PART I: DATA HANDLING

Data are often presented in a form that is not immediately clear. The reader can then either ignore the data, analyse them himself, or return them to their author for *him* to analyse. In the last two cases it helps to know what can be done to make the data easier to comprehend.

Basic procedures for clarifying a table of numbers are described in Chapter 1. Chapter 2 deals with the situation where one already has previous experience of similar data.

After the patterns of the data have been established, the results need to be *communicated*. Here summary figures are more effective than large bodies of numbers. This limits the role of tables and graphs, as is discussed in Chapter 3.

Numerical results cannot be interpreted in isolation but have to be compared with norms and theoretical concepts. This is illustrated in Chapter 4.

CHAPTER 1

Averages and Layout

We start with a table of undigested data. In this chapter we discuss ways of understanding and communicating such data better.

TABLE 1.1 Data in Four Areas and Eight Three-Month Periods in 1969-1970

	13-15	16-18	19-21	22-24	25-27	28-30	31-33	34-36
A	97.63	92.24	100.90	90.39	95.69	94.44	91.13	97.81
B	48.29	42.31	49.98	39.09	46.38	49.74	41.74	37.39
c	75.23	75.16	100.11	74.23	74.23	76.97	71.64	76.47
D	49.69	57.21	80.19	51.09	52.88	49.41	59.32	52.56

The aim is for the reader to be able to see the pattern in the data, which in the present form of the table is not clear, There are a number of useful procedures and guidelines for achieving this, mainly by

(a) improving the display of the data,

(b) developing an explicit summary or "model",

(c) checking on the deviations of the readings from this model.

The readings in the table are not identified so that we can at this stage concentrate on the numerical patterns-about the only case in this book to be treated like that-but the rows represent four geographical areas and the columns eight three-month periods in 1969 and 1970.

1.1 Initial Visual Scanning

The first step is to gain a quick visual impression of the data so as to see the wood for the trees and to be better prepared for a more detailed analysis.

However, Table 1.1 is not easy to take in. One difficulty is the lack of differentiation between the column headings (months 13–15, 16–18, etc.) and the figures in the body of the table. Drawing a separating line, as in Table 1.1a, makes the numbers easier on the eye. This is a common example of the improvements which are possible in the layout of data.

A second problem with the table is the lack of any visual or conceptual focus. What are we supposed to be comparing with what: one area with

TABLE 1.1a Separating off the Headings

	13-15	16-18	19-21	22-24	25-27	28-30	31-33	34-36
A	97.63	92.24	100.90	90.39	95.69	94.44	91.13	97.81
B	48.29	42.31	49.98	39.09	46.38	49.74	41.74	37.39
C	75.23	75.16	100.11	74.23	74.23	76.97	71.66	76.47
D	49,. 69	57.21	80.19	51.09	52.88	49.41	59.32	52.56

another, one period with the next, one *year* with the next, or what? All we can see is a jumble of varied numbers.

But suppose we calculate the average of all the readings, which is about 70. Bearing this one figure in mind, we now see that in the first column Region A is much higher than the average, Region B lower, Region C quite near, and Region D lower again. Looking at the second column, the same pattern seems to recur. We therefore begin to see that the figures in the rows are broadly similar. In Region A the figures are mainly in the 90's, those in Region B are much lower, mostly in the 40's, those in Region C are with one exception in the 70's, and in Region D the figures are mostly in the 50's. This rough summary has been reached by using only the first digit or so in each number, a normal process in the kind of mental arithmetic that is used to scan and interpret numerical data.

In analysing extensive data we can often start with some sub-group. We will use this selective approach here and look only at the data for 1969 in the remainder of this chapter, analysing that for 1970 in Chapter 2. With many kinds of data, one year tends to be like another. If this turns out to be untrue in the present case, our analysis of the 1970 data in Chapter 2 will show it up. At this point the reader may wish to pause and analyse the data in more detail before reading on.

1.2 **Basic Lay-out**

Table 1.2 is an intermediate working table of the 1969 readings that illustrates improvements in the physical layout of the data. It introduces better labelling and table grids and reduces the number of digits in the readings.

TABLE 1.2 Better Labels, Table Grids and Two Significant Figures

Area	Quarters (1969) I	II	III	IV
No	98	92	101	90
so	48	42	50	39
Ea	75	75	100	74
We	50	57	80	51

Better Labels. The abstract area-codes and clumsy period-labels have been replaced with self-explanatory and memorable ones. Labels in tables should aim to be self-explanatory. Usually some mnemonic device is possible, such as *t* for time, *h* for height, and No for North (single letters are only essential for use in *equations)* instead of *x*, y and z.

Table Grids. To guide the eye, ruled lines and gaps of white space have been introduced between certain columns and rows. Usually table-grids are drawn to separate labels and summary figures from the body of the data, while white space on its own is reserved for lesser demarcations or groups of figures. It is worthwhile to draw lines even with rough working-tables and computer print-out because they make the data so much easier on the eye. (Vertical lines are expensive to set with printed matter but not with typed tables. Lines become especially helpful with typed tables because there is less scope for other variations in lay-out, e.g. the use of different type-faces.)

Significant Figures. Perhaps the most effective change made in Table 1.2 has been rounding the readings to two significant figures. In our first rough scrutiny of the data we only looked at the first digit or so in each number, but a widely applicable rule is to work with about *two* significant digits: defined as digits that vary from one number to another.

Recording only two digits has the advantage of making the data clearer to the eye. Mental arithmetic becomes easier to do. Extra digits often hinder one in seeing the data because the least important part of a number is the part looked at or spoken last (e.g. one reads 97.63 as ninety-seven point six-*three*).

There is a loss of accuracy in rounding to two significant digits, but this is usually negligible. For example, in the original 1969 data, the total range of variation is from 39.09 to 100.90, or about 160%. The two decimal places account for only 1 or 2% of that variation. It is unlikely that anyone will be able to interpret such small differences, especially in the early stages of an analysis. The primary need is to see the main patterns in the data. Once the initial analysis is completed, one can see whether the results call for a greater degree of precision. If so, a re-run with an extra digit can be made, But this rarely occurs, the possible exceptions being when a completely new kind of data is analysed.

The effect of such changes of lay-out can be fairly startling. Not only is it easier to see which figures are higher and which lower in Table 1.2, but with hindsight even Table 1.1 begins to look clearer.

1.3 **Summary Figures**

Having improved the presentation of the data, we now turn to analysing the actual information contained in them. Table 1.3 introduces summary

figures through row and column totals. These help us to see that QI and QII, for example, are very similar, both in their totals, 271 and 266, and in the individual figures, 98 and 92, 48 and 44, etc.

TABLE 1.3 *Row* and Column Totals

	Quarters (1969)				
Area	I	II	III	IV	Total
No	98	92	101	90	381
So	48	42	50	39	179
Ea	75	75	100	74	324
We	50	57	80	51	238
Total	271	266	331	254	1122

However, totals are clumsy because they are expressed on different scales of measurement than the original readings. For example, the grand total of 1,122 in the bottom right-hand corner is not directly comparable with any other figure. When analysing data it is best to stick to one type of figure, like the quarterly ones, and to avoid introducing what are effectively new variables. (If a gross annual total is required for some particular purpose, it is easy to multiply a quarterly rate by 4.)

The totals can be reduced to the original scale of measurement by forming *averages*—dividing the totals by the number of readings added together (here 16 for the overall total and 4 for the others). Because the averages have the same visual shape as the individual readings, it may be better to separate them off with full table grids, as in Table 1.3a, rather than merely with extra white space as was done in Table 1.3.

TABLE 1.3a Row and Column Averages

	Quarters				
Area	I	II	III	IV	Av.
No	98	92	101	90	95
So	48	42	50	39	45
Ea	75	75	100	74	81
We	50	57	80	51	60
Average	68	67	83¼	64	70

These averages give us a better feel of the data. Firstly, we see that none of the quarterly averages differs greatly from the overall average of 70, although Quarter III is relatively high at 83. Secondly, we note that the four area averages differ markedly, from a low of 45 to a high of 95. Thirdly, we can see that the area averages reflect quite well what occurs in most of the individual

quarters. The Northern quarterly results are *all* about 95; the Southern ones are all about 45; and while the figures seem more variable in the East and West, *most* of them do not vary much from quarter to quarter.

1.4 More Lay-out: Rows, Columns, and Ordering by Size

Now that we are beginning to understand the data we can rearrange the lay-out accordingly. This is usually worth doing both as an intermediate step, when analysing the data further, and when presenting the data in final form to others.

It is easier to see the relative lack of quarter-by-quarter variation, e.g. 98, 92, 101, 90, and the average 95 in the North, if the figures are written in columns instead of in rows. Table 1.4 accordingly reverses the columns and rows.

TABLE 1.4 Approximately Constant Columns, with Exceptions
(Rows and Columns from Table 1.3a interchanged)

1969	Area No	So	Ea	We	Av.
Q I	98	48	75	50	68
Q II	92	42	75	57	67
Q III	101	50	100	80	83
Q IV	90	39	74	51	64
Average	95	45	81	60	70

In running down the North column in Table 1.4, the eye can read off the consecutive digits in the "tens" column as 9, 9, 10, 9, 9, largely bypassing the figures in the "units" column. In contrast, reading the data for North *horizontally* in Table 1.3a, the eye had to scan 9, 8, blank, 9, 2, blank, 101, blank, and so on.

The new lay-out also makes any exceptions stand out more clearly. We have already noticed that the East and West figures are more variable, but now the table shows that there are in fact two exceptional figures in the data, the 100 in QIII for the East and the 80 in QIII for the West. Running down these columns in Table 1.4 we read 7,7, **10,** 7, 8, and 5,5, **8**, 5, 6.

This visual gain is lost again if the rows of the table are more widely spaced, as in Table 1.4a. The more luxurious lay-out with extra white space fails to guide the eye.

The numerical nature of the data can also help to determine the best *order* of the columns or rows of the table. Rather than keep to their predetermined

TABLE 1.4a The Rows in Double Spacing

1969	Area No	So	Ea	We	Av.
Q I	98	48	75	50	68
Q II	92	42	75	57	67
Q III	101	50	100	80	83
Q IV	90	39	74	51	64
Average	95	45	81	60	70

and possibly accidental order, the columns have been rearranged in Table 1.4b by the size of the averages, i.e. as North, East, West and South.

TABLE 1.4b Columns Rearranged in Order of their Average Size

1969	Area No	Ea	We	So	Av.
Q I	98	75	50	48	68
Q II	92	75	57	42	67
Q III	101	100	80	50	83
Q IV	90	74	51	39	64
Average	95	81	60	45	70

This gives the table a clearer structure. Knowing that the column *averages* decrease from left to right, we can readily check whether the individual rows also decrease. In our example we can see that they do and that a marginal case like the first two figures in QIII also stands out clearly, as would any real exception. Doing the same thing in Table 1.4 we would have had to check each row against the less memorable high-low-high-low pattern of the column averages 95–45–81–60. The quantitative detail would have been almost impossible to sort out.

The columns can be arranged either in decreasing or increasing order of size; i.e. either starting with the big values as here, or in line with common practice in plotting graphs. Apparently there is no strong perceptual argument either way for columns. But with the rows of a table it helps to put rows with larger numbers above rows with smaller ones, since it is easier to subtract mentally that way.

However, in our example there is no point in rearrangingthe order of the rows because there is little variation between the quarterly averages. The "natural" order of the four quarters here might in any case seem sacrosanct. But using the dimensions of a table to represent the previously unknown pattern of the data can be more useful to the reader than repeating the well-known order of some row or column labels (e.g. everyone already knows that QII follows QI).

There can, however, be other considerations in all this. For example, if there are many tables with the same format, usually the overriding concern is to keep the same lay-out to facilitate visual comparison between the different tables This can also apply to the interchanging of rows and columns (although it may still be useful for the analyst to do this as an intermediate step).

1.5 Averages and Exceptions

Averages are the main tool for summarising extensive numerical data. However, the four area averages in Table 1.4b are not good summaries because the data are dissimilar. In the North and South all the readings are close to the averages of 95 and 45, but in the East and West they are not, mainly because of the two exceptional QIII readings. Therefore it is misleading to compare the four areas merely in terms of their averages.

When there is a small number of exceptional readings, the data are easier to describe if one excludes these exceptions from the main summary figures. This is done in Table 1.5, which gives "adjusted" averages for the East and West. We now have more effective summaries of the data-the readings are generally about 95 in the North, 75 in the East, 53 in the West, and 45 in the South, with two large exceptions for QIII in the East and West.

TABLE 1.5 The **Adjusted Area Averages**

(QIII in East and West excluded)

	Area				
1969	No	Ea	We	So	Av.
Q I	98	75	50	48	68
Q II	92	75	57	42	67
Q III	101	(100)	(80)	50	(83)
Q IV	90	74	51	39	64
Average	95	75*	53*	45	67*

* Excluding QIII

Having noted the exceptional readings, we have to start checking on the reasons for them. The most likely explanation is a computing or clerical

error (Twyman's Law that "any figure that looks interesting or different is usually wrong"). In practice we would of course know what the readings in the table referred to, e.g. the sales of Product X, the number of working-days lost due to absenteeism, or the incidence of measles, and could therefore try to find out whether there had been anything special that year to cause such exceptions, such as a price-cut, a large strike, or an epidemic, and whether equally high readings also occurred the year before. (If we are completely new to the data, we can often ask old Joe next door who has been around for 20 years and knows everything.)

Whether or not we find an explanation., the exceptional figures are best excluded from the basic summary figures. This might seem misleading (as if the analyst were trying to mislead himself), but the QIII readings will be exceptional no matter how we report them. We can either give four comparable averages with two exceptions, or four averages that are *not* comparable and two exceptions. No one is trying to hide anything; the exceptional readings stand out even more from the new averages than from the old ones simply because they were excluded. (Chapter 2 will show whether these adjusted averages are successful when checking them against additional data. This is the ultimate test of the usefulness of any description of data.)

If the readings being summarized are not all more or less constant (as in each column here), allowing for exceptional figures will be more difficult. Thus the data in the rows are more complex than in the columns, but at least the three row averages for QI, QII and QIV also provide good summaries. The averages are similar and we can also see that the scatter of readings about each average takes the same form. These row averages therefore perform the useful role of showing where the distributions of readings quarter-by-quarter are the same and of making the exceptional QIII stand out clearly.

With the four area averages we now have a possible summary or "model" of the main features of the observed data. This model says that the quarterly figures are generally about 95 in the North, 75 in the East, 53 in the West and 45 in the South. (The model or "theory" here is obviously on a very low level, but this is because of the narrow range of the data and not because of the modelling process itself.)

1.4 Deviations from the Model -The Final Step

The final step is to examine the "fit" of this theoretical model, i.e. the differences between the observed readings and the area averages (e.g. $98 - 95 = 3$ in QI in the North), as shown in Table 1.6.

Here once again we face an array of numbers that are new and largely undigested. But the data are already somewhat better organised than in Table 1.1 and we know something about their background. For example, we

TABLE 1.6 Deviations Between the Observed **and Theoretical** Figures

1969	Area No	Ea	We	So
Q I	3	0	-3	3
Q II	-3	0	4	-3
Q III	6	25	27	5
Q IV	-5	-1	-2	-6

know about the two exceptions in QIII, and that, compared with a. range from an average of 95 in the North to one of 45 in the South, most of the deviations are small. Judging the deviations on their own, the variation of the numbers in Table 1.6 is however quite marked, e.g. from $+3$ to -3 to $+6$ to -5 in the first column, and so on. (Expressing these readings to more digits, e.g. as 2.6, -2.8, 5.9, -4.6, etc., let alone to a *second* place of decimals, would therefore have been pointless.)

In analysing the deviations we first work out row and column averages as in Table 1.6a, again excluding the exceptional values in QIII.

TABLE 1.6a **Averages**

1969	Area No	Ea	We	So	Av.*
Q I	3	0	-3	3	1
Q II	-3	0	4	-3	0
Q III	6	(25)	(27)	5	6*
Q IV	-5	-1	-2	-6	-4
Average	0	0*	o*	0	0

*Excluding QIII in East and West

The *column* averages are all zero because we used the area averages as our model. Therefore positive and negative deviations must balance out. The area averages in Table 1.6a tell us only that the arithmetic is right.

The *row* averages are more variable. The QIII average of 6 reflects individual readings that are positive, and the QIV average of -4 reflects individual readings that are all negative. But there are no such patterns in the *first* two quarters: in QI, the readings in the North and South are positive and that in the West is negative while the opposite occurs in QII, negative readings in the North and South and a positive one in the West. Not too much should therefore be made of the apparent pattern in QIII and QIV, since it

has to be interpreted in conjunction with the *absence* of a pattern in QI and QII.

Concluding that there is no general pattern of *any* kind in the table would, however, be premature. The signs in the first column alternate systematically + − + -. The same occurs in the South, and there is something of this pattern even in the West (and virtually *no* variation in the East anyway). Successive readings over time-usually called a "time-series"-can contain a tendency to be specially related to each other. This is called serial correlation. In our example the serial correlation is negative. A relatively high reading is followed by a low one, and a low one is followed by a high. (For sales data, for example, this might reflect a tendency to over-sell in one quarter, leading to an excess of stock, and so followed by under-ordering in the next quarter.) With only 16 readings for 1969 it is impossible to tell whether the apparent pattern here is real. But when we analyse the 1970 data in the next chapter we can readily check whether the pattern generalises.

1.7 The Size of the Deviations

In analysing the deviations of the original readings from their averages? some are positive and some negative and they necessarily have an overall average of zero. This does not help to describe the data. Instead, we need a measure of the size of the deviations irrespective of their signs.

For example, in the North the straightforward average of the four deviations is 0 to the nearest whole number ($3 - 3 + 6 - 5 = 1$, divided by 4). But the average *ignoring the sign* is 4, i.e. ($3 + 3 + 6 + 5 = 17$ divided by 4). This tells us how *big* the deviations generally are. Table 1.7 gives the average size of the deviations for each area, ignoring the negative signs.

TABLE 1.7 The Average Size of the Deviations

1969	Area No	Ea	We	So	Av.
Av. Size	4	0*	3*	4	3

*Excluding QIII in East and West

The overall average size of the deviations is 3 (again ignoring the two QIII exceptions). The areas do not vary much in this respect, except that the values in the East are low. Similarly, visual inspection of Table 1.6a shows that the average size of the deviations in each cluarter is also about 3, except that in QIII even the two "normal" readings are on the large side.

Thus, apart perhaps from the high values in QIII and the low ones in the East, the average size of the deviations does not vary dramatically by area

or by quarter. The overall average of 3 is a fairly good summary. (Such a measure of the average size of the deviations ignoring sign is usually known as the *mean deviation.* This and other measures of statistical scatter are discussed in Chapter 11.)

1.8 A Full Description

In Section 1.2 at the beginning of this chapter we started with the 16 quarterly readings for 1969, as reproduced in Table 1.8.

TABLE 1.8 The Original Data for 1969

	13-15	16-18	19-21	22-24
A (North)	97.63	92.24	100.90	90.39
B (South)	48.29	42.31	49.98	39.09
C (East)	75.23	75.16	100.11	74.23
D (West)	49.69	57.21	80.19	51.09

Now that we have achieved some understanding of the data, they seem fairly clear even in their original form. In addition, we have reduced the data to seven summary figures.

Four area averages: North 95, East 75, West 53, South 45.

An apparently irregular quarterly variation of about 3 units about these averages.

Two exceptionally high values in QIII (months 19–21) in the East and West, differing from the area averages by about 25 units.

Seven summary figures, the four area averages, one measure ofaverage-scatter, and two large exceptions, may not seem like much of a reduction of 16 initial readings. One reason for this limited success is that the data analysed so far were not, in any case, very extensive. Another reason is that we still have not accounted for the main regularity that we have found, the large systematic differences between the four areas, nor for the residual scatter, i.e. the deviations of the individual readings. What the analysis *has* achieved is to bring into focus the fact that it is these features which need to be explained. Before attempting any deeper explanation of these features it is worth checking whether they generalise. The 1969 data started with month 13 so there should be at least one year's previous data; but since Table 1.1 already gave the *following* year's data, we shall analyse this in the next chapter.

1.9 Summary

Most data reach us in undigested form. In this chapter we have illustrated various analytic steps that can be used to see the pattern in a table of numbers and to communicate the results. The main guide-lines are as follows.

(i) See if the number of digits shown can be reduced. (Mental arithmetic is difficult with more than two significant digits, i.e. ones which vary.)
(ii) Use self-explanatory, memorable symbols and labels (e.g. t for time, h for heights, and No for North, not x, y and z).
(iii) Separate different types of items or sub-groupings in a table by grid-lines or white space.
(iv) Use averages to help focus the eye when examining any array of numbers.
(v) Ensure that figures that must be compared are close together.
(vi) Use the *columns* of a table for figures that need to be compared, to make both the regularities and the exceptions easier to see.
(vii) See if the columns or rows of a table can be ordered "graphically" to reflect their average size, thus making it easier to see patterns in the body of the table.
(viii) Avoid introducing new variables or scales whenever possible (e.g. use averages rather than totals as summary figures).
(ix) Note any dramatic exceptional values separately and exclude them from the main summary figures.
(x) Summarise irregular aspects of the data statistically, e.g. by a measure of the average size of the deviations.

In general, one should aim to present simple, well-digested patterns and summaries. The alternative amounts to leaving the analysis to the reader.

The analytic process which has been described here is organic rather than mechanical. One step is usually taken because of the patterns revealed by the previous steps. This makes the analysis laborious and slow, but helps to ensure that the final result models the actual structure of the data. Any subsequent analysis of similar data should then be relatively quick and straightforward. This is illustrated in Chapter 2.

CHAPTER 1 EXERCISES

Exercise 1A. Alternative Analyses of the Data

Discuss alternative ways of handling the original data analysed in the present chapter.

Discussion.

Popular approaches to such time-series data include the use of graphs, the calculations of percentage changes or shares, and the use of indices or moving averages. (A widely used alternative is to leave the raw data for the reader to sort out.)

Graphs. Figure 1.1 shows the 32 readings for 1969 and 1970 graphed. The QIII blips in the East and West show up clearly. 'They typically dominate the picture, but the generally steady levels and the tendency for slightly lower values in QIV in 1969 can also be discerned.

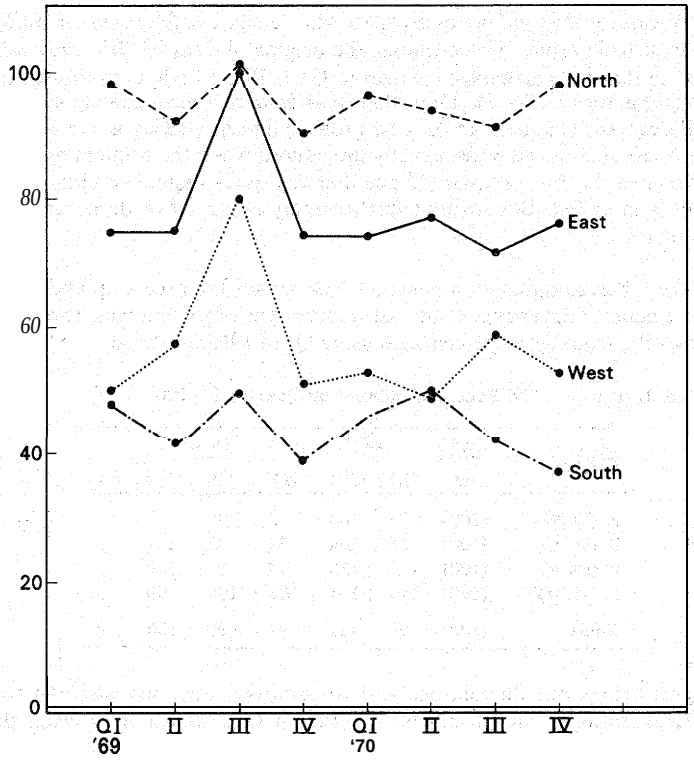

Figure 1.1 Plotting the Four Areas over Eight Quarters

Graphs are, however, relatively time-consuming to construct. Reading-off any detailed *quantitative* information is difficult, if not impossible. Nor does a graph provide a succinct and usable *summary* of the data.

Percentage Changes. Table 1.9 shows the data expressed as percentages of the preceding quarter (a popular alternative is to show percentage changes from the same quarter a year before).

TABLE 1.9 **Quarter-by-Quarter Percentage** Changes

Region		1969 QI	QII	QIII	QIV	1970 QI	QII	etc.
A	(North)	-	94	109	89	107	99	.
B	(South)		- 88	118	78	119	107	.
C	(East)	-	100	133	74	100	104	.
D	(West)	-	115	140	64	104	93	.

Any feeling for the actual data is lost in this form of presentation, e.g. that the North is bigger than the South. The exceptionally high values in QIII show up, but they also recur as exceptionally *low* values in QIV. With such percentages, every variation in the original figures shows up

twice, once as up and once as down (or vice versa). The figures are therefore difficult to interpret. For example, the original figures in QIV 1969 were low, so that the percentage changes to QI 1970 are high, even though the actual figures (95, 76, 54, 46) in that quarter were almost exactly average.

Percentage changes can only be properly interpreted by going back to the original data. They are usually introduced when the original readings have been shown to so many digits that any quick mental scrutiny of the data is inhibited. But without further analysis they provide no effective summary.

Indices. Percentaging on a constant base avoids the excess up-and-down movements of the change-from-last-quarter type of percentages. Table 1.10 shows the readings as percentages using QI of 1969 as a base.

TABLE 1.10 The Readings Indexed on Quarter I = 100

Region	1969 QI	QII	QIII	Q I V	1970 QI	QII	etc.
A (North)	(100)	94	103	93	98	97	.
B (South)	(100)	88	104	81	96	103	.
C (East)	(100)	100	133	99	99	102	.
D (West)	(100)	115	161	103	106	99	.
Total	(100)	99	122	94	99	100	.

This brings out fluctuations and trends over time, and also the two QIII exceptions, but it does little better in this respect than when the original data were shown to two digits. All feeling for the size of the original readings is once more lost (e.g. that the North is twice as big as the South). The choice of base figures is also often arbitrary, and becomes more so as time progresses.

Regional Shares. Table 1.11 gives each area as a percentage of the national total that quarter. (Such share figures can be meaningful with "additive" quantities like numbers of people, sales of a product or rainfall, but not with ages, prices or temperatures.)

TABLE 1.11 Regional Shares

Region	1969 QI	QII	QIII	QIV	Av.
	%	%	%	%	
A (North)	36	35	31	36	34
B (South)	18	16	15	15	16
C (East)	28	28	30	29	29
D (West)	18	21	24	20	21

The general stability of the figures is made clear. It is also easy to see that the North is just over twice as large as the South. Most of this visual clarity, however, is lost in the common practice of showing percentage figures to one decimal place (i.e. as "per mille"), as in Table l.lla.

TABLE 1.11a The Percentages **to** One Decimal Place

Region	1969				
	QI	QII	QIII	QIV	Av.
A (North)	36.0	34.6	30.5	35.5	33.9
B (South)	17.8	15.9	15.1	15.3	16.0
C (East)	27.8	28.1	30.1	29.1	28.9
D (West)	18.3	21.4	24.2	20.1	21.2

Exceptional results do not show up well in either table because they influence the total on which the percentages are based. This can also influence other figures. For example, the North had its lowest share in QIII, 31%, but its highest *absolute* figure, 101 (see Table 1.8). Percentage shares are best used when there are large but proportional variations in the absolute figures from column to column or row to row. The shares are then more or less steady.

Moving Averages. These are a device for smoothing away some of the fluctuations that occur in time-series data. In our example one would first calculate each area average over the four quarters QI to QIV of 1969 (e.g. 95.4 in the North). Then one would calculate the area averages over the next four quarters QII 1969 to QI 1970 (e.g. 94.7 in the North), the area averages over QIII 1969 to QII 1970 (95.2 in the North), and so on.

The figures for each area are then very steady, but they are difficult to interpret because an increase from one reading to the next could be due to a low figure being dropped or a high figure being added in forming the next average. (A better alternative is to summarise the regularities in the data by an overall average or trend line, with deviations. The fitting of such lines is discussed in Part II.)

Exercise 1B. Another Example

What is the main pattern of the following four pairs of readings?

	1956	1957	1958	1959
x:	108	60	89	51
Y:	206	158	187	149

Discussion.

Applying the various guide-lines of Chapter 1 (e.g. arranging in order of size, averaging, and rounding) helps to bring out the pattern which is otherwise not immediately obvious.

For example, rearranging the X values in order of their size and using single spacing gives

	1959	*1957*	1958	1956
X:	51	60	89	108
Y:	149	158	187	206

This makes it easy to see that the X and Y values follow the same trend. Departing from the natural order of the years may appear peculiar but would be done without second thoughts in plotting the data as in Figure 1.2. (Plotting such a graph is more laborious than simply re-arranging the numbers in a new table, but it also shows clearly that X and Y vary together.)

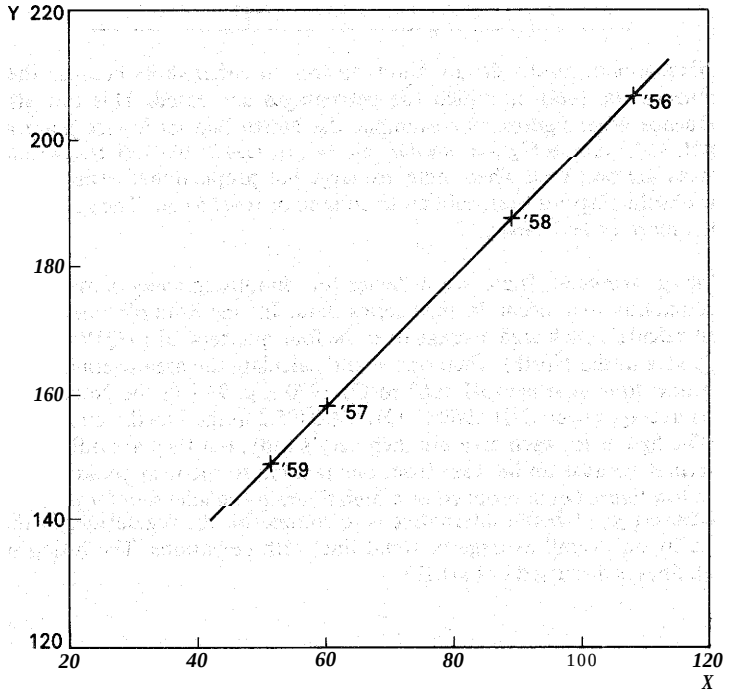

Figure 1.2 Plotting Y against X

Introducing averages shows up the *quantitative* form of the relationship:

	1959	1957	1958	1956	Av.
X:	51	60	89	108	77
Y:	149	158	187	206	175
Av.	100	109	138	157	126

The average values of Y and X differ by 98. With this figure in mind we can scan the individual years to see if the differences are higher or lower. We see that the differences are in fact all 98 (e.g. 149-151 in 1959), so that

$$Y = X + 98.$$

That is the pattern.

Rounding off to the nearest 10 can be done without serious loss of information and also vastly clarifies the pattern:

	1959	1957	1958	1956
x:	50	60	90	110
Y:	1.50	160	190	210

Approximately, therefore, X and Y differ by 100.

However, even after getting to know it, this result remains fairly difficult to see in the original data:

	1956	1957	1958	1959
X:	108	60	89	51
Y:	206	158	187	149

This is because we mostly find it hard to subtract large numbers from smaller ones written above them. Rearranging the rows with the larger numbers on top (and in single spacing once more) helps:

	1956	1957	1958	1959
Y:	206	158	187	149
x:	108	60	89	51

In conclusion, since the simple pattern in these four pairs of readings was relatively difficult to see in the original form, one wonders what regularities are being missed in *other* data due to clumsy lay-out and presentation.

Exercise 1C. The Current Economic Climate

Each week the American journal *Business Week* publishes certain "Figures of the Week". Discuss the production figures shown in Table 1.12 from the issue published April 24, 1971 (reproduced by permission).

TABLE 1.12 Business Week's "Figures of the Week" (April 24, 1971)

Production	1957-59 average	Year ago	Month ago	Week ago	Latest week
Raw steel *[Amer. Iron & Steel Inst., thous. of net tons]*	1,860	2,686	2,844	2,932	2,906
Automobiles *[Ward's Automotive Reports]*	102,264	139,838	190,160	146,400	169,941
Electric power *[Edison Elec. Inst., millions of kilowatt-hours]*	*12,385*	*27,200*	*26,735*	*28,633*	*28,111*
Crude oil, refinery runs *[Amer. Pet. Inst., daily av., thous. of bbl.]*	*7,852*	*10,556*	*11,014*	*11,132*	*11,315*
Bituminous coal *[Bureau of Mines, daily av., thous. of net tons]*	1,437	2,011	2,114	2,140	2,164
Paperboard *[Amer. Paper Inst., thous. of tons]*	*284.3*	*521.0*	*513.5*	*523.8*	*503.0*

Discussion.

The apparent purpose of the magazine's "Figures of the Week" is to help non-technical businessmen and administrators to judge the current economic climate with a quick indication of recent trends. However, the

data are expressed to too many digits to be taken in easily by eye. (Such detail could not even influence the most painstaking economic analysis.)

Table 1.12a shows the data reduced to two digits. Decimal points are avoided by using suitable units, like 100,000 tons of steel instead of million tons. The use of odd-looking units (like 10,000 cars) could be obnoxious in other cases, but it does not seem to matter here since the units quoted from *Business Week* are not self-explanatory to the lay-user anyway: e.g. tons of paper board but *net* tons of steel, bulk barrels of oil (bbl), and "daily averages" without saying whether these are based on a 5- or 7-day week.

TABLE 1.12a The Figures Rounded and Ordered by Size

Weekly Production Figures	1957-59 average	Year ago	Month ago	Week ago	Latest week
Paperboard (in 10,000 tons)	28	52	51	52	50
Raw Steel (in 100,000 net tons)	19	27	28	29	29
Bitum. Coal (daily av., 100,000 tons)	14	20	21	21	21
Electricity (in billion kwh)	12	27	30	29	28
Automobiles (in 10,000)	10	14	19	15	15
Crude Oil (daily av. in million bbl)	8	11	11	11	11

In Table 1.12a the rows have been arranged in decreasing order of size, using the 1957-59 averages as the norm because these appear every week. This ordering depends on the arbitrary choice of measurement units, but that is unimportant. What matters is that it is now visually easier to concentrate on the figures.

For example, we can now see quite easily that since 1957-59 production of the different items has increased by between 50 % and 100% and that the four figures for the last year are rather similar. Since it is better to place figures with slight variation in columns, we can interchange the columns and rows, as in Table 1.12b, and move the information about measurement units below the table.

TABLE 1.12b Rowe and Columns Interchanged

	Paper	Steel	Coal	Electr.	Autos	Oil
1957-59 average	28	19	14	12	10	8
Year ago	52	27	20	27	14	11
Month ago	51	28	21	30	19	11
Week ago	52	29	21	29	15	11
Latest week	50	29	21	28	15	11

Paperboard in 10,000 tons; Raw Steel in 100,000 net tons; Daily average of Bituminous Coal in 100,000 net tons; Electricity in billion kwh; Au tomobiles in 10,000; Daily average crude oil refinery runs in mill. bbl.

Now it is clear that the oil production figures have been 11 for the past year (but have they perhaps been rounded too much?), that auto production was exceptionally high a month ago (a printing error, the end of a strike, or what?), that the electricity figures fluctuated up and down, etc.

In this lay-out it is a matter of choice whether to give the 1957-59 base-line at the top or bottom of the table; the real question now is whether this base-line tells us anything useful anyway.

Exercise 1D. The National Income

Discuss the first five rows of National Income figures from the 1971 edition of the Government Statistical Service's "Britain's Economy in Figures" (a 3 x 5 inch pocket-card):.

Government Statistical Service

BRITAIN'S ECONOMY IN FIGURES

	1964	1969	1970
Population and Manpower **000s**			
Home population (June):			
United Kingdom (94,216 sq. miles)	**54,008**	**55,534**	**55,711**
England (50,333 sq. miles)	**44,658**	**46,102**	**46,254**
Wales (8,017 sq. miles)	**2,676**	**2,725**	**2,734**
Scotland (30,414 sq. miles)	**5,206**	**5,195**	**5,199**
N. Ireland (5,452 sq. miles)	**1,458**	**1,513**	**1,524**
Working population (June)	**25,849**	**25,802**	**25,637**
Wholly unemployed (average)	**405**	**582**	**619**
Full-time students in higher education in Gt. Britain (Autumn of previous year)	**239**	410	427
National Income **£ million**			
Gross national product at current factor cost	29,319	38,913	**42,667**
Consumers' expenditure	21,459	**28,683**	**31,124**
Gross fixed investment	**5,857**	**8,024**	**8,742**
Public	**2,580**	3,714	4,004
Private	**3,277**	4,310	4,700
Income from employment	**19,702**	27,141	30,246
Gross trading profits of companies	**4,591**	4,948	5,036
Economic growth (gdp at constant factor cost 1963= 100)	105.6	118.4	120.4
Public Expenditure **£million** (years **ending 31 March)**			
Defence	1,811	**2,245**	**2,206**
Education	1,393	**2,301**	**2,513**
Health and welfare (excl. N. Ireland)	1,107	**1,770**	1,931
Social security benefits (excl. N. Ireland)	1,976	3,294	3,538
Roads and public lighting (excl. N. Ireland)	365	590	648
Overseas aid (excl. N. Ireland)	156	177	192

		1964	1969	1970
Production				
Index of industrial production 1963 = 100		108.3	122.9	124.0
Output per head 1963 = 100		106.5	124.9	128.1
coal m. tons		193.6	150.5	142.3
Gas-total available m. therms		3,417	5,517	6,350
North Sea m. therms		—	1,001	3,586
Electricity-total generated m. kWh		165,445	219,087	228,894
From nuclear power m. kWh		5,341	25,271	21,871
Crude oil processed m. tons		58.5	90.3	100.3
Merchant ships under construction (end-year) 000 gross tons		1,616	1,617	1,612
Man-made fibres million lb.		825	1,221	1,321*
Passenger cars 000s		1,868	1,717*	1,641
Transport (Great Britain)				
Index of vehicle miles travelled on roads 1963 = 100		112	144	152
Index of ton miles of inland goods (road and rail) 1963 = 100		111	127	132
Licences current:				
Total road vehicles 000s		12,370	14,753	14,951
Motor cars 000s		8,247	11,228	11,516
Mileage of motorways in use (at 1 April)		292	622	662
Prices, Incomes and Money 1963 = 100				
Retail prices		103.3	127.2	135.3
Average earnings		107.1	147.4	165.0
Real personal disposable income per head		103.2	109.4	112.8
Money supply (M3) (end-year)		105.4	145.3	159.0
Balance of payments	£ million			
Current account balance		−395	+437	+631
Visible		-519	−141	+3
Invisible		+124	+578	+628
Total currencyflow		−695	+743	+1,287
Level of official reserves	end Dec.	827	1,053	1,178
Total external assets	end Dec.	13,145	19,375	n.a.
Total external liabilities	end Dec.	11,350	16,685	n.a.
Overseas Trade	£ million			
Total value of exports fob†		4,565	7,339	8,063
Sterling area		1,554	2,040	2,218
EEC		963	1,530	1,754
EFTA		641	1,080	1,277
North America		623	1,223	1,231

	1964	1969	1970
Total value of imports cif	5,696	8,315	9,051
Sterling area	1,874	2,404	2,460
EEC	941	1,609	1,822
EFTA	748	1,248	1,406
North America	1,109	1,639	1,857
Food, drink and tobacco	1,771	1,930	2,052
Raw materials and fuel	1,702	2,159	2,310
Semi-manufactures	1,324	2,302	2,509
Manufactured goods	837	1,835	2,070

* 53 weeks. † Excluding net adjustment for under-recording.

Further copies can be obtained from the Central Statistical Office, Great George Street, London S.W1, or Regional Offices of the Department of Trade and Industry.

Prepared by the Central Statistical Office and the Central Office of Information, May 1971.

Printed in England for Her Majesty's Stationery Office by Stephen Austin and Sons Limited, Hertford.

(Reproduced by permission of The Central Statistical *Office)*

Discussion.

This particular card is a little out-of-date and has since been improved, but it illustrates how the business of presenting almost unintelligibly detailed figures has not always been restricted to business journals or to one side of the Atlantic.

For an "at-a-glance" card, figures to the nearest thousand million are enough and help one to see by simple mental arithmetic that GNP has increased by 50% from 1964 to 1970 (43 – 29 = 14, which is half of 28), and by 10% since 1969 (43 – 39 = 4, which is about 10% of 43).

National Income (£'000 million)	1964	1969	1970
Gross National Product	29	39	43
Consumers' Expenditure	72%	74%	72%
Gross fixed investment	6	8	9
Public sector	44%	46%	46%

The percentages have been introduced not because they are inherently a good thing to use with economic data, but because they remain almost constant despite the increases in GNP and gross investment, and thus reflect simple aspects of the data in a simple manner. We now have figures which are memorable, like consumers' expenditure being roughly a constant 73% of GNP, and public investment being roughly 45% of total capital investment. If it is felt that minor variations are important, such as public investment going up from 44 to 46%, then this can now at least be ***seen.*** (Because the pocket-card requires reasonably explicit descriptions of each item, rearranging the rows and columns seems in this case impossible, even if the pattern of the data suggests it.)

If we now move on to the more up-to-date information of this kind, the 1973 edition of the GSS card entitled "United Kingdom in Figures" (a 10 x 5 inch folding pocket-card), we find several changes.

Two are that the figures are shown to three digits only and that breakdowns are shown as percentages. Thus the comparable National Income figures, as given on the 1973 card, are

	1951	1966	1972
Gross Domestic Product (£'000 mn.)	12.6	32.8	52.6
Used for-Consumption %	87.5	80.6	79.8
-Investment %	12.5	19.4	20.2

Much has been gained since 1971 but the third digit still inhibits any simple visual interpretation, compared with the following version:

	1951	1966	1972
Gross Domestic Product (£'000 m)	13	33	53
Used for-Consumption %	88	81	80
-Investment %	12	19	20

Another change from the 1971 card is one of definitions, e.g. something called "Consumption" at over 80% instead of "Consumer expenditure" at 72% or so. Such problems over definitions (presumably *public* consumption has been added in), make it especially doubtful whether *three* digits on such a card can serve any useful purpose.

Exercise 1E. Elegant Variations

Improve the following report on the incidence of fatal accidents in air travel (from *The Guardian*, July 31, 1971):

> "In 1950... the world's airlines carried 30 million people. There were 27 fatal crashes that year, and just over three passengers were killed for every 100 million miles flown. ... Ten years later the number of passengers had more than tripled, there were 33 crashes, and the rate in terms of miles had nearly halved. ... The world's air travellers now number 300 million annually and each can reckon to fly 125 million miles before the statistics catch up with him."

Discussion.

The incidence of fatal accidents is compared at roughly ten-year intervals, but the results are expressed in a different form each time:

1950: Three passengers killed per 100 million miles flown.

1960: The rate (presumably of deaths) in terms of miles had nearly halved.

1971: Each air traveller "can reckon to fly 125 million miles before the statistics catch up with him."

The comparisons can be clarified by expressing each result in the same terms, giving approximately:

1950: 1 fatality per 30 million miles flown,
1960: 1 fatality per 60 million miles flown,
1971: 1 fatality per 120 million miles flown.

The results are simple: The fatality rate per mile has approximately halved every 10 years or so. (Similarly, the number of passengers has roughly *tripled* every 10 years.) Now one can add stylistic gloss to make for easier reading.

Exercise 1F. TakingOne'sOwnMedicine

Few statistical writers discuss the virtues of working with rounded figures, but Robert Golde (1966a) has stressed the small loss in accuracy of simply "dropping-off" digits. He used the following example for the number 21,742:

Digits Dropped	New Number	Approximate Loss in Accuracy
2	21,740	0.01%
42	21,700	0.19%
742	21,000	3.41%

Discuss the example.

Discussion.

The accuracy lost in "dropping-off" the last three digits of the number 21,742 could have been reported as 3 % instead of 3.41%! Neither the second nor the first decimal can matter. No one would *not* drop the digits if the loss were 3.5 %, but do so if the loss were only 3.4%.

If Golde had used the normal procedure of rounding to the nearest number, 22,000 (instead of merely dropping off the last digits to 21,000), the error would have been about 250 out of 22,000, or only 1%:

Digits Dropped	New Number	Approximate Loss in Accuracy
2	21,740	.01%
42	21,700	.2%
742	22,000	1%

Although the lay-out in the last column above is unorthodox, it can be effective. Certainly the visual obfuscation of writing numbers less than 1 with an initial zero (e.g. 0.01) can often be avoided.

Exercise 1G. What's Wrong?

Golde has quoted elsewhere (1966b) the company president who said:

> "The first few times my controller sent me a report with figures rounded to the nearest $1,000 I felt very uneasy. The report seemed lacking and incomplete. Now that I have grown used to the shorter figures, I find I mentally round all figures I look at down to two or three digits."

Why did the company president feel uncomfortable?

Discussion.

Rounded figures like 95 in the North and 45 in the South seem naked because it is easy to see what they say (the figures in the North are about twice those in the South) but it is not obvious what we ought to do about it. In contrast, figures like 97.63 and 48.29 are comforting because although we may not know what they *mean*, they are at least very precise.

But why should anyone like the company president still be faced with figures which need to be rounded mentally? Figures ought to be presented in suitable form in the first place, consisting only of what matters. The reader's role is to return undigested figures to their originator with rude comments, rather than do his arithmetic for him.

Exercise 1H. Company Reports

Below are given the basic figures in a company's Annual Report, printed in "big, easily-read type and organised to highlight the guts of the business" (Foy, 1973).

But the data are still not easy to take in. How much higher are the 1972 sales than those the year before, and how does this increase compare with those in the other figures?

1972 AT A GLANCE

	1972	1971
Sales..	$172,045,539	$153,220,890
Operating profit.	$11,612,434	$ 8,790,576
As percent of sales.	6.7%	5.7%
Net income.	$ 6,009,155	$ 4,248,645
As percent of sales.	3.5%	2.8%
As percent of investment.	4.8%	3.5%
Earnings per share of common stock.	$1.87	$1.32
Dividends paid per share of common stock	$1.03	$1.00
Taxes paid per share of common stock.	$6.23	$5.24
Market price range of stock during year.	$29\frac{7}{8}$–$17\frac{1}{8}$	$23\frac{1}{2}$–$14\frac{1}{4}$

Discussion.

Rounding helps as usual (including dropping the stock-market's archaic habit of quoting price-ranges in eighths of a dollar). Table I.13 begins to

make some mental arithmetic on the figures possible (e.g. one can see that sales have gone up by about 20 million, which is over 10%, but that operating profits and the range of stock-prices have gone up by over 20 %).

The new table is smaller, but its lay-out generally helps to guide the reader's eye. It is not spread across the page, the rounding allows figures which are to be compared to be close to each other, single spacing is used to indicate which item belongs to which, "ditto" marks indicate the similarity of the "per share" ratios, and the % symbol triggers visual recognition.

Has some of the rounding perhaps been overdone? For an "At a glance" table, showing operating profit as growing from 6% to 7% tells a good story well. (Anyone who wants the percentages to a further place, 5.9% and 6.8 %, can work them out from the dollar figures given.)

TABLE 1.13 An Improved Format

	1971	1972
Sales ($'000)	150,000	170,000
Operating Profit ($'000)	8,800	11,600
As % of sales	6	7
Net Income ($'000)	4,200	6,000
As % of sales	3	4
As % of investment	4	5
Earnings per share	$1.30	$1.90
Dividends " "	$1.00	$1.03
Taxes " "	$5.20	$6.30
Range of stock price	$14-23	$17-29

Nobody is saying that $2 or 3 million, e.g. the difference between $172 million and $170 million, isn't money. It ought therefore to be noted down somewhere, for the record (so that nobody makes off with it). But given a $20 million or so increase in sales, $2 million more or less does not affect any *conclusions* to be drawn from these figures.

The real problem is that two years' figures are not enough anyway. Was I972 an exceptionally high year or 197 1 exceptionally low? To see the full picture, a longer series of data is needed. Adequate digestion and clear presentation of much more extensive data will then be even more necessary.

Exercise 11. Further Examples

From the viewpoint of this chapter examine the last memorandum involving numerical data which you prepared and/or received. Also examine the last scientific paper, technical report, or financial or technical article involving numbers which you read. Is there room for improvement? Would it be easy to achieve? Would it facilitate understanding and communication of the information?

CHAPTER 2

Using Prior Knowledge

Now that we know the results of the data analysed in Chapter 1, we can analyse any additional data for that variable with some prior knowledge.

The notion of using prior knowledge in analysing data is a fundamental one and runs through the remainder of this book. Since our aim is to provide information for future use, the use of prior information should be the common situation. The reason it may not always seem so is that previous data are often left in an undigested, and hence unusable, state.

2.1 New **Data and Prior Knowledge**

The new data to be analysed here are the readings for 1970 which were given earlier in Table 1.1. They are now set out in Table 2.1 with the lay-out and manner developed in Chapter 1.

TABLE 2.1 The 1970 Data in the Same Layout as the 1969 Data (Table 1.5)

1970	Area				
	No	Ea	We	So	Av.
Q I	96	74	53	46	67
Q II	94	77	49	50	68
Q III	91	72	59	42	66
Q IV	98	76	53	37	66
Average	95	75	54	44	67

If we remember the 1969 results, the new results follow quickly.

(i) The 1970 area averages are virtually the same as for 1949: 95, 75, 53, and 45 (as is also demonstrated in Table 2.1a below).

(ii) There are no systematic quarterly differences, just as in 1969.

(iii) The deviations of the individual quarterly figures from the area averages are about 3 units ,on average, again as in 1969.

(iv) There are no exceptionally high figures in Quarter III in the East and West, or at any point in 1970.

TABLE *2.* **1a The** *1969* **and** *1970* **Quarterly Averages for each Area**

	No	Ea	We	So	Av.
1969	*95*	75*	53*	45	67*
1970	95	75	54	44	67

* Excluding QIII in 1969

The new results are accompanied by more conviction just because they are largely the same as in 1969. This analysis has also been much easier to do than the original one because we already know something about this kind of data. This is an advantage of being something of an expert.

It might seem that the analysis was easier only because there is so much agreement between the two sets of data. But it was equally easy to establish that the two exceptionally high values in the East and West in QIII of 1969 did *not* recur in 1970. (It was therefore a good decision to treat them separately in the last chapter. But it would have been a good decision even if the exceptions *had* recurred, because then they clearly would have to be treated separately !) Furthermore, it is also easy to establish that hardly any of the other smaller deviations in 1969 occur again in 1970, as we shall now see.

2.2 Irregular Deviations

Table 2.2 sets out the deviations of the 1970 quarterly figures from the area averages (i.e. 96 − 95 = 1 for QI in the North, etc.).

TABLE **2.2** 'Deviations in 1970 **from the Area Averages**

1970	Area				
	No	Ea	We	So	Av.
Q I					
Q II	•11	-12	•1·5	26	00
Q III	-4	-3	5	-2	-1
Q IV	3	1	-1	-7	-1
Average	0	0	0	0	0

There is no systematic quarter-by-quarter variation and of necessity the area averages are all zero. Therefore only the individual deviations need to be scrutinised further. At first sight these appear to be mainly irregular. However, some of the individual deviations may recur from year to year and be regular in that sense.

For the 1969 deviations (repeated here in Table 2.2a), we recall that they were rather high in QIII and consistently low in QIV-possible seasonal trends-and that there also seemed to be a + − + − pattern, or negative "serial correlation". A glance at the 1970 deviations in the previous table shows that none of this recurs. In the 1970 data the only apparent regularities are that the QI deviations are all small and that the *East* column has a − + − + pattern, and neither of these things occurred in 1969.

TABLE **2.2a** The 1969 Deviation6 (Table **1.6a)**

1969	Area				Av.
	No	Ea	We	So	
Q I	3	0	-3	3	1
Q II	-3	0	4	-3	0
Q III	6	(25)	(27)	5	6*
Q IV	-5	-1	-2	-6	-4
Average	0	0*	0*	0	0*

* Excluding QIII

To check whether other "local sub-patterns" are consistent from one year to another, we can average the deviations over the two years. Two deviations which are inconsistent (one negative and one positive) tend then to cancel and produce a relatively small average. In contrast, consistent results reinforce each other and show up even more strongly. (A common alternative form of this kind of analysis is to average the original readings over the two years

TABLE **2.2b** The 1969 and 1970 **Deviations** Averaged

(From Tables 2.2 and 2.2a)

Average of '69 and '70	Area				Av.
	No	Ea	We	So	
Q I	2	0	-2	2	1
Q II	-2	1	0	2	0
Q III	2	-3*	5*	2	1
Q IV	-1	0	-1	-6	-2
Average	0	0	0	0	0

* Excluding QIII '69

and then to examine *their* variability.) Table 2.2b shows the average deviations over the two years.

Only two of the readings are more than 3 and the overall average has dropped to 2 from the yearly averages of 3. This shows that most of the deviations were inconsistent. However, the reading of -6 for QIV in the South stands out strikingly as a consistent result. With this as the sole exception, the analysis has shown that the residual deviations from the basic model are irregular-not only from quarter to quarter and area to area, but also from year to year. It is this *irregularity* of the readings that appears to be generalisable.

2.3 Empirical Generalisation

The analyses of the data for 1969 and 1970 have led to the empirical generalisation that the quarterly readings are generally about 95 in the North, 75 in the East, 53 in the West, and 44 in the South, that the deviations from this model are apparently irregular, and that these deviations average about $\pm$ 3 ("plus or minus 3").

This model is now known to hold under a fairly wide range of circumstances.

(i) In each of two years (1969 and 1970), despite all the other things that varied from one year to the other.
(ii) In the different quarters of the year, irrespective of the season, etc.
(iii) For different numerical values of the variable ranging from 39 in the South to 101 in the North.
(iv) Despite the large and unrepeated 1969 deviation of about 25 units in Quarter III in the East and West.

The resulting generalisation is limited only by the range of conditions covered. In order to extend it, all we have to establish is whether the same model of 95 $\pm$ 3 in the North, 75 $\pm$ 3 in the East, etc. also holds for data in other years and other countries, and for other related variables. If it does not, we must determine what sort of generalisable differences there are. But we no longer need to tackle each new set of data from the beginning. Use of prior knowledge avoids having to re-invent the wheel every time.

We can also use these results for prediction, e.g. that the 1969/70 patterns will hold for other years. If the prediction is successful, we have a further generalisation covering the new conditions as well.

2.4 The Beginnings of Explanation

The results so far have shown which factors do not matter, such as the seasons of the year, the year itself, and so on. Such conclusions about things

that are *not* related are very simple and powerful. But we also need to explain the main variation in the data, i.e. the systematic differences among the four area results 95, 75, 53, and 44. Why is it 95 in the North and only 44 in the South?

To determine this we need to relate the area results (which refer to millions of units) to some other characteristic of each area. One possibility that springs to mind is the *size* of each area, e.g. the number of potential users of Product X, or the number of employed. Perhaps the results in the North are higher than in the South because there are more people in the North. Table 2.3 shows that this is so. Given the size of the populations in the four areas, about 30, 25, 20, and 15 millions, we can see that the average incidence is about 3 units.

TABLE 2.3 The Area Averages and Population Size

In millions	Area				
	No	Ea	We	So	Av.
Average Quarter	95	75*	53*	44	67
Population	30	25	20	15	22
Incidence per capita	3	3	3	3	3

* Excluding QIII '69

We now have a very simple summary statement: 3 per capita. This relationship with population size accounts for the area differences. It also accounts for the relative stability of the quarter-by-quarter figures, given that populations are generally stable over time.

2.5 **Summary**

In this chapter, the data for 1970 have been analysed against the background of the 1969 results obtained in Chapter 1.

The analysis illustrates two kinds of results which can stem from this use of prior knowledge :

(i) Further generalisation of previous results : here the recurring lack of systematic quarterly differences, the same area averages, and same size quarterly deviations from the area averages, at about ± 3.

(ii) Contradiction of previous results: the two exceptional 1969 QIII results do not recur in 1970, and the deviations of the individual quarters do not generally recur in 1970. (Instead it is the *irregularity* of the deviations that generalises.)

Combining the results for 1970 with those of 1969 has led to increased understanding and conviction. This is particularly so for the factors which

we have learnt do *not* affect the results, such as the season of the year, the year itself, and anything else that changed from one year to the other.

The analysis of the 1970 data has also been much simpler to perform than the "first-time" analysis of the 1969 data in Chapter 1. This is because we already knew the 1969 results. It is one advantage of being something of an expert.

CHAPTER 2 EXERCISES

Exercise 2A. Predicting from Two Readings

The results for QIII in the East were 100 in 1969 and 72 in 1970. Discuss what prediction can be made about QIII in the East for 1971 solely from these two readings.

Discussion.

Because the readings are so discrepant and there are only two of them there is no prediction that one can confidently make. For example, the 1969 and 1970 readings are both "100 to the nearest 100", so that one might predict that the 1971 result should also be "about 100". But equally one could predict

about 81 (the average of 100 and 72),
about 40 (a drop of 30 units per year),
about 150 (based on the notion of some increasing cyclical up-and-down movement over the years).

One's confidence or "strength of belief" in a prediction can be assessed by imagining its outcome. Suppose a reading of 150 were actually observed in 1971. This could not be interpreted as a major discrepancy because the two initial readings were not enough to establish a pattern in the first place. Similarly, a reading of 81 could not be judged to be "as expected" even though it is the average of the two previous results; neither of the given readings lies near this average, so there is no firm reason why the third reading should.

Now suppose we had a third reading to start with, e.g. for 1968. This would begin to provide more predictive information. Three examples are:

	Given Data			
	1968	1969	1970	Prediction for 1971
Case (a)	150	100	72	About 50
Case (b)	79	100	72	About 85 $\pm$ 15
Case (c)	20	100	72	Between 0 and 150

Any 1971 outcome markedly different from that predicted, for example 150 in Case (a) or 500 in Case (c), would seem discrepant with the apparent pattern of the results. But three readings are still a thin base for prediction.

We can see this because *with hindsight,* a sequence of 150, 100, 72 and 150 in Case (a) and even 20, 100, 72, 500 in Case (c), would not appear at all impossible.

Yet another reading, say for 1971, could strengthen the picture:

	Given Data				
	1968	1969	1970	1971	Prediction for 1972
Case (a)	150	100	*72*	*50*	About 30
Case (b)	79	100	72	78	About 82 $\pm$ 15
Case (c)	20	100	72	120	Between 0 and 150, say

If we *now* had an outcome of 150 in Case (a), it would look peculiar even with hindsight; something different has occurred to change the pattern.

Exercise2B. MorePriorKnowledge

What prediction can be made about QIII in the East if the 1969 and 1970 readings in all areas and all quarters are taken into account?

Discussion.

There is no trend in the general pattern of 1969-70 results. Instead there is an irregular quarterly scatter of ± 3 about each area average, which in the East was 75. The 1969 QIII reading of 100 in the East was an exception.

Therefore for any other year we predict QIII in the East will be 75 ± 3. (Any result well outside these limits would be discrepant with the directly relevant data for the East and with the pattern in all the other areas as well.)

Although the only "normal" reading for QIII in the East was 72 this is not the best prediction for QIII. If the QIII Eastern values in other years were generally about 72, then QIII in the East would be consistently lower than the three other quarters. This would be discrepant with all the other available findings for the other areas. Hence the yearly average of 75 for the East is the appropriate prediction even in QIII.

Exercise2C. *BusinessWeek's* "LatestWeek"

In Exercise 1.C we discussed *Business Week's* report of the latest week's production data, accompanied by four earlier figures as interpretative background. The magazine also gave the following commentary:

> "General Motors scheduled only light overtime while Ford shut down operations at four assembly plants. Small declines occurred in steel and electricity. Crude oil refinery runs rose slightly."

The figures are shown here in the layout used in the earlier exercise. How can the results given in the week of April 24 be interpreted?

TABLE 2.4 Business Week's Figures, April 24 1970

	Paper	Steel	Coal	Electr.	Autos	Oil
1957-59 average	28	19	14	12	10	8
Year ago	52	27	20	27	14	11
Month ago	51	28	21	30	19	11
Week ago	52	29	21	29	15	11
Latest week	50	29	22	28	15	11

Discussion.

These figures cannot be interpreted without additional information such as:

the "normal" week-by-week changes that occur,
whether the single week's figure "a year ago" was normal,
whether special events account for any of the *earlier* figures, e.g. the very high auto figure of 19 "a month ago",
whether the economy is sensitive to weekly changes such as a 2% increase in oil production or a 1% drop in steel, and if so, in what way.

Exercise 2D. More *Business Week* Data

From back issues of *Business Week*, a continuous sequence of weekly figures can be reconstructed, as illustrated in Table 2.5 for the first sixteen weeks of 1971. How does this additional information help in interpreting the latest week's figures?

TABLE 2.5 Week-by-Week Figures in 1971

1971 Week ending		Production (per week) Paper	Electr.	Steel	Coal	Autos	Oil
January	2	24	29	24	20	6	11
	9	43	31	24	20	17	11
	16	50	31	25	19	19	11
	23	51	32	26	21	19	11
	30	50	31	27	20	18	11
February	6	50	32	26	20	19	11
	13	50	32	27	18	18	11
	20	51	30	27	18	20	11
	27	51	30	28	19	19	11
March	6	52	30	28	20	18	11
	13	51	30	28	20	19	11
	20	51	30	28	21	19	11
	27	51	30	29	21	19	11
April	3	52	29	29	22	18	11
	10	52	29	29	21	15	11
"Latest"	17	50	28	29	22	15	11
Average		49	30	27	20	17	11

Discussion

The new table firstly helps us to judge what is normal. We can now see that the "Month Ago" Automobile figure of 19 in Table 2.4 was not exceptionally high (as was thought in Exercise 1.C). Instead, Table 2.5 shows that this figure (March 27) is normal and that it is the other three figures in Table 2.4 which were unusually low. (Presumably there were short working-weeks for Easter in the weeks of April 10 and 17, 1971, and Easter "a year ago".) Similarly, the Paper and Automobile figures for the week of January 2 are seen to be low and seem to be tied with Christmas working arrangements in these two industries (this also occurred in the previous year's figures).

Next we can try to see whether this fuller background can help the businessman judge the current economic climate from the latest week's figures ("Shall I invest in the new plant this week, next week, sometime, never?"). We consider three industries with different properties in the early part of 1971: Steel (which shows a steady upward trend), Automobiles (which fluctuate), and Oil (which does not vary).

Steel: the steady upward trend in 1971 is new-it did not occur the year before. The trend however does not predict anything similar for the other five industries, because none of *them* show an upward trend. The figures therefore show that more steel is being produced, but we are given no back-log of other such cases to indicate what else this implies. (It might also be that the latest figure itself signals a levelling-off, but if so, we do not know what *that* would imply.)

Automobiles: the figures fluctuate and show no trend. Suitably adjusted to account for fluctuations due to short working-weeks, strikes, inventory adjustments, heavy overtime coming on, heavy overtime being eased off, and other factors annotated in *Business Week*, the 1971 figures might well be steady. Therefore, it is not clear what the latest weekly figures tell us.

Oil: the figures to two digits do not vary in 1971 (nor did they do so in 1970). Thus, the latest week's figures tell us nothing new. (The figures might appear to have been unduly rounded, since there is *no* variation at all, but to one decimal place they still show no trend: *e.g.* the five 4-week averages for Table 2.5 are 10.8, 11.0, 11.1, 10.8 and 10.9.)

It follows that, even in the context of additional data, the latest week's figures say nothing very clear about the general economic climate. Changes of 2% up or 3% down in one week as commented on by *Business Week*, "small declines occurred in steel and electricity and crude oil refinery runs rose slightly", therefore seem uninterpretable. Showing the figures to 3 or more digits is pointless and even the oil figures have probably not been rounded too much.

The data are too complicated for a layman to interpret immediately. But it should not be too difficult for a professional analyst to derive some understanding to pass on to the reader. For example, the steady rise in steel production in the first four months of 1971 may have some long-term "lagged" implications. Looking at previous years and other countries, a skilled analyst might for example establish that such an increase, unaccompanied by similar increases in other industries, generally represents something like a build-up in stocks preparing for an expected stoppage during the wage-negotiations later in the year.

If, however, skilled economists have been unable to determine any such generalisable patterns, then it is useless to expect the lay reader of *Business Week* to discover them himself.

Exercise 2E. Prior Information in Practice

Examine how prior information was used in the last two memoranda involving numerical data that you either received or reported, and the last two scientific papers or technical reports that you read. Did they make any detailed numerical use of previous results? If so, did it lead to increased confidence and understanding? If not, how would it have helped?

Exercise 2F. Rounding to Two Digits

Most people would be fairly happy to round readings to *three* significant digits, but baulk at *two.* People fear that too much information might be lost. Is the difference important?

Discussion.

Unfortunately, the difference between two and three digits is a big one. Most of us can do mental arithmetic with two-digit numbers, and we can memorise them. But we cannot do this with three-digit numbers (except by first mentally rounding them down to two !).

A simple example is percentages expressed to one decimal place. Few of us can divide 35.2 % by 17.8 % in our heads. But almost all can see at a glance that 35 % is almost twice 18 %. Being able to see patterns, structures and relationships is what really matters in understanding and interpreting data.

What then should we do in practice-should we round to two digits despite our fears?

With data that are being looked at for the very first time (as in the main example of Chapter 1), we will not know beforehand what degree of precision is really required. So we can *record* the data to three or more digits, just in case they might be needed. But the first analysis, to get the *feel* of the data, can best be done with two digits.

After this initial analysis we can then go back and rework the data with three or more digits, now that we know what we are looking for. But, in practice, one finds that this is only rarely called for!

In all other cases, i.e. when there is already prior knowledge of that type of data, we should be able to judge *beforehand* what precision is needed. If, in the past, useful interpretations or conclusions have been made from the variations in the third digit (e.g. 35.2 % as against 35.3 %), then three or more digits are called for. But this is not the case if the third digit is used only because there might be something interesting.

Data can always be recorded or stored somewhere in great precision, in case they are needed for further, more sensitive, analysis and research. But in trying to *communicate* data, a balance has to be struck between the possibility that occasionally something of possible interest is missed through undue rounding and the likelihood of not seeing *anything* in the data if they are presented in too much detail.

Ultimately the decision is one for each user of the data and each analyst to make for himself.

CHAPTER 3

Tables and Graphs

We have seen in the two preceding chapters ways in which extensive data can be reduced to a few summary figures. Tables of data were used for working purposes, but no tables were needed to communicate the final results. The same applies in other situations. Indeed, the more complex the data, the less can the task of understanding the undigested data be left to the reader.

The rule "No report should contain tables-with some exceptions" means that if the analyst ***does*** present a table of extensive data, he should have a specific (i.e. exceptional) purpose in mind. Two such purposes are to illustrate a summary statement or to provide a record of the detailed data for possible ***later*** analysis.

Sometimes graphs are considered easier to understand than tables of numbers. But graphs also require a clear purpose. They are discussed in the later part of this chapter.

3.1 Tables to Illustrate

Communication of key nuinerical results can be aided by ***displaying*** the figures, e.g. as

North 95, East 75, West 53, South 44,

instead of hiding them in the text. Displays of this kind, or even a mini-table like

Average deviation : 3 units

can make the results more memorable and easier to find again later when skimming the pages.

Tables for Proof

Sometimes tables of detailed data are reported for *proof*, since people do not always trust the analyst. Did he add the figures correctly? Has he given

misleading averages? A "bare-bones" description, such as 95 in the North, 75 in the East, etc., can certainly leave room for doubt.

However, the analyst has some alternatives to providing tabular proof in the way he summarises his results.

(a) He can show that he is aware his data contain scatter, and that he knows roughly how big it is, by mentioning that the quarterly figures deviate by an average of about 3 units. (At least he should say, or imply, that the deviations are small.)

(b) He can demonstrate that he has actually looked at the data in some detail by mentioning that the scatter is apparently irregular, and that its size does not vary much from area to area or quarter to quarter.

(c) He can show that he has sorted out the nature of the data and that he is not trying to hide anything by mentioning the two large QIII exceptions. (Reference to large exceptions-or to the *absence of* any special features-is also a quick way of implying the previous steps without having to spell them out explicitly.)

All this can then be usefully illustrated by a clear and well-digested tabular *extract* of the available data, here for *one* year as in Table 3.1, not for *both* years. This makes the summary results easier to comprehend and accept. It shows what kinds of averages are referred to, and the reader can mentally note how the quarterly readings do in fact differ from these by something like the stated 3 units (i.e. 3, – 3, 6, – 5 in the North; 0, 0, – 1 in the East and so on).

TABLE 3.1 The 1969 Area Results Quarter-by-Quarter

1969	Area				
	No	Ea	We	So	Av.
Q I	98	75	50	48	68
Q II	92	75	57	42	67
Q III	101	(100)	(80)	50	(83)
Q IV	90	74	51	39	64
Average	95	75*	53*	45	67*

* Excluding QIII

Detailed data "for proof" are usually requested only when very simple analyses are presented (e.g. averages and percentages). Few readers expect to be shown the full raw data when *complex* analysis techniques have been used, but the reader should still want some assurance that the analyst himself has looked at the data. This is not provided by simply reporting all the facts. Adequate forms of summary reporting and illustration therefore become particularly important.

Tables to Dramatise

Some repetition or redundancy generally improves communication. The fact that the quarterly figures in our example varied relatively little is brought out in Table 3.2.

TABLE 3.2 The Fit of the Theoretical Model
(Observed and Theoretical Values)

	Area				
	No	Ea	We	so	Av.
	Ob Th	Ob Th	Ob Th	Ob Th	Ob Th
1969 Q I	98 95	75 75	50 53	48 44	68 67
Q II	92 95	75 75	57 53	42 44	67 67
Q III	101 95	100 75	80 53	50 44	83 67
Q IV	90 95	74 75	51 53	39 44	64 67
1970 Q I	96 95	74 75	53 53	46 44	67 67
Q II	94 95	77 75	49 53	50 44	68 67
Q III	91 95	72 75	59 53	42 44	66 67
Q IV	98 95	76 75	53 53	37 44	66 67
Average	95 95	75* 75	53* 53	44 44	67* 67

* Excluding QIII in 1969

Here the area averages are explicitly promoted to the status of a theoretical model and shown alongside the observed figures. This makes the general similarity of the observed quarterly figures more apparent. Such a demonstration of the agreement between observed and theoretical figures becomes even more important when the theoretical figures themselves vary, as many cases later in this book will illustrate.

Dramatisation of this kind requires a well laid-out table. It need be done only on a small illustrative basis. Graphs can also be effective here, as long as they present a single, clear-cut story-line.

What Makes a Good Table?

In Table 3.2 the reader still has to work out the differences between the observed and theoretical figures (e.g. $98 - 95 = 3, 92 - 95 = -3$, etc.). This appears to negate the "Don't leave it to the reader" precept. The reader's task might seem easier if the differences between each pair of figures were given *explicitly, e.g.* as the "D" values 3, -3, 6, etc. in Table 3.3.

However, this table is far too complex. If the intention were to give the reader an easy view of the "difference" figures, making his eye, jump three columns to compare like with like is not much help. Placing different kinds

TABLE *3.3* A Table showing the Observed and Theoretical Results (Ob and Th) and the Differences (D) between Them

1969	Area														
	No			Ea			We			So			Av.		
	Ob	Th	D	Ob	Th	D	Ob	Th	D	Ob	Th	D	Ob	Th	D
Q I	98	95	3	75	75	0	50	53	-3	48	45	3	68	67	1
Q II	92	95	-3	75	75	0	57	53	4	42	45	-3	67	67	0
Q III	101	95	6	(100)	75	*	(80)	53	*	50	45	5	(83)	67	*
Q IV	90	95	-5	74	75	-1	51	53	-2	39	45	-6	64	67	-3
Av.	95	95	0	75*	75	0	53*	53	0	45	45	0	67*	67	0

* Excluding QIII

of figures next to each other, and separating ones that need comparison, should generally be avoided.

The golden rule is that a table is good if in looking at it the *next* steps are easy to do. In Table 3.3 this is not so; there is nothing simple that can be done with it. In contrast, the mental arithmetic needed in reading Table 3.2 is easy because the observed and theoretical readings are adjacent and only one or two digits are involved. For example, it is simple to make a visual check of the summary statement that the differences are irregular and average about 3. Furthermore, actually working out these differences and running the eye down the North column to see that 3, -3, 6, -5, 1, — 1, etc. *do* average at about 3 (as stated in the text) is a good learning process. It helps the reader to assimilate the information more deeply.

In more complex situations the differences between the observed and theoretical figures might have to be shown explicitly. Then a separate display, as in Table 3.3a, tends to be best. Here again it is easy for the reader to take the next step mentally, like seeing that the QIV values are all negative, and those in the East all small.

TABLE 3.3a The Differences between the Observed and Theoretical Figures

1969	Area				
	No	Ea	We	So	Av.
Q I					
Q II	-33	0	-3	-33	0
		0	4		
QIII	6	(25)	(27)	5	6
QIV	-5	-1	-2	-6	-4
Average	0	0	0	0	0*

*Average size ignoring sign = 3

3.2 Tables for the Record

Instead of presenting data to aid immediate communication, many tables set out data for possible subsequent use. Extreme examples are census reports and other official statistics where data are published which nobody has as yet attempted to analyse. The data are provided for the record, so that anybody can refer to them at some future stage.

Storage of undigested data is becoming even more common nowadays with the development of computerised data banks and information systems. The idea is to provide "all the facts at the touch of a button". But this can be overdone. Great sums are invested in elaborate methods of data-retrieval, even though no one knew what to do with the data before they were lost in the first place. (A fraction of this expenditure devoted to some ***analysis*** of the data would often yield results that could be stored along more old-fashioned lines, in filing-cabinets or books, or that could even simply be *remembered.*)

Data for Reanalysis and Research

Less extreme examples of data storage are cases where the data ***have*** already been analysed and summarised but the full details are provided in case somebody wants to reanalyse them, usually for research purposes. In this situation, the data need to be given in an Appendix at most, or copies of the data might be kept and made available to the occasional enquirer. They need not be reproduced in every report.

Data One Does Not Understand

Many published tables of data are ones the author did not understand. He is in effect saying to the reader "Here are some data. I do not understand them. Perhaps you can." This is on the whole not the way to address either one's colleagues or the public at large. If the ***author*** has not been able to summarise and communicate the main patterns of the data, what chance is there for anyone else to do so?

However, there are situations where this seems a legitimate approach. Table 3.3a of the quarterly deviations was given deliberately to emphasise that one could not see any pattern. The readings ***appeared*** irregular, but perhaps somebody else could make something of them?

Another example of undigested data given legitimately was the presentation of the summary figures:

North 95, East 75, West 53, South 44.

Here, one was saying that one does not know why these figures differ from each other. But at least they represent an effective summary of much more

extensive data, and the potential pay-off is there: somebody might suggest checking the size of each area. As noted in Chapter 2, this leads to the simple result of 3 units per capita. In this case, reporting data one does not understand led to progress. However, this usually occurs only when the figures are already well-digested summary results rather than the raw, and unorganised, original observations.

3.3 Graph

Graphs have two uses: for working-purposes and for presenting final results. In the first case, the analyst may plot the readings in the early stages of the analysis to gain a quick visual impression of whether a relationship is linear, as in Figure 3.1A, and whether the scatter is "homoscedastic" (equal scatter all along the line) or heteroscedastic, as in Figure 3.1B. Such graphs are for private use. They may be rough (drawn on the back of an envelope) and will be superseded by a proper analysis of the data, leading to summary figures or models.

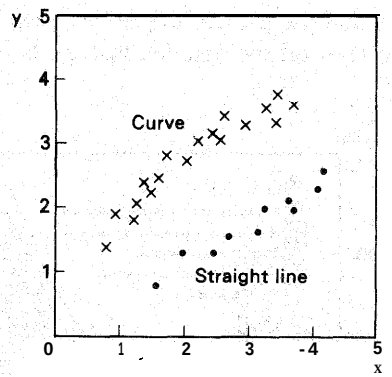

Figure 3.1A Plotting *y* against x

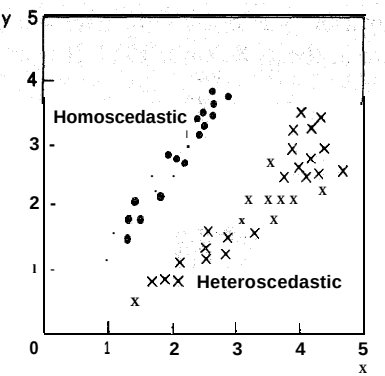

Figure 3".1B Constant or Increasing Scatter

The use of graphs when presenting final results is either to simplify or enliven them: in a word, to make the results more *graphic.* To achieve this both the limitations and the strengths of graphs need to be understood.

Figures 3.2A and 3.2B give monthly temperature and rainfall results for England and Wales from Facts *in Focus* (1972), a handbook compiled by the U.K. Central Statistical Office. It is clear that it is hotter in the summer and rains less in the spring. What else do the graphs tell us quickly? That the 1970 temperatures were mostly close to normal, but the 1970 monthly rainfall figures differed a good deal from the longer-term averages'? Is *that* the kind of thing the chartist meant us to see? If so, should it be up to the *reader*

to note that although the November rainfall was exceptionally heavy, this may have been more than made up for" by the relatively dry months since May?

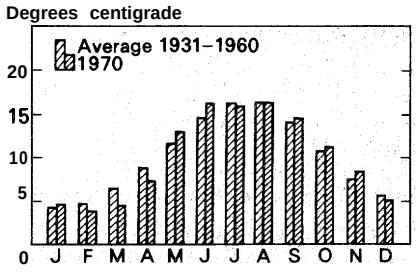

Figure 3.2A Monthly Temperatures

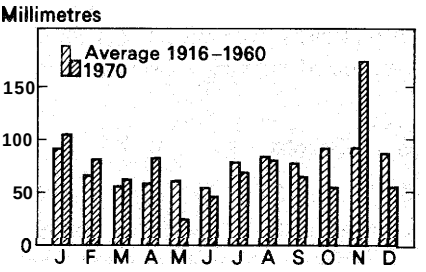

Figure 3.2B Monthly Rainfall

A graph must have a well-digested message. If the analyst does not know what the figures are saying, then this will be the message he communicates to the reader.

Two further examples are Figures 3.3A and 3.3B, giving the age-breakdown of men in the U.K. over 4 decades and the growth in public expenditure in the U.K. since 1953. It is hard to read off the detailed figures. Is the

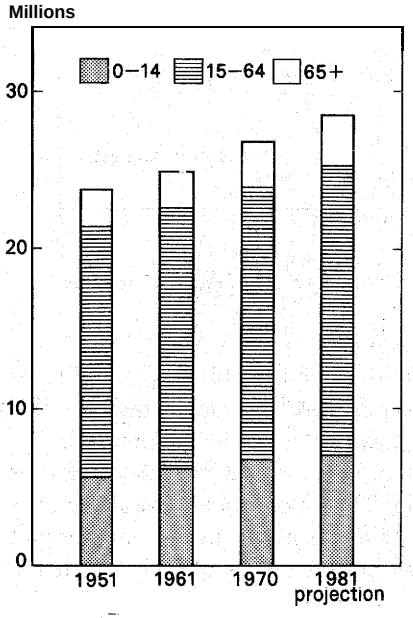

Figure 3.3A The Age-Distribution of Males

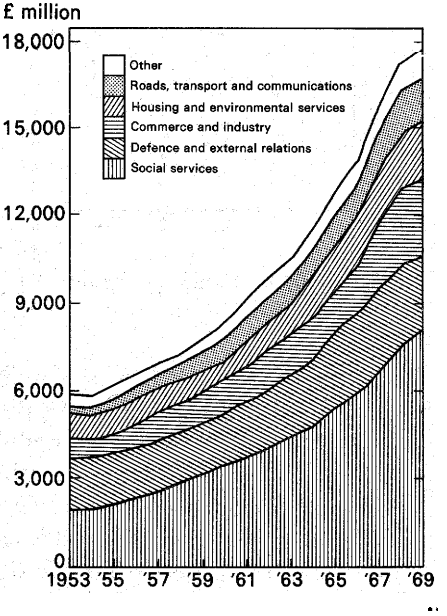

Figure 3.3B Public Expenditure

proportion of old men less than it used to be? Which of the areas of public expenditure have grown more than proportionally, and which less? These graphs illustrate the difficulty of taking in *quantitative* information from a graph. (In the forthcoming revised edition of *Facts in Focus,* Figure 3.3A has been dropped completely and Figure 3.3B replaced by a table plus a chart showing the detailed position in just one particular year.)

In general, graphs can demonstrate a *qualitative* result-e-g. "Public Expenditure has gone up", or "Most people are aged 15 to 64", or "It is hotter in the summer". But graphs cannot communicate *quantitative* results (although it is this quantitative detail which is generally so laborious to plot initially). Even with pictorial devices, such as the population data in Figure 3.4A from the *Pocket Data Book: USA 1971* or the "pie-chart" in Figure 3.4B of the Incidence of Poverty in the U.S., we do not get our information by counting little men or trying to assess the size of the angles. Instead we look at the *numbers.*

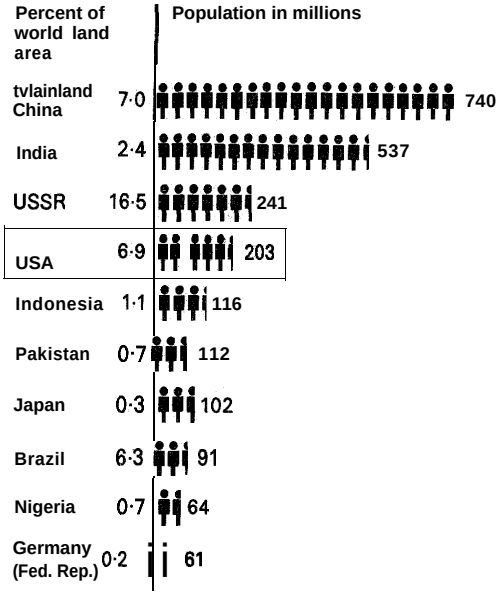

Figure 3.4A The Ten Largest Countries

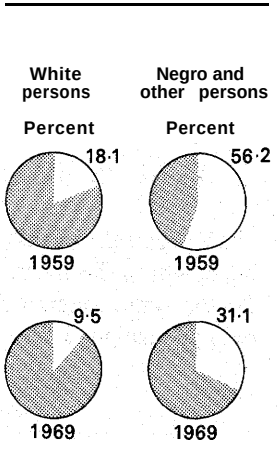

Figure 3.4B The Incidence of Poverty

Graphs are generally useless for detailed analytic purposes, since it is impossible to manipulate any of the readings. In contrast, with a table it is easy to form averages, to take differences, etc. More important still, graphs seldom present a succinct and memorable summary of the results, except at a very broad qualitative level, like "It is hotter in the summer".

The Strengths of Graphs

The kind of story a graph can tell well is that something is *constant.* This is a qualitative property that is easy to take in visually. Thus Figure 3.5A, showing that B is bigger than A, is not a very good graph because it does not show clearly how *much* bigger B is (twice as big, *more* than twice, or *less* than twice?). Figure 3.5B is much more effective because it is clear to the eye that the increase from A to B is about the same as that from B to C, i.e. more or less constant.

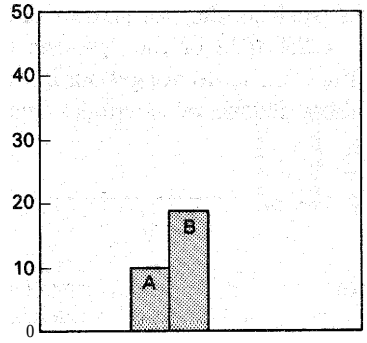

Figure 3.5A How Much Bigger?

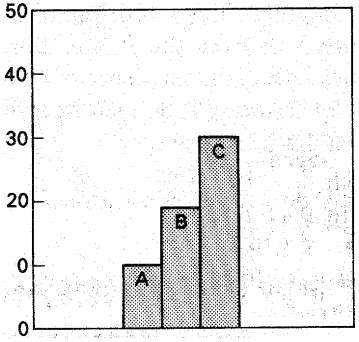

Figure 3.5B Equal Differences

The typical example of a good visual story is a straight line, as in Figure 3.6A, where the increase in *y* for any unit increase in *x* is always the same. By the same token a simple *curve* is also easy to appreciate visually as *not* being a straight line.

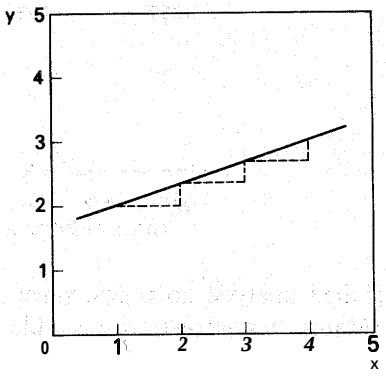

Figure 3.6A Equal Increases

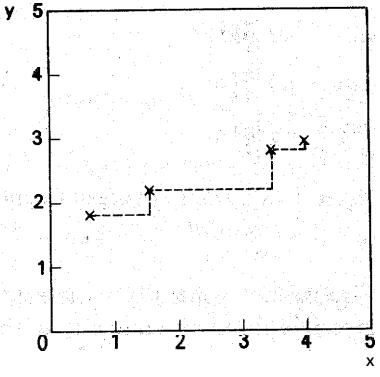

Figure 3.6B Proportional Increases

To be really effective a graph has to show that one variable is more or less constant ***despite another variable varying a great deal.*** The precise amount of variation involved is unimportant-a graph cannot easily communicate that anyway. Figure 3.6B illustrates how this holds for a straight line. It is visually easy to see that for any two points the ratio of the difference in y to *x* is the same, irrespective of how far apart the points are.

It follows that the form of a graph should generally depend on the numerical pattern of the data to be represented, and not on the nature or meaning of the variables as such. For example, in Figure 3.7A, the number of families with high incomes has remained roughly the same over the 30-year period, but the way it is charted does not make this particularly clear. Putting the high income group at the ***bottom*** of each column, as in Figure 3.7B, makes the pattern much more obvious.

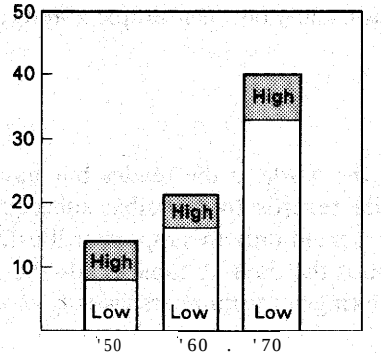

Figure 3.7A High and Low Incomes

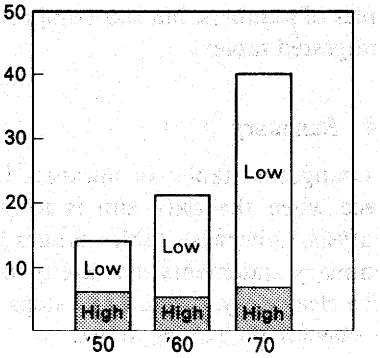

Figure 3.7B Constant Numbers with High Incomes

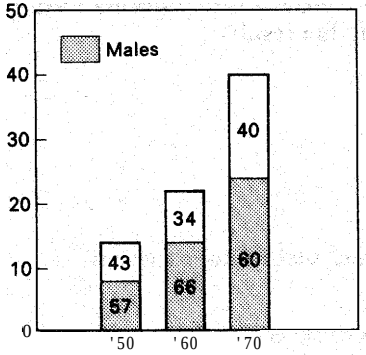

Figure 3.8A Males and Females

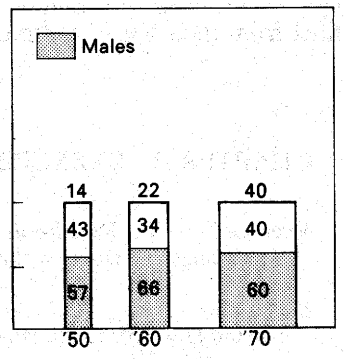

Figure 3.8B Approximately Constant Proportions

In Figure 3.8A it is not the number of men but their *proportion* that has remained fairly constant, at about 60% over the three decades. But neither this nor the degree of variation about 60% is altogether clear from the figure. In contrast, Figure 3.8B makes real use of the *graphic* tool by showing the approximately constant proportion of males as a line that remains more or less at the same level, despite the large increase in the total numbers. Figure 3.8B would tell its story *without* any of the numbers being written in, but these help to make the result more memorable: i.e. that despite the three-fold increase in population from 14 to 40, the proportion of males has stayed at about 60%.

Once we have a clear story, it is no longer so obvious that we need a graph to demonstrate or dramatise it. A few words may do ("about 60%"). If it is not clear what a graph is meant to say, probably more analysis of the data is needed. It is often said that graphs are easier for people to understand than tables of numbers, but this comparison is usually between simple graphs and undigested tables.

3.4 Summary

Undigested tables of numbers leave the work to the reader but have a place when the only aim is to provide records for possible subsequent analysis. Otherwise, tables should be presented only to support or illustrate summary statements that are made about the data. A table needs then to tell a clear story, and the next steps in looking at the numbers visually should be easy for the reader to do.

Graphs also need a simple story-line to communicate. They can be used to highlight a *qualitative* feature of the data. In particular, they excel at showing that one aspect of the data is more or less constant although another varies a great deal.

The use of tables and graphs for the analyst's own working-purposes is distinct from their use in communicating the results.

CHAPTER 3 EXERCISES

Exercise 3A. The Function of Tables

Go through the tables in this book and establish their function.

Discussion.

The tables will be found generally to be either:

records of data for analysis
intermediate working-tables
tables to illustrate results.

The crucial test for the latter purpose is that the tables could be omitted with only slight adjustments to the wording of the text. Their role of aiding communication usually relies on deliberate redundancy.

Exercise 3B. DatafortheRecord

Table 3.4 reproduces Table 1 from the Bureau of the Census's *Pocket Data Book: USA 1971*, a typical table of detailed records. Discuss the nature and role of such a table.

TABLE 3.4 **U.S. Population** and Area from 1970

Year	Resident Population: Number (1,000)	Resident Population: Percent increase over prior year shown	Resident Population: Per square mile of land area	Area (1,000 sq. mi.): Land	Area (1,000 sq. mi.): Water
1790	3,929	(X)	4.5	865	24
1800	5,308	35.1	6.1	865	24
1810	7,240	36.4	4.3	1,682	34
1820	9,638	33.1	5.5	1,749	39
1830	12,866	33.5	7.4	1,749	39
1840	17,069	32.7	9.8	1,749	39
1850	23,192	35.9	7.9	2,940	53
1860	31,443	35.6	10.6	2,970	53
1870	39,818	26.6	13.4	2,970	53
1880	50,156	26.0	16.9	2,970	53
1890	62,948	25.5	21.2	2,970	53
1900	75,995	20.7	25.6	2,970	53
1910	91,972	21.0	31.0	2,970	53
1920	105,711	14.9	35.6	2,969	53
1930	122,775	16.1	41.2	2,977	45
1940	131,669	7.2	44.2	2,977	45
1950	151,326	14.5	42.6	3,552	63
1960	179,323	18.5	50.5	3,541	74
1970	203,185	13.3	57.4	3,541	74

X Not applicable.

Discussion.

The primary purpose of official statistics is to record facts accurately. But not much is gained by recording data to six digits or the like. While it can be argued that very precise figures ought to be available somewhere, a general guide like a *Pocket Data Book* does not seem the right place. The aim here is to create some *understanding* of the data and generally to help the user.

This is recognised in the table by giving two columns of *derived* statistics, the percentage increase and the population per square mile. Each quantity could easily be calculated from the other figures recorded but has been given explicitly to aid the reader. But these derived figures are also given in too much detail to see the pattern easily.

Table 3.4a gives the figures in rounded form. We now see thatsince the founding of the United States,

the population has increased 50-fold (from about 4 million to 200 million),

the area has increased about 4-fold (from almost 900,000 square miles to 3.5 million),

the density per square mile has increased 12-fold (from about 5 to almost 60).

The population increased by about 35% every 10 years until 1860, then by about 25 % until the end of the 19th century, and by about 15% this century, with a low of 7% over the 1930's.

TABLE 3.4a The U.S. Population and Area Data to 1 or 2 Significant Digits

	Resident Population			Area (1,000 sq. mile)	
Year	Millions	% Increase	Per sq. mile	Land	Water
1790	4		5	860	20
1800	5	35	6	860	20
10	7	36	4	1,700	30
20	10	35	6	1,700	40
30	13	34	7	1,700	40
40	17	33	10	1,700	40
50	23	36	8	2,900	50
60	31	36	11	3,000	50
70	40	27	13	3,000	50
80	50	26	17	3,000	50
90	63	26	21	3,000	50
1900	76	21	26	3,000	50
10	92	21	31	3,000	50
20	110	15	36	3,000	50
30	120	16	41	3,000	40
40	130	7	44	3,000	40
50	150	15	43	3,500	60
60	180	18	51	3,500	70
70	200	13	57	3,500	70

These simple overall results are the kind it is important to have well embedded in one's mind when looking at comparable figures for individual states, regions or cities, or for other countries, or at data for other related variables.

If may also be noted that the two columns of derived statistics are now hardly necessary. The main recorded figures are easy enough to use for rough mental arithmetic. For example, anyone can see that between 1890 and 1900, say, the population had increased by about 20% (i.e. 76 − 63 = 13, which is a fifth of 65), and that the density had increased to 76 million persons over 3 million square miles, which is about 25 per square mile.

Of course, something has been lost in the rounding, and the question is whether it matters. Perhaps rounding the water area to just one significant digit was too much. Nonetheless, it is still noticeable from Tab 3.4a that between 1920 and 1930 somebody lost about 10,000 square miles of

U.S. water (or 8,000 to be more exact, from Table 3.4). The fact that the landmass of the States *increased* at the same time by 8,000 square miles can admittedly no longer be discerned in the rounded figures. One's guess is that these changes were due to a minor redefinition of land and water; but it is doubtful whether that is what one should be learning from Table 1 of the U.S. *Pocket Data Book.*

Exercise 3C. An Information System

In dealing with unemployment or balance of payment results, sales or production data, etc., the latest figures are usually examined with great care and the following type of table may be circulated or published.

	Latest Quarter	Previous Quarter	Quarter Year Ago
North	92	98	96
East	72	76	74
West	57	53	53
South	48	37	46

Suppose that these figures report the same kind of data that we discussed in Chapters 1 and 2. What is the value of providing a table of such up-to-date information?

Discussion.

The interpretation of such data is usually left to the reader, although some commentary might be given, e.g. that the latest quarter's results in the North and East are down and those in the West and South are up.

This makes no explicit use of the earlier analysis of previous results, which showed that quarterly figures varied irregularly by about ± 3 around the area averages of 95, 75, 53 and 44. It follows from these results that the latest quarterly figures are well within these usual limits of variation. Since one has never been able to interpret such variations in the past (not even with hindsight well after the event) the new table of figures is literally meaningless. All one need report is that the figures are normal.

Exercise 3D. A Graphical Representation

Later on in this book (in Exercise 10G of Chapter 10) the figures in Table 3.5 are discussed. Without going into the fuller meaning of the data here, develop a graphical representation.

TABLE 3.5 Attitudinal Data from Chapter 10

	% of the Population Using	"Right Taste"		"Convenient"	
		Users of Stated Brand	Non-Users of the Brand	Users of Stated Brand	Non-Users of the Brand
Brand A	50	67	6	19	3
Brand C	30	62	4	55	48
Brand B	10	69	5	17	2
Brand D	5	60	3	17	2

Discussion.

To represent these data graphically, we do not necessarily have to understand their meaning, but we must be able to see the patterns in them.

We can note certain regularities, e.g. that the figures in each "Users of Stated Brand" column are generally similar to each other and much higher than those in the Non-Users column, despite the fact that the numbers of users (and non-users) varies markedly among the brands.

Figure 3.9 brings out these results graphically. It shows which figures are similar and makes a dramatic exception stand out (like "Convenient" for both users and non-users of Brand B), without letting the differing incidence of users distract attention.

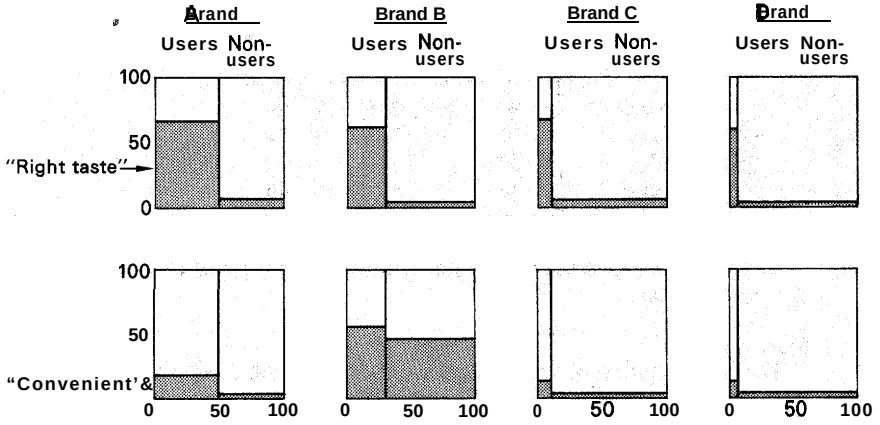

Figure 3.9 Approximately Constant Proportions of Users and Non-Users of a Brand Holding an Attitude, with Exceptions

Exercise 3E. Misleading Graphs

It is commonly alleged that graphs can be used to mislead by not showing the zero-points on the vertical scale. Thus a very small increase, as in Figure 3.10A, appears very large in Figure 3.10B.

Discuss this and some other misleading graphs.

Discussion.

Like most other things, graphs *can* be used to mislead, but anybody who does not look at the scales is likely to misunderstand *any* graph. It would be pointless to plot the data as in Figure 3.10A if the readings never varied by more than a few units from 80.

Changing the units on the *horizontal* scale can also influence one's view of the data, as shown in Figure 3.11A and 3.11B. In Figure 3.11A there appear to be some important increases and decreases, but in Figure 3.11B these same movements are merely irregular errors.

The danger with graphs is not so much misleading others as misleading oneself. Figure 3.12A shows the percentage of illegitimate births over the last century in England and Wales, Northern Ireland, and Scotland (from

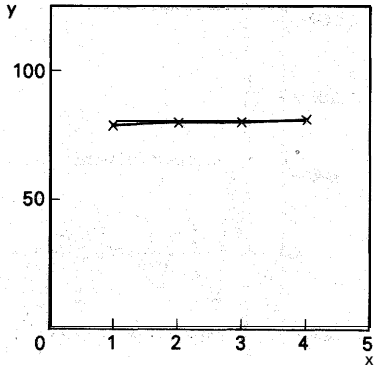

Figure 3.10A The Wrong Scale of *y*

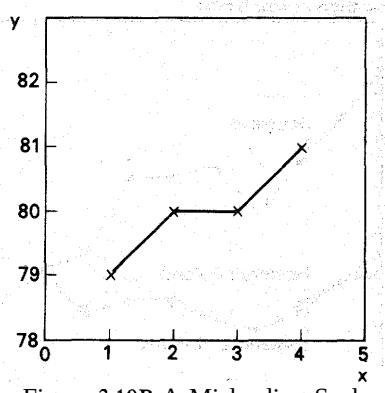

Figure 3.10B A Misleading Scale of y?

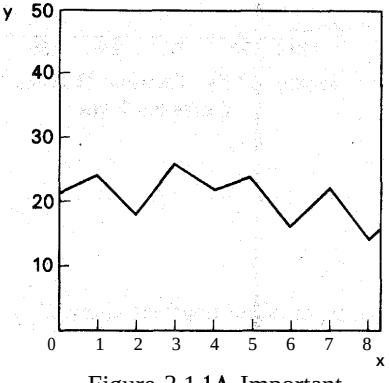

Figure 3.1 1A Important Variations

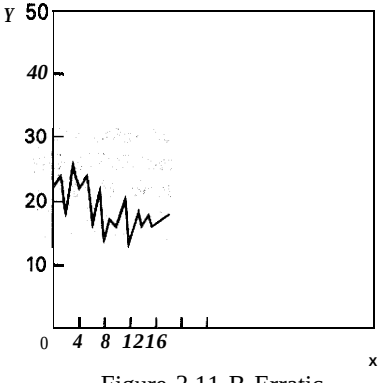

Figure 3.11 B Erratic Fluctuations

Britain in Figures, Sillitoe, 1971). Most people would agree that the graph shows that illegitimacy rates dropped markedly from 1870 until about 1960 and then (in more permissive times?) rose again sharply.

But in England and Wales (by far the biggest of the three regions) the illegitimacy rate hardly declined at all (presumably young Scottish and Irish girls flocked to London during th'30s). Yet somehow the graph does not make this clear.

Figure 3.12B shows two financial indices over a period of 15 years. The visual impression is that the two indices go together. But this is partially misleading. Not only are the later blips in the continuous line far more extreme than those of the broken line (whereas earlier the scales of the two were more similar), but the phasing of the variations is variable. Sometimes the two lines vary together, sometimes the continuous line leads the broken one by a year, sometimes it Iags by a year. Yet one does not immediately "see" this when looking at the graph (and many governments behave as if they had not noticed it either).

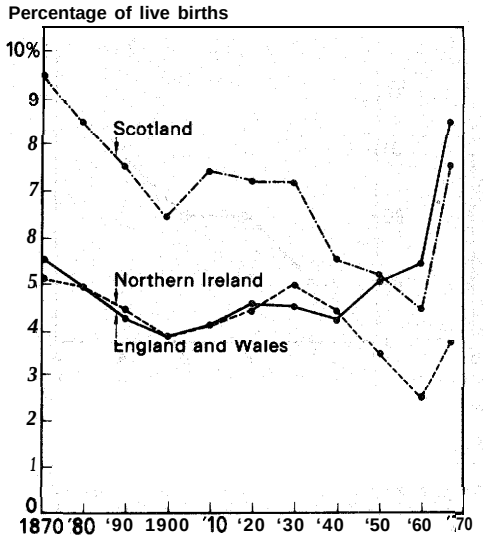

Figure 3.12A Illegitimate Births. (Reproduced with permission from Alan F. Sillitoe, *Britain in Figures, A Handbook of Social Statistics* (Pelican Original 1971). Copyright © Alan F. Sillitoe, 1971.)

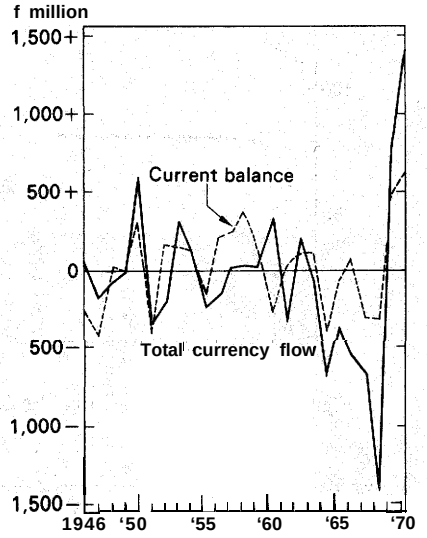

Figure 3.12B Current Balance and Currency Flow

Exercise 3F. Graphsto Dramatise

Figure 3.13A shows time-series for two types of employment statistics *(Facts in Focus,* 1972). It is obvious that when unemployment is high, unfilled vacancies are low, and vice versa. Would it be advantageous to plot the unfilled vacancies as in Figure 3.13B, in effect reversing the scale and plotting a quantity like (800,000 minus Unfilled Vacancies), so that the variations go parallel to those for the figures of unemployed?

Discussion.

Plotting the data in parallel as in Figure 3.13B would help if the aim were to show the quantitative details of the data, e.g. that a particular variation in unemployment is larger or smaller, earlier or later, than the corresponding variation in unfilled vacancies. But a graph will not succeed in communicating such detail until the analyst himself has worked out precisely what the patterns are.

The original Figure 3.13A may in fact be better precisely because it stops us looking at quantitative details, as one is tempted to do in Figure 3.13B, which may obscure the main picture. It also avoids using a rather arbitrary-looking measure like (800,000 minus Unfilled Vacancies). Further, since it is so easy for the reader to work out that the two lines in Figure 3.13A are in broad agreement, this acts as a *learning* device. (The reader is pleased with himself for having worked it out.) Thus for presenting the one point, that unemployment and unfilled vacancies tend to be negatively correlated, Figure 3.13A may be more effective.

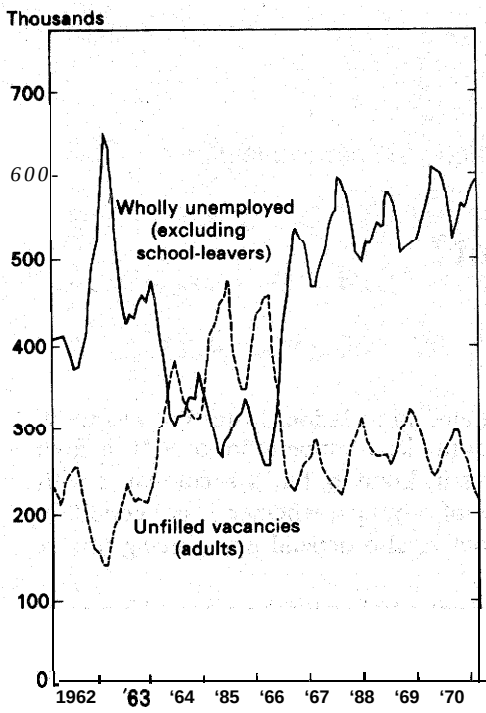

Figure 3.13A Vacancies and Unemployed

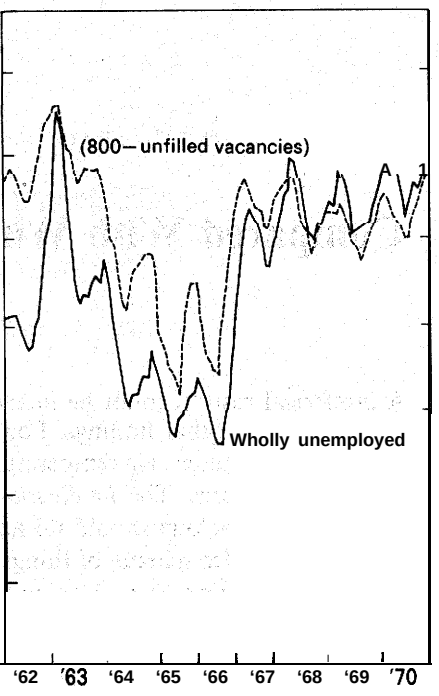

Figure 3.13B (800,000 — Vacancies) and Unemployed

CHAPTER 4

Compared With What?

A numerical result cannot be interpreted in isolation. Instead, it has to be compared with other findings. For example, a temperature of 92°F is high compared with most air temperatures in London, but low compared with body temperatures. The *implications* of any fact, whether it is "good" or "bad", and what one should do about it, also depend upon seeing how it fits into the wider scheme of things.

In this chapter we try to develop a greater feeling for the practical interpretation of an observed result. Any new finding needs to be interpreted against empirically-based norms or concepts. The crucial step is to have found usable patterns and relationships by analysing other measurements of the same type.

4.1 **A Simplifying Concept**

Many problems arise year after year. The answers, if only we knew them, should therefore also be similar year after year.

This changes the emphasis from collecting new data to organising and using past information. The process can be illustrated by the ordinary problem of forecasting how long it will take to drive from town A to town B (cf. Ehrenberg, 1967). Such forecasts affect many practical decisions: what form of transport to use, when to leave A to arrive in B at a certain time, whether to stay overnight, when to return, whether to go at all, etc.

The standard market research or administrative method of providing information for these decisions is to collect facts: **e.g.** to arrange for some investigators to drive routinely between towns A and B, C and D, etc., and regularly report their latest driving times. This would produce the kinds of large arrays of tabulated statistics which are well-known. Usually any extensive analysis of the data is precluded by the high cost of collecting it and the demand for quick, practical decisions. Since even the latest facts are already just out-of-date, it seems difficult to wait for (and pay for) a lengthy analysis.

At some stage however, a more systematic understanding of the data may evolve, or perhaps a sudden discovery is made. The data that the travelling investigators have been collecting can generally be described by dividing the distance between each pair of towns by the number 35. This figure represents a new theoretical concept called the "average speed".

Forecasting now becomes much easier and cheaper: one simply looks up the mileage between any two towns and divides by 35. Most of the forecasts are successful because the notion that average speeds are approximately constant is based on driving times obtained under a wide range of conditions. But there are errors, and some of them are large.

Using the simple law "time = distance/35", further study leads to categorisation of the forecasting errors. It develops that these are broadly of four kinds.

(a) Errors attached to the type of road; more accurate forecasts result by making simple generalisable allowances for this in the calculations (e.g. an average speed of 55 mph for motorways, adding 30 minutes for leaving or entering a conurbation, etc.).
(b) Errors related to particular driving occasions, such as weekend traffic, being affected by fog or ice, making better speed by travelling in the dead of night, etc.
(c) Errors related to the specific driver or car, such as a fast driver or a mechanical breakdown.
(d) Errors that cannot be accounted for yet.

In none of these situations would knowledge of the latest up-to-the-minute driving times be of any help.

Successful forecasting of driving times therefore requires three types of information: (i) the generalisable law that "time taken = distance/35", (ii) generalisable adjustments concerning motorways, conurbations, etc., and (iii) specific facts about distances, types of roads, weather forecasts, etc., to slot into the laws and sub-laws. No direct measurements of recent driving times are needed.

The practical man might consider it too risky to rely on a mathematical formula and correction factors; he would rather have the facts. Yet, in practice, no one would dream of running routine surveys to measure the latest just-out-of-date driving times, unless he were doing academic research. In practical life, we use generalisations and concepts rather than isolated facts.

4.2 High, Low or Normal: The Relationship Between I and U

Much decision-taking depends upon the interpretation of some specific figure or number. For this a background of organised previous knowledge and norms is needed.

Consider having to assess the future progress of Brand X, a frequently bought consumer product launched several months ago. Suppose that in a sample survey housewives are asked which brands in the product-field they intend to buy and that about 14% say they intend to buy Brand X. By itself this figure means little. Thus it does not follow automatically that 14% of housewives *will* buy Brand X. More people may say they will buy new brands than actually do or can.

A second finding in the survey is that only 3% of the sample claim to be currently using Brand X. Now we have *two* figures, as shown in Table 4.1. One obvious interpretation is that if so many more people say they intend to buy Brand X than are currently using it, future prospects for the brand must be good.

TABLE 4.1 Intentions-to-Buy I and Brand Usage U for the New Brand X

I:	% of Sample who Intend to Buy X	=	14%
U	: " " " Currently Using	X =	3%

But is this interpretation so obvious? Do we really know that intentions-to-buy at 14% are unusually high compared with a current usage level of 3%? Do we know that a relatively high level of expressed intentions-to-buy is generally followed by an *increase* in the purchasing level of the brand? Both these notions may seem intuitively acceptable but are in fact wrong. To interpret what the readings in Table 4.1 actually mean we need more empirically based understanding of these variables.

I *is High.* The I and U results for the other brands in the product-field are also known from the survey. The average I/U ratio is about 2/1. Comparing the Brand X ratio of 14/3 with this seems to confirm the common-sense impression that the observed value of I for Brand X is high. This might also seem to imply a good subsequent sales potential.

However, this comparison has been made against the *average* value of I/U for the other brands, where the variation was in fact from $1\frac{1}{2}/1$ to 6/1. First we have to explain this large variation for the other brands before we can understand the I/U value for Brand X and its possible sales implications.

This is where earlier research results come in. Some previous basic research (Bird *et al.,* 1966) had shown that the level of I tends to vary with the square-root of the usage score U. Thus the values of I/,/U for different brands in a product-field were approximately constant at some value K, i.e.

$$\frac{I}{\sqrt{U}} = K.$$

In other words, $I = K\sqrt{U}$, and this relationship had been found to hold within an average of 3 percentage points under a wide range of conditions, as is summarised in Table 4.2.

TABLE 4.2 *The* **Conditions** under which the **Relationship** $I = K\sqrt{U}$ **has** been Found to Hold

- For frequently-bought non-durable branded goods in over twenty different product-fields (both food and non-food).
- For large brands and for smaller brands in each product-field.
- For some American data as well as for the main British data.

- For differing demographic sub-groups of the population.
- For different points in time, ranging over five years.
- For a variety of different forms of Usage questions.
- For certain different forms of Intentions-to-Buy questions.

- For established brands, whether with steady or with varying Usage levels.
- For successful new brands (i.e. ones with increasing Usage levels), except that some 5 to 8% fewer people then expressed an Intention -to -Buy.
- For old, dying brands (i. e. ones with slowly but steadily decreasing Usage levels), except that relatively more people than normal then expressed an Intention-to-Buy.

The relationship explains why the ratio I/U varies so much between different brands. If $I/\sqrt{U}$ is constant, then I/U cannot be constant. It must be bigger for small values of U than for large ones. Perhaps this will explain the high ratio of 14/3 for Brand X.

I *is* Low. We can work out that $K = 11.5$ in the equation $I/,/U = K$ for our product-field from the average ratio of I and $\sqrt{U}$ for the other brands observed in the sample survey. Accordingly, a brand with a 3% usage level should have an intention-to-buy level of $11.5\sqrt{3} = 20\%$. It follows that the observed level of 14 % for Brand X is not high, as thought before? but *low.*

This new finding is, however, still not of practical use. We need to know under what conditions such low values are generally found and their implications in terms of sales potential.

I *is Normal.* One of the specific findings in Table 4.2 provides the explanation we need. Newly launched brands generally had I levels 5 to 8 percentage points lower than established brands in the same product-field. Since Brand X is a new brand, its observed I value of 14%, 6 points lower than the theoretical value of 20%, is therefore *normal* for a new brand.

We now know that the result observed for Brand X does not tell us anything special at all ; we cannot use it to predict the brand's future sales potential. This often happens with observed data. Once we understand an area of study, we find either that most observations turn out to be normal and

predictable, or that there is no discernible pattern at all. This may be unexciting but it is inevitable. Forecasting is not as easy as asking a simple question in a market research survey.

4.3 The Role of Research

What is therefore needed is basic' research to build up a background of knowledge in order to understand what a particular measurement means. Its verbal or operational definition is not enough. For example, asking people their intention-to-buy does not necessarily mean what one thinks it says. This is not peculiar to attitudinal measures or even to social research in general. In physics, we place some mercury in a glass tube and expect it to measure something. This is an odd thing to do. Sometimes it measures temperature (in a thermometer), sometimes pressure (in a barometer), and so on. It all depends on the detailed circumstances and what has been found out about them through past observation and analysis. Research means not only collecting new data but also theoretical analysis of past data and the use of previous knowledge (e.g. looking up books and references to see what has already been found out about such measurements).

The analysis of data depends on establishing simple patterns and relationships. By far the simplest pattern is that something is more or less constant, i.e. unrelated to the other variables in the situation. In some cases, like that "average speed = distance travelled/time taken" is relatively constant, the resulting concept of "average speed" seems to make intuitive sense, at least with hindsight. In other cases, like that $I/\sqrt{U} = K$, the "constant" aspect of the data arises merely as an empirical finding.

In fact, the latter result raises many new questions for further research. Why does the intention-to-buy variable relate to the usage level of different brands as $I = K\sqrt{U}$? How does I relate to subsequent *changes* in U? Why does a new and subsequently successful brand tend to have relatively few people expressing an intention-to-buy, when in fact increasing numbers buy it? Why does an old and slowly dying brand have relatively many people expressing an intention-to-buy, when fewer and fewer people subsequently buy it?

The answers to such questions will be touched on in the Exercises and help us to understand better the on-going nature of research. However, in this book our main concern is not so much the general conduct of such research but the handling of the resultant data.

4.4 Summary

A fact has to be compared with other facts before it can be interpreted. We must establish whether an observed number is "high", "low" or merely

"normal"* Interpretive norms can be established only after the analysis of past data; this is a matter of either experience or of deliberate research. It involves the development of new concepts and the uncovering of simple patterns and relationships that generalise under a wide range of circumstances.

Simply reporting the observed facts is not enough. When a good statistician is asked "How is your wife?" he replies "Compared with whom?" Many people, however, prostitute their data by merely treating them as "better than nothing" instead of looking for any deeper relationships.

CHAPTER 4 EXERCISES

Exercise4A. An Awful Warning

"Fifteen people died on the roads over Christmas." Is that fact worth reporting?

Discussion.

The government used to issue the number of road deaths at Christmas every year as a warning against drink. Then it was recognised that the numbers were well below normal (fewer people drive at Christmas).

Exercise4B. Catch Them Early

A widely held view in television circles is that if a popular programme is screened early in the evening it will lead to higher audiences on that channel for the rest of the evening.

Facts quoted in support of this view are that, typically, a programme shown at 10 pm on a Monday will attract 32 % of the audience watching the 7 pm programme that evening on the same channel, but only 18 % -of those *not* viewing the earlier programme. It therefore seems that the more people who can be induced to watch at 7 pm, the higher the audience will be at 10 pm. But is the 32% attracted from 7 pm really a high figure?

Discussion.

No, the figure of ***32%*** is not especially high. The 10 pm Monday programme would attract 32% of the audience of the other programme even if this had been shown on a different day altogether.

Research has shown that there is a general phenomenon known as "channel loyalty", namely that even programmes shown on different days on the same channel tend to share the same audience. Only if two programmes are virtually adjacent on the same evening is there-any extra overlap in their audiences (Goodhardt *et al.*, 1975).

Exercise 4C. ASuccessfulTreatment

A group of patients is treated by a certain method. If 64% recover, is the treatment successful?

Discussion.

We are given here only a single result. The question therefore is, how this compares with other treatments or no treatment? Perhaps the spontaneous recovery rate of patients with this illness was 85 %. (Then we might ask whether only chronic or abnormally severe cases were given the treatment.)

Exercise 4D. "Only 70% or As Many as 70 %?"

Brand M is a frequently bought consumer product. During the peak-season quarter it was purchased by only about 70% of those who had bought it in the preceding off-peak quarter, but its total sales increased by a third. Does this imply a catastrophic decrease in brand-loyalty, temporarily hidden by the seasonal increase in sales?

Discussion.

What is the normal number of repeat-buyers? Does the 70% figure for Brand M mean *only* 70 % or us *many as* **70** %?

The incidence of repeat-buyers of a frequently bought branded good can be closely predicted when there is no trend in the brand's sales level (e.g. Ehrenberg, 1972). The prediction for Brand M is that about 67 % will buy it again under no-trend conditions. The observed level of 70% therefore seems normal and there is no evidence of any loss of brand-loyalty.

But this figure was observed when sales followed a seasonal trend. This implies that off-season buyers were not affected by the seasonal upswing! The market for this product must therefore be segmented into two types of consumers, narnely year-round buyers who are quite unaffected by the seasonal trend, and peak-season-only buyers: a result which was unexpected but has since been confirmed.

Exercise4E. ExplainingtheRelationshipbetweenIandU (Section **4.2)**

Why does the relationship between the percentage I of consumers who express an Intention-to-Buy a brand and the percentage U using the brand follow the form $I = K\sqrt{U}$? Furthermore, why is I relatively low for a new and subsequently successful brand, yet relatively high for an old and slowly dying brand?

Discussion.

Answers to these questions obviously required further analysis or research. They can be found by turning to the reference already cited in Section 4.2 (Bird *et al.*, 1966).

Given that I varies with U, the level of expressed intentions-to-buy among users and non-users was examined. Table 4.3 gives a typical result in some detail. The percentage of consumers who expressed an intention-to-buy decreased as more time elapsed since they last used the brand. Current users are likely to say they will buy the brand again ; past users are less likely to. It follows from these results-that if Brand A has twice as many current users as Brand B, A has fewer *non-users* and therefore less of a contribution to its intentions score from this source. Intentions-to-buy must therefore vary *less* than proportionately to the usage level. This accounts for the non-linear relationship between I and U. (The fact that the equation uses the *square root* of U has as yet no rationale.)

TABLE 4.3 Intentions-to-Buy Amongst Current and Former Users of a Particular Brand

	Intending to buy
Current users	95%
Used in last 6 months	45%
Used more than 6 months ago	10%
Never Tried*	5%

*Possibly including some who were users long ago

The reason that I is relatively low for a new brand is that it has fewer former users than an established brand, and therefore relatively fewer people expressing an intention-to-buy, given its usage level. In contrast, an old, slowly dying brand has a long tail of former users some of whom express an intention-to-buy it.

Exercise 4F. **Intentions-to-Buy and Future Usage**

Do consumers' expressed intentions-to-buy branded goods predict future changes in their usage behaviour?

Discussion.

Expressed intentions-to-buy vary with current and past usage. They will therefore also successfully predict the level of future usage, as long as there is no general change in the usage pattern. (But it might be easier to assert that the levels of current and past users will not change!) The data discussed so far provide no direct evidence of how expressed intentions-to-buy relate to *changes* in future behaviour, because these had not been measured. Additional research is therefore needed.

Some results have been reported in the reference already cited. For example, it was established whether or not users or non-users of a brand in 1963 also used it in 1964, giving four groups, as shown in Table 4.4. The body of the table shows the percentage of each group who had in 1963 expressed an intention-to-buy the brand.

TABLE 4.4 The **Percentage of Users and Non-Users of a Brand in 1963 and in 1964 expressing an Intention-to-Buy in 1963**

% who in 1963 expressed an intention to buy the brand	Users in 1963	Non-Users in 1963
Users in 1964	75%	27%
Non-Users in 1964	77%	15%

The results show that these 1963 intentions were hardly related to people's subsequent actions in 1964. Over 70% of users in 1963 said they intended to buy, and this turned out to be the case, irrespective of whether they actually bought the brand in 1964 or not—75% and 77%. In contrast, far fewer of the *non-users* in 1963 said they intended to buy, and this also showed relatively little difference by whether or not they did so in 1964-27% and 15%. (The lower percentage among non-users in *both* years is probably due to this group including people who never used *any* brand in the product-class in question.)

The conclusion is that expressed intentions-to-buy do not predict future changes in behaviour, even though they firmly reflect current and past behaviour.

Exercise 4G. Information for Decision-making

How does practical decision-making depend on information?

Discussion.

A classic illustration is the case of the two shoe-manufacturers many years ago, who each sent a salesman to Africa to explore the possible market for boots. One cabled back, "Splendid market for boots-nobody wears any!" The other, "No market for boots-nobody wears any!"

The new information is obviously relevant, but does not necessarily tell one what to do. Nonetheless, it can still be valuable. Firstly, if it *is* decided to go into Africa, it is helpful to know the market conditions there; i.e. the information helps to *execute* the decision once it is made. Secondly, the new information can be combined with *other* information to guide the decision-maker's judgment. A shoe-manufacturer whose past expertise has been producing highly competitive versions of well-established lines might be thought to have a harder task in Africa than one who has specialised in getting new fashions adopted.

Even so, there could be no clear prediction that the latter manufacturer would actually succeed in Africa. To make a valid prediction one would need many past cases where entering a new market *had* succeeded under more or less similar conditions. But if such empirical evidence were available, the situation would no longer be regarded as a major decision-problem since the answer would be obvious.