from earlier in the chapter that $R_1=33$, $R_2=261.5$, $R_3=117$ and $R_4=116.5$ with $n_1=7$, $n_2=13$, $n_3=8$ and $n_4=4$.

Pair (i,j)	$\lvert D_{ij}\rvert = \left\lvert \frac{R_i}{n_i} - \frac{R_j}{n_j} \right\rvert$	
(1,2)	$\lvert D_{12}\rvert = \left\lvert \frac{33}{7} - \frac{261.5}{13} \right\rvert$	$= 15.40$
(1,3)	$\lvert D_{13}\rvert = \left\lvert \frac{33}{7} - \frac{117}{8} \right\rvert$	$= 9.91$
(1,4)	$\lvert D_{14}\rvert = \left\lvert \frac{33}{7} - \frac{116.5}{4} \right\rvert$	$= 24.41$
(2,3)	$\lvert D_{23}\rvert = \left\lvert \frac{261.5}{13} - \frac{117}{8} \right\rvert$	$= 5.49$
(2,4)	$\lvert D_{24}\rvert = \left\lvert \frac{261.5}{13} - \frac{116.5}{4} \right\rvert$	$= 9.01$
(3,4)	$\lvert D_{34}\rvert = \left\lvert \frac{117}{8} - \frac{116.5}{4} \right\rvert$	$= 14.50$

Now it can be shown that, under H_0, the standard deviation of any difference D_{ij} is given by

$$\sigma_{ij} = \sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

and that the mean of D_{ij} is 0. It is also the case that D_{ij} is approximately normally distributed, and so if we divide D_{ij} by σ_{ij} we obtain a statistic T_{ij} which, under H_0, has approximately the standard normal distribution. Recalling that $N=32$, we obtain:

i,j	$\lvert D_{ij}\rvert$	σ_{ij}	$\lvert T_{ij}\rvert$
1,2	15.40	$\sqrt{\frac{32\times33}{12}\left(\frac{1}{7}+\frac{1}{13}\right)} = 4.40$	3.50
1,3	9.91	$\sqrt{\frac{32\times33}{12}\left(\frac{1}{7}+\frac{1}{8}\right)} = 4.86$	2.04
1,4	24.41	$\sqrt{\frac{32\times33}{12}\left(\frac{1}{7}+\frac{1}{4}\right)} = 5.88$	4.15

Continued

i,j	$\lvert D_{ij}\rvert$	σ_{ij}	$\lvert T_{ij}\rvert$
2,3	5.49	$\sqrt{\frac{32\times33}{12}\left(\frac{1}{13}+\frac{1}{8}\right)} = 4.22$	1.30
2,4	9.01	$\sqrt{\frac{32\times33}{12}\left(\frac{1}{13}+\frac{1}{4}\right)} = 5.36$	1.68
3.4	14.50	$\sqrt{\frac{32\times33}{12}\left(\frac{1}{8}+\frac{1}{4}\right)} = 5.74$	2.53

CRITICAL REGIONS

As we mentioned in the previous section, the statistic T_{ij} is approximately distributed as a standard normal variable. So, since large values of T_{ij} will tend to support the two-sided alternative hypothesis, H_0 will be rejected if

$$|T_{ij}| \geqslant z$$

where z is the percentage point corresponding to some prescribed probability in the standard normal distribution. Let us think how we should choose this probability. If we were only carrying out a single test, this probability would be $\frac{1}{2}\alpha$ (since the test is two-sided). However, in fact we are carrying out six tests here (and in general $\frac{1}{2}k(k-1)$ tests will be performed). Sharing out the risk equally between these tests, the appropriate probability is thus $\frac{1}{2}\alpha/\frac{1}{2}k(k-1) = \alpha/k(k-1)$.

So, taking $\alpha=20\%$ and remembering that $k=4$, we seek that value of z corresponding to an upper probability of $\alpha/k(k-1) = 0.20/12 = 0.0167$. From Table A, we obtain $z=2.128$. Therefore, the H_0 corresponding to ith and jth populations will be rejected if $|T_{ij}|\geqslant 2.128$.

To save you the trouble of carrying out this kind of calculation, we give below a table of such critical values of z for $k=3$ to 10 and $\alpha=5\%$ to 30%. From this, you can immediately check that for $\alpha=20\%$ and $k=4$, the critical value is indeed 2.128.

	Significance level					
k	5%	10%	15%	20%	25%	30%
3	2.394	2.128	1.960	1.834	1.731	1.645
4	2.638	2.394	2.241	2.128	2.037	1.960
5	2.807	2.576	2.432	2.326	2.241	2.170
6	2.935	2.713	2.576	2.475	2.394	2.326
7	3.038	2.823	2.690	2.593	2.515	2.450
8	3.125	2.913	2.785	2.690	2.615	2.552
9	3.200	2.990	2.823	2.772	2.700	2.658
10	3.260	3.060	2.933	2.844	2.764	2.713

SOLUTION

Comparing the successive values of $|T_{ij}|$ with the critical value of 2.128, we find significant differences between the following pairs:

(a) 1 and 2, i.e. Japan and Britain;
(b) 1 and 4, i.e. Japan and Italy;
(c) 3 and 4, i.e. West Germany and Italy.

Furthermore, from the direction of the differences, we have evidence that the likelihood of breakdowns is higher for British cars than Japanese, and for Italian cars than Japanese and for Italian cars than West German.

Results from such a multiple comparisons investigation are often presented in a diagram like the one in Figure 13.1. The populations are written down according to the order of the mean ranks, and lines are drawn to join together any populations that the procedure has failed to separate. These results confirm our previous feelings on these data, and this should give us confidence in the ability of Dunn's procedure to provide a valid analysis of other data, which may not be so easy to interpret at first sight.

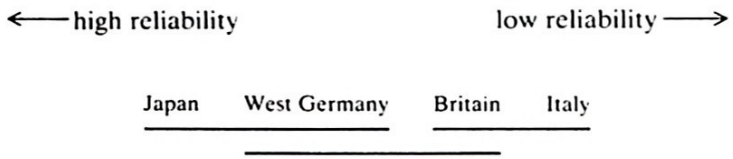


Figure 13.1

Ties

We have ignored ties in applying Dunn's procedure to our problem. However, a correction for ties exists and consists of a familiar type of modification to the standard deviation σ_{ij}. The corrected standard deviation σ^*_{ij} is given by:

$$\sigma^*_{ij} = \sqrt{\left(\frac{N(N+1)}{12} - \frac{\sum t^3 - \sum t}{12(N-1)}\right)\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

where t is the number of ties for a given (tied) value and the summation is taken over all sets of tied values in the data. With our practical problem, the correction is slight; the difference between:

$N(N+1)/12$ and $N(N+1)/12 - (\sum t^3 - \sum t)/12(N-1)$ being only 0.01.

Assignments

1 Water pollution

In a study of the pollution of English inland waterways, five pike were caught at each of seven locations. The amount of pollution present was

assessed by measuring the concentration of copper in the livers of the fish. The table below gives the logarithm of the copper concentration.

Windermere	0.187	0.836	0.704	0.938	0.124
Grasmere	0.449	0.769	0.301	0.045	0.846
River Stour	0.628	0.193	0.810	0.000	0.855
Wimbourne St Giles	0.412	0.286	0.497	0.417	0.337
River Avon	0.243	0.258	−0.276	−0.538	0.041
River Leam	0.134	0.281	0.529	0.305	0.459
River Kennet	0.471	0.371	0.297	0.691	0.535

Task: Investigate these data to examine differences in average copper concentrations between the seven locations. Should there be evidence to suggest such differences, investigate further using Dunn's multiple comparisons procedure.

2 Baby food

A paediatrician conducted an experiment to investigate the claim that three different brands of milk powder (designated A, B and C for anonymity) produce different average weight increases in babies. A sample of 35 babies, all born at a local maternity hospital, were randomly assigned to be fed from birth on one of the brands of milk powder. Their increases in weight (kg) over a fixed period of time are recorded in the following table.

A	B	C
3.24	3.42	3.73
2.22	2.37	3.56
2.40	2.59	3.27
2.58	3.57	2.51
2.40	2.36	3.30
3.49	3.61	3.12
2.68	2.94	3.11
2.15	3.86	3.05
2.30	2.43	3.78
3.21	2.76	2.40
2.86	2.36	2.65
	2.86	3.21

Task: Use an appropriate procedure to test the null hypothesis that the average increases in weight are the same for the three brands of milk powder.

Program 13.1 The Kruskal–Wallis test

```
100 REM *** KRUSKAL-WALLIS TEST ***

110 REM ** INPUT THE NUMBER OF SAMPLES **
120 INPUT "NUMBER OF SAMPLES ( >2 ) ";K
130 DIM M(K),R(K)
140 LET N=0

150 REM ** INPUT SIZE OF SAMPLES & OBSERVATIONS **
160 FOR I=1 TO K
170 PRINT "INPUT SIZE OF SAMPLE ";I
180 INPUT M(I)
190 LET N=N+M(I)
200 NEXT I
210 PRINT
220 PRINT
230 DIM D(N),RK(N)
240 LET L=0
250 FOR I=1 TO K
260 PRINT "INPUT OBSERVATIONS FOR SAMPLE ";I
270 FOR J=1 TO M(I)
280 LET L=L+1
290 INPUT D(L)
300 NEXT J
310 NEXT I
320 GO SUB 1000

330 REM ** CALCULATION OF THE TEST STATISTIC **
340 LET S=0
350 LET L=0
360 FOR I=1 TO K
370 FOR J=1 TO M(I)
380 LET L=L+1
390 LET R(I)=R(I)+RK(L)
400 NEXT J
410 NEXT I
420 FOR I=1 TO K
430 LET S=S +R(I)^2/M(I)
440 NEXT I
450 LET H=12*S/N/(N+1)-3*(N+1)
460 PRINT
470 PRINT "VALUE OF KRUSKAL-WALLIS STATISTIC = ";H

480 REM *** DUNN'S MULTIPLE COMPARISONS ***
490 PRINT
500 PRINT "DIFFERENCES IN MEAN RANKS FOR DUNN'S MULTIPLE COMPARISONS :"
510 PRINT
520 PRINT
530 FOR I=1 TO K
540 LET R(I)=R(I)/M(I)
550 NEXT I
560 FOR I=1 TO K-1
570 FOR J=I+1 TO K
580 LET DF=ABS(R(I)-R(J))/SQRT(N*(N+1)*(1/M(I)+1/M(J))/12)
590 PRINT "DIFFERENCE BETWEEN ";I;" & ";J;" IS ";DF
600 NEXT J
610 NEXT I

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

14

More than two (k) related samples

Introduction

The main practical problem we looked at in Chapter 13 was the following: on the basis of samples of cars from Italy, West Germany, Britain and Japan, is there evidence to suggest that these four countries (i.e. populations) produce cars of differing average reliability?

In this chapter we are still asking the same type of question, i.e. is there evidence of real differences between populations? The essential change between this and the last chapter is in the way the populations are sampled. Previously the samples have been independent of each other, i.e. there has been absolutely no connection between the observations in the different samples. In practice, though, it is often the case that observations are related in some way across the samples. For example, suppose a nurse took four readings of the pulses of three patients, these readings being taken at specific times of the day. Typically we might have:

	Patient		
Time (h)	Kathryn	Fiona	Richard
0600	60	62	55
1000	65	66	65
1800	70	80	75
2200	65	72	68

The observations in the $k=3$ samples of pulse readings are not independent, but are related by a common feature: the time of day. It looks as if the time of day has an important influence on the readings – the pulse rates are relatively low first thing in the morning and high in the early evening. Therefore, the effectiveness of our analysis of these data will be reduced unless we take the time of day into account.

The idea of samples being related is not new to us. In Chapter 8 we dealt with the 'matched-pairs' situation, i.e. where the data in two samples were 'matched' or related in pairs. This led to a substantially

different analysis from that of two independent samples in Chapter 5. So our topic in this chapter of 'k related samples' is an extension of the matched-pairs concept to more than two samples.

Let us consider another example of k related samples. A seed producer wishes to compare the yields of five ($5=k$) varieties of wheat. Four areas or 'blocks' of land were selected for the experiment and each seed variety was sown on equal-sized plots in each block. The table shows the yields (in kg/plot) of the five varieties:

	Variety of seed				
Block of land	Early Cropper	Wet Resistant	Big Cropper	Prairie Special	New Standard
1	46.7	48.2	50.6	45.6	45.2
2	42.8	43.7	45.7	47.3	39.8
3	50.1	49.1	51.9	48.1	44.6
4	49.2	50.6	52.3	47.6	42.7

Naturally, the producer is interested in finding out if there are any systematic differences between the yields of the five varieties.

Readers who have been trained in 'orthodox' statistical methods will recognize that both of the above examples could be analysed using the classical F-test for two-way analysis of variance, provided of course that the user is willing and able to take the big step of assuming that the samples were taken from normal distributions. However, consider the following situation: the five finalists in a seaside beauty contest were placed in order of beauty by three judges. The results were:

	Contestant				
Judge	A	B	C	D	E
Mayor	4	2	1	5	3
Local celebrity	3	1	2	4	5
Last year's winner	5	1	2	3	2

We see that the data consist simply of ranks and there is no way that they can be considered normally distributed! So, use of the classical F-test is out of the question in this case and we *must* use a distribution-free procedure to deal with such data. Even if, as previously, we do have numerical data available (rather than ranks), but the normality assumption is unjustified, the numbers can of course be converted to ranks and the same distribution-free procedure is still appropriate and valid.

One of the earliest distribution-free tests was designed for the problems we have just seen. Milton Friedman (1937), the renowned economist,

introduced a test which now bears his name. This has become well known and widely used, for not only are the required computations fairly easy, but also the test has good performance over a wide range of conditions. We shall examine Freidman's test in detail in this chapter.

Not surprisingly, since 1937 a number of distribution-free competitors to Friedman's test have been devised. However, not many can compete with regard to both power and ease of computation. Nevertheless, we present later in the chapter two very simple procedures – one is particularly easy to apply but has only moderate power, while the other is slightly more involved but yields greater power (sometimes exceeding the power of Friedman's test). Both of these tests are based on the so-called matching principle, which was first investigated by Montemort in 1708, but has only recently been applied in the current context. One reason for introducing these 'match tests' is that they can be applied to quite a wide range of k-sample problems for which it is otherwise difficult to obtain distribution-free procedures.

A note on terminology

A characteristic feature of the k-related-samples problem is that the data may be presented in a tableau of k columns and n rows (say). Many authors refer to the columns as 'treatments' and the rows as 'blocks'; this jargon stems from one of the first main areas of application of the k-related-samples problem, which was in agricultural experiments. It is also for this reason that the term 'randomized block design' is sometimes used (often incorrectly) to describe the situations of this chapter.

FRIEDMAN'S TEST

Introduction

Friedman's test is arithmetically easier than the Kruskal–Wallis test as one has now only to rank the observations within each row rather than ranking all the observations in the (often large) combined sample. As usual with distribution-free procedures, the main theory assumes there are no ties in the data, implying that the samples are drawn from populations having continuous distributions. However, this assumption is frequently relaxed, particularly as the test can easily tolerate a moderate number of ties.

PRACTICAL PROBLEM

A popular weekly photographic magazine reviewed a 200 mm telephoto lens from each of four top Japanese manufacturers. For each of the four lenses, the resolution (in lines per millimetre) at various apertures was recorded. Our concern here is whether the data suggest differences in

performance of lenses from the four manufacturers.

The readings in the four samples are related by the apertures, so we are indeed dealing with a problem of related, rather than independent, samples. The reader might justifiably consider this to be a poor experiment where only one lens from each manufacturer is examined. We agree! It would certainly have been preferable to examine more than one lens from each manufacturer; this would have helped to overcome problems caused by manufacturing variations. However, the staff of the magazine designed the experiment – not us! Had the experiment been improved in this way, the identical procedure to that described below could have been carried out on *averages* of the results.

HYPOTHESES

The null hypothesis H_0 is that the four manufacturers produce 200 mm lenses of equal average performance. The alternative hypothesis H_1 is that there are some differences between the average performances of the lenses. Clearly, if evidence is found to support this alternative hypothesis, then a further multiple comparisons investigation may be needed in order to ascertain which of the differences between lenses are significantly large.

A summary of the hypotheses is:

H_0: There is no difference in the average performances of 200 mm lenses from the four manufacturers.

H_1: There are some differences in the average quality of 200 mm lenses from the four manufacturers.

DATA

The table shows the resolutions recorded for the lenses from the four manufacturers (which we shall refer to as A, B, C and D to preserve anonymity) at apertures of *f*/2.8, 4, 5.6, 8 and 11.

	Manufacturer			
Aperture	A	B	C	D
2.8	83	77	88	87
4	92	96	110	109
5.6	124	119	129	125
8	127	123	132	129
11	124	117	122	123

A quick glance at these data might well immediately make you doubt the truth of H_0 – but we shall see!

STATISTIC

The observations *within each row* are ranked from 1 to k (where $k=4$ in this problem). Now, if H_0 is true, we would expect the ranks 1, 2, 3 and 4 to be fairly evenly distributed among the four manufacturers. On the other hand, if there are substantial differences then we would expect to find at least one manufacturer getting consistently high or low ranks. In other words, if H_0 is true then there should be little difference between the average ranks assigned to the four manufacturers, whereas if H_1 is true we would expect to see some relatively large differences between these average ranks.

We will denote the mean of the ranks in columns 1 to k by $\overline{R}_1$, $\overline{R}_2, \ldots, \overline{R}_k$. Observe also that since the k 'treatments' (lenses, in our problem) are ranked from 1 to k for each row (i.e. for each aperture), the average overall rank $\overline{R}$, say, must be equal to $\frac{1}{2}(k+1)$ since this is the average of the numbers $1, 2, \ldots, k$. In Friedman's test we compare each of the average ranks $\overline{R}_1, \overline{R}_2, \ldots, \overline{R}_k$ with the overall average rank $\overline{R}$ by forming the differences $\overline{R}_1-\overline{R}, \overline{R}_2-\overline{R}, \ldots, \overline{R}_k-\overline{R}$. Friedman's statistic is then based on the sum of the squares of these differences, namely

$$S = \sum_{j=1}^{k} (\overline{R}_j-\overline{R})^2$$

If H_0 is true, this statistic should be fairly small, for then the $\overline{R}_j$ values will generally be close to $\overline{R}$. But if there are sizeable differences between the column mean ranks, thus indicating the truth of H_1, then S will be relatively large.

Let us put this idea into practice with the lenses data. Although we shall evolve Friedman's statistic by following the above idea, we shall later give an alternative formula, which most people find more convenient to use.

Ranking the observations within each row gives the following table of ranks. (Notice how these ranks immediately reveal a clearer picture about the relative performance of the lenses even before we embark on the test itself.)

	A	B	C	D
	2	1	4	3
	1	2	4	3
	2	1	4	3
	2	1	4	3
	4	1	2	3
sum of ranks	11	6	18	15
mean, $\overline{R}_j$	2.2	1.2	3.6	3.0

The overall mean $\overline{R}$ of the ranks is $\frac{1}{2}(k+1) = \frac{1}{2}(4+1) = 2\frac{1}{2}$. So we calculate S as follows:

$$S = \sum_{j=1}^{4} (\overline{R}_j - \overline{R})^2 = (2.2-2.5)^2+(1.2-2.5)^2+(3.6-2.5)^2+(3.0-2.5)^2$$

$$= 3.24$$

Friedman's statistic M is not actually S itself but the following multiple of S:

$$M = \frac{12n}{k(k+1)} S$$

In our example, $k=4$ and $n=5$, and so we have

$$M = \frac{12\times5}{4\times5} S = 3S = 3\times3.24 = 9.72$$

The reason we use this particular multiple of S is that the resulting statistic M then behaves, when H_0 is true, approximately like a chi-squared variable. This allows us to use chi-squared tables to obtain approximate critical values for cases not covered by Table O (or the table on page 34 in *EST*).

The alternative version of Friedman's statistic mentioned above, which often shortens the computation, is

$$M = \frac{12}{nk(k+1)} \sum_{j=1}^{k} R_j^2 - 3n(k+1)$$

where R_j is the sum of the ranks in column j. Referring back to the table of ranks we see that $R_1=11$, $R_2=6$, $R_3=18$ and $R_4=15$. Hence the alternative formula gives

$$M = \frac{12}{5\times4\times5} (112+62+182+152) - 3\times5\times5 = 9.72$$

just as before.

CRITICAL REGIONS

In the previous section, we observed that large values of the sum of squares, and hence M, would support the claim of the alternative hypothesis that there are differences in average performance between the

manufacturers. Thus H_0 will be rejected in favour of H_1 if

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

Critical values of M are tabulated on page 34 of *EST* for $k=3$ to 6 and n up to 25. Table O is an abbreviated version of this table.

For values of k or n not included in either table, approximate critical values are obtainable from the chi-squared distribution (Table B) using $k-1$ degrees of freedom and α_1 significance levels (for k between 3 and 6, they are also given in Table O).

In our particular problem, $k=4$ and $n=5$, and so we are able to use Table O. The critical regions for $\alpha=5\%$ and 1% are $M \geqslant 7.800$ and $M \geqslant 9.960$ respectively.

SOLUTION

A cursory glance at the ranked data suggests that a rejection of H_0 might be appropriate, and indeed this turns out to be so. The value $M=9.72$ is in the 5% critical region, though it does not quite make the 1% region. (In fact, the more detailed tables in *EST* show that the result is significant at the 2% level.) So we have quite strong evidence to support the claim that there are some differences in the quality of lenses from the four manufacturers.

Discussion

In our particular problem, we applied Friedman's statistic to the lenses data to see whether there is evidence of differences in average quality between the four manufacturers. However, one of the advantages of organizing the data so that the samples are related is that we can check for significant differences between the rows (apertures) as well as between the columns (manufacturers). In other words, for our problem we could also test the null hypothesis that there are no average differences in the lens performances throughout the aperture range against the alternative hypothesis that the average performance does differ according to the aperture used. In this case we would rank the data within each column, giving the following table of ranks:

	Manufacturer			
Aperture	A	B	C	D
2.8	1	1	1	1
4	2	2	2	2
5.6	3.5	4	4	4
8	5	5	5	5
11	3.5	3	3	3

Friedman's test would then be applied with $k=5$ and $n=4$. Most photographers would agree with this alternative hypothesis, and you may like to verify that Friedman's test does indeed find strong evidence to support it.

Null distributions

As in the previous chapter, the null hypothesis claims that there are no differences between the averages of the k populations. As usual, we further assume there to be no other differences either – except here, for differences caused by the 'rows' factor. Since we are only ranking the observations within the rows, we may argue that, under H_0, in every row, each of the $k!$ possible arrangements of the ranks 1,2, . . . ,k is equally likely to occur. Since we have n rows each with k ranks, there are altogether $(k!)^n$ possible configurations of ranks. By calculating the value of M for each of these configurations we can build up the frequency distribution for M; the null distribution is then obtained by dividing these frequencies by the total number of configurations (i.e. by $(k!)^n$).

We illustrate this procedure with the case of $k=3$ and $n=3$. Although there are $(k!)^n=(3!)^3=216$ configurations of three rows of three ranks, we need only consider the $(k!)^{n-1}=6^2=36$ of them obtained by fixing the first row as 1 2 3. (Can you see why? Note that, if you swap the columns around in any given configuration, the value of M remains unchanged.) To save space, rather than listing all 36 distinct possibilities, we present just a few to give you the idea. So a partial table of ranks for $k=3$, $n=3$ is as follows:

1 2 3	1 2 3	1 2 3	1 2 3	1 2 3	1 2 3
1 2 3	1 2 3	1 2 3	1 2 3	1 2 3	1 2 3
1 2 3	1 3 2	2 1 3	2 3 1	3 1 2	3 2 1
R_j = 369	378	459	477	558	567
M = 6	4.667	4.667	2	2	0.667
1 2 3	1 2 3	1 2 3	1 2 3	1 2 3	1 2 3
1 3 2	1 3 2	1 3 2	1 3 2	1 3 2	1 3 2
1 2 3	1 3 2	2 1 3	2 3 1	3 1 2	3 2 1
R_j = 378	387	468	486	567	576
M = 4.667	4.667	2.667	2.667	0.667	0.667

Continuing in this way until all 36 configurations have been recorded, we eventually obtain the following frequency distribution of M:

Value of M	Frequency
6	1
4.667	6
2.667	6
2	6
0.667	15
0	2

From this table we can deduce that, under H_0,

$$\begin{aligned}
\text{Prob}(M=0) &= 2/36\\
\text{Prob}(M=0.667) &= 15/36\\
\text{Prob}(M=2) &= 6/36\\
\text{Prob}(M=2.667) &= 6/36\\
\text{Prob}(M=4.667) &= 6/36\\
\text{Prob}(M=6) &= 1/36
\end{aligned}$$

Using these probabilities, we are able to construct a table of upper-tail probabilities Prob($M \geq m$), which, since the critical regions for this test are of the form $M \geq$constant, can then be used to obtain the critical values. Thus the table of Prob($M \geq m$) for $k=3, n=3$ is as follows:

m	Prob($M \geq m$)
6	0.0278
4.667	0.1945
2.667	0.3612
2	0.5279
0.667	0.9446
0	1

The table on page 34 of *EST* gives $M \geq 6$ (i.e. $M=6$) as the best conservative critical region for $\alpha=10\%$ and 5%, and gives no region for $\alpha=2\%$ and 1%. The above table confirms these entries; the next larger critical region, $M \geq 4.667$, corresponds to a significance level of nearly 20%, and so cannot be considered. The probability that $M=6$ is 0.0278, i.e. 2.78%, which is of course less than 5% and 10% but greater than 2% and 1%. (Table O just gives details for $\alpha=5\%$ and 1%.)

The above process is obviously very tedious, but recurrence methods may be used to carry out the computations more efficiently. There are also some excellent approximate methods available, and these two approaches have been used to provide Table O and the table in *EST*.

Large sample sizes

As we mentioned in the 'Critical regions' section, for values of k or n not included in Table O, approximate critical values can be obtained from the chi-squared distribution using $k-1$ degrees of freedom. Thus, for example, if $k=7$ then the chi-squared distribution with $k-1=6$ degrees of freedom gives approximate critical values for $\alpha=5\%$ and 1% as 12.592 and 16.812 respectively.

Treatment of ties

We are not usually troubled much by tied observations when using Friedman's test because of the fact that comparisons are now only being made within rows rather than over the data as a whole as in the Kruskal–Wallis test.

If equal observations are found in the same row, then we assign to them average rank values as usual. With the number of such ties normally being very small, this is usually sufficient to deal with the problem.

However, if, exceptionally, there are a lot of such tied observations, then the null distribution of M is noticeably affected, and we shall need to take this into

account. The usual method is to compute a familiar-looking correction factor C and then modify the Friedman statistic to $M^*=M/C$.

The correction factor C is calculated as follows. Denoting by t the number of observations contributing to a tie, we form $t^*=t^3-t$. (Remember we are now only concerned with ties within rows, and so it will probably be rare for t to exceed 3.) We go through all the rows, adding up such t^* for any ties found, and denote the total of these t^* by T. Then the correction factor C is given by

$$C = 1 - \frac{T}{n(k^3-k)}$$

To illustrate the procedure, consider the following data, where $k=4$ and $n=3$:

	Treatment			
Row	A	B	C	D
1	8.5	8.9	8.8	8.8
2	8.2	8.4	8.2	8.2
3	8.9	9.1	9.1	8.9

These can be ranked as follows:

	Treatment			
Row	A	B	C	D
1	1	4	2½	2½
2	2	4	2	2
3	1½	3½	3½	1½
rank sums	4½	11½	8	6

Then the usual calculation for Friedman's statistic gives $M=5.5$.

Table O shows the 5% critical region to be $M\geqslant 7.4$, so 5.5 is not significant. However, let us now compute the correction factor C in order to evaluate the modified statistic, M^*. In the first row there is just one pair of tied observations, and so $t=2$ and $t^*=6$. In the second row, three observations are tied, giving $t=3$ and $t^*=24$; and in the final row, there are two pairs of tied observations and so $t^*=6$ twice. Hence $t=6+24+6+6=42$, and so

$$C = 1 - \frac{T}{n(k^3-k)} = 1 - \frac{42}{3(4^3-4)} = \frac{23}{30} = 0.7667$$

This gives $M^* = 5.5/0.7667 = 7.17$. This is close to, but does not quite reach, the 5% critical region. However, from page 34 of *EST*, we see that the 10% critical region is $M\geqslant 6.6$ (or rather $M^*\geqslant 6.6$), and so the modified statistic is well inside this region indicating that, perhaps there is after all, some significant evidence here.

THE MATCH TESTS

The authors have developed two simple tests for k related samples; the tests are based on a well known 'matching' principle. To begin, let us take a look at the general ideas on which the match statistics (which we shall call M_1 and M_2) are founded. Consider the following set of ranks from data arranged in five columns and two rows (which corresponds to $k=5$ treatments and $n=2$ rows in the current context):

1	3	5	4	2
2	4	5	1	3

We say there is a match between two ranks in the same column if those ranks are identical. In the above table, there is just one match; it occurs between the '5's in the third column. For the M_2 statistic we also use *near-matches*; we have a *near-match* between two corresponding ranks if they differ by *one*. In the above table, there are three near-matches; they occur in columns 1, 2 and 5.

The statistic M_1 uses only matches and not near-matches. The test based on M_1 may be described as 'quick and easy' since, besides involving only very simple computations, it has the added advantage that critical values are almost independent of the number of column (treatments) in the experiment. Not surprisingly, the power of this test is lower than that of the test based on the M_2 statistic, which incorporates the extra information of near-matches. In fact, the power of the M_2 test has been shown to be close to (and sometimes better than) Friedman's test.

Let us see how these ideas can be applied to the practical problem of the camera lens manufacturers.

HYPOTHESES

These are the same as before.

STATISTICS

We recall that the ranks of the data are:

	Manufacturer			
Aperture	A	B	C	D
2.8	2	1	4	3
4	1	2	4	3
5.6	2	1	4	3
8	2	1	4	3
11	4	1	2	3

Now, if the lenses differ markedly, we would expect some of the columns to contain mainly high or low ranks. This would result in a relatively large number of matches and/or near-matches. On the other hand, if H_0 is true, we would expect there to be only a small number of matches and/or near-matches.

Now let us see the details, beginning with the M_1 statistic. We first count the number of matches obtained when rows 2, 3, 4 and 5 are compared in turn with row 1; in other words, we take the top row as a 'reference row', and count the number of times that ranks in this reference row are repeated in the same columns in the remaining rows. Doing this, we find a total of 2+4+4+2=12 matches. Next, row 2 is used as the reference row and the remaining rows 3, 4 and 5 are compared with it; this gives 2+2+1=5 matches. Repeating the process with row 3 as the reference row gives 4+2=6 matches and finally, with row 4 as the reference row we get 2 matches. The value of M_1 is given by the total number of matches from all these comparisons. So we obtain

$$M_1 = 12+5+6+2 = 25$$

Now we consider the second statistic M_2. We follow a similar procedure for doing the comparisons but this time we record the number of near-matches as well as the number of matches. We give these values in the following table.

Reference row	Number of matches	Number of near matches
1	2 + 4 + 4 + 2 = 12	2 + 0 + 0 + 0 = 2
2	2 + 2 + 1 = 5	2 + 2 + 1 = 5
3	4 + 2 = 6	0 + 0 = 0
4	2 = 2	0 = 0
totals	25	7

M_2 is defined as *the number of matches plus half the number of near-matches* (since near-matches represent 'near-misses' it is only right that they do not contribute the same amount as matches). So we have,

$$M_2 = (\text{number of matches}) + \tfrac{1}{2}(\text{number of near-matches})$$
$$= 25 + 3\tfrac{1}{2} = 28\tfrac{1}{2}$$

CRITICAL REGIONS

M_1 test

Since it is large values of M_1 that are significant, the null hypothesis is rejected in favour of the alternative hypothesis if

$$M_1 \geqslant \text{critical value}$$

Critical values of M_1 are given in Table P for $k=3$ to 6 and n up to 20.

In our problem, we have $k=4$ and $n=5$ and so we see that the critical regions for $\alpha=5\%$ and 1% are $M_1 \geqslant 17$ and $M_1 \geqslant 20$ respectively.

M_2 test

As with M_1, large values of M_2 support the alternative hypothesis, and so H_0 is rejected in favour of H_1 if

$$M_2 \geqslant \text{critical value}$$

Table Q gives critical values of M_2 for $k=3$ to 6 and n up to 20.

In our problem, we have $k=4$ and $n=5$, and so Table Q shows that H_0 is rejected in favour of H_1 if $M_2 \geqslant 23$ and $M_2 \geqslant 26.5$ at the 5% and 1% significance levels respectively.

SOLUTION

Both match tests, M_1 and the more powerful M_2, give results that are significant at better than the 1% level. In other words, both tests indicate strong evidence supporting the claim that the lens manufacturers produce lenses of differing average qualities.

Discussion

It is interesting to note that both the match tests give results that are significant at the 1% level whereas Friedman's test only gave significance at about the 2% level. One should not be too surprised by this; after all, the match statistics and Friedman's statistic are using different aspects of the information in the ranks. It is not at all unusual for a good 'quick' test to produce better levels of significance than more 'orthodox' tests.

The match tests, which, at the time of writing, have only recently been developed, appear to have much in their favour. The computations are simple, and approximate critical values are available in a relatively 'compact' form. These are properties shared by some of the quick tests in earlier chapters, e.g. Tukey's test in Chapter 5. As with these other quick tests, therefore, the match tests may prove particularly attractive to less numerately skilled workers.

The matching principle also has a further advantage, of perhaps even more importance. This is that it can be applied to more complex experiments for which, at present, other distribution-free methods currently seem to be either unworkable or even unavailable (for details see Worthington 1982). An illustration of the match tests to more complex situations is given in Chapter 16.

Treatment of ties

We follow the usual procedure of giving tied observations a rank equal to the average of the ranks which would otherwise have been assigned to those tied observations. Then we need to decide how much these average ranks should

contribute (if anything) to the values of M_1 and M_2. We shall discuss the two statistics in turn.

Suppose that we are comparing two observations whose ranks are r_1 and r_2. A reasonable assessment of the contributions to M_1 is the following:

$$\text{if} \quad |r_1-r_2| \begin{cases} = 0 & \text{then contribute 1 (as usual)} \\ = \frac{1}{2} & \text{then contribute } \frac{1}{2} \\ > \frac{1}{2} & \text{then contribute 0} \end{cases}$$

To illustrate this procedure, consider the following data for seven columns and two rows:

Row 1	2.1	9.5	11.0	9.5	5.4	9.5	9.5
Row 2	3.2	6.1	10.3	6.1	4.0	6.1	8.0

These can be ranked as follows:

Row 1	1	$4\frac{1}{2}$	7	$4\frac{1}{2}$	2	$4\frac{1}{2}$	$4\frac{1}{2}$
Row 2	1	4	7	4	2	4	6
Contribution to M_1	1	$\frac{1}{2}$	1	$\frac{1}{2}$	1	$\frac{1}{2}$	0

which gives $M_1=4\frac{1}{2}$.

For the M_2 statistic we recommend the following procedure when ties are present:

$$\text{if} \quad |r_1-r_2| \begin{cases} = 0 & \text{then contribute 1} \\ = \frac{1}{2} \\ = 1 & \text{then contribute } \frac{1}{2} \\ = 1\frac{1}{2} \\ > 1\frac{1}{2} & \text{then contribute 0} \end{cases}$$

Let us see what this procedure produces for M_2 using the same data as above. We have the following ranks:

Row 1	1	$4\frac{1}{2}$	7	$4\frac{1}{2}$	2	$4\frac{1}{2}$	$4\frac{1}{2}$
Row 2	1	4	7	4	2	4	6
Contribution to M_2	1	$\frac{1}{2}$	1	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{2}$

which gives $M_2=5$.

Null distributions

The argument by which we derive the null distributions of the match statistics is the same as with Friedman's statistic. That is, in each row, any one of the $k!$ possible arrangements of the k ranks is equally likely to occur under H_0. So we may build up a frequency distribution for M_1 and M_2 in the case of n rows of k ranks by calculating the value of M_1 and M_2 for each of the $(k!)^n$ possible

configuraions. (Or again, without loss of generality, one may 'fix' the first row, leaving $(k!)^{n-1}$ configurations to examine.) The null probability distribution is then found by dividing the frequencies by the total number of such configurations, $(k!)^{n-1}$. Since this procedure is so like that for Friedman's statistic, we leave the details for you to check. As an example, the null distributions of M_1 and M_2 for $k=3$ and $n=3$ are given below for you to verify:

m	0	2	3	5	9
Prob($M_1=m$)	0.056	0.500	0.167	0.250	0.028
Prob($M_1 \geq$m)	1	0.944	0.444	0.278	0.028

m	3	4	5	7	9
Prob($M_2=m$)	0.056	0.333	0.417	0.167	0.028
Prob($M_2 \geq m$)	1	0.944	0.611	0.194	0.028

Large sample sizes

M_1 *test* For values of k or n not included in Table P, two useful approximate methods of finding critical values are available:

(a) If $k \geq 4$ then approximate critical values can be obtained from the Poisson distribution of mean $\frac{1}{2}n(n-1)$. Clearly, these critical values are not dependent on the number of columns k. Below we present a table of critical values for $n=3$ to 20 at significance levels $\alpha=5\%$ and 1% based on this approximation.

	Significance level			Significance level	
n	5%	1%	n	5%	1%
3	7	9	12	81	87
4	11	13	13	94	100
5	16	19	14	108	115
6	23	26	15	123	131
7	30	33	16	139	147
8	38	42	17	156	165
9	47	52	18	175	183
10	57	62	19	194	203
11	68	74	20	214	224

(b) If n is large then M_1 is approximately normally distributed with mean and variance both equal to $\frac{1}{2}n(n-1)$. So if z denotes the boundary of the right-hand tail area of α in the standard normal distribution, for example $z=1.6449$ if $\alpha=5\%$ (see Table A), then we approximate the 5% critical value for M_1 as

$$1.6449\sqrt{\tfrac{1}{2}n(n-1)} + \tfrac{1}{2}n(n-1) + \tfrac{1}{2}$$

the '$+\frac{1}{2}$' being incorporated as a continuity correction. So, for example, with $n=20$, $\frac{1}{2}n(n-1) = \frac{1}{2}\times 20\times 19 = 190$ and thus the 5% approximate critical value is

$1.6449\sqrt{190} + 190 + \frac{1}{2} = 213.7 = 214$ effectively, as shown in the table.

M_2 test For $k \geq 4$, approximate critical values of M_2 may be obtained using the normal distribution. For a significance level α and with z defined as above, the approximate critical value for M_2 is given by

$$z\sigma + \mu + \tfrac{1}{2}$$

where $\mu = \frac{1}{2}n(n-1)(2-1/k)$ and

$$\sigma = \sqrt{\frac{n(n-1)(3k^3-9k^2+6k+22)}{4k^2(k-1)}}$$

So, for example, with $k=4$ and $n=10$, $\mu=78.75$ and $\sigma=6.64$, giving a 5% approximate critical value as $1.6449 \times 6.64 + 78.75 + \frac{1}{2} = 90.2 = 91$ effectively.

MULTIPLE COMPARISONS

Conclusions such as 'there is evidence at a 5% (or 1%) level of significance to support the claim that the k population averages are not all equal' should not be regarded as the end of the story. Again, it is natural to follow such conclusions by looking for evidence of 'which populations are different from which?'. In Chapter 13, we used Dunn's multiple comparison method to answer such a question, and we will use the same method here.

With k populations under scrutiny, there are $\frac{1}{2}k(k-1)$ paired comparisons to be performed. Hence there are $\frac{1}{2}k(k-1)$ sets of hypotheses of the form:

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

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

The same comments regarding the overall significance level applies in this situation as in Chapter 13.

PRACTICAL PROBLEM

To illustrate the idea of multiple comparisons with k related samples, we return to the data on the four lens manufacturers. This is an appropriate situation to explore further since we have previously shown there to be evidence of differences in lens quality between the manufacturers.

HYPOTHESES

Appropriate hypotheses for this problem are:

$H_0^{(ij)}$: The ith and jth manufacturers produce lenses of equal average performance.

$H_1^{(ij)}$: Lenses from the ith and jth manufacturers differ in average quality.

where $i,j=1,2,3$ and 4 and $i<j$.

DATA

For convenience, we reproduce the ranked data below:

	A	B	C	D
	2	1	4	3
	1	2	4	3
	2	1	4	3
	2	1	4	3
	4	1	2	3
rank sums	11	6	18	15

STATISTIC

As we are working with a table of ranks, we base our procedure on the absolute differences $|D_{ij}| = |R_i - R_j|$, where R_i and R_j are the rank sums of the ith and jth columns, to examine a possible difference between the ith and jth populations. Unlike the equivalent procedure for k independent samples (see Ch. 13), there is no need to use mean ranks here since the number of ranks in each sample is always the same (namely, n). If there is no real difference between the two populations, $|D_{ij}|$ is likely to be small; on the other hand, a substantial difference between the ith and jth populations will tend to produce a large value of $|D_{ij}|$.

The $\frac{1}{2}k(k-1)=6$ differences for the above data are:

$$\begin{aligned}
|D_{12}| &= |R_1 - R_2| = 5 \\
|D_{13}| &= |R_1 - R_3| = 7 \\
|D_{14}| &= |R_1 - R_4| = 4 \\
|D_{23}| &= |R_2 - R_3| = 12 \\
|D_{24}| &= |R_2 - R_4| = 9 \\
|D_{34}| &= |R_3 - R_4| = 3
\end{aligned}$$

As with the multiple comparison procedure for k independent samples, the differences D_{ij} are approximately normally distributed, now with mean 0 and standard deviation $\sqrt{nk(k+1)/6}$. So if we divide each D_{ij} by the standard deviation $\sqrt{5\times4\times5/6} = 4.0824$ we obtain a 'standardized' statistic T_{ij}. The values of $|T_{ij}|$ are:

$$|T_{12}| = 1.2247$$
$$|T_{13}| = 1.7146$$
$$|T_{14}| = 0.9798$$
$$|T_{23}| = 2.9394$$
$$|T_{24}| = 2.2045$$
$$|T_{34}| = 0.7348$$

CRITICAL REGIONS

Since we conclude that the ith and jth populations differ if $|D_{ij}|=|R_i-R_j|$ is large, and consequently if $|T_{ij}|$ is large, the critical regions will take the form

$$|T_{ij}| \geqslant z$$

where z is the standard normal value corresponding to a single-tail probability of $\alpha/k(k-1)$. (The argument for this being the probability is identical to that in Chapter 13.)

For our example, taking $\alpha=20\%$, z must correspond to a probability of $\alpha/k(k-1)=0.20/12=0.067$. Table A gives $z=2.128$ and so $H_0^{(ij)}$ is rejected in favour of $H_1^{(ij)}$ if

$$|T_{ij}| \geqslant 2.128$$

SOLUTION

Comparing each of the six values of $|T_{ij}|$ reveals that there are significant differences between manufacturers:

(a) B and C, with the lenses from C having a higher resolution than those from B;
(b) B and D, with the lenses from D having a higher resolution than those from B.

We can represent this result pictorially as in Figure 14.1.



Figure 14.1

Assignments

1 Pulse readings

Consider the situation presented at the beginning of this chapter. A nurse took four readings of the pulses of three patients, the readings being obtained at specific times of the day. The data given were:

	Patient		
Time (h)	Kathryn	Fiona	Richard
0600	60	62	55
1000	65	66	65
1800	70	80	75
2200	65	72	68

Task: Analyse these data to see whether pulse readings differ significantly between (a) the patients and (b) the times of day.

2 Animal diet

In 1968, Koch and Sen ('Some aspects of the statistical analysis of the mixed model', *Biometrics* **24**, 27–48) examined the results of experiments undertaken at the Department of Pathology, Duke University Medical Centre, North Carolina by Dr N. Kaufmann and Dr J. V. Klavins. In one of their experiments, 16 animals were randomly placed into one of two groups – an experimental group which received ethionine in their diets, and a control group. The liver of each animal was divided into two parts, one of which was treated with radioactive iron and oxygen, and the other with radioactive iron and nitrogen. The data consist of the amount of iron absorbed by the variously treated liver portions; in the table the treatments are denoted by EO (ethionine–oxygen), EN (ethionine–nitrogen), CO (control–oxygen) and CN (control–nitrogen).

Pair	EO	EN	CO	CN
1	38.43	31.47	36.09	32.53
2	36.09	29.89	34.01	27.73
3	34.49	34.50	36.54	29.51
4	37.44	38.86	39.87	33.03
5	35.53	32.69	33.38	29.88
6	32.35	32.69	36.07	29.29
7	31.54	31.89	35.88	31.53
8	33.37	33.26	34.17	30.16

Task: Analyse these data to see whether there are significant differences in the amounts of iron absorbed in livers subjected to the four

treatments. Should you conclude that differences do exist, then perform Dunn's multiple comparison procedure.

3 Chemical analysis

Four students are each given two samples of four specimens to analyse. The data give the percentage by weight of a certain ingredient.

	Student			
Specimen	A	B	C	D
1	12.0	13.8	16.5	18.0
	16.2	15.8	12.0	15.0
2	18.2	18.5	15.9	14.5
	13.6	15.0	19.7	19.7
3	16.6	16.7	18.0	18.0
	16.8	15.2	18.4	18.1
4	17.0	18.7	17.6	20.2
	17.6	18.3	17.5	19.8

Task: Replace each of the 16 pairs of results with an average (usually the mean or median; here, with just two observations to average, the mean and the median are equal) and then carry out an analysis to see whether there are significant differences in the results between (a) the students and (b) the specimens.

Program 14.1 Friedman's test

```
100 REM *** FRIEDMAN'S  TEST ***

110 REM ** INPUT NUMBER OF TREATMENTS AND BLOCKS **
120 INPUT "NUMBER OF TREATMENTS AND BLOCKS ";K,N
130 DIM A(N,K),RK(N,K),R(K)

140 REM ** INPUT DATA **
150 FOR I=1 TO N
160 PRINT "INPUT THE OBSERVATIONS FOR BLOCK ";I
170 FOR J=1 TO K
180 INPUT A(I,J)
190 NEXT J
200 NEXT I

210 REM ** RANK DATA **
220 GO SUB 1000
```

```
230 REM ** CALCULATION OF TEST STATISTIC **
240 FOR I=1 TO K
250 LET R(I)=0
260 NEXT I
270 LET SUM=0
280 FOR I=1 TO K
290 FOR J=1 TO N
300 LET R(I)=RK(J,I)+R(I)
310 NEXT J
320 LET SUM=SUM+R(I)^2
330 NEXT I
340 LET M=12*SUM/(N*K*(K+1))-3*N*(K+1)
350 PRINT
360 PRINT "VALUE OF FRIEDMAN'S STATISTIC IS ";M

370 REM ** MULTIPLE COMPARISONS **
380 PRINT
390 PRINT "THE RANK DIFFERENCES FOR MULTIPLE COMPARISONS ARE:"
400 FOR I=1 TO K-1
410 FOR J=I+1 TO K
420 PRINT "D";I;J;" = ";ABS(R(I)-R(J))
430 NEXT J
440 NEXT I
450 PRINT
460 INPUT "INPUT THE VALUE OF Z ";Z
470 LET X=Z*SQRT(N*K*(K+1)/6)
480 PRINT "REJECT Ho IF ANY DIFFERENCE IS GREATER THAN ";X
490 STOP

1000 REM ** WITHIN-BLOCK RANKS **
1010 FOR I1 = 1 TO N
1020 FOR I=1 TO K
1030 LET T=0
1040 LET S=0
1050 LET C=A(I1,I)
1060 FOR J=1 TO K
1070 IF A(I1,J) > C THEN GO TO 1120
1080 IF A(I1,J) < C THEN GO TO 1110
1090 LET S=S+1
1100 GO TO 1120
1110 LET T=T+1

1120 NEXT J
1130 LET RK(I1,I)=T + (S + 1)/2
1140 NEXT I
1150 NEXT I1
1160 RETURN
```

Program 14.2 The match tests

```
100 REM *** MATCH TESTS - REPLACE LINES 230-360 OF PROGRAM 14.1 WITH: ***


230 REM ** CALCULATION OF TEST STATISTIC **
240 LET M1=0
250 LET M2=0
260 FOR I=1 TO N-1
270 FOR J=I+1 TO N
280 FOR G=1 TO K
290 IF RK(I,G)=RK(J,G) THEN LET M1=M1+1
300 LET RD=ABS(RK(I,G)-RK(J,G))
310 IF RD=0.5 THEN LET M1=M1+0.5
320 IF RD>0 AND RD<=1.5 THEN LET M2=M2+0.5
330 NEXT G
340 NEXT J
345 NEXT I
350 LET M2=M2+M1
360 PRINT "VALUES OF M1 AND M2 ARE ";M1;" AND ";M2

500 REM ** APPROXIMATE CRITICAL VALUES **
510 IF N<=20 AND K<=6 THEN STOP
520 INPUT "PERCENTAGE POINT OF STANDARD NORMAL DISTRIBUTION ";Z

530 REM ** VALUES FOR M1 **
540 LET B=N*(N-1)/2
550 PRINT "APPROXIMATE CRITICAL VALUE FOR M1 = ";Z*SQRT(B)+B+0.5

560 REM ** VALUES FOR M2 **
570 LET M=B*(2-1/K)
580 LET SD=SQRT(B*(3*K^3-9*K^2+6*K+22)/(4*K^2*(K-1)))
590 PRINT "APPROXIMATE CRITICAL VALUE FOR M2 = ";Z*SD+M+0.5
```

15

More than two (k) samples – ordered alternative hypothesis

Introduction

When discussing two-sample problems in Chapter 5, we saw that the form of the alternative hypothesis could be one of two possibilities:

(a) H_1: $\phi_1 \neq \phi_2$
(b) H_1: $\phi_1 > \phi_2$ (or $\phi_1 < \phi_2$)

where ϕ_1 and ϕ_2 are the two population medians.

The first type, corresponding to a two-sided test, is a general statement that there is some difference between the two population medians. In the k-sample problems of the previous two chapters, the equivalent alternative hypothesis is:

H_1: Not all of $\phi_1, \phi_2, \ldots, \phi_k$ are equal

where $\phi_1, \phi_2, \ldots, \phi_k$ are the k population medians. That is to say, we have a general statement that there are *some* differences among the k population averages.

Now consider the second type of alternative hypothesis, which corresponds to one-sided tests. Here we have a *more specific* statement saying that the median of one population is greater than the median of the other. Is there an equivalent statement for the k-sample problem? The answer is yes; we can have an alternative hypothesis of the form:

H_1: $\phi_1 > \phi_2 > \ldots > \phi_k$ (or $\phi_1 < \phi_2 < \ldots < \phi_k$).

This alternative hypothesis therefore specifies the *order* of any differences which may exist between the population medians. Such a statement is generally known as an *ordered* alternative hypothesis so as to clearly distinguish this situation from that described by the more *general* alternative hypothesis in the previous two chapters. (More accurately, this alternative hypothesis should be written using $\geqslant$ instead of $>$, with the proviso that at least one of the inequalities should be strict. For convenience, the simpler form of H_1 will be used in the text, but this type

of interpretation will always be appropriate. Note also that if the desired H_1 refers to some other ordering of the populations, it can be expressed in the form indicated in the text by simply renaming (renumbering) the populations and samples.)

The decision to have an ordered alternative hypothesis rather than a general alternative is dictated by the circumstances of the situation being examined. For example, the 'treatments' might be a number of increasingly expensive forms of advertising, and a natural alternative hypothesis might be that sales increase according to the advertising costs. Or we might simply want to test our own personal opinions as to the order of differences that exist. Or, in an agricultural experiment, we might want to test the alternative hypothesis that average yields increase according to the amount of a certain nutrient added to the growing compound.

In this chapter, we shall consider procedures for dealing with ordered alternatives for both independent and related samples.

THE TERPSTRA–JONCKHEERE TEST FOR k (>2) INDEPENDENT SAMPLES

Introduction

Although there are quite a few procedures available for the testing k independent samples with an ordered alternative hypothesis, the best (combining both ease and power) is probably T. J. Terpstra's (1952) test. The same test was also developed, quite independently, by A. R. Jonckheere (1954). Jonckheere also pointed out an interesting application of such a test to time series studies where observations on some items are taken at each of k successive time points and evidence is sought of an upward or downward trend.

Before proceeding further, we should remind ourselves just why the Kruskal–Wallis test is inappropriate for testing an ordered H_1 of the form

$$H_1: \phi_1 > \phi_2 > \ldots > \phi_k \qquad (\text{or } \phi_1 < \phi_2 < \ldots < \phi_k).$$

The Kruskal–Wallis test will reject H_0 whenever there is sufficient difference between the rank sums, *irrespective* of their order; thus a significant result in the Kruskal–Wallis test cannot be held as evidence for a *particular* ordered H_1.

PRACTICAL PROBLEM

It is often presumed that more expensive brands of goods are of a higher quality than their less expensive counterparts, i.e. quality increases with

price. From each of three stores we bought six 'own-brand' light bulbs. The stores were categorized as:

A a national 'low-priced' supermarket
B an international general stores
C a high-class department store

HYPOTHESES

In view of our comment regarding price and quality, we are clearly interested in the claim that the average lifetime of bulbs is in direct proportion to their price. If we denote the average lifetime of bulbs from the three stores by m_A, m_B and m_C respectively, the hypotheses may be concisely written as:

$$H_0: m_A = m_B = m_C$$
$$H_1: m_A < m_B < m_C$$

Thus, as usual, H_0 expresses the belief that there are no differences in quality (as measured by average lifetimes) between the three brands. But now H_1 is a much more specific statement about the relative performance of the bulbs, rather than merely saying that H_0 is untrue.

DATA

The lifetimes, in hours, of the three samples of bulbs are given below. Unfortunately, one of brand A's bulbs was broken before it could be tested, so our sample sizes are $n_1=5$, $n_2=6$ and $n_3=6$. We have presented the data so that the order of the samples (A, B, C) corresponds to that postulated in H_1.

A	B	C
619	343	3670
35	2437	2860
126	409	502
2031	267	2008
215	1953	5004
	1804	4782

It is well known that the lifetimes of light bulbs are well modelled by non-symmetric distributions called exponential distributions, and so attempts to analyse such data using classical techniques, which rely on normality, would be untrustworthy.

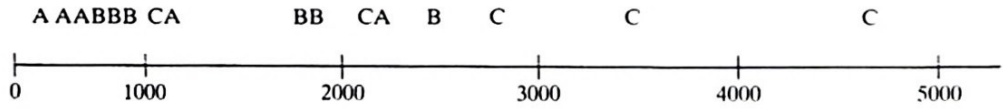
STATISTIC

The Terpstra–Jonckheere statistic is formed by calculating the Mann–Whitney U statistic for each possible pair of samples and adding all the resulting U values. As in the one-sided version of the Mann–Whitney test in Chapter 5, the particular U statistic to be calculated in each case is the one which should yield *small* values if the alternative hypothesis is true. So, in our particular case where we have three samples and are testing H_0 against the alternative H_1: $m_A<m_B<m_C$, we will have the following three U statistics to add:

(a) U_{BA} = the number of (A,B) pairs with the B less than the A;
(b) U_{CA} = the number of (A,C) pairs with the C less than the A;
(c) U_{CB} = the number of (B,C) pairs with the C less than the B.

Now if H_1 is true, then some or all of these U statistics will tend to be small, and so therefore will their sum $W=U_{BA}+U_{CA}+U_{CB}$; otherwise W will tend to be rather larger. (Actually, Terpstra and Jonckheere used a different, but equivalent, version of W; our definition generally involves simpler arithmetic.)

To ease the calculation of W, we adopt a similar technique to that used for the Mann–Whitney statistic in Chapter 5, which is particularly useful for large samples. On a roughly linear scale we indicate, using the appropriate sample letter, the position of the observations. This is shown below.



Then

$$U_{BA} = 0 + 0 + 0 + 3 + 5 = 8$$
$$U_{CA} = 0 + 0 + 0 + 1 + 2 = 3$$
$$U_{CB} = 0 + 0 + 0 + 1 + 1 + 2 = 4$$

Hence the value of the test statistic is $W = 8 + 3 + 4 = 15$.

CRITICAL REGIONS

We know that it is small values of W that support H_1, and so the critical regions are of the form

$$W \leqslant \text{critical value}$$

Table R gives critical values of W for various values of k and sample sizes.

In our example, $k=3$, $n_1=5$, $n_2=6$ and $n_3=6$, and so critical regions for significant levels of 5% and 1% are $W \leqslant 28$ and $W \leqslant 21$ respectively.

SOLUTION

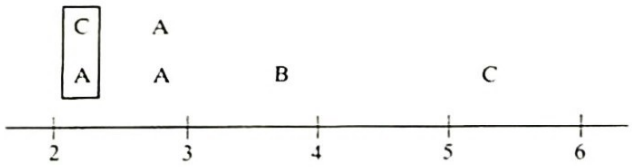
The value of W obtained above was $W=15$, and this lies well in the 1% critical region. A glance at the data will convince the reader that a result of this sort was to be expected! So the test has given us very strong evidence in this instance to support the claim that the average lifetime of the 'own-brand' bulbs increases with price.

Ties

With the Terpstra–Jonckheere test, tied observations are only troublesome if they occur in different samples; ties within the same sample may be treated as 'ordinary' observations (why?). When ties do occur between observations in different samples, the contribution to the Mann–Whitney count is taken as $\frac{1}{2}$, as in Chapter 5. As an example, consider the following three small samples:

A	B	C
2.6	4.1	2.6
3.2		5.7
3.2		

Again portraying the data on an approximate linear scale we have:



where the tie at 2.6 is indicated by a box. Presuming H_1 is as in our main example, we compute

$$\begin{aligned} U_{BA} &= 0 + 0 + 0 = 0 \\ U_{CA} &= \tfrac{1}{2} + 1 + 1 = 2\tfrac{1}{2} \\ U_{CB} &= 1 \qquad\quad = 1 \end{aligned}$$

giving $W = 0+2\frac{1}{2}+1 = 3\frac{1}{2}$.

Null distributions

The principle for deriving the null distribution of W follows the familiar pattern. Interpreting H_0 as implying that the data come from identical (continuous) distributions, all possible letter sequences are then equally likely to occur. As

usual, the principle is that W is computed from each such letter sequence, a frequency distribution is thereby built up, and the null distribution obtained by dividing each frequency by the total frequency. Also as usual, the trouble is that this total frequency gets very large, even for quite small sample sizes; for example, in the above illustration with $k=3$, $n_1=5$ and $n_2=n_3=6$, the number of letter sequences to be considered is

$$\frac{N!}{n_1!n_2!n_3!} = \frac{17!}{5!6!6!} = 5717712$$

Furthermore, although the reasoning behind this null distribution is similar to that for the Kruskal–Wallis statistic, we cannot use a symmetry argument to reduce the work; for example,

A	B	C	and	A	B	C
1,2	3,4	5,6		5,6	1,2	3,4

give the same value for the Kruskal–Wallis statistic but have W values of $0+0+0=0$ and $4+4+0=8$ respectively.

Fortunately, modern computers, aided by a recursion formula given in Jonckheere's paper, have produced a set of critical values for quite a wide range of sample sizes.

Large sample sizes

It can be shown that, as the sample sizes n_i increase, the null distribution of W gets close to normality. The mean of the distribution is

$$\mu=(N^2-\sum n_i^2)/4$$

and its standard deviation is

$$\sigma=\sqrt{[N^2(2N+3)-\sum n_i^2(2n_i+3)]/72}$$

This can be used to obtain approximate critical values when the n_i are beyond the range of Table R. For a significance level α, we reject H_0 in favour of H_1 if

$$W \leqslant \mu - z\sigma - \tfrac{1}{2}$$

where z is the corresponding (right-hand tail) standard normal value and the $\frac{1}{2}$ is the continuity correction.

To see how well this approximation performs, again consider $k=3$, $n_1=5$, $n_2=n_3=6$. We find

$$\mu = [17^2-(25+36+36)]/4=48$$

and

$$\sigma = \sqrt{[17^2\times37-(25\times13+36\times15+36\times15)]/72} = 11.36$$

For $\alpha=5\%$, $z=1.6449$ and so H_0 is rejected if $W \leqslant 48-1.6449\times 11.36-0.5 = 28.82$. The exact (conservative) critical value is 28, which in practice corresponds exactly with this approximate value. Thus, even for sample sizes as low as these, the approximation works well.

k (>2) RELATED SAMPLES

In the previous chapter we used Friedman's test and the match tests to analyse the problem of k related samples. When the nature of the practical problem demands an ordered alternative hypothesis, these tests are inappropriate for exactly the same reason as given earlier in the chapter concerning the Kruskal–Wallis test. Instead, we shall analyse such problems using a test due to E. B. Page (1963) and two tests based on the matching principle used in the previous chapter. The match tests, recently developed by one of the authors, are particularly quick and easy to use.

PAGE'S TEST

Introduction

Page's test is a well known distribution-free test for ordered alternatives in the k-related-samples problem. It is easy to apply and its underlying logic is straightforward. The assumptions of the test are as for Friedman's test, namely that the data be in the form of a 'randomized block design' of, say, k columns (treatments) and n rows (blocks). The theory also assumes the populations have continuous distributions; but, as before, this assumption can be relaxed in practice if it does not result in too many ties.

PRACTICAL PROBLEM

An international company is dissatisfied with the quality of output from six of its subsidiary companies. These companies manufacture various porcelain articles ranging from cups and saucers to toilets. The parent company instigates a four-month training programme for the managers of these subsidiaries. If the programme is effective, the quality should rise over the months. No doubt the parent company will interpret any such rise as validating their training programme.

HYPOTHESES

The null hypothesis is of the same form as for any k-related-samples problem: that is, there is no difference in the average quality of the

output over the five-month period (observations were also included for the month immediately preceding the start of the programme so as to have a standard or control with which to compare subsequent observations). Since the company is interested in seeing whether the training programme is effective, the appropriate alternative hypothesis is that the average quality of the output is increasing over time. If we denote the average quality of the output for the five months by m_1, m_2, m_3, m_4 and m_5 respectively, the hypotheses may be written as:

$$H_0: m_1=m_2=m_3=m_4=m_5.$$
$$H_1: m_1<m_2<m_3<m_4<m_5.$$

(Note: It is important for the construction of the test statistic that H_1 be in this particular form, i.e. $m_1<m_2<\ldots<m_k$, rather than $m_1>m_2>\ldots>m_k$ or any other of the possibilities. If the required ordering is not this particular one, then the populations and samples must be renumbered so that it becomes so.)

DATA

The observations record the number of acceptable items as a percentage of the total output for each subsidiary. Month 1 indicates the control month, and months 2 to 5 cover the training period.

	Month				
Subsidiary	1	2	3	4	5
A	62.59	67.11	73.02	92.34	83.48
B	55.75	63.03	63.93	71.61	93.73
C	65.88	69.89	82.53	77.33	84.82
D	66.13	66.79	85.37	76.72	98.88
E	65.73	57.85	63.08	71.75	87.04
F	78.36	76.17	82.97	91.49	89.61

STATISTIC

As with Friedman's statistic, the observations within each block are ranked from 1 to k (1 to 5 in our example). Evidently, if H_0 is true, these ranks should be evenly distributed among the five columns. On the other hand, if H_1 is true (i.e. $m_1<m_2<m_3<m_4<m_5$), we would expect the ranks to tend to increase as we go from column 1 to column 5. Page's test statistic is obtained by summing the ranks in each column to produce the rank sums $R_1, R_2, \ldots, R_k$ and then multiplying R_1 by 1, R_2 by 2, R_3 by 3 and so on, and finally adding these k quantities to obtain:

$$L = R_1 + 2R_2 + 3R_3 + \ldots + kR_k$$

or, more concisely,

$$L = \sum_{i=1}^{k} iR_i$$

Thus the rank sums that we expect to be largest when H_1 is true have been multiplied by the biggest constants, while those expected to be the smallest have been multiplied by the smaller constants.

Let us now calculate Page's statistic for the data in our practical problem. The table of ranks for the data is given below; the rank-sums R_i are also shown.

	Month				
Subsidiary	1	2	3	4	5
A	1	2	3	5	4
B	1	2	3	4	5
C	1	2	4	3	5
D	1	2	4	3	5
E	3	1	2	4	5
F	2	1	3	5	4
rank sum, R_i	9	10	19	24	28

Hence

$$L = \sum_{i=1}^{5} iR_i$$

$$= 1\times9 + 2\times20 + 3\times19 + 4\times24 + 5\times28$$
$$= 9 + 20 + 57 + 96 + 140$$
$$= 312$$

CRITICAL REGIONS

Clearly it is large values of L that support H_1. Thus H_0 will be rejected in favour of H_1 if

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

Table S gives critical values of L for $k=3$ to 6 and n up to 20.

In our practical problem, $k=5$ and $n=6$, and so the critical regions for $\alpha=5\%$ and 1% are $L\geqslant291$ and $L\geqslant299$ respectively.

SOLUTION

The calculated value of $L=312$ certainly falls well into the 1% critical region. It would appear that the company has quite conclusive evidence supporting the effectiveness of their programme.

Null distributions

The argument is almost identical to that for Friedman's statistic in Chapter 14. The only difference is that we cannot reduce the $(k!)^n$ configurations to $(k!)^{n-1}$ by 'fixing' the first block. For example, in the simple case where $k=3$ and $n=2$, the two configurations

$$\begin{array}{ccc} 1 & 2 & 3 \\ 2 & 1 & 3 \end{array} \quad \text{and} \quad \begin{array}{ccc} 2 & 3 & 1 \\ 1 & 3 & 2 \end{array}$$

would automatically result in the same value of Friedman's M, but give $L=27$ and $L=24$ respectively.

Large sample sizes

For values of n beyond the range given in Table S, we can use the fact that, for large values of n, L is approximately normally distributed. Under H_0, the mean of L is

$$\mu=nk(k+1)^2/4$$

and its standard deviation is

$$\sigma=\sqrt{nk^2(k+1)(k^2-1)/144}$$

So, for a given significance level α, the approximate critical region is given by

$$L \geqslant \mu + z\alpha + \tfrac{1}{2}$$

where z is the corresponding (right-hand tail) value from the standard normal distribution and the $\frac{1}{2}$ is the continuity correction.

To see how well this approximation performs, consider the values $k=5$ and $n=6$ as in our practical problem. The values of μ and σ are 270 and 12.247 respectively. Hence the approximate critical regions are given by

$$L \geqslant 270 + 12.247z + \tfrac{1}{2}$$

For $\alpha=5\%$ and 1%, for which $z=1.6449$ and 2.3263 respectively, we obtain $L\geqslant 290.65$ and $L\geqslant 298.99$ which correspond exactly, in practice, with the exact (conservative) critical regions of $L\geqslant 291$ and $L\geqslant 299$.

THE MATCH TESTS FOR ORDERED ALTERNATIVES

Introduction

The matching principle, introduced in the last chapter, can be applied in an obvious way to the case of ordered alternatives. We shall present two statistics, L_1 and L_2; the first is based solely on matches while the other incorporates the idea of near-matches. Not only are the tests based on these statistics simple to use, but they also possess very respectable power, particularly L_2, which utilizes more information than does L_1. The simple L_1 has the attraction of (approximate) critical values which depend only on n, the number of rows; i.e. these approximate critical values are effectively independent of the number of treatments.

PRACTICAL PROBLEM, HYPOTHESES AND DATA

These are as for Page's test.

STATISTICS

As previously, we shall describe the nature of the statistics with reference to the practical problem.

Both statistics rely on the fact that the strongest evidence for the truth of H_1 is found when column 1 mainly has ranks of 1, column 2 has ranks of 2, column 3 has ranks of 3, and so on. All we do is to compare the ranks in each row in turn with the ranks predicted under H_1 (for example, the predicted ranks could be 1,2,3, . . . ,k). In each comparison, we record the number of matches and near-matches. The two statistics, L_1 and L_2 are then defined as:

$$L_1 = \text{total number of matches}$$
$$L_2 = L_1 + \tfrac{1}{2}(\text{number of near-matches})$$

Let us now apply these simple statistics to the practical problem. Below we reproduce the ranked data in addition to the predicted ranks.

Predicted ranks	1	2	3	4	5
Subsidiary	Actual ranks				
A	1	2	3	4	5
B	1	2	3	4	5
C	1	2	4	3	5
D	1	2	4	3	5
E	3	1	2	4	5
F	2	1	3	5	4

Performing the comparisons row by row, we obtain:

$$L_1 = 5 + 5 + 3 + 3 + 2 + 1 = 19$$

and

$$L_2 = 19 + \tfrac{1}{2}(0 + 0 + 2 + 2 + 2 + 4) = 24$$

CRITICAL REGIONS

The L_1 test

Clearly, if the ranks of the observed data follow the predicted order, then this results in a high value of L_1, i.e. it is large values of L_1 that support H_1. Thus we reject H_0 in favour of H_1 if

$$L_1 \geqslant \text{critical value}$$

Table T gives critical values of L_1 for $k=3$ to 6 and n up to 20.

In our problem, $k=5$ and $n=6$, and so critical regions corresponding to $\alpha=5\%$ and 1% are $L_1 \geqslant 11$ and $L_1 \geqslant 13$ respectively.

The L_2 test

Similarly, large values of L_2 provide evidence in support of H_1. Thus critical regions are of the form

$$L_2 \geqslant \text{critical value}$$

Table U gives the critical values of L_2 for $k=3$ to 6 and n up to 20.

For our problem where $k=5$ and $n=6$, the critical regions corresponding to $\alpha=5\%$ and 1% are $L_2 \geqslant 15$ and $L_2 \geqslant 17$ respectively.

SOLUTION

Both the L_1 test and the L_2 test give results that lie well into their 1% critical regions. These conclusions are consistent with those obtained using Page's test.

Discussion

It is interesting to note that the process of looking for matches helps to bring our notice to those columns that 'disagree' with the predicted order. For example, there appears to be a slight disagreement in columns 3 and 4. In other words, applying these simple tests brings us 'closer to the data' – which is no bad thing in this age of automation.

The fact that $L2$ utilizes more information than L_1 is reflected in the relative performances of the corresponding tests. Computer simulation studies have shown that L_2 has consistently higher power than L_1, although even the simple test based on L_1 performs very creditably (see Worthington 1982).

Ties

Tied observations are tackled in a similar manner as with the M_1 and M_2 tests of the previous chapter. As usual, ties are assigned mean ranks. In the following discussion, r_i refers to a rank in the ith column.

When using the L_1 test, the contributions to L_1 are as follows:

$$\text{if} \quad |r_i - i| \begin{cases} = 0 & \text{then contribute } 1 \\ = \frac{1}{2} & \text{then contribute } \frac{1}{2} \\ > \frac{1}{2} & \text{then contribute } 0 \end{cases}$$

To illustrate this rule, consider the following block of ranks:

Predicted	1	2	3	4	5
Observed	1	$2\frac{1}{2}$	4	$2\frac{1}{2}$	5
Contribution to L_1	1	$\frac{1}{2}$	0	0	1

When using the L_2 test, the contributions are as follows:

$$\text{if} \quad |r_i - i| \begin{cases} = 0 & \\ = \frac{1}{2} & \\ = 1 & \\ = 1\frac{1}{2} & \end{cases} \quad \begin{cases} \text{then contribute } 1 \\ \text{then contribute } \frac{1}{2} \\ \text{then contribute } 0 \end{cases}$$

The above ranks give

$$L_2 = 1 + \tfrac{1}{2} + \tfrac{1}{2} + \tfrac{1}{2} + 1 = 3\tfrac{1}{2}$$

Null distributions

If H_0 is true, thus implying that the k populations have identical (continuous) distributions, we would expect the ranks to be fairly evenly spread between the treatments. Thus, for given values of k and n, the null distributions of L_1 or L_2 may be obtained by computing their value for each of the $(k!)^n$ rank configurations.

For example, with $k=3$ and $n=1$ there are $3!=6$ different rank configurations and, with the predicted order under H_1 being 1 2 3, the null distribution of L_2 is derived as follows:

Predicted ranks	1 2 3	1 2 3	1 2 3	1 2 3	1 2 3	1 2 3
Possible ranks	1 2 3	1 3 2	2 1 3	2 3 1	3 1 2	3 2 1
Value of L_2	$1+1+1$ $=3$	$1+\frac{1}{2}+\frac{1}{2}$ $=2$	$\frac{1}{2}+\frac{1}{2}+1$ $=2$	$\frac{1}{2}+\frac{1}{2}+0$ $=1$	$0+\frac{1}{2}+\frac{1}{2}$ $=1$	$0+1+0$ $=1$

Hence the null distribution is

$$\begin{aligned} \text{Prob}(L_2=1) &= 3/6 \\ \text{Prob}(L_2=2) &= 2/6 \\ \text{Prob}(L_2=3) &= 1/6 \end{aligned}$$

Once the null distributions for $n=1$ are found, subsequent distributions for $n>1$ are easily obtained by a process called convolution. This has enabled extensive tables of exact critical values to be produced.

Large samples

The L_1 *test* If k is large, then it can be shown that the null distribution of L_1 is approximately Poisson with a mean of n. This enables approximate critical values to be obtained for values of k not included in Table T; and indeed these approximate critical values are thus *independent* of k. Below is a table of these approximate critical values for $\alpha=5\%$ and 1%.

	α	
n	5%	1%
2	6	7
3	7	9
4	9	10
5	10	12
6	11	13
7	13	15
8	14	16
9	15	18
10	16	19

Incidentally, the exact critical values for our practical problem where $k=5$ and $n=6$ are precisely the same as those obtained from this table of approximate values.

The L_2 *test* For values of n greater than those included in Table U, we may obtain critical values of L_2 using a normal approximation. The mean and standard deviation of L_2 are given by

$$\mu=n(2-1/k)$$

and

$$\sigma = \sqrt{\frac{n}{k}\left(\frac{3(k-2)}{2}+\frac{1}{k(k-1)}\right)}$$

So, for a given significance level α, the approximate critical region is

$$L_2 \geqslant \mu+z\sigma+\tfrac{1}{2}$$

where z is the corresponding (right-hand tail) percentage point of the standard normal distribution and the $\frac{1}{2}$ is the continuity correction. To see how this approximation performs, consider again the case when $k=5$ and $n=6$. With these values we find that $\mu=10.8$ and $\sigma=2.337$. The approximate critical regions are then $L_2\geqslant 10.8+2.337z+\frac{1}{2}$. For $\alpha=5\%$ and 1%, we obtain $L_2\geqslant 15.14$ and

$L_2 \geqslant 16.74$. The exact (conservative) critical regions are $L_2 \geqslant 15$ and $L_2 \geqslant 17$ respectively; so here the approximation has produced the correct 1% region but is over-conservative as regards the 5% region.

Assignments

1 Light conditions

An experiment was conducted to investigate the effect of different lighting conditions on manual dexterity. Four lighting conditions were chosen for the test:

A Light from a window
B Fluorescent light
C Tungsten bulb light
D Candle light

Eighteen people volunteered to take part in the experiment. They were randomly assigned to four groups and each group was subjected to one of the lighting conditions. The time (seconds) taken by each person to complete a simple manual task is given in the following table.

A	B	C	D
29	63	41	41
43	15	19	59
44	27	37	36
46	30	27	36
	31		40

Task: Choose an appropriate test to investigate the claim that the average time to complete the manual task increases with worsening light conditions.

2 Food deprivation

In 1972 G. J. Syme and J. S. Pollard ('The relation between differences in level of food deprivation and dominance in food getting in rats', *Psychon. Sci.* **29**, 297–8) presented the results of their experiment into the effect of food deprivation in rats. The table below gives the amount of food (g) eaten by eight rats under three levels of food deprivation:

	Hours of food deprivation		
Rat	0	24	72
1	3.5	5.9	13.9
2	3.7	8.1	12.6
3	1.6	8.1	8.1
4	2.5	8.6	6.8
5	2.8	8.1	14.3
6	2.0	5.9	4.2
7	5.9	9.5	14.5
8	2.5	7.9	7.9

The second table gives the times (seconds) spent at the food trough by the rats:

	Hours of food deprivation		
Rat	0	24	72
1	570	706	840
2	589	520	781
3	267	784	817
4	425	753	799
5	666	861	849
6	653	816	810
7	746	790	854
8	785	810	797

Task: Investigate the claims that (i) the amount of food eaten tends to increase with length of deprivation, and (ii) the time spent at the trough tends to increase with length of deprivation.

Program 15.1 Terpstra–Jonckheere test for ordered alternatives

```
100 REM *** TERPSTRA-JONCKHEERE TEST ***

110 REM ** INPUT NUMBER OF SAMPLES **
120 INPUT "NUMBER OF SAMPLES (>2) ";K
130 DIM N(K)
140 LET GN=0

150 REM ** INPUT DATA **
160 FOR I=1 TO K
170 PRINT "INPUT SIZE OF SAMPLE ";I
180 INPUT N(I)
190 LET GN=GN+N(I)
200 NEXT I
210 PRINT
220 DIM D(GN,2)
230 LET L=0
```

```
240 FOR I=1 TO K
250 PRINT "INPUT DATA FOR SAMPLE ";I
260 FOR J=1 TO N(I)
270 LET L=L+1
280 INPUT D(L,1)
290 LET D(L,2)=I
300 NEXT J
310 NEXT I

320 REM ** CALCULATION OF TEST STATISTIC **
330 LET U=0
340 FOR I=1 TO K-1
350 FOR I1=I+1 TO K
360 FOR J=1 TO GN
370 IF D(J,2)<>I THEN GO TO 430
380 FOR J1=1 TO GN
390 IF D(J1,2)<>I1 THEN GO TO 420
400 IF D(J,1)>D(J1,1) THEN LET U=U+1
410 IF D(J,1)=D(J1,1) THEN LET U=U+.5
420 NEXT J1
430 NEXT J
440 NEXT I1
450 NEXT I
460 PRINT "VALUE OF THE TEST STATISTIC IS ";U
470 PRINT
480 IF K<=6 AND N<=10 THEN STOP

490 REM ** APPROXIMATE CRITICAL VALUES **
500 LET S1=0
510 LET S2=0
520 FOR I=1 TO K
530 LET S1=S1+N(I)^2
540 LET S2=S2+(2*N(I)+3)*N(I)^2
550 NEXT I
560 LET M=(GN^2-S1)/4
570 LET S=SQRT((GN^2*(2*GN+3)-S2)/72)
580 INPUT "VALUE OF Z: ";Z
590 LET C=M-Z*S+.5
600 PRINT "APPROXIMATE CRITICAL VALUE IS ";C
```

Program 15.2 Page's test for ordered alternatives

```
100 REM *** PAGE'S TEST FOR ORDERED ALTERNATIVES ***

110 REM ** INPUT NUMBER OF TREATMENTS AND BLOCKS **
120 INPUT "NUMBER OF TREATMENTS AND BLOCKS ? ";K,N
130 DIM A(N,K),RK(N,K)

140 REM ** INPUT DATA **
150 FOR I=1 TO N
160 PRINT "INPUT THE OBSERVATIONS FOR BLOCK ";I
170 FOR J=1 TO K
180 INPUT A(I,J)
190 NEXT J
200 NEXT I
210 GO SUB 1000
```

Continued overleaf

Program 15.2 Continued

```
220 REM ** CALCULATION OF TEST STATISTIC **
230 DIM R(K)
240 FOR I = 1 TO K
250 LET R(I) = 0
260 FOR J = 1 TO N
270 LET R(I) = R(I) + RK(J,I)
280 NEXT J
290 NEXT I
300 LET L = 0
310 FOR I=1 TO K
320 LET L = L + I*R(I)
330 NEXT I
340 PRINT "VALUE OF PAGE'S STATISTIC = ";L
350 PRINT
360 IF N <= 20 AND K <= 6 THEN STOP

370 REM ** APPROXIMATE CRITICAL VALUES **
380 INPUT "PERCENTAGE POINT OF STANDARD NORMAL DISTRIBUTION ";Z
390 LET LC = N*(K^3-K)*(Z/SQRT(N*(K-1))+3*(K+1)/(K-1))/12
400 PRINT "APPROXIMATE CRITICAL VALUE = ";LC

1000 REM ** WITHIN-BLOCK RANKS - USE LINES 1000-1160 OF PROGRAM 14.1 **
```

Program 15.3 Match tests for ordered alternatives

```
100 REM *** MATCH TESTS FOR ORDERED ALTERNATIVES ***

110 REM ** INPUT NUMBER OF TREATMENTS AND BLOCKS **
120 INPUT "NUMBER OF TREATMENTS AND BLOCKS ? ";K,N
130 DIM A(N,K),RK(N,K)

140 REM ** INPUT DATA **
150 FOR I=1 TO N
160 PRINT "INPUT THE OBSERVATIONS FOR BLOCK ";I
170 FOR J=1 TO K
180 INPUT A(I,J)
190 NEXT J
200 NEXT I
210 GO SUB 1000

220 REM ** CALCULATION OF TEST STATISTIC **
230 LET L1=0
240 LET L2=0
250 FOR I=1 TO K
260 FOR J=1 TO N
270 LET RD=ABS(RK(J,I)-I)
280 IF RD=0 THEN LET L1=L1+1
290 IF RD=0.5 THEN LET L1=L1+0.5
300 IF RD>0 AND RD<=1.5 THEN LET L2=L2+0.5
310 LET L2=L2+L1
320 NEXT J
330 NEXT I
340 PRINT "VALUE OF MATCH L1 STATISTIC = ";L1
350 PRINT "VALUE OF MATCH L2 STATISTIC = ";L2
360 PRINT
370 IF N<=20 AND K<=6 THEN STOP

380 REM ** APPROXIMATE CRITICAL VALUES **
390 INPUT "PERCENTAGE POINT OF STANDARD NORMAL DISTRIBUTION ";Z
```

```
400 REM ** VALUES FOR L1  **
410 LET LC=SQRT(N)*Z+N+0.5
420 PRINT "APPROXIMATE CRITICAL VALUE FOR L1 = ";LC

430 REM ** VALUES FOR L2 **
440 LET M=N*(2-1/K)
450 LET SD=SQRT(N*(3*(K-2)/2+1/K/(K-1))/K)
460 PRINT "APPROXIMATE CRITICAL VALUE FOR L2 = ";M+Z*SD+0.5

1000 REM ** WITHIN-BLOCK RANKS - USE LINES 1000-1160 OF PROGRAM 14.1 **
```

16

Tests for interaction and latin squares designs

Introduction

In the first part of this chapter, we present procedures for detecting whether two variables or *factors* (such as the dosage of a medicine and the amount of alcohol consumed) *interact* with each other. Naturally, interaction between two variables can indicate erratic behaviour patterns, and so it is important to test for its presence.

As in Chapter 14, the second part of this chapter deals with the *k*-related samples problem. However, here the observations are affected by a third factor and, in the particular situation we consider, the data are arranged according to the so-called latin squares design. As we shall see, with this situation we can examine *three* sets of hypotheses – one for each of the three factors. To analyse such a design we present a procedure in which either Friedman's test or the match tests of Chapter 14 may be used.

INTERACTION

The layman's meaning of the term *interaction* is 'the action of one on another', which suggests that some quantity is involved with another quantity (or quantities) in a possibly complex manner. Interaction has the same kind of interpretation in statistics. For example, in an experiment on the effects of two variables, these two variables in conjunction with each other may lead to dramatically different results than if they are considered singly. For instance, suppose they represent quantities of alcohol and aspirins consumed by a foolish volunteer!

Consider an experiment in which three grades of seed (A=poor, B=medium and C=best-quality) are subjected to four different strengths of fertilizer (I=20, II=30, III=45 and IV=60 kg/acre). The kind of results that common sense might predict, i.e. yield increasing with higher seed quality and also yield increasing with the strength of fertilizer, are shown below, for yields (kg/m^2) obtained:

Fertilizer strength	Seed quality		
	A	B	C
I	2	5	8
II	3	6	9
III	5	8	10
IV	6	11	17

But now compare these responses with those from a similar experiment conducted on a different (though with similar characteristics) plot of land.

Fertilizer strength	Seed quality		
	A	B	C
I	2	7	3
II	4	3	8
III	6	5	4
IV	10	6	9

Although we would have not expected exactly the same yields as before, we would have expected a similar kind of behaviour. Figure 16.1 shows the behavioural patterns for the two experiments. In the second experiment, *something* has caused this interaction between the seed quality and fertilizer.

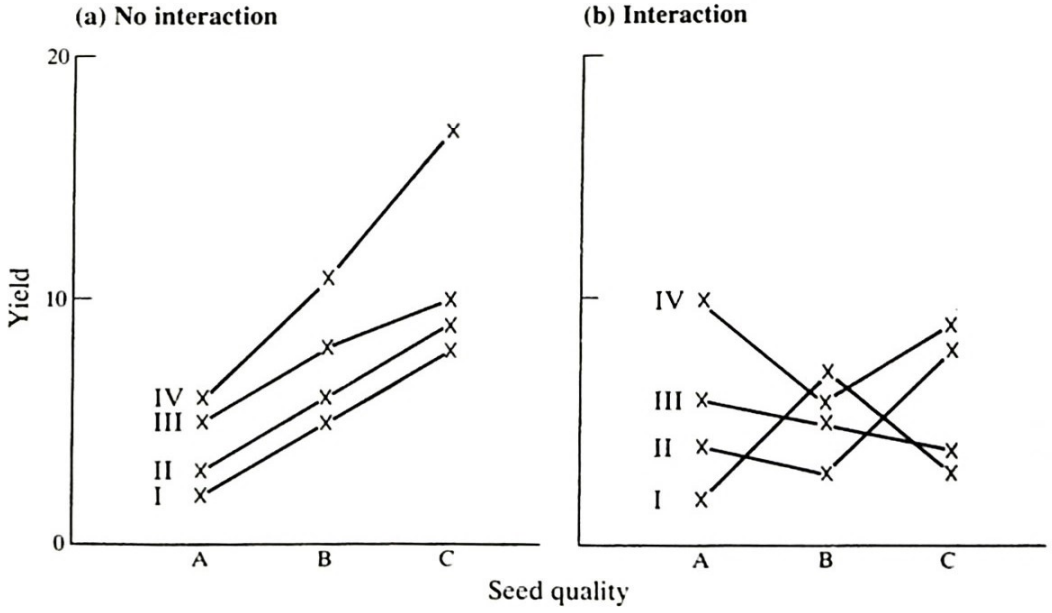


Figure 16.1

Naturally, experiments could involve three, or more, factors (or treatments). Then we could investigate the presence of interaction between any two of the factors, or between any three, and so on. However, in this book we confine our thoughts to experiments with just *two* factors (as in the above illustration). Having said that, even in experiments with only two factors, we do need to distinguish between two types of experiment – those with *linked* observations and those with *unlinked* observations. To see what we mean by these terms, consider the following tables. At first glance, these tables appear to be identical – however, look more closely.

Linked observations		Treatment		
		A	B	C
I	1	7.6	7.2	8.2
	2	6.9	7.3	7.8
II	1	5.6	7.1	7.2
	2	6.8	6.7	8.0
III	1	7.4	8.2	7.4
	2	6.9	7.9	7.5

Unlinked observations	Treatment		
	A	B	C
I	7.6	7.2	8.2
	6.9	7.3	7.8
II	5.6	7.1	7.2
	6.8	6.7	8.0
III	7.4	8.2	7.4
	6.9	7.9	7.5

In the first table, the two observations for each treatment combination have some kind of natural order, which we have indicated by assigning an 'observation number', 1 or 2, to each. This natural ordering may arise by observation 1 relating to day 1 of the experiment and observation 2 to day 2, or it may be that the observation number refers to the particular person obtaining the observation. Whatever the reason, this ordering allows us to link the observations in a certain way; for example, the 7.6 (A,I) is linked with the 7.2 (B,I) and also to the 8.2 (C,I); likewise, the 6.8 (A,II), the 6.7 (B,II) and the 8.0 (C,II) are linked. In the second table, there is no ordering of the observations – the order in which the observations are recorded is immaterial.

It is interesting to note that some authors suggest that it is good enough

to impose some arbitrary ordering; this then allows them to use a test like Wilcoxon's interaction test (see below). Clearly, the danger with this practice is that someone else could choose a different ordering and then possibly end up with a quite different conclusion. For situations where the observations are unlinked, we shall present procedures based on the match tests of Chapter 14.

WILCOXON'S INTERACTION TEST

Introduction

Frank Wilcoxon (1949) wrote an interesting booklet introducing a number of supposedly quick statistical methods. One of these was a technique for testing for interaction when the observations are linked. His idea was to reduce the data (by a sequence of subtractions) to leave only the interaction effect (if any). He then applied Friedman's test to these reduced data.

In common with most other tests for interaction, Wilcoxon's procedure becomes tedious when the two factors (treatments) are divided into many components (levels of treatment). For this reason, we shall illustrate this technique using a small-size experiment, and then indicate the necessary extensions for larger experiments.

PRACTICAL PROBLEM

To back up a new advertising campaign, a paint manufacturer tests dye. for their resistance to fading when subjected to ultraviolet (UV) light. Three technicians each perform the test on red and blue paint, those colours being chosen because of their diverse pigment structure.

HYPOTHESES

The main concern is probably with the effect of increasing the intensity of the UV light, and then with the differences in fading between the two colours. However, a sensible move is first to check for the presence of interaction between the colour and the UV light. So our chosen hypotheses are:

H_0: There is no interaction between paint colour and UV light.
H_1: There is interaction between paint colour and UV light.

DATA

The technicians recorded the percentage shift in the wavelength of light reflected from the painted surface. Three levels of UV light intensity were

used (low, medium and high), each for 200 h. The data are given in the following table.

		UV intensity		
Paint	Technician	Low	Medium	High
red	1	4.9	7.2	7.3
	2	5.2	7.3	7.1
	3	6.1	6.9	7.4
blue	1	5.2	7.2	6.9
	2	4.8	7.7	7.1
	3	4.6	7.4	6.8

STATISTIC

Since we are interested in testing for interaction between paint colour and UV light, we subtract from each 'red' observation the corresponding 'blue' observation. This gives the following table of R−B (red–blue) differences:

	UV intensity		
Technician	Low	Medium	High
1	−0.3	0	−0.4
2	0.4	−0.4	0
3	1.5	−0.5	0.6

Now, if the null hypothesis is true, the R−B differences should be roughly the same along each row. Thus if the differences within each row are ranked (from 1 to 3, in this example) and then Friedman's statistic calculated, this should result in a small value of M. On the other hand, if H_1 is true, then M will tend to be large, as then the ranks of the R−B differences will tend to exhibit some patterns in the table.

Ranking the above data and proceeding to compute Friedman's statistic, we have:

	Ranks		
	2	3	1
	3	1	2
	3	1	2
rank sums	8	5	5

Hence,

$$M = \frac{12}{nk(k+1)} \sum_{j=1}^{3} R_j^2 - 3n(k+1)$$

$$= \frac{12}{3\times3\times4}(64+25+25) - 3\times3\times4 = 2$$

CRITICAL REGIONS

As we have seen, Wilcoxon's method is to modify the observed data in such a way that Friedman's test becomes a test for interaction. This means that critical values can be obtained from Table O (at least in this size of experiment, see the 'Discussion' section below). Thus critical regions are of the form:

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

In our problem, $k=3$ and $n=3$, giving the $\alpha=5\%$ critical region a $M \geqslant 6.000$ (with such a small value of n, no 1% critical region is available).

SOLUTION

Since the calculated value of $M=2$ does not lie in the 5% critical region, there is no evidence of interaction between paint colour and UV light.

Discussion

In our example, it was easy enough to apply Wilcoxon's procedure. However, with larger experiments, the test statistic is made up of several components, each of which comes from applying Friedman's procedure to a set of ranks. For example, suppose our practical problem had involved *three* colours, say red, blue and yellow. In this case, Wilcoxon's procedure is as follows:

Step 1 As before, derive the differences R−B and calculate Friedman's statistic (call it m_1).

Step 2 Form a table of differences R+B−2Y and calculate Friedman's statistic (call this one m_2) from the ranks.

Step 3 The final test statistic is $M=m_1+m_2$.

This test statistic is approximately distributed as a chi-squared variable with $2(k-1)$ degrees of freedom, which enables appropriate critical values to be obtained.

If there were *four* colours in the experiment, say red, blue, yellow and green, the final test statistic is then $M=m_1+m_2+m_3$ where m_1 and m_2 are calculated as explained above, and m_3 is calculated from the table of differences R+B+Y−3G.