7 Parametric and non-parametric tests

Parametric tests

As discussed in Chapter 5, the *t*-test and the variance-ratio test make certain assumptions about the underlying population distributions of the data on which they are used; for example that they are normal. Such tests are often called 'parametric' as these assumptions are about population parameters. (*Parameters* are measures computed from all the observations in a population – examples are the population mean and standard deviation. *Statistics* are measures computed from a sample, in order to estimate parameters.)

Parametric tests are often *robust*, in that they are relatively unaffected by violations of these assumptions (see p. 75). But some situations arise where there are markedly non-normal distributions, or where the data collected are in the form of rankings (first, second, etc.) rather than the scores.

A range of tests, commonly referred to as *non-parametric* (or *rank order*) tests, have been developed which can be used in these situations. They can also be used where it would be appropriate to use parametric tests (while you can't convert a rank into a score, scores can be turned into ranks). In fact some enthusiasts for non-parametric tests urge their virtually universal use, but there are some disadvantages which are discussed below (p. 122).

The Mann-Whitney and Wilcoxon tests

The Mann-Whitney and Wilcoxon tests are two non-parametric tests, which do a very similar job to the independent samples and

correlated samples *t*-tests respectively. While their use means that you can have a more relaxed approach to the type of data to be analysed it is worth stressing that an experiment which is to be analysed using a non-parametric test needs just as careful attention to points of experimental design, randomization, etc., as one to be analysed using a parametric test.

These tests are based on orderings or rankings of the data. Suppose that a person is asked to rank in order of preference eight foods, four of which are savoury and four sweet. If they rank the four sweet foods as first, second, third and fourth, then it would appear likely that they prefer sweet foods to savoury foods. The mathematical basis for this is straightforward. If we had the names of the foods written on cards and shuffled thoroughly then what would be the chances that the four 'sweet' cards would be turned over first? This probability can be calculated quite simply and is pretty low. Randomization of all possible orderings forms the basis of the Mann-Whitney and Wilcoxon tests. The data can be in the form of rankings, as discussed above, or actual measures can be taken which are then converted into ranks.

A non-parametric test does not usually test exactly the same thing as the corresponding parametric test. Effectively, the general procedure is the same as in the food example just considered. We start with the null hypothesis that each of the orders in which the set of eight cards might be turned over is equally likely to occur. If the IV has no effect on the DV, this will be the case. Then, as in other tests, if the results obtained in an actual experiment are highly improbable given that the above hypothesis is true, we come to the decision that the IV does affect the DV. Another way of putting the distinction between what the parametric and non-parametric tests actually test is to say that, whereas the *t*-test tests for a specific difference in the means of the population, the corresponding non-parametric test is a general test of whether or not the populations are the same.

In general, non-parametric tests tend to be less sensitive at detecting an effect of the independent variable on the dependent variable. Another way of saying this is that in situations where both types of test are appropriate, the **power efficiency** of the non-parametric test is lower than its parametric counterpart. To detect

any given effect at a specified significance level, a larger sample size is required for the non-parametric test than the parametric test. This is expressed as

$$\text{power efficiency of test A compared with test B} = \frac{N_B}{N_A} \times 100,$$

where N_A is the sample size needed to show a statistically significant effect at the 5 per cent level for test A, and N_B is the sample size needed to show a statistically significant effect at the 5 per cent level for test B.

Mann-Whitney test

This is the non-parametric counterpart of the independent samples *t*-test for equality of means. Hence, it is appropriate for use with independent-samples designs. It is based on a statistic U which is linked to the sum of the ranks of each of the conditions.

A step-by-step procedure and worked example are given for use with the small-sample case. Table H is used in connection with this and covers situations where the samples have twenty or fewer cases (if the samples are unequal in size, as is possible, of course, with the independent samples design, then the larger of the samples should be twenty or fewer).

Step-by-step procedure

Mann-Whitney test – small-sample* case

For **independent samples designs:** use instead of uncorrelated *t*-test if data is either (a) in the form of ranks or (b) obviously non-normal or (c) there is an obvious difference in the variance of the two groups

Step 1 Rank data (taking both groups together) giving rank 1 to the lowest score, and so on

Step 2 Find the sum of the ranks for the smaller sample – A in the example opposite – (if both samples are the same size, find the sum of ranks of sample A). Call this T

Step 3 Find $U = N_A N_B + \frac{N_A(N_A + 1)}{2} - T$,

where N_A is the number of scores in the smaller sample (or, if both samples are the same size, the sample whose ranks were totalled to find T)

Step 4 Find $U' = N_A N_B - U$

Step 5 Look up the *smaller* of U and U' in Table H. There is a significant difference if the observed value is equal to or less than the table value

Step 6 Translate the result of the test back in terms of the experiment

Treatment of ties

Give the mean rank to the tied observations. Thus, if 2 scores of 10 tie for 5th and 6th ranks, give each score a rank of 5·5. If 3 scores of 18 tie for 12th, 13th and 14th ranks, give each score a rank of 13. The test should not be used if there is a large proportion of ties.

* Not more than 20 observations in either set of scores

Worked example

Mann-Whitney test – small-sample case

Solution times of anagrams under condition A (single category) or condition B (multiple categories) were as follows:

A	*B*
3	23
5	37
97	64
12	24
	14

Step 1

A	*B*
1	5
2	7
9	8
3	6
	4

Step 2 $T = 1 + 2 + 9 + 3 = 15$

Step 3 $N_A = 4, \quad N_B = 5$

$$U = (4 \times 5) + \frac{(4 \times 5)}{2} - 15$$

$$= 15$$

Step 4 $U' = (4 \times 5) - U$

$= 20 - 15$

$= 5$

Step 5 As U' is less than U, look up U' in Table H. Table value for $N_A = 4$, $N_B = 5$ is 1. The observed value (5) is not equal to or less than the table value, therefore there is not significant evidence that the scores under the two conditions differ

Step 6 Anagram solution times under the two conditions (single and multiple categories) do not differ significantly

Mann-Whitney test – large-sample test

The sampling distribution of the statistic U approaches the normal distribution when the sample size becomes large. The procedure is to obtain U as in the small-sample case (steps 1–3 inclusive). The standard deviation of U can then be found as

$$SD_U = \sqrt{\frac{N_A N_B(N_A + N_B + 1)}{12}}$$

and a z-score (see p. 54) as

$$z = \left(U - \frac{N_A N_B}{2}\right) \div SD_U.$$

If we are dealing with a two-tailed test, then the observed z is significant at the 5 per cent level if it exceeds 1·96. For a one-tailed test, 5 per cent significance is attained if z exceeds 1·64 (check these in Table D if you are in doubt).

The ranking procedure can become quite laborious with large samples. Partly for this reason and partly because violations of the assumptions behind parametric statistics become less important for large samples, the Mann-Whitney test tends to be restricted to use with relatively small samples.

The Wilcoxon test

This is the non-parametric counterpart to the correlated samples t-test for equality of means. It is suitable for use with the matched pairs or repeated measures designs. We have already considered a non-parametric test which is appropriate for use in this design (the sign test, p. 35).

The Wilcoxon test is intermediate between the sign test and the correlated samples t-test in the amount of information which is extracted from the data. In the sign test we only take into account the sign of the difference between a pair of scores; in the t-test the actual size of the difference is used in computation. The Wilcoxon test uses the sign of the difference and additionally *orders* the sizes of these differences.

As one might expect, the sign test is low in power efficiency (what is power efficiency? See p. 113), the Wilcoxon test intermediate and the correlated samples *t*-test most efficient. However, there is in fact only a small difference in the power efficiencies of Wilcoxon and correlated samples *t*-tests in situations where either could be used.

The Wilcoxon test is similar both in rationale and in computation to the Mann-Whitney test. It is based on a statistic *T*, derived from the sum of the ranks for the differences in the data pairs in the less frequent direction. The step by step procedure and worked example for small samples (taken as twenty-five pairs or fewer) give full details.

Step-by-step procedure

Wilcoxon test – small-sample* case

For matched pairs or repeated measures designs: use instead of a correlated *t*-test if either (a) the differences between treatments can only be ranked in size or (b) the data is obviously non-normal or (c) there is an obvious difference in the variance of the two groups

Step 1 Obtain the difference between each pair of readings, taking sign into account

Step 2 Rank order these differences (ignoring the sign), giving rank 1 to the smallest difference

Step 3 Obtain *T*, the sum of the ranks for differences with the less frequent sign

Step 4 Consult Table J. If the observed *T* is equal to or less than the table value, then there is a significant difference between the two conditions

Step 4 Translate the result of the test back in terms of the experiment

* Not more than 25 pairs of scores

Worked example

Wilcoxon – small-sample case

Eight pairs of twins were tested in complex reaction time situations; one member of each pair was tested after drinking 3 double whiskies, the other member was completely sober. The following reaction times were recorded:

Sober group	*Whisky group*	**Step 1** *Differences*	**Step 2** *Ranks*
310	300	−10	1
340	320	−20	2
290	360	70	5
270	320	50	4
370	540	170	6
330	360	30	3
320	680	360	7
320	1120	800	8

Step 3 Less frequent sign of difference is negative, $T = 1 + 2 = 3$

Step 4 From Table J, when $N = 8$, $T = 4$. As the observed value of T is less than the table value, there is a significant difference between the two conditions

Step 5 Complex reaction time scores are significantly higher after drinking 3 double whiskies than when sober

Wilcoxon test – large-sample case

As with the Mann-Whitney test, the sampling distribution of the statistic (in this case T) approaches the normal distribution as the sample size becomes large. Having obtained T as in the small-sample (Steps 1–3 inclusive), the standard deviation of T is found as

$$SD_T = \sqrt{\frac{N(N+1)(2N+1)}{24}}$$

and a z-score as

$$z = \left\{T - \frac{N(N+1)}{4}\right\} \div SD_T.$$

The significance decisions are identical to those for the Mann-Whitney large-sample case. Thus, if we have a two-tailed test, the observed z is significant at the 5 per cent level if it exceeds 1·96. For the one-tailed test, significance is attained if z exceeds 1·64. However, as with the Mann-Whitney test, and for the same reasons, the Wilcoxon test tends to be restricted to use with relatively small samples.

Comparison of Mann-Whitney and Wilcoxon with *t*-test

The power efficiency of the Mann-Whitney and Wilcoxon tests, whilst usually somewhat lower than the corresponding t-test, compares very favourably with it. The Mann-Whitney and Wilcoxon tests can be used in situations where the t-test would be inappropriate (e.g. where the assumptions of the t-test obviously do not apply). In other words, they are capable of wider application.

Different statisticians give different advice as to the relative merits of parametric and non-parametric tests. The non-parametric camp claim that their tests are simpler to compute, have fewer assumptions and can be used more widely. The parametric camp claim that their tests are robust with respect to violations of their assumptions and have greater power efficiency.

The strategy recommended here is to use the t-test *unless* the data is in the form of ranks, *or* where the sample is small, and

either the distribution is obviously non-normal or there are obviously large differences in variance.

However, if you are particularly pressed for time or have a large number of analyses to do there is nothing particularly wicked nor inappropriate about using non-parametric statistics, even in cases where *t*-tests might have been used.