The statistic M is now approximately distributed as a chi-squared variable with $3(k-1)$ degrees of freedom.

In general, with p colours $C_1, C_2, \ldots, C_p$ the test statistic M is given by $m_1+m_2+\ldots+m_{p-1}$, where m_i is the Friedman statistic based on differences of the form $C_1+C_2+\ldots+C_i-(i-1)C_{i+1}$, for $i=2,3,\ldots,p-1$. You will see that the ith set of such differences provide data on the possible interaction between C_{i+1} and the previous colours $C_1, C_2, \ldots, C_i$. Under the null hypothesis of no interactions, M is approximately distributed as chi-squared with $(p-1)(k-1)$ degrees of freedom.

MATCH TESTS FOR INTERACTION

Introduction

As we discussed at the beginning of the chapter, our study of interaction is divided between those experiments with linked observations and those with unlinked observations. Wilcoxon's method just described is probably the simplest of many distribution-free procedures for the former case. Unfortunately, *simple* procedures for dealing with the case of unlinked observations are scarce. However, in this section, we present such a technique, which involves the match ideas. As with the previous match tests, we give two tests; one is based solely on the number of matches, while the other incorporates near-matches.

PRACTICAL PROBLEM

An international cycling team is concerned with the road-holding capabilities of various brands of tyres. It seems that individual members disagree about the optimum tyre pressure for wet conditions; some members argue that certain makes of tyre respond better if the pressure is lower than normal with the opposite being true for the other brands. Other team members scorn this idea and keep the tyre pressure normal regardless of the brand of tyre.

The team manager happened to mention his team's views to a member of the research department of a famous bicycle manufacturer. The researcher, being intrigued, offers to conduct an experiment in the company's test laboratory.

HYPOTHESES

Naturally, the team members are interested in finding out about the differences in performance between the tyres. However, because of the conflicting arguments, the researcher suspects that there may be some

interaction between the various makes of tyre and tyre pressures. So his hypotheses for this situation are:

H_0: There is no interaction between tyre pressure and brand of tyre.
H_1: There is some interaction between tyre pressure and brand of tyre.

DATA

Apparently, four different brands of tyre are used by members of the team, and so the researcher selects these brands for the test; we shall denote the brands by A, B, C and D to preserve anonymity. After a discussion with the team about tyre pressures, four tyre pressures were used in the experiment: 20, 40, 50 and 70 lb/sq. in. For each combination of pressure and brand, the experiment was performed three times.

The data record the speed (mph) at which slipping occurred on a standard inclined, wet surface.

	Brand of tyre			
Pressure	A	B	C	D
20	33.16	30.89	35.38	33.88
	30.23	28.38	33.86	32.53
	30.19	30.49	34.77	33.10
40	29.05	31.66	31.64	34.51
	29.72	34.23	30.42	34.89
	31.48	34.44	31.82	34.84
50	32.96	29.00	33.20	32.42
	31.59	32.68	32.36	32.01
	30.98	33.14	35.07	33.29
70	27.84	33.32	31.76	34.07
	29.43	33.15	32.83	35.16
	29.90	32.52	30.76	34.26

STATISTICS

Because of the somewhat intangible nature of interaction, you might suppose that any test procedure would be involved and lengthy. To some extent this is true; the procedure we now present is possibly the simplest available, but even so you will find it tedious to do without a calculator.

The basic idea is to remove the effects of the rows (pressures) and columns (tyres) to leave data that represent just interaction. Before doing this, we replace the group of observations for each combination of pressure and brand by their mean (the observation means). This gives the following data on which we apply further manipulations.

Exhibit 1006
Page 332 of 453

	Brand of tyre				
Pressure	A	B	C	D	Row means
20	31.19	29.92	34.64	33.17	32.23
40	30.08	33.44	31.29	34.75	32.39
50	31.84	31.61	33.54	32.57	32.39
70	29.06	33.00	31.78	34.50	32.08
Column means	30.54	31.99	32.81	33.75	32.27 = overall mean

Notice that we have also calculated the row means, column means and the overall mean in readiness for the next stage.

From this reduced set of data, we obtain the interaction component as:

interaction
= observation mean − row mean − column mean + overall mean

So, for example, taking the top-left observation, we calculate

$$31.19 - 32.23 - 30.54 + 32.27 = 0.69$$

Doing this for all 16 values gives:

	Brand of tyre			
Pressure	A	B	C	D
20	0.69	−2.03	1.87	−0.54
40	−0.58	1.33	−1.64	0.88
50	1.18	−0.50	0.61	−1.30
70	−1.29	1.20	−0.84	0.94

The rest is now easy! The data in this table are ranked by row and by column to give two tables of ranks:

Row ranks			
3	1	4	2
2	4	1	3
4	2	3	1
1	4	2	3

Column ranks			
3	1	4	2
2	4	1	3
4	2	3	1
1	3	2	4

To each of these tables we apply one of the match tests of Chapter 14;

either M_1 or M_2. Suppose M_1 is chosen; then we calculate two values of the statistic, M_{1r} and M_{1c} where the r and c denote that the value is obtained from the row ranks and column ranks respectively. The test statistic S_1 is the sum of these two values. For our problem we obtain $M_{1r}=2$ and $M_{1c}=0$, giving $S_1=2$.

Similarly, if M_2 is chosen we calculate two values M_{2r} and M_{2c} giving a test statistic $S_2=M_{2r}+M_{2c}$. Our problem gives $M_{2r}=7$ and $M_{2c}=6$ from which $S_2=13$.

CRITICAL REGIONS

If H_1 is true then the presence of interaction in the form of erratic behaviour patterns will tend to give small values of S_1 and S_2; for example, the no-interaction situation shown in Figure 16.1a would tend to give a high value of S_1 (or S_2) whereas the interaction situation in Figure 16.1b would tend to produce a low value of S_1 (or S_2). Thus critical regions are of the form:

$$\text{test statistic} \leqslant \text{critical value}$$

Tables V and W give critical values of S_1 and S_2 respectively.

In our problem, $k=4$ and $n=4$ and so the 5% and 1% critical regions for S_1 are $S_1 \leqslant 6$ and $S_1 \leqslant 5$ respectively, while the corresponding critical regions for S_2 are $S_2 \leqslant 16$ and $S_2 \leqslant 15$ respectively.

CONCLUSION

Our conclusion is the same whether we use S_1 or S_2, as both give values that lie in the 1% critical region. Thus there is strong evidence to support the researcher's suspicion regarding interaction between tyre pressure and make of tyre.

Large experiment sizes

S_1 *test* Approximate critical values can be obtained using a normal approximation. If we have k columns and n rows in the experiment, then an approximate critical value for an $\alpha\%$ level of significance is the integer part of:

$$\tfrac{1}{2}[n(n-1) + k(k-1)] - z\sqrt{\tfrac{1}{2}[n(n-1)+k(k-1)]} - \tfrac{1}{2}$$

where z is the appropriate percentage point of the standard normal distribution (e.g. for $\alpha=5\%$, $z=1.6449$) and the $\frac{1}{2}$ is the continuity correction.

S_2 *test* Approximate critical values can be obtained using a normal approximation. If the experiment has k columns and n rows, then an approximate critical value for an $\alpha\%$ level of significance is the value of:

$$\tfrac{1}{2}\left[n(n-1)\left(2-\frac{1}{n}\right)+k(k-1)\left(2-\frac{1}{n}\right)\right]-$$
$$z\sqrt{\frac{n(n-1)(3k^3-9k^2+6k+22)}{4k^2(k-1)}+\frac{k(k-1)(3n^3-n^2+6n+22)}{4n^2(n-1)}}+\tfrac{1}{2}$$

to the nearest 0.5 (since S_2 has values which are multiples of 0.5), and where z is as above.

LATIN SQUARES DESIGNS

Introduction

As in Chapter 14, this part of the chapter deals with the k-related samples problem, but here the observations are related by an extra variable so that now *three* factors are involved. There are many ways of analysing such experiments. The simplest method is to use the so-called latin squares design, which is an extension of the randomized block design we mentioned in Chapter 14. This design allows us to test for differences between the averages of the k populations and also for differences among each of the other two factors; thus we have three sets of hypotheses, some or all of which we may choose to investigate.

The following example illustrates the concept of a latin square design. A company measures productivity on five different jobs (factor 1) during a particular week, the five working days constituting factor 2. Five different working conditions (factor 3) are used, which are:

A complete silence
B piped-in popular music
C piped-in country and western music
D piped-in classic music
E 'request' music

By constructing a so-called 5×5 latin squares design, each combination of these three factors can be represented. A typical 5×5 design for this problem is given below:

	Job type				
Day	1	2	3	4	5
Mon	A 137	B 121	C 118	D 113	E 108
Tues	B 126	C 130	D 135	E 137	A 129
Wed	C 126	D 127	E 147	A 129	B 139
Thurs	D 113	E 118	A 113	B 112	C 104
Fri	E 117	A 115	B 116	C 110	D 118

So, for example, the figure underlined represents the productivity on Friday for job type 4 and under working conditions C. The assignment of letters (which here represent working conditions) to the rows (days) and columns (jobs) is not unique; another valid arrangement of letters is the following:

	Job type				
Day	1	2	3	4	5
Mon	A	C	E	B	D
Tues	B	A	C	D	E
Wed	E	D	B	C	A
Thurs	C	E	D	A	B
Fri	D	B	A	E	C

The essential feature is that each row and column contains no duplication of letters. Thus, for example, the following 5×5 table would not be a valid latin square design.

	Job type				
Day	1	2	3	4	5
Mon	A	C	E	B	D
Tues	B	A	C	D	E
Wed	A	B	B	C	A
Thurs	C	E	E	A	C
Fri	D	B	A	E	C

An obvious restriction of such designs is that each of the three factors must be represented by the same number of different *levels*; in the above example, there were five job types, five days and five working conditions. Another, not so obvious, restriction is that the presence of interactions (say, between day *and* job type or between day and job type *and* working conditions) cannot be detected.

Latin square designs were probably devised for agricultural experiments where the k populations represented k treatments (fertilizers, different varieties of seed, etc.) and the other two factors were simply the coordinates of a plot in a field.

In this section, we present a procedure for analysing latin squares designs by Friedman's test or the match tests of Chapter 14.

PRACTICAL PROBLEM

The R&D section of a well known tyre manufacturer conducted an experiment to find out whether differences existed among the average life

of four brands of tyre. The brands are denoted by T_1 (their own leading tyre), T_2, T_3 and T_4. The tyres were tested under four different climatic conditions: C_1=Arctic, C_2=British summer, C_3=British winter and C_4=hot and dry. Four different road conditions were used: R_1=rough, R_2=newly tarmacadamed, R_3=concrete, and R_4=stone-chippings surface.

HYPOTHESES

In this situation with three factors (tyres, climate and road conditions), we have in fact three possible sets of hypotheses to explore. Although the company's main concern is with differences in average life between the tyres, the design of the experiment also makes it possible to test for differences caused by the climatic conditions and by the road surfaces. These hypotheses are:

(I) H_0: There are no differences in average tyre wear between the four brands of tyre.
H_1: There are differences in average tyre wear between the four brands of tyre.

(II) H_0: There are no differences in average tyre wear between the four climatic conditions.
H_1: There are differences in average tyre wear between the four climatic conditions.

(III) H_0: There are no differences in average tyre wear between the four road surfaces.
H_1: There are differences in average tyre wear between the four road surfaces.

DATA

For each of the three factors, the amount of tread wear (mm) over a given mileage was recorded. The data are presented in the following 4×4 latin square design.

	Brand of tyre			
Climate	T_1	T_2	T_3	T_4
C_1	R_1 3.2	R_2 1.4	R_3 0.4	R_4 2.6
C_2	R_2 2.5	R_3 2.3	R_4 2.8	R_1 3.3
C_3	R_3 2.4	R_4 2.7	R_1 2.9	R_2 0.9
C_4	R_4 3.1	R_1 3.7	R_2 0.6	R_3 0.5

Exhibit 1006
Page 337 of 453

METHOD

From the above latin square table, we construct three other tables, one for each of the combinations: (1) brand × climate, (2) brand × road, and (3) climate × road. These tables are given below.

	Brand			
Climate	T_1	T_2	T_3	T_4
C_1	3.2	1.4	0.4	2.6
C_2	2.5	2.3	2.8	3.3
C_3	2.4	2.7	2.9	0.9
C_4	3.1	3.7	0.6	0.5

	Brand			
Road	T_1	T_2	T_3	T_4
R_1	3.2	3.7	2.9	3.3
R_2	2.5	1.4	0.6	0.9
R_3	2.4	2.3	0.4	0.5
R_4	3.1	2.7	2.8	2.6

	Climate			
Road	C_1	C_2	C_3	C_4
R_1	3.2	3.3	2.9	3.7
R_2	1.4	2.5	0.9	0.6
R_3	0.4	2.3	2.4	0.5
R_4	2.6	2.8	2.7	3.1

We shall now test each set of hypotheses using both Friedman's test and the match tests.

Hypotheses (I)

Using the first table, we rank the observations within each row. This produces the following table of ranks.

	Brand			
	T_1	T_2	T_3	T_4
	4	2	1	3
	2	1	3	4
	2	3	4	1
	3	4	2	1
rank sums	11	10	10	9

From this table of ranks we immediately obtain

$$M_1 = 0 + 1 + 1 = 2$$
$$M_2 = 2 + \tfrac{1}{2}(5 + 3 + 2) = 7$$

and Friedman's statistic is

$$M = \frac{12}{nk(k+1)} \sum_{j=1}^{4} R_j^2 - 3n(k+1)$$

$$= \frac{12}{4\times4\times5}(112 + 102 + 102 + 92) - 3\times4\times5$$

$$= 0.3$$

Alternatively, the test statistics could have been produced from the second table, again by ranking within each row. This time the ranks are:

	Brand			
	T_1	T_2	T_3	T_4
	2	4	1	3
	4	3	1	2
	4	3	1	2
	4	2	3	1
rank sums	14	12	6	8

Hence, M_1=8, M_2=12 and M=6.

Hypotheses (II)

To test these hypotheses we may use either the first table (ranking the columns) or the third table (ranking the rows). The first table gives M_1=2, M_2=7 and M=0.3, while the third table gives M_1=6, M_2=10.5 and M=3.0.

Hypotheses (III)

To test these hypotheses we may use either the second or third table, ranking the columns in each case. Both tables give M_1=18, M_2=21 and M=11.1.

CRITICAL REGIONS

Naturally, critical regions are just as for Friedman's test and the match tests. Hence the 5% and 1% critical regions are respectively:

for Friedman's test (from Table O)

$$M \geqslant 7.800 \quad \text{and} \quad M \geqslant 9.600$$

for the M_1 test (from Table P)

$$M_1 \geqslant 12 \quad \text{and} \quad M_1 \geqslant 15$$

for the M_2 test (from Table Q)

$$M_2 \geqslant 15.0 \quad \text{and} \quad M_2 \geqslant 18.0$$

SOLUTION

Hypotheses (I)
There is no evidence to reject H_0 and so we have no evidence to claim there are differences between the four brands of tyre.

Hypotheses (II)
None of the test values lie in their respective critical regions, and so we have no evidence to support H_1.

Hypotheses (III)
All three tests give results that lie in the 1% critical region, thus providing strong evidence to support the claim that there are differences in the average tyre wear due to different road surfaces.

Discussion
Since the test statistics can be calculated from one of two tables of ranks, it is quite feasible that one value could be significant and the other non-significant. It seems that this conflict arises when interaction exists between the factors and in this case a further investigation would be desirable.

Ties and large sample sizes
The comments made in Chapter 14 apply here.

Program 16.1 Match tests for interaction

```
100 REM *** MATCH TESTS FOR INTERACTION ***

110 REM ** INPUT DATA **
120 INPUT "NUMBER OF ROWS & COLUMNS ? ";N,K
130 INPUT "NUMBER OF REPLICATIONS ?";NR
140 DIM X(N,K,NR),A(N,K),RW(N),CL(K),Y(N,K),RK(N,K)
150 LET S = 0
160 FOR I = 1 TO N
170 FOR J = 1 TO K
```

Continued overleaf

Program 16.1 Continued

```
180 LET A(I,J) = 0
190 NEXT J
200 NEXT I

210 REM ** INPUT OBSERVATIONS **
220 FOR I = 1 TO N
230 FOR J = 1 TO K
240 PRINT "OBSERVATIONS FOR ROW ";I;" & COLUMN ";J
250 FOR L = 1 TO NR
260 INPUT X(I,J,L)
270 LET A(I,J) = A(I,J) + X(I,J,L)/NR
280 LET S = S + X(I,J,L)
290 NEXT L
300 PRINT
310 NEXT J
320 NEXT I

325 REM ** CALCULATION OF TEST STATISTICS **
330 FOR I=1 TO N
340 LET RW(I) = 0
350 NEXT I
360 FOR I = 1 TO K
370 LET CL(I) = 0
380 NEXT I
390 FOR I = 1 TO N
400 FOR J = 1 TO K
410 LET RW(I) = RW(I) + A(I,J)/K
420 LET CL(J) = CL(J) + A(I,J)/N
430 NEXT J
440 NEXT I
450 FOR I = 1 TO N
460 FOR J = 1 TO K
470 A(I,J) = A(I,J) - RW(I) - CL(J) + S/N/K/NR
475 LET Y(I,J)=A(I,J)
480 NEXT J
490 NEXT I
500 GO SUB 1000
510 GO SUB 2300
520 LET S2=M2
530 LET S1=M1
540 FOR I=1 TO N
550 FOR J=1 TO K
560 LET A(J,I)=Y(I,J)
570 NEXT J
580 NEXT I
590 LET F=N
600 LET N=K
610 LET K=N
620 GO SUB 1000
630 GO SUB 2300

635 REM ** OUTPUT OF VALUES OF S1 AND S2 **
640 LET S2=S2+M2
650 LET S1=S1+M1
660 PRINT S2,S1
670 STOP

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

2300 REM ** CALCULATION OF M1 AND M2 - USE LINES 230-360 OF PROGRAM 14.2 BY **
2305 REM ** ADDING A '0' TO  THE LINE NUMBERS **
```

PART VI

Miscellaneous procedures

The three chapters in this part of the book present a variety of tests and procedures. In Chapter 17, we introduce tests based on what are known as 'runs'; these ideas will not only analyse two-sample situations such as those in Chapters 5 and 6 but will also deal with the problem of 'non-randomness' in a sequence of observations.

Chapter 18 presents more tests using frequency data. First, we discuss Fisher's 'exact test' for analysing 2×2 contingency tables and then we lead into Mood's two-sample median test followed by the Mood–Westenberg two-sample dispersion test.

The final chapter describes some common distribution-free methods for constructing confidence intervals for the population median and for the difference of two population medians.

17
Tests based on runs

Introduction

In this chapter, we describe a number of tests based on numbers of runs. The one-sample version provides an extremely useful check for certain features of non-randomness in a sequence of observations, while the two-sample version has a remarkably wide range of general applications.

The term 'run', in the sense relevant to this chapter, is quite familiar to us, for we have all heard the expression 'a run of good (or bad) luck' applied to someone's fortunes. Indeed, it is likely that the statistical usage of the word 'run' originated in gambling situations.

Suppose that a person tossing a coin 20 times generates the following sequence of heads (H) and tails (T):

T H T T T T T T T T T T T H H H T T T

We see that this sequence includes a long run of tails (of length 12, from the third to the fourteenth toss inclusive), which, in a gambling game, could well represent a run of good or bad luck according to the rules of the game. Scanning the whole sequence from left to right, we start with a 'run' of just one 'T', which is followed by a run of one 'H', then a run of 12 'T's, followed by a run of three 'H's, and finishing with a run of three 'T's; five runs altogether. Note that a 'run' may consist of only one observation, as is the case with both the first two runs in this example.

In the context of statistical testing, we might ask: can the above sequence of 'H's and 'T's reasonably be regarded as random – or is there evidence of an element of 'luck' (good or bad) or some other non-random influence? The term 'random' here means effectively that there is no tendency for any discernible pattern to emerge.

As we shall see, the number of runs in the sequence can help answer this question and, of course, this technique can also be applied to many situations other than gambling games. For example, it is usually important to check for non-randomness on a quality control chart; perhaps there is a long run of observations above the central line, giving a picture such as that in Figure 17.1.

Representing values above and below the central line by A and B respectively, we have the sequence:

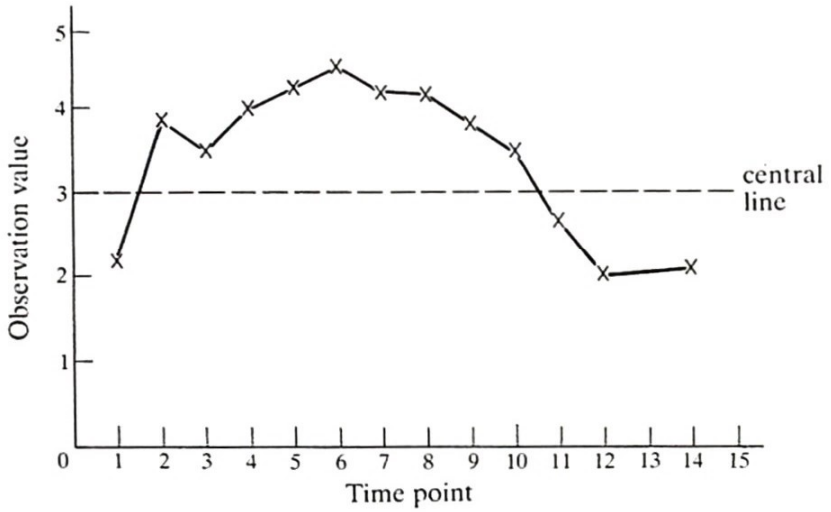


Figure 17.1

B A A A A A A A A A B . . .

Surely we should be suspicious of such a long run (of length 9) of 'A's – perhaps a machine characteristic was wrongly set over this period of time.

In this latter type of example, an alternative 'runs' analysis may be used to check for non-randomness. This is based on the number of 'runs up and/or down' – runs of steadily rising or falling values. We shall see an example of this later in the chapter.

We conclude the chapter with a particularly quick-and-easy two-sample test, again based on runs, designed for detecting general differences between two populations. This is therefore a simple test for the task tackled in Chapter 7 by the Kolmogorov–Smirnov test and, although its power cannot match that of the Kolmogorov–Smirnov test, its simplicity makes it a useful test to have available.

RUNS TEST FOR RANDOMNESS

Introduction

One of the appealing features of the runs test for randomness (or rather, lack of randomness, which is what we want the test to be able to discover) is its 'common sense' nature. Most people would be immediately suspicious of, for example, a sequence of coin tossings which produces too few runs:

H H H H H H H H T T T T T T T T

or too many runs: ·

H T H T H T H T H T H T H T H T H T

Both of these extremes indicate in an obvious way a complete lack of randomness, for each has a clear pattern. (Although, at first glance, the second sequence may strike you as being random, in fact heads and tails alternate strictly throughout the sequence – and that is actually just as unlikely an occurrence as the first sequence.)

The runs test we discuss here requires a minimum of formal assumptions. All we require is that each observation be capable of classification as one type or the other.

PRACTICAL PROBLEM

A manufacturer of ceramic wash-basins also makes the mixer taps for the basins. An important component of these taps is a brass T-connector which (if water is not to spurt everywhere!) should form a 'perfect' fit with the incoming pipes. Clearly, it is necessary that the diameters of these connectors should be made with minimal variation from some target value determined by the designer. A control chart was set up in order to monitor the machine producing these junctions. Now, in a quality control chart, there are many warning signs of something untoward happening to the process. One such possible sign is a lack of randomness – perhaps a high value is always followed by a low value and vice versa (which might indicate the presence of some self-compensating effect), or there might be correlation between adjacent observations which may result in long sequences above and below the average value.

HYPOTHESES

In this simple test for randomness (or rather lack of randomness), our hypotheses are just:

H_0: The sample values come from a random sequence.
H_1: The sample values do not come from a random sequence.

DATA

To establish a control chart for the sample means, 20 samples of four T-junctions were taken at 15 min intervals. The diameters were measured and the mean of the diameters (mm) for the 'vertical' part of the T-junction are given below.

Sample number	1	2	3	4	5	6	7	8	9	10
Mean	15.0	14.8	14.9	15.0	15.1	15.3	15.0	14.9	14.7	15.1
Sample number	11	12	13	14	15	16	17	18	19	20
Mean	15.2	15.3	14.9	15.2	15.1	15.0	14.9	14.8	14.7	14.8

STATISTIC

To apply a runs test for investigating whether there is a lack of randomness about the overall mean value of 14.985, we simply record a '+' for each sample mean above 14.985 and a '−' for each value below 14.985. This gives the following sequence of '+'s and '−'s:

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

Now, as we saw in the introductory discussion, a lack of randomness can show itself by the sequence of signs having either too few or too many runs. This immediately suggests that a suitable test statistic is simply:

$$T = \text{total number of runs}$$

In order to use the table of critical values, we must also note the numbers n_1 and n_2 of + and − signs, respectively. In this case we have $n_1=11$ and $n_2=9$.

CRITICAL REGIONS

Since it is very low or very high values of T that lead us to reject H_0, the critical regions are of the form:

$$T \leqslant \text{lower critical value}$$

or

$$T \geqslant \text{upper critical value}$$

These lower and upper critical values are given in Table X for $\alpha_2=5\%$ and 1% and various values of n_1 and n_2.

We have $n_1=11$ and $n_2=9$, and so at $\alpha_2=5\%$ we reject H_0 in favour of H_1 if $T\leqslant 6$ or $T\geqslant 16$.

SOLUTION

The number of runs T is equal to 8, comprising runs of length

1 of +
2 of −
4 of +
2 of −
3 of +
1 of −
3 of +
4 of −

This value, $T=8$, does not fall in the $\alpha_2=5\%$ critical region, and so we have no evidence to support H_1. Of course, this is good news for the company since a contrary decision would presumably result in some investigation on the apparent non-randomness.

Tied values

It might happen that one or more of the observed values is the same as the mean value, in which case we cannot record the usual + or − sign. If the number of such cases is small compared with the total length of the sequence, then it is usual to ignore these cases and proceed as normal with the remaining + and − signs.

(Note: A more sophisticated approach is to consider all possible sequences arising from recording such tied observations as + or − in turn, and averaging the numbers of runs resulting from such sequences.)

One-sided tests

Suppose that the situation under study is such that the only type of non-randomness likely to occur tends to make the sequence alternate too rapidly (a self-compensating effect is an obvious example). In this case our alternative hypothesis may be expressed something like:

H_1: The process's output tends to 'zigzag' more than would be expected under true randomness.

Critical regions are of the form

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

with α_1 upper critical values being obtained from Table X.

Conversely, it might be that we are only interested in types of non-randomness that tend to produce longer (and therefore fewer) runs than usual; a common cause of this is positive correlation (as opposed to independence) between adjacent observations. In this case the appropriate alternative hypothesis might be:

H_1: The process's output tends to be 'smoother' than would be expected under true randomnesses.

Now the critical regions are of the form:

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

with the α_1 lower critical values being obtained from Table X.

Large sample sizes

Should either n_1 or n_2 be greater than 20, then approximate critical values may be obtained as follows. It can be shown that, for large values of n_1 and n_2, T is approximately normally distributed with mean

$$\mu = 1 + 2n_1n_2/N$$

and standard deviation

$$\sigma = \sqrt{2n_1n_2(2n_1n_2 - N)/(N^2(N-1))}$$

So approximate lower and upper critical values (T_L and T_U) are given by

$$T_L = \mu - z\sigma - \tfrac{1}{2}$$

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

where z is the appropriate percentage point in the standard normal distribution and the '$\frac{1}{2}$' is the continuity correction. For example, if $n_1=20$ and $n_2=20$ then $\mu = 21$ and $\sigma = 3.12$ and so, for $\alpha_2=5\%$ (for which $z=1.96$) the critical values are

$$T_L = 21 - 1.96\times 3.12 - \tfrac{1}{2} = 15$$

$$T_U = 21 + 1.96\times 3.12 + \tfrac{1}{2} = 27 \qquad \text{(when rounded appropriately)}$$

Table X gives the best conservative critical values as 14 and 28 respectively; so, even with n_1 and n_2 as low as 20, the approximation works well.

Runs up and down

Provided our data are recorded as numerical values rather than just as 'H'/'T', etc., it is possible to examine a sequence of observations for another aspect of non-randomness, namely 'runs up' and 'runs down'. To see what this idea involves, consider the following sequence of values:

5.6	6.2	6.5	7.6	8.0	6.1	7.5	4.8	3.2	2.6

The plot of these values in Figure 17.2 clearly shows the run up (of length 4) from 5.6 to 8.0, the run down (of length 1) to 6.1, and so on. (Note that we do not include the first point in counting the length of the first run up since we cannot tell whether or not it belongs to that run without previous history.)

Of course, we can easily study such runs (up or down) without drawing a graph. All we need to do is to record a + or a − if a value is greater or less than its predecessor, respectively. So, applying this idea to the above sequence gives:

5.6	6.2	6.5	7.6	8.0	6.1	7.5	4.8	3.2	2.6
	+	+	+	+	−	+	−	−	−

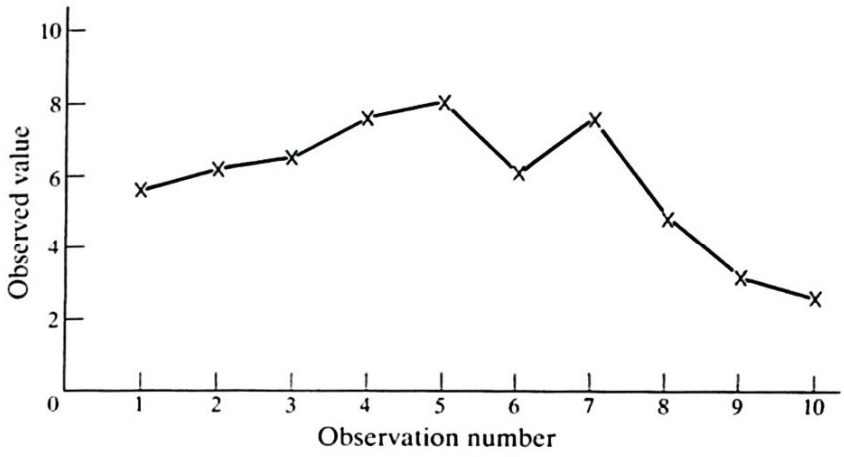


Figure 17.2

where the '+'s and '−'s denote moves up and down respectively.

Intuitively, in a 'random' sequence we would expect moderate numbers of runs up and runs down. But, a sequence that has a general upward trend will tend to have quite a number of long runs up, with only the occasional short run down (and naturally, a sequence with a downward trend will exhibit long runs down). In either case, therefore, the numbers of runs up and runs down will be unusually small. On the other hand, a sequence having particularly erratic behaviour will tend to have many short runs of each type.

As you probably expect, the test statistic used for this feature of randomness is the total number of runs (up or down), or, in other words, just the number of runs of '+'s and '−'s in the above sequence. So it now all looks very much like the previous runs test. But beware! These '+'s and '−'s are different from before, and the corresponding critical values are generally different as well; so we shall need a new table of critical values for this test.

PRACTICAL PROBLEM

The problem of the brass T-connectors for taps (see the previous section) can be examined for this different exhibition of non-randomness, i.e. the tendency for too long or too short runs up and runs down. Too few runs up and runs down might indicate a trend or some broad oscillatory tendency, whereas too many indicates a zigzag effect.

HYPOTHESES

The hypotheses for the situation just discussed may be summarized as:

H_0: The averages occur 'at random'.
H_1: The averages have one or more non-random features, such as trends, oscillations or too much zigzag behaviour.

DATA

The data are as before.

STATISTIC

As we have already seen, a suitable test statistic (U) for this situation is given by

$$U = \text{number of runs (up or down)}$$

CRITICAL REGIONS

Clearly, the alternative hypothesis will be supported if there are unusually few or unusually many runs, and so H_0 will be rejected if

$$U \leqslant \text{lower critical value}$$

or

$$U \geqslant \text{upper critical value}$$

Table Y gives critical values for samples up to $n=25$, where n is the length of the sequence of data.

We have $n=20$ in our problem and so we shall reject H_0 at $\alpha_2=5\%$ and 1% if $U \leqslant 8$ or $U \geqslant 17$ and $U \leqslant 7$ or $U \geqslant 18$ respectively.

SOLUTION

By simply observing whether an observation is greater than or less than its predecessor, we obtain the following sequence of signs:

15.0	14.8	14.9	15.0	15.1	15.3	15.0	14.9	14.7	15.1
	−	+	+	+	+	−	−	−	+
15.2	15.3	14.9	15.2	15.1	15.0	14.9	14.8	14.7	14.8
+	+	−	+	−	−	−	−	−	+

Immediately we see that $U=8$, which is significant at $\alpha_2=5\%$. Thus there is some evidence to support H_1; indeed a glance at the data reveals two 'longish' runs (a run up of length 4 and a run down of length 5), which might warrant some further investigation.

Tied values

Equal values in the sequence only pose a problem if they are consecutive. If a particular observation is equal to its predecessor, we shall indicate this by a zero instead of by + or − (since a zero is neither positive nor negative). Then, if the number of zeros in the sequence of signs is small, we shall simply ignore them and reduce the value of n accordingly. So, for example, consider the following sequence of values:

5.6	6.1	2.8	2.8	2.8	3.4	5.8	6.1	4.8
	+	−	0	0	+	+	+	−

Ignoring the zeros gives the test sequence

+	−	+	+	+	−

with $U=4$ and $n=6$.

One-sided tests

As discussed earlier, some types of non-randomness tend to make the number of runs up and down lower than usual and others make the number larger than usual. These situations result in sequences that might be described as smoother or more erratic, respectively, than expected under the null hypothesis of randomness. If we are only interested in our test diagnosing one of these types of non-randomness, we can of course produce a one-sided version of the test. In the former case, the type of non-randomness of interest tends to produce long runs up or runs down, and therefore relatively few of them.

The appropriate alternative hypothesis might (as earlier in this chapter) be expressed as:

H_1: The process's output tends to be 'smoother' than would be expected under true randomness.

In this case, the α_1 lower critical values from Table Y are used and the critical regions have the form:

$$U \leqslant \text{lower critical region}$$

Likewise, we might be interested in the alternative hypothesis:

H_1: The process's output tends to 'zigzag' more than would be expected under true randomness.

Then the α_1 upper critical values from Table Y are used and the corresponding critical regions are:

$$U \geqslant \text{upper critical value}$$

Large sample sizes

Should the sample size n be greater than 25, then approximate critical values may be obtained using a normal distribution approximation. It can be shown that, for large values of n, U is approximately normally distributed with mean $\mu=(2n-1)/3$ and standard deviation $\sigma=\sqrt{(16n-29)/90}$. Thus, approximate lower and upper critical values U_L and U_U are given by

$$U_L = \mu - z\sigma - \tfrac{1}{2}$$

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

where z is the appropriate percentage point from the standard normal distribution and the '$\frac{1}{2}$' is the continuity correction.

You might like to check that, if we take $n=25$, we obtain $\mu=16.33$ and $\sigma=2.03$. Now for $\alpha_2=5\%$ the value of z is 1.96 and so we obtain

$$\mu - z\sigma - \tfrac{1}{2} = 11.85$$

and

$$\mu + z\sigma + \tfrac{1}{2} = 20.80$$

giving $U_L=11$ and $U_U=21$. These are the same values as those in Table Y.

THE WALD–WOLFOWITZ TWO-SAMPLE RUNS TEST

A. Wald and J. Wolfowitz (1940) proposed a simple two-sample test based on the idea of runs. A very useful feature of their test is that (if the sample sizes are large enough) it can detect differences in averages, or spread, or in any other important aspect between the two populations being studied. However, as might be expected of a test that is both simple and ambitious, its power is not great. Consequently, if one's prime interest is in detecting differences in average, the tests of Chapter 5 are to be preferred. Or, if the main interest is in detecting differences in spread, we would prefer the tests in Chapter 6. However, one problem with these previous tests is that a difference of a type not being catered for can harm their ability to carry out the task for which they are designed; for example, the power of both the Mann–Whitney and Tukey tests for detecting differences in location can be severely reduced if the populations also differ substantially in spread. On the other hand, the Wald–Wolfowitz test might well have greater power in these and other situations of more general differences. The Wald–Wolfowitz test can therefore be regarded as a fairly quick-and-easy alternative to the two-sample Kolmogorov–Smirnov test (see Ch. 7), which also has power for general types of differences.

However, because of this very generality, there could conceivably be

Exhibit 1006
Page 353 of 453

difficulties in interpreting significant results from this test. It is easy enough with the 'special-purpose' tests; if we have a significant result from the Mann–Whitney test we know that this indicates a difference in location and, similarly, a significant result from the Siegel–Tukey test definitely indicates a difference in spread. But significant results from the Wald–Wolfowitz test can be caused by any kind of substantial difference. In practice, however, it is usually easy to make an informal judgement of the reason for rejection of the null hypothesis by taking a quick look at the letter sequence. Incidentally, this discussion also implies that the notion of one-sided tests is not relevant to the Wald–Wolfowitz test.

Let us now get an idea of the way the test works. As usual, we will denote observations from our two populations by A and B and consider the letter sequence thus obtained. However, now we concentrate on the number of runs in the letter sequence. Thus, if we have the letter sequence:

A A B B A B B A B A A

then we find a total of seven runs.

If the two samples are from the same population, then we would expect the 'A's and 'B's in the letter sequence to be 'well mixed', so producing a relatively large number of runs. On the other hand, suppose the two populations differ in location; their distributions might be as shown in Figure 17.3.

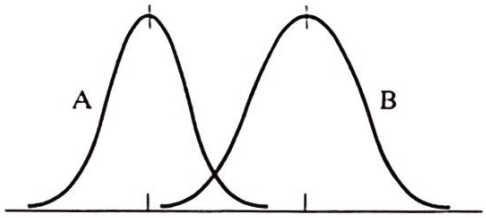


Figure 17.3

A possible letter sequence for this situation could be

A A A A A B A B B B B

which contains relatively few runs.

Now consider two populations which instead differ with respect to spread. Their distributions might be something like the ones shown in Figure 17.4. A possible letter sequence for this could be

A A A B B B B A A A A A A

which is again a sequence with few runs.

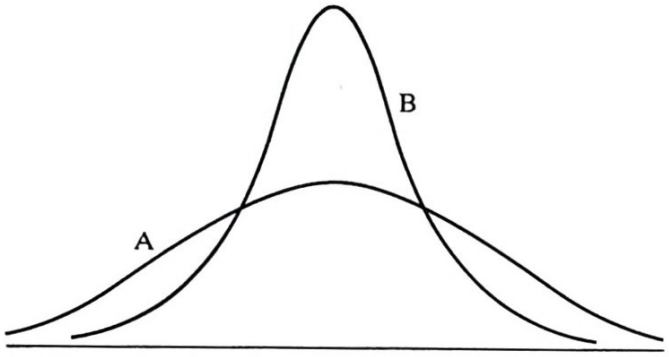


Figure 17.4

Finally suppose the populations differ in both location and spread, as in Figure 17.5. A typical letter sequence is

A A A A A B B B B A A

yet again a sequence with few runs.

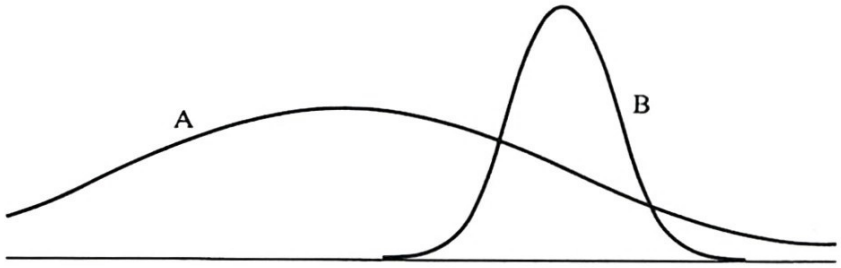


Figure 17.5

PRACTICAL PROBLEM

We will illustrate the Wald–Wolfowitz runs test with the stockbroker's data from Chapter 6, where we were looking for differences in riskiness between fundamental and technical investment methods.

HYPOTHESES

We use the same form of hypotheses as in Chapter 6:

H_0: There is no difference in 'riskiness' between the two methods.
H_1: There is some difference in 'riskiness' between the two methods.

DATA

The percentage gains and losses were as follows:

Fundamental (sample A)	+5	+20	−24	−9	+50	+32	+15	−34
	−19	+28	+41	+11				
Technical (sample B)	−4	+19	+21	+13	+2	−7	+24	+10
	+26	+14	+6	−10	−21	+36	+7	

STATISTIC

As described above, the rationale of the Wald–Wolfowitz test is to count the runs in the combined letter sequence to give a test statistic:

$$T = \text{total number of runs}$$

The combined letter sequence for our data is given below.

−34	−24	−21	−19	−10	−9	−7	−4	2	5	6	7	10	
A	A	B	A	B	A	B	B	B	A	B	B	B	
11	13	14	15	19	20	21	24	26	28	32	36	41	50
A	B	B	A	B	A	B	B	B	A	A	B	A	A

The number of runs is $T=17$, and the sample sizes are $n_1=12$ and $n_2=15$.

CRITICAL REGIONS

As we have seen, the effect of substantial differences in location and/or spread between the two populations is a tendency to reduce the number of runs. Consequently, critical regions for this test are always of the form

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

where the lower critical values for $\alpha=5\%$ are given in Table X. (Note that since the Wald–Wolfowitz test cannot distinguish the direction of a difference, we use α without a subscript to denote the significance level and use the significant levels from the α_2 entries in the table.)

With sample sizes $n_1=12$ and $n_2=15$, the 5% critical region is $T\leqslant 8$.

SOLUTION

Clearly, the calculated value of $T=17$ lies nowhere near the 5% critical region, and so the Wald–Wolfowitz test provides no evidence of any difference in riskiness between the two investment methods.

Discussion

The Wald–Wolfowitz test is very useful in principle (because it tackles a general and widespread task), but is less so in practice (because of low power). To compensate for its relatively low power, it is probably wise in general not to use it unless sample sizes are moderately large – certainly into double figures.

Large sample sizes

The approximate critical values for large sample sizes are as for the ordinary runs test using the expression for T_L.

Assignments

In addition to trying out the various two-sample procedures given in this chapter on the assignments in Chapters 5 and 6, we give below a problem from quality control to investigate for non-randomness.

1 Quality control

In a particular manufacturing process, items can be classified as being defective or non-defective. Samples of 50 items are taken every 6 h and the percentage number of defectives calculated; the results of 25 such samples are given below.

Sample number	1	2	3	4	5	6	7	8	9	10	11	12	13
Percentage defective	8	12	7	5	12	12	4	11	4	7	5	6	8
Sample number	14	15	16	17	18	19	20	21	22	23	24	25	
Percentage defective	5	10	6	8	9	11	6	10	7	5	4	10	

Task: Carry out runs tests to check these data for non-randomness.

18

More tests using frequency data

Introduction

The chi-squared test in Chapter 12 was designed to detect association between two variables that were both measured on a relatively crude 'categorized' scale rather than an ordinary numerical scale. Thus the data used for analysis were the frequency counts of the number of observations falling into each combination of the categories. The statistic for this test was

$$\chi^2 = \sum \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}}$$

where the summation covered all combinations of categories (i.e. all the cells in the contingency table) and the expected frequencies were calculated to reflect the statement in the null hypothesis H_0 of no association between the variables. Evidence to support the alternative hypothesis of association between the variables was indicated by large values of χ^2, caused by large discrepancies between the observed and expected frequencies. Consequently, the appropriate critical regions for rejecting H_0 were of the form

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

However, as we indicated in Chapter 12, there are some unpleasant theoretical difficulties in finding exact critical values; indeed, in most situations, they do not exist! If the sample size is large enough, then the difficulties are effectively removed by the fortunate fact that approximate critical values are obtainable from chi-squared tables. In this chapter, we first study an alternative procedure, which may be used when the sample size is too small for this approximation to be trusted; this procedure is usually known as Fisher's exact test. Unfortunately, it has two drawbacks, namely (a) it is only appropriate for 2×2 contingency tables (i.e. where each variable can only be classified into two categories), and (b) the calculations of critical values depend on an assumption that is not usually met in practice! However, later in the chapter we cover two particular applications where the assumption is satisfied; the first is a particular case

of Mood's median test, which we introduced in Chapter 13, and the other is a two-sample test for differences in spread, known as the Mood–Westenberg interquartile test.

FISHER'S EXACT TEST

Introduction

Fisher's exact test is a method of analysing 2×2 contingency tables which may be carried out even when the sample size is too small for the chi-squared approximation to be valid. It is called an 'exact' test because, unlike the analyses we have previously seen, probability distributions are based on exact computations rather than chi-squared approximations. However, as we mentioned above, there is also a validity problem for Fisher's test, which is that the calculations providing the critical values are conditional on both the row and column totals (the marginal totals) being *fixed*. Of course, in most situations they are not fixed (i.e. they are not known prior to the collection of data), so traditional practice has generally been to ignore this fact simply because this is the only technique available for analysing a 2×2 table based on a small sample! However, it should be noted that the two tests that complete this chapter, Mood's median test and the Mood–Westenberg interquartile test, are special cases where the condition on the marginal totals is satisfied.

PRACTICAL PROBLEM

An international drug company is developing a new treatment for leg ulcers. The most commonly used treatment has been to cleanse and then cover the infected area; the response to such treatment is generally slow with, in some cases, several months elapsing before improvement is observed. The company hopes that their new drug will accelerate the healing process.

The company's R&D section were asked to carry out a pilot survey on a prototype of the drug. Depending on the results of the survey, the company may be encouraged either to continue the development of the drug or alternatively to abandon it.

HYPOTHESES

Since the company is interested in seeing whether or not the results of the pilot survey indicate the new treatment to be better than the old, appropriate hypotheses are:

H_0: There is no difference, on average, between responses to the standard treatment and the new drug.

H_1: Response to the new drug is better, on average, than that for the standard treatment.

A more specific statement of the hypotheses, more closely reflecting the sampling experiment being carried out, is:

H_0: The average proportions of people responding in some given time are the same for both types of treatment.
H_1: The average proportion of people responding to the new drug in the given time is higher than that responding to the standard treatment.

It is interesting to note that here we have a one-sided H_1; this is in contrast to the chi-squared test of Chapter 12, where the natural alternative hypothesis is the general, or two-sided, alternative of 'H_0 is not true'. Using Fisher's test, both types of situation may easily be dealt with.

DATA

Eighteen people suffering from leg ulcers were asked to participate in the pilot survey. Ten were randomly chosen for the new treatment; the rest were given the standard treatment. The numbers responding (and not responding) to treatment after two weeks were recorded in the following contingency table.

	Responding	Not responding	
New treatment	8	2	10
Standard treatment	3	5	8
	11	7	18

(Note: although the two row totals can be regarded as 'fixed' by the design of the experiment, the same can certainly not be said of the column totals of 11 and 7. That is to say, another trial of the same experiment – with the same design and sample size – would again be bound to yield row totals of 10 and 8, but the column totals could be any two frequencies that add up to 18. So, as we warned earlier, the precise conditions required by the theory are not always met in practice.)

STATISTIC AND CRITICAL REGIONS

The identification of the statistic for Fisher's test and of the appropriate critical region for the one-sided test is carried out by the simple steps

shown below. Be assured that when you are familiar with them, you will find these steps much quicker to execute than to describe!

STATISTIC

Step 1 Examine the four marginal totals (the two row totals and the two column totals) to find the smallest of them; we shall denote this value by m. In the table above, $m=7$; it is the second of the two column totals. (If there is more than one marginal total taking the smallest value, then simply choose any one of them).

Step 2 Think of what is likely to happen to the two frequencies making up the value of m when the alternative hypothesis H_1 is true compared to when the null hypothesis H_0 is true. Under H_1, one of them would be expected to be smaller than under H_0, the other being larger. The required test statistic is simply the former of these, i.e. that frequency which you would expect to be smaller under H_1 than under H_0.

In our example, $m=7$ is the total frequency in the second column, which represents the frequencies of those failing to respond to the new treatment and to the standard treatment respectively. If H_1 is true, i.e. the new drug is better than the standard treatment, we would clearly expect the number failing to respond under the new treatment to be relatively low and the number not responding to the standard treatment to be relatively high compared with the figures we might expect under H_0. The *former* frequency, i.e. the number failing to respond to the new treatment, is therefore our statistic; we will denote it by f. Inspecting the contingency table, we find that $f=2$.

CRITICAL REGIONS

Step 3 Denote the other marginal frequency to which f contributes by M. In our example, f is the frequency in the first row and the second column. We already have m representing the total frequency in the second column ($m=7$), and so M is the total frequency in the first row: $M=10$.

Step 4 Denote the total sample size by N; here $N=18$. Now look in the section of Table Z headed by $N=18$, and find the entry with the relevant values of m and M in the first and second columns of the table, respectively. Remembering that we have $m=7$ and $M=10$, you will find that it is in fact the sixth entry from the end of this section of the table.

Step 5 Read off the values in the third and fourth columns; these are $C=1$ and $\alpha=1.8\%$.

Step 6 Compare the obtained value of f with C. Remembering that the smaller the value of f, the more significant it is, the possible

conclusions are as follows (these are relevant to one-sided tests, as in our practical problem).

(a) If $f>C$ then the result is not significant at the 5% significance level, i.e. the test has failed to find evidence against H_0.
(b) If $f<C$ then the result is significant at (better than) the $\frac{1}{2}$% significance level, which means that the test has found very strong evidence in favour of H_1.
(c) If $f=C$ then the result is significant at the $\frac{1}{2}\alpha$% significance level, where α is the percentage read from the fourth column. (The reason for the $\frac{1}{2}\alpha$% rather than α% is that the significance levels in the table apply directly to two-sided rather than one-sided tests; see below for details on two-sided applications.)

SOLUTION

In our example, we have $f=2$ and $C=1$. Thus conclusion (a) is appropriate; the result is not significant at the 5% level, and so we have no evidence to support H_1. Of course, the fact that f is only one greater than C is some indication of the truth of H_1, so the result of the experiment should not be taken as a definite recommendation that the project be abandoned.

Two-sided tests

It is very easy to adapt the above procedure for two-sided tests. In this case step 1 is as above but, instead of choosing a specific one of the two frequencies contributing to m as the test statistic, we consider *both* of them in turn. Formally, you could say that we carry out steps 3 to 6 above for both the frequencies; usually, however, it is immediately apparent that at least one of these frequencies is too big for consideration, i.e. it will not get anywhere near the corresponding (low) tabulated value of C.

Since the significance levels for two-sided tests are double those of the corresponding one-sided tests, all significance levels mentioned in the above three possible conclusions must be doubled, i.e. replace the 5% by 10% in (a), the $\frac{1}{2}$% by 1% in (b), and of course the $\frac{1}{2}\alpha$% by α% in (c).

An example of the use of this procedure for two-sided tests is given below in the section dealing with the Mood–Westenberg test.

MOOD'S TWO-SAMPLE MEDIAN TEST

Introduction

In Chapter 13, we introduced Mood's median test in a form suitable for any number ($k\geq 2$) of samples. It is a quick, convenient test for differences in location; and the fact that it is a straightforward application

of a standard technique (i.e. analysing a contingency table using the chi-squared statistic), which is already familiar to many who have attended introductory statistics courses, makes it particularly attractive.

However, as we know, this 'attractive' feature is not so attractive when the samples are small, for then the critical values obtained from the chi-squared tables are less reliable. But, when we have just two samples, we find ourselves in a situation for which Fisher's exact test is appropriate; indeed, as we shall see, it is particularly appropriate because the condition of the marginal totals being fixed, which was not satisfied in the above example, turns out to hold in the case of this two-sample median test. We shall soon see why.

PRACTICAL PROBLEM, HYPOTHESES AND DATA

Let us return to the problem and data used in Chapter 5 involving the proportions of British and Japanese cars breaking down in the first year. The null hypothesis H_0 of equal proportions was being tested against the one-sided alternative hypothesis H_1 that the proportion of British cars breaking down is greater. The data (percentages of different models breaking down) were as follows:

Japanese	3	7	15	10	4	6	4	7				
British	19	11	36	8	25	23	38	14	17	41	25	21

STATISTIC

First we find the overall median of these data. We have 20 observations altogether, so the median is halfway between the 10th and 11th values when arranged in ascending order. The reader may easily confirm that the relevant values are 14 and 15 respectively, giving the median value of $14\frac{1}{2}$. It is then apparent that, of the eight Japanese models, only one of the percentages exceeds $14\frac{1}{2}$, and consequently we arrive at the following contingency table.

	Below overall median	Above overall median	
Japanese	7	1	8
British	3	9	12
	10	10	20

We now proceed as in the previous example. We shall write down the details in terms of the same six steps as before.

Step 1 The smallest marginal total is $m=8$, being the total frequency in the first row (i.e. the number of Japanese models in the survey).

Step 2 If H_1 is true (i.e. Japanese cars are better), we would expect the number of Japanese models with percentages breaking down below and above the overall median to be high and low respectively; the second percentage is therefore the one we want, giving $f=1$.

CRITICAL REGIONS

Step 3 $f=1$ is in the first row (having total frequency $m=8$) and second column, which has total frequency of 10. Thus $M=10$.

Step 4 The total frequency is $N=20$.

Step 5 Entering Table Z at $N=20$, $m=8$ and $M=10$, we read $C=1$ and $\alpha=2.0\%$.

SOLUTION

Step 6 Conclusion (c) is appropriate in this case, i.e. there is evidence in favour of H_1 (that Japanese cars are better, on average, than British cars) at the $\frac{1}{2}\alpha=\frac{1}{2}\times 2.0 = 1\%$ significance level, which of course represents strong evidence for the truth of H_1.

Remarks

First, let us check why the condition of fixed marginals holds with Mood's two sample median test whereas it did not with the first example in this chapter. Clearly the row totals have to be 8 and 12 as these are just the two sample sizes. (For the same reason, the row totals in the previous example were also fixed rather than being variable.) But, in this case, the column totals are also fixed; they have to be 10 each because of the definition of the overall median – half of the observations have to be below the overall median and the other half above. Consequently, the condition of fixed marginals is satisfied with Mood's test.

The condition of fixed marginals has another interesting consequence. You will recall our mention in Chapter 12 of Yates' correction, which for many years was an adjustment generally made to the chi-squared statistic in 2×2 tables. As we discussed in that chapter, research during the 1960s showed that this correction was not useful in general. The reason was that the theory leading to Yates' correction was again based on the condition of fixed marginal totals. As we have just pointed out, this condition is not met in most situations describable by 2×2 tables. However, we have just shown that this condition is satisfied in the case of Mood's two-sample test. The conclusion is that, when the sample sizes are large enough for the chi-squared (approximate) critical values to be used, which means that in effect both sample sizes should be at least 10, the use of Yates' correction does improve the accuracy of the test. This correction consists of reducing the differences between the observed (Ob.) and expected (Ex.) frequencies in the numerator of the chi-squared statistic before squaring, i.e. the corrected statistic is now

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

The same is true for the Mood–Westenberg test to follow, so we will wait until then for an illustration.

Finally, a slight problem arises if the total number of observations is odd, for then of course the overall median is an actual observation. The easiest thing to do in such a case is simply to ignore that median observation, and to reduce by one the corresponding sample size.

THE MOOD–WESTENBERG TWO-SAMPLE DISPERSION TEST

J. Westenberg (1948) published a simple procedure for testing whether or not two populations have the same spread. Because the test can be presented in a form similar to Mood's two-sample test, we shall refer to it as the Mood–Westenberg test.

As with the Siegel–Tukey test (see Ch. 6) for this situation, it is assumed that the two populations have roughly the same average. However, unlike the Siegel–Tukey test, the Mood–Westenberg test does not involve a ranking procedure and so is somewhat quicker to perform. We will illustrate the Mood–Westenberg procedure with the investment problem of Chapter 6.

HYPOTHESES

As in Chapter 6, the hypotheses for this problem are as follows:

H_0: There is no difference in 'riskiness' between the two methods.
H_1: There is some difference in spread between the two methods.

DATA

For convenience, we reproduce the data below.

Fundamental (sample A)	+5	+20	−24	−9	+50	+32		
	+15	−34	−19	+28	+41	+11		
Technical (sample B)	−4	+19	+21	+13	+2	−7	+24	+10
	+26	+14	+6	−10	−21	+36	+7	

STATISTIC

The idea is really quite simple. First, we combine the two samples (as we did for the Siegel–Tukey test) and then obtain the lower quartile (Q_L) and the upper quartile (Q_U) of this combined sample. Now, if H_0 is true, i.e. there is no difference in spread between the two populations, then we would expect to have (roughly) equal number of observations from each

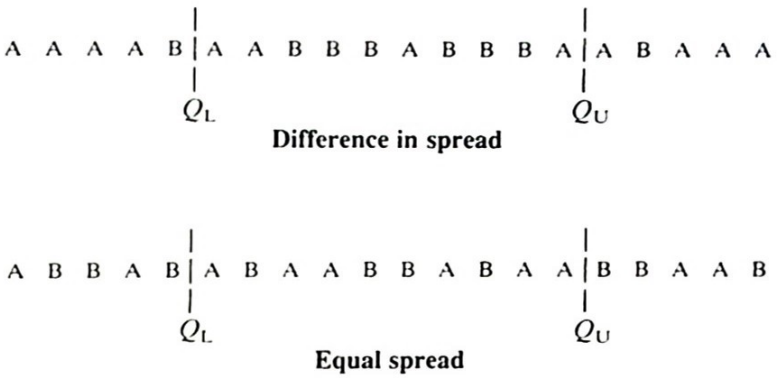


Figure 18.1

sample between Q_L and Q_U and outside this range. This idea is illustrated in Figure 8.1, which also shows typical letter sequences from the two situations.

So all we have to do is to count the numbers of 'A's and 'B's between Q_L and Q_U, i.e. within and outside the interquartile range; we will denote the numbers between Q_L and Q_U by m_A and m_B respectively. This information is conveniently expressed in the form of a 2×2 contingency table; such a table is shown below, where n_A and n_B are the two sample sizes from the respective populations.

	Sample		
	A	B	
Between Q_L and Q_U	m_A	m_B	$m_A + m_B$
Outside Q_L and Q_U	$n_A - m_A$	$n_B - m_B$	$n_A + n_B - (m_A + m_B)$
	n_A	n_b	$n_A + n_B$

We now follow the procedure given for the previous tests in this chapter, but first we recommend writing down the observations on a rough linear scale (as we have done with other two-sample tests). This we have done in Figure 18.2. The values of Q_L and Q_U are easily calculated using the following rules. If N is the total sample size then

(a) if $N/4$ is not a whole number then Q_L and Q_U are the 'next whole number larger than $N/4$' values from each end of the ordered sample;
(b) if $N/4$ is a whole number then Q_L and Q_U are taken to be equal to half the sum of the '$N/4$ and $N/4+1$' values from each end of the ordered sample.

Exhibit 1006
Page 366 of 453

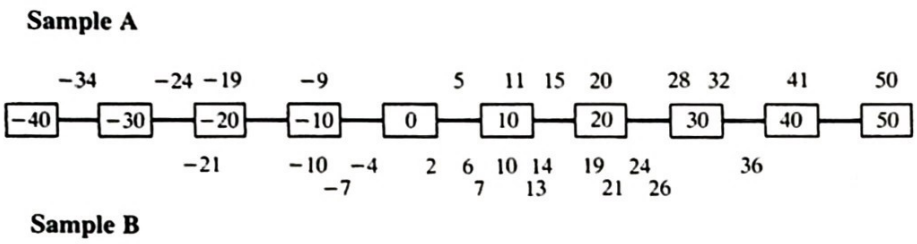


Figure 18.2

In our problem, $N/4=25/4=6.25$ is not a whole number and so rule (a) applies: Q_L is the seventh value from the left and Q_U the seventh value from the right side of the ordered sample. Thus $Q_L=-7$ and $Q_U=24$, both of which are actual data values (in fact, both are from sample B) and so, following the remark on page 343, we discard these values from the counting procedure. We see that we have four A values and nine B values in the interquartile range; the resulting contingency table is given below.

	Sample		
	A	B	
Between Q_L and Q_U	4	9	13
Outside Q_L and Q_U	8	4	12
	12	13	25

We now follow the steps for a two-sided test given on pages 340 and 341.

Step 1 Examining the four marginal totals, we find that the smallest is 12; there are two totals with this value, and we arbitrarily choose the first column total; thus $m=12$.

Step 2 Since we are dealing with a two-sided test, we consider in turn the two values, 4 and 8, which make up the column total $m=12$.

Steps 3, 4 and 5 Consider $f=4$. We have $M=13$, this being the other marginal total to which f contributes. Now with $N=25$, Table Z gives $C=3$ and $\alpha=2.6\%$.

Consider $f=8$. Now $M=12$ and so, with $N=25$, Table Z gives $C=3$ and $\alpha=6.8\%$.

Clearly, the only value of f worth considering is $f=4$ (being the only one close to its corresponding critical value).

Exhibit 1006
Page 367 of 453

CRITICAL REGION AND CONCLUSION

Following step 6, we have $f=4 > c=3$, and so the result is not significant at the 10% significance level (remembering to double α as discussed in the 'Two-sided tests' section on page 341). In other words, we have no worthwhile evidence to support the alternative hypothesis of a difference in spread between the two investment methods. It is interesting to note that this conclusion is in agreement with the conclusion from the Siegel–Tukey test of Chapter 6.

19
Confidence intervals

Introduction

If you are familiar with books dealing with what we have variously called the classical, standard or traditional approach to statistics, you will probably have noticed a very important difference in emphasis in this book (in addition to the obvious concentration on distribution-free methods). So far, we have dealt exclusively with statistical *tests*, whereas in the traditional approach there is usually a comparable amount of attention devoted to *estimation*, and in particular to the formulation of *confidence intervals*. Indeed, in recent years, some statisticians have spoken for the abolition of statistical tests in favour of estimation methods, claiming that the job done by a test can also be done by an estimation method and, in some senses, the estimation method executes the task more accurately and comprehensively. While not subscribing to such an extreme view, the authors do appreciate these arguments in the classical context, and indeed regard them as a consequence of the lack of generality of the traditional, as opposed to the distribution-free, approach. Nevertheless, there are some situations covered earlier in this book which are more or less conducive to the construction of confidence intervals, and the book would not be complete without some mention of such techniques.

In order to understand why estimation methods are less relevant in a distribution-free context, it is opportune here to spell out in a little more detail the fundamental difference between the two approaches. The techniques in this book, which we have always referred to as 'distribution-free', are often termed 'non-parametric' elsewhere. We prefer the former term because it highlights quite unambiguously the essential feature of the approach (i.e. the lack of distributional assumptions on the populations being sampled), whereas 'non-parametric' is less immediate in its meaning and, in some senses, is not a strictly accurate adjective. On the other hand, most traditional statistical methods are definitely 'parametric', i.e. they are defined entirely in terms of *parameters* (such as the mean and standard deviation) of the underlying distributions.

As opposed to the distribution-free nature of the techniques in this book, traditional methods generally require the assumption that the data are drawn from a specific type, or 'family', of distributions, such as normal distributions (or perhaps binomial or exponential, etc.). The

members of these families are 'referenced' or 'indexed' by one or more such parameters so that the specification of the values of these parameters then implies complete knowledge of the distribution; at least, as long as the initial assumptions are correct!

Traditional statistical inference is purely concerned with using the data to give information about these unknown parameters. For example, the parameters of a normal distribution are its mean μ and standard deviation σ, the parameters of a binomial distribution are n and p, and there are a number of families of distributions (e.g. Poisson and exponential) which are indexed by a single parameter, such as the mean. When a family of distributions needs more than one parameter to index its individual members, it may be that in a particular application all but one of the parameters are known (actually or presumed), so that the statistical problem is just concerned with inferences about a single parameter.

There are usually one or more 'obvious' ways of estimating this parameter, in particular by its sample equivalent. Thus one could estimate the population's mean by the mean of the sample, or more relevantly in the current context of distribution-free methods, we could estimate the population's median by the sample median; such single-value estimates are called *point estimates*. The main problem with a point estimate is that no direct indication is given of how close one might expect this estimate to be to the true value. If there is a way of assessing the precision of the estimator, then for many purposes the most sensible way to present the information is by means of a *confidence interval*. As the name implies, this is an interval within which we can be 'confident' that the true value lies. If the basic estimator has high precision (i.e. gives results that are consistently close to the true value) then the corresponding confidence interval will be narrow; but if the estimator is not particularly good (for example, because the sample is too small or because the underlying distribution has large spread) then the confidence interval will be wide. In this context, confidence is quantified by a *confidence level*, which we shall denote by γ, and which represents the level of certainty of the interval containing the true value. The most common (nominal) confidence levels are γ=95% and 99%. Recall that with tests we use 'best conservative' critical regions, i.e. regions whose corresponding significance levels are the nearest available to the nominal level (of, say, 5% or 1%) without exceeding it. In the same way, the actual confidence level of any interval that we shall produce will be the nearest available to the nominal level (of, say, 95% or 99%) but erring 'on the safe side' – that is, the true probability of the interval containing the correct value will be *no less than* the nominal level.

When the statistical problem is simply concerned with inference about a simple parameter (such as the population median), then there is usually an automatic mathematical equivalence between confidence intervals and

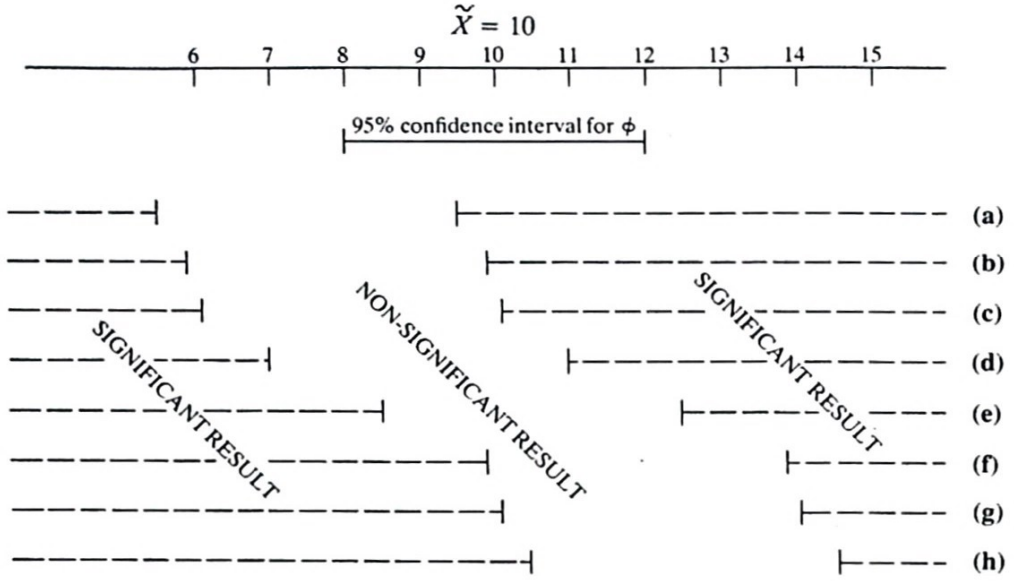


Figure 19.1 Critical regions (broken lines) for testing null hyotheses: (a) $\phi = 7.5$; (b) $\phi = 7.9$; (c) $\phi = 8.1$; (d) $\phi = 9.0$; (e) $\phi = 10.5$; (f) $\phi = 11.9$; (g) $\phi = 12.1$; (h) $\phi = 12.5$.

tests, particularly with two-sided tests. This equivalence is explained by the following rule. Consider testing the null hypothesis H_0 that the parameter is equal to some specific value against the general (two-sided) alternative. *Then H_0 is rejected at the α_2 significance level if and only if this specified value is* not *included in the ordinary confidence interval of level $(100-\alpha_2)\%$*. So, for instance, any value v_0 contained within the γ=95% confidence interval is such that the two-sided test of the null hypothesis that the parameter is equal to v_0 would not give a significant result.

As an example, suppose that the median $\tilde{X}$ of a sample is equal to 10, and that this happens to produce a 95% confidence interval for the population median ϕ of (8,12). Then the 5% critical regions for testing various null hypotheses for the value of ϕ might be as shown in Figure 19.1. Note that X=10 lies *in* the critical regions in cases (a), (b), (g) and (h), which are the cases where the hypothesized value of ϕ is below 8 or above 12, i.e. outside the confidence interval.

In spite of this apparently simple rule, it turns out that, whereas many distribution-free tests are quick and convenient to use, the same cannot be said of distribution-free confidence intervals. Often the notion of confidence intervals is irrelevant anyway, since the practical situation being analysed may be so general that in no way can a *single* parameter describe it. But, supposing this is possible, distribution-free confidence intervals are often tedious to calculate even for quite small samples, while for larger samples the methods can be extremely cumbersome and impracticable except with the aid of a computer (see the programs at the end of this chapter).

The first of the three methods to be described in this chapter is, however, an exception – this is perhaps not surprising since its theory is linked to that of the sign test. It is a method for finding confidence intervals for a population median ϕ. The second technique concerns the estimation of the difference in location (particularly medians) between two populations; that technique is related to the Mann–Whitney test. Finally we have a method based on Wilcoxon's signed-rank test, which therefore is most usually applied to estimating the average change in a matched-pairs situation (although it is also applicable to estimating the median of a single population if that population's distribution is symmetric – which, if you remember, was the requirement demanded in Chapter 3 if Wilcoxon's test was to be used).

CONFIDENCE INTERVALS FOR THE MEDIAN

Introduction

This unusually easy method for finding confidence intervals for the population median ϕ is based on ideas related to those involved in the sign test. Indeed we shall use the critical values for the sign test (Table C or page 29 of *EST*) as an integral part of the procedure.

PRACTICAL PROBLEM

We shall refer to the problem used to illustrate the sign test in Chapter 3. The data given were the times to process orders in an administrative process; for convenience, we reproduce the data below.

16 h 30 min	14 h 00 min	5 h 40 min	9 h 10 min
11 h 45 min	4 h 20 min	7 h 55 min	10 h 15 min
7 h 45 min	16h 05 min	10 h 05 min	7 h 30 min
9 h 15 min	11 h 55 min	9 h 25 min	10 h 35 min
8 h 20 min	12 h 15 min		

Let us now try to find a confidence interval for the median processing time ϕ; we shall use a confidence coefficient $\gamma=95\%$, which means there is *at least* a 95% chance that the resulting interval will include the true value of ϕ.

METHOD

As may be implied by the rule given in the introduction, we enter the table of critical values for the sign test at the α_2 value given by $\alpha_2=100\%-\gamma$; here we have $\gamma=95\%$, and so the relevant value of α_2 is 5%. For $n=18$, the tabulated entry is 4, but let us denote it by x in general.

The confidence interval is then found as follows. Delete from the sample the x smallest and the x largest values. The required confidence interval is then simply *the interval from the smallest to the largest values that remain*.

Checking with the data given above, and recalling that $x=4$ in this case, the four highest times are 16 h 30 min, 16 h 05 min, 14 h 00 min and 12 h 15 min. After deleting these times, the highest time left is 11 h 55 min. The four shortest times are 4 h 20 min, 4 h 35 min, 6 h 30 min and 8 h 05 min; after deleting these times, the shortest time remaining is 8 h 10 min. Hence the required confidence interval for the median processing time is (8 h 10 min, 11 h 55 min), and we are 95% confident (in the sense defined above) that the median processing time lies in this range.

Discussion

This technique shown for obtaining confidence intervals for ϕ is very easy to carry out. This is certainly not true for most other distribution-free confidence interval methods – including the remaining two to be presented in this chapter, in spite of the fact that they are also based on very well known distribution-free tests.

There is a further point worth mentioning. Classical methods for finding confidence intervals for a population mean μ are restricted to producing intervals having the point estimate (viz. the sample mean $\overline{X}$) at their centre. There is no such restriction here. In the above case, the natural point estimate of ϕ is the sample median $\tilde{X}=\frac{1}{2}$(9 h 25 min + 10 h 05 min) = 9 h 45 min, whereas the centre of the confidence interval is at 10 h 02$\frac{1}{2}$ min.

Theory

The easiest way to demonstrate the theory of this technique is first to illustrate it with the above numerical example and then to generalize.

Let us recall the details of the sign test and the interpretation of a critical value. Suppose we were carrying out a *two-sided* test of the null hypothesis H_0 that the median ϕ is equal to some value, m say. (In Chapter 3, m was 8 h.) We represented each observation in the sample by a + or − sign according as to whether it was greater or less than m (to keep this explanation as simple as possible, presume for the moment that there are no observations equal to m). The test statistic was the number of + or − signs, whichever turned out to be the smaller. How exactly, therefore, do we interpret the fact that the critical value corresponding to the $\alpha_2=5\%$ significance level is 4? Referring to the basic formulation of statistical tests in Chapter 2 (if necessary), the meaning is as follows. If H_0 is true, i.e. $\phi=m$, *the total probability of there being 4 or less observations less than m, or 4 or less observations greater than m, is less than or equal to* $\alpha_2=5\%$. In fact, using the binomial tables on page 11 of *EST*, this probability is $2\times0.0154=0.0308$ or 3.08%.

So that we can write this italicized statement in a form more appropriate for the current purpose, it is convenient to introduce a standard notation for the *order statistics* in a sample. This is simply that $X_{(1)}$ represents the minimum value in the

Exhibit 1006
Page 373 of 453

sample, $X_{(2)}$ the next smallest, $X_{(3)}$ the third smallest, and so on up to $X_{(n)}$ which is the maximum observation in the sample. With the data in our example, $X_{(1)}$ = 4 h 20 min, $X_{(2)}$ = 4 h 35 min, $X_{(3)}$ = 6 h 30 min, $X_{(4)}$ = 8 h 05 min and so on up to $X_{(18)}$ = 16 h 30 min.

Referring to the italicized statement above, note that if there are four or less observations less than m, this implies that $X_{(5)}$ must *exceed* m. Similarly, if there are four or less observations that exceed m, this implies that $X_{(14)}$ is *less* than m. We therefore conclude that the probability of either $X_{(5)}>m$ *or* $X_{(14)}<m$ is no greater than α_2=5% (3.08% in this case). By simple logic (or by elementary laws of probability), this statement can be reversed to claim that the probability of both $X_{(5)}\leq m$ *and* $X_{(14)}\geq m$ is no less than $100\%-\alpha_2$, i.e. no less than 95%, and in fact is equal to $(100-3.08)\%=96.92\%$ in our example. But the event now being referred to can be written $X_{(5)}\leq m\leq X_{(14)}$, i.e. m lies in the interval $(X_{(5)}, X_{(14)})$. Therefore, finally, remembering that this argument depends on the truth of H_0: $\phi=m$, it follows that we can claim to have at least 95% confidence that ϕ lies between $X_{(5)}$ and $X_{(14)}$ – and this is precisely what was claimed in the illustration above.

The generalization now follows in a straightforward manner. Given a nominal confidence coefficient $\gamma\%$, we enter the table of critical values of the sign test at $\alpha_2=(100-\gamma)\%$, reading off the value c, say. This implies that the probability of there being c or fewer observations less than the true median ϕ, or of there being c or fewer observations greater than ϕ, totals at most α_2. This statement can be reversed to produce the fact that the probability of $X_{(c+1)}\leq\phi\leq X_{(n-c)}$ is at least $100\%-\alpha_2=\gamma\%$ – and this is precisely what is required for $(X_{(c+1)}, X_{(n-c)})$ to be the required confidence interval.

One can also generalize the above use of the binomial tables to find the *exact* confidence level of any particular interval. The theory both of the sign test and o this method of producing confidence intervals is based on the fact that, in random sample of size n from a continuous distribution, the number S o observations exceeding the median (or less than the median) is binomially distributed with parameters n and $p=\frac{1}{2}$. We used this in Chapter 3 to find the exact significance level corresponding to any critical region. The method for finding an exact confidence level is as follows. Let the critical value found from Table C (or page 29 in *EST*) be c; then from the cumulative probability tables of the binomial distribution (pages 5, 7, 9 and 11 in *EST*) find $F(c)$, i.e. Prob$(S\leq c)$. (This is therefore the exact value of α_1, the significance level of the one-sided test having critical region $S\leq c$.) Now double α_1 to obtain α_2 and calculate $\gamma=100\%-\alpha_2$; this is the required exact confidence level γ.

Large sample sizes

For sample sizes n greater than those in Table C (or the table on page 29 of *EST*), we can use the normal approximation to the binomial distribution. This gives extremely accurate results, providing that the continuity correction is used. In fact, the normal approximation is accurate for much smaller values of n, and it can therefore also be used to estimate the exact confidence level of an interval for those cases not covered by Table C. We demonstrate this idea below.

We have already seen in Chapter 3 how to use the normal approximation to provide critical values for the sign test. So there is really nothing new to learn

here. In order to produce a confidence interval of nominal level γ, we find the critical value that would be appropriate for a two-sided test of significance level $\alpha_2=100\%-\gamma$, and use it as shown above.

Suppose, for illustration, we had a sample of size $n=120$ and wanted to produce a 99% confidence interval for ϕ; $\gamma=99\%$ corresponds to $\alpha_2=1\%$ and thus to $\alpha_1=\frac{1}{2}\%$. We would like to find the value of c for which $\alpha_1=\text{Prob}(S\leqslant c)\approx\Phi[(c+\frac{1}{2}-\frac{1}{2}n)/\frac{1}{2}\sqrt{n}]$ (see Ch. 3); or rather, seeing that we cannot expect to do this exactly, we need to find the value of c such that this probability is just under $\alpha_1=\frac{1}{2}\%$. From Table A, we find that $\frac{1}{2}\%=\Phi(-2.5758)$, and so we solve the equation

$$\frac{c+\frac{1}{2}-\frac{1}{2}n}{\frac{1}{2}\sqrt{n}} = -2.5758$$

for c, remembering that $n=120$. This gives

$$c=-(\tfrac{1}{2}\sqrt{n})\times 2.5758-\tfrac{1}{2}+\tfrac{1}{2}n=45.39$$

This implies that if we take $c=45$, $\text{Prob}(S\leqslant c)$ will be *less* than $\frac{1}{2}\%$ while if we take $c=46$, $\text{Prob}(S\leqslant c)$ will *exceed* $\frac{1}{2}\%$. Therefore 45 is the critical value that would have been tabulated if the tables had gone as far as $n=120$. Thus the required confidence interval is $(X_{(45+1)}, X_{(120-45)})$, i.e. $(X_{(46)}, X_{(75)})$.

Finally, we show how to use the normal approximation to estimate the exact confidence level of an interval. For illustration, we will use the same case as earlier, i.e. the interval $(X_{(5)}, X_{(14)})$ when $n=18$, which was the interval appropriate for the nominal confidence level $\gamma=95\%$. We have already evaluated the exact confidence level, using binomial tables, as 96.92%; let us see how close the normal approximation method comes.

The method approximates $\text{Prob}(S\leqslant c)=\text{Prob}(S\leqslant 4)$ by

$$\Phi\left(\frac{(4+\frac{1}{2}-\frac{1}{2}\times 18)}{\frac{1}{2}\sqrt{18}}\right) = \Phi(-2.121)$$

From Table A, this is equal to 0.0170 or 1.70%. This plays the role of α_1, so that $\alpha_2=2\alpha_1=3.40\%$ giving $\gamma=(100-3.40)\%=96.60\%$. This differs from the true value by only 0.32%.

CONFIDENCE INTERVALS FOR LOCATION DIFFERENCES

Introduction

As we have seen, a confidence interval is usually constructed in association with a particular test; we have just seen a method of constructing an interval for the population median ϕ based on the sign test. Recall the basic principle that the confidence interval contains all values of the quantity being estimated which would not be rejected by a

Exhibit 1006
Page 375 of 453

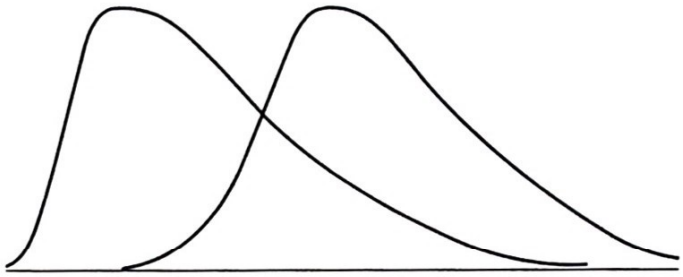


Figure 19.2

two-sided test; the confidence level γ of the interval and the significance level α_2 of the test are complementary, i.e. $\gamma+\alpha_2=100\%$. The 'quantity' being estimated in this section is the difference in location between the two populations that have been sampled. If we assume that the populations' distributions may differ *only* in location, then one distribution is simply a shifted version of the other, as for example in Figure 19.2. It is that shift which is being estimated and, under the above assumption, it is equal to the difference between the two medians, or means (if they exist) or, indeed, the values of any well defined measure of location.

The test associated with the confidence interval method described in this section is the Mann–Whitney test. After first going through the mechanics of the method, we shall then discuss how and why the basic principle just mentioned applies in this case. As we warned you at the beginning of the chapter, the technique is a little unwieldy! It can be carried out graphically or arithmetically; we shall cover both, as it is a matter of personal taste which one prefers.

PRACTICAL PROBLEM

We shall illustrate the technique with the situation and data used in Chapter 5. So now we are attempting to find confidence intervals for the difference between median percentages of the British and Japanese cars breaking down during their first year of use. This is a case where confidence intervals may well be of more interest than the test, because it was virtually certain from the data that H_0 had to be rejected, even without carrying out the tests. The data for this problem, i.e. number of breakdowns, are reproduced below for convenience.

Japanese	3	7	15	10	4	6	4	7				
British	19	11	36	8	25	23	38	14	17	41	25	21

GRAPHICAL METHOD

In Figure 19.3, we have set up scales on ordinary graph paper with the x axis covering the values in sample B and the y axis covering the values in

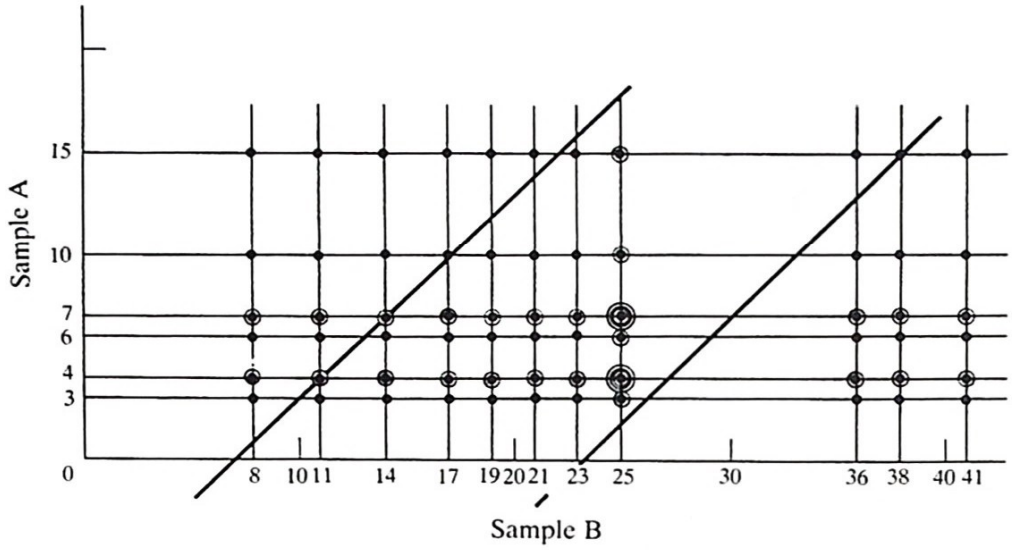


Figure 19.3

sample A. The scales *must* be the same on both axes. We mark sample B's values on the x axis and draw verticals through those points; similarly we mark sample A's values on the y axis and draw horizontal lines through them. If all the observations in each sample are distinct from each other (i.e. there are no tied observations within either sample), there are $n_A n_B$ points of intersection between the vertical and horizontal lines just drawn. If there are ties, the corresponding points should be circled one or more times to indicate such duplication. (Note that in Figure 19.3, points corresponding to pairs of tied values in *both* samples are circled three times, corresponding to quadruple points of intersection.) Thus the total number of points and circles in the diagram is $n_A n_B = 8 \times 12 = 96$. comprising 66 points, 24 single circles and two triple circles (i.e. six circles).

Suppose we want the confidence level γ to be at least 95%. Similarly to finding intervals for ϕ using the sign test, we enter the Mann–Whitney table (Table G or page 30 in *EST*) in the $\alpha_2 = 5\%$ column (remember that $\gamma + \alpha_2 = 100\%$). With sample sizes 8 and 12, the critical value is 22.

Now consider a line, angled at 45° to the horizontal, approaching the graph from the top left-hand corner, and let it travel into the picture until it reaches the place where the number of points above or on it first exceeds 22. This is the upper 45° line shown in Figure 19.3. Similarly, consider a 45° line approaching the graph from the bottom right-hand corner, and move it into the picture until the number of points below or on it first exceeds 22; this is the lower 45° line in Figure 19.3.

A 45° line has an equation of the form $x - y = \text{constant}$, the constant being found by subtracting the y coordinate from the x coordinate of any point on the line. By this means, it is easily seen that the two lines in