Because there is a tie at the point that previously gave the maximum difference, D^* is in fact now reduced; it is now given by the difference 42−16=26, and this value is no longer significant at the α_2=10% level. You might like to check that the above rule concerning the graphical method yields the same result.

The alternative method (which can be tricky to apply if there are a lot of ties) is to go back to the basic principle of ordering ties in all possible ways and averaging the values of the statistic thus obtained. Ties that occur other than in regions where the difference is relatively large may be ignored for this purpose. In the above case, we may therefore ignore the first tie (at 111 h). Ordering the second tie as AB leaves D^* at 26 as above, but ordering it as BA gives D^*=32, as we obtained when the data were recorded more accurately. Here we now average these two possible values for D^*, giving D^*=29. This value is only just outside the α_2=10% critical region of $D^* \geqslant 30$.

Null distributions

As usual, null distributions may be found by considering all possible letter sequences. For illustration, let us use the same sampling situation as in the Mann–Whitney test, i.e. n_A=3 and n_B=6. First, consider one of the one-sided statistics, say $D_{AB}=\max(F_A-F_B)$. This statistic can only take on the seven values 0, 1/6, 2/6, 3/6, 4/6, 5/6 and 1 with respective frequencies of 12, 18, 25, 15, 10, 3 and 1. Since the total frequency is $\binom{9}{3}$=84, this gives

$$\text{Prob}(D_{AB}=1) = 1/84 = 0.0119$$

$$\text{Prob}(D_{AB}\geqslant 5/6) = 4/84 = 0.04762$$

$$\text{Prob}(D_{AB}\geqslant 4/6) = 14/84 = 0.1667$$

You may confirm that these results agree with the critical values for these sample sizes given in Table K, remembering that the values tabulated relate to $D^*=n_An_BD=18D$ and that α_1 significance levels are appropriate.

You might like to check through the 84 letter sequences again, but this time calculating $D_{BA}=\max(F_B-F_A)$. You should obtain precisely the same distribution as before. (Can you see why?)

Finally you can obtain the distribution of the two-sided D, which may be defined as $\max(D_{AB}, D_{BA})$. You will find that this takes the values 1/6, 2/6, 3/6, 4/6, 5/6 and 1 with frequencies 1, 26, 29, 20, 6 and 2 respectively.

If you have worked through all the null distribution exercises, you will notice some differences here compared with previous distributions, such as those for the sign, Wilcoxon signed-rank and Mann–Whitney statistics. In particular, the null distribution of the one-sided Kolmogorov–Smirnov statistic is *not* symmetric. This has two important ramifications. First, the relationship between the one-sided and two-sided statistics is no longer straightforward. Previously it has always been the case that the value of a two-sided statistic has always been immediately derivable from the value of either of the one-sided statistics; this is no longer true here. So whereas previously the two-sided statistic has taken only half the values of the related one-sided statistic, but accordingly with twice the probability, this relationship also does not now hold. Fortunately, however, this probability relationship *does* hold where it matters, i.e. in the right-hand tail, and usually to such an extent that it *does* remain true that $\alpha_2=2\alpha_1$ for significance levels of

interest. So, writing D_1 to indicate either of the one-sided statistics, D_{AB} or D_{BA}, we have

$$\text{Prob}(D=1) = 2\times\text{Prob}(D_1=1) = 2/84$$

and

$$\text{Prob}(D\geq 5/6) = 2\times\text{Prob}(D_1\geq 5/6) = 8/84$$

and even

$$\text{Prob}(D\geq 4/6) = 2\times\text{Prob}(D_1\geq 4/6) = 28/84$$

However, $\text{Prob}(D\geq 3/6) = 57/84$, whereas $2\times\text{Prob}(D_1\geq 3/6) = 58/84$.

The second effect of the non-symmetry affects the asymptotic distribution, which we discuss in the next section.

Large sample sizes

It can be shown that the non-symmetry of the null distribution does not diminish as sample sizes increase (as might have been hoped!), so that certainly the asymptotic distribution cannot be normal. In fact it was shown by L. A. Goodman (1954) that the distribution of $4n_An_BD_1^2/N$ is approximately chi-squared with two degrees of freedom, where again D_1 is either of D_{AB} or D_{BA}. As an example, we find from Table B that the $\alpha_1=5\%$ critical value of the chi-squared distribution with two degrees of freedom is 5.991. Hence the approximate $\alpha_1=5\%$ critical value of D_1 is

$$\tfrac{1}{2}\sqrt{5.991N/n_An_B} = 1.2238\sqrt{N/n_An_B}$$

In the table below (which we presented in Chapter 4 for the Kolmogorov–Smirnov goodness-of-fit test), we give the multipliers of $\sqrt{N/n_An_B}$ required for the usual levels of significance:

α_1	5%	2½%	1%	½%
α_2	10%	5%	2%	1%
	1.2238	1.3581	1.5174	1.6276

As usual, let us illustrate this approximation with $n_A=n_B=25$ (the largest sample sizes in Table K). The approximation gives the $\alpha_1=5\%$ critical value as $1.2238\sqrt{50/(25\times 25)}=0.3461$. Remembering that Table K refers to $D^*=n_An_BD$, we multiply this value by 625 to obtain 216 (to the nearest integer). Now the entry in Table K is 225, which at first sight looks a long way from 216. But recall that D_1 (or D) must be a multiple of 1/25 (since $n_A=n_B=25$) and so 25^2D_1 must be a multiple of 25. Therefore the next larger critical value to the approximate one we calculated must be 225, which is the same as the value given in Table K. In other words, the critical region $D^*\geq 216$ is actually the same as $D^*\geq 225$.

Assignments

1 Soil analysis

Analyse the soil data given in Assignment 1 of Chapter 6.

2 Football results

J. R. Gillett and D. J. Colwell (*Mathematical Gazette*, 1983) investigated data provided by 100 football matches between England and Scotland. The data are:

Number of goals per match	0	1	2	3	4	5	6	7	8	9
Frequency for England	17	39	22	6	7	7	0	1	0	1
Frequency for Scotland	20	32	27	13	3	3	1	1	0	0

Task: Use the Kolmogorov–Smirnov test to see whether there is any significant difference between the results for the two teams.

3 Shopping habits

A nationwide supermarket store commissioned a survey on the shopping habits of its customers. Part of this survey investigated the rate at which customers arrived at the stores. The data below give the times between consecutive arrivals at two stores commening at 0900 h (in fact, these results are from a particular Saturday morning and are just part of total data collected on that day):

Store A

0 min 3 s	0 min 19 s	1 min 53 s	0 min 28 s	0 min 9 s
0 min 20 s	2 min 24 s	0 min 34 s	1 min 46 s	0 min 15 s
0 min 37 s	0 min 16 s	0 min 19 s	1 min 3 s	1 min 30 s
1 min 10 s	0 min 12 s	0 min 12 s	0 min 55 s	0 min 32 s

Store B

0 min 55 s	3 min 47 s	0 min 26 s	1 min 20 s	1 min 35 s
1 min 5 s	1 min 0 s	0 min 2 s	0 min 24 s	2 min 10 s
0 min 17 s	1 min 10 s	1 min 28 s	1 min 40 s	2 min 24 s
0 min 22 s	2 min 46 s			

Task: Construct a dot diagram and/or a histogram for each set of data. Hence choose an appropriate test to see whether there is evidence to suggest that one store is busier than the other.

Program 7.1 The Kolmogorov–Smirnov two-sample test for general differences

```
100 REM *** KOLMOGOROV-SMIRNOV TWO-SAMPLE TEST ***

110 REM ** INPUT THE SAMPLES SIZES **
120 INPUT "SIZE OF THE FIRST SAMPLE: ";NA
130 INPUT "SIZE OF THE SECOND SAMPLE: ";NB
140 LET N=NA+NB
150 DIM D(N),B(NB),FA(N),FB(N),CA(NA),CB(NB),F(N)

160 REM ** INPUT THE DATA **
170 PRINT "INPUT THE OBSERVATIONS (IN ASCENDING ORDER) & FREQUENCIES FOR
    THE FIRST SAMPLE"
180 LET CF=0
190 FOR I=1 TO NA
200 INPUT D(I),FQ
210 LET F(I)=1
220 LET CF=CF+FQ
230 LET CA(I)=CF
240 NEXT I
250 PRINT
260 PRINT "INPUT THE OBSERVATIONS (IN ASCENDING ORDER) & FRQUENCIES FOR
    THE SECOND SAMPLE"
270 LET CF=0
280 FOR I=1 TO NB
290 INPUT B(I),FQ
300 LET F(I+NA)=2
310 LET CF=CF+FQ
320 LET CB(I)=CF
330 NEXT I

340 REM ** MERGE SAMPLES AND SORT **
350 FOR I=1 TO NB
360 LET D(NA+I)=B(I)
370 NEXT I
380 GO SUB 1000

390 REM ** CALCULATION OF TEST STATISTIC **
400 LET Y=0
410 LET Z=0
420 LET CA(0)=Y
430 LET CB(0)=Z
440 FOR I=1 TO N
450 IF F(I)<>1 THEN GO TO 500
460 LET Y=Y+1
470 LET FA(I)=CA(Y)*CB(NB)
480 LET FB(I)=CB(Z)*CA(NA)
490 GO TO 530
500 LET Z=Z+1
510 LET FA(I)=CA(Y)*CB(NB)
520 LET FB(I)=CB(Z)*CA(NA)
530 NEXT I
540 LET ND=0
550 LET PD=0
560 LET DS=0
570 FOR I=1 TO N
580 LET DF=FA(I)-FB(I)
590 IF DF<ND THEN LET ND=DF
600 IF DF>PD THEN LET PD=DF
610 IF ABS(DF)>DS THEN LET DS=ABS(DF)
```

```
620 NEXT I

630 REM ** OUTPUT FOR THE TWO-SIDED ALTERNATIVE **
640 PRINT "VALUE OF TEST STATISTIC FOR TWO-SIDED ALT. = ";DS
650 PRINT

660 REM ** OUTPUT FOR ONE-SIDED ALTERNATIVE **
670 PRINT "VALUE OF TEST STATISTIC FOR A>B IS ";PD," AND FOR A<B IS
";ABS(ND)

1000 REM ** SORT DATA  - USE LINES 1000-1190 OF PROGRAM 4.2 **
```

8
Two related samples

Introduction

We come now to a rather paradoxical situation. We shall have two samples to analyse, but it will turn out that ordinary two-sample methods are invalid and we shall need to have recourse to *one-sample* methods instead. This, of course, has the distinct advantage that, presuming you have already read Chapter 3, you will find virtually nothing new to learn in this chapter!

The particular two-sample situation now under discussion is where the data are in the form of so-called 'matched pairs'. So far in the book, we have always been assuming that all the observations available for analysis are 'statistically independent' of each other, i.e. there is no relationship or connection between any of them apart from the fact that items in any sample originate from the same source or population. We are now going to consider the situation where there is some kind of pairing, natural or otherwise, between observations in the two samples. For example, the two samples might consist of readings of blood pressure, or weight, or reaction time, etc., on the same group of patients before and after some medical treatment. In this situation the data might be:

	Patient					
Response	A	B	C	D	E	F
Before treatment	86	83	78	90	92	77
After treatment	82	81	92	79	92	82

Or on the other hand, an educational researcher might carry out an experiment on pairs of twins (or at least siblings, if it proves too difficult to find enough twins!), teaching some material to one member of each pair by traditional methods and to the other member of each pair by microcomputer. You will see that the resulting form of the data is essentially the same as above – there is a definite pairing between observations in the two samples. Thus the two samples are not truly *random* samples (which implies that *all* the observations are statistically independent of each other) because of such pairwise dependence between them. Consequently, ordinary two-sample tests, such as the Mann–

Whitney and Tukey tests, are inappropriate for analysing such data. Indeed, not only are they invalid for use with matched-pairs data but, even if they were mistakenly applied in this context, they would have little power.

WILCOXON'S SIGNED-RANK TEST FOR MATCHED PAIRS

Introduction

The only valid way to deal with data that are in the form of matched pairs is to regard each matched pair as a single 'sampling unit', and to represent it by an appropriate summary of its original pair of values. Then the data will indeed consist of just *one* sample. It will follow that both the tests in Chapter 3 are immediately applicable. In particular, under the null hypothesis H_0 that the two population distributions are identical, it turns out that this 'summary' value automatically has a *symmetric* null distribution, so that Wilcoxon's signed-rank test may always be used for this kind of application. Because of this, some books introduce the signed-rank test in this context rather than in the single-sample situation. We shall therefore concentrate on Wilcoxon's test here, though bearing in mind that the sign test is also available as a quick-and-easy alternative method.

PRACTICAL PROBLEM

A regional water authority has been carrying out some new pollution-control measures on one of the main rivers under its control. One way of measuring pollution levels is by 'biochemical oxygen demand' (BOD). BOD was measured at 12 sites before the new controls were implemented, and then again four years later at the same sites. Some parts of the river tend to be more heavily polluted than others because of the positions of waste discharge stations and of confluent tributaries of varying pollution.

HYPOTHESES

The water authority is seeking definite evidence that its new controls have been effective, i.e. that average pollution levels have fallen. This therefore defines H_1, with H_0 being that there has been no average change in pollution levels. So we have:

H_0: There is no difference in the average pollution levels before and after implementation of the controls.
H_1: Average pollution levels are greater before implementation of controls than afterwards.

Exhibit 1006
Page 182 of 453

DATA

The BOD values at the 12 sites before and subsequent to implementation of the pollution controls were as follows:

Site	1	2	3	4	5	6	7	8	9	10	11	12
Before controls	17.4	15.7	12.9	9.8	13.4	18.7	13.9	11.0	5.4	10.4	16.4	5.6
Four years later	13.6	10.1	10.3	9.2	11.1	20.4	10.4	11.4	4.9	8.9	11.2	4.8

STATISTIC

Because of the differing environments of the various sites, the BOD readings from different sites are not directly comparable with each other. However, we may certainly compare the two readings at individual sites to see whether there is evidence of improvement after implementation of the controls. So we calculate the *differences* between each pair of readings by subtracting the later reading from the earlier one. Under H_0, the true average (i.e. median) of these differences will be zero, but if H_1 is true then the true average difference will be *positive*. Treating the 12 differences as a single sample, therefore, we may test H_0 against H_1 using the Wilcoxon signed-rank statistic as in Chapter 3.

You will recall that, since H_1 is one-sided, we define the statistic T as *either* the sum of ranks of positive differences *or* the sum of ranks of negative differences, according as which choice would be *expected* to be the *smaller* if H_1 were true. The appropriate choice in this problem is the rank sum of the negative differences.

Note that, in the two-sided case, we would simply choose T to be whichever of the two sums turned out to be the smaller, as opposed to the one-sided case in which the choice is governed by the alternative hypothesis and not by the actual data.

CRITICAL REGIONS

Critical values are obtained from Table D (or page 29 of *EST*) as before, with α_1 significance levels referring to one-sided tests as usual.

In this example with $n=12$, the $\alpha_1=5\%$, 1% and $\frac{1}{2}\%$ critical regions are $T\leqslant 17$, $T\leqslant 9$ and $T\leqslant 7$ respectively.

SOLUTION

The differences (before minus after) in BOD readings at the 12 sites, and their ranks (ignoring signs), are:

Sites	1	2	3	4	5	6	7	8	9	10	11	12
Differences	3.8	5.6	2.6	0.6	2.3	−1.7	3.5	−0.4	0.5	1.5	5.2	0.8
Ranks	10	12	8	3	7	6	9	1	2	5	11	4
Signs of differences	+	+	+	+	+	−	+	−	+	+	+	+

The rank sum of the negative differences is $T=6+1=7$. This value is in fact the boundary of the critical region for $\alpha_1=\frac{1}{2}\%$, and so we have found very strong evidence of a reduction in average pollution levels along the river.

Discussion

Although we have stated that ordinary two-sample tests such as the Mann–Whitney and Tukey tests are invalid when the data are in the form of matched pairs, you might nevertheless find it instructive to try them out on the BOD data. You will discover that neither test rejects H_0 at even the $\alpha_1=5\%$ level. Compare the details, and in particular think carefully about *why* the Mann–Whitney test fails to reject H_0 whereas the Wilcoxon signed-rank test has found almost conclusive evidence of an improvement.

The reason for the invalidity of ordinary two-sample tests is mentioned again in the 'Null distributions' section below. It is shown in the same section why the Wilcoxon test may *always* be used with matched-pairs data, whereas it has limited applicability to one-sample situations.

One-sided and two-sided tests

See the 'Statistic' section above.

Treatment of zero or tied differences

As in Chapter 3.

Null distributions

The reason for the invalidity of ordinary two-sample tests in the present context is simply that the statistical dependence between the members of each matched pair results in the various possible letter sequences not now being equally likely even when H_0 is true. This is because of the tendency for the A and B corresponding to the members of any matched pair to be relatively close together in the letter sequence.

More importantly, let us examine why the Wilcoxon test *is* always valid when analysing matched pairs, unlike when the test is applied to a single sample. As with ordinary two-sample situations, we will take the null hypothesis H_0 to imply that the two components (X_A and Y_B, say) of any matched pair have the same distribution as each other. (It is quite conceivable, however, that each matched pair may have a different such distribution.) We will first prove that, for any given matched pair, the distribution of the difference X_A-Y_B is symmetric about zero if H_0 is true. Such symmetry could be expressed, for example, by requiring $\text{Prob}(X_A-Y_B>d)$ to be equal to $\text{Prob}(X_A-Y_B<-d)$ for all d (see Fig. 8.1). Now,

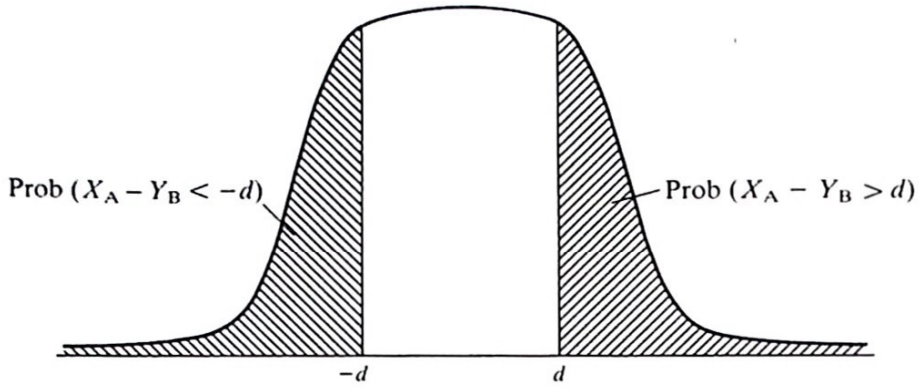


Figure 8.1

$\text{Prob}(X_A - Y_B > d) = \text{Prob}[-(X_A - Y_B) < -d] = \text{Prob}(Y_B - Y_A < -d)$. The crux of the argument is that, since X_A and Y_B have identical distributions, $X_A - Y_B$ and $Y_B - X_A$ must have the same distribution as each other. Consequently, $\text{Prob}(Y_B - X_A < -d) = \text{Prob}(X_A - Y_B < -d)$, and the symmetry condition is thus proved.

Now although the distributions of differences need not be the same for all matched pairs, it follows that they must always be symmetric. Consequently, the basic criterion required for the computation of the null distribution in Chapter 3 is satisfied, i.e. that the probability behaviour of positive and negative differences are identical. Thus (ignoring zero differences) there is a 50–50 chance of the largest difference being positive or negative; and there is a 50–50 chance of the second largest difference being positive or negative; and so on for each and every difference. The null distribution of T is therefore computable as in Chapter 3, and so the tables of critical values (Table D and *EST* page 29) are valid quite generally for matched-pairs data.

Large sample sizes

Exactly as in Chapter 3 for the Wilcoxon signed-rank test.

THE SIGN TEST FOR MATCHED PAIRS

The sign test can of course be applied to the single sample of differences instead of Wilcoxon's test. This provides us with a quicker and simpler test, but one which is not generally as powerful as the signed-rank test for the same reasons as discussed in chapter 3. In the BOD data there are two negative differences out of 12 and, checking with the table of critical values for the sign test (Table C), we see that this is significant at the $\alpha_1 = 2\frac{1}{2}\%$ level but not at the $\alpha_1 = 1\%$ level. The sign test does therefore also find evidence of an improvement in average pollution levels, but the evidence is not as strong as that found by using Wilcoxon's signed-rank test.

Assignments

1 Calf feeding

An experiment was carried out to compare two schemes, A and B, of calf feeding. Scheme A was a new and untried method that was expected to produce better results (in terms of gain in weight) than scheme B, which was a more traditional method of calf feeding. Thirteen pairs of identical twin calves were used in the experiment. One of each pair was fed according to scheme A and the other according to scheme B. The gains in weight (kg) of the calves are given in the following table.

Pair	1	2	3	4	5	6	7	8
Scheme A	20.1	19.5	19.0	21.1	23.1	22.6	18.9	22.8
Scheme B	21.2	18.7	19.0	20.8	19.9	21.4	17.9	23.1

Continued

Pair	9	10	11	12	13
Scheme A	27.1	19.8	21.7	18.9	20.4
Scheme B	24.3	18.5	20.3	18.7	19.4

Task: Use a matched-pairs test to see whether these data support the claim that scheme B is better than A.

2 Distribution of soil particles

In 1973, D. R. Nielsen *et al.* ('Spatial variability of field-measured soilwater properties' *Hilgardia* **42**, 215–59) published the results of a survey on the distribution of particle size in soil. Part of their data is given below, showing percentage of sand and silt at 12 depths:

Depth	Sand	Silt	Depth	Sand	Silt
1	27.3	25.3	7	25.2	28.4
2	31.2	25.4	8	23.6	31.4
3	25.0	26.6	9	33.4	30.3
4	30.1	27.4	10	36.4	28.3
5	33.6	25.1	11	36.4	29.4
6	34.9	24.2	12	20.7	36.7

Task: Use an appropriate procedure to test the null hypothesis that there is no difference in the average percentages of sand and silt.

Program 8.1 Wilcoxon's matched pairs test

```
100 REM *** WILCOXON SIGNED-RANK MATCHED-PAIRS TEST ***

110 REM ** INPUT DATA **
120 INPUT "SAMPLE SIZE  ";N
130 DIM A(N),B(N),D(N),RK(N)
140 PRINT "INPUT THE OBSERVATIONS FOR SAMPLE A & B:"
150 FOR I=1 TO N
160 INPUT A(I),B(I)
170 NEXT I

180 REM ** DIFFERENCES CALCULATED **
190 FOR I=1 TO N
200 LET A(I)=A(I)-B(I)
210 NEXT I

220 REM ** REMOVE ZERO DIFFERENCES **
230 LET NC=0
240 FOR I=1 TO N
250 IF A(I)=0 THEN GO TO 280
260 LET NC=NC+1
270 LET A(NC)=A(I)
280 NEXT I
290 LET N=NC
300 FOR I=1 TO N
310 LET D(I)=ABS(A(I))
320 NEXT I
330 GO SUB 1000

340 REM ** CALCULATION OF RANKS SUMS **
350 LET SN=0
360 LET SP=0
370 FOR I=1 TO N
380 IF A(I)<0 THEN LET SN=SN+RK(I)
390 IF A(I)>0 THEN LET SP=SP+RK(I)
400 NEXT I
410 PRINT
420 PRINT "SUM OF THE POSITIVE RANKS =";SP
430 PRINT "SUM OF THE NEGATIVE RANKS = ";SN
440 PRINT
450 IF N<=50 THEN STOP

460 REM ** CALCULATION OF APPROXIMATE CRITICAL VALUES (N > 50) **
470 INPUT "PERCENTAGE POINT OF STANDARD NORMAL DISTRIBUTION ";Z
480 LET SD=SQRT(N*(N+1)*(2*N+1)/24)
490 LET MN=N*(N+1)/4
500 PRINT "CRITICAL VALUE = ";INT(MN-Z*SD-.5)

1000 REM ** RANK SAMPLE - USE LINES 1000-1140 OF PROGRAM 3.1 **
```

PART IV

Correlation and regression

Chapter 8 dealt with the situation where data occur in the form of 'matched pairs'. This is where we have two samples of equal size but, instead of all the observations being statistically independent of each other, there is some kind of link or connection between corresponding members of the samples. Thus ordinary two-sample methods, which assume the samples are truly *random*, do not apply.

The situation of matched pairs is not far removed from the problems to be studied in this part of the book. In the next three chapters, we cover distribution-free approaches to correlation and regression, topics for which again there exists a considerable non-distribution-free literature. In brief, when we have some data from an ostensibly matched-pairs situation, there are two things to investigate. First, *is* there a matching at all, i.e. is there a pairwise dependence between the samples? Secondly, if there is such a dependence, how can we use it to predict or forecast the value of one member of a matched pair if we know the value of the other one? These two problems are respectively the problems of *correlation* and *regression*.

Although both correlation and regression techniques can be extended to three or more variables, in this book we discuss only the problems arising with two variables. In Chapters 9 and 10, we present some standard distribution-free procedures for dealing respectively with the problems of correlation and regression, while in Chapter 11 we discuss some quick methods for these problems. Chapter 12 deals with correlation between non-numeric variables or, as it is commonly known, the analysis of contingency tables.

9
Correlation

Introduction

The classical approach to correlation is very 'model-dependent'. That is, if a particular kind of formula is proposed to represent the relationship between two components of the matched pairs, then the methods examine the likely validity or otherwise of this type of formula. The methods of *rank correlation* described here are rather less restrictive. They examine simply whether there is a tendency for the two components of the matched pairs to increase and decrease together (*positive* correlation) or, alternatively, for one to decrease as the other increases and vice versa (*negative* correlation); either kind of effect is known as *monotonicity*. The tests to be introduced can be either two-sided (testing for either kind of correlation) or one-sided (testing for *either* positive *or* negative correlation as specified in the alternative hypothesis).

We cannot really refer to the distribution-free and non-distribution-free approaches to correlation as 'modern' and 'traditional' respectively, since both techniques covered in this chapter have a very long history. Spearman's version of rank correlation was published long ago (Spearman, 1904); and the other method, usually attributed to a paper by Kendall (1938), can in fact also be traced back to the turn of the century.

SPEARMAN'S RANK CORRELATION COEFFICIENT

Introduction

Several ways of measuring correlation have been devised. The resulting 'correlation coefficients' are conventionally defined so that they must lie between -1 and $+1$; $+1$ represents extreme evidence of positive correlation (perfect agreement), -1 represents extreme evidence of negative correlation (perfect disagreement), while values near 0 tend to occur if there is little or no such correlation between the two variables.

Spearman's correlation coefficient r_S is of this type. Denoting the data by (X_1,Y_1), (X_2,Y_2), . . . , (X_n,Y_n), r_S will equal $+1$ if and only if it is *strictly* true that 'the larger X is, then the larger Y is' throughout the sample. Conversely, if it is strictly true that 'the larger X is, the *smaller* Y is', then r_S will equal -1. It is unusual to obtain data where such

relationships hold exactly (if so, the reality of the situation is often so obvious that a formal test is unnecessary!). But, as you would expect, a tendency towards either of these extremes will produce relatively large positive or negative values of r_S (i.e. close to +1 or −1), and otherwise r_S will be closer to 0. Such behaviour thus provides an obvious basis for a test for correlation.

PRACTICAL PROBLEM

A business organization specializes in selling (or renting) copying machines. The managing director is analysing a full year's records of the 15 salesmen employed by a certain branch of the company. The records include details of the number of machines placed by each salesman, and the number of calls or visits made by each salesman to various potential customers during the year. The managing director is an advocate of high activity on the part of the salesmen, i.e. he is keen on them increasing their number of visits! A brave member of the sales team suggests that it is not really the number of visits that matters but rather the *type* of visit. Effectively, there are two types of visit: simple visits and demonstrations ('demos'). In a simple visit, the salesman is armed with brochures, advertising and other literature; whereas, as the name implies, a demo consists of an actual on-site demonstration of one of the copying machines. To an extent, the salesmen are given freedom of choice as regards the number and type of visits that they make. However, the managing director favours simple visits, as they are more flexible and cost less.

The salesmen's records include not only the total numbers of visits but also the numbers of demos made. The data are given below. We want to know: what conclusions should the managing director draw?

HYPOTHESES

Assuming that it is unlikely for sales to *fall* if either the total number of simple visits or the number of demos increase, we could test for positive correlation between *either* of these numbers and the number of sales, i.e. the tests will be one-sided.

In the illustration of Spearman's test we shall consider whether there is evidence of positive correlation between the number of simple visits and the number of sales (the second problem of correlation between the number of demos and the number of sales we leave as an exercise). So the hypotheses of concern to us are:

H_0: There is no correlation between the number of simple visits and the number of sales.

H_1: There is some positive correlation between the number of simple visits and the number of sales.

DATA

The relevant extract from the records is as follows:

Salesman	Demos	Simple visits	Total visits	Sales
1	104	265	369	25
2	161	380	541	47
3	156	270	426	40
4	96	224	320	17
5	149	343	492	49
6	143	359	502	39
7	113	216	329	33
8	142	324	466	30
9	115	381	496	31
10	175	242	417	44
11	135	233	368	34
12	145	350	495	43
13	137	276	413	35
14	151	391	542	42
15	126	290	416	36

Before proceeding with a formal analysis of any such data, it is usually fruitful to depict the values graphically on what is commonly called a *scatterplot*. The scatterplots for our data are shown in Figures 9.1 and 9.2. Such plots often give us a 'feel' for whether there is any correlation between the variables – and, if there is, what type it is (e.g. linear, curved, etc.). Figure 9.1 appears to indicate little, if any, correlation between the number of simple visits and sales. But Figure 9.2 reveals the

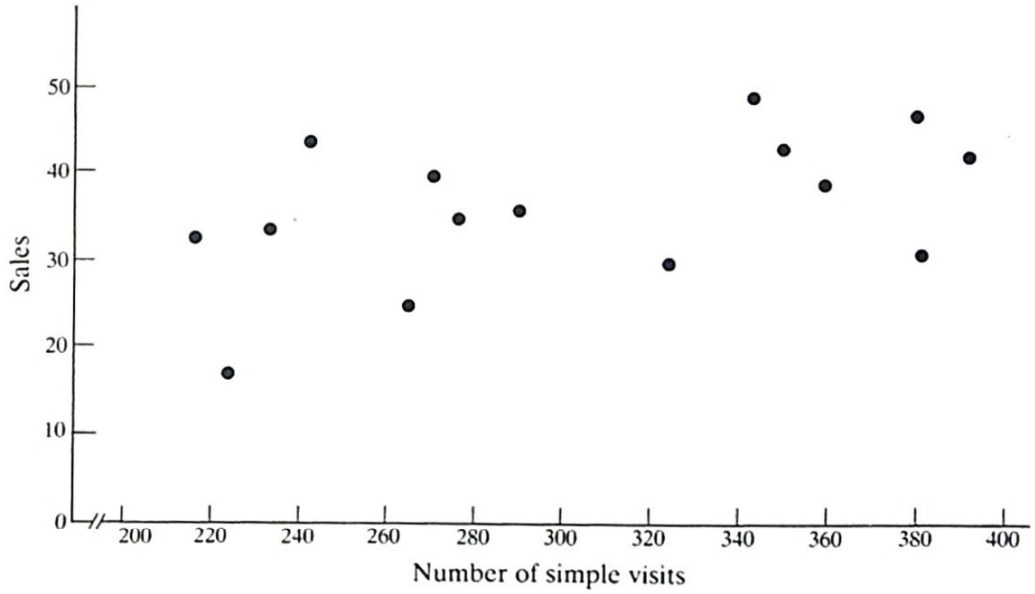


Figure 9.1 Scatterplot of simple visits vs. sales.

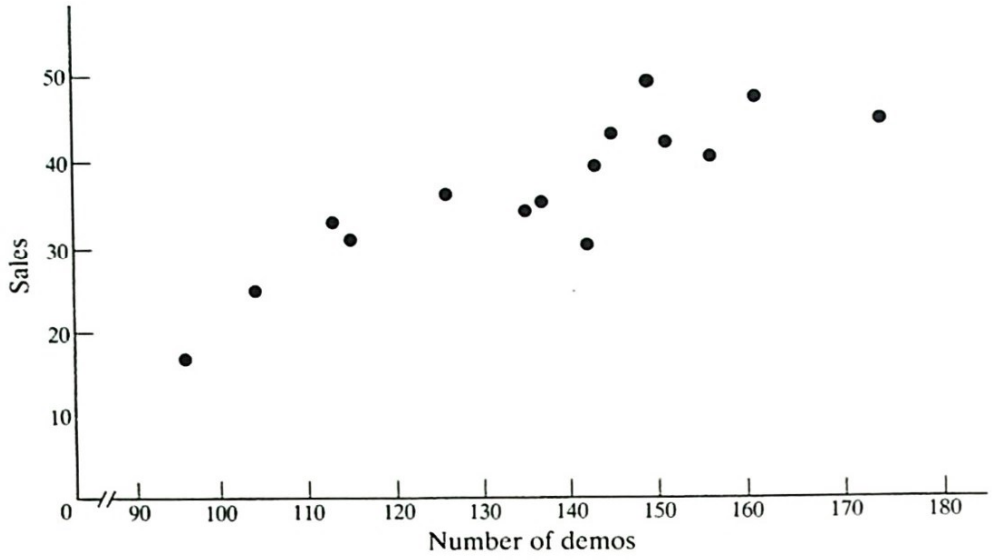


Figure 9.2 Scatterplot of demos vs. sales.

presence of quite strong correlation between the number of demos and sales. It will be interesting to see if the test statistics reflect these indications from the scatterplots.

STATISTIC

Let X_i and Y_i denote the two variables whose correlation coefficient we wish to calculate; so here X_i represents the number of simple visits and Y_i the number of sales by the ith salesman. We first rank both the X_i and Y_i from 1 (for the lowest value) to n (for the highest), where n is the number of pairs of observations; $n=15$ in this example. We shall denote these ranks by $r(X_i)$ and $r(Y_i)$ respectively. The *difference* between the ranks for each pair is then calculated as follows:

Salesman,	i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Visits rank,	$r(X_i)$	5	13	6	2	10	12	1	9	14	4	3	11	7	15	8
Sales rank,	$r(Y_i)$	2	14	10	1	15	9	5	3	4	13	6	12	7	11	8
Difference,	d_i	+3	−1	−4	+1	−5	+3	−4	+6	+10	−9	−3	−1	0	+4	0

Now if the two variables are positively correlated, then the rankings will be rather similar and the rank differences $d_i=r(X_i)-r(Y_i)$ will tend to be small in size. Indeed the strongest indication of positive correlation is when the two sets of ranks are identical, in which case all the rank differences will be zero. Perhaps your first thoughts about trying to summarize these differences would be to average them; but, as you can

easily check, the sum of the differences (presuming their signs are taken into account) is always equal to zero, so their mean is not very informative! However, as seen in Chapter 1 when we introduced the sample variance, a common method in statistics of combining a set of numbers that sum to zero is to form the *sum of their squares*:

$$d_1^2 + d_2^2 + \ldots + d_n^2 = \sum_{i=1}^{n} d_i^2 = D^2 \qquad \text{(say)}$$

This is obviously 0 when the two sets of ranks are identical. D^2 takes its maximum possible value when the ranks are completely in the 'wrong' order (i.e. the observation whose X rank is 1 has Y rank n, the one whose X rank is 2 has Y rank $n-1$, and so on); it can be shown that this maximum value of D^2 is $(n^3-n)/3$.

Exercise Confirm for $n=15$ that this formula does indeed give the value of D^2, when the ranks are completely opposite to each other, and then confirm this formula for some even value of n of your own choice.

D^2 could be used directly as a statistic for the current problem. However, bearing in mind what was said in the introduction about correlaton coefficients usually being constructed to have values between -1 and $+1$, Spearman's rank correlation coefficient is in fact defined by:

$$r_S = 1 - \frac{6D^2}{n^3-n}$$

Obviously, when $D^2 = 0$ (the strongest evidence of positive correlation) then $r_S = +1$; and when $D^2 = (n^3-n)/3$ (its largest value and the strongest evidence of negative correlation) then $r_S = -1$; otherwise r_S lies between -1 and $+1$.

CRITICAL REGIONS

Since the one-sided alternative of positive correlation is supported by values of r_S near $+1$, the required critical regions are of the form

$$r_S \geqslant \text{critical value}$$

Table L (or the first table on page 40 in *EST*) gives the critical values for r_S, as always, α_1 significance levels relate to one-sided tests.

In fact, in the tables the significance levels are denoted by α_1^R; the superscript R refers to the one-sided test having a critical region in the right-hand tail of the distribution of r_S. Had we been testing the opposite one-sided alternative, that of *negative* correlation, critical regions would

have consisted of large negative values, i.e. in the left-hand tail and would be of the form

$$r_S \leqslant -\text{tabulated value}$$

So, in our current problem (with $n=15$), the critical regions corresponding to significance levels $\alpha_1=5\%$, $2\frac{1}{2}\%$, 1% and $\frac{1}{2}\%$ are respectively $r_s \geqslant 0.4464$, $r_S \geqslant 0.5214$, $r_S \geqslant 0.6036$ and $r_S \geqslant 0.6536$.

SOLUTION

Using the table of rank differences constructed earlier, we obtain:

$$\begin{aligned} D^2 = \sum_{i=1}^{15} d_i^2 &= (+3)^2 + (-1)^2 + (-4)^2 + (+1)^2 + (-5)^2 + (+3)^2 \\ &\quad + (-4)^2 + (+6)^2 + (+10)^2 + (-9)^2 \\ &\quad + (-3)^2 + (-1)^2 + 0^2 + (+4)^2 + 0^2 \\ &= 320 \end{aligned}$$

Then

$$r_S = 1 - \frac{6\times 320}{15^3-15} = 1 - \frac{1920}{3360} = 0.4286$$

This is just outside the $\alpha_1=5\%$ critical region. So there is only a little evidence of positive correlation between the number of simple visits and the number of sales. This conclusion agrees with the initial view we had of the scatterplot in Figure 9.1.

Exercise Now calculate Spearman's correlation coefficient using the number of demos as the X variable and with the number of sales still as the Y variable.

Discussion

You should have found the value of r_S for the number of demos and number of sales in the exercise to be 0.8714, which is well into the $\alpha_1=\frac{1}{2}\%$ critical region; so much so that it would clearly be significant for much more stringent values of α_1 (also, this result truly confirms the visual evidence in the scatterplot in Figure 9.2). Thus the salesman's claim that the *type* of visit (i.e. demo as opposed to simple visit) affects sales is well supported by these results. The managing director will hopefully be convinced by the above results that sales are increased more predictably by increasing the number of demos than by increasing the number of simple visits. However, the point remains that demos are expensive to arrange; to increase their number would mean increasing the number of machines to be set aside for demonstration purposes, thereby severely reducing the potential income from those machines and considerably increasing transport, insurance and other costs. To examine whether the increased cost would be worth while, he needs to ascertain how the sales are likely to increase for a certain

increased number of demos. This is the kind of problem tackled by the *regression* techniques in Chapter 10. Thus correlation and regression are somewhat complementary techniques: regression methods can be carried out whatever the data look like, but correlation methods indicate whether they are likely to be of much value.

There are a number of quite tricky points of interpretation concerning correlation, and we shall review these in the 'Discussion' section at the end of the chapter.

Alternative derivation of r_S

This section will be of main interest to readers who are familiar with the standard 'parametric' approach to correlation, which uses Pearson's product-moment coefficient of linear correlation, r. This measures the tendency for the X and Y variables to be not just monotonically related (as with Spearman's coefficient) but for that monotonic relationship to be specifically *linear*. The formula for r can be written in various ways; here are two of the most important forms:

$$r = \frac{\Sigma(X_i-\bar{X})\,(Y_i-\bar{Y})}{\Sigma(X_i-\bar{X})^2\Sigma(Y_i-\bar{Y})^2} = \frac{\Sigma X_iY_i - n\bar{X}\bar{Y}}{(\Sigma X_i^2-n\bar{X}^2)\,(\Sigma Y_i^2-n\bar{Y}^2)}$$

In these formulae, Σ represents the sum for i going from 1 to n, i.e. $\sum_{i=1}^{n}$, and $\bar{X}$ and $\bar{Y}$ are the sample means, $\bar{X} = \Sigma X_i/n$ and $\bar{Y} = \Sigma Y_i/n$. Since r is a measure of *linear* correlation, it only takes the value $+1$ if the data lie exactly on some *line* of positive slope, and -1 if they lie on a *line* of negative slope. Unlike r_S, any other monotonic relationship will give a value of r that lies between, rather than at, these extremes. Nevertheless, r_S can be defined in a way that is very closely related to r. In brief, r_S can be computed from exactly the same formula as r, *but using the ranks* $r(X_i)$ *and* $r(Y_i)$ *instead of the actual values* X_i *and* Y_i.

We will first confirm this numerically for the copier salesmen's data and then give a general proof. Using the second form of the formulae for r, we are therefore claiming that

$$r_S = \frac{\Sigma r(X_i)r(Y_i) - nr(\bar{X})r(\bar{Y})}{[\Sigma(X_i)^2-nr(\bar{X})^2]\,[\Sigma r(Y_i)^2-nr(\bar{Y})^2]}$$

where $r(X_i)$ and $r(Y_i)$ are the ranks tabulated in the 'Statistic' section earlier in this chapter. Now both the $r(X_i)$ and the $r(Y_i)$ consist of permutations of 1,2,3, . . . , 14,15 $(=n)$, so immediately we know that $r(\bar{X})=r(\bar{Y})=8$; you may also check that $\Sigma r(X_i)^2 = \Sigma r(Y_i)^2 = 1^2+2^2+3^2+\ldots+14^2+15^2=1240$. All that needs working out from the actual arrangement of ranks is

$$\Sigma r(X_i)r(Y_i) = (5\times2)+(13\times14)+(6\times10)+\ldots+(15\times11)+(8\times8)=1080$$

(again refer to the table in the 'Statistic' section). Substituting all these figures into the above formula gives

$$\frac{1080-15(8)^2}{1240-15(8)^2} = \frac{120}{280} = 0.4286$$

as above. As an exercise, you might like to try using this alternative method for calculating r_S using the number of demos as the X variable instead of the number of simple visits.

For the general proof, we first quote two well known mathematical results:

$$r(\bar{X}) = r(\bar{Y}) = (1+2+3+\ldots+n)/n = \tfrac{1}{2}(n+1)$$

and

$$\Sigma r(X_i)^2 = \Sigma r(Y_i)^2 = 1^2+2^2+\ldots+n^2 = n(n+1)\ (2n+1)/6$$

Then, the denominator is

$$n(n+1)\ (2n+1)/6-n(n+1)^2/4 = (n^3-\mathrm{n})/12 \qquad (1)$$

Now consider

$$D^2=\Sigma d_i^2 = \Sigma[r(X_i)-r(Y_i)]^2 = \Sigma r(X_i)^2 + \Sigma r(Y_i)^2 - 2\Sigma r(X_i)r(Y_i)$$

Again, because

$$\Sigma r(X_i)^2 = \Sigma r(Y_i)^2 = n(n+1)\ (2n+1)/6$$

this gives

$$\Sigma r(X_i)r(Y_i) = n(n+1)\ (2n+1)/6 - D^2/2 \qquad (2)$$

Dividing (2) by (1) gives the formula quoted earlier for r_S.

We have previously stated (rather than proved) that r_S lies between -1 and $+1$, and is only equal to -1 when the X_i and Y_i are ranked completely oppositely. That $-1 \leqslant r_S \leqslant +1$ now follows immediately, for this inequality is always true for r, and r_S has been seen to be a particular calculation of r. Also, it is obvious that identical rankings give $D^2 = 0$, and hence $r_S = +1$. Moreover, a plot of the points corresponding to completely opposite rankings, $(1,n)$, $(2,n-1)$, $(3,n-2)$,. . . , $(n,1)$, gives a line of negative slope (having equation $x+y=n+1$), and therefore immediately this has to be the case where $r_S=-1$.

One-sided and two-sided tests

The null distribution of r_S is symmetric, so critical regions appropriate to tests for negative correlation are

$$r_S \leqslant -\text{tabulated value}$$

where the tabulated value is found from Table L using the α_1^R significance levels.

For two-sided tests, where H_1 includes both negative and positive correlation, critical regions are of the form

$$|r_S| \geqslant \text{critical value}$$

corresponding to α_2 significance levels.

Treatment of ties

Tied observations are given average ranks as usual. Thus, for example, if salesman 8 had made 31 sales instead of 30, he would then be tied with salesman 9, and instead of these two having ranks 3 and 4 respectively, they would now both be ranked as $3\frac{1}{2}$. Or, further, suppose that salesman 7 had also managed only 31 sales instead of 33; then salesmen 7, 8 and 9 would all be ranked at 4. You might like to check that r_S becomes 0.4205 and 0.4607 respectively in these two cases.

Null distributions

If H_0 were true then all rank orderings of the sales figures Y_i would be equally likely, irrespective of the numbers X_i of simple visits.

For illustration, take $n=5$, so that $(r(Y_1), r(Y_2), r(Y_3), r(Y_4), r(Y_5))$ takes on one of the $5!=120$ permutations (orderings) of (1,2,3,4,5). Without loss of generality, let the X values be in natural order, i.e. $r(X_1)=1$, $r(X_2)=2$, $r(X_3)=3$, $r(X_4)=4$ and $r(X_5)=5$. Then clearly the only arrangement of the Y ranks to give $D^2=0$ (and hence $r_S=1$) is also (1,2,3,4,5).

Next you may check that there are four cases yielding $D^2=2$, i.e. $r_S=0.9$; these are (2,1,3,4,5), (1,3,2,4,5), (1,2,4,3,5) and (1,2,3,5,4). There are three cases yielding $D^2=4$, i.e. $r_S=0.8$; these are (2,1,4,3,5), (2,1,3,5,4) and (1,3,2,5,4). Since small values of D^2 (i.e. large positive values of r_S) constitute one-sided critical regions, and $\text{Prob}(r_S=1.0)=1/120=0.83\%$, $\text{Prob}(r_S\geqslant 0.9)=5/120=4.17\%$ and $\text{Prob}(r_S\geqslant 0.8)=8/120=6.67\%$, we have to take $r_S=1.0$ as the best conservative critical region for both $\alpha_1=1\%$ and $2\frac{1}{2}\%$, and $r_S\geqslant 0.9$ for $\alpha_1=5\%$; there is no suitable region for $\alpha_1=\frac{1}{2}\%$. These regions are consistent with the results in Table L.

Large sample sizes

If H_0 is true and n is large, the distribution of r_S is approximately normal with mean 0 and standard deviation $1/\sqrt{(n-1)}$. So, if z is the appropriate percentage point of the standard normal distribution (see Table A) for a particular significance level, the approximate critical value for r_S is given by $z/\sqrt{(n-1)}$. Thus, for example, taking $n=50$ (the largest sample size in Table L), and recalling that $z=1.96$ if $\alpha_2=5\%$, the corresponding approximate critical region is given by $r_S\geqslant 1.96/\sqrt{49}=0.2800$, which is very close to 0.2791, the value given in Table L.

Convergence to normality for r_S is somewhat slower than with some other distribution-free statistics, but the approximation is certainly satisfactory for $n>50$, which is all we need.

KENDALL'S RANK CORRELATION COEFFICIENT

Introduction

This alternative approach to rank correlation seems to have neither any great advantages nor disadvantages compared with Spearman's, although the latter is perhaps the slightly better known and is said by some to be easier to compute for large samples. On the other hand, Kendall's

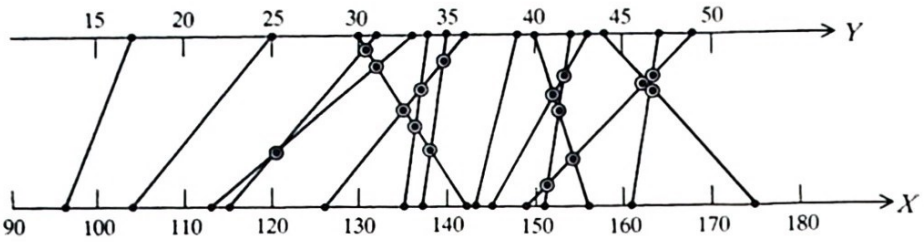


Figure 9.3

coefficient can be computed by a simple graphical, rather than arithmetical, process and also has some theoretical points in its favour.

PRACTICAL PROBLEM, HYPOTHESES AND DATA

We shall use the same situation and data as for the test based on Spearman's coefficient. However, for the detailed illustration we shall take the X variable here to be the number of demos.

STATISTIC

We first present the simple graphical method. The X_i and Y_i values are plotted on two parallel axes and then a line is drawn joining each X_i to its corresponding Y_i. This operation need not be carried out with great accuracy; a freehand sketch will do, particularly if n is fairly small. Similarly, the scales do not have to be chosen according to any particular criterion – anything convenient will do. Figure 9.3 shows our drawing for the data. Now we count the number of times that pairs of lines intersect in the drawing, i.e the points marked by a circle; you will see that there are 16. Note that all points of intersection are considered distinct from each other; if some such points are going to coincide, then this should be avoided by slightly curving the relevant lines in the diagram.

Kendall's approach to rank correlation is based on a comparison of all possible pairs of (X,Y) values with each other. Consider comparing (X_1,Y_1) with (X_2,Y_2), i.e. (104,25) with (161,47). We note that the first of these pairs has both the smaller X value and the smaller Y value; this is an example of a so-called *concordant* pair. Similarly, (X_1,Y_1) is concordant with (X_3,Y_3). When comparing (X_1,Y_1) with (X_4,Y_4), i.e. (104,25) with (96,17), we see that the first pair now has both the *larger* X value and the *larger* Y value; such a pair is also called concordant. In fact all the pairs turn out to be concordant with (X_1,Y_1), which, as we shall see, is reflected by the fact that there are no points of intersection on the line joining $Y_1=25$ to $X_1=104$ in Figure 9.3. Now, if we start comparing $(X_2,Y_2)=(161,47)$ with all the other pairs, we find it to be concordant with both (X_3,Y_3) and (X_4,Y_4). However, when comparing (X_2,Y_2) with (X_5,Y_5), i.e. (161,47) with (149,49), it is now the case that the first pair

has the lower Y value but the higher X value. When this occurs (or vice versa) the pair is said to be *discordant*. Notice in Figure 9.3 that the line joining 47 to 161 does intersect with the line joining 49 to 149. It is then easy to generalize this argument to show that each discordant pair results in a (circled) crossing point in the figure, but that concordant pairs do not.

Kendall's rank correlation coefficient is based on the difference between the total number of concordant (N_C) and discordant (N_D) pairs. N_D is given by the number of crossings in Figure 9.3, while N_C+N_D is equal to the total number of pairs compared, and it can be shown that this number is $\frac{1}{2}n(n-1)$ (check it with some small values of n if you cannot see why this should be so). Hence,

$$N_C=\tfrac{1}{2}n(n-1)-N_D$$

In order to be consistent with the general properties of correlation coefficients (i.e. that they lie between -1 and $+1$, with exceptional cases -1 and $+1$ having the same interpretation as with r_S), Kendall's coefficient, usually denoted by the Greek letter (tau) τ (pronounced taw or tow, as in towel), is defined as

$$\tau = \frac{N_C-N_D}{\frac{1}{2}n(n-1)}$$

The reasoning is as follows. If the X_i and Y_i have the same rank orde. (the strongest evidence for positive correlation), then there will be no crossings in the drawing, i.e. $N_D=0$, $N_C=\frac{1}{2}n(n-1)$ and so $\tau=+1$. If the X and Y values are ranked completely opposite to each other, then all pairs will be discordant, so that $N_D=\frac{1}{2}n(n-1)$, $N_C=0$ and $\tau=-1$.

Since $N_C=\frac{1}{2}n(n-1)-N_D$, it is easy to show that τ may also be written as

$$\tau = 1 - \frac{4N_D}{n(n-1)}$$

This is a particularly convenient form to use after having calculated N_D as the number of crossings in the drawing.

You may now see why no great accuracy is needed when drawing the diagram. The actual values in the data do not matter, just the *order* in which they occur – as with all distribution-free methods.

Exercise It follows that we shall get the same value of N_D, and hence of τ, for any data (X_i,Y_i) giving the arrangement of ranks $(r(X_i),r(Y_i))$ as in our example. In particular, the ranks themselves can be used for this purpose! Just to check,

form a table of ranks $(r(X_i),r(Y_i))$ and join corresponding points in this diagram; confirm that we do get the same answer as before.

CRITICAL REGIONS

As with Spearman's r_S statistic, large values indicate positive correlation, and so, for a one-sided alternative hypothesis of positive correlation, critical regions are of the form:

$$\tau \geqslant \text{critical value}$$

Table M gives critical values for τ, and a fuller table is available on page 40 of *EST*; the layout of these tables is precisely as for r_S.

With $n=15$, the required one-sided critical regions are $\tau \geqslant 0.3333$, $\tau \geqslant 0.3905$, $\tau \geqslant 0.4667$ and $\tau \geqslant 0.5048$ for $\alpha_1 = 5\%$, $2\frac{1}{2}\%$, 1% and $\frac{1}{2}\%$ respectively.

SOLUTION

Since we have found from the diagram in Figure 9.3 that $N_D=16$, we can calculate τ as

$$\tau = 1 - \frac{4N_D}{n(n-1)} = 1 - \frac{4\times 16}{15\times 14} = 0.6952$$

Since this value of τ lies well inside even the $\alpha_1 = \frac{1}{2}\%$ critical region, we conclude with virtual certainty that there is a positive correlation between numbers of demos and sales; this was also the conclusion arrived at using Spearman's coefficient.

Exercise Calculate τ now using the number of simple visits as the X variable. You should obtain $\tau=0.3143$, which is just outside the $\alpha_1=5\%$ critical region; so the conclusion is essentially the same as that obtained using r_S.

If the sample size is at all large, the graphical method becomes messy and the arithmetical technique shown below is then preferable; and indeed, you may wish to use this method for small values of n also.

Discussion

Both Spearman's and Kendall's approaches to rank correlation are well known and widely used. So it is natural to ask which of the two is preferable. As indicated in the introduction, there seems to be no easy choice between them. For sample sizes less than about 20, Spearman's r_s may have a slight edge over Kendall's τ (see the 'Null distributions' section below). The coefficients involve similar amounts of computation; Kendall's τ is slightly quicker to compute for small to moderate sample sizes, but Spearman's r_S is probably quicker for larger samples. Kendall's coefficient has some rather nice theoretical aspects not shared

by Spearman's. For example, (a) its null distribution converges to normality faster than that of r_S (see the 'Large samples' section below); (b) τ is an estimator of an easily identifiable characteristic of the populations, namely the difference between the probabilities of concordance and discordance; and (c) there is a quite well known version of τ that can be used in more complex (multivariate) situations (in fact, a similar extension to so-called partial correlation can also be applied to r_S, but this does not seem to be often used).

It is therefore very much a case of taste and/or tradition as to which of the coefficients is used. However, in Chapter 11 we offer an alternative approach to distribution-free correlation which is rather quicker to execute than either of those described in this chapter.

Arithmetical calculation of τ

We rearrange the data so that the values of one of the variables, say X, are in ascending order (particularly for large n, this may be the major part of the whole operation). Then take each Y value in turn, and count how many of the 'Y's to its right are smaller. You will see that the sum of these counts gives N_D; you can then proceed to calculate τ as before.

This arithmetical method is illustrated on the same data as before.

Salesman,	i	4	1	7	9	15	11	13	8	6
Demos,	X_i	96	104	113	115	126	135	137	142	143
Sales,	Y_i	17	25	33	31	36	34	35	30	39
Count		0	0	2	1	3	1	1	0	0

Salesman,	i	12	5	14	3	2	10
Demos.	X_i	145	149	151	156	161	175
Sales,	Y_i	43	49	42	40	47	44
Count		2	4	1	0	1	–

For the counts on the final row we first check $Y_1=17$ against *all* the other 'Y's (as they all lie to the right of of the 17); in fact they are all greater than 17, resulting in the first count of 0 on the bottom row. Then we check the next Y, i.e. $Y_2=25$, against all the 'Y's lying to its right; again they are all greater than 25, resulting in the second count of 0 in the bottom row. Then we check $Y_3=33$ with all the 'Y's to its right; in this case there are two (31 and 30) less than 33, and this produces the count of 2 as printed in the table. Carrying on in this way, all the counts may be confirmed; and adding them gives $N_D=16$, as above.

Note that there is some similarity between this counting method and that for the Mann–Whitney U (Ch. 5). Note also that this arithmetical calculation may, if preferred, be carried out on the ranks $r(X_i)$ and $r(Y_i)$ in place of the original numbers X_i and Y_i. (As an exercise, try it and see!)

One-sided and two-sided tests

Reread the corresponding paragraph for r_S, earlier in the chapter, replacing r_S by τ throughout.

Treatment of ties

Simple ties are dealt with in the graphical method easily enough by counting $\frac{1}{2}$ for any crossing that lies on the X or the Y 'axes', but multiple ties can cause difficulties. With the arithmetical calculation, however, there is an easy rule: count $\frac{1}{2}$ for any pair whose X values and/or Y values are tied.

Exercise Given the following data, verify that $\tau=17/21=0.8095$.

X	1	3	3	6	7	7	10
Y	2	4	5	5	10	11	11

(If you get $N_D=\frac{1}{2}+\frac{1}{2}$, giving $\tau=19/21=0.9048$, you have probably not allowed for the ties in the X values.)

Null distributions

Again, the principle is essentially the same as for r_S. You may like to confirm that, when $n=5$, we obtain $\text{Prob}(\tau=1)=1/120$, $\text{Prob}(\tau\geqslant 0.8)=5/120$ and $\text{Prob}(\tau\geqslant 0.6)=4/120$, and that the critical values in Table M are in accord with these results.

It is worth pointing out that there are just 11 possible values of τ with $n=5$ (-1, -0.8, -0.6, . . . ,0.8, 1) whereas there were 21 possible values of r_S (-1, -0.9, -0.8, . . . ,0.9, 1). This difference between the number of possible values of the two test statistics increases markedly with n, so that τ is in fact a cruder representation of the data than is r_S. This may well explain, at least in part, the apparent slight superiority in power of Spearman's test over Kendall's.

Large sample sizes

The null distribution of τ when n is large is approximately normal with mean $\mu=0$ and standard deviation $\sigma=\sqrt{2(2n+5)/9n(n-1)}$, and the convergence to this asymptotic distribution is faster than was the case with r_S. Thus approximate critical values may be obtained using $z\sigma$, where z is the appropriate percentage point of the standard normal distribution.

So, for example, with $n=50$ and $\alpha_2=5\%$ (for which $z=1.96$), we obtain an approximate critical value of $1.96\sqrt{2\times 105/(9\times 50\times 49)} = 0.1913$, which compares very favourably with the critical value 0.1918 given in Table M. This is of similar accuracy to the Spearman example, but if you go further out into the tails of the distribution (i.e. take smaller values of α) you will find the Kendall approximation to be the more accurate.

Further discussion on correlation

There are a number of matters concerning correlation that sometimes tend to trap the unwary into misinterpreting the data and the results of the significance tests. This section is an attempt to make the unwary less so.

(a) It is perhaps natural to imagine that a strongly significant result from either of the two tests in this chapter implies, and is implied by, a high correlation between the X and Y variables. Neither implication is correct! The test simply judges whether or not the null hypothesis of zero correlation is consistent with the

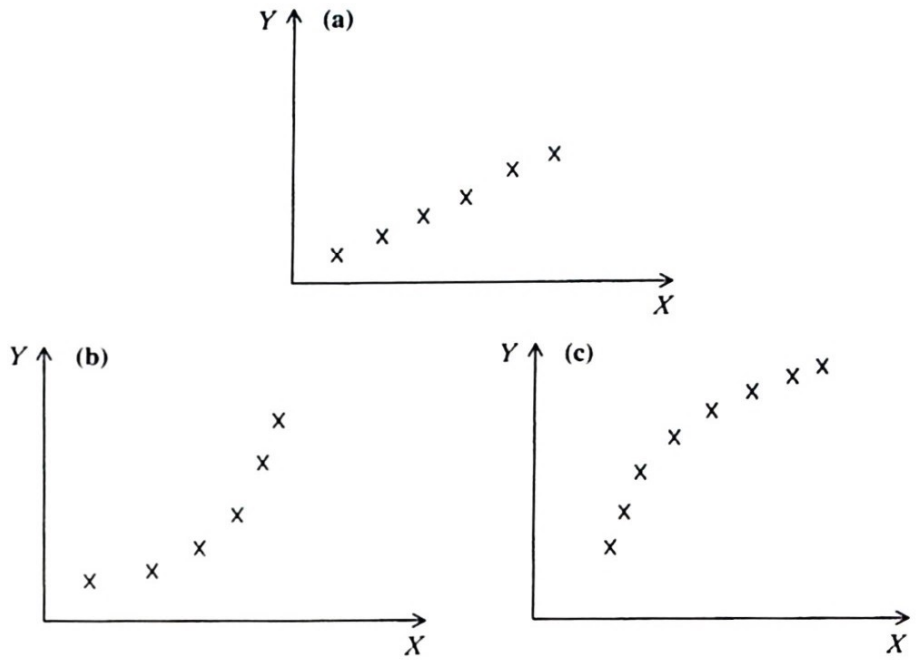


Figure 9.4

data. Thus, if n is large, a pretty small value of r_S or τ may be highly significant. If you check the tables you will see, for example, that with $n=50$ a value of 0.4 for either coefficient is extraordinarily significant. Such a value indicates nothing more than a rather loose relationship between X and Y; but, on the other hand, it is high enough virtually to ensure that some such relationship does exist. However, if n is small, quite high values of r_S or τ may not be significant; for example, if $n=6$, a value of around 0.7 fails to reject H_0. Similar difficulties can arise with hypotheses tests in other areas of statistics, but somehow in practice they seem to cause most problems with correlation.

(b) Even if r_S or τ turn out to be zero or very small, that does not necessarily imply that X and Y are unrelated; in fact, they can still be very strongly related. Remember that r_S and τ are measuring the extent of a monotonic relationship between X and Y (and the traditional correlation coefficient r is even more specific as it measures the extent of *linear* relationship). Thus if the data are exactly related but in a *non*-monotonic way, it is quite conceivable for any or all of the three coefficients to be close to zero.

(c) As mentioned in (b), r_S and τ measure the tendency towards *monotonicity*, not just linearity. This is an automatic consequence of the distribution-free philosophy. As with other distribution-free statistics, neither r_S nor τ relates directly to the precise numerical values in the data, but only indirectly via their ranks. This has to imply that the rank correlation coefficients measure the tendency for X and Y to increase together (or for one to decrease as the other increases), *irrespective of whether the relationship is linear or not*. Thus r_S and τ will both equal +1 in all three scatterplots shown in Figure 9.4, whereas r would be less than 1 in (b) and (c). Whether this is good or bad depends on context. When exploring the possibility of two variables being related, it is probably better not to

be too specific, so that the rank correlation methods will then be preferable. However, the simplest useful method of predicting the Y that will result from a given value of X is by a linear equation (see Ch. 10):

$$\text{predicted } Y = a + bX$$

So r would be the more appropriate measure if we wished to examine whether such a prediction equation might be appropriate.

(d) A very common error is to assume that high correlation between two variables implies that one of them has a *causal* effect on the other. This error is possibly encouraged by some of the standard terminology in the subject; for example, in the prediction equation above, Y is often referred to as the *dependent* variable. However, even if the X and Y variables are well related numerically, there may still be no causal, physical or other connection between them. For example, there are some well known data that show extremely high correlation between electricity consumption and birth rate! In this and many other situations, the real reason is that there are *other* variables or factors having a direct effect on both variables being considered and, without such extraneous influences, there might well be no relationship of any importance between the two variables with which we are concerned.

RANK CORRELATION WITH NON-NUMERIC DATA

Both r_S and τ essentially measure the correlation between two variables through their ranks. A useful consequence of this feature is that we can measure the correlation between variables that are not measured or counted – all we need to be able to do is *rank* the data. We illustrate this idea by applying Spearman's r_s to a rather interesting problem.

PRACTICAL PROBLEM

A colleague of ours has two teenage daughters who, naturally, have quite a number of suitors. Just for fun, he asked us to devise a simple test to see whether his daughters are likely to compete for the same type of suitor. We drew up a list of attributes of males (trying to cast ourselves as females!) and then told the girls to rank these attributes from 1 (for the most important) to 10 (the least important). Of course, no conferring was allowed.

HYPOTHESES

Since our colleague is particularly interested in whether his daughters are likely to compete with each other (i.e. have highly correlated attitudes as measured by these attributes), a one-sided alternative hypothesis of positive correlation is appropriate. Thus the hypotheses may be stated as follows:

H_0: There is no correlation between the two girls' attitudes to the given attributes.

H_1: There is some positive correlation between the girls.

DATA

The list of the attributes and the ranks given by the two girls, Kathryn and Fiona, are shown below.

	Ranks	
Attribute	Fiona	Kathryn
personality	2	1
sense of humour	1	2
wealth	4	6
personal hygiene	5	5
good cook	10	7
good looks	8	9
height	7	8
age/maturity	6	3
sex appeal	9	10
dependability	3	4

CRITICAL REGIONS

Consulting Table L with $n=10$ gives the $\alpha_1=5\%$ and 1% critical regions as $r_S \geqslant 0.5636$ and $r_S \geqslant 0.7455$ respectively.

CALCULATION OF THE STATISTIC

Since we already have the ranks, the calculation of r_S is easily performed as in the following table:

	Ranks			
Attribute	Fiona	Kathryn	d_i	d_i^2
personality	2	1	1	1
sense of humour	1	2	−1	1
wealth	4	6	−2	4
personal hygiene	5	5	0	0
good cook	10	7	3	9
good looks	8	9	−1	1
height	7	8	−1	1
age/maturity	6	3	3	9
sex appeal	9	10	−1	1
dependability	3	4	−1	1

Summing the d_i^2 values gives $D^2 = 28$, and so r_s is:

$$r_S = 1 - \frac{6D^2}{n^3 - n} = 1 - \frac{6 \times 28}{990} = 0.8303$$

CONCLUSION

This result is well into the $\alpha_1 = 1\%$ critical region, and so it seems that our colleague is in for an interesting time watching his daughters compete for the same males!

This example illustrates very well the effectiveness of rank correlation on problems where the data are 'non-numeric'; and this, of course, is a definite advantage over Pearson's classical measure of correlation.

Assignments

1 Diabetes

In 1979, G. M. Reaven and R. G. Miller ('An attempt to define the nature of chemical diabetes using a multidimensional analysis' *Diabetologia* **16**, 17–24) presented the results of an experiment concerning diabetes in non-obese adults. Part of the information collected was the relative weight and the insulin resistance of each subject. These data for the first 20 subjects are given below.

Patient	Relative weight	Insulin resistance	Patient	Relative weight	Insulin resistance
1	0.81	55	11	0.90	53
2	0.95	76	12	0.73	66
3	0.94	105	13	0.96	142
4	1.04	108	14	0.84	93
5	1.00	143	15	0.74	68
6	0.76	165	16	0.98	102
7	0.91	119	17	1.10	76
8	1.10	105	18	0.85	37
9	0.99	98	19	0.83	60
10	0.78	94	20	0.93	50

Task: Investigate whether these data provide evidence of positive correlation between relative weight and insulin resistance.

2 Strange correlation

The table below gives the number of postgraduate awards in medical sciences and the death rate per million from tuberculosis for the years 1959–69 (source: *Annual Abstract of Statistics* 1970).

Year	Number of awards	Tuberculosis death rate
1959	277	83
1960	318	74
1961	382	71
1962	441	65
1963	486	62
1964	597	52
1965	750	47
1966	738	48
1967	849	42
1968	932	43
1969	976	38

Task: Show that these data provide strong evidence of negative correlation between the number of awards and the tuberculosis death rate. Explain this 'strange' result.

Program 9.1 Spearman's rank correlation coefficient

```
100 REM *** SPEARMAN'S RANK CORRELATION COEFFICIENT ***

110 REM ** INPUT NUMBER OF PAIRS OF OBSERVATIONS **
120 INPUT "HOW MANY PAIRS OF OBSERVATIONS ? ";N
130 DIM X(N),Y(N),D(N),RX(N),RY(N),RK(N)
140 PRINT

150 REM ** INPUT OBSERVATIONS **
160 PRINT "INPUT THE PAIRS OF OBSERVATIONS "
170 FOR I=1 TO N
180 INPUT X(I),Y(I)
190 LET D(I)=X(I)
200 NEXT I

210 REM ** RANK  THE X-SAMPLE **
220 GO SUB 1000
230 FOR I=1 TO N
240 LET RX(I)=RK(I)
250 LET D(I)=Y(I)
260 NEXT I

270 REM ** RANK  THE Y-SAMPLE **
280 GO SUB 1000
290 FOR I=1 TO N
300 LET RY(I)=RK(I)
310 NEXT I

320 REM ** CALCULATION OF SPEARMAN'S RHO **
330 LET P=0
340 FOR I=1 TO N
350 P=P+(RX(I)-RY(I))^2
360 NEXT I
370 LET R=1-6*P/N/(N*N-1)
380 PRINT "VALUE OF SPEARMAN'S RHO = ";R
390 IF N<=50 THEN STOP
```

Continued overleaf

Program 9.1 Continued

```
400 REM ** CRITICAL VALUES FOR LARGE SAMPLE SIZES (N > 50) **
410 INPUT "PERCENTAGE POINT OF STANDARD NORMAL DISTRIBUTION ? ";Z
420 LET SD=1/SQRT(N-1)
430 PRINT "APPROXIMATE CRITICAL VALUE = ";INT(10000*Z*SD+.5)/10000

1000 REM ** RANK SUBROUTINE - USE LINES 1000-1140 OF PROGRAM 3.1 **
```

Program 9.2 Kendall's rank correlation coefficient

```
100 REM *** KENDALL'S RANK CORRELATION COEFFICIENT ***

110 REM ** INPUT NUMBER OF PAIRS OF OBSERVATIONS **
120 INPUT "HOW MANY PAIRS OF OBSERVATIONS ? ";N
130 DIM X(N),Y(N)
140 PRINT

150 REM ** INPUT OBSERVATIONS **
160 PRINT "INPUT THE PAIRS OF OBSERVATIONS "
170 FOR I=1 TO N180 INPUT X(I),Y(I)
190 NEXT I

200 REM ** CALCULATION OF KENDALL'S TOR **
210 LET ND=0
220 FOR I=1 TO N-1
230 FOR J=I+1 TO N
240 IF X(I)<X(J) AND Y(I)>Y(J) THEN LET ND=ND+1
250 IF X(I)>X(J) AND Y(I)<Y(J) THEN LET ND=ND+1
260 IF X(I)=X(J) OR Y(I)=Y(J) THEN LET ND=ND+0.5
270 NEXT J
280 NEXT I
290 LET T=1-4*ND/(N*(N-1))
300 PRINT
310 PRINT "VALUE OF KENDALL'S TAU = ";T
320 PRINT

330 REM ** APPROXIMATE CRITICAL VALUES FOR LARGE SAMPLES **
340 LET SD=SQRT(2*(2*N+5)/(9*N*(N-1)))
350 INPUT "PERCENTAGE POINT OF STANDARD NORMAL DISTRIBUTION ? ";Z
360 PRINT "APPROXIMATE CRITICAL VALUE = ";INT(10000*Z*SD+.5)/10000
```

10
Regression

Introduction

In our discussion of correlation in Chapter 9, we tested for the existence of some kind of relationship between two variables. Within our distribution-free environment, the most natural type of correlation of interest is monotonicity, whereas in the classical approach one tests for linearity or some other more detailed mathematical interdependence.

A significant result from such a test is evidence that a relationship of the relevant type is present. It rarely indicates a *perfect* relationship in the sense that the value of one variable, say Y, can be exactly predicted if the value of the other variable, say X, is known. The relationship is rather between X and the *average* value of Y; the *actual* value of Y is made up of this average value plus some random variation. The best we can do, therefore, after having established that X and Y are correlated, is to attempt to *estimate* the average value of Y in terms of the value of X. This is the problem of **regression**.

We should point out that, in general, one can always go through the motions of calculating a detailed regression relationship; but it is the corresponding correlation test that tells us whether the regression is likely to be useful or not (according as to whether or not the test gives a significant result).

The origins of the use of the word 'regression' in this context go back some 100 years to the geneticist, Francis Galton. In studying the relationship between the heights of parents and their offspring, he noticed two rather unsurprising things. One was that parents' and offsprings' heights are positively correlated (tall parents tend to have tall children, and short parents short children). His other observation was a little more subtle. Let us consider, say, a child of tall parents. Then, although from what we have just said, this child would be expected to grow up to have an above-average height, the fact is that, on average, the child would *not* reach the height of the corresponding parent. In other words, he or she would tend to be closer to the population mean than that parent. Similarly, a child of short parents, though likely to be below-average height, tends to be taller than the corresponding parent, being again closer to the population mean. Galton referred to this tendency of offsprings' heights to revert towards the population mean as '*regression* towards the mean', that is 'moving back towards the mean'.

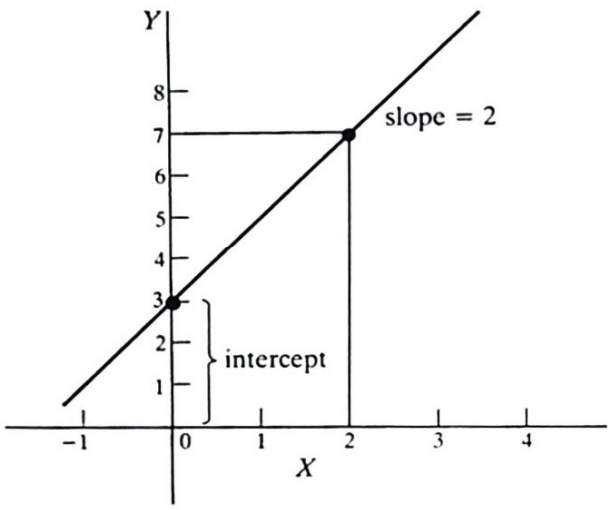


Figure 10.1

Practical applications of regression are common. For example, an economist may be interested in the relationship between share prices and the level of investment; a biologist may wish to establish a relationship between an animal's body weight and its food consumption; or a keep-fit instructor may want to know the relationship between a person's time to run a fixed distance and his heart-rate recovery. In each case, the idea is to establish a relationship either graphically or in the form of an equation which can be used to predict values of one variable, conventionally denoted Y and often called the *dependent* variable, given values of the other variable X, the so-called *independent* variable. For example, if the relation between them is believed to be $Y=4+3X$, then the predicted value of Y (or rather of its average) when $X=6$ is $Y=4+18=22$.

The nature of these relationships could be of many forms or *models*. For instance, one might presume that the data are best modelled by the following:

(a) A *linear* relationship, represented by the equation

$$Y = a + bX$$

where a and b are respectively the *intercept* and the *slope* of the line. For example, $Y=3+2X$ has the graphical form shown in Figure 10.1.

(b) An *exponential* relationship, represented by an equation of the type

$$Y = ab^X$$

For example, $Y=3(2^X)$ has the graphical form shown in Figure 10.2.

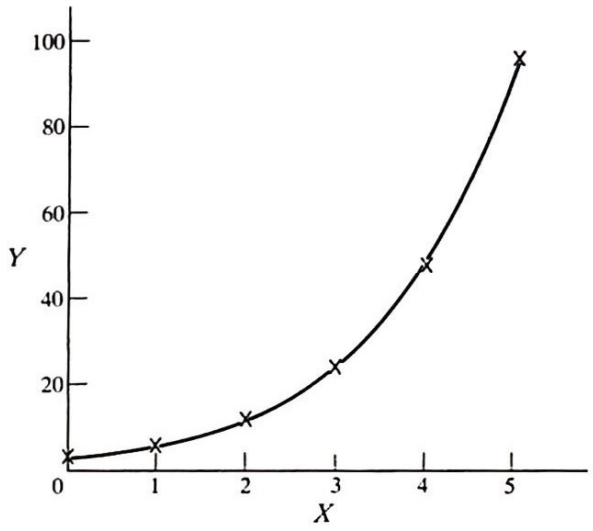


Figure 10.2

It is important to realize that we do not usually expect the value of Y itself to be given by such equations, but it is feasible for the *average* value of Y, for some given value of X, to be represented in this way. Whatever relationship is appropriate to the data, some technique needs to be used to *estimate* the values of a and b from the observed data.

It is obviously necessary to have some idea about which type of relationship best models the data. A simple, but effective, method of doing this is to construct a scatterplot, a technique that we met in Chapter 9. Figure 10.3 gives some indications of the kind of thing to expect.

In the first part of this chapter, we look at a method of estimating a and b when the underlying model is assumed to be linear (see Fig. 10.3a); the so-called *linear regression model*. We will also discuss hypothesis tests for inferences regarding the unknown parameter b (the slope of the regression line). After this, we examine a method for *monotonic* regression, i.e. which requires only the assumption of monotonicity rather than any more specific form of relationship – in other words, the technique is substantially *model-free*.

Before looking at these methods, it is worth while to mention something about the classical approach to linear regression. This is based on the 'least-squares' method, so-called because it is based on minimizing the sum of the squares of errors (or *residuals*) between the actual values of Y and the values of Y predicted by the regression. Figure 10.4 shows these residuals for a 'reasonable-looking' line.

The least-squares method is widely used and, as long as the underlying model is valid, it is without doubt a sound technique *provided there are no 'rogue' points (or outliers)* in the data. Unfortunately, outliers are often present; they can be caused by a number of reasons, for instance human

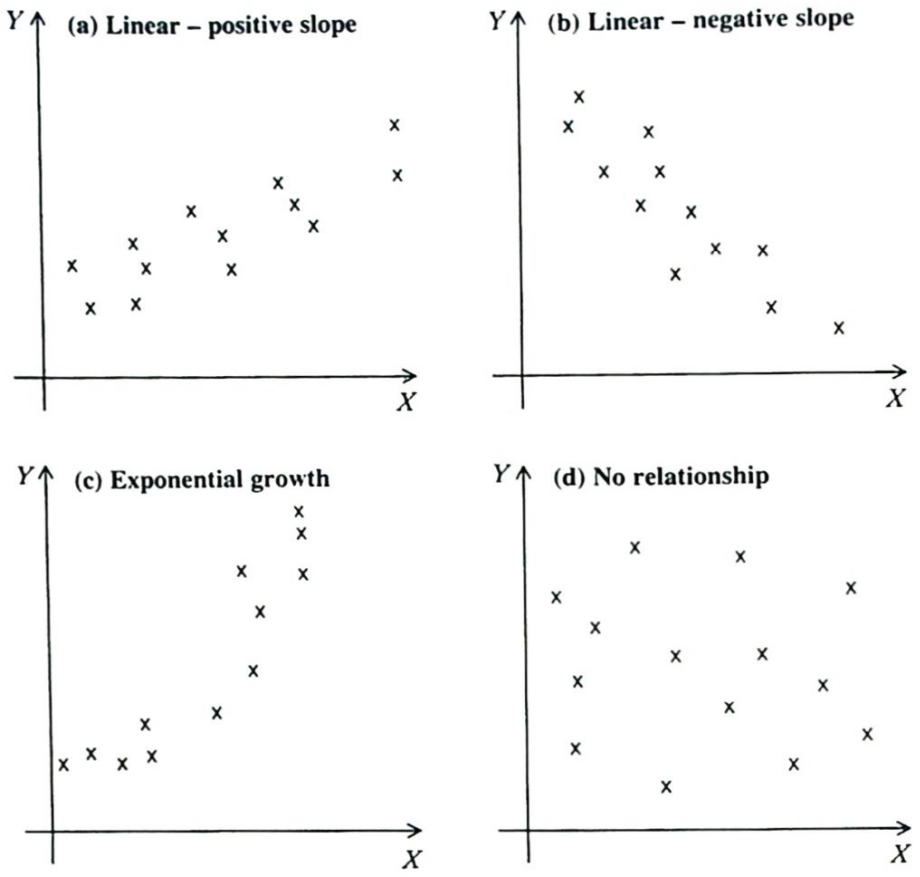


Figure 10.3

error, machine error, untypical conditions and so on. Any such problem points can seriously disturb the least-squares regression line and thus cause it to be quite unrepresentative of the actual relationship between the two variables. Figure 10.5 illustrates what can happen when there are just two rogue points in the data. Common sense indicates that the freehand line drawn in Figure 10.5 is far more representative of the underlying process than is the least-squares line. Fortunately, the regression methods in this chapter, and the next one, are relatively 'robust' in the sense that they are nowhere as susceptible to rogue values as is the traditional least-squares method.

THEIL'S 'INCOMPLETE' METHOD

Introduction

Theil's original method for estimating the unknown parameters a and b dates back to 1950 (Theil 1950). However, the amount of calculation required in this method can be quite substantial. With n data points, the

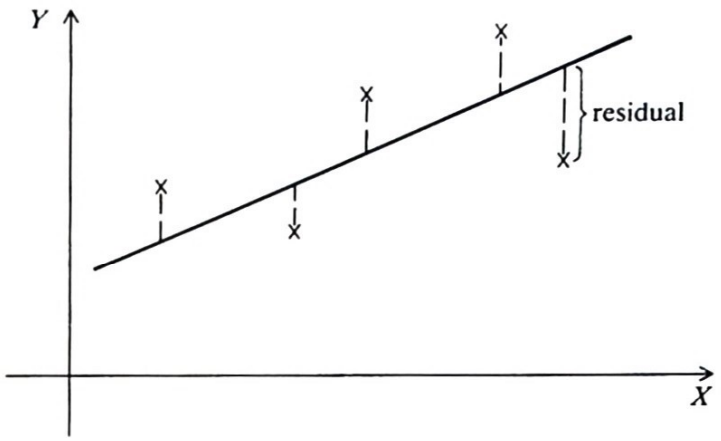


Figure 10.4

number of computations needed is $\frac{1}{2}n(n-1)$; for example, with only 10 data points there are 45 calculations, while with 20 points there are 190. Although the calculations involved are fairly simple, the fact that there can be so many of them makes Theil's original technique rather unappealing. It is for this reason that his method was later modified in order to substantially reduce the number of calculations whilst still maintaining most of the effectiveness of the original technique. This modification, which we discuss below, is often called Theil's 'incomplete' method; the effect of the modification is such that, for example, with 10 or 11 data points only five simple calculations are needed, and only 10 calculations are needed with 20 or 21 points.

Following our usual pattern, we will present this technique through a practical problem, and later we will consider tests for the intercept a and slope b.

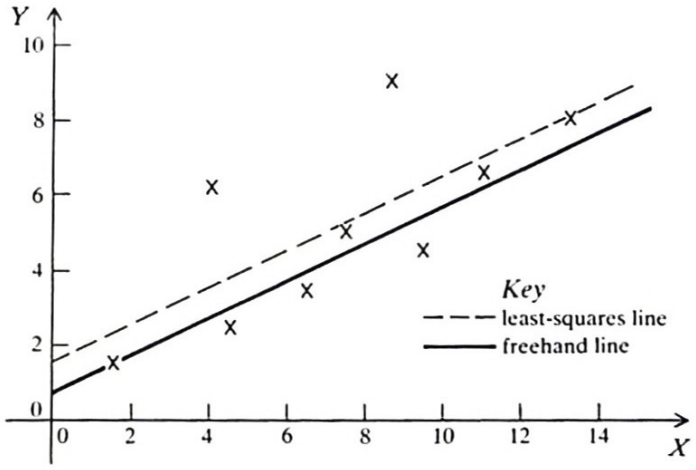


Figure 10.5

PRACTICAL PROBLEM

A geologist conducted an experiment on one of Scotland's largest rivers. He took samples of pebbles from the riverbed at approximately 0.5 km intervals from the source to the first (major) tributary. The volume of each pebble was measured and its distance from the source recorded. He reasoned that the larger stones would tend to be located close to the source, with the stone size decreasing downstream. If this were the case, he hoped this investigation would lead to a linear relationship between the average volume of stones (we call this the Y variable) and the distance from the source (the X variable).

DATA

Fifteen samples were taken along a 7 km stretch of the river from the source. The results obtained were:

Distance, X	0.2	0.5	0.9	1.4	2.1	2.7	3.0	3.6
Volume, Y	50	48	42	36	34	32	29	31
Distance, X	4.1	4.4	5.0	5.6	6.2	6.6	7.1	
Volume, Y	28	24	19	17	10	12	9.5	

PROCEDURE

Before looking at the rationale of Theil's method, let us review the notion of a *slope* or *gradient* of a line. The slope measures the steepness of the line and, for example, the slope of the line joining points (X_1,Y_1) and (X_2,Y_2) is measured by

$$B_{12} = \frac{Y_2 - Y_1}{X_2 - X_1}$$

In practical problems, there are usually more than two points, which, because of random variation, do not all lie on one line. Our basic assumption in regression is that without the random variation the points *would* all lie on a line $Y=a+bX$ – and it is these constants a and b we wish to estimate. For example, the situation may be like that shown in Figure 10.6, where the line drawn represents the theoretical line $Y=a+bX$ which we seek to estimate.

Suppose we were to calculate the slopes of the three lines joining:

(X_1,Y_1) and (X_4,Y_4), i.e.

$$B_{1,4} = \frac{Y_4 - Y_1}{X_4 - X_1}$$

(X_2,Y_2) and (X_5,Y_5), i.e.

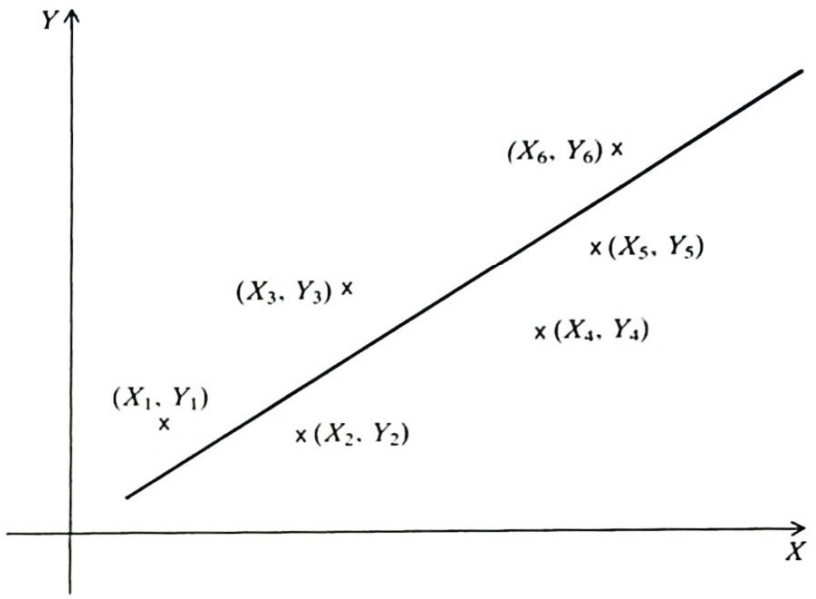


Figure 10.6

$$B_{2,5} = \frac{Y_5 - Y_2}{X_5 - X_2}$$

(X_3,Y_3) and (X_6,Y_6), i.e.

$$B_{3,6} = \frac{Y_6 - Y_3}{X_6 - X_3}$$

These are shown as broken lines in Figure 10.7. Of course, we could have considered other possibilities, say (X_2,Y_2) joining (X_3,Y_3); but we have chosen those particular lines used in Theil's incomplete method.

Clearly, each of these slopes are different; but if we take an average, in particular the *median*, of these slopes then we should obtain a value that is not too far from the true value of b. This is Theil's incomplete method, and it can be summarized as follows:

Step 1 Suppose we have n data points (X_1,Y_1), (X_2,Y_2), . . . , (X_n,Y_n). Arrange these points in ascending order of the X values. (Note: this step is automatic for our data since X represents the increasing distance from the source.)

Step 2 Divide the ordered data at the halfway mark to give two equal groups; if there is an odd number of points then simply ignore the middle point.

Step 3 Calculate the values of the slopes of the lines joining

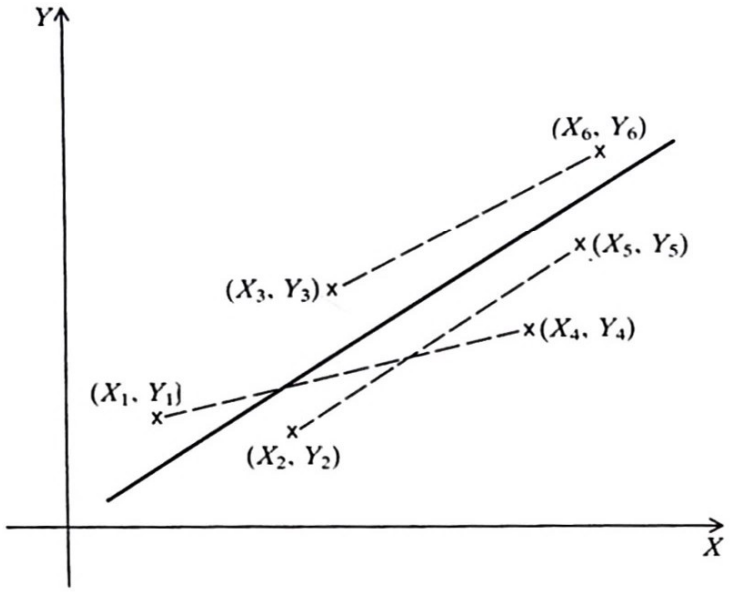


Figure 10.7

(a) the first points of the two groups,
(b) the second points of the two groups,
and so on to the last points of the two groups.

Step 4 Calculate the median of these slopes as the estimate B of the unknown slope b.

The application of this method to our practical problem is quite simple. Since step 1 is already done we proceed immediately to step 2:

Step 2 The two groups of data points are as follows (note that the middle point (3.6,31) is discarded to give two equal groups of points):

Group 1	X	0.2	0.5	0.9	1.4	2.1	2.7	3.0
	Y	50	48	42	36	34	32	29
Point number		1	2	3	4	5	6	7
Group 2	X	4.1	4.4	5.0	5.6	6.2	6.6	7.1
	Y	28	24	19	17	10	12	9.5
Point number		8	9	10	11	12	13	14

Step 3 The individual slopes are

$$B_{1,8} = \frac{28 - 50}{4.1 - 0.2} = -5.64$$

$$B_{2,9} = \frac{24 - 48}{4.4 - 0.5} = -6.15$$

$$B_{3,10} = \frac{19 - 42}{5.0 - 0.9} = -5.61$$

$$B_{4,11} = \frac{17 - 36}{5.6 - 1.4} = -4.52$$

$$B_{5,12} = \frac{10 - 34}{6.2 - 2.1} = -5.85$$

$$B_{7,14} = \frac{9.5 - 29}{7.1 - 3.0} = -4.76$$

Step 4 Arranging these seven values in ascending order gives the sequence:

$$\begin{array}{lllll} -6.15 & -5.85 & -5.64 & -5.61 & -5.12 \\ -4.76 & -4.52 & & & \end{array}$$

from which we immediately find the median slope to be -5.61. This is our estimate B for the unknown slope b.

Now recall that we are attempting to fit an equation of the form $Y=a+bX$ to the observed data. We have just obtained the estimate $B = -5.61$ of b. To estimate the value of a, Theil proposed that, for each data point (all 15 in our problem), the quantity $A_i=Y_i-BX_i$ ($i=1,2,\ldots,n=15$) should be calculated. The required estimate A of a is then taken to be the median of the A_i values. For our data we obtain:

$$\begin{aligned} A_1 &= 50 + 5.61\times0.2 = 51.12 \\ A_2 &= 48 + 5.61\times0.5 = 50.81 \\ A_3 &= 42 + 5.61\times0.9 = 47.05 \\ A_4 &= 36 + 5.61\times1.4 = 43.85 \\ A_5 &= 34 + 5.61\times2.1 = 45.78 \\ A_6 &= 32 + 5.61\times2.7 = 47.15 \\ A_7 &= 29 + 5.61\times3.0 = 45.83 \\ A_8 &= 31 + 5.61\times3.6 = 51.20 \\ A_9 &= 28 + 5.61\times4.1 = 51.00 \\ A_{10} &= 24 + 5.61\times4.4 = 48.68 \\ A_{11} &= 19 + 5.61\times5.0 = 47.05 \\ A_{12} &= 17 + 5.61\times5.6 = 48.42 \\ A_{13} &= 10 + 5.61\times6.2 = 44.78 \end{aligned}$$

$$A_{14} = 12 + 5.61 \times 6.6 = 49.03$$
$$A_{15} = 9.5 + 5.61 \times 7.1 = 49.33$$

You may easily verify that the median of the A_i values is 48.42 and so, finally, our estimated regression line is $Y = 48.42 - 5.61X$.

(You should note that this model implies there are no stones at distances greater than 48.47/5.61 = 8.64 km from the source. If this is unrealistic, then the assumption of a linear model is probably wrong, and perhaps the method appropriate to the general monotonic model, described later, should be used instead.)

Figure 10.8 shows the data points and this estimated regression line.

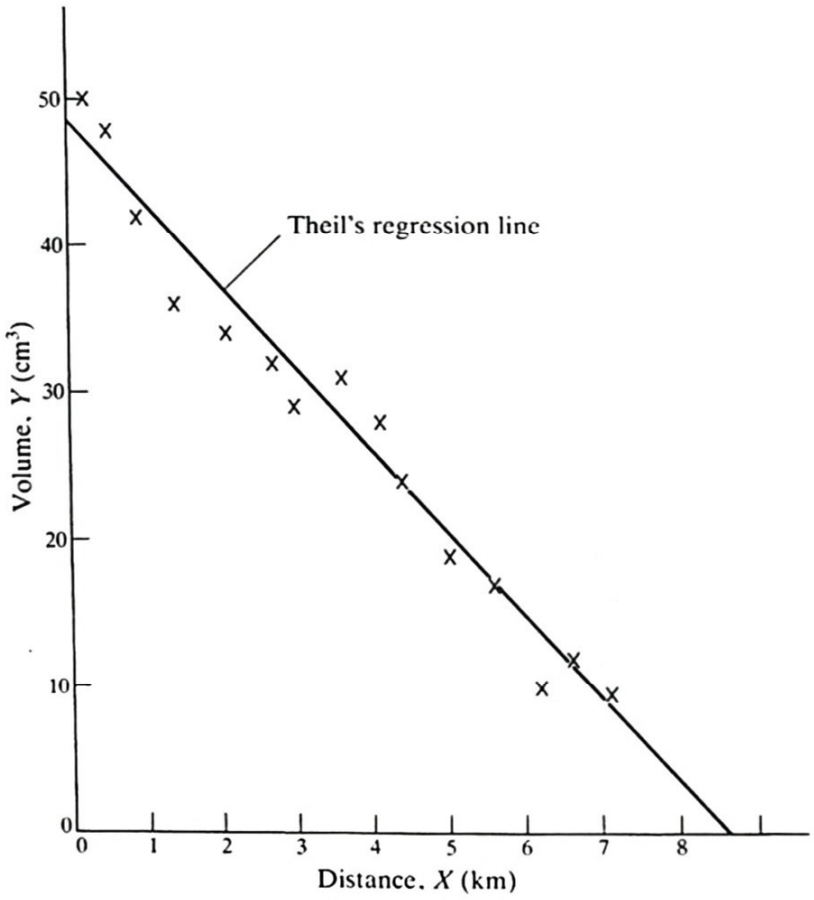


Figure 10.8

Hypothesis test for the slope, b

Perhaps the most frequent type of hypothesis test in linear regression problems has the null hypothesis H_0: $b=0$. Now, *if* b is equal to 0 then the regression line reduces to just $Y=A$, i.e. the prediction of Y is constant for all values of X. So a rejection of this H_0 effectively

constitutes evidence that useful predictions can be made from the regression line, subject of course to the linear model being appropriate. Sometimes, a look at the scatterplot is sufficient to judge whether B, the estimate of b, is significantly different from zero. However, we may also wish to see whether there is evidence that b is different from some other particular value of interest. Thus for example, in our practical problem we may wish to test the null hypothesis H_0: $b=-4$, i.e. that the volume of stones decreases by a factor of 4 as the distance from the source increases, against the alternative hypothesis of H_1 that the factor is different from 4. We already know that our estimated value of b is $B = -5.61$; but there is almost bound to be some error in this estimate, and more formal work needs to be carried out to ascertain whether our estimate is reasonable under the H_0 of $b=-4$, or whether the difference is sufficient to constitute evidence against this H_0.

Fortunately there is a simple test for such a hypothesis. It is based on Spearman's rank correlation coefficient, which we met in the previous chapter.

PRACTICAL PROBLEM

We shall use the scenario based on the geology experiment introduced earlier in the chapter.

HYPOTHESES

For the purpose of illustrating the test procedure, we will suppose that the geologist's initial thoughts (which were based on previous experiments on similar rivers) led him to believe that the volume of stones would decrease by a factor of 4 with increasing distance from the river's source; thus the null hypothesis is that the slope b of the regression line is -4. Since he has no reason to suspect that the slope should be substantially less than -4 on the one hand or greater than -4 on the other, it seems best to select a two-sided alternative hypothesis to the effect that the slope is not equal to -4. Thus we have:

$$H_0: b = -4.$$

$$H_1: b \neq -4.$$

DATA

The data are as before.

TEST PROCEDURE

For each data pair (X_i, Y_i) for $i=1,2,\ldots,n$, we calculate the difference $T_i=Y_i-bX_i$, which in our example is $T_i=Y_i+4X_i$ for $i=1,2,\ldots,15$. Now

if H_0 is true there will be no strong pattern (i.e. no significant correlation) between the X_i and the T_i values. But if H_0 is false, there will be correlation (either negative or positive) between these two quantities. So all we need to do is to perform Spearman's rank correlation test on the X and the T values. For our data, the 15 differences are given by

$$T_1 = 50+4\times0.2 = 50.8$$

$$T_2 = 48+4\times0.5 = 50.0$$

. . . and so on

The following table gives the results of these calculations and those required for Spearman's correlation coefficient.

			Ranks			
Point	X_i	T_i	$r(X_i)$	$r(T_i)$	d_i	d_i^2
1	0.2	50.8	1	15	−14	196
2	0.5	50.0	2	14	−12	144
3	0.9	45.6	3	12	−9	81
4	1.4	41.6	4	6½	−2.5	6.25
5	2.1	42.1	5	8	−3	9
6	2.7	42.8	6	9	−3	9
7	3.0	41.0	7	5	2	4
8	3.6	45.4	8	11	−3	9
9	4.1	44.4	9	10	−1	1
10	4.4	41.6	10	6½	3.5	12.25
11	5.0	39.0	11	4	7	49
12	5.6	49.4	12	13	−1	1
13	6.2	34.8	13	1	12	144
14	6.6	38.4	14	3	11	121
15	7.1	37.9	15	2	13	169

Using Spearman's rank correlation formula

$$r_S = 1 - \frac{6D^2}{n^3-n}$$

where in this case $D^2=955.5$, we obtain

$$r_S=1 - \frac{6\times955.5}{15^3-15} = -0.706$$

CRITICAL REGIONS

Consulting the table of critical values for Spearman's rank correlation coefficient (Table L) with $n=15$, we obtain the $\alpha_2=5\%$ and 1% critical regions as $|r_S|\geq0.5214$ and $|r_S|\geq0.6536$ respectively.

SOLUTION

The calculated value of $r_S = -0.706$ falls in the $\alpha_2 = 1\%$ critical region, giving us very strong evidence to support H_1. It would appear that the geologist's hunch, at least with regard to the stones in this river is wrong; the true factor of decrease would appear to be greater than he surmised.

THE IMAN–CONOVER METHOD FOR MONOTONIC REGRESSION

Theil's regression method is particularly robust in the sense that it is not unduly influenced by outliers, whether they be caused by the underlying distribution of one of the variables or are just rogue points. This is a substantial advantage over the standard regression methods based on the least-squares principle. However, it is still a *linear* regression method in that it assumes the underlying relationship between the values of X and Y is linear, i.e. $Y = a + bX$.

In regression work, this assumption can be generalized for, as we have previously mentioned, instead of a linear relationship an exponential model might be assumed. There are other possibilities, such as a quadratic relationship ($Y = a + bX + cX^2$), or a cubic, or a higher-powered polynomial.

It seems more in the spirit of the development of distribution-free methods to seek regression techniques that are not only distribution-free but *model-free* also, i.e. techniques that do not require much in the way of assumptions about the particular form of the regression relationship. Such methods only began to be developed in the 1970s, and the particular technique described here was published by Iman and Conover (1979) quite recently. The only assumption required for this technique is that the underlying relationship between X and Y be *monotonic*, which is exactly the kind of general relationship measured and tested by the rank correlation methods in Chapter 9. In fact, as we shall see, there is a strong link between this regression method and Spearman's r_S, and this is entirely analogous to the link between classical regression and Pearson's correlation coefficient.

PRACTICAL PROBLEM AND DATA

We return to the dependence of sales on 'demos' and 'simple visits' used throughout Chapter 9. In particular, since the tests in Chapter 9 strongly indicated some positive correlation between demos and sales, let us focus our attention on predicting the number of sales in terms of the number of demos made by a typical salesman. For convenience we repeat the relevant data below.

Salesman	Demos	Sales
1	104	25
2	161	47
3	156	40
4	96	17
5	149	49
6	143	39
7	113	33
8	142	30
9	115	31
10	175	44
11	135	34
12	145	43
13	137	35
14	151	42
15	126	36

The Iman–Conover method produces a *graph* (rather than an equation) from which the (average) number Y of sales can be estimated from the number X of demos. No assumption is made about the relationship except that it is monotonic, which, in effect, means that the average number of sales does increase with the number of demos, and, of course, we already have strong evidence for this from our work in Chapter 9. Note that, while this assumption is eminently reasonable, the assumption required by Theil's method that the relationship is *linear* might not be so reasonable since, for example, if a salesman substantially increases his number of demos, the length, quality and effectiveness of those demos might deteriorate so that 'sales per demo' might be expected to suffer. This method deals quite *automatically* with such effects.

METHOD

In this section we demonstrate the mechanics of the method, leaving the justification to the 'Theory' section below. The operation may seem a little involved, but remember that it is tackling the rather ambitious task of predicting Y from X using very few assumptions about the underlying relationships.

Step 1 We work with a regression equation that relates the ranks $r(X)$ and $r(Y)$ of X and Y respectively; this equation is

$$1r(Y) = r_S r(X) + \tfrac{1}{2}(n+1)(1-r_S)$$

where r_S is the appropriate value of Spearman's rank correlation coeffficient. From Chapter 9, the value of r_S in this problem is 0.8714 and so with $n=15$, the above rank *regression* equation becomes

$$r(Y) = 0.8714r(X) + 1.0288$$

Step 2 Now we construct a graph of the Y_i (number of sales) against their ranks $r(Y_i)$. The table below shows the Y_i and their corresponding $r(Y_i)$, arranged in ascending order, from which the graph in Figure 10.9 can be easily formed.

Salesman, i	4	1	8	9	7	11	13	15	6	3	14	12	10	2	5
Y_i	17	25	30	31	33	34	35	36	39	40	42	43	44	47	49
$r(Y_i)$	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15

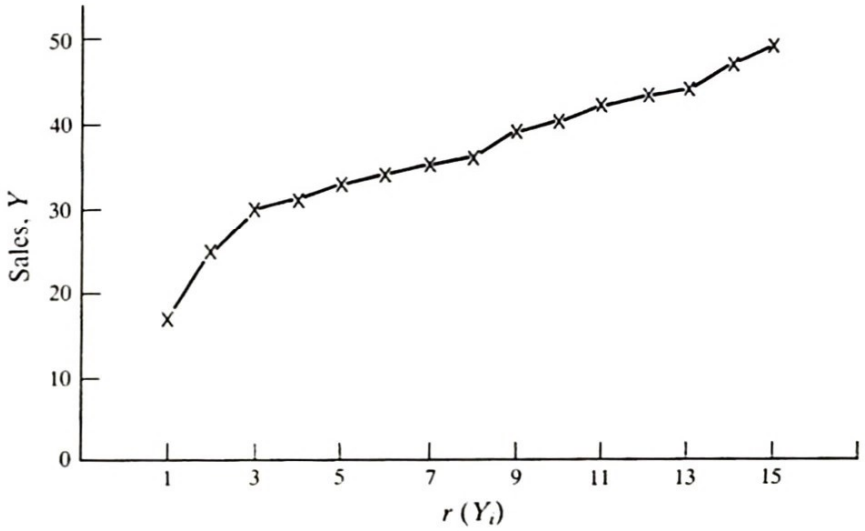


Figure 10.9

Step 3 We now establish a relationship between the original X values and the predicted average Y values (under the assumption of monotonicity) as follows:

(a) Relate each X value to its rank $r(X)$.
(b) For each $r(X)=1,2,3, \ldots ,n$ ($n=15$ here), compute $r(Y)$ from the rank regression equation.
(c) For each $r(Y)$ thus found, use Figure 10.9 to obtain the corresponding estimated average Y value.

The numerical details are as follows, where the points have been rearranged for convenience according to ascending X values.

X	$r(X)$	$r(Y)$	Y
96	1	1.90	24.20
104	2	2.77	28.86
113	3	3.64	30.64
115	4	4.51	32.02
126	5	5.39	33.39
135	6	7.13	35.13
137	7	7.13	35.13
142	8	8.00	36.00
143	9	8.87	38.61
145	10	9.74	39.74
149	11	10.61	41.22
151	12	11.49	42.49
156	13	12.36	43.36
161	14	13.23	44.69
175	15	14.10	47.20

Step 4 A graph, shown in Figure 10.10, is now plotted of these predicted average Y values against the X values. The figure also shows the original data points. This graph is the required estimated monotonic regression relationship between Y and X.

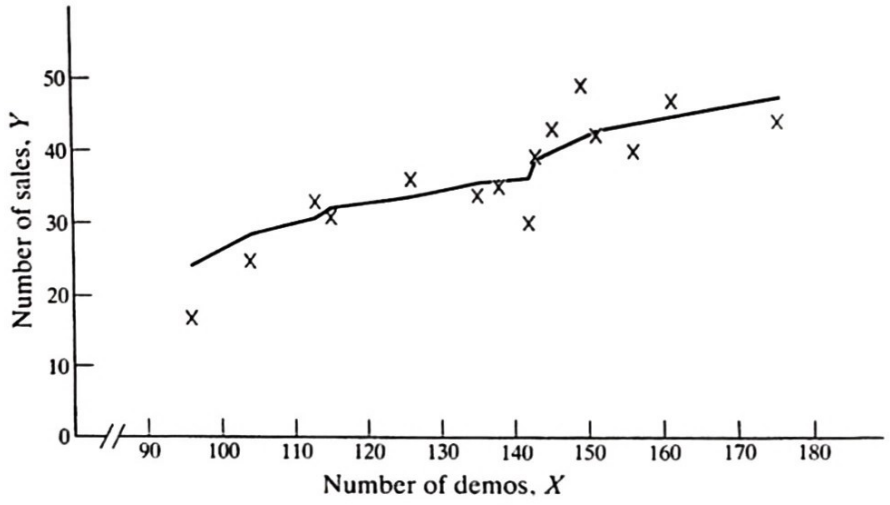


Figure 10.10

Thus, for example, the estimated average sales Y for a salesman carrying out 120 demos can be read off Figure 10.10 as about 32.6. Note that estimates of Y are *only* available for 'X's within the range of X values contained in the data; extrapolation outside this range is not possible. Even with other methods, which do not have this restriction, extrapolation is often a very dangerous exercise since there is no evidence available in the data that the model being assumed is valid outside this range. In a sense, therefore, this restriction might be a blessing in disguise!