1

Univariate Statistics 1: Summarizing Data with Histograms and Boxplots

Example: DNA Exonerations

Histograms

The Five-Number Summary

Summary of the Five-Number Summary and Boxplots

Conclusions

Exercises

Further Reading

The art of statistics is both about discerning patterns in data and about communicating information about these patterns to an audience. Statistics is an art, but that does not mean that anything goes. Like other artists you need to learn technical skills and guidelines in order for your art to be any good. To take an extreme example: go to GOOGLE and IMAGE and put in 'Jackson Pollock'. Jackson Pollock was considered one of America's best twentieth-century artists and was most well known for a brand of abstract expressionism where he appeared to drip paint in a chaotic and undisciplined manner over a

canvas. However, his technical abilities are clearly shown in his earlier paintings, and it was only with these skills that he could venture into an unexplored artistic genre. This book will not turn you into the Jackson Pollock of statistics, but it will help you to learn the basic tools of the trade and how to apply them. While painters, sculptors and poets have certain tools at their disposal, as a statistical artist you have various tools to facilitate both the discovery and the dissemination of your findings. Statistics is not just about what you can do with data; it is also about how you describe what you found to your expected audience. Therefore, your toolbox must include knowledge about your audience, as well as the more traditional tools like a pen and paper, and some computer software.[1]

This book introduces a language that allows us to talk about statistics, and science more generally. This is not a completely foreign language. Statistical phrases permeate our daily lives. Usually these are not the 'formal' statistics that appear in statistics books and in scientific reports, but they are embedded, very innocently, in our conversations. Examples include phrases like 'I will *probably* have a bagel today' and 'It takes *about* 20 minutes to cook rice'. The aims of this book are to enhance your awareness of these natural language statistics, to allow you to translate these into 'formal' statistics and, in so doing, to enable you to conduct, interpret and describe these statistics.

Consider the two examples mentioned above. Regardless of how likely you think it is that you will have a bagel today, you know roughly what the above statement means. When we use words like 'probably' we are not usually worried about the precise meaning of the phrase. Translating from natural language to formal statistics often involves becoming more precise. Here we might say that the *probability* of having a bagel is more than 0.50 or 50%. Probability is at the heart of statistics and will be described throughout this book. If you had a standard deck of 52 cards, shuffled them thoroughly and were about to draw one card, the probability of it being red is 0.50. So using this analogy, the above statement means that it is more likely that you will have a bagel than randomly choosing a red card from a well-shuffled deck of cards.

The second statement, 'It takes *about* 20 minutes to cook rice', is a statistical phrase because of the word 'about'. Depending on the amount and type of rice, the initial heat of the water, the type of stove and even the altitude at which you are cooking, the amount of time it takes to cook rice is not constant, but varies. Translating this into statistics it becomes 'Twenty minutes is the *central tendency* for the time to cook rice, but the exact time may vary from this'. 'Central tendency' is what the statisticians would call the instructions written on the side of the rice box suggesting how long to cook the rice. It is the value that, across all situations, the rice manufacturers think is the best guess for proper cooking time. There are different and more precise ways of calculating the central tendency including the median, which is discussed in this chapter, and the mean, which is discussed in Chapter 2.

---

1 This book is not tied to any specific statistics software. The accompanying web page provides examples from two of the main packages (mainly SPSS and R). The web also includes the data sets used in this book and much-other useful information.

For most of you, the main concern with regards to statistics is not to help you to become a better rice chef, but how statistics are used and reported in the social and behavioural sciences. The point of these examples is to show how frequently statistics are encountered in our lives. During the course of your studies you will come across other 'everyday statistics' and also more formal statistics. This book describes various procedures for creating these statistics.

EXAMPLE: DNA EXONERATIONS

Imagine you are walking home one evening. You can hear police sirens in the background, but you don't think much of them. A police officer approaches and asks you a few questions. A woman has been raped and the police are looking for her attacker. You say you were at a friend's house and have been walking home. The police officer takes your name and contact details, and you go home. The next day another officer arrives at your home, and tells you that you match a rough description that the victim gave of the culprit. They ask you if you will take part in an identification parade. You agree, after all, you're not guilty; the victim won't choose you. Perhaps you would be less calm if you knew what the US Attorney General, Janet Reno, said in the preface to a report about eyewitness accuracy: 'Even the most honest and objective people can make mistakes in recalling and interpreting a witnessed event' (Technical Working Group for Eyewitness Evidence, 1999: iii). The victim identifies you as her assailant, and because jurors trust eyewitness testimony (a lot more than they should), you are convicted and spend years in prison. You may not feel lucky, but in one way you are. The crime that you were falsely convicted of is one that often includes a biological marker, semen. A DNA test is done, which shows that you are not the culprit, and, after some further legal arguments, you are eventually exonerated and released.

Your case is a tragedy of justice, but you are not alone. The Innocence Project in the US reports hundreds of people who have been falsely convicted but later exonerated based on DNA evidence (www.innocenceproject.com). We will look at the first 163 which we downloaded on 17 November 2005. Each of these individuals' cases is a tragedy, and it is important that when you report your statistics you do not lose sight of the meaning of each case. Each individual spent years in prison, falsely accused. As voiced by Uncle Tupelo: 'Handcuffs hurt worse when you've done nothing wrong' ('*Grindstone*' by Farrar and Tweedy).

The length of time in prison of these 163 people (the data file, dnayears.sav, is on this book's website) will be used to illustrate some of the basic statistical concepts and graphs.

Each of the individuals in the DNA file is a *case*. The *sample* is composed of the 163 cases. The larger *population* in this example would be all falsely convicted individuals exonerated by DNA evidence. There is information about several attributes for each of the

Table 1.1 *The DNA cases from the Innocence Project*

$Caseno_i$	$firstn_i$	$lastn_i$	$state_i$	$year1_i$	$year2_i$	$time_i$
1	Gary	Dotson	Illinois	1979	1989	10
2	David	Vasquez	Virginia	1985	1989	4
3	Edward	Green	DC	1989	1990	1
⋮	⋮	⋮	⋮	⋮	⋮	⋮
162	Leo	Waters	North Carolina	1981	2005	24
163	George	Rodriquez	Texas	1987	2005	18

cases. Each of these attributes is called a *variable*. For this example there are seven variables: the case number, the person's first and last name, the state where they were convicted, the year they were convicted, the year they were released, and the time between conviction and release. Each person has a *value* for each variable, thus for the first person, Gary Dotson, the value for state is 'Illinois' and for time is 10 years. Most of the values that are used in this book are numeric, but the values can also be words, pictures, etc. The way that we will refer to variables is by giving them a name that describes them, writing them in *italics*, and including a subscript which tells us that people may have different values for this attribute. So, the variables $state_i$ and $time_i$ refer to the variables denoting the state in which the person was convicted and the time they spent in prison. The subscript *i* shows that there are different values for these variables, the *i* referring to different people in the sample. If you are referring to the first person the subscript 1 is used. Thus, $state_1$= 'Illinois' and $time_1$=10 years. For numeric values it is important to include the units of measurement so that it is clear that Gary Dotson spent 10 years in prison, rather than, say, 10 months in prison.

The values for all the people in the sample, when placed together, form a *data set*. Most of the common statistical packages hold the data set in a spreadsheet format, like Table 1.1. Each row represents a single individual. The '⋮' means that the values for cases 4 to 161 are not included. It is a big data set, so would take up a lot of room to print and would be difficult to get a summary feeling for the data. This is one of the purposes of statistics, to identify useful summary information and to describe this to others.

One of the major objectives of statistics is to accurately summarize large quantities of data so that the reader can understand the overall patterns of responses. Two main types of techniques for summarizing data will be described in this chapter. The first technique is a histogram. Several variations are discussed. First a dot histogram and a stem-and-leaf diagram are shown. Then we present a generic histogram. The second name histogram technique is based on the *Five-point summary* and is called a box-and-whiskers plot (or just boxplot). Both of these methods are appropriate for describing *quantitative data* (where the variable itself is on a numerical scale, such as number of years imprisoned, or score on a measure of anxiety symptoms). Methods for describing *qualitative data* (data that describe category membership such as being in the Republican Party versus the US Democratic Party or as having cats versus dogs) are described in Chapter 3.

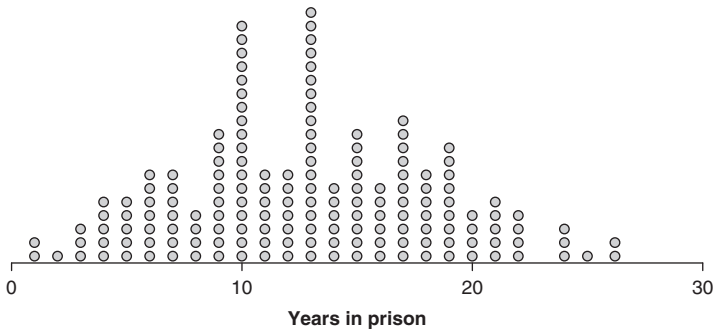

Figure 1.1 A dot histogram of the amount of time spent in prison for the 163 people from the DNA data file

HISTOGRAMS

We go through a series of histograms that vary in how much information is embedded within the histogram. The first type is a dot histogram. Here each individual is shown with a dot. The stem-and-leaf diagram is the next type. Here, numeric information is included. While this type is much lauded by statisticians, it is not as popular as the final histogram we present, the generic histogram. The generic histogram, or just histogram, is the most commonly used type. *Finally, we present a name histogram as an example of an extension to the generic histogram.*

The Dot Histogram

The macro information is usually what you are trying to stress, so we will consider other graphical devices that focus the readers' attention on this information before returning to a graph which includes both types of information, but less of each. Each Observation is manced with a dot on the histogram. This can be done in a word processing package as the name histogram was, or it can be done in some statistics packages. This means that more precise numeric information is provided.

Stem-and-Leaf Diagram

The dots in Figure 1.1 each represent a person. Their placement shows how long that person spent in prison, but the dot itself provides no other information. A stem-and-leaf diagram (Tukey, 1972, 1977) allows the individual 'dot' to provide information. A stem-and-leaf for the DNA data is shown in Figure 1.2. The variable is divided into two-year bins. The numbers on the far left are the number of people in each bin. The next number is the first digit. Each

Freq.	Stem	Leaf
2	0	**1**1
4	0	2333
10	0	4444455555
14	0	66666667777777
14	0	88889999999999
25	1	0000000000000000001111111
26	1	22222223333333333333333333
16	1	4444445555555555
17	1	6666667777777777777
16	1	8888888999999999
9	2	000011111
4	2	2222
4	2	4445
2	2	6**6**

Figure 1.2. A stem-and-leaf diagram of the amount of time spent in prison for the 163 people from the DNA data file. The values in **bold** are the minimum, maximum, the median and the two hinges, which are the five numbers of the five-number summary and described in the next section

digit on the right stands for an individual person, and gives the value for that person. Thus, there are two people who spent one year in prison, one who spent two years in prison, and three who spent three years in prison.

Tukey (1977) goes into much detail about how to make these plots, how to use them to help check the data, and what can be added to them. Until recently they had been used rather sparingly, but have become more popular because they are often used when reporting *meta-analyses*. These are studies which combine the results of different studies. A particular statistical result from each study would be represented by each digit.

In Figure 1.2 multiple stems are the same because the bins are only two years wide. In many stem-and-leaf diagrams each stem is unique. For example, consider the following data from Wright and Osborne (2005) on 80 people's scores on a dissociation measure. Dissociation, which means having difficulty integrating mental images, thoughts, emotions and memories into consciousness (the word 'spaciness' is sometimes used, but this does not capture the full meaning of the term), has scores which can vary from 0 to 100. The stem-and-leaf diagram, as printed by SPSS, is shown in Figure 1.3. It shows that there were four people with scores less than 10 (scores of 3, 7, 9 and 9), a couple of scores of 60 or above, but that most of the scores are between 20 and 60.

The Generic Histogram

The two types of histogram shown above are alternatives to the generic histogram. If you squint looking at any of these three, this is what a generic histogram is. It focuses solely

Freq.	Stem	Leaf
4	0 .	3799
4	1 .	0267
18	2 .	113455556688888899
20	3 .	01112222223357888889
18	4 .	011113556667788889
14	5 .	00002223346779
2	6 .	04

Figure 1.3 A stem-and-leaf diagram of the amount of self-reported dissociation (0–100 scale) for 80 participants (from Wright & Osborne, 2005)

on the macro information. The DNA data, with two-year bins, is shown in Figure 1.4. If doing this by hand, the *x*-axis scale is made in the same way as with the other histograms. You need to calculate the number of people in each bin (which can be done with the 'Frequencies' command in many statistics packages). You draw the *y*-axis from 0 to above the maximum number in any bin. You then draw a horizontal line for each bin corresponding to the number of people in those bins. Use this to make a rectangular box. Make sure that you label the axes properly. All of the main statistics packages allow you to make

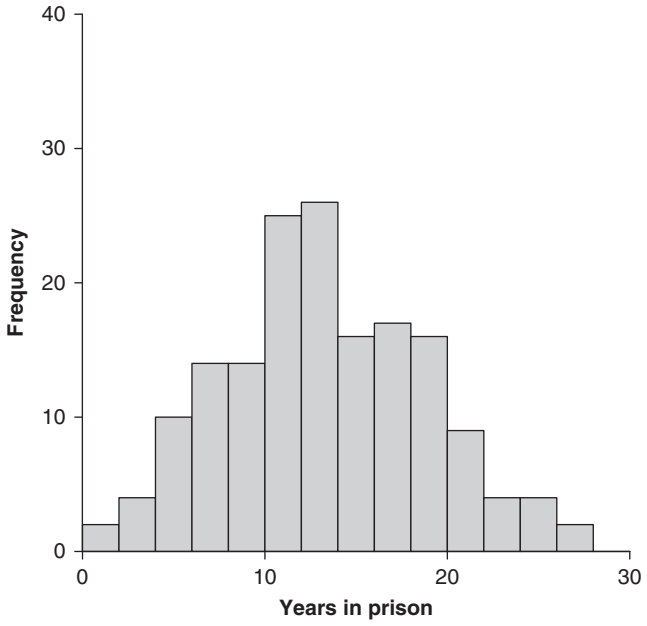

Figure 1.4 A histogram of the amount of time spent in prison for the 163 people from the DNA data file

histograms. The most common mistake is using a 'bar chart' option instead of the histogram option. Bar charts are discussed in Chapter 3. While some of them look like histograms, they are a different graphical technique (Wilkinson, 2005). Notice with the histogram, the bars are touching one another, which denotes that the data are quantitative.

Deciding about Bins

When making a histogram the critical decision is the size of the bins. Many software programs set the number of bins by default, according to the number of cases and variety of their responses. Often, the program's default settings are fine but if needed you can adjust their selected number of bins.

As you increase the number of bins, you increase the amount of numeric information, but sometimes this is providing too much information and it breaks Grice's maxim of quantity. Figures 1.5 (a–c) show the same data as Figure 1.4 but with bins of one year, four years or eight years. Figure 1.5 (a) probably provides too much information. Readers may concentrate on the peaks at 10 and 13 years, and the dips at 11 and 12 years, which probably are not important. Figure 1.5 (c) provides too little information. The reader would probably want to have more precision. Figure 1.5 (b) provides about the right amount of information for most readers. Either this or Figure 1.4 (bin width of two years) is probably the best.[2]

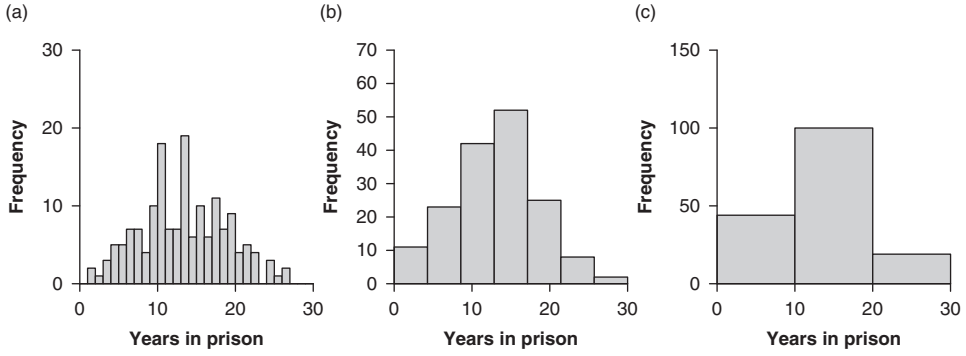

Figure 1.5 (a–c) Histograms for the DNA data cases with bin widths of (a) one year, (b) four years and (c) eight years

2 There are statistical procedures that provide a guide for how wide the bins should be. Wand (1997) describes several of these, and different computer packages use different algorithms. Some of these are complicated and all require some statistical concepts that we have yet to encounter. For the present purpose it is just worth knowing that these procedures suggest widths of between two and four years.

A Name Histogram

A phrase often attributed to, but probably never said by, Stalin is: 'A single death is a tragedy, a million deaths is a statistic' (http://www.time.com/time/personoftheyear/archive/photohistory/stalin.html; 25 October 2005). There is a distrust of statistics because numbers can appear to deny the humanity deserving of each individual represented by those numbers. Turning data, like that in Table 1.1, into a graph, can sometimes remove the humanity from the individual datum (datum is the singular of data). The final histogram we show helps to highlight the importance of each of the individual data points. We call it a *name histogram* and it is based on Tufte's micro/macro distinction (1990, see pp. 140–1). This is like the dot histogram shown in Figure 1.1, but instead of printing dots we print the person's name. We made this in Word (details on the book's web page). We show it to emphasise the types of graphs that you can produce.

Summary of Histograms

Several types of histogram were shown. They vary in how much information they provide. It is important to think about how much information you wish to provide to your audience. The dot histogram shows each case as a dot, but provides no further information. It is useful because it makes clear the number of people in each bin. The stem-and-leaf diagram also provides this, but gives extra information about the numeric value. Stem-and-leaf diagrams are liked by methodologists, and appear in scientific journals, but not in the popular press. Hopefully this will increase. The generic histogram, which does show up in the popular press, focuses only on the macro level. The name histogram provides lots of individual information and for examples like the DNA cases reinforces to your audience the gravity of each datum. It is also useful with smaller data sets, but providing the names for all 163 cases is too much for most purposes.

The information provided in all histograms is about the distribution of the variable, for most of the graphs in this chapter this is the variable $time_i$. The distribution of a variable is an important aspect of the variable. These graphs show that there were a few people who just spent a couple of years in prison for crimes that they did not commit, and a few who spent more than 20 years in prison, but that most spent between five and 20 years in prison. They also show us if any particular values are more prevalent than others. The histograms also reveal the shape of the distribution. Many distributions have a shape like that shown in these figures, where most of the values are in the centre and the figure looks kind of 'bell shaped'. One particular shape that has these characteristics is called the Normal distribution. We will discuss this in more detail in Chapter 5.

Figure 1.6 A name histogram for the data in Table 1.1

THE FIVE-NUMBER SUMMARY

The five-number summary shows five numbers that are used in a popular graphical technique called a boxplot (or box and whiskers). The five numbers are the minimum, maximum, the middle number (called the median), the number that splits the lowest 25% of the sample from the highest 75% of the sample (the first quartile), and the number that splits the lowest 75% of the sample from the highest 25% of the sample (the third quartile). The focus in this section will be on what these values mean and how to use them in graphs. Computing them is straightforward for certain sample sizes, but complex for most sample sizes. Nowadays computers are usually used to calculate them.

Statistics is about reducing the vast amount of data into a smaller amount of information. While the histograms allow important aspects of the data to be understood, they are not reducing the data. The techniques described in this section will allow to you calculate five important values that can be used to summarize an entire distribution. The five numbers of the five-number summary are: the minimum and the maximum, the median, and the two quartiles (sometimes called 'hinges').[3]

The *minimum* is the lowest number. Suppose you went to a playground and wanted to see if it was age-appropriate for a young relative and that there were nine children of the following ages (and some adults):

Values:	2	4	8	3	8	7	2	2	12

First, we sort these from smallest to largest:

Values:	2	2	2	3	4	7	8	8	12
Ranks:	1	2	3	4	5	6	7	8	9

↑	↑	↑	↑	↑
Minimum	First quartile	Median	Third quartile	Maximum

Each value has a rank. Because some children are the same ages (at least when their ages are just in years), we have just assigned the lowest number (2 years old) with a rank of 1, the second number (2) with a rank of 2, and the third number (2) with a

3 The definitions for certain statistics evolve. What this chapter is guided in large part by Tukey's discussion, his term 'hinge' is slightly different from 'quartile'. Nowadays most people use these terms interchangeably.

rank of 3, and similarly with the two cases with the value 8. The minimum value is the one with a rank of 1. This is the value 2 years. The *maximum* is the largest number and has a rank equal to the total number of cases. The total number of cases is often denoted with the letter n, so in this example $n = 9$. The case with rank = 9 has a value of 12 years old so this is the maximum. The *median* is the middle value. With $n = 9$, the median is the value with rank 5. Here the value is 4 years. When n is odd, the rank of the median is $(n + 1)/2$, here $(9 + 1)/2 = 10/2 = 5$. Notice this is the ranking of the value i.e., the fifth. Value, which in this example equals 4 years old. This is observed such that half the people score above it and half score below it. The situation when n is even is discussed below. The quartiles are the 'medians' of the lower and upper halves of the data. They separate out the upper and lowest 25% of the values. The word 'quartile' is based on 'quarter'; the quartiles divide the data into quarters. The five-point summary can be written as:

$n = 9$

	4	Units = years
2	8	IQR = 6
2	12	Range = 10

Within the box are the five numbers. The median is shown and then below it the quartiles and then below them the extreme values (the minimum and maximum). Some additional information is often included with the five-number summary. The number of cases is shown above the box. To the right of the box are three additional pieces of information. First, the units of the variables, here years of age, is printed. Below this the interquartile range (IQR) is shown.[4] This is the difference in values between the upper and lower quartiles and can be calculated from the information in the box ($8 – 2 = 6$). Below this is the range, the difference between the maximum and minimum values ($12 – 2 = 10$). The IQR and the range will always be positive values.

When you have larger datasets it can be easier to calculate the five-number summary using a stem-and-leaf diagram. Figure 1.7 redraws the stem-and-leaf diagram from Figure 1.3, but marking off the quartiles (technically, the median is a quartile since it takes all three points to divide the variable into quarters). Each quarter has 40 cases. When presented like this it is easy to see the five numbers. The five-number summary is shown in Figure 1.7. The quartiles and the IQR are important concepts; they show that approximately half of the cases lie between nine and 17.

4 You may come across the *semi-interquartile* range, which is the IQR divided by 2.

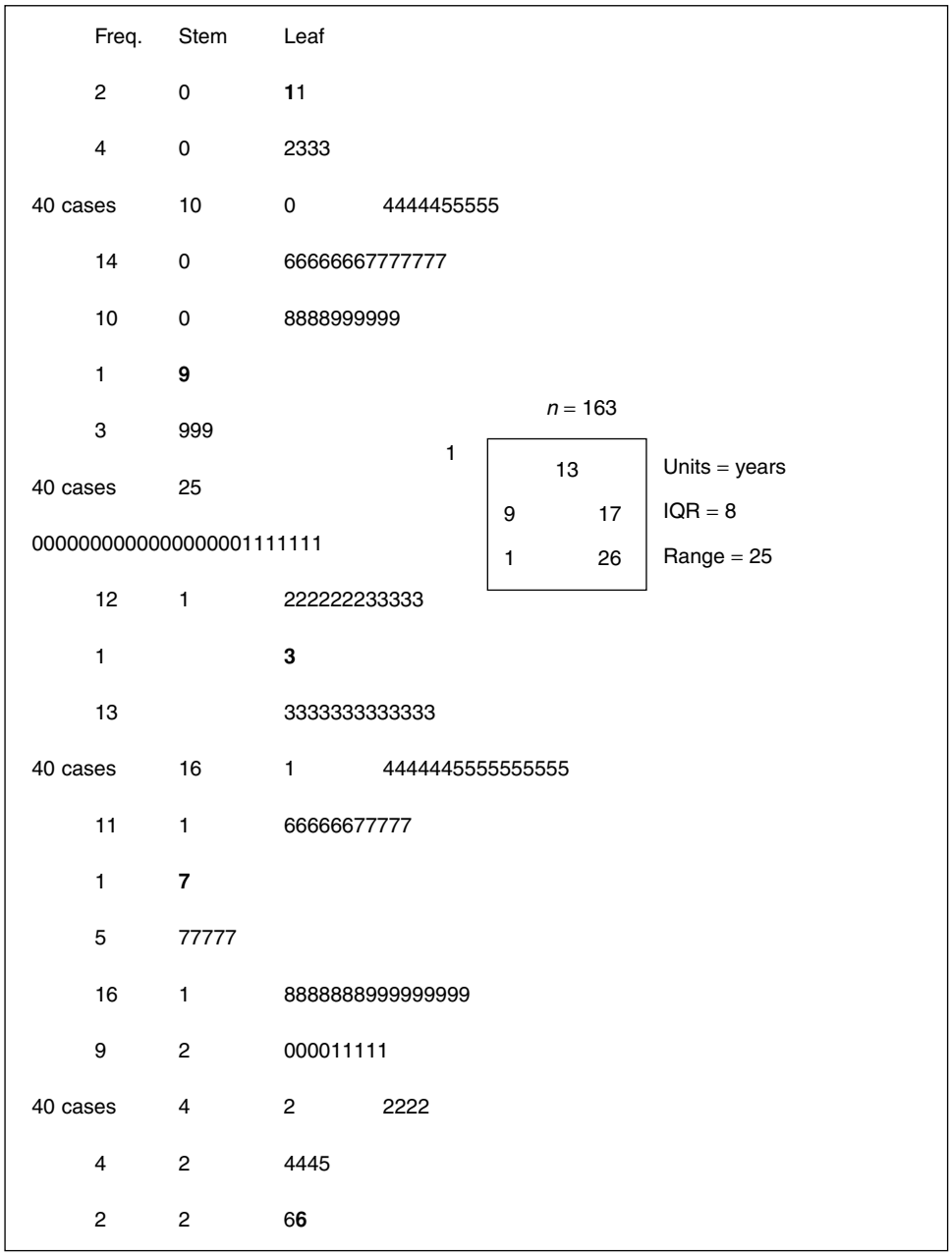

Figure 1.7 A stem-and-leaf diagram, split by quartiles, and a five-point summary box for the DNA data

Having 163 cases here makes calculating the five-number summary relatively simple because these can be divided into four equally sized groups of 40 plus the median and two quartiles. When there is an even number of cases the median is more difficult to calculate. Consider calculations of the median. The formula, $(n + 1)/2$, produces non-whole numbers. If there are 10 values, the median rank is 5.5. To calculate the median you need to take the average of the 5th ranked case and the 6th ranked case. If the data are:

Values:	1	3	3	5	7	9	9	11	13	17
Ranks:	1	2	3	4	5	6	7	8	9	10

↑
Median

The 5th case has the value 7 and the 6th the value 9. Thus, the median is equal to halfway between these, or 8. This is called the mid-rank of the two numbers. In Chapter 2 you will learn about a statistic called the 'mean'. It is also the mean of these two numbers. Half the values fall above eight-years-old and half fall below eight-years-old.

Similarly, finding the quartiles is more difficult and there are some subtle differences in how some packages and books define quartiles and IQR. Luckily, statistics computer programs calculate these values for you, which leaves us to focus on the concepts. The concepts are: the first quartile separates the lowest 25% of the data from the rest, the third quartile separates the highest 25% of the data from the rest, and the IQR includes the middle 50% of the data. The rank for the first quartile is $n/4$ and for the third quartile is $3n/4$, both of which will usually be a non-whole numbers. For ease when doing these by hand, you should round the first quartile up to the next highest rank, and round the third quartile down to the next lowest rank. For example, with $n = 10$, the rank of first quartile is $10/4 = 2.5$ so the 3rd value (3) should be used, and the rank of the third quartile is $3\times10/4 = 7.5$, so the 7th value (9) should be used.

While quartiles divide the values into four equally sized groups, for different purposes you may wish to divide the values into different numbers of groups. Quartiles are one of the most useful. Another popular method is to divide the values into 100 equally sized groups. These are called *percentiles* (dividing the data into percentages). You may hear parents saying their child is in the 98th percentile on some standardized test. This means only 1–2% of children who take this test do better.

Box-and-Whiskers Plots, or Boxplots

Figure 1.8 shows Ruby. On one of our web pages we described how this cunning feline devised the box-and-whiskers plot (as seemed obvious from the photo), not the great statistician John Tukey. Many people wrote who appeared not have grasped *sarcasm*.

Figure 1.8. Resolving the debate about the origin of the box-and-whiskers plot: Ruby versus Tukey

While there were some precursors (for example, Spear, 1952, as cited in Tufte, 2001), let us use these pages to set the record straight. Ruby is a loveable cat, but is insignificant in the history of statistics. John Tukey (1977) is generally credited as the creator of box-and-whiskers plots. In fact, aspects of Tukey's original description of the box-and-whiskers plot and a plot which he called a schematic plot (Tukey, 1972: Fig. 18.8) are often incorporated together, and go under the more general term *boxplots* (McGill et al., 1978). Therefore we will use this term.

The boxplot is a graph of the five-number summary. We will describe two versions, one created by the computer (an actual boxplot) and a simplified version that can be drawn easily by hand, called a quartile plot (Tufte, 2001). Details for drawing a boxplot are shown in Box 1.2. Many of the computer packages use slightly different ways of making boxplots (Reese, 2005), but they all produce the same basic diagrams. A rectangular box is drawn from the lower quartile to the upper quartile and a vertical line is placed within the box to denote the median. The possible length of the whiskers are defined as 1.5 x IQR. For the DNA data the IQR is 8 so the maximum whisker length is 12. The whiskers *could* go from the lower quartile, 9, to –3 and from the upper quartile, 17, to 29. But, the whiskers stop at the most extreme observed value that is within the possible length. Here it stops at 1 and 26. The points where these whiskers end are called *adjacent* points. A boxplot for these data is shown in Figure 1.9.

Tufte (2001) is the most influential person on creating good graphs. He described how the basic boxplot was difficult to draw by hand and also used more ink than is necessary (he stresses that good graphs should use as little ink as possible). He pointed out that a boxplot could be drawn without the box. He describes several alternatives. The one we like

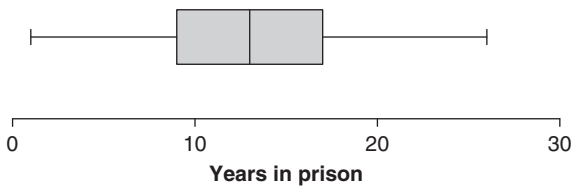

Figure 1.9 A boxplot for the DNA data. This shows the five-number summary in graphical form

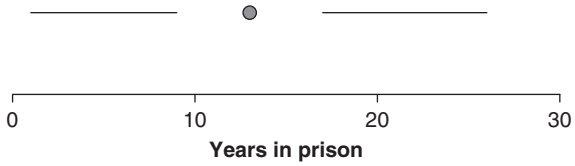

Figure 1.10 A simplified boxplot called a quartile plot (Tufte, 2001). When drawing by hand this is easier to draw than a standard boxplot

best replaces the median with a dot, leaves out the box, but keeps the whiskers. Tufte calls this a 'quartile plot'. An example is shown in Figure 1.10.

When there are values outside the whiskers they should be shown. Sometimes it is useful to distinguish cases just outside the whiskers from those far outside the whiskers. Tukey (1977) gives various rules for doing this, calling points more than 1.5 IQR from the median as *outside* points and those 3.0 IQR from the median as *far outside* points. Suppose the following were annual salaries in $1000 for 25 people:

Values:	3	12	12	15	15	15	15	15	18	21	22	22	22	22	22	27	28	30	30	30	32	33	38	75	150
Ranks:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25

The five-number summary is:

[$n = 25$]

	22	Units = $1000
15	30	IQR = 15
3	150	Range = 147

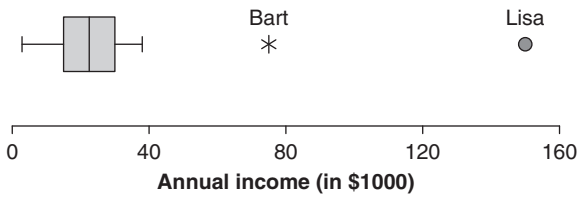

Figure 1.11 A boxplot of annual income (in $1000) for 25 people. The outside point (Bart) and far outside point (Lisa) are both labelled. These are usually referred to as *outliers*

The resulting boxplot is shown in Figure 1.11. The possible whisker length is $1.5 \times$ IQR. The lower adjacent point is $3000 and the upper adjacent point is $38,000, so this is where the whiskers end. The value $75,000 is more than 1.5 IQR from the median so it is an outside point, and is labelled with a star. The value $150,000 is more than 3.0 IQR from the median so it is a far outside point. It is denoted with a circle. Particularly with small data sets it is often useful to label these points. This allows some of the micro information in Figure 1.1 to be included. People are likely to be most interested in these extreme points so this provides micro information where it is most useful.

Box 1.2 Making a Boxplot by Hand

There are some slight variations that can be used when constructing a boxplot. Here are the rules for a fairly simple version. Before beginning this you should have sorted all the data in order from lowest to highest, found the median, the upper and lower quartiles, and the IQR.

1 On graph paper, make the horizontal axis so that the values cover the range of values in the same way as was done for histograms in Chapter 2.
2 Draw short vertical lines, above the axis, to denote the medians for the different groups (or the median if you have a single group).
3 Draw a short vertical line for each quartile, and join these as a box (see Figure 1.9).
4 Calculate the values for the whiskers as follows:

Lower whisker = lower quartile 1.5 IQR
Higher whisker = higher quartile 1.5 IQR

Draw these as lines extending from the box to the furthest data point (called the adjacent point) that is within this 'inner fence'.
5 Denote any data points outside these whiskers, but within 3.0 IQR of the median, with some character (a star in Figure 1.11), and those beyond this point with a different character (a circle in Figure 1.11).

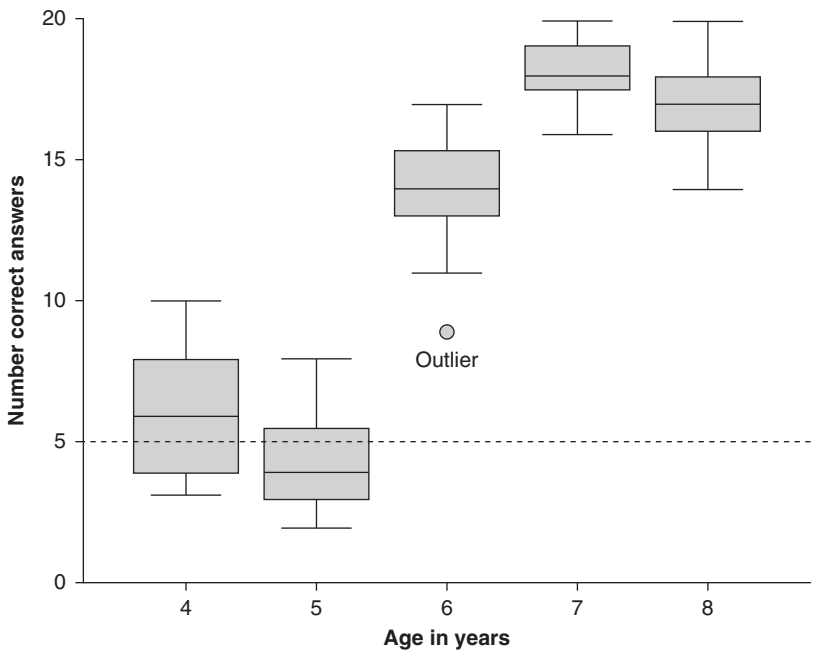

Figure 1.12 Boxplots for the number of correct answers on a test by the participants' age. The dashed line shows the prediction for random guessing

Boxplots are most useful when comparing groups. Suppose a developmental psychologist asked 20 children from each of five different age groups to solve some mathematics problems. Suppose there were 20 multiple-choice questions and each question had four options. The data might look like that depicted in Figure 1.12. We have changed this to vertical boxplots; most packages give you the choice of having them horizontal or vertical. If people were guessing randomly, they would probably get somewhere near five correct answers because the probability of correctly guessing for any individual question is 25% (this level is shown with a dashed line in the figure). Most four- and five-year-olds are at about this level, so clearly have not mastered whichever mathematical techniques are involved. The seven- and eight-year-olds get most of the questions right. They are at the *ceiling*, meaning they cannot do much better. The scores on these mathematics problems differentiate among the six-year-old children. Some do poorly, some do well, and most are in the middle. If an educational psychologist wanted a test to discriminate among children, these problems would be most useful with the six-year-olds. Figure 1.13 shows this graph using the quartile plots (Tufte, 2001). When there are several groups this is often useful. Further, we have added a dashed line between all the medians. This is appropriate if the grouping variable is based on some type of scale, like age.

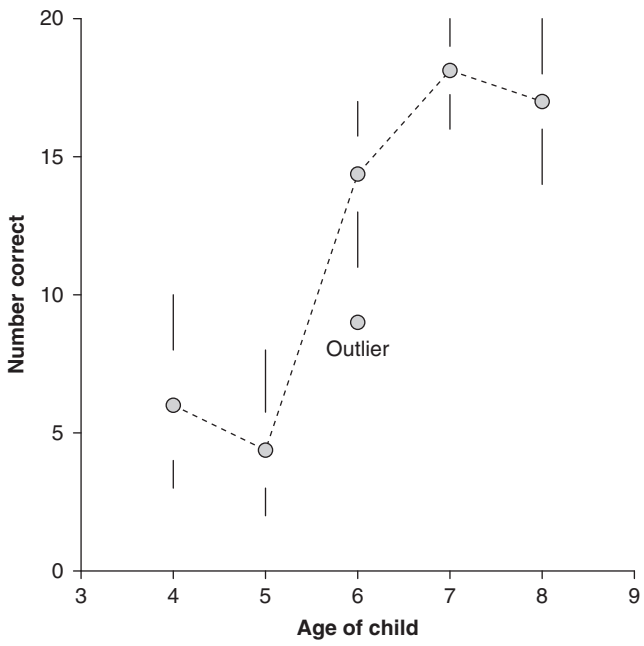

Figure 1.13 Quartile plots, as developed by Tufte (2001), showing the same data as Figure 1.12

SUMMARY OF THE FIVE-NUMBER SUMMARY AND BOXPLOTS

Histograms provide much information about the distribution of a data set. Often it is useful to summarize this information in a few numbers, in order to give the reader more detail about the distribution. The five-number summary is a good set of numbers for this. They summarize the data well and are easy to explain to other people. Most people know what the minimum and maximum are (the lowest and highest numbers). The median is the number with half of the people scored above this value and half of the people scored below it. The lower quartile has 25% of the values lower and the upper quartile has 25% of the values higher. The area between the lower and upper quartiles contains 50% of the data.

The *boxplot* refers to a collection of techniques based on the five-number summary. Most statistics packages produce them. The box shows where about 50% of the cases are. The whiskers are a length which Tukey chose that should include most of the remaining data. Those outside the whiskers are outliers. Tukey distinguished outside points from far outside points, though often people just label all of these as outliers. Reese (2005) notes that boxplots frequently occur in scientific publications, because they accurately and clearly display important characteristics about the distribution and allow different groups of people to be easily compared, but at present they are not included in many popular

magazines and newspapers. He argues that this should change because the five-number summary can be easily comprehended by most people.

CONCLUSIONS

Histograms and boxplots both convey information about single variables. They are univariate (uni, from the Latin *unus* meaning one) procedures. How much of the information that is displayed in a histogram depends on the type of histogram and the bin size. Procedures like the stem-and-leaf diagram display more precise information than the generic histogram. As the width of the bin increases, information is lost. With boxplots most of the information is lost. It is assumed that the key points of the distribution can be summarized with a small set of numbers and including information on only a few outlying cases. It is possible to display histograms and boxplots together, as in Figure 1.14. This shows the distribution of scores on a variable that measures dissociation on a 0–100 scale and it shows that most people have low scores, but a few have high scores.

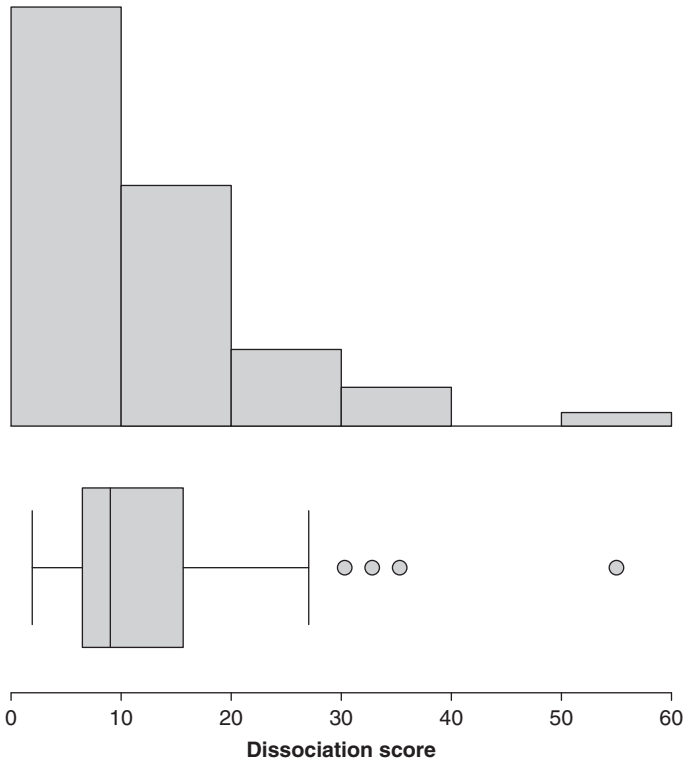

Figure 1.14 A histogram and a boxplot for responses to DES II (Wright & Loftus, 1999), a questionnaire measuring dissociative experiences

EXERCISES

1.1 How is Grice's (1975) maxim of quantity relevant to graphing data? (*Hint*: see discussion of histograms.)

1.2 Why is the number of bins, or the bin width, an important decision when constructing a histogram?

1.3 Find an 'everyday' statistical phrase from a newspaper. Say what the phrase means and why it is a statistical phrase.

1.4 Describe what the following four terms mean: variables, values, sample and cases. How can you write the value of a variable for a single case?

1.5 Four types of histograms were introduced. They were: dot plots, stem-and-leaf diagrams, generic histograms and name histograms. When would each be used? What additional information is included in the name histogram that is not in the dot histogram? Which procedure would be best if you had only 10 cases? Which if you had 2000 cases?

1.6 What are the five numbers in the five-number summary and what does each mean? What other information is usually printed with the five-number summary?

1.7 If the distribution provides more information than the five-number summary, why is five-number summary of value?

1.8 When might you prefer to use boxplots as opposed to histograms?

1.9 In the table below there is a restaurant bill for 10 people. They decided to split the bill evenly for their dinners.

(a) Make a histogram of these data. Say which type of histogram you made.
(b) If you were asked how much it costs for a typical meal at this restaurant, what amount would you give? What is the statistical term for it called?

Meal prices at DK's Curry Palace (prices are in UK pounds; (£1 ≈ $2, where '≈' means approximately)

Name	Price
Louise	£4.50
Steve	£13.00
Alice	£5.50
Susan	£4.50
Dave	£6.00
Joanne	£5.50
Mel	£5.00
Tom	£5.00
Andy	£5.50
Alexa	£6.00

(Continued)

(Continued)

1.10 If you had data on 18,000 people's scores on a life quality measure, which type of histogram would you use?

1.11 The following exam marks, out of 100, were awarded to 23 students:

54 65 63 75 81 32 0 69 48 38 19 68 55 67 70 72 76 0 74 47 61 65 88

(a) Find the five-number summary.
(b) Make a boxplot. Say which type you have made.

FURTHER READING

Brief online works:

Gould, S.J. (1985). The median isn't the message. *Discover*, *6* (June), 40–2.

Available on several websites including: http://cancerguide.org/median_not_msg.html. All the praise given to this essay is well deserved.

Tufte, E. (2003). PowerPoint is evil. *Wired*, 11.09, http://www.wired.com/wired/archive/11.09/ppt2.html.

The title really says it all. What would happen if Microsoft could afford lawyers? An updated version is in his 2006 book *Beautiful Evidence*. See also http://www.edwardtufte.com/bboard/q-and-a-fetch-msg?msg_id=0001yB&topic_id=1.

wikipedia has pages on many of the concepts, like histogram, boxplot, five-number summary, median and quartile. Although *wikipedia* gets the occasional bad press, it is usually accurate and a good source of information. Most other web resources are less reliable. In general, to learn about academic subjects it is best to use academic websites.

Books

There are several good books about graphing. We decided to be very selective and choose two.

Tufte, E.R. (2001). *The Visual Display of Quantitative Information* (2nd ed.). Cheshire, CT: Graphics Press.

A truly marvellous book! See http://www.edwardtufte.com/tufte/ for more on Tufte.

Tukey, J.W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

This is a 'how-to' book which allowed serious scientists to talk about graphics. Tukey was one of the top twentieth-century scientists. Although there is an emphasis on constructing graphs by hand, which thanks to computers is no longer necessary, the logic and fluency of this text is still excellent. Unfortunately it is out of print, but most university libraries will have copies.