STATISTICS WITHOUT MATHEMATICS

retweet this

Click here to post this on Twitter

SAGE

Los Angeles | London | New Delhi
Singapore | Washington DC

SAGE Publications Ltd
1 Oliver's Yard
55 City Road
London EC1Y 1SP

SAGE Publications Inc.
2455 Teller Road
Thousand Oaks, California 91320

SAGE Publications India Pvt Ltd
B 1/I 1 Mohan Cooperative Industrial Area
Mathura Road
New Delhi 110 044

SAGE Publications Asia-Pacific Pte Ltd
3 Church Street
#10-04 Samsung Hub
Singapore 049483

Editor: Jai Seaman
Assistant editor: James Piper
Production editor: Victoria Nicholas
Copyeditor: Richard Leigh
Proofreader: Andy Baxter
Indexer: David Rudeforth
Marketing manager: Sally Ransom
Cover design: Shaun Mercier
Typeset by: C&M Digitals (P) Ltd, Chennai, India
Printed in India at Replika Press Pvt Ltd

FSC
www.fsc.org
MIX
Paper from
responsible sources
FSC® C016779

Library of Congress Control Number: 2015931451

British Library Cataloguing in Publication data

A catalogue record for this book is available from
the British Library

ISBN 978-1-4739-0244-2
ISBN 978-1-4739-0245-9 (pbk)

1

PICTURING VARIABILITY

Summary

Variation is a central idea in Statistics. It may be represented pictorially as a histogram or frequency curve or by numbers in a table. This chapter is designed to introduce the reader to patterns of variation which occur in the world around us and, in particular, to their pictorial representation. The ideas are illustrated by a range of examples, including sentence lengths and spam messages.

Introduction

There is one key idea which underlies most of Statistics. This is the idea of *variation*. If we make measurements of almost anything there will be variation in the results. Any doubts which the reader may have on this score should have been dispelled by the end of the next chapter. Understanding this simple fact and learning how to describe it is half the battle. Variation can be represented in several ways, but here we shall rely entirely on diagrams or pictures. Once we learn how to interpret the pattern of variation revealed by a picture we shall be able to understand the ideas which lie behind most statistical procedures.

Before proceeding, let us pause to elaborate on the key idea a little more, because the claim that *variation* is the central thing may seem radical and surprising to the unprepared reader. In many ways our whole culture, especially as seen through the media, ignores variation. What matters for them is the typical or 'average' or 'representative' case. They are continually inviting us to accept simple statements which end up with some phrase, often implied if not actually stated, such as 'other things being equal'. We all know that other things will not be equal. Things do vary – we experience that variation all the time – and so we have to get back to the reality underlying these simplifications wished on us by the media. The first – and a very important – part of the book is, therefore, taken up with looking at how things actually vary and what that variation means. Only when this idea is securely fixed shall we come to look at why they vary and only then at ways of summarising what is represented by that variation.

It might be argued that there are two basic ideas behind Statistics, not one, and the second idea is *uncertainty*. Our reason for leaving this aside, until much later in the book, is that it is little more than the first idea in a different guise. If something varies we are bound to be uncertain about what values we shall observe next. Conversely, if an outcome is uncertain there will be variation in successive values of it which occur. For if there were no uncertainty we would know exactly what to expect, and that is the same as saying that successive values will not vary! The approach via uncertainty is particularly attractive to mathematicians and accounts, perhaps, for the unattractiveness of the subject to those who find mathematics difficult. It is therefore worth seeing how far we can go without it.

Once the idea of variation is firmly fixed we can go on to look at what is known as *covariation*. This is a natural extension of the idea, and with it we move into the realm of relationships among things that vary – and come to the extremely important notions of correlation and causation.

In the second half of the book, we shall see that there may be some point in actually creating variation artificially, because it enables us to understand how to interpret the sources of real variation. This enables us to compare actual variation with what might have happened.

Picturing Variation

Let us begin with a collection of numbers. It does not much matter what they represent, but actually they are the numbers of words per sentence in the leading article of the London *Times* newspaper on 15 April 2011.

27, 23, 33, 15, 21, 38, 16, 15, 4, 19, 21, 9, 33, 41, 10, 30, 35, 19, 17, 31, 33, 17, 22, 10, 22, 29, 35

It is immediately obvious that the number of words varies considerably over the range covered, but there are no very long sentences – with more than 50 words, for example – and few that are very short. It is much easier to take in the general picture if we arrange them in order as follows:

> 4, 9, 10, 10, 15, 15, 16, 17, 17, 19, 19, 21, 21, 22, 22, 23, 27, 29, 30, 31, 33, 33, 33, 35, 35, 38, 41

We can now see immediately that the numbers range from 4 to 41. This allows us to take in much more at a single glance – for example, that although the range is quite large, there seems to be quite a lot of numbers concentrated somewhere in the middle of the range.

A big step on the way to picturing the variation is to place the numbers on a scale as we have done in Figure 1.1. This enables us to see the spacing and, thus, where values are concentrated.

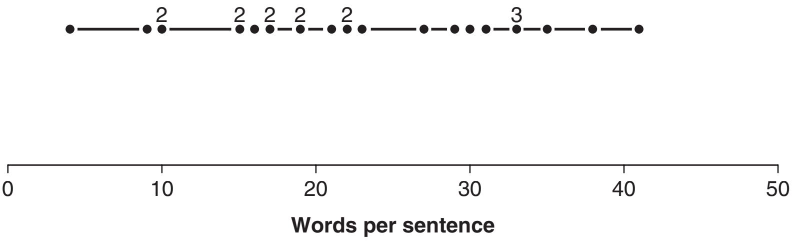

Figure 1.1 Variation of words per sentence for Times(1) data

The numbers printed above some of the dots are frequencies. For example, there were two sentences having 10 words each so a '2' appears above the dot plotted at '10 words'. Figure 1.1 confirms our impression of a concentration around 20 and another in the low 30s, and that is perhaps as far as it is worth going with such a small number of sentences. However, if we had a much larger number of sentences, say 200 or 1000, it would be very tedious to construct a diagram like Figure 1.1 and we might wonder whether there was an easier way to get an overall picture of the variation.

There is, and, paradoxically as it may seem, we can get a clearer view of the general pattern by throwing away some of the detail. And in doing this we meet one of the most important ideas of Statistics, which is that there is often a trade-off between detail and generality. In the present case we shall illustrate this even though, with such a small number of sentences, it might seem rather like 'gilding the lily'. Instead of plotting the lengths as in Figure 1.1, let us merely record how many numbers occur in successive intervals as follows. The choice of interval is somewhat arbitrary, but suppose we merely record how many sentences have lengths in the intervals 0–5, 6–10, 11–15 and so on. The result is as follows:

0–5	6–10	11–15	16–20	21–25	26–30	31–35	36–40	41–45	46–50
1	3	2	5	5	3	6	1	1	0

This is called a *frequency distribution* for the obvious reason that it tells us how the frequency of occurrence of different lengths is distributed across the range. The clustering in the low 30s and around 20 is made more obvious in this way. The final and key step is to present this information pictorially, as in Figure 1.2, in what is called a *histogram*. What we have done is to construct a series of rectangles to replace the frequencies, with the sizes of the rectangles matching the frequencies. The word 'histogram' has a classical origin referring to the height of the rectangles whose base covers the interval for which the frequency is counted.

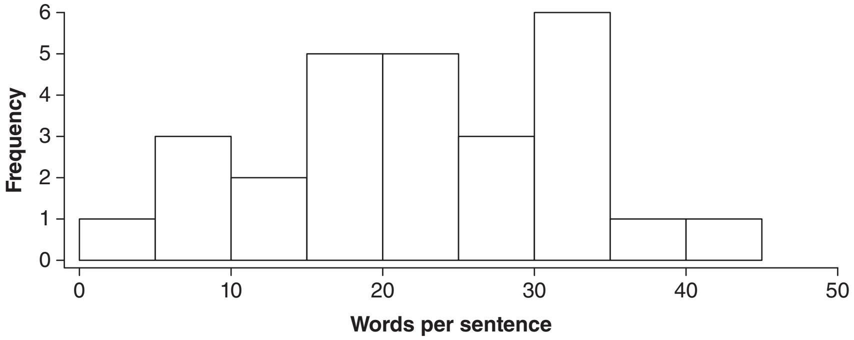

Figure 1.2 Histogram of sentence length frequency for Times(1) data

The histogram provides a *picture* of the variation in sentence length, and the whole of this book revolves around interpreting the *patterns* revealed by histograms. We note, in passing, that there is some ambiguity in this figure about the category into which a sentence length of 20, say, goes. This is an important practical matter, which is taken care of automatically by R, but it need not detain us here, and subsequently, because we are concerned with the overall shape of the distribution which is not significantly affected.

In order to fix the idea behind this example and to see, incidentally, that patterns of variation may themselves vary, we shall repeat the same exercise using the sentence lengths in the second leading article in the same newspaper on the same day. There are rather more sentences in this case and, in the order of occurrence, they were as follows:

> 9, 11, 16, 22, 18, 21, 11, 24, 12, 21, 5, 17, 5, 15, 15, 15, 20, 17, 13, 20,
> 16, 7, 38, 38, 17, 19, 23, 11, 12, 17, 9, 12, 12

It is immediately obvious that they tend to be rather shorter than in the previous article, and this fact becomes more obvious if we list them straight away in order as follows:

> 5, 5, 7, 9, 9, 11, 11, 11, 12, 12, 12, 12, 13, 15, 15, 15, 16, 16, 17, 17, 17,
> 17, 18, 19, 20, 20, 21, 21, 22, 23, 24, 38, 38

The range is much the same as before, extending from 5 to 38 in this case, but the concentration is now between 10 and 20. This becomes even more obvious when we plot the spacings as in Figure 1.3.

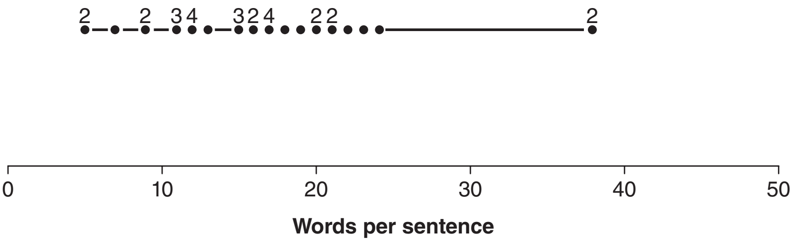

Figure 1.3 Variation of words per sentence for Times(2) data

Finally, we express the data in the form of a histogram in Figure 1.4.

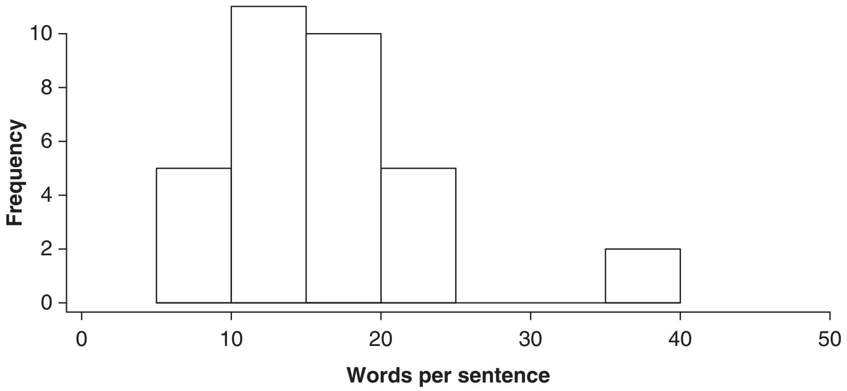

Figure 1.4 Histogram of sentence length frequency for Times(2) data

Both Figures 1.3 and 1.4 show a rather different picture for the second *Times* extract. In this case, apart from the two outliers of 38 words, there is a concentration of frequency in the neighbourhood of 15 words. This shows that the sentence lengths tend to be shorter for the second extract than for the first.

Hitherto we have been concerned with how to picture variability and we have concluded that the histogram achieves what we wanted. There are, of course, other ways of telling the same story and we shall briefly mention some of them at the end of this chapter. First, however, we look at a few other empirical distributions. This will enable us to appreciate the variety of shapes which occur in practice and it also helps us to start thinking about what lies behind these shapes. In short, we would like to know why frequency distributions take on the shape that they do. This leads naturally to the question of what we can infer from the shape of the distribution about the mechanism by which it was generated.

Our first example actually consists of two distributions. They were collected from my own computer, and the reader may care to collect distributions in a similar way. A great deal of traffic on the Internet consists of what is called 'spam'. Spam messages are sent out in bulk by advertisers of one sort or another to many computers and are usually of no interest whatsoever to the recipient. To save the time and effort of looking at all these messages, computers are usually provided with 'spam filters' whose purpose is to extract what is judged to be spam and place it in a 'bin' where it can be quickly scanned by the recipient before deletion.

The number of spam messages varies from day to day. What is the pattern of such variability? To answer this question I collected the daily numbers of spam messages for about a month early in 2013, and later in the year I repeated the exercise for a similar period. Histograms for the two periods are given in Figure 1.5.

Two things about this figure are quite striking. First, the shape of the histogram at the top is very much like that which we found for the word length distribution in Figure 1.2, yet the subject matters are totally different. This raises the hope that there may be a few commonly occurring shapes; if so, it might greatly simplify the matter of interpretation of frequency distributions. Secondly, the histograms in the two parts of Figure 1.5 are quite different even though they refer to the same thing (spam messages). One immediately wonders why there should be this striking difference. There is in fact a good explanation for this difference, but the curious reader will have to wait until Chapter 2 to find out what it is. Note that the vertical scales of frequency are not the same in the two cases. This is because the scales have been arranged to make the histograms about the same size. Size does not matter here because it is the 'shapes' that we are comparing.

We now turn to two distributions which show, among other things, that the shapes already observed by no means exhaust the possibilities.

Unlike the earlier examples, the distribution in Figure 1.6 has a high concentration of frequencies at the beginning, and these tail away rapidly with one extreme value just short of 600 seconds.

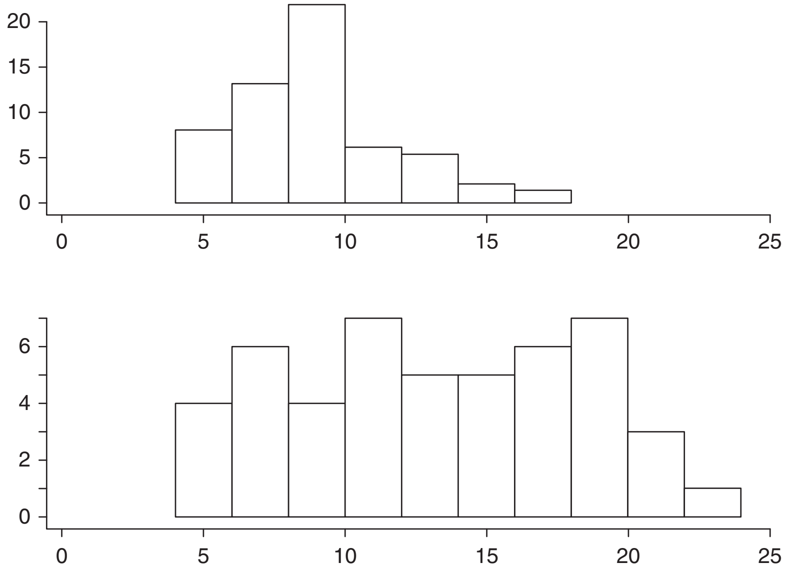

Figure 1.5 Histograms for spam(1) (top) and spam(2) (bottom)

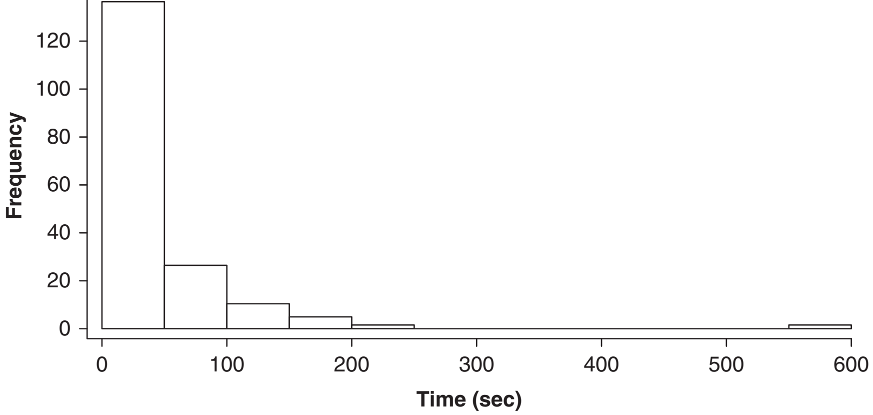

Figure 1.6 Time intervals (seconds) between passing vehicles

The second distribution (Figure 1.7) is somewhat similar in shape but arises in a totally different way, and one does not need to know the exact meaning of 'nearest neighbour' to appreciate this.

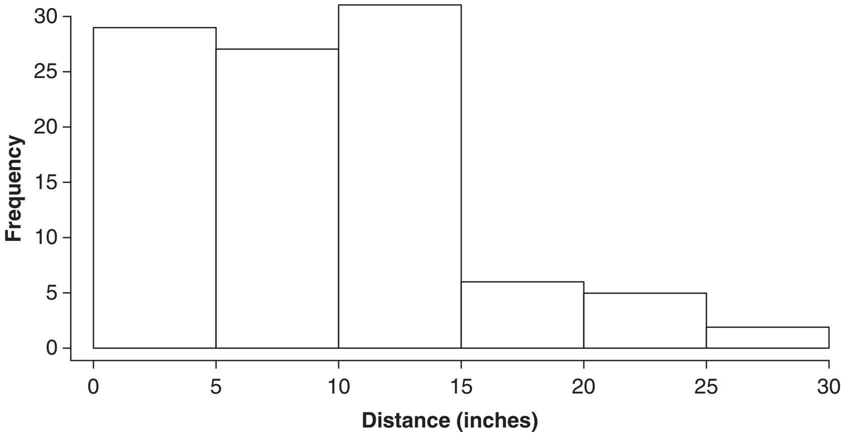

Figure 1.7 Histogram of nearest neighbour distances for seedlings

Figure 1.7 may seem a rather odd choice as an example because it concerns the distribution in space of seedlings growing in the vicinity of sycamore trees. Its interest for us is twofold. Spatial distributions are very important in social science, but real examples are so hedged about with qualifications that it has proved very difficult to find examples to illustrate the point we wish to make at the present elementary level. Furthermore, there is a certain arbitrariness about what we have chosen to measure, and this prepares the ground for a further discussion of this example in the next chapter. Looking at the distribution in Figure 1.7, we notice that it is heavily skewed to the smaller distances but it is not so extreme as the traffic data of Figure 1.6. For the moment we simply note that it forms an interesting addition to the collection of distributions we have met so far.

Our final example may also seem rather strange because it is derived from a source which may well be unfamiliar to most readers. It concerns the distribution of random numbers, the very name of which involves the concept of randomness which will not arise until the later part of this book. The reader must therefore accept, for the moment, that such numbers play a central role in Statistics and that it is useful to know something about their distribution. Imagine that all numbers between 0 and 99 are recorded on slips of paper or on balls and then put into a lottery. Balls are drawn in the manner of a lottery and the number on each is recorded before being returned. We could then construct a frequency distribution of the numbers drawn in the usual way. Figure 1.8 provides the resulting frequency distribution for 200 numbers drawn in a fashion equivalent to that which we have just described.

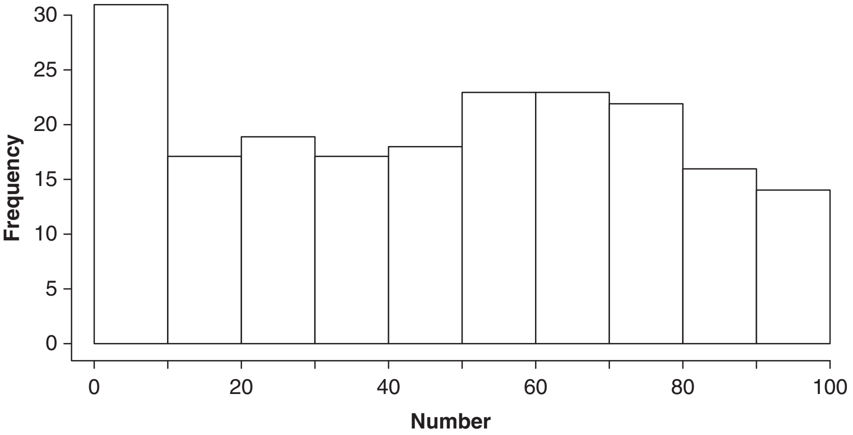

Figure 1.8 Histogram of 200 random numbers between 0 and 99

The histogram has been constructed so that each number group includes ten successive numbers 0–9, 10–19 and so on. We might have guessed that the frequency in each group would be around 20 because the 200 numbers should be evenly distributed. On average that is roughly true, but the actual frequencies range from 15 to 33. These discrepancies raise questions which we cannot answer at the moment, but the main point to notice at this stage is that the broad shape of the distribution is unlike any we have met so far. There is no tendency to level off at the ends and there is no clear peak. It is, in fact, fairly uniform.

These remarks lead us to reflect on the way that we have been using the word 'shape' in the present chapter. When describing the various histograms we might speak of 'having a hump' or 'tailing off' at the ends. In doing this we have, perhaps unconsciously, been 'smoothing out' the ups and downs and imagining a broad underlying pattern. This, in fact, is a characteristic feature of statistical thinking. We are looking for broad patterns which appear to be more 'fundamental', in some not very precisely defined sense.

Coming back to the idea of a frequency distribution, we might wonder whether, if we took a much larger number of random numbers, the irregularities we observed in Figure 1.8 might be 'ironed out' to some extent. This can be investigated empirically by taking a much larger collection of random numbers and seeing what effect this has on the shape of the histogram. The results of one such trial are given in Figure 1.9.

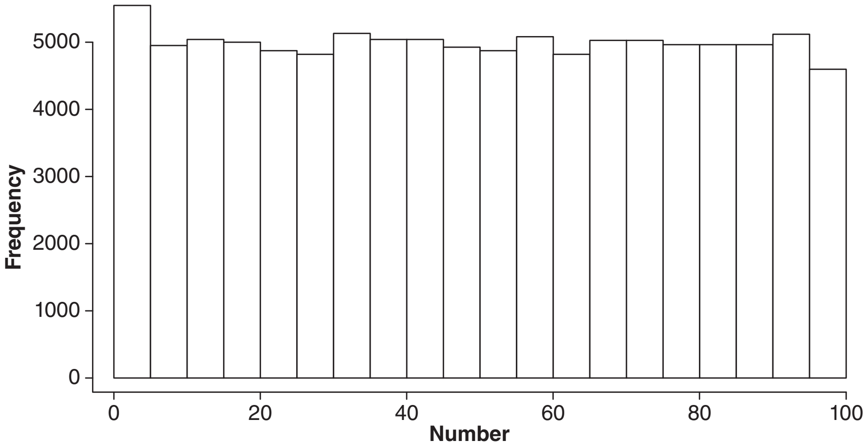

Figure 1.9 Histogram of 100,000 random numbers between 0 and 99

The distribution now appears much more 'uniform', which is a name we might justifiably give to the shape of this distribution.

It seems reasonable to conjecture that patterns might, usually at least, conform to a smooth outline if only we take a large enough number of observations. There are good reasons for this expectation, as we shall see later, but before leaving the matter for the time being, we shall give one example of a histogram whose outline does conform quite closely to a smooth curve.

The data set relating to intervals between traffic was only one of six such sets. The histogram for the distribution in the combined data set is given in Figure 1.10. A fairly fine grouping has been chosen in this case, and although the outline is still somewhat jagged, especially at the upper end, it is, nevertheless suggestive of a fairly smooth declining curve. This point is made even more clearly if we compare this histogram with Figure 1.11, which uses the same grouping intervals to show the data of Figure 1.6.

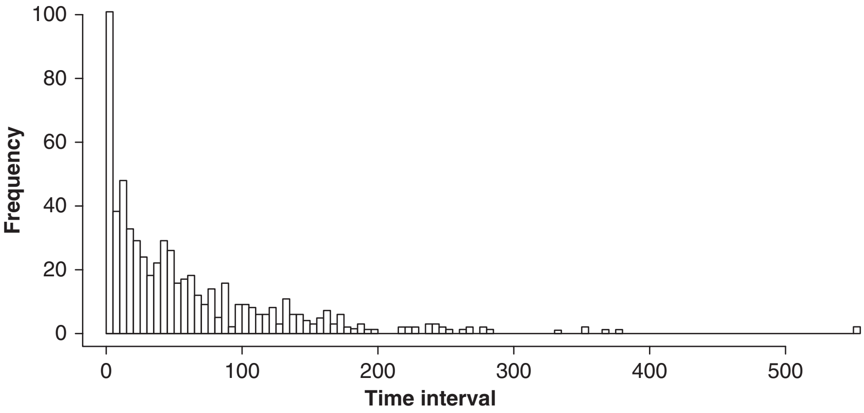

Figure 1.10 Combined distributions for intervals between vehicles

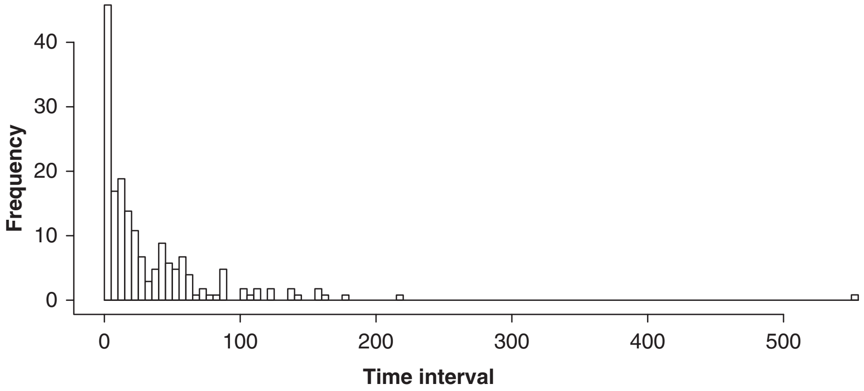

Figure 1.11 Distribution of intervals between vehicles: data of Figure 1.6 with 100 intervals

Roughly speaking, it appears that the larger the data set the easier it will be to describe the outline by reference to a curve with a simple shape. So far, at least, this is a purely empirical observation and there is no reason to expect that this will always or necessarily be the case. However, it will be convenient to get into the habit of imagining the histogram to be approximated by a smooth curve whose shape can be simply described. For example, and it is a very simple example, the distribution of Figure 1.8 may be described as a rectangle. The general name used for such a curve is the *frequency curve*. In later chapters it will often be more convenient to talk about frequency curves rather than histograms, and many of the diagrams given there will show such curves because it is easier to comprehend the message in that way. Nothing more is implied, we emphasise again, than would be conveyed by the equivalent histograms.

Other Ways of Picturing Variability

There are other ways of picturing variability, some of which we shall briefly describe, but none is as widely useful, for our purposes, as the histogram.

The *frequency polygon* is little more than a histogram in another guise. It may be obtained from the histogram by drawing lines connecting the mid-points of the tops of the rectangles which form the histogram. This has the effect of smoothing out the edges of the histogram, thus making it more closely resemble what we have called the frequency curve. The choice between the two is largely a matter of taste, but we prefer the histogram because it can be so easily constructed by any statistical software – especially R which has been used in this book.

The *cumulative frequency distribution* is equivalent to the frequency distribution in the sense that it conveys the same information in a different form. For our purposes this is less good at giving a visual picture of the variation. Instead of listing the individual frequencies, we add them up in a fashion which is best explained by an example. We use the first frequency distribution on sentence length given at the beginning of the chapter. This was set out as in Table 1.1.

Table 1.1 Frequency distribution and cumulative frequency distribution of Times(1) data

	0–5	**6–10**	**11–15**	**16–20**	**21–25**	**26–30**	**31–35**	**36–40**	**41–45**	**46–50**
Frequency	1	3	2	5	5	3	6	1	1	0
Cumulative frequency	1	4	6	11	16	19	26	27	28	28

Each number in the last row is obtained by adding the frequency in the top line to the number immediately to its left. The number 19 for the 26–30 group is thus the cumulative frequency up to and including that group – and similarly for the other entries. We can plot each cumulative frequency against the upper boundary of its associated group. Thus 19 is the number of sentence lengths which have lengths less than or equal to 30. The result of doing this is the cumulative frequency distribution. It does not convey information about the pattern of variability as conveniently as the histogram, as the reader may readily verify by constructing cumulative frequency distributions for some of the data sets used in this chapter.

There is nothing special about adding up frequencies from the left-hand side. It could be done just as well from the right. In some applications, where the variable is a life time, the result of doing this is known as the *survival curve* for obvious reasons.

The *boxplot* can be thought of as a summarisation of the frequency distribution. When describing such distributions we have commented on where the bulk of the frequency is located and said something about the range. The boxplot extracts such information and presents it pictorially as in Figure 1.12.

The 'box' in the centre spans the region where the middle half of the frequency is located. The bold horizontal line marks the centre and the extremes are marked by the horizontal lines connected by dotted lines. The boxplot conveys, pictorially, a good deal of the information contained in the complete distribution, but it cannot compete with the frequency distribution as far as the treatment of this book is concerned.

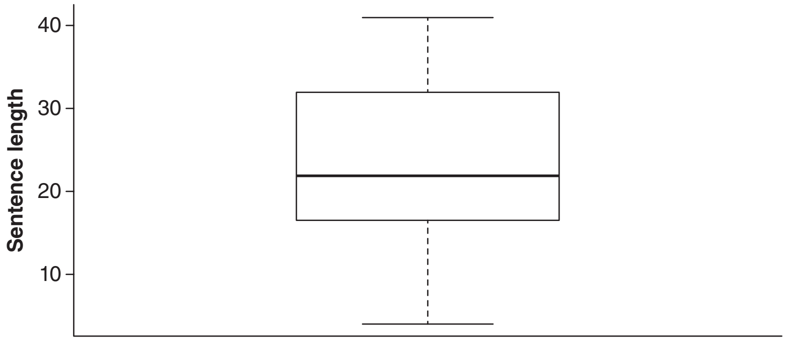

Figure 1.12 Boxplot for Times(1) data

There are other ways of presenting a picture of variability, of which the *stem and leaf* plot is one example, but this is, essentially, a kind of histogram.