STATISTICS IN MEDICINE
Statist. Med. 2008; **27**:462–468
Published online 9 May 2007 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2912

Sample size calculation for the Wilcoxon–Mann–Whitney test adjusting for ties

Yan D. Zhao[1,*,†], Dewi Rahardja[2] and Yongming Qu[1]

[1]*Eli Lilly and Company, Indianapolis, IN 46285, U.S.A.*
[2]*Fresenius Medical Care—North America, Lexington, MA 02421, U.S.A.*

SUMMARY

In this paper we study sample size calculation methods for the asymptotic Wilcoxon–Mann–Whitney test for data with or without ties. The existing methods are applicable either to data with ties or to data without ties but not to both cases. While the existing methods developed for data without ties perform well, the methods developed for data with ties have limitations in that they are either applicable to proportional odds alternatives or have computational difficulties. We propose a new method which has a closed-form formula and therefore is very easy to calculate. In addition, the new method can be applied to both data with or without ties. Simulations have demonstrated that the new sample size formula performs very well as the corresponding actual powers are close to the nominal powers. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: adjusting for ties; ordinal data; power calculation; sample size calculation; Wilcoxon–Mann–Whitney test

1. INTRODUCTION

The Wilcoxon–Mann–Whitney (WMW) test, also known as the two-sample Wilcoxon rank sum test [1] or the Mann–Whitney test [2], is the most popular nonparametric test for comparing two groups of observations on a continuous or ordered categorical (ordinal) variable when there is no underlying distributional assumption imposed on the data. In this paper we consider sample size calculation for the asymptotic WMW test, particularly for data with ties such as ordinal data.

The existing sample size calculation methods for the WMW test are applicable either to data with ties or to data without ties but not to both cases. The sample size calculations have been satisfactorily studied for data without ties such as continuous data in References [3–5]. However, when the response variable is ordinal, the existing methods have limitations. For example, although Whitehead [6] and Kolassa [7] provided asymptotic methods for calculating sample sizes under

*Correspondence to: Yan D. Zhao, Lilly Corporate Center, DC 6054, Indianapolis, IN 46285, U.S.A.
†E-mail: yzhao@lilly.com

Received 25 September 2006
Accepted 22 March 2007

Copyright © 2007 John Wiley & Sons, Ltd.

proportional odds alternatives, Lee *et al*. [8] showed that their methods can either underestimate or overestimate the sample sizes when the alternatives violate the proportional odds assumption. Hilton and Mehta [9] developed an alternative exact method under any general stochastically ordered alternative hypothesis. However, this method is computationally explosive and so is restricted to studies involving small sample sizes and a small number of ordered categories. To overcome this limitation, the method was later enhanced by Rabbee *et al*. [10]. Nonetheless, the enhanced method used a normal approximation to the sample proportions in a two-way contingency table, which may fail when the sample proportions are close to or equal to zero. Finally, Lesaffre *et al*. [11] proposed a bootstrapping method which is not applicable when no previous pilot study is available.

In this paper we propose a new method with a closed-form formula for calculating sample size for the asymptotic WMW test. This new method can be used in continuous data with or without ties as well as ordinal data where ties are nonignorable. To obtain this formula, we first calculate the means and standard deviations of the WMW test under both the null and the general alternative hypotheses. Then, we calculate the sample size by solving a general sample size formula. The new method is not only easy to compute but also works for any general alternative hypothesis.

The rest of the article is organized as follows. In Section 2 we review the WMW test and develop the new sample size method. In Section 3 we apply the method to a real data example and assess its performance. Some general conclusion is presented in Section 4.

2. THE SAMPLE SIZE FORMULA

2.1. The Wilcoxon–Mann–Whitney test

Suppose we observe two independent random samples $X_1, \ldots, X_m$ from Group A with cumulative distribution function (CDF) F_X and $Y_1, \ldots, Y_n$ from Group B with CDF F_Y. To facilitate discussion and to present the data, we order all the distinct observations from the smallest to the largest into D categories $C_1 < C_2 < \cdots < C_D$ as displayed in Table I. The entries in Table I are presented as the number of observations such as m_1 and the row percentages such as $p_1 = m_1/m$ enclosed in the parentheses. The row totals are m and n for Group A and B, respectively; the column totals are $M_1, \ldots, M_D$ for the D categories; and the total sample size is N.

To quantify the difference between the two groups, we define the competing probability $\pi = \Pr(X > Y) + 0.5 \Pr(X = Y)$, where X and Y are random variables with CDF F_X and F_Y, respectively. Then, the following null hypothesis indicates there is no difference between the two groups,

$$H_0 : \pi = 0.5 \tag{1}$$

Table I. Data displayed as $2 \times$ D contingency table.

Group	C_1	C_2	...	C_D	Total
A	$m_1(p_1)$	$m_2(p_2)$	...	$m_D(p_D)$	m
B	$n_1(q_1)$	$n_2(q_2)$	...	$n_D(q_D)$	n
Total	M_1	M_2	...	M_D	N

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

To test this null hypothesis, the WMW test first computes an unbiased estimator $\hat{\pi}$ of the competing probability π,

$$\hat{\pi} = (mn)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n} \delta(X_i - Y_j) \tag{2}$$

where

$$\delta(t) = \begin{cases} 1 & \text{if } t > 0 \\ 0.5 & \text{if } t = 0 \\ 0 & \text{if } t < 0 \end{cases} \tag{3}$$

Second, an estimate of σ_0^2 $(= V_0[\hat{\pi}])$, the variance of $\hat{\pi}$ under the null hypothesis (1), can be calculated as [3, p. 20]

$$\hat{\sigma}_0^2 = \hat{V}_0[\hat{\pi}] = \frac{N+1}{12mn} - \frac{1}{12N(N-1)mn} \sum_{c=1}^{D} (M_c^3 - M_c) \tag{4}$$

Note that when there are no ties, equation (4) is simply reduced to be the first term $(N+1)/(12mn)$. Finally, the WMW tests the null hypothesis in (1) by constructing a z-statistic:

$$z_0 = \frac{\hat{\pi} - 0.5}{\hat{\sigma}_0} \sim \mathrm{N}(0, 1) \tag{5}$$

There is practical guidance in the literature on how large the numbers of m and n should be in order for the normal approximation in (5) to work. For example, Siegel and Castellan [12] recommended the following: $m = 3$ or 4 and $n > 12$; $m > 4$ and $n > 10$.

2.2. The sample size calculation method

To compute the total sample size N, we assume the treatment fraction $t = n/N$ is known. In addition, we assume the proportions $p_1, \ldots, p_D$ and $q_1, \ldots, q_D$ in Table I are known. The treatment fraction t and the proportions $p_1, \ldots, p_D$ and $q_1, \ldots, q_D$ can be obtained from the literature, from previous studies, or from clinically relevant specifications. Finally, we rewrite all the counts in Table I in terms of the unknown total samples size N, the known treatment fraction t, and the known proportions $p_1, \ldots, p_D$ and $q_1, \ldots, q_D$ as follows for $c = 1, \ldots, D$:

$$\begin{gathered} m = N(1-t), \quad n = Nt, \quad m_c = mp_c = N(1-t)p_c, \quad n_c = nq_c = Ntq_c \\ M_c = m_c + n_c = N[(1-t)p_c + tq_c] \end{gathered} \tag{6}$$

In what follows we develop our new method for calculating the sample size N. For a given type I error α and power $1 - \beta$, the general sample size N for a two-sided test statistic $\hat{\pi}$ that is asymptotically normally distributed can be obtained by solving the following equation:

$$\left(\frac{\mu_1 - \mu_0}{\sigma_0} \right)^2 = \left(Z_{\alpha/2} + \frac{\sigma_1}{\sigma_0} Z_\beta \right)^2 \tag{7}$$

where μ_0, σ_0 and μ_1, σ_1 are the means and standard deviations of $\hat{\pi}$ under the null and the alternative hypotheses, respectively, and $Z_{\alpha/2}$ is the upper $(\alpha/2)$th quantile of the standard normal distribution.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

Although formulas to calculate the quantity σ_1 in equation (7) are available in DeLong *et al.* [13], these formulas can be tedious to compute. In order to simplify equation (7), we assume equal variance:$\sigma_0 = \sigma_1$. This assumption was used by Lehmann [3] and Noether [4] in their development of sample size formulas for the WMW test. In addition, this assumption has been shown in References [14, 15] to work very well in their development of sample size formulas for the van Elteren test—a stratified WMW test. With this equal variance assumption, equation (7) becomes

$$\left(\frac{\mu_1 - \mu_0}{\sigma_0}\right)^2 = (Z_{\alpha/2} + Z_\beta)^2 \tag{8}$$

First, the quantity μ_0 in equation (8) has a simple expression:

$$\mu_0 = E_0[\hat{\pi}] = 0.5 \tag{9}$$

Next, dropping the second term (M_c) from the summand in equation (4) and then combining the resulting equation with equation (6) produce an estimate of σ_0^2

$$\hat{\sigma}_0^2 = \hat{V}_0[\hat{\pi}] \approx \frac{1}{12Nt(1-t)}\left(1 - \sum_{c=1}^{D}((1-t)p_c + tq_c)^3\right) \tag{10}$$

Finally, using the fact that $\mu_1 = E_1[\hat{\pi}] = \Pr(X>Y) + 0.5\Pr(X=Y)$ and equation (6), we obtain the following estimate of μ_1:

$$\begin{aligned}\hat{\mu}_1 = \hat{\pi} &= (mn)^{-1}\sum_{i=1}^{m}\sum_{j=1}^{n}\delta(X_i - Y_j) \\ &= (mn)^{-1}\left(\sum_{c=2}^{D} m_c \sum_{d=1}^{c-1} n_d\right) + 0.5(mn)^{-1}\sum_{c=1}^{D} m_c n_c = \sum_{c=2}^{D} p_c \sum_{d=1}^{c-1} q_d + 0.5\sum_{c=1}^{D} p_c q_c\end{aligned} \tag{11}$$

Combining equations (8)–(11), we obtain the sample size formula for the WMW test:

$$N = \frac{(Z_{\alpha/2} + Z_\beta)^2\left(1 - \sum_{c=1}^{D}((1-t)p_c + tq_c)^3\right)}{12t(1-t)\left(\sum_{c=2}^{D} p_c \sum_{d=1}^{c-1} q_d + 0.5\sum_{c=1}^{D} p_c q_c - 0.5\right)^2} \tag{12}$$

We have the following observation on equation (12). First, for ordinal data, our method has advantage over existing methods because they use normal approximations to the sample multinomial distributions and we do not use such approximations. Consequently, our method can be used even when there are zero counts in Table I. Second, for continuous data that do not have ties, the number of categories D is equal to the total sample size N, the counts m_c and n_c in Table I become either 0 or 1, and M_c becomes 1 for $c = 1, \ldots, D$. As a result, equation (10) is simplified to be $\hat{\sigma}_0^2 = 1/[12Nt(1-t)]$ and the sample size formula (12) becomes

$$N = \frac{(Z_{\alpha/2} + Z_\beta)^2}{12t(1-t)(\hat{\mu}_1 - 0.5)^2} = \frac{(Z_{\alpha/2} + Z_\beta)^2}{12t(1-t)\left(\sum_{c=2}^{D} p_c \sum_{d=1}^{c-1} q_d + 0.5\sum_{c=1}^{D} p_c q_c - 0.5\right)^2} \tag{13}$$

The sample size formula (13) suggests that for continuous data without ties, one can either (1) compute N by using an intelligent guess of $\hat{\mu}_1$ and the first part of (13), even when Table I is not

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

available, or (2) compute N by the second part of (13) when Table I is available and most likely constructed from existing data. The sample size formula (13) is very similar to equation (4) in Wang *et al*. [5] except that they did not assume equal variance $\sigma_0 = \sigma_1$. In addition, formula (13) is exactly the same as equation (28) in Zhao [14] when the number of strata is one. Because Zhao [14] has demonstrated using simulations that formula (13) performed very well for continuous data, we conclude that our method will also work for continuous data without ties. Finally, the sample size formula (12) is also applicable to data with a mixture of continuous and ordinal data such as data with probability density function (PDF)

$$f(x) = 0.25 \times \mathrm{U}(0, 1) + 0.25 \times I\{x = 2\} + 0.25 \times I\{x = 3\} + 0.25 \times \mathrm{U}(4, 5)$$

where $\mathrm{U}(a, b)$ is the PDF of the uniform distribution in interval (a, b) and I is the indicator function. In this case one needs to obtain Table I from existing data in order to compute N using formula (12).

3. A REAL DATA EXAMPLE

In this section we apply our sample size calculation method to a real data example with ordinal data where ties are common and nonignorable. Consider a 6-year follow-up study by Bender and Grouven [16] where 613 type-1 diabetic patients were studied for associations between smoking and retinopathy status. Retinopathy status is measured by a response variable with three categories: no retinopathy (0); non-proliferative retinopathy (1); and advanced retinopathy or blind (2). Smoking status is a binary variable which equals to 1 if the patient smoked during the study and 0 otherwise.

Bender and Grouver [16] assessed the association between retinopathy and smoking controlling for confounding factors such as diabetes durations, glycosylated hemoglobin, and diastolic blood pressure. However, for the sake of illustrating the sample size calculation, here we only focus on the effects of smoking on the retinopathy status without controlling for the confounding factors. The observed counts and proportions in the 2×3 table of smoking status by retinopathy status are displayed in Table II.

The following are some analysis results about this data. The WMW test on these data is not significant ($p > 0.3$). In addition, the competing probability π that a non-smoker has a better retinopathy status than a smoker is estimated as $\hat{\pi} = 0.515$ according to equation (2). The 95 per cent confidence interval for π is $(0.478, 0.557)$, which is calculated according to the method developed in DeLong *et al.* [13]. Finally, Bender and Grouven [16] noted that the proportional odds assumption is inappropriate for this data because a goodness-of-fit test produced a p-value of 0.017. Therefore, sample size methods based on proportional odds assumption should not be used here.

Table II. Bender: retinopathy category by smoking status for 613 diabetic patients.

	Retinopathy status			
Group	None	Non-proliferative	Advanced	Total
Non-smoking	191 (66%)	42 (15%)	55 (19%)	288 (100%)
Smoking	197 (61%)	76 (23%)	52 (16%)	325 (100%)
Total	388	118	107	613

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

Table III. Total sample size N based on nominal 80 per cent power and the corresponding actual power for various alternatives in the retinopathy study.

			$t = 0.53$		$t = 0.95$	
Non-smokers	Smokers	π	Sample size	Actual power	Sample size	Actual power
(0.66, 0.15, 0.19)	(0.61, 0.23, 0.16)	0.515	8390	0.795	45 264	0.796
(0.66, 0.15, 0.19)	(0.61, 0.19, 0.20)	0.522	3997	0.798	21 597	0.802
(0.66, 0.15, 0.19)	(0.61, 0.14, 0.25)	0.530	2073	0.802	11 174	0.814
(0.66, 0.15, 0.19)	(0.58, 0.23, 0.19)	0.532	1878	0.806	10 264	0.803
(0.66, 0.15, 0.19)	(0.58, 0.20, 0.22)	0.538	1401	0.799	7665	0.807
(0.66, 0.15, 0.19)	(0.58, 0.15, 0.27)	0.546	929	0.803	5067	0.818
(0.66, 0.15, 0.19)	(0.55, 0.23, 0.22)	0.550	817	0.796	4506	0.809
(0.66, 0.15, 0.19)	(0.55, 0.20, 0.25)	0.555	671	0.803	3702	0.815
(0.66, 0.15, 0.19)	(0.55, 0.15, 0.30)	0.563	502	0.808	2753	0.822
(0.66, 0.15, 0.19)	(0.55, 0.00, 0.45)	0.589	249	0.805	1303	0.847
(0.66, 0.15, 0.19)	(0.45, 0.00, 0.55)	0.646	96	0.811	484	0.844
(0.66, 0.15, 0.19)	(0.40, 0.00, 0.60)	0.675	68	0.817	331	0.857

$\pi = \Pr(\text{smoker} > \text{non-smoker}) + 0.5 \times \Pr(\text{smoker} = \text{non-smoker})$.

Table III presents the sample sizes computed by the sample size formula (12) and the corresponding actual powers computed *via* simulations. In the calculations, we fix the type I error level α at 0.05 and the power $(1 - \beta)$ at 80 per cent. We also fix the response proportions in non-smokers at observed proportions (0.66, 0.15, 0.19). For the response proportions in smokers, we consider the observed proportions (0.61, 0.23, 0.16) as well as other choices. Note that the last three sets of proportions for smokers in Table III includes a zero cell which cannot be handled by existing sample size calculation methods for the WMW test. We consider two values for the treatment fraction (t). The first value for t is chosen as 53 per cent (325/613) which is from the data, and the second value for t is chosen as 0.95 to evaluate the robustness of the sample size calculation method with respect to t. The actual powers are computed based on 10 000 Monte Carlo simulations.

When $t = 0.53$, Table III shows that all the actual powers are close to the nominal 80 per cent, even when the total sample size falls below 100. When $t = 0.95$ which is close to 1 and rare in practice, the actual powers are still close to the nominal 80 per cent for relatively large sample sizes; however, the actual powers can be slightly greater than the nominal when sample sizes get smaller.

4. CONCLUSION

In this paper we have proposed a sample size calculation method for the asymptotic WMW test adjusting for ties. Our method is applicable to continuous response variables with or without ties as well as discrete ordinal response variables. In addition, our sample size formula is a closed-form function which is very easy to compute. Simulations have demonstrated that our sample size formula produces a sample size that has a power very close to the nominal power when the treatment fraction t is around 0.5 which is common in practice. However, the sample size formula should be used with caution when in rare cases where the treatment fraction t is close to 0 or 1 and the sample size is relatively small.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim

ACKNOWLEDGEMENTS

The authors thank the editors and the referees for their constructive comments and suggestions which have led to a much improved presentation of this paper.

REFERENCES

1. Wilcoxon F. Individual comparisons by ranking methods. *Biometrics* 1945; **1**:80–83.
2. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 1947; **18**:50–60.
3. Lehmann EL. *Nonparametrics*: *Statistical Methods based on Ranks*. Holden-Day: San Francisco, 1975.
4. Noether GE. Sample size determination for some common nonparametric tests. *Journal of the American Statistical Association* 1987; **82**(398):645–647.
5. Wang H, Chen B, Chow S-C. Sample size determination based on rank tests in clinical trials. *Journal of Biopharmaceutical Statistics* 2003; **13**(4):735–751.
6. Whitehead J. Sample size calculation for ordered categorical data. *Statistics in Medicine* 1993; **12**:2257–2271.
7. Kolassa J. A comparison of size and power calculations for the Wilcoxon statistics for ordered categorical data. *Statistics in Medicine* 1995; **14**:1577–1581.
8. Lee M-K, Song H-H, Kang S-H, Ahn C-W. The determination of sample sizes in the comparison of two multinomial proportions from ordered categories. *Biometrical Journal* 2002; **44**(4):395–409.
9. Hilton JF, Mehta CR. Power and sample size calculations for exact conditional tests with ordered categorical data. *Biometrics* 1993; **49**:609–616.
10. Rabbee N, Coull BA, Mehta C. Power and sample size for ordered categorical data. *Statistical Methods in Medical Research* 2003; **12**:73–84.
11. Lesaffre E, Scheys I, Frohlich J, Bluhmki E. Calculation of power and sample size with bounded outcome scores. *Statistics in Medicine* 1993; **12**:1063–1078.
12. Siegel S, Castellan NJ. *Nonparametric Statistics for the Behavioural Sciences*. McGraw-Hill: New York, 1988.
13. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; **44**(3):837–845.
14. Zhao YD. Sample size estimation for the van Elteren test—a stratified Wilcoxon–Mann–Whitney test. *Statistics in Medicine* 2006; **25**:2675–2687.
15. Zhao YD, Qu Y, Rahardja D. Power approximation for the van Elteren test based on location-scale family of distributions. *Journal of Biopharmaceutical Statistics* 2006; **16**:805–813.
16. Bender R, Grouven U. Using binary logistic regression models for ordinal data with non-proportional odds. *Journal of Clinical Epidemiology* 1998; **51**:809–816.

Copyright © 2007 John Wiley & Sons, Ltd.

DOI: 10.1002/sim