8 Some useful non-parametric tests

8.1 Introduction

In this chapter, we examine three non-parametric tests which are useful in situations where the conditions for the parametric z- or t-tests are not met. These are the Mann–Whitney U-test, the Wilcoxon signed-ranks test and the sign test. As we shall see, each has a rather different range of applicability. A further non-parametric test, the chi-square test, is so important in linguistic statistics that the whole of chapter 9 will be devoted to it.

8.2 The Mann-Whitney *U*-test

The Mann–Whitney U-test is useful when we wish to know if two independent sets of data show a significant overall difference in the magnitude of the variable we are interested in, but we cannot use the z- or t-tests because the assumptions relating to level of measurement, sample size, normality or equality of variance are not valid. The test assumes only an ordinal level of measurement, since it is based on the ranking of scores. It is almost as powerful as the t-test, and so is a very useful alternative, especially as the test statistic is easy to calculate.

Under the null hypothesis, the two samples we are comparing come from populations with the same distribution. If we combine the samples and then assign ranks to each of the observations, allocating the rank 1 to the smallest, 2 to the next and so on, we shall expect, under the null hypothesis, that the scores from the two samples will be spread randomly in the rank ordering, so that the sum of ranks for each sample will be similar. If, on the other hand, there is a significant difference between the samples, we shall expect one sample to contribute values mainly to the upper part of the rank list, the other mainly to the lower part. The rank sums for the two samples will thus be considerably different. The Mann–Whitney U-test is used to calculate the probability of finding any given difference in rank sums under the null hypothesis that the samples were drawn from populations with the same distribution. As with other tests, the significance attached to the calculated value of the test statistic depends on the significance level chosen, and on whether the test is directional or non-directional.

We shall now consider how to calculate the test statistic U. Let the number of scores in the smaller group (if the groups are unequal in size) be N_1, and the number in the larger group N_2. Obviously, if the groups are equal in size, $N_1 = N_2$. We now rank the whole combined set of $N_1 + N_2$ scores from lowest (rank 1) to highest. If there is more than one occurrence of the same score (that is, 'tied' ranks), each occurrence is given the mean of the ranks which would have been allocated if there had not been a tie. For example, in the series 1, 2, 3, 3, 3, 5, the three occurrences of the score 3 would have occupied ranks 3, 4 and 5 if there had been no tie, and they are therefore each given the mean of these ranks (that is, 4). Since three ranks have been used, however, the next number (the 5 in our series) will have the rank of 6. We now obtain the sum of ranks (R_1) for the smaller sample. If the samples are of equal size. either may be used. The value of R_1, together with those of N_1 and N_2, is substituted in the following expression for U_1:

$$U_1 = N_1 N_2 + \frac{N_1(N_1 + 1)}{2} - R_1.$$

We can now do the same for the larger group:

$$U_2 = N_1 N_2 + \frac{N_2(N_2 + 1)}{2} - R_2.$$

However, it can be shown, by rather tedious but fairly elementary algebra, that

$$U_2 = N_1 N_2 - U_1.$$

This formula is clearly more convenient for computation. We now take the smaller of U_1 and U_2, and call it U. The distribution of U is known for various values of N_1 and N_2, and table A4 of appendix 1 gives the critical values for the 5 and 1 per cent levels in a non-directional test, or for the 2.5 and 0.5 per cent levels (that is, $p \leqslant 0.025$ and $p \leqslant 0.005$) in a directional test. The critical region contains U values lying *below* the critical value, so that if our calculated value is *smaller* than or equal to the critical value we can reject the null hypothesis.

We shall now consider an example of the application of the Mann–Whitney test. Suppose that we have a group of 17 subjects for an investigation into the coherence of English texts. The subjects are allocated randomly to two groups, one of 9 and the other of 8 people. One group is given a short piece of text to read; the other group is given a different version of the text, in which various aspects of sentence linkage have been changed. The subjects are asked to grade the text they have read on a scale of coherence from 0 (totally incoherent) to 10 (totally coherent). The investigator wishes to know whether there is any significant difference between the two sets of ratings at the 5 per cent level in a non-directional test. The results are as shown in table 8.1.

We have two independent groups here, and an ordinal level of measurement, since we should not want to claim that coherence can be quantified in such a way that ratings of, say, 2 and 3 represent exactly the same amount of difference as ratings of 3 and 4, or 4 and 5, although we can say that a rating of 4 represents

Table 8.1 Coherence ratings for two versions of a text

Original text *($N_2 = 9$)*	*Altered text* *($N_1 = 8$)*
7	7
8	4
6	5
9	6
10	8
7	5
7	5
8	7
8	

greater coherence than one of 3, and so on. The conditions for a *t*-test do not, therefore, apply, but the non-parametric Mann–Whitney test can be used.

We first rank the combined scores, giving an average rank to tied scores, as in table 8.2. The rank sum for the smaller group (R_1) is 47. We can now use this value to find U_1:

$$U_1 = N_1 N_2 + \frac{N_1(N_1+1)}{2} - R_1$$

$$= (8 \times 9) + \frac{8(8+1)}{2} - 47 = 61.$$

Using the simpler formula for U_2, we have

$$U_2 = N_1 N_2 - U_1 = (8 \times 9) - 61 = 72 - 61 = 11.$$

Since U_2 is the smaller of the two values, we have $U = U_2 = 11$. Looking in table A4 for $N_1 = 8$ and $N_2 = 9$, we find that the critical value of U for the 5 per cent level in a non-directional test is 15. Since our calculated value is less than this, we can reject the null hypothesis, and claim that there is a significant difference between the two sets of ratings.

Above values of about 20 for N_1 and N_2, the test statistic U conforms to an approximately normal distribution. For large

Table 8.2 Ranks for combined scores in coherence test

Original text	*Rank*	*Altered text*	*Rank*
7	9	7	9
8	13.5	4	1
6	5.5	5	3
9	16	6	5.5
10	17	8	13.5
7	9	5	3
7	9	5	3
8	13.5	7	9
8	13.5		
			$R_1 = 47$

samples, then, we calculate a z value as follows:

$$z = \frac{U - N_1 N_2/2}{\sqrt{\dfrac{N_1 N_2 (N_1 + N_2 + 1)}{12}}}.$$

If the calculated value of z is greater than or equal to the critical value for the required significance level, as determined from table A2, we may reject the null hypothesis.

Let us suppose that we have calculated U for two samples, one of 30 and one of 35 scores, and have obtained a value of 618. We calculate z as follows:

$$z = \frac{618 - (30 \times 35)/2}{\sqrt{\dfrac{30 \times 35 \times (30 + 35 + 1)}{12}}} = 1.22.$$

If the test is non-directional and at the 5 per cent level, the critical value of z is 1.96. Since the calculated value is smaller than this, we cannot reject the null hypothesis.

8.3 The Wilcoxon signed-ranks test

We saw above that the Mann–Whitney U-test can be used as a non-parametric counterpart of the t-test for independent samples, in conditions where the parametric test is inappropriate. The non-parametric equivalent of the t-test for correlated samples is the Wilcoxon signed-ranks test. This test assumes that we can rank differences between paired observations. Strictly speaking, then, it requires an interval level of measurement, since we need to be able to say that one difference is greater than another, and to calculate differences meaningfully the variable has to be measured in units of some kind. The test is a useful alternative to the t-test, however, since it is almost as powerful, and, being non-parametric, makes no assumptions about the shape of the distribution.

The data for the test will consist of a number of pairs of scores, each derived from a single subject, or from a pair of matched subjects. Under the null hypothesis that there is no difference in the distributions of the populations from which the samples are drawn, we should expect some of the differences between members of pairs to be positive and an approximately equal number to be negative. Furthermore, we should expect the sum of the positive differences to be roughly equal to the sum of the negative differences. If, on the other hand, the two sets of scores are representative of populations with different distributions, we should expect an imbalance in the numbers, and the sums, of positive and negative differences. The Wilcoxon signed-ranks test assesses the significance of the imbalance. The rationale of the test is thus very simple, and similar to that of the Mann–Whitney test for independent samples.

In order to calculate the test statistic, we first find the difference between each pair of scores, subtracting consistently (second from first, or vice versa) and recording the signs. We then rank the absolute values of the differences, ignoring the sign. If two scores in a pair are the same (that is, if the difference is zero), that pair is ignored altogether. If two values of the difference are tied, they are given the mean of the ranks they would have had if they had been different in value (compare the similar procedure in the Mann–Whitney test). Each rank is now given the sign of the difference it corresponds to. The sum of the positive ranks is found, and also that of the negative ranks. The smaller of these two sums is the test statistic W. Table A5 of appendix 1 gives the critical values of W for the 5 and 1 per cent significance levels in a non-directional test, or for the 2.5 and 0.5 per cent levels in a directional test, for up to 25 pairs of scores. The value of W must be *smaller* than or equal to the critical value if the null hypothesis is to be rejected. It should be noted that in taking an appropriate value for the number of pairs of scores N, pairs which are tied, and so have been discarded, are not counted.

As an example of the use of the Wilcoxon signed-ranks test, consider the situation where we tabulate the numbers of errors made by a group of 10 subjects in translating two passages of English, of equal length, into French. We wish to test, at the 5 per cent level, whether there is any significant difference between the two sets of scores. Since we are not predicting the direction of any such difference, a non-directional test will be appropriate. The scores observed are given in table 8.3.

We first calculate the differences, as shown in table 8.4, and rank them, giving a mean rank in the case of ties. We give each rank the sign of the difference it corresponds to. The pair with zero difference is dropped from the analysis, and N, the number

Table 8.3 Errors made in translating two passages into French

Subject no.	*Errors in passage A*	*Errors in passage B*
1	8	10
2	7	6
3	4	4
4	2	5
5	4	7
6	10	11
7	17	15
8	3	6
9	2	3
10	11	14

Table 8.4 Differences and ranks for translation error scores

Passage A	*Passage B*	*A−B*	*Rank*
8	10	−2	−4.5
7	6	+1	+2
4	4	0	–
2	5	−3	−7.5
4	7	−3	−7.5
10	11	−1	−2
17	15	+2	+4.5
3	6	−3	−7.5
2	3	−1	−2
11	14	−3	−7.5

of pairs, is decreased accordingly to 9. The sum of the positive ranks is 6.5, while that of the negative ranks is 38.5. We take the smaller of these, 6.5, as our value for W. From table A5, the critical value of W for $N = 9$ and a 5 per cent significance level in a non-directional test is 5. Since our value of W is larger than this, we cannot reject the null hypothesis, and must conclude that no significant difference has been demonstrated.

Where the number of pairs is greater than about 20, the distribution of W is almost normal, and a z-score can be calculated from the computed W value according to the following expression:

$$z = \frac{W - N(N+1)/4}{\sqrt{\dfrac{N(N+1)(2N+1)}{24}}}.$$

Let us consider the case where a value of $W = 209$ has been calculated for a set of data with $N = 35$. We then have

$$z = \frac{209 - (35 \times 36)/4}{\sqrt{\dfrac{35 \times 36 \times \{(2 \times 35) + 1\}}{24}}} = -1.74.$$

We can ignore the sign, which is normally negative. We know that values of 1.96 and 1.64 are required at the 5 per cent level for a non-directional or a directional test, respectively. Therefore if our test was non-directional we must conclude that no significant difference has been demonstrated. If, on the other hand, we had made a directional prediction, we could claim significance at the 5 per cent level.

8.4 The sign test

As we have seen, the Wilcoxon signed-ranks test requires a fairly high level of measurement. In many kinds of investigation of interest to the linguist, only an ordinal level can be achieved. Consider, for example, the case where informants are asked to rate two sentences on a scale of acceptability, or politeness, or some similar variable. Such variables (like coherence, discussed in relation to our example of the Mann–Whitney test in section 8.2) cannot really be measured in units with equal intervals, and we cannot attach much importance to the magnitude of differences between ratings; we could, however, claim that a sentence rated as, say, 4 on an acceptability scale had been rated as more acceptable than a sentence given a rating of 3. We can take into account the direction of the differences between pairs, even though we cannot use their magnitude. The loss of information incurred in ignoring the magnitude of the differences means that the so-called sign test, based on these principles, is less powerful than the Wilcoxon signed-ranks test, which does take the magnitude into account.

Under the null hypothesis that there is no difference in the distributions of the populations from which the samples are derived, we should expect, as we saw in discussing the Wilcoxon test, that the number of positive differences between the members of pairs of correlated scores would be roughly equal to the number of negative differences. The sign test computes the probability of obtaining any particular degree of deviation from this equality, under the null hypothesis. The situation is similar to that involved in the problem we looked at in section 1.2: calculating the probabilities of various combinations of males and females in a sample from a village with equal numbers of men and women. We saw that the so-called binomial distribution could be used to calculate these probabilities. Similarly, in the case of the sign test, we can compute the probability of obtaining 6 positive and 4 negative differences for 10 pairs, and so on.

To carry out the sign test, we first record the sign of the difference for each pair of scores, subtracting consistently. Tied scores are dropped from the analysis, and the number of pairs (N) reduced accordingly. We now find the number of pairs with the *less* frequent sign, and call it x. Table A6 of appendix 1 gives critical values of x for values of N between 5 and 25, in a directional or non-directional test. If the computed value of x is *smaller* than or equal to the critical value, we may reject the null hypothesis.

As an example of a situation where the sign test would be appropriate, let us consider the case where a group of subjects has been asked to rate a sentence on a scale of acceptability from 0 (totally unacceptable) to 5 (totally acceptable) for (a) informal spoken English and (b) formal written English. The investigator predicts that the sentence will be judged as more acceptable in informal spoken than in formal written English. The scores are given in table 8.5.

We first obtain the sign of each difference, and discount tied scores, as shown in table 8.6. We have 3 negative and 10 positive differences, so that $x = 3$ and $N = 13$. The critical value of x at the 5 per cent level for $N = 13$ in a directional test is 3. Since the calculated value of x is equal to the critical value, we can reject the null hypothesis at the 5 per cent level, and conclude that there

Table 8.5 Acceptability ratings for a sentence in informal spoken and formal written English

Subject no.	*Informal spoken*	*Formal written*
1	5	5
2	4	2
3	5	3
4	4	4
5	3	1
6	2	3
7	4	3
8	5	1
9	4	2
10	2	3
11	4	2
12	4	3
13	5	3
14	3	5
15	3	0

Table 8.6 Sign of differences in acceptability scores

Informal spoken	*Formal written*	*Sign of (informal − formal)*
5	5	0
4	2	+
5	3	+
4	4	0
3	1	+
2	3	−
4	3	+
5	1	+
4	2	+
2	3	−
4	2	+
4	3	+
5	3	+
3	5	−
3	0	+

is a significant difference between the two sets of scores, which is clearly in the predicted direction.

If N is greater than about 25, it can be shown that an expression closely related to x is normally distributed. We can calculate a z value using the following expression, and refer it to the table of areas under the normal curve as usual:

$$z = \frac{N - 2x - 1}{\sqrt{N}}.$$

Let us imagine that we have 100 pairs of (non-tied) scores, and that 42 of the differences are positive and 58 negative. Then $x = 42$, and

$$z = \frac{100 - (2 \times 42) - 1}{\sqrt{100}} = 1.50.$$

This falls below the critical value for either a directional or a non-directional test at the 5 per cent level, so we should have to conclude that no significant difference between the two sets of scores had been demonstrated at this level.

8.5 Deciding which test to use: a summary

We conclude this chapter with a flowchart (figure 8.1) showing the steps to be taken in deciding which of the three tests discussed in this chapter should be used in a given case. For a much more detailed treatment of tests involving ranking, interested readers are referred to Meddis (1984).

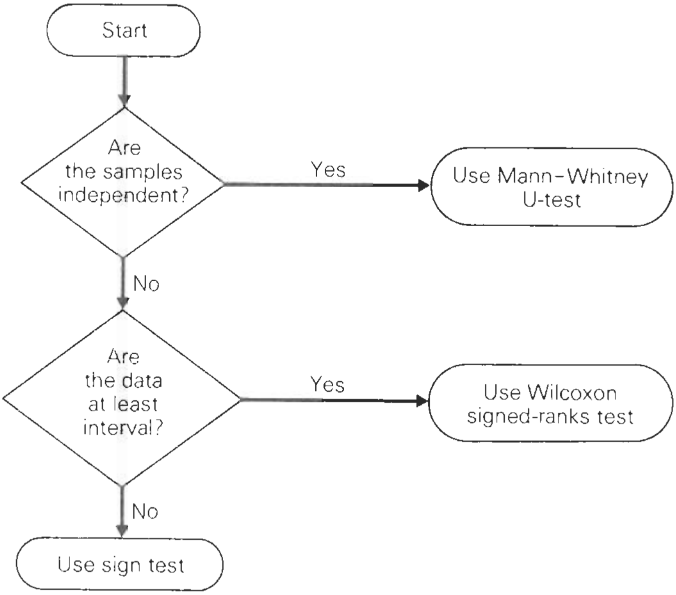
Start
Are the samples independent?
Yes
Use Mann–Whitney U-test
No
Are the data at least interval?
Yes
Use Wilcoxon signed-ranks test
No
Use sign test

Figure 8.1 Deciding which test to use

Exercises

1 The following represent a teacher's assessments of reading skill for two groups of children:

Group 1	*Group 2*
8	4
6	6
3	3
5	3
8	7
7	7
7	5
6	5
5	4
6	4
6	6
7	5
8	6
	5
	4

Test, at the 2.5 per cent significance level, the hypothesis that the first group has a higher level of reading skill than the second.

2 Two sets of ten sentences, differing in their degree of syntactic complexity, are read individually, in randomised order, to a group of 30 informants, who are asked to repeat each sentence after a fixed time interval. The number of correctly remembered sentences in each set, for each informant, is as follows:

Set 1	*Set 2*	*Set 1*	*Set 2*	*Set 1*	*Set 2*
8	5	10	5	8	7
7	7	5	6	7	7
4	9	4	3	6	9
7	3	10	9	9	3
6	4	8	10	4	6
5	5	4	4	3	1
2	0	7	5	8	7
6	4	8	6	7	8
9	7	6	6	5	1
8	6	5	2	6	2

Using a non-parametric test, test the hypothesis that the two sets of scores differ significantly at the 5 per cent level.

3 An experiment is performed to test the effect of certain linguistic features on the politeness of two sentences in a particular social context. Fifteen informants are asked to rate the two sentences on a scale from 1 (very impolite) to 5 (very polite), with the following results:

Informant no.	*Sentence 1*	*Sentence 2*
1	1	3
2	2	2
3	1	4
4	2	3
5	3	1
6	2	4
7	1	1
8	2	3
9	3	5
10	1	3
11	2	3
12	1	4
13	2	1
14	2	4
15	1	3

Test the hypothesis that sentence 2 is rated as more polite than sentence 1, at the 5 per cent level.

4 Two comparable groups of children, whose native language is not English, are taught to read English by two different methods, and their reading fluency is then assessed on a scale from 1 to 10. The scores are as follows:

Method A	*Method B*
6 4 8 3 4 5 5 6	5 5 8 9 7 5 6 7
4 4 7 6 6 2 3	8 3 3 7 8 6 7

Do the two methods give signficantly different results at the 5 per cent level?