Chapter 10.
Rank Score Tests

Population distributions are characterized, or defined, by parameters such as the mean and standard deviation. For skew distributions we would need to know other parameters such as the degree of skewness before the distribution could be identified uniquely, but the mean and standard deviation identify the Normal distribution uniquely. The *t* test described earlier depends for its validity on an assumption that the data originate from a Normally distributed population, and, when two groups are compared, the difference between the two samples arises simply because they differ only in their mean value. However, if we were concerned that the data did not originate from a Normally distributed population, then there are tests available which do not make use of this assumption. Because the data are no longer Normally distributed, the distribution cannot be characterized by a few parameters, and so the tests are often called "non-parametric". This is somewhat of a misnomer because, as we shall see, to be able to say anything useful about the population we must compare parameters. As was mentioned in Chapter 5, if the sample sizes in both groups are large lack of Normality is of less concern, and the large sample tests described in that chapter would apply.

Wilcoxon signed rank sum test

Wilcoxon and Mann and Whitney described rank sum tests, which have been shown to be the same. Convention has now ascribed the Wilcoxon test to paired data and the Mann-Whitney U test to unpaired data.

Boogert *et. al.*[1] (data also given in Shott[2] used ultrasound to record fetal movements before and after chorionic villus sampling. The percentage of time the fetus spent moving is given in Table 10.1 for ten pregnant women.

Table 10.1 Wilcoxon test on fetal movement before and after Chorionic Villus Sampling[1, 2]

Patient no (1)	Before (2)	After (3)	Difference (4)	Rank (5)	Signed (6)
1	25	18	7	9	9
2	24	27	-3	5.5	-5.5
3	28	25	3	5.5	5.5
4	15	20	-5	8	-8

5	20	17	3	5.5	5.5
6	23	24	-1	1.5	-1.5
7	21	24	-3	5.5	-5.5
8	20	22	-2	3	-3
9	20	19	1	1.5	1.5
10	27	19	8	10	10

If we are concerned that the differences in percentage of time spent moving are unlikely to be Normally distributed we could use the Wilcoxon signed rank test using the following assumptions:

1. The paired differences are independent.
2. The differences come from a symmetrical distribution.

We do not need to perform a test to ensure that the differences come from a symmetrical distribution: an "eyeball" test will suffice. A plot of the differences in column (4) of Table 10.1 is given in Figure 10.1. and shows that distribution of the differences is plausibly symmetrical. The differences are then ranked in column 5 (negative values are ignored and zero values omitted). When two or more differences are identical each is allotted the point half way between the ranks they would fill if distinct, irrespective of the plus or minus sign. For instance, the differences of -1 (patient 6) and +1 (patient 9) fill ranks 1 and 2. As (1 + 2)/2 = 1.5, they are allotted rank 1.5. In column (6) the ranks are repeated for column (5), but to each is attached the sign of the difference from column (4). A useful check is that the sum of the ranks must add to n(n + 1)/2. In this case 10(10 + 1)/2 = 55.

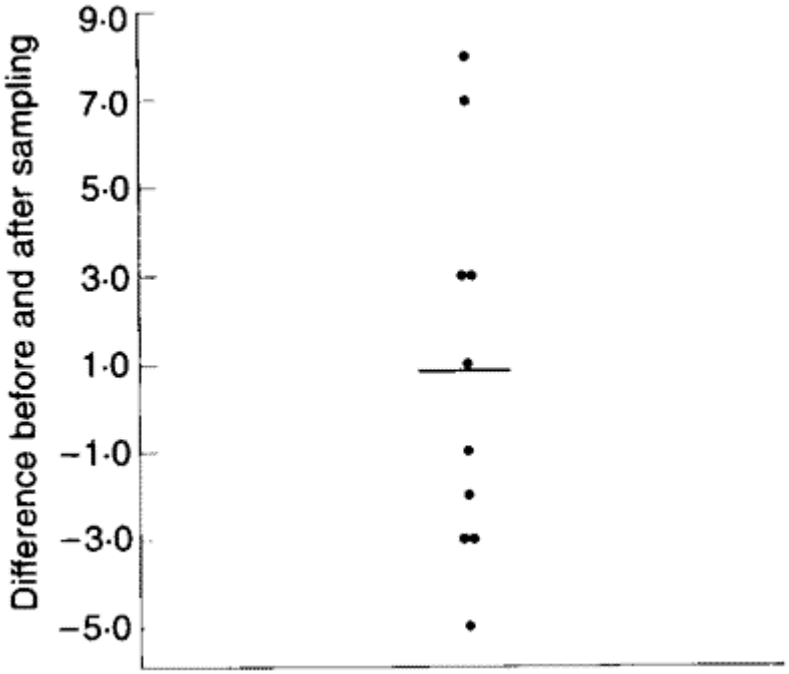

Figure 10.1 Plot of differences in fetal movement with mean value

The numbers representing the positive ranks and the negative ranks in column (6) are added up separately and only the smaller of the two totals is used. Irrespective of its sign, the total is referred to Table D (Appendix) against the number of pairs used in the investigation. Rank totals *larger* than those in the Table are nonsignificant at the level of probability shown. In this case the smaller of the ranks is 23.5. This is larger than the number (8) given for ten pairs in Table D and so the result is not significant. A confidence interval for the interval is described by Campbell and Gardner[2] and Gardner and Altman[4] . and is easily obtained from the programs CIA[5] or MINITAB.[6] The median difference is zero. CIA gives the 95% confidence interval as -2.50 to 4.00. This is quite narrow and so from this small study we can conclude that we have little evidence that chorionic villus sampling alters the movement of the fetus.

Note, perhaps contrary to intuition, that the Wilcoxon test, although a rank test, may give a different value if the data are transformed, say by taking logarithms. Thus it may be worth plotting the distribution of the differences for a number of transformations to see if they make the distribution appear more symmetrical.

Unpaired samples

A senior registrar in the rheumatology clinic of a district hospital has designed a clinical trial of a new drug for rheumatoid arthritis.

Twenty patients were randomized into two groups of ten to receive either the standard therapy A or a new treatment, B. The plasma globulin fractions after treatment are listed in Table 10.2

Table 10.2 Plasma globulin fraction after randomization to treatments A or B

Treatment A	38	26	29	41	36	31	32	30	35	33
Treatment B	45	28	27	38	40	42	39	39	40	45

We wish to test whether the new treatment has changed the plasma globulin, and we are worried about the assumption of Normality.

The first step is to plot the data (see Figure 10.2).

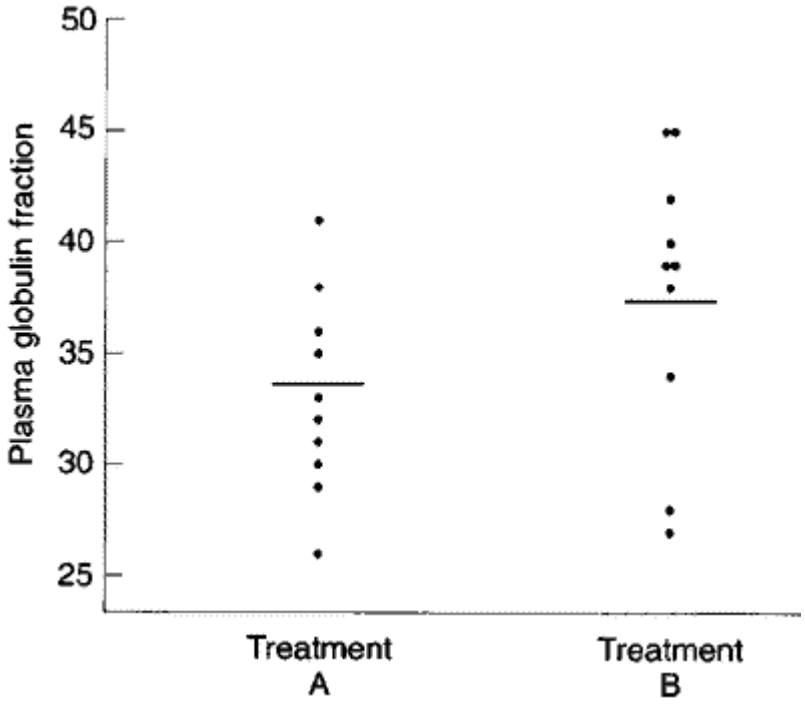

Figure 10.2 Plasma globulin fraction after treatments A or B with mean values.

The clinician was concerned about the lack of Normality of the underlying distribution of the data and so decided to use a nonparametric test. The appropriate test is the Mann-Whitney U test and is computed as follows.

The observations in the two samples are combined into a single series and ranked in order but in the ranking the figures from one sample must be distinguished from those of the other. The data appear as set out in Table 10.3 . To save space they have been set out in two columns, but a single ranking is done. The figures for sample B are set in bold type. Again the sum of the ranks is n(n + 1)/2.

Table 10.3 Combined results of Table 10.2

Globulin fraction	Rank	Globulin fraction	Rank
26	1	36	11
27	2	38	12.5
28	3	38	12.5
29	4	39	14.5
30	5	39	14.5
31	6	40	16
32	7	41	17

33	8	42	18
34	9	45	19.5
35	10	45	19.5

Totals of ranks: sample A, 81.5; sample B, 128.5

The ranks for the two samples are now added separately, and the smaller total is used. It is referred to Table E (Appendix), with n_1 equal to the number of observations in one sample and n *2*equal to the number of observations in the other sample. In this case they both equal 10. At n_1 = 10 and n_2 = 10 the upper part of the Table shows the figure 78. The smaller total of the ranks is 81.5. Since this is slightly larger than 78 it does not reach the 5% level of probability. The result is therefore not significant at that level. In the lower part of , which gives the figures for the 1% level of probability, the figure for n_1 = 10 and n_2 = 10 is 71. As expected, the result is further from that than the 5% figure of 78.

To calculate a meaningful confidence interval we assume that if the two samples come from different populations the distribution of these populations differs only in that one appears shifted to the left or right of the other. This means, for example, that we do not expect one sample to be strongly right skewed and one to be strongly left skewed. If the assumption is reasonable then a confidence interval for the median difference can be calculated[3,4]. Note that the computer program does not calculate the difference in medians, but rather the median of all possible differences between the two samples. This is usually close to the median difference and has theoretical advantages. From CIA we find that the difference in medians is -5.5 and the approximate 95% confidence interval is - 10 to 1.0. As might be expected from the significance test this interval includes zero. Although this result is not significant it would be unwise to conclude that there was no evidence that treatments A and B differed because the confidence interval is quite wide. This suggests that a larger study should be planned.

If the two samples are of unequal size a further calculation is needed after the ranking has been carried out as in Table 10.3 .

Let n_1 = number of patients or objects in the smaller sample and T_1 the total of the ranks for that sample. Let n_2 number of patients or objects in the larger sample. Then calculate T_2 from the following formula:

$$T_2 = (n_1 + n_2 + 1) - T_1$$

Finally enter Table E with the smaller of T_1 or T_2 . As before, only totals smaller than the critical points in are significant. See Exercise 10.2 for an example of this method.

If there are only a few ties, that is if two or more values in the data are equal (say less than 10% of the data) then for sample sizes outside the range of we can calculate

$$z = \frac{|(T_1 - n_1(n_1 + n_2 + 1)/2|}{[[\sqrt{[n_1 n_2(n_1 + n_2 + 1)/12}}$$

On the null hypothesis that the two samples come from the same population, z is approximately Normally distributed, mean zero and standard deviation one, and can be referred to Table A (Appendix) to calculate the P value.

From the data of Table 10.2 we obtain

$$z = \frac{|81.5 - 10 \times 21/2|}{\sqrt{(10 \times 10 \times 21/2)}}$$

and from Table A we find that P is about 0.075, which corroborates the earlier result.

The advantages of these tests based on ranking are that they can be safely used on data that are not at all Normally distributed, that they are quick to carry out, and that no calculator is needed. Non-Normally distributed data can sometimes be transformed by the use of logarithms or some other method to make them Normally distributed, and a *t* test performed. Consequently the best procedure to adopt may require careful thought. The extent and nature of the difference between two samples is often brought out more clearly by standard deviations and *t* tests than by non-parametric tests.

Common questions

Non-parametric tests are valid for both non-Normally distributed data and Normally distributed data, so why not use them all the time?

It would seem prudent to use non-parametric tests in all cases, which would save one the bother of testing for Normality. Parametric tests are preferred, however, for the following reasons:

1. As I have tried to emphasize in this book, we are rarely interested in a significance test alone; we would like to say something about the population from which the samples came, and this is best done with estimates of parameters and confidence intervals.
2. It is difficult to do flexible modeling with non-parametric tests, for example allowing for confounding factors using multiple regression (see Chapter 11).

Do non-parametric tests compare medians?

It is a commonly held belief that a Mann-Whitney U test is in fact a test for differences in

medians. However, two groups could have the same median and yet have a significant Mann-Whitney U test. Consider the following data for two groups, each with 100 observations. Group 1: 98 (0), 1, 2; Group 2: 51 (0), 1, 48 (2). The median in both cases is 0, but from the Mann-Whitney test P<0.000 1.

Only if we are prepared to make the additional assumption that the difference in the two groups is simply a shift in location (that is, the distribution of the data in one group is simply shifted by a fixed amount from the other) can we say that the test is a test of the difference in medians. However, if the groups have the same distribution, then a shift in location will move medians and means by the same amount and so the difference in medians is the same as the difference in means. Thus the Mann-Whitney U test is also a test for the difference in means.

How is the Mann- Whitney U test related to the t test?

If one were to input the ranks of the data rather than the data themselves into a two sample t test program, the P value obtained would be very close to that produced by a Mann-Whitney U test.

References

1. Boogert A, Manhigh A, Visser GHA. The immediate effects of chronic villus sampling on fetal movements. *Am J Obstet Gynecol* 1987; *157* :137-9.
2. Shott S. *Statistics for Health Professionals.* Philadelphia: WB Saunders, 1990.
3. Campbell MJ, Gardner MJ. Calculating confidence intervals for some non-parametric analyses. *BMJ* l988; *296* :l369-7l.
4. Gardner MJ, Altman DG. *Statistics with Confidence . Confidence Intervals and Statistical Guidelines.* London: BMJ Publishing Group, 1989.
5. Gardner MJ, Gardner SB, Winter PD. *CIA (Confidence Interval Analysis)* . London: BMJ Publishing Group, 1989.
6. Ryan BF, Joiner BL, Ryan TA. *Minitab Handbook* , 2nd ed. Boston: Duxbury Press, 1985.

Exercises

Exercise 10.1 A new treatment in the form of tablets for the prophylaxis of migraine has been introduced, to be taken before an impending attack. Twelve patients agree to try this remedy in addition to the usual general measures they take, subject to advice from their doctor on the taking of analgesics also.

A crossover trial with identical placebo tablets is carried out over a period of 8 months. The numbers of attacks experienced by each patient on, first, the new treatment and, secondly, the placebo were as follows: patient (1) 4 and 2; patient (2) 12 and 6; patient (3) 6 and 6; patient (4) 3 and 5; patient (5)15 and 9; patient (6) 10 and 11; patient (7) 2 and 4; patient (8) 5 and 6; patient (9)11 and 3; patient (10) 4 and 7; patient (11) 6 and 0; patient (12) 2 and 5. In a Wilcoxon rank sum test what is the smaller total of ranks? Is it significant at the 5% level?

Exercise 10.2 Another doctor carried out a similar pilot study with this preparation on 12 patients, giving the same placebo to ten other patients. The numbers of migraine attacks experienced by the patients over a period of 6 months were as follows.

Group receiving new preparation: patient (1) 8; (2) 6; (3) 0; (4) 3; (5) 14; (6) 5; (7) 11; (8) 2

Group receiving placebo: patient (9) 7; (10) 10; (11) 4; (12) 11; (13) 2; (14) 8; (15) 8; (16) 6; (17)1; (18) 5.

In a Mann-Whitney two sample test what is the smaller total of ranks? Which sample of patients provides it? Is the difference significant at the 5% level?