thus obtained must be a sensible 'compromise' value for the statistic.

If there are several tied values in the data, this procedure could become tedious, but at least it provides a principle on which to work. Fortunately, in the Mann–Whitney test, this mean value of the statistic may be found by an alternative, and much quicker, method. If using the rank-sum calculation, tied observations should be given the average rank of the positions they cover in the letter sequence, just as we did for Wilcoxon's signed-rank test in Chapter 3. Or, if using the Mann–Whitney pairs-count method, the same result is obtained by regarding every tied (A,B) pair as simply contributing $\frac{1}{2}$ to whichever count is being carried out. To illustrate this, let us slightly alter the numbers in our example. Suppose that the best-performing British model suffered a 7% breakdown rate instead of the original 8%, and also that the model originally recorded as 14% is now 15%. The ordered combined samples and the corresponding letter sequence are now as follows, where tied values (involving items from both samples) are boxed:

Percentage	3	4	4	6	7	7	7	10	11	15	15	17	19	21
	A	A	A	A	A	A	B	A	B	A	B	B	B	B
Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Percentage	23	25	25	36	38	41								
	B	B	B	B	B	B								
Position	15	16	17	18	19	20								

Consider first the Wilcoxon rank-sum calculation. The three tied values at 7 occupy positions 5, 6 and 7 in the letter sequence, so we assign the average rank (5+6+7)/3=6 to all three. The two tied values at 15 occupy positions 10 and 11, and so we give each a rank of $10\frac{1}{2}$. So now

$$R_A = 1+2+3+4+6+6+8+10\tfrac{1}{2} = 40\tfrac{1}{2}$$

giving $U=40\frac{1}{2}-\frac{1}{2}\times 8\times 9 = 4\frac{1}{2}$, which is of course still extremely significant.

Alternatively, consider the direct calculation of U as the number of (A,B) pairs with the B less than the A. If we compute this as above by summing the number of 'B's less than each A, we now say there is $\frac{1}{2}$ of a B preceding each of the two 'A's at value 7, and $2\frac{1}{2}$ 'B's preceding the A at 15. This gives:

$$U = 0+0+0+0+0+\tfrac{1}{2}+\tfrac{1}{2}+1+2\tfrac{1}{2} = 4\tfrac{1}{2}$$

as before.

In practice we treat ties in this way, whether they arise from inaccurate measurements of data from continuous distributions or because the population distributions are in fact discrete.

Null distributions

To compute the null distribution we need a stronger form of the null hypothesis than that stated earlier, namely one that states that the populations from which the two samples are drawn have identical continuous probability distributions.

Equivalently, the 'A's and the 'B's can effectively be regarded as being drawn from the same source. In other words, given the $N=n_A+n_B$ numbers constituting the data, all allocations of n_A of these numbers to sample A and the rest to sample B have the same probability. This implies that, *when H_0 is true, all possible letter sequences* (having n_A 'A's and n_B 'B's) *are equally likely*. (As usual in theoretical considerations, we assume there to be no ties.)

This important idea provides the principle upon which the null distribution of *any* two-sample distribution-free statistic may be calculated. There are

$$\binom{N}{n_A} = \binom{N}{n_B} = \frac{N!}{n_A!n_B!}$$

possible letter sequences. (This follows since, for example, each letter sequence depends on the choice of the n_A out of N positions into which the 'A's are placed; the number of ways of choosing n_A positions out of N available is $\binom{N}{n_A}$.) If the test statistic is computed for each of them, then the probability of the statistic taking any specified value is given by the proportion of times it takes that value in these $\binom{N}{n_A}$ sequences. Of course this is an impracticable proposition if the sample sizes are large. But it is viable for smallish sample sizes, especially with a computer; more efficient methods, usually involving recurrence relations, can often be devised for finding these probabilities for larger samples.

Consider, for example, the one-sided Mann–Whitney test with sample sizes $n_A=3$ and $n_B=6$, where U is defined as the number of (A,B) pairs having the A greater than the B. The number of possible letter sequences is $\binom{9}{3} = 84$. As a partial illustration, let us calculate the probabilities of $U=0$, 1, 2 and 3. There is clearly only one letter sequence giving $U=0$, namely AAABBBBBB. There is also only one giving $U=1$, namely AABABBBBB. Two can be found giving $U=2$: they are AABBABBBB and ABAABBBBB. There are three with $U=3$: AABBBABBB, ABABABBBB and BAAABBBBB. Thus $\text{Prob}(U=0) = \text{Prob}(U=1) = 1/84$, $\text{Prob}(U=2) = 2/84$ and $\text{Prob}(U=3) = 3/84$. As critical regions are of the form $U \leqslant$ critical value, let us check the validity of the entries for these sample sizes in Table G. Our calculations give

$$\begin{aligned}
\text{Prob}(U=0) &= 1/84 = 0.0119 \\
\text{Prob}(U\leqslant 1) &= \text{Prob}(U=0)+\text{Prob}(U=1) = 2/84 = 0.0238 \\
\text{Prob}(U\leqslant 2) &= 4/84 = 0.0476 \\
\text{Prob}(U\leqslant 3) &= 7/84 = 0.0833
\end{aligned}$$

Thus there are no critical regions available for $\alpha_1=1\%$ or $\frac{1}{2}\%$, but the best conservative critical regions for $\alpha_1=2\frac{1}{2}\%$ and 5% do exist and are $U\leqslant 1$ and $U\leqslant 2$ respectively. These results concur with the tabulated critical values in Table G.

Large sample sizes

The complete null distribution of U for the one-sided test with sample sizes $n_A=3$ and $n_B=6$ is sketched in Figure 5.2. You will see immediately (as in the similar example for the Wilcoxon signed-rank test) that not only is it symmetric but its shape is very reminiscent of the normal distribution. Indeed, as sample sizes increase, it becomes ever closer to the normal distribution. The mean of the

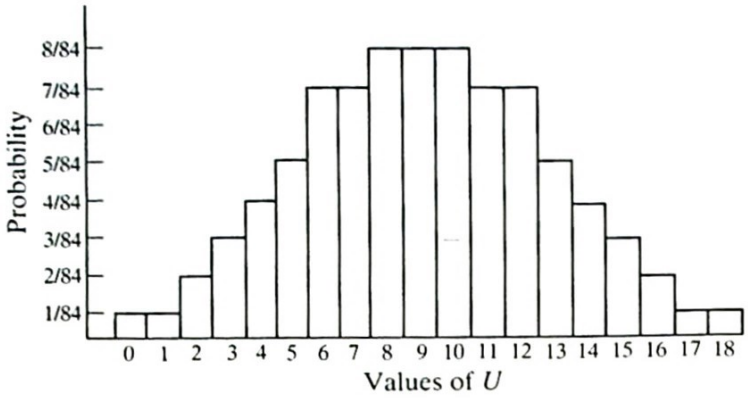


Figure 5.2

distribution is $\mu=\frac{1}{2}n_An_B$ (obviously, since the distribution is symmetric and ranges from 0 to n_An_B) and the standard deviation is $\sigma=\sqrt{n_An_B(N+1)/12}$.

To illustrate the use of the resulting normal approximation we consider the final case given in Table G, namely $n_A=n_B=20$. The value of μ is then $\frac{1}{2}n_An_B=200$ and σ is 36.97. To find what the approximation produces at $\alpha_1=2\frac{1}{2}\%$ (equivalent to $\alpha_2=5\%$ significance level), we calculate $X=\mu-1.96\sigma-\frac{1}{2}$ = $200-1.96\times36.97-\frac{1}{2}$ = 127.04 (the '$-\frac{1}{2}$' is the continuity correction) and then round down; this gives 127, which agrees with the tabulated value.

Strictly speaking, the above formula for the standard deviation is inaccurate if the data include tied observations. Usually this is not troublesome unless there are a lot of ties. However, here are the details of the necessary correction in case you ever need it.

Denote the number of observations tied at a particular value by t (so $t=2$ or 3 or 4 or . . .), and calculate $t^*=t^3-t$ (i.e. 6 or 24 or 60 or . . .). Add up the t^* values for all groups of tied observations, and denote the sum by T. For example, if there are four pairs of tied observations, two sets of three tied observations and one set of four tied observations, then $T=4\times6+2\times24+1\times60=132$. The standard deviation of the null distribution is then calculated using

$$\sqrt{\frac{n_An_B(N^3-N-T)}{12N(N-1)}}$$

(You might like to check that, if you put $T=0$ in this formula, you do of course get the simpler expression for the standard deviation quoted above.)

Relationship between the rank sums and U*

We have stated that the two possible values of U^* (U_A and U_b) are given respectively by $R_A-n_A(n_A+1)$ and $R_B-n_B(n_B+1)$. To see why these relationships hold, let us compare the calculation of R_A from our data with the corresponding computation of U_A:

U_A	0	0	0	0	0	0		1			3											(U_A=4)
Letter sequence	A	A	A	A	A	A	B	A	B	B	A	B	B	B	B	B	B	B	B	B	B	
R_A	1	2	3	4	5	6		8			11											(R_A=40)

If you compare the ranks included in the computation of R_A with the corresponding contributions to U_A, you will see that they differ respectively by 1, 2, 3, . . . , 7 and 8; i.e.

$$\begin{aligned} U_A &= 0 + 0 + 0 + 0 + 0 + 0 + 1 + 3 \\ &= (1-1)+(2-2)+(3-3)+(4-4)+(5-5)+(6-6)+(8-7)+(11-8) \\ &= 1+2+3+4+5+6+8+11 - (1+2+3+4+5+6+7+8) \\ &= R_A - \tfrac{1}{2}\times 8\times 9 \\ &= R_A - \tfrac{1}{2}n_A(n_A+1) \end{aligned}$$

as stated above. We have used here the well known mathematical result that, for any positive integer n, $1+2+3+ \ldots +n=\frac{1}{2}n(n+1)$.

Having seen this pattern in a numerical example, it is then easy to generalize the argument. It depends on the fact that the rank of the ith smallest A is equal to i plus the number of 'B's less than that A (for all i from 1 to n_A). You may check that precisely the same pattern holds even if there are tied observations.

TUKEY'S QUICK TEST

Introduction

The years after World War II saw considerable development of distribution-free statistical methods. Nevertheless, at a meeting of the Industrial Applications Section of the Royal Statistical Society held in Nottingham in 1956, a Mr W. E. Duckworth was reputed as speaking 'vigorously' of the need for a further procedure that would be easier both to use and to teach than those so far available. Less than three years later, Professor John W. Tukey (1959) published, in the very first issue of the journal *Technometrics*, a paper entitled 'A quick, compact, two-sample test to Duckworth's specifications'! The 'quick' in the title is justified by the fact that the statistic is extremely simple to calculate, irrespective of the sample size, and in particular usually requires the ordering of only a few of the observations rather than all of them as in the

Mann–Whitney test. The 'compact' refers to the unusual property that, in most applications, no tables of critical values are needed!

PRACTICAL PROBLEMS, HYPOTHESES AND DATA

These are just the same as those used to illustrate the Mann–Whitney test.

STATISTIC

When H_0 is true we would expect the 'A's and 'B's in the letter sequence to be well mixed, i.e. with no patterns or trends of any kind being discernible (except by occasional, unlucky coincidence).

The most obvious effect on the letter sequence when there is a real difference in location is a tendency for 'A's to accumulate at one end of the sequence, with 'B's in the majority at the other end. So, for example, if the true average (median) in population B is higher than the average in population A, we would expect to see mostly 'A's at the left-hand end and 'B's at the right-hand end. Recall that this is just the type of letter sequence resulting from the car breakdowns survey; in this case, the truth of H_1 would lead us to expect the sequence to begin with an A and end with a B, as did indeed happen.

Tukey's test begins, and sometimes ends, with a simple step motivated by this fact. If we are testing H_0 against the one-sided alternative H_1 that the median of population B is higher than the median of population A, we first check to see whether the overall maximum value is from sample B and the overall minimum value is from sample A. If this is not true, then the test ends then and there with the decision not to reject H_0 in favour of H_1. We could say that the Tukey statistic, denoted by *Ty*, is zero in these circumstances.

On the other hand, when the overall minimum and maximum observations do come from the 'correct' samples according to H_1, the statistic *Ty* is simply defined as:

the number of 'A's that are less than the smallest value in sample B

plus

the number of 'B's that are greater than the largest value in sample A

In other words, referring to the letter sequence, *Ty* is the sum of the lengths of the two extreme runs – the initial run of 'A's and the final run of 'B's.

If the alternative hypothesis H_1 were that the median of population B is *less* than the median of population A, the letters A and B should be interchanged throughout the last two paragraphs.

It should be easy to see that the computation of *Ty* may easily be carried out without needing to order all the data, i.e. without deriving the complete letter sequence. We shall illustrate this in the 'Solution' section below.

CRITICAL VALUES

The remarkable feature of this test is not so much the simplicity in calculating *Ty* but the simple form of the critical values (at least for most practical purposes). It turns out that, as long as the sample sizes are not very small nor very different from each other, the $\alpha_1=5\%$ and 1% critical regions may be taken as $Ty\geq 6$ and $Ty\geq 9$ respectively, regardless of the actual sample sizes! Thus you need only commit these two numbers to memory, and you then have a test that is indeed 'compact' in the sense described above.

These simple critical values work very well for ratios of sample sizes from 1 up to $1\frac{1}{2}$ or a little higher. For example, they are valid for the numerical example being studied where $n_A=8$, $n_B=12$ and the ratio is $1\frac{1}{2}$. If the sample sizes differ more substantially, some simple amendments to the critical values are available and are mentioned in the 'Null distributions' section below. However, to be on the safe side, we have included a full table of critical values in Table H, which may be consulted if you do not trust these easy guides!

SOLUTION

As already stated, the overall minimum and maximum observations in the car breakdowns data are from samples A and B respectively, consistent with what is to be expected under the one-sided H_1 that Japanese cars have the greater reliability. Thus *Ty* is given by the sum of the lengths of the two extreme runs in the letter sequence. Now recall that the letter sequence was

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

We therefore see that the initial run of 'A's has six members and the end run of 'B's contains nine members. Thus $Ty=6+9=15$. Recalling that the $\alpha_1=5\%$ and 1% critical regions are $Ty\geq 6$ and $Ty\geq 9$ we see that, as with the Mann–Whitney test, this result is extremely significant; it is so far into the $\alpha_1=1\%$ critical region that it would clearly still be significant at much smaller values of α_1. Thus again, even using this very quick-and-easy test, we have quite conclusive evidence of the truth of H_1.

We have mentioned earlier the distinct advantage of Tukey's test that the complete letter sequence is usually not needed. Let us pretend therefore that we only have the original unordered data and see how to

calculate *Ty* quickly *without* ordering them. (You will then realize that this method is easy to apply even with large samples, where otherwise it would take a long time to produce the complete ordering.)

Step 1. Check through the two samples separately, noting the maximum and minimum observations in each sample (say $A_{\min}$, $A_{\max}$, $B_{\min}$ and $B_{\max}$).

Step 2. Immediately we may check whether the overall minimum and maximum observations come from the 'correct' samples as expected under H_1; if not, the test ends right here with the decision not to reject H_0 in favour of H_1.

Step 3. Suppose now that H_1 is as in our example, so that we proceed only if the overall minimum and maximum observations are $A_{\min}$ and $B_{\max}$ respectively. (Note that, if the problem had involved the opposite H_1, i.e. that the median of population A were greater than the median of population B ($\phi_A > \phi_B$), then A and B need interchanging throughout this step.) Trace through sample A, noting all values (we know there is at least one) that are less than $B_{\min}$; similarly trace through sample B, noting *all* values that exceed $A_{\max}$. *Ty* is then the sum of these two counts.

Illustrating this procedure with the car breakdowns data, we find:

Step 1. $A_{\min}=3$, $A_{\max}=15$, $B_{\min}=8$ and $B_{\max}=41$.

Step 2. The overall minimum and maximum observations are $3=A_{\min}$ and $41=B_{\max}$; as these are expected under H_1, we proceed to step 3.

Step 3. Tracing through sample A we find six observations less than $B_{\min}=8$, namely 3, 7, 4, 6, 4 and 7; and in sample B we find nine observations that exceed $A_{\max}=15$; these are 19, 36, 25, 23, 38, 17, 41, 25 and 21. Thus $Ty=6+9$, as before.

Discussion

Do not be misled into thinking that because Tukey's test is so easy to execute, then it cannot really be a serious statistical procedure! It is an entirely valid distribution-free test and, while its power is usually not up to that of such sophisticated tests as the Mann–Whitney, it does perform quite well in this respect compared with some other tests that have been proposed elsewhere. Of course, if sample sizes are small, the Mann–Whitney test is quite quick to execute also, and so would probably then be preferred. Otherwise, Tukey's test is an attractive alternative technique, particularly if you come across some data to analyse when you do not have this book or any statistical tables easily to hand!

One-sided and two-sided tests

If we are testing H_0 against the two-sided H_1: $\phi_A \neq \phi_B$ (i.e. that the population medians are unequal, without specifying the direction of the difference), then this

H_1 will be supported by letter sequences that have a preponderance of 'A's at the left-hand end and 'B's at the right-hand end, or vice versa. Thus *Ty* should now be taken as zero *only* if the overall minimum and maximum observations come from the same sample. If they come from different samples, then *Ty* is calculated as the sum of the lengths of extreme runs in the letter sequence as before (or again it is easy to see that this computation may be carried out without forming the letter sequence, as described in the 'Solution' section above).

Under the terms described previously, the critical values 6 and 9 now correspond to significance levels $\alpha_2=10\%$ and 2%. Under similar terms, the critical values corresponding to $\alpha_2=5\%$, 1% and $\frac{1}{2}\%$ may be taken as 7, 10 and 13 respectively, which are also very easy to remember and are often referred to as the '7–10–13' rule. However, as shown in the 'Null distributions' section below, for large sample sizes 7 is not strictly conservative at the $\alpha_2=5\%$ level, although it certainly is the nearest critical value if the sample sizes are exactly or approximately equal. Of course, critical values may also be obtained from Table H as usual.

How do we deal with ties?

It is easiest to discuss ties with reference to the letter sequence in spite of the fact that, in practice, it is not usually necessary to form the complete letter sequence to carry out Tukey's test.

First, it is apparent that ties will only affect the test if they occur at or near the extremes of the letter sequence. In fact, the computation of *Ty* proceeds as above unless there are 'A's tied with B_{min} or B_{max}, or 'B's tied with A_{min} or A_{max}; any ties more central in the letter sequence do not affect the value of *Ty*. Consequently it is usually very easy to deal with any relevant ties by the process mentioned earlier – breaking them in all possible ways and averaging the resulting values of the statistic. We therefore merely demonstrate the principle with a few examples, rather than producing any complicated general rules. In these examples, ties are indicated by boxed letters, and it is presumed that the sample sizes satisfy the conditions for using the easily remembered critical values.

(a) A A A A $\boxed{\text{A A B}}$ ... A B B

Here the tie could be ordered as AAB, ABA or BAA, resulting in *Ty* values of 8, 7 or 6 respectively (unless H_1 is $\phi_A>\phi_B$ in which case $Ty=0$, of course). As the average of these is 7, we can just reject H_0 at the $\alpha_2=5\%$ level in favour of H_1: $\phi_A\neq\phi_B$ or at the $\alpha_1=2\frac{1}{2}\%$ level in favour of the alternative H_1: $\phi_A<\phi_B$.

(b) A A A A $\boxed{\text{A A B}}$ ... $\boxed{\text{A B}}$ B B

The first tie can be ordered in three ways as before, so the initial run of 'A's is of length 4, 5 or 6. The tie near the right-hand end makes the run of 'B's either 2 or 3, according as it is ordered BA or AB; the average rank is $2\frac{1}{2}$, yielding $Ty=5+2\frac{1}{2}=\frac{1}{2}$, a slightly more significant result than before.

(c) A A A A A A A A B ... $\boxed{\text{A B}}$

A tie at either extreme usually results in a non-significant result, even if there is a very long run at the other end. The reason is of course that at least one ordering of the tie must give $Ty=0$. Here we obtain $Ty=8+1=9$ if the tie is ordered AB, but $Ty=0$ if it is ordered BA; the average of these values is only $4\frac{1}{2}$, which is not significant.

Null distributions (including large samples)

Although we are usually avoiding presenting much mathematical detail about null distributions, it seems worth making an exception in this case, because the asymptotic null distribution of Ty (i.e. the null distribution as the sample sizes tend to infinity) is easy to establish.

Mathematical statisticians speak of 'asymptotic' probabilities and probability distributions when considering what happens to probabilities as sample sizes 'tend to infinity', for which their notation is $\rightarrow \infty$. If you are unfamiliar with such ideas and language, simply think of the sample sizes being very large (say, a few million!), and you will see that the results to be derived are not absolutely exactly accurate, but are very close indeed to being true.

Suppose that the sample sizes $n_A, n_B \rightarrow \infty$ with $n_A/N \rightarrow p$ and $n_B/N \rightarrow q=1-p$ where, as before, $N=n_A+n_B$. (Of course, $p=q=\frac{1}{2}$ gives the case of equal sample sizes.) Consider the one-sided test where only letter sequences beginning with an A and ending with a B are considered. For any $h \geq 2$, we now compute the symptotic value of Prob($Ty \geq h$) as $n_A, n_B \rightarrow \infty$ in the following manner. The letter sequences giving rise to the event $Ty \geq h$ may be partitioned into h types:

Type	Letter sequence	
1	A B . . .	. . . (at least $h-1$ 'B's)
2	A A B . . .	. . . (at least $h-2$ 'B's)
3	A A A B . . .	. . . (at least $h-3$ 'B's)
⋮	⋮	⋮
$h-2$	($h-2$ 'A's) b . . .	. . . (at least 2 'B's)
$h-1$	($h-1$ 'A's) B . . .	. . . (at least 1 'B')
h	(at least h 'A's) . . .	. . . (at least 1 'B')

Now, if n_A and n_B are large compared with h, the probability of an A or B being in any prescribed position in the letter sequence will be approximately p or q respectively, almost independently of any other prescribed placings. So the probabilities of occurrence of these types of letter sequence are given approximately by the simple multiplication rule of probability (see Ch. 1) as:

Type of sequence	Asymptotic probability
1	pq^h
2	p^2q^{h-1}
3	p^3q^{h-2}
⋮	⋮
$h-2$	$p^{h-2}q^3$
$h-1$	$p^{h-1}p^2$
h	p^hq

Thus, asymptotically,

$$\text{Prob}(Ty \geq h) = pq(q^{h-1} + pq^{h-1} + p^2q^{h-3} + \ldots + p^{h-3}q^2 + p^{h-2}q + p^{h-1})$$

By using a well known formula* for the sum of a finite geometric series, this gives, for any $h \geq 2$,

$$\text{Prob}(Ty \geq h) = \frac{pq(q^h - p^h)}{q - p}$$

as long as $p \neq q$. When $p = q = \frac{1}{2}$, it is easily seen that the probability of each of the h types of letter sequence is $2^{-(h+1)}$, giving $h2^{-(h+1)}$ altogether. In the case of the two-sided Tukey test, all of these sequences may also be reversed (beginning with a B and ending with an A), so that all the probabilities are then doubled.

Let us look at the case $p = q = \frac{1}{2}$, which corresponds to equal sample sizes, in more detail. Here is a table giving some of the probabilities for a one-sided test:

h	5	6	7	8	9	10
$h2^{-(h+1)}$	0.078125	0.046875	0.027344	0.015625	0.008789	0.004883

We therefore confirm that, for large equal-sized samples, 6 and 9 are indeed the best conservative critical values for $\alpha_1 = 5\%$ and 1% respectively. Also, 10 is the best conservative value for $\alpha_1 = \frac{1}{2}\%$ (corresponding also to $\alpha_2 = 1\%$), but $\text{Prob}(Ty \geq 7)$ is just over 0.025, and so is not quite the best conservative region for $\alpha_1 = 2\frac{1}{2}\%$ (corresponding also to $\alpha_2 = 5\%$). However, it turns out that, for finite sample sizes, the true probabilities are less than these asymptotic ones, and $\text{Prob}(Ty \geq 7)$ is actually less than 0.025 for equal sample sizes up to 24. Even for larger sample sizes, $\text{Prob}(Ty \geq 7)$ is a lot closer to 0.025 than is $\text{Prob}(Ty \geq 8)$, and this is why Tukey recommended that 7 be used as the most appropriate critical value at this level.

For unequal sample sizes, the probabilities $\text{Prob}(Ty \geq h)$ are generally larger than when the sizes are equal, so the critical values may need increasing in these circumstances. A study of the tables of critical values and of the approximate probabilities when $p \neq q$ can yield some approximate amendments. Let us denote by r the larger sample size divided by the smaller, and consider first $1 \leq r \leq 3$; it is fairly unusual to have one sample more than three times as big as the other. Then the following rules give good approximations to the critical values:

$\alpha_1 = 5\%$, $\alpha_2 = 10\%$	6 for $1 \leq r < 7/4$ 7 for $7/4 \leq r \leq 3$	$\alpha_1 = 1\%$, $\alpha_2 = 2\%$	9 for $1 \leq r < 5/3$ 10 for $5/3 \leq r < 9/4$ 11 for $9/4 \leq r < 11/4$ 12 for $11/4 \leq r \leq 3$
$\alpha_1 = 2\frac{1}{2}\%$, $\alpha_2 = 5\%$	7 for $1 \leq r < 4/3$ 8 for $4/3 \leq r < 13/6$ 9 for $13/6 \leq r \leq 3$	$\alpha_1 = \frac{1}{2}\%$, $\alpha_2 = 1\%$	10 for $1 \leq r < 3/2$ 11 for $3/2 < r \leq 2$ 12 for $2 \leq r < 5/2$ 13 for $5/2 \leq r \leq 3$

* $1 + x + x^2 + \ldots + x^{k-1} = (1 - x^k)/(1 - x)$.

We do not recommend you to try to commit these to memory! However, they are brief enough for you to note them down in your diary, so that you can carry out this test even in the absence of detailed tables. The rules give best conservative critical values in most cases and usually the nearest critical values otherwise. (Tukey did suggest rather simpler rules than these, but unfortunately they do not seem to be very accurate.)

If required, asymptotic probabilities for unequal sample sizes may easily be found on a calculator using the above formula for Prob($Ty \geq h$) with $p = n_A/N$ and $q = n_B/N$.

Assignments

1 Lung functions

In 1984, Burghuber, Bergmann *et al.* (*Respiration* **45** (2)) reported the results of an experiment which investigated various lung functions in two groups of patients. Group 1 contained patients whose pulmonary artery pressure was less than 20 mmHg; group 2 contained patients with a pathologically elevated mean pulmonary artery pressure.

Below we reproduce the data for four lung functions, namely vital capacity (VC), forced expiratory volume (FEV), total lung capacity (TLC) and intrathoric gas volume (IGV). All the results are in litres.

	VC	FEV	TLC	IGV
Group 1	2.87	0.53	7.26	5.56
	2.26	1.53	4.87	3.01
	4.98	2.67	8.09	5.74
	3.91	0.73	9.61	7.42
	3.58	2.10	6.22	3.92
	4.33	3.50	6.41	3.18
	3.91	1.13	7.74	5.41
Group 2	4.68	1.51	7.75	5.38
	3.76	1.43	7.28	5.55
	1.51	0.40	8.46	7.48
	2.87	1.67	4.80	2.27
	3.60	2.53	6.68	3.98
	2.94	0.43	9.11	7.80
	2.08	1.60	4.42	2.69

Task: Perform appropriate two-sample tests to examine the differences between the two groups for each of the four lung functions.

2 Vibration measurements

Two machines (X and Y) are used to measure a vibration property of a particular rubber product. The data below give the vibrational force (kg) of samples from each machine.

Machine X	9.6	11.3	14.2	17.0	17.8	18.6	19.3	20.2	22.1	23.0	23.2
	27.3	31.7	34.0	34.8	36.0	43.3	44.3	53.9	59.1	63.5	
Machine Y	19.5	25.6	27.0	33.5	37.6	38.4	39.4	40.3	42.7	45.2	48.3
	48.9	50.1	51.3	51.6	52.2	55.5	86.2				

Task: The items entering each machine are assumed to have equal characteristics so that any differences apparent in the data are due to the machines. Is there any evidence of the two machines producing different results?

3 Children's diet

A dietician is conducting a study on the effect on children's growth of different diets. Part of the study is investigating two religious sects, A and B. Although both sects are vegetarians, A eat only vegetables from below the ground while B only eat vegetables from above the ground. Samples of young boys are selected from each sect. The height and weight of the boys are measured at regular intervals over a number of years.

The heights (cm) of the boys at the age of 12 are as follows:

Sect A	140	140	140	143	135	144	156	149	146	148	144	153
Sect B	151	154	139	159	159	148	140	157	145			

Task: Is there any evidence of a difference in average height between 12-year-old boys from the sect A and the sect B?

Program 5.1 Mann–Whitney two-sample location test

```
100 REM MANN-WHITNEY TEST

110 REM ** INPUT SAMPLES **
120 INPUT "SIZE OF FIRST SAMPLE ";NA
130 INPUT "SIZE OF SECOND SAMPLE ";NB
140 PRINT
150 DIM A(NA),B(NB)
160 PRINT "INPUT FIRST SAMPLE "
170 FOR I=1 TO NA
180 INPUT A(I);"  ";
190 NEXT I
200 PRINT
210 PRINT "INPUT SECOND SAMPLE "
220 FOR I=1 TO NB
230 INPUT B(I);"  ";
240 NEXT I
250 PRINT
```

Continued overleaf

Program 5.1 Continued

```
260 REM ** CALCULATION OF THE TEST STATISTIC **
270 U=0
280 FOR I=1 TO NA
290 FOR J=1 TO NB
300 IF A(I)<B(J) THEN U=U+1
310 IF A(I)=B(J) THEN U=U+0.5
320 NEXT J
330 NEXT I
340 IF U>NA*NB/2 THEN U=NA*NB-U
350 PRINT "THE VALUE OF THE MANN-WHITNEY STATISTIC FOR NA = ";NA;" AND NB
= ";NB;" IS ";U
360 IF NA<=20 AND NB<=20 THEN STOP

370 REM ** CALCULATION OF APPROXIMATE CRITICAL VALUES **
380 INPUT "PERCENTAGE POINT OF STANDARD NORMAL DISTRIBUTION ";Z
390 LET SD=SQRT(NA*NB*(NA+NB)/12)
400 LET MN=NA*NB/2
410 PRINT "CRITICAL VALUE = ";INT(MN-Z*SD-.5)
```

6

Two independent samples – tests for differences in spread

Introduction

In this chapter we turn our attention to another important kind of difference that may exist between two populations. Tentatively assuming that the two populations being investigated have (at least roughly) equal medians, what now concerns us is whether these populations have differing amounts of *spread* (i.e. dispersion, or variability) about that common median. In many situations, it may be just as important to examine this problem as it was the problem of differences in location (medians) in Chapter 5. For example, consider two machines used for filling the 500 g size of Corn Flakes packets. There are bound to be controls on the machines that can alter the average contents of the packets, and these may be set to ensure that the median is a little over 500 g, say around 505 g. But if this results in one machine producing packets whose contents generally vary between 500 g and 510 g, and the other producing packets whose contents vary between 475 g and 535 g, it is fairly obvious which machine is to be preferred! In classical statistics, such variability is usually measured by the standard deviation; however, in distribution-free statistics there is no such clearly preferred measure, and so we shall just use the qualitative term 'spread' when referring to such variability in populations.

As in most other chapters we shall present two tests – a reasonably sophisticated one and a quick-and-easy one. Indeed there are very strong links here with Chapter 5, since the two tests in this chapter are very closely associated, both in philosophy and execution, with the Mann–Whitney and Tukey tests respectively. Because of this connection, in the unlikely event that you have come to this chapter before reading Chapter 5, we would ask you now to go back and read that chapter before continuing with this one. Further, if you have only read the normal-sized type of Chapter 5, please also in addition read the 'One-sided and two-sided tests' section, since the practical problem to be used in this chapter will require a two-sided test, as opposed to the problem in Chapter 5 which involved a one-sided test.

It is worth mentioning at this point that we are not restricted to testing *just* for location differences (as in Ch. 5) nor *just* for differences in spread

(as in this chapter). We might be interested in differences in *both* of these features simultaneously or in *more general* differences that affect the overall shape of the population distributions. Appropriate tests for these more general problems are presented in Chapter 7. (Nevertheless, it is advantageous to isolate the more specific kind of difference of interest if possible, and then use the tests of this chapter or Chapter 5, since they will usually be better able to detect differences of these particular types than will the tests of Chapter 7.)

THE SIEGEL–TUKEY TEST

Introduction

In Chapter 5 we saw that, when there is a location difference between the two populations being compared, the natural effect on the letter sequence is to have 'A's monopolizing one end of the sequence and 'B's monopolizing the other end. Let us now consider what the effect on the letter sequence will be if there is a difference in spread. Consider the two population distributions sketched in Figure 6.1, which clearly differ considerably in spread.

It is easy to visualize how this difference will be reflected in the letter sequence; one of the letters (corresponding to the population having the 'wider' spread) will tend to monopolize both ends of the sequence, while the majority of occurrences of the other letter will be towards the centre. So for the situation illustrated in Figure 6.1, the letter sequence might look something like:

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

Now recall that in the Wilcoxon version of the Mann–Whitney test, we simply rank the observations from 1 to $N=n_A+n_B$, from the smallest to the largest, and then sum the ranks of one of the two samples. This rank sum is then unusually small or large if the sample concerned occurs mainly to the left-hand or right-hand end of the letter sequence (i.e. when there is evidence of a location difference), and tends to be between these

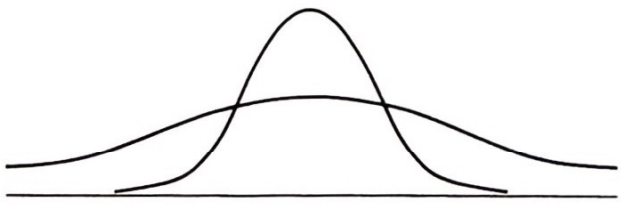


Figure 6.1

two extremes otherwise. What we now need for detecting differences in spread is some alternative ranking or scoring system which will assign, say, large values at *both* extremes of the letter sequence and small values towards the centre (or vice versa). Then if any one of the two letters, say B, does predominate in both extremes, thus indicating a wider spread, the sum of the scores for that sample will be particularly large. Such a sum can therefore form the basis of a useful test for differences in spread.

A number of ingenious ways of applying this principle have been devised. A. M. Mood (1954) published a test where the observations in the letter sequence are ranked as usual from 1 at the left to N at the right, but where each such rank r is then converted to a 'score' of $[r-\frac{1}{2}(N+1)]^2$. With $\frac{1}{2}(N+1)$ being halfway between 1 and N, it is clear that these scores are indeed small in the centre of the letter sequence and large in both extremes. A. R. Ansari and R. A. Bradley (1960) introduced a test involving simpler arithmetic; the positions in the letter sequence are ranked from 1 *at each end* to N or $(N+1)$ (according as N is even or odd) in the middle – these 'ranks' are therefore small in the extremes and large in the middle. Again the test statistic is formed by summing these ranks for either sample.

Unfortunately, both of these tests need special tables of critical values. However, at the same time S. Siegel and J. W. Tukey (1960) produced a test that is very similar in spirit to the Ansari–Bradley test but which has the added convenience of using the Mann–Whitney tables to provide the critical values. Indeed, apart from one vital aspect, the computation of the statistic follows a very similar pattern to that of the Wilcoxon rank-sum version of the Mann–Whitney test. It is mainly because of this rather neat tie-up with previous work that we have selected the Siegel–Tukey test as the most appropriate rank method to present here.

PRACTICAL PROBLEM

Three methods used by investment advisors for analysing and forecasting movements in the stock market are 'fundamental analysis', 'technical analysis' and 'beta analysis'. Beta analysis concentrates on the 'riskiness' of a share, seeking to identify *aggressive* shares and *defensive* shares which are suitable for adventurous and cautious investors respectively. A stockbroker is interested in studying the other two methods of analysis, essentially from a similar point of view. He is satisfied that, 'on the average', or 'in the long run', there is not much to choose between fundamental and technical analysis. What he wants to investigate is whether the methods tend to advise different proportions of aggressive and defensive shares, i.e. whether one method produces more risky advice than the other. (It should be emphasized that 'risky' in this context may not be altogether a bad thing, for although the chances of substantial losses are greater, so are the chances of substantial profits.)

In order to investigate this matter, the stockbroker obtains advice from all his investment advisors on a particular day, telling them that the investments are to be made for a fixed term of six months and then sold. There are 27 advisors in all, 12 favouring fundamental analysis and 15 technical analysis. True to his word, he then monitors the resulting portfolios and six months later calculates the percentage profit or loss in each case.

HYPOTHESES

If one type of advice is more risky than the other, the corresponding percentages will tend to be spread relatively widely. The stockbroker is open-minded about which, if either, type of analysis is likely to prove the more risky. He thus tests the null hypothesis of equality of the percentage spreads against the two-sided alternative that there is some difference in the spreads. So we have:

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

DATA

The percentage gains and losses were as follows:

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

STATISTIC

The tie-up with the Wilcoxon rank-sum calculation stems from the fact that the same set of ranks, 1, 2, 3, . . . , $N-1$, N, is used; but the test is converted to a test for differences in spread rather than location simply by reordering these ranks in such a way as to reflect the above argument. The ordered data and corresponding letter sequence, along with the reordered ranks, are as follows:

Data	−34	−24	−21	−19	−10	−9	−7	−4	2	5	6	7	10	11
Sequence	A	A	B	A	B	A	B	B	B	A	B	B	B	A
Ranks	1	4	5	8	9	12	13	16	17	20	21	24	25	27
Data	13	14	15	19	20	21	24	26	28	32	36	41	50	
Sequence	B	B	A	B	A	B	B	B	A	A	B	A	A	
Ranks	26	23	22	19	18	15	14	11	10	7	6	3	2	

You will see that the reordering scheme is to rank the smallest value as 1, then the largest value as 2 and the one next to it as 3, then the next two values at the left-hand end as 4 and 5, then the next two at the right-hand end as 6 and 7, and so on, progressing in pairs towards the middle of the sequence. (The reason for *not* ranking *one* at a time alternatively at the ends to produce

1 3 5 7 . . . . . 8 6 4 2

is that this makes observations at the left-hand end 'more important' than the corresponding observations at the right-hand end. The Siegel–Tukey scheme is not exactly symmetric, but is the 'fairest' possible share-out of the ranks.)

A rank-sum statistic is then calculated in the usual way, i.e. by summing the ranks for one of the two samples at your choice. This is then converted to a U^* value, and finally to U precisely as was done with the two-sided version of the Mann–Whitney test.

CRITICAL REGIONS

As in the Mann–Whitney test, critical regions are of the form

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

where the critical values are read from the same table as for the Mann–Whitney test, i.e. Table G (or page 30 in *EST*). From this table, we find that for our problem the α_2=5% and 1% critical regions are $U \leqslant 49$ and $U \leqslant 37$ respectively.

SOLUTION

We note that the rank sum for sample A will probably be the easier to calculate, since it is the smaller sample and also it has several values near the ends of the letter sequence. We therefore compute:

$$R_A = 1+4+8+12+20+27+22+18+10+7+3+2 = 134$$

Then, as in Chapter 5, we convert R_A to a U^* value by subtracting $\frac{1}{2}n_A(n_A+1)$:

$$\begin{aligned} U^* &= R_A - \tfrac{1}{2}n_A(n_A+1) \\ &= 134 - \tfrac{1}{2}\times 12\times 13 = 56 \end{aligned}$$

Had we computed R_B instead, and then subtracted $\frac{1}{2}n_B(n_B+1)$, we would

have obtained the other possible value of U^*. As in the (two-sided) Mann–Whitney test, the required final statistic U is the smaller of these two possible values of U^*. To check whether our U^* value (i.e. 56) is the required smaller one, remember that the two values add up to $n_A n_B = 180$, implying that the required value is less than (or, exceptionally, equal to) $\frac{1}{2}n_A n_B = 90$. Therefore, as expected, we have obtained the smaller value, and so the value of U is 56. (You might like to confirm that $R_B = 244$, which gives a U^* value of 124, verifying that the two possible U^* values do indeed sum to $n_A n_B = 180$.)

Checking our value of $U=56$ against the details of critical regions, we find that it is some distance from the $\alpha_2 = 5\%$ critical region. Indeed, from the table of critical values on page 30 of *EST*, we see that $U=56$ is just outside the $\alpha_2 = 10\%$ critical region ($U \leqslant 55$). This is interpreted as either no evidence or, at best, a very low level of evidence in favour of H_1.

Discussion

The Siegel–Tukey test has a couple of disadvantages compared with the other two tests (Mood's and Ansari–Bradley's) mentioned in the introduction. First, it is not exactly symmetric with respect to the letter sequence; thus, if the letter sequence were reversed, it is quite likely that the resulting value of U would not be quite the same. Secondly, the power of the Siegel–Tukey test is a little less than that of Mood's test.

However, both of these 'drawbacks' have only minor effects on the test in practice, and are effectively outweighed in importance by the convenience of being able to use the familiar Mann–Whitney tables to obtain the critical values.

One-sided and two-sided tests

In the case of a one-sided alternative hypothesis H_1, such as 'technical analysis is more risky than fundamental analysis', we proceed as in the Mann–Whitney test to choose a specific calculation for U^* which will tend to give *small* values if this H_1 is true, and to allocate this value to U directly *without* going through the intermediate step of comparing it with $\frac{1}{2}n_A n_B$. If this particular H_1 is true in our particular problem, then the 'B's will tend to predominate in the extremes of the letter sequence. Consequently, we would calculate U specifically using R_B in this case, i.e. $U = R_B - \frac{1}{2}n_B(n_B+1)$. Its value is then checked against the α_1 critical values in the Mann–Whitney table (Table G) to see if it is significantly small. In the above case we obtain $U=124$, which would of course furnish no evidence at all for rejecting H_0 against *this* one-sided H_1 since critical regions are of the form $U \leqslant$ critical value.

Note that, as always, the choice of statistic in this one-sided test is made purely *on the basis of what H_1 suggests*, preferably before the data are inspected (in order to prevent any possibility of the data influencing the choice of hypotheses). Recall from Chapter 2 that H_1 is chosen to interpret properly the practical question being asked, and should therefore not be influenced by the data, which, after all, are being used to judge the truth or otherwise of the hypothesis.

How do we deal with ties?
As previously, the 'ranks' of tied values are simply averaged. In this case, of course, this may well involve averaging ranks such as 1 and 4 to give $2\frac{1}{2}$, which would occur if the two smallest observations were equal.

Null distributions
The particularly clever feature of the Siegel–Tukey statistic is that the null distribution, and hence the critical values, are identical to the Mann–Whitney test. The reason is as follows.

As we know, the principle under which null distributions for two-sample statistics are computed is that, under H_0, all of the possible $\binom{N}{n_A} = \binom{N}{n_B}$ letter sequences are equally likely to occur. In terms of ranks in the Wilcoxon version of the test, this means that (for example) all of the $\binom{N}{n_B}$ possible selections of n_B ranks for sample B out of 1, 2, . . . , N occur with equal probability. In the Siegel–Tukey test, the only difference is that the N ranks are laid out in a different order. So it will still happen that, when all of the $\binom{N}{n_B}$ possible letter sequences are generated, each possible allocation of n_B ranks to sample B occurs once and once only – and thus again each occurs with equal probability. Consequently we obtain the same frequency distribution for values of R_B as before, and hence of U^* and U, so we finish up with precisely the same (null) probability distribution and critical values.

Exercise To convince yourself of this, write out the $\binom{7}{3} = 35$ possible letter sequences for $n_A=3$ and $n_B=4$, and find the frequency distributions of R_A and R_B using first the ordinary Wilcoxon ranks and then the Siegel–Tukey ranks.

Large sample sizes
The related comments for the Mann–Whitney test in Chapter 5 apply here also.

A weakness of distribution-free dispersion tests
Although the Kolmogorov–Smirnov and Rosenbaum tests in Chapter 7 deal with situations where there may be differences in both location and dispersion, there is one particular kind of problem that we have not mentioned and which in fact does not seem to have any good distribution-free solution. This is the problem of detecting a dispersion difference *irrespective* of whether there is a location difference. In classical statistics, this problem is solved very neatly by the F-test since the F-statistic remains completely unchanged if either or both samples have constants added or subtracted from them (representing changes in location). Frustratingly, there is, as yet, no such neat equivalent distribution-free method. Several attempts have been made to solve the problem, but all resulting tests suffer from being rather unpowerful or not truly distribution-free, or both. A naively obvious thing to try is to shift one of the samples so that the resulting two samples both have the same median. But, because that would make letter sequences such as AA. . .AABB. . .BB completely impossible, clearly the basic theory of null distributions is violated, and so any resulting tests are not distribution-free.

It is particularly unfortunate that there appears to be no good distribution-free solution to this problem since several researchers have shown that non-normality can upset the behaviour of the F-statistic to a very considerable extent. The best

attempt at such a distribution-free method appears to be Moses' 'rank-like' tests, and the interested reader should consult *Nonparametric statistical methods* by M. Hollander and D. A. Wolfe (Wiley, 1973) for details.

A TUKEY-TYPE TEST FOR DIFFERENCE IN SPREAD

Introduction

Having read in Chapter 5 about Tukey's quick test for location differences, you may well be able to anticipate what statistic is going to be used here. We need a statistic of similar nature to that used in Chapter 5, but amended so as to be relevant to differences in spread rather than location. Make a guess as to what it is, and then check in the 'Statistic' section below to see whether you are right.

Unfortunately, only one of the two big advantages of Tukey's location test is retained here. The relevant statistic is still extremely easy to compute (if anything, it is even quicker than the location statistic), and again it does not require the production of the complete letter sequence. However the feature of easily memorable critical values does not hold in this case, at least not to any useful extent. It is true that for large equal-sized samples, the values used in Chapter 5 apply, i.e. 6 for $\alpha_1=5\%$ and $\alpha_2=10\%$, 7 for $\alpha_1=2\frac{1}{2}\%$ and $\alpha_2=5\%$, etc. However, the samples need to be larger than in the location tests for these values to apply, and also the critical values are now much more sensitive to the ratio of sample sizes.

One cannot, therefore, usually apply these tests without recourse to tables of critical values, and accordingly these are provided in Table I. (These tables have been specially computed for this book and, to our knowledge, are not available elsewhere.)

PRACTICAL PROBLEM, HYPOTHESES AND DATA

We shall use the same problem and data here as for the Siegel–Tukey test.

STATISTIC

Referring to the letter sequence, the phenomenon that would most obviously indicate a difference in spread is the predominance of one of the samples in *both* extremes. (This is opposite to the location case where, if you remember, the effect we looked for was the predominance of one sample in one extreme and the other sample in the other extreme.) Consequently, in this case the appropriate statistic, *Ts* say, is the sum of the lengths of the extreme runs in the letter sequence, *as long as these are from the same sample*; otherwise $Ts=0$.

You can therefore immediately find the value of *Ts* by referring to the letter sequence produced earlier. However, *Ts* may be found without deriving the whole letter sequence, and we shall therefore calculate *Ts* in the 'Solution' section below directly from the original data, imagining that we have not previously gone to the trouble of ordering all the data.

CRITICAL REGIONS

There is a crucial aspect in which the tables of critical values (Table I) for *Ts* differ from those of other tests. When the sample sizes are unequal, it should be apparent that the values of *Ts* needed for significance will depend very much on whether it is the larger or smaller sample that is found in the extremes. For example, suppose $n_A=5$ and $n_B=20$; then if the overall minimum and maximum observations come from sample A, so that *Ts* is being calculated from the 'A's, only a small value of *Ts* is needed to achieve significance since large values (greater than $n_A=5$) are impossible under any circumstances. In fact, with such a discrepancy in sample sizes, to have the overall minimum and maximum coming from sample A is quite significant by itself: in this case $Ts=2$ is actually significant at the $\alpha_2=10\%$ level! On the other hand, if H_0 is true, it is very likely not only that the overall minimum and maximum will come from the larger sample B but that the extreme runs of 'B's will be quite long – simply because there are far more 'B's than 'A's in the sequence. We shall see from the table that *Ts* needs to be as large as 15 for significance at the $\alpha_2=10\%$ level if being calculated from the 'B's.

Accordingly, unlike other tests, we do not enter Table I with n_1 and n_2 representing the smaller and larger sample sizes. Instead, n_1 represents the size of that sample *from which Ts is calculated*, i.e. which contains the overall minimum and maximum observations in the data. (Remember that if one sample provides the overall minimum and the other the overall maximum, then *Ts* is defined as 0 and H_0 cannot be rejected in favour of a difference in spread.)

In our practical problem, *Ts* is calculated from sample A (see the 'Solution' section below) and so $n_1=12$ and $n_2=15$. From Table I, we then see that the critical regions for $\alpha_2=10\%$, 5% and 1% are respectively $Ts\geq5$, $Ts\geq6$ and $Ts\geq7$.

SOLUTION

The method for calculating *Ts* directly from the data and without producing the letter sequence may be summarized as follows:

Step 1 Find the minimum and maximum observations in each sample, denoted by A_{min}, A_{max}, B_{min} and B_{max}.

Step 2 Immediately we may check whether the overall minimum and maximum observations are from the same sample or not. If they

are not, the test ends at this step with the decision not to reject H_0 in favour of a difference in spread.

Step 3 Suppose that the overall minimum and maximum come from sample B (i.e. $B_{min}<A_{min}$ and $B_{max}>A_{max}$). Trace through sample B, noting every observation that is less than A_{min} and every observation that exceeds A_{max}. The value of the test statistic *Ts* is the total number of observations thus noted.

Applying this procedure to the investments data from earlier in the chapter gives:

Step 1 $A_{min} = -34$, $A_{max} = +50$, $B_{min} = -21$ and $B_{max} = +36$.

Step 2 The overall minimum and maximum observations both come from sample A, so we proceed to step 3.

Step 3 We find just two observations (−24 and −34) in sample A that are less than $B_{min}=-21$, and there are also two (+50 and +41) that exceed $B_{max}=+36$. Thus $Ts=2+2=4$.

A quick glance at the critical regions above shows that we have obtained a very similar result to that using the Siegel–Tukey test; i.e. just failing to reject H_0 at the $\alpha_2=10\%$ significance level.

Discussion

Apart from definitely needing tables of critical values, this test has similar attractive properties to Tukey's location test. As one would expect, its power is not so good as the Siegel–Tukey test. However, if anything, the gap between its power and that of the Siegel–Tukey test is smaller than that between Tukey's location test and the Mann–Whitney test. It is therefore a very worthwhile alternative technique for testing for differences in spread.

One-sided and two-sided tests

A one-sided H_1 specifies which population is to have the greater variability if H_0 is untrue. In this case, *Ts* will only be taken as non-zero if the overall minimum and maximum observations come from the corresponding sample. Thus, with the practical problem used in this chapter, the H_1 specifying fundamental analysis as having the greater riskiness would result in $Ts=4$ again, and this now only just fails to be significant at the $\alpha_1=5\%$ level. Had H_1 specified technical analysis as having the greater riskiness, we would have stopped the test at step 2 with no evidence to reject H_0 in favour of this H_1.

Tied observations

The problem of ties is treated similarly as in the location test, so we will just give three examples for you to check. We assume that we are carrying out either a two-sided test or a one-sided test where we are looking for 'A's in the extremes. Again, boxes show the ties.

(a) A A A A [A A B] . . . B A A $Ts=7$
(b) A A A A [A A B] . . . [A B] A A $Ts=7\frac{1}{2}$
(c) A A A A B A A A A B . . . [A B] $Ts=4\frac{1}{2}$

Null distributions (including large samples)
The theory for asymptotic null distributions is even easier than for the location test. As before, let $n_A, n_B \to \infty$ with $n_A/N \to q$. Now consider the one-sided test, which looks for evidence of population A being the more widely spread. Then, the event $Ts \geq h$ may be partitioned into the following types of letter sequence:

Type	Letter sequence	
1	A B . . .	. . . (at least $h-1$ 'A's)
2	A A B . . .	. . . (at least $h-2$ 'A's)
3	A A A B . . .	. . . (at least $h-3$ 'A's)
⋮	⋮	⋮
$h-2$	($h-2$ 'A's) B . . .	. . . (at least 2 'A's)
$h-1$	($h-1$ 'A's) B . . .	. . . (at least 1 'A')
h	(at least h 'A's) . . .	. . . (at least 1'A')

Letter sequences of types 1 to $h-1$ each prescribe the positions of h 'A's and a B, whereas letter sequences of type h prescribe $h+1$ 'A's and no 'B's.

Thus, asymptotically, $\text{Prob}(Ts \geq h) = (h-1)p^h q + p^{h+1}$, for $h \geq 2$,

$$\text{i.e. } \text{Prob}(Ts \geq h) = p^h[1+(h-2)q]$$

If $p=q=\frac{1}{2}$, this gives $h2^{-(h+1)}$ as in the location test, but the formula is quite different otherwise, as might be expected from the completely different tables of critical values. Also, as mentioned earlier, even in the case of equal-sized samples, the samples need to be quite large for the asymptotic values to be appropriate; for example, with samples of size 20, the $\alpha_2=10\%$ and 5% values are 6 and 7 as expected, but the 2% and 1% values are 8 and 9 instead of the asymptotic 9 and 10 respectively.

Assignments

1 Soil analysis
In 1979, Wright and Wilson ('On the analysis of soil variability, with an example from Spain' *Geoderma* **22**, 297–313) conducted a study into soil composition in the province of Murcia, Spain. Eight sites were selected and 11 samples from each site were analysed for their silt and clay content. The table below gives the percentage contents of silt (s) and clay (c) for each sample.

Site		Percentage contents from 11 samples										
1	s	46.2	36.0	47.3	40.8	30.9	34.9	39.8	48.1	35.6	48.8	45.2
	c	30.3	27.6	40.9	32.2	33.7	26.6	26.1	34.2	25.4	35.4	48.7
2	s	40.0	48.9	48.7	44.5	30.3	40.1	46.4	42.3	34.0	41.9	34.1
	c	35.9	32.8	36.5	37.7	34.3	35.1	36.2	37.7	28.4	28.4	39.8
3	s	41.9	40.7	44.0	40.7	32.3	37.0	44.3	41.8	41.4	41.5	29.7
	c	34.0	36.6	40.0	30.1	38.6	30.0	28.7	34.5	34.6	34.7	32.8
4	s	41.1	40.4	39.9	41.1	31.9	43.0	42.0	40.3	42.2	50.7	33.4
	c	48.3	49.6	40.4	43.0	49.0	49.1	39.6	44.5	42.2	38.2	59.6
5	s	48.6	50.2	51.2	47.0	42.8	46.6	46.7	48.3	47.1	48.8	38.3
	c	44.3	45.1	44.4	44.7	52.1	42.0	41.8	46.0	46.3	37.1	54.9
6	s	43.7	41.0	44.4	44.6	35.7	50.3	44.5	42.5	48.6	48.5	35.8
	c	37.0	31.3	34.1	29.7	39.1	36.1	32.4	33.4	38.3	36.7	39.9
7	s	47.0	46.4	46.3	47.1	36.8	54.6	43.0	43.7	43.7	45.1	36.1
	c	38.3	35.4	42.6	38.3	45.4	34.0	44.6	38.6	41.9	44.0	54.3
8	s	48.0	47.9	49.9	48.2	40.6	49.5	46.4	47.7	48.9	47.0	37.1
	c	40.1	38.6	38.1	39.8	46.0	39.4	41.6	41.4	39.2	39.5	52.2

Tasks: Is there evidence of differences in spread between the silt and clay contents of soil in this region of Spain? Perform an analysis to see for which sites there is strong evidence of the average contents of silt and clay being different.

2 Swimming ability

Twenty children who enrolled in a swimming class for the handicapped were randomly assigned to one of two groups. Group A, the control group, were taught by traditional methods while group B, the experimental group, were taught by the new method. At the end of the course, each child's swimming ability was assessed by a judge who was unaware of the method by which the child was taught. The results were as follows:

Group A	66	86	80	78	77	63	62	87	75	84
Group B	95	85	56	46	91	79	45	41	54	94

Task: Is there any evidence to suggest that one method produces more erratic results than the other?

Program 6.1 Siegel–Tukey two-sample test for spread

```
100 REM *** SIEGEL-TUKEY TEST FOR SPREAD DIFFERENCES ***

110 REM ** INPUT DATA **
120 INPUT "SAMPLE SIZE FOR A: ";NA
130 INPUT "SAMPLE SIZE FOR B: ";NB
140 LET N=NA+NB
150 DIM D(N),RK(N),B(NB),F(N)
160 PRINT "INPUT THE FIRST SAMPLE:"
170 FOR I=1 TO NA
180 INPUT D(I)
190 LET F(I)=1
200 NEXT I
210 PRINT "INPUT THE SECOND SAMPLE:"
220 FOR I=1 TO NB
230 INPUT B(I)
240 LET F(I)=2
250 NEXT I

260 REM ** MERGE SAMPLES  AND SORT **
270 FOR I=1 TO NB
280 LET D(NA+I)=B(I)
290 NEXT I
300 GO SUB 1000

310 REM ** ASSIGNMENT OF THE RANKS **
320 LET K=INT(N/2)
330 LET R=1
340 LET L=2
350 FOR I=1 TO K STEP 2
360 LET RK(I)=R
370 LET RK(N-I+1)=L
375 IF I=K THEN GO TO 430
380 LET RK(I+1)=R+3
390 LET RK(N-I)=L+1
400 LET R=R+4
410 LET L=L+4
420 NEXT I
430 IF INT(N/2)*2-N<>0 THEN LET RK((N+1)/2)=N

440 REM ** ADJUSTMENT FOR TIES **
450 LET I=1
460 LET B=I
470 LET S=I
480 IF D(I)<>D(I+1) THEN GO TO 530
490 LET S=S+1
500 LET I=I+1
510 IF I=N THEN GO TO 530
520 GO TO 480
530 LET RS=0
540 LET NT=S-B+1
550 FOR J=B TO S
560 LET RS=RK(J)+RS
570 NEXT J
580 FOR J=B TO S
590 LET D(J)=RS/NT
600 NEXT J
610 LET I=I+1
```

Continued overleaf

Program 6.1 Continued

```
615 IF NT=1 AND I=N THEN LET A(N,1)=2
620 IF I>=N THEN GO TO 640
630 GO TO 460

640 REM ** CALCULATION OF THE RANK SUMS **
650 LET RX=0
660 LET RY=0
670 FOR I=1 TO N
680 IF F(I)=1 THEN LET RX=RX+D(I)
690 IF F(I)=2 THEN LET RY=RY+D(I)
700 NEXT I
710 PRINT
720 PRINT "RANK-SUM FOR THE FIRST SAMPLE = ";RX
730 PRINT "RANK-SUM FOR THE SECOND SAMPLE = ";RY
740 STOP

750 REM ** FOR APPROXIMATE CRITICAL VALUES SEE LINES 370-410 OF PROGRAM
 5.1 **

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

7

Two independent samples – tests for general differences

Introduction

In the previous two chapters, we have presented tests designed to detect differences in location and spread respectively. It has been assumed (explicitly or implicitly) that if the two populations being sampled *are* different then they differ solely, or at least primarily, in the specified manner, i.e. either in location or in spread but *not* in both, and *not* in any other manner, such as in the shape of distributions. In general, all the tests in Chapters 5 and 6 suffer a reduction in power when used in these latter cases.

But it is quite common for populations to differ in more general ways. Therefore, we need tests that can cope with more general alternative hypotheses than we have so far used. In particular, it is very common for the *spread* (variability) in a population to be directly related to the location (the median, say) of the population.

Consider, for example, two production processes, A and B, from which samples are taken periodically to check for defective items. In order to make the point, assume for the moment that process A is much better than process B so that, for instance, the average number of defectives found in samples from process A is about 2 and in samples from process B is about 10. Under some reasonable assumptions it turns out that the numbers of defectives from process B might typically range between about 3 and 18, with less than a 1% chance of being outside this range. The corresponding range from process A would be from 0 to about 6. (The source of these details is explained in the 'Discussion' section for Rosenbaum's test below.) *Automatically*, the difference in average numbers for defectives implies a difference in variability or spread of the numbers of defectives. Now admittedly, with such a huge difference between the processes, the tests of Chapter 5 would still be highly likely to give significant results. But note the effect of a change in spread. If the spread in the results from process B around the average of 10 had been the same as that from process A about its average of 2, the range of results from process B would have been from about 8 to 14 (rather than the much wider 3 to 18 previously stated). Then, if this were so, *all* the results from process B would probably have exceeded *all* the results from

process A, giving rise to the most extreme letter sequence:

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

However, because of B's greater spread, there is instead almost bound to be a fair amount of overlap between the 'A's and 'B's, so that the level of significance will be nowhere near as good as if B's spread had stayed as small as A's. If the processes differ, but not by as much as in the discussion so far, the result may well be to deny a rejection of H_0 at all. In other words, the power of the Mann–Whitney and Tukey tests would have been reduced because of the change in spread. It is also worth mentioning that the tests of Chapter 6 have virtually *no* power at all in such situations. In spite of the spread of process B's results being well over double those of process A, which in ordinary circumstances would make the Siegel–Tukey and Tukey-type tests of Chapter 6 very likely to reject H_0, the difference in *averages* renders these tests virtually powerless – they will be almost completely unable to reject H_0 in favour of a difference in spread.

If there were also other differences between the populations, this could exacerbate the problem still further. We therefore present two tests in this chapter which are effective in more general situations. We reverse the order used previously and present a quick-and-easy test first. This test, due to Rosenbaum (1965), is very effective in situations where an increased average implies increased spread, as in the above illustration. But if differences may be of a yet more general nature, then use of the second test to be described, the Kolmogorov–Smirnov test, is recommended instead.

ROSENBAUM'S TEST

Introduction

Let us think more carefully about the effect of the letter sequence when an increase in average is accompanied by an increase in spread. The story can be told via three simple sketches. In Figure 7.1 the horizontal lines indicate the 'effective ranges' of three pairs of populations, which might be defined, say, as ranges in which 99% of the observations in the population lie. The medians of the pair of populations are unchanged in the three cases, but the variations do change.

In case (a), both populations have relatively small spread, in case (b) both have larger spread, and in case (c) A has the smaller spread and B the larger spread. What are we likely to see in the letter sequence for the three cases? In case (a) we get the usual effect described in Chapter 5, i.e. mostly 'A's at the left-hand end and mostly 'B's at the right, so that

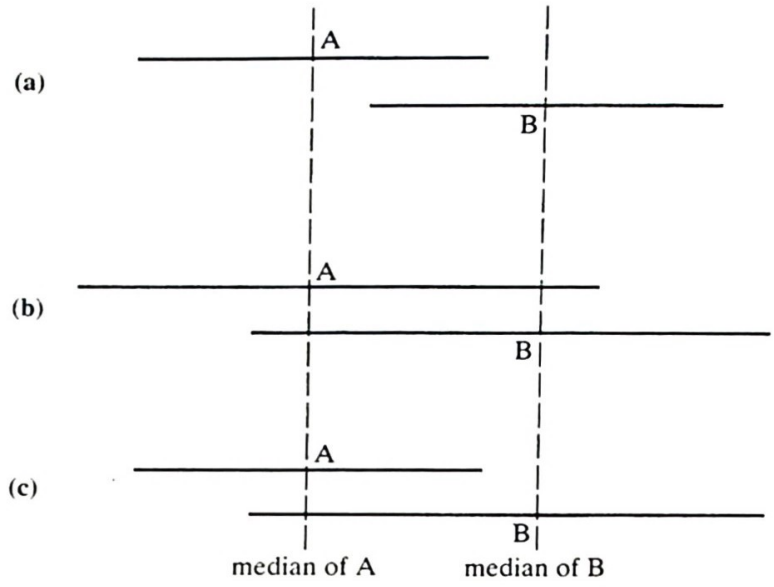


Figure 7.1

both the Mann–Whitney and Tukey tests will perform well. The situation is similar in case (b), though the larger spreads reduce the effect somewhat, so that the tests will not be quite so powerful. However, case (c) is rather different; there is less tendency for 'A's to predominate at the left-hand end than for either of the first two cases. But, at the other end, 'B's predominance is *much greater* than in either of the other two cases.

So, bearing in mind that Rosenbaum's test is of the quick-and-easy type, and is based on the length of extreme runs in the letter sequence, what do you think the test statistic will be?

PRACTICAL PROBLEM

A well known television commercial demonstrates the superiority of a certain make (A) of long-life batteries over ordinary batteries by showing how much longer mechanical toys powered by the long-life batteries keep working compared with those powered by conventional batteries. Another company (B) also starts manufacturing long-life batteries. A consumer organization decides to compare the two types of long-life batteries by carrying out a similar test. Its researchers purchase 20 of the toys and equip 10 of them with A's long-life batteries and the other 10 with B's. The toys are set running and their total operating times are recorded.

During the experiment, it is observed that six of the toys, two in the A sample and four in the B sample, operate *much* longer than the other toys. On subsequent examination it is found that these six toys, though externally similar to the others, are differently designed and in particular

are lighter, thus causing less strain on the batteries. It was therefore decided to eliminate these six readings from the test.

HYPOTHESES

The researchers are examining whether there are any systematic differences in performance between the two types of batteries. Their prime interest is to check for a difference in average running time of the toys powered by the two types. However, it is common for the spread of operating times and 'times to failure', etc., to depend upon average times, so they fully expect that if one type of battery is superior, then this may be accompanied by a wider spread of operating times. The researchers have no idea as to which type may be the better, so the test must be two-sided with hypotheses of the form:

H_0: Batteries A and B perform equally on average.
H_1: Batteries A and B do not perform equally on average.

DATA

The 14 observed operating times retained for analysis were as follows:

Batteries A		
110 h 35 min	120 h 25 min	136 h 20 min
129 h 05 min	139 h 10 min	99 h 25 min

Batteries B			
100 h 55 min	116 h 00 min	119 h 35 min	94 h 40 min
105 h 10 min	114 h 40 min	102 h 55 min	110 h 40 min

STATISTIC

Recall the discussion on case (c) in the 'Introduction'. When the spread is related to the average, it seems sensible that the test statistic should concentrate on the *right*-hand end of the letter sequence and ignore the left-hand end. So Rosenbaum's test statistic R is therefore simply the length of the extreme run (either of 'A's or 'B's) at the right-hand end of the letter sequence.

As with the previous quick tests, it is of course unnecessary to derive the complete letter sequence. All that is needed is the following procedure. Identify the largest observation in each sample. Then if the largest overall observation is from A, R is the number of 'A's greater than the largest B. Or if the largest overall observation is from B, then R is the number of 'B's greater than the largest A.

CRITICAL REGIONS

Since large values of R support H_1, critical regions are of the form

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

Critical values for the usual significance levels are given in Table J. As in the Tukey-type spread test in Chapter 6, it is important to enter the tables with n_1 representing the size of the sample from which R is calculated, with n_2 representing the other sample size.

In the above example, R is in fact calculated from the 'A's (see the 'Solution' section below), so with $n_1=6$ and $n_2=8$ the critical regions are $R\geqslant 4$ for $\alpha_2=10\%$ or 5% and $R\geqslant 5$ for $\alpha_2=2\%$ or 1%.

SOLUTION

The largest running time is the 139 h 10 min from sample A; so R is the number of 'A's greater than the largest B, which is 119 h 35 min. All we now do is to count the number of times in sample A that exceed 119 h 35 min; there are four (120 h 25 min, 136 h 20 min, 129 h 05 min and 139 h 10 min). Hence $R=4$ and this result is significant at the $\alpha_2=5\%$ significance level. The researchers have therefore found reasonable evidence of a difference between average performances of the two makes of batteries.

Discussion

If you care to form the letter sequence for the above data and carry out the test of the previous two chapters, you will find the Mann–Whitney U to be equal to 11, Tukey's Ty to be 5, the Siegel–Tukey U to be 13 and Ts to be 0. *None* of these four results are significant, even at the 10% level. Rosenbaum's test is therefore seen to be very effective when spread increases with average. It also has some power when the spread of a population does not depend on the average, but in this case the tests of Chapter 6 are to be preferred.

If you are familiar with Poisson and exponential distributions, you will know that these are important examples where the standard deviation is very dependent on the mean of the distribution. In fact, in the exponential distribution the standard deviation is *equal* to the mean, and in the Poisson distribution the standard deviation is the square root of the mean. Data likely to come from these distributions are therefore very appropriate for analysis by Rosenbaum's test.

Poisson distributions are well known to model counts accurately in many kinds of random processes (e.g. numbers of shoppers arriving at check-out points) and to approximate many binomial distributions. The illustration in the 'Introduction' concerning numbers of defectives is a case in point, and you may check the statements concerning the likely ranges of these numbers with the details of the Poisson distributions having means equal to 2 and 10 with *EST* pages 14 and 15. Exponential distributions, and other associated distributions such as gamma and Erlang distributions, are often used to model lifetimes, times to failure, times

between arrivals of shoppers, etc. Thus, situations for which Rosenbaum's test is relevant are very common.

It is interesting to note that Rosenbaum first presented this test as a test for differences in spread! Of course, it does have some power when differences between populations are purely in spread, but it certainly seems much more appropriate for the context in which it is presented here.

One-sided and two-sided tests

There are no prizes for guessing what the one-sided version of Rosenbaum's test consists of! If H_1 specifies the population that is to have the larger average (and spread), rejection of H_0 is only considered if the corresponding sample contains the overall maximum – so that R is computed from observations in that sample. In this case, α_1 significance levels should be used.

Thus, for example, if the researchers in our practical problem had *suspected* that A was superior to B (perhaps because it is the more expensive of the two), then the hypotheses would have read:

H_0: Batteries A and B have equal performance on the average.
H_1: On the average, battery A has superior performance to battery B.

Then if the overall maximum is from A, we proceed to calculate R as the number of 'A's greater than the largest B. But if the overall maximum is from B, the value of R is taken as 0 and we are unable to reject H_0 in favour of this H_1.

Treatment of ties

In the unlikely event that both samples have the same maximum value, one would not normally consider rejecting H_0. This is perhaps not universally true, but the only exceptions would be when there is a huge disparity between sample sizes, and this should only rarely apply in practice. Otherwise, only a tie that occurs at the lower end of the extreme run need be considered, and this would be dealt with as in the Tukey tests by considering all orderings of those tied observations and averaging the resulting values of R.

As a rather unlikely example, suppose that the right-hand end of the letter sequence appears thus:

$$\ldots \boxed{\text{A A B B B}}\ \text{A A A}$$

with the box, as usual, denoting tied observations. Considering all 10 possible permutations of the tied observations, six give $R=3$, three give $R=4$ and just one gives $R=5$. Thus we would calculate R as $(6\times3+3\times4+1\times5)/10=3.5$.

Null distributions (including large samples)

The theory of Rosenbaum's R statistic is extremely easy, for any size samples. If the sample sizes are n_A and n_B, with $N=n_A+n_B$, the probability that R is formed from 'A's and takes a value of at least h is the probability of there being (at least) h 'A's at the right-hand end of the letter sequence, and this is

$$\frac{n_A}{N}\times\frac{n_A-1}{N-1}\times\ \ldots\ \times\frac{n_A-h+1}{N-h+1}=\frac{n_A!(N-h)!}{N!(n_A-h)!}$$

Asymptotically, if n_A, $n_B \to \infty$ with $n_A/N \to p$, this expression simply approaches p^h. And this simple expression is extremely convenient to use if we need to derive critical values for sample sizes not included in Table J; in the large majority of cases, it will produce the 'correct' best conservative value.

As an example, suppose $n_A=20$ and $n_B=60$ and the overall maximum observation is an A. The following table compares the exact values of Prob($R \geq h$) with the asymptotic approximation (which is just 4^{-h}):

h	1	2	3	4
Exact probability	0.2500	0.0601	0.0139	0.00306
Asymptotic probability	0.2500	0.0625	0.0156	0.00391

Note that, apart from $h=1$, the exact probabilities are all less than the asymptotic probabilities. This is always the case, and so use of the p^h approximation will give critical values erring, if at all, on the conservative side.

THE KOLMOGOROV–SMIRNOV TEST

Introduction

The Kolmogorov–Smirnov test is the best known of several distribution-free procedures which compare two sample cdfs in order to test for differences of *any kind* between the distributions of the populations from which the samples have been taken.

As we saw in Chapter 1, the sample cdf is an approximation of the true cdf of the corresponding population – though, admittedly, a rather crude one if the sample size is small. This fact has already been used in the Kolmogorov–Smirnov goodness-of-fit test in Chapter 4. Any kind of substantial difference between the two distributions will result in a corresponding difference between their cdfs, which should in turn show up as a significantly large difference between the sample cdfs. Such differences may be in location and/or spread or may be more general differences in shape of the distributions.

Various ways of comparing two sample cdfs have been suggested, and the Kolmogorov–Smirnov method is almost certainly the simplest – it uses the *maximum vertical difference* between them as the test statistic (compare the Kolmogorov–Smirnov goodness-of-fit statistic in Chapter 4). The goodness-of-fit test was introduced by A. N. Kolmogorov (1933), and this particular two-sample application was devised by N. V. Smirnov (1939).

PRACTICAL PROBLEM, HYPOTHESES AND DATA

We shall illustrate the Kolmogorov–Smirnov test with the comparison of the two makes of long-life batteries. As is often the case with applications

of this test, the prime interest is probably to see if there is a difference in average. There may be a subsidiary interest in other types of difference and/or the realization that other types of difference may reduce the effectiveness of pure location difference tests such as those in Chapter 5. However, it will be of interest also to see how this test deals with the practical problems studied in Chapters 5 and 6, and we shall do this in the 'Discussion' section below.

STATISTIC

As stated in the introduction, the Kolmogorov–Smirnov test statistic is simply the maximum vertical difference between the two sample cdfs. Recall that the cdf for a sample of size n is the step function that rises by $1/n$ at each observation, always starting at 0 and finishing up at 1. The cdf for sample A thus rises in steps of 1/6 and that of sample B in steps of 1/8. The two cdfs are given below:

Sample A			
99 h 25 min	110 h 35 min	120 h 25 min	
1/6	2/6	3/6	
129 h 05 min	136 h 20 min	139 h 10 min	
4/6	5/6	6/6=1	
Sample B			
94 h 40 min	100 h 55 min	102 h 55 min	105 h 10 min
1/8	2/8	3/8	4/8
110 h 40 min	114 h 40 min	116 h 00 min	119 h 35 min
5/8	6/8	7/8	8/8=1

In Figure 7.2, we have drawn the two sample cdfs and have indicated the maximum difference D; it turns out to be equal to 2/3. Actually there is no real need to draw these cdfs. Since D depends only on the order of the 'A's and 'B's, and not on their exact values, it can be computed directly from the letter sequence. Study Figure 7.2 and imagine changing some of the values slightly, but not enough to alter the letter sequence; then you should soon be able to convince yourself that this is true. So, having written down the letter sequence, one could write in the values of the cdfs of samples A and B, respectively, above and below the sequence as follows:

Sample A cdf	0	1/6	1/6	1/6	1/6	2/6	2/6	2/6	2/6	2/6	3/6	4/6	5/6	1
	B	A	B	B	B	A	B	B	B	B	A	A	A	A
Sample B cdf	1/8	1/8	2/8	3/8	4/8	4/8	5/8	6/8	7/8	1	1	1	1	1

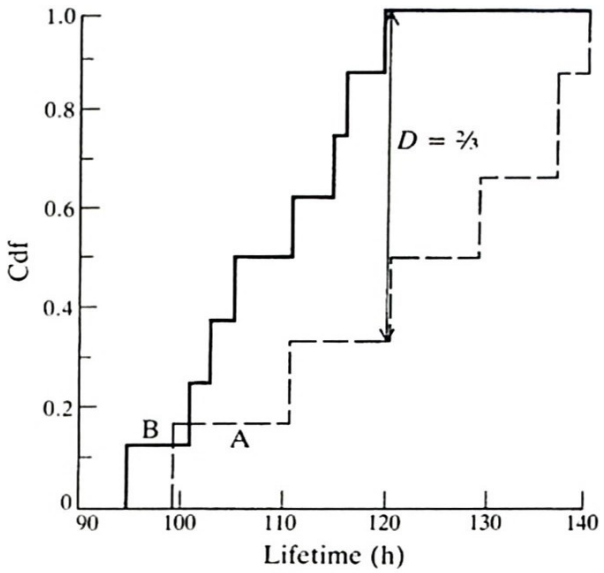


Figure 7.2

Then we compare the two rows of fractions to locate the maximum difference between them, which gives $D = 1-2/6 = 2/3$, as before.

It can sometimes be a little tricky to pinpoint the maximum difference in this latter method when the sample sizes (n_A and n_B), i.e. the denominators of the fractions, are different. But note that, if we multiply all the fractions by $n_A n_B$, the effect is to change them all into integers; and obviously it is a lot easier to find the maximum difference between two rows of integers rather than between two rows of fractions. In practice, the production of the two rows of integers is more easily described as follows. Above the letter sequence, increase the 'score' by n_B (=8 here) at each occurrence of an A; whereas below the letter sequence increase the 'score' by n_A (=6 here) at each occurrence of a B. To check the accuracy, make sure that the final 'score' in both rows is $n_A n_B$ (=6×8 = 48 here):

48 × cdf of sample A	0	8	8	8	8	16	16	16	16	16	24	32	40	48
	B	A	B	B	B	A	B	B	B	B	A	A	A	A
48 × cdf of sample B	6	6	12	18	24	24	30	36	42	48	48	48	48	48

The maximum difference, say D^*, between these rows of integers is now easily located where marked, with $D^*=48-16=32$. Clearly, $D^*=n_A n_B D$,

so we can then evaluate D as $D^*/n_A n_B = 32/48 = 2/3$.

However, in practice, because of the form of the tables of critical values, it is not even necessary to convert D^* into D in order to carry out the test.

CRITICAL REGIONS

Because $D^* = n_A n_B D$ must take integer values only, it is simpler to print tables of critical values directly in terms of D^*, as this avoids the messy fractions or decimals that would be involved in tables for D. Thus, if you have evaluated D by the graphical method or by the fractions calculation using the letter sequence, you will need to multiply D by the product of the sample sizes, $n_A n_B$, to obtain D^* before using the tables. Clearly, large values of D^* will cause H_0 to be rejected and so the critical regions are of the form

$$D^* \geqslant \text{critical value}$$

Critical values for D^* are given in Table K (and on page 31 of *EST*). With $n_1=6$ and $n_2=8$ the $\alpha_2=10\%$, 5%, 2% and 1% critical regions are, respectively, $D^*\geqslant 30$, $D^*\geqslant 34$, $D^*\geqslant 40$ and $D^*\geqslant 40$ (corresponding to $D\geqslant 5/8$, $D\geqslant 17/24$, $D\geqslant 5/6$ and $D\geqslant 5/6$).

SOLUTION

We have already calculated the value of D^* (and D) above and obtained $D^*=32$, which we see is significant at a level somewhere between 5% and 10%.

Discussion

Despite its more sophisticated nature, the Kolmogorov–Smirnov test has not achieved quite as good a significance level as Rosenbaum's quick test on these data. This should not be too much of a surprise, bearing in mind the design and purpose of the two tests – and indeed this is an opportune place to survey all the tests of the last three chapters.

The tests of Chapter 5 are appropriate when the difference between the populations is wholly or mainly in location. The tests in Chapter 6 are appropriate when the difference is wholly or mainly in spread. Rosenbaum's test is designed for cases where an upward move in location is accompanied by an increase in spread. Although there is some overlap in sensitivity between these various tests, by and large none of them work well except in the specific situations for which they were designed. The Kolmogorov–Smirnov test is a 'catch-all' test. It has some power in all of the three situations for which the other tests are relevant, and also has power for other kinds of differences. However, it is in the nature of things that such a general-purpose test could not be expected to have such high power in a *specific* situation as other tests that are designed especially for that situation. This is not an inviolable rule, but it is certainly true more often than not. (If it were not, then presumably we would find ourselves using the

Kolmogorov–Smirnov test universally in preference to the other tests.) So the real place of the Kolmogorov–Smirnov test is in situations where it is perhaps unclear what kind of differences to expect between the populations, or where expected differences do not fit into the usual categories of location or spread.

You might care to try out the Kolmogorov–Smirnov test on the problems in Chapters 5 and 6. For the Japanese vs. British car breakdowns data, $D^*=76$ with $n_1=8$ and $n_2=12$; this is extremely significant of a difference between the two, as would be expected with these data. Using the data in Chapter 6 we obtain $D^*=48$ with $n_1=12$ and $n_2=15$; this is not a significant result. Again this is what we might have anticipated since even the two tests of Chapter 6 found only very slight evidence of a difference in spread (which in this case was the kind of difference most expected).

It is worth mentioning that D^* (or D) is much easier to compute if the samples happen to be equal-sized, say $n_A=n_B=n$. For then both sample cdfs rise in steps of $1/n$ and so, even using the fractional calculation above, it is simple to spot where the maximum difference D occurs. However, it is even simpler to take advantage of the common step size by working in steps of 1 (instead of $1/n$ or n) and then finally dividing the maximum difference thus found by n to obtain D or multiplying by n to obtain D^*. Indeed, having derived the letter sequence, the statistic can easily be calculated 'in the head' in this case. To illustrate this, let us ignore the final two operating times in sample B, so that we then have two samples of size 6. The resulting letter sequence, along with the calculation using a step size of 1 is as follows:

0	1	1	1	2	2	2	2	3	4	5	6
B	A	B	B	A	B	B	B	A	A	A	A
1	1	2	3	3	4	5	6	6	6	6	6

It can be immediately seen that the maximum difference here is $6-2=4$, so that $D=4/6=2/3$ (*coincidentally*, exactly the same as with the original data) and $D^*=24$ (which, with sample sizes both equal to 6, is not significant at the 10% level). The quick form of calculation, which can be carried out without writing anything down except for the letter sequence, is to trace through the letter sequence, keeping a mental tally of, say, 'number of 'B's minus number of 'A's', and noting the maximum value of this tally, irrespective of sign. We write it down here to demonstrate the idea, but you will see that this is indeed an operation which can easily be done in your head:

	B	A	B	B	A	B	B	B	A	A	A	A
Number of 'B's − number of 'A's	+1	0	+1	+2	+1	+2	+3	+4	+3	+2	+1	0

The maximum size of this tally is 4, which can then be simply converted to D or D^* as above.

Finally, apart from those tests (mentioned in the introduction) that compare sample cdfs by other criteria, there are still further tests designed for recognizing general rather than specific differences. Perhaps the best-known of these is the Wald–Wolfowitz number-of-runs test, and this test is described in Chapter 17.

One-sided and two-sided tests

Seeing that the Kolmogorov–Smirnov test is deliberately designed to have power for diagnosing differences of all types, it may seem a contradiction in terms to refer to a one-sided Kolmogorov–Smirnov test. However, sometimes one does use an alternative hypothesis to the effect that, say, 'population A is *stochastically greater* than population B'. This is a generalization of the concept of A being located above B, or of the median of A being greater than the median of B – there being now no implication that the distributions are the same except for a location difference. If this alternative hypothesis is true, the effect is that the true cdf of B will rise earlier than that of A, and of course the hope is that this phenomenon will be reflected in the sample cdfs. Therefore the appropriate statistic is now D_{BA}, defined as the maximum non-negative value of $F_B - F_A$, where F_A and F_B are the cdfs of samples A and B. Equivalently, this is the maximum vertical difference between the graphs over that section (if any) where $F_B > F_A$. The value of D_{BA} needs multiplying by $n_A n_B$ to compare with the tabulated values or, as above, it is easiest in practice to work with an integer calculation which produces $n_A n_B D_{BA}$ directly. For this one-sided version of the statistic, α_1 significance levels are appropriate as usual. Such a one-sided test would have been appropriate when analysing the data of Chapter 5; there, in fact, we would have based the test on D_{AB}, the maximum positive value of $F_A - F_B$. You should calculate the value of D_{AB} and compare your conclusion with that in Chapter 5.

Treatment of ties

Tied observations may only cause problems with the Kolmogorov–Smirnov test if they occur in the vicinity of the maximum difference. There are two methods for dealing with ties. One is very easy but has the disadvantage of possibly giving the statistic a slightly lower value than it should have; however, the error is not usually of a size to make any substantial difference to the conclusion. Graphically, the method amounts simply to *ignoring* the points where ties occur for the purposes of locating the maximum difference. Arithmetically, the method amounts to considering differences only at the *end* of ties.

To illustrate the idea, suppose that the operating times of the toys had only been measured to the nearest hour. The data would then have been:

A	111	120	136	129	139	99		
B	101	116	120	95	105	115	103	111

The corresponding letter sequence is:

B A B B B [A B] B B [A B] A A A,

where boxed letters indicate ties. Using the integer calculation for D^* we proceed thus:

0	8	8	8	8	16	16	16	24	32	40	48	
B	A	B	B	B	[A B]	B	B	[A B]	A	A	A	
6	6	12	18	24	30	36	42	48	48	48	48	