PART III: STATISTICAL VARIATION

This part of the book concentrates on ways of summarising and interpreting irregular variation within a single set of readings.

Chapter 11 discusses how the individual readings in a set of data can be arranged as a frequency distribution and be further summarised by measures of their average size and scatter. Observed frequency distributions can take different shapes and these can often be described by mathematical functions. Chapter 12 discusses the main examples : the "Normal", the "Poisson", and various "Binomial" distributions.

The concept of probability is often helpful in analysing irregular variation. Chapter 13 outlines this approach, in which the observed variability of the data is regarded as if it were random. This leads to models of a probabilistic or "stochastic" kind.

Chapters 14 and 15 describe and evaluate techniques of correlation, regression and factor analysis which are widely used in statistical analyses. These methods aim to describe the way different variables vary together within a single set of readings. (In Part II we considered how different variables vary together for *more* than one set of readings.)

The topics covered in this part of the book are of more restricted practical relevance than those in Parts I and II. For example, some subject-areas almost never use frequency distributions or probability methods but others do so a great deal. Experienced readers will know which topics are most relevant to their areas of interest. But many readers will need to gain some broad familiarity with all the topics covered here, and this they should be able to extract.

CHAPTER 11

Summary Measures

So far we have been dealing with variations between different sets of data. This is "controlled" variation. For example, when we were analysing heights and weights, we first took boys of different age-groups and looked at the systematic variation between them. We controlled the sex and age differences.

But if we simply took a group of children and knew nothing else about them except their heights, the differences in the readings would be "uncontrolled" variation. We now concentrate on this kind of uncontrolled variation within a single set of readings.

Reporting uncontrolled data in full is clumsy, especially when there are many readings. This chapter discusses how to look at such data as frequency distributions and to reduce these to meaningful summaries of average size and scatter.

11.1 Empirical Frequency Distributions

As a small numerical example, consider the following set of eight readings:

$$3, 4, 8, 3, 6, 1, 3, 4.$$

To see these better, we can arrange them in order of size:

$$1, 3, 3, 3, 4, 4, 6, 8.$$

To summarise further, we can express the readings as a *frequency distribution,* one 1, no 2's, three 3's, etc., as in Table 1.1.

TABLE 11.1 The Frequency Distribution of the 8 Readings

	Value								The Total Number of Readings
	1	2	3	4	5	6	7	8	
Frequency:	1	0	3	2	0	1	0	1	8

This kind of frequency distribution might be a sufficient description of an isolated set of readings, but it is still too clumsy if we want to compare different sets of data.

One factor which usually differs is the number of readings in each set of data. For example, Table 11.2 compares our original 8 readings with a second set of 40 readings. The eye cannot immediately take in the detailed similarities and differences in the two sets of data.

TABLE 11.2 Two Sets of Readings to Compare

	Value								The Total Number of Readings
	1	2	3	4	5	6	7	8	
1st set	1	0	3	2	0	1	0	1	8
2nd set	5	1	14	10	1	4	1	4	40

We can eliminate the visual confusion by showing the observed frequencies as *proportions* of the total number of readings in each set, as in Table 11.2a.

TABLE 11.2a The Two Frequency Distributions Expressed as Proportions

	Value								Total
	1	2	3	4	5	6	7	8	
1st set	.12	.00	.38	.25	.00	.12	.00	.12	0.99*
2nd set	.12	.03	.35	.25	.03	.10	.03	.10	1.01*

* Not 1.00 due to rounding errors

Now it is clear that the two sets of readings have similar properties. Table 11.2b further simplifies the visual comparisons by expressing the frequencies as percentages and avoiding the clumsy decimal point. (But in mathematical work it is simpler to work in *proportions.)*

TABLE 11.2b The Frequency Distributions Expressed as Percentages

	Value								Total
	1	2	3	4	5	6	7	8	
1st set %	12	0	38	25	0	12	0	12	99
2nd set %	12	3	35	25	3	10	3	10	101

Table 11.2c simplifies the comparison even more by arranging the frequencies in broad groups (although here, with only four effective categories, the grouping has possibly been overdone, especially for the large 3-4 category). But the table is still not succinct enough to provide a memorable summary of the data. To reduce the data still further we need a summary like an *average.*

TABLE 11.2c The Two Frequency Distributions in Grouped Categories

	Values				
	1-2	3-4	5-6	7-8	Total
1st set	% 12	63	12	12	99
2nd set	% 15	60	13	13	101

11.2 Summaries of Average Size

Averages are the main tool of statistical analysis. But some averages are good and some can be misleading. To judge whether simple averages describe the data adequately, one first needs to calculate them and then check the figures against the data. This also helps one to see what the data themselves are like. The golden rule in looking at data is "Average *before* you look".

The three main types of average are the *mode*, the *median*, and the ***mean*** (or arithmetic average). We illustrate these with our original set of eight readings, repeated in Table 11.3.

TABLE 11.3 The Frequency Distribution of the 8 Readings

(From Table 11.1)

	Values								
	1	2	3	4	5	6	7	8	Total
Frequency	1	0	3	2	0	1	0	1	8

The ***mode*** is defined as the most frequent reading. For the data in Table 11.3 the mode is 3, there are more 3's than any other reading. Sometimes the mode is difficult to determine precisely because there is more than one "most frequent" reading, or the grouping obscures the mode, as in Table 11.2c.

The ***median*** is the value that is exceeded by half the readings. This is not always clear. For example, in Table 11.3, 4 out of 8 readings are greater than 3, and 4 out of 8 are less than 4, so that each of these could be the median, as defined. In such cases the median is conventionally taken at 3.5, half-way between the largest and smallest possible values.

Finally, the ***mean*** or arithmetic "average" of the readings is their sum divided by the total number of readings. In our example this is $32/8 = 4$.

Thus for the 8 readings in Table 11.3 we have

the Mode at 3,
the Median at 3.5,
the Mean at 4.

Although the three measures are conceptually quite different, they take similar values in this example. Such cases occur mainly with distributions that are approximately symmetrical and have their mode or "hump" in the middle. Figure 11.1 gives an illustration for the heights of a group of lo-year-old boys.

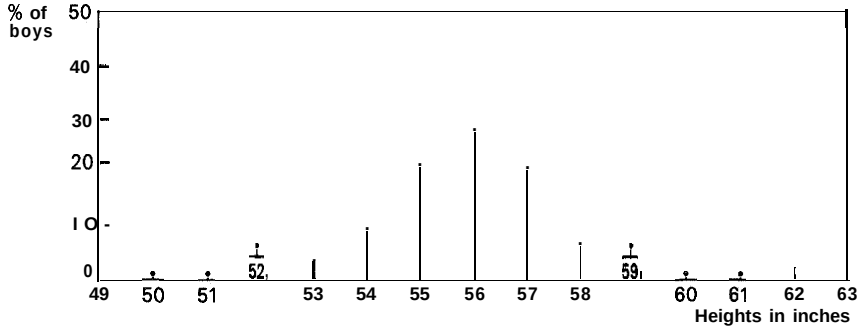

Figure 11.1 The Heights of 10-Year-Old Boys

These kinds of symmetrical distribution are exceptionally important. They are virtually the only kind of statistical data that can be described in a simple standardised way. The mean, median and mode will always be approximately equal. Therefore a single number has all the different meanings of these measures and most of the readings will fall close to this value.

This situation should be implied whenever an unqualified average is reported. For example, if the average age of a group of people is described as 30 years, the implication should be that about half are younger than 30, and that most of them are close to 30. 'If the data are more complex than this, one needs to say so.

The Average of Non-symmetrical Distributions

As a contrast, Table 11.4 sets out the reading frequencies of a monthly magazine. Most people either read all or almost all of the issues, or else they read none or almost none. Very few people read about half of the issues.

TABLE 11.4 The Number of Issues of the Monthly Magazine X Read per Year
(Percentages from Figure 11.2)

	Number of Issues Read													All Adults
	0	1	2	3	4	5	6	7	8	9	10	11	12	
% of Adults	40	15	5	2	1	1	1	0	1	1	3	10	20	100

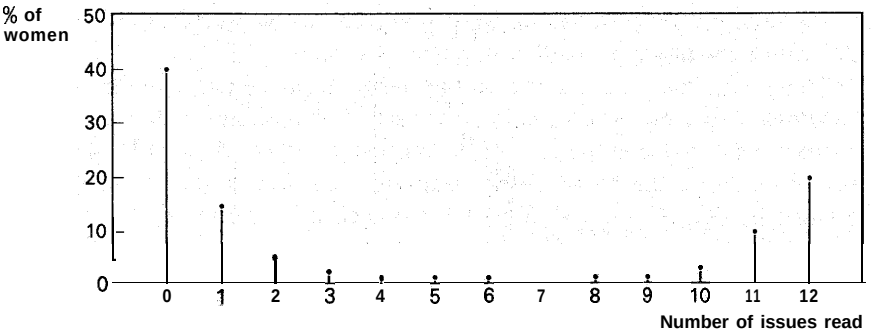

Figure 11.2 The Number of Issues of a Monthly Magazine Read per Year

The data therefore fall into a U-shaped distribution, as Figure 11.2 illustrates.

The median of the distribution is 1, since 40% of the people read no issue and 45% read 2 or more. If we regard the mode as a "locally" most common reading, this distribution has two modes, 0 and 12. Finally, the average number of issues read per person per year is $440/100 = 4.4$. We therefore have

two Modes at 0 and 12,
the Median at 1,
the Mean at 4.4.

Much of the confusion about averages arises from situations like this. The mean, median and mode take very different values, but sometimes one type of average is interpreted as if it were another. The mean here is in no sense a typical or modal reading: hardly anybody reads 4 or 5 issues and nobody reads 4.4 issues. Again, half the population does *not* read more than the mean number of issues, 63% read 4 or fewer issues and only 37% read 5 or more.

There is no simple routine way to describe such data. They require a *tailor-made* description: e.g. that about 55% of the people read no or only 1 issue and about 30% read 11 or 12 issues. This would summarise the main characteristics of these particular data quite well.

The Pre-eminence of the Mean

The mean is the most commonly used measure of average size. This is because it provides a summary which can generally be used in further analyses of the data where the median and mode cannot. For example, when combining two sub-groups of readings into a single group, the mode and the median of the sub-groups cannot be related to the corresponding values

of the total group. Instead, one has to go back to the individual readings to determine the mode or median of the total group.

There is no such difficulty with the mean. Suppose we have one set of 8 readings with a mean of 4 and another set of 12 readings with a mean of 6. Since a mean value is the sum of the individual values divided by the number of readings, the mean of the combined sets will be (8 x 4) + (12 x 6) divided by (8 + 12) or $104/20 = 5.2$. No such calculations are possible for the median or the mode.

However, the latter two measures are not meaningless. It can be helpful to know the most frequent or "modal" reading, and also to see where the 50 :50 or median division comes in a set of data. But the median and the mode do not provide summaries which can be widely used. The distinction is between interesting concepts and usable summaries.

Another property of the arithmetic mean is that the *total* value of the readings can easily be calculated by multiplying the mean by the number of readings. Such totals can sometimes be of practical relevance, For example, an average rainfall of 2 inches per month adds up to 24 inches in the whole year. (There is no corresponding benefit in knowing that the average height of a group of boys is 57 inches, "laid end to end ..."; nor does it follow from a mean readership of 4 magazine issues per person that a total of 2 million copies of magazine X were sold in a city of 500,000, since some copies will have been read by more than one person.)

113 Measures of Scatter

When summarising uncontrolled variation one must also describe how far the readings are from the mean. There are several different statistical measures for this. One is the *mean deviation,* already widely used in Parts I and II.

The Mean Deviation

In our example of eight readings

$$1,\ 3,\ 3,\ \ 3,\ \ 4,\ \ 4,\ \ 6,\ \ 8,$$

the individual readings deviate from the mean, 4, by

$$-3,\ -1,\ -1,\ -1,\ \ 0,\ \ 0,\ \ 2,\ \ 4$$

units respectively. The total or average of these deviations is necessarily zero because the negative deviations balance the positive ones. This provides a useful check on one's arithmetic.

To summarise the scatter of the deviations we need to measure their average size. One simple procedure is to ignore the negative signs and take

the average of the numbers, which is 12/8 or 1.5. This is the *mean deviation.* (Sometimes this is also called the "mean absolute deviation", where "absolute" signifies that one is considering the sheer *size* of the deviations, irrespective of plus or minus signs. The *modulus* symbol, two vertical lines as in |x|, described verbally as "mod x", is also used to signify absolute values.)

The mean deviation is a very simple and useful measure in numerical analyses of data. For example, in all the analyses in Parts I and II of this book it seemed natural to look at the average (absolute) size of the deviations from the various "models" that were fitted.

The Standard Deviation and the Variance

The standard deviation and its square, the variance, are the conventional measures of scatter used in statistical theory. They are much easier to manipulate mathematically and thus more useful in theoretical work, but are conceptually less obvious than the mean deviation and more complicated to compute with pencil and paper.

In our example of eight readings, we are typically faced with positive and negative deviations from the mean :

$$-3, \quad -1, \ -1, \ -1, \ 0, \ 0, \quad 2, \quad 4.$$

The standard deviation is based on the idea of squaring all the deviations to eliminate any negative signs. Three steps are involved : (1) squaring the deviations, (2) taking the average of the squares, and (3) taking the square root of this average in order to return to the original scale of measurement. Expressed as a formula, we have

$$\text{Standard deviation} = \sqrt{\left\{\frac{\text{Sum of (Deviations)}^2}{\text{No. of readings}}\right\}}.$$

With the eight deviations above, the squares are

$$9, \quad 1, \ 1, \ 1, \ 0, \quad 4, \quad 16.$$

The average of these squared deviations is $32/8 = 4$. The standard deviation is the square root of this, or $\sqrt{4} = 2$. (It is sometimes also called the "root mean square".)

The formula for the standard deviation is generally slightly adjusted for technical reasons in the theory of statistical sampling. Thus the sum of the squared deviations is divided by one less than the number of readings or $n - 1$, i.e.

$$\text{Standard deviation} = \sqrt{\left\{\frac{\text{Sum of (Deviations)}^2}{n-1}\right\}}.$$

This makes no difference to the numerical result if the number of readings, n, is at all large. Even for our small numerical example with n = 8, the

standard deviation using the divisor $(n-1)$ is only about 5% different, 2.1 instead of 2. To all intents and purposes we can therefore think of the standard deviation as an average, but use the formula with $(n-1)$.

The square of the standard deviation is defined as the *variance*, and is a useful mathematical quantity. It is the intermediate stage in calculating the standard deviation, i.e. the average of the squared deviations:

$$\text{variance} = \frac{\textit{Sum of (Deviations)}^2}{n-1}.$$

The Coefficient of Variation

The coefficient of variation is a related measure that is sometimes used to express the standard deviation as a percentage of the mean instead of in the original units of measurement:

$$\text{Coefficient of Variation} = \frac{100 \times \text{Standard Deviation}}{\text{Mean}}.$$

For example, with a mean of 4 and a standard deviation of 2, as in the case of our 8 readings, the standard deviation is half the size of the mean or 50%:

$$\text{Coefficient of Variation} = \frac{2 \times 100}{4} = 50\%.$$

This approach is most useful when the scatter in different sets of readings increases proportionally with the mean values of the readings. For example, in Table 11.5, instead of having to report and remember quite different numerical values of the standard deviation (e.g. .5 for set A, 1.2 for set B, etc.), we can adequately summarise the size of the scatter with a single "constant" figure of about 25 %.

TABLE **11.5** Approximately Constant Coefficients of Variation in Five Different Sets of Data

	Sets of Data				
	A	B	C	D	E
Mean Value	2	5	10	40	120
Standard Deviation	.5	1.2	3	9	31
Coefficient of Variation	25%	24%	30%	22%	26%

In contrast, Table 11.5a describes five other sets of data where the size of the scatter increases only minimally with the mean. Here it is simpler to report the approximately constant standard deviations of 5 units than the different coefficients of variation.

The method of reporting does not affect the *interpretation* of the data. It is still easy to see that Set L with a mean of 4, has a relatively large standard deviation of 4 (the coefficient of variation being about 100%) whereas Set Q, with a mean of 80, has relatively small scatter (the coefficient of variation being only about 8%).

TABLE 11.5a Approximately Constant Standard Deviations in Five Other Sets of Data

	Sets of Data				
	L	M	N	P	Q
Mean Value	4	8	12	25	80
Standard Deviation	4	3	5	5	6
Coefficient of Variation	100%	37%	42%	20%	8%

The Range

The final measure of scatter to consider is the *range.* This is defined as the difference between the highest and lowest readings in the data, an obvious, common-sense type of measure. For the eight readings

$$1, 3, 3, 3, 4, 4, 6, 8,$$

the range is $8 - 1 = 7$. Reported together with the mean of 4, the range of 7 helps to give a good feel of these data; one likes to know how different the largest and smallest readings in a group are (although giving the two extreme values of 1 and 8 seems even more informative).

But like medians and modes, the range has serious disadvantages in detailed analytic work.

(i) It depends on the two extreme values and is therefore very sensitive to odd outlying readings. (But it provides a good check on the *occurrence* of any unusually high or low values.)

(ii) With large numbers of readings not already ordered by size, searching for the highest and lowest values is laborious.

(iii) The numerical value of the range depends on the number of readings. (Adding another reading may *increase* the range but can never decrease it.) This makes it difficult to use when comparing the scatter of sets of data with different numbers of readings. (In contrast, the mean and standard deviations are *averages,* and thus independent of the number of readings.)

(iv) The range of some combined set of readings cannot be calculated from the ranges of each of the sub-sets being grouped (or vice versa). Instead, one has to go back to the raw data and look for the highest and lowest values in each set.

A variation on the range is to exclude one or more of the extreme readings. This can give the measure more stability. The best-known example is the "inter-quartile range", the difference between the two quartiles. (These are the two values below and above which 25 % of the readings lie, and thus are akin to the median). But such measures are seldom used in practice.

The Descriptive Meaning of Measures of Scatter

In our example of 8 readings the mean is 4 and the range is 7. If we reported just these two figures, the description should imply that the readings are more or less systematically distributed about the mean of 4, from 0 or 1 to 7 or 8. But the readings *could* extend from 3 to 10 and still have a mean of 4 and a range of 7, as follows

3, 3, 3, 3, 3, 3, 4, 10.

If this were the case, it would be misleading to report the mean and the range alone. One would need to describe the data in more detail, i.e. mostly 3's, with one high value at 10.

Similar considerations apply to the use of other measures of scatter. For example, the mean deviation and standard deviation tell us the "average" size of the deviations from the mean, but do not indicate how the individual deviations are distributed. Are most of the deviations of about this average size, or are half much greater and half less, or are most of them small with one or two very large deviations, or what?

The only simple description arises with symmetrical humpbacked distributions, like the one illustrated in Figure 11.1. The mean value here is 56 and the mean deviation is 1.2. Most of the readings fall quite close to the mean. Well over half the individual deviations from 56 are smaller than the *mean* deviation, and only about 10 % are bigger than *two* mean deviations.

This is a common pattern for symmetrical humpbacked distributions and it is discussed in more detail in Chapter 12. For such distributions either the mean deviation or the standard deviation adequately summarizes the proportion of readings that lie in any particular range of values. But for distributions that are *not* symmetrical and humpbacked (like the U-shaped one in Figure 11.2), it is not possible to describe in any general way how many readings lie within one mean deviation of the mean, or more than two mean deviations away, and so on. More elaborate methods are needed to describe such variation. These also are discussed in the next chapter.

11.4 Summary

The "shape" of a set of readings can be seen by arranging them in order of size and grouping them in convenient intervals. It is easier to compare sets of data with different numbers of readings if the frequencies of each value are

expressed as proportions or percentages of the total number of readings in each set.

The data can be summarised further by calculating the average size of the readings. The mean or arithmetic average is the preferred measure because it can readily be used in further analyses, e.g. when combining or separating different sets of data. Two other measures of average size are the mode (the most frequent reading) and the median (the value exceeded by half the readings). These are helpful concepts but do not provide summaries which in practice are usable in further analyses.

The mean, the mode and the median tend to be approximately equal in humpbacked, symmetrical distributions because most of the readings lie relatively close to the mean. Such symmetrical distributions are therefore easy to summarise. But quoting the mean value on its own can be misleading with other kinds of distributions.

The mean deviation, the standard deviation and its square, the variance, are commonly used measures to describe the *scatter* of readings. These are different ways of summarising how far the individual readings differ from the mean. The mean deviation is the average of all the deviations, ignoring any minus signs. It is arithmetically and conceptually easy to use. The variance eliminates the minus signs by squaring the deviations and *then* averaging. The standard deviation is the square root of the variance. These two measures are easier to use in mathematical analysis.

CHAPTER 11 EXERCISES

Exercise 11A. A Logical Sequence

What is a logical sequence in describing statistical data?

Discussion.

The successive steps in describing data are best kept independent of each other. A typical sequence is

- n the number of readings;
- m the mean, which is independent of n;
- s the standard deviation of the readings from the mean, which is independent of n and m;
- f the "shape" of the distribution (to be summarised by some mathematical function), which is independent of the values of n, m, and s.

Exercise 11B. The Greek Sigma Notations

Statistical texts often use the Greek symbols σ (small sigma) and Σ (capital sigma). What do these mean?

Discussion.

A common notation is to use the Greek letters for population values and Roman letters for sample values. For example, σ means the standard deviation of a set or "population" of readings, and s means the standard deviation of a sample of n readings from that population.

If sampling is not involved, then regardless of the number of readings, the data effectively represent the whole population. Strictly speaking, Greek letters are then the appropriate ones to use. However, Roman letters are simpler to use in practice. The context should make the situation clear.

Capital sigma, Σ, is used for a completely different purpose. It denotes the *sum* of the relevant readings. Thus for n readings of x, say x_1, x_2, x_3, etc. up to x_n,

$$\Sigma x = x_1 + x_2 + x_3 + \cdots + x_n.$$

Sometimes this is written as

$$\sum_{1}^{n} x$$

when there is possible doubt about how many terms are being summed, or as

$$\Sigma x_i,$$

where x_i stands for the ith reading, i taking all possible values from 1 to n. This may be written still more explicitly as

$$\sum_{i=1}^{n} x_i,$$

i.e. the summation of x_i for all values of i from 1 to n.

Exercise 11C. Deviations from the Mean

Prove that in a set of readings the average deviation from the mean is zero.

Discussion.

If you have n readings of x, the sum of the deviations from the mean $\bar{x}$, can be written as $\Sigma (x - \bar{x})$. Here $\bar{x} = \Sigma x/n$, or the total of the readings, Σx, divided by their number, n.

$$\begin{aligned} \text{Now } \Sigma (x - \bar{x}) &= (x_1 - \bar{x}) + (x_2 - \bar{x}) + (x_3 - \bar{x}) + \cdots + (x_n - \bar{x}) \\ &= (x_1 + x_2 + x_3 + \cdots x_n) - n\bar{x} \\ &= (\Sigma x) - n\bar{x} \\ &= \Sigma x - n \Sigma x/n \\ &= 0. \end{aligned}$$

Exercise 11D. Combining Two Averages

In Section 11.2 we noted that the mean of two combined sets of readings could be calculated from the two separate means. Express this in algebraic terms.

Discussion.

If we have two sets of n_1 and n_2 readings with means m_1 and m_2, the numerical *totals* of each set are n_1m_1 and n_2m_2. The total of the combined set of readings is therefore $(n_1m_1+n_2m_2)$. *Thus* the mean is

$$\frac{n_1m_1 + n_2m_2}{n_1 + n_2}.$$

(There are three commonly used symbols for the mean of a set of readings:

(i) $\bar{x}$, described as "x bar", for any set of readings,
(ii) m, usually for a sample mean,
(iii) μ, the Greek "mu", for the mean of a population.

Because the suffices in our example referred to different sets of data rather than individual readings, we used m instead of $\bar{x}$.)

Exercise 11E. A Computing Short-cut

The expression $\{\Sigma(x^2) - n\bar{x}^2\}/(n-1)$ is a short-cut formula for calculating the variance. Prove and discuss it.

Discussion.

The variance of a set of n readings is the average of the squared deviations from their mean. Consider the sum of the squared deviations:

$$\Sigma(x - \bar{x})^2 = (x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2.$$

Now

$$\begin{aligned}(x_1 - \bar{x})^2 &= (x_1 - \bar{x})(x_1 - \bar{x}) \\ &= x_1^2 - x_1\bar{x} - \bar{x}x_1 + \bar{x}^2 \\ &= x_1^2 - 2x_1\bar{x} + \bar{x}^2.\end{aligned}$$

Similarly

$$(x_2 - \bar{x})^2 = (x_2^2 - 2x_2\bar{x} + \bar{x}^2)$$

Therefore

$$\begin{aligned}\Sigma(x - \bar{x})^2 &= \Sigma(x^2 - 2x\bar{x} + \bar{x}^2) \\ &= \Sigma(x^2) - \Sigma(2x\bar{x}) + \Sigma(\bar{x}^2) \\ &= \Sigma(x^2) - 2\bar{x}\Sigma x + \Sigma(\bar{x}^2) \\ &= \Sigma(x^2) - 2\bar{x}n\bar{x} + n\bar{x}^2 \\ &= \Sigma(x^2) - 2n\bar{x}^2 + n\bar{x}^2 \\ &= \Sigma(x^2) - n\bar{x}^2.\end{aligned}$$

The sum of squared deviations can therefore be written as

$$\text{Sum}(x^2) - n\bar{x}^2$$

or, because $n\bar{x}^2 = n(\Sigma x)^2/n^2$, as

$$\text{Sum}(x^2) - (\text{Sum } x)^2/n.$$

Both forms are usually easier to calculate than the basic expression $\text{Sum}(x - \bar{x})^2$, especially when using a desk or pocket calculating machine.

The reason is that the standard formula requires calculation of all the deviations, $(x - \bar{x})$, before squaring them. With the short-cut formulae one can square the original readings x, sum these squares, and sum the x-readings (all in one single operation on a machine with a "revolution counter").

For example, consider the 5 readings

$$4, \quad 25, \quad 6, \quad 9, \quad 12.$$

The mean is $\bar{x} = 11.2$, and the deviations are

$$-7.2, \quad 13.8, \quad -5.2, \quad -2.2, \quad .8.$$

If we use the standard formula for the variance, $\Sigma (x - \bar{x})^2/(n-1)$, we first have to write down the deviations, then square them, and then sum these squares, giving $51.84 + 190.44 + 27.04 + 4.84 + 0.64 = 274.80$. (The last two steps can be done in one operation on a calculating machine.) Hence we get a variance of $274.80/4 = 68.7$, and a standard deviation of 8.3.

But if we use the short-cut formula we avoid computing the deviations and introducing negative numbers with their extra error possibilities. Squaring the original readings and summing them gives $16 + 625 + 36 + 81 + 144 = 902$. (The same computation will also give the total of the readings, 56, if there is a "revolution counter".) The sum of squared deviations is therefore $902 - (56)^2/5 = 274.8$, as before.

An additional short-cut which is convenient on many calculating machines is to calculate $\Sigma(x^2) = 902$ and $\Sigma x = 56$ in one operation, then multiply $\Sigma(x^2)$ by n $(902 \times 5 = 4{,}510)$, and then subtract $(\Sigma x)^2$, i.e. $4{,}510 - (56)^2 = 1{,}374$. This gives $n\Sigma (x - \bar{x})^2$ or 5×274.8 in this case. To obtain the variance we therefore divide by $n(n-1) = 20$, giving 68.7 as before. The only number that needs to be written down is the final answer.

A *Warning*. While the short-cut formula avoids calculation of the individual deviations, it also by-passes any chance of *looking* at these deviations to get the "feel" of the data, to spot unusually large deviations, etc.: all the things which need to be done when first handling new data. The short-cut formula should therefore only be used when one is familiar with the kind of data in question and when the resulting standard deviation agrees with prior expectations (implying no aberrant outliers in the new data).

Exercise 11F. The Standard versus the Mean Deviation

What are the relative advantages and disadvantages of these two measures?

Discussion.

The advantages of the standard deviation (and of the variance) are that they are much easier to use in mathematical theory than the mean deviation.

When calculating the mean deviation it is easy for the human mind to eliminate the minus signs before averaging, but it is very difficult for formal mathematics to cope with this. It involves examining each deviation separately and then doing something different according to each deviation's sign. This operation is difficult to incorporate in any mathematical procedures.

In contrast, squaring each deviation when calculating the variance is complex to the human mind unless the numbers are very simple. Yet it is

mathematically easy because the same process (squaring) occurs regardless of the sign of the initial number. The outcome is always a positive number.

For example, it is possible to have a short-cut formula for the variance but no corresponding simplification is feasible for the mean deviation. The same applies to many other mathematical applications of these measures (e.g. the "analysis of variance" discussed in Chapters 18 and 19).

But conceptually and numerically the mean deviation is the simpler measure. In the numerical example in the last exercise, the mean deviation is the average of the numbers 7.2, 13.8, 5.2, 2.2, 0.8, which is $29.2/5 = 5.8$. It is even simpler to compute if the deviations are rounded to the nearest whole number. But the corresponding pencil and paper calculations for the standard deviation involve squaring and taking square-roots. The numerical analyses in Parts I and II would have been much more laborious if carried out in terms of the standard deviation.

There is, however, no major problem of having to choose between the two measures. In the most important case, for *Normal* Distributions, the two measures are directly equivalent to each other:

$$1 \text{ standard deviation} = 1.25 \text{ mean deviation},$$

as will be seen in the next chapter. One can therefore calculate the mean deviation but translate it into standard deviations when necessary (like changing from inches to centimetres). In other cases, *neither* measure has a direct descriptive meaning.

Exercise 11G. Outliers

With an exceptionally high or low value in a set of readings, how sensitive are

(i) the mean, median, and mode?

(ii) the mean deviation, standard deviation, variance, and range?

Discussion.

Consider two simple sets of 1,000 readings A and B:

A: 100 3's, **800** 4's, 100 5's,
B: 100 3's, 799 4's, 100 5's, 1 1,000.

The "outlier" at 1,000 in B might be a measurement or recording error. These are often very dramatic and are important to spot and to eliminate from the main analysis.

The mean, median and mode of the two sets of data are

	A	B
Mean	4	5
Median	4	4
Mode	4	4

Only the mean is at all sensitive to the outlier. The single reading of 1,000 in Set B increased it by 25%, but even here the effect is not dramatic. Unless the means of other such sets of data are generally very close to 4, one would probably not react to the value of 5 as implying an aberrant value.

The mean deviation, standard deviation, variance and range of the two sets are approximately

	A	B
Mean Deviation :	0.2	2
Standard Deviation :	0.4	3
Variance:	0.2	10
Range:	2	997

In terms of absolute increase, the standard deviation is a little more sensitive than the mean deviation. The variance is even more sensitive because the odd outlying value is *squared* before averaging, and thus becomes more dominant. The range is clearly too sensitive to act as a measure of "average" or typical scatter at all, but it is highly efficient for actually spotting outliers.

Exercise 11H. The Divisor $(n-1)$ for the Variance

Discuss the practical and theoretical implications of using the divisor $(n-1)$ instead of n in the variance formula, i.e.

$$\frac{\Sigma (x-\bar{x})^2}{n-1} \text{ instead of } \frac{\Sigma (x-\bar{x})^2}{n}.$$

Discussion.

From a "practical" point of view, n is the better divisor because it is easier to comprehend. One can see the formula represents the average of the squared deviations from the mean. In contrast, the expression Sum $(x-\bar{x})^2/(n-1)$, has to be accepted as a "formula".

Yet we have seen that using the divisor $(n-1)$ instead of n makes virtually no difference *numerically*, except when the number of readings is much smaller than 10.

The reasons for using the $(n-1)$ formula are entirely theoretical. They arise especially in statistical sampling theory, and also in procedures like the "analysis of variance" (discussed in Chapters 18 and 19) where the mathematics is far simpler when using $(n-1)$.

By using the divisor $(n-1)$ the value of the variance becomes independent of the number of readings. (In statistical sampling theory this is like having unbiased estimators.) As an illustration consider a small set of 3 readings

$$2, 4, 6.$$

The mean is 4 and the sum of the squared deviations from the mean is

$$(2-4)^2 + (4-4)^2 + (6-4)^2 = 8.$$

Dividing by $(n-1)$ the variance is 4, Dividing by (n) the variance is $2\frac{2}{3}$,

Now, consider a different number of readings of the same kind of data, for example all possible sub-groups of *two* readings from the data:

$$2 \text{ and } 4, 4 \text{ and } 6 \text{ and } 2 \text{ and } 6.$$

These three sets have the following variances:

	Divisor	
	$(n-1)=1$	$n=2$
2 and 4	2	1
4 and 6	2	1
2 and 6	8	4
	4	2

Using the divisor $(n-1)$, the average value of the variances is 4, the same as the variance of the original set of three readings. But using the divisor n, the average variance of the pairs of readings is somewhat smaller than that for the original set when using n there. This result generalises to any set of data.

The theoretical advantage of $(n-1)$ is a very strong one in much advanced work, but in practice one could still approximate and use the conceptually simpler divisor, n. However, even in practice, writing the formula with $(n-1)$ is now very widespread.

Exercise 111. Degrees of Freedom

Can the theoretical considerations for the divisor $(n-1)$ be explained in commonsense terms?

Discussion.

In general, n variable quantities $x_1, x_2, x_3, \ldots, x_n$ can vary in n different ways, i.e. each variable can take any value, independently of the other variables. One can thus say the data has n ways in which to vary, or n "degrees of freedom", a notion due to Sir Ronald Fisher.

If we take the n deviations from the mean $\bar{x}$:

$$(x_1-\bar{x}), (x_2-\bar{x}), (x_3-\bar{x}), \ldots, (x_n-\bar{x}),$$

only $(n-1)$ of these quantities can vary independently. The last one is determined by the others because the deviations from the mean have to add to zero (see Exercise 11C).

As an illustration consider two readings x_1 and x_2. Each can vary any way one likes. Now consider two particular values $x_1=3$ and $x_2=7$. The mean is 5 and the deviations are

$$-2 \text{ and } +2.$$

Given the first deviation of -2, the other *must* be $+2$. Only one of the two deviations can vary independently.

Similarly, with three readings, only two of the deviations from the mean can vary independently. They determine the third deviation. This generalises for any number of readings. Having calculated the mean, we have effectively "used up" the independence of one of the readings.

Since there is no reason to identify one particular reading as being "used up", the same idea is expressed as having used "one degree of freedom" in the whole set of data.

More generally, one degree of freedom has to be subtracted for each constant or coefficient in a model that is fitted to the data. This idea is very helpful in various parts of statistical theory, as we shall see in Parts IV and V. It also provides a kind of reason why the average of the squared deviations from the mean is formed as if there were only $(n-1)$ readings.

CHAPTER 12

Frequency Distributions

Many phenomena are highly irregular when observed at the individual level, but when analysed in groups the patterns often become systematic and generalisable. For example, we cannot predict an individual's income unless we know other facts about him, such as his age, occupation, etc. But for any *group* of people a regular pattern tends to appear. In general *most* will have relatively low incomes, some will have higher incomes, and a small number will have very high incomes, as illustrated in Figure 12.1. This kind of "skew" pattern occurs generally with income data for groups of fairly similar people, and in that sense it is predictable.

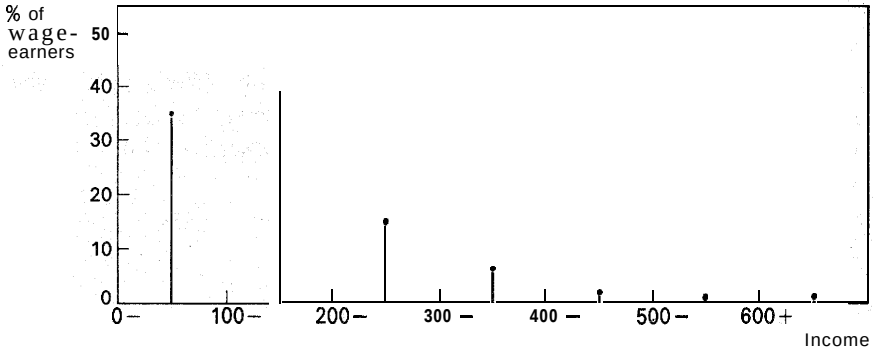

Figure 12.1 Wage-Earners' Incomes (Grouped in Units of 100)

We saw another case of this in the last chapter, with the number of monthly magazine issues read in a year. It may be impossible to predict the number read by any one person, but for any group of people the U-shaped distribution in Figure 12.2 occurs widely. It is the occurrence of such *statistical regularities* which makes the study of individually irregular data important.

When the same type of observed frequency distribution arises often, it is worth modelling by a mathematical formula. Then the theoretical formula can be used to compare and summarise different sets of data.

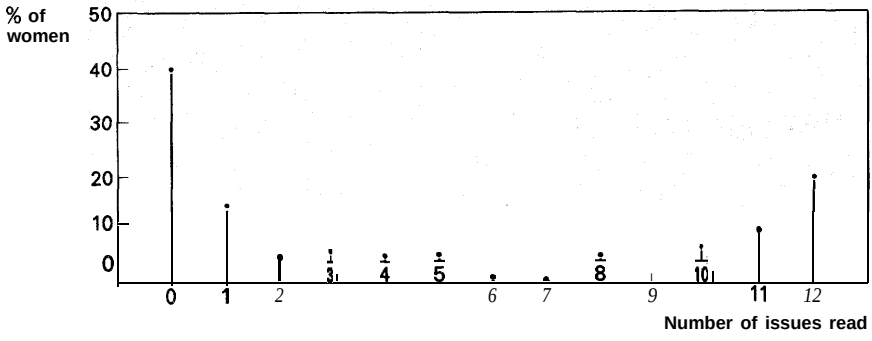

Figure 12.2 The U-Shaped Distribution for Magazine Readership (Figure 11.2)

The number of theoretical frequency distributions in common use is fairly small. This chapter deals with three main cases: the Normal, the Poisson, and various versions of the Binomial. Sometimes the mathematics is fairly complicated, but this need not matter much in practice. One mostly uses numerical tables or computer programmes, and simple verbal shorthand like "it's a Normal Distribution" often effectively summarizes the mathematics.

12.1 **The Normal Distribution**

Figure 12.3 shows a set of readings, the heights of lo-year-old boys, which are grouped symmetrically around a single modal value, where most of the readings lie close to this central point. Such data can often be approximated by a particular mathematical formula called the Normal Distribution.

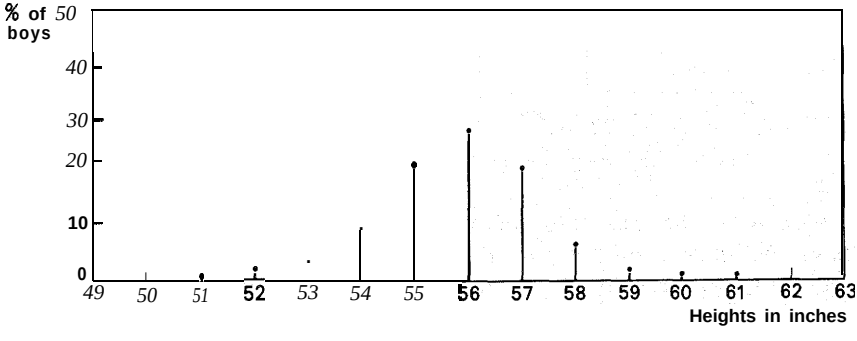

Figure 12.3 The Heights of 10-Year-Old Boys (Figure 11.1 repeated)

If the variability of the observed phenomenon is due to a large variety of independent factors, then an approximately Normal Distribution must

result. This is a very common situation. For example, errors of measurement often follow this distribution. In the 19th Century it was called "the normal curve of errors", and the name Normal stuck. (The distribution was discovered in 1711 by de Moivre in England. Sometimes it is called the Gaussian Distribution, after the 19th Century German mathematician, C. F. Gauss.)

The Normal is an exceptionally simple distribution because it always takes the same shape. Describing this shape in terms of the standard deviation as a measure of scatter, a Normal Distribution has

68% of its readings between ± 1 s.d. from the mean,
95% of its readings between ± 2 s.d. from the mean,
99.7% of its readings between ± 3 s.d. from the mean.

Figure 12.4 shows how a Normal Distribution is composed. Thus mean values and the size of scatter may differ, but just about two-thirds of the readings will ***always*** lie between ± 1 standard deviation from the mean of the data. This makes it unnecessary in practice to refer to the mathematical formula for the distribution (see Exercise 12K).

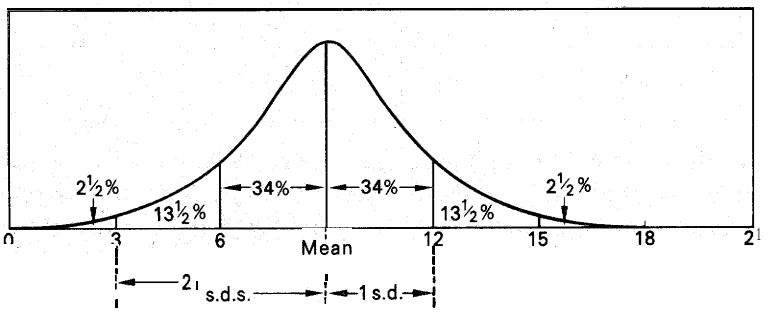

Figure 12.4 A Normal Distribution with Mean 9 and Standard Deviation 3

Theoretically the Normal Distribution ranges from minus infinity to plus infinity. But because only 0.3% of the readings lie more than ± 3 standard deviations from the mean, the distribution can approximate data that lie within some fairly restricted range.

In all Normal Distributions there is a simple relationship between the standard deviation and the mean deviation:

mean deviation = 0.8 standard deviation.

Thus the Normal Distribution can also be described as

58% of the readings between ± 1 m.d. from the mean,
90% of the readings between ± 2 m.d. from the mean,
98% of the readings between ± 3 m.d. from the mean.

The remarkable thing about the Normal curve is that it can be described for most purposes by three values like this, or the equivalent ones for the standard deviation. (In statistical sampling, theory, Chapter 18, the 1% and 0.1% values of $\pm$ 2.6 s.d. and 3.3 s.d. are conventionally referred to as well.) One can therefore fully summarize appropriate data by reporting that it follows a Normal Distribution with a certain mean and standard deviation (or mean deviation).

Numerical Examples

The Normal Distribution approximates so much observed data that it is important to be familiar with its role in summarising and comparing different sets of data. The following numerical examples illustrate this.

Table 12.1 reproduces the 8 readings discussed in the last chapter. The mean was $32/8 = 4$ and the mean deviation $12/8 = 1.5$. Even with such a small and "jumpy" set of data, the Normal Distribution works adequately.

TABLE 12.1 The Frequency Distribution of 8 Readings from Chapter 11

	Value							
	1	2	3	4	5	6	7	8
Frequency	1	0	3	2	0	1	0	1

It is easy to see that 5 out of 8 readings, or about 62%, lie within $\pm$ 1 mean deviation and about 87% of the readings lie within 2 mean deviations of the mean. (A typical grouping problem is whether or not to count the reading of "1" within the 2 mean deviation limit. Such problems can usually be resolved on commonsense grounds.)

In this case the two observed percentages are close to the theoretical values of 58% and 90% for the Normal Distribution, and hence there is useful correspondence. Even with such a small number of readings, the data can therefore be described as being approximately Normal, with a mean of 4 and a mean deviation of 1.5. From this summary the full data could be reconstructed to within a close degree of approximation. About 60 % of the readings must take the values 3, 4 or 5, and so on.

Table 12.2 gives another set of readings. The mean is $180/30 = 6$ and the mean deviation is $48/30 = 1.6$.

These data are also approximately Normal because 16/30 (53 %) of the readings lie within $\pm$ 1 mean deviation and 28/30 (93 %) lie within ± 2 mean deviations of the mean. These are again within a few percentage points of the theoretical figures for a Normal Distribution. The degree of approximation involved is usually insignificant considering the typical practical use of the theoretical distribution in comparing two such sets of readings.

TABLE 12.2 A Set of 30 Readings

	Value									
	1	2	3	4	5	6	7	8	9	10
Frequency	0	0	4	3	6	6	4	3	2	2

Table 12.3 expresses the frequencies in the two sets as percentages, to eliminate the differences in the total numbers of readings. Even so, the direct comparison is still not easy to do. Just how do the two sets of readings differ and in what ways are they similar?

TABLE 12.3 Comparison of the Two Sets of Data from Tables 12.1 and 12.2
(Frequencies expressed in percentage form)

	Value									
	1	2	3	4	5	6	7	8	9	10
8 readings from T. 12.1 %	13	0	37	25	0	13	0	13	0	0
30 readings from T. 12.2 %	0	0	13	10	20	20	13	10	7	7

The fact that both sets of readings are approximately Normal makes the comparison much easier, as shown in Table 12.4. The degree of approximation in the fit of the Normal Distribution does not affect the conclusion that the readings in the second set are generally about 2 units higher than those in the first.

TABLE 12.4 Summary of the Two Sets of Data in Table 12.3

	Distribution (Approximate)	Mean	Mean Deviation
8 readings from T. 12.1	Normal	4	1.5
30 readings from T. 12.2	Normal	6	1.6

TABLE 12.5 Summary of the Two Sets of Data and Comparison with the Normal Distribution

	% of readings within	
	± One m.d.	± Two m. d.
The 8 readings	62%	87%
The 30 readings	53%	93%
The Normal Distribution	58%	90%

Table 12.5 illustrates the similarity of both observed distributions to the theoretical Normal Distribution, after allowing for the difference in means. Saying that both sets of readings are Normal with mean deviations of about $1\frac{1}{2}$, but that their means differ at 4 and 6, clearly communicates the essence of the data.

12.2 The Poisson Distribution

We now turn to data that count the number of times some event occurs in a given interval of time or space. The simplest theoretical distribution for this type of counting data is called the Poisson, named after an 18th Century French mathematician.

Poisson Distributions can arise when the events occur more or less independently of each other with a constant average frequency of occurrence. A classical illustration concerns the number of Prussian soldiers kicked to death by a horse (von Bortkewitsch, 1898). The chance of such a death happening is rare, and on the whole one case will not be connected with another. In records for ten army corps over twenty years there was a total of 122 such deaths, on average .61 per corps per year.

Out of the total of 200 possible annual readings there were 109 cases where no deaths occurred in a corps, 65 cases of one death, 22 cases of 2 deaths, etc. The detailed distribution of deaths can be closely fitted by a Poisson Distribution with a mean of .61, as Table 12.6 shows. We therefore have a succinct summary of the data.

TABLE 12.6 The Numher of Soldiers Killed Annually by Horse-Kicks in a Prussian Cavalry Corps, and the Theoretical Poisson Distribution

(10 Corps over *20* years)

	Number of Deaths per Year						No. of "Corps-Years"
	0	1	2	3	4	5+	
Observed	109	65	22	3	1	-	200
Poisson	109	66	20	4	1	.1	200

Another example of a Poisson Distribution, reported more recently, is the incidence of "major strikes" per week in the United Kingdom from 1948 to 1959 (Kendall, 1961). In that 626-week period there were 563 strikes, or about .90 per week. Table 12.7 shows there were 252 weeks with no strike, 229 weeks with one strike, 109 weeks with two, 28 weeks with three, 8 weeks with four, and none with more. These numbers approximate a theoretical Poisson Distribution with a mean of .90.

Poisson Distributions differ from each other only in one parameter, their mean values, usually denoted by m. A particularly simple property of the

TABLE 12.7 The Fit of the Poisson Distribution for the Occurrence of 'Major Strikes' in the U.K. per Week in 1948-1959

	Number of strikes per week						Total number
	0	1	2	3	4	5+	of weeks
Observed	252	229	109	28	8	0	626
Poisson	254	229	103	31	8	1	626

Poisson is that its variance equals its mean, i.e.

$$\text{mean} = m = \text{variance}.$$

A simple way to estimate the theoretical Poisson frequencies is to equate the number of readings taking the valuer to (m/r) times the number taking the value $(r - 1)$. For example, if the number of weeks with zero strikes is 252 out of a total of 626, the theoretical proportion of weeks with one strike is $252(.90/1) = 227$. The number of weeks with 2 strikes is $227(.90/2) = 102$, and so on. These estimates are almost identical to the theoretical frequencies in Table 12.7. (The figures in the table were calculated by a fractionally more accurate method outlined in Exercise 12F. The easier formula works well if the number of zeros is relatively large, as here.)

The Poisson Distribution can- take different shapes, depending on the value of the mean. It can be reverse-J-shaped, like our two examples, or humpbacked if the mean is high. This makes it impossible to give any routine description in terms of the standard deviation, as one can do with the Normal Distribution. But if the observed mean and variance of a set of data are approximately equal, then it is worth calculating the theoretical Poisson frequencies to see if they fit. For example, with the data on strikes, the observed mean is .90 and the observed variance is .86. These are close enough to indicate that a Poisson Distribution with a mean $m = .90$ should give a good fit.

The Normal Approximation to Poisson. Poisson Distributions with relatively high means take a humpbacked and increasingly symmetrical shape, tending to resemble a *Normal* Distribution. This makes the data even easier to handle.

TABLE 12.8 The Distribution of the Number of Yeast Cells in Each of 400 Squares, and the Poisson Distribution

	Number of Cells														No. of
	0	1	2	3	4	5	6	7	8	9	10	11	12	13+	Squares
Observed	-	20	43	53	86	70	54	37	18	10	5	2	2	-	400
Poisson	4	17	41	63	74	70	54	36	21	11	5	2	1	1	400

Table 12.8 gives an example with a mean of about 5. It sets out the number of yeast cells counted in each of 400 squares into which a square millimetre was divided (Fisher, 1950). The observed mean is 4.7 and the variance is 4.5, so the observed distribution should closely fit a Poisson, which it does.

The observed distribution has a mode at 4 which falls close to the mean at 4.7 as does the median (with 202 readings below 4.7 and 198 above it). These characteristics indicate that a Normal Distribution should also fit the data. To check this quickly, we can calculate the theoretical standard deviation by taking the square root of the mean, which equals the variance in a Poisson. This is 2.2. (The standard deviation of the data could have been calculated directly, giving the value 2.1, but theory is faster!) According to the main characteristics of a Normal Distribution we should have

		Normal	Observed
± 1 s.d., i.e. between	2.5 to 6.9:	68%	263/400 = 66%
± 2 s.d., i.e. between	0.3 to 9.1:	95%	391/400 = 98%
± 3 s.d., i.e. between	− 1.9 to 11.3:	99.7%	398/400 = 99.5%

These observed data can therefore be adequately modelled by either the Poisson or the Normal Distribution, even though the Poisson is discrete (taking whole numbers only) and the Normal is continuous (taking fractional numbers also). This is typical of the way different kinds of approximate solutions can be used.

A variety of natural phenomena follow the Poisson Distribution to a fair degree of approximation. Apart from our examples, the Poisson describes certain cases in cosmic radiation in physics, breakdowns in telephone equipment, and certain other forms of "accident". However, in many situations the simple conditions for the Poisson Distribution, that successive events occur independently of each other but at a constant average rate, do not quite hold. In fact, the very simplicity of the Poisson Distribution with its single parameter makes it rather inflexible. The main importance of the Poisson is as a constituent of more complex models.

12.3 The Negative Binomial Distribution

Not all cases of accidents follow a Poisson Distribution because every person does not have the same chance of an accident. For example, in the case of the Prussian soldiers, if some belonged to cavalry corps and others to infantry corps, they would have different exposure to horses and *systematically* different degrees of risk. The incidence of deaths then might *not* follow a Poisson Distribution because the corps might have different mean number of deaths over the years, and the Poisson requires a constant mean rate.

Again, some people might be accident-prone, *inherently* pre-conditioned to have more accidents than others, regardless of how many accidents they had previously. Alternatively, having one accident might make a person more likely to have another, which is called a "contagious" or "learning" type of phenomenon, because something like it also occurs in the spread of certain diseases and in psychological learning situations. In certain circumstances both these situations could lead to a frequency distribution called the Negative Binomial Distribution, discovered by de Montmort in about 1700. (The name Negative Binomial stems from a mathematical technicality which usually has no direct practical implications.)

One example occurs with consumer purchases of frequently bought branded goods. Table 12.9 shows the number of times a sample of housewives bought Corn Flakes in a given six-months period (Charlton *et* ***al.,*** 1972). Thirty-nine per cent of the sample did not buy Corn Flakes at all, 14 % bought once, 10 % bought twice, and so on.

TABLE 12.9 The Numbers of Purchases of Corn Flakes Made in 24 Weeks
(Number of Households out of a Sample of 491)

Number of Purchases	0 1 2 3 4	5 6 7 8 9	10 11 12 13
Households buying	19371 49 28 20	22 14 16 11 9	12 7 7 4
Number of Purchases	14 15 16 17 18	19 20 21 22 23	24 29 36 37
Households buying	1 3 5 4 2	2 2 2 2 1	2 1 1 -

The basic statistics (to 3 digits for working purposes) are that the mean = 3.43 and the variance = 26.9. Since the variance is so much bigger than the mean, the distribution certainly is not Poisson. However, the Negative Binomial Distribution (NBD) gives a good fit, as shown in Table 12.9a. This occurs very generally for such purchasing data (e.g. Ehrenberg, 1972).

TABLE 12.9a The Fit of the Negative Binomial Distribution

24 Weeks		Number of Purchases										
		0	1	2	3	4	5	6	7	8	9	10+
Cornflakes buyers	Observed %	39	14	10	6	4	4	3	3	2	2	12
	NBD %	35	16	10	7	6	5	3	3	2	2	11

The Negative Binomial Distribution is either reverse-J-shaped, as here, with most of the readings at 0, some at 1, less at 2, etc., or, if the mean is large, humpbacked with a long positive tail to the right. If the mean of the NBD is denoted by m, the variance is defined as

$$\text{variance} = m(1 + m/k),$$

where the quantity k is always positive but generally not integral. Thus the variance in an NBD is always larger than the mean.

The value of the parameter k can be calculated by equating the variance of the observed distribution to the expression $m(1+ m/k)$. For the data in Table 12.9 this gives $k = SO$. Given the value of k, one can estimate the theoretical frequencies of the NBD with an easy-to-use formula. It relates the proportion p_r of r occurrences to the proportion p_{r-1} of $(r-1)$ occurrences :

$$p_r = \left(\frac{m}{m+k}\right)\left(1-\frac{1-k}{r}\right)p_{r-1}.$$

Using the data in Table 12.9 as an example, we can estimate p_1, the frequency of *single* purchases of Corn Flakes, from $p_0=.39$, the proportion of nonbuyers, by computing

$$p_1 = \left(\frac{3.43}{3.43+.50}\right)\left(1-\frac{1-\cdot 50}{1}\right).39 = .17 \text{ or } 17\%,$$

and so on for p_2, p_3, etc. (This works well only if the observed p_0 is reasonably large.) The results are fairly similar to those in Table 12.9a, which were calculated by a somewhat more complex method (discussed in Exercise 121).

The Negative Binomial Distribution has had a fairly wide range of practical applications, e.g. in analysing bus drivers' accidents and in the ecological distribution of the number of species of a particular type of plant. In the case of consumer purchasing, the underlying explanation seems to be proneness rather than a learning type of phenomenon. That is, some households are consistently more likely than others to buy Corn Flakes, instead of each purchase of Corn Flakes increasing the tendency to buy it again. Underlying models will be discussed further in Chapter 13. Here we are concerned with using such mathematical distributions simply to describe and summarise observed frequency distributions.

12.4 **The Binomial Distribution**

Theoretically, the number of occurrences or events being counted in the Poisson and Negative Binomial Distributions can extend indefinitely to plus infinity, but in most cases the theoretical frequency of high numbers is very small. Thus observations extending only to some finite upper limit can also be accommodated by the distributions.

However, there are situations where the maximum number of occurrences has a clear-cut and attainable upper limit. For example, the number of patients who improve in a clinical trial of a drug cannot exceed the total number of patients in the trial.

Such data can often be represented by the Positive Binomial Distribution, usually referred to simply as the Binomial Distribution. (Sometimes this is also called the Bernouilli Distribution, after the 17th Century Swiss mathematician Jacob Bernouilli.) The phrase "bi-nomen" means "two names". In a binomial classification the observed items are sorted into two categories, such as Success or Failure, Boy or Girl, Yes or No, etc.

A more general case is the Multinomial Distribution, where the classification consists of more than two classes: e.g. single, married, widowed or divorced ; or Yes, No, or Don't Know responses to a question ; etc. The statistical theory of Multinomial Distributions is similar to the Binomial. The basic mathematics used is permutations and combinations. (The resemblance between the Positive and the *Negative* Binomial Distributions lies only in the form of certain basic mathematical formulae.)

An example of a Binomial Distribution is the incidence of boys and girls in families of a given size. If the sex of successive children in a family is determined independently for each child and there is no difference in the tendency towards boys or girls among different families, then the proportion of families with n children who have r boys should be given by the Binomial formula

$$\frac{n!}{(n-r)!r!}p^r q^{n-r}.$$

Here p is the overall proportion of boys and $q = (1-p)$ is the overall proportion of girls. (The "factorial" symbol ! in $n!$ stands for the product $n(n-1)(n-2)\ldots 3 \times 2 \times 1$. Similarly $(n-r)!$ stands for $(n-r)(n-r-1)(n-r-2)\ldots 3 \times 2 \times 1$. The expression $0!$, e.g. the value of $(n-r)!$ when $r = n$ boys, equals 1.)

Table 12.10 gives data on the observed percentages of some 3-children families who had either 3, 2, 1, or no boys (Geissler, 1889). The overall proportion of boys is $p = .51$, and n equals 3. The theoretical values calculated by Geissler from the above formula fit well (based on his value of $p = .5147676$, and multiplying by 100 to give percentages).

Table 12.11 illustrates the distributions for $n = 1$, 3, and 5. With one-child families, 51% are boys, but with 5-child families, only 3.7 % are all boys.

TABLE 12.10 The Binomial Distribution of the Number of Boys in 3-children Families

	Number of Boys *in* Family				Average Proportion of Boys
	3	2	1	0	
Observed %	14	38	36	12	.51
Binomial %	14	39	36	11	.51

While the three observed distributions are very different, they can all be closely fitted by theoretical Binomial Distributions with a proportion $p = .51$ (or Geissler's value 5147676). Only the value of n varies. The data can therefore be succinctly summarised as being Binomial with an average of 51% boys.

TABLE 12.11 The Distribution of Boys in 1-, 3- and 5-child Families and the Theoretical Binomial Frequencies
(Expressed as Percentages)

Number of Children in Family	Number of Boys in Family						Average Percentage of Boys
One			One	None			
Observed %			51.5	48.5			51
Binomial %			51.5	48.5			51
Three		Three	Two	One	None		
Observed %		13.9	38.3	36.2	11.6		51
Binomial %		13.6	38.6	36.4	11.4		51
Five	Five	Four	Three	Two	One	None	
Observed %	3.7	16.7	32.2	30.5	14.0	2.9	51
Binomial %	3.6	17.0	32.1	30.3	14.3	2.7	51

The Binomial Distribution depends on the two parameters p and n. These differ in their status. The size of the individual groups being examined, n, is generally determined before the observations are collected, whereas p is determined by the data itself. If the data's two categories (boy : girl, success : failure, etc.) are scored 1 and 0, the mean and variance of the distribution are

$$\text{mean} = np,$$

$$\text{variance} = npq.$$

Since, q is a proportion less than 1, the variance of a Binomial is always less than its mean.

When the parameter p is at or near .50, the distribution will be approximately symmetrical. But when p markedly differs from .50, the Binomial Distribution is quite skew. For example, with an industrial batch production process, each batch might consist of 20 items, $n = 20$, and an average of 10% of the items might be faulty, so $p = .1$. If one fault occurs independently of another and with a consistent probability of .1, the theoretical Binomial Distribution shown in Table 12.12 will result.

TABLE 12.12 The Binomial Distribution for Faults in Batches of n = 20 Items, with on average p = . 1 Faults

	Number of Faults per Batch								
	0	1	2	3	4	5	6	7	8-20
% of all batches	12	27	29	19	9	3	.9	.2	.05

Twelve batches per 100 will be free of faults, 27 will contain 1 fault, and so on, with virtually no batches (5 in 10,000) having 8 or more faults. This kind of result is used in industrial quality-control inspection schemes.

The Poisson and Normal Approximations to the Binomial

In certain cases the Binomial Distribution approximates the Normal and Poisson Distributions. When the Binomial parameter p is very small, for rare events, then $q = (1 - p)$ will be nearly 1. Thus the variance of the Binomial Distribution, npq, will be virtually equal to its mean, np. This is characteristic of the *Poisson* Distribution and in such cases the Binomial closely approximates the Poisson.

When the Binomial parameter n is large, or even for small n when the proportion p is near the .50 mark, the distribution is approximately symmetrical. Then it can usually be well represented by a *Normal* Distribution. (It was in this context that de Moivre first discovered the Normal Distribution, one of the fundamental steps in theoretical statistics.) For example, for families with $n = 5$ children in Table 12.11, the mean is about 2.6 and the theoretical standard deviation is $\sqrt{(5 \times .51 \times .49)} =$ about 1.1. About 63 % of the readings lie within one standard deviation of the mean and 93 % lie within two standard deviations. This compares with the Normal values of 68 % and 95 %. In such cases, one can use the simpler calculations of the Normal Distribution to describe the data.

12.5 The Beta Binomial

Close examination of Table 12.11 shows that in the 3- and 5-child families, the observed numbers of *all* boys or *all* girls are slightly higher than the theoretical values. In his classic book *Statistical Methods* (1950), Fisher quoted the data for S-child families (shown here in Table 12.13) and noted a similar excess. The original data (Geissler, 1889) show that this excess in fact generalises for all family sizes, i.e. the discrepancies are systematic. Thus the fit of the Binomial Distribution to these data is close, but not perfect.

TABLE 12.13 The Distribution of Boys in 8-child Families

	Number of Boys in Family								
	8	7	6	5	4	3		210	
Observed %	.6	3.9	12.4	22.2	27.9	19.8	9.9	2.7	.4
Binomial %	.5	3.7	12.2	23.1	27.2	20.5	9.7	2.6	.3

The Binomial Distribution should fit if the sex of each baby is determined independently and if the average incidence of boys shows no systematic trends (e.g. between different types of families, first-born and later children, winter and summer babies, etc.). But it is not self-evident that the sex of babies behaves like this. For example, the sex of a first child might have affected the chemical balance of the mother's hormones, in turn affecting the conception or survival of a subsequent baby of the opposite sex. Or there might have been social or economic pressures to have at least one boy, leading to a higher proportion of boys in ***smaller*** families. The evidence does not suggest that either of these possibilities occurred in this data, but the excess of all-boy or all-girl families shows that *some* special factor was at work.

One possible factor is the incidence of identical twins, but Fisher noted these did not occur frequently enough to account for all the observed discrepancies. Another possibility is that the underlying incidence of boys might vary among families. It could be .60 for one, .52 for another, .43 for a third, and so on, instead of .51 for all.

If this were true, the sex of babies could still be determined independently and at a constant proportion *within* ***eachfamily, so*** that the incidence of boys and girls would still follow a Binomial Distribution for each family. But the required model would then consist of a "mixture" of Binomial Distributions, each with a different proportion ***p.*** These different values of ***p*** would then follow a frequency distribution across different families. If the distribution of p-values were of the so-called "Beta" type, the resulting distribution of the number of boys would be of the ***Beta-Binomial*** form (also known as the Negative Hypergeometric Distribution).

Table 12.13a shows the fit of the Beta-Binomial to the data for S-child families.

TABLE 12.13a The Fit of the Beta-Binomial for the 8-child Families

	Number of Boys in Family								
	8	7	6	5	4	3	2	1	0
Observed %	.6	3.9	12.4	22.2	27.9	19.8	9.9	2.7	.4
Beta-Binomial %	.6	4.0	12.5	22.9	26.7	20.4	9.9	2.8	.4

To one decimal place there is no excess of all-boy or all-girl families, but working to two decimal places there is still a small systematic excess of about .04, that occurs consistently for all family sizes from 2 to 12. This must be due to an additional factor (perhaps now identical twins). The Beta-Binomial therefore provides an improved, but still not perfect model for the data.

The Beta-Binomial Distribution has three parameters (see Exercise 12J for technical details) which makes it more flexible than the simple Binomial. Depending on the values of the parameters, the Beta-Binomial can take a variety of different shapes. For example, it can be reverse-J-shaped with a mode at 0, it can be humpbacked, or it can be U-shaped with two modes at 0 and *n*. Such a variety of shapes can occur in the same empirical context, e.g. for the number of different episodes of a television programme which viewers see (Goodhardt *et al.*, 1975).

Table 12.14 illustrates a U-shaped distribution. The readings show how many different issues of the weekly magazine *Woman* were bought in 12 weeks.

TABLE 12.14 A U-Shaped Distribution and the Fit of the Beta-Binomial

	Number of Issues Bought												
	0	1	2	3	4	5	6	7	8	9	10	11	12
Observed %	85	4.2	1.7	.8	.4	.3	.2	.4	.3	.5	1.1	1.2	3.8
Beta-Binomial %	87	2.0	1.1	.8	.7	.6	.5	.6	.6	.6	.7	1.1	3.8

The distribution is rather a *skew* U: 85% of the sample bought none, 4% bought 1 issue, 1.7% bought 2, and less than 1% bought from 3 to 9 issues. Then the frecluencies increase again, with 1.1%, 1.2%, and 3.8% buying 10, 11 or all 12 issues. Such a pattern is very common in readership data. (People either buy all or almost all of the issues, or they buy none or almost none; very few readers buy about half.) None of the other theoretical distributions discussed in this chapter can describe such U-shaped data. While the Beta-Binomial gives a close fit there are some significant discrepancies. For example, more people bought 1 or 2 copies than the model predicts. Thus, the data are more complex than even a relatively sophisticated distribution like the Beta-Binomial can fully describe. There may be another, yet more complicated, mathematical function which could fit such data better.

But there is also another approach to such data. Perhaps some of the complicating factors could be handled by more *direct* analysis, instead of by fitting more comphcated mathematical models. For example, there are two kinds of purchasers, those buying regularly by subscription and those buying occasional copies at newsagents or newstands; such groups could be analysed separately. Again, whilst the Beta-Binomial could be expected to hold if all twelve issues sold the same number of copies, there may have

been one or two "bumper issues" in the 12-week period, explaining the "excess" of purchasers of one or two copies.

12.6 Other Distributions

There are many other mathematical distributions that can be fitted to empirical data, but few have been applied as widely as the examples already described.

One distribution of particular theoretical importance is the Gamma-Distribution. This is related to the Beta-Distribution (which is the distribution of the ratio of two Gamma variables), and has various applications in the theory of statistical sampling and for "stochastic" models of data. This will be discussed in later chapters.

Sometimes the scale of measurement of an observed variable can be transformed so that apparently complex data can be modelled by one of the simpler distributions. The main example is the Lognormal Distribution (e.g. Aitcheson and Brown, 1957). This arises with certain kinds of skew data with only a few high values (like income distributions), where logarithms of the readings may follow a Normal Distribution.

12.7 Summary

Statistical frequency distributions are theoretical formulae that describe observed distributions of readings. They are particularly useful when the same form of distribution occurs for different sets of data,

The most widely used distribution is the Normal, which describes symmetrical, humpbacked distributions. It is exceptionally simple because it always takes the same shape, e.g. 68 % of the readings lie within ± 1 standard deviation of the mean.

The Poisson, Binomial, and Negative Binomial Distributions are useful in cases where the occurrence of an event is counted. They refer to data with different degrees of scatter. The variance is equal to the mean for the Poisson Distribution, but it is always smaller than the mean for the Binomial and larger than the mean for the Negative Binomial.

CHAPTER 12 EXERCISES

(Exercises 12F onwards deal with relatively technical matters in fitting frequency distributions.)

Exercise 12A. The Use of a Distribution

What is the point of fitting complicated formulae like the Negative Binomial Distribution?

Discussion.

The pay-off comes when the same formula describes different sets of data. To illustrate, Table 12.15 gives the distributions of purchases of Corn Flakes and Puffed Wheat over 12-week and 24-week periods (Charlton *et al.*, 1972).

TABLE ***12.15*** Frequency Distributions of Purchase of Corn Flakes and Puffed Wheat in Different Length Time-Periods

(% of households buying)

	Number of Purchases											
	0	1	2	3	4	5	6	7	8	9	10+	Total
Corn Flakes												
% buying in 24 weeks	39	14	10	6	4	4	2	3	2	2	14	100
" " " 12 "	51	15	8	6	5	5	2	3	1	2	2	100
Puffed Wheat												
% buying in 24 weeks	84	9.6	2.4	1.0	.6	.6	.4	.2	0	.2	1.0	100
" " " 12 "	90	6.3	1.4	.6	0	.4	.2	.2	0	.4	.2	100

The four distributions differ markedly, yet they can all be fitted by an NBD (as was illustrated in Table 12.9a for the 24-week Corn Flakes data). Only the means and variances differ, as Table 12.16 shows. Since the occurrence of the NBD for purchasing data is a very general finding (e.g. Ehrenberg, 1959, 1972), these two parameters are all one needs to summarise and distinguish the different sets of data.

TABLE 12.16 The Means and Variances of the Four Distributions

	Distribution (approx.)	Mean	Variance
Corn Flakes			
24 weeks	NBD	3.4	27
12 weeks	NBD	1.8	8
Puffed Wheat			
24 weeks	NBD	.4	2.7
12 weeks	NBD	.2	.9

Exercise 12B. The Parameters of a Distribution

Although the means of the NBD distributions in Table 12.16 are not "typical" values, they do have a physical meaning. They represent the relative sales levels or market-shares of the brands. But the *variances* of these skew distributions have no descriptive meaning. Are there other characteristic values which would be more useful in describing the data?

Discussion.

The parameters of a distribution are values which serve to identify a particular distribution. Different distributions of the same type can thus

be distinguished by the numerical values of these parameters, e.g. two Normal Distributions can have different means and/or variances. (Different *types* of distributions, e.g. a J-shaped Poisson and a symmetrical Normal, are more difficult to compare. It is like trying to compare a straight line with a curve.)

The NBD has two parameters, in the sense that two numbers will differentiate one NBD from another. But one does not have to use the mean and variance: various other aspects of the data can also be used. Table 12.17 sets out three alternatives for the breakfast cereal data.

The first is the percentage of the sample buying at all (i.e. the penetration "b" from Chapters 9 and 10, or 100 minus the percentage of zeros) and the second is the average number of purchases per buyer, or w. These parameters give one a good descriptive "feel" of the data-how many people buy the item at all in the time-period, and how often on average they do so.

TABLE 12.17 Other Parameters of the Four Distributions

	% of sample buying at all	Av. number of purchases per buyer	k
Corn Flakes			
24 weeks	61	5.6	.50
12 weeks	49	3.7	.55
Puffed Wheat			
24 weeks	16	2.6	.08
12 weeks	10	2.2	.06

Furthermore, these parameters follow useful relationships. Multiplying b by w for a brand reproduces its overall mean level of purchases, as given in Table 12.16 (on a per 100 household basis). Again, the values of b and w for different brands are linked, because $w(1-b)$, or $w(100-b)$ for percentages, is approximately constant as we saw in Chapter 10. Finally, the results in time periods of different lengths are linked, as will be shown in Chapter 13.

The third parameter, k, is a more abstract quantity that arises from the mathematical formula for the frequency p_r of the NBD. But it too has a useful descriptive property. It appears that for a given brand the value of k hardly varies in different length time-periods. It is about .5 for Corn Flakes and about .07 for Puffed Wheat (the variation in the values is small compared with the other differences in the data). This is a general property of k which makes it a very simple parameter to use; generally only *one* k-value has to be specified for each brand, irrespective of the length of time-period analysed.

Thus different parameters can have different descriptive advantages. Each type of distribution has a minimum number of parameters that need to be determined in order to "fix" it. But one can use more than this minimum number for different purposes and these numerical values will then be interrelated.

Most of the common distributions have a minimum of two parameters. The Poisson is unusual because it is fully specified by a single characteristic value. This could be the mean m, or the variance (which is equal to m) or the proportion of zeros (e^{-m}). Clearly, all these different values are mathematically equivalent to each other.

The fact that the Poisson Distribution has only one parameter makes it particularly simple, but also rather inflexible. It can only fit if the observed variance is (approximately) equal to the mean. In the NBD, the variance is always *greater* than the mean, and the parameter k determines this difference since the variance is equal to $m(1+m/k)$.

In the ordinary (or positive) Binomial Distribution, the variance, npq, is always *less* than its mean, np, by a factor q, which depends directly on the mean since $q = 1 - p$. Thus the Binomial Distribution is also relatively inflexible. There are not many practical situations, outside artificial games of chance, where it gives a good fit. The Binomial has its widest applicability as an ingredient of more general distributions, such as the Beta-Binomial in Section 12.5.

Exercise 12C. Deviations from a Model

Discuss the use of frequency distributions in summarising the deviations of observed data from a theoretical model.

Discussion.

Summarising the irregular deviations from a model is one of the most common uses of statistical methods. Examples occurred with the quarterly readings in the four areas in Chapters 1 and 2, and with the deviations of the age-group means from the relationships like log w $= .02h + .76$ in Part II.

Such deviations can commonly be summarised by the Normal Distribution. The deviations are often due to a large variety of independent factors or "errors", which is the situation in which the Normal Distribution tends to arise.

As an example, the 30 quarterly deviations for 1969 and 1970 in Tables 2.2 and 2.2a (excluding the two exceptional QIII values) were

-7, -6, -5, -5, -4, -3, -3, -3, -3, -2, 2, -1, -1, -1, -1, -1, 0, 0, 1, 1, **2, 2,** 3, 3, 3, 4, 5, 5, 6, 6.

They have a mean of 0 and a mean deviation of 3. If they follow a Normal Distribution, we would expect about 90% of the readings to lie within ± 6 and 58% to lie within ± 3. Because of rounding, the deviations are grouped at integer values, and there is also a particular clustering at -3, at -1, and at $+3$. This has to be allowed for by "smoothing" the data. Thus the percentage of observed values lying between ± 6 are 97% and 87% (depending on whether the 6's are included or excluded), an average of 92% which is close to 90%. Similarly, the observed percentages between ± 3 are 67% and 43%, averaging 55%, which is close to the theoretical 58%.

Given that direct empirical analysis has shown the deviation to be apparently *irregular*, a description of the deviations as approximately following a Normal Distribution with mean 0 and mean deviation 3 would therefore allow one to reconstruct the data rather closely.

Exercise 12D. Exceptional Deviations

How can exceptionally large deviations be dealt with?

Discussion.

If the deviations from a model follow a generalisable pattern such as a Normal Distribution, this can be used to judge values which appear exceptional.

Thus in Chapter 1, two readings (for QIII in the East and West) gave deviations of 25 and 27 which were about 8 and 9 times the mean deviation of 3 of the remaining 30 readings. Since with a Normal Distribution only 1 in 1,000 readings lie more than even just *four* mean deviations from the mean, the data can no longer be described as being approximately Normal.

This does not prove that the two readings are necessarily wrong. But they *are* exceptional. It is easier to describe the data by saying that there are 30 readings which are approximately Normal (as usually happens) with a mean zero and mean deviation 3, plus the two large exceptions.

Only *exceptionally* large deviations need to be reported separately. A "border-line" deviation, say 3 or 4 times the mean deviation, would not markedly affect the Normal approximation and therefore need not be excluded. If there are more than "a few" exceptions they have themselves to be summarised statistically, unless they form generalisable patterns (e.g. that QIII in the East and West is *always* about 25 units high, every year).

Exercise 12E. Fitting the Positive Binomial Distribution

Illustrate the numerical calculation for the Binomial Distribution.

Discussion.

For a Binomial Distribution with parameters n and p, the proportion of readings taking the value r is

$$\frac{n!}{(n-r)!r!}p^r q^{n-r}.$$

There is no short-cut to working out these values, except that one can usually use the Normal approximation for large n and the Poisson approximation for very small p.

In the past, extensive tables were published giving the Binomial proportions for different values of n and p, but now the proportions are usually generated as needed by simple computer programmes. However, working out small examples by hand helps to provide better understanding.

As such an example, we shall calculate the theoretical proportions for 3, 2, 1, and 0 boys in three-children families for the observed data in Table 12.11. Because the Binomial formula involves many multiplications of p by q, rounding-off errors tend to build up. It is therefore better to work with p-values to 3 or 4 digits in the detailed calculations.

Taking Geissler's value of $p = .5148$ (rounded from .5147676 to 4 digits), and remembering that both O! and any number raised to the power of 0 are equal to 1, we have for the theoretical proportion of families with

3 boys

$$p_3 = \frac{3!}{0!3!}(.5148)^3(.4852)^0$$

$$= \frac{3 \times 2 \times 1}{1(3 \text{ x } 2 \text{ x } 1)} \text{ x } .1364 \text{ x } 1$$

$$= .136, \text{ or } 13.6\%.$$

The proportion of families having 2 boys is

$$p_2 = \frac{3!}{1!2!}(.5148)^2(.4852)^1$$

$$= \frac{3 \times 2 \times 1}{1(2 \text{ x } 1)} \text{ x } .2650 \text{ x } .4852$$

$$= 386, \text{ or } 38.6\%.$$

Similarly, $p_1 = 36.4\,\%$ and $p_0 = 11.4\,\%$. A simple check of the calculations is that the sum of the proportions, $p_3 + p_2 + p_1 + p_0$, should equal 1.

Exercises 12F onwards deal with relatively technical matters

Exercise 12F. The Mathematics of the Poisson Distribution

What is the mathematical formulation of the Poisson Distribution?

Discussion.

In Section 12.2 we noted that in a Poisson Distribution with mean m, the proportion of readings taking the value r is m/r times the proportion taking the value $(r-1)$. We can write this as the "recurrence formula"

$$p_r = \frac{m}{r} p_{r-1}.$$

It follows that if the proportion of zeros $p_0 = \text{k}$, some empirical constant, then the various proportions p_0, p_1, p_2, etc. must take the form

$$p_0 = \text{k},$$

$$p_1 = m\text{k},$$

$$p_2 = m^2\text{k}/2,$$

$$p_3 = m^3\text{k}/3 \cdot 2 \cdot 1 = m^3\text{k}/3!$$

$$p_r = m^r\text{k}/r!, \text{ etc.}$$

The sum of all the proportions $p_0, p_1, \ldots$ must be unity, so that $\Sigma\,(\text{k} + m\text{k} + m^2\text{k}/2 + \cdots + m^r\text{k}/r! + \cdots) = \text{k}\,\Sigma\,(1 + m + m^2/2 + m^3/3! + \cdots + m^r/r! + \cdots) = 1$, The series in brackets is well-known in elementary algebra as the exponential series. The sum equals the expression e^m, where e is an absolute constant, approximately 2.718, which arises in certain kinds of mathematics (e.g. in connection with logarithms).

It follows that $ke^m = 1$, so that k must equal e^{-m}. The Poisson frequencies are therefore

$$p_0 = e^{-m},$$
$$p_1 = m\,e^{-m},$$
$$p_2 = m^2\,e^{-m}/2,$$
$$p_3 = m^3\,e^{-m}/3!, \quad \text{etc.},$$

with the rth term being

$$p_r = m^r\,e^{-m}/r!.$$

The mean value of the Poisson Distribution, i.e. the readings 0, 1, 2, etc. multiplied by the proportion of times they occur, is therefore given by

$$(0 \times e^{-m} + 1 \times m\,e^{-m} + 2 \times m^2\,e^{-m}/2 + 3 \times m^3\,e^{-m}/3! + \cdots + r \times m^r\,e^{-m}/r! + \cdots)$$
$$= (0 + m\,e^{-m} + m^2\,e^{-m} + m^3\,e^{-m}/2! + \cdots + m^r\,e^{-m}/(r-1)! + \cdots)$$
$$= m\,e^{-m}(1 + m + m^2/2! + \cdots + m^{r-1}/(r-1)! + \cdots),$$

taking $m\,e^{-m}$ outside the brackets. The new terms inside the brackets are again an exponential series and hence add to l^m. Therefore the mean of a Poisson is $m \times 1$, i.e.

$$m.$$

To determine the variance of the Poisson Distribution, we need to calculate $\Sigma(r-m)^2 p_r$ for all values of r from 0 upwards. By arguing along the lines of Exercise 11E, we can show this equals $\Sigma(r^2 p_r) - m^2$. The average of $r^2 p_r$ can be seen to be $m^2 + m$, if we write it as $\{r(r-1)+r\}p_r$ and work along the same lines as we did for the mean in the previous paragraph. Thus the variance of the Poisson is $m^2 + m - m^2$, i.e.

$$m.$$

Exercise 12G. Calculating the Poisson Frequencies

What is the best way to calculate the numerical values of the Poisson frequencies, $p_r = m^r\,e^{-m}/r!$?

Discussion.

The only complex part of the Poisson formula is the exponential expression e^{-m}, where e = 2.718. This can be worked out using logarithms. For example, if $m = .61$, as in the case of the Prussian soldiers (Table 12.6), we look up the logarithm to base 10 of *2.718,* which is .434 and multiply it by $-m = .61$, giving $-.265$. The antilogarithm of this number (written in the usual logarithm form as i.735) is given in logarithmic tables as .543. Once we have the value of e^{-m}, the rest follows simply. We can note that $p_0 = e^{-m}$ and multiply by m/r for successive values of r. Thus $p_1 = .61 \times .543 = .331$, $p_2 = .61 \times .331/2 = .1O1$, etc.

One can shorten the calculations by using a table like Table 12.18, giving values of e^{-m} for selected values of m. Interpolation for intermediate values of m is easy because of the additive property of exponents.

TABLE 12.18 Values of e^{-m}

m	e^{-m}	m	e^{-m}	m	e^{-m}
.01	.990	.1	.905	1	.368
.02	.980	.2	.818	2	.135
.03	.970	.3	.741	3	.050
.04	.961	.4	.670	4	.018
.05	.951	.5	.607	5	.007
.06	.942	.6	.549	6	.002
.07	.932	.7	.497	7	.001
.08	.923	.8	.449	8	.000
.09	.914	.9	.407	9	.000

Thus for $m = .61$, we can write

$$\begin{aligned} e^{-0.61} &= e^{-0.6} \times e^{-0.01} \\ &= .549 \times .990, \\ &= .544, \end{aligned}$$

which is the same as before (within rounding errors).

We used a slightly different method to fit the Poisson Distribution to the strike data in Section 12.2. There we simply started with the observed number of zeros and multiplied this by m to obtain an estimate of p_1, by $m/2$ to obtain an estimate of p_2, and so on. This method is easier because it avoids calculating e^{-m}, but it gives slightly different results.

A theoretical distribution rarely fits perfectly (especially with sample data). Thus the two methods of fitting will not give identical results because the observed and theoretical numbers of zeros will not be exactly equal. The quicker recurrence formula gives estimated frequencies p that do not quite add to 1.0. This method is simpler but not as accurate, and only works well if the proportion of zeros is high.

The shorter method can usually not be used for Poisson data with a mean above 1 (like that in Table 12.8) because the theoretical number of zeros is small and the *observed* number might even be nought. In such cases it is better first to calculate e^{-m} from the mean and $e^{-m}m^r/r$ for some value of r near the mean, and then to use the recurrence formula for both higher and lower values of r. This reduces the effects of rounding-off errors. When the mean is much larger than 1, the Normal approximation can be used, as with the yeast-cells data in Table 12.8.

Exercise 12H. The Exponential Distribution

What is the time interval between successive events in a Poisson Distribution?

Discussion.

The Poisson Distribution gives the frequencies with which different numbers of events will occur in time intervals of a given length. Since this

number varies (sometimes 2 events a week, sometimes none, sometimes 1 or 3, etc.) the amount of time between *successive* events must also vary. This is an important concept in many practical applications, e.g. for queues (the "waiting-time" for patients at a hospital or for aircraft landing at airports), breakdowns in equipment (the "life" of electronic equipment), learning processes (the time taken to learn particular repetitive tasks), and reaction times in chemistry.

It can be shown mathematically that if the occurrences follow a Poisson Distribution with mean m, the time interval t from any given instant till the next event follows a distribution of the form

$$1 - e^{-tm}.$$

This is the Exponential Distribution, which is so-called because the variable occurs as an *exponent* (here of the number e). The distribution of time intervals between successive events can then be deduced; the *average* of the distribution is l/m, and its variance $1/m^2$.

The particular characteristic of exponential functions is that they transform *additive* properties into multiplicative ones. Exponential distributions occur for example in biological growth situations ("exponential growth"), where in successive equal time intervals things often grow proportionately to their size.

Exercise 121. The Negative Binomial Distribution

Why are there different ways of estimating the theoretical frequencies of the Negative Binomial Distribution?

Discussion.

In Section 12.3 we calculated the NBD parameter k by equating the variance of the observed distribution to the theoretical variance formula $m(1 + m/k)$, where m was the observed value of the mean. But the theoretical frequencies in Table 12.9 were calculated by using a different and somewhat more complex method of estimating k.

Different fitting methods will give the same results if the observed data follow the theoretical distribution *exactly*. But this rarely occurs. Most theoretical models are at best close approximations to the data, and with *sample* data additional fluctuations occur.

Fitting distributions by the mean and the varianceis a well-established procedure in statistics, particularly for the Normal Distribution. It is known as the *method of moments* (the mean and variance being technically known as the first two "moments" of a frequency distribution). But with a highly skew distribution an occasional exceptional value can markedly influence the variance (see Exercise11G). Consequently, use of the variance can lead to relatively unreliable estimates of other parameters, such as k in the NBD.

An alternative way of estimating k is by equating the observed proportion of zeros p_0 to the theoretical NBD value for the number of zeros, which is $(1 + m/k)^{-k}$. The resulting equation cannot be solved for k by direct algebra, but simple methods are described in the literature (e.g. Chatfield, 1969; Ehrenberg, 1972). This method is statistically more efficient for data with many zeros than the method of moments.

The general formula for the proportion p_r of values r in an NBD with mean m and parameter k is

$$p_r = \frac{(k+r-1)!}{r!(k-1)!}\left(\frac{m+k}{k}\right)^{-k}\left(\frac{m}{m+k}\right)^{r}.$$

(Strictly speaking, factorial expressions like $(k+r+1)!$ should be expressed as Gamma-functions, since k is usually non-integral and factorials are not defined for such values.) The parameter k is called the "exponent" because the above expression arises as the expansion of the second term in the binomial expression

$$\left(\frac{m+k}{k}\right)^{-k}\left(1-\frac{m}{m+k}\right)^{-k},$$

where the exponent k has a negative sign. Hence the name of the distribution. (The expansion of a binomial expression will be discussed more fully for the positive Binomial in Exercise 13L.)

Exercise 12J. The Beta-Binomial Distribution

By reference to the statistical literature, set out the basic formulae of the Beta-Binomial Distribution discussed in Section 12.5.

Discussion.

The Beta-Binomial Distribution has three parameters. They are n, the fixed size of the phenomenon being examined (e.g. the number of children in a family in Table 12.13a, or the number of weeks in Table 12.14), and two quantities α and β, which depend on the observed data.

The formula for the Beta-Binomial proportion of observations where there are r occurrences out of n (e.g. r boys in a family of n children) is:

$$\frac{n!}{(n-r)!r!}\,\frac{(\alpha+r-1)!(n+\beta-r-1)!}{(n+\alpha+\beta-1)!}\,\frac{(\alpha+\beta-1)!}{(\alpha-1)!(\beta-1)!},$$

(where the factorial should strictly be written as Gamma-functions for non-integral α and β). The mean and variance are given by

$$\text{mean} = n\alpha/(\alpha+\beta)$$

$$\text{variance} = n\alpha\beta(n+\alpha+\beta)/(\alpha+\beta)^2(1+\alpha+\beta).$$

The theoretical frequencies in Table 12.13a were found by equating these values to the observed mean and variance (using the method of moments), and solving for α and β with $n=8$.

An alternative way of fitting the distribution is to equate the theoretical expressions for the mean $n\alpha/(\alpha+\beta)$ and the number of zeros, $\{(n+\beta-1)!\,(\alpha+\beta-1)!\}/\{(n+\alpha+\beta-1)!(\beta-1)!\}$, to the observed values and solve for α and β. If the incidence of zeros is large, as in Table 12.14, this method is statistically more efficient but cumbersome to deal with (Chatfield and Goodhardt, 1970).

Exercise 12K. The Normal Distribution

What is the mathematical formula for the Normal Distribution? Comment on its use.

Discussion.

For the Normal Distribution with mean μ and standard deviation σ, the proportion of readings in the interval $\mu - k\sigma$ to $\mu + k\sigma$ (for any positive value k) is given by the integral

$$\frac{1}{\sqrt{(2\pi\sigma^2)}} \int e^{-(x-\mu)^2/2\sigma^2} \, dx.$$

from $(\mu - k\sigma)$ to $(\mu + k\sigma)$.

Table 12.19 gives some selected values of this function (more detailed tables are given in most textbooks of statistics).

TABLE *12.19* *The* Descriptive Characteristics of the Normal Distribution
(Selected values)

± Distance from the mean	The proportion of readings lying within the stated limits
± .5 s.d.	38%
± .8 s.d.	58%
± 1.0 s.d.	68%
± 1.6 s.d.	89%
± 2.0 s.d.	95%*
± 2.6 s.d.	99%*
± 3.0 s.d.	99.7%
± 3.3 s.d.	99.9%*

*Conventional 5%, 1% and .1% significance levels

Thus about 40% of the readings lie within $.5\sigma$ on either side of the mean, and only 1 in a thousand lie more than 3.3σ away.

Practical statisticians or data analysts virtually never deal directly with the theoretical formula or even with the detailed numerical results in the book. The reason is that all Normal Distributions take the same shape, therefore given their mean, their standard deviation, and something like the values in the table, one can describe the data for almost all practical purposes.

Exercise 12L. The Combination of Frequency Distributions

What is the frequency distribution when two sets of data are combined each of which follows a specified frequency distribution (e.g. heights of boys and heights of girls)?

Discussion

In general there is no simple answer. For example, combining two Normal variables with means 5 and 20 will generally give a complex distribution with two humps or modes, one at 5 and the other at 20.

Conversely, if one comes across such a complex-looking distribution in practice, it can often be best analysed by separating the data into two sub-groups, each of which follows a simple, unimodal distribution.

However, there are cases where "mixtures" of different distributions lead to a simple result One case was the Beta-Binomial in Section 12.5, where different Binomial Distributions for individual families were combined. A similar case arose with the Negative Binomial Distribution of purchases in Section 12.3, which was a mixture of different Poisson Distributions. (This is outlined further in Chapter 13.)

Exercise 12M. Adding Different Variables

What is the frequency distribution of $(x+y)$, where x and y are two different variables for the same items (e.g., heights and weights of some boys)?

Discussion.

Adding the two variables here is like adding peoples' salaries in successive weeks, or their salaries plus their *other* forms of income.

The combination of different variables is one of the more important parts of statistical theory and practice. In general, there are no straightforward answers. But cases where simple answers occur are exceptionally important. They arise particularly where the variables are *independent* of each other. Examples are:

(i) Two independent Normal Distributions for x and y with means μ_x and μ_y, and standard deviations σ_x and σ_y. The sum $(x+y)$ will be Normally distributed with mean $(\mu_x+\mu_y)$ and standard deviation $\sqrt{(\sigma_x^2+\sigma_y^2)}$. This is a fundamental result. In particular, it follows that the distribution of the *mean* of x + y, i.e. $(x+y)/2$, will 'be Normally distributed with mean $(\mu_x+\mu_y)/2$ and standard deviation $\sqrt{\{(\sigma_x^2+\sigma_y^2)/4\}}$. If $\sigma_x^2=\sigma_y^2=\sigma^2$, the standard deviation of $(x+y)/2$ will therefore be $\sigma/\sqrt{2}$. This is the basis of sampling theory, as discussed in Part IV.

(ii) Two independent Poisson Distributions for x and y with means μ_x and μ_y. The sum $(x+y)$ will also follow a Poisson Distribution, with mean $(\mu_x+\mu_y)$.

(iii) Two independent Negative Binomial Distributions with means μ_x and μ_y and parameters k_x and k_y. If $\mu_x/k_x=\mu_y/k_y=p/k$, then $(x+y)$ will be distributed as an NBD with mean $(\mu_x+\mu_y)$ and a second parameter (k_n+k_y) yielding the same ratio.

CHAPTER 13

Probability Models

The concept of probability is linked to the notion of randomness. Both are theoretical abstractions and cannot be directly observed. The two concepts are difficult to pin down precisely, but we all have some idea of what they mean and they have useful practical applications.

Probabilities apply to individual items or events that occur on a more or less chance or random basis. This means that individually the items or events occur in no discernible pattern, but in the aggregate, or over a long term, they tend to occur in certain proportions. For example, the sex of any one baby seems to be determined in a random way; in any sequence of births we can see no girl-boy-girl order or the like. But in the aggregate we generally find that 51% of babies are boys.

In everyday terms probabilities are mainly used in predicting events that are unknown and uncertain, such as the sex of an unborn child or whether it will rain tomorrow. But there are several other distinct uses of probability mathematics. These mostly occur in relatively advanced forms of analysis. However, some familiarity with probability and the notions of randomness and chance is useful at many levels of statistical work. In this chapter we shall be primarily concerned with the use of probabilities in describing and interrelating facts which are known but irregular.

13.1 **The Probability Concept**

The probability of a particular event occurring can be any number from 0 to 1. If the event is part of a steady series, then the probability of the given outcome should equal the proportion of times that event occurs in the series. That is how the numerical value of the probability is often arrived at in the first place.

Compared with the use of theoretical frequency distributions and proportions discussed in Chapter 12, probability models are simply an alternative descriptive language. For example, we used the theoretical Binomial Distribution to say that 14% of three-child families have three boys, 39 %

have two boys, 36% have one boy, and 11% have no boys. Alternatively, we could express this in the language of probabilities and say that the probability of any such family having three boys is .14, two boys .39, etc. The crucial difference is that proportions are characteristics of a group of readings, while probabilities are corresponding statements about each individual reading.

Probabilities have two practical advantages over proportions. Firstly, they are easier to interrelate mathematically, particularly in complex situations. (Since proportions are always expressed in terms of a particular set of data, they need to be redefined if that set is subdivided, combined, etc.) Secondly, while theoretical frequency distributions merely *describe* the data, probabilities can also imply an underlying process since they are tied to the notion of randomness. Thus looking at data in terms of the individual readings often allows one to develop some deeper theoretical understanding of why the observed phenomena occur the way they do.

To use the probability concept one must first satisfy the essential condition of knowing (or assuming) how the probabilities of different events interrelate. The simplest case is when the probabilities are taken to be independent of each other. For instance, if the probabilities of successive children being boys are independent and always .51 (or, more precisely, .5148), then it follows that the probability of a family having 3 boys should be about $.51 \times .51 \times .51 = 51^3$. Thus 51 out of the 100 first-born will on average be boys; for 51% of these, i.e. about 26, the *second*-born will be boys; and for 51% of these, i.e. 13, the *third*-born will be boys. But the data need not behave like that. Whether probability statements provide useful descriptions, or what *kinds* of probability statements do so, depends on the nature of the particular data.

Suppose that 60% of consumers buy Corn Flakes on their next purchase of breakfast cereals. It is then not necessarily useful (or correct) to say that all consumers have the same probability of .6 of buying Corn Flakes. If all the 60% buy Corn Flakes again on their second purchase, then it begins to look as though there might be two kinds of people: some (60%) who *always* buy Corn Flakes, and some (40%) who never buy.

But suppose instead that the data show that only 60% of those who bought Corn Flakes on the first purchase buy them again on their second purchase, and that, similarly, of those who did not buy Corn Flakes on their first purchase 60% buy them on *their* second purchase. Then we might summarise a relatively complex situation by saying that *any* consumer has a 60% chance, or a .6 probability, of buying Corn Flakes. This would have to be true whether the buyer belongs to a group who already bought them 100 times in the past or only once. Attaching a probability statement to each

individual is then easier and simpler than making statements about all possible groups.

A more complex situation where the probability of an event depends only on the directly preceding event, is called a simple first-order Markov process (after the Russian mathematician, A. A. Markov). This is one of the simpler *stochastic* processes, which are defined as more or less random or probabilistic forms of behaviour that involve linked sequences of events, especially ones linked over time. The word "stochastic" comes from the Greek *stochos*, meaning guess, i.e. "what will happen next'?". (In practice, consumer behaviour follows a more complex stochastic pattern where the probability of repeat-buying is related to the frequency of purchase.)

Irregular or "as-if random" phenomena take many forms. In the next section we briefly outline how probability models are used to describe irregular phenomena with independent probabilities.

13.2 Independent Events

When analysing irregular data we deal with three distinct entities. One is the actual observations which combine into an empirical distribution. The second is a theoretical or mathematical frequency distribution which is fitted to describe the data to within some degree of approximation. The third is a probability model which speaks in terms of the individual observations and implies an underlying process to account for the theoretical distribution.

For example, in Chapter 12 we looked at data on the incidence of major strikes per week in the United Kingdom from 1948 to 1959. The observations approximated a theoretical Poisson Distribution with a mean of .90. As already mentioned in Section 12.2, Poisson Distributions can arise when events occur independently of each other and with a constant average frequency (i.e. with no trend in their probability of occurrence). Therefore we can infer that the process underlying the incidence of major strikes might be one of more or less independent events with constant probability, where the occurrence of a major strike is not influenced by when the previous one occurred. This is called a Poisson Process. By using this probability model we may gain more understanding of the system, for example, whether the occurrence of a "run" of several strikes is no more than one might expect to occur occasionally, on the basis of chance in a Poisson Process, or whether it implies some special causative factor.

As another example, suppose that an observation can be affected by a large number of small irregular factors acting additionally and independently of each other, e.g. various sources of "error". Then it can be shown mathematically that as the number of such factors increases, different observations of this type tend to follow a Normal Distribution. This theoretical result is

known as the *Central Limit Theorem* and is of exceptional importance in statistics. If we observe a Normal Distribution, the theorem suggests a possible underlying mechanism, of many independent chance errors. And since many phenomena are known to consist of or to be influenced by diverse and more or less independent factors, it also explains why the Normal Distribution occurs so widely.

Again, we know that the observed incidence of boys and girls in families approximates a Binomial Distribution. This implies that it follows a Bernouilli Process, named after Jacob Bernouilli. Here the requirements are that for a fixed number of possible events (e.g. the number of children in a family), the occurrence of a particular event (a child being a boy) must have a constant probability and that the outcomes of different events must be independent of each other. Thus the process implies that the sex of successive children in a family is determined independently and that the probability of either sex is constant.

We therefore have a choice of descriptions for the incidence of boys and girls. We can model the observed data either with a Binomial Distribution, simply describing the data in terms of groups and proportions, or with a Bernouilli Process, speaking in terms of individuals and probabilities and also implying an underlying mechanism. But we can only apply the probability model to aspects of the data that behave irregularly, with no systematic patterns. Thus one must first establish *empirically* that to the limits of current knowledge the events effectively occur in an apparently chance or random manner.

There is no inherent logical reason why a Bernouilli Process should occur for such data. If instead of looking at *family* grouping, we observed children at different schools, or groups of children playing together, the sexes of successive children would not appear to be independent. One school might be all boys, another all girls, and a model of independent probabilities for successive children in each school would not describe such data at all. Similarly, children mostly play together in groups of the same sex, at least until the age of about 15. After that, groups of2 tend to be mixed more often than a 50: 50 model would predict, whilst *larger* groups observed talking, or eating together, or playing games tend still to be predominantly of one sex, The sex distributions in such cases therefore do not follow anything like a Bernouilli Process.

Even for the sex distribution in *families* the Binomial Distribution does not fit exactly. We saw in Chapter 12 that the theoretical values slightly, but consistently, underestimated the number of all-girl or all-boy families, Thus it is purely an *empirical* finding that a model of independent probabilities approximately describes the distribution of boys and girls in families. We can only say *empirically* that the sex of successive children in families acts almost as if it operates independently with constant probabilities.

When a probability model fits well this does not prove that the underlying physical process is *really* probabilistic or *really* random. All we have is a situation where the data appear sufficiently irregular for it to be useful to describe them *as if* they were random. We have no base in theory for any assumptions of independence or randomness. Both independence and randomness are abstract concepts which can never be fully established in observational data but which can supply a concise "as-if" model.

So-called games of chance are a common example of confusion here. There is no inherent reason why tossing a coin should be a chance phenomenon and follow a Bernouilh Process. Heads do not necessarily come up randomly and independently in successive throws with a constant probability of $p = \frac{1}{2}$ (or some other fixed value if the coin is biased).

If a coin is placed tails-up on one's hand, tossed gently so it turns over just once, and then caught horizontally, it will show heads every time. But if the coin is tossed so that it turns a large number of times, differing on each occasion, heads or tails cannot be predicted. It is an *empirical* finding that no one has yet developed the skill to make the outcome regular under such conditions, or to predict the variations from one toss to the next (otherwise they would be demonstrating their skill on television). Therefore tossing a coin is only an "as-if" random process, and this description depends on empirical observations that under the stated conditions there are no regular patterns in the results.

13.3 **Stochastic Models**

Irregular events which do not appear to follow simple *independent* probability processes have to be described by more complex stochastic models. These use the idea of conditional probabilities, where the probability of an event may be influenced by a previous event or depend on other factors. For example, one labour strike could trigger off others (although this does not usually seem to happen), or there could be some common factor (a large increase in the cost of living, some new governmental legislation, or a politically-inspired "plot") which causes more strikes to occur at certain times than would occur "by chance" under a Poisson Process with independent probabilities. The Poisson *Distribution* would then not give a good fit to the data.

The Beta-Binomial Distribution mentioned in Section 12.5 was an example of a more complex stochastic model and arose because the simple Binomial Distribution did not entirely give a good fit. The possible underlying process suggested was that the probability of boys varied among families instead of being constant at .51 for all families.

Other stochastic models arise with the Negative Binomial Distribution which has already been mentioned. Models involving the NBD have been

used in a variety of applications, including the occurrence of accidents, the spread of animals or plants in ecology, and buyer behaviour (e.g. Greenwood and Yule, 1920; Fisher *et al.*, 1943; Ehrenberg, 1972). Once such a stochastic model has been successfully fitted, it is easier to interrelate many different aspects of the observed data because one is dealing with probabilities and individual readings rather than with proportions tied to specific sets of data.

For example, we have observed that the number of purchases of any particular brand of frequently bought goods in a time period follows an NBD, as noted in Table 12.9a. However, there are at least two different underlying processes that might cause an NBD to occur. One is the "contagious" or "learning" type of situation. Here everybody starts with the same probability, but once the event occurs for someone the probability of recurrence (of buying again) increases for that individual.

The second possible underlying process is the "heterogeneous" or "proneness" model. Here the event occurs with different probabilities for different individuals, but the probabilities do not change over time. To differentiate the two possibilities, one must look at other characteristics of the observed data.

To illustrate this procedure, we consider the "heterogeneous" process which can lead to an NBD. The model involves a mixture of different Poisson Distributions (as do various stochastic processes, e.g. Haight, 1967). It assumes that an individual's purchases of a brand over successive periods of time (e.g. weeks) follow a Poisson Processs, and that different consumers' long-run average rates of purchasing μ (Greek "mu" = the means of the different Poissons) follow a Gamma-Distribution, as shown in Table 13.1. Thus the distribution in each row is assumed Poisson (with mean μ_A, μ_B, etc.) and the distribution of the μ's in the last column is assumed Gamma.

TABLE 13.1 Schema of the Poisson-Gamma Model Leading to Negative Binomial Distributions

Consumer	Successive Weeks 1	2	3	4	5	6	.	.	.	Long-run Averages	Distribution
A	x	x	x	x	x	x	x	.	.	μ_A	Poisson
B	x	x	x	x	x	x	x	.	.	μ_B	Poisson
C	x	x	x	x	x	x	x	.	.	μ_C	Poisson
D	x	x	x	x	x	x	x	.	.	μ_D	Poisson
	x	x	x	x	x	x	x	.	.		.
	.	.	.	.	.	.	.	.	.	.	
	.	.	.	.	.	.	.	.	.		.
Distribution	NBD		NBD		NBD		NBD	. . . . .		Gamma	

This model cannot be validated *directly*, since in practice one cannot observe the individual consumers' long-run averages, μ, and check the form of their distribution. (Nor can one fully check out the Poisson assumption, as applying in the long run). However, from this mixed Poisson-Gamma model a variety of theoretical deductions can be made mathematically. If these approximate the relevant aspects of the observed data, they support the validity of this particular theoretical model and at the same time make it a *useful* one, in that it describes and integrates all these different aspects of the data in one theoretical formulation.

The first theoretical deduction is that the number of purchases made by the different consumers in any given time-period should follow a Negative Binomial Distribution. This is of course the observation we started off with, that consumer purchases tend to follow an NBD. We have here the typical backwards-and-forwards or chicken-and-egg process in theory-building. We were considering the Poisson-Gamma model only because we already *knew* that an NBD tends to fit.

But there are additional deductions. For example, the model implies that the distribution should be an NBD in *any* time-period, e.g. a week as for any column shown in Table 13.1, or for a month as in aggregating the data for 4 weeks, and so on. And this also is what is found in practice (with some deviations in very short time-periods).

Next, the Poisson-Gamma model says that the NBD parameter k should be constant in different length time-periods. This is what is found in practice (e.g. Table 12.17 in Exercise 12B). In contrast, the "contagious" model mentioned above, whilst also leading to an NBD, says that k should vary in direct proportion with the length of the time-period. This therefore provides a very sharp differentiation between the two types of underlying processes.

A further deduction is that in different length time-periods the average purchase frequency of the brand should increase less than proportionately to the length of the time-period, as we saw in Chapter 10. The quantitative

TABLE 13.2 Observed Values of Average Purchase Frequencies and Theoretical NBD Predictions from the 24-Weekly Values

	Average Purchase Frequency per Buyer in				
	24 weeks	12 weeks		4 weeks	
	Obs.	Obs.	Theo.	Obs.	Theo.
Corn Flakes	5.7	3.5	3.7	1.8	2.1
Weetabix	5.7	3.8	3.7	2.0	2.1
Shredded Wheat	4.4	3.0	2.9	1.7	1.8
Sugar Puffs	3.4	2.4	2.4	1.5	1.6
Puffed Wheat	2.6	2.1	1.9	1.6	1.4
Average	4.4	3.0	2.9	1.7	1.8

details can be summarised by the approximate formula $(w_T - 1) = (w_t - 1)(T/t)^{.82}$, where w_T and w_t are the average purchase frequencies per buyer in time-periods of length T and t. Table 13.2 compares this theoretical formula with the observed data for breakfast cereals.

Other examples arise if we examine repeat buying in successive equal time-periods. For instance, consumers who buy a given brand in the second period but not the first ("new buyers") generally buy it about 1.4 times on average, as the Poisson-Gamma model predicts (e.g. Ehrenberg and Pyatt, 1971, pp. 25 and 70; Ehrenberg, 1972).

These examples illustrate three major points about stochastic models. Firstly, a single model can both describe and interrelate a great variety of different results.

Secondly, although the model is probabilistic (e.g. a mixture of Poisson Processes) the general user needs no explicit probabilistic mathematics to apply the results, but just straightforward averages or the like (as in Table 13.2).

Thirdly, although all the results were reached by using a stochastic or probabilistic theory, this does not imply that a consumer actually makes purchase decisions on a random or chance basis. It only says that purchases by a variety of different consumers appear to be sufficiently irregular so that in the aggregate they can be successfully summarised by a probability model, *as if* they were in certain respects probabilistic.

13.4 Probability and Uncertain Events

Until now we have concentrated on the use of probability mathematics to model different kinds of irregular variations in observed data. Another use is to try to quantify "degrees of belief" about some uncertain hypothesis or assertion, e.g. that Homer was blind, that it will rain tomorrow, or that one's next child will be a boy.

Instead of interrelating observed data on the apparently irregular incidence of boys and girls, we may have to say or do something before the fact is observed, i.e. before the child is born. We may then attach a probability of .51 to the child being a boy.

This is an assertion about one's state of uncertainty, not about the child. The statement can be interpreted empirically by referring to the frequency with which similar statements would be correct. Thus in asserting that an uncertain event has a probability of .51, one may imply that such statements (about *any* event) should prove correct on just over half the occasions.

If a topic has been studied extensively, like the incidence of male and female babies, there is clearly an empirical basis for determining such predictive probabilities. But in other cases it is difficult to establish the correct probability values to use. This difficulty is recognised by the term "subjective

probabilities" that is widely used in this context (i.e. "guessed" probability levels rather than objectively established ones). An even greater problem is knowing whether the different uncertain events being studied are independent or, if not, how their probabilities of occurrence are interrelated.

Using probability mathematics to deal with uncertain events requires extensive prior knowledge about the kind of events involved and their interrelations. Without such knowledge probability applications would involve making arbitrary assumptions. (The so-called "Bayesian" approach to decision-theory can involve many such difficulties, but its discussion is outside the intended scope of this book.)

Statistical sampling (discussed in Part IV) applies probability mathematics to uncertainty in a different way. Here the chance element and the independence of successive observations are deliberately introduced into the data by the physical operation of random sampling.

13.5 **Summary**

The probability concept can be useful when dealing with irregular or uncertain phenomena. In this chapter we have discussed its use for describing irregular observed data, such as statistical frequency distributions. The main steps are to assign a probability value to the individual observation and to specify the independence or interdependence between the probabilities of different observations.

The descriptive use of probabilities does not imply that the phenomena really occur by chance, but only that they appear so irregular that they can be successfully described *as if* they were random. Simple models of independent events can provide an underlying rationale for observed frequency distributions like the Normal, Poisson, and Binomial. Other stochastic processes can be used to model more complex phenomena, where the probability of one event is influenced by the occurrence of another event.

CHAPTER **13** EXERCISES

(Exercises 13L onwards illustrate some of the more technical uses of probabilities in dealing with frequency distributions and stochastic models.)

Exercise 13A. Exclusive Events

If the probability of rain on a certain day is .4, what is the probability of dry weather?

Discussion.

If the only possibilities are "rain" and "dry", the probability of dry weather must be .6 since the probability of *something* happening must be 1, i.e. Probability of Rain + Probability of Dry = 1.

If additional categories were possible, e.g. "mild drizzle" with a probability of .05 and "no record available" with a probability of .01, the probability of a dry day would have to be only .54 (so that .40 + .05 + .01 + .54 = 1.00). These probabilities should mean that in a long run of days, about 40% tend to have rain, about 5% have some drizzle, about 1% have no record, and about 54% are known to be dry.

Whether it is useful to think of any specific day as having such individual probabilities depends on the absence of predictable patterns. For example: these probability statements would make no sense in a tropical country that had a rainy season of about 150 consecutive days (40%) and drought the rest of the year. Whether the loss of records on about 1% of days is effectively random is also open to question. (Perhaps records are mainly lost during exceptionally heavy downpours or are not as regularly kept at weekends.)

Exercise 13B. Independent Events

If the probability of having arthritis in a lifetime is .3 and that of having measles is .8, and if the two events are independent, what is the probability of having both?

Discussion.

If a proportion .8 of the population have measles, then on the independence criterion, .8 (80%) of those having arthritis should also have measles. The probability of having both is therefore .8 of .3, or .24. Thus 24% of the population should have both. Independence here means for example that the incidence of the one phenomenon does not affect the incidence of the other.

The criterion of independence is relatively straightforward mathematically, but it can never be fully established empirically since it means independence in all *possible* respects, i.e. in every possible sub-group of the population. However, if all *available* cross-analyses have shown no dependence between the two variables, a probabilistic model of independence could provide a useful description. (One kind of complication that can arise is that the incidence of an illness generally depends on how long one lives and that different illnesses tend to occur at very different ages.)

Exercise 13C. Non-independence

If in the duplication-of-purchase law $b_{XY} = Db_X b_Y$ of Section 9.3 the coefficient D is 1, is purchasing of Brand X independent of purchasing of Y? (Here b_X and b_Y are the proportions of the population who buy X and Y at least once in the analysis-period, and b_{XY} is the proportion who buy both X and Y, each at least once.)

Discussion.

If the coefficient D is greater than 1, b_{XY}/b_Y is greater than b_X, i.e. the proportion of buyers of Y who also buy X is greater than the proportion of the whole population who buy X. There is then a tendency for buying of X to go with buying of Y, and so buying of X is not independent from that of Y. Similarly, if D is less than 1, buying of Y *inhibits* buying of X (b_{XY}/b_Y is less than b_X) and the two are not independent.

If additional categories were possible, e.g. "mild drizzle" with a probability of .05 and "no record available" with a probability of .01, the probability of a dry day would have to be only .54 (so that .40 + .05 + .01 + .54 = 1.00). These probabilities should mean that in a long run of days, about 40% tend to have rain, about 5% have some drizzle, about 1% have no record, and about 54% are known to be dry.

Whether it is useful to think of any specific day as having such individual probabilities depends on the absence of predictable patterns. For example, these probability statements would make no sense in a tropical country that had a rainy season of about 150 consecutive days (40%) and drought the rest of the year. Whether the loss of records on about 1% of days is effectively random is also open to question. (Perhaps records are mainly lost during exceptionally heavy downpours or are not as regularly kept at weekends.)

Exercise 13B. Independent Events

If the probability of having arthritis in a lifetime is .3 and that of having measles is .8, and if the two events are independent, what is the probability of having both?

Discussion.

If a proportion .8 of the population have measles, then on the independence criterion, .8 (80%) of those having arthritis should also have measles. The probability of having both is therefore .8 of .3, or .24. Thus 24% of the population should have both. Independence here means for example that the incidence of the one phenomenon does not affect the incidence of the other.

The criterion of independence is relatively straightforward mathematically, but it can never be fully established empirically since it means independence in all *possible* respects, i.e. in every possible sub-group of the population. However, if all *available* cross-analyses have shown no dependence between the two variables, a probabilistic model of independence could provide a useful description. (One kind of complication that can arise is that the incidence of an illness generally depends on how long one lives and that different illnesses tend to occur at very different ages.)

Exercise 13C. Non-independence

If in the duplication-of-purchase law $b_{XY} = Db_Xb_Y$ of Section 9.3 the coefficient D is 1, is purchasing of Brand X independent of purchasing of Y? (Here b_X and b_Y are the proportions of the population who buy X and Y at least once in the analysis-period, and b_{XY} is the proportion who buy both X and Y, each at least once.)

Discussion.

If the coefficient D is greater than 1, b_{XY}/b_Y is greater than b_X, i.e. the proportion of buyers of Y who also buy X is greater than the proportion of the whole population who buy X. There is then a tendency for buying of X to go with buying of Y, and so buying of X is not independent from that of Y. Similarly, if D is less than 1, buying of Y *inhibits* buying of X (b_{XY}/b_Y is less than b_X) and the two are not independent.

The first question in Exercise 13D refers to all 100,000 families. Since 25,000 have two boys, we can say the probability of a family having two boys is $25{,}000/100{,}000 = 1/4$.

The second question concerns those families for whom we know one child is a boy. This excludes the girl-girl families and leaves 75,000. Of these, 25,000 still have two boys. Hence the probability of one of the families having two boys is $1/3$.

The third question concerns those families where one *particular* child is known to be a boy, e.g. the one nearest the door. Excluding the possibility that both children are equally close to the door, this refers to 50,000 families. In the other 50,000 a *girl* will be nearest the door, on the supposition that boys and girls are distributed independently in *all* respects. Of the first 50,000 families, 25,000 have two boys. Thus the probability of one of these families having two boys is $1/2$.

*Comparison of the discussions in this and the last exercise indicates how the language of probabilities can be more concise than that of proportions.

Exercise 13F. An Unusual Occurrence?

A family with six children has all boys. Given that the probability of boys is .51, does this imply some special causal factor?

Discussion.

If 51% of children are boys, and if the occurrence of boys amongst successive children is independent, one would expect about 1.8% of all six-children families to have all boys (i.e. $.51^6$).

If approximately this percentage is found on examining large numbers of families, the occurrence of six boys in one particular family would not necessarily signify any special factor. It is what one expects to see happen "by chance" (in 1.8% of all families) if the incidence of boys is random.

However, the model of the independent random occurrence of boys or girls is only a theoretical abstraction. The sex of the individual child is presumably determined by specific factors and the facts have shown (Chapter 12) that a simple Bernouilli Process is not quite true, We therefore cannot strictly *prove* by an appeal to probabilities that the occurrence of 6 boys in a particular family is not due to some special factor.

While the argument is therefore slightly less powerful, the empirical facts still show that the incidence of boys and girls in families of all sizes closely approximates Binomial Distributions with $p = .51$. Except for the small extent to which the Binomial model does not fit, the incidence of boys is therefore consistent with the hypothesis that it occurs on an *as-if random* basis. Thus no special factors have to be invoked to account for a particular family-pattern such as six boys (compared with the factors which account for *mixed* families in their observed proportions and for all-girl families).

Exercise 13G. Deviations from a Model

Discuss the way observed deviations from a theoretical model can be described in probabilistic terms.

Discussion.

Deviations from a model are often irregular and tend to follow a Normal Distribution, as discussed in Exercise 12C. However, there is no defined set or "population" of such readings; each deviation is distinct. For example, in the data in Chapter 1 the deviation in QI in the West was -3, and this was considered quite separately from the deviation of 5 for QIII of the next year, which was analysed in Chapter 2.

This is a situation where the use of probabilities rather than proportions is particularly appropriate (i.e. statements about individual readings rather than about groups of readings). In attempting to describe the Normality of the data we do not have to say that in some arbitrary group of readings, 68% lie within $\pm$ one standard deviation of the mean. Instead, we can make the following sequence of assumptions.

(i) That any particular deviation can be regarded as coming from a certain type of probability distribution (say Normal) with mean 0 and standard deviation σ.

(ii) That the probability distributions for each of the other deviations takes the same form (e.g. Normal with mean 0 and standard deviation σ).

(iii) That the individual deviations are independent of each other.

We can test these assumptions by checking whether *any* group of deviations broadly fit a Normal Distribution with mean 0 and standard deviation σ. (By assumption (iii) it does not matter which group we select for this, as long as the selection does not depend on the observed values themselves, e.g. excluding all large positive values.)

The use of probabilities in dealing with deviations from a model leads to statements which are logically far more precise (and hence easier to manipulate) than the use of proportions and relative frequencies.

Exercise 13H. Theoretical Assumptions

Reference to statistical analyses in the scientific literature shows that in discussing a theoretical model (e.g. the simple straight line $y = ax + b$) it is often assumed that the deviations from the model are independent and Normally distributed, with mean 0 and a certain standard deviation which is constant all along the line. Is such an assumption justified?

Discussion.

The assumption is justified only if it is consistent with the facts, i.e. if it is based on direct empirical analysis of the specific data, or of a range of previous similar data.

But the assumption is often made before any data have been analysed. Assuming independence of the deviations, or even a zero mean, could then be quite inappropriate since the theoretical model itself might not fit. For example, if a linear model has been specified but the empirical relationship is curved, the deviations from any linear equation fitted will not be independent of each other, or Normal, or have a zero mean (except more or less accidentally), or have constant scatter all along the line.

Exercise 13I. "AtHaphazard"

In discussing the strength of heredity, Fisher (1950, p. 191) said he assumed for the sake of the argument that "any environmental effects are distributed at haphazard". What is involved?

Discussion.

When natural phenomena are considered to act haphazardly, the usual implication is that one knows little about them. There is therefore no reason to believe they lack patterns, let alone that they are "random".

But if detailed analysis of the observed phenomena has shown them to be effectively irregular, they can be usefully described by a stochastic (i.e. quasi-random) model. However, the randomness is only a property of the model, not of the empirical phenomena, and can therefore not be assumed *a priori.*

Exercise 13J. What are Random Independent Events?

Ask a friend to call heads or tails when throwing a coin in an effectively random manner for about 10 throws. Comment on the results.

Discussion.

General experience suggests that your friend will not make the same call, e.g. "heads", more than about five times in succession. He will change to "tails" for one or more throws, and then probably back again to heads, and so on. Yet with random throws of an unbiased coin, calling "heads" all the time would tend to be right about half the time, and there is no way of consistently doing better.

People tend to vary their calls because

(i) they do not believe the quasi-randomness of the throwing and/or the lack of bias of the coin; or
(ii) they want to demonstrate free will; or
(iii) they want to make a "game" of it; or
(iv) they do not understand the nature of randomness and, in particular, that of independent random events (the outcome of one event being independent of the outcomes of previous ones).

Exercise 13K. The Central Limit Theorem

Can you prove that the sum of a large number of small, independent, random variables will tend to be Normally distributed?

Discussion.

Suppose for simplicity that all the random variables have zero mean. It is then easy to see at an intuitive level that the sum of these variables must follow a humpbacked and more or less symmetrical distribution. Since a large number of variables are involved, about half the values in any particular instance will be positive and about half negative. Their sum will therefore tend to be near zero, Relatively large positive or negative values will occur only rarely.

However, it is very difficult to prove mathematically that the distribution tends to the *Normal* form. The reason for this difficulty is easy to see.

Our starting-point was very general: the sum of a large number of independent random variables that can follow any form of distribution. In contrast, the final result is very specific: that the probability of taking the value x should be proportional to $e^{(x-\mu)^2/2\sigma^2}/\sqrt{(2\pi\sigma^2)}$, the so-called "probability density function" of a Normal Distribution with mean μ and variance σ^2. The connection between the two must therefore be complex. (Although already surmised by the great French mathematician Laplace around 1800, the first rigorous proof of the Central Limit Theorem was only given in 1901, by the Russian mathematician Liapounov. It has since been extended and refined.)

Exercises 13L onwards deal with relatively technical matters.

Exercise 13L. The Binomial Expansion

The Binomial Frequency Distribution gives the probabilities of r occurrences out of n observations, for an event with probability p, by expanding the Binomial formula $(p+q)^n$. What does this mean?

Discussion.

Consider $n = 2$. Then $(p+q)^2 = (p+q)(p+q) = p^2 + 2pq + q^2$. The three terms in this expression equal the probabilities that in two observations the event with probability p occurs both times (p^2), only once ($2pq$, either in the first or the second observation, hence the factor 2), or *neither* time (q^2).

More generally, the terms of the corresponding expansion of the Binomial expression $(p+q)^n = p^n + np^{n-1}q + [n(n-1)/2]p^{n-2}q^2 + \cdots + [n!/(n-r)!r!]p^{n-r}q^r + \cdots + q^n$ give the probabilities of $n, (n-1), (n-2), \ldots, (n-r)$ etc. occurrences out of n observations, as already noted in Section 12.4. (The expression $n!$ stands for $n(n-1)(n-2)\ldots 3 \times 2 \times 1$.)

The expression $(p+q)^n$ is clearly a very neat way of summarising the terms of the Binomial Frequency Distribution, But there are some drawbacks.

In a Binomial situation, the two probabilities p and q are interrelated, since the probability q of the event not occurring is $1-p$, so that $p+q=1$. In considering the expression $(p+q)^n$, we could therefore write $(p+q)^n = (1)^n = 1$. For example, with $p = .6$, we have $(.6+.4)^n = 1^n = 1$. This does not get us anywhere, and the mathematician says "It is not what I am interested in". Instead, he expands $(p+q)^n$ into its constituent terms, as outlined above, *before* taking note that $p+q=1$. (The sum of these terms then necessarily adds to 1, which merely reflects that the probability of *some* outcome, $n, n-1, n-2, \ldots, 3, 2, 1$, or 0 occurrences, is 1.)

What the mathematician wants us to focus on is the individual term in the expansion, which for the rth term happens to give the probability of $(n+1-r)$ occurrences and $(r-1)$ non-occurrences. But "picking-out" this rth term can be a rather clumsy thing to do, especially when we note that in any *real* situation the terms are all simply *numbers*. For example for $n = 3$ and $p = .6$, the four binomial terms in expanding the expression

$(.6 + .4)^3$ are

$$.216, .432, .288, .064.$$

These have to be written in a clearly established sequence (or appropriately printed out by a computer) if we are to know which terms refer to 3, 2, 1 and 0 occurrences. The numbers lack any identifying labels.

The situation gets worse in more complex cases, such as when there are *two* binomial characteristics. Consider for example families of n children, where the incidence of boys and girls have probabilities p and q, and the incidence of being born on a weekday and at the weekend have probabilities a and b. Then the expansion of the product of two binomial expressions, i.e. $(p+q)^n(a+b)^n$ will give terms like

$$\frac{n!}{(n-r)!r!}p^{n-r}q^r\frac{n!}{(n-s)!s!}a^{n-s}b$$

for the probability of a family having r girls and s children born at the weekend. Here "picking out the term for r and s" from all possible n x n terms begins to be quite complex, especially when in any practical instance all the terms are merely numerical. The only direct way to differentiate the terms is by tabulating them in some neat manner. Doing *theoretical* mathematics with an expression like $(p+q)^n(a+b)^n$ is equally difficult. Some way of labelling the terms seems to be called for.

Exercise 13M. Probability Generating Functions

Is there a way of overcoming the problem of identifying terms in the expansion of $(p+q)^n$?

Discussion.

What is needed is a form of mathematical *labelling*. For a single-child family, consider the expression

$$pu^1 + qu^0$$

where p is the probability of the child being a boy and $q = (1-p)$ is the probability of its being a girl. This type of expression is called a "probability generating function" (p.g.f.) because it "generates" the probabilities for the incidence of boys and girls. The quantity u is a mathematical labelling device technically called a "dummy variable". It takes no real values, but makes it easy to pick out the terms wanted. The coefficient of u^1 gives the probability of 1 boy, that of u^0 the probability of 0 boys.

Similarly, for families of 2 children we can write $(pu^1+qu^0)^2 = p^2u^2 + 2pqu^1 + q^2u^0$, where the coefficient of u^r gives the probability of r boys, with r = 2, 1 or 0. More generally we can write the probability generating function of the Binomial with parameters p and q either as

$$(pu^1 + qu^0)^n,$$

or more concisely (if less explicitly)

$$(pu + q)^n,$$

since $u^1 = u$ and $u^0 = 1$.

There is now no problem in picking out terms. Thus for $n=3$ and $p=.6$, as in the last exercise, we have $(.6u^1+.4u^0)^3=.216u^3+.432u^2+.288u^1+.064u^0$. The dummy variable "$u$" with its exponent r acts as a clear label for $r=3$, 2, 1, and 0.

Although the function $(pu+q)^n$ might seem more complex than the straightforward Binomial expression $(p+q)^n$, the introduction of the dummy variable in fact *simplifies* the ensuing mathematics. Writing $(pu+q)^n$ also eliminates the temptation to add $p+q=1$ in $(p+q)^n$.

For the situation with *two* Binomial variables we write the p.g.f. with two different dummy variables, say u and v:

$$(pu+q)^n(av+b)^n.$$

Here the general term, for $(n-r)$ boys and r girls, and for $(n-s)$ weekday births and s weekend ones, is

$$\left\{\frac{n!}{(n-r)!r!}p^{n-r}q^r u^{n-r}\right\}\left\{\frac{n!}{(n-s)!s!}a^{n-s}b^s v^{n-s}\right\}$$

It is now simpler to identify the corresponding probability as the coefficient of $u^{n-r}v^{n-s}$.

Exercise 13N. The Negative Binomial Distribution

What is the probability-generating-function of the Negative Binomial Distribution?

Discussion.

In Exercise 121 we noted that the Negative Binomial Distribution arose from expanding the second term in the expression

$$\left(\frac{m+k}{k}\right)^{-k}\left(1-\frac{m}{m+k}\right)^{-k}.$$

Although the formula looks different, it is directly equivalent to the positive Binomial except that the sign of the exponent is changed. This can be seen by multiplying the first term into the second, giving

$$\left(\frac{m+k}{k}-\frac{m}{k}\right)^{-k},$$

where $(m+k)/k$ is equivalent to p and m/k to q. As before, the inside of the bracket adds to 1. It is useful to express this as a probability generating function by introducing a dummy variable, say u

$$\left\{1+\frac{m}{k}-\frac{m}{k}u\right\}^{-k}, \quad \text{or} \quad \left\{1+\frac{m}{k}(1-u)\right\}^{-k}.$$

Expanding this expression in powers of u gives the probability of observing r occurrences as the coefficient of u^r.

The usefulness of the p.g.f. approach can be illustrated by a very simple and powerful extension of the above formula to more than one time-period.

Consider consumer purchasing data in *two* time-periods, of length T_1 and T_2. Then the expression

$$\left\{1 + \frac{m}{k}[T_1(1 - u_1) + T_2(1 - u_2)]\right\}^{-k}$$

is the p.g.f. of the *bivariate* NBD, where m refers to the average rate of purchasing in a period of some unit length (e.g. a week). When this expression is multiplied out in powers of u_1 and u_2, the coefficient of the term $u_1^r u_2^s$ will give the probability of a consumer making r purchases in the first period and s purchases in the second period.

More generally still, the p.g.f. of the *multivariate* NBD is given by the succinct expression

$$\{1 + m \Sigma T_i(1 - u_i)/k\}^{-k}.$$

Here the summation Σ is over t different time-periods of lengths $T_1, T_2, \ldots, T_t$. The results mentioned in Section 13.3 (for the NBD as such, for k, for repeat-buying, for "new" buyers, and for the average purchase rates w_T in different length time-periods) all stem from this single expression.

Exercise 13P. The Poisson-Gamma Hypothesis

What is the justification for the Poisson-Gamma formulation of the NBD model outlined in Section 13.3?

Discussion.

The important justification of such a model is *indirect.* The model works in the sense that deductions such as those illustrated in Section 13.3 fit the facts. This has been found to a close degree of approximation in many thousands of different cases. Certain systematic deviations also occur (especially relating to very short time-periods) which are also increasingly well-understood. The model therefore provides a workable and' highly generalisable summary of complex data.

But none of this is a *direct* justification of the model's two basic assumptions of Poisson and Gamma Distributions. As already noted in Section 13.3, these assumptions cannot be checked directly because one cannot have empirical data extending over an indefinitely long time-period (as the model specifies).

The Poisson assumption can, however, be checked on a *sample* basis, i.e. for limited succession of time-periods like the four quarters of a year. Here the Poisson assumption of independent random events has been shown to hold well. This fits in with commonsense experience; *precisely* how many purchases of Corn Flakes, say, one makes in one quarter will hardly depend on precisely how many one made in the previous quarter (and it is the variation about one's *average* number of purchases per quarter which the Poisson aims to model). Exceptions occur in very short time-periods like a week or less where one is less likely to buy in the middle of the night than during the day, or *just* after one has already bought the item. But in many respects the NBD model is not sensitive to such deviations from the Poisson assumption (Ehrenberg, 1972; Chatfield and Goodhardt, 1973).

The Gamma-Distribution assumption, although also not testable directly, has been given a mathematical justification by Goodhardt and Chatfield (1973). This is derived from other kinds of empirical observations altogether. Thus, given that purchasing of one brand is approximately independent of purchasing of another brand (as illustrated by the duplication of purchase law, $b_{XY} = Db_X b_Y$, with D near to 1) and given that the average rate of product purchase is independent of the brand bought (Table 9.10), it follows from some powerful mathematics that different consumers' average purchasing frequencies for any one brand *must* follow a Gamma-Distribution. This kind of result justifies the initial assumption and also links up the Poisson-Gamma model for a single brand with a range of other results concerning switching between different brands. Nonetheless, the direct application of the whole Poisson-Gamma model rests not on this kind of justification at all, but" on the extent to which it actually works in practice.

CHAPTER 14

Correlation and Regression

In this chapter we discuss two statistical techniques which are often used in analysing relationships between variables: correlation and regression.

The data dealt with differ from those in Part II. There we always had two or more sets of data for analysis, e.g. the heights and weights of groups of children of different ages, sexes, races, etc., as illustrated in Figure 14.1A. In contrast, correlation and regression generally deal with a *single* set of readings. The individual readings are differentiated only by their values in the two variables, as shown in Figure 14.1B.

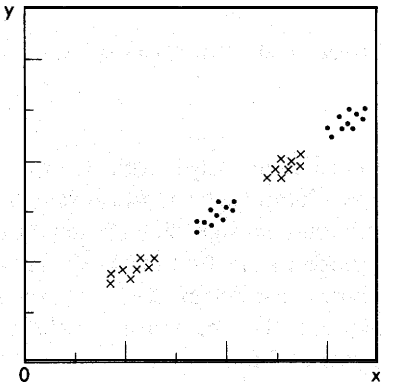

Figure 14.1A Different Sets of Readings

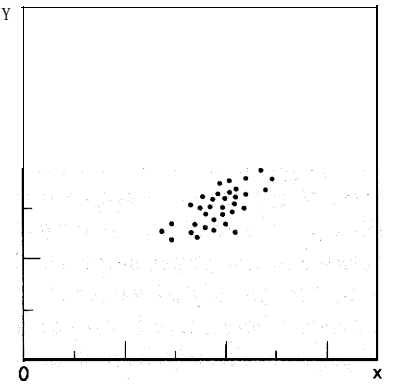

Figure 14.1 B. A Single Set of Readings

14.1 The Correlation Coefficient

Correlation coefficients are indices that measure the *strength* of a relationship. (The general idea of correlation was largely developed by the biologist Sir Francis Galton in the 1880s, but the particular form now in general use, the "product-moment correlation coefficient", was introduced by Karl

Pearson in 1898.) To provide a simple numerical illustration, we consider five pairs of readings in Table 14.1. Clearly there is some tendency for high values of y to go with high values of x, as Figure 14.2A also shows. The correlation coefficient aims to *measure* this tendency.

TABLE 14.1 Five Pairs of **Readings** in the **Variables** x and y

x:	1,	2,	2,	4,	6
y:	17,	11,	23,	19,	30

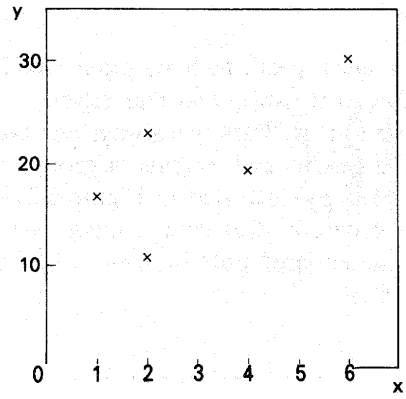

Figure 14.2A The Readings from Table 14.1

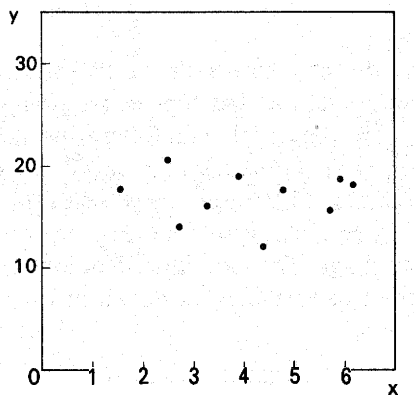

Figure 14.2B Uncorrelated Readings

The possible values of the correlation coefficient range from $+1$ to -1 for complete positive or negative correlation. Complete correlation describes a situation where all the readings lie exactly on a straight line having either a positive or a negative slope. (A negative slope means that the lower values of y go with the higher values of x.) Correlation coefficients near 0 represent situations where there is no particular tendency for the x and y values to vary together linearly, as illustrated in Figure 14.2B.

The formula for the correlation coefficient, usually denoted by r, is

$$r = \frac{\text{Covariance of } x \text{ and y}}{\sqrt{(\text{variance } x)}\sqrt{(\text{variance y})}}$$

The covariance is the product $(x - \bar{x})(y - \bar{y})$ of the deviations of each pair of readings x and y, from the means $\bar{x}$ and $\bar{y}$, averaged across all pairs of readings, or Sum $(x - \bar{x})(y - \bar{y})/(n - 1)$. If a pair of x and y readings are both greater or both smaller than the corresponding means, then the product of $(x - \bar{x})(y - \bar{y})$ will be positive. But if x, say, is greater than the mean and

y lower, the product will be negative. The average of the products therefore reflects the extent to which high x goes with high y and low x with low y.

For the five pairs of readings in Table 14.1 the means are $\bar{x} = 3$ and $\bar{y} = 20$. The covariance of x and y is therefore the average value of

$$\begin{aligned}
&(1-3)(17-20) + (2-3)(11-20) + (2-3)(23-20) \\
&\quad + (4-3)(19-20) + (6-3)(30-20) \\
&= (-2)(-3) + (-1)(-9) + (-1)(3) + (1)(-1) + (3)(10) \\
&= 6 + 9 - 3 - 1 + 30 \\
&= 41.
\end{aligned}$$

To find the average for n pairs of readings we divide by $(n-1)$, just as we do when computing variances, as discussed in Chapter 11. In our example this gives

$$\text{Covariance (xy)} = 41/4 = 10.25.$$

The numerical value of the covariance depends on the scales of measurement of x and y (e.g. inches versus feet, as can be seen by multiplying the x-values in Table 14.1 by 12 and recalculating the covariance), This effect is eliminated by dividing the covariance by the standard deviations of x and y. This results in the correlation coefficient as defined in the formula above. For the five pairs of readings in Table 14.1, the variance of x is $\text{Sum}(x-\bar{x})^2/(n-1) = \text{Average}\ \{(1-3)^2 + (2-3)^2 + (2-3)^2 + (4-3)^2 + (6-3)^2\} = 4$, and the variance of y is SO. The correlation coefficient for the five pairs of readings is therefore

$$r = \frac{10.25}{\sqrt{4}\sqrt{50}} \doteq \frac{10}{14} \doteq .7,$$

or .72, to be more precise.

The correlation coefficient is relatively tedious to calculate, especially with more extensive data. Exercise 14Q gives a convenient computing formula for use with a desk machine or pocket calculator ; but with extensive data it is now customary to use a computer.

14.2 Interpreting the Correlation Coefficient

To see how the correlation coefficient reflects the strength of the relationship between x and y, suppose the linear equation $y = 3x + 10$ has been fitted to the data in Table 14.1. For any particular value of x, say $\hat{x}$, the

equation gives an estimated value $\hat{y}$, namely $3\hat{x}+10$. For example, when $\hat{x}=1, \hat{y}=13$. There will then be deviations between the observed and the estimated values of y, i.e. $(y-3\hat{x}-10)$, which are called the "residuals". The variance of these deviations is called the "residual variance". For $y=3x+10$, the deviations for the five y readings are 4, -5, 7, -3, -2 the residual variance is 24.5, or a residual *standard deviation* of 4.95.

If the relationship between x and y is strong, the residual variance will be small compared to the variance of y (i.e. the average squared deviation of the observed y-values from their *overall* mean $\bar{y}$). This is then reflected by a high value of the correlation coefficient. The connection is that the square of the correlation equals 1 minus the ratio of the residual variance to the variance of y, i.e.

$$r^2 = 1 - \frac{\text{Residual variance}}{\text{Variance of y}}.$$

This can be rewritten as

$$\text{Residual variance} = \text{Variance of } y(1-r^2).$$

Thus $(1-r^2)$ measures the extent to which the relationship has reduced the variance of the y-readings. This is often referred to as x having "accounted for" a proportion r^2 of the variance of y, leaving $(1-r^2)$ "unaccounted for".

To use a less abstract measure of scatter than the variance we can take square roots. This gives

$$\text{standard deviation of residuals} = (\text{standard dev. of } y)\sqrt{(1-r^2)}.$$

(Variances and standard deviations are more useful in the mathematics of correlation and regression analysis than the *mean deviations* which were used in Part II.)

The variance of the y-data in Table 14.1 is 50, so that the standard deviation of y is 7.1. We have already calculated the correlation coefficient as .72, so the residual standard deviation should be

$$7.1\sqrt{(1-.72^2)} = 4.9 \doteqdot 5.$$

This is close to the value of 5.1 that we worked out directly from the data. (Strictly speaking, the residual variance or standard deviation can only be derived from the correlation coefficient when the deviations are from a "regression" equation, as described in Section 14.3 below. But, in practice, the derivation holds approximately for almost any reasonable equation that is fitted to the data.)

The "unexplained scatter" of the y readings about the equation $y=3x+10$ is therefore smaller than the original scatter of y : a standard deviation of about 5 compared to one of 7. But the reduction is not very large. Correlations need to be very high to reduce the residual variation

dramatically. This is illustrated in Table 14.2. A relationship with a correlation coefficient of .5 reduces the y-scatter by only about 13 %. Even a correlation of .95 reduces the y-scatter by only about 70 %.

TABLE 14.2 The Standard Deviation of the Residuals as a Percentage of the Standard Deviation of y

	Correlation r						
	.1	.3	.5	.7	.9	.95	.99
$100\sqrt{(1 - r^2)}$	99	95	87	71	44	31	14

Comparing Different Correlation Coefficients

A major drawback of using correlation coefficients is that although they are measures of scatter, they do not actually show whether or not two different sets of data have the *same* scatter. This is because correlations measure the residual variance relative to the *total variation* in the data, and the two sets may differ in this respect, If the two correlations are the same, it does not follow that the residual scatter is the same. And if the two correlations are different, it does not follow that the residual scatter is different.

The older literature on correlation analysis gives correction formulae to allow for differences in the total scatter of y (or in x). But it is much easier to compare the residual variances in the different sets of data directly, without using correlations at all. For example, in Figure 14.3A the average size of the residual scatter is clearly the same, even though the two correlations differ radically.

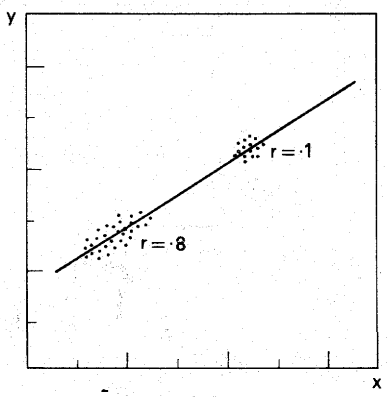

Figure 14.3A Similar Scatter but Different Correlations

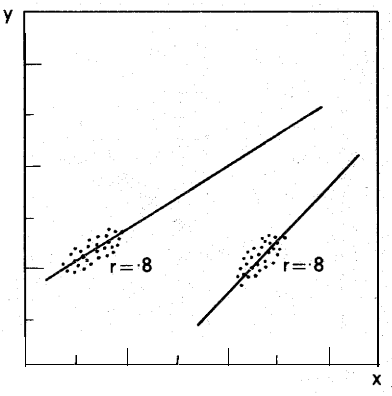

Figure 14.3B Equal Correlations but Different Equations

The correlation coefficient also does not tell us whether the actual *relationships* between x and y are the same in two different sets of data. In the two sets of data in Figure 14.3B for example, the two correlations and the sizes of the residual scatter are the same, but the two relationships are different.

Having calculated the correlation coefficient for a given set of data, it is not at all clear what one can usefully do with it. In particular, it is of no help for prediction or for comparing different sets of data.

Usually it is not difficult to see that in a given set of data there is a positive relationship, with some scatter. What then does it add to say that the numerical value of the correlation coefficient is .6? It does not tell us what the relationship is. It also does not tell us how big the scatter is, except that it is relatively small compared with the observed variation in *y*, whatever *that* may be. (If the latter is reported as well, it is simpler for predictive or comparative purpose to give the size of the residual scatter directly and let anyone who wants to do so take its ratio to that of the y-variation.)

14.3 **Regression Equations**

The primary need in analysing a relationship between two variables is to describe how y actually varies with x. This means we have to specify an equation that somehow describes this variation. With scattered readings we also have to describe the scatter about the equation. Furthermore, various equations can then give reasonable fits, as Figure 14.4A illustrates. One criterion for choosing among such alternatives is by the degree to which they fit the data.

With the five pairs of readings in our example, the equation $y = 3x + 10$ had a residual standard deviation of about 5, or 5.1 to be more exact. But

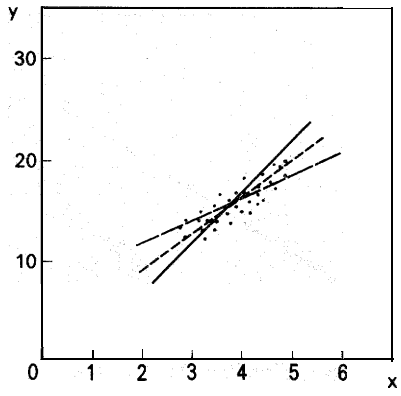

Figure 14.4A Three Possible Equations

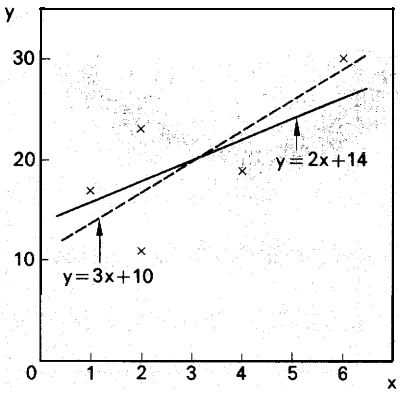

Figure 14.4B Two Equations for the Data in Figure 14.2A

Table 14.3 shows that an equation like $y = 2x + 14$ gives almost the same degree of fit. The residuals $(y - 2x - 14)$ have a standard deviation of 5.0. Figure 14.4B illustrates these equations. They look rather different from each other and the problem is which such equation to choose.

TABLE 14.3 The Fit of the Equation $y = 2x + 14$

(Data from Table 14.1)

						Average
x	1	2	2	4	6	3
y	17	11	23	19	30	20
2x + 14	16	18	18	22	26	20
y - 2x - 14	1	-7	5	-3	4	0

One possibility is to pick the equation that has the lowest residual variance. Because the variance is the average of all the squared deviations from the fitted line, the criterion involved here is called the "least-squares" principle.

The equation giving the minimum residual variance in the y-direction is called the regression of y on x. If we are fitting a linear equation, this must be of the form $(y - \bar{y}) = a(x - \bar{x})$, since the line has to go through the means $(\bar{x}, \bar{y})$ of the data. It can then be shown mathematically that the slope-coefficient a must equal the ratio of the covariance (xy) to the variance of x, or

$$a = \frac{\text{cov } xy}{\text{var } x}.$$

For our numerical example, the means are 20 and 3, the covariance is 10.25, and the variance of x is 4. The regression equation of y on x is therefore

$$(y - 20) = \frac{10.25}{4}(x - 3).$$

This reduces to

$$y = 2.6x + 12.2.$$

Table 14.4 shows the fit of this equation. The variance of the residuals is $94.96/4 = 23.7$ and their standard deviation is 4.9. The fit is therefore closer

TABLE 14.4 The Fit of the Regression Equation $y = 2.6x + 12.2$

						Average
x	1	2	2	4	6	3
y	17	11	23	19	30	20
2.6x + 12.2	14.8	17.4	17.4	22.6	27.8	20
y - 2.6x - 12.2	2.2	-6.4	5.6	-3.6	2.2	0

than for the equations mentioned earlier. In the "least-squares" sense it is the best-fitting line.

Exercise 14Q gives a short-cut formula for calculating the slope-coefficient of a regression equation. Regression equations for extensive data are now usually calculated on a computer.

The Regression of x on y

The regression equation just discussed provides a "best fit" to the data in the y-direction. It is the equation which minimizes the sum of the squared deviations $(y - ax - b)^2$. But this is a somewhat arbitrary criterion. For example, Figure 14.5A shows one could also consider the deviations of any point $(\hat{x}, \hat{y})$ *perpendicular* to the line, or deviations in the x-direction.

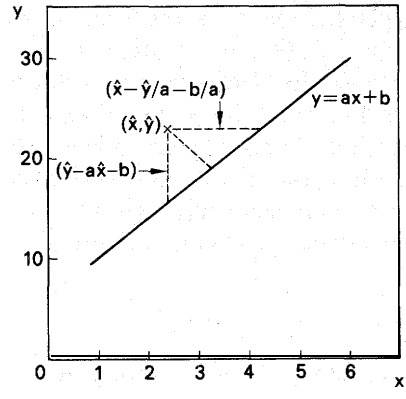

Figure 14.5A Vertical, Horizontal, and Perpendicular Deviations

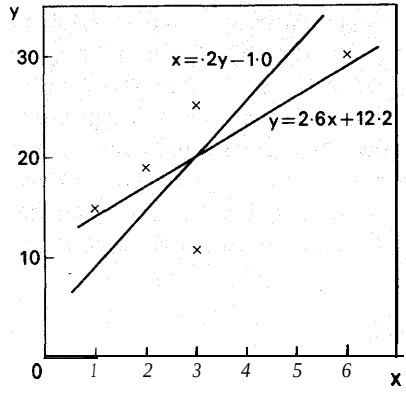

Figure 14.5B The Regressions of y on x and of x on y

Fitting a line by minimising the ***perpendicular*** deviations might seem an intuitively attractive approach, but is has a crippling disadvantage. Changing the units of one variable (as from inches to feet) results in a completely different equation. This approach is therefore not used in practice.

But if minimising the deviations in the ***y-direction*** seemed attractive, minimising the deviations in the ***x-direction*** must also be. The two things are different, and so are the results. Thus corresponding to the regression equation of y on x

$$(y - \bar{y}) = \frac{\text{cov}\,(xy)}{\text{var}\,(x)}(x - \bar{x}),$$

the regression of x ***on y*** is

$$(x - \bar{x}) = \frac{\text{cov}\,(xy)}{\text{var}\,(y)}(y - \bar{y}).$$

This equation is the best "least-squares" fit in the sense of minimising the deviations in the x-direction.

For our numerical example, this formula gives

$$x = .2y - 1.0.$$

Multiplying by 5 makes this equation read $5x = y - 5$, or $y = 5x + 5$. This is numerically quite different from the regression equation of y on x, $y = 2.6x + 12.2$ which we calculated earlier. The two lines are shown in Figure 14.5B. For x on y, a unit change in x corresponds to a 5-unit change in y; for y on x, a unit change in x corresponds to only a 2.6-unit change in y.

In general, any given set of data has two different regression lines. For the user this poses problems. Most statistical textbooks state that the regression of y on x is the best equation for predicting y from x; and that the regression of x on y is the best equation for predicting x from y. This idea is usually not explained any further and may seem confusing. We now examine it.

14.4 **The Non-comparability of Regression Equations**

If an equation is to give a correct prediction, it must hold for the new data about which the prediction is being made. But under what circumstances do the regression equations for one set of data also hold for another set of data?

Consider the height and weight data for Birmingham children referred to in Part II. These children had been classified into boys and girls, nine yearly age-groups from 5 to 13, and three different social classes. There were therefore 27 different sets of data for boys and 27 different sets for girls, a total of 54.

Each set of data has two regression equations. Figure 14.6A illustrates them for the middle social-class boys aged 5, 9 and 13 years. Table 14.5

TABLE 14.5 The Regressions of Weight on Height and of Height on Weight for Each Age-Group of Birmingham Boys in Social Class (3)

Age	Regressions of	
	Weight on height	Height on weight
5	w = 1.9h - 40	h = .35w + 28
6	w = 2.2h - 58	h = .31w + 32
7	w = 2.3h - 59	h = .29w + 33
8	W = 2.7h - 78	h = .23w + 37
9	W = 3.0h - 94	h = .19w + 40
10	W = 3.2h - 105	h = .21w + 40
11	W = 3.6h - 125	h = .16w + 43
12	w = 3.8h - 138	h = .19w + 42
13	w = 4.0h - 148	h = .16w + 45

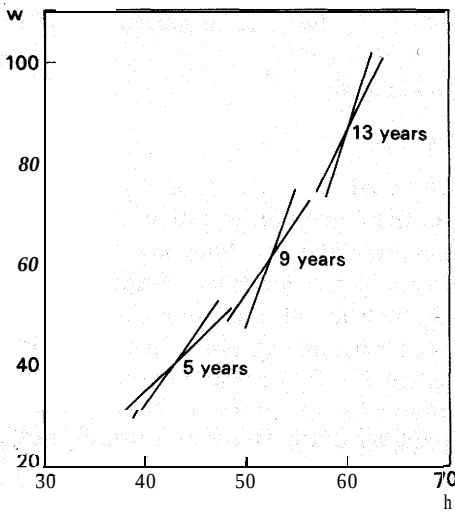

Figure 14.6A *h* on w and w on *h* in Yearly Age-Groups

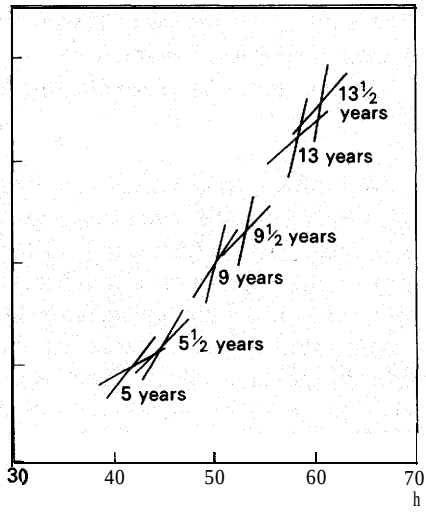

Figure 14.6B The Two Regressions for Half-Yearly Age-Groups

gives both regression equations for ***each*** age-group (based on equations reported in kilogrammes and millimetres by Healy, 1952, and on data provided by Clements, 1954).

We can see that the equations generally differ from each other. (For ease of comparability Table 14.5a shows the regressions of height on weight also in the form $w = ah + b$, i.e. having divided through by the initial slope-coefficients.) Thus for our 54 sets of data we have a total of 108 different regression equations. Had we analysed the data for half-yearly instead of

TABLE **14.5a** The Equations for **Social** Class **(3) with** the Regressions of Height on Weight **written** as w = ah + b

Age	Regressions of	
	Weight on height	Height on weight
5	w = 1.9h - 40	w = 2.9h - 8
6	w = 2.2h - 58	w = 3.2h - 10
7	w = 2.3h - 59	w = 3.4h - 11
8	w = 2.7h - 78	w = 4.3h - 16
9	w = 3.0h - 94	w = 5.3h - 21
10	w = 3.2h - 105	w = 4.7h - 19
11	w = 3.6h - 125	w = 6.2h - 27
12	w = 3.8h - 138	w = 5.3h - 22
13	w = 4.0h - 148	w = 6.2h - 28

yearly age-groups, we would have had 216 different regression equations, all even more different from each other, as Figure 14.6B illustrates.

There is nothing exceptional about these results. The height/weight data are not unusually complex. Indeed, we saw in Part II that a single generalisable relationship exists. The reason why regression analysis yields such complex results lies in the technique of analysis.

The two regression equations for a given set of readings go through the means of that data. Another set of data will usually have different means. Hence it will have two different regression equations, as Figure 14.7A shows. The two regression equations for the *first* set of data cannot both go through the means of the second set of data, because two different straight lines can only have *one* point in common. (Some special cases are discussed in Exercises 14C and D.)

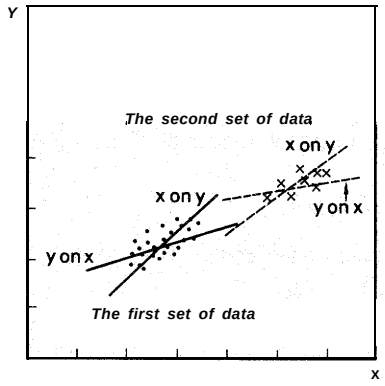

Figure 14.7A The Regressions of y on *x* and x on y in Two Different Sets of Data

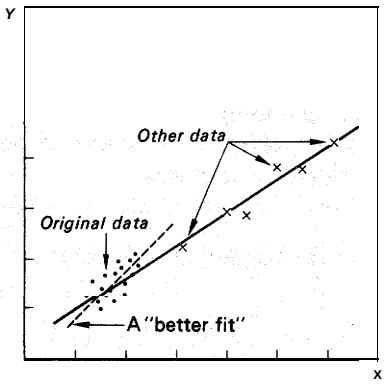

Figure 14.7B Generalization versus Better Fit

We therefore have the theorem that, in general, a regression equation fitted to one set of data cannot hold again for any other data. This is what theory says. And in practice no one has claimed anything different. There appear to be no cases quoted in any textbook where a regression equation fitted to one set of readings *has* held for another, different, set of data.

There is nothing in the theory of regression analysis that says it *should* lead to generalisable results. For an equation to be a best fit to one set of data does not say that it also needs to hold for other data. And with *two* possible regression lines for one set of data (y on *x* and *x* on y), it is generally impossible for *both* lines to hold again. They are the wrong type of equations for this purpose and no success has ever been claimed in this direction.

Residual Scatter Versus Empirical Generalisation

The basic problem is how to choose one equation from the various alternatives which give reasonable fits to a given set of readings. Regression analysis chooses an equation which gives a certain "best fit" for that particular set of data. But if one wants to use the equation for prediction to another set of data, then one wants the equation to hold also for this other set of data. That is quite a different criterion. In the one case we aim at the equation which fits one set of data *best;* in the other we aim at the equation, if any, which fits two (or more) sets of data *at all.*

There is no major conflict between fit and generalisation, if generalisation is put first. Faced with a choice among the many different equations that fit a set of data, we can choose the equation which generalises to *other* data. Figure 14.7B shows how an equation can give a good fit for the original set of data and for other data. It does not necessarily provide quite the "best" fit for either set in least-squares regression terms, but the fit is *good.* Although the regression equation gives the "best" fit, there are many other equations which are almost equally good. (That is precisely why there is a problem of choosing among such alternatives!)

As an example, we return to our five pairs of readings. The regression equation of y on x was

$$y = 2.6x + 12 \pm 4.9.$$

But at the beginning of Section 14.2 we also considered other equations like

$$y = 2x + 14 \pm 5.0,$$

$$y = 3x + 10 \pm 5.1.$$

The last term in each equation is the standard deviation of the residual scatter, given to two digits to point up the small differences that exist. Although the regression equation gives the "best fit", the residual scatters of the other equations are barely larger. We can also consider a fourth equation with a substantially higher slope-coefficient,

$$y = 4x + 8 \pm 5.7,$$

but even here the residual scatter is not *that* much larger than for the regression equation. Given that y varies almost 20 units from 11 to 30, could one say that the equation $y = 4x + 8 \pm 5.7$ is "wrong" whereas $y = 2.6x - 12 \pm 4.9$ is "right", just because of a 0.8 unit difference in the residual scatter?

The important conclusion is that different equations can give a very similar degree of fit. The criterion of "fit" is very inefficient in differentiating between equations. Therefore when faced with choosing between alternative equations we can select the equation that generalises to other data. This need not greatly affect the fit to the original set of data. This is the approach we discussed in Part II.

14.5 Regression with One Variable Controlled

Regression analysis is also used for data where one variable, say x, is controlled. This means that readings of y are observed only for specific values of x.

For example, we might select children aged 5, 9, 11, 14, and 15 years and measure their heights. Age A would then be a controlled variable. The data would consist of vertical arrays of the different h-values for each age selected, as in Figure 14.8A. Sometimes the controlled variable can be directly manipulated, e.g. with certain kinds of physical apparatus, in clinical trials testing different dosage levels of a drug, and in many other experimental situations.

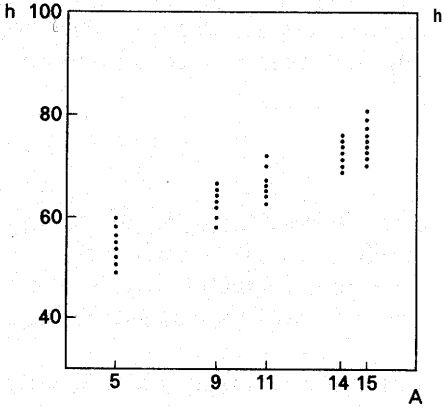

Figure 14.8A One Variable Controlled

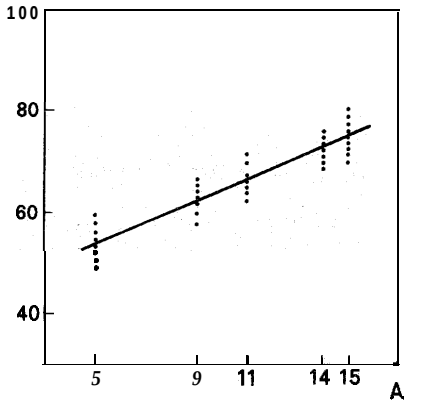

Figure 14.8B The Regression of h on A

Here the least-squares regression principle can only be applied in one direction, in our example minimising the variance of the h-values about the line to be fitted. Thus regression analysis gives only one equation in this case. But the principle of minimising scatter is really not needed to fix this line. For linear data the straight line can be determined without any extraneous principle because there is no choice between different possible equations. The line must go through the mean values of h for each controlled value of A, as Figure 14.8B shows.

The situation is therefore similar to that discussed in Part II. We have more than one distinct set of data and the fitted equation is an empirical generalisation that has to hold for each set of data. The only special factor here is that each set of readings has one variable that does not vary (so all the values of A in each set are equal to their mean, $\bar{A}$).

If the means $(\bar{h}, \bar{A})$ of the different sets of data lie on a straight line, then *that* is the relationship between *h* and A. The scatter of the individual h-values

does not determine which line to fit. (Least-squares regression would give this answer also, but virtually any "reasonable" method of fitting a line gives the same result in this case.)

The problem of choosing one equation from various alternatives only arises if the mean values of the different sets of data do not lie exactly on a straight line. But then the problem is one of fitting a straight line as a deliberate approximation to non-linear data. An initial working-solution can be determined as before, calculating the slope from the two extreme sets of data and putting the line through the overall means of x and y. (As an alternative, the regression of y on x could be calculated as a first working-solution, if it is easy to compute. But there is nothing "best" or especially attractive about this line. In fact the statistical requirements of linear regression theory are not even fulfilled, because the data are strictly non-linear.)

An initial working-solution fitted in such a situation is usually of no lasting importance. It may well need to be adjusted when further data become available, as we have seen earlier.

14.6 Errors in the Variables

A further problem with regression analysis arises if there are unsystematic errors of measurement in the data. The coefficients of the regression of y on x are affected by errors of measurement in x. In this situation instead of just two regression equations, the theory effectively distinguishes four different equations for any one set of data.

If x is subject to errors of measurement, which usually happens, then the regression of y on the observed values of x is generally not the same line as the regression of y on the "true" (error-free) values of x. Such problems do not arise with the lawlike relationships discussed in Part II. Using that type of analysis, the line is fitted to the mean values of each set of readings and therefore remains the same despite different kinds of error in the data. Only the residual scatter is affected by errors, which is as it should be.

With a *controlled* variable a special form of measurement error can arise. For example, if a drug dosage of 5 cubic centimetres is prescribed for a certain type of patient in a clinical trial, the "observed" values will all appear equal. But the actual amounts administered to different patients, the "true" dosages, may vary somewhat about this value.

The same applies to the height and *age* data. Figure 14.9A shows the mean heights of the five controlled age-groups. But the 5-year olds are not all exactly 5 years, that is only their *nominal* age when rounded to the nearest year. In practice, the ages will range between 5, and 6 years, or between $4\frac{1}{2}$ and $5\frac{1}{2}$, depending on how the rounding is done. (There may also be some "real" error of measurement if a G-year-old child is wrongly classified as being 5.) Figure 14.9B shows the true picture, which brings us back again to

the general case of two variables discussed in Part II. Errors in the controlled variable are therefore only a special case of this. As long as the errors are irregular and unbiased (i.e. average out at zero), they do not affect the relationship fitted to the mean values.

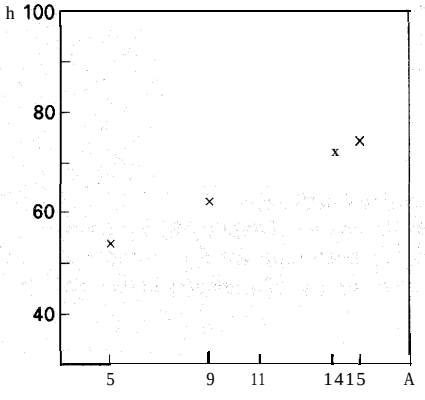

Figure 14.9A The Means of the h-arrays

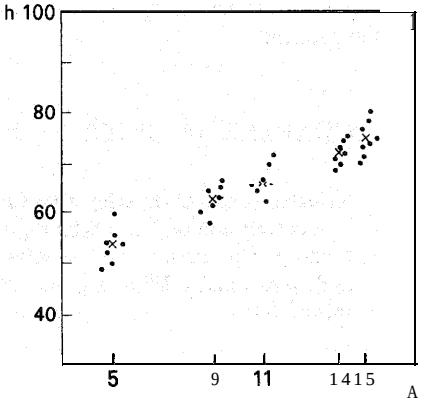

Figure 14.9B "Error" in the Controlled Variable

14.7 Summary

Correlation coefficients and regression equations are two statistical techniques that are widely used in analysing the relationship between two variables.

The correlation coefficient, denoted by the symbol *r*, is an index that measures how closely the two variables are related in a given set of data. But it only describes this covariation in relation to the total variation in y or *x* in that particular data. Two sets of data can therefore have the same correlation coefficients, but the scatter about the fitted relationships (the residual variances) can be of different sizes and the relationships themselves can also be different.

A regression equation is the equation that gives the "best fit" to the particular set of data being analysed. But there is more than one "best" equation. The regression of y on *x* minimises the residual variations in the y direction, and the regression of *x* on y minimises the residual variations in the *x* direction. These two regression equations are different for any set of data. Most statistical textbooks claim the regression of y on *x* is best for predicting y from *x*, and the regression of *x* on y is best for predicting *x* from y. But it is not made clear what this means in practice, nor is there any reason why it should be so.

The problem with regression equations is that the regressions for one set of data generally no not fit any *other* set of data. The reason is that the

two linear regression equations for one set of data cannot both go through the means of the other set of data. Regression equations are therefore not useful when building empirical generalisations, nor is this usually claimed for them.

Personally, I have not found either correlation or regression analysis of practical value. But they are widely used and therefore need to be described and evaluated.

CHAPTER 14 EXERCISES

Exercise 14A. Comparing Two Correlation Coefficients

A correlation of .8 has been reported for one set of data in the variables x and y. The same value has subsequently been reported for another set of data in x and y. What can one say about the x/y relationships in the two sets of data?

Discussion.

In both cases, high values of x tend to go with high values of y since each correlation is positive. However, the form of the relationships is not necessarily the same, it could be $y = 38x + 5$ in one case and $y = 0.1x - 10$ in the other. Nor is the size of the scatter about the relationships necessarily the same.

Alternatively, the relationships and/or the size of the scatter in the two sets of data *could* be the same. We cannot tell simply from the correlations.

Exercise 14B. Two Different Correlations

The correlation between the heights and weights of some 8-year-old children is .3 and that for children ranging from 5 to 10 years is .9. What does this difference mean?

Discussion.

The correlation for the 8-year olds *might* be lower because the scatter of their individual readings about a fitted height/weight equation is much larger than for the 5- to lo-year olds. But from previous knowledge it is clear that boys in a six-year age-group (from 5 to 10) differ far more from each other in their heights and weights than boys in a one-year age-group. Even if the two sets of data had the same size residual scatter, we would therefore expect a much higher correlation for the 5- to lo-year olds. (The correlations do not of course tell us whether the relationships between the two variables are the same in the two sets of data.)

Exercise 14C. Comparing Different Regressions

In Section 14.3 we fitted the two regressions

$$y = 2.6x + 12,$$

$$x = 0.2y - 1.0,$$

to the 5 pairs of readings in the numerical example (e.g. Tables 14.1 and 14.4). Table 14.6 gives a second set of data. Compare the regression lines of the two sets of data.

TABLE 14.6 A Second Set of Data

						Av.
x	11	12	12	14	16	13
y	7	41	53	44	60	50

Discussion.

The means of x and y in the new data are 13 and 50. The variances and covariance are 4, 50, and 10.25, the same as for the previous data.

The two regressions for the new data are therefore

$$(y - 50) = \frac{10.25}{4}(x - 13),$$

$$(x - 13) = \frac{10.25}{50}(y - 50),$$

giving

$$y = 2.6x + 16,$$

$$x = 0.2y - 2.8.$$

Both regressions differ from the corresponding earlier ones, although only in their intercept-coefficients.

Whenever the correlations of two sets of data are equal and the two pairs of variances are either equal or in the same ratio, the two pairs of regressions will be *parallel* to each other, as here. *The two pairs of regressions will be the same only if the two sets of means* (x, y) *are also the same,* as Figure 14.10 illustrates.

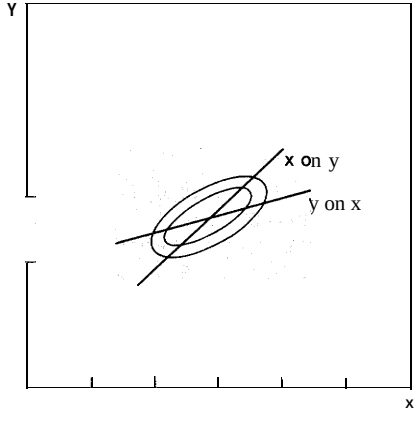

Figure 14.10 Two Pairs of Regressions the Same

The simplest case is of course when both sets of data are identical, i.e. with the same means, same correlations, and same standard deviations. The regressions fitted to the two sets of data will then be the same. But so would *any* line that was fitted by any method.

However, if the means are different in two sets of data, then at least one of the regression lines (y on x or x on y) has to be different for the two sets.

The conditions just outlined for equal or parallel regressions do not form a part of predicting a relationship. For example, if we use Boyle's Law, $PV = C$, to predict Pressure P for a certain value of V, we are only predicting that the relationship $PV \doteqdot C$ will hold again, *not* that the new data will also have the same means as the initial data and its standard deviations in the same ratio.

Exercise 14D. One Regression the Same

Can at least *one* regression be the same in different sets of data?

Discussion.

Figure 14.11 illustrates that it is technically feasible for one regression (say y on x) to be the same in sets of data with different means.

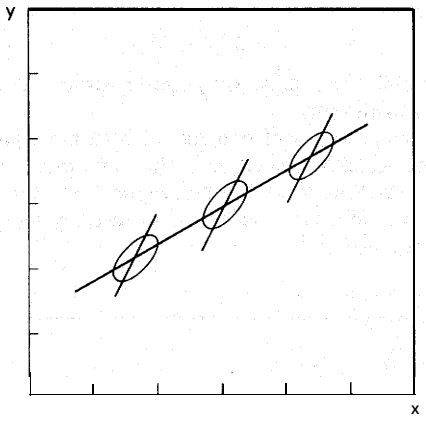

Figure 14.11 One Regression the Same

This case is not discussed in statistical literature. This seems right, because it appears to be technically "trivial", i.e. it could occur only accidentally. For example, suppose each set of the data in Figure 14.11 consisted of yearly age-groups. If each were divided into two sub-groups, such as boys and girls or *half-yearly* age-groups, with different means, then *both* regressions would differ from one sub-set to another. (Figures 14.6A and B provided relevant illustrations.)

Therefore the general conclusion remains. If two sets of data have different means, then the regressions fitted to one set of data will virtually never hold for the other set. To reach that conclusion one does not even have to analyse the data.

Exercise 14E. Two or More Sets of Data

How can one fit a regression equation to two or more sets of data?

Discussion.

The problem of fitting a regression equation to more than one set of data is not discussed in the general statistical literature.

Fitting separate regressions to each set of data generally gives the non-comparable results already illustrated.

Pooling the data into one group does not improve the situation. The result would be influenced by the arbitrary number of readings in the initial groups. Also, pooled data would not generally meet the theoretical requirements of statistical regression analysis (e.g. Normal Distributions, see also Exercise 14 P).

In any case, pooling is unnecessary. Any line fitting more than one set of data must go through the means of each, or at least approximately so. Knowing that is enough to determine a straight line if the data are in fact linear. No special principle of "least squares" or the like is needed.

Exercise 14F. Fitting a Regression to Mean Values

Can regression analysis be used to fit an equation to the mean values of different sets of data?

Discussion.

If the mean values lie exactly on a straight line, there is no problem of finding the "best fit". Only one equation is possible.

If the means do not lie on a straight line, as in Figure 14.12, then the theoretical requirements of regression analysis would not be satisfied and, strictly speaking, fitting a linear equation would be wrong (Exercise 14P). Regression is concerned with statistical analysis of *irregular* variations.

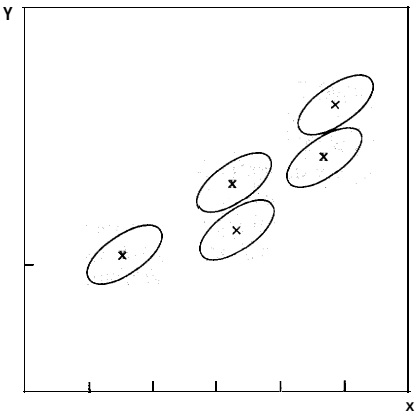

Figure 14.12 Means Not on a Straight Line

Fitting a straight line in this case would be a deliberate oversimplification and approximation of *systematic* deviations. Statistical literature does not discuss forcing a linear regression equation to fit data which are significantly non-linear.

Exercise 14G. Using a Regression Equation

Is it common to use a regression equation fitted to one set of data in the analysis of another set of data?

Discussion.

There is an apparent lack of such cases discussed in the general statistical literature. A seeming exception to prove the rule is a recent paper by Scott (1973), relating test-tank data for certain ship models to their subsequent performance on trial. (There must be other such cases, but they are not common.)

In Scott's paper, regression equations fitted to data from one test-tank laboratory held for data from another laboratory. But the two sets of data were rather similar (see Exercise 14C), so *any* type of line fitted to one would also fit the other. (The deeper methodological problem is that the "best fit" regression lines for the first set of data were not the *best* fit lines for the second set. It is therefore not strictly a case of *regression* lines holding again.)

Exercise 14H. Predictions for a Single Child

Could the height/weight regression equation for a specific age-group be used to assess whether a particular child of that age has the correct weight for its height, rather than using the more general height/weight relationship?

Discussion.

This question was answered some time ago as follows (Ehrenberg, 1968, p. 235):

"I have a daughter aged 7 who weighs 58.4 lbs. and stands 51.2 inches high.

"The regression of weight on height for 7-year-olds (in Table 14.5) is $w = 2.3h - 59$, so that. my daughter is .3 lbs. overweight. The regression of height on weight for 7-year-olds is $h = .29w + 33$, so that at 51.2 inches my daughter is 1.3 inches too tall for her weight.

"Being both too heavy for her height *and* too tall for her weight may be less confusing to statisticians than to my daughter, but she will be 8 years old in a day or two and then everything will be different anyway. Eight-year-olds are generally taller and heavier than 7-year-olds, and the regressions necessarily differ.

"For 8-year-olds, the regression of weight on height is $w = 2.7h - 78$, so that in a day or two my daughter will be 1.6 lbs. *underweight* for her height, compared with being .3 lbs. overweight now. (This of course does not happen with the lawlike relationship log w $= .02h + .76$, which holds

for both the 7- and the 8-year-olds.) All this leaves me with a conviction that statisticians never actually *use* the regression equations which they calculate."

A fundamental question is why should one use the 7-year-old data for Birmingham boys in 1947 for a comparison with one's 7-year-old London daughter in 1968? Given that the regression equations differed for all the age-groups in 1947, why should a comparison with the same *age-group* but different time, place, and sex be relevant? Yet a comparison with the regression equations for any other age would have led to different results, too.

Only a generalisable relationship that has been found to hold under a wide range of conditions, such as time, place, sex, age, etc., can be used as a yardstick or norm. Only for such an equation do we have evidence that the differences between then and now do not matter.

Exercise 141. The Least-squares Principle

What justifications are there for adopting the least-squares principle?

Discussion.

In terms of obtaining simple results, there would appear to be no justification.

On *a priori* grounds, Lindley (1947) in a major review paper said that the method of least-squares "is not easily justified except in certain cases" without saying what they are. Kendall (1951) in another review paper quotes Lindley to the effect that "there is something to be said for accepting the principle of least-squares in its own right" but he does not say what this something is. Anscombe (1967) tried more recently to justify some modifications of the least-squares approach and was reduced to ambiguous phrases like "it seems desirable", "a satisfactory solution", "there is reason to think", "it is natural to eliminate", and "we need not be abashed". Anscombe also said that those who disapprove of least-squares regression methods altogether have "undoubtedly much good reason on their side".

No one seems to have given a clear reason for adopting the regression method. In the past, the practical man has accepted the theoreticians' judgment that the least-squares approach will give the "best" solution; and the theoretician has believed that this kind of "best fit" is what the practical man wants. But although a physicist, for example, will want to know the scatter of observed readings about Boyle's Law, he never needed the scatter-minded statistician to tell him what the law actually is.

Exercise 14J. Collecting the Data

Does the way data in two variables are collected match the way regression analysis then deals with the data?

Discussion.

No. Collecting data in two variables usually consists of collecting a number of undifferentiated "items" (e.g. all boys aged 11 in a certain

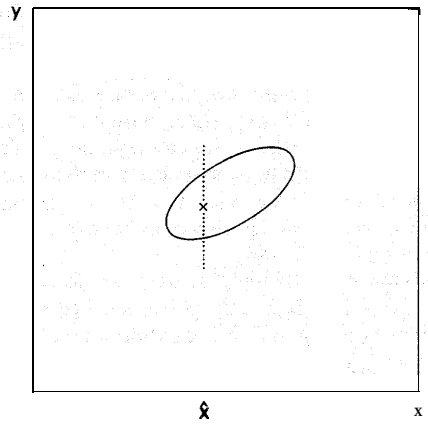

Figure 14.13 Values of y for a Selected Value of $\hat{x}$

school) and then measuring or recording their x and y values, as is illustrated by the ellipse in Figure 14.13.

The procedure differs from that used in the *analysis*, if regression is used. There we first select all items with x-values equal to some specific value $\hat{x}$. Only *then* do we look at their y-values. Figure 14.13 illustrates this selective approach in the analysis. (The case where the data are *collected* by first controlling one variable, x say, at certain specific values, has been discussed in Section 14.5.)

Exercise 14K. Significant Relationships

How do we judge whether there is a significant correlation or regression?

Discussion.

When fitting an equation to two variables x and *y*, whether by regression or other methods, the basic question is whether the observed y-values deviate less from the fitted line than they do from their own overall mean.

If the difference between the two types of deviation is small, then the relationship is weak. (As noted earlier, a correlation of 0.5 means that the residual standard deviation is only about 13 % smaller than the standard deviation of y from its mean.)

If we are dealing with *sample* data, a correlation may appear in the data due to errors of random sampling, when no such relationship exists in the population as a whole. This can be tested by statistical means, as discussed in Chapter 18, Section 18.4.

But there is no special merit in finding a relationship in one's data, especially if it means overinterpreting minor fluctuations in the data. It could be just as effective, and a far simpler result, to establish that y is *not* related to *x*, or that there is only a very weak relationship with a great deal of scatter-

Exercise 14L. Perpendicular Deviations

Illustrate the statement in Section 14.3 that the linear equation obtained by minimising perpendicular deviations is influenced by the choice of units.

Discussion.

Figures 14.14A and B show a simple illustration provided by G. J. Goodhardt. Suppose we have 4 readings of height and weight, as in Figure 14.14A. The line with the least perpendicular deviations is the vertical line shown, h = 0.5 yards. (This is visually obvious and can also be proved mathematically.)

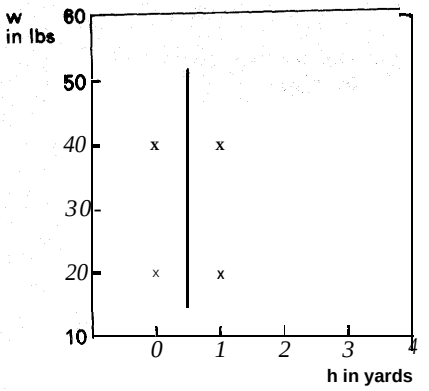

Figure 14.14A $h = \frac{1}{2}$ yard

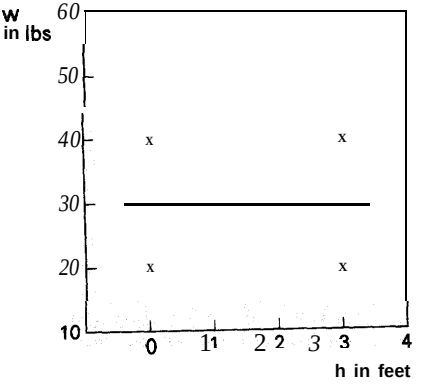

Figure 14.14B w = 30 lbs

In Figure 14.14B the scale of measurement for height has been changed from yards to feet. The best-fitting "perpendicular" line here is w = 30 lbs, quite different from the previous line. This example is an artificial one, but the same kind of effect occurs with any type of data.

Exercise 14M. Why are There Two Regression Lines?

Why does the regression of y on x not give the best-fitting equation in terms of deviations in both the x- and y-directions?

Discussion.

The regression of y on x is the straight line that gives the minimum variance of deviations in the y-direction. Similarly, the regression of x on y gives the minimum variance of deviations in the x-direction.

These two lines cannot be the same (unless all the points lie exactly on a straight line anyway). Figure 14.15 illustrates the problem. For a single point $(\hat{x}, \hat{y})$ and two particular lines, A and *B*, the figure shows that line A is closer to the point than line *B* in the vertical y-direction. In contrast, line *B* is closer to the point than *A* in the horizontal x-direction.

It follows that for a large number of observed points, the line minimising the (squared) deviations in one direction need not be the same as that

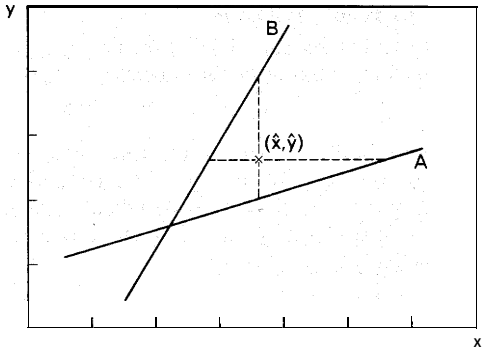

Figure 14.15 Closer to Line *A* in the y-direction and to Line *B* in the x-direction

minimising the deviations in the other direction. This can be proved rigorously.

Mathematically there is no doubt that there have to be two different regression lines, if a regression is defined as minimising the residual variance of $(\hat{y} - ax - b)$ or $(\hat{x} - cy - d)$. The question is whether this criterion is useful in summarising empirical data. It has been said that anyone who finds least-squares regression analysis easy cannot have understood it at all. But fitting a straight line to some data should not be that difficult.

Exercise 14N. A Special Case?

In Section 14.5, data with one variable controlled were said to be special cases of the situation considered in Part II, where equations were fitted to the means of different sets of data. But isn't this also a special case of regression analysis?

Discussion.

It is a special case of lawlike relationships because nothing is different. The fact that the x readings do not vary in each set of data affects nothing else.

In contrast, general bivariate regression analysis deals with one undifferentiated set of readings and yields *two* regression equations, whereas here we have two or more sets of readings and these yield only *one* equation. This is therefore not what is usually meant by something being merely "a special case" of a more general form of analysis.

Exercise 140. Best for Prediction

In Section 14.3 it was said that the regression of y on x is usually claimed to be best for predicting y from x, and the regression of x on y best for predicting x from y. What does the word "prediction" mean here?

Discussion.

Lindley (1947) defines the situation carefully. (No other writer seems to discuss the question so explicitly.) Lindley says in effect that in regression theory the prediction in question is for a new reading from the population which has already been observed and analysed.

But if we know that the reading comes from a particular population whose characteristcs have already been established, there is no problem of prediction. (The new observation must fit, within the usual errors of sampling.) Conventional statistical theory apparently uses the word "prediction" differently from the ordinary scientist or man in the street, who generally uses the word to refer to new data about which nothing is known *directly* when the prediction is being made.

Exercise 14P. Statistical Conditions

Are there technical restrictions on the kinds of data for which correlations and regressions can be calculated?

Discussion.

Formal statistical theory generally stipulates that for correlation and regression analysis the data should follow a Bivariate Normal Distribution. This means that the distributions of x readings and y readings should each be Normal, and that the distribution of the y-values corresponding to any particular value of x should also be Normal. If plotted on a graph the data should look roughly like that shown in Figure 14.1B (an egg-shaped ellipse).

This restriction is partly because of statistical sampling theory, and partly so that the standard deviations used in regression analysis have a descriptive meaning in terms of Normal Distributions. The restriction also guarantees that the mean values of y for a given value of x lie on the regression line (except for error with sample data). This is another way of defining the regression equation.

Statistical theory does not provide alternative analytic procedures for *non-Normal* data. However, correlation coefficients and regression equations tend to be calculated even when the data do not strictly satisfy the requirements, like the numerical example in the main text.

Exercise 14Q. A Computing Formula for the Covariance

Devise a short-cut formula for calculating the *covariance.*

Discussion.

The covariance of x and y is defined as the average cross-product of the x and y deviations from their means, $\bar{x}$ and $\bar{y}$; i.e.

$$\text{Covariance } (x, y) = \frac{\text{Sum } (x - \bar{x})(y - \bar{y})}{n - 1}$$

where n is the sample size. This formula entails first working out all the n individual deviations $(x - \bar{x})$ and $(y - \bar{y})$, and then cross-multiplying

them and adding. As a short-cut, Sum $(x-\bar{x})(y-\bar{y})$ can be calculated from the sum of the cross-products of the individual readings, Sum (xy), minus a simple "correction term" for the means, $n\bar{x}\bar{y}$. Thus,

$$\text{Covariance } (x, y) = \{\text{Sum } (xy) - n\bar{x}\bar{y}\}/(n-1).$$

(The mathematical proof is the same as for the short-cut formula for the variance in Exercise 11E.) For the numerical example in Table 14.1 this short-cut formula gives Sum (xy) $= 341$, $n\bar{x}\bar{y} = 300$. Hence the covariance is $41/4 = 10.25$, as before.

From this simple expression and the corresponding ones for the variances of x and y, the values of the correlation and regression coefficients can be calculated more easily. (Deviations of readings from their means, of the form $(x-\bar{x})$ and $(y-\bar{y})$, are the basis of factors called "moments" in mathematical statistics. Hence this particular type of correlation coefficient is named the "product-moment" correlation, which distinguishes it from many others invented in the earlier part of the century that are no longer used.)

CHAPTER 15

Multivariate Techniques

The most commonly used statistical techniques for analysing data in more than two variables are multiple regression and component or factor analysis. These are extensions of the regression and correlation methods used to analyse a *pair* of variables.

A multiple regression equation is derived by applying the least-squares principle between one variable and several others on which it might depend. The method has been most highly developed in econometrics, a statistically oriented branch of economics. Component or factor analysis on the other hand aims to group a large number of undifferentiated variables into sub-patterns or factors.

A common feature of these procedures is that the results depend on the particular version of the analysis technique adopted. The various arithmetical procedures need not be described in detail here because there are simple-to-use computer packages. Instead, we can concentrate on a broad description and an evaluation of the results.

The methods are often used when seeking to establish interrelationships among large numbers of variables in a single set of previously undigested data. They do not appear capable of providing generalisable results, nor is this usually claimed for them. I therefore do not find either the methods or the results useful, but that is of course a personal view. Since the methods are widely used in certain branches of statistics they require description and evaluation.

15.1 Multiple Regression

Suppose we want to analyse the yield levels of a certain crop over a number of years. Yield could depend on many factors : rainfall, temperature, the amount of fertiliser applied, the amount of sunshine, etc. So we might seek a multivariate equation. For illustrative purposes here we shall only consider two explanatory or "independent" variables, fertiliser f and temperature t, in relation to yield y. Table 15.1 gives readings for five successive

TABLE 15.1 Crop-Yield, Temperature and Amount of Fertiliser in 5 Successive Years

	Year					
	'61	'62	'63	'64	'65	Average
Yield, y	150	170	100	150	180	150
Temperature, t	11	12	5	7	15	10
Fertiliser, f	19	16	15	17	23	18

years, from 1941 to 1965. (Arranging the data in ascending order of y, say, would show clearly that there is strong correlation between all three variables. The problem is to *describe* the relationships in question.)

In general, no reasonably simple equation will estimate the crop-yields exactly. But the regression method will minimise the residual deviations, i.e. give the "best" tit for this particular set of data. Using this method, we want to determine a linear equation of the form

$$y = at + bf + c.$$

The coefficients a, b, and c have to be determined at those values which minimise the variance of the deviations $(y - at - bf - c)$. The coefficients a and b can be calculated from the correlation coefficients r_{yt}, r_{yf}, r_{ft}, and the standard deviations s_y, s_t, s_f of the variables. Thus, the coefficient a in our example, called the partial regression of y on t, is

$$a = \frac{r_{yt} - r_{yf} r_{ft}}{1 - r_{ft}^2} \cdot \frac{s_y}{s_t}.$$

This expression looks complex but is straightforward to use. (The calculations are now often done automatically on a computer, but Exercise 15A outlines the arithmetic.) If f and t are not correlated, the coefficients a and b simplify to the ordinary regression coefficients of y on t and y on f. The intercept-coefficient, here c, is determined by making the equation go through the overall mean values of all three variables (i.e. $c = \bar{y} - a\bar{t} - b\bar{f}$).

Applied to the data in Table 15.1 these formulae give the multiple regression equation

$$y = 8t - f + 88.$$

Table 15.2 shows the fit of this equation. The residual standard deviation over the five years is 14, which is fairly small compared with the 80-unit range in crop-yields. Because it is a regression equation, this is the smallest residual standard deviation that could be achieved by any linear equation among these three variables in this set of data.

TABLE 15.2 The Fit of the Multiple Regression Equation
$y = 8t - f + 88$

	Year '61	'62	'63	'64	'65	Average
Yield						
Observed, y	150	170	100	150	180	150
$8t - f + 88$	157	168	113	127	185	150
$y - 8t + f - 88$	-7	2	-13	23	-5	0*

* Standard deviation: 14

The fit of a multiple regression equation is often reported in a different manner, namely in terms of the multiple correlation coefficient, R. This measure is defined as the correlation between the observed values of y and the estimated values, $(8t - f + 88)$. It can be interpreted like the ordinary product-moment correlation coefficient in Section 14.2 as

$$R^2 = 1 - \frac{\text{residual variance}}{\text{variance of y}}.$$

In Table 15.2 the variance of the residuals $(y - 8t + f - 88)$ is 194 and the variance of y is 950. Thus the value of R^2 is $1 - 194/950 = .8$, and R is about .9. The fitted equation therefore accounts for about 80% of the variation among the variables ("variation" being conventionally measured in terms of the variance).

Interpreting the Regression Equation

In a purely "mathematical" sense, the interpretation of the equation $y = 8t - f + 88$ is straightforward. If f does not vary, then a 1-degree increase in temperature will increase yield by 8 units. If t does not vary, then a 1-unit increase in f will decrease y by 1 unit. If *both* f and t increase by 1 unit each, the total effect will be the sum of these separate effects: an increase in y of 7 units.

But the equation does not necessarily mean this in real life. The data analysed gives no evidence of how y varies with t when f is stable. Nor does it give any reason why the equation should state correctly the effects of *joint* variations in f and t, namely that an increase in fertiliser and a drop in temperature act *independently* of each other (e.g. that a 5-unit increase in f and a 2-degree drop in t will decrease yield by 21). The only *observed* increases in fertiliser occurred when the temperature also increased,

The negative coefficient of -1 for f in the equation appears to say that increased fertiliser will decrease yield, which might seem like nonsense. But if the five sets of readings are the only data we have, this objection is not

tenable. The chemical in question could have some deleterious effects, either intrinsically or because of some impurity, or because the applications are at too high a level and "burn" the plants. Or the fertiliser may encourage lush growth but diminish the actual crop that is harvested (e.g. a lot of straw but little wheat). Or the negative coefficient may be due to other factors altogether, quite unconnected with the effects of the fertiliser as such.

We need more information about the effects of various levels of fertiliser on this crop at various temperatures before we can begin to know how to interpret these phenomena, let along any equation that may be fitted. The particular equation fitted here is a multiple regression one, the "best" fit for this particular set of data, but there is no reason why the same coefficients should hold for any other data we may analyse next.

AlternativeEquations

Any doubts one may have about the interpretation of the regression equation are confirmed by the existence of alternative equations which also give fairly good fits to the data but not in the minimum residual-variance sense of regression analysis.

One example is the equation

$$y = 5t + 2f + 64,$$

as shown in Table 15.3. It does not matter how this was derived. The question is how well it works for the present and, ultimately, for other data.

TABLE 15.3 The Fit of the Alternative Equation
y = 5t + 2f + 64

	Year					
	'61	'62	'63	'64	'65	Average
Yield:						
Observed, y	150	170	100	150	180	150
5t + 2f + 64	157	156	119	133	185	150
y - 5t - 2f - 64	-7	14	-19	17	-5	0*

* Standard deviation : 15

The difference in the degree of fit of the two equations is small-a residual standard deviation of 15 here and 14 for the regression, compared with the 80-unit range of y-values in the data (and an R still of .9). It is especially small considering the fact that the regression equation implies an 8 unit increase in crop-yield per degree rise in temperature and the other only a 5-unit increase, and that the regression implies a 1 unit *decrease* in yield for each unit increase of fertiliser and the other a 2-unit *increase.* (This has nothing to do with the small number of readings in the example. The same variability

in coefficients would occur for similar data which extend over many more years or come from other sources.)

So we again see that markedly different equations can give approximately the same degree of fit. This occurs especially if the "independent variables", here fertiliser level and temperature, are themselves correlated, as in the present instance. The lowest values of both f and t occur in 1963, and the highest in 1965. (In econometrics this is referred to as "cohnearity" of the variables.)

Problems arise over which equation to fit and how to interpret its coefficients because multiple regression analysis seeks to find the "best" answer to a complex problem by analysing an isolated set of data. Irrespective of how limited or incomplete the data, the solution is always "best", with scant regard to whether it is any good. (A danger with high-powered salesmanship is that it often deludes the salesman himself.) These difficulties tend to be eliminated by building generalisable relationships from an increasingly wide range of different sets of data, as was described in Part II, where the additional data determine the equation to be fitted.

15.2 Component and Factor Analysis

Component and factor analysis are related techniques based on the correlation coefficient. They are used when there are a large number of different types of measurements for a given set of items (e.g. height H, weight W, girth G, leg length L, etc. for a sample of people). Because factor analysis started with the analysis of intelligence tests in psychology, the different measurements are often called "test variables".

These techniques aim to structure the data by reducing the numerous test variables to a smaller number of variables (components or factors) which account for most of the variation in the given data. A component is simply a linear equation of the test variables

$$\text{Component} = aH + bW + cG + dL + \text{etc.},$$

i.e. a weighted average. The numerical coefficients a, b, c, etc. are any numbers one cares to choose (subject only to the technical condition that the component must have a unit variance). A "factor" is a somewhat more complex version of the same model, with an "error term" added. For simplicity we talk here mostly of *component* analysis.

The difference between these techniques and multiple regression is that there are no direct observations of the component or factor variable on the left of the equation. Instead, component and factor analysis ***create*** new variables, usually by using some special analytic techniques. Having created these new variables, the analyst then has to find out what they are and what they mean.

An Illustrative Example

Suppose we have a sample of children and data giving six body measurements for each. The input for the analysis is a table of the correlation coefficients between all pairs of test variables, usually called a correlation matrix, as in Table 15.4. Each correlation between different variables necessarily occurs twice in the table, e.g. *L* with *H* and *H* with *L*. The diagonal entries are necessarily unity, e.g. the correlation of *H* with *H*. In practice the input may be 20 or more variables, with several hundred correlation coefficients. The purpose of the analysis is to find some sort of grouping or underlying structure for the test variables.

TABLE 15.4 The Correlations Between Six Body Measurements for a Group of Children

	H	L	A	C	G	W
Height	(1)	.7	.9	.3	.2	.8
Leg Length	.7	(1)	.8	.4	.3	.6
Arm Length	.9	.8	(1)	.4	.5	.7
Chest Circ.	.3	.4	.4	(1)	.8	.3
Girth	.2	.3	.5	.8	(1)	.4
Weight	.8	.6	.7	.3	.4	(1)

To help in this process a number of new variables, the components or factors, are created. The first step is to derive the correlations of each new component with the original test variables. These correlations are called the "loadings" of the test variables on the component.

Factor and component analysis each has many different versions and the results depend on which version is being used. We shall first illustrate the general nature of the approach in terms of a component extracted by a method akin to "centroid analysis", which for many years was the most popular form of factor analysis. Suppose we calculate the average of each row of the correlation matrix in Table 15.4. This gives a series of six numbers, .65, .63, .72, etc., shown to one digit (.7, .6, .7, etc.) in Table 15.5. These new numbers must each be less than or equal to 1 and it is therefore possible for each to be correlation coefficients.

We now define a new variable, the component *M* say, as that linear function of the six test variables which has these six numbers as its correlation coefficients with the six test variables. The new variable *M* therefore has correlations of about .7 with height, .6 with leg length, etc.

To understand this new variable better we note that all the test variables are "standardised" in this analysis. This means that the scales of measurement are changed so that each test variable has a mean of 0 and a standard

TABLE 15.5 The "Loadings" of the First Component M

	The First Component M
Height	.7
Leg Length	.6
Arm Length	.7
Chest	.5
Girth	.6
Weight	.6

deviation of 1. This is the same as changing from inches to centimetres, except that the new unit of measurement is chosen on the basis of the specific data. For example, if h is the height in inches for the sample of children with a mean of 45 and a standard deviation of 3, then the standardised height variable H used in the analysis is

$$H = \frac{\text{h} - \text{mean}}{\text{stand. dev.}} = \frac{h - 45}{3}.$$

A child 39 inches high will then have a standardised score of $H = -2$, i.e. it will be 2 standard deviations below the mean height. This standardising applies to all the test variables and also, as a matter of definition, to all the components.

We can now form the regression equations of the test variables on the component M. Since all the standard deviations equal 1, the slope-coefficient a in a regression equation is equal to the correlation coefficient. Thus

$$a = \frac{\text{cov}\,(x, y)}{\text{var}\,x} \quad \text{and} \quad r = \frac{\text{cov}\,(x, y)}{\sqrt{\text{var x}}\sqrt{\text{var } y}},$$

and since for standardised variables $\text{var x} = \sqrt{\text{var x}} = \sqrt{\text{var y}} = 1$,

$$r = a.$$

Because all the means are now zero, we simply have

$$H = .7M$$
$$L = .6M$$
$$A = .7M, \text{ etc.}$$

From the general formula for a regression equation that the residual variance = variance $y(1 - r^2)$, we know that r^2 is the proportion of the variance of y that is accounted for by the relationship with x, leaving $(1 - r^2)$ unaccounted for. Therefore the component M accounts for $.7^2 = .5$ of the observed variance of H, for $.6^2 = .36$ of the variance of L, for $.7^2$ of the variance of A, and so on. This one component accounts for a total of Sum $(r^2) = (.7^2 + .6^2 + .7^2 + .5^2 + .6^2 + .6^2) = 2.3$ or almost 40% of the

total observed variance of all six standardised test variables (which is 6 since each has a variance of 1).

To account for the remaining-60 % of the observed variation we need further components. It is generally found that a small number of components can account for the greater part of the variance. In effect it is one of the main considerations of component (and factor) analysis to find the least possible number. of components to account for the generally large number of test variables.

A common method for doing this is to extract the *Principal Components.* The way this works is that the first principal component accounts for more of the total variance of the test variables than any other linear component could. From the remaining variation the second principal component is extracted, and again accounts for more of that variation than any other component could. And so on. (By definition, the principal components are uncorrelated.)

The component *M* in Table 15.5 accounted for 40% of the total variance but Principal Component analysis can do better than that. Table 15.6 gives the loadings of the first three principal components, labelled I, II, and III.

TABLE 15.6 The Loadings of the First Three Principal component8 I, II, and III

	Principal Components		
	I	II	III
Height	.9	-.4	.1
Leg Length	.8	-.2	.4
Arm Length	.9	-.2	.1
Chest	.6	.7	.1
Girth	.6	.7	.2
Weight	.8	-.2	.4
Sum of (Loadings)^2	3.6	1.3	.4

Principal component I accounts for as much as 3.6/6 or 60% of the total observed variance. Component II accounts for about 22% and component III accounts for about 7%. There must be six components for six variables, and because the first components here were chosen to be so big, the remaining three must be small. Each can only account for an average of about 4% of the total variance (making a total of 100% for all six). Such small components are then ignored. It is in this way that component analysis finishes up with fewer components than test variables. In practice, where the number of test variables is usually 20 or more, only 3, 4, or 5 components tend to be extracted (or considered further) and hence there is a marked reduction in the number of variables.

Interpretation and Rotation

Now that we have extracted some components, e.g. the principal components, or "centroid" components like M, the next step in the analysis must be to interpret them. Principal component I accounts for 60% of the total observed variation (80% of the variance of H, 64% of the variance of L, etc.). This might seem fine, but we still do not know what this new variable actually is.

The interpretation of a component (or factor) is usually done by looking at the loadings and noting which test variables are highly correlated with the component. The component is then named accordingly. *This is usually the end of the analysis.*

For example, the correlations in Table 15.6 with component I are all large and positive, so that it is highly correlated with all six body measurements: if the values of the six measures for a particular person are large, the value of Component I will be large. Therefore it can be regarded as a measure of "Size". Component II is positively correlated with Girth and Chest Circumference but negatively with Height, etc., so it may be considered a measure of "Shape". Component III is more difficult to interpret in this way.

Difficulties in interpreting components are not uncommon. When they occur, alternative solutions to those first extracted (e.g. the first three principal components) are often sought. It is usual to adopt some linear functions of the initial components, e.g.

$$\text{New Component} = \hat{a}\text{I} + \hat{b}\text{II} + \hat{c}\text{III},$$

where $\hat{a}, \hat{b}$, and $\hat{c}$ are any numbers one cares to choose, subject to the technical restriction that the new component is again a standardised variable with unit variance. (Given that the initial components were uncorrelated with each other, or "orthogonal", the required condition is that $\hat{a}^2 + \hat{b}^2 + \hat{c}^2 = 1$.)

A change from one set of components to another is called a "rotation" because the technical procedures involved were initially developed by graphical or geometrical means. Rotation is a major part of modern component (or factor) analysis.

Alternative Solutions

There is in fact no unique solution in component (or factor) analysis. Other components can account for the observed correlations equally well. (One does not even have to start with principal components.) We shall now consider some alternative solutions for the data on the six body measurements.

Table 15.7 illustrates two other components, P and Q. (At this stage it does not matter how these new variables were derived.) Following the usual procedure of interpreting components by inspecting their loadings, we see that P is highly correlated with all the variables. It therefore looks like

another measure of "Size" (but not the same as the "Size" component I in Table 15.6). Component Q is correlated .6 with Height, only .2 with Leg Length, and *negatively*, $-.6$ and $-.5$, with Chest Circumference and Girth. The bigger the chest and waist, the lower the score on Q. It may therefore be called a measure of "Thinness".

TABLE 15.7 The Loadings on Components P and Q

	Components	
	P "Size"	Q "Thinness"
Height	.8	.6
Leg Length	.7	.2
Arm Length	.8	.4
Chest	.8	-.6
Girth	.6	-.5
Weight	.7	.4

These two components do not look very different from the principal components I and II in Table 15.6 except for a change of signs in Q, and they account for almost as much of the observed variation, namely 4.6, as against 4.9 for I and II. But their make-up is not the same, as we shall see later.

Table 15.8 gives another pair of components, labelled T and S. They are "rotations" or linear functions of P and Q, namely

$$T = .8P + .6Q,$$
$$S = .8P - .6Q.$$

TABLE 15.8 The Loadings of Components T and S

	Components	
	T "Tallness"	S "Chestiness"
Height	1.0	.3
Leg Length	.7	.4
Arm Length	.9	.4
Chest	.3	1.0
Girth	.2	.8
Weight	.8	.3

To interpret T and S we note that T is very highly loaded in Height (a correlation of 1, to the nearest first decimal). It may therefore be regarded as a factor of "Tallness". Correspondingly, component S may be regarded as a measure of "Chestiness".

Table 15.9 gives still more examples, X, *Y*, and *Z*, but these are not functions or "rotations" of an earlier solution. Component X is perhaps another measure of general "Size", or maybe better called "Tallness". Component Y, unlike any of the previous components, differentiates between Height and Leg Length (a positive correlation of .4 with *L* and a negative one of $-.4$ with *H*). It is therefore a measure of "Legginess" (being long in the leg relative to one's height). Similarly, component *Z* differentiates between Chest Circumference and Waist Girth, so it can be interpreted as reflecting an "Hour-Glass" shape (big chest and small waist or vice versa). However, Components Y and *Z* only account for 7% and 6% of the total variance.

As noted earlier, the analysis usually stops when the components or factors have been *named* and some kind of structure has been imposed on the initial data.

TABLE 15.9 A Three-Factor Solution

	Component		
	X "Size"	Y "Legginess"	Z "Hour-Glass"
Height	.9	-.4	.2
Leg Length	.9	.4	.2
Arm Length	.9	-.1	-.2
Chest	.4	.1	.3
Girth	.4	.1	-.3
Weight	.4	-.3	-.2

Components and Factors

Instead of concluding the formal analysis when the components have been "named", each individual in the sample could now be given a score or value on each of these new variables. These scores can be calculated from the linear equations in the initial test variables that define the components, as stated at the very beginning of this section. But in practice this is seldom done.

Factor analysis differs in this respect in that the new variables cannot actually be calculated, This is because of differences in the definition of the techniques. But since in component analysis the values of the new variables are seldom calculated, this difference between the two techniques may not be crucial. In other respects the two techniques are rather similar. Factor analysis used to be practised more in the first half of the century, but component analysis has become far more popular now, perhaps because the computational difficulties connected with principal component analysis have been by-passed by the computer.

As already stated, component scores are seldom calculated (or used) in any subsequent work. But it is helpful to illustrate their nature. The principal

components in Table 15.6 are given in terms of the six body measurements by the equations

$$\text{I} = .12H + .11\text{L} + .13A + .09C + .09G + .11W,$$

$$\text{II} = .27H + .14L + .12A - .46C - .47G + .15W,$$

$$\text{III} = .23H - 1.47L - .27A - .44C + .54G + 1.47W.$$

(The coefficients are derived by matrix algebra from the initial correlation matrix in Table 15.4 and the loadings matrix in Table 15.6.) Each principal component is therefore a function of all or most of the variables. This is typical of most components that are reported in practice (and implicitly, of many factors).

Some of the alternative solutions given above were, however, deliberately chosen to differ in this respect. For example, components P and Q in Table 15.7 are functions of only two variables, Height H and Chest Circumference C,

$$P = .6(H + C),$$

$$\text{Q} = .8(H - \text{C}).$$

The components S and T in Table 15.8 are "rotations" of P and Q, i.e. linear functions of the form

$$S = .8P + .6Q,$$

$$T = .8P - .6Q,$$

so that

$$S = .8 \text{ x } .6(H + \text{C}) + .6 \text{ x } .8(H - \text{C}) = H,$$

$$T = .8 \text{ x } .6(H + \text{C}) - .6 \text{ x } .8(H - \text{C}) = \text{C},$$

to a close degree of approximation. Thus any of the original variables can itself be a "component".

All of the alternative solutions discussed earlier were derived by deliberately choosing some function of the initial variables. For example, the "Legginess" component Y in Table 15.9 is simply $1.3(H - L)$, and the "Hour-Glass" component Z is $1.6(C - \text{G})$. It is therefore not surprising that these two components correlated with the test variables in the way they did.

There is nothing wrong with such "arbitrary" methods of constructing components. A component is merely *any* linear function of the test variables. The justification and meaning of any particular component lies in what it is empirically found to measure.

Problems of Interpretation and Use

Component analysis and factor analysis are highly quantitative techniques, but the interpretation of their results tends to be *qualitative*, i.e. some

new variables are "named". Whether it is useful to have an analysis which merely asserts that body measurements are mainly made up of three "factors", "Size", "Shape", and "Legginess", must at best remain a matter of personal judgment.

In general, even factor analysts make little *numerical* use of their results in further work. (Instead, they carry out new factor analyses.) There are several reasons for this.

One is that the "parsimony" achieved in reducing a large number of test variables to a few components generally does not result in any of the original variables actually being discarded, but just in adding additional ones, the new components. (As we have seen, components I, II, and III typically are functions of *all* the initial variables.) Moreover, before the analysis these new variables were altogether unknown quantities. Not even their existence was known. Therefore no previous results exist with which these components can be directly compared. This is what occurs in every such analysis.

Next, the concern with the proportion of the total variance that each component accounts for leads to difficulties. In an analysis such as that of Table 15.9, the "Legginess" component would generally be discarded because it accounted for only a small proportion of the total observed variation. But this depends entirely on which other test variables were included in the analysis, a more or less arbitrary decision, If other measures of leg length (e.g. length of shin and length of thigh bone) had been included, legginess would be a numerically "important" factor. (In this particular respect, component and factor analysis have been likened to a sewer : "What you get out of it depends on what you put into it.".)

Factor and component analysis are often said to be useful at the early stages of analysing new kinds of data, when nothing much is known. But such a degree of ignorance cannot occur often. Furthermore, the techniques do not facilitate comparisons across other sets of data and the building up of generalizable quantitative knowledge. Both the input and the output (the "loadings") are correlation coefficients. As seen in the last chapter, if the correlations or loadings in two studies are the same, it does not follow that the relationships in question are the same. The process of standardising the test variables makes comparisons even more difficult. In one set of data a variable might need to be divided by a standard deviation of 3, in the next by a standard deviation of 5, and so on. Thus one cannot readily compare the different results, nor is this generally claimed for the techniques.

An alternative approach to multivariate data is to start with more structured data. For instance one could study the systematic interrelationships between different body measurements for different groups of children (differing by race, age, sex, etc.). Chapter 9 gave examples where the aims were broadly the same as here, to reduce a large number of variables to simpler patterns. Only the methods and the results differed.

The results of a factor or component analysis are usually not altogether "obvious" even after the analysis, especially with regard to the quantitative details. One is therefore relying on the particular version of the analytic technique used to "justify" one's findings (e.g. "as shown by factor analysis with 'Maximax' type of rotation"). While component and factor analysis are often regarded as *objective* ways of reducing large numbers of variables, in fact they are not. The choice of variables to put into the analysis, whether to use factor analysis or component analysis, the choice of a particular version of these techniques, the choice of a particular type of "rotation", and the interpretation of the results are all *subjective.* There is nothing inherently wrong with subjectivity since judgment has to be exercised in many problems in data analysis. But the adoption of some arbitrary analytic technique should not be regarded as objective justification for one's results.

15.3 Other Multivariate Techniques

Other statistical techniques of multivariate analysis include discriminant analysis, cluster analysis and multi-dimensional or non-metric scaling. These techniques are as yet less widely used than multiple regression and component or factor analysis.

Cluster analysis and multi-dimensional scaling are currently modish and have received considerable impetus through the increasing availability of computer programmes which take the drudgery out of the arithmetical calculations. Where factor analysis seeks to establish groups of test variables which measure the same thing (i.e. are "highly loaded" in the same factor), cluster analysis aims to find groups of items (e.g. people) that are relatively similar to each other. (An analogy from biological taxonomy would be the classifying of plants or animals into species, but this particular set of results was not achieved by any method of *statistical* cluster analysis.) Multi-dimensional scaling also seeks to establish groupings in data, but it uses less information (or fewer assumptions) than the other techniques.

However, when facing many variables, the usual analysis problem is not that we have little information, but that we do not know how to interpret and integrate the information that we already possess. None of these techniques seem to have shown themselves capable of building a growing body of coherent knowledge.

Discriminant analysis is an older technique with a rather different purpose. The question posed is whether two sets of objects differ from each other. A typical example from anthropology is two sets of skulls found in different locations. Are they different or do they stem from the same kind of people? For any one measurement (e.g. depth of cranium) it may not be clear whether the two sets of skulls differ "significantly". The two observed *means* differ, but there is a good deal of overlapping scatter in the readings from individual

skulls. Discriminant analysis then aims to construct a linear function of a number of such measurements so that the mean of this new function discriminates "best" between the two sets of data, relative to the scatter of values for individual items.

If the difference between the means is "significant" when subjected to a statistical test of significance, then the two sets are regarded as coming from different populations. However, since the original data do not usually derive from any explicit random sampling from a larger population, it is not altogether clear what this kind of procedure really signifies.

15.4 Summary

Multiple regression and component or factor analysis are relatively advanced statistical techniques used to structure the relationships between large numbers of variables in a single set of data.

Multiple regression gives the "best fit" to a single set of multivariate data, and component and factor analyses create a small number of completely *new* variables to explain the large number that have been observed.

Both types of technique suffer from difficulties in interpreting the results. Since the analyst can choose among many variations in the techniques, they are less "objective" than might appear. They also lack the discipline of having to obtain reproducible and generalisable results, since the techniques are not designed to lead to such findings.

CHAPTER 15 EXERCISES

Exercise 15A. Calculating the Multiple Regression Equation

Using the data in Table 15.1 and the formula for the partial regression coefficient in Section 15.1, calculate the multiple regression equation $y = 8t - f + 88$. Also calculate the multiple correlation coefficient $R = 0.9$.

Discussion.

The partial regression coefficient of y on t, with f "partialled out" or "held constant", was defined as

$$\frac{r_{yt} - r_{yf}r_{tf}}{1 - r_{ft}^2} \times \frac{s_y}{s_t}.$$

We therefore need to calculate the pair-wise correlation coefficients r_{yt}, r_{yf}, and r_{ft} and the standard deviations s_y and s_t. Applying the formula of Exercise 14Q to the data in Table 15.1, the reader should obtain the values

correlation between y and t, $r_{yt} = .92$,

correlation between y and f, $r_{yf} = .67$,

correlation between f and t, $r_{ft} = .79$,

and $s_y = 30.8$, $s_t = 4.0$. The partial regression coefficient of y on t is therefore

$$= \frac{.92 - .67 \times .79}{1 - .79^2} \times \frac{30.8}{4.0}$$

$$= 8.0.$$

Similarly, the partial regression coefficient of y on f with t partialled out is

$$\frac{.67 - .92 \times .79}{1 - .79^2} \times \frac{30.8}{3.2} = -1.5,$$

$= -1$ to the nearest whole number for simplicity.

(Rounding this coefficient to -2 makes virtually no difference to the fit of the equation.) The intercept-coefficient is calculated by substituting the means of three variables in the equation with these two coefficients, i.e.

$$150 = 8 \times 10 - 1 \times 18 + \text{intercept-coefficient},$$

so that its value is 88.

The multiple correlation R can also be calculated from the pair-wise correlation coefficients. The formula given in specialist texts for the multiple correlation of y with f and t is

$$R^2 = \frac{r_{yt}^2 + r_{yf}^2 - 2r_{yt}r_{yf}r_{ft}}{1 - r_{tf}^2}$$

$$= \frac{.92^2 + .67^2 - 2 \times .92 \times .67 \times .79}{1 - .79^2}$$

$$= .84,$$

so that R is about .9. (In Section 15.1 we calculated the same value of R from an alternative formula in terms of the variance of the residual deviations.)

Exercise 15B. The Linear Multiple Regression Model

What are the main assumptions in the linear multiple regression model?

Discussion.

First, that the observed relationships are (approximately) linear. If they are not, this can often be overcome by transforming one or more of the variables.

Second, that there is no interaction among the independent variables. In our example this means that the effect of a unit change in fertiliser f does not depend on the temperature t. But there is no reason why this should be the case. Indeed, at normal temperatures, application of a suitable fertiliser will have a positive effect on yield, but at extreme temperatures nothing grows so the level of fertiliser will have no effect at all. In less extreme cases the two factors may still interact. If there is interaction, it will show up in systematic patterns of the deviations from the fitted line. Sometimes this can be overcome by non-linear transformations of scale.

Third, that one has selected the main variables that matter. If not, this will show up in any other set of data because the relationship between y, f, and t will be very different. This should mean that some other factor or factors has not been allowed for.

Fourth, that the deviations or "errors" from the fitted line fulfil certain statistical requirements, without which the least-squares regression method is not technically valid. For example, there has to be "homoscedasticity" (i.e. the average size of the "errors" has to be constant all along the line). Next, the "independent" variables, t and f, have to be free from any errors of measurement or the like (which is seldom true). The errors in the *dependent* have to be normally distributed and independent of each other, or "serially uncorrelated", which is often not the case in time-series, for example. (More complex regression procedures have been developed, especially in econometrics, for situations where some of these assumptions are not true.)

Finally, that each variable covers a good deal of its relevant range of variation for various *different* values of each of the other variables. Otherwise the equation mostly consists of extrapolations beyond the observed data.

Exercise 15C. The Best-fitting Line for Other Data

The multiple regression equation is the "best-fitting" line for the given data in the least-squares sense. Will it also provide the best fit for other data?

Discussion.

Suppose the equation in question also holds for another set of data, in that it goes through the new mean values and has irregular deviations of about the same size as in the original data. Then there will almost certainly be other linear equations which can fit the new data better (such as the multiple regression equation for that set). There is no reason why the original line should be the best-fitting one for the new data, even when it fits at all. This implies that mechanical application of a "best-fit" criterion gives results which generally cannot hold for other data. There is nothing in the literature of statistics, either in the theory or in the results, to indicate the contrary. As Corlett (1963) has put it in his *Ballade of Multiple Regression*:

> "Your optimum only is bonum
> for the data you've fitted it to!"

Exercise 15D. A Text-book Example

In his *Principles of Econometrics*, Henri Theil (1971, p. 101) gives for his basic example of multiple regression some time-series data on textile consumption (C), real per capita income (I), and the relative price of textiles (P) in the Netherlands from 1923 to 1939 (indexed on 1925 and transformed to logarithms, as shown in Table 15.10).

The question Theil poses is that since classical demand theory indicates real income and relative prices are the variables that determine the consumption of various commodities, to what extent does statistical data show these variables account for the variation in textile consumption over time? Answer his question.

TABLE 15.10 Time-Series of Dutch Textile Consumption

Year	Volume of Textile Consumption per Capita* (1)	Real Income per Capita* (2)	Relative Price of Textiles* (3)	Logarithms of Columns (1)-(3)		
				$\log_{10}(1)$ (4)	$\log_{10}(2)$ (5)	$\log_{10}(3)$ (6)
1923	99.2	96.7	101.0	1.9965 1	1.98543	2.00432
1924	99.0	98.1	100.1	1.99564	1.99167	2.00043
1925	100.0	100.0	100.0	2.00000	2.00000	2.00000
1926	111.6	104.9	90.6	2.04766	2.02078	1.95713
1927	122.2	104.9	86.5	2.08707	2.02078	1.93702
1928	117.6	109.5	89.7	2.0704 1	2.03941	1.95279
1929	121.1	110.8	90.6	2.08314	2.04454	1.95713
1930	136.0	112.3	82.8	2.13354	2.05038	1.91803
1931	154.2	109.3	70.1	2.18808	2.03862	1.84572
1932	153.6	105.3	65.4	2.18639	2.02243	1.81558
1933	158.5	101.7	61.3	2.20003	2.00732	1.78746
1934	140.6	95.4	62.5	2.14799	1.97955	1.79588
1935	136.2	96.4	63.6	2.13418	1.98408	1.80346
1936	168.0	97.6	52.6	2.22531	1.98945	1.72099
1937	154.3	102.4	59.7	2.18837	2.01030	1.77597
1938	149.0	101.6	59.5	2.17319	2.00689	1.77452
1939	165.5	103.8	61.3	2.21880	2.01620	1.78746

*Index base 1925 = 100.

Discussion.

The multiple regression equation for the data (in terms of the log variables) is

$$\log c = 1.14 \log I - .83 \log P + 1.37.$$

The reader can calculate this from the data, using the formulae set out in Exercise 15A. The equation fits to within a residual standard deviation of about .014 in $\log C$ units, which is small compared with the range of variation of log C from 2.00 to 2.22.

Exercise 15E. Alternative Equations

The multiple regression equation gives the "best" fit to the textile data, but there must be other equations that fit almost as well. Derive one.

Discussion.

Suppose we look at the last three columns of Table 15.10:

Column (4) increases fairly steadily from 2.00 to 2.22.

Column (5) varies little and irregularly between 1.99 and 2.02,

Column (6) decreases almost steadily from 2.00 to 1.79.

Columns (4) and (6) therefore complement each other. They run from 2.0 to about 2.2 and from 2.0 to about 1.8, and the two figures generally add to about 4. So we have approximately that

$$\log c + \log P \doteqdot 4.$$

Table 15.11 shows this more clearly (in the form $\log CP \doteqdot 4$).

TABLE 15.11 Textile Consumption, Price, and Income

(Netherlands, 1923-1939)

Year	log c	log P	log CP*	log I	log CP/I**
'23'	2.00	2.00	4.00	1.99	2.01
'24	2.00	2.00	4.00	1.99	2.02
'25	2.00	2.00	4.00	2.00	2.00
'26	2.05	1.96	4.01	2.02	1.99
'27	2.09	1.94	4.03	2.02	2.01
'28	2.07	1.95	4.02	2.04	1.98
'29	2.08	1.94	4.04	2.04	2.00
'30	2.13	1.92	4.05	2.05	2.00
'31	2.19	1.85	4.04	2.04	2.00
'32	2.19	1.82	4.01	2.02	1.99
'33	2.20	1.74	3.99	2.01	1.98
'34	2.15	1.80	3.95	1.98	1.97
'35	2.13	1.80	3.93	1.98	1.95
'36	2.23	1.72	3.95	1.99	1.96
'37	2.19	1.78	3.97	2.01	1.96
'38	2.17	1.77	3.94	2.01	1.93
'39	2.22	1.79	4.01	2.02	1.99
Av.	2.12	1.87	*3.99*	*2.01*	1.98

* log CP = log C + log P. ** log CP/I = log C + log P - log I.

We also note from Table 15.11 that the small variations that exist in this sum tend to be in line with the small variations in the income variable, log *I*. If we subtract log I from log C+ log *P*, we have a quantity that varies even less. It is more or less constant at about 1.98. Hence we have the equation

$$\log C + \log P - \log I = 1.98, \quad \text{or} \quad \log C = \log I - \log P + 1.98.$$

The residual standard deviation of this equation is .025, compared with .014 for the multiple regression.

The new equation is similar to the multiple regression shown in the last exercise, but much simpler. The intercept-coefficient, 1.98, is the only coefficient that is not unity. If rounded this coefficient is 2.0, which is the log of 100. It is primarily a scale-factor caused by all the variables being indexed as "1925 = 100". (If the indices had been expressed as "1925 = 1.00", this coefficient would effectively be unity as well.)

Exercise 15F. The New Equation $\log C = \log I - \log P + 1.98$

Consider the meaning of this new relationship.

Discussion.

Eliminating the logarithms, the relationship reads $CP = 96I$ in the indexed variables C, P, and I ("1925 = 100"). Expressed in terms of real data, the relationship will therefore be

$$CP = kI,$$

where k is some numerical coefficient which depends on the units of measurement of C, P, and I. (The multiple regression would correspondingly read $CP^{.83} = \hat{k}I^{1.14}$, where $\hat{k}$ is a numerical coefficient reflecting the antilog of the intercept-coefficient 1.37 and the change to the original units of measurement.)

The product *CP* stands for the per capita volume of textile purchases times their relative price, and thus equals per capita *expenditure* on textiles. The relationship $CP = kI$ therefore says that per capita expenditure on textiles was an approximately constant proportion of consumer income.

However, real income varied very little over the 17 years in Table 15.11 (the fluctuations average at about 5%). We therefore have as a simpler approximate result that *CP*, per capita expenditure on textiles, was approximately *constant*.

(One further question is to what extent the adjustments of incomes in turning actual incomes into "real" incomes reflect the variation in price levels used to adjust textile prices to "relative" prices? An examination of this point may also throw light on the small decreasing trend in the quantity ($\log C + \log P - \log I$), in the last column of Table 15.11.)

Because of the simplicity of the $CP = kI$ relationship, it is relatively easy to move away from the "1924 = 100" type of index and "adjusted" figures to examine the data more openly in terms of simple systems of relationships. We may also be moving from curve-fitting to economics.

The data here may seem exceptionally simple. But if multiple regression in skilled hands produces such results in a *simple* case, what can it do in more complex ones?

Exercise 15G. Other Things Being Equal

Under what conditions can the fitted equations be applied?

Discussion.

Theil (1971, p. 116) interprets the multiple regression equation

$$\log C = 1.14 \log I - .83 \log P + 1.37$$

as meaning that when real incomes go up by 1per cent and price goes down by 2 per cent, textile consumption will go up by about $1.14 - 2(-.83) = 2.8$ per cent, if all other determining factors remain the same (the common "ceteris paribus" assumption of economics). Since he does not establish what these other factors are, this assumption can only mean that the equation will hold in all those other situations where it will hold.

There is no reason for such a situation ever to arise. For example, if the initial analysis had been from 1923 to 1938 instead of to 1939,

multiple regression analysis would have given an equation with different coefficients. Such *different* equations cannot all hold for other data.

In contrast, the coefficients of the equation $CP = kI$ do not depend on the almost haphazard selection of readings. Therefore it is possible for this equation to generalise. Whether it does is a matter for empirical research. No doubt there will be cases where it does not hold, where the relationship between the variables is different. This one also has to determine and explain.

Its simplicity makes it easy to compare the equation $CP = kI$ with other cases, e.g. for Dutch textiles in other years, for textiles in other countries, for other products in Holland or in other countries, when volume goes *down* (as well as for the present case when volume went up), when (real) income levels change, and so on. In this way, we can build empirically based knowledge of the conditions under which the result can be applied and of those where it cannot.

Exercise 15H. Prior Knowledge

How can any other data on consumption, price, and income now be analysed?

Discussion.

In introducing the Dutch textile example, Theil stated that the classical demand theory of economics indicates real income and relative prices as the variables which determine the consumption of a commodity (see Exercise 15D). Now we have more specific theoretical knowledge: for the Dutch textile data, $CP \doteq kI$. The question is whether this equation holds for the new data. If it does not, the question is what the discrepancies are like, what an alternative relationship would look like, and how far would this new equation generalise? There is no longer any need for a "blind search" procedure (such as looking for a "best fit") when faced with another set of data.

Exercise 15I. Factor and Component Analysis

What is the main difference between factor analysis and component analysis?

Discussion.

A component is a multivariate linear equation (or weighted average) of all the standardised observed variables, i.e.

$$\text{Component} = aH + bW + cG + \cdots,$$

where a, b, and care numerical coefficients. Different methods of component analysis will determine different numerical coefficients.

In contrast, a factor is a new variable that cannot be directly expressed in terms of the observed data. This makes it more difficult to deal with.

Exercise 15J. Creating New Variables

Are there any constraints on the kinds of components or factors one may construct?

Discussion.

By definition any factor or component is a new variable. Any linear combination of the test variables may be put forward as a new component. There is not even any restriction to use the given test variables, which are only an arbitrary collection anyway.

Exercise 15K. The Comparison of Different Component Analyses

Table 15.12 reproduces the loadings of two components, *P* ("Size") and Q ("Thinness"), on the six test variables.

TABLE 15.12 The Loadings on Components P and Q

	Components	
	P "Size"	Q "Thinness"
Height	.8	.6
Leg Length	.7	.2
Arm Length	.8	.4
Chest	.8	-.6
Girth	.6	-.5
Weight	.7	.4

Suppose that the same set of loadings on these six variables were found in another study, for two components provisionally labelled D and E. Are *P* and *D* the same, and Q and E?

Discussion.

Each component analysis says that a variable exists, *D* in one case, *P* in the other, which has correlations .8 with *H*, .7 with *L*, etc. It does not follow that *D* and *P* are the same since correlations cannot determine that sort of thing. Many different variables can have correlations of .8 with *H*, .7 with *L*, etc. These variables may be correlated to some extent, but they need not (and generally will not) be the same-we cannot tell. (It is possible for two variables, X and *Y*, each to have a correlation of .7 with a third variable, *Z*, but to be completely uncorrelated with each other.)

The comparison problem is made worse because all the variables in component or factor analysis are standardised. Thus the heights in one study may have a mean of 45 inches and a standard deviation of 3 inches. The standardised height variable *H* in the resulting factor analysis is treated as

$$H = (\text{height in inches} - 45)/3.$$

In the second study the heights may have a mean of 60 inches and a standard deviation of 2 inches. The standardised height variable $\hat{H}$ in that component analysis is then

$$\hat{H} = (\text{height in inches} - 60)/2.$$

The two standardised height variables, Hand $\hat{H}$, are therefore not the same. A child 5 1 inches tall will have a score of $(51-45)/3=2$ for H and of $(51-60)/2=-4\frac{1}{2}$ for $\hat{H}$. If component P has a correlation of .8 with H, and component $\hat{D}$ has the same correlation of .8 with $\hat{H}$, it certainly is no evidence that D and P are the same (i.e. that any particular individual would have the same score on both D and P).

In principle, when interpreting different component analyses it might be possible to make allowances for the different means and standard deviations of the test variables in each set of data. But in practice these values are usually not even reported. Furthermore, a component is generally a linear function of *all* the test variables, but sometimes one or more test variables are changed from one study to another. It is then generally impossible for any component in the first study to be the same as a component in the next study.

The literature of factor and component analysis does not usually claim either that there are any procedures for comparing results from different studies, or that the quantitative results of different analyses are the same.

Exercise 15L. Useful for an Initial Look?

It is frequently suggested that factor analysis is useful for sorting out data where there is no prior knowledge of the relationships between the variables.

Discussion.

If there is no prior knowledge, there can be no way of judging whether one type of factor solution is more "meaningful" than another. If such an interpretive judgment of the results is nonetheless made, e.g. that it is "reasonable" for a factor to be heavily loaded in both chest circumference and girth, then the implied knowledge could have been used to analyse the given data in the first place. ("The correlation between C and G in Table 15.4 is high, now that we look!")

Exercise 15M. The Analysis of Structured Data

How can the techniques of multivariate analysis discussed in this chapter be applied to the kinds of multivariate data discussed in Chapters 9 and 10?

Discussion.

Techniques like multiple regression and factor analysis apply to a single *unstructured* set of data. There is no way of using the techniques on two or more sets of (structured) data. In Chapters 9 and 10 we had readings on the different variables for different *groups* of items, e.g.

(i) body measurements for children of different ages, races, etc.,

(ii) size measurements for trees from different rootstocks,

(iii) attitudinal responses to different brands, among users and non-users of each, etc.

The only way to apply the present techniques to such data is to *pool* the different sets ofreadings, thus losing one's prior knowledge of the differences between them.

If the experimenter or analyst has used his prior understanding of the phenomena to obtain data in a structured form, he does not need the techniques of multivariate statistical analysis discussed in this chapter. The procedures in Part II will suffice.

Exercise 15N. The Naming of Factors or Components

What is the point of a form of analysis which merely creates new variables?

Discussion.

It would be unfair to dismiss component or factor analysis outright just because it merely creates new variables and calls them names. For example, it *could* be valuable to establish that human intelligence is made up of "general ability", "verbal ability", "spatial ability", "arithmetical ability", and other such factors. Again, an art critic might distinguish paintings according to the extent to which they have "a strong line", "a sense of composition", "a sensitive use of colour", together with more specific factors like "a Rembrandesque handling of light", "a Rubenesque feel for flesh tones", etc.

Such concepts are of practical use if they achieve a certain consensus. In particular, in scientific work one does not accept a single analyst's results or opinions. Instead, detailed numerical agreement of the results of different investigators' empirical results is needed, and it is this which factor and component analysis have failed to supply.