Chapter 1

Introduction

chapter 1

RESEARCH IS PERFORMED to find answers to questions: what events from their lives do people remember best?; can we judge people's occupations from the way they dress?; what effect does tiredness have on our performance of different tasks? To help us develop answers to these questions we often collect data. We distinguish between two types of data: quantitative and qualitative. Quantitative data concerns numbers or quantities that we have collected using measuring devices such as timers, performance tests or questionnaires. Qualitative data concerns accounts, descriptions and explanations – linguistic rather than numeric data. Most researchers focus on either quantitative or qualitative data collection analysis (and this book is exclusively concerned with the former) but ultimately it is a combination of the two that will provide the fullest insight into our research questions. Consider students undertaking an examination. We might collect information on how many hours they spend studying, how many books they have read and how well they perform in the examination (quantitative data) but we might also ask them for their own explanations of how well they studied, how motivated they were and why, along with what they thought about the experience of taking the examination (qualitative data).

Sometimes, but not often, it is possible to look at that research data and see what it is telling us. Usually, however, the implications of the data are not so obvious, especially when we have collected a large amount of data in numeric form. Simply looking at lots and lots of numbers is usually uninformative and possibly confusing. We need to draw from it the relevant information for the research question posed. This is where statistics can help us. A mass of data can be described and summarised or different sets of data can be compared by the calculation of appropriate statistics. Thus statistical analysis should not be seen as either incomprehensible or esoteric, but as a useful technique for helping the researcher in finding answers to the questions set.

Much of this book is about the various statistics we calculate. Whilst we shall see in Chapter 5 that it has a technical definition, a statistic is essentially a number that has been systematically obtained. A 'total' is a statistic. We can find a total for the number of apples in a bowl or children in a school: we just add them up. Some statistics are easy to obtain (such as the number of fingers on my left hand) whereas others are a little more difficult to work out (such as the F-ratio in the analysis of variance – something we shall be looking at later in the book). However, the purpose of calculating these statistics is to tell us something we want to know: are girls performing better than boys at school?; which of two types of cola do people prefer? It is not the calculation of statistics that is intrinsically interesting (we have computers to do this) but what the statistic tells us about the questions we are interested in. However, the ability to choose the appropriate statistic, and the ability to see whether our calculations are correct or not, are both crucial factors in obtaining a valid answer to our questions, rather than making an error: we don't want to do the statistical equivalent of asking the time and being told it's Tuesday.

We invariably need to calculate statistics when we undertake certain forms of research and having an understanding of what they are and why we calculate them can make us much better able to critically analyse the work of others. If someone informs you that the statistical analysis of their research shows that pigs can fly, and people sometimes do make wild claims as a result of their research, then you might be sceptical about their choice or use of statistics. However, there are many cases where the claims are not so obviously in error yet a simple knowledge of statistical analysis can reveal a flaw.

The purpose of this book is to explain the logic behind statistical techniques, when you would use them and how you would calculate them. Often the latter tends to dominate one's experience, and there is a desire to

just get the thing worked out, but with calculators and computers it is easy to put data into an analysis but less easy to know we have done it correctly. It is understanding why one is calculating a particular statistic that is of crucial importance to data analysis.

The book begins with an explanation of the statistics that help us to describe data, examining what 'frequency distributions' can show us and which summary statistics we can calculate. It then moves on to the importance of the 'normal distribution' and hypothesis testing. The difference between populations and samples is considered along with the use of information from samples to estimate the details of populations. Subsequently the various techniques are introduced that allow us to compare data from different samples.

The book can be read straight through to see the way in which the statistical tests have been developed. These tests all have a logical basis, and explanations are provided for the particular formulae that we use for our calculations. Alternatively, the book can be dipped in and out of, providing enough information on each test so that readers requiring a specific analysis can see why it has been developed and undertake an analysis on their own data by following the worked examples provided.

The final chapter provides an introduction to the model underlying many of our statistical tests. In the explanation of this model we can see why many statistical tests require a particular set of assumptions. Whilst this chapter does not contain any new statistical techniques to learn it is hoped that the reader who does tackle this chapter will gain a deeper understanding of the principles underlying statistical techniques which can lead to a greater appreciation of what in practice is happening when carrying out a statistical analysis.

Chapter 2

Descriptive statistics

chapter 2

A MAJOR REASON FOR CALCULATING statistics is to describe and summarise a set of data. A mass of numbers is not usually very informative so we need to find ways of abstracting the key information that allows us to present the data in a clear and comprehensible form. In this chapter we shall be looking at an example of a collection of data and considering the best way of describing and summarising it.

One hundred students sit an examination. After the examination the papers are marked and given a score out of one hundred. You are given the results and asked to present them to a committee that monitors examination performance. You are faced with the following marks:

22	65	49	56	59	34	9	56	48	62
55	52	78	61	50	62	45	51	61	60
54	58	59	47	50	62	44	55	52	80
51	49	58	46	32	59	57	57	45	56
90	53	56	53	55	55	41	64	33	0
38	57	62	15	48	54	60	50	54	59
67	58	60	43	37	54	59	63	68	60
46	52	56	32	75	57	58	47	45	52
55	51	50	50	69	63	64	49	56	52
37	60	71	26	30	57	56	55	58	61

Fortunately, you are told the sort of questions the committee might ask:

- Can you describe the results of the examination?
- Can you give us a brief summary of them?
- What is the average mark?
- What is the spread of scores?
- What is the highest and lowest mark?
- Here are last year's results, how do this year's compare?

You sit looking at the above table. The answers to the questions are not obvious from the 'raw' data, that is, the original data before any statistics have been calculated. We need to do something to make it clearer. The first thing that we can do is to list the data in order, from lowest to highest:

0	9	15	22	26	30	32	32	33	34
37	37	38	41	43	44	45	45	45	46
46	47	47	48	48	49	49	49	50	50
50	50	50	51	51	51	52	52	52	52
52	53	53	54	54	54	54	55	55	55
55	55	55	56	56	56	56	56	56	56
57	57	57	57	57	58	58	58	58	58
59	59	59	59	59	60	60	60	60	60
61	61	61	62	62	62	62	63	63	64
64	65	67	68	69	71	75	78	80	90

With this ordering certain things are more apparent: we can now see the lowest and highest scores more easily, with the scores falling between 0 and 90.

Another thing we can do to improve our presentation is to add up the number of people who achieved the same mark. We work out the frequency of each mark. For example, 5 people scored 52 and only 1 scored 69. When we do this it allows us to see that the most 'popular' mark was 56 with a frequency of 7. We should not forget that there are a number of possible marks that no one achieved: no one scored 8 or 35 for example, so each of these marks has a frequency of 0.

We can present this information in graphical form if we convert it to a histogram, where the frequency of a mark is represented as a vertical bar. In the histogram, shown in Figure 2.1, we list out all the possible marks that a

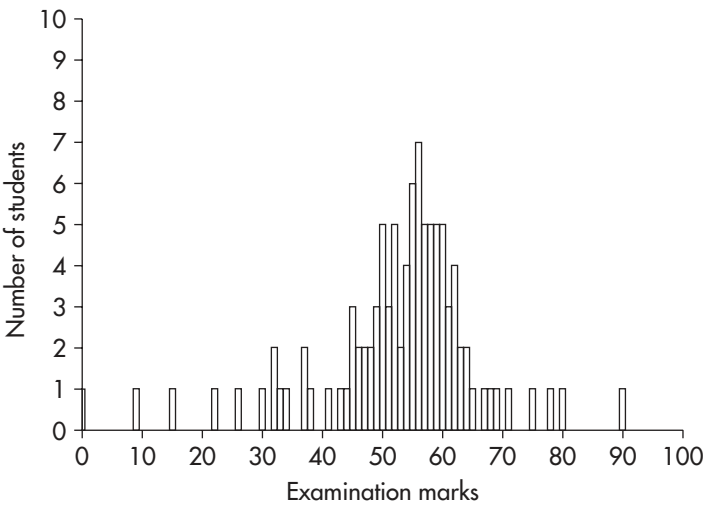

FIGURE 2.1 Frequency distribution of the examination results

student could get, 0 to 100, and draw a bar above each mark, with the length of the bar corresponding to the frequency of the mark in the set of results. For a mark of 55 we draw a bar of length 6 (as 6 students obtained a mark of 55) and for 64 we draw a bar of length 2. This gives a clear visual presentation of the results.

This histogram is called a frequency distribution, as we can see how the marks are distributed across the range of possible marks. Frequency distributions are very important in statistical analysis as they provide the basic representation of our information. The frequency distribution is a clear infor-mative chart, providing us with a way of showing the pattern of the marks we obtained: their distribution across the range of possible values. We might wish to present the frequency distribution to the committee as it provides us with a graphical representation of the marks. But what it doesn't do is provide us with a summary of the findings.

Measures of 'central tendency'

Is there a single mark that best represents the results? Can we provide the committee with a typical mark to summarise the findings? The most reasonable mark to use here is a central or middle mark. In statistical terms we are trying to find a measure of central tendency. The question we are now faced with is: what is the central position in our frequency distribution?

One answer is simply to select the most frequent mark, the longest bar in the histogram. This statistic is called the mode. As you can see from Figure 2.1 the longest bar is at the mark of 56, where seven people obtained this result in the examination. In this case 56 appears to be a reasonable estimate of a central mark. However, the mode is not often used as a measure of central tendency for a number of reasons. First, what do we do if there were two marks each having the same high frequency? What if seven people had scored 52 and seven 56, which one would we choose? Second, there will be occasions where the mode clearly does not represent a central mark. Imagine that we had ten very weak students who all scored zero in the examination, yet the rest of the distribution was the same as in Figure 2.1. Even though there would be a clustering of the marks in the 50s our mode would be zero. In this case the mode would be a poor measure of central tendency.

Another measure of central tendency that is used more often than the mode is the median. This is the score that comes in the middle of the list when we have ordered it from lowest to highest. If we had nine students in all then the median would be the fifth mark in the list. However, we have

one hundred students and, with an even number, there is no middle mark. The middle lies halfway between the fiftieth and fifty-first marks. In our example the fiftieth and fifty-first marks are both 55, so the median is 55. (If the fiftieth and fifty-first marks had been different the median would be halfway between them. We would simply add them up and divide by two to get our median value.[1])

The median is a good measure of central tendency as it picks up the score in the middle position of the distribution. Its weakness, if indeed it is a weakness, is that, like the mode, it does not use all the information given by the marks. The median is simply the score where we cut our list into two halves. The marks either side of the median could be anything below or above the median respectively. If we found that someone who had had been given a mark of 9 in the examination really had a mark of 29 or 39, correcting this score would not change the median as 55 would still be the middle mark in the list. The median would stay the same even if a number of marks were changed (as long as a mark below the median was not changed to a value higher than the median or vice versa). The median doesn't take account of the values of all the scores, only the value of the score at the middle position.

Whilst we might regard the median as a better choice of a central value than the mode, as it finds the score at the middle position rather than the most frequent score, there is a third measure of central tendency that is used far more often than either of the above two measures. This is the mean.

We express the formula for calculating the mean using special symbols. We use the Greek letter μ (pronounced 'mu') for the mean, the Greek letter capital sigma, $\sum$, to mean 'the sum of' (or 'add up'), X to indicate a score (in our example, an examination mark) and N for the number of scores. The symbols $\sum X$ means 'add up all the scores'. The mean, μ is the sum of the scores divided by N:

$$\mu = \frac{\sum X}{N}$$

When we talk of an 'average' we are usually referring to the mean (although the word 'average' is often used much more loosely than the word 'mean' which has its statistical definition). To calculate the mean we add up all the marks and divide them by the number of students. Adding up all the marks we arrive at 5262. Dividing this by 100 gives us a mean of 52.62.

One way of thinking about the mean is by analogy with a see-saw. Imagine that the horizontal axis of our frequency distribution is a beam of

wood going from 0 to 100 in length. Each of the marks is a student sitting on the beam at the position specified by their mark (so there are seven students sitting on the beam at 56 and one at 75 etc.). Where would you have to put a supporting post under the beam to make a perfectly balanced see-saw? The answer is at the mean position. We can see it as the value that balances the scores either side of it. Any change in the marks (we move a student along the beam) results in a change in the mean (the see-saw will tip to one side unless we move the supporting post to a new position to restore balance). So the mean is a statistic that is sensitive to all the scores about it, unlike the median, as we saw above.

There is another point about the mean that we can see from the see-saw analogy; that is, the mean is very sensitive to extreme values. A very large score or a very small score will have a greater effect on where the support post ends up than a mark in the middle of the distribution. If you have a number of people sitting on a balanced see-saw it tips up much more easily if a new person sits on an end rather than near the middle. Thus, the mean position, like the supporting post of a see-saw, is determined both by the number of scores and also by their distance from it.

Comparing measures of central tendency

In our example we now have three measures of central tendency, a mode of 56, a median of 55 and a mean of 52.62. Which do we choose? The answer is: whichever we want. We simply choose the one that best represents a central value in our distribution, for our purpose. Usually this results in us picking the mean as it takes into account all the scores but there are occasions when we choose the mode or median.

The mode is quick and easy to determine once we have created the frequency distribution, so we might use it as a 'rough and ready measure' without the need for further calculation. Also we cannot calculate the median or mean with some types of data. For example, if I am planning a trip for a group of friends and I suggest a range of places to visit, I'll probably select the place chosen by the largest number. Note that we cannot calculate a mean or a median here as the names of places cannot be put in numeric order or added up.

We use the median when we have an abnormally large or small value in our frequency distribution, which would result in the mean giving us a rather distorted value for the central tendency. As an example, six aircraft have the following maximum speeds: 450 km/h, 480 km/h, 500 km/h,

530 km/h, 600 km/h and 1100 km/h. We can see that most have a maximum speed around 500 km/h but the inclusion of the supersonic aircraft with a speed of 1100 km/h gives us a mean of 610 km/h. This number might not be appropriate to use as a central value as 610 km/h is faster than five out of the six aircraft can fly. If we take the median, which is 515 km/h (halfway between 500 and 530) we have a more representative value for our central point.

However, in most cases of data collection the mean is the measure chosen. We shall see further reasons for the importance of the mean in Chapter 5.

Measures of 'spread'

So far we have charted our data on a frequency distribution and found measures of central tendency. Another useful statistic for summarising the data is a measure of 'spread'. It is important for a number of reasons to find out how spread out the scores are. Two groups of students taking the same examination could produce different frequency distributions yet the means might be the same. How then can we express the difference in the distributions? It is almost certain that the marks for one group of students are more spread out than the other. A small spread of results in a study is often seen as a good thing, as it indicates that all the people (or whatever produces the scores) are behaving similarly, and hence the mean value represents the scores very well. A large spread may be a problem as it indicates that there are large differences between the individual scores and the mean is therefore not so representative. Thus, we want a statistic that gives us a small number when the scores are clustered together and a large number when the scores are spread out.

The range

The simplest measure of spread is the range. The range is the difference between the highest and lowest scores. In our example the highest score is a mark of 90 and the lowest is 0. The range is therefore 90.

This measure is a little crude, it sets the boundaries to the scores but does not tell us anything about their general spread. Indeed, even if our marks were evenly spread between 0 and 90 rather than clustered in the 50s, our range would still be 90. The range uses information from only two

scores, the rest could be anything between, so it is rather limited in what it tells us.

Quartiles

Another way of looking at the spread is to calculate quartiles. We saw earlier that the median cuts the ordered data into halves; the quartiles simply cut the ordered data into quarters. The first quartile indicates the score one quarter of the way up the list from the lowest. The second quartile indicates the score two quarters up the list. It does not take very much to realise that the second quartile is halfway up the list and is therefore the median. The third quartile is the score three quarters up the list. The fourth quartile is all the way to the end of the list and so it is the highest score.

From our ordered list of examination results, one quarter along the list of a hundred scores lies between the twenty-fifth and the twenty-sixth person's marks, so the first quartile is midway between 48 and 49, which is 48.5. We already know that the second quartile (between the fiftieth and fifty-first person's marks) is 55 as we worked out the median above. The third quartile is three quarters along the list so is between the seventy-fifth and seventy-sixth person's marks: this is 59.5.[2] And of course the fourth quartile is 90, as it's the highest score. If we use the symbol Q for quartile, we have $Q_1 = 48.5$, $Q_2 = 55$, $Q_3 = 59.5$, $Q_4 = 90$.

A slightly more sophisticated measure of spread than the range is the interquartile range: that is the difference between the third and first quartile, $Q_3 - Q_1$. In our example this is $59.5 - 48.5 = 11$. This is the range of half the scores, those in the middle of the distribution. The reason why the interquartile range is used is that, unlike the range, it is not going to be affected by one particularly high or low score and may represent the spread of the distribution more appropriately. (Some people use the semi-interquartile range, which is simply half the interquartile range. In our example this is 5.5.)

Calculating quartiles is quite useful as it can tell us a few interesting things about the distribution, in particular whether the distribution is symmetric about the median within the interquartile range. $Q_2 - Q_1$ tells us the range of the quarter of scores below the median and $Q_3 - Q_2$ tells us the range of the quarter of the scores above the median. In our example the first is 6.5 and the second is 4.5. We have the scores bunched closer together in the quarter above the median than in the quarter below the median, as 4.5 is a smaller range than 6.5, for the same number of scores.

It is worth noting here how each new statistic tells us something different about the data. It may be something we already know by looking at the distribution but often the statistic makes it clearer and more explicit, with a number attached. However, these statistics do not miraculously appear. They have been created by people attempting to find ways of best describing their data. When we wish to describe our data we choose the most appropriate statistic for our purposes.

Variation

Calculating quartiles does not use all the information available from the scores in the data, and again, as in our discussion of the median, some scores could be different and we would still end up with the same interquartile range. The question therefore is whether we can devise a measure of spread that takes into account each and every score. It is in answer to this question that a number of measures of spread have been developed. The common feature of them is that they all begin with the mean (once again indicating the importance of the mean). Their logic is as follows. If we take the mean as our 'central' position then we can compare each of the scores with the mean and find out how far each score varies or deviates from it. If we add up the deviation of each of the scores from the mean we will have a measure of the total variability in the data. If we want to we can then divide this total by the number of scores to find the average deviation of a score from the mean.

We can calculate the deviation of a score from the mean by simply working out $X - \mu$, where X is a score and μ is the mean. We can do this for every score. However, we have a problem: when we add them up to find the total variability, the deviations tend to cancel each other out. In our example, a mark of 55 gives a deviation from the mean of $55 - 52.62 = +2.38$ and a mark of 50 gives a deviation from the mean of $50 - 52.62 = -2.62$. If we add up these deviations we get 2.38 plus −2.62, which equals −0.24. Due to the minus sign, two scores, both over two marks from the mean, end up giving a deviation of less than one when added up. We do not want this; it is not a statistic that reflects the variability as it really is. Indeed, as the mean is the position of 'balance' in the scores, adding up all the deviations will give us a total of zero as all the positive deviations exactly cancel out the negative deviations. As the sum of the deviations of our scores always turns out to be zero whatever scores we have, it is useless as a statistic as it certainly does not provide us with a measure of how spread out the scores are.

When we consider it, all that the minus sign of a deviation is telling us is that the score is lower than the mean. We are not actually interested in whether the score is higher or lower than the mean only how far away it is from the mean. What we need to do is to find a way of adding up the deviations so that they do not cancel each other out, so that we end up with a reasonable estimate of the real variability of the scores. There are two solutions:

1 Absolute deviation

We can solve our problem by ignoring the minus sign altogether and treat all the deviations as positive. If we get a deviation of −2.62 we call it +2.62. We put two vertical lines round a formula to indicate that we take the absolute value, that is, ignore a minus sign in the solution and treat it as positive. Absolute deviation is $|X - \mu|$. We add up the deviations for all the scores. To find the average deviation we divide it by the number of scores, denoted by N. We call this the mean absolute deviation and represent it by the following formula:

$$\text{Mean absolute deviation} = \frac{\sum |X - \mu|}{N}$$

For our examination results the mean absolute deviation is 9.15.

2 Variance

An alternative solution to taking absolute values is to square the deviations, as the square of a number is always positive. The square of −2.16 is +4.67. We then add up the square of each of the deviations to produce a sums of squares: $\sum(X - \mu)^2$. This formula can be can be translated into English as: 'find the deviation of each score from the mean, square each deviation, then add up the squared deviations'. We can then divide this figure by the number of scores (N) to find the average of the squared deviations. This value is called the variance.

$$\text{Variance} = \frac{\sum(X - \mu)^2}{N}$$

In our example the variance is 176.52.

The variance gives us a figure for the average variability of the scores about the mean, expressed as squared deviations. It also does what we want: gives us a large figure for scores that are spread out and a smaller one for scores that are closer together. Interestingly, as it is dealing with squared deviations, the variance gives more weight to extreme scores. For example, a score that deviates by 2 from the mean will contribute 4 to the variance but a score 4 away from the mean will contribute 16 to the variance, so even though the second score is only twice as far from the mean as the first it contributes four times as much to the variance.

If we just wanted a measure of variability then variance is fine. However, note that the figure we calculated of 176.52 cannot be placed on the frequency distribution as a distance from the mean. This is because the variance is the average of the squared deviations, rather than the average deviation. To bring the statistic back to the terms we started with we need to find the square root. (As we squared the deviations earlier to get rid of the minus signs we need to 'undo' this now it has served its purpose.) We call this statistic, the square root of the variance, the standard deviation and represent it by the symbol σ (the lower case Greek letter sigma).

$$\text{Standard deviation, } \sigma = \sqrt{\frac{\sum(X - \mu)^2}{N}}$$

A simple example will show how we calculate the standard deviation. Imagine that we only had 4 scores 2, 2, 3, 5 in our data. The mean is 3. We work out σ as follows:

Score X	*Deviation* $X - \mu$	*Squared deviation* $(X - \mu)^2$
2	−1	1
2	−1	1
3	0	0
5	2	4
		$\sum(X - \mu)^2 = 6$

Dividing the sums of squares ($\sum(X - \mu)^2 = 6$) by the number of scores ($N = 4$) gives a variance of 1.5. Taking the square root of 1.5 gives us a

standard deviation of $\sigma = 1.22$. In the examination example the standard deviation of the one hundred marks is 13.29.

The standard deviation gives us a measure of spread about the mean. In many cases most of the scores (about two-thirds) will lie within one standard deviation less than and one standard deviation greater than the mean, that is, in the range $X - \sigma$ to $X + \sigma$. The standard deviation gives us a measure of the 'standard' distance of a score from the mean in this set of data.

WARNING! The above formulae for variance and standard deviation are used when are interested in these data *only*. When our data is a subset or a sample of a larger set of data that we want it to *represent* then we use slightly different formulae, the same as the above except that we would divide the sums of squares by the degrees of freedom *df*, rather than the sample size *n*, where $df = n - 1$. If it was the case that our one hundred students were not the complete set we were interested in but were a sample drawn from one thousand students taking the examination then we would use the different formulae:

$$\text{Variance} = \frac{\sum (X - \mu)^2}{n - 1}$$

$$\text{Standard deviation, } \sigma = \sqrt{\frac{\sum (X - \mu)^2}{n - 1}}$$

The reason for the difference is explained in Chapter 5 when we consider samples. Most of the time we use the formulae with $n - 1$ (the degrees of freedom) rather than n (the sample size) as most of the time we wish to use our samples to generalise to a wider population rather than treating our data as all that we are interested in.

Comparison of measures of spread

As with the measures of central tendency, the measure of spread that is most useful depends on the reasons for calculating it.

The range and interquartile range are both easy measures to calculate, giving a limited but potentially adequate measure of spread. Their weakness is that they do not take into account all the scores in the data and may be limited in their ability to represent the true variability of the scores. The range, in particular, may not reflect the general spread of results very well if there is one very low or very high score.

The variance is a good measure of the variability in the data. It uses all the scores and will give a small number if all the scores cluster round the mean and a large number if they are spread out. As we shall see in the chapters on the analysis of variance this statistic is extremely important in some statistical analyses. However, when we are describing a set of data the variance may not be particularly useful as a description of the spread of scores as the number it produces is not of the same order as the scores. It is expressed in terms of the squared deviations from the mean. In our example the variance of 176.52 appears large but this may be because it is expressed in terms of marks squared, not marks.

The mean absolute deviation and the standard deviation are both good descriptive statistics of the spread of a set of scores. They both use the information from all the scores and both produce a number that expresses an 'average' deviation from the mean in the terms we want (in our example: marks). As they are expressed in the same terms as the scores they are easy to understand. We can, if we wish, plot these figures as a distance from the mean on the frequency distribution, so they can be graphically represented as well.

Why is it that the spread of a set of results is almost always expressed, in research reports, as the standard deviation and rarely as the mean absolute deviation? If the data we are describing is all we are interested in then there is not a compelling argument. However, there is a distinct advantage of using the standard deviation when our data is a sample of a larger set (a population) that we wish it to represent. In our example, the 100 students were the only ones we were interested in. If, however, 1000 students had taken the examination and our 100 were a representative sample then we would want to use the standard deviation. The reasons, which are dealt with in Chapter 5, concern samples representing populations and the use of sample statistics to estimate population values.

Describing a set of data: in conclusion

When describing a set of data we want to summarise the frequency distribution by two measures, one indicating a central value indicating the 'average' score and a second to indicate the spread of the scores. The two most commonly used statistics for these measures, because of their usefulness, are the mean and the standard deviation. We can summarise the examination results by the following statistics: mean = 52.62 marks, standard deviation = 13.29 marks.

Comparing two sets of data with descriptive statistics

Summary statistics neatly and briefly describe the data but in most cases people want to use the information to make certain points. In our example a committee member might be concerned about possible falling standards or the effect of a change in the student selection procedure. The summary statistics can then be used to help make a decision about such questions. Notice that the points raised by the committee member both require a comparison with the previous year's results. The calculation of statistics is often used to go beyond description to allow us to answer specific research questions and this invariably involves comparing different sets of results.

For our example, the previous year's results for the same examination, where 100 students also sat the examination, are shown below. Note that it would not be easy to see any similarities or differences between the results for the two years very well by simply looking at the two tables of raw data. Both years have a mixture of marks in them and, whilst we might be able to pick out certain interesting results, such as the highest and lowest, the tables do not provide a good way of allowing us to make any comparisons between the two sets of data.

24	56	54	56	55	43	55	52	45	58
54	52	65	50	60	57	47	62	7	58
51	60	53	81	59	61	56	63	57	49
68	61	39	59	49	63	54	60	57	60
66	53	36	50	59	52	37	70	66	30
61	50	55	55	65	58	51	22	68	57
87	64	50	35	56	54	60	72	58	51
46	62	56	15	63	59	39	60	58	76
65	36	4	59	57	53	49	69	64	53
38	58	48	58	66	62	56	54	61	63

Again, if we order the data and create a frequency distribution we might begin to see where any differences lie. Figure 2.2 shows the frequency distribution of last year's results. We can compare Figure 2.2 with Figure 2.1 by eye. The distribution of results looks similar over the two years. This in itself might be useful evidence indicating a year on year consistency in performance. However, simply looking by eye cannot really tell us how similar the two distributions are, as we may miss subtle differences between them. This is where statistics can help.

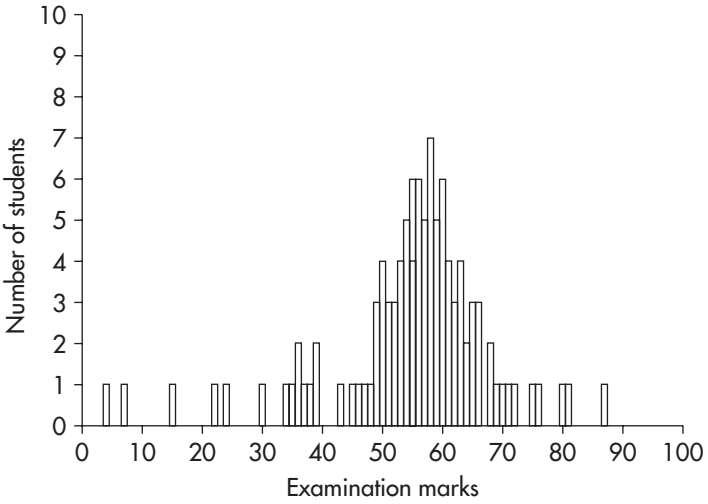

FIGURE 2.2 Frequency distribution of last year's examination results

If we consider measures of central tendency first, we can compare the two years directly:

	Last year	*This year*
Mode	58	56
Median	56.5	55.0
Mean	54.25	52.62

We can see that all three measures have dropped a little since last year. The mode could easily change by the effect of just a few students and so in this case is not the most useful statistic. The median does indicate that the central point was higher last year. The mean value shows a drop of 1.63 from last year to this. It may not seem a lot but remember the mean takes into account all the students, so there is a reduction of 1.63 marks per student. Now this could be due to a number of factors that are worth further investigation, such as: are there less able students this year or is it a harder examination this year? Before we do that we want to eliminate a simple alternative reason. Maybe last year there were a few particular good students

or this year a few poor ones. These occur now and again and do not indicate that the general standards are changing. The way we can look at this is by considering the spread: maybe the spread was wider in one of the two years indicating a greater mixture of student ability in that year?

We can compare the various measures of spread:

	Last year	*This year*
Range	83	90
Interquartile range	10.5	11.0
Mean absolute deviation	8.82	9.15
Variance	169.93	176.52
Standard deviation	13.04	13.29

There was a narrower range last year with no one scoring as low as 0 or as high as 90, but there was not much difference in the interquartile range and, more particularly, the standard deviations are not very different from each other. It might be worth investigating further to see why there was the reduction in the mean performance. Notice that these results alone cannot distinguish between reasons for a difference, they can only be used to argue that one has occurred. The reason for the slightly lower marks, be it lower ability students, a harder paper, stricter marking or whatever, requires the skill of the researcher to find out.

As can be seen from the above figures, the mean and the standard deviation are generally the most informative statistics for a particular distribution. These are the statistics that are most commonly chosen, but there may be occasions when you think that other statistics are more appropriate or will tell you more accurately what you want to know. This leads to an important point: it is NOT worth calculating statistics until you know why you are doing it and what you want the statistics to show. It could be that the raw data tells you all you need to know, so do not bother calculating statistics. However, most of the time it is not possible to see the important characteristics of the data without some further analysis. Calculating the appropriate statistics can help you decide the answers to the questions you are asking. The difficulty in describing and analysing data is NOT calculating the statistics (we have computer programs that do this) but in knowing the questions you wish to find answers for, and the statistics that help inform those answers.

Note also that calculating statistics only gives you information. It is up to you how you interpret and use that information. A difference in means, or standard deviations, might be useful information, but that is all. Calculating statistics will not explain similarities and differences between distributions. What the statistics do is to provide us with pieces of information we can work with: they are tools to be used for our own purposes. After that we must use our judgement.

Details on how to produce statistics to describe a set of data using the SPSS computer statistical package can be found in Chapter 3 of Hinton *et al*. (2004).

Some important information about numbers

Up to now we have been calculating statistics using sets of examination results. This is fine as examination results are the types of number that it makes sense to calculate means and other statistics on. But this is not the case for all types of number. We need to know what type of data we have before we know what statistics we are able to calculate.

Nominal data

Sometimes numbers are used like names. For example, in a sports squad of 22 players the number 15 on the back of a player's shirt simply allows us to identify him or her during play. It does not mean that player number 15 is better than players 1 to 14 or worse than players 16 to 22. It is meaningless to calculate statistics on these numbers as they are only nominal, used as names.

When we categorise someone or something we can use numbers to label the categories. For example, if we classify people by eye colour we might choose to label brown as 1, blue as 2, green as 3 and so on. Notice that the numbers are arbitrarily assigned to colours: we could have chosen other numbers or assigned the same numbers in a different way. The use of these numbers is nominal. We cannot use these numbers to calculate statistics: it is nonsense to say that the mean of a brown eyed person (1) and a green eyed person (3) is a blue eyed person (2)!

Ordinal data

We can use numbers to define an order of performance. For example, Susan is the best chess player in the class, followed by Robert, Marie and Peter. We can give Susan the top rank of 1, Robert 2, Marie 3 and Peter 4. These numbers tell us the rank order but little else. They do NOT tell us that the difference between 1 and 2 (Susan and Robert) is the same as the difference between 3 and 4 (Marie and Peter) despite there only being one place between them in the ranks. Susan could be the best player for her age in the country whereas the other three might not be as good as others of their age from nearby schools. Because of this we cannot calculate means and standard deviations on ordinal data. Chapter 16 discusses ordinal data further and considers how we can calculate statistics with it.

Interval and ratio data

Time, speed, distance and temperature can all be measured on interval scales and we have clocks, speedometers, tape measures and thermometers to do it. They are called interval scales because the differences between the consecutive numbers are of equal intervals: the difference between 1 and 2 is the same as the difference between 3 and 4 or 10 and 11. Unlike an ordinal scale where these could be different, on an interval scale they are all the same. For example, the difference between 6 and 7 minutes is the same as the difference between 20 and 21 minutes, it is 1 minute in both cases. When our numbers come from a scale with equal intervals then we can calculate means and standard deviations.

Ratio data is a special kind of interval data. With interval data the zero value can be arbitrary, such as the position of zero on some temperature scales: the Fahrenheit zero is at a different position to that of the Celsius scale, whereas with ratio data zero actually indicates the point where 'nothing' is scored on the scale, such as zero on a speedometer when there is no movement, and so this zero means the same thing regardless of whether we are measuring in miles per hour or kilometres per second. We can illustrate the difference in the following example. In an examination there are 100 questions of equal difficulty and students are required to get at least 50 correct answers to pass the examination. The examiner could choose to label the pass mark as zero. A score of 0 indicates 50 correct answers, +1 indicates 51 correct answers, −1 indicates 49 correct answers and so on. This is an interval scale with an arbitrary zero: the examiner chose where to

put it. Now let us consider the same examination where zero indicates no correct answers and the pass mark is given a score of 50. This time the zero is nonarbitrary as it specifies a score of 'nothing' in terms of examination performance. Here the interval scale becomes a ratio scale.

Only on a ratio scale, with a genuine zero, can we make claims to do with ratios, such as: Susan's score is twice as good as John's, Robyn's score is one third of Peter's. If Susan had scored 80 and John 40 on a ratio scale examination then her score really is twice John's score. On an interval scale with zero set arbitrarily at 50 their scores are 30 and −10. With the interval scale we would not have been able to make the ratio judgements appropriately.

Many of our statistics require interval or ratio data. In the majority of the book (up to Chapter 16) we shall be considering only data that is interval or ratio as these types of data allow us to perform the largest range of statistical tests. For this reason, researchers often choose to collect interval or ratio data for analysis. With human subjects research often focuses on how fast or how accurately a task can be performed, where both *speed* and *accuracy* can be measured on ratio scales.

Chapter 3

Standard scores

chapter 3

Comparing scores from different distributions

If you received a mark of 58 in an examination would you know how well you had done relative to the other candidates? Were you the best in the class or the worst? Clearly this is a case where you need further information. With the mean and standard deviation you can begin to answer these questions. If the mean is 52 and the standard deviation is 5 then your score is one of the best. If, however, the mean is 59 and the standard deviation is 3 then you are a little below average but as the scores are clustered around 59 there are likely to be a lot of students with similar results.

If you took two examinations and received a 58 for Psychology and a 49 for Statistics, which would you be most pleased with? You might want to use these results to help a decision on which subject to major in. You could choose the 58 because it is numerically higher. But if you found out that everyone else who took the Psychology examination scored over 60 and all the others who took the Statistics examination scored under 45 then you might change your mind. Even though you received a higher mark for Psychology the distributions of the two sets of scores are different. It could be that the Statistics examination is especially hard this year and 49 is a very high mark compared to the rest of the class, whereas 58 in Psychology might be a relatively low mark.

You then find out that, for the Statistics examination, the mean is 45 and the standard deviation is 4, and for Psychology the mean is 55 and the standard deviation is 6. This at least tells you that you are above average in both subjects but it doesn't tell you which yielded the higher class position.

To compare two scores that come from different distributions we need to standardise them. We do this by calculating a statistic called, not surprisingly, a standard score (or z score). This expresses the score relative to the mean in terms of the standard deviation. Thus a score of 58 is 3 away from a mean of 55. With a standard deviation of 6, this distance is 3/6th of the standard deviation. The score is half a standard deviation from the mean. Essentially the standard score tells us how many standard deviations the score is from the mean of the distribution. We calculate the standard score using the following formula:

The standard score, $z = \dfrac{X - \mu}{\sigma}$,

where X is the score to be standardised, μ is the mean and σ is the standard deviation of the distribution.

Standard scores can be compared, because, no matter what your distribution is like to start with, converting scores to z scores always results in a distribution of z scores with a mean of 0 and a standard deviation of 1. If the examination scores are converted to standard scores then we can compare them and see which examination result gives the higher class position.

In Psychology, $X = 58$, $\mu = 55$, $\sigma = 6$:

$$z = \frac{X - \mu}{\sigma} = \frac{58 - 55}{6} = \frac{3}{6} = 0.5$$

In Statistics, $X = 49$, $\mu = 45$, $\sigma = 4$:

$$z = \frac{X - \mu}{\sigma} = \frac{49 - 45}{4} = \frac{4}{4} = 1.0$$

In Psychology you are half a standard deviation above the mean and in Statistics you are one standard deviation above the mean. The higher z score for Statistics means that you are higher in the class results for Statistics than you are for Psychology.

In the previous chapter we compared two sets of examination results, from this year and last year. Notice that this year a score of 59 gives the following z score:

$$z = \frac{59 - 52.62}{13.29} = \frac{6.38}{13.29} = 0.48$$

For last year's distribution a score of 59 produces the following z score:

$$z = \frac{59 - 54.25}{13.04} = \frac{4.75}{13.04} = 0.36$$

From these two z scores we can see that a score of 59 is higher up the distribution this year ($z = 0.48$) than last year ($z = 0.36$), so 59 is a better score this year than last, possibly because the examination is harder this year (or one of the other reasons cited earlier).

The Normal Distribution

If I decided to collect data on, say, women's heights I might initially measure the height of a large number of women and plot the results as a frequency distribution on a histogram. What would the distribution look like? I start by choosing the steps for my histogram, i.e. deciding on the range of values to include for each bar. I'll choose 5 centimetre steps and include in the same bar all the women whose height falls within a particular 5 cm band. (To stop overlapping bands, the band includes heights from the lowest point of the band up to but not including the highest point of the band: for example, the band 160 cm to 165 cm covers the women's heights from 160 cm up to but not including 165 cm, so the woman whose height is exactly 165 falls into one band only – the 165 cm to 170 cm band.) When I have collected the data and added up all the women whose height falls within each 5 cm band I would find lots of women whose height was between 160 and 165 cm, or between 165 cm and 170 cm but not many between 135 cm and 140 cm or between 185 cm and 190 cm. There are not as many very short or very tall women compared to those in-between. In fact, the distribution would probably look like the histogram in Figure 3.1. Notice the distribution has a hump in the middle and tails off symmetrically either side.

If I then kept on measuring more and more women and also made my steps smaller (instead of 5 cm I choose 2 cm bands, then I plot

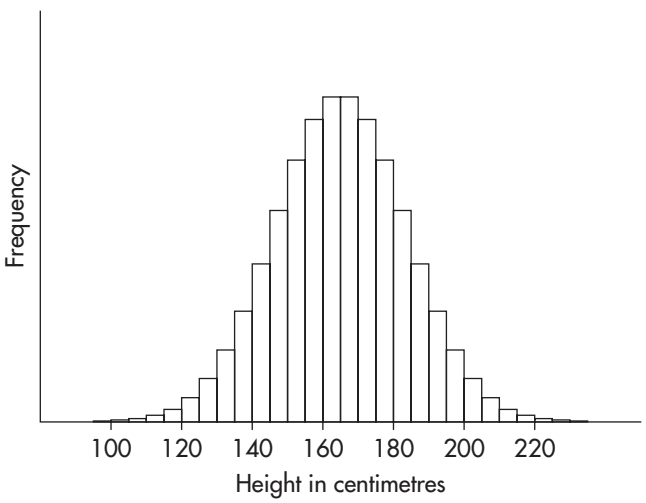

FIGURE 3.1 The distribution of women's height: histogram

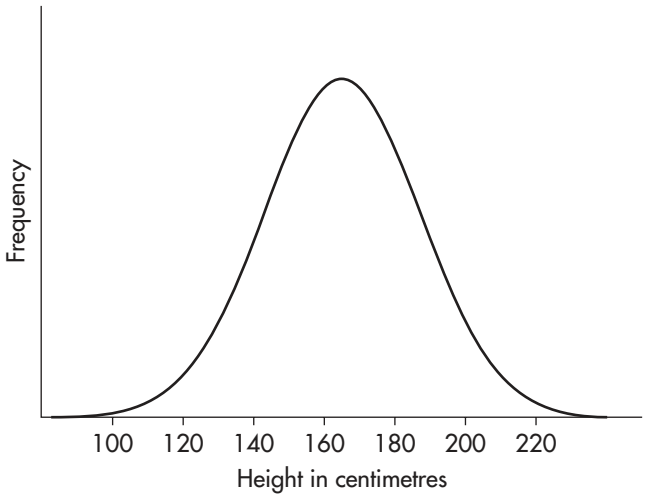

FIGURE 3.2 The distribution of women's height

the heights within 1 cm bands, then 0.5 cm and so on until my bands become extremely small) I would end up with a very large number of women's heights plotted on a histogram with very small steps. Eventually my histogram would become a smooth curve, as in Figure 3.2.

It is remarkable how many times we end up with this same bell-shaped curve, irrespective of which variable we are studying, be it women's heights, the foot size of ten year old boys or the gestation period of babies. As the curve is produced so often it is called the Normal Distribution. The interesting and very useful feature of this curve is that it is actually quite simple to express mathematically and can be calculated using only the mean and standard deviation. That is, we can work out the formula for a normal distribution precisely with only the knowledge of the mean and standard deviation.[3]

The normal distribution is very important for statistical analysis for the following reasons.

1 Many of the things we study and measure in our research (although not all) are assumed to be come from a population of scores that are normally distributed (such as women's heights). If we took all the men in the population we would expect to get normal distributions for height, weight, foot size, etc. We would expect normal distributions for the women's data as well.

2 Many of the statistical tests that we shall be examining in the course of the book make the assumption that the distributions they are investigating are normally distributed. Indeed these tests rely on this assumption: without it the logic of the test fails.

3 Interestingly, even if a distribution is not a normal distribution, when we take a large number of samples of the same size and plot their means on a frequency distribution this distribution tends to become a normal distribution. This again is extremely useful for statistical analysis.

These points are examined further in Chapter 5 when we consider samples, but the important thing to note here is that we have a lot of useful information when we know the mean and standard deviation of a set of scores and also that the distribution of the scores is a normal distribution.

The Standard Normal Distribution

As it is such a useful distribution people have drawn up tables of the normal distribution. However, the values would be different for all the various means and standard deviations we could get, and we would end up with lots and lots of different tables. So the values in the table are for a normal distribution with a mean of 0 and a standard deviation of 1. This normal distribution is called the Standard Normal Distribution.

If scores come from a normal distribution (such as height, weight) then converting the scores to standard scores (z scores) converts the distribution to the standard normal distribution. When we convert a score from a normal distribution to a z score we can then look up the z score in the standard normal distribution tables. This is given in Table A.1 of the Appendix. This information can be remarkably useful in statistical analysis.

The table tells us how many scores in the distribution are higher than the score we are examining. It does this by providing us with a figure for the area under the standard normal curve beyond the z scores, shown in Figure 3.3. The area underneath the whole curve is 1 (we have one whole area, like a whole cake before we cut it into portions) and the z score (like the knife cutting the cake) cuts it into two portions and the table tells us what proportion of the whole area we have cut off beyond the z score. If we subtract this value from 1 we know how much of the area is below the z score. Also, as the curve is symmetrical the mean value cuts the area into halves (so there is 0.5 of the area above the mean and 0.5 below).

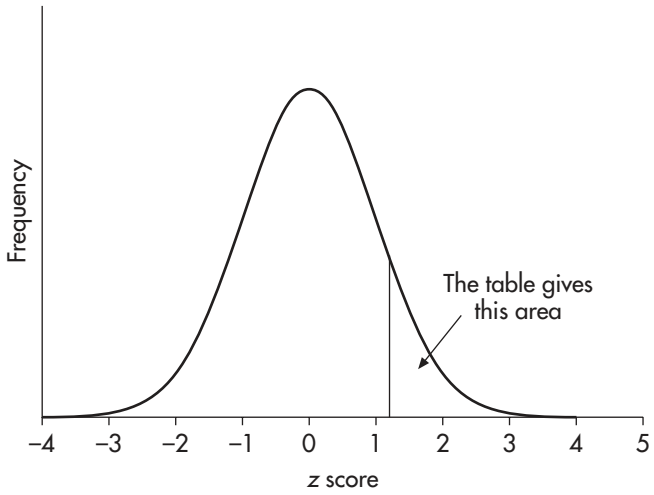

FIGURE 3.3 The Standard Normal Distribution

In this case, proportions are linked to probability. I am a 180 cm tall man. Let us assume, for the sake of this example, that the proportion of men taller than me in the population is a fifth. Now one fifth is 1 divided by 5, which equals 0.2, so we can express the proportion of men taller than me as 0.2 of the whole population. From this information I also know that the chance or probability of finding a man taller than me in the population is also one in five or 0.2. The area under the standard normal distribution curve is linked to probability in this way. The whole area under the curve (1) is linked to the probability of 1. Probability values range from 0 to 1. A probability of 1 is a certainty that something is the case. There is a certainty that any man I find will have a height somewhere on the men's height distribution so the whole area (1) is certain to include him. A probability of 0 is a certainty that something is not the case. The probability of finding a man twice my height (360 cm!) is so small as to be virtually zero. As we move from a probability of 0 to a probability of 1 we go from taking none of the area to taking larger and larger portions until we have the whole area.

When people talk about the chances of something happening they do not often talk in terms of probabilities ('the probability of me passing the examination is 0.5'), rather they prefer to use percentages ('I've a 50 per cent chance of passing the examination'). There is a simple relationship between probabilities and percentages, a percentage is a probability multiplied by 100. Thus, a probability of 0.3 is the same as a 30 per cent chance.

By looking at the area under the standard normal distribution curve above or below a z score we are able to obtain the probability of finding a score from the distribution larger or smaller than the score we have selected. In this way we are able to work out a whole range of interesting probabilities concerning scores from a normal distribution.

An example of using the standard normal distribution table

The distribution of scores in a Statistics examination is a normal distribution with a mean of 45 and a standard deviation of 4. You receive a mark of 49.

(a) What is the probability of someone scoring higher than you?
(b) What percentage of people are above the mean but lower than you?

As we have a normal distribution, the calculation of z scores will convert the distribution to the standard normal distribution. The score of 49 gives a z score as follows:

$$z = \frac{x - \mu}{\sigma} = \frac{49 - 45}{4} = 1$$

The standard normal distribution table (Table A.1 in the Appendix) will give us the probability of a score greater than a z score of 1. We look up the z score of 1.00 in the table and get a figure of 0.1587, so the probability of a score greater than 49 is 0.1587. (This means that you are in the top 16 per cent of the class, as $0.1587 \times 100 = 15.87$ per cent of the scores are better than yours.)

We know the area above the mean is 0.5 (half of the area) and the probability of a score greater than a z score of 1.00 is 0.1587, so if we subtract 0.1587 from 0.5 we will find the probability of a score above the mean and below your score: $0.5 - 0.1587 = 0.3413$. If we multiply this by 100 we will obtain the percentage: $0.3413 \times 100 = 34.13$ per cent. There are 34.13 per cent of the scores lower than your score but above the mean.

z scores of less than zero

If you calculate a z score and it turns out to be a minus number, all this means is that the score is less than the mean. As you can see from the

standard normal distribution table you cannot look up negative z scores. However, as we have seen, the normal distribution is symmetrical so the proportion of scores *greater than*, say +1.52, is the same as the proportion of scores *less than* −1.52. If you wish to look up a minus number in the table ignore the minus and look up the number. The figure you get from the table now tells you the probability of a score *less* than the z score. To find the proportion of scores greater than the z score subtract the table figure from 1. For example, if we calculated a z score of −1, this means the score is below the mean. We cannot look up −1 in the tables. We ignore the minus and look up 1 in the table. The probability value is 0.1587. This tells us that the probability of a score *lower* than a z score of −1 is 0.1587 and the probability of a z score *greater* than −1 is $1 - 0.1587 = 0.8413$.