Discussion

As stated earlier, this regression method is tackling a somewhat ambitious task in estimating a completely general monotonic relationship, especially with only 15 data points. With rather more data, and especially if the X values are more equally spaced than here (which is often the case), the method can be very powerful and accurate. Indeed, even in this illustration the result has much to commend it. A particularly interesting feature here is the abrupt change at around $X=142$ which, incidentally, indicates that a linear model would probably be invalid. This might indicate an actual change in the underlying relationship, which perhaps could be caused by the existence of two different types or groups of salesmen in the data. Of course, the effect might be spurious, but the method does at least enable such possibilities to be recognized. The best way to further the investigation is, if possible, to obtain more data and see whether or not the phenomenon is reinforced.

As a general comment, the ability of this method to work with any monotonic relationship is both unusual and useful. It is perhaps unfortunate that the final graph used for predicting Y from X is not more smooth, but that is a somewhat cosmetic problem and is a virtually necessary consequence of the wide generality of the method. Hand calculation becomes more tedious if n is at all large, but the procedure is easily computerized (see Program 10.2).

Theory

The theory of the Iman–Conover method involves several ideas from standard least-squares regression. So, if you are not familiar with this, you might be well advised to give this section a miss!

Recall that r_S may be found by calculating the ordinary linear correlation coefficient r from the *ranks* of the observations. Similarly, the rank regression equation is obtained by applying the standard least-squares principle to the *ranks*. When predicting Y from the simple linear equation $Y=A+BX$, the least-squares theory produces the coefficients:

$$B = \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sum X_i^2 - n\bar{X}^2}$$

$$A = \bar{Y} - B\bar{X}$$

where $\bar{X}$ and $\bar{Y}$ are the sample means $\sum X_i/n$ and $\sum Y_i/n$ respectively, and all the sums $\sum$ run from 1 to n. Using the ranks $r(X_i)$ and $r(Y_i)$ therefore, we predict the average $r(Y)$ for a given $r(X)$ by $r(Y) = a + br(X)$, where

$$b = \frac{\sum r(X_i)r(Y_i) - n\overline{r(X)}\,\overline{r(Y)}}{\sum r(X_i)^2 - n\overline{r(X)}^2}$$

$$a = \overline{r(Y)} - b\overline{r(X)}$$

In Chapter 9 we saw that $\overline{r(X)}=\overline{r(Y)}=\frac{1}{2}(n+1)$ and $\sum r(X)^2=\sum r(Y)^2 = n(n+1)(2n+1)/6$. In particular, since $\overline{r(X)}=\overline{r(Y)}$ and $\sum r(X)^2=\sum r(Y)^2$, it follows

that $b=r_S$; and then immediately $a=\frac{1}{2}(n+1)(1-r_S)$, from which the rank regression equation follows.

An obvious consequence is that if the data do not contain convincing evidence of a monotonic relationship, the slope b will be small; this is in complete analogy to the way the slope in standard regression analysis is proportional to Pearson's linear correlation coefficient, r. On the other hand, if $r_S=+1$ or -1, i.e. the data are completely monotonic, this method then reduces simply to plotting and joining the actual (X_i,Y_i); this is analogous to the fact that if $r=+1$ or -1 (so that the data points all lie on a line) then this line *is* the regression line. In this exceptional situation, our distribution-free method also produces this very same line.

The motivation for the procedure lies in the fact that, if the two variables (X,Y) have a completely monotonic relationship, their ranks $(r(X),r(Y))$ will also have a linear relationship with $r(X)=r(Y)$ or $r(X)+r(Y)=n+1$ in the case of increasing or decreasing relationships respectively. Also, scatter around the monotonic pattern results in corresponding scatter in the ranks. It is common practice in traditional regression methods to deal with, for example, a double-logarithmic model $\log Y=a+b \log X$ by transforming to new variables $W=\log Y$ and $Z=\log X$ in order to give the linear prediction model $W=a+bZ$, and then using the standard least-squares formulae on these new variables. Having obtained the prediction equation, one then simply transforms back to the original variables X and Y using $X=e^Z$ and $Y=e^W$. Here we are similarly using ranks for a transformation to linearity. Again, as in the logarithmic example, having carried out the linear regression on the transformed variables (the ranks), we finally transform back to the original (X,Y) variables. This is all that the procedure of this section amounts to.

Assignments

1 Wind chill

In 1975, J. J. Wiorkowski ('The wind chill index – a case study in table sleuthing', *Am. Statist.* **29**, 154–7) investigated the relationship between wind speed and the chilling effect of wind. The table opposite below presents the National Oceanic and Atmospheric Administration's wind chill equivalent temperatures.

The wind chill equivalent temperature attempts to combine the joint cooling effects of actual temperature (°F) and wind speed (mph) into one 'equivalent' temperature, which indicates the actual cooling effect on human flesh. For example, if the actual temperature is 20°F and the wind speed is 15 mph, then the human body would 'sense' a temperature of −6°F.

Task: For each wind speed (5 to 50 mph), obtain the equation of the regression line of chill index (Y) on actual temperature (X).

2 Weights and heights

The table gives the weights (kg) and heights (cm) of 42 male students who enrolled on a business studies course in a certain college in 1978.

Wind speed (mph)	Wind chill equivalent temperatures (°F)																
calm	35	30	25	20	15	10	5	0	−5	−10	−15	−20	−25	−30	−35	−40	−45
5	33	27	21	16	12	7	1	−6	−11	−15	−20	−26	−31	−35	−41	−47	−54
10	21	16	9	2	−2	−9	−15	−22	−27	−31	−38	−45	−52	−58	−64	−70	−77
15	16	11	1	−6	−11	−18	−25	−33	−40	−45	−51	−60	−65	−70	−78	−85	−90
20	12	3	−4	−9	−17	−24	−32	−40	−46	−52	−60	−68	−76	−81	−88	−96	−103
25	7	0	−7	−15	−22	−29	−37	−45	−52	−58	−67	−75	−83	−89	−96	−104	−112
30	5	−2	−11	−18	−26	−33	−41	−49	−56	−63	−70	−78	−87	−94	−101	−109	−117
35	3	−4	−13	−20	−27	−35	−43	−52	−60	−67	−72	−83	−90	−98	−105	−113	−123
40	1	−4	−15	−22	−29	−36	−45	−54	−62	−69	−76	−87	−94	−101	−107	−116	−128
45	1	−6	−17	−24	−31	−38	−46	−54	−63	−70	−78	−87	−94	−101	−108	−118	−128
50	0	−7	−17	−24	−31	−38	−47	−56	−63	−70	−79	−88	−96	−103	−110	−120	−128

Source: 1973 Associated Press Almanac, p. 171.

Weight	Height	Weight	Height	Weight	Height
74.8	189.0	70.0	188.0	63.6	181.0
61.4	170.0	73.2	188.0	73.6	178.0
70.0	189.0	70.0	180.0	76.4	177.0
68.0	182.0	63.6	178.0	60.3	179.5
108.0	180.0	63.6	178.0	60.3	172.0
67.0	187.0	66.8	175.0	60.3	176.0
70.0	180.0	79.5	185.0	70.0	167.0
70.0	173.0	79.5	182.0	66.7	177.0
90.9	170.0	60.0	170.0	70.0	177.0
66.8	183.0	73.6	167.0	66.2	165.0
89.0	188.0	71.6	178.0	63.2	179.0
58.6	179.0	67.3	178.0	62.0	167.0
66.8	174.0	63.6	171.0	72.0	184.0
68.4	187.5	65.5	172.0	70.0	176.0

Task: Use Theil's incomplete method to obtain a regression line for predicting the weight of a student from his height.

3 Vehicle tyres

A laboratory test of vehicle tyres is designed to investigate the effect of speed on the heat produced in the tyre material. Tyres are run on a test rig at chosen speeds under controlled conditions of load, inflation pressure and ambient temperature. The table below gives the average temperature (°C) at speeds ranging from 10 to 70 mph.

Speed	10	15	20	25	30	35	40	45	50	55	60	65	70
Temp	46	53	56	68	65	79	77	85	107	106	124	131	137

Task: Fit a monotonic regression relationship to these results and hence predict the temperature at speeds of (a) 28 mph, (b) 38 mph and (c) 68 mph.

Program 10.1 Theil's incomplete regression method

```
100 REM *** THEIL'S INCOMPLETE REGRESSION METHOD ***

110 REM ** INPUT DATA **
120 INPUT "NUMBER OF DATA-PAIRS: ";ND
130 LET NL=INT(ND/2)
140 IF 2*NL-ND<>0 THEN L=NL+1
150 IF 2*NL-ND=0 THEN L=NL
160 DIM X(ND),Y(ND),A(ND),D(ND)
170 PRINT "INPUT THE DATA, X-VALUE THEN Y-VALUE"
180 FOR I=1 TO ND
190 INPUT X(I),Y(I)
200 NEXT I
```

```
210 REM ** SORT THE X-VALUES TAKING THE Y'S **
220 LET N=ND
230 GO SUB 1000

240 REM ** CALCULATE THE GRADIENTS **
250 FOR I=1 TO NL
260 LET D(I)=(Y(L+I)-Y(I))/(X(L+I)-X(I))
270 NEXT I

280 REM ** CALCULATE MEDIAN OF THE GRADIENTS **
290 LET N=NL
300 GO SUB 2000
310 LET J1=INT(NL/2)+1
320 LET J2=NL-INT(NL/2)
330 LET M=(D(J1)+D(J2))/2

340 REM ** CALCULATION OF THE INTERCEPT **
350 FOR I=1 TO ND
360 LET D(I)=Y(I)-M*X(I)
370 NEXT I
380 LET N=ND
390 GO SUB 2000
400 LET J1=INT(ND/2)+1
410 LET J2=ND-INT(ND/2)
420 LET C=(D(J1)+D(J2))/2

430 REM ** OUTPUT OF THE REGRESSION LINE **
440 PRINT "GRADIENT = ";M
450 PRINT "INTERCEPT = ";C
460 STOP

1000 REM ** SORT THE X-VALUES WITH THE Y'S: USE LINES 1000-1190 OF **
1010 REM ** PROGRAM 4.2, REPLACING D(*) WITH X(*) AND F(*) WITH YO(*) **

2000 REM ** SORT **
2010 LET N1=N
2020 LET N1=INT(N1/2)
2030 IF N1=0 THEN GO TO 2160
2040 FOR I=1 TO N-N1
2050 FOR J=I TO 1 STEP -N1
2060 LET J1=J+N1
2070 IF D(J)>D(J1) THEN GO TO 2100
2080 LET J=0
2090 GO TO 2130

2100 LET T=D(J)
2110 LET D(J)=D(J1)
2120 LET D(J1)=T
2130 NEXT J
2140 NEXT I
2150 GO TO 2020
2160 RETURN
```

Program 10.2 The Iman–Conover monotonic regression method

```
100 REM ** IMAN-CONOVER MONOTONIC REGRESSION METHOD **

110 REM ** INPUT THE DATA **
120 INPUT "THE NUMBER OF PAIRS ";N
130 DIM D(N),X(N),Y(N),RX(N),YP(N),PRY(N),YO(N)
140 PRINT "INPUT THE DATA, X-VALUE THEN Y-VALUE "
150 FOR I=1 TO N
160 INPUT X(I),Y(I)
170 LET YO(I)=Y(I)
180 NEXT I

190 REM ** RANK SAMPLE **
200 FOR I=1 TO N
210 LET T=0
220 LET S=0
230 LET C=X(I)
235 FOR J=1 TO N
240 NEXT I
250 IF X(J)>C THEN GO TO 300
260 IF X(J)<C THEN GO TO 290
270 LET S=S+1
280 GO TO 300
290 LET T=T+1
300 NEXT J
310 LET RX(I)=T+(S+1)/2
320 NEXT I
330 REM ** END OF RANKING **

340 REM ** SORT THE X-VALUES WITH THE Y'S **
350 GO SUB 1000

360 REM ** SORT THE Y-VALUES **
370 FOR I=1 TO N
380 LET D(I)=Y(I)390 NEXT I
400 GO SUB 2000
410 FOR I=1 TO N
420 LET Y(I)=D(I)
430 NEXT I

440 REM ** CALCULATE RANK REGRESSION EQUATION **
450 INPUT "SPEARMAN'S COEFFICIENT: ";RS
460 FOR I=1 TO N
470 LET PRY(I)=RS*I+(N+1)*(1-RS)/2
480 NEXT I

490 REM ** SORT THE PREDICTED Y-RANKS **
500 FOR I=1 TO N
510 LET D(I)=PRY(I)
520 NEXT I
530 GO SUB 2000
540 FOR I=1 TO N
550 LET PRY(I)=D(I)
560 NEXT I

570 REM ** OUTPUT OF PREDICTED Y FOR THE GIVEN X-VALUES **
580 FOR I=1 TO N
590 LET L=INT(PRY(I))
600 LET U=L+1
```

```
610 IF RS>0 THEN LET YP(I)=Y(L)+(Y(U)-Y(L))*(PRY(I)-L)
620 IF RS<0 THEN LET YP(N-I+1)=Y(L)+(Y(U)-Y(L))*(PRY(I)-L)
630 NEXT I
640 PRINT "X","Y","PRED. Y"
650 FOR I=1 TO N
660 PRINT X(I),YO(I),YP(I)
670 NEXT I

680 REM ** PREDICTED VALUE OF Y FOR A GIVEN X-VALUE **
690 INPUT "DO YOU WISH TO PREDICT A Y-VALUE? ";A$
700 IF A$="N" THEN STOP
710 INPUT "VALUE OF X: ";NX
720 FOR I=1 TO N
730 IF NX>X(I) THEN GO TO 750
740 GO TO 760
750 NEXT I
760 LET C=I-1
770 LET PY=YP(C)+(YP(C+1)-YP(C))*(NX-X(C))/(X(C+1)-X(C))
780 PRINT "PREDICTED VALUE OF Y GIVEN X=";NX;" IS ";PY
790 GO TO 690
800 STOP

1000 REM ** SORT THE X-VALUES : USE LINES 1000-1140 OF PROGRAM 4.2 **

2000 REM ** SORT - USE LINES 2000-2160 OF PROGRAM 10.1 **
```

11

Quick methods for correlation and regression

Introduction

All the methods presented in the previous two chapters involve fairly substantial computations, and become tedious to execute with more than moderate-sized samples. In this chapter we present alternative methods for both correlation and regression, which are rather quicker and easier.

We begin the chapter with a test for correlation, which is another example of a 'Tukey-type' test in the sense that it only considers observations in the extremes of the samples and completely ignores the rest. This is followed by a regression method, which was introduced by Brown and Mood (1951). It is essentially a graphical procedure that produces a good prediction line very quickly. We complete the chapter with 'quick' tests for the slope and intercept of the regression line, which again use Tukey-type ideas.

THE OLMSTEAD–TUKEY 'CORNER' TEST

Introduction

This very interesting test was published by Olmstead and Tukey (1947) and, in our opinion, deserves wider usage than it has so far enjoyed. Unless the sample size n is very small, it is quicker to carry out than Spearman's or Kendall's tests of Chapter 9. Also, it has reasonable power properties, often quite comparable with those of previous tests. Furthermore, it has critical values that are largely independent of n; thus it is easy to memorize sufficient information to permit the test usually to be carried out without needing full tables of critical values (as is the case with Tukey's quick test for location differences given in Chapter 5).

PRACTICAL PROBLEM, HYPOTHESES AND DATA

We return to studying the relationship between sales and simple visits and demos used in Chapter 9. For the moment, we concentrate on the connection between the numbers of simple visits and sales. Recall from

Chapter 9 that the alternative hypothesis was of positive correlation between these two quantities.

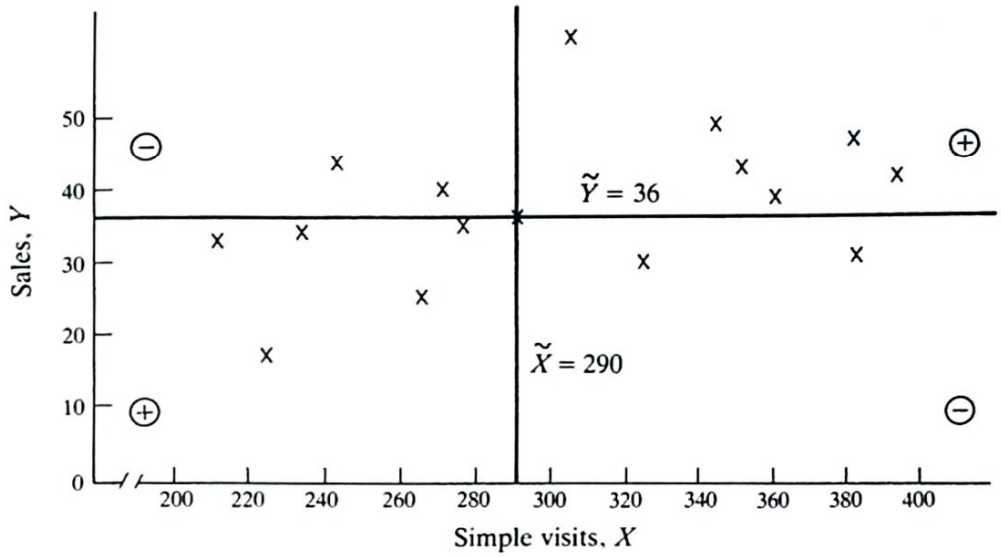


Figure 11.1

STATISTIC

It is easiest to introduce the test statistic with reference to a scatterplot, although in practice the diagram may not be necessary.

After drawing the scatterplot of the data, we compute the medians $\tilde{X}$ and $\tilde{Y}$ of the X_i and Y_i respectively. If n is large, this can be the most time-consuming part of the whole operation, and so we discuss later in this chapter how to speed up the computation of the medians. Once $\tilde{X}$ and $\tilde{Y}$ have been found, we draw the lines $X=\tilde{X}$ and $Y=\tilde{Y}$ on the diagram. (It may or may not happen that the point of intersection $(\tilde{X},\tilde{Y})$ is one of the data points; in the current case, this does happen.) These are, of course, vertical and horizontal lines respectively, which divide the picture into four regions called *quadrants*. Figure 11.1 shows the scatterplot together with these median lines.

The top-right and bottom-left quadrants are termed *positive* (+) quadrants and the other two are called *negative* (−) quadrants. The quadrants are referred to in this way because, if there is positive correlation between the X and Y variables, there will tend to be more points in the + quadrants than in the − quadrants, and vice versa. This follows since positive correlation implies that relatively large values of X and Y will occur together and also that relatively small values of X and Y will occur together; and such points will lie respectively in the top-right and bottom-left quadrants, i.e. the two + quadrants. Similarly, negative correlation increases the proportion of points lying in the negative

quadrants. On the other hand, if H_0 is true, every point should have equal probabilities of falling in a + or − quadrant. Therefore a suitable statistic for testing H_0 against an H_1 of positive correlation, negative correlation, or either, should reflect whether there is a tendency for the positive or negative quadrants to predominate, as opposed to the points being fairly equally divided between them.

Therefore, one obvious statistic would be a count of the number of points in the positive or negative quadrants, thus producing a form of sign test for correlation. However, the Olmstead–Tukey test presented here is quicker to carry out, particularly for large samples, and seems to have superior power properties over this sign test.

To develop the ideas of the Olmstead–Tukey test, consider a vertical line approaching the diagram from the right-hand side (Fig. 11.2a). Note that the first point that is 'hit' by this line is in the top-right quadrants (point A in Fig. 11.2b), i.e. in a + quadrant. Let the vertical line continue moving to the left, noting any further '+' points hit (if any), and stopping as soon as a '−' point is hit. In our example, the next point (B) hit is in the − quadrant, so the line is stopped straightaway at this point (Fig. 11.2c). We record the number, C_1 say, of '+' points crossed by the line; in this example, $C_1=1$.

If, instead, the line had first met a '−' point, we would then have continued its progress until it was stopped by a '+' point; C_1 would then have been given by *minus* the number of '−' points encountered by the line before it is stopped by the '+' point.

We now repeat the operation, first with a vertical line approaching from the *left*, then with a *horizontal* line approaching from *above* the diagram, and finally with a horizontal line approaching from *below*. Let the counts obtained in these three cases be C_2, $C3$ and C_4 respectively. (In practice, the order in which the four operations are carried out is unimportant.) In our example, a '+' point is encountered first in each one of the four operations, so all of C_1, C_2, C_3 and C_4 are positive. But *do not forget* that when and if a '−' point is the first point hit, the corresponding count must be taken as *negative*.

The test statistic C is given by the sum of these counts:

$$C = C_1 + C_2 + C_3 + C_4$$

Minor difficulties may occur if the data contain ties; as usual, we discuss this possibility in a later section. An associated difficulty may occur if n is an *odd* number. In this case there is bound to be a point lying on the median line $X=\tilde{X}$, and similarly there must be a point lying on $Y=\tilde{Y}$. Sometimes, as in the current example, these may be the *same* point, i.e. $(\tilde{X},\tilde{Y})$, which is the intersection of the two median lines, though more usually they will be two separate points. The easiest thing to do with such points as regards the calculation of C is to ignore their

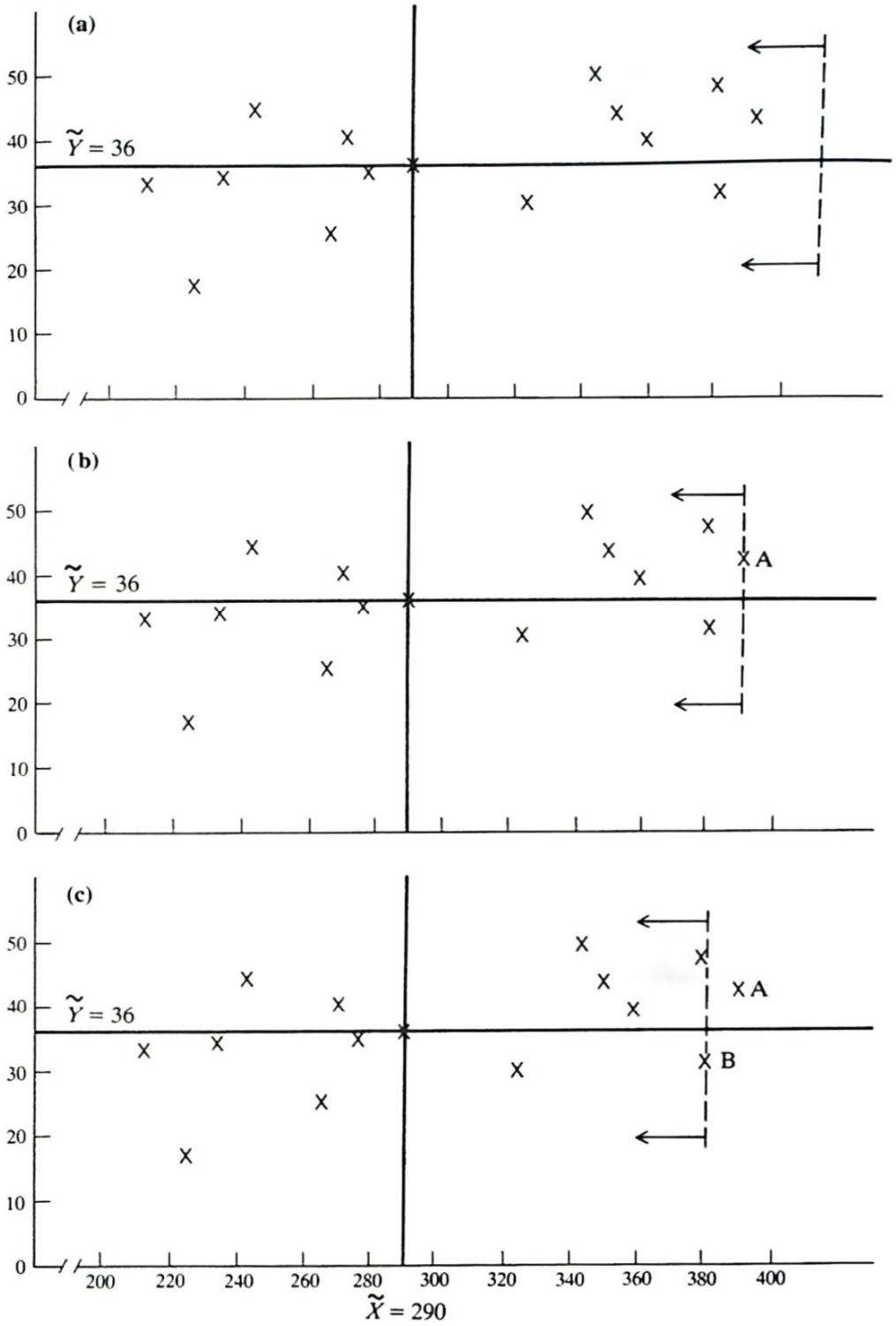

(a)
$\tilde{Y} = 36$
(b)
$\tilde{Y} = 36$
A
(c)
$\tilde{Y} = 36$
A
B
0
10
20
30
40
50
200
220
240
260
280
300
320
340
360
380
400
$\tilde{X} = 290$

Figure 11.2

existence completely. (Olmstead and Tukey's advice was more complicated. Denoting the two points as $(\bar{X}, Y_0)$ and $(X_0, \bar{Y})$, they also required that these points be deleted from consideration but that they be replaced by the artificial single point (X_0, Y_0).)

As a final detail, in cases of strong correlation it is possible for all the points to lie in the two positive or two negative quadrants, so that there is never anything to stop our moving horizontal and vertical lines according to the above rules. In this case, we simply define C as equal to plus or minus (as appropriate) twice the number of points in the diagram (again ignoring any points on the median lines).

CRITICAL REGIONS

The nature of the critical regions is clear. The more points occurring in the + quadrants, the more support there is for positive correlation. In such a case, most or all of the four counts C_1, C_2, C_3 and C_4 will generally be positive and possibly quite large, leading to relatively large positive values of the statistic C. Therefore, critical regions when testing for positive correlation are of the form:

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

One of the particularly nice features of this test is that critical values are virtually independent of the sample size. It is sufficient to provide only the following very brief table of critical values.

α_1	5%	2½%	1%	½%
α_2	10%	5%	2%	1%
Critical value	9	11	13	15 or 14
Valid for	$n \geqslant 6$	$n \geqslant 8$	$n \geqslant 10$	$10 \leqslant n \leqslant 47$ or $n \geqslant 48$

In this table, n represents the number of points available to contribute to the counts, i.e. it is either equal to or slightly less than the total number of points in the scatterplot. (Points excluded are those on the median lines $X = \bar{X}$ and/or $Y = \bar{Y}$; there will be one or more such points if the sample size is odd and/or there are ties at the median values.) It is worth emphasizing that this table is not just a rough approximation to the critical values – it is completely accurate except in a tiny minority of cases; and in those rare cases, the significance levels corresponding to the critical regions indicated in the table are so close to the nominal levels that the discrepancies are of no practical importance. However, note that, if n is very small, suitable critical regions may not exist (see the final row of the table).

It follows from the table that, since our practical problem involves a one-sided test, the 5% and 1% critical regions are $C \geqslant 9$ and $C \geqslant 13$

respectively (n is equal to 14 in this example, and so there are no validity problems).

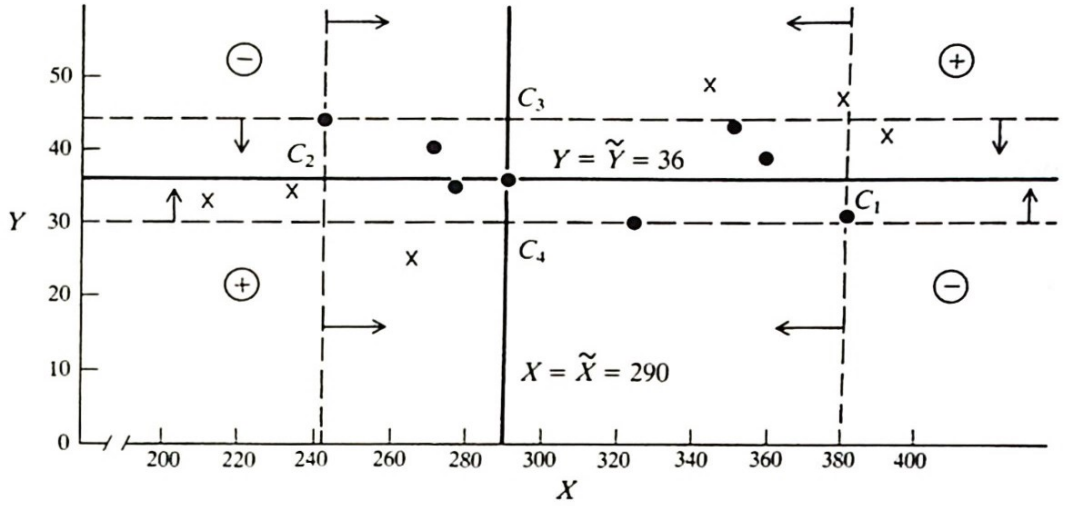


Figure 11.3

SOLUTION

Figure 11.3 shows the scatterplot again, but this time with the finishing positions of the two vertical and horizontal lines. To clarify matters, only those points which contribute to one or more of the counts C_1, C_2, C_3 and C_4 are indicated by a '×'; the others are indicated by dots. We have $C_1=+1$, $C_2=+3$, $C_3=+2$ and $C_4=+2$, giving $C=1+3+2+2=8$. This point is just outside the $\alpha_1=5\%$ critical region $C\geqslant 9$, which is wholly consistent with the results obtained from the other correlation tests in Chapter 9. Note that it is quite usual for some points to contribute to two of the four counts.

Exercise Carry out the corner test using the number of demos as the X variable. You will find the more usual situation here concerning points on the median lines, i.e. there are two different points involved. Also, you will find that, except for the two points on the median lines, all the remaining 13 points are in the positive quadrants; thus the statistic C, by virtue of the special rule for this extreme case, takes the value $2\times 13=26$, which of course is virtually conclusive evidence of there being positive correlation between the two variables.

Discussion

For a relatively quick test, the Olmstead–Tukey test has good power properties, often being very comparable with Spearman's and Kendall's tests. Of course, it is not a *very* quick test to perform, since it still involves finding the sample medians; however, as we have mentioned, a method is shown later to help with this operation. Also, when n is large, it would be quite a lengthy matter to form the

complete scatterplot; but in fact it is not necessary to do this. One could, for example, select from the data a few points (say, six each) having the largest and smallest X and Y values. If just these points are plotted, this is usually sufficient to find the values of C_1, C_2, C_3 and C_4 immediately (unless any of these counts are particularly large, in which case we will need to select a few more appropriate points from the data). Adopting this approach will generally make the calculation quite speedy, however large n happens to be.

One-sided and two-sided tests

We have been dealing with the version of the test appropriate for when H_1 stipulates positive correlation. If H_1 had stipulated negative correlation, this alternative hypothesis would be supported by large negative values of C. Therefore appropriate critical regions would then be

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

Thus for example, $C \leqslant -9$ is the appropriate critical region for the $\alpha_1 = 5\%$ significance level.

For testing against the two-sided alternative hypothesis of correlation of either type, critical regions are of the form:

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

using α_2 significance levels.

Treatment of ties and median values

The discussion has so far ignored the possibility of tied observations, except in the case of ties occurring at median values; in this special case we ignore all such points, and reduce the value of n accordingly.

The kind of tie that can cause difficulties is where one (or more) of the moving horizontal or vertical lines used to evaluate the counts C_1, C_2, C_3 and C_4 encounters two or more points simultaneously. If these points all lie in the same quadrant there is no difficulty. However, consider the first five points met by the vertical line approaching from the right in Figure 11.4. The line meets two '+' points first, and then finds three points, two above and one below the X axis, all of which have the same X coordinate. Olmstead and Tukey's suggestion for this situation was the following: In the tie, call those points in the 'correct' quadrant, i.e. the quadrant that is contributing to the count (the + quadrant in this case), 'favourable' and the others 'unfavourable'. Then the tied points are regarded as contributing

$$\frac{\text{number of favourable points}}{1+\text{number of unfavourable points}}$$

to the score (either + or − according as 'favourable' means + or − respectively).

In the above example, there are two favourable and one unfavourable, so the contribution is $2/(1+1) = 1$, giving $C_1 = 3$ altogether.

Finally, should the moving horizontal or vertical line hit both + and − points at

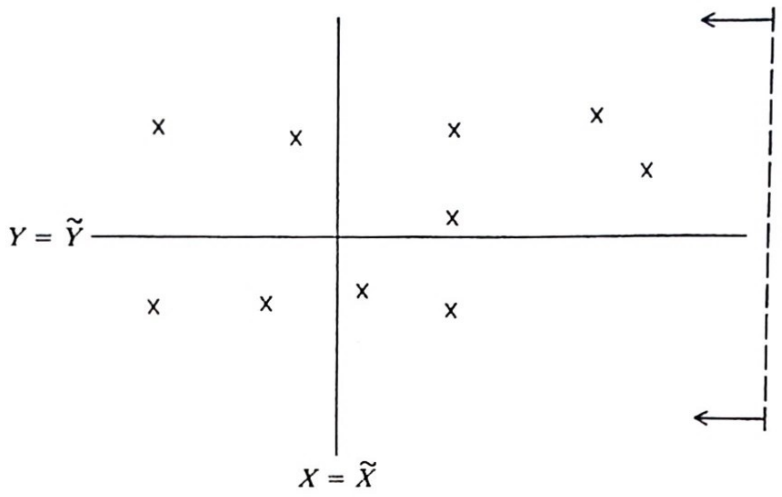


Figure 11.4

its first impact with the scatterplot (as with the vertical line approaching from the left in Figure 11.4), the corresponding count (C_2 in this case) is taken to be 0.

Null distributions

As usual, the values of the individual counts C_1, C_2, C_3 and C_4 and the overall C do not depend on the actual values of the observations (X_1,Y_1), (X_2,Y_2), . . . , (X_n,Y_n) only on their ranks $(r(X_1),r(Y_1))$, $(r(X_2),r(Y_2))$, . . . , $(r(X_n),r(Y_n))$, using the notation of the previous two chapters. Without loss of generality, we may reorder the data according to their X values so that $r(X_1)=1$, $r(X_2)=2$, . . . , $r(X_n)=n$. For the purpose of calculating null distributions we assume, as usual, that there are no ties. Then under H_0 all the $n!$ possible permutations of $r(Y_1)$, $r(Y_2)$, . . . , $r(Y_n)$ are equally likely. The null distribution may be obtained in the familiar way by writing down these $n!$ permutations, calculating C in each case, and building up a frequency distribution. As a simple example, if $n=4$, you might like to confirm that the $4!=24$ permutations give four cases of $C=+8$, one of $C=+4$, four of $C=+2$, six of $C=0$, four of $C=-2$, one of $C=-4$ and four of $C=-8$. (You will need to sketch 24 four-point scatterplots to do this!) Thus, in the one-sided case relative to the 'Practical problem', the smallest critical region is $C=+8$, but as the probability of this is $4/24=16.67\%$, there is certainly no $\alpha_1=5\%$ (or smaller) critical region, and this is as indicated by the final 'validity' row in the above table.

Large samples

The above table is appropriate for all sample sizes.

Quick calculation of medians

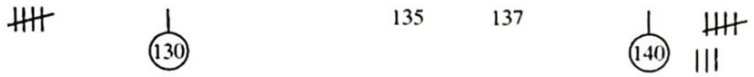
In Chapter 5, we suggested a way of speeding up the operation of ordering (ranking) the observations in a sample. In particular, this method can of course be used to find the median. However, the method does a lot more than just find the median so if the median is really all that we want, we should be able to reduce the

work involved. One obvious way of doing this is first to look briefly through the data and suggest some interval within which we are fairly sure the median lies. Then we scan through the data, ordering all those observations that lie in the interval, but merely tallying (i.e. counting) the observations lying to the left and right of the interval. Presuming the median does lie in the interval, it will now be very easy to locate it. Even if it does not, it presumably will not lie far outside the interval, and a further quick scan through the data to identify a few values just outside the interval should enable the median to be quickly found.

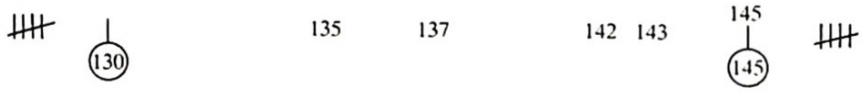
Even with only 15 points, as in our current example, this method should be noticeably quicker than completely ordering all the observations, and for large samples it produces a very considerable improvement. Let us illustrate the method by finding the median number of demos. Here again are the numbers of demos done by the 15 salesmen:

104	161	156	96	149	143	113	142
115	175	135	145	137	151	126	

A quick glance through the data shows that the number of demos ranges between 96 and 175, with no noticeable asymmetry, so it would seem reasonable to guess that the median lies between, say, 130 and 140. Scanning through the 15 values, we place numbers in the interval on a roughly linear scale, and simply tally the others. Doing this we obtain the following picture:



This shows we have just missed trapping the median value. We want the 8th (= the 8th smallest) value, and this lies just to the right of 140; for we have five values less than 130, two between 130 and 140 and eight greater than 140. So, we simply rescan the data, now looking for the smallest value larger than 140, and this is almost immediately seen to be 142. Of course, had we chosen the interval as, say (130,145), this second scan would have been unnecessary as we would then have had:



from which the median value of 142 is immediately obtained.

BROWN AND MOOD'S QUICK REGRESSION METHOD

Introduction

G. W. Brown and A. M. Mood (1951) introduced a simple but effective technique for quickly fitting a straight line $Y=a+bX$ to a set of points. Although they presented it as a numerical procedure, the technique seems best suited to a graphical approach; this makes it even faster to

carry out and reduces computation to a minimum. Also, as it is based on medians, the method is quite robust against the occurrence of 'rogue values' in the data. After presenting their regression method, we shall take a look at a simple method for testing hypotheses concerning the slope b and the intercept a.

PRACTICAL PROBLEM AND DATA

To illustrate Brown and Mood's regression method, we will use the geology problem discussed in the previous chapter. For convenience, we reproduce the data of that problem below.

Distance, X (km)	0.2	0.5	0.9	1.4	2.1	2.7	3.0	3.6
Volume, Y (cm^3)	50	48	42	36	34	32	29	31
Distance, X (km)	4.1	4.4	5.0	5.6	6.2	6.6	7.1	
Volume, Y (cm^3)	28	24	19	17	10	12	9.5	

PROCEDURE

Brown and Mood's procedure is best described by the following steps.

Step 1 Construct the scatterplot of the data.

Step 2 Draw a vertical line through $\tilde{X}$, the median of the 'X's; in our example, we have $\tilde{X}=3.6$ (see Fig. 11.5).

Step 3 Find the X median and the Y median of the points lying to the left of the median line just drawn (i.e. $\tilde{X}=3.6$, in this example); we will call these medians $\tilde{X}_1$ and $\tilde{Y}_1$. In our case, $\tilde{X}_1=1.4$ and $\tilde{Y}_1=36$.

Step 4 Now find the X median and the Y median of the points lying to the right of the median line; call them $\tilde{X}_2$ and $\tilde{Y}_2$. In our case, $\tilde{X}_2=5.6$ and $\tilde{Y}_2=17$.

(It so happens here that both $(\tilde{X}_1,\tilde{Y}_1)$ and $(\tilde{X}_2,\tilde{Y}_2)$ are actual data points. This tends to happen if there is a strong monotonic relationship in the data but, in general, this does not happen.) We can now use the line defined by $(\tilde{X}_1,\tilde{Y}_1)$ and $(\tilde{X}_2,\tilde{Y}_2)$ as a very quick and simple regression line – this is the broken line in Figure 11.6. More usually it is regarded as the *first approximation* to the final regression line.

To improve on this first approximation, we consider the *deviations* or differences between the actual Y values and those Y values predicted from the regression line (i.e. the vertical distance from the actual data point to the line). Brown and Mood reasoned that for a particularly good regression line, the *median* of these deviations in both groups of points (those lying to the left and right of the median line $X=\tilde{X}$) should be zero.

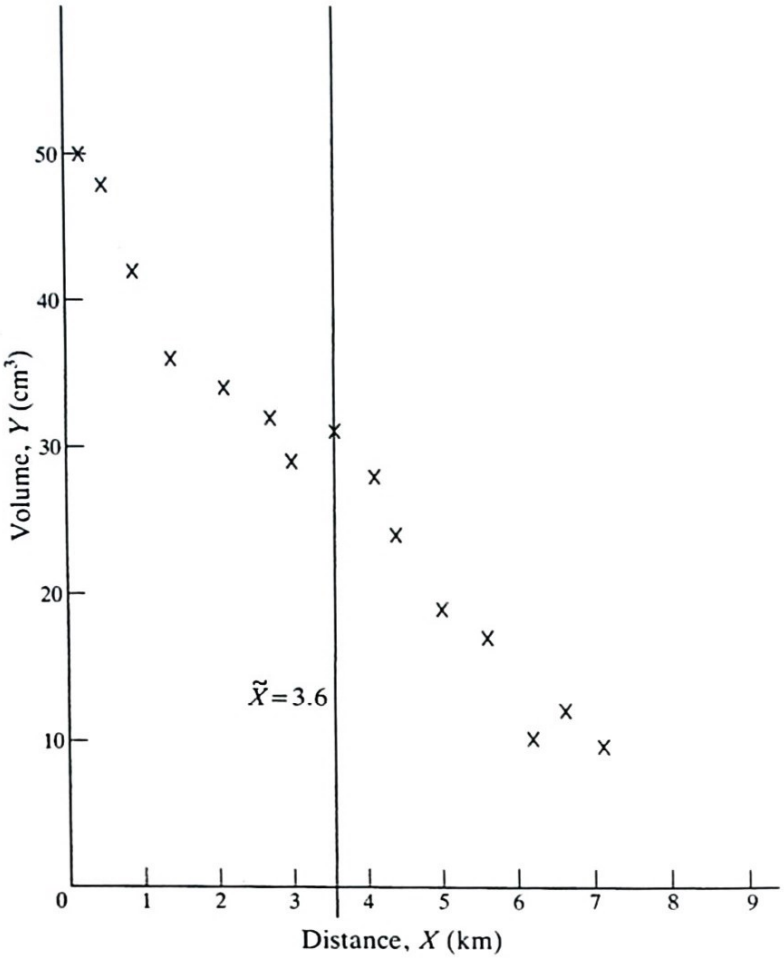


Figure 11.5

However, using the above first approximation to the regression line, we find the deviations to be:

Left-hand group	8.50	8.00	3.75	0	1.00	2.00	0.25	(median=2)
Right-hand group	4.00	1.50	−0.75	0	−4.25	−2.50	−0.75	(median=−0.75)

This confirms what we see in Figure 11.6; the first approximation is not too good for this case since, in particular, all six points in the left-hand group (apart from the one point on the line) lie *above* the line, corresponding to positive deviations.

The final version of the regression line (which is most easily obtained using a transparent ruler) is found by shifting the line around until, as suggested above, the median of the deviations in both groups is equal to 0. This needs some experimenting but, with practice, it is quickly achieved. Basically, it is a matter of ensuring that, in each group, there are as many points above the line as below; if there are an odd number of

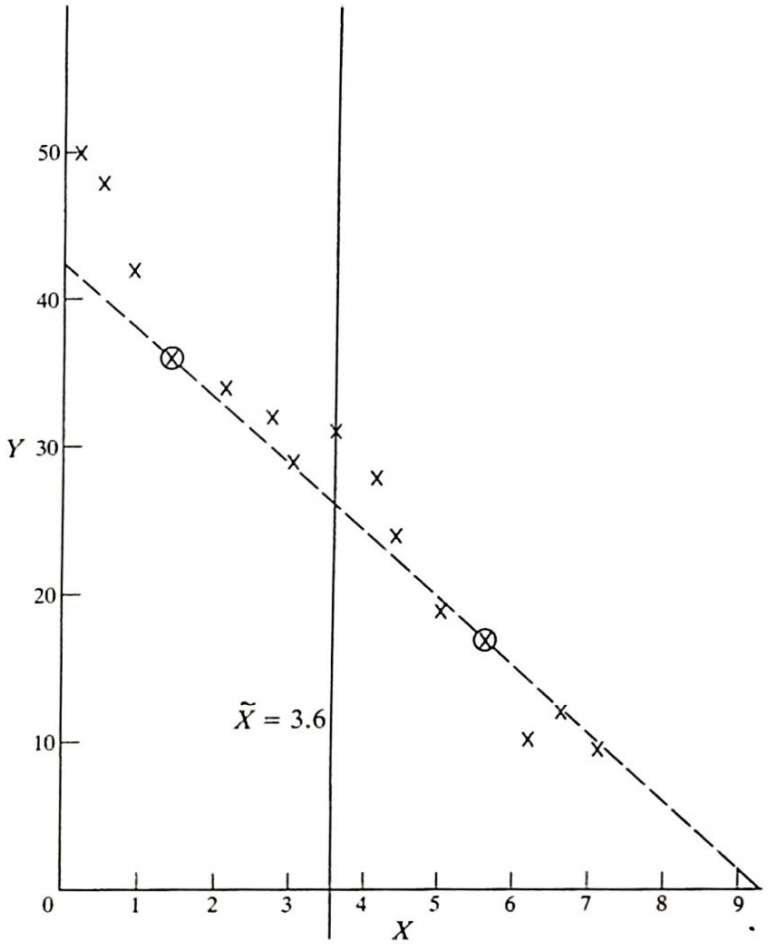


Figure 11.6

points in the group, the line will go through one of them but if the number of points is even, then a little judgement is needed to make the middle two deviations the same size (one above and one below the line).

In our case, the required line goes through the points (2.7,32) and (6.6,12), as drawn in Figure 11.7. Predictions for Y given an X value can now be directly obtained from this line.

If preferred, one can carry out a little algebra to discover that the equation of the line going through the points (2.7,32) and (6.6,12) is $Y=45.846-5.128X$. The deviations in the two groups may then be calculated as:

Left-hand group	5.18	4.75	0.77	−2.67	−1.08	0	1.46
Right-hand group	3.18	0.72	−1.21	−0.13	−4.05	0	0.06

Note that the median deviation in each group is zero, as required.

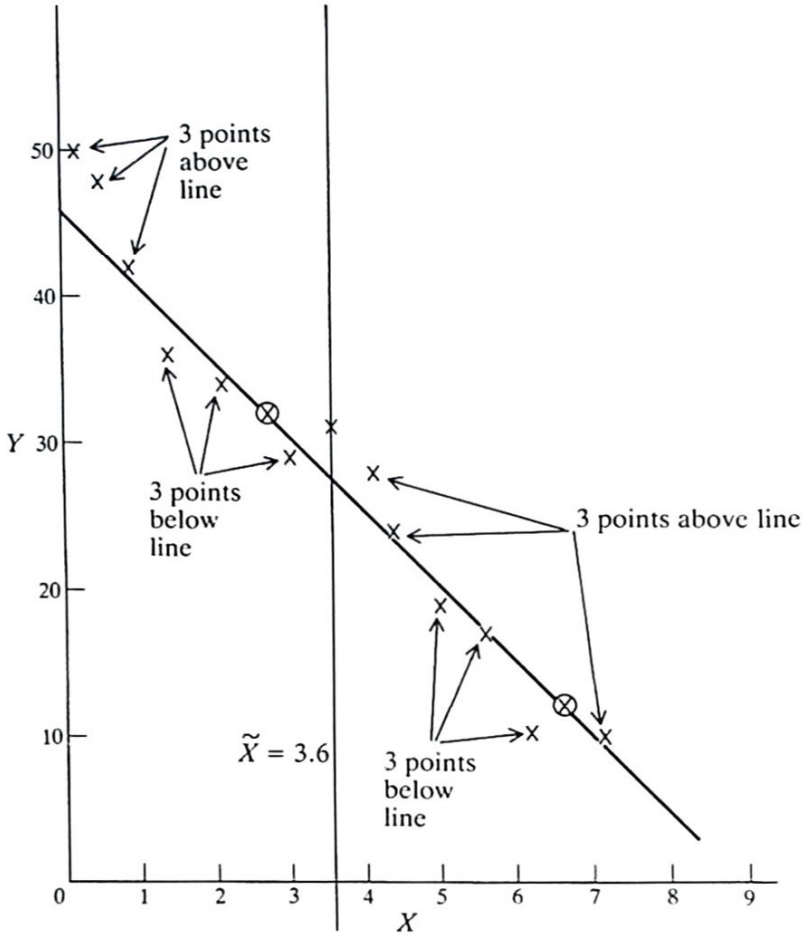


Figure 11.7

A QUICK HYPOTHESIS TEST FOR THE SLOPE

Introduction

We now present an interesting procedure for a quick and easy hypothesis test on the slope b of a regression line. A similar test may also be designed for testing a hypothesis about the intercept a and this is presented at the end of the chapter. After drawing the scatterplot, plus a simple construction on this diagram, the test is really just a simple applicaton of Tukey's quick test of Chapter 5.

PROBLEM AND DATA

We shall study a familiar situation, namely the relationship between the volumes of stones and their distances from the river source.

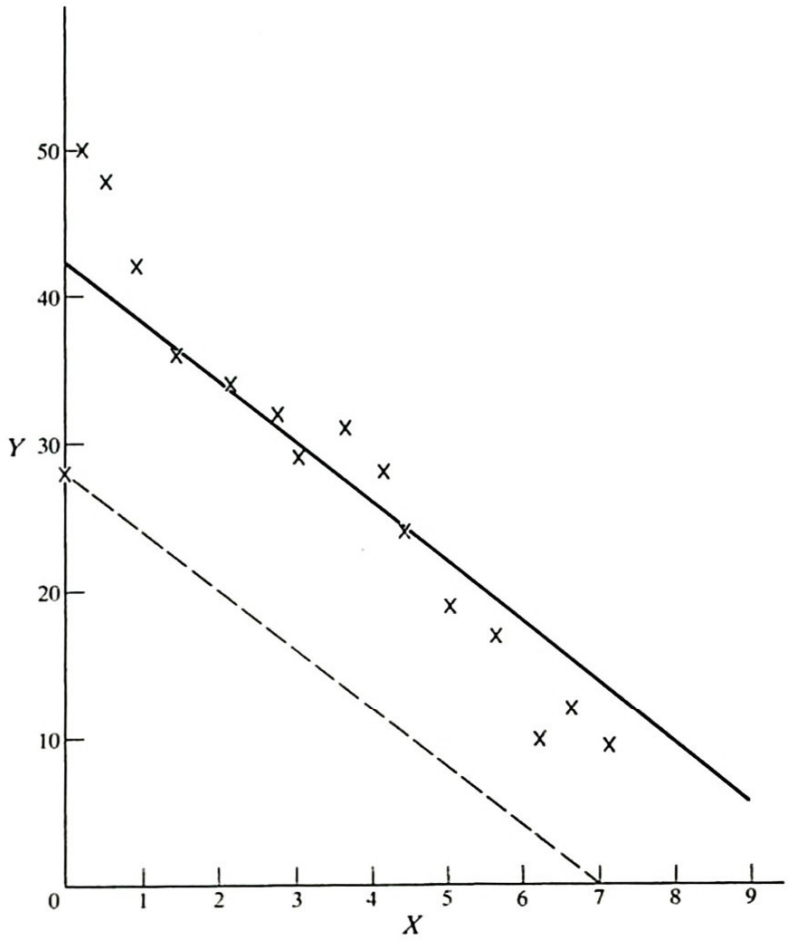


Figure 11.8

HYPOTHESES

When analysing the same problem in Chapter 10 we used the hypotheses:

H_0: $b = -4$.

H_1: $b \neq -4$.

We shall use the same hypotheses here, and it will be interesting to see how this quick test compares with the more powerful test in the previous chapter.

STATISTIC

On the scatterplot (Fig. 11.8), we first draw a 'guide line', which is any line having the hypothesized slope ($b=-4$ in this example); this is the broken line in this figure. Then, retaining this slope, we move the line up or down *until there are as many points above it as below* – this line is also

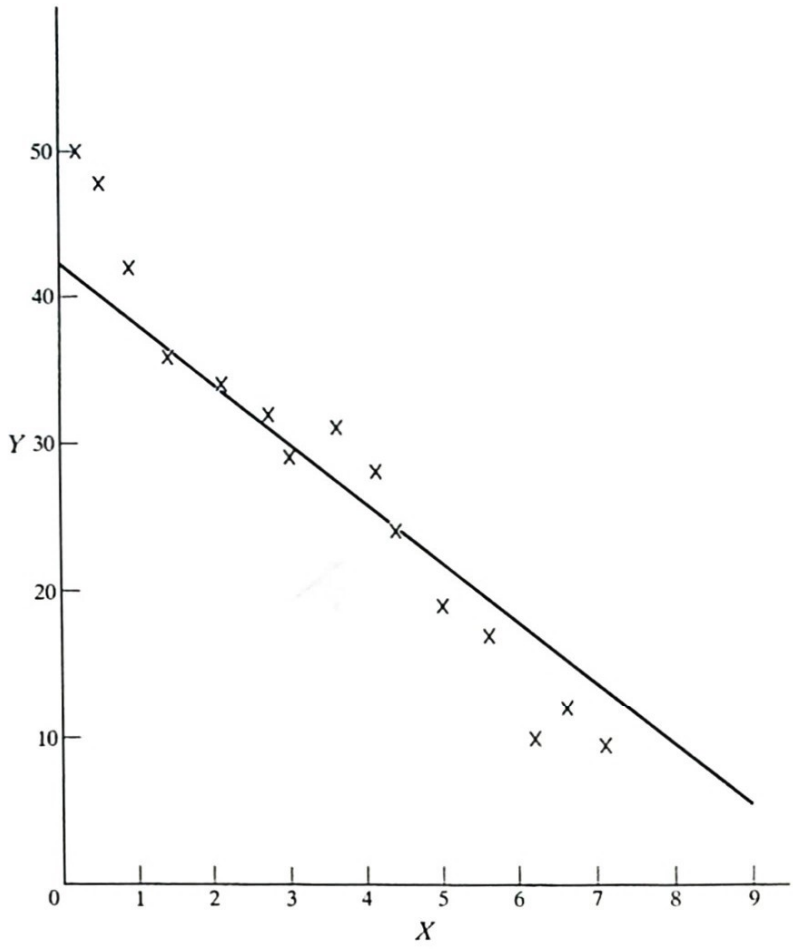


Figure 11.9

drawn in Figure 11.8. Sometimes, as *in this example*, this cannot be achieved exactly, but we get as close to equality as possible. Here, the best we can do is to have seven points above and eight points below. This line is shown in Figure 11.9. Now denote each point above the line as a '+', and each point below the line as '−'. We write down the appropriate sequence of '+'s and '−'s going from left to right in Figure 11.9. By this process, we obtain the following sequence of signs:

$$+ \quad + \quad + \quad - \quad + \quad + \quad - \quad + \quad + \quad - \quad - \quad - \quad - \quad - \quad -$$

Regarding the '+'s and '−'s as two samples, we then carry out the usual Tukey quick test.

CRITICAL REGIONS

These are exactly as referred to previously in Chapter 5, with the critical values being obtained from Table H. Thus with $n_1=7$ and $n_2=8$, we reject

H_0 in favour of H_1 at the α_2=5% and 1% significance levels if $Ty \geqslant 7$ and $Ty \geqslant 9$ respectively.

SOLUTION

There is a run of three '+'s at the left-hand end and six '−'s at the right-hand end, giving a value of the test statistic of $Ty=3+6=9$. This result therefore is just significant at the α_2=1% level.

Discussion

The result just obtained is not quite as conclusive as that in Chapter 10, but it is clearly not far behind. In view of the very quick and easy nature of this procedure, we seem to be getting good power relative to the effort required to carry out the test.

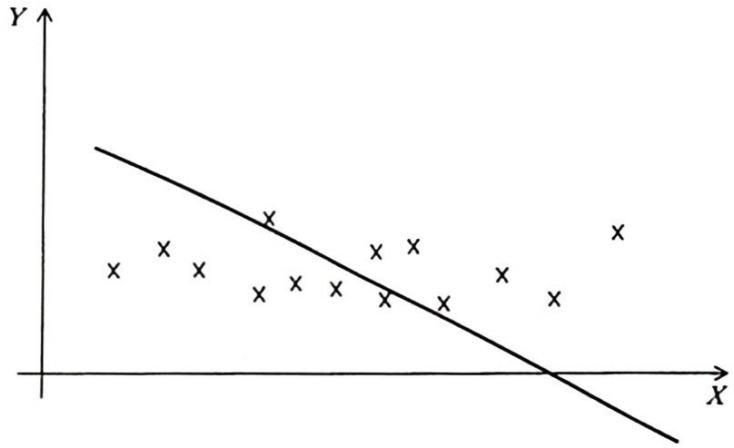


Figure 11.10

One-sided and two-sided tests

The test is based on the fact that if the hypothesized slope is seriously in error then there will be a tendency for deviations at the left side of the diagram to be negative and those on the right to be positive, or vice versa. A one-sided test, such as for H_0: $b=-4$ against H_1: $b>4$ will be based on the appropriate selection from these two alternatives. For this example, $b>-4$ implies the line is *less* steep than implied by H_0 (or even of positive slope). Imagine the situation if the true value of b were substantially greater than −4, say $b=0$. We would then obtain a diagram something like Figure 11.10, where most of the deviations on the left of the diagram are negative (−) and those on the right are positive (+).

Consequently, we would only allow the Tukey statistic to count for this one-sided test if there is a '−' at the left of the sequence and a '+' at the right. Subject to this, α_1 significance levels should be used (i.e. 7 for $\alpha_1=2\frac{1}{2}\%$, 10 for $\alpha_1=\frac{1}{2}\%$, etc., unless n is small). In the current case, the signs are the opposite way round and so, unsurprisingly, there is no evidence in our data for rejecting H_0: $b=-4$ in favour of the alternative H_1: $b>-4$; the evidence in fact points the other way.

Null distributions and large samples

There is virtually nothing to add to the information given in Chapter 5. Note that the procedure automatically provides equal, or nearly equal, numbers of '+'s and '−'s, which is the situation when the '7–10–13' rule applies best for critical values, as long as the samples sizes are not too small. So, if the number n of data points available is, say, at least 20, we also have the advantage that we do not need to refer to Table H for the critical values.

A QUICK HYPOTHESIS TEST FOR THE INTERCEPT

Introduction

We now conclude this chapter with a quick-and-easy test on the intercept ($=a$) of the regression line. This test is very similar to that just given for the slope (b), and so we really have two tests for the price of one!

PRACTICAL PROBLEM AND DATA

As before, we shall use the geologist's problem, i.e. the relationship between the volume of stones and distance from the river source.

HYPOTHESES

Using his previous experience with similar rivers, the geologist believes that, near the river source, the average volume of stones should be about 42 cm^3, i.e. the intercept a of the regression line should be 42. It is certainly of interest to him to see whether or not this is a reasonable value for this particular river, and so appropriate hypotheses are:

H_0: $a = 42$.
H_1: $a \neq 42$.

STATISTIC

On the scatterplot (Fig. 1.11), we draw a reference line through the point $(0,a)=(0,42)$ (this is shown as a broken line in the figure). We then simply rotate this line about the point (0,42) until there are an equal number of points above as below the line (as with the previous test, this process is easily done with the aid of a transparent ruler). This line is drawn in Figure 11.11; notice that it passes through the point (4.4,24) to give seven points above and seven below the line.

As with the previous test, we denote each point above the line as a '+' and each point below as a '−'. Writing down the sequence of '+'s and '−'s obtained when going from left to right in Figure 11.11 gives:

+ + + − + + − + + − − − − −

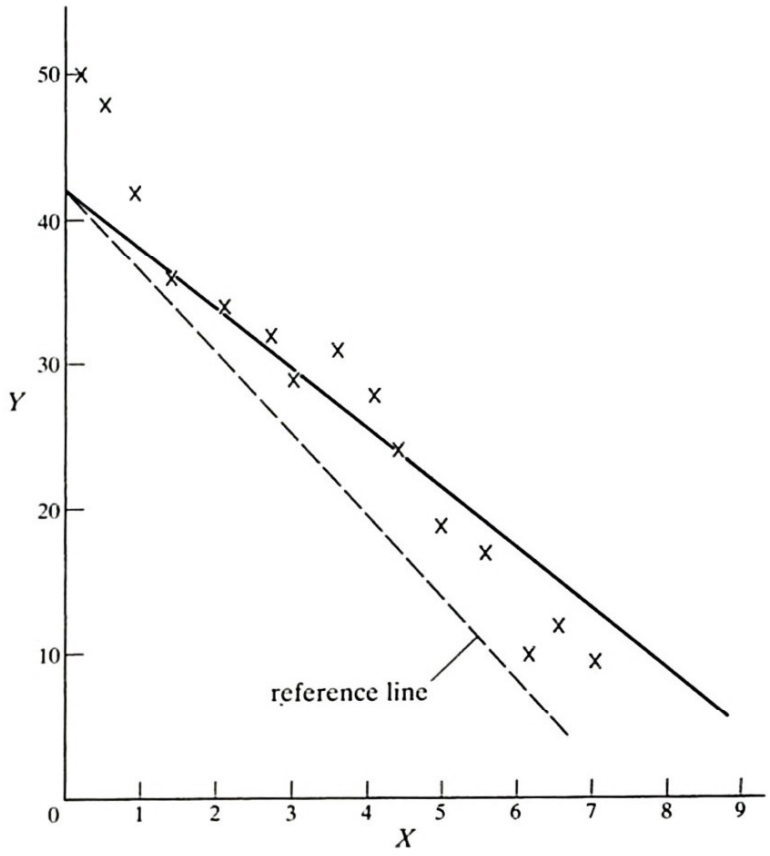


Figure 11.11

As before, we regard the '+'s and '−'s as two samples and carry out the usual Tukey two-sample test.

CRITICAL REGIONS

With $n_1=7$ and $n_2=7$, Table H gives the $\alpha_2=5\%$ and 1% critical regions as $Ty \geq 7$ and $Ty \geq 9$ respectively.

SOLUTION

There is a run of three '+'s at the left-hand end and five '−'s at the right-hand end, giving $Ty=3+5=8$. So this result is significant at the $\alpha_2=5\%$ level but not at the 1% level, i.e. we have some (but not strong) evidence in support of the alternative hypothesis that the volume of stones near (or at) the source is not equal to 42 cm^3.

One-sided and two-sided tests and large samples

The same comments apply here as with the previous test.

Assignments

In addition to the assignments of Chapters 9 and 10, we include the following for this chapter.

1 Medical data

In 1957, W. H. Gibson and G. H. Jowett (' "Three group" regression analysis. Part I. Simple regression analysis', *Appl. Statist.* **6**, 114–22) presented a regression analysis of some medical data. In fact, the data are ideal for the techniques presented in this chapter. In the table below, X is the diastolic blood pressue (mmHg) and Y is the heart weight (g) of men dying from cerebral haemorrhage.

X	121	120	95	123	140	112	92	100	102	91
Y	521	465	352	455	490	388	301	395	375	418
X	143	94	108	78	170	80	179	110	120	163
Y	567	264	503	320	458	399	524	465	432	602

Task: Use the Olmstead-Tukey method to test for positive correlation in these data. Apply Brown and Mood's regression technique to obtain a suitable regression line.

12

Association between two variables

Introduction

In this chapter we are concerned with a problem that is similar to testing for correlation, i.e. whether two variables, or quantities, are related to each other. The essential difference from our previous investigations into correlation is that the 'quantities' of interest may not be numerical; indeed, even if they are numerical then they are usually measured on a relatively crude scale so that we may expect several of the data to take on the same recorded values as each other. In this situation, we refer to 'association' between quantities rather than 'correlation'.

Problems of association occur in many disciplines, such as educational studies, social sciences, pure sciences, medical research and so on. For instance, a sociologist may be concerned as to whether the type of crime a person commits is associated with his social background; medical researchers ask 'is there any association between food colourings and stomach cancer?'; a market researcher may be interested in whether a person's sex or age is associated with preferences in packaging colours.

The most common statistical technique for analysing such problems is attributed to Karl Pearson's (1900) work around the turn of the century. He used what is essentially a chi-squared goodness-of-fit test since, as we shall see, the method involves comparing a set of observed frequencies with a set of 'expected' frequencies calculated under the null hypothesis of no association between the two quantities.

At this point, it will be useful for you to see a typical set of data on which Pearson's association test may be used. In fact, the data we give below were quoted by Pearson in 1909 and concern the drinking habits of criminals and the type of crime committed.

Offence	Drinker	Abstainer
arson	50	43
rape	88	62
violence	155	110
stealing	379	300
coining	18	14
fraud	63	144

Clearly, these data comprise 'observed frequencies'; for example, 144 people who committed fraud were abstainers. These data were collected in an attempt to answer the question 'Is the type of crime committed associated with drinking habits?'.

It is usual to refer to such a tabular layout of frequencies as an $r \times c$ *contingency table*,* where r represents the number of rows and c the number of columns in the table. Thus the above data are said to be in a 6×2 contingency table. In view of this terminology, many authors refer to the 'analysis of contingency tables' when dealing with this kind of situation.

THE CHI-SQUARED TEST

PRACTICAL PROBLEM

In the early 1980s, a household survey was conducted on the inhabitants of four regions of South Wales (Rhiubina, Cyn Coed, Splott/Adamstown and Canton/Riverside). There were many aspects to this survey, and the one we have chosen to investigate is that arising from the question 'How satisfied are you with the level of facilities for sport activities in your region?'.

As we work through the practical problem, you will see the resemblance between the techniques of this chapter and the χ^2 test of Chapter 4.

HYPOTHESES

The above question was posed to see whether people's opinion on sporting facilities depends on where they live. The alternative hypothesis is then a statement to this effect, with the null hypothesis saying that a person's opinion on sporting facilities is *independent* of where they live. Thus the hypotheses may be stated as:

H_0: Opinion on sporting facilities in South Wales is independent of the region.

H_1: There is some association between regions in South Wales and opinions on sporting facilities.

DATA

The table below shows the number of people in each of the four regions who were 'very satisfied', 'satisfied', 'not very satisfied' or 'dissatisfied' with the level of sporting facilities in their area.

* The term contingency means 'depending on chance'.

Opinion	Rhiubina	Cyn Coed	Splott	Canton
very satisfied	7	7	7	3
satisfied	18	37	24	30
not very satisfied	19	19	10	19
dissatisfied	15	8	4	21

Notice that the data form a 4×4 contingency table.

STATISTIC

As we mentioned earlier, Pearson's test for independence (no association) between two variables is essentially the same as his goodness-of-fit test; therefore, the test statistic is of the form

$$\chi^2 = \frac{\sum(\text{Ob.} - \text{Ex.})^2}{\text{Ex.}}$$

the sum being taken over all the $r \times c$ positions in the table. However, for contingency tables, the frequencies expected (Ex.) under H_0 are obtained in a different way from those for the goodness-of-fit test in Chapter 4. In order to explain how these expected frequencies are calculated, we now reproduce the data together with the *marginal* row and column totals and also the *grand* total.

Opinion	Rhiubina	Cyn Coed	Splott	Canton	Row totals
very satisfied	7	7	7	3	24
satisfied	18	37	24	30	109
not very satisfied	19	19	10	19	67
dissatisfied	15	8	4	21	48
Column totals	59	71	45	73	248

Now, what we would 'expect' of these frequencies if the null hypothesis of independence between the row and column factors were true? First, the pattern (i.e. proportions) of responses 'very satisfied', 'satisfied', 'not very satisfied' and 'dissatisfied' should be the same in each of the four regions; as in Chapter 4, we would not expect this to be *exactly* true in our sample, but we would expect it 'on the average' or 'in the long term'. Also, there is no reason for us to 'expect' different row and column totals from those in the observed data. Therefore we need to calculate expected frequencies that (i) satisfy the above rule concerning proportions, and (ii) retain the row and column marginal totals in the table.

Fortunately, there is a very simple rule which gives such expected

frequencies. To find the expected frequency for any position in the table, i.e. for any given row and column, the rule is to multiply together the corresponding row and column marginal totals and then divide by the grand total, or more concisely:

$$\text{expected frequency} = \frac{\text{row total} \times \text{column total}}{\text{grand total}}$$

This formula can be derived by the following argument. Consider the category 'very satisfied *and* in Rhiubina'. If H_0 *is true* and so 'opinion' is independent of 'region', we can calculate the probability of someone being in this category by using the multiplication law given in Chapter 1: if events A and B are independent then

$$\text{Prob(A } \textit{and} \text{ B)} = \text{Prob(A) Prob(B)}$$

so that

$$\begin{aligned}&\text{Prob(very satisfied } \textit{and} \text{ in Rhiubina)}\\ &= \text{Prob(very satisfied) Prob(in Rhiubina)}\end{aligned}$$

Now, since 24 out of the grand total of 248 are 'very satisfied' we estimate that

$$\text{Prob(very satisfied)} = 24/248$$

and, likewise,

$$\text{Prob(in Rhiubina)} = 59/248$$

Thus, if H_0 is true we have

$$\begin{aligned}\text{Prob(very satisfied } \textit{and} \text{ in Rhiubina)} &= (24/248 \times 59/248)\\ &= \frac{\text{row total}}{\text{grand total}} \times \frac{\text{column total}}{\text{grand total}}\end{aligned}$$

Hence the number 'expected' in this category if H_0 is true is found by multiplying this probability by the grand total, which gives the formula quoted above.

Below are the tables of expected frequencies, before and after the arithmetic is carried out. Before the actual calculations are performed the table is as follows:

Opinion	Rhiubina	Cyn Coed	Splott	Canton	Row totals
very satisfied	$\frac{24\times59}{248}$	$\frac{24\times71}{248}$	$\frac{24\times45}{248}$	$\frac{24\times73}{248}$	24
satisfied	$\frac{109\times59}{248}$	$\frac{109\times71}{248}$	$\frac{109\times45}{248}$	$\frac{109\times73}{248}$	109
not very satisfied	$\frac{67\times59}{248}$	$\frac{67\times71}{248}$	$\frac{67\times45}{248}$	$\frac{67\times73}{248}$	67
dissatisfied	$\frac{48\times59}{248}$	$\frac{48\times71}{248}$	$\frac{48\times45}{248}$	$\frac{48\times73}{248}$	48
Column totals	59	71	45	73	248

After the calculations are carried out, we have:

Opinion	Rhiubina	Cyn Coed	Splott	Canton	Row totals
very satisfied	5.7	6.9	4.4	7.1	24
satisfied	25.9	31.2	19.8	32.1	109
not very satisfied	15.9	19.2	12.2	19.7	67
dissatisfied	11.5	13.7	8.6	14.1	48
Column totals	59	71	45	73	248

By examining the first of these tables, you should see that the desired conditions are exactly satisfied; they are only approximately satisfied in the second table because of rounding. (Note that there is no reason to round the expected frequencies to integers; to do so would only result in the test being less accurate.)

It is worth noting that, because of the retention of marginal totals, some of the arithmetic could be shortened. For example, suppose we have computed the first three entries, 5.7, 25.9 and 15.9 in the first column; then, since we know the four entries in this column should add up to 59, we could simply calculate the final entry as

$$59 - 5.7 - 25.9 - 15.9 = 11.5.$$

We could do the same in the second and third columns. *All* of the entries in the final column could be computed by a similar 'indirect' method, using the knowledge about the row totals. The details might differ slightly in practice because of rounding errors.

The minimum number of 'direct' calculations (nine in our problem) is called the *number of degrees of freedom*. In the general case of a contingency table having r rows and c columns, there are $(r-1)(c-1)$ degrees of freedom (df).

CRITICAL REGIONS

Since large discrepancies between Ob. and Ex. support the alternative hypothesis and give a large value of χ^2, we reject H_0 in favour of H_1 if

$$\chi^2 \geqslant \text{critical value}$$

As we saw in Chapter 4, χ^2 does not have a precise null distribution – it varies according to circumstances (even for a fixed sample size). Fortunately, under certain provisos, it does not vary to any great extent, and it can be well approximated by the chi-squared distribution having $(r-1)(c-1)$ degrees of freedom.

Using this approximation, Table B gives the 5% and 1% critical regions for our problem with 9 df as $\chi^2 \geqslant 16.919$ and $\chi^2 \geqslant 21.666$ respectively.

SOLUTION

The calculation of $\chi^2 = \sum(\text{Ob.}-\text{Ex.})^2/\text{Ex.}$ is quite straightforward once the expected frequencies (Ex.) have been found. The details of the calculation are shown in the following table.

Ob.	Ex.	Ob.−Ex.	(Ob.−Ex.)²	(Ob.−Ex.)²/Ex.
7	5.7	1.3	1.69	0.30
18	25.9	−7.9	62.41	0.41
19	15.9	3.1	9.61	0.60
15	11.5	3.5	12.25	1.07
7	6.9	0.1	0.01	0.00
37	31.2	5.8	33.64	1.08
19	19.2	−0.2	0.04	0.00
8	13.7	−5.7	32.49	2.37
7	4.4	2.6	6.76	1.54
24	19.8	4.2	17.64	0.89
10	12.2	−2.2	4.84	0.40
4	8.6	−4.6	21.16	2.46
3	7.1	−4.1	16.81	2.37
30	32.1	−2.1	4.41	0.14
19	19.7	−0.7	0.49	0.02
21	14.1	6.9	47.61	3.38
			total =	19.03

The value of $\chi^2=19.03$ is certainly significant at the 5% level; in fact, consulting the more detailed tables in *EST* reveals that this value is just significant at the $2\frac{1}{2}$% level (the critical value for which is 19.023). So it seems that the data contain fairly strong evidence to support the alternative hypothesis that a person's opinion on the level of sporting facilities in South Wales does depend on where that person lives.

DISCUSSION

Since H_0 has been rejected, we might wish to examine the data further to find out more about this dependence. However, this is not an easy job! An exploratory method is simply to look at the column of values for $(\text{Ob.}-\text{Ex.})^2/\text{Ex.}$ to see which contribute most to the value of χ^2. If we do this and select those categories for which $(\text{Ob.}-\text{Ex.})^2/\text{Ex.}>2$ (say) we obtain:

Rhiubina *and* satisfied
Cyn Coed *and* dissatisfied
Splott *and* dissatisfied
Riverside *and* very satisfied
Riverside *and* dissatisfied

So, on this fairly subjective basis, the lack of independence shows up most in just these five categories. Since three of these (including the final one which corresponds to the largest value of $(\text{Ob.}-\text{Ex.})^2/\text{Ex.}$) involve the factor 'dissatisfied', in particular, we might conclude that dissatisfaction depends upon region.

A note on small expected frequencies

As with Pearson's 'chi-squared' goodness-of-fit test, small expected frequencies can invalidate the critical values obtained from the chi-squared distribution. A generally accepted guideline is the following:

(a) no more than one-fifth of the expected frequencies should be less than 5, and
(b) no expected frequency should be less than 1.

If these conditions are not satisfied then we usually combine categories to ensure that the conditions are fulfilled. In our practical problem only one expected frequency, the 4.4, was less than 5 and so we were quite justified in proceeding with the usual calculation of χ^2. However, another question on the same survey was 'How important do you think that hobbies and leisure pursuits are?'. The data resulting from this question are given below, where the figures in brackets denote the corresponding expected frequencies.

Opinion	Rhiubina	Cyn Coed	Splott	Canton	
very important	36 (29.2)	41 (37.0)	17 (29.2)	37 (35.6)	131
important	18 (19.9)	27 (25.1)	24 (19.9)	20 (24.1)	89
quite important	4 (8.7)	8 (11.0)	15 (8.7)	12 (10.6)	39
not very important	2 (2.2)	0 (2.9)	4 (2.2)	4 (2.7)	10
	60	76	60	73	269

In this case, all four of the expected frequencies for 'not very important' are less than 5 and this, of course, represents 25% of all the expected frequencies. The remedy in this instance is to combine 'not very important' with 'quite important' (let us call the combined class 'less important') to give the following table:

Opinion	Rhiubina	Cyn Coed	Splott	Canton	
very important	36 (29.2)	41 (37.0)	17 (29.2)	37 (35.6)	131
important	18 (19.9)	27 (25.1)	24 (19.9)	20 (24.1)	89
less important	6 (10.9)	8 (13.0)	19 (10.9)	16 (13.3)	49
	60	76	60	73	269

The analysis of this 3×4 contingency table would then proceed as before.

Of course, we might have had the low frequencies all in one column, in which case we would have combined that column with another; in this example the choice of column to combine with might be quite arbitrary unless it is possible to relate the regions by geographical or some other factor. But, in the case of the 'not very important' row it was certainly more meaningful to combine it with the 'quite important' row than with any other since it was the closest in meaning.

Yates' correction

In an attempt to improve the accuracy of the chi-squared approximation, especially in the case of 2×2 contingency tables, the following amended version of the statistic was once recommended:

$$\chi^2 = \sum \frac{(|\text{Ob.} - \text{Ex.}| - \frac{1}{2})^2}{\text{Ex.}}$$

the $\frac{1}{2}$ representing a continuity correction. You may still find this modified version in some books. However, research carried out in the 1960s showed that this alteration was just as likely to worsen the approximation as to improve it in standard contingency table applications, and therefore we do not recommend its use except in the particular situations covered in Chapter 18.

Assignments

1 Recreation habits

Three other questions and their responses from the survey on recreation habits in South Wales are given below.

(i) How important do you consider physical recreation to be?

	Rhiubina	Cyn Coed	Splott/ Adamstown	Canton/ Riverside
very important	36	37	18	37
important	19	26	21	20
quite important	5	9	15	10
not important	3	4	7	6

(ii) How satisfied are you with the level of facilities for leisure activities in your area?

	Rhiubina	Cyn Coed	Splot/ Adamstown	Canton/ Riverside
very satisfied	5	7	4	5
satisfied	24	42	24	32
not very satisfied	11	19	16	15
dissatisfied	11	18	1	21

(iii) Do you think that participation in leisure and recreation improves your health?

	Rhiubina	Cyn Coed	Splott/ Adamstown	Canton/ Riverside
yes	38	62	51	70
no	0	4	2	0
don't know	7	10	6	3

Task: Analyse these contingency tables using the technique given in the chapter.

Program 12.1 Chi-squared test for association

```
100 REM *** CHI-SQUARED TEST FOR ASSOCIATION ***

110 REM ** INPUT SIZE OF CONTINGENCY TABLE **
120 INPUT "NUMBER OF ROWS & COLUMNS ? ";NR,NC
130 PRINT
140 DIM R(NR),C(NC),OB(NR,NC),EX(NR,NC)
150 FOR I=1 TO NR
160 LET R(I)=0
170 NEXT I
180 FOR I=1 TO NC
190 LET C(I)=0
200 NEXT I
210 LET GT=0

220 REM ** INPUT OBSERVED FREQUENCIES **
230 PRINT "INPUT THE OBSERVED FREQUENCIES FOR :"
240 FOR I=1 TO NR
250 PRINT "ROW ";I
260 FOR J=1 TO NC
270 INPUT OB(I,J)
280 LET R(I)=R(I)+OB(I,J)
290 LET C(J)=C(J)+OB(I,J)
300 NEXT J
310 LET GT=GT+R(I)
320 NEXT I

330 REM ** CALCULATION OF EXPECTED FREQUENCIES **
340 LET F=0
350 FOR I=1 TO NR
360 FOR J=1 TO NC
370 LET EX(I,J)=R(I)*C(J)/GT
380 IF EX(I,J) < 5 THEN LET F=F+1
390 NEXT J
400 NEXT I
410 PRINT

420 REM ** LOW-FREQUENCIES WARNING **
430 IF F=0 THEN GO TO 460
440 PRINT "WARNING ... THERE ARE ";F;" EXPECTED FREQUENCIES LESS THAN 5"
450 PRINT

460 REM ** CALCULATION OF THE TEST STATISTIC **
470 LET Q=0
480 FOR I=1 TO NR
490 FOR J=1 TO NC
500 LET Q=Q+(OB(I,J)-EX(I,J))^2/EX(I,J)
510 NEXT J
520 NEXT I
530 PRINT "VALUE OF THE TEST STATISTIC = ";Q
```

PART V

Tests involving more than two samples

This part of the book is concerned with comparisons between *more than two* populations on the basis of samples drawn from them.

In Chapter 13, 14 and 15 we present tests for detecting differences in location between k (>2) populations. In Chapter 13, the situation is an extension of the two-independent-samples situation of Chapter 5, while Chapters 14 and 15 are an extension to k (>2) samples of the matched-pairs concept of Chapter 8.

The essential difference between Chapters 14 and 15 lies with the alternative hypotheses. In Chapter 15, the alternative hypothesis is quite specific about which populations are 'better' than which, that is, the populations are 'ordered'; however, Chapter 14 deals with the more general style of alternative hypothesis which simply states that there are *some* differences between the averages of the k populations without attempting to say which averages are larger or smaller.

The first part of Chapter 16 tackles a different sort of problem. Rather than simply looking for differences in average between the populations, we look for evidence of more complicated kinds of behaviour between the populations known as *interactions*. A typical example is where we are interested in the speed of a chemical reaction as related to the heat and pressure at which the reaction takes place. Since, for example, 10°C increase in temperature will have a greater effect at a high pressure than at a low pressure, such a situation cannot be dealt with in terms of average speed of reaction at a particular temperature nor at a particular pressure – the 'interactive' effect of temperature and pressure must be taken into account.

In the second part of Chapter 16, we use distribution-free methods to analyse the so-called latin squares design.

13

More than two (k) independent samples

Introduction

In Chapter 5, we used the information in two independent samples to look for evidence that the two populations from which the samples were taken had different averages (medians, usually). In this chapter, we deal with similar investigations, but now we have independent samples from three, four, five or even more populations. For example, suppose the consumer magazine in Chapter 5 had conducted its survey on cars from *four* countries, say Britain, Japan, West Germany and Italy. Then we would be looking for evidence of differences in reliability in cars from these four countries.

In classical statistics, this type of problem is called *one-way analysis of variance*. The analyses involve F-tests which rely for their validity on the observations being from normal distributions. The term 'analysis of variance' is not appropriate in distribution-free statistics since 'variance' is not relevant in this more general context. We shall simply refer to the problem of *k independent samples*. As usual, our procedures will not depend on any normality assumptions.

An early distribution-free test for k independent samples was an extension to Mood's (1950) median test; this is just a chi-squared test (see Ch. 12) based on the numbers of observations in each sample lying above and below the overall median. Then in 1952, W. H. Kruskal and W. A. Wallis (Kruskal 1952, Kruskal & Wallis (1952) introduced a test that is now by far the best-known distribution-free procedure for this situation. Both of these procedures for k independent samples are described in this chapter.

We also deal with another important issue in this chapter. Whenever we have evidence to support the alternative hypothesis that the averages of the k populations are differernt, we naturally want to know which populations are 'better' and which are 'worse'. To deal with this, we shall use O. J. Dunn's (1964) *multiple comparisons* procedure.

THE KRUSKAL–WALLIS TEST

Introduction

This popular test for the k-independent-samples problem is effectively an extension of the Mann–Whitney test (see Ch. 5). As you might deduce, therefore, it has very good power and is reasonably easy to execute unless the sample sizes are large (then you could always use the computer program – Program 13.1). The only requirements for the validity of the test are that the samples are independent of each other (meaning that the observations in any sample are not related to the observations in any other sample in any way) and, at least for theoretical purposes, that the populations are continuously distributed so that we are not troubled with ties.

It is interesting to note that the Kruskal–Wallis test was derived from the F-test (which is the classical test under normality for this situation) by replacing the actual observations by their ranks; this is analogous to the way that Spearman's rank correlation coefficient r_S was derived from the classical product-moment correlation coefficient r.

PRACTICAL PROBLEM

A survey was conducted by a consumer organization (in fact, the same organization we first mentioned in Chapter 5) on cars from Britain, Japan, West Germany and Italy. As before, they published a table giving the percentages of new cars that broke down in their first year. Our concern here is whether the published data suggest differences in reliability in cars from these different countries.

HYPOTHESES

The null hypothesis H_0 is that cars from the four countries are equally reliable with respect to breakdowns. In other words, H_0 says that the average percentage breakdowns is the same for all four countries. The alternative hypothesis H_1 is that there are some differences between the average percentage breakdowns. If it turns out that there is evidence to support H_1, then a further investigation would be useful to ascertain the relative performance of each country's cars – such a *multiple comparisons* investigation will be considered in a later section.

A summary of the hypotheses is:

H_0: There is no difference in the average reliability between cars from the four countries.

H_1: There are some differences in average reliabilities between cars from the four countries.

As with distribution-free procedures, the theory of the Kruskal–Wallis test is based on a more specific null hypothesis, namely that the populations are identically distributed (that is, identical in all respects, rather than just having the same average). Effectively, we assume that the populations are identical in *all* respects except, possibly, location. Thus it is usual to describe the hypotheses as something like:

H_0: The k populations have the same average (median).
H_1: There are some differences in average (median) between the k populations.

DATA

The survey gave breakdown information on 13 British models, seven Japanese, eight West German and four Italian. The percentages of cars of each type breaking down during their first year were:

Britain	21	33	12	28	41	39	24	29	30	19	27	38	23
Japan	9	13	6	3	5	10	4						
West Germany	18	35	8	17	22	20	37	11					
Italy	41	41	48	34									

As with the two-sample case, the conclusion seems apparent from a cursory glance at the data! Nevertheless, let us see if the test supports our intuition.

STATISTIC

The general idea behind the Kruskal–Wallis test is quite simple. Consider, for the moment, all the samples combined into one sample of size $N=n_1+n_2+\ldots+n_k$, where n_i is the size of the ith sample ($i=1,2,\ldots,k$); in our example $N=32$, with $k=4$ and $n_1=13$, $n_2=7$, $n_3=8$ and $n_4=4$. These N observations are then ranked from 1 to N. Now, if H_0 is true, then the $N=32$ ranks should be randomly distributed between the four groups; in other words, each group should have its fair share of low, medium and high ranks. But if H_1 is true, then we would perhaps expect one group to have mainly low ranks, another group to have medium-sized ranks and so on. Thus, if H_0 is true, the mean of the ranks assigned to one group would not differ much from the mean of the ranks for any other group. The Kruskal–Wallis statistic is based on comparing each group's mean rank with the mean of all the ranks. In practice, this is achieved by calculating the sum of the squares of the differences

$$\text{group mean rank} - \text{overall mean rank}$$

with these squares being weighted by the appropriate sample size to

compensate for the effect of unequal sample sizes.

Let us put this idea into practice with the car breakdown data. First we need to rank the combined samples, yet at the same time keep track of which ranks belong to which sample. This is effectively achieved by a technique similar to that used for the Mann–Whitney test: the observations are roughly placed on a linear scale and then the assignment of ranks is a trivial matter. In Table 13.1, the ranks are shown in brackets (note that, as usual, tied values are assigned average ranks).

From this table it is easy to obtain R_i, the sum of the ranks for sample i (i=1,2,3,4) and hence $\overline{R}_i$, the mean of the ranks for sample i. We obtain

R_i	33	261.5	117	116.5
$\overline{R}_i=R_i/n_i$	4.71	20.12	14.63	29.13

The overall mean rank, $\overline{R}$, is (1+2+. . .+32)/32 = 16.5. The weighted sum of the squared deviations is then given by:

$$\sum_{i=1}^{4} n_i(\overline{R}_i-\overline{R})^2=n_1(R_1-\overline{R})^2+n_2(R_2-\overline{R})^2+n_3(R_3-\overline{R})^2+n_4(R_4-\overline{R})^2$$

$$=7(4.71-16.5)^2+13(20.12-16.5)^2+8(14.63-16.5)^2+4(29.13-16.5)^2$$
$$=1809.43$$

To produce the Kruskal–Wallis statistic, we multiply this result by

$$\frac{12}{N(N+1)}=\frac{12}{32\times33}$$

So we obtain the statistic:

$$H=\frac{12}{N(N+1)}\sum_{i=1}^{k} n_i(R_i-\overline{R})^2$$

$$=\frac{12}{32\times33}\times 1809.43$$

$$=20.56$$

In practice, it is more convenient to use an alternative, but equivalent, formula for H, namely:

$$H=\frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i}-3(N+1)$$

Table 13.1

	1 (Japan)	2 (Britain)	3 (W. Germany)	4 (Italy)
0 →				
	3 (1)			
	4 (2)			
	5 (3)			
	6 (4)			
			8 (5)	
	9 (6)			
10 →	10 (7)			
			11 (8)	
		12 (9)		
	13 (10)			
			17 (11)	
			18 (12)	
		19 (13)		
20 →			20 (14)	
		21 (15)		
			22 (16)	
		23 (17)		
		24 (18)		
		27 (19)		
		28 (20)		
		29 (21)		
30 →		30 (22)		
		33 (23)		34 (24)
			35 (25)	
			37 (26)	
		38 (27)		
		39 (28)		
40 →				
		41 (29.5)		41 (29.5)
				46 (31)
				48 (32)
50 →				

where again R_i is the sum of the ranks in the ith group. Performing the calculation with this formula gives

$$H = \frac{12}{32\times33}\left(\frac{33^2}{7}+\frac{261.5^2}{13}+\frac{117^2}{8}+\frac{116.5^2}{4}\right) - 3\times33$$
$$= 20.56$$

as before.

CRITICAL REGIONS

Our discussion in the previous section clearly indicates that the alternative hypothesis is supported by large values of H, arising from an uneven distribution of ranks among the groups, which leads to relatively large differences between the mean ranks. Thus H_0 will be rejected in favour of H_1 if

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

Critical values of H are tabulated on pages 32 to 34 of *EST* for a large range of cases; Table N is an abbreviated version of this table. For values of k or for sample sizes not included in either table, approximate critical values are obtained from the chi-squared distribution (Table B) using $k-1$ degrees of freedom.

The sample sizes in our example are beyond the range of Table N, and therefore we need to obtain approximate critical values from Table B. The number of degrees of freedom for our problem is $k-1=4-1=3$. Hence, the appropriate approximate critical regions for $\alpha=5\%$ and 1% are given by $H\geqslant7.815$ and $H\geqslant11.345$ respectively.

Note that, since the terms 'one-sided' and 'two-sided' have no meaning in the k-sample context, we simply use α without any suffices to denote the significance level.

SOLUTION

Given these particular data, we should not be too surprised to discover that $H=20.56$ is well into the 1% critical region. Indeed it should be clear that the result is significant at very stringent levels (the detailed chi-squared tables in *EST* show the result to be in the 0.1% critical region), yielding essentially conclusive evidence of differences. So the Kruskal–Wallis test has firmly supported our intuitive feelings on these data.

Discussion

In the 'Discussion' section for the Mann–Whitney test, we remarked that the consensus of practical considerations seems to favour its use above all other appropriate distribution-free tests for two independent samples. The same is true

of the Kruskal–Wallis test when considering k independent samples; when sampling from a normal distribution its power is almost equal to that of the classical F-test. Also, it is interesting to note that the Kurskal–Wallis test is much more reliable than the F-test when outliers are present in the data.

It is a pity that, unless the sample sizes are small, the Kruskal–Wallis test is rather tedious to calculate by hand, although our systematic layout of the data certainly helps, as does the computer program (Program 13.1). Having said this, it should be observed that the F-test is, if anything, even more tedious and, of course, it does not have the distribution-free features of the Kruskal–Wallis test.

Ties

One of the underlying assumptions of the Kruskal–Wallis test is that the samples are taken from continuous populations so as to avoid the problem of tied observations. However, ties in data do occur, perhaps because of rounding errors in recording the observations or because the data were in fact taken from discrete distributions. When they do occur, they are dealt with by using average ranks, as with the two observations of 41 in the numerical example above. Whatever the cause of ties, their presence affects the null distribution of H and so, strictly speaking, the usual critical values are not valid. If the number of ties is small then H is calculated and used in the usual way (which is just what we did in analysing the car breakdowns data).

However, should there be a large number of ties, then it is as well to incorporate a correction factor. The value of H calculated from

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} - 3(N+1)$$

is divided by

$$C = 1 - \frac{\sum(t^3 - t)}{N(N^2 - 1)} = 1 - \frac{\sum t^3 - \sum t}{N(N^2 - 1)}$$

to give the statistic $H^* = H/C$, which is approximately chi-squared distributed with $k-1$ degrees of freedom. In this formula, t is the number of ties for a given (tied) value and the summation is taken over all sets of tied values. An example will clarify the procedure, with which we are already familiar from Chapter 5. Consider the following three samples A, B and C:

A	B	C
4	14	10
9	10	14
10	9	14
	12	8

The rank sums are easily obtained from the usual layout:

	A	B	C
0 →			
	4 (1)		
			8 (2)
	9 (3.5)	9 (3.5)	
10 →	10 (6)	10 (6)	10 (6)
		12 (8)	
		14 (10)	14 (10)
			14 (10)
20 →			

They are as follows:

i	1	2	3
Rank sums R_i	10.5	27.5	28
n_i	3	4	4

Using

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{3} \frac{R_i}{n_i} - 3(N+1) \text{ gives } H = 2.346.$$

However, as there are three different tied values, namely 9, 10 and 14, it may be worth using the refinement recommended above. Now, 9 occurs twice, giving $t=2$; 10 and 14 each occur three times, giving $t=3$ for each. To calculate the correction C it is helpful to construct the following table:

Observations	t	t^3
9	2	8
10	3	27
14	3	27

Then summing over the three ties, we have $\sum t=8$ and $\sum t^3=62$, from which

$$C = 1 - \frac{(\Sigma t^3 - \Sigma t)}{N(N^2 - 1)} = 1 - \frac{(62 - 8)}{11 \times 120}$$

i.e.

$$C = 0.959$$

Hence the 'corrected' value H^* of the Kruskal–Wallis statistic is

$$H^* = \frac{H}{C} = \frac{2.346}{0.959} = 2.446$$

which, as you can see, is not very different from H.

Incidentally, recalling that the car breakdown data contained a few ties, yo might like to confirm that the correction for these data is C=0.9998, which gives $H^* = H/C = 20.43/0.9998 = 20.43$ (i.e. unchanged to two decimal places). Clearly we were justified in ignoring the correction in this case!

Null distributions

The null hypothesis for the k-independent-sample problem states that the populations from which the samples are drawn have identical, continuous distributions (although as we stated previously, the hypotheses are usually worded in terms of differences in averages, everything else being assumed equal). In other words, under H_0 the k samples can be regarded as all coming from the same population, and so all arrangements of the $N=n_1+n_2+\ldots+n_k$ observations are equally likely; it is on this principle that the null distributions of H are derived. Correspondingly, all the arrangements of the N ranks into groups of sizes n_1, n_2, . ,n_k are equally likely under H_0. There are

$$\frac{N!}{n_1!n_2!\ldots n_k!}$$

such arrangements and, in order to obtain the distribution of H, we need to compute H for each of these arrangements. Even for moderate values of N, this is not only tedious but even beyond the limits of many computers. The size of the problem can be envisaged if we just look at a small-scale problem; with k=4 and $n_1=n_2=n_3=n_4=4$ there are 63 063 000 such arrangements! Even the existence of recurrence relations for the distributions of H (see Kruskal & Wallis 1952) does not greatly help matters, although they do permit some progress. However, we can easily demonstrate how to derive a null distribution of H for a small value of N. Consider k=3, $n_1=n_2=2$ and $n_3=1$, giving N=5. For this case there are 30 arrangements to consider. These are given below, together with the corresponding values of H. (In fact, we need to list only 15 arrangements since, for example, (1,2), (3,4), 5 and (3,4), (1,2), 5 obviously give the same value of H.)

Ranks of sample			
1	2	3	H
1,2	3,4	5	3.6
1,2	3,5	4	3.0
1,2	4,5	3	3.6
1,3	2,4	5	2.4
1,3	2,5	4	1.4
1,3	2,5	4	3.0
1,4	2,3	5	2.0
1,4	2,5	3	0.4
1,4	3,5	2	1.4
1,5	2,3	4	0.6
1,5	3,4	2	0.6
1,5	2,4	3	0
2,3	4,5	1	3.6
2,4	3,5	1	2.4
2,5	3,4	1	2.0

Thus,

$$\begin{aligned}
\text{Prob}(H=0) &= 1/15 \\
\text{Prob}(H=0.4) &= 1/15 \\
\text{Prob}(H=0.6) &= 2/15 \\
\text{Prob}(H=1.4) &= 2/15 \\
\text{Prob}(H=2.0) &= 2/15 \\
\text{Prob}(H=2.4) &= 2/15 \\
\text{Prob}(H=3.0) &= 2/15 \\
\text{Prob}(H=3.6) &= 3/15
\end{aligned}$$

From these probabilities, we construct a table of 'upper-tail' probabilities $\text{Prob}(H \geq h)$ from which critical values would normally be obtained:

h	$\text{Prob}(H \geq h)$
0	1
0.4	0.9333
0.6	0.8667
1.4	0.7333
2.0	0.6000
2.4	0.4667
3.0	0.3333
3.6	0.2000

In fact, in this case, because of the high probabilities associated with the high values of H, no meaningful critical values can be obtained; the smallest 'round

figure' significance level for which a critical region exists is $\alpha=20\%$! This explains the blank entry for $k=3$, $n_1=n_2=2$ and $n_3=1$ on page 32 of *EST*.

THE MEDIAN TEST

Mood (1950) introduced his *median test* for two independent samples, and this has since been extended to deal with k independent samples. It is much quicker to use than the Kruskal–Wallis test, since all that is necessary is to count the number of observations that are above and below the median of the combined sample. However, it obviously uses the data more crudely than does the Kruskal–Wallis procedure, and so it will generally be somewhat less powerful. As usual, therefore, we are giving you the choice of a test that is particularly easy to carry out or one that is a little more demanding but has greater sensitivity. A drawback of the median test is that, strictly speaking, it should not be used unless the sample sizes are at least 10; it is for this reason that we have chosen a different practical problem to illustrate the test.

PRACTICAL PROBLEM

The manager of a local sports club has devised four new training methods suitable for marathon runners. He asks for volunteers to participate in an experiment to determine whether there is any effective difference in the training schemes. Forty-nine members agreed to cooperate with the manager. Four groups of roughly equal sizes were randomly selected from the 49; each group was then randomly assigned to a particular training method. At the end of three months, the volunteers are to be timed on a trial marathon run.

HYPOTHESES

The hypotheses suggested by the nature of the experiment are simply:

H_0: The four training methods produce the same average running times.

H_1: There are some differences in average running times produced by the four training methods.

DATA

For various reasons, not all the volunteers completed their three-month training course. The times to run the trial marathon for the 44 who did complete the course are given below, for training methods A, B, C and D:

A	B	C	D
3 h 32 min	3 h 51 min	4 h 13 min	4 h 39 min
3 h 1 min	2 h 56 min	2 h 54 min	3 h 39 min
3 h 43 min	3 h 6 min	3 h 46 min	4 h 32 min
4 h 59 min	2 h 49 min	3 h 27 min	3 h 11 min
5 h 31 min	2 h 54 min	2 h 58 min	2 h 50 min
4 h 15 min	3 h 11 min	3 h 58 min	4 h 54 min
4 h 27 min	3 h 00 min	4 h 30 min	3 h 11 min
5 h 4 min	3 h 5 min	4 h 21 min	4 h 25 min
3 h 37 min	2 h 55 min	4 h 1 min	4 h 16 min
4 h 6 min	3 h 28 min	3 h 10 min	4 h 46 min
3 h 38 min	2 h 58 min		4 h 58 min
4 h 16 min			

STATISTIC

The idea behind this test is very simple. We first find the median of the whole set of observations; we shall refer to this as the *overall median*. For each sample we count the number of observations that lie above and below the overall median. If H_0 is true we would expect that each sample should have roughly the same number of observations above and below this median. So, if we construct a $2\times k$ contingency table from these counts, then we can test H_0 with the usual χ^2 test (see Ch. 12). The form of the table is shown below and the test statistic χ^2 is again $\sum(\text{Ob.}-\text{Ex.})^2/\text{Ex.}$ (summed over all $2k$ cells in the table) where the expected frequencies are calculated as before from the formula:

$$\text{Ex.} = \frac{\text{row total} \times \text{column total}}{\text{grand total}}$$

Sample	1	2	3	. . .	k
Number above overall median	A_1	A_2	A_3	. . .	A_k
Number below overall median	B_1	B_2	B_3	. . .	B_k
Totals	n_1	n_2	n_3	. . .	n_k

Now consider the data in our problem. Since there are $N=44$ observations altogether, the overall median is the average of the 22nd and 23rd values; this is easily found using the quick method given in Chapter 11, and its value is 3 hr 41 min. The contingency table is then as follows:

Training method	A	B	C	D	
Number above 3 h 41 min	8	1	6	7	22
Number below 3 h 41 min	4	10	4	4	22
	12	11	10	10	44

The calculation for χ^2 is:

Ob.	Ex.	$(\text{Ob.}-\text{Ex.})^2/\text{Ex.}$
8	6	0.667
4	6	0.667
1	5.5	3.682
10	5.5	3.682
6	5	0.200
4	5	0.200
7	5.5	0.409
4	5.5	0.409
		$9.916 = \chi^2$

CRITICAL REGIONS

Approximate critical values are obtained from the chi-squared distribution (Table B) with $k-1$ degrees of freedom. As with the usual contingency table analysis, we reject H_0 in favour of H_1 if

$$\chi^2 \geqslant \text{critical value}$$

Here we have $(r-1)(c-1)=(2-1)(4-1)=3$ degrees of freedom, and so the appropriate (approximate) critical regions for $\alpha=5\%$ and 1% are given by $\chi^2 \geqslant 7.815$ and $\chi^2 \geqslant 11.345$ respectively.

SOLUTION

The calculated value of χ^2 is 9.916, which lies well in the 5% critical region but not in the 1% region. In other words, there is some evidence to support the alternative hypothesis that there are some differences in the average running times produced by the training methods. In fact, if we look at the values of $(\text{Ob.}-\text{Ex.})^2/\text{Ex.}$, we see that the greatest difference between Ob. and Ex. occurs with method B, and seeing that 10 out of the 11 times with this method are below the median, we are certainly tempted to say that this is a superior training method.

Discussion

Although, as we have said earlier, Mood's median test treats the data rather crudely, it is still quite capable of producing highly significant results. The

desirability of using it only when sample sizes are at least 10 is not a serious restriction, since the Kruskal–Wallis test itself is fairly easy and quick to carry out when sample sizes are small. But when sample sizes are large, the Kruskal–Wallis test becomes very tedious to execute, whereas the median test is still quite quick, particularly if the technique in Chapter 11 is used to find the overall median. The median test is also far more convenient to use if there are a lot of ties in the data, since problems of averaging ranks and associated effects on the null distribution are of little importance in this test.

As an exercise, we suggest you apply Mood's test to the car breakdown data. Although the smallness of the sample sizes implies that the details of the critical regions may be rather inaccurate, you will find that $\chi^2=16.77$, which is a clear indication of very significant evidence in support of H_1, a conclusion that is very similar to that from the Kruskal–Wallis test.

MULTIPLE COMPARISONS

So far in this chapter, our concern has been the search for evidence of *some* differences between the averages (say the medians) of the k populations under investigation. If a test (be it the Kruskal–Wallis or the median test) finds such evidence, then it is natural to follow up by asking: 'Which populations can be most confidently claimed to have different averages from each other?' This query is answered by using a so-called *multiple comparisons* technique.

There are several such techniques available. We have chosen one due to O. J. Dunn (1964), mainly because of its simplicity and effectiveness, although its conclusions do tend to be on the 'conservative' side. In this procedure $\frac{1}{2}k(k-1)=p$ comparisons are made in order to judge the level of evidence for pairs of populations differing. Thus we have p sets of hypotheses of the form:

$H_0^{(ij)}$: There is no difference between the ith and jth populations.
$H_1^{(ij)}$: There is some difference between the ith and jth populations.

with $i, j=1,2, \ldots ,k$ and $i<j$.

It is usual to use one overall level of significance α for all of these statements rather than considering there to be a separate level for each. In this case, α is interpreted as the overall probability of making at least one error (i.e. saying there is a difference) in the p comparisons when in fact no differences exist among the k populations. Dunn argues for α to be taken greater than the 'usual' maximum value of 5% so that any differences among the comparisons can be easily detected. The reason why we dare think of using what would otherwise be such unrealistically high significance levels here is that we agree at the outset only to invoke this multiple comparisons procedure when we have already had a significant result from the Kruskal–Wallis or the median test. Since this

automatically provides us with a high degree of protection from even entering the procedure when the 'overall H_0' is true (i.e. when the whole set of p pairwise comparison H_0's are true or, in other words, when there are no differences in averages between the k populations), there is no harm caused by using Dunn's suggested high α levels. Significance levels of 10%, 15%, 20% or even 25% are therefore recommended. As a general rule, the higher the value of k the larger the value of α that should be used, for there are then larger numbers of pairwise comparisons amongst which to share out the risk of making an error of the type mentioned above. (In the problem discussed below, $k=4$ and we shall use $\alpha=20\%$.)

PRACTICAL PROBLEM

We use the car breakdowns data again; this situation is very suitable for Dunn's procedure since the Kruskal–Wallis test gave very strong evidence of some differences between the four populations, i.e. between the average reliabilities of cars from the four countries involved in the survey.

HYPOTHESES

The six $(=\frac{1}{2}k(k-1)=\frac{1}{2}\times4\times3)$ sets of hypotheses for this multiple comparison test are:

$H_0^{(ij)}$: The ith and jth country's cars have the same average number of breakdowns.
$H_1^{(ij)}$: The ith and jth country's cars have different average number of breakdowns.

with $(i,j)=(1,2)$, $(1,3)$, $(1,4)$, $(2,3)$, $(2,4)$ and $(3,4)$ and where, for convenience, we have attached the numerical labels 1,2,3 and 4 to Japan, Britain, West Germany and Italy respectively.

STATISTIC

An obvious way to attempt to identify differences in the averages is to compare the means of the ranks assigned to the samples. Dunn's procedure does just this. The $p=\frac{1}{2}k(k-1)=\frac{1}{2}\times4\times3=6$ absolute differences

$$|D_{ij}| = \left| \frac{R_i}{n_i} - \frac{R_j}{n_j} \right|$$

are first calculated for each of the six pairs of values of (i,j), where R_i and R_j are the sums of the ranks in the ith and jth samples. Clearly if there is a real difference between the ith and jth populations, then D_{ij} will tend to be large. The six differences for our problem are calculated below; recall

$3\times2=6\cdot\frac{1}{2}$