 = 0.65$$

The coefficient of determination can be interpreted as an asymmetric PRE measure of association, much like the PRE measures we encountered in Chapters 6–7. In fact, it has a logic very similar to lambda, but applied to interval/ratio data. We make predictions about the expected value of the dependent variable *without* any information about the independent variable. We then make predictions *with* knowledge of the independent variable and compare the error rates.

For example, if we have to guess the number of civil disturbances in each city, and all we know is that the average number of disturbances *for all* five cities is 10.2, the best guess we can make is to say that the number of civil disturbances in *each* city is 10.2, regardless of the actual unemployment rate. In other words, we draw a straight horizontal line at this value as the regression line through the scatter plot (Figure 12.9).

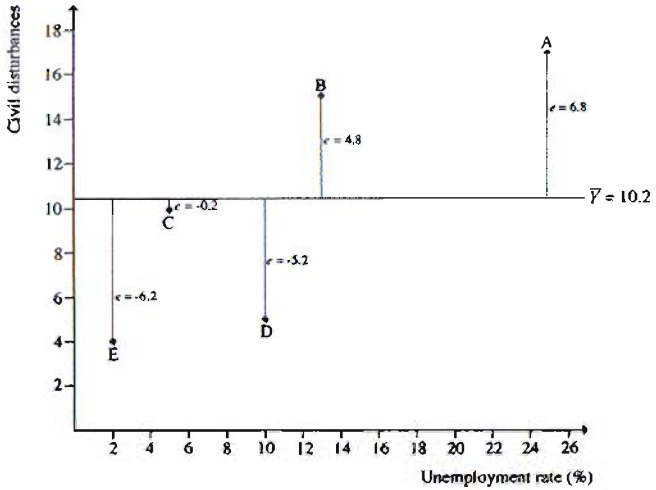


Figure 12.9 Regression line without knowledge of the independent variable

This horizontal line is the line we draw if there is no correlation between these two variables; knowing whether the unemployment rate is high or low will not cause me to change my expected number of civil disturbances from the average. Sometimes this line comes very close to the mark. For city C we see that this line predicts, at an unemployment rate of 5 percent, there will be 10.2 civil disturbances. There were in fact 10 civil disturbances producing an error (e) for this city of only –0.2. However, in other instances we make a large error using this line. For city A, at an unemployment rate of 25 percent we again predict 10.2 civil disturbances, but in fact there were 17, producing an error of 6.8.

Now we compare these errors with the errors we make when predicting on the basis of the ordinary least squares regression line (Figure 12.10). Does the OLS line substantially improve our guesswork?

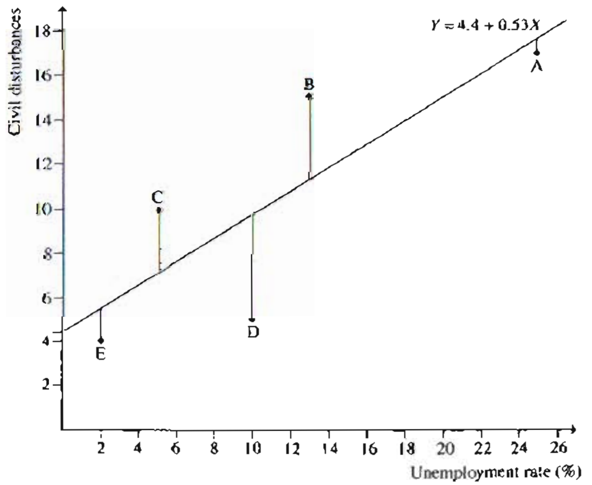


Figure 12.10 Regression line with knowledge of the independent variable

We can see that if there is a close correlation between these two variables, the least squares regression line will reduce the error rate. The gaps between the data points and the line will be much smaller when using the least squares regression line than when using the horizontal line based on the assumption of no correlation. It is precisely this aspect of the regression line that the coefficient of determination captures. A value for r^2 of 0.65 indicates that the least squares regression line explains 65 percent of the variance of the dependent variable relative to the variance explained by the horizontal line. This is a substantial reduction in the error rate.

It may pay to stop at this point and discuss the difference between r and r^2 since they are very closely related. The correlation coefficient is a standardized measure of the relationship between two variables; that is, it indicates the extent to which a change in one variable will be associated with a change in another variable. Thus r (like b) is primarily a tool for prediction. The coefficient of determination, on the other hand, is a PRE measure of the amount of variation explained by a regression line, and therefore gives a sense of how much *confidence* we should place in the accuracy of our predictions.

Plots, correlation, and regression using SPSS

The data from this example have been entered into SPSS. To generate the results we obtained above on SPSS, we can use either of two separate commands, each of which produce different amounts of information, neither of which is (unfortunately) completely ideal. One command generates a graphical description of the data in the form of scatter plot along with a regression line and value for r^2 (but not the inferential statistics that we will discuss in Chapter 26). The other command provides the numerical descriptions in the form of the regression equations and correlation coefficients, along with the inferential statistics, but not the scatterplot.

Generating an interactive scatter plot with a regression line and statistics

To obtain a simple scatter plot of the data, we use the procedures given in Table 12.3 and Figure 12.11, which also present the output from this set of commands. Note that point 4 is optional, but with graphs that only have few data points, such as this one, it is helpful to label the plots with an identification variable (such as city letters in this instance) to help us better 'read' the graph.

Table 12.3 Interactive scatter plots using SPSS (file: Ch12-1.sav)

	SPSS command/action	Comments
1	From the menu select **Graphs/Interactive/Scatterplot**	This brings up the **Create Scatterplot** dialog box.
2	Drag **Number of civil disturbances** into the empty box on the vertical arrow	This pastes **Number of civil disturbances** as the variable to be displayed on the *Y*-axis (dependent)
3	Drag **Unemployment rate** into the empty box on the horizontal arrow	This pastes **Unemployment rate** as variable to be displayed on the *X*-axis (independent)
4	Drag **city** into the box next to **Label Cases By:** (optional)	This will place the city label next to each of the points on the scatter plot
5	Click on the **Fit** option	
6	From the drop-down menu next to **Method:** select **Regression**	This will fit the OLS regression line in the scatter plot and also the regression equation with the graph
7	Click on **OK**	

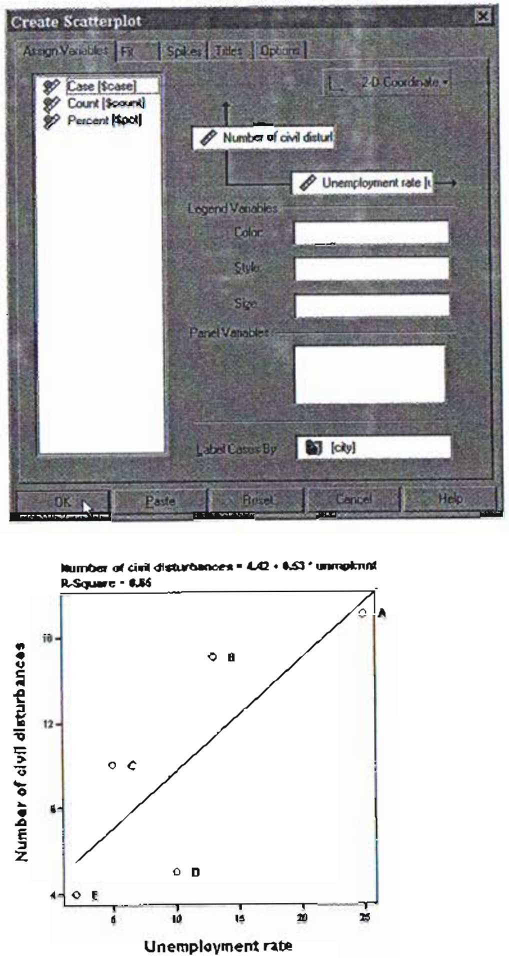


Figure 12.11 The Simple Scatterplot dialog box and output

Regression statistics

To generate the statistics behind this regression line we follow the procedure in Table 12.4 and Figure 12.12.

A wealth of output is generated as a result of this command (Figure 12.13), much of which is beyond the scope of this book. The two parts of the output that concern us now are the tables headed **Model Summary** and **Coefficients**.

Table 12.4 Regression with curve estimation using SPSS (file: Ch12-1.sav)

	SPSS command/action	Comments
1	From the menu select **Analyze/Regression/Linear**	This brings up the **Linear Regression** dialog box
2	Click on **Number of civil disturbances**	
3	Click on ▸ that points to the **Dependent:** target variable list	This pastes **Number of civil disturbances** as the dependent variable
4	Click on **Unemployment rate**	This highlights **Unemployment rate**
5	Click on ▸ that points to the **Independent(s):** target variables list	This pastes **Unemployment rate** as the independent variable
6	Click on **OK**	

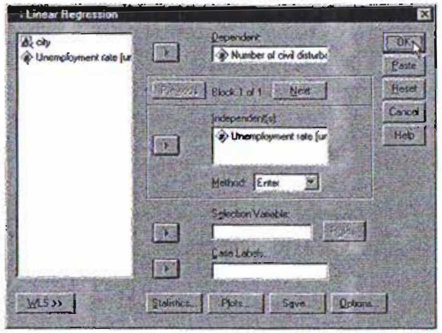


Figure 12.12 The **Linear Regression** dialog box

In the **Model Summary** table we see that Pearson's product moment correlation coefficient, R, is .807, and the coefficient of determination, R Square, is .651, which are the same as our hand calculations. The important part of the **Coefficients** table is the column headed **B** under Unstandardized Coefficients. This tells us that:

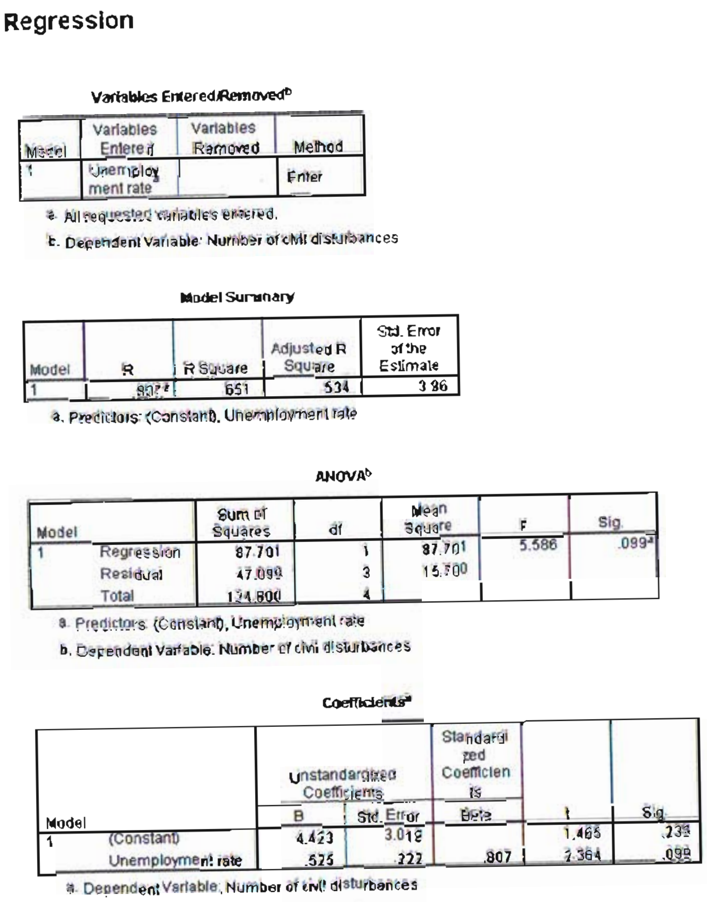
- the value for the *Y*-intercept (which we called *a* in the analysis above but SPSS calls the Constant) is 4.423, and
- the slope of the regression line, which is the coefficient for Unemployment rate, is .525.

Again these are the same values we calculated by hand. The figure under Standardized Coefficients, .807, should look familiar: this is the value for Pearson's *r*, which was also given to us in the other table.

In addition to SPSS there are a number of web pages listed at the following address that allow you to perform correlation and regression analysis:

- members.aol.com/johnp71/javastat.html#Regression

These pages are usually limited by the amount of data points that can be entered; at most some of these pages allow for 84 points to be entered.



Regression

Variables Entered/Removed^b

Model	Variables Entered	Variables Removed	Method
1	Unemployment rate^a		Enter

a. All requested variables entered.

b. Dependent Variable: Number of civil disturbances

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.807^a	.651	.534	3.96

a. Predictors: (Constant), Unemployment rate

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	87.701	1	87.701	5.586	.099^a
	Residual	47.099	3	15.700		
	Total	134.800	4			

a. Predictors: (Constant), Unemployment rate

b. Dependent Variable: Number of civil disturbances

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4.423	3.018		1.465	.239
	Unemployment rate	.525	.222	.807	2.364	.099

a. Dependent Variable: Number of civil disturbances

Figure 12.13 SPSS Regression command output

Example

A museum keeps track of the number of visitors on randomly selected days across the year, in order to help it plan for crowds. It suspects that the daily temperature is a good predictor of the number of people who will pass through on any given day. The data on the daily temperature, measured in degrees Celsius, and the number of people attending, together with the calculations needed to construct a regression line, are included in Table 12.5.

We can use this information to calculate the mean for each variable:

$$\bar{X} = \frac{\Sigma X_i}{n} = \frac{422}{20} = 21.1$$

$$\bar{Y} = \frac{\Sigma Y_i}{n} = \frac{9102}{20} = 455.1$$

Table 12.5 Calculations for the slope of the regression line

Temperature, X_i	People, Y_i	X_i^2	Y_i^2	X_iY_i
13	501	169	251,001	6513
28	175	784	30,625	4900
32	390	1024	152,100	12,480
20	452	400	204,304	9040
11	550	121	302,500	6050
15	734	225	538,756	11,010
9	620	81	384,400	5580
33	199	1089	39,601	6567
16	390	256	152,100	6240
29	223	841	49,729	6467
12	768	144	589,824	9216
15	679	225	461,041	10,185
18	410	324	168,100	7380
26	320	676	102,400	8320
18	590	324	348,100	10,620
17	650	289	422,500	11,050
27	258	729	66,564	6966
32	201	1024	40,401	6432
28	458	784	209,764	12,824
23	534	529	285,156	12,282
$\Sigma X_i = 422$	$\Sigma Y_i = 9102$	$\Sigma X_i^2 = 10{,}038$	$\Sigma Y_i^2 = 4{,}798{,}966$	$\Sigma(X_iY_i) = 170{,}122$

These figures produce the equation for the slope of the regression line:

$$b = \frac{n\Sigma(X_iY_i) - (\Sigma X_i)(\Sigma Y_i)}{n\Sigma X_i^2 - (\Sigma X_i)^2} = \frac{20(170{,}122) - 422(9102)}{20(10{,}038) - (422)^2} = -19$$

The value for *a* will be 863:

$$a = \bar{Y} - b\bar{X} = 455.1 - (-19.34)(21.1) = 863$$

The OLS regression line therefore is defined by the following equation:

$$Y = 863 - 19X$$

If we want to use less mathematics and use plain words rather than symbols, this equation is:

estimated number of patrons = 863 – 19(daily temperature)

A negative correlation exists between the temperature and the number of people attending the museum. We predict that for every degree that the temperature increases, 19 fewer people will attend the museum. The value for *a* indicates that when the temperature falls to zero the museum should expect 863 visitors.

To assess the strength of this relationship, and the confidence we can place in the predictions based on it, we also calculate the correlation coefficient and then the coefficient of determination. These indicate that there is a strong negative relationship and that the OLS regression line explains a high proportion of the variance in the data, allowing the museum to make confident predictions.

$$r = \frac{n\Sigma(X_iY_i) - (\Sigma X_i)(\Sigma Y_i)}{\sqrt{\left[n\Sigma X_i^2 - (\Sigma X_i)^2\right]\left[n\Sigma Y_i^2 - (\Sigma Y_i)^2\right]}}$$

$$= \frac{20(170{,}122) - 422(9102)}{\sqrt{\left[20(10{,}038) - (422)^2\right]\left[20(4{,}798{,}966) - (9102)^2\right]}}$$

$$= -0.8$$

$$r^2 = (-0.8)^2 = 0.64$$

The assumptions behind regression analysis

We have used the concept of least squares regression to derive a measure of correlation between two interval/ratio-level variables. However, implicit in the use of OLS are certain assumptions, which, if violated, will mean that this will not be the best rule for fitting a line through a scatter plot. It is worth noting these assumptions, although a more detailed discussion would take us too far from the needs of this book.

Linear relationships

Least squares regression assumes that the line of best fit is a straight one, or in more technical terms that there is a **linear relationship**. However, this is not always the case (Figure 12.14).

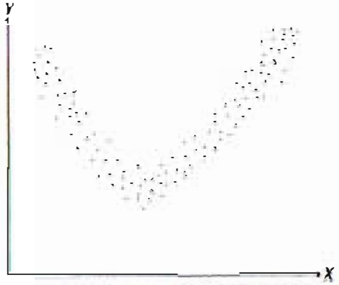


Figure 12.14 A non-linear relationship

It is clear that the line of best fit for this scatter plot will be curvilinear. We can ask SPSS to fit a regression line through these data points, and it will give us the best straight line, but clearly a straight line is not the best line!

Stability

Looking back at the example regarding the relationship between unemployment and civil unrest, the range of values for the independent variable was 2–22 percent. It might be tempting to use the regression line fitted for these data to predict the number of civil disturbances in a city with an unemployment rate of 30 percent. In other words, we might try to project the regression line out past the right-edge of the scatter plot when employing it as a tool for prediction. To do this we have to assume that the relationship is **stable**: that the correlation coefficient will be the same for the whole range of values over which we want to make predictions. This is analogous to the concept of consistency when looking at a crosstab.

This can sometimes be a very dangerous assumption. The statistics we have generated apply just to the cases for which we have information, and to extend their domain to cases for which we don't have information requires some justification. It may be, for example, that when unemployment rates hit a certain threshold level, such as 25 percent, the crime rate jumps up dramatically.

The other aspect of stability relates to time. Unlike the physical sciences, a relationship between two variables in the social sciences is not always the same over time. The relationship between force and mass seems relatively permanent, but the relationship between unemployment and civil disturbance may not be, because history brings about changes to social *institutions* that may alter the character of the relationship. For example, governments may respond to a strong relationship between unemployment and civil disturbance by creating new social institutions such as income support schemes and community programs that could soften the effect of unemployment. Using the information from one historical period for another historical period may therefore be inaccurate.

Homoscedasticity

The strict definition of homoscedasticity is that the variance of the error terms (residuals) of a regression line is constant. The best way to explain this is through an illustration (Figure 12.15).

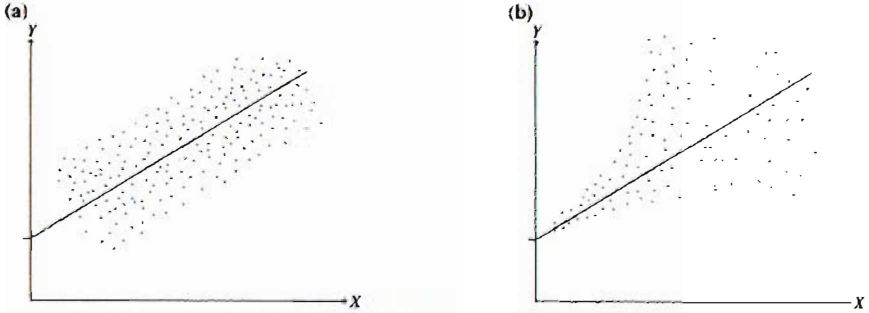


Figure 12.15 Regression where the error terms are **(a)** homoscedastic and **(b)** heteroscedastic

In Figure 12.15(a) we can see that the spread of the data points around the regression line is fairly constant over the length of the regression line. The data points form a 'cigar shape' around the line. In Figure 12.15(b), though, the data points lie far away from the line at one end, and gradually get closer as the value of the independent variable decreases. Graph (a) is the case of homoscedasticity, whereas graph (b) shows heteroscedasticity. The presence of heteroscedasticity causes any significance test on the value of r to be invalid, so that we are not able to generalize from a sample result to the population. Usually a simple inspection of a scatter plot will be sufficient to detect whether this assumption is valid.

Reversibility

This is not so much an assumption regarding the construction of a regression line but rather an assumption in its use. A positive correlation, for example, implies that when the value of an independent variable increases, the value of the dependent variable increases as well, and that when it decreases the dependent variable decreases as well. However, it is not always the case that the same relationship holds for increases as it does for decreases. We all know that there is a positive correlation between income levels and consumption levels: when we have more to spend we spend more! A researcher may look at a period of rising income levels and calculate a value for the regression coefficient (b) of 0.8: when income increases by $100, consumption will go up by $80. Can this researcher then argue that if income decreases by the same amount, consumption levels will go back to where they were before the initial increase? The answer is no. Most people adjust their spending patterns to the higher income level, and do not tend to give it up very easily, even if income falls again. People go into debt or sell off assets in order to maintain the higher spending patterns they have become accustomed to, so that the correlation observed in one direction may not be the same as that observed in the other direction.

Spearman's rank-order correlation coefficient

In Chapter 7 we discussed the use of crosstabs and measures of association as means of describing a relationship between two variables measured at least on ordinal scales. These techniques apply in situations *where the scales do not have too many categories* (five or less categories is a good rule of thumb for the appropriate range of scores). In this chapter thus far we have discussed an alternative set of statistics we can use to describe the relationship between two variables, where both variables are measured at the interval/ratio level and the data have *many* different values. There are situations, though, where we have ordinal scales for two variables and each have a wide range of possible scores and we are reluctant to collapse these scores down to a few categories just to be able to fit the data into a crosstab. This is especially the case where the underlying variables are continuous. An example is an attitude scale. People's attitude to the quality of health care services, for example, is essentially a continuous variable. We may try to capture this intrinsic characteristic of the variable by ensuring that there are a wide number of scores on the scale we use to measure the variation that exists in people's attitude to health care services, from one extreme of 'very unfavorable' to the other extreme of 'very favorable'. We may in fact have a 10 point scale with these two extremes at either end. If we collapse these scores into a smaller number of categories in order to 'fit' them into a crosstab we will lose the scale's sensitivity to small differences in people's attitudes.

Where we have two ordinal scales with a large number of scores (or one ordinal and one interval/ratio) we can describe a relationship between the two variables using **Spearman's rank-order correlation coefficient**, which is also known as **Spearman's rho**. Spearman's rho, in fact, is a particular application of Pearson's correlation coefficient, which we discussed above for relating variables measured on interval/ratio scales that have many values. We can use the logic of Pearson's r, even though the raw data come from ordinal scales, by working with the *ranks* rather than with the original data. In other words while the raw scores may be ordinal, *the ranks of these scores are interval/ratio*, and hence we can calculate correlation coefficients on these ranks.

The basic logic underlying rho is the same as that for other PRE measures of association, in so far as it tries to predict the ranking of pairs of cases on the dependent variable given their ranking on the independent variable. To illustrate the calculation of rho, we will work through the following hypothetical example. A physiotherapist uses a new treatment on a group of patients and is interested in whether their age affects their ability to respond to the treatment. After taking into account a number of other variables, such as the severity of the injury, each patient is given a mobility score out of 15, according to his or her ability to perform a number of tasks.

The results of the study are shown in Table 12.6, along with the rank of each person in terms of each variable. Notice in Table 12.6 that Jordan and Alana had the same mobility score so they each are assigned the average rank of 7.5. To calculate the value for rho we first calculate the difference in rank for each person, D, and then square these differences, D^2. The last step is to enter these results into the equation for rho, which produces a rank-order correlation coefficient of –0.8.

Table 12.6 Calculating rank differences

Patient	Age	Rank on age	Mobility score	Rank on mobility	Rank difference, D	D^2
Danielle	23	1	14	15	1-15 = -14	196
Christine	25	2	15	16	2-16 = -14	196
Leanne	28	3	12	13	3-13 = -10	100
Marie	30	4	8	5	–1	1
Erin	35	5	13	14	–9	81
Ben	37	6	10	10	–4	16
Luke	38	7	11	12	–5	25
Sophie	39	8	8	5	3	9
Elli	40	9	10	10	–1	1
Jordan	41	10	9	7.5	2.5	6.25
Timothy	45	11	10	10	1	1
Alana	50	12	9	7.5	4.5	20.25
Amanda	52	13	7	3	10	100
Lisa	55	14	8	5	9	81
Stacey	60	15	4	1	14	196
Chloe	62	16	6	2	14	196
						$\Sigma D^2 = 1225.5$

$$r_s = 1 - \frac{6\Sigma D^2}{n(n^2 - 1)} = 1 - \frac{6(1225.5)}{16(16^2 - 1)} = -0.8$$

Spearman's rho is a PRE measure, and therefore has a concrete interpretation. A value of 0.8 indicates a strong correlation between these two variables, and the negative sign indicates that this is a negative correlation. In other words, increase in age strongly reduces the effect of the treatment. The older the patient, the less benefit received from the program.

Spearman's rho using SPSS

The commands to calculate rho for these data are shown in Table 12.7 and Figure 12.16 along with the output (note that this is also a third way by which we can generate Pearson's correlation coefficient, in addition to the two commands we used above).

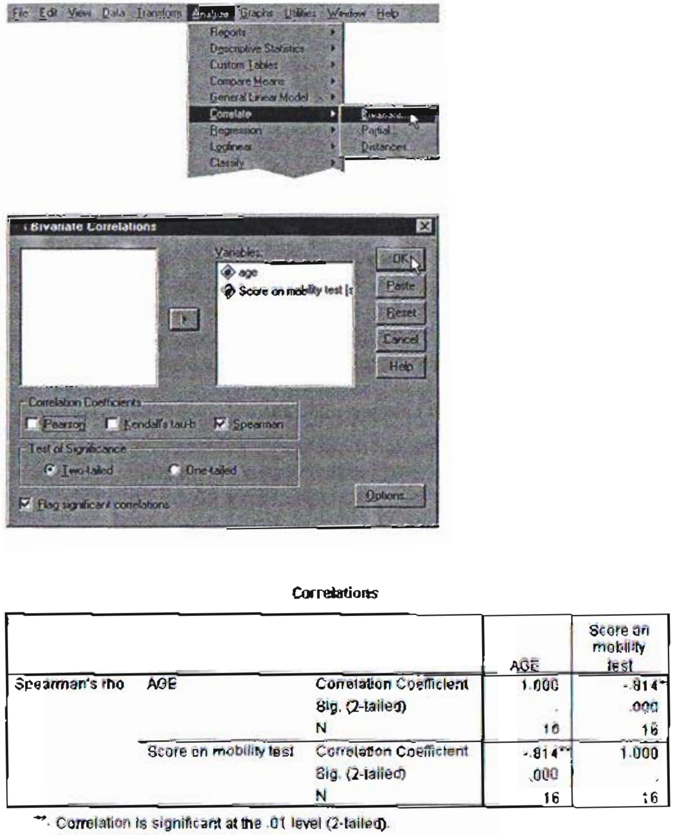
What does all this mean? SPSS calculates the correlation coefficient for each variable with itself and all the other variables we pasted into the target variable list in the dialog box. Looking at the first row of the **Correlations** table we see that age has a correlation coefficient with itself of 1.000; any variable by definition is perfectly correlated with itself. AGE has a correlation coefficient with Score on mobility test of –0.814, which is the same as our hand calculation. Note the minus sign indicating a negative correlation: as age increases, mobility scores decrease.

The second row of the **Correlations** table does the same thing in reverse. It gives the correlation coefficient for Score on mobility test correlated with age which is –0.814. In other words, since rho is a symmetric measure of association it does not matter which way we view the direction of causality (age to mobility or vice versa) since the value calculated will be the same. AGE is also correlated with itself, which produces a perfect correlation of 1.

The table also provides a row of information titled Sig. (2-tailed). This deals with issues we will discuss in Chapter 26, where we will refer to this output. For those who are already familiar with the logic of statistical inference, or have read ahead and are coming back to this chapter, I will quickly explain this portion of the output. Although we have significance of .000 this does not mean a zero significance. The exact probability is less than 5 in 10,000 (i.e. $p < 0.0005$), which SPSS has rounded off to .000. Thus this probability should be read as 'less than 1 in 50,000', which is clearly a significant result. The strong relationship we have detected in the sample is due to such a relationship holding in the population, and not just due to sampling error.

Table 12.7 Generating Spearman's rho on SPSS (file: Ch12-2.sav)

SPSS command/action	Comments
1 From the menu select **Analyze/Correlate/ Bivariate**	This brings up the **Bivariate Correlations** dialog box. You will notice an area called **Correlation Coefficients**, with the box next to **Pearson** selected. This is the default setting. Pearson's coefficient is applicable to interval/ratio data, so is not appropriate here where at least one variable is ordinal
2 Click on **age** in the source variable list	This highlights **age**
3 Click on ▸	This pastes **age** into the **Variables:** target list
4 Click on **Score on mobility test** in the source variable list	This highlights **Score on mobility test**
5 Click on ▸	This pastes **Score on mobility test** into the **Variables:** target list
6 Click on the box next to **Pearson**	This removes ✓ from the tick-box so that this measure of correlation is no longer selected
7 Click on the box next to **Spearman**	This replaces ✓ in the tick-box so that this measure of correlation is selected
8 Click on **OK**	



Correlation Coefficients: □ Pearson □ Kendall's tau-b ☑ Spearman

Test of Significance: ● Two-tailed ○ One-tailed

☑ Flag significant correlations

Correlations

			AGE	Score on mobility test
Spearman's rho	AGE	Correlation Coefficient	1.000	-.814**
		Sig. (2-tailed)	.	.000
		N	16	16
	Score on mobility test	Correlation Coefficient	-.814**	1.000
		Sig. (2-tailed)	.000	.
		N	16	16

**. Correlation is significant at the .01 level (2-tailed).

Figure 12.16 The SPSS **Bivariate Correlations** command, dialog box and output

Example

An instructor is interested in whether the heavy use of formal exams as a form of assessment is biased against students who might perform better under different exam conditions, such as verbal presentations. A group of 15 students is selected and each student is assessed in terms of their verbal presentation skills and in terms of their formal examination skills. These 15 students are rank-ordered on each of these variables as indicated in Table 12.8, along with the calculations we need to generate rho.

Table 12.8 Calculating rank differences

Student	Rank on exam	Rank on presentation	Rank difference D	D^2
1	4	15	−11	121
2	6	3	3	9
3	9	14	−5	25
4	12	9	3	9
5	3	10	−7	49
6	13	11	2	4
7	5	6	−1	1
8	1	4	−3	9
9	14	8	6	36
10	2	1	1	1
11	10	2	8	64
12	7	5	2	4
13	15	7	8	64
14	8	12	−4	16
15	11	13	−2	4
				$\Sigma D^2 = 416$

$$r_s = 1 - \frac{6\Sigma D^2}{n\left(n^2-1\right)} = 1 - \frac{6(416)}{15\left(15^2-1\right)} = -0.26$$

This indicates a weak, negative association between the two types of skills. The instructor might therefore conclude that exams are not a good indicator of other forms of learning skills: students who perform poorly in exams might perform well in verbal presentations. Similarly, students who do well in exams might not relatively do all that well when other skills are required. A mix of assessment methods might give a better indication of students' learning.

Correlation where the independent variable is categorical: Eta

Before completing this chapter on correlation and regression, one last variation of the correlation coefficient is worth discussing. This is **eta**, which is a PRE asymmetric measure of correlation where the dependent variable is measured on an interval/ratio scale and the independent variable is categorical. Eta is therefore extremely useful in situations where we want to compare groups defined by a nominal scale in terms of some interval/ratio scale. An example is comparing males and females in terms of age in whole years. We can calculate a range of univariate descriptive statistics such as the mean and median for each group and compare these (as we discussed in Chapter 9). As an alternative, or in addition to these comparisons, we can use eta to measure the correlation between a person's sex and their age.

Eta will only range between 0 and 1, since the categories of the independent variable are treated as unordered (i.e. essentially nominal), and it is therefore not appropriate to talk of the relationship being either positive of negative in direction. We can generate eta in SPSS under the **Analyze/Descriptive Statistics/Crosstabs/Statistics** sub-command. This is unfortunate, since it is unlikely that we would want to generate a crosstab on data for which eta is applicable; an interval/ratio dependent variable will usually result in a crosstab with far too many rows. It would have been preferable for SPSS to offer eta as an option under the **Analyze/Bivariate/Correlations** command, but this is not the case.

Summary

This chapter has introduced the concepts of correlation and regression. But we have only just skimmed the surface. We could spend a whole course discussing this topic alone, and still not give it adequate treatment. Moreover, you will have noticed that there were many options within SPSS that we did not explore, sticking only to the bare minimum needed to get the results we were after. It is not within the scope of this book to pursue these issues in more detail – we only want to introduce the key concepts and methods. There are many other books that delve into regression analysis in far more depth. Nevertheless, the key ideas hopefully have emerged by sticking to the basics and not elaborating further on more advanced topics. Having digested this much, the task of absorbing the more advanced material may prove a little easier.

Exercises

12.1 Why should we draw a scatter plot of data before undertaking regression analysis?

12.2 What does the Y-intercept of a regression line indicate?

12.3 What is the principle used for drawing the line of best fit through a scatter plot?

12.4 For each of the following regression equations, state the direction of the relationship:

(a) $Y = 30 + 42X$ (b) $Y = 30 - 0.38X$ (c) $Y = -0.5 + 0.38X$
(d) $Y = -0.5$ (e) $Y = -0.5X$

12.5 Graph each of the following equations on graph paper. On your graph indicate the Y-intercept and the slope:

(a) $Y = 30 + 42X$ (b) $Y = 30 - 0.38X$ (c) $Y = -0.5 + 0.38X$
(d) $Y = -0.5$ (e) $Y = -0.5X$

12.6 Explain the difference between the correlation coefficient, r, and the coefficient of the regression line, b.

12.7 (a) Using graph paper draw a scatter plot for the following data:

X	5	6	9	10	10	13	15	18	22	27
Y	35	28	30	22	28	28	20	21	15	18

(b) Looking at the data, what do you expect the sign in front of the coefficient to be (i.e. is there a positive or negative correlation)?

(c) Draw a regression line through these data using the naked eye. Determine the equation for your line, and predict Y for $X = 12$.

(d) Calculate the least squares regression line through these data. What is the least squares estimate for Y when $X = 12$?

(e) What is the sum of the squared errors for your freehand line and the OLS line? Which of these sums is smallest?

(f) Enter these data on SPSS and run the regression command to confirm your results.

12.8 A regression line is plotted through data on life expectancy in years and government expenditure on health care per head of population (in $'000) for a group of developing nations. Life expectancy is considered the dependent variable and expenditure the independent variable. The equation for the regression line is $Y = 40 + 0.7X$.

(a) What will life expectancy be if the government spends no money on health?

(b) What will life expectancy be if the government spends $30,000 per head on health?

(c) Can you say that there is a strong relationship between the two variables?

12.9 A university lecturer in statistics wants to emphasize to her students the value of study to exam performance. The lecturer monitors the amount of time in minutes that all 11 students in the class spend in the library per week, and the final grades for each student. The figures are recorded in the following table:

Library time (minutes)	Exam score
41	52
30	44
39	48
48	65
55	62
58	60
65	74
80	79
94	80
100	90
120	86

The lecturer analyzes these data using the regression command in SPSS. Enter these data yourself into SPSS and from the output answer the following questions:

(a) Write down the equation for the OLS regression line for these data.
(b) What is the strength and direction of the relationship between these variables?
(c) Will a student who spends no time in the library fail?
(d) A student wants to use this information to work out the minimum amount of time the student needs to spend in the library in order to get a bare pass grade (50). What is the minimum amount of time he needs to spend in the library? Can the student be very confident in this prediction? What is the problem with using the regression line for such a purpose?
(e) Draw a scatter plot of these data to check that it was appropriate for the lecturer to use linear regression.
(f) Use these data to calculate by hand the same values presented in the SPSS output and check the results are the same.

12.10 A real-estate agent wants to explore the factors affecting the selling price of a house. The agent believes that the main factor explaining differences in selling prices is house size. In this model which variable is cast as the independent and which is the dependent? The agent collects data on these two variables, with the results:

Selling price ($,000)	House size (squares)
260	20
240	15
245	20
210	13
230	18
242	14
295	28
235	16
287	24
252	20
270	23
275	25

(a) Calculate all the relevant statistics needed to assess the agent's model, both by hand and by SPSS.
(b) In SPSS create another column to enter the data for the selling price of houses, but this time enter the data without rounding to the nearest $,000, i.e. enter 250,000, 240,000, 243,000, etc Recalculate your regression statistics. What, if anything, is different. Interpret any changes.

12.11 Enter into SPSS the data we used in the example above relating visitors to a museum with the daily temperature, and generate all the relevant descriptive statistics.

12.12 A study investigates the factors that may lead to a reduction in the number of working days lost due to illness at a certain factory. Ten people are studied and the following information about their respective number of hours of exercise per week and the number of working days they were absent due to illness is recorded in the following table:

Hours of exercise	Days lost
3	12
8	10
1	10
0	15
0	18
4	7
7	7
2	14
5	9
0	16
9	8
3	10

(a) What is the correlation between these two variables?
(b) If someone exercises 8 hours a week, how many days do you predict that they will be absent from work due to illness over the course of the year?

12.13 Using the **Employee data** file can we say that beginning salary is a good predictor of current salary?

12.14 From the **World95** data file that comes with SPSS, a social worker finds the correlation between female life expectancy and birth rate per 1000 people to be –0.862. What does this mean? Use SPSS to determine the full regression equation for this relationship and interpret the results.

12.15 Eleven countries are rank-ordered in terms of two variables: infant mortality rate and expenditure on the military as a proportion of national income. These ranks are:

Country	Rank on infant mortality	Rank on military spending
A	9	8
B	4	5
C	6	6
D	2	2
E	7	11
F	3	4
G	10	7
H	5	3
I	8	9
J	1	1
K	11	10

(a) Calculate Spearman's rank-order correlation coefficient for these data. What can you conclude about the relationship between these variables?
(b) Enter these data into SPSS and calculate rho to check your results.

12.16 Does price reflect quality? When people pay more for something are they actually getting something better? To assess this, a number of expert judges are asked to taste and rank 15 wines whose identity and price are not disclosed to them. The wine rated 15 is considered the highest quality, while the wine scoring 1 is considered the most inferior. The rank of each wine according to the judges and its price is listed in the following table:

Quality	Price
1	3.00
2	4.00
3	5.50
4	5.90
5	11.99
6	6.80
7	7.50
8	7.00
9	18.00
10	3.50
11	11.30
12	12.00
13	7.00
14	4.50
15	13.00

Calculate Spearman's rho to assess the nature of any relationship between quality and price. Check your answer by calculating rho with SPSS.

12.17 A group of ten runners is interested in whether running ability is associated with age. They record their ages in years and also their order in finishing a run. The results are:

Name	Age	Place
Kenny	54	4
Schuey	46	1
Scotty	29	6
Pat	28	2
Garth	25	3
Les	36	9
Michael	15	10
Garry	26	5
Linda	38	8
George	34	7

Calculate Spearman's correlation coefficient to assess whether there is any relationship between age and running ability. Enter these data on SPSS to assess your answer.

13

Multiple regression

In Chapter 12, Exercise 12.10, we considered the following problem. A real-estate agent wants to explore the factors affecting the selling price of a house. The agent collects data on these two variables for 12 houses, with the results given in Table 13.1.

Table 13.1 House size and selling prices

Selling price ($,000)	House size (squares)
260	20
240	15
245	20
210	13
230	18
242	14
295	28
235	16
287	24
252	20
270	23
275	25

The purpose of the analysis is to determine the selling price, which is the dependent variable. The agent believes that the main factor explaining the variation in selling prices is the variation in house sizes. As discussed in Chapter 1, we call this a **model** of the factors determining the sale price of a house, since it is a theoretical depiction of a relationship that may or may not hold up to empirical scrutiny. Let us compare for example two houses from the sample, such as the house that sold for $252,000 and the one that sold for $230,000. Indeed, we find the more expensive house is also the larger house, so that these two houses seem to be consistent with the agent's model. Does this relationship hold true for all 12 houses?

A simple regression analysis on these data from Exercise 12.10, using the method of ordinary least squares, produces the following results:

$$Y = 157 + 4.88X$$

$$r^2 = 0.85$$

On the basis of these results we can conclude the following:

- There is a positive relationship between house size and selling price.
- For every one square increase in house size the selling price increases by $4880.
- The relationship is strong and highly reliable for making predictions.
- The variation in house size does not *perfectly* predict selling price. The coefficient of determination is high (0.85), but not equal to one. Therefore other factors also affect the sale price of houses in our sample.

This last point means that on a scatter plot of the data in Table 13.1 not all the data points lie right on the regression line, as evident in Figure 13.1, which presents an SPSS-generated scatter plot with the regression line for these data.

Graph

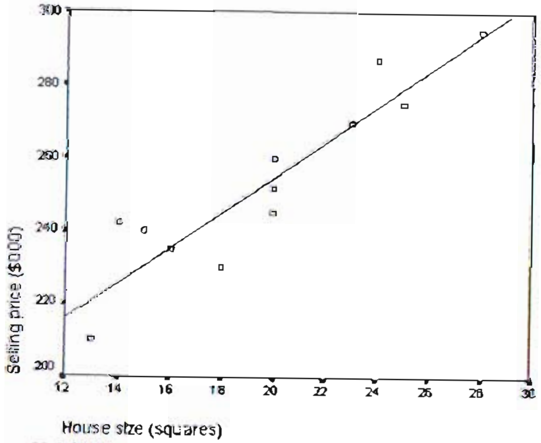


Figure 13.1 SPSS scatter plot with regression line

The actual sale price for any given house can in fact be expressed by the following equation:

$$\text{Selling price} = a + b(\text{house size}) + e$$

This equation states that the sale price of houses varies primarily because of differences in their size, but also because of random factors, represented by the **error term** (e). The error term expresses the difference between what we predict the price of a house will be, given its size and what it actually sells for.

We should stop for a moment and be clear about what we mean by the 'error term' and 'random variation'. We can all agree that many factors affect the specific price at which a house sells. It would not be hard to provide a long list of such factors. Our bivariate model of the sale price argues that among all these factors there is one variable – house size – that plays a major role in determining sale price and it does so in a systematic and consistent way. This is why we have singled out the variable 'house size' and given it an explicit position in the equation. But we also do not want to ignore all the other factors. The error term bundles up all these other factors, factors that affect the sale price of houses in a haphazard, unsystematic way. One house may have sold at a high price because the estate agent was particularly aggressive in his or her sales pitch; another house may have sold well because the buyers particularly liked the color scheme; still another may have sold for a low price because that particular vendor had to sell quickly in order to repay a bad debt. It is because these and other factors spring up randomly from one sale to the next that we do not treat them as separate independent variables, but allow the error term to capture their collective influences. These random factors sometimes cause the sale price to be higher than what we predict based on knowledge of the house's size, and sometimes they cause the sale price to be below the predicted value. Knowing a house's size will allow us to predict a value for sale price that will be close to the mark, but we concede that for any given house the effect of these random factors will mean that the actual sale price will not necessarily equal the predicted sale price.

Introduction to multiple regression

We may, however, regard the bivariate model as overly simplistic. We may feel that there are factors other than house size that are not random, but which operate in a systematic way to cause sale prices of houses to vary independently of their size. In other words, if we compare two houses *of the same size*, the difference in their respective sale prices is not only due to random factors such as those we just discussed. We have three houses in our sample, for example, that are each 20 squares in size. One sold for $260,000, the other $245,000, and the third for $252,000. Why the differences in sale price? We could put faith in our bivariate model and argue that random factors explain these differences, or we could argue that another model that allows for the operation of other variables to systematically affect sale price offers a better explanation.

We may, for instance, believe that the *age* of a house also (partly) explains its sale price. That is, the age of a house is not a variable that may occasionally impact on the sale price of a house, but instead is a common factor that *systematically* impacts on the prices that houses sell for. Our new model may hold that it is reasonable to expect that the older the house the cheaper will be its price; we expect a *negative relationship* between house prices and age.

If we suspect this to be the case, we need to extend our regression analysis to include the operation of this other variable, much in the same way that in the previous chapter we extended our simple bivariate crosstab analysis to account for the possible effects of third variables. When working with interval/ratio data (as we have here) this is the task of **multivariate regression**.

Multivariate regression investigates the relationship between two or more independent variables on a single dependent variable.

With this new multivariate model in mind we collect data in Table 13.2 for the ages (in years) of the 12 houses we originally surveyed (this example is adapted from A. Slevanathan *et al.*, 1994, *Australian Business Statistics*, Melbourne: Thomas Nelson).

Table 13.2 Selling price, house size, and age of 12 houses

Selling price ($,000)	House size (squares)	Age (in years)
260	20	5
240	15	12
245	20	9
210	13	15
230	18	9
242	14	7
295	28	1
235	16	12
287	24	2
252	20	5
270	23	5
275	25	5

Generally we can express the relationship between any dependent variable and any (k) number of independent variables in the following way:

$$Y = a + b_1X_1 + b_2X_2 + \ldots + b_kX_k + e$$

For the specific example we are investigating we therefore can represent the model of the relationship in the following terms:

$$\text{Selling price} = a + b_1(\text{house size}) - b_2(\text{age}) + e$$

In other words, we believe that the sale price of a house is pulled in one direction or another by its age and its size. We expect an old house that is also relatively small to have its price pulled in a downward direction through the independent operation of both age and size. Conversely, we expect a new house that is also relatively large to have its price pulled upwards. In other instances, house size and age may be pulling in opposite directions.

In our equation we still allow for random factors to have an influence so that age and price do not *exactly* determine sale price in every instance. But if this multivariate model is a better explanation of house selling prices than the bivariate model, *the amount of variation left over to be explained by the error term will be much smaller than in the bivariate model we started with.* If, on the other hand, introducing age into the equation does not reduce the proportion of sale price variation attributed to the error term then knowing a house's age does not improve our ability to predict its sale price. We have information on a variable that is not helpful in statistically accounting for the dependent variable.

The task of multivariate regression is to try and apportion the variation in house prices to each of these competing 'pulls' on the dependent variable. Does one dominate the determination of selling price, such that we can say age or size is clearly more important, or do they have similar influences? And what is the role left over to random factors? Multivariate analysis, through the calculation of the **regression coefficients** and the **partial correlations** for each variable, gives us precise measures of the respective influence of these independent variables on the dependent variable.

It is possible to use formulas to calculate the regression coefficients between each of these independent variables and the dependent variable. However, these techniques are very cumbersome, and with large data sets, overwhelmingly time consuming. No one today would consider conducting multiple regression by hand. To save ourselves the hassle we will leave it to SPSS to conduct the calculations, and we will simply interpret the results.

Multiple regression with SPSS

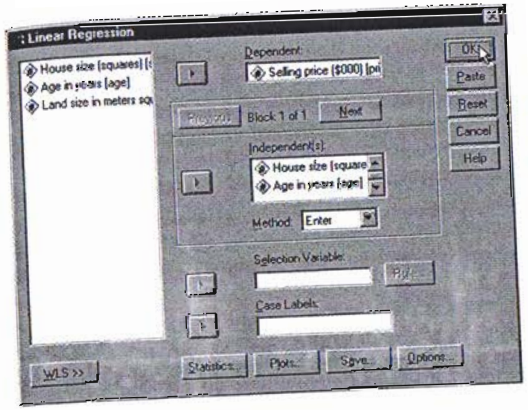
The procedure for calculating the equation statistics for multiple regression is the same as that for simple bivariate regression from Chapter 12, except for the fact that we paste more than one variable into the **Independent(s)** variable target list. This procedure is presented in Table 13.3 and Figure 13.2. Figure 13.2 also presents the output from this procedure.

Table 13.3 Multiple regression using SPSS (file: **Ch13.sav**)

SPSS command/action	Comments
1 From the menu select **Analyze/Regression/Linear**	This brings up the **Linear Regression** dialog box
2 Click on **Selling price** in the source variable list	This highlights **Selling price**
3 Click on the ▸ that points to the **Dependent:** target variable list	This pastes **Selling price** as the dependent variable
4 Click on **House size** in the source variable list and while holding down the **Shift** key click on **Age in years**	This highlights both **House size** and **Age in years**
5 Click on the ▸ that points to the **Independent(s):** target variable list	This pastes both **House size** and **Age in years** as the independent variables
6 Click on **OK**	

A great deal of information is presented in the SPSS regression output (some of it repeated several times); we will concentrate on just the most important parts.

- The table headed **Variables Entered/Removed** provides a simple verbal description of the model(s) we are estimating. It is possible in SPSS to run several multiple regressions simultaneously, using different combinations of independent variables to see which combination 'best' explains the variation in the dependent variable. Here we have only estimated one model, called Model 1, which uses the variables Age in years and House size in squares as the predictors of the dependent variable, Selling price ($000). We will detail the various options that SPSS provides for entering the selected independent variables into the regression model below.



Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Age in years, House size (squares)		Enter

a. All requested variables entered.

b. Dependent Variable: Selling price ($000)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.959[a]	.919	.901	7.80

a. Predictors: (Constant), Age in years, House size (squares)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6248.759	2	3124.379	51.298	.000[a]
	Residual	548.158	8	60.906		
	Total	6796.917	11			

a. Predictors: (Constant), Age in years, House size (squares)

b. Dependent Variable: Selling price ($000)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	224.290	26.222		8.553	.000
	House size (squares)	2.578	.973	.487	2.650	.026
	Age in years	-2.974	1.076	-.503	-2.764	.022

a. Dependent Variable: Selling price ($000)

Figure 13.2 SPSS **Linear Regression** dialog box and output

- The second table, headed **Model Summary**, provides the correlation coefficient, which indicates the strength of the relationship between the combination of independent variables in the model and the dependent variable. The value for R of .959 indicates a very strong relationship. R is the multivariate equivalent for the bivariate correlation coefficient, *r*. Of more interest is the value for Adjusted R Square (the coefficient of multiple determination) which is .901. When we used the bivariate model to explain selling price (correlating it only with house size) the value for the coefficient of determination was 0.85. When we use both house size *and* the age of the house to predict selling price the coefficient of determination rises to .901. This indicates that our ability to explain (or predict) the selling price of a house has increased when we also have information about its age as well as its size. Part of the variation in sale price that we had previously attributed to random factors is actually due to the systematic effect of age. Note that in multiple regression we use the Adjusted R Square rather than the simple R Square, since the latter may overestimate the extent to which our sample data explain the variance in the dependent variable, partly because it is affected by the number of variables included in the model.
- The next table, headed **ANOVA**, contains inferential statistics that allow us to make an inference from the sample to the population of all houses. We are at the moment concentrating just on the descriptive statistics for the sample, so we will skip this part of the output for now, and return to it in the discussion below.
- The table headed **Coefficients** provides the elements of the regression equation we have estimated, which can be written:

$$\text{Selling price (\$,000)} = 224.29 + 2.578(\text{house size in squares}) - 2.974(\text{age in years})$$

When reading a regression equation it is important to keep in mind the units of measurement in which the variables have been measured. Here we see that for every one square increase in house size, the selling price increases by $2578. Independently of this relationship we also find that for every one year increase in the age of a house, its selling price decreases (note the negative sign) by $2974.

To give this slightly more practical meaning let us assume that we are now presented with a house that is going up for sale. We measure it as being 15 squares in size and also 5 years old. What do we predict it will sell for? According to the equation it will sell for $248,090:

$$\begin{aligned}\text{Selling price (\$,000)} &= 224.29 + 2.578(15) - 2.974(5)\\ &= 224.29 + 38.67 - 14.87 = \$248.09\end{aligned}$$

Of course we do not expect $248,090 to be the exact price realized when the house is actually sold, because random factors will still play a role. But given the high value of the coefficient of determination, these random factors should not cause the actual sale price to deviate much from this predicted value.

- It is difficult to use the regression coefficients to assess the *relative* importance of each independent variable in determining the value of the dependent variable, since each independent variable is measured with different units (one is measured in years, the other in squares). If we measured house size in another unit, such as square feet, the regression coefficient for this variable will be different because of the unit of measurement. In other words, we cannot say that, because the coefficient for house size is 2.578, whereas the coefficient for age is –2.974, age is a more powerful force acting on selling price. The **Coefficients** table therefore provides a column of **Standardized Coefficients**. Without going into the details of how these standardized coefficients, also called beta-weights, are calculated, we simply note that they 'wash out' the effect of the units of measurement. We can see that age (–.508) has a slightly stronger 'pull' on sale price than house size (.487).

Table 13.4 summarizes the role that the various measures generated by SPSS play.

Table 13.4 Interpretation of SPSS output

Regression coefficient	Allows us to make predictions for the dependent variable based on the values of the independent variables, in terms of the original units of measurement
Standardized coefficient	Allows us to distinguish the relative importance of each independent variable in determining the value of the dependent variable
R	Indicates the strength of the relationship between the combination of independent variables and the dependent variable
Adjusted R-squared	Indicates the amount of variation in the dependent variable explained by the combination of independent variables in the model, thereby indicating whether the model is a good predictor of the dependent variable

Testing for the significance of the multivariate model

You may have noticed in the output some of the inferential statistics we will come across in later chapters. Although we have yet to deal with inferential statistics, we will note their general meaning here so that we have a relatively complete coverage of multiple regression output. After covering these statistics in more detail in later chapters you may wish to return to this section. Inferential statistics tell us whether we can generalize from a sample result, such as that in our example, to the population from which the sample is drawn. Will the relationship between selling price and house size and age still hold if we surveyed all houses sold in the area?

The critical information for this inference test is contained in the table headed **ANOVA**. SPSS conducts an *F*-test on the whole model, which tests the hypothesis that the correlation coefficients for all the variables included in the model are zero. In this example, the *F*-statistic for the model has a significance level of 0.000. This tells us that at least one of the correlations between each of the independent variables and the dependent variable is not equal to zero in the population.

This conclusion is confirmed in the **Coefficients** table, where we can see that the *t*-statistics for each independent variable are significant at the 0.05 level. Thus we use the *F*-test to see whether at least some of the independent variables in our model are significant, and the *t*-statistics for each individual variable indicate which ones are significant.

Alternative methods for selecting variables in the regression model

In multiple regression analysis we enter a 'block' of independent variables that we want to model in the **Independent(s):** list. Depending on the number of independent variables in this block, there will be numerous combinations of these variables that could be included in the regression model. In our example, we only have two independent variables with which to predict selling price: age and house size. Two independent variables, though, give us three potential models: with each of the variables on their own and with the two variables together.

The **Method:** option in the **Linear Regression** dialog box (Figure 13.2) provides alternative means by which the variables in the block of independents are included in the regression model:

- **Enter**. This is the default setting and produces a single regression model that includes all the variables in the block.
- **Stepwise**. This adds *or* removes variables in a number of steps, depending on the extent to which such addition or removal will increase R-squared. In essence, it finds the 'best' combination of variables in the block that maximize R-squared. The Stepwise method will be discussed in more detail below. It is especially useful in exploratory analysis, or where predictive accuracy as such is desired, but can be used atheoretically in a 'fishing' expedition to discover statistical relationships that have no substantive basis.

- **Forward**. This is a stepwise method, where variables are added based on their relative semi-partial correlation coefficients with the dependent variable.
- **Backward**. This is a stepwise method, where all the variables in the block are entered in the model in one step and those that do not make a significant contribution to predicting the dependent variable are then removed.
- **Remove**. Used only in hierarchical regression, which we will cover below. All the variables in the block are removed in one step, based on their collective impact on the R-Squared.

Stepwise regression

The previous SPSS example used the **Enter** method for generating the regression model, which uses all the variables in the independents block. This method of variable inclusion is generally favored, since it requires us to think in advance of the statistical analysis what our theory suggests about the nature of the relationships in which we are interested.

There are instances, however, where multiple regression analysis is used for more practical purposes than testing theoretical models. We may be purely interested in having a model with predictive accuracy, without being interested too much about the underlying theoretical understanding of why the model has such predictive accuracy. Similarly, our theory may suggest a small set of potential independent variables, but is not prescriptive as to which members of this small set of variables will actually make up the model in a given context. Having determined a 'short-list' of variables we believe may influence the dependent variable (on the basis of theory or past research), we can then use the stepwise regression method we are about to detail to select the specific variables that actually do have significant influence.

For example, our real-estate agent in the example we have used is probably not too interested in the underlying causal structure of variables that determine the sale price of a house. She may just want to know with the highest confidence what the likely sale price will be, given certain features of a house. Imagine that she has observed the calculation above and yet believes that we have still left out other important factors that determine the selling price of houses in the area. Despite the high explanatory power of our model with only two independent variables, the agent may argue that our ability to predict the sale price of houses will be even further improved if we include the size of the land as *another* independent variable. The agent therefore goes back and gathers the data for the 12 houses we are analyzing, measuring the land area in meters squared (Table 13.5).

Table 13.5 Selling price, house size, age, and land size of 12 houses

Selling price ($,000)	House size (squares)	Age (in years)	Land size (meters squared)
260	20	5	420
240	15	12	640
245	20	9	600
210	13	15	590
230	18	9	700
242	14	7	720
295	28	1	624
235	16	12	590
287	24	2	710
252	20	5	630
270	23	5	700
275	25	5	710

With three independent variables that can be used in various combinations, we now have seven models to potentially explain the sale price of houses:

$$\text{selling price} = a + b_1(\text{house size}) + e$$
$$\text{selling price} = a + b_1(\text{age}) + e$$
$$\text{selling price} = a + b_1(\text{land size}) + e$$
$$\text{selling price} = a + b_1(\text{house size}) + b_2(\text{age}) + e$$
$$\text{selling price} = a + b_1(\text{house size}) + b_2(\text{land size}) + e$$
$$\text{selling price} = a + b_1(\text{age}) + b_2(\text{land size}) + e$$
$$\text{selling price} = a + b_1(\text{house size}) + b_2(\text{age}) + b_3(\text{land size}) + e$$

We discussed above that the way we judge whether a variable adds to the explanatory power of a model is by looking at the impact its inclusion has on the value for R-squared. If the value for R-squared increases significantly when a variable is added to the model, then the extra information provided by this variable increases the model's ability to explain the variation in sale price.

One way to decide between the various models is to undertake separate linear regressions based on the particular combination of independent variables we want to include. We can then compare the R-squared values to see the extent to which our ability to explain the variation in sale price is maximized by each combination of independent variables. For example, if we do conduct a multiple regression including land size R-squared is 0.922, which is the same as that for the model with only age and house size. In other words, land size does not increase our ability to explain selling prices; the time and effort in measuring this variable is wasted.

The problem with this approach is that it is tedious to run separate regressions for each of the possible models we can construct. It is also difficult to judge how much of an increase in R-squared justifies the inclusion of a variable in our model. A way of assessing all the possible combinations of variables is to use the variable inclusion method of **stepwise regression**, which determines the combination of possible independent variables that best explains the dependent variable. It does this by adding in and taking out variables from the calculations according to whether each makes a statistically significant change to the value of R-squared.

But before illustrating how this is done, we need to again raise a word of caution. We can potentially provide SPSS with a whole list of variables that may or may not affect a particular dependent variable, and then run a stepwise regression on SPSS to find the 'best' combination. This kind of fishing expedition is not appropriate since it selects variables based on statistical results alone. We should try, where appropriate, to be guided by our theories of the world and/or past research as to the variables to consider for analysis.

To choose the stepwise option we follow the procedures listed in Table 13.3, but also click on the **Method:** option in the **Linear Regression** dialog box (Figure 13.3). The output is presented in Figure 13.4.

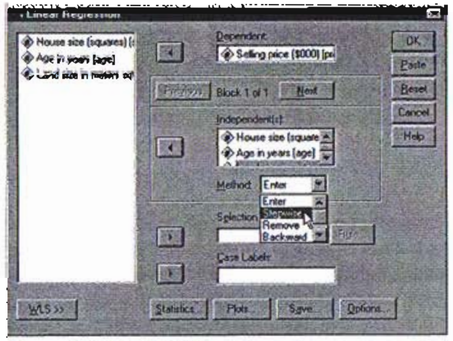
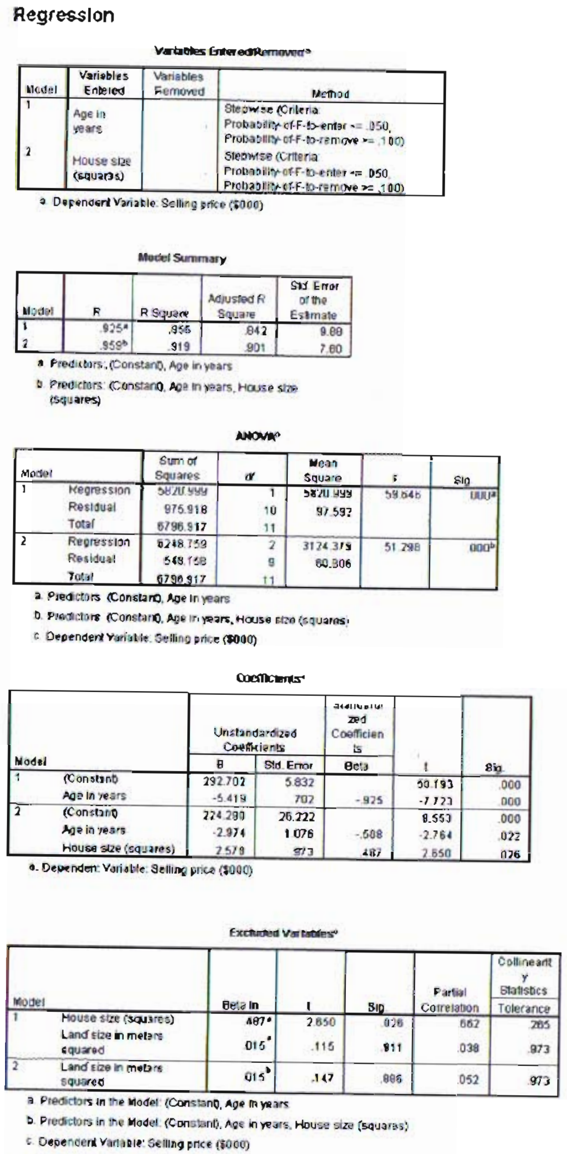


Figure 13.3 The SPSS Stepwise option



Regression

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	Age in years	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100)
2	House size (squares)	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100)

a. Dependent Variable: Selling price ($000)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.925^a	.856	.842	9.88
2	.959^b	.919	.901	7.80

a. Predictors: (Constant), Age in years

b. Predictors: (Constant), Age in years, House size (squares)

ANOVA^c

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5820.999	1	5820.999	59.646	.000^a
	Residual	975.918	10	97.592		
	Total	6796.917	11			
2	Regression	6248.759	2	3124.379	51.298	.000^b
	Residual	548.158	9	60.906		
	Total	6796.917	11			

a. Predictors: (Constant), Age in years

b. Predictors: (Constant), Age in years, House size (squares)

c. Dependent Variable: Selling price ($000)

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	292.702	5.832		50.193	.000
	Age in years	-5.419	.702	-.925	-7.723	.000
2	(Constant)	224.290	26.222		8.553	.000
	Age in years	-2.974	1.076	-.508	-2.764	.022
	House size (squares)	2.579	.973	.487	2.650	.026

a. Dependent Variable: Selling price ($000)

Excluded Variables^c

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	House size (squares)	.487^a	2.650	.026	.662	.285
	Land size in meters squared	.015^a	.115	.911	.038	.973
2	Land size in meters squared	.015^b	.147	.886	.052	.973

a. Predictors in the Model: (Constant), Age in years

b. Predictors in the Model: (Constant), Age in years, House size (squares)

c. Dependent Variable: Selling price ($000)

Figure 13.4 The SPSS **Stepwise** regression output

In the first table headed **Variables Entered/Removed** we see that SPSS has generated two models from the three variables we suggested: one with **Age in years** only (our original bivariate model) which SPSS calls **Model 1**, and another with age and **House size (squares)** which SPSS calls **Model 2**. The rest of the output is exactly the same as that we generated separately before for each of these models; here we have the two models presented in the same analysis.

The new part of the output is the last table, headed **Excluded Variables**. This tells us that on the basis of the *F*-test on changes in R-squared, land size is not a useful variable to include in any of the models. All of this means that complicating the model by adding this new variable does not 'buy' us any more accuracy in terms of estimating the dependent variable. Parsimony suggests that we should leave it out of the picture.

Extending the basic regression analysis: Adding categorical independent variables

Regression analysis is usually conducted with variables measured on interval/ratio scales, but it is possible to include in a number of ways categorical independent variables in a regression analysis.

1. A simple bivariate scatterplot can incorporate the possible impact of a third categorical variable by running *for each category of the third variable* separate plots with regression lines and statistics. For example, I may be interested in the relationship between years of education and the salary which employees are paid when they commence work. I might believe that the nature of such a relationship is affected by the sex of the employee, such that men receive a higher pay-off for education than women. To account for this possibility I can run separate regression analyses; one relating education and starting salary for males and one relating education and starting salary for females. I can then compare the regression coefficients to assess the extent to which an increase in education will 'buy' an increase in starting salary for men as compared to women. In SPSS this can be done through the **Graphs/Interactive/Scatterplot** command, by placing the relevant categorical variable (such as sex of employee) into the **Panel Variables:** box, and then under the **Fit** option selecting **Subgroups** under **Fit Lines For**.
2. Where the categorical variable is binomial it can be added directly as an independent variable in multiple regression. Thus in the example we just discussed, comparing male and female employees in terms of starting salaries, sex of employee can be added along with years of education in the same regression model. In analyzing the regression coefficients it is necessary to keep in mind the coding scheme for the categorical variable so that the pattern of any relationship can be properly interpreted. For example, if we find the coefficient for sex of employee to be positive in value (greater than zero), and females are coded 1 and males coded 2, then being male rather than female produces a positive impact on starting salary for any given level of education. If on the other hand males were coded as 1 and females coded 2, a positive sign will indicate that being female increases starting salary, thereby confounding our expectations. The actual value of the regression coefficient measures the amount that salary increases for one sex over the other.
3. Where the categorical variable is multinomial (i.e. has more than two categories) it can be indirectly included in a multiple regression by first transforming it into a number of **dummy variables**. An example is the best way to understand the nature of dummy variables. Assume we want to asses the impact that ethnicity has on starting salary along with education level. Ethnicity has three categories: 'English-speaking', 'non-English speaking European', and 'other non-English speaking'. From this one variable I can create three dummy variables:

   - English-speaking or not,
   - non-English speaking European or not, and
   - other non-English speaking or not.

   In other words, for each category of the original variable, a separate dummy variable is created indicating whether a case falls into that category (coded with 1) or not (coded with 0). In SPSS dummy variables can be created through the **Transform/Recode** command, whereby the value for the relevant category of the old variable is recoded as 1 and all other old values are coded as 0. The dummy variables are then added as separate independent variables in the multiple regression and the results are assessed in a similar manner to the interpretation of a binomial categorical independent variable. If we find, however, that our regression model is largely comprised of such dummy variables, rather than variables measured on interval/ratio scales, it might be worth using multivariate analysis better suited to such data, such as logistic regression.

Further extensions to the basic regression analysis: Hierarchical regression

There are many more elaborate extensions of the basic regression model such as Cox regression and two-stage least squares regression. Introducing these forms of regression analysis will take us beyond the aims of this text; any advanced statistics text will provide a detailed guide for those wishing to pursue these topics. One extension of the basic multiple regression analysis is worth mentioning, though, since it appears as an option in the **Linear Regression** dialog box which we have already covered in some depth. This is **hierarchical regression**, whereby separate blocks of independent variables are entered into the regression analysis in sequential stages. Hierarchical regression is used where we believe, on theoretical grounds, that there is a particular causal structure among groups (blocks) of independent variables. An example from G. Francis, 2004, *Introduction to SPSS for Windows*. Sydney: Pearson Education, pp.120–2, illustrates this type of regression (this text should be consulted for a more detail explanation of this procedure and the associated SPSS output). In predicting English achievement of students, we believe that socio-economic status, sex, attentiveness in class, and frequency of English homework are all useful predictors. However, we believe that socio-economic status and sex of students are 'background' variables that affect the behavioural variables of attentiveness and homework frequency, which then affect English achievement.

To enter these two blocks, each comprising two variables, in an hierarchical order, we enter the first block of background variables into the **Independent(s):** list in the **Linear Regression** dialog box, then click on Next, and then enter the second block of behavioural variables.

The assumptions behind multiple regression

While multiple regression is a powerful tool for assessing the impact of many independent variables on a dependent variable, there are a number of assumptions behind it that limit its applicability. All the assumptions we covered in the discussion of bivariate regression in Chapter 12 still apply in the case of multiple regression.

1. The dependent variable is measured on an interval/ratio scale.
2. The independent variables are measured on interval/ratio scales or are binomial (although some argue that ordinal scales with many points will produce valid results).
3. Observations for each case in the study are independent of the observations for the other cases in the study.
4. The relationship between the independent variables and the dependent variable is linear.
5. The **error terms are normally distributed** for each combination of the values of the independent variables.
6. The error terms are homoscedastic (i.e. are of equal variance).

To this list, though, we must add another very important assumption. Multiple regression *assumes* that each of the independent variables is independent of each other (there is no **multicollinearity**). In the example we used in this chapter for predicting sale price of houses, this can be depicted as shown in Figure 13.5.

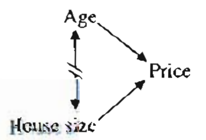


Figure 13.5 The assumption of no multicollinearity

Age and house size each affect price *but do not affect each other*. This may seem fairly reasonable for these particular variables: if a house was suddenly enlarged, this would not also suddenly make it older or younger! Similarly, as a house grows older it does not usually grow larger or smaller.

This assumption underlying multiple regression makes it a little more restricted than the multivariate techniques we looked at in the previous chapter. There we used multivariate analysis to *determine* which model out of a range of models best explains the relationship between three or more variables. With regression analysis we *assume* a specific model.

Exercises

13.1 The study described in Exercise 12.12 investigating the factors that cause employees to be absent due to illness at a certain factory is extended to include data on the employees' ages.

Hours of exercise	Days lost	Age in years
3	12	36
8	10	35
1	10	54
0	15	42
0	18	41
4	7	25
7	7	32
2	14	39
5	9	43
0	16	29
9	8	32
3	10	50

(a) Which variable is the dependent variable?
(b) What do you expect the sign in front of the independent variables to be?
(c) Enter these data into SPSS and conduct a multiple regression. What is the regression equation?
(d) Has the inclusion of age added anything to our ability to predict number of working hours lost due to illness?

13.2 In Exercise 12.14 you were asked to generate, from the **World95** data file that comes with SPSS, the regression equation relating female life expectancy and birth rate per 1000 people. Are there any other variables in the data file that you feel should be included in the equation? Test your model by running the appropriate regression on SPSS.

13.3 Using the **Employee data** file, select variables you think will be good predictors of current salary, and conduct a stepwise regression to see which ones are actually worth including in your model.

PART 4

Inferential statistics: Tests for a mean

14
Sampling distributions

So far, we have looked at ways of summarizing information; we collect measurements from a set of cases and then reduce these hundreds (sometimes thousands) of numbers into descriptive statistics such as the mean or standard deviation. We have seen how such descriptive statistics provide a useful summary of the overall distribution, drawing out those features of a distribution that will help us answer our research question.

If the set of cases from which we take a measurement includes all the possible cases of interest – the **population** – data analysis ends with the calculation of these descriptive measures. An investigation that includes every member of the population is a **census** and the descriptive statistics for a population are **parameters**.

A parameter is a statistic that describes some feature of a population.

When using mathematical notation, parameters are denoted with Greek symbols, such as μ and σ for the population mean and standard deviation respectively.

Sometimes we actually have information about the whole population of interest, such as when a government agency conducts a census of people and can tell us the age distribution of the entire population at a certain date. Other times we don't have information about the population – it is out there but we just can't get our hands on it. Therefore, in research we often work with a smaller sub-set; a **sample** of the population. The descriptive measures used to summarize a sample are **sample statistics**. These sample statistics are denoted, in mathematical shorthand, with Roman letters: $\bar{X}$ for the sample mean, and s for the sample standard deviation.

There are several reasons why we may draw a sample rather than conduct a complete census:

- Samples are usually cheaper and quicker.
- It is sometimes impossible to locate all the members of a population, either because a complete list of the population is unavailable, or because some of its members are difficult to reach or unwilling to participate in the study.
- Research sometimes destroys the units of analysis so that a census would destroy the population. For example, a factory might be interested in a quality control check of the batteries it produces. Testing that the products have sufficient battery life may involve running the units down until they are out of power, a process that will cause bankruptcy if it is applied to all the batteries that the firm produced.
- Sometimes sampling is more accurate. If there is reason to believe that the survey process generates errors, then a full-scale census may amplify these errors. For example, assembling the research team required to undertake a census may lead to inexperienced survey staff being used to collect data, whereas a smaller team might be better trained and more experienced (see Lipstein, B. 1974. In defense of small samples. *Journal of Advertising Research*, February, p. 35).

For whatever reason sampling is undertaken, a central problem arises. Are the descriptive statistics we get from a *sample* the same as the corresponding statistics we would get if a complete and accurate census was undertaken? Are the sample statistics in some sense 'representative' of the population from which the sample is drawn? Even though we may do everything in our power to draw a 'representative' sample from a population, the operation of **random variation** may cause the sample to be 'off'. On what basis then can we make a valid generalization from the sample to the population?

For example, we might sample a group of 120 people from a certain area and ask each their age in years. Here the variable of interest (age) is measured at the interval/ratio level. We can describe the information contained in the data by calculating a measure of central tendency to give a sense of the average score; by calculating a measure of dispersion to give a sense of the spread of scores around the average; and by drawing a graph to give us an overall impression of the distribution. These are not the only ways of describing a distribution (as we have seen) but will often satisfy many of our research questions.

This information might be interesting in itself, but usually we compile information about a sample because we have another issue to address: what is the average age of *all* people in this area? If the average age for this sample is 36 years, can I generalize from this to the whole population? This is where the operation of random variation may cause us to feel uneasy about making such generalizations from the sample statistics. How can we be sure that our sample did not *by chance* include a few disproportionately old or disproportionately young people, in relation to the population? We address this problem with **inferential statistics**.

Inferential statistics are the numerical techniques for making conclusions about a population based on the information obtained from a random sample drawn from that population.

To undertake statistical inference we generate three separate sets of numbers:

1. *Raw data*. These are the measurements taken from each case for a variable (e.g. the age of each person, measured in years). This will often be a very large set of numbers, depending on the actual sample size.
2. *Sample statistics*. These are the descriptive statistics that summarize the raw data obtained from the sample (e.g. the mean, standard deviation, or frequency distribution).
3. *Inferential statistics*. These help us to make a decision about the characteristics of the population based on the sample statistics.

Although the detailed steps involved in making an inference vary from situation to situation, we use the same general procedure, which involves generating these three sets of numbers. This procedure is illustrated in Figure 14.1.

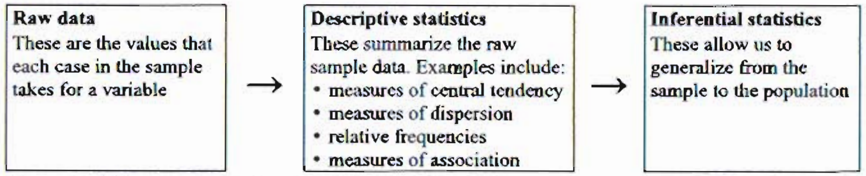

Raw data
These are the values that each case in the sample takes for a variable
→
Descriptive statistics
These summarize the raw sample data. Examples include:
• measures of central tendency
• measures of dispersion
• relative frequencies
• measures of association
→
Inferential statistics
These allow us to generalize from the sample to the population

Figure 14.1 The process of inferential analysis

Random samples

The most important condition that must apply if we are to use inferential statistics to generalize from a sample to a population is that the sample must be **randomly selected** from the population.

Random selection is a sampling method where each member of the population has the same chance of being selected in the sample.

A telephone survey of the population is not perfectly random. Only people in a household with a telephone at the time of the survey have a chance of being included. This excludes the homeless and households without a phone. Similarly, it gives households with more than one telephone number a greater chance of being included. In fact, very few surveys will be perfectly random in terms of the strict definition. The important consideration is whether the deviation from random selection is likely systematically to over-represent or under-represent cases of interest such that the results will have a **bias**. A biased sample favors the selection of some members of the population over others.

Sometimes there are good reasons to deviate from simple random selection by using **stratified random sampling**. A stratified random sample is used on a population that has easily discernible strata. Each stratum is a segment of the population that we suspect is homogenous in terms of the variable we want to measure. We first predetermine the proportion of the total sample that will come from each stratum. We then randomly select cases from *within* each stratum. For example, we might feel that men are similar to each other in terms of a particular variable and that women are also similar to each other in terms of this variable, but there is a difference between men and women. Thus we might stratify a sample to ensure that 50 percent of the sample is women and 50 percent men. We then randomly select the required number of women and required number of men.

Random sampling is often called **probability sampling**. But there is a whole range of non-probability (non-random) sampling techniques, such as **snow-ball sampling**. Snow-ball sampling involves selecting cases on the basis of information provided by previously studied cases. Such a sampling method is particularly useful when conducting research on close-knit populations that are difficult to get to, or whose exact size and composition cannot be known in advance.

There is no inherent reason why probability sampling should be considered 'better' than non-probability sampling. Each method is appropriate for different research questions, and sometimes a research question will be better addressed by choosing a non-probability sampling method. One of the implications of using a non-probability sampling method, though, is that we cannot use the inferential statistics we are about to learn. This is not necessarily a bad thing, and other ways of interpreting information are as valid as statistical inference, and sometimes more so.

Unfortunately, the professional and academic worlds do not always see it this way. Research seems to acquire a 'scientific' look when dressed in terms of inferential statistics, and often research is forced into this framework just to suit the fashion. Inferential statistics are sometimes calculated on samples that are not randomly selected. In other instances, the research project is structured in such a way as to make inferential statistics applicable, even though other methods may have been more insightful. This is a problem with the practice of research that raises broader issues than can be dealt with here. All we will do now is issue a word of caution: the choice of research methods should never be undertaken on the basis of the technique to be used for analyzing data. It should be chosen on the basis of best addressing the research problem at hand, and if that happens to involve the kind of statistical analyses we will be learning below, then we will know how to deal with it. If not, then the project is not lost. It simply means other avenues should be pursued.

The sampling distribution of a sample statistic

Inferential statistics only apply to random samples because the central tool used to make inferences is based on the assumption of random sampling. This tool is the **sampling distribution of a sample statistic**. Before defining the sampling distribution, we will illustrate the idea behind its construction through a very simple experiment. Assume that we have a board that consists of rows of nails that are evenly spaced and protrude from the board (Figure 14.2).

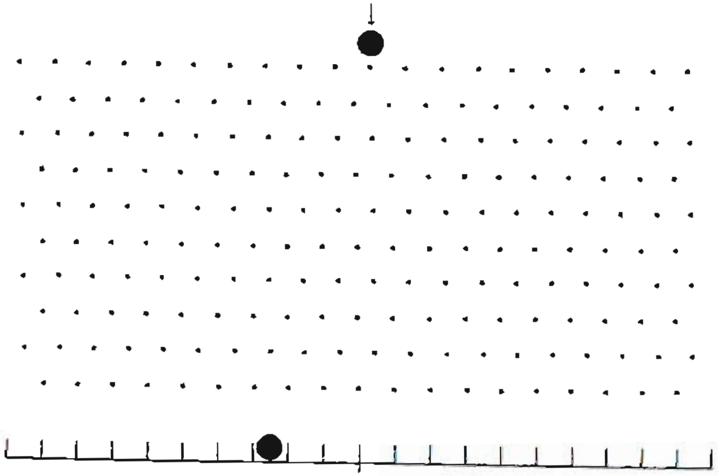


Figure 14.2

A ball is dropped directly above the middle nail in the top row and allowed to find its way down to the bottom. The path the ball takes will depend on a whole host of factors, but eventually the ball will bounce around and emerge somewhere at the bottom. The point at which any *individual* ball will fall is a random event. However, if I dropped 100 identical balls from the same position and let each find its way down the rows of nails to pile up at the bottom, we might get a distribution that looks like Figure 14.3.

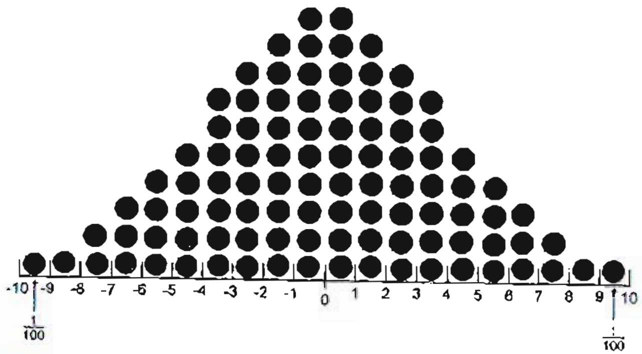


Figure 14.3 The distribution of repeated random drops

Most balls will bounce around, but since they are dropped from the same point plenty of balls will pile up in the center. But not all the balls will travel this path. Some will just happen to bounce to the left of each nail more often than they bounce to the right, and therefore emerge over to one side, and some will happen to keep bouncing to the right more often and come out on the other side. In fact the occasional odd ball will land way out to the left (position –10) or way out to the right (position 10). But we can see that the chances of a ball landing way out to the left, if allowed to fall freely, is only 1 in 100. In other words, although the location of any individual ball is a random event, the shape of the overall distribution of repeated drops is not random – it has a definite shape.

What has all this got to do with research statistics and inference? To see how the same logic applies in the 'real' world rather than just with balls and nails, let's go back to the example where people from a small community are surveyed and their age in years recorded. The parameters for this population of 1200 are:

$$\mu = 35 \text{ years}, \sigma = 13 \text{ years}$$

Let us assume, however, that we do not survey all 1200 members of this community. Instead we carry out the following experiment. We randomly select 120 people and ask only these 120 their respective ages and calculate their average age. We then put these people back into the community and randomly select another 120 residents (which may include members of the first sample). We proceed to draw a third sample of 120 residents. We keep doing this over and over again taking a random sample of 120 community members and calculating the average age for each *sample*.

This should sound a little like the experiment of dropping 100 balls down the board and seeing where they land, except instead of balls, we are taking samples and seeing where the sample means 'fall'. I have actually performed this hypothetical experiment (not with real people but using SPSS, as will be illustrated below), and the results of these 20 repeated random samples are displayed in Table 14.1, in the order in which they were generated, rounding to the nearest decimal point. These results are also plotted in Figure 14.4 to show the spread of sample means.

Table 14.1 Distribution of 20 random sample means (n = 120)

Sample number	Sample mean
1	34.7
2	35.9
3	35.5
4	34.7
5	34.5
6	35.4
7	35.7
8	34.6
9	37.4
10	35.3
11	34.1
12	35.5
13	34.9
14	36.2
15	35.6
16	35.0
17	35.1
18	36.4
19	35.6
20	33.6

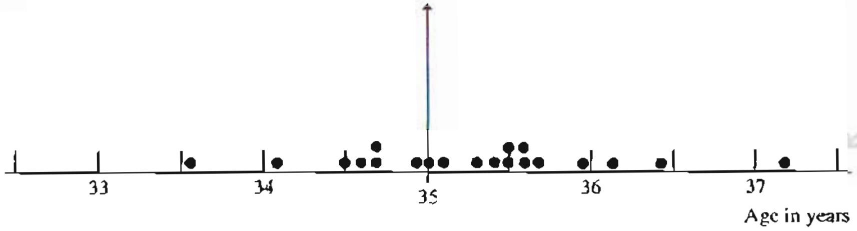


Figure 14.4 Distribution of 20 random sample means (n = 20)

We can see that most of the results are clustered around the population value of 35 years, with a few scores a bit further out and one 'extreme' score of 37.4 years. This is obviously a sample that just happened by chance, through the operation of random variation, to include a few relatively older members of the community. Even so it is interesting that despite the fact that the individual ages of the 1200 people in the community range from 2 years to 69 years of age, the *means of the samples* have a very narrow range of values. Nearly half of the 20 samples I took produced mean ages within half a year of the 'true' population average. This gives us some sense of the value and reliability of random samples.

Let us push this hypothetical example a little further, and imagine that we theoretically take an *infinite* number of random samples of equal size from this population and observe the distribution of all of these sample means. The pattern we have already observed with just 20 random samples will be reinforced. Most of the samples will cluster around the population parameter, with the occasional sample result falling relatively further to one side or the other of the distribution. Such a distribution is a **sampling distribution**.

A sampling distribution is the theoretical probability distribution of an infinite number of sample outcomes for a statistic, using random samples of equal size.

A sampling distribution is a *theoretical distribution* in that it is a construct derived on the basis of a logical exercise – the result that will follow *if* we could take an infinite number of random samples of equal size. The distribution of a sample and the distribution of a population, on the other hand, are *empirical distributions* in the sense that they exist in the 'real world'.

Here we are dealing with the **sampling distribution of sample *means*** since it is the distribution of all the means obtained from repeated random samples. This sampling distribution of sample means will have three very important properties:

1. *The mean of the sampling distribution is equal to the population mean.* In other words, the average of the averages ($\mu_{\bar{X}}$) will be the **same as** the population mean. This is written formally in the following way:

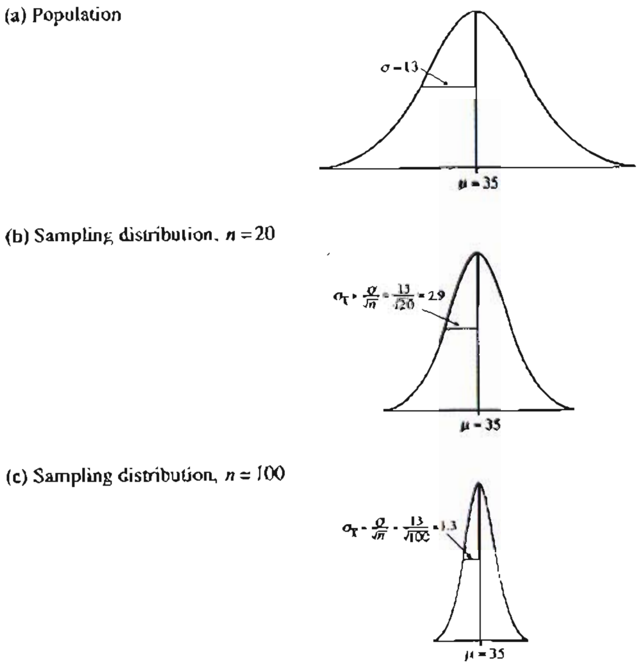
$$\mu_{\bar{X}} = \mu$$

2. *The standard error will be related to the standard deviation for the population.* The standard deviation of the sampling distribution is known as the **standard error** ($\sigma_{\bar{X}}$), and its value is affected by the sample size and the amount of variation in the population. If we are only taking a sample of five people, and one of the people in this small sample happens to be 60 years of age, the average for this sample will be greatly affected by this one score. In other words, we expect small samples to be less reliable than large samples, since they have a higher probability of producing a very wide dispersion of results. If our sample size is 200 the effect of one large score will be diluted by a greater number of cases that are closer to the population mean. So repeated large samples will be clustered closer to the population value; they will be more reliable. Similarly, if we were drawing samples from a population where age spreads from 2 years of age to 102 years of age, the range of scores we would get from these samples will be much greater than if we were sampling from a population where age only ranged between 20 and 30 years. *The more homogeneous the population, the more tightly clustered will be random samples drawn from that population.* These two factors are captured by the following formula for the standard error:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

3. *The sampling distribution will be normally distributed.* The proportion of samples that will fall within a certain range of values will be given by the standard normal distribution.

These features of the sampling distribution of sample means are illustrated in Figure 14.5.


(a) Population

(b) Sampling distribution, $n = 20$

(c) Sampling distribution, $n = 100$

Figure 14.5 Sampling distributions with different sample sizes

Figure 14.5(a) displays the distribution of all 1200 people which is the population of the community. Figure 14.5(b) is the sampling distribution of sample means for samples of size $n = 20$. In other words, it is the distribution of means we will get if we repeatedly sample 20 people from this community. Figure 14.5(c) is the sampling distribution of sample means for samples of size $n = 100$. We can see that both sampling distributions will be centered on the population mean of 35 years. Both will also be normally distributed. However, the standard error for each sampling distribution will vary. With repeated samples of size $n = 20$ there is a greater spread of sample means, with a standard error of 2.9 years, whereas with the larger samples the sample results are clustered more tightly around the population value. Both sampling distributions are normal, in that 68 percent of all cases fall within one standard deviation from the mean. But for the sampling distribution where $n = 20$ this range will be between 32.1 years and 37.9 years:

$$35 \pm 2.9 = 32.1 \text{ and } 37.9 \text{ years}$$

whereas for the second sampling distribution this range will be much narrower, having a lower limit of 33.7 years and an upper limit of 36.3 years:

$$35 \pm 1.3 = 33.7 \text{ and } 36.3 \text{ years}$$

The central limit theorem

We have looked at the properties of a sampling distribution derived from a population that is normally distributed. In particular, the sampling distribution will also be normal. However, there are few populations in the social world that are even approximately normal. What if the ages of the 1200 people in our small community are distributed as shown in Figure 14.6?

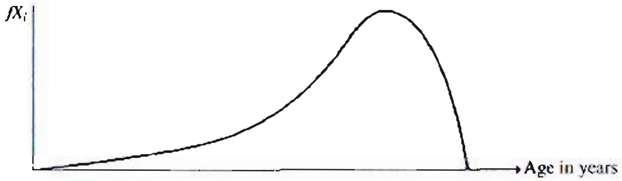


Figure 14.6 A skewed distribution

The distribution is skewed to the left, indicating that there are relatively more older people than younger people in this community. It would seem that repeated random samples from this skewed distribution will produce a skewed sampling distribution as well. However, this is not so. According to one of the key principles in statistics, the **central limit theorem** states that under certain conditions the sampling distribution will be normal, even though the population distribution from which the samples are drawn is not normal.

The **central limit theorem** states that if an infinite number of random samples of equal size are selected from a population, the sampling distribution of the sample means will approach a normal distribution as sample size approaches infinity.

The population may be non-normal, yet repeated sampling will (theoretically) generate a normal sampling distribution. In fact, the sample size does not have to be as large as suggested in the formal statement of the theorem: once the sample size is greater than 100, the sampling distribution of sample means will be approximately normal.

Generating random samples using SPSS

We can generate repeated random samples on SPSS to see the spread of sample means. In fact this is how I got the results presented in Table 14.1. This is a fairly repetitive procedure, since we need to generate a large number of random sample means. There are two steps repeated in sequence over and over. The first is to select a random sample, and the second is to calculate the mean for the sample.

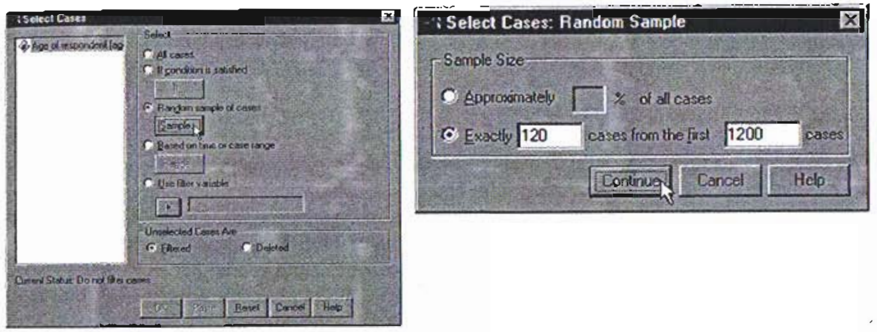
Selecting a random sample

Using the data that have been entered on the ages of the 1200 residents of our hypothetical community, the first step is to ask SPSS to randomly select a certain number of cases, which in this instance will be 120 (Table 14.2, Figure 14.7).

When you have completed the commands listed in Table 14.2 and refer to the **Data Editor** window you will see that SPSS has placed a slash through most of the numbers in the shaded column on the left of the page. These cases are the ones that are not included in the calculation of the mean – the ones that have not been randomly selected. Similarly, you will notice that SPSS has created a new 'variable', which it calls **filter_$**. We do not actually use the filter ourselves, even though it will appear in variable lists. SPSS uses this variable to choose some cases in the sample and ignore others, by assigning a value of 1 to cases without a slash through their case number and 0 to those that have been 'slashed'.

Table 14.2 Generating repeated random samples on SPSS (file: **Ch14.sav**)

SPSS command/action	Comments
1 From the menu select **Data/Select Cases**	This brings up the **Select Cases** dialog box. The SPSS default setting is to use all cases, indicated by ● in the radio button next to **All cases**
2 Select **Random sample of cases** by clicking on the small circle next to this option	A ● will appear in the radio button next to **Random sample of cases** and the text below it will darken
3 Click on the **Sample** button	This brings up the **Select Cases: Random Sample** dialog box. This gives us the option of selecting a certain percentage of cases, or a certain number of cases. Here we want a certain number of cases (120)
4 Click on the small circle next to **Exactly**	The cursor will jump to the box next to **Exactly**
5 Type 120	This is the size of the sample we wish to draw
6 Type 1200 in the box next to **from the first**	This is the total number of cases from which we want to draw the sample
7 Click on **Continue**	
8 Click on **OK**	


Select Cases
Select
○ All cases
○ If condition is satisfied
● Random sample of cases
○ Based on time or case range
○ Use filter variable
Unselected Cases Are
● Filtered ○ Deleted
Current Status: Do not filter cases

Select Cases: Random Sample
Sample Size
○ Approximately □ % of all cases
● Exactly 120 cases from the first 1200 cases

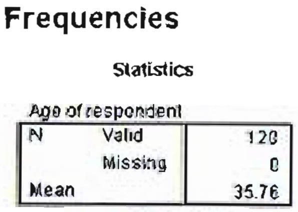
Figure 14.7 The **Select Cases** and **Random Sample** dialog boxes

Calculating the sample mean

The next step is to ask SPSS to calculate the mean for this sample using the **Analyze/Descriptive Statistics/Frequencies** command we learnt in Chapter 4. It might be helpful to select *only* the mean in this option, so that we do not get a frequency table and other descriptive statistics for each repeated sample, since this will generate more output than is necessary for our purpose here.

Repeating the sampling procedure

Running these two commands in sequence will generate a mean for the randomly selected sample of 120 cases. To draw another random sample all that is required is that we select the **Data/Select Cases** command and then click directly on **OK**. It is not necessary to again tell SPSS to randomly select 120 cases – it will automatically repeat the previous set of instructions and choose a new sample of 120. Similarly, by selecting **Analyze/Descriptive Statistics/ Frequencies** and then clicking on **OK** another sample mean will be calculated without having to reselect all the options within this command. You will get 20 SPSS Statistics tables that each look like the one in Figure 14.8, but each will have a different value for the mean.



Frequencies

Statistics

Age of respondent

N	Valid	120
	Missing	0
Mean		35.76

Figure 14.8 SPSS mean for a random sample

Your own set of 20 results will have different values for each mean, since we are working with random samples. These results do not constitute a true sampling distribution, since there are only 20 samples, whereas a *sampling* distribution is theoretically the distribution of an *infinite* number of random samples. Despite this a general pattern should emerge from your repeated sampling procedure:

- Most of the sample results will be very close to the population value of 35 years. There will be some variation around this, but most sample results will be clustered around the population parameter.
- You should get one or two sample means that are relatively a great distance from the population parameter of 35. There is always a possibility that an individual sample may produce an 'odd' result, but most samples will tend to be 'true' to the population value.

Summary

We have spent a great deal of time in this chapter dealing with abstract theoretical concepts. In particular we have played around with a thought experiment: what if we could take an infinite number of samples of equal size from a certain population, and calculate the mean for each of these samples? At some point the critical reader will have thought 'but who gets to take an infinite number of samples?' Usually a social or health researcher only gets to take one sample from a population and has to determine what the population looks like from that one sample. What use is the sampling distribution then? In the next chapters we will see that it is the basis on which inferences can be made from a single random sample to a population.

Exercises

14.1 What is the difference between a parameter and a sample statistic?

14.2 What is the difference between descriptive statistics and inferential statistics?

14.3 What is random variation? How does it affect our ability to make a generalization from a sample to a population?

14.4 State whether each of the following statements is true or false:

(a) The reliability of random sample means depends on the size of the sample, the variance of the population, and the size of the population.
(b) The means of random samples will cluster around the population mean.
(c) The standard deviation of random sample means will be greater than the standard deviation of the population from which they are drawn.
(d) The sampling distribution of sample means will be normal only if they are drawn from a normal population.

14.5 If the mean of a normal population is 40, what will the mean of the sampling distribution be with $n = 30$; with $n = 120$?

14.6 What is meant by the standard error? Will it be equal to, greater than, or less than the standard deviation for the population? Why?

14.7 Sketch the sampling distribution of sample means when $n = 30$ and when $n = 200$. In what way are these two distributions different, and in what way are they similar?

14.8 A teacher wants to evaluate a course by surveying registered students. The teacher writes the letters in the alphabet on separate pieces of paper and selects the one with G written on it out of a hat. The teacher therefore selects all students in class whose last names begins with G. In what ways, if any, is this sampling method non-random?

14.9 A library wants to assess the condition of the books in its possession. It randomly selects Thursday, and examines the condition of all books returned to the library on the following Thursday. In what ways, if any, is this sampling method non-random?

14.10 Describe a research project that might use the process of stratified random sampling.

14.11 Why is the central limit theorem so important to research?

14.12 Using the data for the age distribution of the community of 1200 people, draw another 20 random samples, this time using sample sizes of 30. How does the spread of results differ from that in the text, where sample size was 120?

15

Introduction to hypothesis testing and the one sample z-test for a mean

In research we are often interested in whether a population parameter, such as the population mean, has a *specific* value. The information we have collected about this parameter, however, is usually obtained from a sample rather than a census and we therefore have to make an inference from the sample to the population.

Before turning to a detailed description of the way we make such an inference, we will pose this problem in a slightly different way: as a problem of betting on a two-horse race. Assume you are at a racetrack and about to place a bet on an upcoming race that only has two horses running. From the form guide you know that one of these horses will win one race in every 100: will you put your money on it? Probably not. If the odds of this horse winning are 1-in-20 races, will you bet on it? Maybe. Essentially, inferential statistics involve the same mental exercise – two 'runners' are lined up against each other, and the odds of one of these runners 'winning' are calculated. We then decide which one we will bet on.

The reason we have to gamble is that, as we have seen in previous chapters, information from a random sample is not always an accurate reflection of the population from which the sample is drawn. To see this we will work with the 1200 people and their ages in years that we have introduced in earlier chapters. We know that this population has an average age of 35 years and standard deviation of 13 years.

We are also told that a sample of 150 people has an average age of 32 years:

$$\overline{X} = 32 \text{ years}$$

We want to know whether this sample did or did not come from the population of 1200. There is a difference of 3 years between the sample and this population. Does this difference of 3 years suggest that this sample came from another population or did it come from this population with a mean age of 35, and the difference of 3 years is due to random variation when sampling? In other words, there are two possible explanations as to why a sample result may differ from a population that we suspect it may have been drawn from.

The first explanation is that *the sample did come from the population but the sample just happened to select, by chance, a lot of younger people*. We will call this explanation of our sample result the 'null hypothesis of no difference'. Mathematically we write this as:

$$H_0: \mu = 35 \text{ years}$$

An alternative explanation is that *the sample came from another population whose average age is not equal to 35 years*. We call this the 'alternative hypothesis'. Symbolically, we write:

$$H_a \ \mu \neq 35 \text{ years}$$

These two hypotheses are mutually exclusive: if one is right the other is wrong. Either the sample came from the population whose average age is 35 or it did not. This is like the two-horse race where only one can win. We do not know which is correct: each statement is just an *hypothesis* that may or may not be right. If I now said the chances that the null hypothesis of no difference is correct are 1-in-100, will you bet on it? What about if the odds were 1-in-10? Inferential statistics provide us with these odds.

The whole hypothesis testing procedure proceeds on the assumption that the null hypothesis of no difference is correct. This may at first seem strange, since usually we undertake research in the hope of discovering a difference. Why then assume no difference? It is because we think this assumption is incorrect that we make it. The logical exercise involved in hypothesis testing is to show that the assumption of no difference is 'inconsistent' with our research findings, thereby leading us to argue that this is an unjustified hypothesis. We try to prove that there is a difference by disproving its opposite – the assumption of no difference. This may seem like the long way to go about reaching a conclusion, but if we work through enough examples in the following chapters we will see that we are *testing an assumption* by seeing whether our research data are 'plausibly' consistent with it.

In the context of the example we have been working with, I may strongly believe that the sample with a mean age of 32 years did not come from the population of 1200 people whose mean age is 35 years. Despite how strongly I believe this to be true, I actually begin by assuming the opposite; what I believe to be untrue. If I can show that it is highly unlikely for a sample with a mean of 32 to be drawn from a population with a mean age of 35, then this starting assumption will not be plausible and I am justified in rejecting it. This is why we talk of 'hypothesis testing' – we put the null hypothesis to the test by comparing our actual sample result to it. And often we want it to fail the test!

Let us then *assume* for the sake of argument that the null hypothesis of no difference is true. We are assuming that the sample has come, despite the difference of 3 years, from the population whose average age is 35. Is the sample result of 32 inconsistent with the assumption that the population average is 35? What is the *probability* of getting by chance a sample that differs from the population value of 35 by 3 years or more?

This is where the sampling distribution of sample means enters the picture. Remember that the sampling distribution is the distribution of means for repeated random samples of equal size. We can therefore refer to the sampling distribution, whose properties we know in detail, to determine the probability of getting a sample mean of 32, *if* the population value is 35. This will give us the odds to allow us to place our bet on either the null hypothesis or alternative hypothesis. Deriving these probabilities is a fairly straightforward (although somewhat tedious) procedure, with which we are now familiar: convert the sample statistic – the mean age – into a *z*-score and look up the associated probability from the table for the area under the standard normal curve in Table A1.

The first step then is to calculate the *z*-score that is associated with our sample result. When calculating such *z*-scores for the purpose of testing a mean we use the following modified formula for *z*:

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

For the sample of 150 people whose mean age is 32 years, the *z*-score is –2.8:

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} = \frac{32 - 35}{13/\sqrt{150}} = -2.8$$

This equation has *standardized* the observed difference of 3 years between the sample score and the hypothesized population value by converting it into a *z*-score. The advantage of 'washing out' the natural units in which the difference is initially measured (in this instance years) is that we can now refer to the table for the area under the standard normal curve (Table 15.1) which is printed in every statistics textbook to determine the probability of getting a *z*-score of 2.8 or more.

Table 15.1 Areas under the standard normal curve

z	Area under curve between both points	Area under curve beyond both points	Area under curve beyond one point
⋮	⋮	⋮	⋮
±2.1	0.964	0.036	0.0180
±2.2	0.972	0.028	0.0140
±2.3	0.979	0.021	0.0105
±2.33	0.980	0.020	0.0100
±2.4	0.984	0.016	0.0080
±2.5	0.988	0.012	0.0060
±2.58	0.990	0.010	0.0050
±2.6	0.991	0.009	0.0045
±2.7	0.993	0.007	0.0035
±2.8	0.995	0.005	0.0025
±2.9	0.996	0.004	0.0020
±3	>0.996	<0.004	<0.0020

You might be wondering why I referred to the column headed *Area under curve beyond both points*, rather than the column headed *Area under curve beyond one point*. The reason for this relates to the way in which I have framed the problem. I am interested in the probability of randomly drawing a sample that differs from the hypothesized population mean of 32 years *by 3 years or more*, since this is the amount of difference we actually have between our sample (with a mean of 32) and the population (with a mean of 35) from which it may have been drawn. Since sampling variation may cause the means of random samples to be either higher or lower than the underlying population mean, a sample may differ by 3 years or more from the hypothesized value either by being 3 years *above* it (a mean of 38 years), or by being 3 years *below* it. We therefore refer to the middle column to determine the probability of drawing, through sampling variation alone, a sample that differs from the hypothesized population value by 3 years or more.

The area under the curve beyond the z-scores of +2.8 or −2.8 is 0.005. This is the probability of drawing, from a population with an average age of 35 years, a sample with an average age that is 3 years or more above or below this mean. In other words, only 5-in-1000 samples will differ from a population mean of 35 years by 3 years or more. We are left with a choice:

- we can still hold that the assumption that this sample came from a population with a mean age of 35 is correct, and explain the sample result as a rare 5-in-1000 events; or
- we can reject the assumption that this sample came from a population with a mean age of 35; the sample statistic is not a 'freak', but instead reflects that the sample is drawn from an underlying population with an average age other than 35 years.

Given the long odds that the first choice is correct, it might be a safer bet to reject the assumption that the sample came from a population with an average age of 35. The difference of 3 years between the sample result and the hypothesized population value is so great that it is unlikely that it came about by random variation when sampling. It instead reflects that we are not sampling from a population with an average age of 35 years.

To illustrate this procedure again, let's suppose the sample of 150 people yielded the result:

$$\bar{X} = 36 \text{ years}$$

Again we will *assume* that this sample came from a population with a mean age of 35 years. Clearly there is again a difference between the sample statistic and the population parameter, this time of 1 year (36 − 35 = 1). There seems to be an apparent conflict between our *assumption* that the sample came from a population with a mean age of 35 and our *observation* that the sample statistic is not exactly equal to the population value. Should this cause us to reject the assumption and argue that the sample came from a different population?

To answer this we need to derive the probability of randomly selecting a sample that differs from a population with an average age of 35 by 1 year or more. We need first to convert the sample result into a z-score:

$$z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{36-35}{13/\sqrt{150}} = 0.9$$

The table for areas under the standard normal curve (Table 15.2) indicates that the probability of obtaining this z-score or greater either side of the mean is 0.368.

Table 15.2 Areas under the standard normal curve

z	Area under curve between both points	Area under curve beyond both points	Area under curve beyond one point
±0.1	0.080	0.920	0.4600
±0.2	0.159	0.841	0.4205
±0.3	0.236	0.764	0.3820
±0.4	0.311	0.689	0.3445
±0.5	0.383	0.617	0.3085
±0.6	0.451	0.549	0.2745
±0.7	0.516	0.484	0.2420
±0.8	0.576	0.424	0.2120
±0.9	0.632	0.368	0.1840
±1	0.683	0.317	0.1585
⋮	⋮	⋮	⋮
±3	>0.996	<0.004	<0.0020

From a population with an average age of 35 nearly 37-in-100 samples will have a mean age that differs from 35 by 1 year or more. Random variation will cause roughly one-third of all samples to vary this much from a population with a mean value of 35. Given such a high probability, we can say that the sample result is simply due to random variation when sampling from a population with a mean age of 35 years.

The material to be presented in later chapters is simply a variation on this theme. These differences, however, do not change the basic method of approach. In fact, we can approach just about any problem of inference using the following five-step procedure:

Step 1: State the null and alternative hypotheses.
Step 2: Choose the test of significance.
Step 3: Describe the sample and derive the *p*-value.
Step 4: Decide at what alpha level, if any, the result is statistically significant.
Step 5: Report results.

Step 1: State the null and alternative hypotheses

We begin our inference procedure by making two, mutually exclusive, hypotheses: the **null hypothesis** and the **alternative hypothesis**. These hypotheses have three crucial elements:

- they identify the population(s) about which we want to make a statement;
- they identify the variable(s) for which we will gather data;
- they identify the relevant descriptive statistic that will be tested.

The null hypothesis of no difference (H_0)

This is a statement that the statistic we are using to describe the population under investigation will equal a specified, predefined value. The null must be clearly capable of being rejected or not rejected; that is, it can be shown to be false. There should be no ambiguity: either the population statistic has a certain value or it does not.

An abbreviated way of writing the null is in mathematical shorthand, depending on the particular descriptive statistic about which we are making an hypothesis. If we are making an hypothesis about the population mean, for example, the general form of the null hypothesis is:

$$H_0\ \mu = X$$

where X is the pre-specified 'test' value. For instance, in the example above we were testing whether $\mu = 35$ years.

Where does this test value stated in the null hypothesis come from? There are usually two different kinds of research questions that will prompt us to investigate whether a population parameter takes on a specific value. The first is where a particular value is chosen for *practical or policy reasons*. For example, a company may decide that anything more than a 5 percent reject rate for its product is commercially unacceptable. It therefore instructs its quality control department to sample 300 randomly selected products and determine whether the reject rate is 5 percent or more. Thus the company is not simply interested in finding whatever the reject rate happens to be; it wants to know whether this rate is specifically 5 percent or more. Similarly, the government may have decided that it will devote extra health resources to any area where the mean age is greater than 40 years. It will therefore want to test specifically whether a sample taken from a particular region indicates whether the whole population of that region is on average 40 years of age or more, as measured by the mean.

The other situation in which we will have a specified test value is where we want *to compare the population under investigation with another population whose parameter value is known.* For example, we want to compare two populations in terms of their respective average amounts of TV watched per day: the population of Australian children between 5 and 12 years of age and the population of British children between the ages of 5 and 12 years. We know from census data that British children watch on average 162 minutes of TV each day, but we only have a sample of children from Australia. We have to make an inference (which is basically a fancy way of saying an educated guess) whether the unknown average amount of TV watched by all Australian children is equal to the known average for British children.

The alternative hypothesis (H_a)

This is a statement that the population parameter *does not* equal the pre-specified value; there is a difference:

$$H_a: \mu \neq X$$

It is commonly argued that the alternative hypothesis, on the basis of theoretical expectation or practical need, may specify a *direction* of difference between the relevant sample statistic and a specific value, rather than simply stating that there is a difference. For example, in the analysis above we operated on the basis that there is no *a priori* reason to believe that the sample comes from a population either on average *younger or older* than 35 years. As a result we were interested in whether the sample result falls in either end of the sampling distribution. However, we might really suspect that the population from which the sample came is on average *younger* than 35 years. Alternatively, we may really believe that this population has a mean age *older* than 35 years. In either case, the alternative hypothesis specifies that there is not only a difference, but also a *direction* of difference. In mathematical notation we respectively write each of these in the following ways:

$$H_a: \mu < 35$$

or

$$H_a: \mu > 35$$

Where we specify a direction of difference in the alternative hypothesis, according to conventional logic of hypothesis testing, we need to halve the two-tail significance we obtain in Step 3 in order to determine the sample result's one-tail significance (these concepts will be explained shortly). For reasons I will discuss below, I do not agree with the use of one-tail tests (regardless of the form of the alternative hypothesis) but I present here the usual implications of specifying a direction of difference in the alternative hypothesis so that you are aware of the 'standard' procedure used in other books.

One thing to note about the alternative hypothesis is that it usually embodies what we really believe to be the 'truth' about the world. As a result it is sometimes referred to as the **research hypothesis**. This confuses people: if we really believe the alternative hypothesis to be an accurate depiction of the world, why do we begin the hypothesis testing procedure on the assumption that the null is correct. As we discussed earlier, we begin with the assumption that the null is correct so that we can 'test' it, and if it fails the test, this lends support to the alternative hypothesis. In other words, we are using the logic of proof by contradiction: we want to provide support for a statement we believe to be true by showing that its opposite is not true!

Step 2: Choose the test of significance

In this chapter we have introduced the most basic significance test, the one sample *z*-test for a mean, but there are many tests available to help us assess the null hypothesis (Table 15.3).

Table 15.3 Tests of significance

Descriptive statistic and number of samples	Test of significance	SPSS Command: Analyze/...
One sample mean	*z*-test for a mean (population variance known)	Not available
	t-test for a mean (population variance unknown)	Compare Means/One sample T Test
Two independent sample means	*t*-test for the equality of two means	Compare Means/Independent-Samples T Test
More than two independent sample means	ANOVA *F*-test for the equality of means	Compare Means/One-Way ANOVA
Two dependent sample means	*t*-test for the mean difference	Compare Means/Paired-Samples T Test
Frequency table for one sample (binomial scale)	*z*-test for a binomial percentage	Nonparametric Tests/Binomial
Frequency table for one sample (multinomial scale)	chi-square test for goodness-of-fit	Nonparametric Tests/Chi-Square
Crosstabulation for two or more independent samples	chi-square test for independence (can also use a *z*-test for proportions on a 2-by-2 table).	Nonparametric Tests/Chi-Square
Crosstabulation for two dependent samples	McNemar chi-square test for change (equivalent to the sign test)	Nonparametric Tests/2 Related Samples
Rank-sum for two independent samples	Wilcoxon *W* test (also known as the *z*-test for rank sums, which is equivalent to the Mann-Whitney *U* test)	Nonparametric Tests/2 Independent Samples
Rank-sum for more than two independent samples	Kruskal-Wallis *H* test	Nonparametric Tests/K Independent Samples
Rank-sum for two dependent samples	Wilcoxon signed-ranks *z*-test	Nonparametric Tests/2 Related Samples
Number of runs in a single sample	*z*-test for randomness	Nonparametric Tests/Runs
Number of runs between two samples	Wald-Wolfowitz *z*-test for the number of runs	Nonparametric Tests/2 Independent Samples
Correlation coefficient	*t*-test for a correlation coefficient	Correlate/Bivariate

Table 15.3 provides a quick guide for selecting the appropriate test of significance, based on these main factors, and the SPSS command for conducting the test. Often these tests of significance are given a shorthand name based on the statistician who first devised them, such as the Wilcoxon test. All of these tests require random samples (or at least reasonably random samples), but they vary according to the information available to the researcher. The most important factors that determine the choice of a test are:

- the descriptive statistic we are testing;
- the number of samples from which inferences are being made;
- whether we have independent or dependent samples.

Each test, in other words, applies in very specific circumstances. These do not exhaust all the possible hypothesis tests available; they present only those that will be covered in this and following chapters. The following chapters are basically organized around these individual tests, so that the conditions under which each is applicable will be clearly delineated.

This chapter will cover the use of a single-sample z-test for a mean. The conditions that allow this test to be used are:

- the desired descriptive statistic for summarizing the sample data is the mean (which itself requires that the data are measured at the interval/ratio level and the population distribution is not highly skewed);
- the variance of the population is known;
- the population is normally distributed along the variable; and/or
- the sample size is large ($n > 100$).

Either of these last two conditions, according to the central limit theorem, will guarantee that the sampling distribution of sample means is normal (you may wish to review the section on the central limit theorem in the previous chapter at this point).

Step 3: Describe the sample and derive the *p*-value

This is the process of calculating the relevant descriptive statistic for the sample as defined by the null hypothesis we are testing. On any given set of data we can usually calculate many different summary statistics, as we discussed in the early chapters of this book. The statistics we *actually* calculate depend on the hypotheses we are testing. Thus if we want to test whether the mean age of a population is 35 years, the relevant statistic to calculate from the sample data is self-evident: it is the *mean* age.

We usually find that the sample statistic does not conform exactly to the value suggested by the null hypothesis. In the example above we hypothesized that the sample came from a population with a mean age of 35 years, yet the sample itself produced a mean age of 32. The mere fact that the sample differs from the value we assume for the population is not necessarily a cause for concern; random samples will regularly produce results different from the population from which they are drawn. The issue is the *probability* of obtaining a particular sample result from a population that has the value specified in the null hypothesis. This is the **significance** of the sample statistic, commonly called the '*p*-value' ('*p*' for 'probability').

To derive this probability we have to first transform the sample statistic into a standardized **test statistic** using the appropriate equation, such as the following equation that transforms a sample mean into a z-score:

$$z_{sample} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

From the table for the areas under the standard normal curve (Table A1) we then determine the probability of obtaining a particular sample z-score if the null hypothesis is true.

sample statistic → test statistic → p-value

In the example above where the sample mean was 32 years, we obtained a z-score of -2.8, which had a p-value of 0.005. This is depicted in Figure 15.1, which displays the sampling distribution of all sample means that could be obtained from a population with a mean age of 35 years.

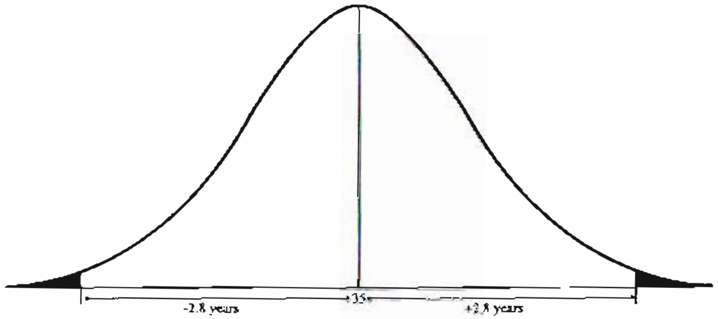


Figure 15.1

The shaded areas, representing 0.005 of the area under the curve, indicate that very few random samples of this size will have a mean 3 years above or below 35 years. In simple terms the sample mean of 32 years is an extremely unlikely outcome to get from a population with a mean of 35 years.

It is common practice at this point in the procedure to decide between the **one-tail** or **two-tail** significance of the result (although for reasons I will discuss later in this chapter, I now regard this distinction as unnecessary). The need to choose between a two-tail and one-tail test is justified with reference to the form of the alternative hypothesis. We noted above that the alternative hypothesis, on the basis of theoretical expectation or practical need, may specify a *direction* of difference between the relevant sample statistic and a specific value, rather than simply stating that there is a difference. For example, we might really suspect that the population from which the sample came is on average *younger* than 35 years. According to the proponents of the use of one-tail tests, we are thereby interested in whether the sample result falls far enough *to the left* of the population mean; we are only interested in whether the result suggests that we have one of those few random samples that will fall in the **left-tail** of the sampling distribution. Similarly, we may really believe that this population has a mean age *older* than 35 years, and therefore need to conduct a **right-tail** test.

In either case, since we are only interested in just one-tail of the sampling distribution, we refer to the column for the 'Area under the curve beyond one point' when using the table for the area under the normal curve to derive the p-value associated with a specific z-score. This will be its **one-tail** significance (we could also simply halve the two-tail significance). Thus in the example we have been working with, a sample mean of 32 years has a two-tail significance of 0.005; it therefore has a one-tail significance of 0.0025.

If we do use the one-tail significance, based on the direction of difference specified in the alternative hypothesis, we need to be careful that we refer to the appropriate tail of the sampling distribution. If the alternative hypothesis holds that the population value will be less than the specified value, the critical region will be in the left tail; if it holds that the population value will be greater than the specified value, the right tail is the relevant one (Table 15.4).

Table 15.4 Choosing a tail for a test

Alternative hypothesis	Tail of the sampling distribution
$H_a: \mu \neq X$	Both
$H_a\ \mu < X$	Left
$H_a: \mu > X$	Right

Left-tail significance is often used when we want to test if some *minimum* requirement has been met, whereas a right-tail significance is often used when we want to test whether some *maximum* limit or standard has not been exceeded. For example, if we wanted to see whether the average life of a piece of hospital equipment is at least 4.5 years, we would use a left-tail test. If we were interested in whether the time taken for a drug to have an effect on a patient was no greater than 1.5 minutes, then we would use a right-tail test.

In later chapters we will see that as an alternative to hand-calculation of test scores followed by reference to a table of critical values, we can use a program such as SPSS to calculate the relevant sample descriptive statistic and also determine the *p*-value for this statistic (Table 15.3). Alternatively, we can turn to calculation pages on the internet that allow us to enter the values we are testing and have calculated for us the relevant results. Many of these resources can be found from the Statpages.net homepage located (at the time of writing) at the following web address:

- members.aol.com/johnp71/javastat.html

Step 4: Decide at what alpha level, if any, the result is statistically significant

In the examples we used above to analyze the age of a sample of people, the decision whether to reject or not to reject the null hypothesis was easy. In the first instance, with a sample mean of 32 years, the probability that this sample came from a population with a mean of 35 was very small; in the second instance with a sample mean of 36 the probability was very large. But what if the sample result falls somewhere in between? At what point does the probability get small enough for us to say that the null hypothesis is not valid? Determining this cut-off point is called choosing the **alpha (α) level**.

There are two broad approaches we can take to this issue. One is the traditional approach that involves determining in advance a critical alpha level that delineates 'high' scores from 'low' scores so that we can decide to reject or not reject the null hypothesis by comparing the sample result to this specific cut-off point.

The other approach, which we will generally follow in this book, is less deterministic than the traditional hypothesis testing method. It involves reporting the *p*-value of the sample statistic and whether this is 'statistically significant' at the lowest of two conventional alpha levels, 0.05 and 0.01 (although occasionally 0.10 and 0.001 are of interest). This method indicates at what alpha level the null hypothesis *can* be rejected, but leaves some room for the reader of the results to judge whether the null hypothesis *should* be rejected or whether a more stringent alpha level should be set (it is interesting that the original formulation of significance testing by R.A. Fisher, 1925, *Statistical Methods for Research Workers*, Oxford University Press: Oxford, advocated this less deterministic approach; see also W.R. Rozeboom, 1960, The fallacy of the null-hypothesis significance test, *Psychological Bulletin*, vol. 57, pp. 416–28 for a powerful critique of the deterministic approach to hypothesis testing).

This method is illustrated in Figure 15.2, which displays various **regions of rejection**, defined by the two common alpha levels of 0.05 and 0.01, which may lead us to reject the null hypothesis.

The **region of rejection**, or **critical region**, is the range of scores that will cause the null hypothesis to be rejected.

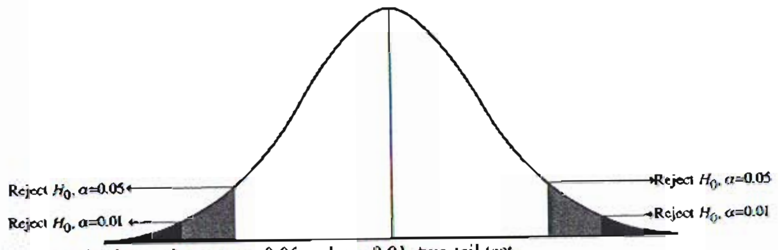


Figure 15.2 Rejection regions at $\alpha = 0.05$ and $\alpha = 0.01$, two-tail test

Figure 15.2 shows the sampling distribution of sample means, as discussed in the previous chapter. It shows that, from a population with a hypothesized mean, 5% (0.05) of all samples will have a mean either greater or smaller than those marked off by the lightly shaded areas. Similarly, 1% (0.01) of all possible random samples will have a mean either greater or lower than the areas marked off by the dark-shaded areas. (Note the relative sizes of the areas in Figure 15.2 are not to scale to allow for easier presentation). Since fewer random samples will have means that are very far from the population value, the rejection regions for $\alpha = 0.01$ are much smaller than those for $\alpha = 0.05$. With these rejection regions in mind, we can plot the *p*-value obtained in Step 3 and indicate if this is statistically significant at various alpha levels.

The critical aspect of this approach is to *provide the p-value of the sample statistic*, so that the importance of the result can be determined at least in part by whoever wishes to use the results, rather than having it prescribed by the person reporting the results. For example, assume we determine that a sample mean has a significance level of $p = 0.03$ (Figure 15.3).

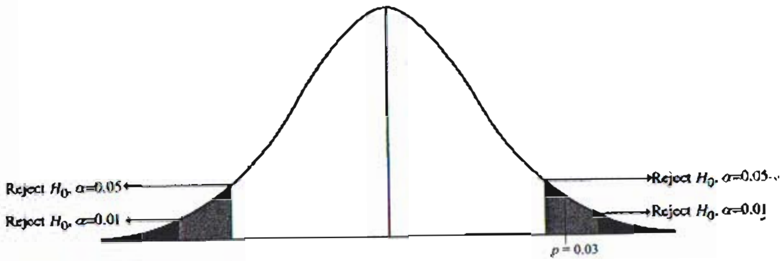


Figure 15.3 Rejection regions for $\alpha = 0.05$ and $\alpha = 0.01$

It is clear that the alpha level will determine whether we reject or do not reject this assumption of no difference. At an alpha level of $\alpha = 0.01$ the difference between the sample and the population value can be attributed to sampling error: do not reject the null. But at an alpha level of $\alpha = 0.05$ the same difference between the sample and the hypothesized parameter value will lead us to reject the null. In this instance we would state that the result is statistically significant at the 0.05 level, but if we also provide the *p*-value of 0.03, the reader is also made aware that the result *is not* statistically significant at the 0.01 level.

Figure 15.4 simplifies the logic of Figure 15.3 by using a single scale of possible probability values that ranges from 0 to 1.

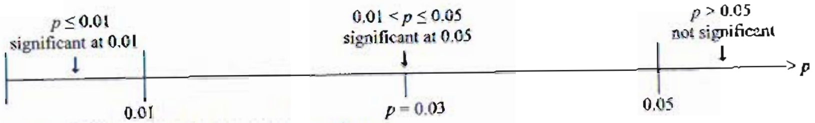


Figure 15.4 Determining statistical significance

You may find it helpful when conducting significance tests in your own research to draw a scale like the one in Figure 15.4 and plot the p-value you have obtained. You can then quickly read off the appropriate conclusion that you should reach.

A common confusion often arises right at this point of decision-making. In observing the so-called 'p-value' of the sample score, students are often dismayed if it proves to be very close to zero. We are used to thinking that small numbers indicate that 'nothing is there' and therefore the difference we suspected or hoped to find has not eventuated. Here the opposite is true. Usually we do want to find a low p-value (lower than the alpha level), since this indicates that the null hypothesis of no difference should be rejected. A very high p-value, on the other hand, indicates that the null hypothesis should not be rejected.

Step 5: Report results

We have detailed the technical steps involved in determining whether and at what level the null hypothesis should be rejected. In *presenting* the results of these procedures, though, we should try to be as non-technical as possible. We should try, for instance, to state our conclusion in plain words and indicate what practical or theoretical meaning the results have beyond whether they lead to us rejecting or not rejecting the null hypothesis. Similarly, the inferential statistics should be presented but should not be the focus of the discussion, which should concentrate on the general meaning of the results. In other words, while the steps involved in getting our results involve some formal and technical procedures, the readers of our results should not be labored with them. We cannot avoid using a little bit of jargon, but we should keep this to a minimum (see G. Francis, 2005, *An Approach to Report Writing in Statistics Courses*, www.stat.auckland.ac.nz/~iase/publications/14/francis.pdf).

To illustrate the way in which we present results let us return to the example of testing for the mean age of our population:

- We begin by stating in general terms what we are investigating. Thus I might introduce my findings by stating "We are interested in whether the mean age for the population is 35 years."
- I then state the relevant descriptive statistics that summarize the sample: "A random sample of 150 people had a mean age of 32 years".
- I discuss the statistical significance of this result and report the relevant test statistics: "The sample mean was statistically significant at the 0.01 level ($z = -2.8$, $p = 0.005$, two-tail)."
- I then interpret this with reference, in plain words, to the hypothesis that I have tested: "We reject the hypothesis that the population from which the sample is drawn has a mean age of 35 years."
- Finally, I should indicate whether any *statistically* significant difference is significant in any other sense; a point I will discuss in more depth below.

As a further example, I would report the results above where the sample produced a mean of 36 years in the following way:

> We are interested in whether the mean age for the population is 35 years. A random sample of 150 people had a mean age of 36 years. This is not statistically significant ($z = 0.9$, $p = 0.368$, two-tail). As a result, despite the sample being slightly older on average than we hypothesized, we cannot reject the possibility that the population from which the sample is drawn has a mean age of 35 years.

Notice that the conclusion is always stated in terms of the null hypothesis: reject or fail to reject. We are deciding whether the null hypothesis is plausibly consistent with a sample result. Samples do not always exactly mirror the populations from which they are drawn, so making an inference from a sample to a population always involves a risk of error. Specifically, whether we choose to reject or not reject a null hypothesis we need to be aware of the difference between a **type I error** (alpha error) and a **type II error** (**beta error**). A **type I error** occurs when the null hypothesis of no difference is rejected, even though in fact there is no difference. In assessing whether the sample in the example above with a mean age of 32 came from a population with a mean age of 35, we rejected the null hypothesis of no difference. The chances of selecting, from a population where the average is 35 years, a sample with an average age of 32 or less is only 5-in-1000. However, we may have actually selected one of those rare 5-in-1000 samples. The sample may indeed have come from a population with an average age of 35 years, but the sample just happened to randomly pick up a few especially young people. There is always a risk of such an event, which is why we speak in terms of probabilities. The question is the chance we are prepared to take of making this error.

A **type II error** occurs when we **fail to reject the null hypothesis** when in fact it is false. For example, where the sample above had an average age of 36, we concluded that it did come from a population with an average age of 35 years. The difference between the sample statistic and the hypothesized parameter value is so small that it can be attributed to random sampling error. However, it may in reality be that the population from which the sample is drawn does not have an average age of 35, but our sample just happened to select some unrepresentative people. The relationship between these two possible error types is summarized in Table 15.5.

Table 15.5 Error types

Decision based on hypothesis test	Truth about population	
	H_0 true	H_a true
Reject H_0	Type I error	Correct decision
Do not reject H_0	Correct decision	Type II error

It is clear that these two error types are the converse of each other so that *reducing the chance of one error occurring increases the chance of the other error occurring*. It is a question of which mistake we most want to avoid, and this depends on the research question. If we are testing a new drug that may have harmful side effects we want to be sure that it actually works. We do not want to make a type I error (conclude that the drug does make a difference when it doesn't) because the consequences could be devastating. The difference in the rate of improvement observed between a test group taking the drug and a control group that is not will have to be very large before we can say that such an improvement is not due to chance (say 1-in-1000). Thus a sample result may be significant at the 0.01 level, yet we may not be prepared to reject the null unless the more demanding alpha level of 0.001 is reached.

In other words, the 'appropriate' balance between these two alternative error types depends on the use to which the results are to be put, and this requires us to provide sufficient information when reporting results to allow a reader to make his or her own judgment about the null hypothesis, given their preparedness to make a type I or type II error. In particular, the exact probability associated with the test statistic (and the test statistic itself) should be reported so that the reader can compare the p-value to the alpha level he or she thinks is warranted in a given context, rather than simply being told that a result 'is significant at the 0.05 level', or words to that effect. If the preceding statement is all that is reported, the sample probability could have been 0.049 or 0.00001 – there is no way of knowing without doing the calculations. This may be frustrating to a reader who feels that an alpha level of 0.01 is warranted in the circumstances rather than the stated alpha of 0.05.

We have seen that we reach one of either two decisions about the null hypothesis of no difference: reject or fail to reject. In either case, we need to ask ourselves whether we have '*proven*' anything. The answer is 'no'! Given this general point about what we can conclude from significance tests, we will explore in turn the specific meaning of each possible conclusion that can be reached about the null hypothesis.

What does it mean when we 'fail to reject the null hypothesis'?

We begin with the presumption that the null hypothesis is true, and then proceed to test this assumption, but researchers are usually interested in rejecting the null. Normally we believe a difference exists; a decision to reject the null is usually the desired outcome (we want a low '*p*-value'). We are using the logic of proof by contradiction: we want to find support for the alternative hypothesis by showing that there is no support for its opposite, the null hypothesis.

Does this mean that if we fail to reject the null, the difference we are searching for does not exist? Not necessarily: failing to reject the null hypothesis of no difference simply means there is *not sufficient evidence* to think that the null hypothesis is wrong. This does not necessarily mean, however, that it is right. There might actually be a difference 'out there', but on the basis of the sample result such a difference has not been detected. This is like the presumption of innocence in criminal law. A defendant is presumed not guilty unless the evidence is strong enough to justify a verdict of guilty. However, when someone has been found not guilty on the strength of the available evidence, it does not mean that the person is in fact innocent: all it means is that, given that either verdict is possible, we do not choose 'guilty' unless stronger evidence comes to light. Similarly, with a verdict of 'no difference', failing to reject the null hypothesis does not mean the alternative is wrong. It simply means that on the basis of the information available, the null can explain the sample result without stretching our notion of reasonable probability.

Therefore, failing to find a significant difference should not be seen as conclusive. If we have good theoretical grounds for suspecting that a difference really does exist, even though a test suggests that it does not, this can be the basis of future research. Maybe the variable has not been operationalized effectively, or the level of measurement does not provide sufficient information, or the sample was not appropriately chosen or was not large enough. In the context of research, inference tests do not prove anything; they are usually evidence in an ongoing discussion or debate that rarely reaches a decisive conclusion.

What does it mean to 'reject the null hypothesis'?

What if our decision is the converse: we reject the null hypothesis? In formal language we say that we have found a *statistically significant* difference. So what? What have we learned about the world, and should we do anything about it? These questions are not ones that hypothesis testing as such can answer. A difference that is *statistically significant* simply indicates that it is unlikely to have come about by random error when sampling from a population defined by the null hypothesis. Whether such a difference is of any *practical or theoretical* importance – whether it is 'significant' in any other sense of the word – is really something we as researchers or policy-makers have to decide for ourselves.

To give this a concrete application assume that I, as a statistics teacher, want to know whether the university should spend more money on computer workshops and hire extra instructors to help students with their statistics classes. The university argues that it will only do this if there is a 'significant' difference between grades in statistics courses and grades in other courses that these students undertake at university. I collect a sample of students and find that their average statistics mark is 59, and compare it with the average for all other courses of 62, and find this to be statistically significant at an alpha level of 0.01. Have I won my argument with the university? Not necessarily. I might consider the difference in average marks to justify the extra expenditure because I think that statistics is very important to a well-rounded education. But the university has every right to say that given all the other possible ways it can spend its money, a difference of 3 marks is something it can live with. The university, in other words, may have no argument with me over the *statistical* difference; that is, it accepts that the difference really is there in the population and not just due to sampling error. However, it may strongly disagree that this difference is of *practical* significance in the sense that it should prompt the university to spend money to close the gap.

This illustrates an all-too-often neglected point. It is not uncommon for researchers simply, and blandly, to state that a result is significant at the 0.05 or 0.01 level without further comment, as if this is all that needs to be said. In fact this should just be the entry point to the more creative and interesting (but usually more difficult) research problem: what does this tell us about the world and what can we do about it? A finding may be statistically significant but does it *matter*? (see D.M. McCloskey and S.T. Ziliak, 1996, The standard error of regression, *Journal of Economic Literature*, March, pp. 97–114).

With all these general considerations in mind we will now turn to an example to familiarize us with the hypothesis testing procedure.

A two-tail *z*-test for a single mean

Suppose that a university is interested in the average academic ability of foreign students in a particular program. In this program, the university knows that the mean grade for all local students is 62 with a standard deviation of 15, and wants to assess whether foreign students constitute a distinct population in terms of their grades.

Step 1: State the null and alternative hypotheses

Are foreign students on average different to the rest of the university population in terms of their average grade? Given this research question we form the following two hypotheses:

H_0: The population of foreign students has the same mean grade as the rest of the university population.

$$H_0\colon \mu = 62$$

H_a: The mean grade of foreign students is different to the mean grade of all other students.

$$H_a\colon \mu \neq 62$$

Step 2: Choose the test of significance

The important factor is that we are interested in the mean grade. Hence the descriptive statistic we calculate to summarize the data is the mean. The university also knows what the standard deviation is for the population of domestic students. These two factors allow us to conduct a *z*-test for a single mean.

Step 3: Describe the sample and derive the p-value

From a random sample of 150 foreign students the mean grade is calculated as 60.5. From this information we calculate the test statistic:

$$z_{sample} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} = \frac{60.5 - 62}{15 / \sqrt{150}} = \frac{-1.5}{1.22} = -1.2$$

We look down the column of *z*-scores in the table for the areas under the standard normal curve until we reach 1.2, and then read across to find the probability under the column for the 'Area under the curve beyond both points'. This gives a *p*-value of 0.23.

Step 4: Decide at what alpha level, if any, the result is statistically significant

It is clear that the sample result, although different from that stated in the null hypothesis, is not 'different enough' to suggest that it came about by more than just sampling error, at any of the conventional alpha levels. From a population with a mean grade of 62, nearly 1-in-4 random samples will have a mean grade 1.5 marks above or below this grade.

Step 5: Report results

The university is interested in whether foreign students do either better or worse than local students in terms of their academic performance. Local students are known to receive a mean grade of 62, with a standard deviation of 15. A random sample of 150 foreign students has a mean grade of 60.5. While this sample mean is lower than the mean grade for local students, we cannot reject the possibility that it is due to sampling error ($z = -1.2$, $p = 0.23$, two-tail), and that foreign students are no different to local students in terms of academic performance, as measured by mean grades.

The debate over one-tail and two-tail tests of significance

Within the field of statistics there is a dispute as to whether a one-tail test should ever be used, regardless of the form of the alternative hypothesis. Yet it has become routine to make this distinction in statistics textbooks without the underlying rationale for it ever being seriously considered. It has become a case of "everyone does it because everyone else does it!" The main argument against the use of a one-tail test is that the decision to use a one-tail test is arbitrary, and can lead to a statement of the alternative hypothesis using directional difference simply as a means of increasing the chance of rejecting the null hypothesis.

To see why, consider Figure 15.5. On a one-tail test, with an alpha level of 0.05, the region of rejection begins at either $z = -1.645$ or $z = +1.645$ but not both, depending on the direction of difference expressed in the alternative hypothesis. On a two-tail test this region has to be split in two because we are interested in a sample result *either* greater or smaller than the population value. This pushes the critical z-score outward to ±1.96. As a result, a sample mean will have to be further from the hypothesized value under a two-tail test before it falls in the region of rejection than under a one-tail test.

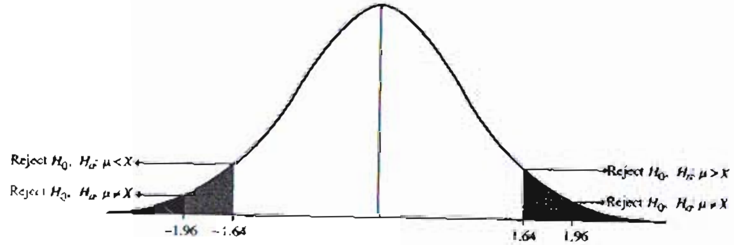


Figure 15.5 Critical regions for one-tail and two-tail tests, $\alpha = 0.05$

That is, since the one-tail significance is always half the two-tail significance, a result that may not lead to the rejection of the null hypothesis using a two-tail test may result in the rejection of the null using a one-tail test at a given alpha level. For example, if the two-tail significance is $p = 0.06$ (do not reject H_0 at $\alpha = 0.05$), the one-tail significance will be $p = 0.03$ (reject H_0 at $\alpha = 0.05$).

This alone should warrant caution in the use of one-tail tests, but the problem goes beyond the need to guard against arbitrary specification of the alternative hypothesis. The use of one-tail tests, in fact, could lead to some very bizarre conclusions about the null. For example, assume we are still testing whether a population has a mean age of 35 years, but we really suspect the population on average is younger than this. We thereby state the alternative hypothesis as $\mu < 35$ years. We analyze our sample and find that the mean age of the sample is 50 years, which has a two-tail p-value of 0.000002. It would be patently absurd not to reject the null hypothesis in this instance, simply because the sample result falls *above* the value specified in the null. The null hypothesis – the population from which the sample is drawn has a mean age of 35 years – is clearly at odds with the data. Yet the use of a one-tail test will cause us to live with the argument that the mean age of the population is 35 years and that we have drawn a 2-in-one-million random sample.

The untenability of this situation is made even more dramatic if we compare it to another outcome where the sample produces a mean age of 33 years, with a p-value of 0.04. Using a one-tail test we now reject the null hypothesis on the basis of a 2-year difference between the sample result and the hypothesized value, whereas previously we did not reject the null on the basis of a 15-year difference (50 – 35 = 15 years).

The point that needs to be borne in mind is that we are testing the null hypothesis as such, not the null hypothesis in relation to the alternative hypothesis. The null hypothesis can be contradicted by results that fall far enough away from its specified value *in either direction*. The direction specified in the alternative hypothesis, I argue, is not relevant to the strict logic of the hypothesis testing procedure, but rather *in determining what we do with the results*. We should always conduct two-tail tests, and if we find that the results are statistically significant we then consider whether the sample result is in the direction that provides evidence for our suppositions. This is similar to considering whether the statistical difference we have observed is large enough to be of any practical or theoretical importance; it should also be in the 'correct' direction. Thus we might find that a sample result is statistically significant, but because it is on the 'wrong side' of the sampling distribution it does not lend support to the argument we would like to make.

Despite these misgivings, I will use one-tail tests in the rest of this book, since they are so ingrained in the conventional methodology of hypothesis testing (one such example follows below). It is therefore important to understand the nature of such tests. In any event, the practical implications of this debate are not great, since any one-tail significance can be easily converted into a two-tail significance simply by doubling it. Thus for those who agree that the use of one-tail tests is never appropriate, they can simply double a one-tail p-value whenever it is derived and compare it with this two-tail significance to alpha.

A one-tail z-test for a single mean

A group of workers in a factory suspect that working conditions are unsafe and have caused them to suffer a high rate of illness. They call in a public health researcher who randomly selects 100 workers and asks each worker how many days work they lost in the previous year as a result of illness. The mean number of days lost was 10 days per worker. Official guidelines suggest that workers in this kind of setting should lose no more than 7 days a year due to illness, with permissible standard deviation of 7.5 days.

The union representing these workers argues that the sample result shows that they come from a population where the rate of illness has exceeded the official guidelines. Management, however, claims that the difference of 3 days (10 – 7 = 3) between the rate of illness in the sample and the official 'benchmark' is so small that it could easily be due to sampling error. Obviously there is some difference between the sample of factory workers and the benchmark of 7 days, but is this difference big enough to suggest that it is more than just random chance?

Step 1: State the hypotheses

H_0: The rate of illness suffered by all workers in this factory equals 7 days (i.e. does not exceed the benchmark):

$$H_0: \mu = 7 \text{ days}$$

H_a: The rate of illness suffered by all workers in this factory is greater than 7 days (i.e. does exceed the benchmark):

$$H_a: \mu > 7 \text{ days}$$

Notice that the alternative hypothesis does not just specify a difference, but also a direction of difference. The workers are only interested in rejecting the null if it shows that they have a *higher* incidence of illness to ground their claim for compensation. This will be important in determining whether to derive the one-tail or two-tail significance in Step 3.

Step 2: Choose the test of significance

We are interested in the mean number of hours lost for a single sample where the population standard deviation is known. We therefore conduct a *z*-test for a single mean.

Step 3: Describe the sample and derive the p-value

The mean number of days lost for a sample of 100 workers is 10 days. We put the sample result and hypothesized population value into the equation for *z* and derive the test statistic:

$$z_{sample} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{10-7}{7.5/\sqrt{100}} = 4$$

From the table for the area under the normal curve, under the column for the 'Area under the curve beyond one point', we find that the significance level associated with this sample *z*-score is $p < 0.0005$.

Step 4: Decide at what alpha level, if any, the result is statistically significant

The 3 days lost above the 7-day benchmark is statistically significant at the 0.01 level.

Step 5: Report results

The union may report the results of the statistical test in the following terms:

> Statistical analysis was conducted on a random sample of 100 workers to assess whether workers in the factory do not have a higher rate of illness than the maximum permissible rate set by government guidelines of 7 days lost due to illness. The sample had a mean number of days lost of 10, which represented three days more than that prescribed by the guidelines. This was found to be statistically significant ($z = 4$, $p < 0.0001$, one-tail). Morevover, the union argues that since three days lost represents a large amount of income foregone and distress to the workers and their families, the results are not just statistically significant, but also require management to take action and reduce the incidence of illness.

Summary

We have just worked through the steps involved in the most basic of hypothesis testing procedures: the *z*-test of a single mean. However, in practice this test is very rarely employed because it requires a great deal of information about the population. We begin with it, though, because it provides the clearest exposition of the process of hypothesis testing. Having learnt this basic procedure we are now able to deal with more complicated situations that are more likely to arise in 'real life'. The next chapters detail the tests to be used in these situations.

Appendix: Hypothesis testing using critical values of the test statistic

In earlier editions of this book, and in many other textbooks, the hypothesis testing procedure included the derivation of the critical scores associated with the critical regions defined by pre-set alpha levels. These critical scores are obtained by referring to the table for the area under the standard normal curve:

$$\alpha \rightarrow z_{critical}$$

Once we derive the critical value for *z* from the alpha level we can then compare this to the sample *z*-score and make a decision. In other words, we have two points of comparison for making a decision about the null hypothesis:

compare z_{sample} with $z_{critical}$

or

compare the *p*-value with the alpha level

Since any given *z*-score is uniquely related to a particular probability, and vice versa, we will get the same answer regardless of the comparison we choose to make.

Certain alpha levels are conventionally chosen in most research contexts, and the associated critical *z*-scores for these conventional levels of significance become familiar through regular use. If you work often enough with inferential statistics the following information (Table 15.6) will eventually be memorized. This is especially so for an alpha level of 0.05, which is by far the most common significance level used in research.

Table 15.6 Common critical scores

α	$z_{critical}$	
	Two-tail test	One-tail test
0.01	+ and − 2.58	+ or − 2.33
0.05	+ and − 1.96	+ or − 1.645
0.10	+ and − 1.645	+ or − 1.28

In this text we have done away with the calculation of critical scores associated with particular alpha levels for two reasons. First, as discussed above, we want to avoid determining the alpha level in advance; instead we now prefer to state the minimum alpha level at which the null can be rejected given the sample *p*-score. Second, deriving the critical scores introduces an unnecessary layer of calculations that makes an already complex procedure more complex. There is nothing of importance in *z*-scores as such (or the other test statistics we will come across in later chapters). They are just a means of deriving probabilities, and since we can compare probabilities directly with various alpha level, deriving their associated critical scores is unnecessary.

Exercises

15.1 Under what conditions is the sampling distribution of sample means normally distributed?

15.2 What is meant by type I and type II errors? How are they related?

15.3 How does the choice of significance level affect the critical region?

15.4 Complete the following table:

Probability	Test	*z*-score
0.230		±1.2
0.100	Two-tail	
0.018		
		±2.1
	Two-tail	±2.3
	One-tail	±3.4

15.5 Sketch the critical region for the following critical scores:

(a) $z > 1.645$ (b) $z < -1.645$ (c) $z > 1.96$ or $z < -1.96$

What is the probability of a type I error associated with each of these critical regions?

15.6 For each of the following sets of results, calculate z_{sample}:

	μ	σ	$\bar{X}$	n
(a)	2.4	0.7	2.3	180
(b)	18	11	16.7	100

15.7 A sample with a mean of 12 years is tested to see whether it comes from a population with a mean of 15 years.

(a) The significance level on a two-tail test proves to be 0.03. Explain in simple words what this indicates.

(b) The significance level on a one-tail test proves to be 0.015. Explain in simple words what this indicates.

15.8 A sample of nurses finds that they work on average 4.3 hours of overtime per week. This is tested to see whether the average amount of overtime worked by all nurses is 0 hours. The significance level proves to be $p = 0.00002$. Does this prove that the sample did not come from a population with a mean number hours of overtime per week of 0?

15.9 A particular judge has acquired a reputation as a 'hanging judge' because he is perceived as imposing harsher penalties for the same sentence. A random sample of 40 cases is taken from trials before this judge that resulted in a guilty verdict for a certain crime. The average jail sentence he imposed for this sample is 27 months. For all crimes of this type the average prison sentence is 24 months, with a standard deviation of 11 months (assume a normal distribution). Is this judge's reputation justified? (Pay close attention to the form of the alternative hypothesis.)

16

The one sample *t*-test for a mean

The previous chapter introduced the logic of hypothesis testing. The careful reader will have noticed that in conducting the one sample *z*-test for a mean we used the population standard deviation to make an inference from the sample mean to the population mean. The careful *and critical* reader will have thought this a peculiar situation: the data used to calculate the standard deviation for the population should also allow us to directly calculate the mean; if we know the standard deviation for the population how can we *not know* the population mean? We should not need to make an inference from the sample to the population mean, but should be able to directly calculate it.

In other words, it is unlikely that we will ever find a situation where we do know the population standard deviation but do not know the population mean. Indeed, SPSS does not even provide an option for a *z*-test for a single mean. Before you suddenly decide that the previous chapter was a waste of time and tear it out of the book, let me justify why we spent so much time learning a test that we are unlikely ever to use in practice. We begin with the one sample *z*-test for a mean because it is the simplest illustration of inferential statistics. Having learnt the basic logic in this, albeit unrealistic, situation, we can then go on and apply it to more relevant, but slightly more complicated, situations. Thus the previous chapter allowed us to sharpen our hypothesis testing 'knife' so that we can use it to 'slice through' more real-life problems.

The tests that follow in the ensuing chapters are all variations of the basic hypothesis testing procedure. We will learn the specific conditions under which each test is relevant. These are the factors we look for in Step 2 of the hypothesis testing procedure to determine the test of significance to employ. The two key factors to consider (although there are others) are, first, the descriptive statistic that is used to summarize the sample data, and, second, the number of samples from which inferences are to be made. This chapter will detail the **one sample *t*-test** for a mean, which is used instead of the one sample *z*-test for a mean in the more common situation where neither the value of the population mean nor the population standard deviation are known.

The Student's *t*-distribution

When we want to make an inference about a population mean but don't know the standard deviation of the population a slight change is required to the basic procedure outlined in the previous chapter. We no longer use the *z*-distribution to derive the *p*-value of the sample statistic. This is because the sampling distribution of sample means will no longer be normal. Instead, the sampling distribution we use is the **Student's *t*-distribution**, and we conduct a ***t*-test**. (It is called the Student's *t*-distribution after W. Gossett who first defined its properties. As an employee of the Guinness brewing company, he was not permitted to publish under his own name. He therefore chose 'the Student' as his alias.)

A *t*-distribution looks like a *z*-distribution in that it is a smooth, unimodal, symmetrical curve. The difference is that a *t*-distribution is 'flatter' than the *z*-distribution. Exactly how much flatter depends on the sample size (Figure 16.1). The *t*-distribution where sample size is 30 has much 'fatter tails'; these tails become thinner for a sample size of 90; and eventually the *t*-distribution is identical to the normal curve when sample size becomes very large (greater than 120).

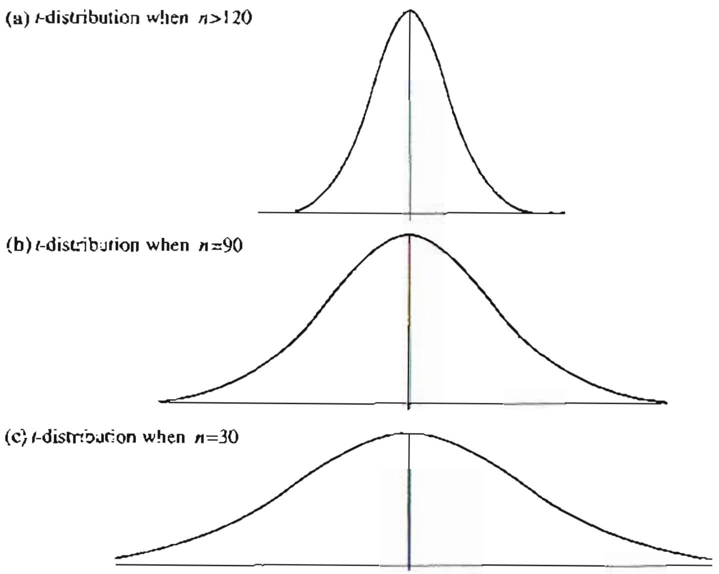


Figure 16.1 t-distributions for sample sizes (a) $n > 120$, (b) $n = 90$, and (c) $n = 30$

The one sample t-test for a mean

We will detail the one sample t-test for a mean by working through an example using the five-step hypothesis testing procedure, indicating as we do the ways in which this test varies from the z-test for a mean.

Assume that the Health Department, in order to decide how much money it should allocate to the local hospital, is interested in whether the average age for the population in the region is over 40 years. Unable to survey the whole area, the Department takes a random sample of 51 people from this population, which yields a sample mean of 43 years and standard deviation of 10 years. Clearly the sample is on average older than 40 years. The Department is reluctant to conclude from this, however, that the *population* from which the sample is drawn is on average over 40 years of age. The Department argues that the sample could easily have come from a population with a mean age of only 40 and the effect of random variation explains the slightly older sample result. We can test this claim using the one sample t-test for a mean.

Step 1: State the null and alternative hypotheses

H_0: The population in this region has a mean age of 40 years.

$$H_0\text{: } \mu = 40 \text{ years}$$

H_a: The population in this region has a mean age greater than 40 years.

$$H_a\text{: } \mu > 40 \text{ years}$$

Notice the inequality in the statement of the alternative hypothesis. Given the Health Department's policy on funding we are not interested in whether the population in this region is on average younger than 40: its funding will only change if we find that the average age of the population is significantly *older* than 40 years.

Step 2: Choose the test of significance

We are interested in the *mean* age for *one* sample. Unlike the examples in the previous chapter, we do not have any information regarding the population standard deviation, so we use the one sample t-test for a mean.

Step 3: Describe the sample and calculate the p-value

When the population standard deviation is unknown we use the following equation for t to calculate the sample score (instead of the equation for z). This equation substitutes the sample standard deviation for the population standard deviation.

$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

We can substitute the data we have in our example to calculate the test statistic:

$$t_{sample} = \frac{\overline{X} - \mu}{s/\sqrt{n}} = \frac{43-40}{10/\sqrt{51}} = 2.1$$

To obtain the p-value for this t-score we refer to the critical values for t-distributions in Table A2, and partially reproduced here as Table 16.1. In order to use this table we first need to determine the **degrees of freedom**. The concept of degrees of freedom can be illustrated with a simple example. If there are five students and their final exam grades must have an average of 10, a restriction has been placed on the range of possible scores these students can get. For example, if the first four marks are 12, 7, 15, and 11, the fifth mark *must be* 5 for the total to produce the average of 10, which is the restriction I have imposed on the data. We have lost one degree of freedom (*df*) because we have imposed a certain result on the data. Instead of n degrees of freedom, where n is the sample size, we have $n - 1$. In this example, we have four degrees of freedom.

A similar correction applies when working with t-tests. The t-test is based on the *assumption that the population standard deviation* (which is unknown) *is equal to the sample standard deviation* (which is known). The imposition of this assumption on the data means we lose one degree of freedom.

$$df = n - 1$$

The degrees of freedom affect the likelihood that any given sample mean will be significantly different from the test value. For any given alpha level, a select number of which are listed across the top of the table, the t-score that will mark off that area under the curve will be 'further out' with small samples (fewer degrees of freedom) than it will be for larger samples (more degrees of freedom). This means that the larger the sample size (and therefore degrees of freedom) the more likely that any difference between the sample mean and the test value will prove to be significant. For example, with a sample of 150 ($df = 149$), the critical regions for a $\alpha = 0.05$ (the lightly-shaded areas in Figure 16.2) begin closer to the test value than for a sample of only 51 ($df = 50$), which has critical regions marked off by the darker shaded areas in Figure 16.2.

Since the t-distribution is 'flatter' with smaller samples, the critical regions lie further out compared with the t-distribution for the larger sample. As a result Table 16.1 provides a set of t-scores and levels of significance for various degrees of freedom (*df*).

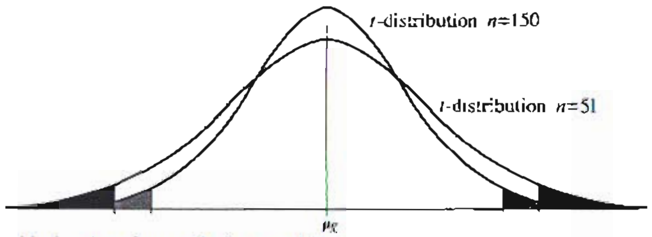


Figure 16.2 Critical regions for sample sizes n = 150 and n = 51, α = 0.05

Table 16.1 Critical values for t-distributions

	Level of significance for one-tail test				
	0.10	0.05	0.02	0.01	0.005
	Level of significance for two-tail test				
df	0.20	0.10	0.05	0.02	0.01
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
⋮	⋮	⋮	⋮	⋮	⋮
35	1.306	1.690	2.030	2.438	2.724
40	1.303	1.684	2.021	2.423	2.704
45	1.301	1.679	2.014	2.412	2.690
50	1.299	1.676	2.009	2.403	2.678
55	1.297	1.673	2.004	2.396	2.668
60	1.296	1.671	2.000	2.390	2.660
70	1.294	1.667	1.994	2.381	2.648
80	1.292	1.664	1.990	2.374	2.639
90	1.291	1.662	1.987	2.368	2.632
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

To derive the p-value for a particular t-score from this table we need to:

1. move down the first column of the table below *df* until we identify the row with the desired number of degrees of freedom, in this instance we identify *df* = 50 (if the degrees of freedom for our sample are omitted from the table, we should locate the row with the closest degrees of freedom *below* the sample *df*);
2. then move across this row until we identify the t-scores between which the sample t-score falls, in this instance our t_{sample} of 2.1 lies between the scores in the table of 2.009 and 2.403 (this is unlike the table for areas under the normal curve in the previous chapter where we could locate the exact z-score for the sample mean);
3. we then move up these columns and read off the associated p-values for these t-scores, choosing the values for either one-tail or two-tail significance according to the specification of the alternative hypothesis. Here the p-score lies between 0.02 and 0.01.

We can also obtain the p-value from various web pages that provide statistical calculation options. Two such pages that will perform a t-test on a sample mean are:

1. *Statistical Applets*, www.assumption.edu/users/avadum/applets/applets.html and click on the **t test: One Sample** option on the left-menu;
2. GraphPad's *QuickCalcs*, graphpad.com/quickcalcs/OneSampleT1.cfm

These pages not only provide the t-score, but also the exact p-value, unlike the table we used in the band calculations, which only provides a range of values between within which the p-value falls. From these pages I determined that the two-tail significance level is 0.037 and the one-tail significance is 0.0185, which falls within the range of p-values we obtained from the table.

Step 4: Decide at what alpha level, if any, the result is statistically significant

The p-value we obtained in Step 3, regardless of whether we use a one-tail or two-tail test, is significant at the 0.05 level (i.e. the p-value is less than 0.05), but not significant at the 0.01 level (i.e. the p-value is not less than 0.01). Although *it is possible* to draw a sample with a mean age of 43 or higher from a population with a mean age of only 40, this will only occur less than five-in-every-hundred times.

Step 5: Report results

Given the results we have obtained, the Health Department may conclude the following:

> Based on a sample of 51 people with a mean age of 43 years and a standard deviation of 10 years, we found the results to be statistically significantly different from 40 years (t = 2.1, p = 0.0185, one-tail). However, the fact that the sample was only 3 years above the benchmark age for increased funding to the local hospital, although statistically significant, is not very large in real terms, and therefore may not justify a large increase in funding to meet the extra health needs of only a slightly older population.

Notice that in this conclusion we have been careful to draw a distinction between statistical significance and practical significance. At what point a mean age greater than 40 years becomes large enough to warrant a major increase in hospital funding (regardless of its statistical significance) is a policy decision for the Health Department and not an issue that statistics can answer.

Notice also that a slightly different conclusion is also open to the Health Department. While the results are significant at the 0.05 level, given that important funding decisions are at stake whereby an increase in funding to one hospital may lead to reduced funding to other hospitals, the Department may only be prepared to reject the null at the 0.01 level, since it wants to minimize the risk of a Type 1 error (rejecting the null when it is in fact correct). By providing the p-score for the sample result such a decision is available to anyone who regards a Type 1 error in this context to be a serious problem.

Looking back at this example we can see that there are some slight changes to the hypothesis testing procedure we introduced in the previous chapter. These changes take account of the fact that we do not know the standard deviation for the population about which we want to make an inference. In particular, we use a slightly different formula in Step 3 to derive the test statistic; and we refer to a slightly different sampling distribution to derive the p-score, one that requires us to consider the degrees of freedom we are working with. Apart from these modifications the procedure is basically the same. In order to familiarize ourselves further with the one sample t-test for a mean, we will now work through a number of examples.

Example

According to AC Nielsen, a market research company, children in Britain between the ages of 5 and 12 years watch on average 196 minutes of TV per day. For the sake of exposition we will assume that this is the value for the population of all British children in this age bracket. A survey is conducted by randomly selecting 20 Australian children within this age group to see if Australian children are significantly different from their British counterparts in terms of the average amount of TV watched per night.

The null hypothesis is that Australian children watch on average the same amount of TV each night as their British counterparts:

$$H_0: \mu = 196 \text{ minutes}$$

The alternative hypothesis is that Australian children on average watch a different amount of TV than their British counterparts:

$$H_a: \mu \neq 196 \text{ minutes}$$

Note that in this research question we are simply interested in whether there is a difference between Australian children and the hypothesized value of 196 minutes. We are not specifically concerned whether Australian children watch significantly more or significantly less, just whether they are different.

The sample of 20 Australian children has a mean of 166 minutes of TV viewing, with a standard deviation of 29 minutes. Substituting this information into the equation for t, we get a sample t-score of: –4.6:

$$t_{sample} = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{166 - 196}{29/\sqrt{20}}$$

$$= -4.6$$

From the table for the critical values for t-distributions, at 19 degrees of freedom, the sample score has a significance level of less than 0.005. It is therefore statistically significant and leads us to reject the statement that Australian and British children watch on average the same amount of television per day. Whether the difference of 30 minutes we observed is of any practical significance is something I will leave for you to consider (that is, do children who watch 30 more minutes of TV per day 'suffer' in any important sense?).

The one sample *t*-test using SPSS

We will now work through this example using SPSS. The data for the 20 Australian children are entered into SPSS. The procedure for generating a one sample t-test on these data is detailed in Table 16.3 and Figure 16.3. Figure 16.3 also shows the SPSS output from this command.

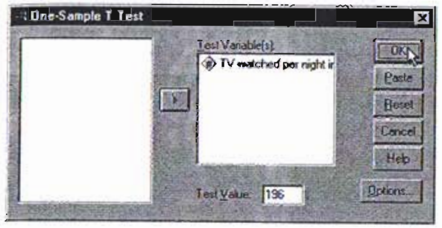
The first part of the output is the **One sample Statistics** table, which provides all the descriptive statistics: the number of cases (20), the mean for the sample (165.85), and the standard deviation for the sample (29.29). The last column for the first table, Std. Error Mean, is the standard deviation of the sampling distribution for t for this number of degrees of freedom. This is the value in the denominator of the equation for t.

The second part of the output is the **One sample Test** table, which contains the results of the inference test. The t-value is –4.603, as we have already calculated, which at 19 degrees of freedom (df) has a two-tail significance of less than 0.0005 (SPSS has rounded this off to 3 decimal places). The difference between the test value of 196 and the sample mean is the Mean Difference of –30.15. This is the numerator of the equation for t.

Given the very low probability of obtaining a sample with a mean of 165.85 or less from a population that watches on average 196 minutes of TV a day, we reject the hypothesis that the population mean is 196 minutes: Australian children do not watch the same amount of TV on average than children in Britain.

Table 16.3 The One sample T Test command using SPSS (file: Ch16.sav)

SPSS command/action	Comments
1 From the menu select Analyze/Compare Means/ **One sample T Test**	This brings up the **One sample T Test** dialog box
2 Select **TV watched per night** from the source variables list	
3 Click on ▸	This pastes **TV watched per night** into the target list headed **Test Variable(s):**
4 In the text-box next to **Test value:** type 196	
5 Click on **OK**	



T-Test

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
TV watched per night in minutes	20	165.85	29.29	6.55

One-Sample Test

	Test Value = 196					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
TV watched per night in minutes	-4.603	19	.000	-30.15	-43.86	-16.44

Figure 16.3 The SPSS One sample T Test dialog box and output

Another way of reaching the same inference about the population is to look at the confidence interval constructed around the sample mean, a topic we shall detail in the following chapter, but which we will briefly discuss here to give a complete meaning to the SPSS output we have generated. This information is provided in the last column of the **One sample Test** table. At a 95 percent confidence level the interval for the difference between the sample and the test value ranges from a lower limit of –43.86 to an upper limit of –16.44. In other words the difference between the average amount of TV watched by the population of all children and the hypothesized value we estimate to lie somewhere in this range, at a 95 percent confidence level. Since this range does not include the value of zero, which would indicate no difference, we can reject the hypothesis of no difference.

Summary

In this chapter we have worked through the one sample t-test for a mean. It is in most respects equivalent to the one sample z-test for a mean, which we introduced in the previous chapter, but is used in the more usual situation where we do not know the population standard deviation. In fact, the two tests are identical where the sample size is greater than 120. If we look at the last line of Table 16.1 for the t-distribution that has the infinity symbol, ∞, the scores should be familiar. For example, on a two-tail test at an alpha of 0.05 the t-score is 1.96. This is exactly the same value for z at this level of significance. In other words, when sample size is greater than 120, the t-distribution and the z-distribution are identical, so that the areas under the respective curves for any given scores are also identical.

An important, but often neglected, assumption behind the use of t-tests needs to be pointed out before moving on. With small samples, the sampling distribution of sample means will have a t-distribution only where the underlying population is normally distributed. This assumption is robust in that the sampling distribution will still approximate a t-distribution even where the population is moderately non-normal. Even so, we should be cautious about conducting a t-test without thinking about the validity of this assumption first. Chapter 22 provides a way of assessing this assumption based on the sample data, and if there is reason to believe that this assumption does not hold, a whole range of non-parametric tests are available. These will be investigated in the later chapters.

Exercises

16.1 What assumption about the distribution of the population underlies the t-test?

16.2 From the table for critical scores for t-distributions, fill in the following table:

t-score	Probability	Test	df
2.015		One-tail	5
	0.02	Two-tail	10
1.708	0.05	One-tail	
	0.05	Two-tail	65
	0.10	One-tail	228

16.3 Conduct a t-test, with $\alpha = 0.05$, on each of the following sets of data:

	Sample mean	s	H_0	H_a	n
(a)	62.4	14.1	$\mu = 68$	$\mu \neq 68$	61
(b)	62.4	14.1	$\mu = 58$	$\mu < 68$	61
(c)	2.3	1.8	$\mu = 3.1$	$\mu \neq 3.1$	25
(d)	2.3	1.8	$\mu = 3.1$	$\mu \neq 3.1$	190
(e)	102	45	$\mu = 98$	$\mu \neq 98$	210
(f)	102	45	$\mu = 90$	$\mu \neq 90$	210

16.4 To gauge the effect of enterprise bargaining agreements, union officials sampled a total of 120 workers from randomly selected enterprises across an industry. The average wage rise in the previous year for these 120 workers was \$1018, with a standard deviation of \$614. The union is worried that its workers have not reached its bargaining aim of securing a wage rise of \$1150. Conduct a two-tail t-test to assess whether this objective has been met.

16.5 The following data are ages at death, in years, for a sample of people who were all born in the same year:

34, 60, 72, 55, 68, 12, 48, 69, 78, 42, 60, 81, 72, 58, 70, 54, 85, 68, 74, 59, 67, 76, 55, 87, 70

(a) Calculate the mean age at death and standard deviation for this sample.

(b) What is the probability of randomly obtaining this sample from a population with an average life expectancy of 70 years?

(c) Enter these data into SPSS and check your answers.

16.6 A health worker wants to gauge the effect of hip fractures on people's ability to walk. On average, people walk at a rate of 1 meter per second. Walking speed for 43 individuals who had suffered a hip fracture 6 months previously averaged 0.44 m/s, with a standard deviation of 0.28 m/s. What should the health worker conclude?

16.7 AC Nielsen has provided the following figures for the average number of minutes of TV watched by children in some selected countries:

Country	Mean viewing time, minutes
Australia	159
Canada	140
Britain	196
Singapore	212

In the text we compared the hypothetical results of a survey of 20 Australian children, which had an average viewing time of 166 minutes and standard deviation of 29 minutes, with the 'population' value for Britain. Compare this sample with the population values for Canada and Singapore, as well as the population value for Australia, and test whether there is a significant difference.

16.8 In Chapter 9 we used the following data for the weekly income of 20 people in a sample:

\$0, \$0, \$250, \$300, \$360, \$375, \$400, \$400, \$400, \$420, \$425, \$450, \$462, \$470, \$475, \$502, \$520, \$560, \$700, \$1020

The mean for these data we calculated to be \$424.45, with a standard deviation of \$216.

(a) Conduct a t-test, with $\alpha = 0.05$, to assess the probability that this sample is drawn from a population with a mean weekly income of \$480.

(b) Enter these data into SPSS, and conduct the same t-test.

16.9 Open the **Employee data** file.

(a) Generate the mean and standard deviation for the current salary of workers in the sample.

(b) Assume that the average salary for all other workers is \$25,000. Conduct by hand (showing all working) a t-test to assess whether there is a significant difference between the employees in this firm and all other employees. State your conclusion in simple terms.

(c) Conduct this test on SPSS and check that your hand calculations conform to the SPSS output.

(d) Assume that the average salary for all other workers is \$33,000. Conduct by hand (showing all working) a one sample t-test to assess whether there is a significant difference between the employees in this firm and all other employees.

(e) Conduct this test on SPSS and check that your hand calculations conform to the SPSS output.

(f) Assume that your research question is whether the employees in this firm are paid significantly *more* than other employees. Will your answer to part (d) be any different? Explain.

17

Inference using estimation and confidence intervals

In the previous two chapters we introduced a method for making an inference about the value of a population mean from the mean of a random sample. This method is the hypothesis testing procedure, which begins with the statement of a null hypothesis that specifies the population mean equals a particular value (e.g. the mean age of the population is 35 years).

This hypothesis testing procedure for making an inference has come under heavy criticism. We discussed some of those criticisms in Chapter 15, but there are others of a more fundamental nature (see Gardner, M.J. and Altman, D.G., 1986, Confidence intervals rather than P values: Estimation rather than hypothesis testing, *British Medical Journal*, March, pp. 746–50). Of most concern has been the fact that the hypothesis testing procedure only tests whether the sample result is significantly different from a *particular* 'test' value specified in the null hypothesis. We discussed the criteria for choosing a test value in Chapter 15, but we can see that there is still an element of arbitrariness in the selection of this test value. Moreover, it is logically possible to find that a sample result is significantly different to many possible 'test' values; equally there is obviously a whole range of values from which the sample result will not be significantly different (at a given alpha level).

For example, on the basis of our sample that had a mean age of 32 years, we could have tested the hypothesis that the population mean is 33 years, 34.5 years, 30.2 years, and so on. Some of these tests will yield very low *p*-scores and some will not. A single test against a single value of 35 years tells us only whether the sample is significantly different to this one score. In this sense, the hypothesis testing procedure is extremely limited.

To overcome this limitation another procedure for making an inference from a sample to a population has been developed. This is the process of **estimation** through the construction of **confidence intervals**. Some see this as an alternative to the hypothesis testing procedure discussed in the previous two chapters. This text, however, presents the estimation procedure as a complimentary procedure for reaching the same conclusions that can be reached by hypothesis testing; it provides additional information for making an inference from a sample result to a population, but also has limitations of its own. Thus if we take the two procedures together we can obtain a more complete picture than if one procedure is used exclusively.

Fortunately, SPSS usually provides the estimation values along with the results of hypothesis tests, so that no additional commands need to be run. The important thing is the meaning of these results and how they help us expand our ability to make an inference from a sample to a population. Before we detail this estimation procedure, we need to first remind ourselves of the main conclusions we reached in Chapter 14 regarding the sampling distribution of sample means (a quick read over that chapter at this point may be helpful in understanding what follows).

The sampling distribution of sample means

The sampling distribution of sample means has three very important properties:

1. *The mean of the sampling distribution is equal to the population mean.* Although the mean of any individual sample may differ from that of the population from which it is drawn, repeated random sample means will cluster around the 'true' population value.

$$\mu_{\bar{X}} = \mu$$

In other words, although individual results will vary from sample to sample, on average the sample means will be equal to that of the population. This property of a sample mean makes it an **unbiased estimator** of the population value.

A sample statistic is **unbiased** if its sampling distribution has a mean equal to the population parameter it is estimating.

2. *The spread of sample results around the population value is affected by the sample size.* The standard deviation of the sampling distribution, called the standard error, is defined by the following equation:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

As sample size increases the standard error of the sampling distribution gets smaller, so that sample results are more *tightly* clustered around the population value. In other words, large samples provide **efficient estimators** of the population values.

3. *The sampling distribution of sample means is normal.* The proportion of sample means that will fall within a certain range of values will be given by the standard normal distribution.

These three properties of the sampling distribution of sample means allow us to refer to the table for the area under the standard normal curve (Appendix A1) in order to gauge the probability that any given sample mean will be within a certain range of values around the population mean. For example, we know that around 95 percent of repeated samples will have a mean within 1.96 standard errors of the population mean (Figure 17.1).

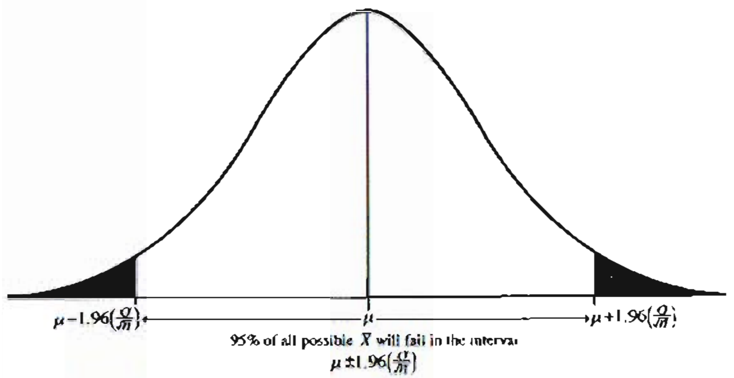


Figure 17.1 The sampling distribution of sample means

This allows us to specify a range or **interval** of scores within which 95 percent of all possible sample means will fall, defined by the formula:

$$\mu \pm 1.96\left(\frac{\sigma}{\sqrt{n}}\right)$$

Estimation

In Figure 17.1 we posed the problem in a certain way. We have a population parameter, and estimate the range of values that 95 percent of all random samples drawn from that population will take. However, in research the problem usually poses itself in a different way. We have a single sample result, and we need to estimate the population value from the sample. Whereas in the hypothesis testing procedure we asked "Does the population mean equal X?", in estimation we ask the broader question "What is the population mean?"

For example, in Chapter 14 we analyzed a population of 1200 residents of a hypothetical community. The mean age for this population is 35 years. We take random samples of size n = 125 from this population, and observe that the averages for each of these samples is not equal to the population parameter, but most of them cluster around the population value. But what if we do not know that the mean age of the population is 35 years, and all we have to work with is *one* of these samples of 125 residents? Let us assume that this one and only sample is the one that has an average age of 34.5 years and our task is to estimate the population parameter (which for the moment we are pretending we do not know) from this one sample result.

In estimating the population parameter we start with an assumption. We assume that the sample actually *falls* within a certain region of the sampling distribution. We assume that the sample mean is not one of those few, very unlikely and extreme results that are very different from the population value. For example, we might feel comfortable with the assumption that this one sample of 125 residents is one of the 95 percent of all *possible* samples that will fall within ±1.96 standard errors from the population mean.

Remember that this is only an assumption: we may have actually drawn one of these freakish samples with a mean very different from the population parameter. We can never know if this is the case, but given the very low probability of this being the case (less than 5-in-100), the assumption seems reasonable. In other words we can be confident that this assumption is correct. In fact, we call this assumed probability the **confidence level**; in this instance we choose a 95 percent confidence level.

Given this assumption – that the sample result is within the range that 95 percent of all possible sample results will fall – we can make an estimate of the population value. We know that the standard deviation of the sampling distribution, called the **standard error**, is:

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}$$

Here, though, we do not know the standard deviation for the population, σ, so we use the sample standard deviation instead, which for this sample equals 13 years.

$$\sigma_{\overline{X}} = \frac{s}{\sqrt{n}} = \frac{13}{\sqrt{125}} = 1.2 \text{ years}$$

As discussed in the previous chapter, the use of the sample standard deviation rather than the (unknown) population standard deviation requires us to use t-scores rather than z-scores to construct our estimate:

- We look up the table for critical values of the t-distribution (Appendix A2), for the number of degrees of freedom we are working with. Here $df = 125 - 1 = 124$. Since this is greater than 120, we refer to the last row of the table;
- We then read off the t-score by referring to the column of values under the equivalent alpha level to our selected confidence level (here 95% = 0.05). In our example the appropriate t-score is 1.96.
- We then use this t-score in the following equation to multiply the standard error.

$$\overline{X} \pm t\left(\frac{s}{\sqrt{n}}\right)$$

What does this mean? The furthest the population parameter can be *below* the sample value such that the sample value remains within the 95 percent region is −1.96 standard errors. This is called the **lower limit** of the estimate. It sets the maximum distance that the population value can be below the sample (Figure 17.2) for that sample to still be within the range of values within which 95 percent of all samples will fall:

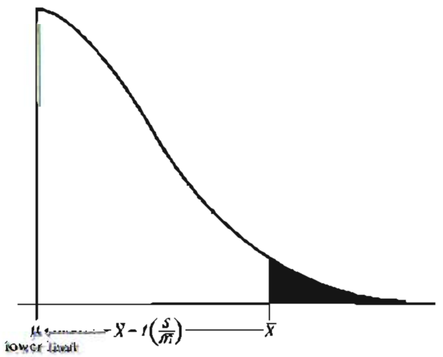


Figure 17.2 The lower limit of the estimate

$$\text{lower limit} = \overline{X} - t\left(\frac{s}{\sqrt{n}}\right)$$

Using similar reasoning, the furthest the population parameter can be *above* the sample value so that the sample value is still within the 95 percent region is +1.96 standard errors. This is called the **upper limit** and is illustrated in Figure 17.3.

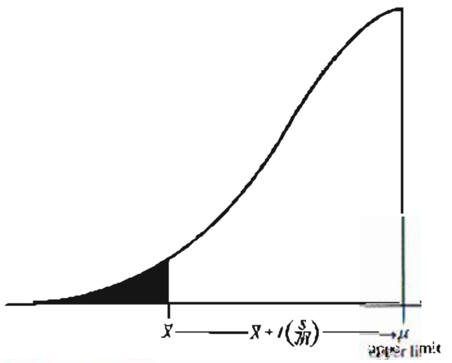


Figure 17.3 The upper limit of the estimate

$$\text{upper limit} = \overline{X} + t\left(\frac{s}{\sqrt{n}}\right)$$

Putting these two pieces of logic together allows us to define a range of values, called a **confidence interval** (*ci*), within which, we estimate, lies the population mean.

A **confidence interval** is the range of values that, it is estimated, includes a population parameter, at a specified level of confidence.

The steps involved in determining the lower limit and the upper limit of a confidence interval can be combined in the following equation (note the value for t may differ from 1.96 according the sample size you are working with):

$$ci = \bar{X} \pm t\left(\frac{s}{\sqrt{n}}\right)$$

This equation simply states that we add and subtract from the sample result a distance defined by the maximum number of t-scores we assume the sample result can be from the population parameter, at the given confidence level. In the example of the age of our residents, the lower and upper limits are:

$$\text{lower limit} = \bar{X} - t\left(\frac{s}{\sqrt{n}}\right) = 34.5 - 1.96\left(\frac{13}{\sqrt{125}}\right) = 32.2 \text{ years}$$

$$\text{upper limit} = \bar{X} + t\left(\frac{s}{\sqrt{n}}\right) = 34.5 + 1.96\left(\frac{13}{\sqrt{125}}\right) = 36.8 \text{ years}$$

We write such an estimate in the following way: 34.5 [32.2, 36.8].

We have constructed a confidence *interval* because in estimating the average age of the population from a single sample we need to allow for the effects of sampling variation. Looking at the estimate we have constructed from this sample, we can see that it includes the actual population average of 35 years, which we pretended we did not know. The confidence interval is accurate in that the range of values between 32.2 and 36.8 years includes this actual population mean. Normally we do not know whether the estimate is accurate, but the confidence level indicates the probability of being accurate.

In fact I have constructed a confidence interval around all the 20 random samples drawn in Chapter 14 from this population. The sample averages have been graphed and the confidence intervals around them drawn in Figure 17.4.

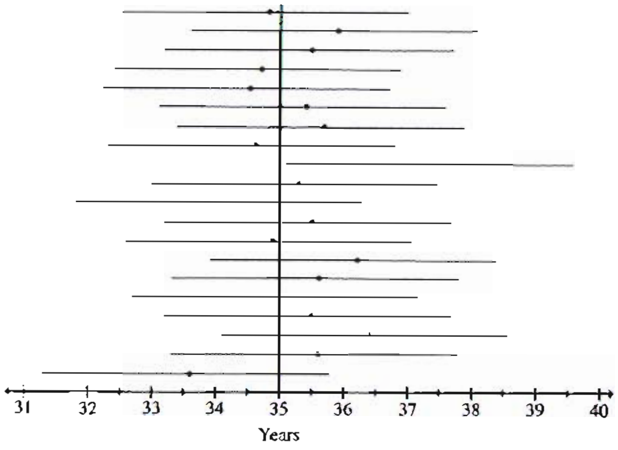


Figure 17.4 Twenty confidence intervals (95% level)

Looking at Figure 17.4 we can see the potential problem with making an estimate using sample results. If the one sample we drew happened to be the one that produced an average age of 37.4 years, our estimate will be inaccurate. The assumption that this is one of the 95 percent of samples that fall within 1.96 standard errors from the population value is invalid: it is one of those 5-in-100 samples that fall a relatively long distance from the population mean. Therefore the interval constructed on the basis of a 95 percent confidence level does not include the parameter of 35 years. We can never know whether this is the case – whether the one sample we do undertake just happens to be 'freakish'. However, we can see from Figure 17.5 that such an event is highly unlikely. In fact, 19 out of the 20 intervals do include the true population value of 35 years, which is in accord with the confidence level of 95 percent.

Another way to think of this is that with a confidence level of 95 percent we are prepared to be wrong only five times in every 100 samples (i.e. 1-in-20). This is the risk we take of not including the population parameter in our interval estimate, given that we have to make an estimate based on a sample that is affected by random variation. This probability of error is known as the alpha level (α), which is simply one minus the confidence level (expressed as a proportion). Thus the 95 percent (0.95) confidence level is the same as an alpha level of α= 0.05. A 90 percent confidence level is the same as an alpha level of α= 0.10, or a risk of being wrong 1 time in every 10.

Changing the confidence level

In this discussion, we chose a confidence level of 95 percent. This is why we multiplied the standard error by a t-score of 1.96, since this defines the region under the sampling distribution that includes 95 percent of repeated sample results, at this number of degrees of freedom. This is the commonly used confidence level, but we can choose either larger or smaller levels, depending on how sure we want to be that our interval has 'taken in' the population mean. The larger the confidence level the more likely that the interval derived from it will include the population mean. If we choose a 99 percent confidence interval, for example, then we are assuming that a given sample mean is one of the 99-in-100 that falls 2.58 standard errors either side of the true mean:

$$ci = \bar{X} \pm 2.58\left(\frac{s}{\sqrt{n}}\right)$$

Making the starting assumption safer, however, by choosing a larger confidence level comes at a cost. In order for us to argue that the sample is one of the 99 percent that fall within a certain region around the true value, that region has to be widened. Rather than multiplying the standard error by t = 1.96, we multiply by t = 2.58. It is like firing an arrow at a target. Making an assumption that an arrow is likely to fall within 1 meter of the bullseye is safer than making the assumption that it will fall within 10 centimeters of the bullseye, but it has come at the cost of some accuracy. Making the target 'bigger' by widening the confidence interval means we are more likely to 'hit it' (i.e. make sure that the interval includes the population value), but we are no longer as precise in our shooting.

To see the effect of choosing different confidence levels we will work through the following example. A random sample of 200 nurses is taken and each nurse asked his or her annual income in whole dollars. These 200 nurses have a mean income of \$35,000, with a standard deviation of \$5000. What is our estimate for the average annual income of all nurses?

With a 95 percent confidence interval the range (rounded to the nearest whole dollar) is:

$$ci = \bar{X} \pm t\left(\frac{s}{\sqrt{n}}\right) = 35,000 \pm 1.96\left(\frac{5000}{\sqrt{200}}\right) = \$35,000 \pm 695$$

The lower and upper limits will be $34,305 and $35,695 respectively:

lower limit: 35,000 – 695 = $34,305

upper limit: 35,000 + 695 = $35,695

We therefore estimate that the average income of all nurses, *with a 95 percent level of confidence*, will lie within the following range:

$$\$34,305 \leq \mu \leq \$35,695$$

The width of this interval (the difference between the upper and lower limits) is $1390 (i.e. 35,695 – 34,305).

With a 99 percent confidence interval, the *t*-score we use in the calculation is 2.58. The calculation will thereby be:

$$ci = \bar{X} \pm t\left(\frac{s}{\sqrt{n}}\right) = 35,000 \pm 2.58\left(\frac{5000}{\sqrt{200}}\right) = 35,000 \pm 915$$

$$= \$35,000\ [34,085,\ 35,915]$$

To be more confident that the interval will actually contain the true population value, it has become much wider; it now ranges from $34,085 to $35,915. The interval width is $1830.

If, on the other hand, I want to be more precise in my estimate I will choose a 90 percent confidence level, but this will be at the higher risk of being wrong. The *t*-score I get from the table is that for alpha = 0.10, which is $t = 1.645$. The confidence interval will be:

$$ci = 35,000 \pm 1.645\left(\frac{5000}{\sqrt{200}}\right) = \$35,000\ [34,415,\ 35,585]$$

The effect of these changes to the confidence level on our estimates is summarized in Table 17.1 and Figure 17.5. Using a smaller confidence level reduces the interval width in which we estimate the population mean lies. However, because this interval width is smaller the chances of being wrong (which is equal to the alpha level) have also increased. Having a narrower range of values increases the chance that it will not include the mean of the population. Making the bullseye on a target smaller allows us to say that we are better archers if we hit it, but it also increases the chances of not hitting it. On the other hand, choosing a confidence level of 99 percent widens the interval estimate so that it is more likely to include the population value, but it may as a result make the estimate meaningless from a theoretical or practical point of view. Knowing that the mean annual income of nurses can be anywhere between $34,085 and $35,915 may actually be saying nothing of practical importance.

Table 17.1 Effect of confidence levels on intervals

Confidence level (%)	*t*-score	Confidence interval	Interval width
90	1.645	$35,000 ± 585	$1170
95	1.96	$35,000 ± 695	$1390
99	2.58	$35,000 ± 915	$1830

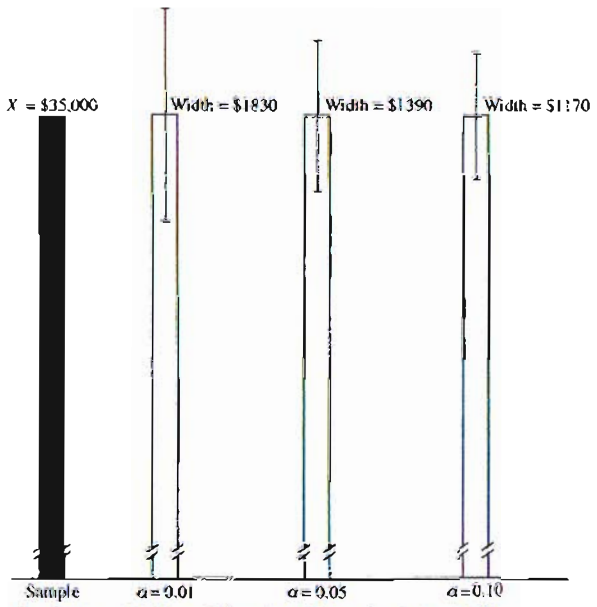


Figure 17.5 Interval estimates with three different confidence levels (n = 200)

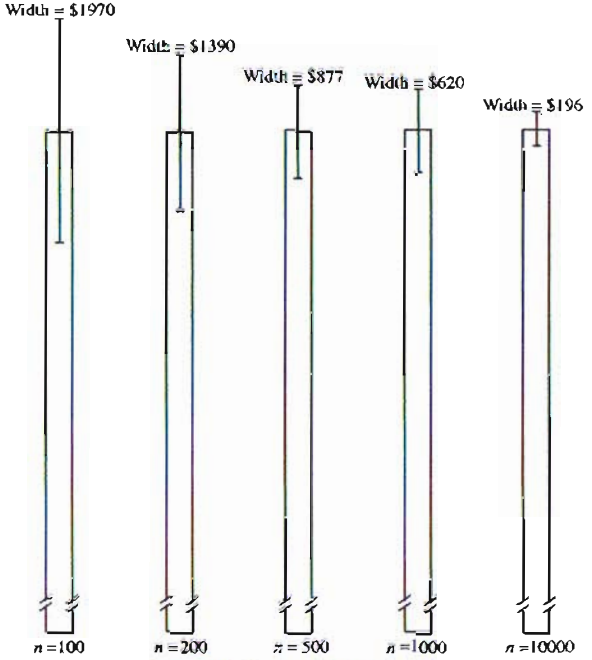


Figure 17.6 Interval estimates for five sample sizes (α = 0.05)

Changing the sample size

Apart from the alpha level, the other factor that will determine the width of the confidence interval is the sample size. If we stick with a confidence level of 95 percent, and only vary the sample size, the width gets smaller (we increase our accuracy) as sample size increases (Table 17.2, Figure 17.6).

Table 17.2 The effect of sample size on interval width ($\alpha = 0.05$)

Sample size	Interval width
100	$1970
200	$1390
500	$877
1000	$620
10,000	$196

One thing to notice about the effect of sample size is that enlarging the sample has its greatest effect on the interval width with small samples. Increasing the sample size from 100 to 200 reduces the interval width by $580, which is more than the $424 reduction in interval width when sample size is expanded from 1000 to 10,000. This is why many social surveys and public opinion polls, even when generalizing to a population of millions, will have sample sizes of only 1200–1400. Samples of this size narrow the confidence interval to a relatively small width, and to increase sample size any further would increase research costs without obtaining much greater accuracy.

Estimation using SPSS

To see how we can use SPSS to generate confidence intervals we will work through the example we introduced in the previous chapter for a sample of 20 children for each of which the amount of TV watched per night is recorded. This sample watches, on average, 165.85 minutes of TV nightly, with a standard deviation of 29.29 minutes. What can we estimate the population mean to be?

If we choose a confidence level of 95 percent (i.e. $\alpha = 0.05$), the appropriate t-score we use in the equations is that for $df = 19$. From the table for critical values of the t-distribution this is $t = 2.093$. The lower and upper limits will be:

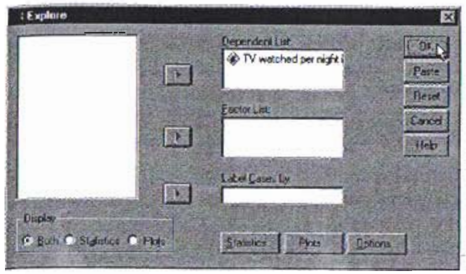
$$\text{lower limit} = \bar{X} - t\left(\frac{s}{\sqrt{n}}\right) = 165.85 - 2.093\left(\frac{29.19}{\sqrt{20}}\right) = 179.6 \text{ minutes}$$

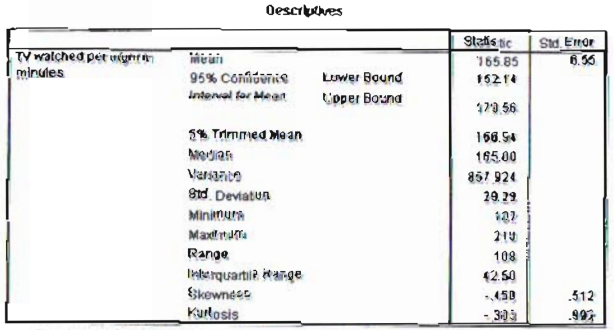
$$\text{upper limit} = \bar{X} + t\left(\frac{s}{\sqrt{n}}\right) = 165.85 + 2.093\left(\frac{29.19}{\sqrt{20}}\right) = 152.1 \text{ minutes}$$

Thus the estimated average amount of TV watched nightly, with a 95 percent confidence level, is 165.85 minutes [152.4, 179.6].

As is the case with many other statistics, SPSS provides a number of ways by which we can calculate this confidence interval for a mean. Three commands are particularly relevant:

1. **Analyze/Descriptive Statistics/Explore**. We introduced this command in Chapter 9 as a way of producing descriptive statistics. If we open the **Ch17.sav** file and enter **TV watched per night** into the **Dependent List:** we will generate a number of pieces of output, mainly presenting the descriptive statistics we discussed in Chapters 9–10. The relevant part of the output for our purposes here is the table headed **Descriptives** (Figure 17.7).





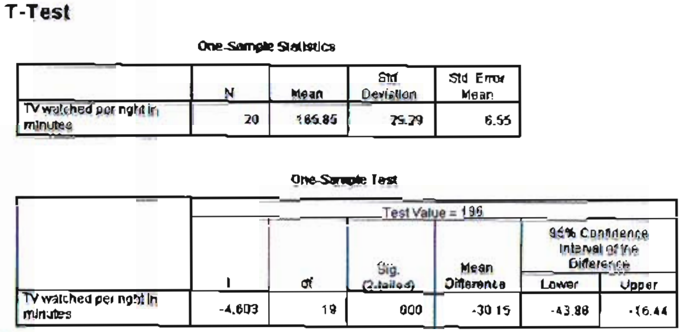
Descriptives

			Statistic	Std. Error
TV watched per night in minutes	Mean		165.85	6.55
	95% Confidence Interval for Mean	Lower Bound	152.14	
		Upper Bound	179.56	
	5% Trimmed Mean		166.54	
	Median		165.00	
	Variance		857.924	
	Std. Deviation		29.29	
	Minimum		107	
	Maximum		215	
	Range		108	
	Interquartile Range		42.50	
	Skewness		-.450	.512
	Kurtosis		-.303	.992

Figure 17.7 SPSS **Explore** dialog box and output

The first three rows of the table provide in turn the mean, the lower bound and upper bound of the 95 percent confidence interval, which is the default confidence level. We can see that the estimate is 165.85 minutes [152.14, 179.56], which conforms to our hand calculations. If we wanted a confidence interval based on a different confidence level, such as 90 percent or 99 percent, we click on the **Statistics** button in the **Explore** dialog box, and type over 95 with the desired level.

2. **Analyze/Compare Means/One-Sample T-Test**. As we shall see in later chapters, confidence intervals are also often generated by SPSS in the course of conducting hypothesis tests. In the previous chapter we noted this when we conducted the one-sample t-test for a mean on the data we have for TV viewing. The output we obtained from that command is presented in Figure 17.8.



T-Test

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
TV watched per night in minutes	20	165.85	29.29	6.55

One-Sample Test

	Test Value = 196					
					95% Confidence Interval of the Difference	
	t	df	Sig. (2-tailed)	Mean Difference	Lower	Upper
TV watched per night in minutes	-4.603	19	.000	-30.15	-43.88	-16.44

Figure 17.8 The SPSS One-Sample T Test output

The estimation information is provided in the last column of the **One-Sample Test** table. At a 95 percent confidence level the interval for the difference between the sample and the test value ranges from a lower limit of –43.86 to an upper limit of –16.44. In other words the difference between the average amount of TV watched by the population of all children and the hypothesized value we estimate to lie somewhere in this range, at a 95 percent confidence level. Since this range does not include the value of *zero*, which would indicate no difference, we can reject the hypothesis of no difference. In fact, if we subtract the values in the Lower box (–43.86) and the Upper box (–16.44) from the test value of 196, we obtain the confidence interval we calculated above by hand and also obtained from the **Analyze** command. Since this confidence interval *does not* include the test value of 196 we reject the hypothesis that the population mean equals 196 at this level of confidence.

3. **Interactive error bar graph.** If we select **Graphs/Interactive/Error Bar** from the SPSS menu the **Create Error Bar Chart** dialog box appears. The minimum information we must provide for this command to be executed is to place a variable for which confidence intervals (called 'error bars' by SPSS) will be constructed into the blank box on the vertical axis arrow. Here we drag **TV watched per night** into the box, since we want the confidence interval for the mean of this variable. We can also adjust the confidence level from the default 95% value by moving the slide-bar next to **Level:** or by simply typing the desired level in the box next to it. An additional option that is worth selecting is under the **Error Bars** tab; selecting **Mean** next to **Bar Labels** will give us the sample mean around which the confidence interval is constructed (Figure 17.9).

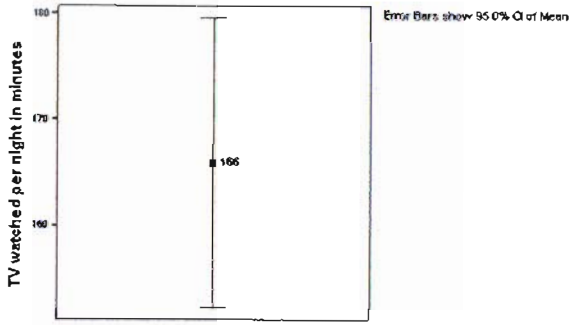


Figure 17.9 An SPSS Interactive Error Bar chart

The confidence interval ranges between the same upper and lower bounds that we calculated above. We should note here in anticipation of the discussion in the next chapter that if we wished to compare the means of more than one group, we place the variable that defines these groups into the blank box on the horizontal arrow in the **Create Error Bar Chart** dialog box.

We can also turn to web-based statistical calculation pages to obtain the confidence interval around a mean. One such page that will calculate a confidence interval around a sample mean is GraphPad's *QuickCalcs*, graphpad.com/quickcalcs/OneSampleT1.cfm. At this page we enter the mean, standard deviation and sample size, and then select 0 under **3. Specify the hypothetical mean value.**

Confidence intervals and hypothesis testing

We pointed out at the start of this chapter that the estimation techniques we have just discussed provide additional information that we do not obtain through the hypothesis testing procedure. In particular, it provides *at a given alpha level* the full range of values against which the sample result will not be significantly different (and by implication the full range against which it will be significantly different). However, in so far as it requires us to specify in advance a particular alpha level, the estimation procedure is more limited than the hypothesis testing procedure for making inferences. The hypothesis testing procedure gives us the exact *p*-score for a sample result, so that we can assess the full range of alpha levels at which a sample result will be significantly different from *a given test value for the population mean*. Thus we should use the information provided by both procedures to report our results.

The last point that is worth noting is that since estimation and hypothesis testing procedures share the same underlying *logic*, they also share the same limitations. In particular, they can each in their own way divert us from a discussion of whether a particular result is substantively significant to a more narrow discussion about whether it is statistically significant (see A.R. Feinstein, 1998, P-values and confidence intervals: Two sides of the same unsatisfactory coin, *Journal of Clinical Epidemiology*, vol. 51, no. 4, pp. 355–60). The substantive significance of any result is always the more important issue; statistical significance is only a small element of that broader discussion.

Exercises

17.1 What is meant by interval estimation?

17.2 Explain what is meant by a confidence level. How do changes in the confidence level affect the width of the interval estimate?

17.3 How does sample size affect the confidence interval?

17.4 How does the population standard deviation affect the width of a confidence interval?

17.5 For each of the examples in the text regarding the age of pre-school children and the amount of TV watched construct interval estimates for 90 percent and 99 percent confidence levels.

17.6 A survey is conducted to measure the length of time, in months, taken for university graduates to gain their first job. Assuming that this is a normally distributed variable, derive the interval estimates for the following sets of graduates, using a 95 percent confidence level:

Degree	Sample size	Mean	Standard deviation
Economics	45	6	2.5
Sociology	35	4	2.0
History	40	4.5	3.0
Statistics	60	3	1.5

17.7 To gauge the effect of wage bargaining agreements, union officials select a sample of 120 workers from randomly selected enterprises across an industry. The average wage rise in the previous year for these 120 workers was $1018, with a standard deviation of $614. Estimate the increase for all workers within this industry (use both 95 percent and 99 percent confidence levels).

17.8 A hospital checks the records of 340 randomly selected patients from the previous year. The average length of stay in the hospital for these patients was 4.3 days, with a standard deviation of 3.1 days.

(a) What would be the estimated average length of stay of all patients in the previous year (at a 99 percent confidence level)?
(b) How would this compare with the average length of stay for all patients in another hospital of 4 days?
(c) What could the hospital do to improve the accuracy of the estimate?

17.9 A study of 120 divorced couples that had been married in the same year found an average length of marriage of 8.5 years, with a standard deviation of 1.2 years. What is the estimate for the average length of marriage for all divorced couples, using a confidence level of 95 percent?

17.10 Open the **Employee data** file and calculate the (a) 90 percent, (b) 95 percent, and (c) 99 percent confidence intervals for employees' current salary.

18

The two samples *t*-test for the equality of means

The tests covered thus far deal with the one sample case. That is, they all involve making an inference about only one population mean: we don't have information about the population, so we infer it from the sample mean. This chapter will introduce hypothesis testing in the two samples case. In the two samples case we ideally want to compare two populations in terms of some descriptive statistic such as a mean. However, we do not know the value of these statistics for either population so we take a sample from each population and make inferences from each of these samples.

For example, in Chapter 16 we worked through an example where we were interested in the average amount of TV watched by Australian and British children between the ages of 5 and 12 years. We wanted to compare the population means, but unfortunately we only had the mean for the population of British children. We did not know the mean for all Australian children, so we took a sample of 20 and made an inference based on the data from this sample (Figure 18.1(a)). Thus in the one sample case we covered in Chapter 16, country of residence was not a variable, since all cases for which I collected data are from the same country (Australia).

What if we do not have information for the population of British kids either? The best we can do is take a random sample of British children as well, and make another inference from this second sample. In such a situation we conduct a **two samples test of significance**. In this instance we conduct a survey of children from each country. Although in practice we may think in terms of one sample, which is made up of both Australian and British children, conceptually we say that we are working with two samples: one from each of the populations we want to compare. That is, although in the actual mechanics of data collection we have one big collection of children who have been surveyed as part of the same research process, when analyzing the data we treat the two groups of children as separate samples (Figure 18.1(b)).

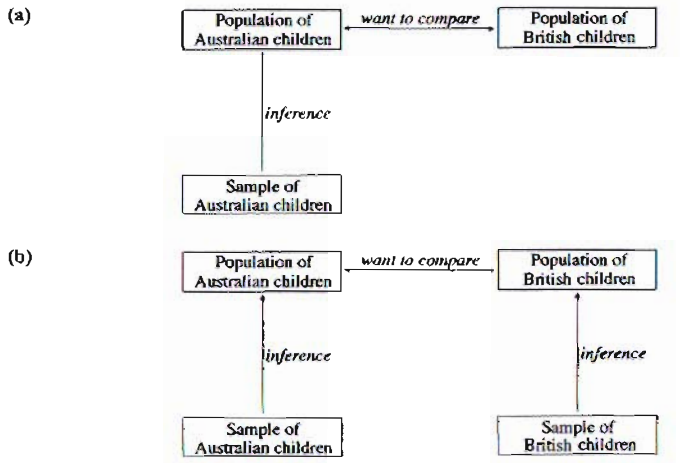

(a)
Population of Australian children
want to compare
Population of British children
inference
Sample of Australian children
(b)
Population of Australian children
want to compare
Population of British children
inference
inference
Sample of Australian children
Sample of British children

Figure 18.1 Hypothesis testing: (a) the one sample case, and (b) the two samples case

In fact, we could extend this to a situation in which we want to compare more than two populations. For example, we might be interested in comparing children from more than two countries in terms of their average amount of TV viewing and only have samples from each of these populations. Working with more than two samples requires a different test of significance that we will analyze in the following chapter. Generally, the choice of inference test is affected by the number of samples from which an inference is made. In particular, it is common practice to distinguish between one sample tests, two samples tests, and tests for more than two samples. When making inferences from more than two samples we speak of tests for *k*-samples, where *k* is a number greater than two. Often the change involved in moving from the one sample to the two samples situation, or to the *k*-sample situation, will not be great, but as an organizing principle it is useful to keep in mind whether the number of samples from which an inference is being made is one, two, or more than two.

Let us look again at the example of comparing Australian and British children in terms of their average amount of TV viewing. Now that children can differ not only in terms of TV viewing but also in terms of where they live, country of residence is a variable. We thus now have data on two variables: country of residence and amount of TV viewing. A child, in other words, can vary from another child in one of two ways; in terms of the country he or she lives in, and/or different in terms of the amount of TV he or she watches.

We use one of these variables to sort cases into distinct samples, based on the populations we want to compare. SPSS calls this a **grouping variable**.

The grouping variable defines the number of samples from which inferences will be made.

The samples are then compared on the basis of another variable, which SPSS calls a **test variable**. Thus, in our example, children are first grouped according to the variable 'Country of residence', since this defines the populations we are interested in, and the two samples thus formed (Australian and British children) are compared in terms of a test variable, 'Amount of TV watched each night'.

In other words, each case (i.e. each child) is assigned two values. The first 'tags' each case as belonging to a group defined by country of residence. The second value is the amount of TV each child watches, which is the variable on which the groups will be compared.

Dependent and independent variables

We can think of this two samples problem according to the notions of independent and dependent variables, which we introduced in Chapter 1. Usually the *grouping variable is the independent variable* and the *test variable is the dependent variable.*

A dependent variable is explained or affected by an independent variable.

In our example of children, we suspect that country of residence somehow affects or causes the amount of TV a child will watch (due possibly to factors such as the weather or the quality of programming in different countries). It is clear that in this situation we have a case of one-way causality that must run from place of residence to TV watching; it is unimaginable that children's TV viewing habits determine where they live! In other instances, however, the choice of appropriate model may be more contentious, as we discussed in Chapter 1 (it may help readers to return to that discussion before proceeding). These considerations involved in organizing data in the two samples case are summarized in Table 18.1 and Figure 18.2.

Table 18.1

Type of variable	SPSS name	Function in inference test
Independent	Grouping variable	Sorts cases into a number of samples to be compared
Dependent	Test variable	Calculated to describe and compare the samples

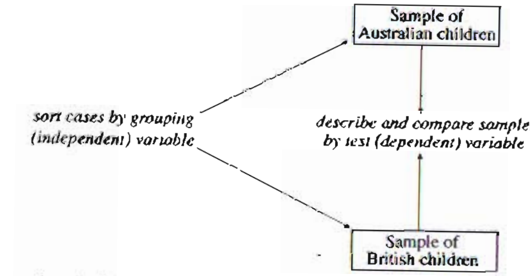

Sample of Australian children
sort cases by grouping (independent) variable
describe and compare samples by test (dependent) variable
Sample of British children

Figure 18.2 Two samples significance test

The sampling distribution of the difference between two means

As with all other hypothesis tests, we begin by *assuming* that the null hypothesis of no difference is correct. On this assumption we build up a sampling distribution of the difference between two sample means. We then use this sampling distribution to determine the probability of getting an observed difference between two sample means from populations with no difference.

For example, let's begin by assuming that the average amount of TV watched by children is the same in both Australia and Britain. This null hypothesis of no difference is formally written as:

$$H_0: \mu_1 = \mu_2$$

or

$$H_0: \mu_1 - \mu_2 = 0$$

If this assumption is true, what will we get if we take repeated samples from each country and calculate the difference in means for each pair of samples? Intuitively, we expect that the most common result will be that the difference is small, if not zero. Since we are assuming no difference between the two populations, we expect the sample means to be equal as well (the three-dot triangle is mathematical shorthand for 'therefore').

$$\bar{X}_1 = \bar{X}_2 \therefore \bar{X}_1 - \bar{X}_2 = 0$$

This is illustrated in Figure 18.3.

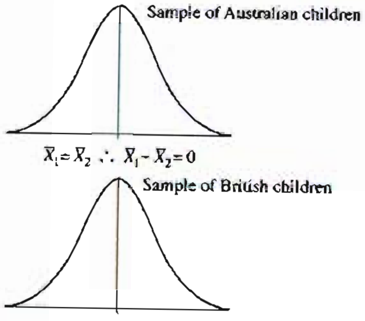

Sample of Australian children
$\bar{X}_1 = \bar{X}_2 \therefore \bar{X}_1 - \bar{X}_2 = 0$
Sample of British children

Figure 18.3 Two samples with means equal

But this will not always be the result. Occasionally we might draw a sample from Australia that has a lower than average amount of TV viewing coupled with a sample from Britain that has a higher than average amount of TV viewing (Figure 18.4).

$$\bar{X}_1 < \bar{X}_2 \;\therefore\; \bar{X}_1 - \bar{X}_2 < 0$$

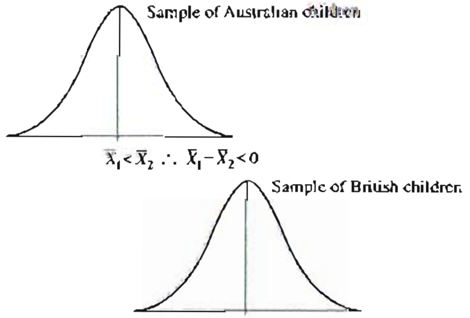


Figure 18.4 Two samples with means unequal

Similarly we might get, through the operation of sampling error, the opposite situation:

$$\bar{X}_1 > \bar{X}_2 \;\therefore\; \bar{X}_1 - \bar{X}_2 > 0$$

If we take a large number of these repeated random samples and calculate the difference between each pair of sample means, we will end up with a sampling distribution of the difference between two sample means that has the following properties:

- It will be a t-distribution:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}-\bar{X}}}$$

- The mean of the difference between sample means will be zero:

$$\mu_{\bar{X}-\bar{X}} = 0$$

- The spread of scores around this mean of zero (the standard error) will be defined by the formula:

$$\sigma_{\bar{X}-\bar{X}} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{n_1+n_2}{n_1 n_2}}$$

This is called the **pooled variance estimate**. This estimate assumes that the populations have equal variances. Sometimes this assumption cannot be sustained, in which case a **separate variance estimate** is used. As we shall see, SPSS will calculate t using each estimate, plus information that allows us to choose one or the other. But when doing hand calculations this pooled variance estimate is generally used since it is much easier to work with, and will usually lead to the same decision being reached as the separate variance estimate.

The two samples t-test for the equality of means

We can use these properties of the sampling distribution to conduct a t-test for the equality of means. Assume our survey consists of 20 Australian children and 20 British children, and the research wants to assess whether TV viewing time is affected by country of residence. (Although it is the situation in this example, the two samples t-test does not require the same number of cases in each sample.). We will work through this example using the five-step hypothesis testing procedure.

Step 1: State the null and alternative hypotheses

H_0: There is no difference in the mean amount of TV watched by children in Australia and Britain.

$$H_0\!: \mu_1 = \mu_2 \text{ or } H_0\!: \mu_1 - \mu_2 = 0$$

H_a: There is a difference in the mean amount of TV watched by children in Australia and Britain.

$$H_a\!: \mu_1 \neq \mu_2 \text{ or } H_a\!: \mu_1 - \mu_2 \neq 0$$

Step 2: Choose the test of significance

The following two factors are relevant in choosing the test of significance:

1. We are making an inference from two samples: a sample of Australian children and a sample of British children. Therefore we need to use a two samples test.

2. The two samples are being compared in terms of the average amount of time spent watching TV. This variable is measured at the interval/ratio level. Therefore the relevant descriptive statistic is the mean for each sample.

These factors lead us to choose the two samples t-test for the equality of means as the relevant test of significance.

Step 3: Describe the sample and derive the p-value

We have the following results (Table 18.2) that describe the data for each sample:

Table 18.2 Descriptive statistics for the samples

Descriptive statistic	Australian sample	British sample
Mean	166 minutes	187 minutes
Standard deviation	29 minutes	30 minutes
Sample size	20	20

The equation for calculating the sample t-score is:

$$t_{sample} = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}-\bar{X}}}$$

where:

$$\sigma_{\bar{X}-\bar{X}} = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{n_1+n_2}{n_1 n_2}}$$

If we substitute the sample data into these equations (where Australian children are sample 1, and British children are sample 2) we get a test statistics of $t = -2.3$:

$$\sigma_{\bar{X}-\bar{X}} = \sqrt{\frac{(20-1)29^2+(20-1)30^2}{20+20-2}}\sqrt{\frac{20+20}{20\times 20}} = 9.3$$

$$t_{sample} = \frac{166-187}{9.3} = -2.3$$

To obtain the p-value for this t-score, we need to consult the table for critical values for the t-distribution (Table 18.3). The number of degrees of freedom we refer to in this table is the sample size minus two (since we have to assume that the sample variances are equal to the unknown population variances, we have imposed two restrictions on the data):

$$df = n - 2 = 40 - 2 = 38$$

The table does not have a row of probabilities for 38 degrees of freedom. In such a situation, we refer to the row for the nearest reported number of degrees of freedom *below* the desired number, which in this instance is 35. With 38 degrees of freedom on a two-tail test, t_{sample} falls between the two stated t-scores of −2.030 and −2.438. The p-value, which falls between the significance levels for these t-scores, is therefore between 0.02 and 0.05.

Table 18.3 Critical values for t-distributions

	Level of significance for one-tail test				
	0.10	0.05	0.02	0.01	0.005
	Level of significance for two-tail test				
df	0.20	0.10	0.05	0.02	0.01
⋮	⋮	⋮	⋮	⋮	⋮
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
35	1.306	1.690	2.030	2.438	2.724
40	1.303	1.684	2.021	2.423	2.704
45	1.301	1.679	2.014	2.412	2.690
50	1.299	1.676	2.009	2.403	2.678
55	1.297	1.673	2.004	2.396	2.668
60	1.296	1.671	2.000	2.390	2.660
⋮	⋮	⋮	⋮	⋮	⋮
∞	1.282	1.645	1.960	2.326	2.576

We can also obtain the p-value from various sites on the internet that provide statistical calculation pages. Two such pages that will perform a t-test on a sample mean are:

1. *Statistical Applets*, www.assumption.edu/users/avadum/applets/applets.html and click on the **t test: Independent Groups** option on the left-menu;
2. GraphPad's *QuickCalcs*, graphpad.com/quickcalcs/ttest1.cfm

These pages not only provide the t-score, but also the exact p-value, unlike Table 18.3, which only provides a range between within which the p-value falls. From these pages I determined that the two-tail significance level is 0.030, which falls within the range of p-values we obtained from the table.

Step 4: Decide at what alpha level, if any, the result is statistically significant

On a two-tail test the p-value of 0.03 is statistically significant at the 0.05 level, but not at the 0.01 level.

Step 5: Report results

The mean number of minutes of TV watched by the sample of 20 British children is 187 minutes, which is 21 minutes higher than the sample of 20 Australian children, and this difference is statistically significant at the 0.05 level ($t = -2.3$, $p = 0.03$, two-tail). Based on these results we can reject the hypothesis that Australian children watch on average the same amount of TV each night as British children.

The two samples *t*-test using SPSS

SPSS calls this test the **Independent samples *t*-test**. The word 'independent' is very important because it raises both conceptual issues for hypothesis testing and practical issues for SPSS coding. We will define independent samples in the following chapter, when we can compare them with dependent samples, since their basic character is most evident when compared with dependent samples.

In SPSS the data for the children have been coded for the two variables. Each of these variables occupies a separate column, so that we have a column of numbers for the amount of TV watched and a column of numbers indicating the country in which each child lives. All independent-samples tests have data entered in the same way: *one column for the test variable and one column for the grouping variable*.

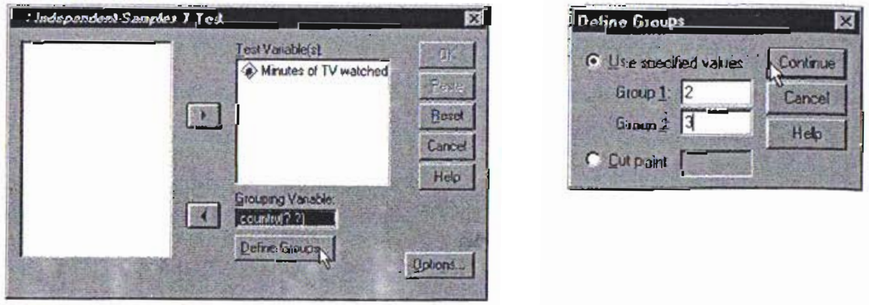
The data for this example also contain information for hypothetical samples of children from Canada and Singapore that will be used in the next chapter where we consider the *k*-independent samples situation. The value labels for each country are:

1 = Singapore
2 = Australia
3 = Britain
4 = Canada

Thus in this example we want to compare values 2 (Australia) and 3 (Britain) for **Country of residence** (Table 18.4 and Figure 18.5, which also presents the output).

Table 18.4 Independent-samples *t*-test using SPSS (file: Ch18.sav)

SPSS command/action	Comments
1 From the menu select **Analyze/Compare Means/ Independent-Samples T Test**	This brings up the Independent-Samples T Test dialog box
2 Click on **Minutes of TV watched** in the source list	This highlights **Minutes of TV watched**
3 Click on the ▸ that points to the **Test Variable(s):** list	This pastes **Minutes of TV watched** into the **Test Variable(s):** list
4 Click on **Country of residence** in the source list	This highlights **Country of residence**
5 Click on the ▸ that points to the **Grouping Variable:** list	This pastes **Country of residence** into the **Grouping Variable:** list. Notice that in this list the variable appears as **country(? ?)**
6 Click on **Define Groups**	This brings up the **Define Groups** box
7 In the area next to **Group 1:** type 2, and in the area next to **Group 2:** type 3	This identifies the two groups to be compared, which are **Australia** and **Britain**
8 Click on **Continue**	
9 Click on **OK**	



T-Test

Group Statistics

	Country of residence	N	Mean	Std. Deviation	Std. Error Mean
Minutes of TV watched per night	australia	20	165.85	29.29	6.55
	britain	20	186.75	29.56	6.61

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means						
									95% Confidence Interval of the Difference	
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	Lower	Upper
Minutes of TV watched per night	Equal variances assumed	.025	.876	-2.246	38	.031	-20.90	9.31	-39.74	-2.06
	Equal variances not assumed			-2.246	37.998	.031	-20.90	9.31	-39.74	-2.06

Figure 18.5 The Independent-Samples T Test dialog box, Define Groups dialog box, and output

The first table headed **Group Statistics** provides the descriptive statistics: the number of cases, the mean, and the standard deviation for each group.

The following table headed **Independent Samples Test** provides the inferential statistics. This table provides information for two different *t*-tests: one where the variances of the population are assumed to be equal and one where the population variances are not assumed to be equal. In calculating the *t*-score in the example above, we assumed that the variances of the two populations being compared were equal. In practical terms this means using the pooled variance estimate in the calculations. However, this may not always be a valid assumption. The validity of this assumption is tested in the columns headed Levene's Test for **Equality of** Variances. The value for F is the ratio of the two sample variances, and if this ratio is not equal to 1, it may reflect an underlying difference in the population variances. If the significance for this *F*-value (in the Sig. column) is less than 0.05 we conclude that the difference in variances observed in the samples reflects a difference in the variances of the populations from which the samples came. In such a situation we refer to the *t*-score in the first row of the table. We therefore use the following rule: read across the first row labelled Equal variances assumed, and

- if we find that the value under Sig. is greater than 0.05 we continue along that line to assess whether the means are significantly different; or
- if we find that the value under Sig. is less than 0.05 we refer to the *t*-test in the next row labelled Equal variances not assumed.

Usually the two estimates will agree with each other in terms of whether to reject or not reject the null (as is the case here), but in strict terms, we should use the relevant estimate, either that for equal or unequal variances. Here the first row is the relevant one. Moving

across the columns we see that the sample *t*-value is −2.246, which, with 38 degrees of freedom, has a two-tail significance of .031. These values all correspond to the values we generated by hand (with some slight differences due to rounding in the hand calculations). We also have a column headed Mean Difference. This is the difference between the two sample means, −20.9, which in the equation used to calculate *t*-scores is represented by $\bar{X}_1 - \bar{X}_2$.

You will also notice that SPSS has generated the lower and upper limits of 95 percent confidence interval for the difference in sample means, which are printed as −39.74 and −2.06 respectively. This allows us to conduct the same inference test, but using the estimation procedures developed in Chapter 17. These lower and upper limits indicate that at a 95 percent level of confidence, the difference between the population means lies somewhere between −39.74 minutes and −2.06 minutes. Since this interval does not include the value of 0, we reject the hypothesis that the population means are equal.

Example

A study is conducted to investigate whether foreign-owned companies on average have a lower rate of conformity to local health and safety codes when compared with locally owned companies. A survey of 50 foreign-owned and 50 domestic companies of similar size and in similar industries is conducted. Inspectors record the number of breaches of health and safety regulations they observe when inspecting these establishments.

Step 1: State the null and alternative hypotheses

H_0: There is no difference in the mean number of breaches between locally owned and foreign-owned firms.

$$H_0: \mu_1 = \mu_2 \text{ or } H_0: \mu_1 - \mu_2 = 0$$

H_a: Foreign-owned firms have a higher mean number of breaches than locally owned firms:

$$H_a: \mu_1 > \mu_2 \text{ or } H_a: \mu_1 - \mu_2 > 0$$

Step 2: Choose the test of significance

We are making an inference from two samples. The two samples are being compared in terms of the average number of breaches of the health and safety code, measured at the interval/ratio level. Therefore the relevant descriptive statistic is the mean of each sample. We therefore use the two samples *t*-test for the equality of means as the relevant test of significance.

Step 3: Describe the sample score and calculate the p-value

On average the 50 foreign firms are found to make 4.2 breaches per firm, with a standard deviation of 1.3. The 50 domestic firms are found to average 3.5 breaches per firm, with a standard deviation of 1.2.

In order to derive the test statistics, we need first to calculate the standard error (assuming equal variances), and from this the sample *t*-score:

$$\sigma_{\bar{X}-\bar{X}} = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}\sqrt{\frac{n_1+n_2}{n_1 n_2}}$$

$$= \sqrt{\frac{(50-1)1.3^2+(50-1)1.2^2}{50+50-2}}\sqrt{\frac{50+50}{50\times 50}} = 0.25$$

$$t_{sample} = \frac{\overline{X}_1 - \overline{X}_2}{\sigma_{\overline{X}-\overline{X}}} = \frac{4.2 - 3.5}{0.25} = 2.8$$

From the table for critical values of the t-distribution, we find that the t_{sample} has a two-tail significance of less than 0.001 and a one-tail significance of less than 0.005.

Step 4: Decide at what alpha level, if any, the result is statistically significant

Regardless of whether we use a one-tail or two-tail test, the difference is significant at the 0.01 alpha level.

Step 5. Report results

The results from a sample of 50 foreign-owned and 50 locally-owned firms suggest that the foreign-owned firms are more likely to breach domestic health and safety regulations. On average the 50 foreign firms are found to make 4.2 breaches per firm, with a standard deviation of 1.3, while the 50 domestic firms are found to average 3.5 breaches per firm, with a standard deviation of 1.2. The difference between the mean for local and the mean for foreign firms is statistically significant at the 0.01 level ($t = 2.8, p < 0.005$).

Exercises

18.1 What assumptions need to be made about the distribution of the populations before an independent-samples t-test is conducted?

18.2 For the following sets of results, test for a significant difference (assuming equal population variance):

		Mean	Standard deviation	Sample size
(a)	Sample 1	72	14.2	35
	Sample 2	76.1	11	50
(b)	Sample 1	2.4	0.9	100
	Sample 2	2.8	0.9	100
(c)	Sample 1	72	14.2	35
	Sample 2	76.1	11	50
(d)	Sample 1	450	80	120
	Sample 2	475	77	100

18.3 A researcher is interested in the effect that place of residency has on the age at which people begin to smoke cigarettes. The researcher divides a randomly selected group of people into 91 rural and 107 urban residents and finds that rural dwellers started smoking at an average age of 15.75 years, with a standard deviation of 2.3 years, whereas the urban dwellers began to smoke at a mean age of 14.63 years, with a standard deviation of 4.1 years. Is there a significant difference (using the pooled variance estimate)?

18.4 A water utility wishes to assess the effectiveness of an advertising campaign to reduce water consumption. Before the campaign the utility randomly selects 100 households throughout a region and records water usage for a morning shower as averaging 87 liters, with a standard deviation of 15 liters. It then randomly selects another 100 households after the campaign. These households average 74 liters per shower, with a standard deviation of 14 liters. Is there a significant difference? What conclusions can the utility make about the advertising campaign? What factors need to be considered when selecting the appropriate test?

18.5 A new form of organic pest control is developed for crop growing. Fifty plots of grain are sprayed with traditional pesticide, whereas 50 are sprayed with the new pest control. The output, in tonnes, of each set of plots, is recorded as follows:

	Old pesticide	Organic pesticide
Mean	1.4	2.20
Standard deviation	0.3	0.35

Conduct a t-test to assess the effectiveness of the new method.

18.6 A study is conducted to investigate the political awareness of children in public (state-funded) and private schools. Twenty-four students from a private school and 20 students from a nearby public school are randomly selected, and asked a series of questions relating to the political system. The mean score for private school students is 46 and for public school students the mean score is 64. Both samples have a standard deviation of 18.5. Conduct an independent-samples t-test for the equality of means to confirm your decisions as to whether the two school systems are significantly different.

18.7 Use the **Employee data** file to determine whether there is a significant difference between the mean current salaries for employees based on minority classification.

19

The *F*-test for the equality of more than two means: Analysis of variance

In Chapter 18 we considered the *t*-test for two independent samples, and tested the assumption that the samples came from populations with the same mean:

$$H_0: \mu_1 = \mu_2$$

We worked through an example where we had a sample of 20 children from Australia and 20 from Britain. Each child was asked how much TV, in minutes, they watched per night. We compared the samples in order to test the null hypothesis that there was no difference in the average amount of TV watched between children from the two countries, illustrated in Figure 19.1.

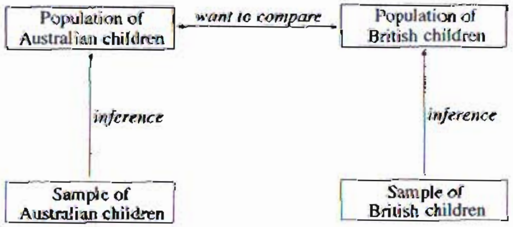

Population of Australian children
want to compare
Population of British children
inference
inference
Sample of Australian children
Sample of British children

Figure 19.1 Hypothesis testing: the two samples case

We call this a two samples problem because we are using two samples to make inferences about each population. However, sometimes the problem we are addressing is slightly wider. Instead of just comparing two countries, we might be interested in comparing the average amount of TV watched by children in several countries. For example, we may have samples of 20 children from Australia, Britain, Canada and Singapore, and want to see if the means for all these four populations are equal:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

This is called the problem of *k* independent samples, where *k* is any number greater than two. Here *k* is four, and this example is illustrated as in Figure 19.2. One way to deal with this problem is to test all the possible two samples combinations. With four samples the maximum number of combinations is six, illustrated in Figure 19.2 by the heavy arrows running from each population to the others:

Australia by Singapore
Australia by Canada
Australia by Britain
Singapore by Canada
Singapore by Britain
Canada by Britain

Thus we can undertake six separate *t*-tests and assess whether there are any significant differences. When we are working with more than two samples, however, we can test for the equality of means all at once using the analysis of variance *F*-test (ANOVA). The reason why a single ANOVA is preferable to multiple *t*-tests is that the risk of making a type I error for the *series* of *t*-tests will be greater than the stated alpha level for each *t*-test. Thus if the alpha level for each *individual t*-test is 0.05, the chance of making a type I error over *all* the *t*-tests that can be conducted for a given number of samples will be greater than 0.05. The ANOVA test, on the other hand, has a stated alpha level equal to the risk of making a type I error.

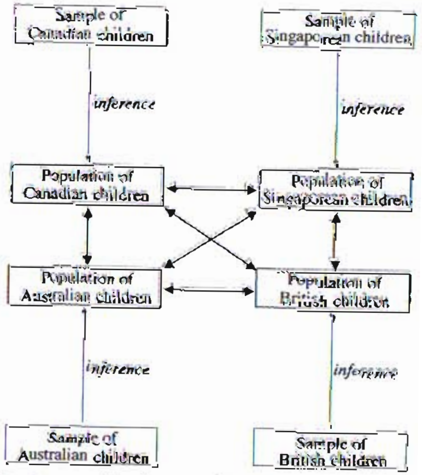

Sample of Canadian children
Sample of Singaporean children
inference
inference
Population of Canadian children
Population of Singaporean children
Population of Australian children
Population of British children
inference
inference
Sample of Australian children
Sample of British children

Figure 19.2 Hypothesis testing for more than two samples

The ANOVA procedure tests the null hypothesis that the samples come from populations whose means are equal. If the null hypothesis is true, samples drawn from such populations will have means roughly equal in value. In the example of children and TV time, the samples will all have roughly similar means, if the null is correct. Of course, we do not expect the sample means to be equal, even if the population means are the same, since random variation will affect the sampling process. The question is whether the size of the differences between the samples are consistent with the assumption of equality between the populations.

Consider the hypothetical sample results for our four groups of children in Table 19.1.

Table 19.1 Descriptive statistics for TV viewed per night, in minutes

TV watched per night	Country			
	Canada	Australia	Britain	Singapore
Mean	127	166	187	203
Standard deviation	27	29	30	26

We can see that there is a good deal of variation *between* the means of the four samples. In fact if we compare the highest with the lowest values, which are the means for Canada and Singapore, we can see a very large difference in average amounts of TV watched. Notice also the row for the standard deviation for each sample. We can see that *within* the sample for each country the results are clustered together, as indicated by the small standard deviations relative to the means. In other words, there are distinct differences from country to country, but similarity within each country. On the face of these statistics we might question the hypothesis that the populations from which these samples came have the same mean.

This logic is exactly the same as that used by ANOVA. It compares the amount of variation between the samples with the amount of variation within each sample – hence the name 'analysis of variance'. Thus, although we are interested in the difference between the means, ANOVA actually works with the variance, which is the square of the standard deviation.

Before working through an ANOVA for our hypothetical survey of children from four countries, we will illustrate the logic behind the test. Consider the two hypothetical sets of distributions in Figure 19.3.

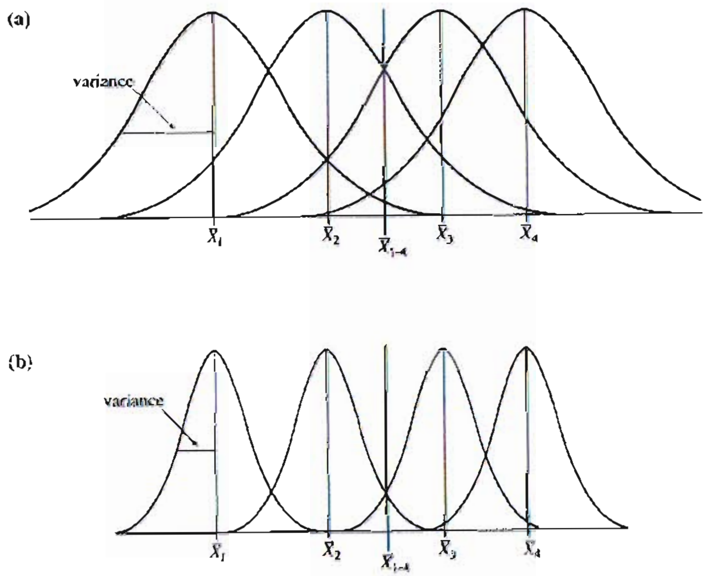


Figure 19.3 (a) Large variance within samples, and (b) Small variance within samples

Four samples are randomly selected and the mean for each is calculated, together with the overall mean when the cases for all four samples are pooled together ($\bar{X}_{1-4}$). In both (a) and (b) we can see that the means are not equal: there is some variance between the sample means. We can also see that while the sample means are the same in the two sets of distributions, there is also an important and obvious difference. In (a) the spread of cases *within* each sample around the sample mean is quite wide, whereas in (b) the variance within each sample is relatively small. Each sample in (b) seems distinct from the others, whereas in (a) there is considerable overlap in the distributions, so that the samples seem to blend into each other. We would be more inclined to consider the second set of samples (b) to come from populations that are different from each other, whereas the first set (a) can be more easily explained as coming from identical populations, with random variation causing the *samples* to differ slightly from each other.

We can capture this difference by calculating two numbers and expressing one as a ratio of the other. The first number is the amount of variance *between* the sample means and the grand total mean. Consider the two sets of sample means shown in Figure 19.4. We can see in Figure 19.4 that in (a) the variance of the sample means around the overall mean (when the samples are pooled together) is small relative to the second situation. Thus the samples in (a) are less likely to form distinct clusters of cases that reflect underlying differences between the populations.

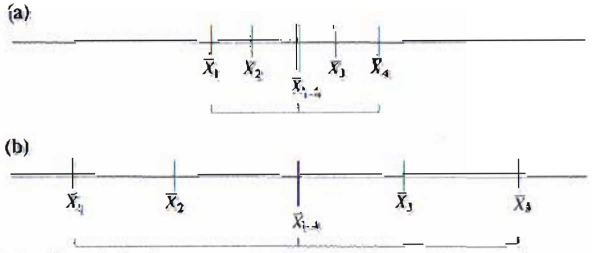


Figure 19.4 (a) Small variation between sample means and overall mean, and (b) Large variation between sample means and overall mean

We cannot jump, however, to this conclusion about the populations just on the basis of the variance between sample means. As we saw in Figure 19.3, the variances within each sample in (a) might be very small, so that each sample forms a distinct 'spike' around each sample mean. The variances within each sample in (b), on the other hand, might be very wide so that the samples still blur into each other, despite the differences between the means. To capture these aspects of the distributions, we need to calculate a second number, which measures this variance within each sample around each sample mean. The extent to which samples will form these distinct spikes around their respective means will be expressed by the ratio of the **variance between samples to the variance within samples**.

The one-way analysis of variance *F*-test

We can now use these general concepts to determine whether there is a significant difference between children in different countries in terms of the average amount of TV they watch. To calculate the relevant test statistic we need to formalize some of these basic concepts. The first is the total amount of variation for the scores of all 80 cases sampled. This is measured by a concept called the **total sum of squares (TSS)**. This is calculated using the formula:

$$TSS = \Sigma(X_i - \bar{X})^2$$

The value for the TSS can be divided into two components. The first is the amount of variation within each sample, called the **sum of squares within (SSW)**. The second is the amount of variation between each sample, called the **sum of squares between (SSB)**:

$$TSS = SSB + SSW$$

Each of these components of the TSS can be calculated in the following way, where $\bar{X}_s$ is the mean for a given sample and n_s is the number of cases in a given sample:

$$SSW = \Sigma(X_i - \bar{X}_s)^2$$

$$SSB = \Sigma n_s(\bar{X}_s - \bar{X})^2$$

These formulas should remind the reader of the formula for the standard deviation, since they embody the same principle that variance relates to the amount of difference between individual scores and the mean. As with the formula for the standard deviation, these definitional formulas can be difficult to work with. In particular, to calculate the TSS, it is easier to work with the formula:

$$TSS = \Sigma X_i^2 - n\bar{X}^2$$

Once we have TSS, we only need to calculate either SSW or SSB, and then use the formula $TSS = SSB + SSW$ to calculate the other. In other words, if we calculate TSS and SSB, we substitute these into the following equation to arrive at SSW:

$$SSW = TSS - SSB$$

To see how this is done we will work through our example with the four samples of 20 children. These calculations are best done by constructing a listed data table (Table 19.2). The score for each case is listed, with the samples placed in separate columns.

Table 19.2 Calculations for ANOVA

Canada		Australia		Britain		Singapore	
X_i	X_i^2	X_i	X_i^2	X_i	X_i^2	X_i	X_i^2
89	7921	102	10,404	124	15,376	156	24,336
92	8464	120	14,400	135	18,225	165	27,225
95	9025	132	17,424	156	24,336	174	30,276
105	11,025	134	17,956	165	27,225	179	32,041
106	11,236	145	21,025	167	27,889	180	32,400
108	11,664	149	22,201	172	29,584	184	33,856
110	12,100	156	24,336	178	31,684	189	35,721
113	12,769	162	26,244	182	33,124	189	35,721
116	13,456	165	27,225	184	33,856	196	38,416
125	15,625	165	27,225	185	34,225	203	41,209
128	16,384	165	27,225	186	34,596	204	41,616
135	18,225	174	30,276	187	34,969	207	42,849
138	19,044	179	32,041	189	35,721	210	44,100
139	19,321	180	32,400	198	39,204	218	47,524
140	19,600	187	34,969	209	43,681	221	48,841
146	21,316	189	35,721	212	44,944	228	51,984
146	21,316	196	38,416	218	47,524	231	53,361
154	23,716	201	40,401	223	49,729	238	56,644
167	27,889	206	42,436	225	50,625	241	58,081
194	37,636	210	44,100	240	57,600	250	62,500
$\Sigma X_i=2546$		$\Sigma X_i=3317$		$\Sigma X_i=3735$		$\Sigma X_i=4063$	
	$\Sigma X_i^2=337,732$		$\Sigma X_i^2=566,425$		$\Sigma X_i^2=714,117$		$\Sigma X_i^2=838,701$

From this information we can calculate the mean for each sample, and the mean for all the samples combined:

$$\overline{X}_{canada} = \frac{2546}{20} = 127.3 \text{ minutes}$$

$$\overline{X}_{australia} = \frac{3317}{20} = 165.85 \text{ minutes}$$

$$\overline{X}_{britain} = \frac{3735}{20} = 186.75 \text{ minutes}$$

$$\overline{X}_{singapore} = \frac{4063}{20} = 203.15 \text{ minutes}$$

$$\overline{X} = \frac{(2546+3317+3735+4063)}{80} = 170.8 \text{ minutes}$$

Using this information we can calculate the TSS, SSB, and SSW:

$$TSS = \Sigma X_i^2 - n\overline{X}^2 = (337,732+566,425+714,117+838,701) - 80(170.8)^2$$
$$= 124,189$$

$$SSB = \Sigma n_i(\overline{X}_s - \overline{X})^2$$
$$= 20(127.3-170.8)^2 + 20(165.85-170.8)^2 + 20(186.75-170.8)^2 + 20(203.15-170.8)^2$$
$$= 64,353$$

$$SSW = TSS - SSB = 124,189 - 64,353 = 59,836$$

The actual test statistic we use to determine the statistical significance of the sample result is the *F*-ratio. We have actually encountered this test statistic before when analyzing SPSS output for a two samples *t*-test. Just as in that case, the *F*-ratio tests for a difference between variances. The *F*-ratio is a ratio of the two variances, the SSB and SSW, each corrected for the appropriate degrees of freedom, where *k* is the number of samples:

$$F_{sample} = \frac{\frac{SSB}{k-1}}{\frac{SSW}{n-k}}$$

Substituting the relevant numbers into this equation we get:

$$F_{sample} = \frac{\frac{SSB}{k-1}}{\frac{SSW}{n-k}} = \frac{\frac{64,353}{4-1}}{\frac{59,836}{80-4}} = 27.25$$

As with the other test statistics we have come across, namely *z*-scores and *t*-scores, we need to obtain the *p*-score for this test statistic in order to decide whether to reject or not reject the null hypothesis. To do this we refer to the table for the distribution of *F* (Table A3), taking into account the following three factors:

1. *The degrees of freedom for the estimate of the variance between samples.* This is the number of samples minus one, and appears in the numerator of the *F*-ratio. The formula with the values for our example is:

$$dfb = k - 1 = 4 - 1 = 3$$

2. *The degrees of freedom for the estimate of the variance within the samples.* This is the total number of cases minus the number of samples, and appears in the denominator of the *F*-ratio:

$$dfw = n - k = 80 - 4 = 76$$

Notice that Table A3 does not have a line for the 'degrees of freedom within' equal to 76. In fact, whole ranges of values are skipped after the first 30. This is because the critical scores do not decrease very much for incremental increases in the degrees of freedom after 30. Where we have degrees of freedom that do not appear in the table, we refer to the closest value that appears in the table *below* the desired number. Here the closest value below 76 that appears in the table is 60.

3. *The alpha level.* Unlike the tables for *z* and *t*-distributions, the table for *F*-scores (Table A3) is produced for a *given* alpha level of 0.05. Thus this table does not allow us to determine whether a sample result is significant at any other alpha level. That is, we would need a different table if the alpha level were not equal to 0.05.

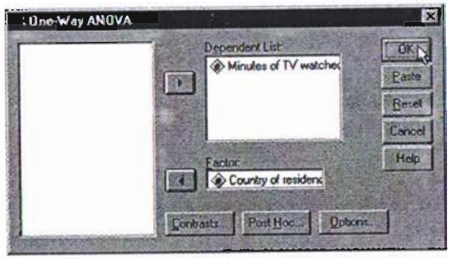
Since the table for *F*-scores is produced for a *given* alpha level of 0.05, we will instead turn directly to SPSS as a means of obtaining the *p*-value for the sample results, since SPSS will give us the exact significance level of the *F*-score.

ANOVA using SPSS

The data from the previous example have been entered into SPSS. The data file has two columns, one for the variable indicating how much TV each child watches, and another indicating their country of residence. To conduct an ANOVA we work through the procedures in Table 19.3 and Figure 19.5, which also presents the results of this set of commands.

Table 19.3 One-Way ANOVA on SPSS (file: Ch19.sav)

SPSS command/action	Comments
1 From the menu select **Analyze/Compare Means/ One-Way ANOVA**	This brings up the **One-Way ANOVA** dialog box
2 Click on **Minutes of TV watched** in the source variable list	This highlights **Minutes of TV watched**
3 Click on ▸ pointing to the box below **Dependent List:**	This pastes **Minutes of TV watched** in the dependent variable target list, which is the variable used to compare the samples
4 Click on **Country of residence** in the source list	This highlights **Country of residence**
5 Click on ▸ pointing to the box below **Factor:**	This pastes **Country of residence** in the **Factor** variable target list, which will form the samples to be compared
6 Click on OK	



Oneway

ANOVA

Minutes of TV watched per night

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	64353.438	3	21451.146	27.246	.000
Within Groups	59835.050	76	787.303		
Total	124188.488	79			

Figure 19.5 The SPSS ANOVA dialog box and ANOVA output

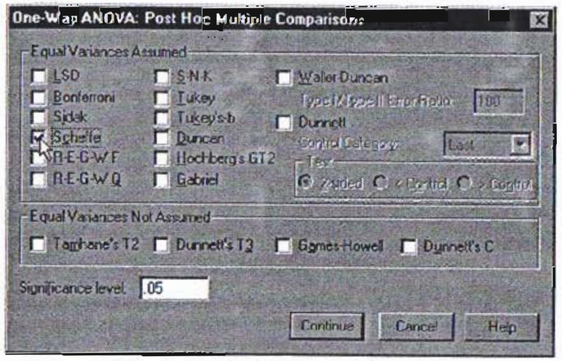
Looking at the SPSS output we can see the results we calculated by hand. The sum of squares between, the sum of squares within, and the total sum of squares are in the first column of the **ANOVA** table, together with the relevant degrees of freedom in the third column. From these, the *F*-ratio is 27.246, which is the same as that calculated above (allowing for rounding). The probability is printed as .000. This *does not* mean that the probability of obtaining an *F*-ratio of 27.246 is zero. SPSS rounds off the probability to 3 decimal places, so that this result is read as 'less than 5 in 10,000'.

We must stop at this point and be clear about what this *F*-test ANOVA has determined. The null hypothesis is that the samples come from populations with the same mean:

$$H_0: \mu_1 = \mu_2 = \mu_3 = \ldots = \mu_k$$

We have found that the *p*-score is so low that we reject the null hypothesis of no difference; by rejecting the null hypothesis, we have decided that *at least one of these populations has a mean that is not equal to the others*. Notice the particular wording of the conclusion: *at least* one population differs from the rest. The *F*-test itself does not tell us which of the populations, and how many, differ. Which of the possible pairwise differences between samples are significant cannot be answered by the *F*-test.

To determine which samples are significantly different, after having performed an *F*-test and rejected the null, we turn to a set of techniques called *post hoc* comparisons. Thus when conducting an *F*-test we normally ask for some follow-up information to be provided, so that if we do discover a statistically significant difference, we can determine which of the populations differ(s) from the others. These are called *post hoc* comparisons. In SPSS *post hoc* comparisons are available as an option in the **One-Way ANOVA** dialog box by clicking on the **Post Hoc** button. This will bring up the **One-Way ANOVA Post Hoc Multiple Comparisons** dialog box (Figure 19.6), which provides us with a range of options for comparing the samples so that we can determine exactly which ones come from populations different from the others.



☑ Scheffe

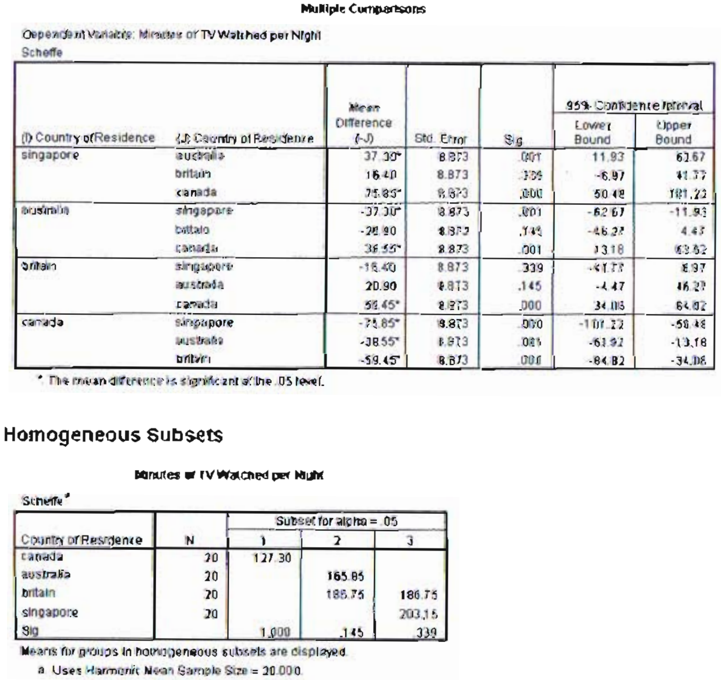
Significance level: .05

Figure 19.6 The Post Hoc Multiple Comparisons dialog box

Unfortunately, there are many *post hoc* comparisons to choose from, each subtly different from the others. We will not explore these subtle differences between the choices; in most situations they will all lead to the same conclusions. The main considerations involved in choosing among the options are:

- whether we can assume equal variances among the populations to be compared;
- whether the samples have equal or roughly equal variances;
- the extent to which we want to minimize type 1 errors.

The SPSS Help function provides a reasonably simple explanation of the *post hoc* comparisons (right-click on each item to bring up the contextual help). When in doubt, the most conservative test should be used; namely, the one that is the least likely to find a significant difference and this usually is the Scheffe *post hoc* comparison, which is selected by clicking on the box next to it. The other advantage of the **Scheffe** test is that it also examines sub-groups formed by various combinations of the samples, rather than just pairwise comparisons. If we select the Scheffe test we will produce the output in Figure 19.7 along with the ANOVA output in Figure 19.5.



Multiple Comparisons

Dependent Variable: Minutes of TV Watched per Night
Scheffe

(I) Country of Residence	(J) Country of Residence	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
singapore	australia	37.30*	8.873	.001	11.93	62.67
	britain	16.40	8.873	.339	-8.97	41.77
	canada	75.85*	8.873	.000	50.48	101.22
australia	singapore	-37.30*	8.873	.001	-62.67	-11.93
	britain	-20.90	8.873	.145	-46.27	4.47
	canada	38.55*	8.873	.001	13.18	63.92
britain	singapore	-16.40	8.873	.339	-41.77	8.97
	australia	20.90	8.873	.145	-4.47	46.27
	canada	59.45*	8.873	.000	34.08	84.82
canada	singapore	-75.85*	8.873	.000	-101.22	-50.48
	australia	-38.55*	8.873	.001	-63.92	-13.18
	britain	-59.45*	8.873	.000	-84.82	-34.08

*. The mean difference is significant at the .05 level.

Homogeneous Subsets

Minutes of TV Watched per Night

Scheffe[a]

Country of Residence	N	Subset for alpha = .05		
		1	2	3
canada	20	127.30		
australia	20		165.85	
britain	20		186.75	186.75
singapore	20			203.15
Sig.		1.000	.145	.339

Means for groups in homogeneous subsets are displayed.
a. Uses Harmonic Mean Sample Size = 20.000.

Figure 19.7 SPSS Post Hoc Multiple Comparisons output

This table provides a comparison of means for each country of residence against each other country of residence. The first rows compare Singapore with each of Australia, Britain, and Canada. The second set of rows compare the mean amount of TV watched by the Australian sample with each of the other three countries, and so on. Notice that this results in the same comparison being repeated. For example, in the first set of rows we see that the difference between the means when comparing Singapore with Australia is 37.3 minutes, and in the second set of rows when comparing Australia with Singapore the mean difference is –37.3 minutes, since this is effectively the same comparison looked at the other way.

The important aspect to this table is Sig. column that provides the exact significance for the difference between any two means. Where this is less than 0.05 SPSS places an * next to the value in the Mean Difference column, indicating a significant difference between the means of the two samples being compared, at the SPSS default significance level of 0.05. Collecting these * together we can see that a significant difference exists between the means for each of the following pairwise comparisons:

Singapore by Australia
Singapore by Canada
Australia by Canada
Britain by Canada

In other words, for each of these pairwise combinations, we can reject the hypothesis that the mean amounts of TV watched per night by children are the same (at the set alpha level).

A similar conclusion can be reached using the Lower Bound and Upper Bound values, presented under the 95% Confidence Interval column in the SPSS output. We can see that where the interval defined by these values does not take in the value of 0 (indicating no difference), an asterisk is next to the mean difference.

Example

Three children are compared in terms of their reading abilities. Each child is asked to complete 12 reading tasks, and the number of mistakes made during each reading task is recorded (Table 19.4). Can we say that these children differ in their readings abilities?

Table 19.4 Number of mistakes per child

Task number	Alexandra	Katherine	Evelyn
1	8	15	12
2	6	9	6
3	14	20	8
4	9	15	9
5	14	6	10
6	8	9	14
7	12	17	16
8	19	12	5
9	6	6	18
10	11	13	21
11	8	13	15
12	15	5	11

Step 1: State the null and alternative hypotheses

H_0: The mean number of mistakes made by each child are equal.

$$H_0: \mu_1 = \mu_2 = \mu_3$$

H_a: The mean number of mistakes made by each child are not all equal.

$$H_a: \mu_1 \neq \mu_2 \neq \mu_3$$

Step 2: Choose the test of significance

The research question is interested in the mean number of mistakes to see if they are equal. We also have three samples, so we are comparing means across more than two samples. The appropriate test is therefore the ANOVA *F*-test for the equality of means.

Step 3: Describe the sample and calculate the p-score

In conducting an ANOVA it is helpful to set up a listed data table with the relevant calculations (Table 19.5). From this information we calculate the mean for each sample, and the mean for all the samples combined.

$$\overline{X}_{alexandra} = \frac{130}{12} = 10.8$$

$$\overline{X}_{katherine} = \frac{140}{12} = 11.7$$

$$\overline{X}_{evelyn} = \frac{145}{12} = 12.1$$

Table 19.5 Calculations for ANOVA

Alexandra		Katherine		Evelyn	
X_i	X_i^2	X_i	X_i^2	X_i	X_i^2
8	64	15	225	12	144
6	36	9	81	6	36
14	196	20	400	8	64
9	81	15	225	9	81
14	196	6	36	10	100
8	64	9	81	14	196
12	144	17	289	16	256
19	361	12	144	5	25
6	36	6	36	18	324
11	121	13	169	21	441
8	64	13	169	15	225
15	225	5	25	11	121
$\Sigma X_i = 130$	$\Sigma X_i^2 = 1588$	$\Sigma X_i = 140$	$\Sigma X_i^2 = 1880$	$\Sigma X_i = 145$	$\Sigma X_i^2 = 2013$

$$\overline{X} = \frac{(130+140+145)}{36} = 11.5$$

These are the descriptive statistics for the sample data. Clearly there is a difference between the samples in terms of the average number of mistakes made. Could this be due to random variation when sampling from populations with no difference?

To determine this we first calculate the TSS and SSB:

$$TSS = \Sigma X_i^2 - n\overline{X}^2 = (1588+1880+2013) - 36(11.5)^2 = 720$$

$$SSB = \Sigma n_s(\overline{X}_s - \overline{X})^2 = 12(10.8 - 11.5)^2 + 12(11.7-11.5)^2 + 12(12.1-11.5)^2 = 10.7$$

$$SSW = TSS - SSB = 720 - 10.7 = 709.3$$

From this we can finally calculate the sample F-statistic that we use in the test of significance:

$$F_{sample} = \frac{\frac{SSB}{k-1}}{\frac{SSW}{n-k}} = \frac{\frac{10.7}{3-1}}{\frac{709.3}{36-3}} = 0.25$$

This F-score has a p-value greater than 0.05; it is smaller than the critical value for F of 3.32, printed in the table for critical values for the F-distribution for an alpha level of 0.05, at these degrees of freedom. We have noted that the table for critical values for the F-distribution only allows us to determine whether a sample result is or is not significant at the 0.05 level. To obtain the exact p-score for the sample F-statistic we can either put the data into SPSS or into a web-based statistical calculation page (a full list of these pages is available at members.aol.com/johnp71/javastat.html#Comparisons). For example, if I enter the data from Table 19.5 into the web-page at:

- www.physics.csbsju.edu/stats/anova_NGROUP_NMAX_form.html

I obtain the following result (some of the specific calculations returned by this page will differ slightly from my hand calculations due to rounding errors):

"The probability of this result, assuming the null hypothesis, is 0.793"

Step 4: Decide at what alpha level, if any, the result is significant

A sample p-score of 0.793 is clearly not significant at any alpha level, indicating that the null hypothesis should not be rejected. Despite the differences in the sample means we cannot say that these reflect differences between any of the underlying populations. The differences, in other words, we attribute to sampling error.

Step 5: Report results

The reading abilities of three children were assessed by comparing the number of mistakes each made on a standard test. The mean number of mistakes made by each child is respectively 10.8, 11.7, and 12.1. However, these differences are not statistically significant ($F = 0.25$, $p = 0.793$).

Summary

We have taken the inference for a mean from the one sample case, to analyzing two independent samples, through to the analysis of more than two samples. In the following chapter we will complete the discussion of making inferences for means by detailing the two-dependent samples case. However, these chapters do not exhaust all the possible forms of analysis for means. A whole class of procedures called General Linear Models exist to handle more complex situations, available under the SPSS **Analyze/General Linear Model** command. Three general classes of GLM are available:

1. **Univariate.** This allows us to analyze the effect that several independent variables have had on a single dependent variable. For example, I might compare two groups in terms of their rested heart rates, and want to see the role that sex, age, and past exercise levels have had on heart rate. I could conduct separate t-test or ANOVAs to assess whether there is a significant difference between men and women, a significant difference between age groups, and between categories of exercise level. The GLM Univariate command allows these comparisons to be made at once, and to analyze interactions between these variables.
2. **Multivariate.** This allows us to analyze differences between groups (defined by one or more variables) across a number of dependent variables. For example, I might want to measure the impact of sex, age, and past exercise levels on heart rate, walking speed, and blood pressure. Here I have three dependent variables, whose distributions are analyzed jointly when assessing the impact of the independent variables.
3. **Repeated Measures.** This is particularly useful in medical/health science research where two groups (control and experimental) are each compared before and after some intervention. Thus the samples are dependent and we want to assess both the before-and-after change (within subjects) and the difference across the groups (between groups).

Exercises

19.1 A comparison is made between five welfare agencies in terms of the average number of cases handled by staff during a month. The research is aimed at finding whether the workload is significantly different between agencies.

(a) Explain why an ANOVA should be used to explore this issue.

(b) State the null hypothesis for this research in words and using mathematical notation.

(c) From the following hypothetical results calculate the F-ratio and make a decision about the null ($a = 0.05$).

Variation	Sum of Squares	Degrees of Freedom
Between Agencies	50	4
Within Agencies	7210	110
Total	7260	114

19.2 A university instructor uses different teaching methods on three separate classes. The instructor wants to assess the relative effectiveness of these methods by testing for a significant difference between the classes. The data on final grades are presented in the following table:

Method A	Method B	Method C
21	28	19
19	28	17
21	23	20
24	27	23
25	31	20
20	38	17
27	34	20
19	32	21
23	29	22
25	28	21
26	30	23

(a) Calculate the mean and standard deviation for each sample. Can you anticipate from these descriptive statistics the result of an ANOVA conducted on these data?

(b) Conduct the ANOVA to assess your expectations.

19.3 The prices ($) of an item are collected from the stores of three separate retail chains:

Chain A	Chain B	Chain C
3.30	3.20	2.99
3.30	3.35	3.00
3.45	3.15	3.30
3.35	3.10	3.45
3.20	2.99	3.40
3.25	3.30	3.25
3.30	3.15	3.25

(a) Can we say that these chains do not price this good differently?

(b) Enter these data in SPSS and confirm your results. Note that you will need two columns: one column to indicate the sample each case fails into, and one column indicating each case's measurement for the dependent variable.

(c) If you find a significant difference, use the Scheffe *post hoc* comparison to determine which group(s) are different.

19.4 The following data were obtained from a hypothetical study of the effects of blood alcohol levels on driving performance. Subjects were randomly assigned into four groups, with each group being assigned a different blood alcohol level. Each group was then measured by time in seconds spent on target when steering a car in a simulated environment.

(a) Using an ANOVA *F*-test, determine whether driving ability is significantly reduced with higher blood alcohol levels.

(b) Enter these data in SPSS and confirm your results. Note that you will need two columns: one column to indicate the sample each case falls into, and one column indicating each case's measurement for the dependent variable.

(c) If you find a significant difference, use the Scheffe *post hoc* comparison to determine which group(s) are different.

Level 1	Level 2	Level 3	Level 4
216	178	180	166
187	144	132	145
166	176	172	148
242	132	137	136
229	188	154	126
276	168	154	176
233	204	176	133
166	187	178	184
208	165	169	155
224	193	188	177
213	201	175	189
254	197	186	165
227	183	179	172
203	176	168	172
206	196	188	179
221	182	176	166
219	202	185	180
220	190	195	176
196	202	177	165
230	188	186	193

19.5 The following data are from a hypothetical sample of 20 children from the USA, representing the number of minutes of TV watched per night:

195	184	165	162	168	196	217	190	212	232
204	205	217	210	230	197	180	192	190	198

(a) How will the addition of this sample to the ANOVA of Australian, British, Canadian, and Singaporean children affect the number of degrees of freedom?

(b) Enter these data into the file with the data for the Australian, British, Canadian, and Singaporean children and recalculate the ANOVA and *post hoc* analysis on SPSS.

(c) What do you conclude about the amount of TV watched between children from different countries?

19.6 Enter into SPSS the data from the example in the text above regarding the reading ability of three children and conduct a comparison of the means to see if there is a significant difference.

19.7 Using the **Employee data** file determine whether there is a significant difference across employment categories in terms of current salary.

20

The two dependent samples *t*-test for the mean difference

Dependent and independent samples

In the previous 2 chapters we looked at inference tests for the mean of two or more **independent samples**.

Independent samples are those where the criteria for selecting the cases that make up one sample do not affect the criteria for selecting cases that make up the other sample(s).

For example, to compare Australian and British children in terms of average amounts of TV watched, we selected any random sample of Australian children and any random sample of British children. However, there are research questions that require us to choose samples that are not independent. We sometimes want to link our samples so that if a certain case is included in one sample this necessitates a specific case being included in the other. Samples that are linked in this way are called **dependent** samples.

Dependent samples are those where the criteria for selecting the cases to make up one sample affect the criteria for selecting cases to make up the other sample(s).

There are generally two situations in which such dependence is required:

1. *When the same subject is observed under two different conditions.* This is often used in a before-and-after experiment (sometimes called a pre-test–post-test design). For example, a new drug may be tested to see its impact on blood pressure. The blood pressure of a group of subjects is taken and then these *same* participants take the drug and their blood pressure is again measured. Obviously, to isolate the effect of the drug, a person who is included in the 'before treatment' sample is also included in the 'after treatment' sample. The measurement for each person in the 'before' sample is then matched with their respective measurement after receiving the new drug to see if it has improved their condition.
2. *When subjects in different samples are linked for some special reason.* An example may be where we want to compare the amount of TV watched by a parent with the amount of TV watched by his or her particular child. If we choose a certain set of parents, we cannot choose any set of children with which to compare them: the sample needs to be comprised of the children of the people making up the parent sample. This is sometimes called a **matched-pairs** technique.

It is clear that in either situation the make-up of one sample determines the make-up of the other sample. The advantage of a dependent samples method is that it controls in a loose fashion for other variables that might affect the dependent variable. For example, consider further the issue of whether parents and children differ in the amount of TV they watch. If we take a random sample of parents and a random sample of *any* children and compare the means for each sample, we might find that there is a statistically significant difference. But this might not be due to family status. There might be another variable, such as socioeconomic status, that affects TV watching, and because our sample of parents has more cases from one socioeconomic group than does the sample of children, a difference has emerged.

It might be safe to assume, however, that any given parent-and-child pair falls into the same socioeconomic group. By taking parent-and-child pairs, therefore, and looking at *the difference for each pair*, the effect of other variables such as socioeconomic status is mitigated. In effect we are saying that all other variables that might determine TV watching are the same for each member of a given pair, and therefore only family relationship differs between them, allowing us to isolate its impact on the dependent variable.

The two dependent samples *t*-test for the mean difference

To illustrate the use of a dependent (or paired) samples *t*-test we will work through the following example. A survey of 10 families is conducted and a parent from each household and a child from each household are each asked to keep a diary of the amount of TV they watch during a set time period For each parent-child pair the amount of TV watched in minutes is recorded (Table 20.1).

Since the variable of interest, amount of TV watching, is measured at the interval/ratio level, the mean for each sample has been calculated. If we were comparing independent samples of adults and children, we would conduct a *t*-test on the difference between these two sample means. This procedure for the *independent samples t*-test can be summarized as follows:

1. Calculate the mean for *each* sample, then
2. Calculate the difference between the two sample means.

However, here we have selected these two samples so that we can match each member of one group with a member of the other. To conduct a dependent samples *t*-test, we reverse the order of the two steps:

1. Calculate the difference for each pair of cases (*D*), then
2. Calculate the mean of the differences $\left(\bar{X}_D\right)$.

To put it even more succinctly, an independent samples *t*-test looks at the *difference between the means*, while a dependent samples *t*-test looks at the *mean of the differences*.

Table 20.1 goes through the first step involved in performing a dependent samples *t*-test by calculating the difference in the amount of TV watched for each pair.

Table 20.1 Amount of TV watched by each household pair

Household	Minutes of TV watched by child	Minutes of TV watched by parent	Difference (*D*)
1	45	23	45–23 = 22
2	56	25	56–25 = 31
3	73	43	73–43 = 30
4	53	26	53–26 = 27
5	27	21	27–21 = 6
6	34	29	34–29 = 5
7	76	32	76–32 = 44
8	21	23	21–23 = –2
9	54	25	54–25 = 29
10	43	21	43–21 = 22
Mean	$\bar{X} = \frac{\Sigma X}{n} = 48.2$	$\bar{X} = \frac{\Sigma X}{n} = 26.8$	$\bar{X}_D = \frac{\Sigma D}{n} = 21.4$

You may notice that the mean difference is equal to the difference between means; this will always be the case. So why go through this alternative procedure for calculating the difference between two means? Although the mean difference will always equal the difference between the means, the variances will not be the same: the *variance* around the mean difference is much smaller than the variance around the difference between means. Because of this we may fail to reject a difference if it is treated as a difference between means, when we would have rejected it if it were treated as a mean difference.

We can see in Table 20.1 that there is on average a difference for each of the pairs that make up the samples. Let us assume that in the population as a whole there is no difference in the amount of TV watched between parents and their respective children. The null hypothesis is written in the following way:

$$H_0: \mu_0 = 0$$

When sampling from such a population, occasionally we might find a parent who watches more TV than his or her child, and occasionally we might find that a child watches a little more than his or her parent, but *if the null hypothesis of no difference is true*, on average the positive differences will cancel out the negative differences. It is not unreasonable to expect that random variation might occasionally result in a few extra households in which the parent watches less TV than the corresponding child, or vice versa, so that the mean difference between the samples is not zero. The bigger the difference between the sample result and the expected result of zero mean difference, though, the less likely that this will be due to random variation and the more likely that it reflects an underlying difference between parents and their children.

In this example the average of the differences is 21.4 minutes. Should this difference between the samples cause us to reject the hypothesis that there is no difference between the populations?

The formulas involved in conducting a *t*-test for the mean difference are:

$$t = \frac{\overline{X}_D}{s_D / \sqrt{n}}$$

where:

$$s_D = \sqrt{\frac{\Sigma D^2 - \frac{(\Sigma D)^2}{n}}{n - !}}$$

Note that *n* refers to the number of pairs, and not the total number of cases. Here $n = 10$, even though we have a total of 20 cases made up of 10 parents and 10 children. The sample score will be 14.2:

$$s_D = \sqrt{\frac{\Sigma D^2 - \frac{(\Sigma D)^2}{n}}{n - 1}} = \sqrt{\frac{6400 - \frac{45,796}{10}}{10 - 1}} = 14.2$$

$$t_{sample} \frac{\overline{X}_D}{s_D / \sqrt{n}} = \frac{21.4}{14.2 / \sqrt{10}} = 4.8$$

We can refer to the table for critical values of the *t*-distribution to obtain the *p*-score for this test statistics, at 9 degrees of freedom (the 10 pairs minus one). We can see that t_{sample} is larger than the largest value reported in the table, which is the *t*-score for $\alpha = 0.01$ (Table 20.2).

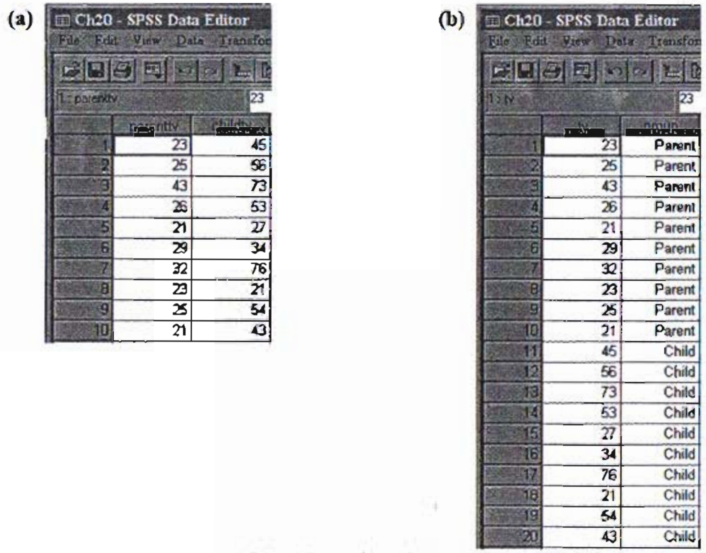
If we enter our data into the statistical calculation page located at the web address, www.physics.csbsju.edu/stats/Paired_t-test_NROW_form.html we find that the exact *p*-score for these data is 0.001. We therefore reject the hypothesis that there is no difference in the mean amount of TV watched by parents and their respective children at the 0.01 level.

Table 20.2 Critical values for *t*-distributions

	Level of significance for one-tail test				
	0.10	0.05	0.02	0.01	0.005
	Level of significance for two-tail test				
df	0.20	0.10	0.05	0.02	0.01
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
⋮	⋮	⋮	⋮	⋮	⋮
∞	1.282	1.645	1.960	2.325	2.576

The two dependent samples *t*-test using SPSS

In order for SPSS to do this same calculation, we first need to note the special way in which data are entered in order to conduct a dependent samples *t*-test. When coding data for paired samples, *each pair has to be treated as one case so that the information for each parent-and-child pair has to appear along the same row of data* (Figure 20.1(a)). The unit of analysis is the pair, not the individual people. Thus, in our example, there are only 10 rows of data.



(a) 1 : parenttv 23

	parenttv	childtv
1	23	45
2	25	56
3	43	73
4	26	53
5	21	27
6	29	34
7	32	76
8	23	21
9	25	54
10	21	43

(b) 1 : tv 23

	tv	group
1	23	Parent
2	25	Parent
3	43	Parent
4	26	Parent
5	21	Parent
6	29	Parent
7	32	Parent
8	23	Parent
9	25	Parent
10	21	Parent
11	45	Child
12	56	Child
13	73	Child
14	53	Child
15	27	Child
16	34	Child
17	76	Child
18	21	Child
19	54	Child
20	43	Child

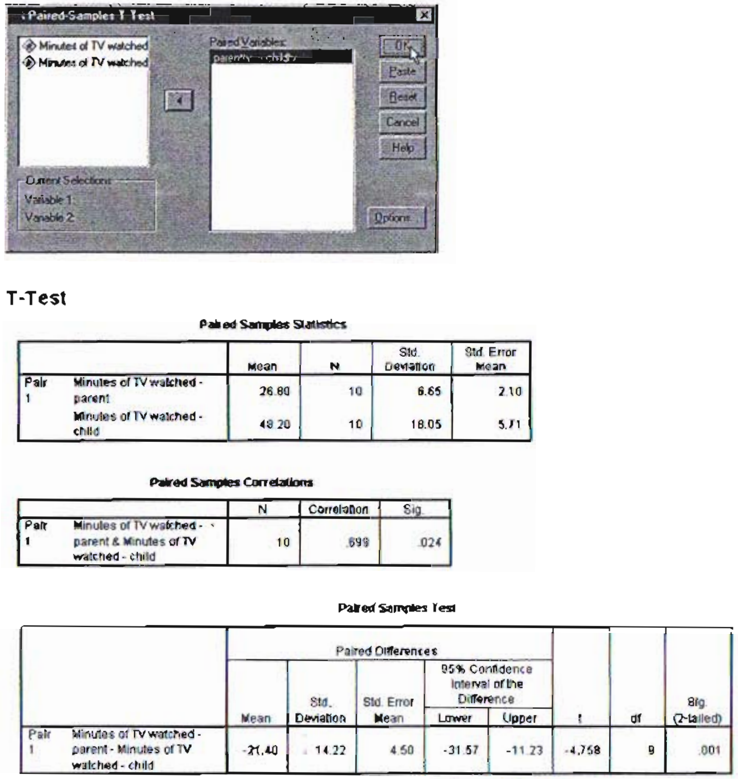
Figure 20.1 SPSS data entered for **(a)** two dependent samples and for **(b)** two independent samples

By placing each pair on the same row of data, we can match responses according to household. This produces a column for the amount of TV the parent watches, which is given the variable name **parenttv**, and a second column for the amount of TV the child watches, which has been given the variable name **childtv**. Thus each row has an entry for the amount of TV the child watches and the amount of TV the parent watches. If, on the other hand, we were treating the two samples as independent, we enter all 20 scores in the *same column*, so that there are 20 rows of data. We would then have a second column for the variable indicating the status of each case within a family – either parent or child (Figure 20.1(b)).

For data entered in the appropriate way for a two dependent samples *t*-test, which SPSS calls a paired-samples *t*-test, we follow the instructions in Table 20.3 (Figure 20.2).

Table 20.3 Paired-samples *t*-test using SPSS (Ch20.sav)

SPSS command/action	Comments
1 Select from the menu **Analyze/Compare Means/ Paired-Samples T Test**	This brings up the **Paired-Samples T Test** dialog box. In the top left of the box will be an area with a list of the variables entered in the data page
2 Click on **Minutes of TV watched – parent**, and then click on **Minutes of TV watched – child** in the source variable list	This highlights the two variables that will be matched
3 Click on ▸	This pastes the highlighted variables into the **Paired Variables:** target list
4 Click on **OK**	



T-Test

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Minutes of TV watched - parent	26.80	10	6.65	2.10
	Minutes of TV watched - child	48.20	10	18.05	5.71

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Minutes of TV watched - parent & Minutes of TV watched - child	10	.699	.024

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
					95% Confidence Interval of the Difference				
		Mean	Std. Deviation	Std. Error Mean	Lower	Upper			
Pair 1	Minutes of TV watched - parent - Minutes of TV watched - child	-21.40	14.22	4.50	-31.57	-11.23	-4.758	9	.001

Figure 20.2 The SPSS Paired-Samples T Test dialog box and output

The output begins with a table called **Paired Samples Statistics**. This provides the descriptive statistics for the paired samples: the mean of 48.20 minutes for the 10 children and 26.80 minutes for the 10 parents. The next table with the correlation information is not relevant to our discussion here. The important table is the last one labeled **Paired Samples Test**. This contains the information on the dependent samples *t*-test, and confirms the calculations above. The mean difference is calculated as –21.4 minutes. The *t*-test for this value is –4.758. From the last column we see that, with 9 degrees of freedom, a mean difference this large or greater will occur, if the null hypothesis of no difference is true, less than one time in every thousand samples (.001). This is well below any normal alpha level, such as 0.05 or 0.01, so we reject the null hypothesis of no difference.

The output also provides the 95 percent confidence interval for the estimate of the difference. The upper limit of the estimate is –11.23 while the lower limit is –31.57. We can use this information to conduct the hypothesis test. Since the interval does not include the value of 0, we can conclude that the difference in the population as to the amount of TV watched by parents and their children is not zero.

Example

A teacher is interested in the effect of a new study technique on the ability of students to complete basic arithmetic. The teacher selects five students and asks them to complete a basic arithmetic test. The teacher then introduces the new study technique and after a month selects the same five students and asks them to complete a similar test. The results are presented in Table 20.4.

Table 20.4 Results of arithmetic test

Student	Time to complete test – pre	Time to complete test – post
Stacey	7.3	6.8
Chloe	8.5	7.9
Billie	6.4	6.0
Alana	9.0	8.4
Timothy	6.9	6.5
Mean	$\bar{X} = 7.62$	$\bar{X} = 7.12$

Initially the teacher treats these as independent samples. The average time for the pre-test is 7.62 minutes while for the post-test it is 7.12 minutes. Using the *independent samples t-test* for the difference between sample means, the teacher obtains a sample *t*-score of 0.75, which is not significant at the 0.05 level. Feeling disheartened that, although the sample results looked promising, the inference test did not reject the possibility that the improvement came about by sampling error, the teacher decides to abandon the new study method.

Fortunately a colleague knows a little more about statistics and realizes that, since the same students make up each sample, a dependent samples test is required for this research design. They work through the data with the following results.

Step 1: State the null and alternative hypotheses

$$H_0: \mu_D = 0$$

$$H_a: \mu_D \neq 0$$

Step 2: Choose the test of significance

Here we are comparing two dependent samples in terms of mean differences. Therefore we use the two dependent samples *t*-test for the mean difference.

Step 3: Describe the sample and calculate the p-score

To help in calculating the mean difference between the samples and the associated *t*-score we construct Table 20.5.

Table 20.5 Calculations for dependent samples t-test

Student	Time to complete test – pre	Time to complete test – post	Difference	D^2
Stacey	7.3	6.8	0.5	0.25
Chloe	8.5	7.9	0.6	0.36
Billie	6.4	6.0	0.4	0.16
Alana	9.0	8.4	0.6	0.36
Timothy	6.9	6.5	0.4	0.16
Sum			$\Sigma D = 2.5$	$\Sigma D^2 = 1.29$
Mean			$\bar{X}_D = 0.5$	

Substituting this information into the equation for the standard error and then for t_{sample} we get a test statistic of 111.8:

$$s_D = \sqrt{\frac{\Sigma D^2 - \frac{(\Sigma D)^2}{n}}{n-1}} = \sqrt{\frac{1.29 - \frac{(2.5)^2}{5}}{5-1}} = 0.1$$

$$t_{sample} = \frac{\bar{X}_D}{s_D/\sqrt{n}} = \frac{0.5}{0.1/\sqrt{5}} = 111.8$$

Step 4: Decide at what alpha level, if any, the result is statistically significant

The t-score, when calculated on the basis of dependent samples rather than independent samples, is now clearly significant at even the extremely low p-score of 0.01 level.

Step 5: Report results

Five children were randomly selected and asked to complete an arithmetic test, upon which they took 7.62 minutes to complete on average. Their teacher then introduced a new study technique and after a month the same students were asked to complete a similar test. The mean on the second test was 7.12 minutes. The reduction in mean completion time is statistically significant, using a dependent samples t-test ($t = 111.8$, $p < 0.01$, two-tail). The teacher can reject the hypothesis that the improvement came about only by random chance.

Exercises

20.1 (a) What is the mean difference for the following 10 pairs of observations?

Pair	Observation 1	Observation 2
1	12	15
2	10	13
3	8	13
4	14	14
5	12	18
6	15	13
7	14	18
8	9	9
9	18	11
10	13	14

(b) What is the standard error?

(c) Conduct a dependent samples t-test on the following data.

20.2 Test the following hypotheses using the data provided:

	H_0	H_a	Mean difference	s_D	n	α
(a)	$\mu_D = 0$	$\mu_D \neq 0$	2.3	1.4	20	0.10
(b)	$\mu_D = 0$	$\mu_D < 0$	-3.2	20	41	0.05

20.3 One hundred and forty patients are given a new treatment for lowering blood pressure. The mean difference between systolic blood pressure for these patients before and after the treatment is –9, with a standard deviation of 8. Given that the drug may have side effects and therefore the need to minimize a Type I error, the treatment will only be adopted if it is significant at a 0.01 level. Should it be adopted?

20.4 A company wants to investigate whether changes in work organization can significantly improve productivity levels. It randomly selects 10 workplaces and measures productivity levels in terms of units per hour produced. It then introduces a program in these workplaces giving workers greater discretion over conditions and job structure, and measures productivity levels 6 months later. The results are presented in the following table:

Workplace	Productivity before change	Productivity after change
1	120	165
2	121	154
3	145	120
4	112	155
5	145	164
6	130	132
7	134	154
8	126	162
9	137	130
10	128	142

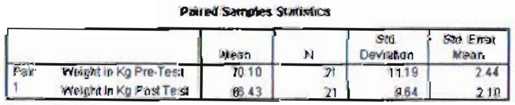
Has the program significantly improved productivity levels (note the form of the alternative hypothesis)?

20.5 The following data list the asking and selling prices (in dollars) for a random sample of 10 three-bedroom homes sold during a certain period:

Home	Asking price ($)	Selling price ($)
1	140,000	144,300
2	172,500	169,800
3	159,900	155,000
4	148,000	150,000
5	129,900	129,900
6	325,000	315,000
7	149,700	146,000
8	147,900	149,200
9	259,000	259,000
10	223,900	219,000

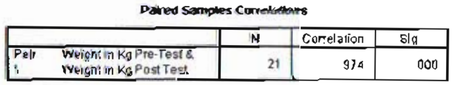
Why is a dependent samples test appropriate in this situation? Using a dependent samples t-test, do people receive the price they want when selling their home? Enter these data in SPSS and conduct this test. Compare the results with your hand calculations.

20.6 A nutritionist is interested in the effect that a particular combination of exercise and diet has on weight loss. The nutritionist selected a group of people and measured their weight in kilograms before and after a program of diet and exercise. A paired-samples t-test was conducted on SPSS with the following results:



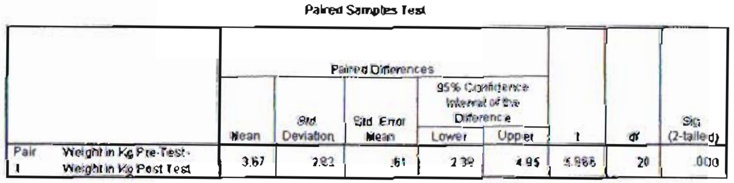
Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Weight in Kg Pre-Test	70.10	21	11.19	2.44
	Weight in Kg Post Test	66.43	21	9.64	2.10



Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	Weight in Kg Pre-Test & Weight in Kg Post Test	21	.974	.000



Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	Weight in Kg Pre-Test - Weight in Kg Post Test	3.67	2.82	.61	2.39	4.95	5.986	20	.000

From this output determine the:

(a) variable names assigned to the before-and-after measurements;
(b) number of pairs in the test;
(c) mean weight for the pre-test sample;
(d) mean weight for the post-test sample;
(e) mean difference between the two samples;
(f) value of t_{sample} and the number of degrees of freedom;
(g) probability of obtaining this mean difference if the null hypothesis of no difference is true;
(h) upper limit of the confidence interval for the estimate of the difference;
(i) lower limit of the confidence interval for the estimate of the difference.
(j) What should the nutritionist conclude about the effect of the program?

20.7 From the previous question if this nutritionist considered an average weight loss of 5 kg or more to be the measure of success of this program, can we say that the program was successful? What does this say about the difference between practical and statistical significance?

20.8 Using the data for the example in the text regarding the study technique to improve mathematical skills, enter the data into SPSS, first to conduct an independent samples test and second to conduct a dependent samples test. What explains the difference?

20.9 Using the **Employee data** file determine whether there has been a significant increase in salaries since employees began working at the company. If the research question was, alternatively, whether any increase was significantly greater than $15,000, what would you conclude?

PART 5

Inferential statistics: Tests for frequency distributions

21

One sample tests for a binomial distribution

The previous chapters looked at inference tests for a **mean**. These procedures apply to research questions that direct our investigation to the central tendency of a distribution, and the variable in which we are interested is measured at the interval/ratio level. We call such tests **parametric tests** because they test hypotheses about population parameters (in this instance the mean). However, there are many instances where we are interested in aspects of a variable's distribution other than its mean, such as its frequency distribution.

Take for example the problem we dealt with in Chapter 16, where the Health Department had a policy of allocating funds to a region depending on whether the average age of the population is over 40 years. Clearly, this policy rule directs our analysis to the *average* value for the variable of interest – age. Assume that the Health Department suddenly changes its policy rule and decides now to provide extra funding to a region's health services only if 20 percent or more of the population in that region is over 40 years of age. Suddenly the mean age of the population becomes irrelevant. We can still calculate the mean, but this will not assist us in making the policy decision about funding. The appropriate way to describe the data to deal with this new policy rule is to divide the sample into those people who are 40 years of age or less and those over 40, and calculate the percentage of people in each category. In effect, we have organized the data into the simplest type of frequency distribution, called a **binomial distribution**.

A binomial (or dichotomous) distribution has only two categories or values.

Data considerations

Some variables are intrinsically measured on a dichotomous scale. A classic example is a coin toss, which has only two possible outcomes: either heads or tails. Similarly, questions in opinion polls that allow only 'Yes/No' responses are dichotomous. Sex is another common example of a variable that intrinsically has a binomial distribution: someone is either male or female.

However, even where a variable does not initially have only two categories, it can be transformed into one that does. In fact, practically any variable measured at any level can be turned into a binomial by collapsing categories.

Nominal scales

A nominal variable that does not intrinsically have only two categories can be collapsed into a binomial by simply specifying the number of cases that fall into an existing category (or combination of categories) or not. For example, a nominal distribution of cases according to religious denomination might begin with five classifications for religion: Catholic, Protestant, Jewish, Orthodox, and Muslim. These can be collapsed into a binomial distribution in one of two ways:

- by referring to the percentage of cases that fall into one of the existing categories or not, such as Catholic and Non-Catholic, or
- by creating two entirely new categories by combining the existing ones, such as Christian and Non-Christian.

Each of these methods of collapsing categories is represented in Tables 21.1–21.4.

Table 21.1 Religious affiliation: original distribution

Religion	Frequency
Catholic	20
Protestant	15
Orthodox	12
Muslim	12
Jewish	7

→

Table 21.2 Religious affiliation: binomial distribution

Religion	Frequency
Catholic	20 (30%)
Non-Catholic	46 (70%)

Table 21.3 Religious affiliation: original distribution

Religion	Frequency
Catholic	20
Protestant	15
Orthodox	14
Muslim	10
Jewish	7

→

Table 21.4 Religious affiliation: binomial distribution

Religion	Frequency
Christian	49 (74%)
Non-Christian	17 (26%)

Ordinal and interval/ratio scales

Ordinal or interval/ratio scales can be collapsed into a binomial distribution by simply specifying the number of cases that fall above or below a particular value on the scale. For example, a list of exam scores can be collapsed into a binomial by selecting 50 percent as the dividing line and organizing the scores into 'pass' and 'fail'. After arranging the data into a binomial distribution and calculated the relevant percentage of cases in each of the two categories, we can then proceed to conduct an inference test on these percentages. To do this we have to know the properties of **the sampling distribution of sample percentages.**

The sampling distribution of sample percentages

In the previous chapters we had a sample mean and we were interested in making an inference from this sample mean to the mean for the population. To make this inference we constructed the sampling distribution of the sample means. This sampling distribution allows us to assess the probability of obtaining our actual sample mean from a population with a specific hypothesized value (the null hypothesis). When working with a binomial distribution, however, the descriptive statistic calculated from the sample is no longer the mean. Instead it is the percentage of cases that fall within one of the two possible categories of the variable. Having calculated the sample percentage we then need to make an inference about the percentage for the population as a whole. Thus we need to explore the properties of the sampling distribution of sample percentages: the distribution of sample percentages that will arise from repeated random samples of equal size.

For example, we might know that 50 percent of all students at a (hypothetical) university are male and 50 percent are female. Despite this, if we take a random sample of 100 university students we will not necessarily get 50 males and 50 females. Random variation will cause some samples to include slightly more females, while other samples will include slightly more males. But most of these repeated samples will have a percentage of each sex either equal or close to 50 percent. In other words, while there is some variation in the distribution of repeated sample percentages, these sample percentages will cluster around the 'true' population value of 50 percent.

If we take an infinite number of random samples of equal size from a population, and calculate the percentage of cases in each that have a certain value for a binomial distribution, the sampling distribution of these sample percentages will have the following properties:

- *The sampling distribution is approximately normal with a median percentage equal to the population value.* It is only *approximately* normal because a binomial scale is discrete, whereas the normal curve is continuous. However, the larger the sample size the more closely the distribution approximates the normal.
- The standard error of the sampling distribution will be defined by the following equation where P_u is the population percentage.

$$\sigma_p = \sqrt{\frac{P_u(1-P_u)}{n}}$$

These two pieces of information are very useful, as we discovered in previous chapters. Knowing the distribution of *all possible* sample percentages that could come from a particular population allows us to calculate the probability of getting any *given* sample result from a population with an hypothesized value. For example, if a *sample* has 60 percent females, we can calculate the probability that this was the result of sampling error when drawing from a *population* that only has 50 percent females. This is exactly the type of question the **one sample *z*-test for a percentage** is designed to answer.

The *z*-test for a binomial percentage

Although we are describing the data by organizing it into a binomial distribution rather than by calculating a mean, the procedures for making an inference from a sample to the population are similar. In practical terms the steps involved in an hypothesis test for a percentage are exactly the same as when conducting an hypothesis test for a mean. We conduct an inference test, much like those in Chapters 15 and 16, on the *percentage of the sample falling in one of the two categories of the binomial*, rather than on the sample mean.

Since the sampling distribution is normal we conduct a *z*-test on the difference between the sample percentage and the test value. The specific formulas used to calculate z_{sample} are (where P_s is the sample percentage and P_u is the population percentage):

$$z_{sample} = \frac{(P_s - 0.5) - P_u}{\sqrt{\frac{P_u(100-P_u)}{n}}} \quad \text{where } P_s > P_u$$

or

$$z_{sample} = \frac{(P_s + 0.5) - P_u}{\sqrt{\frac{P_u(100-P_u)}{n}}} \quad \text{where } P_s < P_u$$

The addition or subtraction of 0.5 to or from the sample percentage in each of these equations is made because, strictly speaking, a binomial distribution is not exactly normal and the addition or subtraction of 0.5 (called a continuity correction) gives us a better approximation. With samples larger than 30 this approximation will be fairly accurate, but with less than 30 the approximation is not accurate and an exact binomial probability test should be used. Many statistics books print tables for the exact binomial distribution for various sample sizes, and these should be referred to in the small sample case rather than the standard normal table. A number of web pages also provide such tables and these are listed at members.aol.com/johnp71/javastat.html#Tables. SPSS automatically calculates an exact binomial probability in the small sample case.

Example

A researcher is concerned that the local area is harder hit by recession than the rest of the country. The researcher knows that the national unemployment rate is 11 percent. The researcher randomly asks 120 local people who are in the labor market if they are unemployed. The raw data are described by constructing the binomial distribution in Table 21.5. Does this indicate that this local area is harder hit by recession?

Table 21.5 Distribution of respondents by employment status

Employment status	Frequency	Percentage
Employed	102	85
Unemployed	18	15
Total	120	100

Step 1: State the null and alternative hypotheses

H_0: The local area has the same percentage of people unemployed as the rest of the nation.

$$H_0\text{: } P_u = 11\%$$

H_a: The local area has a higher percentage of people unemployed than the rest of the nation.

$$H_a\text{: } P_u > 11\%$$

Step 2: Choose the test of significance

The research question is interested in the percentage of people in a category of a binomial distribution (i.e. people unemployed). Therefore we use a single-sample z-test for a percentage.

Step 3: Describe the sample and calculate the p-score

The relevant sample data from the binomial distribution in Table 21.5 are:

$$n = 120$$

$$P_s = \frac{18}{120} \times 100 = 15\%$$

We substitute these data into the equation to obtain the test statistic of $z = 1.2$:

$$z_{sample} = \frac{(P_s - 0.5) - P_u}{\sqrt{\dfrac{P_u(100 - P_u)}{n}}} = \frac{(15 - 0.5) - 11}{\sqrt{\dfrac{11(100 - 11)}{120}}} = 1.2$$

From the table for the area under the standard normal curve, the p-value associated with this z-score is 0.23 (two-tail) or 0.115 (one-tail).

Step 4: Decide at what alpha level, if any, the result is statistically significant

The result we obtain is not statistically significant at any of the conventional alpha levels. There is a high probability that the sample percentage is higher than the hypothesized percentage of 11 percent simply as a result of sampling error.

Step 5: Report results

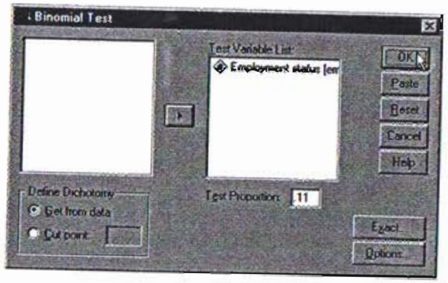
A researcher is concerned that her local area is harder hit by recession than the rest of the country. She randomly asks 120 local people who are in the labor market if they are unemployed and finds 18 people sampled are unemployed (15%), which is higher than the national unemployment rate of 11 percent. Although the *sample* has a higher unemployment rate than the national figure, the difference is not big enough to suggest that the *population* in this area systematically suffers from a higher unemployment rate ($z = 1.2$, $p = 0.23$, two-tail).

The z-test for a binomial percentage using SPSS

In order to work through this test on SPSS, the data for the previous example have been entered into a data file. SPSS calls the test for a percentage a binomial test, and the commands needed to carry out this test are as shown in Table 21.6 and Figure 21.1 along with the output from this analysis.

Table 21.6 The Binomial Test on SPSS (file: **Ch21-1.sav**)

SPSS command/action	Comments
1 From the menu select **Analyze/Nonparametric Tests/ Binomial**	This brings up the **Binomial Test** dialog box
2 From the source variable list click on **Employment status**	
3 Click on ▸	This pastes **Employment status** in the **Test Variable List:**
4 In the square next to **Test Proportion** type 0.11 over 0.50	The default setting is .50, which means that SPSS will compare the sample proportion against 0.5 unless we specify an alternative, as in this case where we specify 0.11 as the test proportion
5 Click on **OK**	

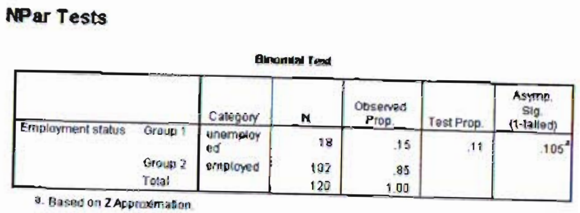


Binomial Test

Test Variable List: Employment status [em

Test Proportion: .11

Define Dichotomy
● Get from data
○ Cut point



NPar Tests

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Asymp. Sig. (1-tailed)
Employment status	Group 1	unemployed	18	.15	.11	.105[a]
	Group 2	employed	102	.85		
	Total		120	1.00		

a. Based on Z Approximation

Figure 21.1 SPSS Binomial Test dialog box and output

Notice that SPSS conducts the test in terms of proportions rather than percentages. This is not such a dramatic difference since we know that any proportion can be easily converted into an equivalent percentage simply by moving the decimal point two places to the right.

This result confirms the calculations above. In the **Binomial Test** table we have a column headed N with the number of cases in each of the categories of the binomial distribution, and then a column headed Observed Prop. indicating the relative frequencies as proportions. The last column headed Asymp. Sig. (1-tailed) is the important one for the purposes of the inference test. Although we are not given the value of z_{sample}, we are given the one-tail probability associated with it. Here the one-tail probability of .105 indicates that if the null were true (the population from which the sample is drawn has an unemployment rate of 11 percent) at least 1-in-10 samples will have an unemployment rate of 15 percent or more. This is not too unlikely: the assumption that the null is true cannot be rejected.

There are four points to notice about the SPSS **Binomial Test** command:

1. SPSS always produces a one-tail test when a test percentage is specified rather than the default value of 0.5. If the alternative hypothesis requires a two-tail test, then we simply double the one-tail probability. If one tail of the sampling distribution, at this *z*-score, contains 0.105 of the area under the curve, two tails will contain 0.21 of the area under the normal curve.
2. SPSS provides two methods by which we define the two groups that make up a binomial distribution. Under **Define Dichotomy** in the **Binomial Test** dialog box the default option is **Get from data**. This is used when the variable we are analyzing already has a binomial distribution, such as employment status in this example. However, sometimes we might be working with a variable that has three or more values or categories. We could use the **Recode** command and create a new variable by collapsing the values into two. This is unnecessary because if we choose the **Cut point** option, we can indicate the point on a scale that will divide a set of cases into a binomial. The value we type into the **Cut point:** box *defines the upper limit of the first group, and the percentage of cases in that group will be compared to the test value*. Thus if I had a range of exam scores and I wanted to analyze the percentage that passed or failed, I would type 49 as the cut point, and SPSS would then calculate the percentage of cases that were less than or equal to this cut point and compare that to a specified test value for failure rate.
3. SPSS often rounds the observed proportion to one decimal place when a test proportion is entered with only one decimal place. This may cause the observed proportion to appear to 'equal' the test proportion, when in fact they are different. You will need to edit the table by clicking on it and selecting **Edit/SPSS Pivot Table Object/Edit**, and then, after selecting the cells in the table you wish to change, change the number of decimal places using the **Format/Cell Properties** command from the menu.
4. SPSS does not give the confidence interval information for the sample proportion, unlike the tests for a mean we discussed in the previous chapters. We will discuss how confidence intervals for a percentage can be calculated below.

Example

A political scientist is interested in whether there has been a change in people's attitudes toward the major political parties that normally contest elections in a particular political system. The researcher groups political parties into two distinct categories: major and non-major parties. At the previous election the percentage of people who voted for one of the major parties was 85 percent. A survey of 300 eligible voters conducted 2 years since that election indicates that 216 (72 percent) plan to vote for one of the major parties at the next election (Table 21.7).

Table 21.7 Support for major political parties

Who do you support?	Last election	Next election
Major parties	85%	72%
Other parties	15%	28%

Can we say that the level of support for the major parties has changed since the last election? Since we are dealing with a situation where $P_s < P_u$ we use the following formula to calculate the test statistic:

$$z_{sample} = \frac{(P_s + 0.5) - P_u}{\sqrt{\frac{P_u(100 - P_u)}{n}}} = \frac{(72 + 0.5) - 85}{\sqrt{\frac{85(100 - 85)}{300}}} = -6.1$$

A test statistic of –6.1 has an extremely low probability (< 0.0001) of occurring by chance from a population where 85 percent of its members plan to vote for one of the major parties at the next election. We can reject the null hypothesis that the percentage of people planning to vote for one of the major parties is the same as that in the previous election.

Estimating a population percentage

Chapter 17 detailed the procedure for estimating from a sample mean a confidence interval within which the population mean falls. A similar procedure can be followed to construct a confidence interval from a sample percentage, within which the (unknown) population percentage falls.

Estimating population percentages is common in public opinion surveys. We often read in newspapers that a certain percentage of eligible voters favor one person over another as preferred Prime Minister or President. This percentage figure is not obtained by surveying all eligible voters, but rather through a sample of eligible voters. We therefore need to estimate the population value from this sample result.

Unfortunately, there is no single equation upon which everyone agrees for constructing the confidence interval around a sample percentage (see R.G. Newcombe, 1998, Two-sided confidence intervals for the single proportion: Comparison of seven methods, *Statistics in Medicine*, vol. 17, pp. 857–72). This may explain why SPSS does not provide such an interval as part of the **Binomial Test** command. To overcome this, Table 21.8 and Table 21.9 provide the sampling errors for various sample sizes and sample 'splits' into the two categories of the binomial distribution, at the 95% and 99% confidence levels respectively, using the adjusted Wald method (see A. Agresti and B.A. Coull, 1998, Approximate is better than 'Exact' for interval estimation of binomial proportions, *The American Statistician*, vol. 52, pp. 119–26). Note that these figures may vary from those that might be presented in other books, internet calculation pages listed at members.aol.com/johnp71/javastat.html#Confidence or statistics programs, as they may use slightly different equations. These differences, however, are usually so small as to be unimportant for practical purposes.

To use these tables:

- we find the row with the closest sample size to the one we are using;
- we find the column with the 'split' across the two categories closest to the one we have in our sample;
- we find the intersection of this row and this column;
- we then add/subtract this percentage to/from the sample percentage to determine the upper/lower bounds of the interval, at that confidence level.

To illustrate the use of these tables, we will use the data for the previous example where 72% of a sample of 300 people surveyed stated they will vote for one of the major parties at the next election. At the 95% confidence level, we read down the rows of Table 21.8 until we reach the row for a sample size of 300, and then read across this row until we reach the value for the 70/30 column, which gives us a sampling error of 5.2%. Thus the lower limit of the confidence interval is 66.8 percent (72 – 5.2) and the upper limit is 77.2 percent (72 + 5.2).

Table 21.8 Sampling errors for a binomial distribution (95% confidence level)

Sample size	Binomial percentage distribution					
	50/50	60/40	70/30	80/20	90/10	95/5
50	13.3	13.1	12.4	11.1	9.0	7.4
100	9.6	9.4	8.9	7.8	6.1	4.8
150	7.9	7.7	7.3	6.4	4.9	3.8
200	6.9	6.7	6.3	5.5	4.3	3.2
250	6.2	6.0	5.7	5.0	3.8	2.9
300	5.6	5.5	5.2	4.5	3.4	2.6
400	4.9	4.8	4.5	3.9	3.0	2.2
500	4.4	4.3	4.0	3.5	2.7	2.0
600	4.0	3.9	3.7	3.2	2.4	1.8
700	3.7	3.6	3.4	3.0	2.2	1.6
800	3.5	3.4	3.2	2.8	2.1	1.5
900	3.3	3.2	3.0	2.6	2.0	1.4
1000	3.1	3.0	2.8	2.5	1.9	1.4
1100	2.9	2.9	2.7	2.4	1.8	1.3
1200	2.8	2.8	2.6	2.3	1.7	1.2
1300	2.7	2.7	2.5	2.2	1.6	1.2
1400	2.6	2.6	2.4	2.1	1.6	1.2
2000	2.2	2.1	2.0	1.8	1.3	1.0
10,000	1.0	1.0	0.9	0.8	0.6	0.4

Table 21.9 Sampling errors for a binomial distribution (99% confidence level)

Sample size	Binomial percentage distribution					
	50/50	60/40	70/30	80/20	90/10	95/5
50	17.6	17.3	16.3	14.6	11.8	9.0
100	12.6	12.4	11.7	10.3	8.1	6.3
150	10.4	10.2	9.6	8.4	6.5	5.0
200	9.0	8.9	8.3	7.3	5.6	4.3
250	8.1	7.9	7.4	6.6	5.1	4.0
300	7.4	7.3	6.8	6.0	4.5	3.4
400	6.4	6.3	5.9	5.2	3.9	2.9
500	5.7	5.6	5.3	4.6	3.5	2.6
600	5.2	5.1	4.8	4.2	3.2	2.4
700	4.9	4.8	4.5	3.9	2.9	2.2
800	4.5	4.5	4.2	3.6	2.8	2.0
900	4.3	4.2	3.9	3.4	2.6	1.9
1000	4.1	4.0	3.7	3.3	2.5	1.8
1100	3.9	3.8	3.6	3.1	2.3	1.7
1200	3.7	3.6	3.4	3.0	2.2	1.6
1300	3.6	3.5	3.3	2.9	2.2	1.6
1400	3.4	3.4	3.2	2.8	2.1	1.5
2000	2.9	2.8	2.6	2.3	1.7	1.3
10,000	1.3	1.3	1.2	1.0	0.8	0.6

We can use confidence intervals obtained from these tables to conduct the hypothesis test we detailed in the previous example. Since the confidence interval of 66.8–77.2% does not include the test value of 85%, we can reject the hypothesis that support for the major political parties has not changed, at the 95% confidence level (alpha = 0.05). In fact, if we determine the confidence interval for the 99% level, we derive an interval ranging from 65.2–78.8%, which still excludes the test value; we reject the null at the 0.01 alpha level.

Before turning to another example to illustrate this method, there are a few points to observe about these tables.

1. As with confidence intervals for a mean, the confidence intervals for a percentage shrink dramatically with increases in the size of small samples, but shrink only fractionally with increases in the size of large samples. Thus an increase in the sample size from 50 to 100 'buys' a substantial increase in accuracy (around 4–5%), whereas an increase in sample size from 1400 to 2000 barely increases accuracy by half a percent. Thus with small samples it can be worth spending extra research money to increase sample size even slightly, but beyond a certain point, around 1200–1400, the cost of increasing the sample size does not generate much accuracy, in terms of smaller confidence intervals.
2. As with other inferential statistics, the more dispersed the data the wider the confidence interval. This accords with common sense. If the population is diverse, then random samples drawn from that population will have a greater range of outcomes. Thus a 50/50 split, which represents the greatest dispersion of data in a binomial distribution, is much wider at any given sample size and confidence level than the corresponding 95/5 split, which indicates a group that is very homogeneous.
3. The confidence interval for a given split and sample size is wider for the 99% level than it is for the corresponding 95% level. Again this accords with common sense; to be more confident that our interval takes in the true population value it has to be much wider.
4. These tables can also be used to determine the sample size required to achieve a desired level of accuracy in estimating a population percentage, based on an assumption about the expected split. Thus they can be used in advance of collecting data to determine how many cases should be included in the study.

The runs test for randomness

The proportion of the sample that falls into one category or the other of a binomial distribution is not the only descriptive statistic we might be interested in. We might have no interest in the question of what proportion of the total sample falls in one category or the other. Instead we might be interested in the *series or sequence of scores*: how each score follows on from the previous one. Usually we look at the sequence of cases with a particular question in mind: is a series of events **random**?

An event is **random** if its outcome in one instance is not affected by the outcome in other instances.

For example, if I toss a coin and the coin comes down 'heads', then if it is an unbiased coin we should not expect the next toss to be more likely to come down 'heads' (or 'tails').

To decide whether the value of a variable in one case is random with respect to the value it takes in other cases, we conduct a *z*-test on the number of sample runs – the **runs test for randomness**.

The idea behind a runs test of randomness is simple. If the outcome of a coin toss is random, and an unbiased coin is tossed and comes up heads, the probability of the next toss being either heads or tails should be 50/50. There should be a fairly even spread of heads and tails after each toss. If any of the following three results occurs from tossing a coin 20 times we might get a little suspicious:

Set 1: T T T T T T T T T T T T T T T T T T T T

Set 2: H H H H H H H H H H H H H H H H H H H H

Set 3: H T H T H T H T H T H T H T H T H T H T

In each set, it seems that each case is not random. In the first two sets of tosses each flip leads to the same result in the next – tails seem to determine tails, and heads seem to determine heads. In the third set of tosses, tails determine heads and vice versa. Either way, the outcome of a coin toss does not appear to be random. But another interpretation could be that each of these outcomes occurred simply on the basis of chance. Coin tosses might be random, but we just happened by chance to get these outcomes.

To decide between these explanations, we describe the results of each set of tosses by calculating the number of **runs**.

A **run** is a sequence of scores that have the same outcome for a variable. A run is preceded and followed by scores that have a different outcome for a variable, or no data.

In short, we look for sequences of like results in the series. In the first two sets of coin tosses above we have 1 run each:

Set 1: $\underbrace{\text{T T T T T T T T T T T T T T T T T T T T}}_{\text{1 run}}$

Set 2: $\underbrace{\text{H H H H H H H H H H H H H H H H H H H H}}_{\text{1 run}}$

In the third set of tosses we have 20 runs:

Set 3: $\underbrace{\underline{\text{H}}\ \underline{\text{T}}\ \underline{\text{H}}\ \underline{\text{T}}\ \underline{\text{H}}\ \underline{\text{T}}\ \underline{\text{H}}\ \underline{\text{T}}\ \underline{\text{H}}\ \underline{\text{T}}\ \underline{\text{H}}\ \underline{\text{T}}\ \underline{\text{H}}\ \underline{\text{T}}\ \underline{\text{H}}\ \underline{\text{T}}\ \underline{\text{H}}\ \underline{\text{T}}\ \underline{\text{H}}\ \underline{\text{T}}}_{\text{20 runs}}$

It is conceivable that I could toss an unbiased coin and get such results – they could happen just by chance. This is the null hypothesis of randomness. However, such results are very unlikely. The probability of getting either 1 run or 20 runs from 20 coin tosses, *if* the toss of a coin is truly random, is extremely low. On average, we expect to get between 1 and 20 runs. In fact, the value we expect to get if the results are random, and which we use in the null hypothesis, is given by the formula:

$$H_0:\ \mu_R = \frac{2n_1n_2}{n} + 1$$

where n_1 is the number of cases with a given value, n_2 is the number of cases with the other value, and n is the total number of cases.

However, even though coin tosses are random, individual samples will not always have this many runs. The spread of possible sample results around the expected value is given by:

$$\sigma_R = \sqrt{\frac{n^2 - 2n}{4(n-1)}}$$

Given this information we can perform a z-test to determine whether the sample value of R is likely to be the result of chance or something systematic. The test statistic is calculated using the following equations, where R is the number of runs in the sample, μ_R is the number of runs expected from repeated sampling, and σ_R is the standard error of the sampling distribution (where the sample is less than 20, the sampling distribution of sample runs will not be approximately normal, and an exact probability test needs to be conducted; in our test we will work with samples larger than 20 where the normal approximation is applicable):

$$z_{sample} = \frac{(R + 0.5) - \mu_R}{\sigma_R} \quad \text{where } R < \mu_R$$

or

$$z_{sample} = \frac{(R - 0.5) - \mu_R}{\sigma_R} \quad \text{where } R > \mu_R$$

We simply follow the hypothesis testing procedure we have learnt and compare the sample z-score and probability with a pre-chosen critical value and decide either to reject or not to reject the null hypothesis. It is important when conducting this test that the data are ordered in the sequence in which they were generated. For example, when looking at time series, as we do below, the data are ordered according to year.

Example

The data upon which a runs test is conducted must be ordered into a sequence in some way. This condition makes this test one that is commonly used to analyze time series data. Time series refers to a sequence of cases occurring over successive time periods. For example, a doctor might be interested in whether the pain associated with a particular condition occurs on random days or whether it occurs over periods extending beyond one day. To assess this the doctor monitors a patient with this condition over a 33-day period, recording whether the pain suffered is high or low. Do days of relatively high pain tend to follow each other and do days of relatively low pain tend to follow each other? Table 21.10 provides the raw data.

Table 21.10 Pain levels for patient

Day	Pain level	Above or below median
1-Mar-2004	High	*Run 1*
2-Mar-2004	High	
3-Mar-2004	Low	*Run 2*
4-Mar-2004	High	*Run 3*
5-Mar-2004	Low	*Run 4*
6-Mar-2004	Low	
7-Mar-2004	Low	
8-Mar-2004	Low	
9-Mar-2004	Low	
10-Mar-2004	Low	
11-Mar-2004	Low	
12-Mar-2004	Low	
13-Apr-2004	Low	
14-Mar-2004	Low	
15-Mar-2004	Low	
16-Mar-2004	Low	
17-Mar-2004	Low	
18-Mar-2004	High	*Run 5*
19-Mar-2004	Low	*Run 6*
20-Mar-2004	High	*Run 7*
21-Mar-2004	Low	*Run 8*
22-Mar-2004	High	*Run 9*
23-Mar-2004	High	
24-Mar-2004	High	
25-Mar-2004	High	
26-Mar-2004	High	
27-Mar-2004	High	
28-Mar-2004	High	
29-Mar-2004	High	
30-Mar-2004	High	
31-Mar-2004	High	
1-Apr-2004	High	
2-Apr-2004	High	

To see whether days of low or high pain occur in 'patches' or are distributed randomly across days, we have also shaded sequences of High pain days and numbered the runs. Thus we are able to describe our sample result by saying that there are 9 runs. How likely is this to occur if the pain level on any given day is random with respect to the level on the previous day? To specify the null hypothesis we need to calculate:

$$\mu_R = \frac{2n_1n_2}{n}+1 = \frac{2(16)(17)}{33}+1 = 17.5$$

Thus the null and alternative hypotheses will be:

$$H_0: \mu_R = 17.5$$

$$H_a: \mu_R \neq 17.5$$

The probability of getting the actual sample result of 9 runs, on the assumption that the null hypothesis is true, can be calculated:

$$z_{sample} = \frac{(R+0.5)-\mu_R}{\sqrt{\frac{n^2-2n}{4(n-1)}}} = \frac{9.5-17.5}{\sqrt{\frac{33^2-2(33)}{4(33-1)}}} = -2.83$$

From the table for the area under the standard normal curve, this z-score has a probability of occurring (on a two-tail test) by chance less than 5 times in 1000. Therefore we reject the null hypothesis of randomness, and argue that the pain does occur in 'blocks' of days.

The runs test using SPSS

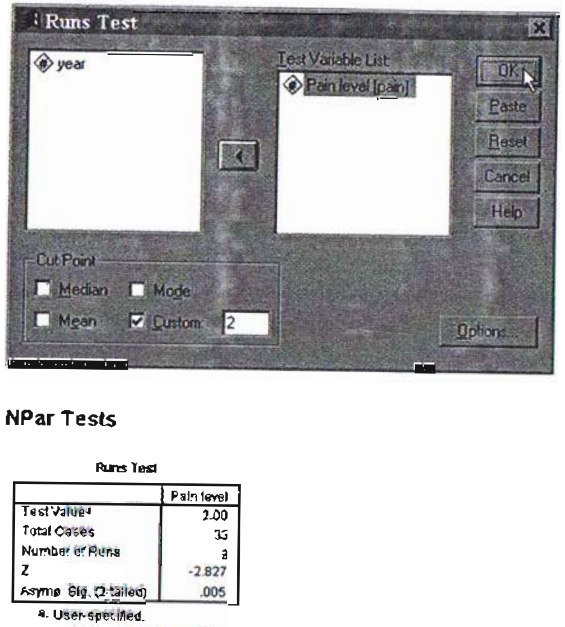
The data from this example have been entered into SPSS and to conduct a runs test we follow the procedure in Table 21.11 and Figure 21.2, which also presents the output from this test. You will notice that SPSS provides a number of **Cut Point** methods for determining the two outcomes that can form a run:

- With categorical data (as is the case here) we use the **Custom** option. Based on the coding scheme for the test variable sequences of values below the cut point will form one run, and sequences equal to or above will form another. With a binomial scale, therefore, we choose the highest of the two values in the coding scheme. Here, with pain coded with '1 = Low pain level' and '2 = High pain level' we enter 2.
- We can also use the **Mode** option if there are more than two categories and we want to assess runs based on whether scores are below the modal category, or equal to or greater than the modal category, according to the values assigned in the coding scheme.
- If we have interval/ratio scales, in addition to using the **Custom** method, we can define the two groups according to whether the scores fall below the median/mean, or equal to or above the median/mean.

This indicates that the Test Value is 2.00, which is the cut point we selected. In effect this is the point that forms the dividing line of a binomial distribution. Provided the data are entered in chronological order, so that the first day on which observations were taken is on the first row of data and so on, SPSS calculates that there are 9 runs. The z-score of −2.827 has a two-tail probability of 0.005, if the null hypothesis of randomness is true. (If we want to convert the two-tail probability into a one-tail probability, we halve its value.) This is so improbable that we reject the null hypothesis.

Table 21.11 The Runs Test on SPSS (file: Ch21-2.sav)

SPSS command/action	Comments
1 Select **Analyze/Nonparametric Tests/Runs**	This brings up the **Runs Test** dialog box
2 Click on **Pain level** in the source variable list	
3 Click on ▸	This pastes **Pain level** into the **Test Variable List:**
4 In the area called **Cut Point** click on the square next to **Custom**	This places ✓ in the check-box to show that the cut point will be specified by the user
5 In the box next to **Custom** we enter 2	This defines the scores that will be identified as forming a run. Cases with values less than the cut point are assigned into one group, and cases with the cut point value or above are assigned into the other group
5 Click on **OK**	



NPar Tests

Runs Test

	Pain level
Test Value[a]	2.00
Total Cases	33
Number of Runs	9
Z	-2.827
Asymp. Sig. (2-tailed)	.005

a. User-specified.

Figure 21.2 SPSS **Runs Test** dialog box and output

Exercises

21.1 In order to estimate the percentage of a population giving a certain response to a survey we need to take a larger sample for larger populations. Is this statement true or false? Why?

21.2 For the following sets of statistics, conduct a z-test of percentages:

(a) $P_u = 52$, $P_s = 61$, $n = 110$
(b) $P_u = 42$, $P_s = 39$, $n = 110$

21.3 A random sample of 900 jail prisoners is surveyed to gauge the success of an in-prison resocialization program. Of the total, 350 stated that the program has been effective in reducing the likelihood of repeat offense. The program's target was a 40 percent success rate in reducing the likelihood of repeat offense.

(a) Using a z-test of percentages, can we say that the program was successful?
(b) Construct a 95 percent confidence interval to estimate the population value. How does this confirm the result of the z-test?

21.4 A survey polls 120 eligible voters the day before an election and 63 state that they will vote for the opposition candidate. This candidate declares that the election is a waste of time since she will clearly win. Is this argument justified? Explain.

21.5 A physiotherapist is interested in whether ankle taping has reduced the incidence of ankle sprains in basketball players. The incidence of ankle sprains in basketball players has been reported to be 8 percent. The physiotherapist randomly selects 360 basketball players who tape their ankles and finds that 11 have sprained their ankles. Does this suggest that taping reduces the incidence of ankle sprain?

21.6 A study of 500 people finds that 56 percent support the decriminalization of marijuana use. What is the 95 percent confidence interval for the percentage of all people in favor of decriminalization? Can we say that a majority of people are in favor of decriminalization?

21.7 A random survey of 60 firms in an industry finds that 12 are not meeting pollution emission control standards. What are the:

(a) 99 percent and
(b) 95 percent confidence intervals for the estimate of all firms in the industry not meeting the standards?

21.8 A hockey team captain has recorded the outcome of 20 coin tosses for the last 20 games. These tosses had the following sequence of results:

heads	tails	heads	heads	tails	tails	tails	heads	heads	heads
tails	heads	heads	tails	tails	heads	tails	tails	tails	tails

(a) Why is a runs test applicable to such data?
(b) Conduct a runs test to see if the outcome of these tosses is random.
(c) Enter these data into SPSS and confirm your results.

21.9 A hospital has kept a tally of the years in which a majority of boys were born and those in which a majority of girls were born. The sequence of results is as follows:

boys	boys	boys	girls	boys	boys	girls	boys	girls
boys	boys	boys	boys	boys	girls	boys	boys	girls
boys	boys	boys	boys	girls				

(a) How many runs describe this sequence?
(b) How many runs will we expect to get if the sex of each child born is purely random with respect to the previous year's outcome?
(c) Can we say that the outcome is a non-random event?
(d) Enter these data on SPSS and conduct a runs test to confirm your own calculations.

21.10 Use the **Employee data** file to determine whether the percentage of employees in the company receiving a current salary of $25,000 or less is not greater than 35 percent (hint: in the **Binomial Test** dialog box use the **Cut point:** option to organize the distribution into a binomial one).

22

One sample tests for a multinomial distribution

The previous chapter discussed the simplest situation for analyzing a frequency distribution, which is the one sample case where the distribution is split into two categories (i.e. a binomial distribution). We hypothesize that the percentage of the population falling into one or the other of the two categories is a specific value and then determine the likelihood of drawing from such a population a sample with the percentage we actually obtain in the course of research. We do this by calculating a z-score and looking up the corresponding probability in the table for areas under the standard normal curve. If the difference between the sample statistic and the hypothesized population percentage is large, the corresponding probability that the sample is drawn from such a population will be low. In short, the question boils down to whether an observed difference between a sample statistic and a hypothesized population value is 'big enough'.

The chi-square goodness-of-fit test

This chapter will extend the analysis of the previous chapter. The previous chapter was interested in a very particular kind of frequency distribution: a scale with only two categories. We often construct a binomial by collapsing categories down into two. But what if responses do not, for example, fall into simple yes/no dichotomies and instead fall into a range of values such as 'strongly agree', 'agree', 'disagree', 'strongly disagree' and we are not prepared to collapse these categories down to two? Such a distribution is called a **multinomial distribution** because it has more than two points on the scale.

Where the research question we are addressing does not direct us to collapse the data down into two categories, but rather directs our attention to the frequency distribution of cases across a wide range of categories or values of a variable, we use the **chi-square goodness-of-fit test** (χ^2 – pronounced 'kigh-square').

The chi-square goodness-of-fit test is a non-parametric test for the multinomial frequency distribution of cases across a range of scores for a single variable.

The nature of the question addressed by the goodness-of-fit test, as opposed to other tests we have encountered, is illustrated in Figure 22.1.

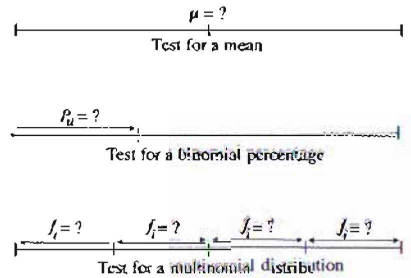


Figure 22.1 Comparison of inference tests

The chi-square goodness-of-fit test analyzes a frequency distribution, which can be constructed for all levels of measurement. We will introduce the goodness-of-fit test as applied to nominal and ordinal data, where data are arranged into discrete categories. We will then show how this test can also be useful in analyzing the frequency distribution of interval/ratio data.

The test is called the 'chi-square' test because the sampling distribution we use to assess the probability of the null being true is a chi-square distribution. (A more detailed explanation of the chi-square distribution is presented in Chapter 23 for the two or more samples case, which is the most common use of the chi-square distribution. It may be helpful to return to the present chapter after reading Chapter 23. The one sample case is presented here to maintain the overall logic of this book, which is to present the one sample test first, before moving to tests for two or more samples.)

The chi-square distribution has the general shape shown in Figure 22.2.

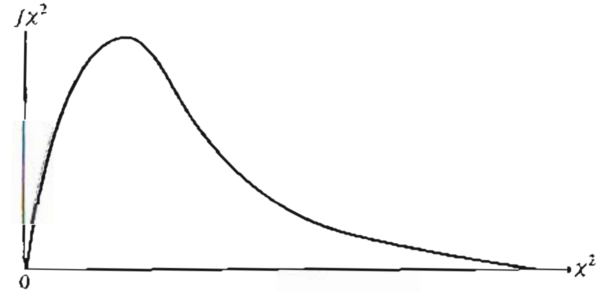


Figure 22.2 Distribution of chi-square (χ^2)

The chi-square distribution is constructed on the same basis as the other sampling distributions we have already encountered: it is the probability distribution of a test statistic we will get from an infinite number of samples of the same size drawn from a population with certain specified features.

To illustrate the goodness-of-fit test we will try to answer the following question: is the crime rate affected by the seasons? Clearly, we are not interested in the average crime rate, but rather the *distribution of crime rates across the range of seasons*. We begin by making an hypothesis about the population distribution: we assume that there is no relationship between crime rates and seasons. On this hypothesis we will *expect* the number of crimes committed in any year to be evenly distributed across the four seasons, where f_e is the expected frequency in each category.

$$f_e = \frac{\text{total number of crimes}}{4}$$

However, in any given year the crime rate might be affected by random events that cause the distribution to be a little bit different from this expected result. In other words, not every sample will conform with this expectation of an exactly equal number of crimes in each season. We can express the difference between the expected value and the observed value by calculating a sample chi-square statistic, where f_e is the expected frequency in each category and f_o is the observed frequency in each category.:

$$\chi^2_{sample} = \sum \frac{(f_o - f_e)^2}{f_e}$$

We can see that if the sample result conforms exactly to the expected result, the value of the sample chi-square (χ^2_{sample}) will be zero: if observed frequencies are the same as expected frequencies then subtracting one from the other will be zero.

What about situations in which the observed distribution is not exactly the same as the expected distribution? Looking at the formula for chi-square we can see that any difference will produce a positive value for the sample chi-square. This is because any difference is squared, thereby eliminating negative values. We can also see that the larger the difference between the observed and expected frequencies, called the **residuals**, the higher the (positive) value of the sample chi-square. The question then becomes at what point does the value of the sample chi-square become so large that it suggests the sample was not selected from a population with a uniform spread of crime rates across seasons?

We follow the same procedure used with other tests. We describe the sample, in this instance by forming a frequency table. We then calculate the test statistic (here it is χ^2_{sample}) and refer to the appropriate table (Table A4) to determine the *p*-score for this test statistic.

For example, if we actually observe the (hypothetical) distribution of crime shown in Table 22.1 can we conclude that crime is indeed affected by the seasons?

Table 22.1 Distribution of crime by season

	Summer	Spring	Winter	Autumn	Total
Observed	300	270	200	250	1020
Expected	255	255	255	255	1020
Residual	45	15	−55	−5	

The expected values are simply the total divided by the number of seasons:

$$f_e = \frac{1020}{4} = 255$$

The row labelled 'Residual' is the difference between the observed and expected values. To get a better picture of the logic behind this test, we have graphed the data in Figure 22.3.

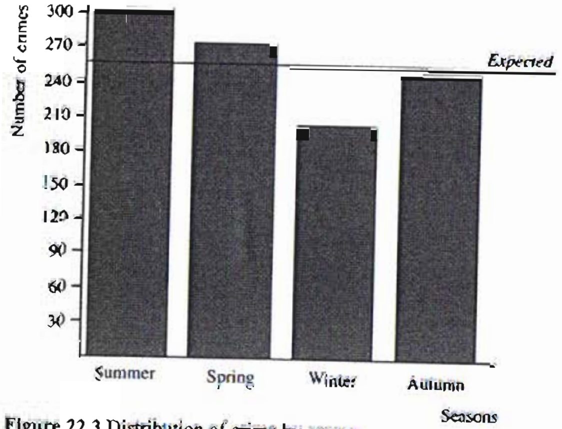


Figure 22.3 Distribution of crime by season

The straight line represents the height that the bars will be if the observed values are equal to the expected values. However, we can see that this is not the case: Summer and Spring have higher than expected values, whereas Winter and Autumn fall short. The gap between the line and each bar is the residual. We then substitute these results into the formula for chi-square:

$$\chi^2_{sample} = \sum\frac{(f_o - f_e)^2}{f_e} = \frac{(30C-255)^2}{255} + \frac{(270-255)^2}{255} + \frac{(20C-255)^2}{255} + \frac{(250-255)^2}{255}$$

$$= 2C.78$$

To find the p-score for a given chi-square value we need to take into account the number of degrees of freedom. For any given distribution the number of degrees of freedom will be:

$$df = k - 1$$

where k is the number of categories. Thus if a variable has four categories, as in this case, the degrees of freedom will be:

$$df = 4 - 1 = 3$$

When we refer to the table for the critical values for chi-square distributions (Appendix Table A4), with 3 degrees of freedom, we see that the highest reported chi-square value is 16.268, which has a significance level of 0.001. Our sample test statistics of 20.78 is larger than this highest reported value, and therefore has a significance level of less than 0.001 (Table 22.2).

Table 22.2 Critical values for chi-square distributions

	Level of significance (α)									
df	0.99	C.90	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.00016	0.0158	0.148	0.455	1.074	1.642	2.706	3.841	6.635	10.827
2	0.0201	0.211	0.713	1.386	2.408	3.219	4.605	5.991	9.210	13.815
3	0.115	0.584	1.424	2.366	3.665	4.642	6.251	7.815	11.341	16.268
4	0.297	1.064	2.195	3.357	4.878	5.989	7.779	9.488	13.277	18.465
5	0.554	1.610	3.000	4.351	6.064	7.289	9.236	11.070	15.086	20.517
6	0.872	2.204	3.828	5.348	7.231	8.558	10.645	12.592	16.812	22.457
7	1.239	2.833	4.671	6.346	8.383	9.803	12.017	14.067	18.475	24.322
8	1.646	3.490	5.527	7.344	9.524	11.030	13.362	15.507	20.090	26.125
9	2.088	4.168	6.393	8.343	10.656	12.242	14.684	16.919	21.666	27.877
10	2.558	4.865	7.267	9.342	11.781	13.442	15.987	18.307	23.209	29.588

The value of the sample chi-square leads us to reject the null hypothesis of an even distribution of crime across seasons.

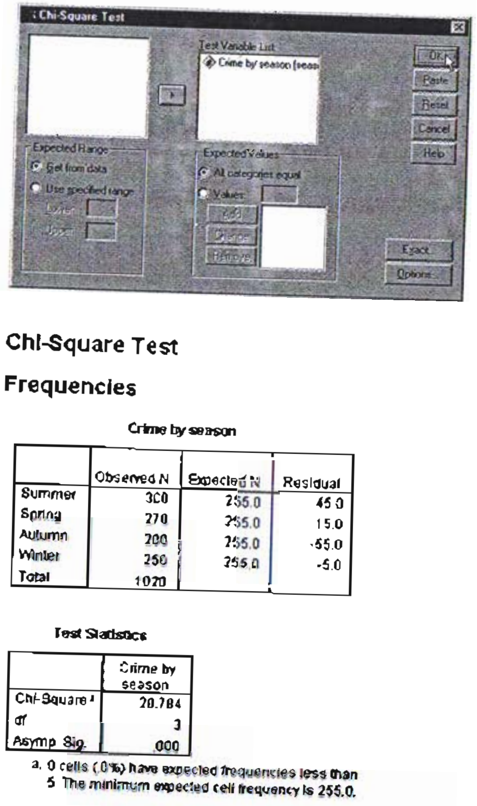
Chi-square goodness-of-fit test using SPSS

The data from this test have been entered into SPSS. This data file comprises a column of 1020 numbers representing the season in which each crime was committed. To conduct a one sample chi-square test on these data we work through the procedure shown in Table 22.3 and Figure 22.4, which also presents the output from this set of instructions.

We can compare these results with the hand calculations above. The table in the output titled **Crime by season** contains the descriptive statistics that summarize the sample. In this case, the distribution of cases across the four seasons is provided in the column headed Observed N. A column of expected frequencies is also generated, based on the assumption that an equal number of cases is expected in each season. The values in the **Expected N** column are subtracted from the **Observed N** column to give the **Residual** values. This is simply a replication of the frequency table we used above, but turned 'on its side' so that the seasons are down the left of the table rather than across the top.

Table 22.3 Chi-square goodness-of-fit test using SPSS (file: Ch22.sav)

SPSS command/action	Comments
1 From the menu select **Analyze/Nonparametric Tests/ Chi –Square**	This brings up the **Chi Square** dialog box
2 Click on **Crime by season** in the source list	This highlights **Crime by season**
3 Click on ▸	This pastes **Crime by season** into the **Test Variable List:**
4 Click on **OK**	



Chi-Square Test

Frequencies

Crime by season

	Observed N	Expected N	Residual
Summer	300	255.0	45.0
Spring	270	255.0	15.0
Autumn	200	255.0	-55.0
Winter	250	255.0	-5.0
Total	1020		

Test Statistics

	Crime by season
Chi-Square [a]	20.784
df	3
Asymp. Sig.	.000

a. 0 cells (.0%) have expected frequencies less than 5. The minimum expected cell frequency is 255.0.

Figure 22.4 The Chi-Square Test dialog box and output

Below **Crime by season** the frequency table is the chi-square **Test Statistics** table. The sample chi-square is 20.784 (as we calculated above), which, with 3 degrees of freedom (df), has a probability of occurring if crime is evenly spread across seasons of less than 5 in every 10,000 samples (SPSS has rounded this to .000). Such a low probability leads us to reject the null hypothesis: crime rates do seem to be related to the seasons.

Notice in the Chi-Square Test dialog box, in the area called **Expected Values** that the radio button next to **All Categories equal** is selected. This is the default setting: SPSS will automatically calculate the number of expected cases in each category by dividing the total by the number of categories, which is what we desired in this example. The categories of the variables, however, do not need to have exactly equal numbers of expected cases. The chi-square test can be used for situations where, for some *a priori* reason, we hypothesize some unequal distribution of cases in the population. To enter user-specified expected values we select **Values:**, enter a value greater than 0 for each category of the test variable, and click on **Add**. Each time an expected value is added, it appears at the bottom of the value list. The order of the values is important: it corresponds to the ascending order of the category values of the test variable. The first value in the list corresponds to the category with the lowest value of the variable, and the last expected frequency added corresponds to the highest value.

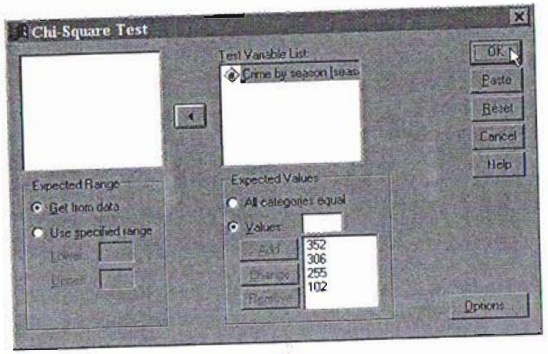
For example, assume that a region adjacent to the one we are investigating has the distribution of crime across seasons presented in the first column of Table 22.4.

Table 22.4 Distribution of crime by season

		Expected number
Season	Expected %	352
Summer	35%	306
Spring	30%	255
Autumn	25%	102
Winter	10%	1020
Total	100%	

We know, in other words, what the distribution of *all* crimes across the seasons is for this nearby region. Can we say that our region has the same distribution of crime? Given this research question we calculate the expected values on the basis of these percentages, producing the expected number of crimes in each season given in the second column. To conduct the chi-square test on SPSS using these expected values we first need to note the values given to each season in the coding scheme, which is 1 for Summer, 2 for Spring, 3 for Autumn, and 4 for Winter. We begin with the category with the lowest value, which is Summer (Figure 22.5).

- We click on **Values:** and type 352, which is the expected number of crimes for Summer.
- We then click on the **Add** button so that it appears in the list of expected frequencies.
- We then type 306, which is the expected frequency for Spring, and click on the **Add** button, and so on for each of the seasons.



Expected Range: ● Get from data ○ Use specified range

Expected Values: ○ All categories equal ● Values:

352
306
255
102

Figure 22.5 Entering expected values

Example

In 1996, 40 percent of sales by a car dealer were four-cylinder cars, 30 percent were six-cylinder, and 30 percent were eight-cylinder. A random sample of sales in recent months produced the distribution shown in Table 22.5.

Table 22.5 Observed sales distribution

Engine type	Observed number of sales
Four-cylinder	42
Six-cylinder	26
Eight-cylinder	12
Total	80

Can we say that this reflects a trend toward smaller cars? First we calculate the expected number of sales, based on the 1996 percentages (Table 22.6).

Table 22.6 Expected sales distribution

Engine type	Expected number of sales
Four-cylinder	$\frac{40}{100} \times 80 = 32$
Six-cylinder	$\frac{30}{100} \times 80 = 24$
Eight-cylinder	$\frac{30}{100} \times 80 = 24$
Total	80

Notice here that we are not expecting an *even* spread of cases across the categories, as we did in the example above regarding crime rates across seasons. This does not alter the test: our decision as to the frequencies to be expected in each category is determined primarily by our research question and the theory that informs it. Given these expected values we then conduct the chi-square test.

Substituting observed and expected frequencies into the formula for chi-square we get 9.29:

$$\chi^2_{sample} = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(42-32)^2}{32} + \frac{(26-24)^2}{24} + \frac{(12-24)^2}{24} = 9.29$$

At two degrees of freedom, we find from the table for critical values of the chi-square distribution that the *p*-score is between 0.01 and 0.001 (Table 22.7)

Table 22.7 Critical values for chi-square distributions

	Level of significance (α)									
df	0.99	0.90	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.00016	0.0158	0.148	0.455	1.074	1.642	2.706	3.841	6.635	10.827
2	0.0201	0.211	0.713	1.386	2.408	3.219	4.605	5.991	9.210	13.815
3	0.115	0.584	1.424	2.366	3.665	4.642	6.251	7.815	11.341	16.268
4	0.297	1.064	2.195	3.357	4.878	5.989	7.779	9.488	13.277	18.465
5	0.554	1.610	3.000	4.351	6.064	7.289	9.236	11.070	15.086	20.517
6	0.872	2.204	3.828	5.348	7.231	8.558	10.645	12.592	16.812	22.457
7	1.239	2.833	4.671	6.346	8.383	9.803	12.017	14.067	18.475	24.322
8	1.646	3.490	5.527	7.344	9.524	11.030	13.362	15.507	20.090	26.125
9	2.088	4.168	6.393	8.343	10.656	12.242	14.684	16.919	21.666	27.877
10	2.558	4.865	7.267	9.342	11.781	13.442	15.987	18.307	23.209	29.588

Looking at the relative frequencies we can see that there is a trend toward smaller cars, and the frequency distribution of car sales by engine size is significantly different (in a statistical sense) to the distribution of engine size in 1996 at the 0.01 level. The change in the distribution of sales suggests a change in the way the car dealer goes about doing business, given the extremely low probability of making a type 1 error: the pattern of car sales does seem to have changed.

The chi-square goodness-of-fit test for normality

We have worked through examples of the chi-square goodness-of-fit test on nominal and ordinal data that fall into discrete categories. Any test that can be applied to nominal and ordinal data, though, can also be applied to the higher levels of measurement of interval/ratio. In the case of interval/ratio data we look at the frequency distribution of cases across the range of values or class intervals, in exactly the same way as when we looked at the distribution across discrete categories.

This logic makes the goodness-of-fit test particularly useful in assessing whether interval/ratio data come from a normal population. In this way, this non-parametric test can be a useful preliminary and complement to parametric tests, which require the assumption that a sample comes from a normal population.

Suppose that we have a sample and we want to assess whether it was drawn from a normal population. Remember from Chapter 11 that a normal distribution is defined by the frequency distribution shown in Table 22.8 and illustrated in Figure 22.6. We can use the percentage values in Table 22.8 to calculate the expected values we use in the formula for chi-square. Notice that unlike the previous example of crime rates, we are not assuming cases are evenly distributed across the categories; instead the expected frequencies are based on the characteristics of the normal curve.

Table 22.8 Distribution of the normal curve

Range of values	Percentage of cases
Further than 2 standard deviations below the mean	2%
Between 1 and 2 standard deviations below the mean	14%
Within 1 standard deviation below the mean	34%
Within 1 standard deviation above the mean	34%
Between 1 and 2 standard deviations above the mean	14%
Further than 2 standard deviations above the mean	2%

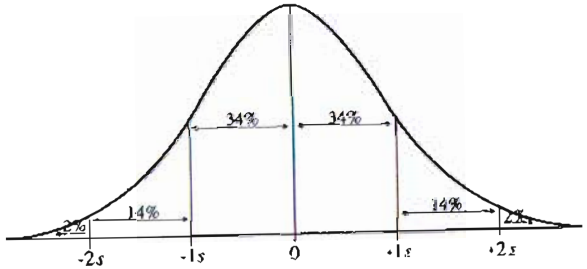


Figure 22.6 Areas under the normal curve

For example, assume that we have a sample of 110 people whose mean age is 45 years, with a standard deviation of 10 years. If this sample is normally distributed we will expect to find the numbers of people within the ranges shown in Table 22.9. The table includes the calculations for the first range to show the method involved.

Table 22.9 Expected distribution of the sample ($n = 110$)

Range of values	Percentage of cases	Number of cases
25 years or less (further than 2 standard deviations below the mean)	2%	$\frac{2}{100} \times 110 = 2.2$
26–35 years (between 1 and 2 standard deviations below the mean)	14%	15.4
36–45 years (within 1 standard deviation below the mean)	34%	37.4
46–55 years (within 1 standard deviation above the mean)	34%	37 4
56–65 years (between 1 and 2 standard deviations above the mean)	14%	15 4
66 years or over (further than 2 standard deviations above the mean)	2%	2 2

However, we might actually get a sample distribution as shown in Table 22.10.

Table 22.10 Observed distribution of the sample

Range of values	Number of cases
25 years or less	5
26–35 years	17
36–45 years	33
46–55 years	33
56–65 years	17
66 years or over	5

There is obviously some difference between the observed and expected values: should this cause us to reject the hypothesis that the population is normally distributed? To answer this we need to calculate chi-square:

$$\chi^2_{sample} = \sum \frac{(f_o - f_e)^2}{f_e}$$

$$= \frac{(5-2\,2)^2}{2.2} + \frac{(17-15.4)^2}{15.4} + \frac{(33-37.4)^2}{37.4} + \frac{(33-37.4)^2}{37.4} + \frac{(17-15.4)^2}{15.4} + \frac{(5-2.2)^2}{2\,2}$$

$$= 8.5$$

The significance level of this chi-square value, with $df = 6 - 1 = 5$, lies between 0.20 and 0.10 (Table 22.11).

Table 22.11 Critical values for chi-square distributions

	Level of significance (α)									
df	0.99	0.90	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.00016	0.0158	0.148	0.455	1.074	1.642	2.706	3.841	6.635	10.827
2	0.0201	0.211	0.713	1.386	2.408	3.219	4.605	5.991	9.210	13.815
3	0.115	0.584	1.424	2.366	3.665	4.642	6.251	7.815	11.341	16.268
4	0.297	1.064	2.195	3.357	4.878	5.989	7.779	9.488	13.277	18.465
5	0.554	1.610	3.000	4.351	6.064	7.289	9.236	11.070	15.086	20.517
6	0.872	2.204	3.828	5.348	7.231	8.558	10.645	12.592	16.812	22.457
7	1.239	2.833	4.671	6.346	8.383	9.803	12.017	14.067	18.475	24.322
8	1.646	3.490	5.527	7.344	9.524	11.030	13.362	15.507	20.090	26.125
9	2.088	4.168	6.393	8.343	10.656	12.242	14.684	16.919	21.666	27.877
10	2.558	4.865	7.267	9.342	11.781	13.442	15.987	18.307	23.209	29.588

The high probability of obtaining these differences between the observed and expected frequencies through sampling error means we do not reject the hypothesis that the sample comes from a normally distributed population.

Summary

We have introduced a new test in this chapter: the chi-square test. Although the chi-square test involves slightly different calculations, it is very similar to the binomial test we looked at in the previous chapter. Whereas the z-test for a binomial percentage only applies to frequency distributions organized into a binomial distribution, the chi-square test is more general in that it applies to frequency distributions with any number of categories (thus the z-test for a binomial percentage can be considered a special case of the chi-square test). This gives the test a wide applicability, especially since (as we will see in Chapter 23), it can be extended in a direct way to the two sample and more than two sample situations.

Exercises

22.1 What will be the number of degrees of freedom, and the value of $\chi^2_{critical}$ at $\alpha = 0.10$ and $\alpha = 0.05$, for a goodness-of-fit test on a variable with:

(a) three categories
(b) five categories
(c) eight categories

22.2 Conduct a goodness-of-fit test on the following data to test the hypothesis that the sample comes from a population with an equal proportion of cases across all categories:

(a)

Value	Number of cases
1	45
2	40
3	55
4	54
5	38

(b)

Value	Number of cases
1	120
2	111
3	119
4	125
5	120
6	127
7	118

22.3 According to a 1991 *Census of Population and Housing*, Australians between the ages of 25 and 34 years had the following distribution according to marital status:

Marital status	Number of persons
Never married	896,206
Married	1,591,010
Separated not divorced	104,296
Divorced	117,673
Widowed	14,216
Total	2,723,401

A survey of 350 residents aged between 25 and 34 is taken in a local area, which had the following distribution according to marital status:

Marital status	% of Sample ($n = 350$)
Never married	40
Married	50
Separated not divorced	6
Divorced	2
Widowed	2
Total	100

Using the census information to calculate the expected values, can we say that this area is significantly different from the rest of the population? In which direction are the differences?

22.4 Ninety people are surveyed and the amount of time they each spend reading each day is measured. The researcher wants to test the assumption that this sample comes from a normal population. The mean for the sample is 45 minutes, with a standard deviation of 15 minutes. The observed distribution of the sample across the following ranges of values is:

Range of values	Number of cases
less than 16 minutes	3
16–30 minutes	15
31–45 minutes	34
46–60 minutes	31
61–75 minutes	5
over 75 minutes	2

Using an alpha level of 0.05, test the assumption of normality for the population. Enter these data into SPSS and conduct the goodness-of-fit test.

22.5 Five schools are compared in terms of the proportion of students that proceed to university. A sample of 50 students who graduated from each school is taken and the number of those who entered university from each school are:

School	Number entering university
School 1	22
School 2	25
School 3	26
School 4	28
School 5	33

(a) Calculate the expected values and then conduct a chi-square goodness of fit test.
(b) What do you conclude about the prospects of entering university from each of the schools?
(c) Enter these data into SPSS and compare the results with your hand calculations.

22.6 Use the **Employee data** file to assess whether the sample data for the company indicates that its employment structure is 'top heavy'. This can be tested by assessing whether there are proportionately more employees in the Manager category than for similar companies. Assume that official data indicate that for similar firms the proportion of cases in each of the employment categories is Clerical 82 percent, Custodial 8 percent, and Managerial 10 percent.

(a) Calculate the expected number of employees in the sample for each employment category, on the assumption that this firm is no different to all others.
(b) Use these expected frequencies to test this assumption on SPSS.

23

The chi-square test for independence

This chapter will look at the technique for conducting an hypothesis test for categorical data arranged in a crosstabulation. This is the **chi-square test for independence**, which is similar to the one-sample test we have already encountered in the previous chapter. It extends the logic of the goodness of fit test to situations where we are assessing whether there is a relationship between two variables arranged in a crosstabulation, and thus is the inferential statistics counterpart to the descriptive statistics we presented in Chapter 5. To understand the place of the chi-square test as one choice in the 'menu' of inference tests available to us it is helpful to review the general criteria for choosing an inference test.

The chi-square test and other tests of significance

The earlier chapters emphasized that the choice of inference test is determined by two main considerations:

1. *The descriptive statistic used to describe the raw data.* This factor is itself usually determined by the research question we want to answer. The research question almost invariably directs our interest to a specific characteristic of the distribution for a given variable. A public health research worker might be concerned with the question of whether a population is on average 'young' or 'old', a research problem that directs one to look at the central tendency of the variable. A political scientist may also be concerned with the age distribution of this population, but the specific interest may be the relative number of people that are above voting age. For this research problem the political scientist will organize the data into a binomial distribution and calculate the proportion of the sample above and below the voting age. Both researchers are interested in the same population, and both have exactly the same raw data in front of them, but their respective research questions decide whether they are interested in the central tendency of the distribution, or the proportion of cases above or below a certain point on the scale
2. *The number of samples to be compared.* We have seen that when we collect data from only one sample we have a certain range of inference tests to choose from. The range of choices is different when we have two samples and therefore need to make an inference about each of the two populations from which the samples are drawn. Similarly, with more than two samples we are then confronted with another range of tests to choose from. For example, when comparing means, a *t*-test for sample means is used with one or two samples, whereas ANOVA is used for more than two samples.

With this discussion in mind, we can now look at the conditions under which the chi-square test is appropriate.

1. *The descriptive statistic upon which the chi-square test for independence is conducted is the frequency distribution contained in a bivariate table.* We investigated the construction and use of bivariate tables that crosstabulate data on two variables in Chapter 5. We saw that crosstabs are a convenient way of summarizing and displaying categorical data when we are interested in the overall frequency distribution of cases across the whole range of categories, rather than just the central tendency. Nominal and ordinal data come 'prepackaged' in categories, and hence crosstabs are a very common way of describing such data (although even in these instances we sometimes need to recode the categories into a smaller number). It should also be remembered, however, that interval/ratio data can be collapsed down into discrete categories, as we do when we organize people's dollar incomes into clusters such as 'low', 'middle', and 'high income' groups. Hence, a crosstab can also potentially be a means of describing data collected on an interval/ratio scale, as well as on nominal and ordinal scales.
2. *The chi-square test is basically the same, regardless of whether we have one, two, or more than two samples.* We have already encountered the chi-square test as a one-sample test for a frequency distribution. Unlike other tests, the chi-square test can be extended to the two samples and more than two samples cases without much modification: we follow the same basic procedure, and use the same formula, regardless of the number of samples being compared (although in the one-sample case it is called a goodness-of-fit-test, whereas with two or more samples it is called a test for independence).

Statistical Independence

We construct crosstabulations to get a visual sense of whether the two variables under investigation are **independent** of each other.

Two variables are statistically **independent** if the classification of cases in terms of one variable is not related to the classification of those cases in terms of the other variable.

Take the example we used in Chapter 5 to construct a crosstab between the sex of students and how they rate their own health (Table 23.1).

Table 23.1 Health rating by sex of students

Health rating	Sex		
	Female	Male	Total
Unhealthy	34 43%	16 16%	50 28%
Healthy	29 36%	27 28%	56 32%
Very healthy	17 21%	54 56%	71 40%
Total	80 100%	97 100%	177 100%

We make a visual, or 'eye ball', inspection of the relative frequencies in each cell of the table and assess whether *in the sample* the two variables are independent or whether in fact some kind of a relationship exists. We observe in the table that there is some relationship between these two variables: males tend to rate their own health more highly than females.

However, our conclusion is based on sample data, and we must therefore be wary that it may be due to sampling error when drawing from populations in which there is no relationship between a student's sex and the way they rate their own health. The chi-square test for independence assesses this possibility.

The chi-square test for independence

The starting point for conducting a chi-square test for independence, as with all inference tests, is the statement of the null and alternative hypotheses. In the example we are using, the hypotheses take the form of:

H_0: Sex of students and health rating are independent of each other
H_a: Sex of students and health rating are not independent of each other

The statement of independence forms the null hypothesis for the test, and if the null is rejected, we conclude that the two variables are not independent. Conversely if we do not reject the null hypothesis we argue that the variables are independent, even though dependence is observed in the samples.

Looking at our actual example, we have determined that the variables are not independent in the sample – there does appear to be some relationship – but can we draw this inference about the populations from which the samples came?

To see how the chi-square test helps us assess whether these two variables are truly independent of each other, even where there is dependence in the samples, we begin by looking at the *frequencies for the row totals* in Table 23.1 (Table 23.2).

Table 23.2 Health rating: all students sampled

Health rating	Total	Percentage
Unhealthy	50	28%
Healthy	56	32%
Very healthy	71	40%
Total	177	100%

These row totals and percentages are the basic reference points from which the chi-square test is conducted. The argument is that if 28 percent of *all* respondents rate themselves as unhealthy, then we should expect 28 percent of *each group* (males and females) to also rate themselves as unhealthy, if the two variables are independent.

> Under the null hypothesis of independence the relative frequencies for each group are expected to be the same as that for the groups combined.

In other words, we expect to find in each cell of the table, if the two variables are independent, the relative frequencies in Table 23.3.

Table 23.3 Expected relative cell frequencies

Health rating	Sex of student		
	Female	Male	Total
Unhealthy	28%	28%	28%
Healthy	32%	32%	32%
Very healthy	40%	40%	40%
Total	100%	100%	100%

However, even if the null hypothesis of independence is true, we should not always expect random samples of females and males to reflect this. For example, we might occasionally draw samples of female and male students and get one of the three separate results shown in Table 23.4.

Table 23.4(a) represents a situation in which the observed percentages very closely reflect the expected percentages, assuming that the two variables are independent. Occasionally we might find the situation shown in Table 23.4(b), where there is a greater variation between the groups, but it is not too great. Table 23.4(c) shows an extreme situation in which we happened to pick up cases from either end of the scale, causing the relative frequencies in the first two columns to diverge a great deal from those in the Total column. Although this is a possibility when randomly sampling from populations where there is no relationship, it is also highly unlikely.

In fact, we can take an infinite number of random samples from populations where the two variables are independent and observe the spread of results. Obviously most would be like Table 23.4(a), some like Table 23.4(b), and very few like Table 23.4(c).

Table 23.4(a)

Health rating	Sex of student		
	Female	Male	Total
Unhealthy	29%	26%	28%
Healthy	33%	32%	32%
Very healthy	38%	42%	40%
Total	100%	100%	100%

Table 23.4(b)

Health rating	Sex of student		
	Female	Male	Total
Unhealthy	21%	35%	28%
Healthy	35%	30%	32%
Very healthy	44%	35%	40%
Total	100%	100%	100%

Table 23.4(c)

Health rating	Sex of student		
	Female	Male	Total
Unhealthy	15%	45%	28%
Healthy	25%	25%	32%
Very healthy	60%	30%	40%
Total	100%	100%	100%

The **chi-square statistic** is a means by which we can capture this difference between observed and expected frequencies.

> The **chi-square statistic** is calculated from the difference between the observed and expected frequencies in each cell of a bivariate table.

> The **chi-square distribution** is the probability distribution of the chi-square statistic for an infinite number of random samples of the same size drawn from populations where the two variables are independent of each other.

The exact formula for calculating chi-square, where f_o is observed cell frequencies and f_e is expected cell frequencies, is:

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

Occasionally we draw samples that are 'true' to the population so that there is no difference between the actual and expected frequencies. In other words, we get cell frequencies like those in Table 23.3. In this case the value of chi-square will be zero:

$$f_o = f_e \rightarrow (f_o - f_e)^2 = 0 \rightarrow \chi^2 = 0$$

This will not be the case for every sample. We will occasionally take samples that, through random chance, do not fully reflect the populations from which they are drawn. The result is that chi-square will take on a positive value:

$$f_o \neq f_e \rightarrow (f_o - f_e)^2 > 0 \rightarrow \chi^2 > 0$$

The greater the difference between the observed frequencies and the expected frequencies, the larger the value of chi-square. In the formula for chi-square, notice that differences between observed and expected frequencies are squared. This ensures that the range of all possible chi-square values must start at zero and increase in a positive direction. Regardless of whether the expected frequency is larger than the observed frequency or vice versa, squaring any difference will produce a positive number. (Since chi-square is calculated on the basis of the difference between expected and actual scores *squared*, and not on the *direction* of difference, there is no sense in which we have to choose between a one-tail or two-tail inference test. All differences between observed and expected scores, regardless of whether they are due to the observed scores being above or below the expected scores, will take on a positive value.)

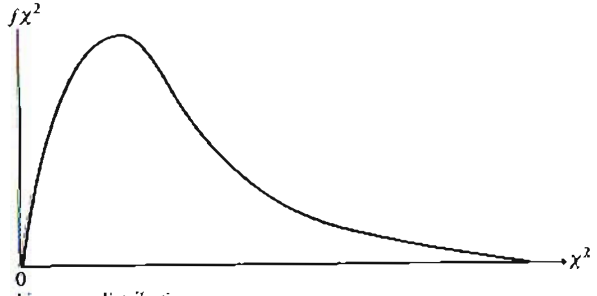


Figure 23.1 The chi-square distribution

The chi-square distribution has a long tail (Figure 23.1), reflecting the fact that it is possible to select random samples that yield a very high value for chi-square, even though the variables are independent, but this is highly improbable. It will be a fluke just to happen to select a sample from one group in which all cases come from one end of the distribution and another sample from the other group that comes from the other end of the distribution, if the null hypothesis of independence is true. Therefore the area under the curve for very large chi-square values is small, reflecting the low probability of this happening by chance.

From the sampling distribution of chi-square we can determine the probability that the difference between observed and expected scores is due to random variation when sampling from populations in which the two variables are independent.

For example, we might find that the samples in Table 23.4(c) above will be drawn only one time in a thousand ($p = 0.001$) if the two variables are independent of each other. This will be considered so unlikely as to warrant us to argue that our assumption about independence should be dropped – there really is a relationship between students' sex and health rating.

We will now use the actual data for the example of students' sex and health rating to provide a concrete illustration of this procedure. The (hypothetical) survey, you will recall, consists of 177 students made up of 80 females and 97 males, with the distribution in terms of health according to that in Table 23.1 above.

The first number in each cell is the actual count of females and males who give themselves a particular health rating. The percentage figure is the number of students in that cell as a percentage of the column total. That is, 43 percent of all females surveyed rate themselves as unhealthy, which is 34 females. On the other hand, only 16 percent of all males rate their own health this poorly.

Another way to visualize the results in Table 23.1 is with a stacked bar graph, which I have generated on SPSS (Figure 23.2).

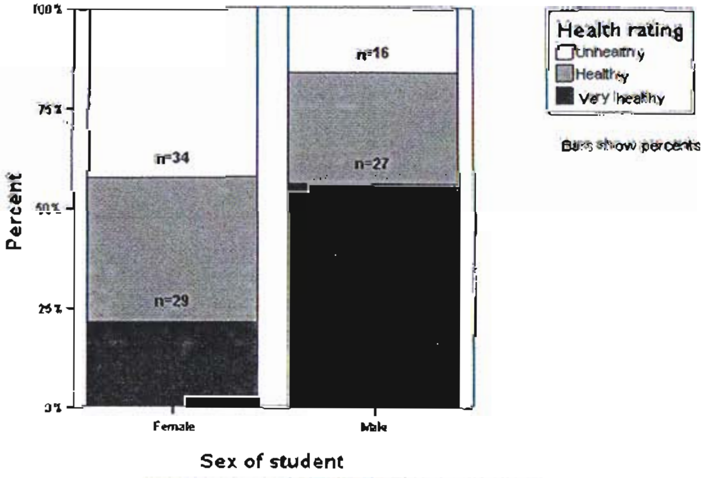


Figure 23.2 A stacked bar chart of health rating by sex of students

We can see that there is obviously a difference between females and males in terms of how they rate their own health, but could this be due to sampling error? To answer this we need to calculate the expected frequencies: the numbers we expect to find in each cell *if* the two variables are independent. These expected values are obtained by taking the percentages in the Total column for each row and applying them to each sex, as illustrated in Table 23.5.

Table 23.5 Sex of students by health rating. Expected frequencies

Health rating	Sex of student		
	Female	Male	Total
Unhealthy	$\frac{28}{100} \times 80 = 22.4$	$\frac{28}{100} \times 97 = 27.2$	28%
Healthy	$\frac{32}{100} \times 80 = 25.6$	$\frac{32}{100} \times 97 = 31$	32%
Very healthy	$\frac{40}{100} \times 80 = 32$	$\frac{40}{100} \times 97 = 38.8$	40%
Total	80	97	100%

The number of respondents we expect to find in each cell, if the variables are independent, is calculated for each cell. For example, if 28 percent of *all* respondents rate themselves as unhealthy, then we expect to find 28 percent of females rating their health this way. There are 80 female students in total, and 28 percent of 80 gives us 22.4 females *expected* to rate themselves as unhealthy. Effectively we are calculating the numbers we would need so that *exactly the same percentage* of females and males give each health rating.

Thus we have two numbers for each cell in the table, one for the expected frequencies based on the null hypothesis, and the other is the actual observed frequencies we obtain from our samples. We show both of these sets of numbers in Table 23.6, with the expected frequencies in brackets. The Totals row and column have been omitted so that we can focus on the values in the cells of the table alone. We can see that in each cell of Table 23.6 there is a difference between the observed and expected frequencies, and we can use the formula for chi-square to express this difference in a single number.

Table 23.6 Sex of students by health rating: Observed and expected frequencies

Health rating	Sex	
	Female	Male
Unhealthy	34 (22.4)	16 (27.2)
Healthy	29 (25.6)	27 (31)
Very healthy	17 (32)	54 (38.8)

Table 23.7 illustrates how we go through the mechanics of calculating the chi-square statistic from these differences.

Table 23.7 Calculations for chi-square

Health rating	Sex	
	Female	Male
Unhealthy	$\chi^2 = \frac{(34-22.4)^2}{22.4} = 6$	$\chi^2 = \frac{(16-27.2)^2}{27.2} = 4.6$
Healthy	$\chi^2 = \frac{(29-25.6)^2}{25.6} = 0.5$	$\chi^2 = \frac{(27-31)^2}{31} = 0.5$
Very healthy	$\chi^2 = \frac{(17-32)^2}{32} = 7$	$\chi^2 = \frac{(54-38.8)^2}{38.8} = 6$

Having calculated these values for each cell we can add them together to get an overall χ^2 for the crosstab as a whole. In other words, the chi-square statistic gathers up these individual values so that we get a single number for the whole table that expresses the fact that the actual sample result does not conform perfectly to the null hypothesis of independence:

$$\chi^2_{sample} = \sum \frac{(f_o - f_e)^2}{f_e} = 6 + 4.6 + 0.5 + 0.5 + 7 + 6$$

$$= 24.6$$

The distribution of chi-square

So we have obtained a value for chi-square of 24.6. What does this tell us? In and of itself it does not tell us a great deal, apart from the fact that it is not equal to zero and therefore indicates that there is some dependence between these variables in the sample data. Whether this should cause us to reject the null hypothesis of independence depends on the probability of obtaining this sample chi-square value of 24.6 from populations where the two variables are independent. To determine this probability we refer to the table for the critical values of chi-square printed as Table A4, and reproduced in part in Table 23.8.

In using this table to work out the probability of obtaining a sample chi-square of 24.6 just by random chance, we need to take into account the degrees of freedom. For any table the number of degrees of freedom will be:

$$df = (r - 1)(c - 1)$$

where r is the number of rows and c is the number of columns. In a 3-by-2 table such as this, therefore, there are 2 degree of freedom.

We can now refer to the table for the critical values of chi-square and determine the relevant probability. To illustrate how this is done a portion of the table is reproduced in Table 23.8.

Table 23.8 Critical values of chi-square

	Level of significance									
df	0.99	0.90	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.00016	0.0158	0.148	0.455	1.074	1.642	2.706	3.841	6.635	10.827
2	0.0201	0.211	0.713	1.386	2.408	3.219	4.605	5.991	9.210	13.815
3	0.115	0.584	1.424	2.366	3.665	4.642	6.251	7.815	11.341	16.268
4	0.297	1.064	2.195	3.357	4.878	5.989	7.779	9.488	13.277	18.465
5	0.554	1.610	3.000	4.351	6.064	7.289	9.236	11.070	15.086	20.517
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
30	14.953	20.599	25.508	29.336	33.530	36.250	40.256	43.773	50.892	59.703

This table is very similar to that for the distribution of t, with the critical values for chi-square in the body of the table, the number of degrees of freedom down the side, and a select set of significance levels across the top. In our example, with 2 degrees of freedom, the sample value of 24.6 lies further out than the largest value presented in the table, which is for a significance level of 0.001. This means the sample p-score is less than 0.001, so we reject the null hypothesis of independence: sex of student and health rating are not independent.

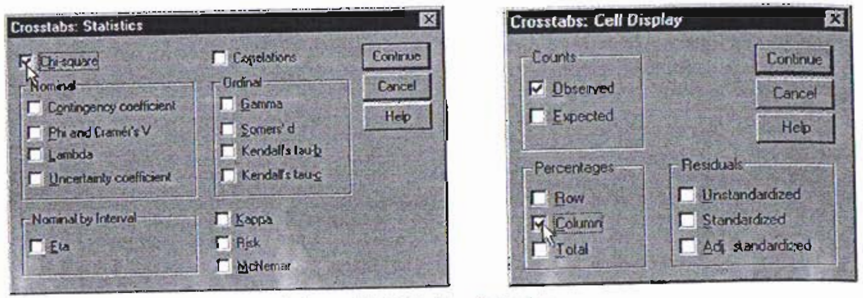
It is of the utmost importance to note, however, that the test itself does not tell us what is the nature of the relationship. All we conclude is that there is some association between these variables. In this instance it is obvious that there must be a one-way relationship from sex of student to health rating. In other instances, the appropriate model of the relationship may be open to debate. The chi-square test will not decide this issue for us. It merely tells us whether the variables are independent. How we choose to characterize any relationship observed is a matter for theoretical debate that statistical analysis can inform, but never decide.

The chi-square test using SPSS

In Chapter 5 we introduced the commands for generating a crosstab on these data in SPSS. The chi-square test appears as an option within the procedure for generating a crosstab, much like the way in which we added lambda to the crosstab in Chapter 6. Table 23.9 and Figure 23.3 repeat the steps for generating a crosstab in SPSS, but with the addition of the relevant chi-square statistic, and also column percentages.

Table 23.9 Generating crosstabs with chi-square on SPSS (file: Ch23.sav)

SPSS command/action	Comments
1 From the menu select **Analyze/ Descriptive Statistics/ Crosstabs**	This brings up the **Crosstabs** dialog box
2 Click on the variable in the source list that will form the rows of the table, in this case **Health rating**	This highlights **Health rating**
3 Click on ▸ that points to the target list headed **Row(s):**	This pastes **Health rating** into the **Row(s):** target list
4 Click on the variable in the source list that will form the columns of the table, in this case **Sex of students**	This highlights **Sex of students**
5 Click on ▸ that points to the target list headed **Column(s):**	This pastes **Sex of students** into the **Column(s):** target list
6 Click on the **Statistics** button	This brings up the **Crosstabs: Statistics** box. In the top-left corner notice **Chi-square** with a tick-box next to it
7 Select **Chi-square** by clicking on the box next to it	This places ✓ in the tick-box to show that it is selected
8 Click on the **Cells** button	This brings up the **Crosstabs: Cell Display** box
9 Select **Column** by clicking on the box next to it	This places ✓ in the tick-box to show that it is selected
10 Click on **Continue**	
11 Click on **OK**	



Crosstabs: Statistics

☑ Chi-square □ Correlations

Nominal: □ Contingency coefficient □ Phi and Cramér's V □ Lambda □ Uncertainty coefficient

Ordinal: □ Gamma □ Somers' d □ Kendall's tau-b □ Kendall's tau-c

Nominal by Interval: □ Eta

□ Kappa □ Risk □ McNemar

Crosstabs: Cell Display

Counts: ☑ Observed □ Expected

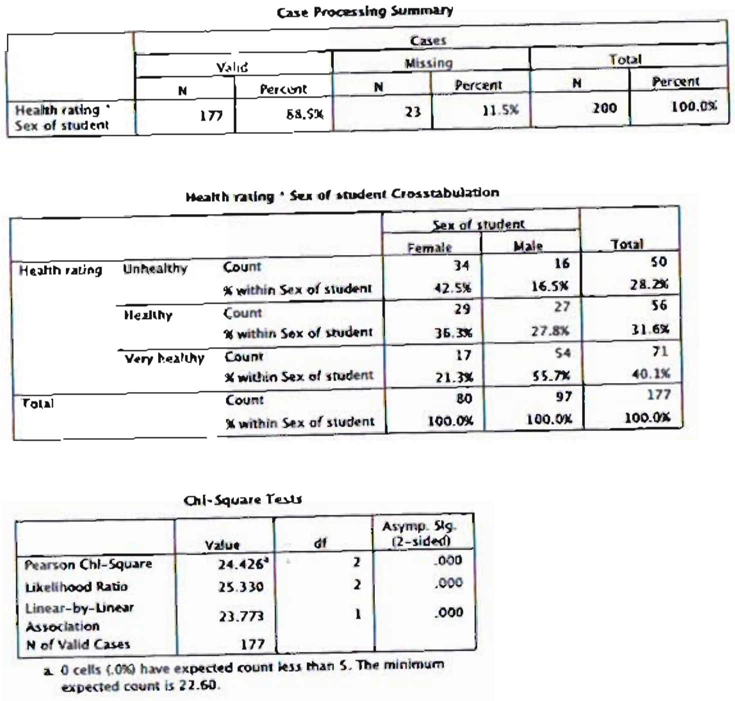
Percentages: □ Row ☑ Column □ Total

Residuals: □ Unstandardized □ Standardized □ Adj. standardized

Figure 23.3 The Crosstabs: Statistics and Cell Display dialog box

This set of commands will produce the necessary information for conducting a chi-square test. Notice, I instructed SPSS to also include the relative column frequencies in the crosstab. This is done by clicking on the Cell button in the **Crosstabs** dialog box, which provides a range of options for information to be printed in each cell of the table. By clicking on the check-box next to **Column** we instruct SPSS also to include the column percentages in the output. While the choice of information to be calculated and printed in each cell is really up to the person conducting the research, and what they think is needed to make a reasonable eyeball assessment, the column percentages allow us to look at the data and make a preliminary judgment as to whether we think the two variables are independent or not.

The output generated from these commands will be as shown in Figure 23.4.



Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Health rating * Sex of student	177	88.5%	23	11.5%	200	100.0%

Health rating * Sex of student Crosstabulation

			Sex of student		Total
			Female	Male	
Health rating	Unhealthy	Count	34	16	50
		% within Sex of student	42.5%	16.5%	28.2%
	Healthy	Count	29	27	56
		% within Sex of student	36.3%	27.8%	31.6%
	Very healthy	Count	17	54	71
		% within Sex of student	21.3%	55.7%	40.1%
Total		Count	80	97	177
		% within Sex of student	100.0%	100.0%	100.0%

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	24.426[a]	2	.000
Likelihood Ratio	25.330	2	.000
Linear-by-Linear Association	23.773	1	.000
N of Valid Cases	177		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 22.60.

Figure 23.4 SPSS chi-square output

We get a table headed **Health rating * Sex of student Crosstabulation** which is an SPSS version of Table 23.1. We then get a table headed **Chi-Square Tests**. The relevant part of this table is the first row labelled **Pearson Chi-Square**. Under **Value** we see 24.426, which is the sample χ^2 value we calculated above (there is some slight difference due to rounding error in calculating the expected frequencies in each cell). With 2 degrees of freedom (df), we see that the significance level printed under **Asymp. Sig. (2-sided)** is reported to be .000, although this really means 'less than 5-in-10,000' chances of obtaining samples with this distribution from populations where the variables are independent. This very low p-value leads us to reject the null hypothesis of independence.

Example

We will work through one more example using the five-step hypothesis testing procedure to see it in the familiar context. We have the data presented in Table 23.10 showing the joint distribution of 800 children in terms of their sex and whether they watch the news on TV.

Table 23.10 Children's TV newswatching by sex

Watch news on TV?	Sex		
	Girl	Boy	Total
No	25	35	60 (7 5%)
Yes	377	363	740 (92.5%)
Total	402	398	800

Step 1: State the null and alternative hypotheses

H_0: Sex and TV newswatching are independent of each other.
H_a: Sex and TV newswatching are not independent of each other.

Step 2: Choose the test of significance

We have sample data arranged in a bivariate table to see if there is a relationship between two variables. This makes the chi-square test for independence the relevant inference test.

Step 3: Describe the sample and calculate the p-score

We have already described the data in the crosstab above. To derive the test statistic from the table so that we can look up the p-score we first need to calculate the expected frequencies based on the Total column percentages in Table 23.11.

Table 23.11 Children's TV newswatching by sex: expected frequencies

Watch news on TV?	Sex		
	Girl	Boy	Total
No	30.2	29.8	60 (7.5%)
Yes	371.8	368.2	740 (92.5%)
Total	402	398	800

$$\chi^2_{sample} = \sum\frac{(f_o - f_e)^2}{f_e} = \frac{(25-30.2)^2}{30.2} + \frac{(35-29.8)^2}{29.8} + \frac{(377-371.8)^2}{371.8} + \frac{(363-368.2)^2}{368.2} \approx 1.9$$

Step 4: Determine at what alpha level, if any, the result is statistically significant

We have a 2-by-2 table, so there is only 1 degree of freedom. We look across this row in Table 23.12 and find that the sample chi-square lies between the critical values for alpha levels of 0.1 and 0.2. The result is therefore not statistically significant and we cannot reject the possibility that there is no relationship between these variables, despite the differences observed in the samples.

Table 23.12 Critical values of chi-square

df	Level of significance									
	0.99	0.90	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.00016	0.0158	0.148	0.455	1.074	1.642	2.706	3.841	6.635	10.827
2	0.0201	0.211	0.713	1.386	2.408	3.219	4.605	5.991	9.210	13.815
3	0.115	0.584	1.424	2.366	3.665	4.642	6.251	7.815	11.341	16.268
4	0.297	1.064	2.195	3.357	4.878	5.989	7.779	9.488	13.277	18.465
5	0.554	1.610	3.000	4.351	6.064	7.289	9.236	11.070	15.086	20.517
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
30	14.953	20.599	25.508	29.336	33.530	36.250	40.256	43.773	50.892	59.703

We can also conduct a chi-square test and obtain the exact significance level by turning to one of the internet-based statistics calculation pages such as those located at:

- www.unc.edu/~preacher/chisq/chisq.htm
- www.physics.csbsju.edu/stats/contingency_NROW_NCOLUMN_form.html

These indicate the exact *p*-score is 0.167, which falls in the range we determined with reference to the table.

Step 5: Report results

A sample of 402 girls and 398 boys were asked if they watch the nightly news on TV. Of the girls, 94 percent watched the news, while 91 percent of boys watched the news. This slight difference was not statistically significant ($\chi^2 = 1.9$, $p = 0.167$, $df = 1$), so that we cannot reject the possibility that the sex of students does not affect the rate at which they watch the TV news.

Example

A random sample of 50 migrants from non-English-speaking backgrounds (NESB) and a random sample of 50 migrants from English-speaking backgrounds (ESB) are asked whether or not they feel they have ever been discriminated against in seeking employment or promotion. We suspect that perception of discrimination is somehow dependent on language background, so that we will form crosstabs with language background as the independent variable and perception of discrimination as the dependent variable. However, this suspicion may not be correct. These two variables may in fact be independent of each other, so that knowing if a migrant is ESB or NESB tells us nothing about whether that migrant feels a stronger or weaker sense of discrimination. The results for all 100 respondents are shown in Table 23.13.

Table 23.13 Perception of discrimination

Discrimination	Total
No	40
Yes	60
Total	100

If the two variables are independent we should expect to find the percentage distribution of 'Yes' and 'No' responses for each migrant group to be the same as that for the two groups combined. Table 23.14 illustrates the simplest way to calculate these expected frequencies.

To calculate the expected frequency for each cell multiply the column total by the row total and divide the product by the total number of cases.

Table 23.14 Expected distribution of responses

Discrimination	Status		
	NESB	ESB	Total
No	$\frac{50\times 40}{100} = 20$	$\frac{50\times 40}{100} = 20$	40
Yes	$\frac{50\times 60}{100} = 30$	$\frac{50\times 60}{100} = 30$	60
Total	50	50	100

However, instead of these expected values, the sample produced the observed frequencies shown in Table 23.15.

Table 23.15 Actual distribution of responses

Discrimination	Migrant status		
	NESB	ESB	Total
No	5	35	40
Yes	45	15	60
Total	50	50	100

We could stop here and let the descriptive statistics contained in these tables speak for themselves. NESB migrants do have a relatively higher perception of being discriminated against than ESB migrants. However, we must remember that because we are only working with samples rather than populations, the result can simply be due to random variation. We might just happen to select a high proportion of NESB migrants who feel discriminated against and/or a slightly lower proportion of ESB migrants who feel discriminated against, even though in the populations there is no difference. This is where the chi-square test helps. Table 23.16 uses the expected and observed values for each cell to calculate their respective contributions to the total chi-square value.

Table 23.16 Calculations for chi-square

Discrimination	Migrant status	
	NESB	ESB
No	$\chi^2 = \frac{(5-20)^2}{20} = 11.25$	$\chi^2 = \frac{(35-20)^2}{20} = 11.25$
Yes	$\chi^2 = \frac{(45-30)^2}{30} = 7.5$	$\chi^2 = \frac{(15-30)^2}{30} = 7.5$

$$\chi^2_{sample} = \sum \frac{(f_o - f_e)^2}{f_e} = \frac{(5-20)^2}{20} + \frac{(35-20)^2}{20} + \frac{(45-30)^2}{30} + \frac{(15-30)^2}{30}$$

$$= 11.25 + 11.25 + 7.5 + 7.5 = 37.5$$

In a 2-by-2 table such as this, there is 1 degree of freedom. Looking at the table for the distribution of chi-square, with 1 degree of freedom, the level of significance is less than 0.001 (Table 23.17). We can therefore say that the probability of getting frequencies such as those we observe, if the two variables are independent, is less than 0.001 (less than one in a thousand). Therefore, we reject the null hypothesis of independence, and argue that the perception of discrimination does systematically differ between migrant groups, such that NESB migrants have a systematically higher perception of discrimination than ESB migrants.

Table 23.17 Critical values of chi-square

df	Level of significance									
	0.99	0.90	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.00016	0.0158	0.148	0.455	1.074	1.642	2.706	3.841	6.635	10.827
2	0.0201	0.211	0.713	1.386	2.408	3.219	4.605	5.991	9.210	13.815
3	0.115	0.584	1.424	2.366	3.665	4.642	6.251	7.815	11.341	16.268
4	0.297	1.064	2.195	3.357	4.878	5.989	7.779	9.488	13.277	18.465
5	0.554	1.610	3.000	4.351	6.064	7.289	9.236	11.070	15.086	20.517
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
30	14.953	20.599	25.508	29.336	33.530	36.250	40.256	43.773	50.892	59.703

Problems with small samples

You may have noticed the footnote attached to the Pearson chi-square value in the SPSS output (Figure 23.4) we generated above:

b. 0 cells (.0%) have expected count less than 5. The minimum expected count is 22.60

This is a check to see whether any cells in the crosstab have an expected frequency of 5 or less. SPSS basically runs through the table and determines how many, and what percentage of, cells have an expected frequency of less than 5. It then indicates what the lowest expected frequency in the table is – in this case 156.12. You can confirm this for yourself by referring to our calculations of expected frequencies in Table 23.6 (allowing for slight rounding error).

The reason why SPSS goes through such a procedure to indicate how many, if any, cells have an expected frequency of less than 5, is because a problem can arise with the use of a chi-square test when working with small samples. If the use of small samples leads to either of the following situations, the chi-square statistic becomes difficult to interpret:

- Any cell in the bivariate table has an expected frequency of less than 1.
- The expected frequency of cases in a large percentage of cells is less than 5. Usually 20 percent of cells is considered too high, but any cells with expected values of less than 5 can create a problem.

If the footnote to the chi-square value in the SPSS output indicates that one of these conditions has been violated, the chi-square test cannot be meaningfully interpreted. In such situations there are some alternatives, depending on the dimensions of the table.

With 2-by-2 tables some writers suggest using Yate's correction for continuity:

$$\chi_c^2 = \sum \frac{\left(\left|f_o - f_e\right| - 0.5\right)^2}{f_e}$$

Other writers suggest that for 2-by-2 tables, Fisher's exact probability test should be used. SPSS calculates both of these alternatives in the relevant situations (See H.T. Reynolds, 1977, *The Analysis of Cross-classification*, London: Free Press, 9–10, for a discussion of these procedures.)

With tables larger than 2-by-2 the only possible solution is to collapse categories together for either or both variables so as to increase expected frequencies. Before doing this, though, we need to justify the procedure because information is lost when categories are collapsed together. Originally there was enough information to say that one case differed from another case in terms of a variable, but if these cases are now in the same category after the original categories are combined, we are saying that such cases are the same. For example, we might need to collapse the four-point scale shown in Figure 23.5 into a two-point scale (in SPSS using the **Transform/Recode** command) in order to avoid small expected frequencies.

Thus cases that were previously classified into separate groups, such as Low and Very low, now are classified in the same group, namely Low. The scale was originally constructed for supposedly good theoretical reasons, and we should be wary of abandoning that scale simply to allow us to use a statistical procedure.

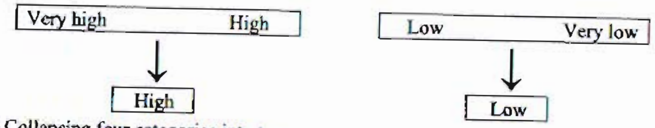


Figure 23.5 Collapsing four categories into two

Problems with large samples

The other main problem with the use of chi-square as a test of independence is that it is especially sensitive to large samples. The chance of finding a significant difference between samples always increases with sample size, regardless of whether we use *z*-tests, or *t*-tests, or *F*-tests, or chi-square. This in itself is not a problem; in fact, we should place greater faith in the results of larger samples rather than small samples, since large samples are more reliable. However, especially with chi-square, we may risk overstating the importance of a statistically significant difference.

To illustrate this problem, imagine that you are looking at two people standing far away. With the naked eye they appear to have the same height. But through a pair of binoculars it is evident that one person is slightly taller than the other. The more powerful the looking device we use to make our observation, the more likely slight differences will be detected. However, this example should also highlight the important distinction between statistical difference and meaningful difference. There may be a statistical difference in height of 1 inch between two people, but for all practical purposes they are as tall as each other. Using too powerful a looking device may complicate a picture by exaggerating slight statistical differences that aren't really worth worrying about in practice. When performing inference tests, increasing sample size has the effect of intensifying the 'looking device' we are employing and thereby accentuating slight differences that may not be important.

Increasing the sample size increases the chance of detecting a statistical difference that smaller samples may attribute to sampling error.

Of all the tests we cover, chi-square is especially sensitive to sample size and might result in a statistically significant difference even though a difference is trivial. To see this, assume that we have respondents grouped according to their respective level of education. Level of education is measured by asking if the respondent has had a university education or not. We are interested in whether this affects enjoyment of work, measured according to whether respondents find their job 'Exciting', 'Routine', or 'Dull'. The distribution, when expressed as percentages of the total for each group is shown in Table 23.18.

Table 23.18 Enjoyment of work by education level: Relative frequencies

Enjoyment of work	University education		
	No	Yes	Total
Dull	47.0%	45.7%	46.8%
Routine	48.2%	47.9%	48.2%
Exciting	4.8%	6.4%	5.0%
Total	100%	100%	100%

If these percentages are derived from a total of 1461, consisting of 1242 people without university education and 219 people with university education, the figures for observed and expected values will be as listed in Table 23.19.

Table 23.19 Enjoyment of work by education level

Enjoyment of work	University education		
	No	Yes	Total
Dull	584 (581.5)	100 (102.5)	684
Routine	599 (598.5)	105 (105.5)	704
Exciting	59 (62)	14 (11)	73
Total	1242	219	1461

Just by looking at the table it is clear that there is little difference between the observed frequencies and the expected frequencies (shown in brackets). As a matter of common sense we will say that the difference between the distribution of those without a university education and those with a university education in terms of job satisfaction is so slight that it could easily be put down to chance: the null hypothesis of independence is not rejected. In fact, the chi-square for this table is:

$$\chi^2_{sample} = 1.08148$$

The probability of getting this by chance alone, with 2 degrees of freedom, is:

$$p_{sample} = 0.58$$

However, if we obtain exactly the same pattern of responses, but from a sample size 10 times as large (n = 14,610) the conclusion is different. The bivariate table will be as shown in Table 23.20.

Table 23.20 Enjoyment of work by education level. Observed and expected frequencies

Enjoyment of work	University education		
	No	Yes	Total
Dull	5840 (5815)	1000 (1025)	6840
Routine	5990 (5985)	1050 (1055)	7040
Exciting	590 (620)	140 (110)	730
Total	12,420	2190	14,610

All we have done is to multiply the value in each cell by a factor of 10. The effect is to also increase the value of chi-square for this table by exactly 10 times the value for that calculated from the previous table:

$$\chi^2_{sample} = 10.8148$$

This is now significant at the 0.01 level: the difference between observed and expected frequencies is large enough to allow us to reject the null hypothesis of independence. The pattern of responses is the same *relatively*, yet the conclusion is reversed. This shows that *any relative difference in frequency distributions can be significant if it comes from sufficiently large samples.*

One possible solution is to do the opposite to that when confronted with a small sample: use even finer scales to measure the dependent and/or independent variables. For example, here we could use more than three possible responses for the question: 'How much do you enjoy your work?' Unfortunately, by the time this problem arises – the data analysis stage of research – it is usually too late to change the scale and resurvey the respondents. At best it is a solution to an anticipated problem, but it does indicate the value of allowing for a wide range of possible responses when working with nominal/ordinal data on large samples.

If this problem is not anticipated and a significant result is obtained that might be due to sample size, then we should look at the percentage distribution of responses alone and make a judgment based on these percentages, without adding the complication of chi-square (i.e. work with the 'naked eye' rather than the statistical binoculars).

To aid this decision, we can refer to the appropriate measure of association and see if these measures indicate a negligible association between the two variables. If we calculate gamma for either of these tables it will equal 0.04, since measures of association are not affected by sample size when relative frequencies stay the same. This indicates that the relationship is so weak as to be negligible; we should not even bother to proceed to determine whether such a trivial relationship derives from a relationship in the population.

Appendix: hypothesis testing for two percentages

This chapter discussed a widely used test of significance – the chi-square test of independence. The reason for its popularity is that it is applicable in situations in which we have categorical (nominal and ordinal) data and we are interested in the frequency distribution across the categories of the variable. This situation is very common in research. The chi-square test looks at the distribution of responses in a bivariate table and assesses whether a pattern of dependence exists. In the case of a 2-by-2 bivariate table (i.e. when both variables are binomial) a z-test of percentages can also be carried out on the same data; in fact, the two tests are equivalent ways of analyzing the same data and yield the same result. Indeed, the z-test of percentages can be considered a special case of the chi-square test, and since it is commonly used in research, it is worth knowing the mechanics of its calculation.

This appendix will work through an example of a z-test of sample percentages and then use a chi-square test to show that the results will be the same.

The z-test for two percentages

A (hypothetical) survey is conducted to investigate the level of support for social welfare reform, and whether this varies by age. Respondents are grouped according to whether they are aged 'under 45' or '45 or over'. Each respondent is also asked whether the government should do more to alleviate poverty. This is put to respondents as a simple 'yes or no' question.

The null hypothesis is that the percentage of under 45s responding 'yes' (P_1) is the same as the percentage of those 45 or over responding 'yes' (P_2):

$$H_0: P_1 = P_2$$

If this is true, samples taken from such populations will usually reflect the equality. In other words, the difference between any two sample percentages, if there is no difference between the populations, should be zero or close to it.

But this will not always be the case. Samples do not always exactly reflect the populations from which they are drawn. Random variation may cause us to pick up a few 'extra' young people who are in favor of welfare reform, and a few 'extra' older people who are opposed, causing the sample percentages to differ considerably. This means that if there is a difference between the two sample percentages, we cannot automatically conclude it reflects an underlying difference in the populations. However, larger differences between the sample percentages are less likely to be due to random chance. The z-test for percentages gives us the precise probability of such unlikely events occurring.

The survey consisted of 600 people under the age of 45 and 400 people aged 45 years or older. The percentage of each group responding 'yes', the government should do more to alleviate poverty, is:

$$\text{under 45: } P_1 = \frac{490}{600} \times 100 = 82\%$$
$$n_1 = 600$$

$$\text{45 or older: } P_2 = \frac{232}{400} \times 100 = 58\%$$
$$n_2 = 400$$

Does this reflect an underlying difference between the age groups on this issue? To determine this we begin with the following formula:

$$P_u = \frac{n_1P_1 + n_2P_2}{n_1 + n_2}$$

This is basically a weighted average of the two sample percentages, a sort of mid-point between the two results. If we substitute the relevant numbers into the equation we get:

$$P_u = \frac{n_1P_1 + n_2P_2}{n_1 + n_2} = \frac{600(82) + 400(58)}{600 + 400} = 72.2\%$$

This calculation allows us to determine the standard error of the sampling distribution of all possible sample differences. One standard error is defined by:

$$\sigma_{p-p} = \sqrt{P_u(100 - P_u)}\sqrt{\frac{n_1 + n_2}{n_1n_2}} = \sqrt{72.2(100 - 72.2)}\sqrt{\frac{600 + 400}{600(400)}} = 29\%$$

The actual difference between our two samples in terms of z-scores is:

$$z_{sample} = \frac{P_1 - P_2}{\sigma_{p-p}} = \frac{82 - 58}{29} = 8.3$$

This z-score is significant at the 0.01 level: we reject the null hypothesis of no difference and argue that support for government assistance to the poor does vary with age.

Chi-square test for independence

The alternative way of analyzing these data is to organize them into a 2-by-2 bivariate table (Table 23.21). The figures in brackets are the expected values based on the percentage of total respondents who said 'yes' or 'no'. Notice that 72.2 percent of all respondents agreed with the need for welfare reform. From this figure we calculate the number of 'under 45' respondents and '45 or over' respondents who are expected to agree. The 72.2 percent is the same figure that popped up in the two-sample z-test for percentages as the reference point for calculating the standard deviation of the sampling distribution.

Table 23.21 Attitude to government policy by age group

Agree	Age group		
	Under 45	45 or over	Total
No	110 (166.8)	168 (111.2)	278 27.8%
Yes	490 (433.2)	232 (288.8)	722 72.2%
Total	600	400	1000

We can substitute these observed and expected frequencies into the equation for chi-square to give us a test statistic of 67:

$$\chi^2 = \sum\frac{(f_o - f_e)^2}{f_e} = \frac{(110 - 166.8)^2}{166.8} + \frac{(168 - 111.2)^2}{111.2} + \frac{(490 - 433.2)^2}{433.2} + \frac{(232 - 288.8)^2}{288.8}$$
$$= 67$$

From the table for the distribution of chi-square the probability of getting this value (or greater) from identical populations is 0.005 – the same as that for the z-test.

The conclusion to draw from this is that while two-sample binomial z-tests are very common, and therefore worth knowing, they are in fact a special case of chi-square. Since the formula for the z-test is more cumbersome, and the logic not as intuitively clear, it is probably best to use chi-square in most situations. Also SPSS cannot conduct two-sample z-tests of proportions, but it can calculate a chi-square on a 2-by-2 table.

Exercises

23.1 How many degrees of freedom are there for tables with each of the following dimensions:

(a) 2 by 4 (b) 4 by 2 (c) 6 by 4 (d) 3 by 5?

23.2 If a chi-square test, with n = 500, produces χ^2 = 24, what will χ^2 be with the same relative distribution of responses, but with:

(a) n = 50 (b) n = 1000?

23.3 For the following table, calculate the expected frequencies for each cell and identify the ones that violate the rules for using chi-square.

	a	b	c	d	Total
a	1	0	6	48	55
b	2	0	7	40	49
Total	3	0	13	88	104

23.4 For the data in Exercise 5.4, which you used to construct a bivariate table, conduct a chi-square test to test your hypotheses about independence. Conduct this test on SPSS and compare the results with your hand calculations.

23.5 In earlier chapters we compared hypothetical samples of children from Australia, Canada, Singapore, and Britain, in terms of the amount of TV they watch. Assume that this variable was not measured at the interval/ratio level, but rather on an ordinal scale. The results of this survey are presented in the following table. Can we say that the amount of TV watched is independent of country of residence?

Amount of TV	Country				
	Canada	Australia	Britain	Singapore	Total
Low	23	25	28	28	104
Medium	32	34	39	33	138
High	28	30	40	35	133
Total	83	89	107	96	375

23.6 A sample of 162 men between the ages of 40 and 65 years is taken and the state of health of each man recorded. Each man is also asked whether he smokes cigarettes on a regular basis. The results are crosstabulated using SPSS, the results of which are shown over the page.

(a) What are the variables and what are their respective levels of measurement?

(b) Should we characterize any possible relationship in terms of one variable being dependent and the other independent? Justify your answer.

(c) Calculate by hand the column percentages and the expected values if the null hypothesis of independence is true, and confirm that they are the same as those in the SPSS table.

Health level * Smoking habit Crosstabulation

			Smoking habit		
			Doesn't smoke	Does smoke	Total
Health level	Poor	Count	13	34	47
		Expected Count	28.1	18.9	47.0
		% within Smoking habit	13.4%	52.3%	29.0%
	Fair	Count	22	19	41
		Expected Count	24.5	16.5	41.0
		% within Smoking habit	22.7%	29.2%	25.3%
	Good	Count	35	9	44
		Expected Count	26.3	17.7	44.0
		% within Smoking habit	36.1%	13.8%	27.2%
	Very good	Count	27	3	30
		Expected Count	18.0	12.0	30.0
		% within Smoking habit	27.8%	4.6%	18.5%
Total		Count	97	65	162
		Expected Count	97.0	65.0	162.0
		% within Smoking habit	100.0%	100.0%	100.0%

(d) Looking at the column percentages, do you think that differences in health level between smokers and non-smokers could be the result of sampling variation rather than a difference in the populations?

(e) Conduct a chi-square test of independence on these data. Does it confirm your answer to (d)?

23.7 The following information was obtained from a survey of 50 'blue-collar' and 50 'white-collar' workers. The survey asked respondents if they could sing the National Anthem from start to finish. The results are 'Blue collar': Yes = 29, No = 21; 'White collar': Yes = 22, No = 28.

(a) Arrange these data into a bivariate table, and conduct a chi-square test of independence.

(b) (optional) Conduct a two-sample test for proportions on the same data and compare your results.

23.8 Use the Employee data file to assess whether minority classification and employment category are independent.

24

Frequency tests for two dependent samples

For each test for independent samples there is usually an analogous test for dependent samples. For example, the independent samples t-test for the equality of two means has its counterpart in the dependent samples t-test for the mean difference. A similar set of tests exists where we are comparing samples across the frequency distribution for a categorical variable. We have looked at the chi-square test for independence, which assumed that the groups formed by the categories of the independent variable for independent samples. Slightly different tests are used when the samples we are comparing are related (see Chapter 20).

This chapter will consider tests that can be applied to dependent samples compared in terms of a binomial scale. These two tests, the McNemar chi-square test for change and the sign test, each of which are actually special applications of test we have already covered to the dependent samples context. These tests compare two dependent samples in terms of their distribution across a binomial variable. These two tests are equivalent, in the sense that they will always produce the same p-value for any given difference between the samples. In the text we will detail the McNemar test, since the SPSS output for this test provides slightly more information than with a sign test. After working through the McNemar test we will conduct a sign test on the same data to show the difference in the presentation of the results.

The McNemar chi-square test for change

The McNemar test applies to two dependent samples that are compared in terms of outcomes for a binomial variable (i.e. a variable that has two possible outcomes). The McNemar test compares the outcome for each case in one sample with the outcome for its respective pair in the other sample. For example, a political scientist might be interested in whether televised debates between political candidates have an effect on voting intentions. The researcher randomly selects 137 people and asks them whether they plan to vote Progressive or Conservative at the forthcoming election, ignoring all other candidates. The researcher then asks the same question of *the same* 138 people after they have watched a televised debate between the Progressive and Conservative candidates.

In comparing each individual in the 'before' stage with his or her own particular response after the debate, there are four possibilities. Table 24.1 and Figure 24.1 illustrate these possibilities.

The McNemar test only considers those pairs for which a change has occurred, and analyzes whether any changes tend to occur in one direction (e.g. Conservative to Progressive) *or* the other (Progressive to Conservative). The total number of pairs registering a change will be cells (b) and (c) in Table 24.1. If the changes induced by watching the TV debate do not favor a shift in one direction or the other, then we should *expect* to find 50 percent of the total number of changes in cell (b), and 50 percent in cell (c).

Table 24.1 Joint distribution of survey results

Before	After	
	Conservative	Progressive
Conservative	No change (a)	Change (b)
Progressive	Change (c)	No change (d)

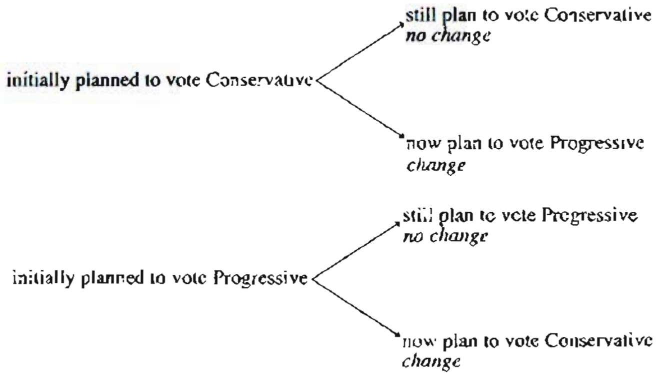

still plan to vote Conservative
no change
initially planned to vote Conservative
now plan to vote Progressive
change
still plan to vote Progressive
no change
initially planned to vote Progressive
now plan to vote Conservative
change

Figure 24.1 Before-and-after voting intentions

Of course, random variation will cause samples to differ from the expected result, even if the debate did not affect the overall opinion of the population. It is possible (although very unlikely) to select a random sample where 90 percent of all pairs registering a change in opinion are in cell (b), even if in the whole population the changes are similar in either direction. The greater the difference between the observed cell frequencies and the expected cell frequencies, however, the less likely that such an event is due to sampling error when sampling from populations where no change has occurred.

This discussion of expected and observed cell frequencies should sound similar to the chi-square test. In fact the McNemar test (with large samples) is a chi-square test for the difference between expected and observed cell frequencies. This test statistic is calculated using the following formula:

$$\chi_M^2 = \frac{(n_1 - n_2 - 1)^2}{n_1 + n_2}$$

where n_1 is the observed number of cases in cell (b) or cell (c), whichever is *largest*; and n_2 is the observed number of cases in cell (b) or cell (c), whichever is *smallest*.

The distribution of responses to this hypothetical study is shown in Table 24.2.

Table 24.2 Voting intentions before and after TV debate

Before	After	
	Conservative	Progressive
Conservative	28	55
Progressive	27	27

We can immediately see that the total number of cases that did not change their opinion (the unshaded cells) is 55:

$$28 + 27 = 55$$

whereas the total number of cases that *did* record a change (the shaded cells) is 82:

$$55 + 27 = 82$$

Obviously the sample result differs from the expected result, but is the difference big enough to warrant rejecting the null hypothesis? Using the formula for the McNemar statistic we get:

$$\chi_M^2 = \frac{(n_1 - n_2 - 1)^2}{n_1 + n_2} = \frac{(55 - 27 - 1)^2}{55 + 27} = 8.89$$

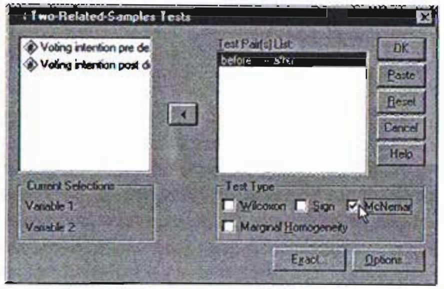
From Table A4 for the critical values of chi-square, with 1 degree of freedom, the *p*-score for this chi-square is less than 0.01 level. This leads us to reject the null hypothesis. The TV debate does have an affect on voting intentions. Looking back at the table of raw numbers, it is clear that the direction of change is from Conservative to Progressive.

The McNemar test using SPSS

Table 24.3 and Figure 24 2 go through the steps involved in conducting a McNemar test on these data.

Table 24.3 McNemar test on SPSS (file: **Ch24.sav**)

SPSS command/action	Comments
1 From the menu select **Analyze/ Nonparametric Tests/2 Related Samples**	This brings up the **Two-Related-Samples Tests** dialog box. You will notice that in the area to the bottom right of the window headed **Test Type** the small square next to **Wilcoxon** is selected. This indicates that the Wilcoxon test for two dependent samples is the default test. Here we want to conduct a McNemar test so we need to 'unselect' Wilcoxon and select McNemar instead
2 Click on the square next to **Wilcoxon**	This removes ✓ from the tick-box
3 Click on the square next to **McNemar**	This places ✓ in the tick-box, indicating that it is the selected test
4 Click on **Voting intention pre debate** and then click on **Voting intention post debate** in the source variables list	These two variable names will be highlighted
5 Click on ▸	This pastes the highlighted two variables into the **Test Pairs:** target list indicating responses for the two variables will be matched
6 Click on **OK**	


Two-Related-Samples Tests

Test Pair(s) List: before -- after

Test Type: □ Wilcoxon □ Sign ☑ McNemar □ Marginal Homogeneity

Figure 24.2 The SPSS **Two-Related-Samples Tests** dialog box

Figure 24.3 presents the output from this set of instructions. The first table in the output contains the descriptive statistics for the sample data, and is basically the same as Table 24.2. The second table labelled **Test Statistics** contains the information for the McNemar chi-square test on the cells in the first table reflecting a change. The difference between the observed and expected frequencies produces a chi-square value for the sample of 8 890. With 1 degree of freedom, the exact probability of getting this sample chi-square just by random variation is .003. This is well below any normal alpha level such as 0.05. The researcher concludes that the TV debate is likely to favor a change in opinion, from Conservative to Progressive.

McNemar Test

Crosstabs

Voting intention pre debate & Voting intention post debate

	Voting intention post debate	
Voting intention pre debate	1	2
1	28	55
2	27	27

Test Statistics[b]

	Voting intention pre debate & Voting intention post debate
N	137
Chi-Square[a]	8.890
Asymp. Sig.	.003

a. Continuity Corrected

b. McNemar Test

Figure 24.3 The SPSS **Two-Related-Samples Tests** dialog box and McNemar test output

Before leaving the discussion of the McNemar test, we should note that since it is a special application of the chi-square test, it also suffers from the same limitations. In particular, from Chapter 23 we know that the chi-square test is only appropriate when expected cell frequencies are 5 or more. This rule applies to the McNemar test, and the same correction is taken. When cell sizes are small, SPSS will automatically use the binomial approximation to the normal curve, and print the two-tail probability associated with this approximation.

The sign test

You will notice that under **Test Type** in the **Two-Related-Samples** dialog box there are three options for conducting an inference test on two dependent samples. One is the McNemar chi-square test that we have just discussed. Another is the Wilcoxon signed-ranks test, which is the default test, and which we will discuss in detail below. The third is the **sign test**. The sign test conducts a binomial *z*-test, much like that detailed in Chapter 21. The pairs in which there is a change in one direction (such as Conservative to Progressive) are given a positive sign, and the pairs in which there is a change in the other direction (such as Progressive to Conservative) are given a negative sign. A binomial *z*-test is then conducted by comparing the proportion of positive changes (or negative changes) with the test proportion of 0.5.

In the above SPSS procedure, if we had selected the sign test rather than the McNemar test under **Test Type** in the **Two-Related-Samples** dialog box, we would obtain the output presented in Figure 24.4.

Sign Test

Frequencies

		N
Voting intention post debate - Voting intention pre debate	Negative Differences[a]	27
	Positive Differences[b]	55
	Ties[c]	55
	Total	137

a. Voting intention post debate < Voting intention pre debate

b. Voting intention post debate > Voting intention pre debate

c. Voting intention pre debate = Voting intention post debate

Test Statistics[a]

	Voting intention post debate - Voting intention pre debate
Z	-2.982
Asymp. Sig. (2-tailed)	.003

a. Sign Test

Figure 24.4 SPSS sign test output

In the first table headed **Frequencies** we have the sample descriptive statistics, which is just another way of presenting the same information as in Table 24.2. The **Test Statistics** table presents the information on the binomial *z*-test for the signs. Note that the two-tailed probability of .003 for the sample *z*-score of -2.982 is the same as that for chi-square in the McNemar test.

The probability obtained through the sign test is always exactly the same as that obtained from a McNemar test applied to the same data. Therefore the same decision is made regarding the null hypothesis, regardless of which test is used. The advantage of the McNemar test is that the crosstab that is generated as part of the SPSS output provides a more detailed breakdown of the pairs than the output that comes with the sign test. This makes it easier to interpret the data since it allows us to see in which direction the changes move.

Example

A study is conducted to investigate attitudes toward computer games. Fifty people are randomly chosen and asked if they believe video games to be of any educational value, with responses restricted to 'yes' or 'no'. After playing a range of video games each person is asked the same question. The distribution of responses is recorded in Table 24.4, with the cells indicating a change in attitude highlighted.

Table 24.4 Attitude to video games before and after playing

After	Before	
	No	Yes
No	15	18
Yes	10	7

Substituting this information into the formula for the McNemar test produces a test statistic of 1.75.

$$\chi^2_M = \frac{(n_1 - n_2 - 1)^2}{n_1 + n_2} = \frac{(18 - 10 - 1)^2}{18 + 10} = 1.75$$

From the distribution for chi-square table, this sample chi-square has a significance level between 0.1 and 0.2. We therefore do not reject the null hypothesis: playing video games does not seem to change people's attitude in one particular way or the other.

Summary

We have observed that the McNemar test and the sign test are essentially the same: SPSS presents them as alternatives yielding slightly different information, however, so we have covered each separately. As an alternative to SPSS there are web pages that can perform the McNemar and Sign tests calculations on data entered, such as the following:

- www.fon.hum.uva.nl/Service/Statistics/McNemars_test.html
- www.fon.hum.uva.nl/Service/Statistics/Sign_Test.html
- home.clara.net/sisa/pairwise.htm

Exercises

24.1 Conduct a McNemar test and sign test on the following data.

(a)

After	Before	
	1	2
1	27	22
2	34	28

(b)

After	Before	
	1	2
1	12	55
2	50	17

(c)

After	Before	
	1	2
1	32	134
2	79	12

24.2 Brothers and sisters are matched and asked if they play regular sport. The results are:

Sister	Brother	
	Yes	No
Yes	18	11
No	16	15

(a) Conduct a McNemar test and sign test to assess whether there is a difference between brothers and sisters in terms of sport playing.

(b) Enter these data in SPSS, and conduct a McNemar test and sign test. Compare the SPSS output with your hand calculations.

PART 6

Inferential statistics: Other tests of significance

25

Rank-order tests for two or more samples

The previous parts of this book concentrated on tests of significance for data described by a mean or a frequency distribution respectively. These tests are very handy because the mean and the frequency distribution are such important descriptive tools in research. However, these by no means exhaust the possible ways of describing a set of sample data. We know from Part 2 that there are other descriptive statistics such as the median and the standard deviation that are important ways of assessing aspects of a distribution that a mean or frequency distribution do not capture. Generally, for each descriptive statistic that we can generate for a random sample, there is also a corresponding inferential statistic that will allow us to generalize from this sample to a population. We have concentrated thus far on tests for means and frequency distributions because these specific ways of data description are particularly common and useful. This part of the book will detail some other tests that rely on other descriptive statistics that are common, but to provide an exhaustive account of all the tests of significance that are potentially applicable to research data would take us beyond the needs of most researchers. This chapter will discuss rank-order tests of significance called the **z-test for the rank sum of two independent samples** (also known for short as the **Wilcoxon W test**), and its very close counterpart, the **Wilcoxon signed-ranks** test for two dependent samples.

Data considerations

Rank-order tests of significance are often used as substitutes for tests for means in situations where the mean is not an appropriate measure of central tendency. This can occur for two main reasons:

1. *The level at which the variable is measured is only ordinal and there are many points on the scale.* In research we do not always work with interval/ratio data but ordinal-level data instead. Sometimes this ordinal data looks interval/ratio. For example, we might construct an 'index of satisfaction', whereby we ask individuals to rate themselves on a scale from 1 to 10, with 1 indicating 'Not at all satisfied' and 10 indicating 'Extremely satisfied':

   1 —— 2 —— 3 —— 4 —— 5 —— 6 —— 7 —— 8 —— 9 —— 10

   Not at all satisfied *Extremely satisfied*

   Such an index is ordinal because the numbers assigned to each group are purely arbitrary. We can just as easily, and just as validiy, label the grades on the index 2, 5, 8, 12, 100, 133, 298, 506, 704, 999, rather than 1, 2, 3, 4, 5, 6, 7, 8, 9, 10. All we need to do in constructing an index is preserve the ranking of cases, since we are not measuring satisfaction by some unit of measurement, as we do when measuring age in years. All we can say is that one case is more or less satisfied than the other; we do not have a unit of measurement that allows us to say *by how much* one case is more or less satisfied than the other. For example, we cannot say that someone with a score of 6 is three times more satisfied than someone with a score of 2. In fact, instead of using numbers to label the categories, we could have used terms like 'Moderately satisfied' and 'Very satisfied' without losing any information at all. The problem is that when we use a long ordinal scale with numbers for labels, like 1, 2, 3, ..., 10, there is the *appearance* of interval/ratio data. This might tempt us to calculate a mean in order to compare two samples that have been measured on this scale. This is, strictly speaking, not a correct procedure.

Unfortunately, calculating a mean on essentially ordinal data is not an infrequent occurrence. Market research companies do this as a standard procedure when describing survey data. Indeed, this writer's own academic institution, the University of New South Wales, has introduced course evaluation measures, much like the above satisfaction scale, and uses the means of such scales to compare student evaluations of courses and instructors. What a score of 5.6 is meant to signify, however, and whether this is different in any meaningful way to a score of say 5.3, is not very obvious. Clearly, even such an august institution as this is not immune from statistical silliness!

2. *We cannot assume that the population is normally distributed.* Even if the level of measurement allows the mean to be calculated as the descriptive statistic for a set of data, to conduct an inference test on this mean requires the additional assumption (especially when working with small samples) that the population is normally distributed. This assumption is sometimes questionable. For example, we know that income in the population is not normally distributed: it is usually skewed to the right. Therefore, it is inappropriate to conduct a test for mean income. Fortunately, there is a range of significance tests such as those we will discuss in this chapter, called **distribution-free** (or **non-parametric**), tests that *do not require any assumption about the shape of the underlying population distribution.*

The rank sum and mean rank as descriptive statistics

To see the logic of the Wilcoxon rank sum test, we need to remind ourselves of the relationship between descriptive and inferential statistics. We begin with the raw data from a sample, and then calculate a descriptive statistic that somehow captures the 'essence' of these data that will help us answer a specific research problem. We then use inferential statistics to see if we can generalize from this sample result to the population. For any of the reasons we have just discussed the mean may not be an appropriate descriptive statistic. We might need to generate a different descriptive statistic from a sample, and then apply our inferential statistics to it. With data measured at least at the ordinal level, we can order cases from lowest to highest according to the 'score' each case receives on the scale. Once arranged in this order, each case can be assigned a **rank** that indicates where in the order it appears: first, second, third, and so on. Think of the way that tennis players are given a ranking, with the best player ranked number one, the second best ranked two, and so on. These numbers do not measure tennis playing ability as such, they merely indicate a position in an ordered series based on tennis playing ability. Just as we can rank-order people according to their tennis ability, we can rank cases according to any variable measured at least at the ordinal level.

To see how we use the **rank sum** and the **mean rank** as descriptive statistics for such data, we will elaborate the example we have used in preceding chapters regarding the TV viewing habits of Australian and British children. Let us assume that in trying to assess whether there is a difference between Australian and British children in terms of their TV watching behavior, the researcher is dissatisfied with using just viewing time measured in minutes as the operationalization of TV viewing behavior. The researcher believes that a child may sit in front of the TV for long periods of time, but this does not indicate the *intensity* with which the child watches TV, the level of interest in what is actually screened.

To incorporate this factor into the measurement of TV watching behavior the researcher observes 20 children from Australia and 20 children from Britain, taking note of their level of attention and their responses to what they see on the screen. Based on these observations, each child is given a score between 0 and 100 indicating their level of intensity of TV viewing. A score of 0 indicates a child who is completely disinterested with what is on TV, while a score of 100 indicates a child who shows an extremely high interest in the TV. The raw data from this research are listed in Table 25.1.

Table 25.1 Scores on viewing intensity index: Raw data from a (hypothetical) survey of children's TV viewing behavior

Australia	Britain
3	1
9	4
12	5
19	10
20	14
25	21
33	24
37	30
38	35
45	37
56	40
58	43
64	50
69	59
73	62
75	65
78	70
80	74
83	76
89	95

Clearly, this listing of the raw data, even when rank-ordered as in this table, is difficult to interpret. One British child shows the least interest in what he or she watches, but another British child is also the most highly engaged. What about the *overall* distribution across the range of scores? Before proceeding, 'eyeball' these data and try to make a judgment about any difference between these two samples in terms of their intensity of TV viewing.

You have probably concluded that the scores for British children tend to be clustered at the low end of the scale (relatively uninterested in TV), while the Australian children tend to be clustered at the other end (relatively interested in TV). We might be inclined to take just the mean for each set of scores and compare them. However, we need to resist this temptation because this is only an ordinal scale and therefore the mean will not 'mean' anything. We might more usefully calculate the median for each sample: I will leave it to you to calculate that the median for the Australian children is 50.5 and for British children it is 38.5. This gives us a better sense of the distribution, but since the median only makes use of the central score(s) of a distribution, rather than all the data points, it has limitations of its own.

A better way of describing these 40 pieces of data in a more digestible way is to assign each case a rank and to sum the ranks for each sample. If one sample tends to cluster at the low end of the scale then the sum of the ranks for this sample will be smaller than that for the other sample.

We first assign ranks to each case in our survey. To do this imagine that all 40 children are lined up with the British child who scored 1 at the head of the line, followed by the Australian child who scored 3, and so on down to the British child who scored 95 at the end of the line. Each child is then given a number, indicating their place, or **rank**, in the line (Table 25.2).

Table 25.2 Scores and ranks on viewing intensity index

Australia Score	Britain Score	Rank
	1	1
3		2
	4	3
	5	4
9		5
	10	6
12		7
	14	8
19		9
20		10
	21	11
	24	12
25		13
	30	14
33		15
	35	16
37	37	17.5
38		19
	40	20
	43	21
45		22
	50	23
56		24
58		25
	59	26
	62	27
64		28
	65	29
69		30
	70	31
73		32
75		33
	74	34
	76	35
78		36
80		37
83		38
89		39
	95	40
$\Sigma R = 441.5$	$\Sigma R = 378.5$	

A problem arises in assigning ranks when two or more cases score the same score for the variable. These are called **tied ranks**.

> To assign ranks to tied cases divide the **sum** of the ranks to be filled by the **number** of ranks to be filled.

For example, an Australian child and British child each scored 37 on the index. These two children occupied positions 17 and 18 in line, so their average rank is 17.5:

$$\text{average rank} = \frac{17+18}{2} = 17.5$$

Having allocated ranks to all the cases we simply then sum them **for each sample**. This produces **rank sums** of 441.5 and 378.5 for Australian and British children respectively. We can now easily compare these two numbers rather than compare the two sets of 20 numbers

that made up the raw data, and make an assessment of our research findings. The higher rank sum for Australian children indicates that they tended to cluster toward the high end of the scale, indicating that they watch TV with more intensity than British children.

In this example we conveniently have two samples with the same number of cases. If we had samples of unequal size, the rank sums would not be so easily compared because they will be affected by the number of cases in each sample rather than just the relative positions of the cases in the rank-ordering. To compensate for this problem with rank sums, an even more meaningful way of describing rank-ordered raw data is to calculate the **mean rank** ($\bar{R}$) for each sample. This is the rank sum for a sample divided by the number of cases in that sample:

$$\bar{R}_{australia} = \frac{441.5}{20} = 22$$

$$\bar{R}_{britain} = \frac{378.5}{20} = 19$$

On average Australian children are 22nd in line, whereas on average British children are 19th in line. We can see by comparing these two numbers, rather than by comparing the original 40 scores from which these mean ranks are derived, that British children watch TV with less interest than Australian children, although the difference does not seem very great.

To sharpen this notion of the rank sum and mean rank as descriptive statistics for long ordinal scales, let us consider the extreme situation depicted in Table 25.3.

Table 25.3 Scores on viewing intensity index

Australia	Britain
48	1
52	3
52	5
56	5
58	8
62	10
65	12
69	15
69	15
72	19
73	20
78	23
79	28
85	29
86	31
86	31
89	38
91	39
92	44
95	46

We can immediately see that if we lined these children up according to their index scores the British children will occupy the first 20 ranks, while the Australian children will occupy ranks 21, 22, 23, ..., 40. The mean ranks for each sample will be 10.5 and 30.5:

$$\bar{R}_{britain} = \frac{1+2+3+...+20}{20} = \frac{210}{20} = 10.5$$

$$\bar{R}_{australia} = \frac{21+22+23+...+40}{20} = \frac{610}{20} = 30.5$$

These two mean ranks clearly and concisely describe the basic difference in the distributions, which is the clustering of cases from one sample at one end of the scale and the clustering of cases from the other sample at the other end of the scale.

The *z*-test for the rank sum for two independent samples

We have observed two samples in Table 25.1 that differ in terms of the variable with which we are comparing them: intensity of viewing TV. In particular, we have found that the sample of Australian children tends to watch TV with more interest than the sample of British children. Can we draw an inference from this to the entire populations of Australian and British children?

Let us assume that in fact there is no difference between the two populations of children in terms of this variable. If there is no difference between these two populations (remember, this is just a hypothesis) we expect that the two *samples* will not differ. It is possible to select randomly two samples that produce the extreme rank sums from Table 25.3, even though there is no difference between the populations. Such a result, however, is highly improbable. If the two populations do not differ, the more likely result is that the sample of Australian and the sample of British children will be evenly spread through the joint distribution. In this case the rank-orders for the two samples will be identical so that each Australian child will tie with a British child on the intensity scale. Where the two samples are evenly spread through the rank-ordering, the rank sums for either sample will be equal to:

$$\mu_W = \frac{1}{2}n_1(n_1 + n_2 + 1)$$

where n_1 is the sample with the fewest cases and n_2 is the sample with the most cases.

In this example, each sample has 20 cases, so if the samples conformed exactly with our hypothesis of no difference between the populations, we will generate rank sums of:

$$\mu_W = \frac{1}{2}n_1(n_1 + n_2 + 1) = \frac{1}{2}20(20 + 20 + 1) = 410$$

The actual rank sums that we observe in our samples do not conform to this, reflecting the fact that one sample tended to cluster higher up the scale than the other. The rank sum for the sample of Australian children is 441.5 and for the sample of British children the rank sum is 378.5. We know, however, that random samples do not always exactly reflect the populations from which they are drawn. Random variation will often cause samples to differ from each other, even though the populations from which they are drawn are not different. What is the probability, in other words, of drawing samples that are as different in their index scores as that which we observe from populations that are not different?

To determine this probability for sample sizes of 20 or more we conduct a *z*-test on the difference between the smallest of the two rank sums (which is given the symbol *W*) and the value for μ_W. The formula for the *z*-value that is the test statistic is:

$$z_{sample} = \frac{W - \mu_W}{\sigma_W}$$

where:

$$\sigma_W = \sqrt{\frac{1}{12}n_1 n_2(n_1 + n_2 + 1)}$$

We substitute our sample results into these equations to determine the sample *z*-score. Here the smallest of the two rank sums is that for British children so that the value for $W = 378.5$:

$$\sigma_W = \sqrt{\frac{1}{12}n_1 n_2(n_1 + n_2 + 1)} = \sqrt{\frac{1}{12}20(20)(20 + 20 + 1)} = 37$$

$$z_{sample} = \frac{W - \mu_W}{\sigma_W} = \frac{378.5 - 410}{37} = -0.85$$

Such a *z*-score is not significant at any usually accepted alpha level. In plain words, although our *samples* differ, the difference is not so great for it to suggest the samples come from *populations* that differ. We cannot reject the hypothesis that the populations of Australian and British children do not differ in terms of TV viewing intensity and the sample difference is due just to random variation.

We should note that the sampling distribution of *W* is only approximately normal, but this a reasonable approximation for sample sizes larger than 20. A table for the exact distribution for *W* should be used for probabilities in the small sample case, a copy of which can be downloaded at fsweb.berry.edu/academic/education/vbissonnette/tables/wilcox_r.pdf. In the small sample case SPSS will automatically conduct an exact test rather than use the normal approximation.

Example

We want to see if people from rural areas are more or less conservative than people from urban areas. We asked a random sample of 22 people from rural areas and 22 people from urban areas a detailed set of questions, and from their responses constructed an 'index of conservatism' which ranges from 0 to 40. A score of 40 indicates someone who is extremely conservative, while a score of 0 indicates someone who is not at all politically conservative. All 44 scores are listed in Table 25.4.

Table 25.4 Scores (and ranks) on conservatism index: Samples of rural and urban residents

Urban	Rural
0 (*1*)	2 (*3*)
1 (*2*)	3 (*4*)
4 (*5*)	6 (*7*)
5 (*6*)	7 (*8*)
10 (*11*)	8 (*9*)
11 (*12*)	9 (*10*)
13 (*14*)	12 (*13*)
14 (*15*)	17 (*18*)
15 (*16*)	18 (*20*)
16 (*17*)	19 (*22*)
18 (*20*)	21 (*24*)
18 (*20*)	22 (*25*)
20 (*23*)	24 (*27*)
23 (*26*)	25 (*28*)
26 (*29*)	28 (*31*)
27 (*30*)	29 (*32*)
31 (*34*)	30 (*33*)
32 (*35*)	33 (*36.5*)
35 (*39.5*)	33 (*36.5*)
35 (*39.5*)	34 (*38*)
37 (*42*)	36 (*41*)
38 (*43*)	39 (*44*)

Now imagine lining up these 44 people from lowest to highest (i.e rank-ordering the cases). An urban resident scored the lowest with 0, and so appears first in line, while a rural dweller had the highest score of 39, and appears at the end of the line. Ranks are assigned to each person (indicated in the brackets) according to their position in the line-up.

Just to remind ourselves of how to assign tied ranks, look at the one rural and two urban dwellers who each scored 18 on the index of conservatism. Together, these three people occupy three spaces, which are 19th, 20th, and 21st in line:

$$\frac{19+20+21}{3} = 20$$

Therefore they are each assigned a rank of 20. Notice that in assigning this rank of 20 to each of these three cases we do not use ranks 19 or 21 for the cases immediately preceding or following them in line.

Can we say that the data in Table 25.4 indicate that urban and rural residents are politically distinct? To make this inference we will work through our five-step hypothesis testing procedure.

Step 1: State the null and alternative hypotheses

In this example with sample sizes of 22, the value of μ_W is:

$$\mu_W = \frac{1}{2}n_1(n_1+n_2+1) = \frac{1}{2}22(22+22+1) = 495$$

Remember, this is the rank sum we will get on average from samples drawn from populations that are no different in terms of the conservatism scale. Therefore the null and alternative hypotheses for this example are:

$$H_0\text{: } \mu_W = 495$$

$$H_a\text{: } \mu_W \neq 495$$

Step 2: Choose the test of significance

In this example we are comparing *two random samples* to see if they differ in terms of their rankings on an ordinal scale with many points. The appropriate test, therefore, is the Wilcoxon *z*-test for the rank sum.

Step 3: Describe the sample and derive the p-score

If we sum and average the ranks for each group, we get descriptive statistics that indicate the relative spread of the two samples in the joint distribution:

$$\Sigma R_1 = 480, \ \bar{R}_1 = 21.82$$

$$\Sigma R_2 = 510, \ \bar{R}_2 = 23.18$$

These rank sums and mean ranks give a sense as to whether one sample is more or less conservative than the other. Here we see that the mean rank for the urban sample is 21.82, whereas for the rural sample it is 23.18. This indicates that urban residents tended to have lower scores on the conservatism scale than rural residents.

The smallest of the two rank sums is that for urban dwellers, so that:

$$W = 480$$

This is obviously different to the value assumed in the null hypothesis, indicating that the samples differ. Can we conclude from this that the *populations* are different as well? The Wilcoxon test analyzes whether 480 is 'different enough' from the expected value of 495 to suggest that there is also a difference between the populations.

The standard error of the sampling distribution of rank sums (σ_W) for this example is:

$$\sigma_W = \sqrt{\frac{1}{12}22 \times 22(22+22+1)} = 42.6$$

The *z*-test for *W* produces the following result:

$$z_{sample} = \frac{W-\mu_W}{\sigma_W} = \frac{480-495}{42.6} = -0.352$$

If we refer to the table for the areas under the standard normal curve, we see that this test statistic of –0.352 has a *p*-score between 0.689 and 0.764 (Table 25.5).

Table 25.5 Area under the standard normal curve

z	Area under curve between both points	Area under curve beyond both points	Area under curve beyond one point
±0.1	0.080	0.920	0.4600
±0.2	0.159	0.841	0.4205
±0.3	0.236	0.764	0.3820
±0.4	0.311	0.689	0.3445
±0.5	0.383	0.617	0.3085
±0.6	0.451	0.549	0.2745
±0.7	0.516	0.484	0.2420
±0.8	0.576	0.424	0.2120
±0.9	0.632	0.368	0.1840
±1	0.683	0.317	0.1585
⋮	⋮	⋮	⋮
±3	>0.996	<0.004	<0.0020

As an alternative to these hand calculations (or to SPSS), we can enter our sample information into the following web calculation pages:

- www.fon.hum.uva.nl/Service/Statistics/Wilcoxon_Test.html, which allows the input of raw data;
- home.clara.net/sisa/ordinal.htm, which requires summarized data, specifically the sample sizes and the smallest of the two rank sums (*W*).

These pages indicate that the exact significance level for these data is $p = 0.64$.

Step 4: Decide at what alpha level, if any, the result is statistically significant

The difference between the two samples is clearly not statistically significant; it has a high probability of occurring as a result of sampling error from population that are no different.

Step 5 Report results

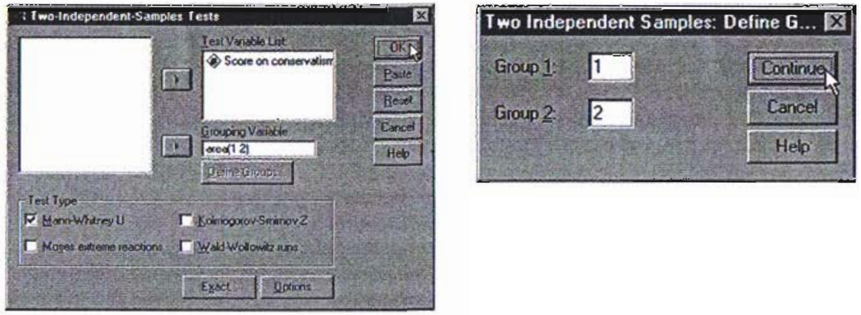
Samples of 22 urban residents and 22 rural residents were a detailed set of questions, and from their responses constructed an 'index of conservatism' which ranges from 0 to 40. A score of 40 indicates someone who is extremely conservative, while a score of 0 indicates someone who is not at all politically conservative. The sample of urban residents was slightly less conservative than the sample from the rural areas, with a mean rank of 21.82 compared to 23.18. This difference, however, was not statistically significant ($z = -0.352$, $p = 0.64$, two-tail). We cannot dismiss the possibility that urban and rural residents are no different in their political orientation and that the sample difference is due to sampling error.

Wilcoxon's rank sum *z*-test using SPSS

To conduct a rank-sum test on these data we follow Table 25.6 (Figure 25.1). The results from this set of instructions are presented in Figure 25.2.

Table 25.6 Wilcoxon's rank sum test using SPSS (file: Ch25-1.sav)

SPSS command/action	Comments
1 From the menu select **Analyze/Nonparametric Tests/2 Independent Samples**	This brings up the **Two-Independent-Samples Tests** dialog box. Notice that in the area for **Test Type** the tick-box next to **Mann-Whitney U** has ✓ indicating that this is the default test. This is the same test as the Wilcoxon. In other words, the Wilcoxon test is the default test which will automatically be generated under this command
2 Click on **Score on conservatism index** in the source variables list	This highlights **Score on conservatism index**
3 Click on ▸ that points to the **Test Variable List:**	This pastes **Score on conservatism index** into the **Test Variable List:**
4 Click on **Area of residence** in the source variables list	This highlights **Area of residence**
5 Click on ▸ that points to the area headed **Grouping Variable:**	This pastes **Area of residence** into the **Grouping Variable:** list. Notice that in this list the variable appears as **area(? ?)**
6 Click on **Define Groups**	This brings up the **Define Groups** box
7 In the area next to **Group 1:** type 1, and in the area next to **Group 2:** type 2	This identifies the two groups to be compared, which are urban and rural residents
8 Click on **OK**	


Two-Independent-Samples Tests

Test Variable List: Score on conservatism

Grouping Variable: area(1 2)

Test Type

☑ Mann-Whitney U □ Kolmogorov-Smirnov Z

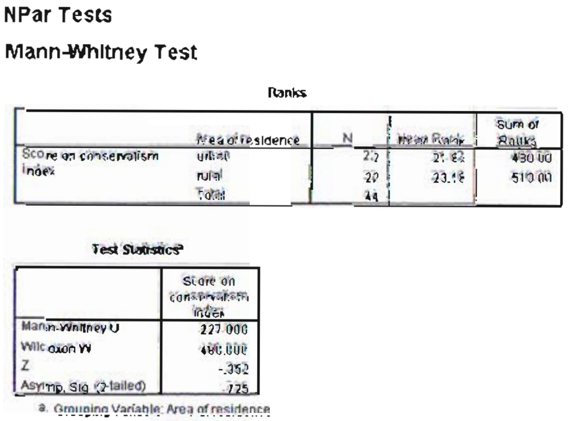
□ Moses extreme reactions □ Wald-Wolfowitz runs

Two Independent Samples: Define G...

Group 1: 1

Group 2: 2

Figure 25.1 The Two-Independent-Samples Tests and Define Groups dialog box


NPar Tests

Mann-Whitney Test

Ranks

	Area of residence	N	Mean Rank	Sum of Ranks
Score on conservatism index	urban	22	21.82	480.00
	rural	22	23.18	510.00
	Total	44		

Test Statistics^a

	Score on conservatism index
Mann-Whitney U	227.000
Wilcoxon W	480.000
Z	-.352
Asymp. Sig. (2-tailed)	.725

a. Grouping Variable: Area of residence

Figure 25.2 SPSS rank-sum test output

The first thing you will notice is that the output is titled **Mann-Whitney Test**. As we show in the Appendix to this chapter (for those interested) the Mann–Whitney and Wilcoxon tests are equivalent ways of reaching the same conclusion.

The **Ranks** table provides the relevant descriptive statistics for the two samples we are comparing: the number of cases in each and in total, the mean ranks, and the sum of ranks. These figures all correspond to the values we calculated by hand above.

Below this descriptive information is the table providing the **Test Statistics**. The footnote indicates the samples have been formed on the basis of their area of residence variable. The value for Wilcoxon W is 480.000 (the smallest of the two rank sums from the **Ranks** table). The value for Z is –.352, which is the same as our sample *z*-score calculated above. This has a two-tail probability, if the null hypothesis of no difference is true, of .725. Clearly, the difference between the samples of rural and urban residents should not be taken to indicate a difference between the population of rural and the population of urban residents. We do not reject the null hypothesis of no difference.

The Wilcoxon signed-ranks *z*-test for two dependent samples

The previous section detailed the process of comparing *independent* samples using the ranks of the cases rather than the raw scores. Similar principles apply when working with ranked data and we want to compare *dependent* samples. The **Wilcoxon signed-ranks test** compares two dependent samples, using the ranks of the *pairs* of scores formed by the matched pairs in the samples. This is analogous to the relationship between the independent samples *t*-test for the equality of means and the dependent samples *t*-test for the mean difference.

For example, assume that the researcher who conducted the McNemar test in Chapter 24 to assess people's attitude to video games is dissatisfied with the results. The researcher suspects that playing video games really does affect a person's attitude to the educational value of such games, and that the simple binomial scale used in the original study was not sensitive enough to detect this change. The researcher therefore conducts another study involving 15 people who are asked to rate on a 10-point scale whether they believe video games have any educational value, with 1 indicating no educational value and 10 indicating very high educational value. Each of these 15 people is then asked to play a variety of video games and again rate whether they believe video games are of educational benefit. What effect does actually playing the game have on opinion?

The scores for each person, before and after playing, are recorded in Table 25.7, together with the difference, for each pair.

Table 25.7 Rating of video games before and after use

Person	Before	After	Difference in scores
1	3	5	+2
2	5	5	0
3	2	8	+6
4	3	4	+1
5	8	7	−1
6	6	3	−3
7	4	4	0
8	7	6	−1
9	2	7	+5
10	6	7	+1
11	1	9	+8
12	9	7	−2
13	8	1	−7
14	5	5	0
15	6	2	−4

The first step is to exclude the cases with no change in scores, which are those shaded in the table. As with the McNemar test, cases that show no change are not used in the analysis. Here cases 2, 7, and 14 record no change in their scores before and after.

It would be tempting simply to calculate an average change in scores and conduct a *t*-test on the difference. However, we are working with an ordinal scale and such averages are not appropriate. Instead we take a slightly more difficult route. We rank the cases, starting with those registering the smallest change in scores (these will be cases 4, 8, 5, and 10, which each registered a change of ±1) and continuing through to the case with the largest change (case 11 with a change of 8). Pairs, that is, are ordered according to the absolute difference between their 'Before' and 'After' scores (Table 25.8).

Table 25.8 Ordering of all non-tied pairs

Pair number	4	5	8	10	1	12	6	15	9	3	13	11
Difference	+1	−1	−1	+1	+2	−2	−3	−4	+5	+6	−7	+8
Rank	2.5	2.5	2.5	2.5	5.5	5.5	7	8	9	10	11	12

Notice that cases that have the same absolute change in scores have been assigned an average rank. For example, four cases each changed their score by one point on the scale. Since collectively these cases occupy ranks 1, 2, 3, and 4, the average rank for these four cases is 2.5:

$$\frac{1+2+3+4}{4} = 2.5$$

If playing video games has no effect on attitudes regarding their educational value, there should not be a tendency for pairs with either positive or negative changes to bunch up at one end of the ranking or the other. Another way of assessing this is to compare the rank sum for pairs registering a positive change in attitude to the rank sum for pairs registering a negative change in attitude. If the positive and negative changes are equally distributed through the ranks, the sum of these ranks will be equal, and can be calculated using the formula:

$$\mu_T = \frac{n(n+1)}{4}$$

The value of μ_T is the rank sum we expect from samples drawn from a population where attitude to video games does not change systematically in one direction or the other, and is the value we use in stating the null hypothesis. For these data we obtain:

$$\mu_T = \frac{12(12+1)}{4} = 39$$

The null hypothesis in this instance will therefore be:

$$H_0: \mu_T = 39$$

However, even if this is the case, random samples drawn from such a population will not always produce a value of 39. We need to compare this hypothesized value with the sample statistic we obtain, and assess whether any difference can be attributed to random variation.

We derive this sample statistic by separating out those cases that have a positive change (increase) in their score after playing the video games from those cases that have a negative change (reduction) in score. We then sum the ranks for each group (Table 25.9).

Table 25.9 Ordering of pairs

Pair number	4			10	1				9	3		11
Difference	+1			+1	+2				+5	+6		+8
Positive rank	2.5			2.5	5.5				9	10		12
Pair number		5	8			12	6	15			13	
Difference		−1	−1			−2	−3	−4			−7	
Negative rank		2.5	2.5			5.5	7	8			11	

In this example we have rank sums of 36.5 and 41.5. What is the probability of obtaining such a sample result if the null hypothesis is true? The sample statistic, called Wilcoxon's *T*, is the smallest rank sum, which in this case is the rank sum for the positives. We conduct a *z*-test on the difference between the value of μ_T and the sample value, *T*, where:

$$z_{sample} = \frac{T - \mu_T}{\sigma_T}$$

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}}$$

If we substitute the data from the example into these equations, we get $z_{sample} = -0.2$:

$$\sigma_T = \sqrt{\frac{n(n+1)(2n+1)}{24}} = \sqrt{\frac{12(12+1)(2\times 12+1)}{24}} = 12.75$$

$$z_{sample} = \frac{T - \mu_T}{\sigma_T} = \frac{36.5 - 39}{12.75} = -0.2$$

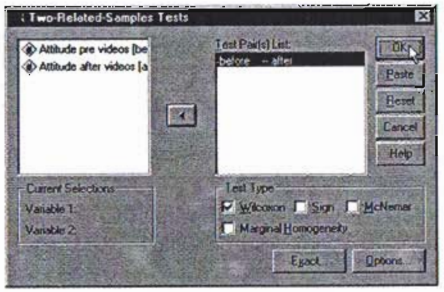
This value for *z*, from the table for the area under the standard normal curve, has a two-tail probability of 0.8445. We cannot reject the null hypothesis since the differences observed in the pairs could easily come about through sampling error when drawing from a population in which playing video games has no effect on attitude to their educational value.

The Wilcoxon signed-ranks test using SPSS

The actions required to conduct this test in SPSS are listed in Table 25.10 and Figure 25.3, which also presents the output from this command.

Table 25.10 Wilcoxon signed-ranks test on SPSS (file: **Ch25-2.sav**)

SPSS command/action	Comments
1 From the menu select **Analyze/ Nonparametric Tests/ 2 Related Samples**	This brings up a window headed **Two-Related-Samples Tests**. You will notice that in the area to the bottom left of the window headed **Test Type** the small square next to **Wilcoxon** is selected. This indicates that the Wilcoxon test for two dependent samples is the default test
2 Click on **after** and while holding down the command key click on **before**	These two variable names will be highlighted
3 Click on ▸	This pastes the highlighted variables into the area headed **Test Pairs List**:
4 Click on **OK**	



Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
Attitude after videos - Attitude pre videos	Negative Ranks	6[a]	6.08	36.50
	Positive Ranks	6[b]	6.92	41.50
	Ties	3[c]		
	Total	15		

a. Attitude after videos < Attitude pre videos

b. Attitude after videos > Attitude pre videos

c. Attitude pre videos = Attitude after videos

Test Statistics[b]

	Attitude after videos - Attitude pre videos
Z	-.197[a]
Asymp. Sig. (2-tailed)	.844

a. Based on negative ranks.

b. Wilcoxon Signed Ranks Test

Figure 25.3 The SPSS **Two-Related-Samples Tests** dialog box and output

The SPSS output gives us the same results as those we calculated by hand. As with many of the other tests we have covered, the first part of the output presents the descriptive statistics for the samples, followed by the information from the inference test. In the table labelled **Ranks** we first see that there are six pairs that registered an increase in score after playing videos (Attitude after videos > Attitude pre videos). There are also six pairs that registered a decrease (Attitude after videos < Attitude pre videos), and three pairs whose score did not change (Attitude after videos = Attitude pre videos).

Second, SPSS calculates the mean ranks and rank sums for the positives and negatives. If we multiply these mean ranks by the number of cases in each group we get the sum of ranks:

$$\Sigma R_- = 6.92 \times 6 = 41.5$$

$$\Sigma R_+ = 6.08 \times 6 = 36.5$$

The **Test Statistics** table, which contains the information on the *z*-test for the rank sums, indicates that we should not reject the null hypothesis, given that the probability of .844 is greater than the alpha level of 0.05. In other words, even if playing video games makes no difference in attitude toward their educational value, we will still get sample results with this amount of difference or greater more than 8 times out of 10.

Other non-parametric tests for two or more samples

This chapter has worked through one of the most common non-parametric tests: the Wilcoxon test for two independent samples. The other common non-parametric test is the chi-square test, which we introduced in previous chapters. A researcher, in fact, can tackle most problems with a sound knowledge of the Wilcoxon and the chi-square tests. However, there are many other non-parametric tests available, to which some reference should be made. Indeed, the attentive reader will have noticed that SPSS offered a number of choices in the Test Type area when conducting a test of two independent samples. This range of choices is further extended when we consider situations where more than two samples are being compared.

Kruskal–Wallis H test on more than two samples

The Wilcoxon test compares two samples in terms of a variable measured at least at the ordinal level. In the example, we had a sample of rural and a sample of urban residents. But what if we have more than two samples that we want to compare? What if we want to compare urban, rural, and semi-rural residents, rather than just urban and rural residents?

One way of doing this is simply to conduct multiple Wilcoxon rank-sum tests, using all the possible combinations of samples:

Urban by Rural
Urban by Semi-Rural
Rural by Semi-Rural

Thus with three samples to compare we will need to undertake three separate two-sample Wilcoxon tests. In practical terms, on SPSS, this will involve specifying under **Define Groups**, one test at a time, each possible combination of values for the grouping variable, and then rerunning the test. This is obviously a cumbersome procedure.

When we have more than two samples, a more direct path is to conduct a Kruskal–Wallis *H* test. The Kruskal–Wallis test compares all possible combinations of the samples in one test. It has very similar logic to the Wilcoxon test, in that it compares rank sums for each sample being compared. The test statistic, though, is no longer a z-score. The Kruskal–Wallis test uses a chi-square test to assess the null hypothesis that the populations have the same distribution in some ordinal scale. It is available in SPSS as part of the **Analyze/Nonparametric Tests/K Independent Samples** command.

The difference between the Wilcoxon W and Kruskal–Wallis H tests is analogous to the difference between a two-samples t-test and an ANOVA. These latter tests compare the relevant number of samples in terms of the difference between means, whereas the W and H tests compare samples in terms of rank sums.

Wald–Wolfowitz runs test

This test uses the same logic as the one-sample runs test we introduced in Chapter 21. It can be used in similar situations to the Wilcoxon test, where the cases in the two samples are pooled and ordered in terms of their scores on an ordinal scale. The number of runs of cases from each same sample is counted, and this number of runs is the sample statistic tested. In the extreme case, using the example above, all 22 rural residents will be at one end of the distribution and all 22 urban residents at the other end, thus forming only two runs. Such a sample result will strongly suggest that the two populations are different in terms of this ordinal scale. On the other hand, if the two samples were scattered throughout the combined distribution, the number of runs will be much higher. The Wald–Wolfowitz runs test conducts a z-test on the difference between the number of runs from the samples and the expected number of runs, if the null hypothesis of no difference is true. It is available in SPSS as part of the **Analyze/Nonparametric Tests/2 Independent Samples** command. One limitation of this test though is that it is seriously affected by tied ranks.

Appendix: the Mann–Whitney *U* test

In generating the results of the Wilcoxon test on SPSS, we actually clicked on the box under **Test Type** next to **Mann-Whitney**. The SPSS output produced, along with the Wilcoxon W, another statistic called a Mann–Whitney U. This is also common in many textbooks, and is based on a slightly different calculation. Since it is a little more complicated than simply looking at the sum of the ranks, and will always result in the same probability value as the Wilcoxon rank-sum test, we have detailed the latter in the text. However, for those who are interested, the logic of the Mann–Whitney U is presented here. In the example, the 44 respondents in the sample were lined up from highest to lowest rank on the conservatism scale. We ask the rural resident who scored 39 to step out of the line and count how many urban residents he or she is ranked above. This of course will be all 22 urban residents. We then ask the second highest ranked rural dweller to step out of the line and count how many urbanites he or she is ahead of in the line. If we get each rural resident to do this and add up all the figures, the number obtained is the sample U-statistic. This sample statistic can be calculated for any sample using the following formula:

$$U = n_1 n_2 + \frac{n_1(n_1+1)}{2} - \Sigma R_1$$

where n_1 is the smaller of the two samples, n_2 is the larger of the two samples, and ΣR_1 is the sum of ranks for the smaller sample.

If the two samples came from populations that were not different in terms of this variable, then we would on average randomly select samples that produced a U-statistic given by the following formula:

$$\mu_U = \frac{n_1 n_2}{2}$$

In this example, the expected value of U will be:

$$\mu_U = \frac{22(22)}{2} = 242$$

From the SPSS output we see that the sample U is 227. We can conduct a z-test to see if the difference between the sample and expected values of U is large enough to warrant the rejection of the null hypothesis.

$$z_{sample} = \frac{U - \mu_u}{\sigma_u}$$

where:

$$\sigma_U = \sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}$$

The z-score obtained will be exactly the same as that derived from conducting a Wilcoxon test on the same data, and therefore, regardless of the test used, the conclusion regarding the null hypothesis will be the same.

Exercises

25.1 Determine the correct rank for the score of 10 in each of the following series:

(a) 2, 9, 17, 10, 11, 6
(b) 2, 9, 17, 10, 11, 6, 8
(c) 2, 9, 17, 10, 10, 11, 6
(d) 2, 9, 17, 10, 10, 11, 6, 8, 11
(e) 3, 20, 15, 10, 22, 4, 10, 9, 16, 10

25.2 Identify and assign the correct rank to the score immediately following 10 in each of the following rank-ordered series:

(a) 2, 6, 9, 10, 11, 17
(b) 2, 6, 8, 9, 10, 11, 17
(c) 2, 6, 9, 10, 10, 11, 17
(d) 2, 6, 8, 9, 10, 10, 11, 11, 17
(e) 3, 4, 9, 10, 10, 10, 15, 16, 20, 22

25.3 When comparing two samples, under what conditions will you use a Wilcoxon z-test for the rank sum rather than a t-test for the equality of means?

25.4 (a) Order the following data, assigning ranks to each case:

Group 1	Group 2
1	12
15	25
12	29
16	8
23	15
9	20
11	7

(b) What are the rank sums and mean ranks for each group?
(c) Which rank sum is the sample statistic for conducting the Wilcoxon test?
(d) Calculate the value for μ_W.
(e) Conduct a Wilcoxon test to assess whether there is a significant difference in ranks.

25.5 A trial is used to evaluate the effectiveness of a specific exercise program to improve standing up performance of individuals who have suffered a stroke. Twenty subjects are randomly assigned to either a treatment or control group, and their individual scores on a Motor Assessment Scale (MAS), which measures standing up performance for stroke patients on a scale of 0 to 6, are recorded:

Treatment group		Control group	
Subject	MAS	Subject	MAS
1	0	11	3
2	4	12	1
3	5	13	0
4	6	14	2
5	4	15	3
6	4	16	6
7	6	17	1
8	3	18	2
9	6	19	2
10	2	20	2

Using the Wilcoxon rank-sum test assess the effectiveness of the exercise program. Enter these data on SPSS and conduct the test.

25.6 Enter into SPSS the data in Table 25.1 for the example in the text for the comparison of Australian and British children.

(a) Conduct a Wilcoxon rank-sum test on these data and compare the results with the calculations in the text.
(b) The following data are the viewing intensities for a sample of 23 American children:

5, 8, 16, 21, 26, 35, 39, 45, 45, 54, 59, 61, 78, 79, 83, 85, 85, 90, 97, 99

Add these data to the SPSS file and conduct another Wilcoxon rank-sum test to see if there is a significant difference between British and American children.
(c) Conduct the same Wilcoxon test by hand and compare your results with the SPSS output.

25.7 Use the **Employee data** file to determine whether there is a significant difference in the starting salaries of employees based on their minority status. Why might we use the Wilcoxon test rather than the two-sample t-test to make this comparison?

25.8 The following are scores of 8 matched pairs in a before-and-after experiment. Use the Wilcoxon signed-ranks test to assess whether there is a difference.

Before	After
75	65
63	67
82	51
37	43
46	47
59	61
39	52
33	85

25.9 Ten people are asked to rate the effectiveness of two training programs, with 1 equal to 'Very poor' and 10 equal to 'Very good'. The responses are summarized in the table over the page.

(a) Can we say that one program is preferred over another, at a 0.01 level of significance?
(b) Enter these data into SPSS and confirm your results.

Program 1	Program 2
3	5
2	6
3	7
2	7
1	4
4	2
5	5
1	8
6	9

26

The *t*-test for a correlation coefficient

The last test we will detail corresponds to the descriptive statistics we covered in Chapter 12. In Chapter 12 we calculated the regression line, and the correlation statistics that go along with it, for a set of cases measured in terms of two interval/ratio scales. We introduced these descriptive statistics in the context of investigating the relationship between unemployment and civil unrest across cities. The result we arrived at was:

$$\text{civil unrest} = 4.4 + 0.53(\text{unemployment rate})$$

$$r = 0.81$$

These statistics tell us that in *our sample* there is a strong, positive association between civil unrest and unemployment rates. But this is a result that obtains in the sample, and therefore might not reflect what is happening *in all cities*. As with any other descriptive statistics that we may calculate for a sample, we need to determine whether the correlation coefficient that describes the sample data reflects the population from which it is drawn. There may be no correlation between these variables in the population of all cities (r_μ) and it is only sampling error that has caused us to select five cities that are not like the rest. We therefore need to conduct an inference test on the value of the correlation coefficient we have obtained.

The *t*-test for Pearson's correlation coefficient

The null hypothesis for this test is that there is no correlation in the population, whereas the alternative hypothesis is that there is some correlation:

$$H_0: r_\mu = 0$$

$$H_a: r_\mu \neq 0$$

Obviously the sample correlation coefficient of 0.81 does not conform to the null hypothesis. But can we reject the hypothesis of no correlation in the population on the basis of this sample result? What is the probability of obtaining a sample of five cities with a correlation between civil disturbances and unemployment of 0.81 from a population where the correlation is zero?

To obtain this probability we conduct a *t*-test, using the following formulas:

$$t_{sample} = \frac{r - r_\mu}{s_r}$$

$$s_r = \sqrt{\frac{1-r^2}{n-2}}$$

If we substitute the sample values for r into this equation, we get $t_{sample} = 2.38$:

$$s_r = \sqrt{\frac{1-r^2}{n-2}} = \sqrt{\frac{1-(0.81)^2}{5-2}} = 0.34$$

$$t_{sample} = \frac{r - r_\mu}{s_r} = \frac{0.81-0}{0.34} = 2.38$$

In determining the *p*-score for this test statistic we refer to Table 26.1, which presents critical values of *t* for a range of degrees of freedom (*df*). For this test *df* = *n* – 2.

Table 26.1 Critical values for *t*-distributions

	Level of significance for one-tail test				
	0.10	0.05	0.02	0.01	0.005
	Level of significance for two-tail test				
df	0.20	0.10	0.05	0.02	0.01
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
⋮	⋮	⋮	⋮	⋮	⋮
∞	1.282	1.645	1.960	2.326	2.576

The *p*-score lies somewhere between 0.10 and 0.05; in fact we can see that it is almost equal to 0.10. It is important to stop and consider what has happened. In the sample we measured a strong positive correlation between unemployment and civil unrest. The inference test tells us that despite this the sample result might be due to chance when sampling from a population where these variables are not correlated. To see why we cannot conclude the sample reflects a relationship in the population, it is helpful to look again at the scatter plot (Figure 26.1).

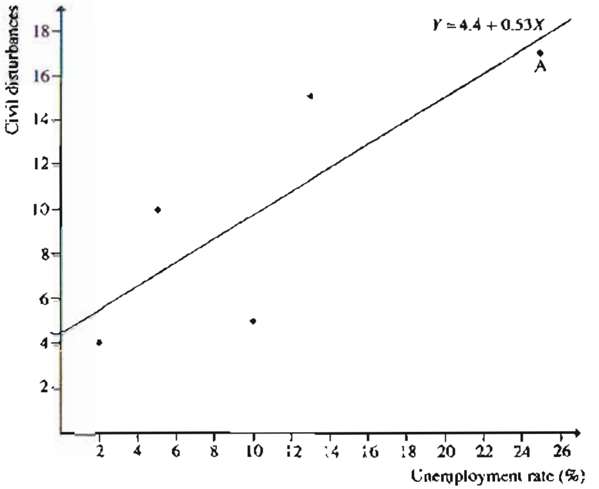


Figure 26.1 The OLS regression line

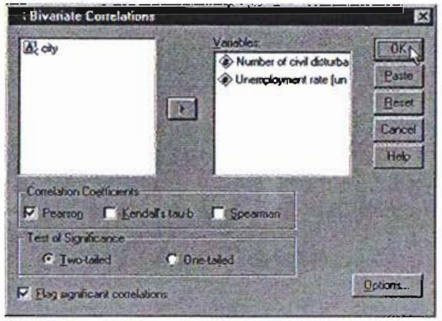
We can see that the regression line has been heavily influenced by the one score for City A with an unemployment rate of 25 percent and 17 civil disturbances. Because we are working with such a small sample ($n = 5$) one extreme case can throw out the results for the whole sample. If this one score was different, the regression line would also be very different. Since small samples are so easily influenced by scores that are outliers for either variable, or for the two variables jointly, even strong correlations may not turn out to be significant when working with very small samples.

Testing the significance of Pearson's correlation coefficient using SPSS

The test of significance for Pearson's product moment correlation coefficient is generated as part of the output when conducting a regression analysis. The procedures we followed in Chapter 10 for generating regression statistics therefore are the same as those for generating the necessary information for conducting an inference test on these statistics. There is also an alternative means by which we can generate the correlation coefficient between two variables and the associated t-score and significance level. This is through the **Bivariate Correlations** command (Table 26.2, Figure 26.2). We use this command if we wish to generate just the correlation coefficient without all the additional information that comes with a complete regression analysis.

Table 26.2 Bivariate Correlation with a t-test using SPSS (file: **Ch26.sav**)

SPSS command/action	Comments
1 From the menu select **Analyze/Correlate/Bivariate**	This brings up the **Bivariate Correlations** window
2 Click on **Number of civil disturbances**	This highlights **Number of civil disturbances**
3 Click on the ▸ that points to the **Variables:** target list	This pastes **Number of civil disturbances** into the **Variables:** target list
4 Click on **Unemployment rate**	This highlights **Unemployment rate**
5 Click on the ▸ that points to **Variables:** target list	This pastes **Unemployment rate** into the **Variables:** target list
6 Click on OK	



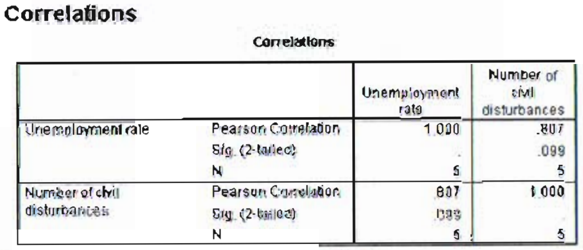
Correlation Coefficients: ☑ Pearson ☐ Kendall's tau-b ☐ Spearman

Test of Significance: ● Two-tailed ○ One-tailed

☑ Flag significant correlations

Figure 26.2 The Bivariate Correlations dialog box

Notice that the radio button under **Test of Significance** and next to **two-tailed** is selected indicating that a two-tail t-test is the default setting. This command will generate the output in Figure 26.3.



Correlations

Correlations

		Unemployment rate	Number of civil disturbances
Unemployment rate	Pearson Correlation	1.000	.807
	Sig. (2-tailed)	.	.099
	N	5	5
Number of civil disturbances	Pearson Correlation	.807	1.000
	Sig. (2-tailed)	.099	.
	N	5	5

Figure 26.3 SPSS bivariate correlation output

With two variables being correlated the table produces four correlation coefficients. One is between Unemployment rate and itself and the other is between Number of civil disturbances and itself, each of which produces a coefficient of 1.000. This is necessarily so since any variable is perfectly correlated with itself. In the first row the table also provides the correlation between Unemployment rate and Number of civil disturbances, which is .807, and the significance of this coefficient which is .099. This indicates that the coefficient, despite its strength, is not significant at the 0.05 level, and therefore could be the result of sampling variation. The next row of the table provides the correlation coefficient between Number of civil disturbances and Unemployment rate, which is exactly the same as that in the first row of the table since it is the same correlation looked at the other way and Pearson's r is symmetric.

The t-test for Spearman's rank-order correlation coefficient

We have, in earlier chapters, learnt the techniques for calculating two different correlation coefficients: Pearson's r and Spearman's rho (r_s). The former is used to investigate the association between two variables measured at the interval/ratio level, whereas the latter is used when at least one of the two variables is measured on an ordinal scale. However, if we look closely at the procedures for calculating the two types of correlation coefficients we see that they are almost identical. The difference is that Pearson's r uses the raw data in the computations, whereas Spearman's rho is calculated on the *ranks* of the data.

Given the basic mathematical equivalence between the two measures of correlation, the test of significance for each is the same. That is, the formula for calculating the sample t-score is the same regardless of whether we are testing for the significance of Pearson's correlation coefficient or Spearman's correlation coefficient, where ρ is the hypothesized value for Spearman's correlation coefficient for the population:

$$t_{sample} = \frac{r_s - \rho}{s_r}$$

$$s_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

To see how we conduct a test for Spearman's rho we will use the five-step hypothesis testing procedure on the example we introduced in Chapter 12.

Step 1: State the null and alternative hypotheses

H_0: There is no correlation between age and mobility scores.

$$H_0 \ \rho = 0$$

H_a: There is a correlation between age and mobility scores.

$$H_a: \rho \neq 0$$

Step 2: Choose the test of significance

Since we are investigating the correlation between two variables measured at the ordinal level, the data have been described by calculating Spearman's rho. The appropriate inference test is therefore the *t*-test for a correlation coefficient.

Step 3: Describe the sample and derive the p-score

The correlation between age and mobility scores for 16 physiotherapy patients is:

$$r_s = -0.8$$

To see whether this correlation might result from random variation when sampling from a population where these variables are not correlated, we first need to calculate the standard error for the sampling distribution of rho:

$$s_r = \sqrt{\frac{1-r_s^2}{n-2}} = \sqrt{\frac{1-(-0.8)^2}{16-2}} = 0.16$$

The sample *t*-score will therefore be –5:

$$t_{sample} = \frac{r_s - \rho}{s_r} = \frac{-0.8-0}{0.16} = -5$$

Step 4: Determine at what alpha level, if any, the result is statistically significant

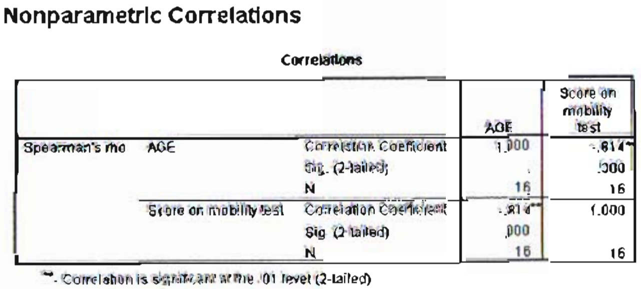
From the table for critical values of the *t*-distribution we see that the test statistic of –5 is significant at even the lowest reported level in the table of 0.01.

Step 5: Report results

A physiotherapist uses a new treatment on a group of 16 patients and is interested in whether their age affects their ability to respond to the treatment. After treatment each patient is given a mobility score out of 15, according to his or her ability to perform a number of tasks. On the basis of the strong, negative relationship we find in the sample of 16 patients (Spearman's rho = –0.8), we reject the hypothesis that there is no correlation between age and mobility scores ($t = -5$, $p < 0.01$, two-tail). These variables do seem to be related such that older patients do not respond as well to the treatment.

Testing the significance of Spearman's correlation coefficient using SPSS

As with testing for the significance of Pearson's *r*, the relevant inferential statistics are automatically generated when we ask SPSS to calculate Spearman's rho. Thus the procedures we introduced in Chapter 12, page 181, provide the relevant information. Here we reproduce the output from that SPSS command so that we interpret the relevant portion of it for hypothesis testing (Figure 26.4).



Nonparametric Correlations

Correlations

			AGE	Score on mobility test
Spearman's rho	AGE	Correlation Coefficient	1.000	-.814**
		Sig. (2-tailed)	.	.000
		N	16	16
	Score on mobility test	Correlation Coefficient	-.814**	1.000
		Sig. (2-tailed)	.000	
		N	16	16

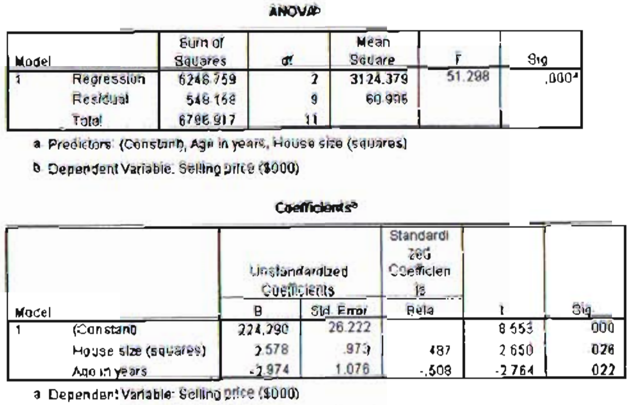
**. Correlation is significant at the .01 level (2-tailed).

Figure 26.4 The SPSS Bivariate Correlations output

SPSS calculates a correlation coefficient between each variable selected and all the other variables selected in the target list, including itself. Thus with only two variables selected in this example, we end up with four correlations: age with mobility score, mobility score with age, age with age, mobility score with mobility score. The correlations for each variable and itself are irrelevant since any variable is always perfectly correlated with itself – hence the value of 1.000 in the SPSS table. The correlation for age and score on mobility test is –.814. This is the same as the correlation for score on mobility test and age, since it is exactly the same relationship. Notice the ** next to this correlation coefficient. As the footnote to the table states, ** signals a value for rho that is significant at the 0.01 level on a two-tail test. In fact we can see that the exact significance is reported to be .000, which indicates that less than 5 in every 10,000 samples drawn from a population where these variables are not related will have a correlation coefficient this strong or stronger.

Testing for significance in multiple regression

A bivariate correlation is reasonably straightforward in terms of testing for the significance of the correlation coefficient. In a multivariate analysis, however, such as that we undertook in Chapter 13, the problem of inference is a little more complex. We will repeat the way we interpret the statistical significance of SPSS multiple regression output that we presented on page 191. The relevant portion of the regression output is again presented as Figure 26.5.



ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6248.759	2	3124.379	51.288	.000^a
	Residual	548.158	9	60.906		
	Total	6796.917	11			

a. Predictors: (Constant), Age in years, House size (squares)

b. Dependent Variable: Selling price ($000)

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	224.290	26.222		8.553	.000
	House size (squares)	2.578	.973	.487	2.650	.026
	Age in years	-2.974	1.076	-.508	-2.764	.022

a. Dependent Variable: Selling price ($000)

Figure 26.5 SPSS Linear Regression output

Exercises

26.1 In Chapter 12 we calculated the rank-order correlation coefficient for a set of 15 students to see if there is a relationship between performance in exams and performance in presentations. The correlation coefficient was found to be –0.26. Assuming that these data came from all the students in the class, do we need to conduct an inference test?

26.2 In Exercise 12.16 you were asked to calculate the value for rho for a sample of wines, relating price to quality. Conduct a *t*-test to assess whether the result reflects a non-zero correlation for the population of all wines. Check your SPSS output to confirm your results.

26.3 A survey of employed workers found that the correlation coefficient between the number of years of post-secondary education and current annual income measured in dollars is 0.54. The sample size for this survey was 140. The significance of this correlation coefficient was tested using a *t*-test, which gave a *t*-value of 7.54. What conclusion should be drawn about the nature of the relationship between these two variables?

26.4 The firm from which the **Employee data** file is generated is interested in whether starting salaries are correlated with current salaries. Generate the necessary information to determine whether any observed correlation in the sample is due to sampling variation or whether it reflects an underlying relationship for all employees in the firm.

Appendix

Table A1 Area under the standard normal curve

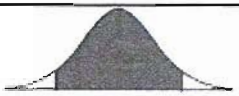
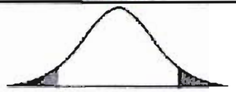
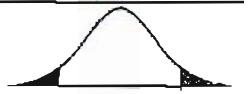
z	Area under curve between both points	Area under curve beyond both points (two tails)	Area under curve beyond one point (one tail)
			
±0.1	0.080	0.920	0.4600
±0.2	0.159	0.841	0.4205
±0.3	0.236	0.764	0.3820
±0.4	0.311	0.689	0.3445
±0.5	0.383	0.617	0.3085
±0.6	0.451	0.549	0.2745
±0.7	0.516	0.484	0.2420
±0.8	0.576	0.424	0.2120
±0.9	0.632	0.368	0.1840
±1	0.683	0.317	0.1585
±1.1	0.729	0.271	0.1355
±1.2	0.770	0.230	0.1150
±1.3	0.806	0.194	0.0970
±1 4	0.838	0.162	0.0810
±1.5	0.866	0.134	0.0670
±1.6	0.890	0.110	0.0550
±1.645	0.900	0.100	0.0500
±1.7	0.911	0.089	0.0445
±1.8	0.928	0.072	0.0360
±1.9	0.943	0.057	0.0290
±1.96	0.950	0.050	0.0250
±2	0.954	0.046	0.0230
±2.1	0.964	0.036	0.0180
±2.2	0.972	0.028	0.0140
±2.3	0.979	0.021	0.0105
±2.33	0.980	0.020	0.0100
±2.4	0.984	0.016	0.0080
±2.5	0.988	0.012	0.0060
±2.58	0.990	0.010	0.0050
±2.6	0.991	0.009	0.0045
±2.7	0.993	0.007	0.0035
±2.8	0.995	0.005	0.0025
±2.9	0.996	0.004	0.0020
±3	0.997	0.003	0.0015
±3.1	0.998	0.002	0.0001
±3.2	0.9986	0.0014	0.0007
±3.3	0.9990	0.0010	0.0005
±3.4	0.9993	0.0007	0.0003
±3.5	0.9995	0.0005	0.00025
±3.6	0.9997	0.0003	0.00015
±3.7	0.9998	0.0002	0.00010
±3.8	0.99986	0.00014	0.00007
±3.9	0 99990	0.00010	0.00005
±4	>0 99990	<0.00010	<0.00005

Table A2 Critical values for *t*-distributions

	Level of significance for one-tail test				
	0.10	0.05	0.02	0.01	0.005
	Level of significance for two-tail test				
df	0.20	0.10	0.05	0.02	0.01
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.340	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
35	1.306	1.690	2.030	2.438	2.724
40	1.303	1.684	2.021	2.423	2.704
45	1.301	1.679	2.014	2.412	2.590
50	1.299	1.676	2.009	2.403	2.678
55	1.297	1.673	2.004	2.396	2.668
60	1.296	1.671	2.000	2.390	2.660
70	1.294	1.667	1.994	2.381	2.648
80	1.292	1.664	1.990	2.374	2.639
90	1.291	1.662	1.987	2.368	2.632
120	1.289	1.658	1.980	2.358	2.617
∞	1.282	1.645	1.960	2.326	2.576

Table A3 Critical values for *F*-distributions ($\alpha = 0.05$)

		Degrees of freedom for estimates of variance between samples $k-1$									
	$n-k$	1	2	3	4	5	6	7	8	9	∞
Degrees of freedom for estimates of variance within samples	1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	254.3
	2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.50
	3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.84	8.81	8.53
	4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.63
	5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.36
	6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	3.67
	7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.23
	8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	2.93
	9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	2.71
	10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.54
	11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.40
	12	4.75	3.88	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.30
	13	4.67	3.80	3.41	3.18	3.02	2.92	2.83	2.77	2.71	2.21
	14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.13
	15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.07
	16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.01
	17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	1.96
	18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	1.92
	19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	1.88
	20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	1.84
	21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	1.81
	22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	1.78
	23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.38	2.32	1.76
	24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	1.73
	25	4.24	3.38	2.99	2.76	2.60	2.49	2.40	2.34	2.28	1.71
	26	4.22	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	1.69
	27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.30	2.25	1.67
	28	4.20	3.34	2.95	2.71	2.56	2.44	2.36	2.29	2.24	1.65
	29	4.18	3.33	2.93	2.70	2.54	2.43	2.35	2.28	2.22	1.64
	30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	1.62
	40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	1.51
	50	4.03	3.18	2.79	2.56	2.38	2.29	2.20	2.13	2.07	1.44
	60	4.00	3.15	2.76	2.52	2.37	2.25	2.17	2.10	2.04	1.39
	80	3.96	3.11	2.72	2.48	2.33	2.21	2.12	1.99	1.91	1.32
	100	3.94	3.09	2.70	2.46	2.30	2.19	2.10	2.03	1.97	1.28
	120	3.92	3.07	2.68	2.45	2.29	2.17	2.09	2.02	1.96	1.25
	∞	3.84	2.99	2.60	2.37	2.21	2.09	2.01	1.94	1.88	1.00

Table A4 Critical values for chi-square distributions

	Level of significance									
df	0.99	0.90	0.70	0.50	0.30	0.20	0.10	0.05	0.01	0.001
1	0.00016	0.0158	0.148	0.455	1.074	1.642	2.706	3.841	6.635	10.827
2	0.0201	0.211	0.713	1.386	2.408	3.219	4.605	5.991	9.210	13.815
3	0.115	0.584	1.424	2.366	3.665	4.642	6.251	7.815	11.341	16.268
4	0.297	1.064	2.195	3.357	4.878	5.989	7.779	9.488	13.277	18.465
5	0.554	1.610	3.000	4.351	6.064	7.289	9.236	11.070	15.086	20.517
6	0.872	2.204	3.828	5.348	7.231	8.558	10.645	12.592	16.812	22.457
7	1.239	2.833	4.671	6.346	8.383	9.803	12.017	14.067	18.475	24.322
8	1.646	3.490	5.527	7.344	9.524	11.030	13.362	15.507	20.090	26.125
9	2.088	4.168	6.393	8.343	10.656	12.242	14.684	16.919	21.666	27.877
10	2.558	4.865	7.267	9.342	11.781	13.442	15.987	18.307	23.209	29.588
11	3.053	5.578	8.148	10.341	12.899	14.631	17.275	19.675	24.725	31.264
12	3.571	6.304	9.034	11.340	14.011	15.812	18.549	21.026	26.217	32.909
13	4.107	7.042	9.926	12.340	15.119	16.985	19.812	22.362	27.688	34.528
14	4.660	7.790	10.821	13.339	16.222	18.151	21.064	23.685	29.141	36.123
15	5.229	8.547	11.721	14.339	17.322	19.311	22.307	24.996	30.578	37.697
16	5.812	9.312	12.624	15.338	18.418	20.465	23.542	26.296	32.000	39.252
17	6.408	10.085	13.531	16.338	19.511	21.615	24.769	27.587	33.409	40.790
18	7.015	10.865	14.440	17.338	20.601	22.760	25.989	28.869	34.805	42.312
19	7.633	11.651	15.352	18.338	21.689	23.900	27.204	30.144	36.191	43.820
20	8.260	12.443	16.266	19.337	22.775	25.038	28.412	31.410	37.566	45.315
21	8.897	13.240	17.182	20.337	23.858	26.171	29.615	32.671	38.932	46.797
22	9.542	14.041	18.101	21.337	24.939	27.301	30.813	33.924	40.289	48.268
23	10.196	14.848	19.021	22.337	26.018	28.429	32.007	35.172	41.638	49.728
24	10.856	15.659	19.943	23.337	27.096	29.553	33.196	36.415	42.980	51.179
25	11.524	16.473	20.867	24.337	28.172	30.675	34.382	37.652	44.314	52.620
26	12.198	17.292	21.792	25.336	29.246	31.795	35.563	38.885	45.642	54.052
27	12.879	18.114	22.719	26.336	30.319	32.912	36.741	40.113	46.963	55.476
28	13.565	18.939	23.647	27.336	31.391	34.027	37.916	41.337	48.278	56.893
29	14.256	19.768	24.577	28.336	32.461	35.139	39.087	42.557	49.588	58.302
30	14.953	20.599	25.508	29.336	33.530	36.250	40.256	43.773	50.892	59.703

Table A5 Sampling errors for a binomial distribution (95% confidence level)

Sample size	Binomial percentage distribution					
	50/50	60/40	70/30	80/20	90/10	95/5
50	13.3	13.1	12.4	11.1	9.0	7.4
100	9.6	9.4	8.9	7.8	6.1	4.8
150	7.9	7.7	7.3	6.4	4.9	3.8
200	6.9	6.7	6.3	5.5	4.3	3.2
250	6.2	6.0	5.7	5.0	3.8	2.9
300	5.6	5.5	5.2	4.5	3.4	2.6
400	4.9	4.8	4.5	3.9	3.0	2.2
500	4.4	4.3	4.0	3.5	2.7	2.0
600	4.0	3.9	3.7	3.2	2.4	1.8
700	3.7	3.6	3.4	3.0	2.2	1.6
800	3.5	3.4	3.2	2.8	2.1	1.5
900	3.3	3.2	3.0	2.6	2.0	1.4
1000	3.1	3.0	2.8	2.5	1.9	1.4
1100	2.9	2.9	2.7	2.4	1.8	1.3
1200	2.8	2.8	2.6	2.3	1.7	1.2
1300	2.7	2.7	2.5	2.2	1.6	1.2
1400	2.6	2.6	2.4	2.1	1.6	1.2
2000	2.2	2.1	2.0	1.8	1.3	1.0
10,000	1.0	1.0	0.9	0.8	0.6	0.4

Table A6 Sampling errors for a binomial distribution (99% confidence level)

Sample size	Binomial percentage distribution					
	50/50	60/40	70/30	80/20	90/10	95/5
50	17.6	17.3	16.3	14.6	11.8	9.0
100	12.6	12.4	11.7	10.3	8.1	6.3
150	10.4	10.2	9.6	8.4	6.5	5.0
200	9.0	8.9	8.3	7.3	5.6	4.3
250	8.1	7.9	7.4	6.6	5.1	4.0
300	7.4	7.3	6.8	6.0	4.5	3.4
400	6.4	6.3	5.9	5.2	3.9	2.9
500	5.7	5.6	5.3	4.6	3.5	2.6
600	5.2	5.1	4.8	4.2	3.2	2.4
700	4.9	4.8	4.5	3.9	2.9	2.2
800	4.5	4.5	4.2	3.6	2.8	2.0
900	4.3	4.2	3.9	3.4	2.6	1.9
1000	4.1	4.0	3.7	3.3	2.5	1.8
1100	3.9	3.8	3.6	3.1	2.3	1.7
1200	3.7	3.6	3.4	3.0	2.2	1.6
1300	3.6	3.5	3.3	2.9	2.2	1.6
1400	3.4	3.4	3.2	2.8	2.1	1.5
2000	2.9	2.8	2.6	2.3	1.7	1.3
10,000	1.3	1.3	1.2	1.0	0.8	0.6

Key equations

The mean of a population: listed data

$$\mu = \frac{\Sigma X_i}{N}$$

N is the size of the population
X_i is each score in a distribution

The mean of a sample: listed data

$$\bar{X} = \frac{\Sigma X_i}{n}$$

n is the size of the sample

The mean of a sample: frequency data

$$\bar{X} = \frac{\Sigma f X_i}{n}$$

f is the frequency of each value in a distribution

The mean of a sample: class intervals

$$\bar{X} = \frac{\Sigma f m}{n}$$

m is the mid-point of a class interval

The standard deviation of a population: listed data

$$\sigma = \sqrt{\frac{\Sigma(X_i - \mu)^2}{N}}$$

The standard deviation of a sample: listed data

$$s = \sqrt{\frac{\Sigma(X_i - \bar{X})^2}{n-1}} \quad \text{or} \quad s = \sqrt{\frac{\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}}{n-1}}$$

The standard deviation of a sample: frequency data

$$s = \sqrt{\frac{\Sigma f X_i^2 - \frac{(\Sigma f X_i)^2}{n}}{n-1}}$$

Coefficient of relative variation

$$CRV = \frac{s}{\bar{X}} \times 100$$

Index of qualitative variation

$$IQV = \frac{\text{observed differences}}{\text{maximum possible differences}}$$

$$\text{maximum possible differences} = \frac{n^2(k-1)}{2k}$$

k is the number of categories

Z-score for describing a population

$$Z = \frac{X_i - \mu}{\sigma}$$

z-score for describing a sample

$$z = \frac{X_i - \bar{X}}{s}$$

Lambda

$$\lambda = \frac{E_1 - E_2}{E_1}$$

E_1 is the number of errors without information for the independent variable
E_2 is the number of errors with information for the independent variable

Cramer's *V*

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}$$

χ^2 is the chi-square statistic for the crosstab
k is the number of rows or the number of columns, whichever is smaller

Gamma

$$G = \frac{N_c - N_d}{N_c + N_d}$$

N_c is the number of concordant pairs
N_d is the number of discordant pairs

Somers' *d*

$$d = \frac{N_c - N_d}{N_c + N_d + T_y}$$

T_y is the number of cases tied on the dependent variable but varying on the independent variable

Kendall's tau-*b*

$$\text{tau-}b = \frac{N_c - N_d}{\sqrt{(N_c + N_d + T_y)(N_c + N_d + T_x)}}$$

T_x is the number of cases tied on the independent variable but varying on the dependent variable

Kendall's tau-*c*

$$\text{tau-}c = \frac{2k(N_c - N_d)}{N^2(k-1)}$$

Spearman's rank-order correlation coefficient

$$r_s = 1 - \frac{6\Sigma D^2}{n(n^2 - 1)}$$

Equation for a straight line

$$Y = a \pm bX$$

Y is the dependent variable
X is the independent variable
a is the *Y*-intercept (the value of *Y* when *X* is zero)
b is the slope of the line
+ indicates positive association
– indicates negative association

Regression coefficient

$$b = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2} \quad \text{or} \quad b = \frac{n\Sigma(X_iY_i) - (\Sigma X_i)(\Sigma Y_i)}{n\Sigma X_i^2 - (\Sigma X_i)^2}$$

Pearson's product moment correlation coefficient

$$r = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\left[(X_i - \bar{X})^2\right]\left[(Y_i - \bar{Y})^2\right]}} \quad \text{or} \quad r = \frac{n\Sigma(X_iY)_i - (\Sigma X_i)(\Sigma Y_i)}{\sqrt{\left[n\Sigma X_i^2 - (\Sigma X_i)^2\right]\left[n\Sigma Y_i^2 - (\Sigma Y_i)^2\right]}}$$

Confidence interval for a mean

$$\text{lower limit} = \bar{X} - z\left(\frac{s}{\sqrt{n}}\right), \quad \text{upper limit} = \bar{X} + z\left(\frac{s}{\sqrt{n}}\right)$$

***z*-test for a single mean**

$$z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

***t*-test for a single mean**

$$t = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

***z*-test for a binomial percentage**

$$z = \frac{(P_s - 0.5) - P_u}{\sqrt{\frac{P_u(100 - P_u)}{n}}} \quad \text{where } P_s > P_u \quad \text{or} \quad z = \frac{(P_s + 0.5) - P_u}{\sqrt{\frac{P_u(100 - P_u)}{n}}} \quad \text{where } P_s < P_u$$

P_u is the population percentage

Runs test

$$z = \frac{(R + 0.5) - \mu_R}{\sigma_R} \quad \text{where } R < \mu_R \quad \text{or} \quad z = \frac{(R - 0.5) - \mu_R}{\sigma_R} \quad \text{where } R > \mu_R$$

$$\mu_R = \frac{2n_1n_2}{n} + 1$$

$$\sigma_R = \sqrt{\frac{n^2 - 2n}{4(n-1)}}$$

R is the number of runs in the sample
n_1 is the number of cases with a given value
n_2 is the number of cases with the other value

Chi-square test for independence and goodness of fit

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e}$$

f_o is the observed frequency in each category
f_e is the expected frequency in each category

$df = (r - 1)(c - 1)$

r is the number of rows
c is the number of columns

The *t*-test for the equality of two means

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sigma_{\bar{X}-\bar{X}}}$$

$$\sigma_{\bar{X}-\bar{X}} = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}\sqrt{\frac{n_1 + n_2}{n_1n_2}} \quad \text{(pooled variance estimate)}$$

ANOVA *F*-test for more than two sample means

$$F = \frac{\frac{SSB}{k-1}}{\frac{SSW}{n-k}}$$

$TSS = SSB + SSW$

$TSS = \Sigma X_i^2 - n\bar{X}^2$

$SSW = \Sigma(X_i - \bar{X}_j)^2$

$$SSB = \Sigma n_s\left(\overline{X}_s - \overline{X}\right)^2$$

$\overline{X}_s$ is the mean for a given sample
n_s is the number of cases in a given sample

The two-sample *z*-test for the rank sum (Wilcoxon's rank-sum test)

$$z = \frac{W - \mu_W}{\sigma_W}$$

$$\mu_W = \frac{1}{2}n_1\left(n_1 + n_2 + 1\right)$$

$$\sigma_W = \sqrt{\frac{1}{12}n_1 n_2\left(n_1 + n_2 + 1\right)}$$

n_1 is the sample with the fewest cases
n_2 is the sample with the most cases

The dependent-samples *t*-test for the mean difference

$$t = \frac{\overline{X}_D}{s_D / \sqrt{n}}$$

$$\overline{X}_D = \frac{\Sigma D}{n}$$

$$s_D = \sqrt{\frac{\Sigma D^2 - \frac{(\Sigma D)^2}{n}}{n - 1}}$$

The McNemar chi-square test for change

$$\chi^2_M = \frac{\left(n_1 - n_2 - 1\right)^2}{n_1 + n_2}$$

n_1 is the observed number of cases in cell (b) or cell (c), whichever is *largest*
n_2 is the observed number of cases in cell (b) or cell (c), whichever is *smallest*

The Wilcoxon signed-ranks *z*-test for dependent samples

$$z = \frac{T - \mu_T}{\sigma_T}$$

$$\sigma_T = \sqrt{\frac{n(n + 1)(2n + 1)}{24}}$$

$$\mu_T = \frac{n(n + 1)}{4}$$

Glossary

Arithmetic mean The sum of all scores in a distribution divided by the total number of cases in the distribution.

Asymmetric measures of association Measures of association whose value depends on which variable is specified as independent and which variable is specified as dependent.

Binomial distribution A distribution that has only two possible values or categories.

Bivariate descriptive statistics A class of statistics that can be used to analyze whether a relationship exists between two variables.

Bivariate table A table that displays the joint frequency distribution for two variables.

Case An entity that displays or possesses the traits of a variable.

Census An investigation that includes every member of the population.

Central limit theorem A theorem which states that if an infinite number of random samples of equal size are selected from a population, the sampling distribution of the sample means will approach a normal distribution as sample size approaches infinity.

Class interval A range of values on a distribution that are grouped together for presentation and analysis.

Coefficient of relative variation A descriptive statistic that expresses the standard deviation of a distribution as a percentage of the mean.

Conceptual definition The use of literal terms to specify the qualities of a variable (also called the **nominal definition**).

Concordant pair Two cases in a joint distribution that are ranked the same on each of the variables.

Confidence level The probability that an interval estimate will include the value of the population parameter being estimated.

Constant An attribute or quality that does not vary from one case to another.

Contingency table See **Bivariate table**.

Continuous variable A variable that can vary in quantity by infinitesimally small degrees.

Coordinate A point on a scatter plot that simultaneously indicates the values a given case takes for each variable.

Critical region The range of scores that will cause the null hypothesis to be rejected at a specified significance level.

Crosstabulation See **Bivariate table**.

Cumulative frequency table A table that shows, for each value in a distribution, the number of cases up to and including that value.

Cumulative relative frequency table A table that shows, for each value in a distribution, the percentage or proportion of the total number of cases up to and including that value.

Dependent samples Samples for which the criterion for inclusion in one sample is affected by the composition of the other samples.

Dependent variables A variable whose distribution is affected or caused by variation in the independent variable.

Descriptive statistics The numerical, graphical, and tabular techniques for organizing, presenting, and analyzing data.

Dichotomous variable A variable that has only two possible values.

Discordant pair Two cases in a joint distribution whose rank on one variable is different to their rank on the other variable.

Discrete variable A variable that has a countable number of values.

Error term See Residual.

Frequency The number of times that a particular score appears in a set of data.

Frequency table A table that reports, for each value of a variable, the number of cases that have that value.

Hypothesis A statement about some characteristic of the distribution of a population.

Hypothesis testing The procedure for deciding whether some aspect of a population distribution has a specified characteristic.

Independence Two variables are independent if the pattern of variation in the scores for one variable is not related to the pattern of variation in the scores for the other variable.

Independent variables A variable whose distribution affects or causes the variation in the dependent variable.

Index of qualitative variation The number of differences between scores in a distribution expressed as a proportion of the total number of possible differences.

Inferential statistics The numerical techniques used for making conclusions about a population distribution, based on the data from a random sample drawn from that population.

Interquartile range The difference between the upper limits of the first quartile and the third quartile; the range for the middle 50 percent of cases in a rank-ordered series.

Interval scale A level of measurement that has units measuring intervals of equal distance between values on the scale.

Mean See **Arithmetic mean.**

Measurement The process of determining and recording which of the possible traits of a variable an individual case exhibits or possesses.

Measures of association Descriptive statistics that indicate the extent to which a change in the value of one variable is related to a change in the value of the other variable.

Measures of central tendency Descriptive statistics that indicate the typical or average value for a distribution.

Measures of dispersion Descriptive statistics that indicate the spread or variety of scores in a distribution.

Median A measure of central tendency which indicates the value in a rank-ordered series that divides the series in half.

Missing cases Cases in a data set for which measurements of a variable have not been taken.

Mode A measure of central tendency; the value in a distribution with the highest frequency.

Multivariate regression A technique that investigates the relationship between two or more independent variables and a single dependent variable.

Nominal definition See **Conceptual definition.**

Nominal scale A level of measurement that only indicates the category of a variable into which a case falls.

Non-parametric test A test of an hypothesis about some feature of a population distribution other than its parameters.

Operational definition The specification of the procedures and criteria for taking a measurement of a variable for individual cases.

Ordinal scale A level of measurement that, in addition to the function of classification, allows cases to be ordered by degree according to measurements of a variable.

Ordinary least squares regression A rule which states that the line of best fit for a linear regression is the one that minimizes the sum of the squared residuals.

Parameter A statistic that describes some feature of a population.

Parametric test A test of an hypothesis about the parameters of a population distribution.

Percentages Statistics that standardize the total number of cases to a base value of 100.

Perfect association A statistical relationship where all cases with a particular value for one variable have a certain value for the other variable.

Population The set of all cases of interest.

Proportions Statistics that standardize the total number of cases to a base value of one.

Random selection A sampling method where each member of the population has the same chance of being selected in the **sample.**

Range The difference between the lowest and highest scores in a distribution.

Rank A number that indicates the position of a case in an ordered series.

Ratio scale A level of measurement which assigns a value of 0 to cases which possess or exhibit no quantity of a variable.

Region of rejection See **Critical Region.**

Regression coefficient A descriptive statistic that indicates by how many units the dependent variable will change, given a one-unit change in the independent variable.

Relative frequencies Statistics that express the number of cases within each value of a variable as a percentage or proportion of the total number of cases.

Residual The difference between the observed and expected value of a variable.

Run A sequence of scores that have the same outcome for a variable. A run is preceded and followed by scores that have a different outcome for a variable, or no data.

Sample A set of cases that does not include every member of the population.

Sampling distribution The theoretical probability distribution of an infinite number of sample outcomes for a statistic, using random samples of equal size.

Scatter plot A graphical technique for describing the joint distribution for two variables.

Standard deviation A measure of dispersion that is the square root of the variance.

Stated class limits The upper and lower bounds of an interval that determine its width.

Symmetric measures of association Measure of association whose strength will be the same regardless of which variable is specified as independent and which variable is specified as dependent.

Type I error The error of rejecting the null hypothesis of no difference when in fact it is correct.

Type II error The error of failing to reject the null hypothesis when in fact it is false.

Valid cases Cases in a data set for which measurements of a variable have been taken.

Variable A condition or quality that can vary from one case to another.

Variance A statistic that expresses the mean deviation of scores from the mean of a distribution.

z-scores Numbers that express the interval between a point and the mean of a normal distribution as a proportion of the standard deviation of that normal distribution.

Answers

1.1 (a) Not exhaustive: no option for people not eligible to vote. Not mutually exclusive: someone can be either of the first two options and did not vote at the last election.

(b) Not exhaustive: needs an 'other category' at least for students enrolled in other courses. Not mutually exclusive: social sciences is a broader category that includes sociology and economics.

(c) Not mutually exclusive: someone can have multiple reasons for enlisting.

1.2

(a) interval/ratio	(b) nominal	(c) nominal
(d) interval/ratio	(e) nominal	(f) nominal
(g) ordinal	(h) interval/ratio	(i) ordinal
(j) nominal	(k) interval/ratio	(l) nominal
(m) ordinal	(n) interval/ratio	(o) ordinal
(p) nominal		

1.5

(a) discrete	(b) continuous	(c) continuous
(d) discrete	(e) continuous	(f) continuous

3.1 A pie chart emphasizes the contribution that the frequency for each category makes to the total, whereas a bar graph emphasizes the frequency of each category relative to each other.

3.2 A bar graph expresses the distribution of discrete variables whereas a histogram expresses the distribution of continuous variables.

3.4 This is continuous interval/ratio data, so that a frequency polygon is the best technique, given the number of values in the distribution. If these data were organized into class intervals a histogram could also be constructed.

3.5 (a)

Price	Frequency
7000–8499	2
8500–9999	3
10000–11499	6
11500–12999	3
13000–14499	1

3.6 The pie graph illustrates the large proportion of migrants from Europe in the total.

3.7 (a) A pie graph will highlight that clerical workers make the most significant contribution, in terms of employment categories, to the total.

(b) You should have a bar chart with three spikes, one for each of the employment categories. The spikes should be divided into males and females. The graph will reveal that women are highly concentrated in clerical positions, whereas men dominate managerial and, especially, custodial positions.

(c) The curve is highly skewed to the right.

4.1 A proportion standardizes totals to a base of 1, whereas a percentage standardizes totals to a base of 100.

4.2 A percentage is calculated using the same formula as a proportion multiplied by 100, ensuring that the percentage will be a higher number (by a factor of 100) than the corresponding proportion.

4.3 (a) 0.01 (1%) (b) 0.13 (13%) (c) 1.24 (124%) (d) 0.0045 (0.45%)

4.4 (a) 12% (0.12) (b) 14.4% (0.144) (c) 167% (1.67) (d) 4.5% (0.045)

4.5

Time to complete fitness trial

Interval	Mid-point	Frequency	Cumulative frequency	Percent	Cumulative percent
1–9	5	0	0	0.0%	0.0%
10–19	14.5	5	5	12.5%	12.5%
20–29	24.5	7	12	17.5%	30.0%
30–39	34.5	14	26	35.0%	65.0%
40–49	44.5	6	32	15.0%	80.0%
50–59	54.5	4	36	10.0%	90.0%
60–69	64.5	0	36	0.0%	90.0%
70–79	74.5	1	37	2.5%	92.5%
80–89	84.5	3	40	7.5%	100.0%

Heart rate in minutes

Interval	Mid-point	Frequency	Cumulative frequency	Percent	Cumulative percent
60–69	64.5	4	4	10.0%	0.0%
70–79	74.5	10	14	25.0%	25.0%
80–89	84.5	14	28	35.0%	60.0%
90–99	94.5	11	39	27.5%	87.5%
100–109	104.5	1	40	2.5%	90.0%

4.6

Region	People attending public libraries	Relative frequency, %	People attending popular music concerts	Relative frequency, %
A	1409	31.7	1166	33.7
B	1142	25.7	870	25.2
C	713	16.1	604	17.5
D	423	9.5	280	8.1
E	497	11.2	332	9.6
F	130	2.9	99	2.9
G	90	2.0	32	0.9
H	38	0.9	74	2.1
Total	4442	100	3456	100

4.10 (a) 104 (b) 21.9% (c) 77% (d) 27%

5.1 The conclusion drawn incorrectly about the causality of the relationship from the observed statistical association. It is more appropriate to regard the causality as running in the opposite direction: the higher injury rate 'causes' the higher number of ambulance officers attending the accident.

5.2 (a)

Dependent	Independent		
	1	2	Total
1	40%	55%	49%
2	60%	45%	51%
Total	100%	100%	100%

(b)

Dependent	Independent			
	1	2	3	Total
1	79%	57%	16%	53%
2	21%	43%	84%	47%
Total	100%	100%	100%	100%

5.3 (a)

Dependent	Independent		
	1	2	Total
1	33%	67%	100%
2	47%	53%	100%
Total	41%	59%	100%

(b)

Dependent	Independent			
	1	2	3	Total
1	53%	38%	9%	100%
2	16%	32%	47%	100%
Total	35%	35%	30%	100%

5.4 (a) It is most likely that since a father's voting preference is formed before his own child's that this is the independent variable and the child's voting preference is the dependent variable. Voting preferences are measured at the ordinal level.

(b)

Own voting preference	Father's voting preference			
	Progressive	Conservative	Other	Total
Progressive	22	4	4	30
Conservative	5	19	6	30
Total	27	23	10	60

(c) Adding column percentages will help determine by eye whether there is any dependence. The pattern of dependence suggests that children tend to vote in a similar way to their respective father.

5.5 After calculating the relative frequencies there appears to be no relationship between country of residence and amount of TV watched.

5.6 (a) Smoking habit is ordinal, and health level is ordinal.

(b) Both of these are behavioral variables so any plausible explanation which has either variable as the independent, or mutually dependent, is permissible.

(c)

Health level	Smoking level		
	Doesn't smoke	Does smoke	Total
Poor	13 13.4%	34 52.3%	47 29.0%
Fair	22 22.7%	19 29.2%	41 25.3%
Good	35 36.1%	9 13.8%	44 27.2%
Very Good	27 27.8%	3 4.6%	30 18.5%
Total	97 59.9%	65 40.1%	162 100%

5.7 (a) 84 (b) 254 (c) 74 (d) 88.1% (e) 0

6.1 An asymmetric measure will be affected by the choice of which variable is specified as the dependent and which variable is specified as independent. A symmetric measure will yield the same value for the strength of association irrespective of the model of the relationship. A symmetric measure is therefore the appropriate one.

6.2 It is important to specify the dependent and independent variables since lambda is an asymmetric measure of association, whose value is therefore affected by this choice. If the pattern of dependence is not thought to be that of one-way dependence, the symmetric version of Lambda should be used.

6.3 (a) Lambda=0.11 (very weak association)
(b) Lambda=0.42 (moderate association)
(c) Lambda=0 (this does not necessarily indicate no association. Looking at the table it is clear that there is some variation between columns, but the modal response for all values of the independent variable is one, causing lambda to equal zero).

6.4 Lambda=0.54. Looking at the table the moderate association is due to the higher proportion of gun owners in favor of capital punishment.

6.5

Can sing anthem?	Job classification		
	Blue collar	White collar	Total
Yes	29	22	51
No	21	28	49
Total	50	50	100

Lambda = 0.12

6.6 The study indicates that the strength of the association has increased in recent times. In a *relative* sense we might say that the association is strong, but this is only in relation to the past studies, rather than in some absolute sense.

6.7 (a) Lambda = 0.19 (with current income dependent)
(b) Lambda = 0.262 (with current income dependent)

7.1 Negative

7.2 Nominal variables do not have a direction of change

7.3 (a) 14(60)=840 (b) 24(8)=192
(c) 19(12+17+20)=931 (d) 16(12+17+10+14+22)=1200

7.4 (a) 14(12)=168 (b) 32(24+12)=1152
(c) 24(32)=768 (d) 11(25+42+19+24)=1210

7.5 (a) 8(14+32)=1472 (b) 32(14+8)=1472
(c) 24(60+12)=1728 (d) 11(6+16+20)=462

7.6 The number of cases tied on the dependent but not on the dependent variable is 5030. The value of Somers' *d* is 0.25.

7.7 Concordant pairs: 26(20+58+15+62) + 22(20+58) + 62(20+15) +62(20) = 9234
Discordant pairs: 12(58+22+62+23) + 15(58+22) + 62(22+23) + 62(22) = 7334
Gamma = 0.11; therefore a very weak, positive relationship between these variables.

7.8 (a) Inspection of the table by eye reveals a negative association, since health level seems to decrease as smoking level increases (it is helpful to calculate the column percentages to see this). This will appear as a negative sign in front of any measure of association calculated on these data.
(b) The value for gamma is –0.69, indicating a moderate to strong negative association.

7.12 Somers' *d* with current income as dependent is 0.794 and Gamma is 0.914 indicating a strong, positive relationship. Tau-*b* is not useful because there is not the same number of columns and rows.

8.1 This is an example of a spurious relationship; there is no theoretical basis for concluding that a direct causal relationship exists between these two variables. Rather they are each determined by a child's general state of development.

8.2 The relationship remains the same for each of the partial tables, indicating that the control variable does not alter the direct relationship between *X* and *Y*.

8.3 The original relationship is not as strong once the control variable is added (by comparing the original gamma with the partial gamma). This indicates that the relationship is partially spurious or intervening, although some direct relationship also exists between age and concern for the environment. This is stronger for conservatives than for liberals.

9.1 No: the numbers on an ordinal scale are values which have no quantitative significance. They are merely labels which preserve the ordering of cases. To calculate a mean we need to perform the mathematical operation of addition and this requires interval/ratio data.

9.2 μ is the mean for a population; $\overline{X}$ is the mean for a sample.

9.3 2

9.4 (a) mean=23.3; median=14 (b) mean=267.4; median=289 (c) mean=2.9; median=2.4

9.5 This student had a lower than average IQ in the first class, and a higher than average IQ for the class the student joined.

9.6 (a) 9, 11, 20, 22, 36, 36, 39, 43, 45, 50, 56, 57, 59, 60, 66, 68, 68, 73, 75, 80, 87
Median=56
(b) 50.5 (rounded to 1 decimal place)
(c) The median is greater than the mean, therefore the distribution is skewed to the left.
(d) Mean=57; Median=56.5

The median is a relatively stable measure of central tendency that is not sensitive to extreme outliers, whereas the mean, by including every value in its calculation, is affected by the addition of one extreme score.

9.7 mean=$33,500; median=$32,500; mode=$22,000

9.8 (a) mean (ungrouped)=29.6 minutes; mean (grouped)=28.25 minutes
median (ungrouped)=31.5 minutes; median (grouped)=31–40 minutes

The differences are due to the fact that class intervals do not provide as much information as a listing of the raw scores. Since we use class mid-points rather than the actual data in calculating the mean, the answer will vary. With median and mode we can only report the class, rather than the specific value.

9.9 Degree of enrolment:
(a) Nominal (b) mode=Arts

Time spent studying in library:
(a) Interval/ratio (b) mean=3.275 hours; median=2 hours; mode=4 hours

Satisfaction with employment:
(a) Ordinal (b) mode=satisfied; median=satisfied

9.10 (a) mean=8.7 years; median=9–12 years; mode=9–12 years
(b) the distribution is skewed to the right

9.11 No; the value that occurs the most is Europe. The mode is not the frequency with which it occurs.

9.12 (a) $17,403

10.1 The advantage of the range is that it is very easy to calculate and everyone understands it. Its disadvantage is that because it only uses two scores it does not use all the information available in a distribution. For the same reason it is very sensitive to extreme values.

10.2 σ is the standard deviation for a population; s is the standard deviation for a sample.

10.3 (a) range=67; standard deviation=24.9
(b) range=332; standard deviation =120.6
(c) range=4; standard deviation =1.4

10.4 (a) range=$60,000; *IQR*=$15,000; standard deviation=$14,183

10.5 (a) The *CRV* for beginning salaries is 46.4%. The *CRV* for current salaries is 49.6%. Therefore current salaries have slightly more variation.
(b) 10 years, 1 month

11.1 (a) 0.097 (b) 0.097 (c) 0.3082 (d) 0.9665 (e) 0.0915 (f) 0.110 (g) 0.050

11.2 (a) ±1 (b) +2.1 (c) −1.645 (d) ±1.5

11.3 (a) 0 (b) −0.8 (c) 2.5 (d) −1.7 (e) 1.3

11.4 The z-score for the poverty line is −0.83. The proportion for $z = -0.8$ is 0.212, and the proportion for $z = -0.9$ is 0.184. Therefore the proportion of all families headed by a single mother also living in poverty is between 0.212 and 0.184 or around 1 in 5.

11.5 $z = -1.6$, and the area under curve is 0.055. Therefore 5.5% of light bulbs last 462 hours or less.

11.6 (a) $z = -1.65$, area under curve is 0.05.
(b) $z = \pm 1.96$; for $z = -1.96$ the selling price is $15,292; for $z = 1.96$ the selling price is $24,308. Therefore the range is $15,292–$24,308.

11.7 (a) $z = 1.4$, probability is 0.081
(b) $z = 1.645$, distance is 48.225 meters

11.8 (a) for 18 years $z = -1.3$, proportion between mean and 18 is 0.403
for 65 years $z = 2.1$, proportion between mean and 65 is 0.486
proportion between 18 and 65 years is 0.403+0.486=0.889
(b) middle 50%: closest probability in Table is 0.516 with $z = \pm 0.7$
for $z = -0.7$ the age is 26, for $z = 0.7$ the age is 45 (both figures rounded to nearest whole year)

11.9 (a) At $1.7 million $z = 1$, which has a one-tail probability of 0.1585
(b) At $1.2 million $z = -1.5$, which has a one-tail probability of 0.067

11.10 At 15 km/h $z = 0.5$, which has a probability of 0.3085. This means that the wind speed will be over 15 km/h 30 percent of the time, which meets the proposal requirements.

12.1 The purpose of drawing a scatter plot is to make judgment about whether the conditions for using a linear regression hold. In particular, we can assess visually whether there is a linear relationship, rather than a curvilinear relationship.

12.2 The *Y*-intercept indicates the expected value for the dependent variable when the independent variable is zero. It is equal to a in the regression equation.

12.3 The principle, often called the ordinary least squares regression line, is to draw a line that minimizes the sum of the squared residuals between each point in a scatter plot and the regression line.

12.4 (a) positive (b) negative (c) positive (d) no relationship (e) negative

12.6 The correlation coefficient is a standardized measure of correlation that ranges from −1 to 1, regardless of the units in which the variables are measured. The coefficient of the regression line indicates the amount of change in the dependent variable expected from a one-unit change in the independent variable. It is therefore sensitive to the units of measurements.

12.7 (d) $Y = 27.165 - 0.15(X)$; when $X = 12$, $Y = 24.885$

12.8 (a) When $X = 0$, $Y = 40$ years
(b) $Y = 40 + 0.7(30) = 61$ years (note that we use 30 in the equation not 30,000, since the units of measurement are $,000).
(c) We cannot use the regression coefficient of +0.7 to assess the strength of the correlation. To do this we need to calculate the correlation coefficient.

12.9 (a) $\hat{Y} = 33.4 + 0.511(X)$
(b) The value for r indicates a strong, positive relationship.
(c) When hours (X) are zero, Y=33.4, indicating a fail.
(d) $50 = 33.4 + 0.511(X)$, $X = 32.6$ hours. The high value of r^2 indicates that the student can be very confident in the prediction. It is wrong to use the regression line in this way because it is not a deterministic relationship: there is an element of error. The student may not actually work when in the library, thinking that just spending the time there will be sufficient.

12.10 (a) $\hat{Y} = 157 + 4.88(X)$, $r = 0.92$, $r^2 = 0.85$
(b) The regression coefficient changes to 4880. Since r and r^2 are standardized coefficients their values are unaffected by the units of measurement.

12.12 (a) $r = -0.77$
(b) days lost = 14.4 − 0.88(hours of exercise); for 8 hours of exercise, days lost = 7.4

12.13 current salary = $1928 + 1.9(beginning salary)
The value for r^2 is 0.755 indicating that using beginning salary to predict current salary will produce reliable predictions.

12.15 Rho = 0.85. There is a strong positive association between these variables.

12.16 Rho = 0.51

12.17 Rho = −0.19

13.1 (a) Days lost
(b) It would be reasonable to suspect that days lost decrease as the amount of exercise increases (negative) and that days lost increases as age increases (positive).
(c) days lost = 16.99 − 0.942(exercise hours) − 0.06(age in years); note that the sign in front of age is not the one expected.
(d) The coefficient for age is not significant, and the value for the adjusted R-squared indicates that it has not improved our predictive ability over the regression equation using exercise alone.

14.1 A sample statistic is a numerical measure of a sample while a parameter is a measure of some feature of a population.

14.2 Descriptive statistics summarize the data from a sample, inferential statistics attempt to generalize from a random sample to the population.

14.3 Random variation is the variation in sample outcomes brought about by random selection from a population. It requires us to use probability theory when generalizing to a population.

14.4 (a) False; it is evident from the equation for the standard error that the size of the population is not a factor affecting the reliability of a sample.
(b) True
(c) False; the standard error is equal to the standard deviation of the population *divided* by the square root of the sample size and therefore must be smaller than the standard deviation of the population.
(d) False; provided the sample size is large (i.e. greater than 12) the central limit theorem states that the sampling distribution of sample means will be normal, even where the population from which the samples are drawn is not normal.

14.5 In either case the mean of the sampling distribution is 40.

14.6 The standard error is the standard deviation of a sampling distribution. It is always smaller than the standard deviation of the population since the effect of any extreme individual scores included in a sample will be muted by more representative scores included in the sample.

14.7 The difference is that where $n = 30$ the distribution has fatter tails than the distribution for $n = 200$; that is, the standard error is smaller in the larger sample. They are similar because they both approximate the normal curve and centered on the population mean.

14.8 It appears to be random because each letter in the hat has an equal chance of being selected; however, since there may not be the same number of students for every letter it does not mean every *student* in the class has an equal chance of being selected. For example, if there were a lot of people with a surname beginning with G in the class the sample would over-represent that particular group.

14.9 The sampling method is random if every book in the library has an equal chance of being borrowed and then returned on a Thursday and there is nothing about Thursday that will influence the condition of books returned on that day.

14.11 The theorem is important because it allows the use of a normal sampling distribution to carry out statistical analysis, even where samples are drawn from non-normal populations, and such populations are very common in social research.

14.12 There is far greater variation in the sample means from the $n = 20$ samples. The spread of scores still should be centered on the population mean.

15.1 The distribution approaches the normal curve as sample size increases towards infinity, as described by the central limit theorem, regardless of the shape of the population distribution.

15.2 Type I error occurs when the null hypothesis is rejected even though it is true; a type II error occurs when the null hypothesis is accepted when a rejection should have been made. The probability of one happening decreases the possibility of the other occurring increases.

15.3 As the significance level is increased the critical region becomes smaller; that is, the bigher the significance level the larger the difference has to be before the null hypothesis is rejected

15.4

Probability	Test	z-score
0.230	Two-tail	±1.2
0.100	Two-tail	±1.645
0.018	One-tall	±2.1
0.021	Two-tail	±2.3
0.0003	One-tail	±3.4

15.5 (a) $z > 1.645$, $\alpha = 0.05$ (b) $z < -1.645$, $\alpha = 0.05$ (c) $z > 1.96$ or $z < -1.96$, $\alpha = 0.05$

15.6 (a) $z = -1.9$ (b) $z = -11.8$

15.7 (a) The probability of selecting, from a population with a mean of 15 years, a random sample with a mean that differs from the population mean by three or more is 0.003.

(b) The probability of drawing, from a population with a mean of 15 years, a random sample with a mean less than the population mean by three or more is 15 in 1000.

15.8 No; significance tests never definitively prove anything about a population. They only indicate the *probability* of drawing a sample with a known mean value from a population with an hypothesized mean value. Even with extremely low significance levels we risk making a type I error.

15.9 H_0: $\mu = 24$, H_a: $\mu > 24$, $\alpha = 0.05$, $z_{sample} = 1.73$, $p = 0.0445$, $z_{critical} = 1.645$

We are using a one-tail (right-tail) test because we are interested in whether this judge has an average greater than the rest. At an alpha level of 0.05 the probability of the judge being the same as other judges is less than the alpha level, leading the null hypothesis to be rejected. Note that an alpha level of 0.01, or on a two-tail test, the sample score will not be significantly different to the hypothesized value.

16.1 The sample is drawn from a normal population.

16.2

t-score	Probability	Test	df
2.015	0.05	One-tail	5
2.764	0.02	Two-tail	10
1.708	0.05	One-tail	25
2.000	0.05	Two-tail	65
1.282	0.10	One-tail	228

16.3 (a) $t = -3.08$ (reject) two-tail
(b) $t = -3.08$ (reject) one-tail
(c) $t = -2.18$ (reject)
(d) $t = -6.13$ (reject)
(e) $t = 1.29$ (fail to reject)
(f) $t = 3.86$ (reject)

16.4 (a) $t_{sample} = -2.35$, so the null hypothesis is rejected, the pay rise has not been achieved. However, at the 0.01 level the null hypothesis is not rejected.

16.5 (a) mean = 63 years; standard deviation = 16.63 years
(b) $p = 0.045$ (around 45 in every thousand)

16.6 $t_{sample} = -12.96$, the null hypothesis is rejected. Hip fractures affect walking speed.

16.7 The following sample scores and decisions regarding the null apply:

Canada: $t_{sample} = 3.85$ (reject)
Singapore: $t_{sample} = -6.87$ (reject)

Australia: $t_{sample} = -1.02$ (do not reject)

16.8 (a) $t_{sample} = -1.151$, on a two-tail test $p = 0.264$; therefore do not reject null hypothesis.

17.1 Interval estimation is the process of inferring the range of values that contain the (unknown) population parameter, together with the probability (confidence level) that this estimate does include the parameter.

17.2 A confidence level is the probability that a particular range of values will include the population parameter. As the confidence level increases the width of the confidence interval also increases, and vice versa.

17.3 As sample size increases the width of the confidence interval becomes smaller.

17.4 The standard deviation of the population alters the width of the confidence interval by affecting the standard error of the estimate. As the standard deviation increases so does the standard error, meaning the confidence interval will also widen.

17.5 Age of pre-school children:
90% confidence level: 3.75 [3.64, 3.86]
99% confidence level: 3.75 [3.57, 3.93]

TV watching:
90% confidence level: 150 [145.24, 154.76]
99% confidence level: 150 [142.49, 157.51]

17.6 Economics: 6 [5.26, 6.74]
Sociology: 4 [3.33, 4.67]
History: 4.5 [3.56, 5.44]
Statistics: 3 [2.62, 3.38]

17.7 Increase for all workers across the industry at 95% is $1018 [$907.68, $1128.32], and at the 99% confidence level is $1018 [$871.65, $1164.35].

17.8 (a) 4.3 days [3.79, 4.82] at 99%.
(b) Compared to the other hospital it is about the same since the confidence interval includes the value of 4 days.
(c) To improve the accuracy of the estimate it could include more people in the sample.

17.9 8.5 years [8.28, 8.72]

17.10 (a) $34,420 [$33,127, $35,712]
(b) $34,420 [$32,878, $35,961]
(c) $34,420 [$32,391, $36,448]

18.1 The samples come from normal populations, and when using the pooled variance estimate, the populations have the same variance.

18.2 (a) $t_{sample} = -1.5$, $df = 83$; do not reject null
(b) $t_{sample} = -3.38$, $df = 238$; reject null
(c) $t_{sample} = -1.5$, $df = 83$; do not reject null
(d) $t_{sample} = -2.5$, $df = 218$; reject null

18.3 $t_{sample} = -2.2$, ($\alpha = 0.05$, two-tail, $df = 196$). Reject null hypothesis.

18.4 $t_{sample} = 3.5$

Reject null hypothesis. Important considerations are the number of samples to be compared, interval/ratio data used to describe a mean, and population standard deviations are unknown.

18.5 $t_{sample} = -12.2$. Reject null hypothesis, the organic pesticide is different and better.

18.6 The sample t-score is 3.2, which is significant at the 0.01 level. Therefore reject the null hypothesis.

19.1 (a) We are comparing more than two samples in terms of a variable measured at the interval/ratio level.

(b) There is no difference in the average number of cases handled at each agency.

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$

(c) The F-ratio is 0.245. At $\alpha = 0.05$, and $dfb = 4$ and $dfw = 106$, $F_{critical} = 2.52$. Therefore the Null hypothesis is not rejected: all means are equal.

19.2 (a) Method A: mean = 22.73, standard deviation = 2.05
Method B: mean = 29.82, standard deviation = 2.87
Method C: mean = 20.27, standard deviation = 3.95

Looking at the means and the standard deviations it seems that only Method C will be significantly different to each of the others.

(b) The F-ratio is 29, which is statistically significant at the 0.01 level.

19.3 The F-score is 24.6, which is significant at the 0.01 level. At least one of the populations has a mean not equal to that of the others.

19.4 The significant difference is between Level 1 and all the other Levels of blood alcohol, but no other combinations.

20.1 (a) Mean difference = −1.3
(b) $s_e = 1.212$
(c) $t_{sample} = -1.07$; do not reject null

20.2 (a) $t_{sample} = 7.16$; reject null
(b) $t_{sample} = -1.012$; do not reject null

20.3 $t_{sample} = 13.3$, which is significant at the 0.01 level. Therefore the treatment should be adopted.

20.4 $H_0: \bar{X} = 0,$

$H_a: \bar{X}_D > 0$

$t_{sample} = 2.5$. Reject the null, the changes in workplace have improved productivity.

20.5 $t_{sample} = 1.4$. The two-tail significance is greater than $\alpha = 0.05$, therefore accept the null hypothesis: people do seem, on average, to get the price they offer.

20.6 (a) Weight in kg Pre Test; Weight in kg Post Test
(b) 21
(c) 70.1 kg
(d) 66.43 kg
(e) 3.67 kg
(f) $t_{sample} = 5.966$, $df = 20$

(g) less than 0.0005 (note that SPSS rounds off to 3 decimal places, so that the probability is not actually equal to zero)
(h) upper limit = 4.5 kg
(i) lower limit = 2.38 kg
(j) Using the *t*-test, the sample value is lower than any **critical** value, therefore reject **the null – the program** is effective **in reducing** weight. We could also refer to the **confidence interval,** which does not **include** the value of 0.

The 95 percent confidence interval **does** not include the value of 5. The range of estimated values for weight loss is **below** the target value; therefore the program is not successful.

20.9 The mean difference is both significantly greater than $0 and also $15,000. We can test the latter by looking at the confidence interval which does not include the test value of $15,000.

21.1 The statement is false. The width **of an interval** estimate is only affected by the sample size, the confidence level, and **the sample** proportion. No other factor enters into the equation for the confidence **interval. Given** these factors the **interval estimate** will be the same regardless of the size of the **population from** which the **sample is drawn.**

21.2

	z_{sample}	Two-tail	One-tail
(a)	1.78	Fail to reject	Reject
(b)	–0.36	Fail to reject	Fail to reject

21.3 (a) z_{sample} = –0.31, **at** α = 0.05, one tailed, therefore the null hypothesis is not rejected: the sample percentage is not significantly different to the target of 40 percent, so that the program was successful.
(b) The confidence interval supports this because 40 percent is inside the 95 percent confidence interval of [35.6%, 42.2%].

21.4 At 95 percent, the confidence interval is [43.6%, 61.4%]. This includes values of less than 50 percent so that the sample does not confirm that the candidate is a certain winner. Similarly, if we conduct a *z*-test using 50 percent as the test value, the sample percentage is not significantly different.

21.5 At an alpha level of 0.05 the *z*-score of –3.08 will lead us to reject the null hypothesis so that taping does reduce ankle sprain injury.

21.6 The confidence interval is [51.6%, 60.4%], at a 95 percent confidence level, meaning that a majority of the population supports decriminalization.

21.7 (a) [5.4%, 34.6%] (b) [8.9%, 31.1%]

21.8 (a) Runs test applicable because the results are in sequence and using a binomial distribution. Runs test is applicable because the research question is interested in whether a *series* of outcomes for a binomial variable is random.
(b) z_{sample} = –0.19, fail to reject.

21.9 (a) 12
(b) 9.9
(c) Not significantly different to the test value; therefore we cannot say the series is non-random.

22.1

	df	α = 0.10	α = 0.05
Three categories	2	4.605	5.991
Five categories	4	7.779	9.488
Eight categories	7	12.017	14.067

22.2 (a) χ^2_{sample} = 5.28, *df* = 4, *p* = 0.35; do not reject null
(b) χ^2_{sample} = 1.33, *df* = 6, *p* = 0.965; do not reject null

22.3 χ^2_{sample} = 40, *df* = 4, *p* < 0.01; reject null

22.4 χ^2_{sample} = 6.246; do not reject null

22.5 (a) 26.8 is the expected value for each school.
(b) The sample chi-square value is 2.49, which is not significant at the 0.05 level. We cannot reject the statement that these schools have the same percentage of students going on to university.

22.6 Expected frequencies are Clerical 389, Custodial 38, and Manager 47. This is significantly different.

23.1 (a) 3 (b) 3 (c) 15 (d) 8

23.2 (a) 0.24 (b) 48

23.3 Expected frequencies:

	a	b	c	d	Total
a	1.59	0	6.87	46.54	55
b	1.41	0	6.13	41.46	49
Total	3	0	13	88	104

The shaded cells violate the rules that expected frequencies should not be less than 5.

23.4 χ^2_{sample} = 20.9, which is significant at the 0.01 level with 2 degrees of freedom.

23.5 χ^2_{sample} = 0.76 (your answer may differ slightly due to rounding); we cannot reject the null hypothesis of independence, since this has a very low *p*-score. There appears to be no relationship between country of residence and amount of TV watched.

23.6 (a) Health level (ordinal), Smoking habit (ordinal).
(e) The significance level for the sample chi-square indicates that we should reject the null hypothesis of independence.

23.7

Can sing anthem?	Job type		
	Blue collar	White collar	Total
Yes	29	22	51
No	21	28	49
Total	50	50	100

χ^2_{sample} = 1.96. With 1 degree of freedom, *p* is between 0.1 and 0.2; we do not reject the null hypothesis.

24.1 (a) χ^2_M = 2.16; do not reject null
(b) χ^2_M = 0.343; do not reject null
(c) χ^2_M = 14.723; reject null

24.2 (a) χ_M^2 = 0.593; do not reject null

25.1 (a) 2, 6, 9, 10, 11, 17; rank is 4
(b) 2, 6, 8, 9, 10, 11, 17; rank is 5
(c) 2, 6, 9, 10, 10, 11, 17; rank is 4.5
(d) 2, 6, 8, 9, 10, 10, 11, 11, 17; rank is 5.5
(e) 3, 4, 9, 10, 10, 10, 15, 16, 20, 22; rank is 5

25.2 In the preceding exercise identify and assign the correct rank to the score immediately following 10 in the rank-ordered series.

(a) 2, 6, 9, 10, 11, 17; 11 is rank 5
(b) 2, 6, 8, 9, 10, 11, 17; 11 is rank is 6
(c) 2, 6, 9, 10, 10, 11, 17; 11 is rank is 6
(d) 2, 6, 8, 9, 10, 10, 11, 11, 17; 11 is rank is 7.5
(e) 3, 4, 9, 10, 10, 10, 15, 16, 20, 22; 15 is rank is 7

25.3 A rank sum test is used when (i) the test variable is measured at the ordinal level, or (ii) the test variable is measured at the interval/ratio level but the samples come from populations that are not normally distributed.

25.4 (a) (Ranks in brackets)

Group 1	Group 2
1 (1)	12 (6.5)
15 (8.5)	25 (13)
12 (6.5)	29 (14)
16 (10)	8 (3)
23 (12)	15 (8.5)
9 (4)	20 (11)
11 (5)	7 (2)

(b) Group 1: 47, Group 2: 58
(c) The smallest rank sum is that for Group 1, $W = 47$
(d) $\mu_W = 52.5$
(e) $z_{sample} = -0.7$, do not reject null hypothesis

25.5 The sample z-score is -2.15, which is significant at the 0.05 level. Therefore reject the null hypothesis that the exercise program makes no difference.

25.8 $z_{sample} = 0.84$, which has a two-tail probability of 0.4; therefore do not reject the null.

25.9 (a) $z_{sample} = -2.31$, which has a two-tail probability of 0.02; therefore do not reject the null at the 0.01 level. We cannot say that one program is preferred over the other.

26.1 No; inference tests only apply when generalizing from random samples to the population. Here we have data for the population so there is no need to make an inference.

26.2 The sample t-value is 2.14.

26.3 We reject the hypothesis that there is no correlation between these two variables in the population.

Index

a. See *Y*-intercept, 166
Abscissa, 43
Alpha (α), 222, 271
Alpha error. See Type I error 225
Alpha level, 250
  def, 247
  null hypothesis and, 225
Alternative hypothesis, 217, 218
Analysis of variance (ANOVA) 266-277
  computation of, 269-272
  example of, 274-277
  logic of, 268
  SPSS procedures for, 272-275
  post hoc comparisons for, 273
  test of significance for, 219
ANOVA. See Analysis of variance
Area above or below a score, 149, 152, 154
Area between two scores, 149, 156
Area under normal curve, 152, 216
Arithmetic Mean. See Mean
Association, 12-13
  asymmetric, 82, 104, 379
  def, 82
  existence of, 12
  measures of. See Measures of association
  negative, 76, 98, 164
  pattern and/or direction of, 75-77
  strength of, 75-76
  perfect, 83, 100
  positive, 76, 98
  symmetric, 82, 106, 112
Average. See Measures of central tendency

b. See Ordinary least-squares regression
Bar chart, 42-44
Beta error. See Type II error
Beta-weights, 192
Bias, of estimator, 242
Binomial distribution, 291-303, 305, 379
  def, 291
  hypothesis test for, 293
  SPSS procedure for, 295-296
Bivariate tables, 71, 316, 379, 394, 501
  association between variables, and 81-93,
  chi-square test, and, 219, 305-308, 317-322
  def, 71
  dimensions of, 72
  independence between variables, and, 12
  SPSS procedures for, 74-75

Causal relationships, 115
Causation, 83
Cells, 20, 34
Census, 203, 379
Central Limit Theorem, 210
Central tendency, measures of. See Measures of central tendency
Charts. See Graphs
Chi-square test, 305-313, 316-331
  calculation of, 306-308
  def, 305
  goodness-of-fit, 305-308
  limitations of, 308-331
  logic of, 306
  SPSS procedures for, 308-310, 323-325
  test for independence, 219, 317-322
Class intervals, 61-64, 127-128, 379
  def, 61
Class limits, 62-64
  for histograms, 44
  real, 62
  stated, 62
Coefficient of Determination (r^2), 170-172
Coefficient of Relative Variation (CRV), 140-141, 379
Conditional relationship, 115-116
Confidence interval, 242-253
  and alpha level, 247
  and sample size, 250
  def, 245
  for a mean, 244
  for a percentage, 297-299
Confidence level, 247, 379
Construct validity, 5
Contingency table. See Crosstabulation
Continuity correction, 328
Continuous variable, 7-8, 379
  def, 7
Control variable, 111

Coordinate, 161
Correlation, 164, 180
Correlation coefficient (*r*), 169, 362
Cramer's *V*, 92
Critical region, 222-224, 226, 379
    def, 222
Critical scores, 230
Crosstabulation, 70-72, 324
    Elaboration of, 110
    SPSS procedure for, 74-75
Cumulative frequency, 60, 379
Cumulative percentage, 60

DATA EDITOR window, SPSS, 19-21
DATA FILES, SPSS, 33-35
Deciles, 64
DEFINE VARIABLES command, 21
Degrees of freedom (*df*),
    for ANOVA, 271
    for chi-square test, 323
    for one sample *t*-test, 237
    for two sample *t*-test, 261
Dependent samples, 280, 379
    def, 280
    hypothesis test for means, 280-281
    hypothesis test for nominal data, 333-340
    hypothesis test for ranked data, 353
Dependent variable, 13, 256
Descriptive statistics, 14-15, 380
    def, 14
Discrete variable, 7-8, 380
    def, 7
Dispersion. See Measures of dispersion
Distribution, 56
    binomial, 291-303
    frequency, 127
    F, 371
    normal, 147-159
    skewed, 128
    symmetric, 128
    t, 233, 370
    z, 151, 215, 369
Distribution-free tests. See Non-parametric
    tests

Efficiency, of estimator, 243
Elaboration of bivariate tables, 110-118
    control variable in, 111
    partial Gamma in, 113
    SPSS procedures for, 112-113
Estimation procedures, 243-253
    bias and efficiency, and 246-247
    interval estimates, controlling width of,
    250
    for sample means, 242
    for sample percentages, 297
Exclusiveness, 8
Exhaustiveness, 8
Expected frequency, 306, 309, 311, 319, 325
Explained variation, 141

F-distribution, 371
F-ratio, 271
Frequency, 55, 380
    cumulative, 60, 379
    def, 55
    expected, 306
    observed, 306
    relative, 57, 381
    simple table, 55-57
Frequency distributions, 55-64
    categories for, 56
    computing mean for, 126
    computing median for, 125
    computing standard deviation for, 138
    for interval/ratio scales, 56
    for nominal scales, 56
    for ordinal variables, 56
    SPSS procedures for, 65-66

Gamma (*G*), 99-101
    computation of, 99
    partial, 113-114
    SPSS procedures for, 103
Goodman-Kruskall *tau*, 92
Goodness-of-fit test. See Chi-square test
Graphs, 39-53
    bar, 42-44
    for comparing two variables, 47-49
    frequency polygon, 45-46
    histogram, 44-45
    pie, 40-42
    scatter plot, 161
Grouping variable, 256
    def, 256

Histogram, 44-45
Homoscedasticity, 178
Hypotheses, 380
Hypothesis test, 214-231, 380
    alpha level in, 222
    analysis of variance, 266-269
    five-step procedure for, 217
    for a mean, 233
    for a percentage, 291-293
    for correlations, 362-364
    for difference between means, 255-261
    for difference between percentages, 331

    for goodness-of-fit, 305-313
    for independence, 316-323
    one sample case, 234-238
    one-tailed, 221, 29
    two sample case, 266
    two-tailed, 221, 227, 228

Independence, 380
    Chi-square test for, 219, 317-322
Independent sample, 280
Independent variable, 12
Index of qualitative variation (IQV), 141-143,
    380
Inferential statistics, 380, 204
    def, 204
Interaction between variables, 116
Interquartile range, 137, 380
    def, 137
Interval estimates, 242-253
    controlling width of, 248
    for means, 242-247
    for percentages, 297
Interval-ratio level of measurement, 10, 11
Intervening relationship, 114-115

Kendall's tau-*b*, 102
Kendall's tau-*c*, 102
Kruskall–Wallis *H* test, 219, 357

Lambda, 84-91
    calculation of, 86, 87
    limitations of, 90-91
    SPSS procedure for, 87-88
Least squares regression line. See Ordinary
    least squares regression line
Level of measurement, 8-11
    interval-ratio, 10-11
    nominal, 8-9, 123
    ordinal, 9-10
    specifying in SPSS, 27-28
Linear relationship, 177
Line, equation for straight, 163-164
Line of best fit. See Ordinary least squares
regression line

Mann–Whitney *U* test, 219, 353, 358-359
Marginals of a bivariate table, 72
Matched pairs. See Dependent samples
Mean, 126, 380
    calculation of, 126
    def, 126
    estimating, 242-247
    for grouped data, 126, 127, 128, 346
    for a population, 126
    relationship to median and mode, 128
    for a sample, 126-127
Measures of association, 81-93, 380
    def, 82
    gamma, 99-101
    Kendall's tau-*b*, 102
    Kendall's tau-*c*, 102
    lambda, 84-91
    for nominal scales, 70-73, 83
    for ordinal scales, 98-102
    Somer's *d*, 101
    Spearman's rho, 179-183, 365-366
    SPSS procedures for, 87-88
Measures of central tendency, 123
    choosing, 123
    def, 124
    mean, 126-128
    median, 125-126
    mode, 124
    SPSS procedures for, 130-133
Measures of dispersion, 136-145, 380
    coefficient of relative variation, 140-141
    def, 136
    index of qualitative variation, 141-145
    interquartile range, 137
    range, 136-137
    SPSS procedures for, 145
    standard deviation, 138-140
Median, 125-126, 380
    def, 125
    for grouped data, 125-126
    relationship to mean and mode, 128
Mid-point, 63, 125-126
MISSING VALUES command, SPSS, 25-26
Mode, 124, 380
    def, 124
    relationship to mean and median, 128
Multiple regression, 187-199, 381
    and categorical variables, 197
    assumptions behind, 198
    SPSS procedures for, 190-192
    stepwise regression, 194-198

Negative association, 164
    def, 76
Nominal level of measurement, 8-9
    def, 8
Non-parametric tests, 344, 381
    Kruskall–Wallis *H* test, 219, 357
    McNemar test, 337-338
    runs test, 299-302
    SPSS procedures for, 74, 87-88, 102-103
    Wilcoxon rank sum test, 219, 343
    Wilcoxon signed ranks test, 348-352

- Normal distribution, 147-159
  - area above or below a score, 149, 151
  - area between two scores, 152, 155
  - area under, 147, 149, 150, 152
  - computing z scores, 151-153
  - *def.* 147
  - table for standard, 150, 151, 369
- Null hypothesis, 217
  - for analysis of variance *F* test, 266-269
  - for chi-square test, 316
  - for goodness-of-fit test, 308
  - for McNemar Test, 335-337
  - probability of rejecting, 217-225, 226-227
  - for one sample *t*-test for a mean, 234-238
  - for runs test, 299-301
  - for two sample *t*-test for means, 259-261
  - for Wilcoxon rank-sum test, 348-355
  - for Wilcoxon signed-ranks test, 355, 356-357
  - for one sample *z*-test for a percentage, 294
  - for two sample *z*-test for percentage, 331

- Observed frequency, 306, 309, 311, 319
- One-tail hypothesis tests, 229
- One way analysis of variance. See Analysis of Variance
- Ordinal scales, 10
  - *def.* 10
- Ordinary least squares (OLS) regression line, 166-169, 170, 381
  - calculation of, 167-168
  - *def.* 166
  - scatter plot and, 161-162
  - slope of, 162-165
  - SPSS procedures for, 172-175
  - *Y*-intercept of, 164
- Ordinate, 43

- Parameter, 203, 381
- Parametric tests, 381
- Partial gamma, 113, 114
- Partial tables, 111
- Pearson's correlation coefficient (*r*), 168-170
  - calculation of, 169-170
  - SPSS procedure for, 364-365
  - testing significance of, 364
- Percentages, 58, 381
  - *def.* 58
  - estimating, 297
  - hypothesis test for one, 293
  - hypothesis test for difference between two, 331
- Perfect association, 83, 100, 381
- Pie graph, 40-42
- Polygon, 45
- Pooled variance estimate, 258
- Population, 3, 203, 381
  - *def.* 3
- Population mean, 374
- Positive association, 76, 164
- *Post hoc* comparison. See Analysis of variance
- Probability sampling. See Random sampling
- Proportional reduction in error, 84
  - for nominal scales, 83
  - for ordinal scales, 98
- Proportion, 58

- Quartiles, 137
- Quintiles, 137

- Random sampling, 204-211
  - SPSS procedures for, 210-211
  - stratified, 205
- Random selection. See Random sampling
- Random variation, 204, 320
- Range, 136-137, 381
  - *def.* 136
- Rank, 9, 344, 381
- Raw data, 204
- Rejection of rejection. See Critical region
- Regression and prediction, 164
- Regression coefficient, 168
- Regression line. See Ordinary least squares regression line
- Relative frequency, 73
  - table, 59, 73
  - polygon, 45-47
  - crosstabulations, 73
- Residuals, 381
  - in chi-square test, 307
  - in regression analysis, 165
- Run, 300, 381
- Runs test, 299-302
  - SPSS procedure for, 302-303

- Sample, 204, 381
  - *def.* 4
  - biased, 205
  - random, 204-205, 381
  - snowball, 205
  - stratified random, 205
- Sample percentage, 292-296
  - estimating, 297-299
  - one sample hypothesis test for, 234-239
  - two sample hypothesis test for, 331
- Sampling distribution, 203-212, 381

- SAVE/SAVE AS commands, SPSS, 30-33
- Scatter plot, 161-162, 172-175, 381
  - *def.* 161
- Significance test. See Hypothesis test
- Skew, 128
- Slope. See Ordinary least squares regression
- Somer's *d*, 101
- Spearman's rho, 179-182
  - calculation of, 180
  - SPSS procedures for, 180-181
- Spurious relationship, 114-115
- Standard deviation, 138-140, 381
  - calculations for, 138
  - SPSS procedure for, 132-133
- Standard error, 208
  - effect of sample size on, 208
  - effect of alpha level on, 247
- Standard normal distribution, 148
- Stated class limits, 62, 382
- Statistical significance, 222, 226
  - interpretation of, 226-227
- Stepwise regression, 194-197
  - SPSS procedure for, 196
- Stratified sample, 205
- Statistics,
  - descriptive, 39-199
  - inferential, 203-367
- Student's *t*-distribution, 236, 233
- Sum of squares between, 269
- Sum of squares within, 269

- Tables,
  - bivariate (contingency), 71-73
  - cumulative frequency, 60
  - relative frequency, 57-60
  - simple frequency, 55-57
  - SPSS procedures for, 65-66
- *t*-distribution, 233, 236
- *t*-test, 280-285
  - for a single mean, 234
  - for correlation coefficient, 362-364
  - for difference between means, 219, 257-258, 280-286
  - for mean difference, 280-283
- Test variable, 256
- Tied ranks, 346
- Time series, 299
- Total sum of squares, 269
- Two-tailed test, 227, 228
- Type I error, 225, 382
- Type II error, 225, 382

- VALUE LABELS command, SPSS, 24-25
- Variable, 3-7, 382
  - conceptual definition, 5
  - continuous, 7
  - dependent, 71, 317
  - discrete, 7
  - independent, 12, 71
  - operational definition, 5
- VARIABLE LABELS command, SPSS, 23

- Wald-Wolfowitz runs test, 358
- Wilcoxon rank-sum test, 219, 353
  - SPSS procedure for, 352
- Wilcoxon signed-rank test, 353-357
  - SPSS procedure for, 356

- *X*-axis, 163

- *Y*-intercept (*a*), 164, 166
- *Y*-axis, 201, 163

- *z*-scores, 151-154, 155-156, 382