CHAPTER 10

Nonparametric Methods

Nonparametric methods are frequently used with small sample sizes as large sample properties of parametric tests then do not apply. Nonparametric methods have less power than their corresponding parametric methods when the assumptions for the parametric methods are met. This is well known, but computing power for nonparametric tests when the assumption of a specific distribution is not met poses special problems. This is because a specified distribution is used in computing probabilities for parametric methods (as seen in previous chapters), with such probabilities being the power values. For these reasons, it is best to use parametric methods if it appears, for a given set of data, that the parametric distribution assumptions are not seriously violated. But when we are at the point of trying to determine sample size for use with a nonparametric method, we don't yet have any data! Therefore, we can't test for any distribution assumption.

In general, we would use a nonparametric approach if we seriously doubt that the distribution assumption for a parametric test is likely to not even be approximately met, such as more than slight nonnormality. We need to consider the type of test being used, however, as some tests are robust to violations of the test assumptions and some are not. Robustness of various statistical methods is discussed by Rasch and Guirard (2004).

We need not be concerned with the actual distribution when the test is performed, but, somewhat paradoxically, a distribution must be assumed when sample size is determined for a specified power because power depends on the actual distribution! Of course, if we knew the actual distribution, we wouldn't use a nonparametric (distribution-free) approach in the first place! Because of these difficulties, one conservative approach would be to inflate the sample size for the corresponding parametric approach. This is discussed in Section 10.2. Certainly any such simple rule-of-thumb would avoid the complexities involved in sample size determination for nonparametric methods, in general.

Sample Size Determination and Power, First Edition. Thomas P. Ryan.

Although the focus in this chapter is on determining sample size when a nonparametric test is used, Walker (2003) considered a Bayesian nonparametric decision theoretic approach to sample size calculations, independent of any specific test. Mumby (2002) pointed out that the power of nonparametric tests has not been discussed to any extent in the literature and advocated the use of simulation for power and sample size determination.

10.1 WILCOXON ONE-SAMPLE TEST

The Wilcoxon (1945) one-sample test, which is often called the Wilcoxon signed rank test, is used to test whether the median of a symmetric distribution is equal to a specified number. (With a normal distribution, the median and mean are the same.) For observations x_i, the test entails computing $|x_i - \text{median}|$ for $i = 1, 2, \ldots, n$, and ranking the absolute values from smallest to largest, while keeping track of which values of $(x_i - \text{median})$ are positive and which are negative. Let T_+ denote the total of the ranks for the positive deviations and let T_- denote the total of the ranks for the negative deviations. Typically, the smaller of the two totals is used as the test statistic, but that doesn't really matter.

If the sample data support the hypothesized median, the two totals should be about the same. If, however, the data do not support the hypothesized median, we would expect the totals to differ considerably.

In general, if a parametric test assumption is normality and a nonparametric test is to be used but normality is specified for the purpose of comparing the sample size with the required sample size for the parametric procedure, the sample sizes will of course differ. For example, it was shown in Example 2.1 that $n = 56$ when H_0: $\mu = 100$ is tested against the alternative H_α: $\mu > 100$, using $\alpha = .05$ and the desired power being .80 for detecting $\mu = 101$ under the assumption $\sigma = 3$. If the nonparametric counterpart, the one-sample Wilcoxon test, is used to test that the median is equal to a specified value (such as 100 in this case), the sample size will be greater than 56 if normality is assumed because the ranks of the data are used relative to the median, rather than the distance from the median being used. Since information from a sample is less when ranks are used rather than the actual observations, the necessary sample size must be greater than when the latter are used.

The term *asymptotic relative efficiency* (A.R.E.) is used in comparing nonparametric tests to the corresponding parametric test in terms of power, computed assuming that the assumptions of the parametric test are met. The A.R.E. is defined as $100 \lim_{n\to\infty}(n_a/n_b)$, with n_a = sample size for the parametric test and n_b = sample size for the nonparametric test, using of course the same null hypothesis and assumed true parameter value, α, and power.

For the Wilcoxon test, the A.R.E. is $3/\pi = 95.5\%$. The term *relative efficiency* is $100 n_a/n_b$. PASS can be used to show that $n = 59$ for this example. The relative efficiency is then 100(56/59) = 94.9%, slightly less than the A.R.E. If the power is increased to .99, the sample sizes are 149 and 142, respectively, with 100(142/149) = 95.3%—very close to the asymptotic result and expectedly closer to 95.5% than with sample sizes of 59 and 56, respectively, since the sample sizes are much larger.

As explained by Conover (1980, p. 291) and Lehmann (2006, p. 80), the A.R.E. of the Wilcoxon test relative to the *t*-test has a lower bound of 0.864, so a conservative approach would be to compute the sample size for the *t*-test and multiply it by (1/0.864) = 1.1574, and use that as the sample size for the Wilcoxon test. That would be a more defensible approach than what is done when sample size determination software is used because the user must specify a distribution, but if the distribution were known, a test could be performed using the appropriate theory for that distribution and it wouldn't be necessary to use the Wilcoxon test!

Noether (1987) gave the sample size expression for a one-sided test as

$$n = \frac{(z_\alpha + z_\beta)^2}{3(p' - 1/2)^2} \tag{10.1}$$

with α being the significance level of the test, z_α and z_β being as defined in previous chapters, and $p' = P(X + X' > \text{median})$, with X and X' denoting two independent observations. Equation (10.1) results from an approximation that is stated "for sufficiently large *n*." The existence of z_α and z_β in Eq. (10.1) implies that a normal approximation (i.e., asymptotic approach) is being used. Such approaches can be used when n is large, but whether or not n is large won't be known until it is calculated. How well Eq. (10.1) works for small n, which is an important question since nonparametric procedures are most often applied to small samples, is apparently unknown. It might not work very well, however, because the presence of z_α and z_β in the numerator means that there is the assumption of approximate normality of the test statistic.

The median in Noether's article is assumed to be zero, without loss of generality, but of course this isn't necessary, in general. This requires that the user have some prior knowledge regarding the actual distribution. Note that if symmetry does exist and the median is that specified by the null hypothesis, this would cause Eq. (10.1) to be undefined as the probability for a positive deviation would be .5, as would the probability of a negative deviation. This is essentially irrelevant, however, because sample size in general is for detecting a state that differs from what is specified in the null hypothesis. It wouldn't make any sense to speak of power if the null hypothesis were true. [See Noether (1987, p. 646) for the derivation of Eq. (10.1), which is an approximate large sample result.]

For the current example and the assumed mean of 101, $X + X'$ has a normal distribution with a mean of 202 (μ^*, say) and a standard deviation, $\sigma^* = \sqrt{18} =$

4.24, with the mean under the null hypothesis equal to 200. Then, $P(X + X' > 200|\mu^* = 202 \text{ and } \sigma^* = 4.24) = .6813$. Then

$$\begin{aligned} n &= \frac{(z_\alpha + z_\beta)^2}{3(p' - 1/2)^2} \\ &= \frac{(1.645 + 0.84)^2}{3(0.681324 - 1/2)^2} \\ &= 62.61 \end{aligned}$$

so $n = 63$ would be used. This differs from the $n = 59$ obtained, as indicated previously, using PASS, but PASS computes the sample size in a different way. Specifically, PASS uses an adjustment from the sample size that would be required for a t-test, following the recommendation of Al-Sunduqchi (1990), who suggested multiplying the sample size required for the t-test (or z-test) by a factor that depends on the assumed distribution. For a normal distribution, as assumed in this example, the multiplier is $\pi/3 = 3.14159/3 = 1.0472$. So PASS obtains $n = 59$ as $56(1.0472) = 58.64$, so $n = 59$ would be used. This is an asymptotic adjustment that is based on the asymptotic relative efficiency of $3/\pi$ that was given previously. Of course, this adjustment should work well when the t-test sample size is large, but perhaps not so well when the t-test sample size is small. This is apparently unknown, however, and needs to be investigated, as does the Noether approximation. Although not specifically addressing this issue, Kolassa (1995) did examine the accuracy of some approximations to sample size and power and concluded that “in some cases one must exercise much care in using the simpler approximations.”

Of course, the real value of nonparametric tests is when normality does *not* exist, but we would want to think twice about using a nonparametric test that was very inferior to the corresponding parametric test in terms of power when approximate normality *did* exist.

Note that determining the sample size to be used for a Wilcoxon one-sample test places considerable demands on the user. Unlike a parametric test of a single mean, which requires the user to state a value for the mean that the user wished to detect with the specified power, the user of the Wilcoxon test must not only specify a value for the median of the distribution (or the mean if PASS is used), but must also specify the *name* of the distribution. In reality, this could be anything; in PASS the options are normal, uniform, double exponential, and logistic. To see how the choice of distribution affects the sample size using PASS, the sample sizes produced under the assumption of a uniform, double exponential, and logistic distribution are 56, 38, and 52, respectively, so choice of distribution does have a sizable effect on the sample size when the choice is the double exponential distribution.

G*Power also has the capability for the Wilcoxon one-sample test. Its options for the assumed distribution are normal, Laplace, and logistic. The sample sizes that it produces for each of these are 60, 39, and 52, respectively. Thus, there is agreement between G*Power and PASS for the logistic distribution and a difference of 1 for a normal distribution. MINITAB, Power and Precision, and nQuery all do not have sample size determination capability for nonparametric tests.

If prior information suggests that the actual distribution probably cannot be adequately represented by any of these four distributions, the user might want to estimate p' using whatever distribution seems appropriate (or using prior knowledge, if it exists), and then use the formula given by Noether (1987). Even if one of these four distributions seemed appropriate in a given application, it might be a good idea to determine the sample size using both approaches and compare the results, especially since the work of Al-Sunduqchi was apparently never published.

Sample size expressions for the one-sample signed rank test were considered by Shieh, Jan, and Randles (2007); Wang, Chen, and Chow (2003) looked at sample size determination for both the one-sample test and the two-sample rank test (Section 10.2).

10.1.1 Wilcoxon Test for Paired Data

The Wilcoxon one-sample test can also be applied to paired data and be the nonparametric counterpart to a paired-*t* test when normality is assumed. That is, the one sample can be a sample of differences, computed by subtracting each of the second set of observations from each of the corresponding observations in the first sample. Then the null hypothesis is the same as when the starting point is just a single sample. This is probably the most common use of the Wilcoxon test.

The set of differences might be correlated, however, so Rosner, Glynn, and Lee (2003, 2006) proposed a modified Wilcoxon test for paired comparisons of clustered data and Rosner and Glynn (2011) proposed sample size determination methods for the test by extending their methods for the regular Wilcoxon test that were given in Rosner and Glynn (2009).

10.2 WILCOXON TWO–SAMPLE TEST (MANN–WHITNEY TEST)

The Wilcoxon two-sample test is more commonly referred to as the Mann–Whitney (1947) test and is sometimes called the Wilcoxon–Mann–Whitney test [as in Rahardja, Zhao, and Qu (2009) and Shieh, Jan, and Randles (2006)]. It is used for two independent samples to test whether the corresponding two populations have the same distribution when it is not reasonable to assume a normal distribution. As such, it is the most commonly used nonparametric test for comparing two populations. Okeh (2009) surveyed five biomedical journals and concluded that the Mann–Whitney two-sample test and the Wilcoxon one-sample test should be used more frequently in medical research, in which nonnormal data are widespread, with much data being ordinal (Rabbee, Coull, Mehta, Patel, and Senchaudhuri, 2003). Posten (1982) studied the power of the test relative to the independent sample t-test for various nonnormal distributions and concluded that the former is superior to the latter, although not for U- and J-shaped distributions. Wang, Chen, and Chow (2003) considered sample size determination for the two Wilcoxon tests and found that the methods work well under various alternative hypotheses for moderate sample sizes.

The two independent samples are merged when the Mann–Whitney test is used for the purpose of ranking the numbers from smallest to largest and the sum of the ranks is computed for each sample. The test statistic is then either the smaller or the larger of the two sums, as it doesn't make any difference which one is used.

As with nonparametric tests in general, there is both an exact test and an asymptotic form of the test. As discussed by Rahardja et al. (2009), the asymptotic version is the one that is most commonly used. One problem with exact tests is that they are generally computationally intensive, with the computations prohibitive for more than small-to-moderate sample sizes. Another problem is that they are generally conservative. For example, for the tests that they considered for a test of association in a small sample, unordered $r \times c$ tables, Lydersen, Pradhan, Senchaudhuri, and Laake (2007) found that "in general, we observe that the significance levels of the exact tests can be substantially lower than α, which demonstrates that exact tests are conservative." Thus, even though the term "*exact* tests" may have a good ring to it, such tests should not automatically be chosen over asymptotic tests. If a large sample size is used, the asymptotic test may be quite satisfactory.

Tied data points and thus tied ranks will often occur, necessitating the use of a method of handling ties. Zhao, Rahardja, and Qu (2008a) proposed such a method, which is incorporated in the R package `samplesize` (see `http://cran.r-project.org/web/packages/samplesize/index.html`).

Most sample size determination software do not have Mann–Whitney test capability, and that is also true for nonparametric tests in general, as noted previously. PASS 11 does have that capability, more or less, but that isn't readily apparent as the routine is not a menu item. This is because it is included as a variation of the parametric t-test, with the sample size computed by adjusting the sample size that would be used for the independent-sample t-test, with the adjustment factor being one of those given by Al-Sunduqchi (1990). Specifically, the appropriate sample size for the independent sample t-test is multiplied by 1 for a uniform distribution; 2/3 for a double exponential distribution; $9/\pi^2$ for logistic distribution; and $\pi/3$ for a normal distribution. It should be kept in mind that those adjustment factors are based on asymptotic theory so they could be off considerably for sample sizes that are not large. The efficacy of these adjustment factors has apparently not been investigated, and similarly, Al-Sunduqchi's (1990) work was apparently never published. Furthermore, since there is no stand-alone Mann–Whitney procedure, there is no adjustment in the case of ties, as in Zhao et al. (2008a).

The need to specify a distribution to determine the sample size for a nonparametric test is disturbing because if the distribution were known, at least approximately, then there wouldn't be a need to use a nonparametric test! To avoid this complication, one recommendation is if the corresponding parametric test is a *t*-test, one should add 15% to what the sample size should be if the *t*-test could be used. This is a commonly offered suggestion, provided that the sample sizes are reasonably large. This rule-of-thumb might be motivated by the fact that the A.R.E. of the Mann–Whitney test relative to the *t*-test cannot be less than 0.864, regardless of the actual distribution (see, e.g., Conover, 1980, p. 291). Since $1/0.864 = 1.157$, this would suggest an adjustment of 15% or 16% as the largest adjustment that should be needed.

Consider the example in Section 3.2 for which the required common sample size for a two-tailed test was given as 16. If we use a Mann–Whitney test with normality assumed for the purpose of determining the relative efficiency, we obtain $n = 17$ using PASS. A slight increase in the sample size would be expected since $n = 16$ is a small sample. We may note that $16/17 = .94$—essentially in line with what we would expect. This would suggest that the proxy for a true Mann–Whitney procedure that is available in PASS may produce reasonable results.

By comparison, nQuery can also be used for the Mann–Whitney test, but the inputs are quite different, as the user must specify $P(X < Y)$, with X and Y being the random variables corresponding to the two groups. A user may have difficulty deciding what value to input since a probability will generally be harder to estimate than a mean or standard deviation. The suggestion given by nQuery is to use the "Assistants" pull-down menu and select "Compute Effect Size." The user then enters each of the two population means and the common standard deviation, and the software computes $P(X < Y)$. For example, if the user enters 50 and 55 for the two population means and 2 for the common standard deviation, the software gives $P(X < Y) = .961$. This is based on the assumption of normality for each distribution as simple hand calculation gives this result since the variance of $(X - Y) = 8$ and $P[(X - Y) = 0 | \mu_1 - \mu_2 = -5]$ leads to $Z = 5/\sqrt{8} = 1.76777$ and $P(Z < 1.76777) = .96145$. Then for a two-sided test with $\alpha = .05$, power $= .90$, and $P(X < Y) = .961$, nQuery gives $n = 9$. Of course, this is a somewhat crude approach to sample size determination because if there was normality and a common standard deviation, a Mann–Whitney test wouldn't be used in the first place! By comparison, G*Power gives a sample size of $n = 10$, using the same inputs, although the user can select a normal distribution as one of the options and doesn't have to specify $P(X < Y)$.

Chakraborti, Hong, and van de Wiel (2006) gave the sample size expression, attributed to Noether (1987), for the Mann–Whitney test for continuous data and equal sample sizes as

$$n = \frac{[z_{\alpha/2} + z_{\beta}]^2}{6(p - 0.5)^2}$$

and briefly explained how it is derived, with p denoting $P(Y > X)$. See also how it is used in Walters (2004). To illustrate, again assume $\alpha = .05$, power $= .90$, and $P(X < Y) = .961$. This produces

$$\begin{aligned} n &= \frac{[1.96 + 1.28]^2}{6(.961 - 0.5)^2} \\ &= 8.23 \end{aligned}$$

so that $n = 9$ would be used, in agreement with the nQuery output. If the desired power had been .80, then $n = 7$ would be used. Although these are very small sample sizes, the inputs used to obtain the .961 probability corresponded to a difference in assumed population means of 1.77 times the standard deviation of $(X - Y)$ and 1.77 is not a small Z-value.

Chakraborti et al. (2006) also gave the sample size that results from solving for sample size from the power expression given by Lehmann (1975) for the Mann–Whitney test, pointing out that it should be more accurate than the expression given by Noether (1987) but it is also more demanding of the user in terms of inputs. Consequently, Chakraborti et al. (2006) proposed two new methods, both of which rely on pilot samples and one method involves bootstrapping.

It should be apparent that a rigorous approach to sample size determination for this test, as well as nonparametric tests in general, will be impossible without (sufficient) data from at least one pilot study. Rahardja et al. (2009) recommended the following:

> For a particular application, we recommend that readers study the plot data carefully and understand the underlying distribution. Then, readers can apply the method most suitable to their data. If possible, we also recommend that readers try various sample size formulas under different assumptions.

Although this is generally sound advice, a considerable amount of data is needed to gain insight into the underlying distribution.

It is well known that a t-test applied to the ranks of the data has the same power and Type I error probability as the Mann–Whitney test. Zimmerman (2011) showed that the power and Type I error probability remain about the same if the number of ranks is reduced from a very large number to a much smaller number by replacing sequences of ranks with a single number.

Divine, Kapke, Havstand, and Joseph (2010) examined the various methods that have been proposed for determining sample size for the Mann–Whitney test and found that under certain conditions the formula given by Zhao et al. (2008a) works just as well as the methods incorporated in nQuery Advisor and SAS 9.2 PROC POWER, although the latter two can be more accurate for certain allocation ratios.

Although exact tests, such as Fisher's exact test, are often computationally prohibitive, some research has been performed on determining sample size for exact tests, such as Hilton and Mehta (1993).

10.2.1 van Elteren Test—A Stratified Mann–Whitney Test

van Elteren (1960) proposed a test that is a form of the Mann–Whitney test and is applicable when the data are stratified and stratified factors are to be accounted for in the analysis. As discussed by Zhao, Rahardja, and Mei (2008b), it has been used in various fields of application, including ecological studies, epidemiological studies, and clinical trials. Qu, Zhao, and Rahardja (2008) compared the Mann–Whitney test when strata are used with the van Elteren test and found that the latter is preferable when the stratum effects are large and the Mann–Whitney test should be used when the effects are small.

Zhao (2006) considered the asymptotic version of the van Elteren test and presented three large-sample size estimation methods, in addition to presenting sample size estimation when the stratum fractions are unknown. Zhao, Qu, and Rahardja (2006) approximated the power of the test when the response data from a new treatment is limited to summary statistics.

10.3 KRUSKAL–WALLIS ONE-WAY ANOVA

This is the counterpart to one-factor ANOVA with normality assumed. The data are converted to ranks, which are used in the computations. PASS uses simulation to arrive at the common sample size and gives suggestions regarding the number of simulations to use, as it does for its many other procedures that use simulation.

Fan, Zhang, and Zhang (2011) used a bootstrap approach to determining sample size. Specifically, they adapted a particular bootstrap power calculation technique, the extended average *X* and *Y* method of Mahoney and Magel (1996), and then generalized the sample size calculation method for the Wilcoxon test given by Hamilton and Collings (1991) to the Kruskal–Wallis test. Sample size determination for this test was also considered by Rasch and Šimečková (2007).

10.4 SIGN TEST

Any test that uses only the sign of the number, relative to some criterion, is certainly going to have less power than tests that use the ranks of the numbers, such as the Wilcoxon one-sample test. Thus, the A.R.E. for the sign test will be less than the A.R.E. for the Wilcoxon one-sample test, and it is therefore presented herein after the sign test.

For a two-sided test of H_0: $\text{median}_0 = c$, the hypothesis would be rejected if there was a disproportionate number of plus signs (values above c) or minus signs (values below c). The objective would be to solve for n such that a rejection region of size approximately α and the desired power results for a specified value of median $\neq c$. For example, if $n = 17$, we can apply the binomial distribution and observe that below 5 and above 12 gives a significance level of $2(.02452) = .04904$, assuming that median $= c$ and there is a symmetric distribution. (Of course, "above 12" and "below 5" each imply the other when there are no values equal to c.) Let $\text{median}_0 = 0$ and median $= 1$, with $\sigma = 1$. Now what is the probability that the number of plus signs exceeds 12 when $\mu = 1$? That probability should be equal to the specified power, which for this example was given as .80.

PASS uses simulation to arrive at the necessary sample size, but that really isn't necessary because the appropriate formula was given by Noether (1987), as well as the development of it. The latter is similar to the development of Eq. (2.3) for the one-sided, one-sample mean test. That equation was given as

$$Z_\alpha + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} - Z_\beta \tag{10.2}$$

which can be rewritten

$$\frac{\mu - \mu_0}{\sigma/\sqrt{n}} = Z_\alpha + Z_\beta \tag{10.3}$$

Noether (1987) defines

$$Q(T) = \left(\frac{\mu(T) - \mu_0(T)}{\sigma_0(T)}\right)^2$$

as the noncentrality factor for the test T, with the sample size determined by setting $Q(T) = (Z_\alpha + Z_\beta)^2$ and solving for n. Note that this result has the same general form as would be obtained by squaring Eq. (10.3) and solving for n and also note that there is an implied assumption that $\sqrt{Q(T)}$ has a normal distribution, at least approximately.

For the sign test, let $S =$ the number of observations that exceed the hypothesized median [following Noether's (1987) notation]. Then $\mu(S) = np$ and $\sigma^2(S) = np(1 - p)$, with $p = P(X > \text{median}_0)$ and the mean and variance being those of the binomial distribution. With $p = 1/2$, by assumption, it follows that $\sigma_0^2(S) = n(1/2)(1 - 1/2) = n/4$. Then

$$Q(S) = \left(\frac{np - n(1/2)}{n/4}\right)^2 = 4n(p - 1/2)^2$$

Setting $Q(S) = (Z_\alpha + Z_\beta)^2$ and solving for n then produces

$$n = \frac{(z_\alpha + z_\beta)^2}{4(p - 1/2)^2} \tag{10.4}$$

Note that the p in Eq. (10.4) is the "true p." That is, it is the probability that an observation exceeds the hypothesized median, given the actual median. As with Noether's sample size expression for the one-sample Wilcoxon test, this expression depends on the assumption of normality of the test statistic.

To illustrate the use of Eq. (10.4), assume that the standard normal distribution, median = mean = 0, is hypothesized, but the mean and median = 1 instead of zero. Then $P(X > 0|\ \mu = 1) = .8413$.

Equation (10.2) applies whether the test is upper-tailed or lower-tailed, but if a two-sided test is to be used, z_α would be replaced by $z_{\alpha/2}$.

Assume a two-sided test. Since $P(X > 0|\ \mu = 1) = .8413$ when normality is assumed with $\sigma = 1$, it follows that

$$\begin{aligned} n &= \frac{(1.96 + 0.84)^2}{4(.8413 - 1/2)^2} \\ &= 16.955 \end{aligned}$$

PASS uses simulation to arrive at the same answer. [If the results differed more than slightly, this would suggest that the normal approximation that is inherent in the expression given by Noether (1987) is probably inadequate.] PASS similarly uses simulation to estimate the power and the documentation contains a table giving the degree of error to be expected in estimating the power for various numbers of simulations. When 100,000 simulations are specified, PASS does that many simulations for various sample sizes and converges to $n = 17$ for this example. The value of α is given as .049, which is correct because, as stated previously, $\alpha = .04904$. (This does not depend on the adequacy of a normal approximation because the value of α is, as indicated previously, the sum of two tail areas of the appropriate binomial distribution.) The value for the power is off slightly, however, because PASS gives .882, whereas the correct value, using the binomial distribution, is .88086. PASS does indicate, though, that the true value is between .880 and .884 for the particular set of simulations that I ran, although it is unnecessary to use simulations to compute power because a true distribution with specified parameter values must be assumed in order to compute power, and the power would be computed directly using that distribution.

Ahn, Hu, and Schucany (2011) extended Noether's formula to binary observations that are dependent within a cluster and a sign test is to be used to test a proportion. Specifically, they gave the sample size expression for each of three standardized test statistics, with those test statistics resulting from the use of equal weights to observations, equal weights to clusters, and optimal weights, respectively, while stating that all three standardized test statistics are the same if cluster sizes are constant.

10.5 McNEMAR'S TEST

McNemar's test is a test for significant changes. Lachenbruch (1992) considered the determination of sample size for that test under certain conditions and compared the results with sample sizes given previously in the literature. Lu and Bean (1995) considered the problem of testing the one-sided equivalence in the sensitivities of two medical diagnostic tests under a matched-pair design and derived conditional and unconditional sample size formulas. They claimed that their approach was superior to that of Lachenbruch (1992). Lachin (1992) compared various methods for determining sample size. PASS utilizes the approach given by Schork and Williams (1980).

10.6 CONTINGENCY TABLES

The simplest type of contingency table is a 2×2 table of counts, with the hypothesis that is tested being that the two classification factors are independent. Fisher's exact test is often used as the method of analysis, although it usually will be inapplicable as all four marginal totals would have to be fixed for the test to be applicable.

Very little research has been performed on determining sample size when a contingency table analysis is to be performed. Lydersen, Fagerland, and Laake (2009) give their recommended tests for testing for association in $2{\times}2$ tables and very briefly discussed power and sample size calculations for those tests. Some research has been performed for determining sample size for such tables. Dai, Li, Kim, Kimball, Jazwinski, and Arnold (2007) determined sample size for $2 \times 2 \times 2$ tables that are to be analyzed using Fisher's exact test, despite the fact that only two of the marginal totals were known.

10.7 QUASI-LIKELIHOOD METHOD

Mahnken (2009) showed that quasi-likelihood methods can be used for sample size determination when the response variable distribution is unknown but the relationship between the mean and variance is known, or assumed, as only knowledge of the variance as a function proportional to the mean is needed.

Sample size determination, which is done iteratively, is performed using an approach that is based on asymptotic arguments.

Numerical results showed that the results were more conservative than results obtained using Monte Carlo (MC) simulation, and that when the actual distribution was either a normal distribution or a mixture distribution, power estimates are different from the MC power estimates by less than .03 in the vast majority of cases.

10.8 RANK CORRELATION COEFFICIENTS

The Pearson correlation coefficient (see Section 5.7) is based on the assumption that the two random variables have a joint bivariate normal distribution, which implies that they have a linear relationship. As pointed out in Section 5.7, the bivariate normal distribution assumption will hardly ever be met and the relationship may not be linear, so it is then a question of how much of a departure there is from these assumptions.

As an alternative, there are various nonparametric correlation coefficients that are well known and have been used extensively. One of these is the Spearman correlation coefficient, which is the Pearson correlation computed using the ranks of the data. Specifically, the data in each sample are ranked from smallest to largest and the Pearson correlation is computed using those ranks.

Another correlation coefficient is Kendall's tau (τ), which is also a measure of agreement between ranks. As an example, assume that four student essays are ranked by each of two judges. There are multiple ways to compute the value of τ but the simplest way is as follows. List the ranks assigned by the first judge in ascending order and list the ranks of the second judge in juxtaposition and before the ranks of the first judge. Assume that the corresponding ranks of the second judge are, in order, 2 3 1 4. Let Q equal the total number of times that a smaller rank is to the right of each rank. That number is obviously 2 if the second judge's ranks are listed first. Kendall's τ is then computed as $\tau = 1 - 4Q/[n(n - 1)]$, with n in this example denoting the number of student essays. Thus, $\tau = 1 - 4(2)/[4(4 - 1)] = .33$.

Bonett and Wright (2000) stated: "Testing the null hypothesis that a population correlation is equal to zero is not always interesting" and instead they considered sample size requirements for Spearman, Kendall, and Pearson correlation coefficients such that a Fisher confidence interval on the population correlation is of a desired width.

10.9 SOFTWARE

Software for sample size determination with nonparametric procedures is, unfortunately, far from plentiful. Chakraborti et al. (2006) noted that none of the most popular statistical packages provide any options for sample size calculations for the Mann–Whitney test, which motivated them to provide a routine in *Mathematica* that implements the methods that they proposed. As indicated in Section 10.2, however, PASS, nQuery, and G*Power all have Mann–Whitney sample size determination capability, and nQuery has a Mann–Whitney routine for ordered categories in addition to its routine for continuous data. Whereas both PASS and G*Power have capability for the Wilcoxon one-sample test, nQuery does not have that capability.

It should be noted that although PASS has considerable nonparametric capability, its procedures are not menu items and consequently may be difficult to find. For example, PASS can determine sample size for the McNemar test, as stated in Section 10.5, but the test is listed under "Proportions" rather than under "Nonparametric" in the main menu.

10.10 SUMMARY

Users of software for nonparametric procedures may be frustrated and confused when they discover that it is necessary to specify a distribution when a nonparametric test is considered, since nonparametric procedures are also referred to as distribution-free procedures. Power cannot be determined without specifying a distribution, however. All things considered, a simple rule-of-thumb that inflates the sample size for the corresponding parametric test by 10% or 15% to arrive at the sample size for each nonparametric test would undoubtedly have considerable practical appeal.

REFERENCES

Ahn, C., F. Hu, and W. R. Schucany (2011). Sample size calculation for clustered binary data with sign tests using different weighting schemes. *Statistics in Biopharmaceutical Research*, **3**(1), 65–72.

Al-Sunduqchi, M. S. (1990). *Determining the Appropriate Sample Size for Inferences Based on the Wilcoxon Statistics*. Unpublished Ph.D. dissertation. Department of Statistics, University of Wyoming, Laramie, Wyoming.

Bonett, D. G. and T. A. Wright (2000). Sample size requirements for estimating Pearson, Kendall, and Spearman correlations. *Psychometrika*, **65**(1), 23–25.

Chakraborti, S., B. Hong, and M. A. van de Wiel (2006). A note on sample size determination for a nonparametric test of location. *Technometrics*, **48**, 88–94.

Conover, W. J. (1980). *Practical Nonparametric Statistics*, 2nd edition. New York: Wiley.

Dai, J., L. Li, S. Kim, B. Kimball, S. M. Jazwinski, and J. Arnold (2007). Exact sample size needed to detect dependence in 2 × 2 × 2 tables. *Biometrics*, **63**, 1245–1252.

Divine, G., A. Kapke, S. Havstad, and C. L. M. Joseph (2010). Exemplary data set sample size calculation for Wilcoxon–Mann–Whitney tests. *Statistics in Medicine*, **29**, 108–115.

Fan, C., D. Zhang, and C.-H. Zhang (2011). On sample size of the Kruskal–Wallis test with application to a mouse peritoneal cavity study. *Biometrics*, **67**, 213–224.

Hamilton, M. A. and B. J. Collings (1991). Determining the appropriate sample size for nonparametric tests for location shift. *Technometrics*, **33**, 327–337.

Hilton, J. F. and C. R. Mehta (1993). Power and sample size calculations for exact conditional tests with ordered categorical data. *Biometrics*, **49**, 609–616.

Kolassa, J. (1995). A comparison of size and power calculations for the Wilcoxon statistics for ordered categorical data. *Statistics in Medicine*, **14**, 1577–1581.

Lachenbruch, P. A. (1992). On the sample size for studies based upon McNemar's test. *Statistics in Medicine*, **11**, 1521–1525.

Lachin, J. M. (1992). Power and sample size evaluation for the McNemar test with application to case–control studies. *Statistics in Medicine*, **11**, 1239–1251.

Lehmann, E. L. (1975, 2006). *Nonparametrics: Statistical Methods Based on Ranks*, first edition and revised edition. New York: Springer.

Lu, Y. and J. A. Bean (1995). On the sample size for one-sided equivalence of sensitivities based upon McNemar's test. *Statistics in Medicine*, **14**, 1831–1839.

Lydersen, S., M. W. Fagerland, and P. Laake (2009). Recommended tests for association in 2 × 2 tables. *Statistics in Medicine*, **28**, 1159–1175.

Lydersen, S., V. Pradhan, P. Senchaudhuri, and P. Laake (2007). Choice of test for association in sample unordered $r \times c$ tables. *Statistics in Medicine*, **26**, 4328–4343.

Mahnken, J. D. (2009). Power and sample size calculations for models from unknown distributions. *Statistics in Biopharmaceutical Research*, **1**(3), 328–336.

Mahoney, M. and R. Magel (1996). Estimation of the power of the Kruskal–Wallis test. *Biometrical Journal*, **38**, 613–630.

Mann, H. B. and D. R. Whitney (1947). On a test of whether one or two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, **18**, 50–60.

Mumby, P. J. (2002). Statistical power of nonparametric tests: A quick guide for designing sampling strategies. *Marine Pollution Bulletin*, **44**(1), 85–87.

Noether, G. E. (1987). Sample size determination for some common nonparametric tests. *Journal of the American Statistical Association*, **82**, 645– 647.

Okeh, U. M. (2009). Statistical analysis of the application of Wilcoxon and Mann–Whitney U test in medical research studies. *Biotechnology and Molecular Biology Reviews*, **4**(6), 128–131.

Posten, H. O. (1982). Two-sample Wilcoxon power over the Pearson system and comparison with the t-test. *Journal of Statistical Computation and Simulation,* **16**(1), 1–18.

Qu, Y., Y. D. Zhao, and D. Rahardja (2008). Wilcoxon–Mann–Whitney test: Stratify or not? *Journal of Biopharmaceutical Statistics*, **18**(6), 1103–1111.

Rabbee, N., B. A. Coull, C. Mehta, N. Patel, and P. Senchaudhuri (2003). Power and sample size for ordered categorical data. *Statistical Methods in Medical Research*, **12**(1), 73–84.

Rahardja, D., Y. D. Zhao, and Y. Qu (2009). Sample size determinations for the Wilcoxon–Mann–Whitney test: A comprehensive review. *Statistics in Biopharmaceutical Research*, **1**(3), 317–322.

Rasch, D. and V. Guiard (2004). The robustness of parametric statistical methods. *Psychology Science*, **46**(2), 175–208.

Rasch, D. and M. Šimečková (2007). The size of experiments for the one-way ANOVA for ordered categorical data. MODA 8, June 4-8.

Rosner, B. and R. J. Glynn (2009). Power and sample size estimation for the Wilcoxon rank sum test with application to comparisons of C statistics from alternative prediction models. *Biometrics*, **65**, 188–197.

Rosner, B. and R. J. Glynn (2011). Power and sample size estimation for the clustered Wilcoxon test. *Biometrics*, **67**(2), 646–653.

Rosner, B., R. J. Glynn, and M.-L. T. Lee (2003). Incorporation of clustering effects for the Wilcoxon rank sum test: A large sample approach. *Biometrics*, **59**, 1089–1098.

Rosner, B., R. J. Glynn, and M.-L. T. Lee (2006). The Wilcoxon signed rank test for paired comparisons of clustered data. *Biometrics*, **62**, 185–192.

Schork, M. and G. Williams (1980). Number of observations required for the comparison of two correlated proportions. *Communications in Statistics—Simulation and Computation*, **B9**(4), 349–357.

Shieh, G., S.-L. Jan, and R. H. Randles (2006). On power and sample size determinations for the Wilcoxon–Mann–Whitney test. *Journal of Nonparametric Statistics*, **18**, 33–43.

Shieh, G., S.-L. Jan, and R. H. Randles (2007). Power and sample size determinations for the Wilcoxon signed-rank test. *Journal of Statistical Computation and Simulation*, **77**(8), 717–724.

van Elteren, P. H. (1960). On the combination of independent two sample tests of Wilcoxon. *Bulletin of the Institute of International Statistics*, **37**, 351–361.

Walker, S. G. (2003). How many samples? A Bayesian nonparametric approach. *The Statistician*, **52**(4), 475–482.

Walters, S. J. (2004). Sample size and power estimation for studies with health related quality of life outcomes: A comparison of four methods using the SF-36. *Health and Quality of Life Outcomes*, **2**, 26.

Wang, H., B. Chen and S.-C. Chow (2003). Sample size determination based on rank tests in clinical trials. *Journal of Biopharmaceutical Statistics*, **13**, 735–751.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, **1**, 80–83.

Zhao, Y. D. (2006). Sample size estimation for the van Elteren test—A stratified Wilcoxon–Mann–Whitney test. *Statistics in Medicine*, **25**, 2675–2687.

Zhao, Y. D., D. Rahardja, and Y. Qu (2008a). Sample size calculation for the Wilcoxon–Mann–Whitney test, adjusting for ties. *Statistics in Medicine*, **27**(3), 462–468.

Zhao, Y. D., D. Rahardja, and Y. Mei (2008b). Sample size estimation for the van Elteren test adjusting for ties. *Journal of Biopharmaceutical Statistics*, **18**(6), 1112–1119.

Zhao, Y., Y. Qu, and D. Rahardja (2006). Power approximation for the van Elteren test based on location-scale family of distributions. *Journal of Biopharmaceutical Statistics*, **16**, 803–815.

Zimmerman, D. W. (2011). Power comparisons of significance tests of location, using scores, ranks, and modular ranks. *British Journal of Mathematical and Statistical Psychology*, **64**(2), 233–243.

EXERCISES

10.1. Justify the use of the 15% rule-of-thumb discussed in Section 10.1.

10.2. What is a major problem when sample size software is used to determine sample size for nonparametric tests?