Chapter 3: Nonparametric Tests

3.1 Introduction

Nonparametric, or *distribution free* tests are so-called because the assumptions underlying their use are "fewer and weaker than those associated with parametric tests" (Siegel & Castellan, 1988, p. 34). To put it another way, nonparametric tests require few if any assumptions about the shapes of the underlying population distributions. For this reason, they are often used in place of parametric tests if/when one feels that the assumptions of the parametric test have been too grossly violated (e.g., if the distributions are too severely skewed). Discussion of some of the more common nonparametric tests follows.

3.2 The Sign test (for 2 repeated/correlated measures)

The sign test is one of the simplest nonparametric tests. It is for use with 2 repeated (or correlated) measures (see the example below), and measurement is assumed to be at least ordinal. For each subject, subtract the 2nd score from the 1st, and write down the *sign* of the difference. (That is write "-" if the difference score is negative, and "+" if it is positive.) The usual null hypothesis for this test is that there is no difference between the two treatments. If this is so, then the number of + signs (or - signs, for that matter) should have a **binomial distribution**[1] with p = .5, and N = the number of subjects. In other words, the sign test is just a binomial test with + and - in place of Head and Tail (or Success and Failure).

EXAMPLE

> A physiologist wants to know if monkeys prefer stimulation of brain area A to stimulation of brain area B. In the experiment, 14 rhesus monkeys are taught to press two bars. When a light comes on, presses on Bar 1 always result in stimulation of area A; and presses on Bar 2 always result in stimulation of area B. After learning to press the bars, the monkeys are tested for 15 minutes, during which time the frequencies for the two bars are recorded. The data are shown in Table 3.1.

To carry out the sign test, we could let our statistic be the number of + signs, which is 3 in this case. The researcher did not predict a particular outcome in this case, but wanted to know if the two conditions differed. Therefore, the alternative hypothesis is **nondirectional**. That is, the alternative hypothesis would be supported by an extreme number of + signs, be it small or large. A middling number of + signs would be consistent with the null.

The sampling distribution of the statistic is the binomial distribution with N = 14 and p = .5. With this distribution, we would find that the probability of **3 or fewer** + signs is .0287. But because the alternative is nondirectional, or **two-tailed**, we must also take into account the probability **11 or more** + signs, which is also .0287. Adding these together, we find that the probability of (3 or fewer) or (11 or more) is .0574. Therefore, if our pre-determined alpha was set at .05, we would not have sufficient evidence to allow rejection of the null hypothesis.

---

[1] The binomial distribution is discussed on pages 37-40 of Norman & Streiner (2nd ed.). It is also discussed in some detail in my chapter on "Probability and Hypothesis Testing" (in the file *prob_hyp.pdf*).

Table 3.1 Number of bar-presses in brain stimulation experiment

Subject	Bar 1	Bar 2	Difference	"Sign" of Difference
1	20	40	-20	-
2	18	25	-7	-
3	24	38	-14	-
4	14	27	-13	-
5	5	31	-26	-
6	26	21	+5	+
7	15	32	-17	-
8	29	38	-9	-
9	15	25	-10	-
10	9	18	-9	-
11	25	32	-7	-
12	31	28	+3	+
13	35	33	+2	+
14	12	29	-17	-

Tied scores

If a subject has the same score in each condition, there will be no sign, because the difference score is zero. In the case of tied scores, some textbook authors recommend dropping those subjects, and reducing N by the appropriate number. This is not the best way to deal with ties, however, because reduction of N can result in the loss of too much power.

A better approach would be as follows: If there is only one subject with tied scores, drop that subject, and reduce N by one. If there are 2 subjects with tied scores, make one a + and one a -. In general, if there is an even number of subjects with tied scores, make half of them + signs, and half - signs. For an odd number of subjects (greater than 1), drop one randomly selected subject, and then proceed as for an even number.

3.3 Wilcoxon Signed-Ranks Test (for 2 repeated/correlated measures)

One obvious problem with the sign test is that it **discards** a lot of information about the data. It takes into account the direction of the difference, but not the *magnitude* of the difference between each pair of scores. The Wilcoxon signed-ranks test is another nonparametric test that can be used for 2 repeated (or correlated) measures when measurement is at least ordinal. But unlike the sign test, it *does* take into account (to some degree, at least) the magnitude of the difference. Let us return to the data used to illustrate the sign test. The 14 difference scores were:

-20, -7, -14, -13, -26, +5, -17, -9, -10, -9, -7, +3, +2, -17

If we sort these on the basis of their absolute values (i.e., disregarding the sign), we get the results shown in Table 3.2. The statistic T is found by calculating the sum of the positive ranks, and the sum of the negative ranks. T is the smaller of these two sums. In this case, therefore, T = 6.

If the null hypothesis is true, the sum of the positive ranks and the sum of the negative ranks are expected to be roughly equal. But if H_0 is false, we expect one of the sums to be quite small--and therefore T is expected to be quite small. The most extreme outcome favourable to rejection of H_0 is T = 0.

Table 3.2 Difference scores ranked by absolute value

Score	Rank	
+2	1	
+3	2	
+5	3	
-7	4.5	Sum of positive ranks = 6
-7	4.5	
-9	6.5	Sum of negative ranks = 99
-9	6.5	
-10	8	**T = 6**
-13	9	
-14	10	
-17	11.5	
-17	11.5	
-20	13	
-26	14	

If we wished to, we could generate the sampling distribution of T (i.e., the distribution of T assuming that the null hypothesis is true), and see if the observed value of T is in the rejection region. This is not necessary, however, because the sampling distribution of T can be found in tables in most introductory level statistics textbooks. When I consulted such a table, I found that for N = 14, and $\alpha = .05$ (2-tailed), the *critical value* of T = 21. The rule is that if T is equal to or less than $T_{critical}$, we can reject the null hypothesis. Therefore, in this example, we would reject the null hypothesis. That is, we would conclude that monkeys prefer stimulation in brain area B to stimulation in area A.

This decision differs from our failure to reject H_0 when we analysed the same data using the sign test. The reason for this difference is that the Wilcoxon signed-ranks test is more *powerful* than the sign test. Why is it more powerful? Because it makes use of more information than does the sign test. But note that it too does discard some information by using *ranks* rather than the scores themselves (like a paired *t*-test does, for example).

Tied scores

When ranking scores, it is customary to deal with tied ranks in the following manner: Give the tied scores the *mean of the ranks they would have if they were not tied*. For example, if you have 2 tied scores that would occupy positions 3 and 4 if they were not tied, give each one a rank of 3.5. If you have 3 scores that would occupy positions 7, 8, and 9, give each one a rank of 8. This procedure is preferred to any other for dealing with tied scores, because the sum of the ranks for a fixed number of scores will be the same regardless of whether or not there are any tied scores.

If there are tied ranks in data you are analysing with the Wilcoxon signed-ranks test, the statistic needs to be adjusted to compensate for the decreased variability of the sampling distribution of T. Siegel and Castellan (1988, p. 94) describe this adjustment, for those who are interested. Note that if you are able to reject H_0 without making the correction, then do not bother, because the correction will increase your chances of rejecting H_0. Note as well that the problem becomes more severe as the number of tied ranks increases.

3.4 Mann-Whitney U Test (for 2 independent samples)

The most basic independent groups design has two groups. These are often called Experimental and Control. Subjects are randomly selected from the population and randomly assigned to two groups. There is *no basis for pairing scores*. Nor is it necessary to have the same number of scores in the two groups.

The *Mann-Whitney U test* is a nonparametric test that can be used to analyse data from a two-group independent groups design when measurement is at least ordinal. It analyses the *degree of separation* (or the amount of overlap) between the Experimental and Control groups.

The *null hypothesis* assumes that the two sets of scores (E and C) are samples from the same population; and therefore, because sampling was random, the two sets of scores *do not differ systematically* from each other.

The *alternative hypothesis*, on the other hand, states that the two sets of scores *do* differ systematically. If the alternative is directional, or one-tailed, it further specifies the direction of the difference (i.e., Group E scores are systematically higher or lower than Group C scores).

The statistic that is calculated is either U or U'.

U_1 = the number of Es less than Cs
U_2 = the number of Cs less than Es

U = the smaller of the two values calculated above
U' = the larger of the two values calculated above

Calculating U directly

When the total number of scores is small, U can be calculated directly by counting the number of Es less than Cs (or Cs less than Es). Consider the following example:

Table 3.3 Data for two independent groups

Group E:	12	17	9	21	
Group C:	8	18	26	15	23

It will be easier to count the number of Es less than Cs (and vice versa) if we rank the data from lowest to highest, and rewrite it as shown in Table 3.4.

Table 3.4 Illustration of direct calculation of the U statistic

Score	Group	Rank	E<C	C<E
8	C	1	0	
9	E	2		1
12	E	3		1
15	C	4	2	
17	E	5		2
18	C	6	3	
21	E	7		3
23	C	8	4	
26	C	9	4	
			13	7

$U = 7$
$U' = 13$

CHECK:
Note that U + U' = n_1n_2. This will always be true, and can be used to check your calculations. In this case, U + U' = 7 + 13 = 20; and $n_1n_2 = 4(5) = 20$.

Calculating U with formulae

When the total number of scores is a bit larger, or if there are tied scores, it may be more convenient to calculate U with the following formulae:

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1 \tag{3.1}$$

$$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2 \tag{3.2}$$

where n_1 = # of scores in group 1
n_2 = # of scores in group 2
R_1 = sum of ranks for group 1
R_2 = sum of ranks for group 2

As before, U = smaller of U_1 and U_2, and U' = larger of U_1 and U_2.

For the data shown above, $R_1 = 2+3+5+7 = 17$; and $R_2 = 1+4+6+8+9 = 28$. Substituting into the formulae, we get:

$$U_1 = 4(5) + \frac{4(4+1)}{2} - 17 = 13 \tag{3.3}$$

$$U_2 = 4(5) + \frac{5(5+1)}{2} - 28 = 7 \tag{3.4}$$

Therefore, $U = 7$ and $U' = 13$.

Making a decision

The next step is deciding whether to reject H_0 or not. In principle, we could generate a probability distribution for U that is conditional on the null hypothesis being true--much like we did when working with the binomial distribution earlier. Fortunately, we do not have to do this, because there are tables in the back of many statistics textbooks that give you the *critical values* of U (or U') for different values of n_1 and n_2, and for various significance levels.

For the case we've been considering, $n_1 = 4$ and $n_2 = 5$. For a two-tailed test with $\alpha = .05$, the critical value of U = 1. In order to reject H_0, the observed value of U would have to be *equal to or less than* the critical value of U. (Note that **maximum separation** of E and C scores is indicated by **U = 0**. As the E and C scores become more mixed, U becomes larger. Therefore, *small* values of U lead to rejection of H_0.) Therefore, we would decide that we cannot reject H_0 in this case.

Tied scores

According to Siegel and Castellan (1988), any ties that involve observations in the same group do not affect the values of U and U'. (Note that Siegel and Castellan refer to this test as the *Wilcoxon-Mann-Whitney Test*, and that the call the statistic *W* rather than *U*.) But if two or more tied ranks involve observations from both groups, then the values of U and U' are affected, and a correction should be applied. See Siegel & Castellan (1988, p. 134) should you ever need more information on this, and note that the problem is particularly severe if you are dealing with the large-sample version of the test, which we have not yet discussed.

3.5 Kruskal-Wallis H-test (for *k* independent samples)

The Kruskal-Wallis H-test goes by various names, including *Kruskal-Wallis one-way analysis of variance by ranks* (e.g., in Siegel & Castellan, 1988). It is for use with *k* independent groups, where *k* is equal to or greater than 3, and measurement is at least ordinal. (When *k* = 2, you would use the Mann-Whitney U-test instead.) Note that because the samples are independent, they can be of different sizes.

The null hypothesis is that the *k* samples come from the same population, or from populations with identical medians. The alternative hypothesis states that not all population medians are equal. It is assumed that the underlying distributions are continuous; but only ordinal measurement is required.

The statistic H (sometimes also called KW) can be calculated in one of two ways:

$$H = \left[\frac{12}{N(N+1)}\right]\sum_{i=1}^{k} n_i\left(\overline{R}_i - \overline{R}_\bullet\right)^2 \tag{3.5}$$

or, the more common computational formula,

$$H = \left[\frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i}\right] - 3(N+1) \tag{3.6}$$

where k = the number of independent samples
n_i = the number of cases in the i^{th} sample
N = the total number of cases
R_i = the sum of the ranks in the i^{th} sample
$\overline{R}_i$ = the mean of the ranks for the i^{th} sample
$\overline{R}_\bullet = \frac{N+1}{2}$ = the mean of all ranks

Example

A student was interested in comparing the effects of four kinds of reinforcement on children's performance on a test of reading comprehension. The four reinforcements used were: (a) praise for correct responses; (b) a jelly bean for each correct response; (c) reproof for incorrect responses; and (d) silence. Four independent groups of children were tested, and each group received only one kind of reinforcement. The measure of performance given below is the number of errors made during the course of testing.

Table 3.5 Data from 4 independent groups

a	b	c	d
68	78	94	54
63	69	82	51
58	58	73	32
51	57	67	74
41	53	66	65
		61	80

The first step in carrying out the Kruskal-Wallis H-test is to rank order all of the scores from lowest to highest. This can be quite laborious work if you try to do it by hand, but is fairly easy if you use a spreadsheet program. Enter all scores in a single column, and enter a group code for each score one column over. For example:

Group	Score
a	68
a	63
a	58
a	51
a	41
b	78
etc.	

When all the data are entered thus, sort the scores (and their codes) from lowest to highest. Now you can enter ranks from 1 to N (taking care to deal with tied scores appropriately). After the scores are ranked, you can sort the data the data by group code, and then calculate the sum and mean of the ranks for each group. I did this for the data shown above, and came up with the following ranks:

Table 3.6 Ranks of data from Table 3.5

a	b	c	d	
15	19	22	6	
11	16	21	3.5	
8.5	8.5	17	1	
3.5	7	14	18	
2	5	13	12	
		10	20	
40.0	55.5	97.0	60.5	**Sum of Ranks**
8.0	11.1	16.2	10.1	**Mean of Ranks**

CHECK: The sum of ranks from 1 to N will always be equal to [N(N+1)]/2. We can use this to check our work to this point. We have 22 scores in total, so the sum of all ranks should be [22(23)]/2 = 253. Similarly, when we add the sum of ranks for each group, we get 40 + 55.5 + 97 + 60.5 = 253. Therefore, the mean of ALL ranks = 253/22 = 11.5.

Now plugging into Equation 3.4 shown above, we get H = 4.856. If the null hypothesis is true, and the *k* samples are drawn from the same population (or populations with identical medians), and if $k > 3$, and all samples have 5 or more scores, then the distribution of H closely approximates the chi-squared distribution with df = *k*-1. The critical value of chi-squared with df=3 and $\alpha = .05$ is 7.82. In order to reject H_0, the obtained value of H would have to be equal to or greater than 7.82. Because it is less, we cannot reject the null hypothesis.

Sampling distribution of H

As described above, when H_0 is true, if $k > 3$, and all samples have 5 or more scores, then the sampling distribution of H is closely approximated by the chi-squared distribution with df = *k*-1. If $k = 3$ and the number of scores in each sample is 5 or fewer, then the chi-squared distribution should not be used. In this case, one should use a table of critical values of H (e.g., Table O in Siegel & Castellan, 1988).

Tied observations

Tied scores are dealt with in the manner described previously (i.e., they are given the mean of the ranks they would receive if they were not tied). The presence of tied scores does affect the variance of the sampling distribution of H. Siegel and Castellan (1988) show a correction that can be applied in the case of tied scores, but go on to observe that its effect is to increase the value of H. Therefore, if you are able to reject H_0 without correcting for ties, there is no need to do the correction. It should only be contemplated when you have failed to reject H_0.

Multiple comparisons

Rejection of H_0 tells you that at least one of the *k* samples is drawn from a population with a median different from the others. But it does not tell you which one, or how many are different. There are procedures for conducting multiple comparisons between treatments, or comparisons of a control condition to all other conditions in order to answer these kinds of questions. Should you ever need to use one of them, consult Siegel and Castellan (1988, pp. 213-215).

3.6 The Jonckheere test for ordered alternatives

The Jonckheere test for ordered alternatives is similar to the Kruskal-Wallis test, but has a more specific alternative hypothesis. The alternative hypothesis for the Kruskal-Wallis test states that all population medians are *not equal*. The more precise alternative hypothesis for the Jonckheere test can be summarised as follows:

$$H_1: \ \theta_1 \leq \theta_2 \ldots \leq \theta_k$$

where the θ's are the population medians. This alternative is tested against a *null hypothesis of no systematic trend across treatments*.

The test can be applied when you have data for *k* independent samples, when measurement is at least ordinal, and when it is possible to specify *a priori* the ordering of the groups. Because the alternative hypothesis specifies the order of the medians, the test is one-tailed.

Siegel and Castellan (1988) use *J* to symbolise the statistic that is calculated. It is sometimes also called the "Mann-Whitney count". As this name implies, *J* is based on the same kind of counting and summing that we saw when calculating the U statistic via the direct method. The mechanics of it become somewhat complicated for the Jonckheere test, so we will not go into it here. (I hope you are not *too* disappointed!) Should you ever need to perform this test, see Program 5 in Appendix II of Siegel and Castellan (1988). Siegel and Castellan also provide a table of critical values of *J* (for small sample tests).

3.7 Friedman ANOVA

This test is sometimes called the *Friedman two-way analysis of variance by ranks*. It is for use with *k* repeated (or correlated) measures where measurement is at least ordinal. The null hypothesis states that all *k* samples are drawn from the same population, or from populations with equal medians.

Example

The table on the left (below) shows reaction time data from 5 subjects, each of whom was tested in 3 conditions (A, B, and C). The Friedman ANOVA uses ranks, and so the first thing we must do is rank order the *k* scores for each subject. The results of this ranking are shown in the table on the right, and the sum of the ranks ($\sum R_i$) for each treatment is shown at the bottom.

Table 3.7 RT data and ranks for 3 levels of a within-subjects variable

Subj	A	B	C
1	386	411	454
2	542	563	556
3	662	667	665
4	453	502	574
5	548	546	575

Subj	A	B	C
1	1	2	3
2	1	3	2
3	1	3	2
4	1	2	3
5	2	1	3
$\sum R_i$	6	11	13

It may be useful at this point to consider what kinds of outcomes are expected if H_0 is true. H_0 states that all of the samples (columns) are drawn from the same population, or from populations with the same median. If so, then the sums (or means) of the ranks for each of the columns should all be roughly equal, because the ranks 1, 2, and 3 would be expected by chance to appear equally often in each column. In this example, the expected □R for each treatment would be 10 if H_0 is true. (In general, the expected sum of ranks for each treatment is N(*k* +1)/2.) The Friedman ANOVA assesses the degree to which the observed □R's depart from the expected □R's. If the departure is too extreme (or not likely due to chance), one concludes by rejecting H_0.

The F_r statistic is calculated as follows:

$$F_r = \left[\frac{12}{Nk(k+1)} \sum_{i=1}^{k} R_i^2 \right] - 3N(k+1) \qquad (3.7)$$

where N = the number of subjects
k = the number of treatments
R_i = the sum of the ranks for the *ith* treatment

Critical values of F_r for various sample sizes and numbers of treatments can be found in tables (e.g., Table M in Siegel & Castellan, 1988). Note that when the number of treatments or subjects is large, the sampling distribution of F_r is closely approximated by the chi-squared distribution with df = k - 1. (Generally, use a table of critical values for F_r if it provides a value for your particular combination of k and N. If either k or N are too large for the table of critical values, then use the chi-squared distribution with df = k -1.)

For the example we've been looking at, the critical values of F_r are 6.40 for α = .05, and 8.40 for α = .01. In order to reject H_0, the obtained value of F_r must be equal to or greater than the critical value. Therefore, we would fail to reject H_0 in this case.

Tied scores

If there are ties among the ranks, the F_r statistic must be corrected, because the sampling distribution changes. The formula that corrects for tied ranks is actually a general formula that also works when there are no ties. However, it is rather complicated, which is why the

simplified version shown above is used when possible. The general formula is not shown here, but can be found in Siegel and Castellan (1988, p. 179), should you need it.

Multiple comparisons

Siegel and Castellan (1988) also give formulae to be use for conducting multiple comparisons and/or comparisons of a control condition to each of the other conditions.

3.8 Large sample versions of nonparametric tests

You may have noticed that the tables of critical values for many nonparametric statistics only go up to sample sizes of about 25-50. If so, perhaps you have wondered what to do when you have sample sizes larger than that, and want to carry out a nonparametric test. Fortunately, it turns out that the sampling distributions of many nonparametric statistics converge on the normal distribution as sample size increases. Because of that, it is possible to carry out a so-called "large-sample" version of the test (which is really a *z*-test) if you know the mean and variance of the sampling distribution for that particular statistic.

Common structure of all z- and t-tests

As I have mentioned before, all *z*- and *t*-tests have a common structure. In general terms:

$$z \text{ (or } t) = \frac{\text{statistic - (parameter} \mid H_0 \text{ is true)}}{\text{standard error of the statistic}} \tag{3.8}$$

When the sampling distribution of the statistic in the numerator is normal, then if the true (population) standard error (SE) of the statistic is known, the computed ratio can be evaluated against the standard normal (z) distribution. If the true standard error of the statistic is not known, then it must be estimated from the sample data, and the proper sampling distribution is a *t*-distribution with some number of degrees of freedom.

Example: Large-sample Mann-Whitney U test

The following facts are known about the sampling distribution of the U statistic used in the Mann-Whitney U test:

$$\mu_U = \frac{n_1 n_2}{2} \tag{3.9}$$

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \tag{3.10}$$

Furthermore, when both sample sizes are greater than about 20, the sampling distribution of U is (for practical purposes) normal. Therefore, under these conditions, one can perform a z-test as follows:

$$z_U = \frac{U - \mu_U}{\sigma_U} \tag{3.11}$$

The obtained value of z_U can be evaluated against a table of the standard normal distribution (e.g., Table A in Norman & Streiner, 2000). Alternatively, one can use software to calculate the p-value for a given z-score, e.g., StaTable from Cytel, which is available here:

http://www.cytel.com/statable/index.html

Example: Large-sample Wilcoxon signed ranks test

The following are known to be true about the sampling distribution of T, the statistic used in the Wilcoxon signed ranks test:

$$\mu_T = \frac{N(N+1)}{4} \tag{3.12}$$

$$\sigma_T = \sqrt{\frac{N(N+1)(2N+1)}{24}} \tag{3.13}$$

If $N > 50$, then the sampling distribution of T is for practical purposes normal. And so, a z-ratio can be computed as follows:

$$z_T = \frac{T - \mu_T}{\sigma_T} \tag{3.14}$$

The obtained value of z_T can be evaluated against a table of the standard normal distribution, or using software as described above.

Example: Large-sample Jonckheere test for ordered alternatives

The mean and standard deviation of the sampling distribution of J are given by the following:

$$\mu_J = \frac{N^2 - \sum_{i=1}^{k} n_i^2}{4} \tag{3.15}$$

$$\sigma_J = \sqrt{\frac{1}{72}\left[N^2(2N+3) - \sum_{i=1}^{k} n_i^2(2n_i+3) \right]} \tag{3.16}$$

where N = total number of observations
n_i = the number of observations in the i^{th} group
k = the number of independent groups

As sample sizes increase, the sampling distribution of J converges on the normal, and so one can perform a *z*-test as follows:

$$z_J = \frac{J - \mu_J}{\sigma_J} \tag{3.17}$$

Example: Large sample sign test

The sampling distribution used in carrying out the sign test is a binomial distribution with $p = q = .5$. The mean of a binomial distribution is equal to Np, and the variance is equal to Npq. As N increases, the binomial distribution converges on the normal distribution (especially when $p = q = .5$). When N is large enough (i.e., greater than 30 or 50, depending on how conservative one is), it is possible to carry out a *z*-test version of the sign test as follows:

$$z = \frac{X - Np}{\sqrt{Npq}} \tag{3.18}$$

You may recall that z^2 is equal to χ^2 with $df = 1$. Therefore,

$$\chi_1^2 = z^2 = \frac{(X - Np)^2}{Npq} \tag{3.19}$$

This formula can be expanded with what Howell (1997) calls "some not-so-obvious algebra" to yield:

$$\chi_1^2 = \frac{(X - Np)^2}{Np} + \frac{(N - X - Nq)^2}{Nq} \tag{3.20}$$

Note that X equals the **observed** number of p-events, and Np equals the **expected** number of p-events under the null hypothesis. Similarly, N-X equals the **observed** number of q-events, and Nq = the **expected** number of q-events under the null hypothesis. Therefore, we can rewrite equation (3.20) in a more familiar looking format as follows:

$$\chi^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} = \sum \frac{(O - E)^2}{E} \tag{3.21}$$

Large-sample z-tests with small samples

Many computerised statistics packages automatically compute the large-sample (*z*-test) version of nonparametric tests, even when the sample sizes are small. Note however, that the z-test is just *an approximation* that can be used when sample sizes are sufficiently large. If the sample sizes are small enough to allow use of a table of critical values for your particular nonparametric statistic, you should always use it rather than a *z*-test.

3.9 Advantages of nonparametric tests

Siegel and Castellan (1988, p. 35) list the following advantages of nonparametric tests:

1. If the sample size is very small, there may be no alternative to using a nonparametric statistical test unless the nature of the population distribution is *known exactly*.

2. Nonparametric tests typically make fewer assumptions about the data and may be more relevant to a particular situation. In addition, the hypothesis tested by the nonparametric test may be more appropriate for the research investigation.

3. Nonparametric tests are available to analyze data which are inherently in ranks as well as data whose seemingly numerical scores have the strength of ranks. That is, the researcher may only be able to say of his or her subjects that one has more or less of the characteristic than another, without being able to say *how much* more or less. For example, in studying such a variable as anxiety, we may be able to state that subject A is more anxious than subject B without knowing at all exactly how much more anxious A is. If data are inherently in ranks, or even if they can be categorized only as plus or minus (more or less, better or worse), they can be treated by nonparametric methods, whereas they cannot be treated by parametric methods unless precarious and, perhaps, unrealistic assumptions are made about the underlying distributions.

4. Nonparametric methods are available to treat data which are simply classificatory or categorical, i.e., are measured in a nominal scale. No parametric technique applies to such data.

5. There are suitable nonparametric statistical tests for treating samples made up of observations from several different populations. Parametric tests often cannot handle such data without requiring us to make seemingly unrealistic assumptions or requiring cumbersome computations.

6. Nonparametric statistical tests are typically much easier to learn and to apply than are parametric tests. In addition, their interpretation often is more direct than the interpretation of parametric tests.

Note that the objection concerning "cumbersome computations" in point number 5 has become less of an issue as computers and statistical software packages become more sophisticated, and more available.

3.10 Disadvantages of nonparametric tests

In closing, I must point out that nonparametric tests do have at least two major disadvantages in comparison to parametric tests. First, **nonparametric tests are less powerful**. Why? Because parametric tests use more of the information available in a set of numbers. Parametric tests make use of information consistent with interval scale measurement, whereas parametric tests typically make use of ordinal information only. As Siegel and Castellan (1988) put it, "nonparametric statistical tests are wasteful."

Second, parametric tests are much more flexible, and allow you to test a greater range of hypotheses. For example, factorial ANOVA designs allow you to test for interactions between variables in a way that is not possible with nonparametric alternatives. There are nonparametric techniques to test for *certain kinds of interactions under certain circumstances*, but these are much more limited than the corresponding parametric techniques.

Therefore, when the assumptions for a parametric test are met, it is generally (but not necessarily always) preferable to use the parametric test rather than a nonparametric test.

Review Questions

1. Which test is more powerful, the sign test, or the Wilcoxon signed ranks test? Explain why.

2. Which test is more powerful, the Wilcoxon signed ranks test, or the *t*-test for correlated samples? Explain why.

For the scenarios described in questions 3-5, identify the nonparametric test that ought to be used.

3. A single group of subjects is tested at 6 levels of an independent variable. You would like to do a repeated measures ANOVA, but cannot because you have violated the assumptions for that analysis. Your data are ordinal.

4. You have 5 independent groups of subjects, with different numbers per group. There is also substantial departure from homogeneity of variance. The null hypothesis states that there are no differences between the groups.

5. You have the same situation described in question 4; and in addition, the alternative hypothesis states that when the mean ranks for the 5 groups are listed from smallest to largest, they will appear in a particular *pre-specified* order.

6. Explain the rationale underlying the large-sample *z*-test version of the Mann-Whitney U-test.

7. Why should you not use the large-sample *z*-test version of a nonparametric test when you have samples small enough to allow use of the small-sample version?

8. Give two reasons why parametric tests are generally preferred to nonparametric tests.

9. Describe the circumstances under which you might use the Kruskal-Wallis test. Under what circumstances would you use the Jonckheere test instead? (HINT: Think about how the alternative hypotheses for these tests differ.)

References

Howell, DC. (1997). *Psychology for statistics*. Duxbury Press.

Siegel, S., & Castellan, N.J. (1988). *Nonparametric statistics for the behavioral sciences* (2nd Ed.). New York, NY: McGraw-Hill.

Appendix

The following is an excerpt from the BMJ Statistics at Square One article on Rank Score Tests, which can be downloaded here:

http://www.bmj.com/collections/statsbk/10.shtml

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.

The entire series of Statistics at Square One chapters is listed here:

http://www.bmj.com/collections/statsbk/