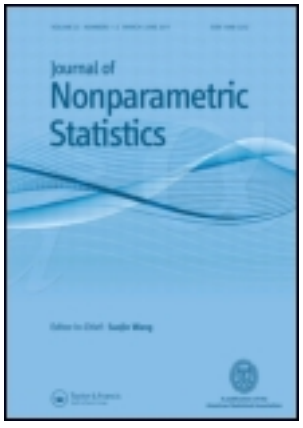

Journal of Nonparametric Statistics
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gnst20

On power and sample size determinations for the Wilcoxon-Mann-Whitney test

Gwowen Shieh [a], Show-li Jan [b] & Ronald H. Randles [c]

[a] Department of Management Science, National Chiao Tung University, Hsinchu, Taiwan, 30050, R.O.C.

[b] Department of Applied Mathematics, Chung Yuan Christian University, Chungli, Taiwan, 32023, R.O.C.

[c] Department of Statistics, University of Florida, Gainesville, FL, 32611-8545, USA

Published online: 25 Apr 2007.

To cite this article: Gwowen Shieh, Show-li Jan & Ronald H. Randles (2006) On power and sample size determinations for the Wilcoxon-Mann-Whitney test, Journal of Nonparametric Statistics, 18:1, 33-43, DOI: 10.1080/10485250500473099

To link to this article: http://dx.doi.org/10.1080/10485250500473099

Nonparametric Statistics
Vol. 18, No. 1, January 2006, 33–43

Taylor & Francis
Taylor & Francis Group

On power and sample size determinations for the Wilcoxon–Mann–Whitney test

GWOWEN SHIEH†, SHOW-LI JAN‡ and RONALD H. RANDLES*§

†Department of Management Science, National Chiao Tung University, Hsinchu, Taiwan 30050, R.O.C.
‡Department of Applied Mathematics, Chung Yuan Christian University, Chungli, Taiwan 32023, R.O.C.
§Department of Statistics, University of Florida, Gainesville, FL 32611-8545, USA

(Received June 2004; in final form November 2005)

This article considers power and sample size calculations for the Wilcoxon–Mann–Whitney test. An exact variance large-sample method is introduced and explicit formulas are derived for the uniform, normal, double exponential and exponential shift models. A Monte Carlo simulation study is conducted to evaluate the exact variance procedure and compare its performance with some other large sample methods. The results show the remarkable accuracy of the suggested formula and the fundamental limitations of the other approximations. Practical considerations in determination of sample size for the two-sample location problem are also discussed.

Keywords: Large-sample approximation; Nonparametric method; Two-sample location problem

1. Introduction

In the two-sample location problem, the Wilcoxon–Mann–Whitney test is a strong competitor to the two-sample t-test because it has good Pitman efficiencies and makes minimal assumptions about the underlying populations. For example, the Wilcoxon–Mann–Whitney test has a null distribution that does not depend on the common distribution of the error terms, the only requirement being that the distribution is continuous. See ref. [1, section 2.4] for a general comparison of the Wilcoxon–Mann–Whitney test with the classical two-sample t-test.

To evaluate the power of the Wilcoxon–Mann–Whitney test, it is necessary to specify the alternatives to the null hypothesis and to derive the corresponding exact distribution of the possible rankings. This tends to be difficult and only a few such computations have been performed. For example, Milton [2], Haynam and Govindarajulu [3], and Ramsey [4] obtained the power of the Wilcoxon–Mann–Whitney test with normal, uniform, exponential and double exponential shift alternatives for rather limited selections of sample size. In view of the difficulty of computing the probabilities of different rankings for most alternatives of interest, one

*Corresponding author. Email: rrandles@stat.ufl.edu

Nonparametric Statistics
ISSN 1048-5252 print/ISSN 1029-0311 online
http://www.tandf.co.uk/journals
DOI: 10.1080/10485250500473099

may wish to use good, simple approximations. It appears that useful results can be obtained from asymptotic normal approximations to the exact distribution of the Wilcoxon–Mann–Whitney statistic. Lehmann [1, section 2.3] and Noether [5] proposed some simplifications to the large sample distribution to avoid the calculations of certain quantities in the asymptotic power function. It is important to note that these two simplifications are only justified when the location shift is small. However, there is no simple guideline that indicates when the location shift is sufficiently small, so that the results are valid. Specifically, Lehmann's [1] power approximation was studied only for normal shift alternatives with one selected set of sample sizes. Essentially, little is known about the performance of the approximation of Lehmann [1] for different sample size settings and non-normal distributions. On the other hand, the accuracy of the modified normal approximation by Noether [5] was studied by Vollandt and Horn [6]. However, their evaluation was made by comparing Noether's approximation with an alternative procedure based on the upper bound for the variance of the Wilcoxon–Mann–Whitney statistic. They concluded that Noether's method is sufficiently reliable with small, medium and large deviations from the null hypothesis. These conclusions may be questionable because Noether's method is only justified for sufficiently small location shift alternatives and may become quite unreliable as the location shift increases. Consequently, comparative analysis of Noether's method and an exact variance large-sample approach can reveal unique types of information impossible to obtain from comparing Noether's procedure with other approximations. More importantly, no research to date has compared the two approximations of Lehmann [1] and Noether [5] together with an exact variance large-sample method on common ground. For a more complete comparison, it is of interest to consider two alternative approaches based on the general lower and upper bounds for the variance of the Wilcoxon–Mann–Whitney statistic given by Birnbaum and Klose [7]. It turns out that these formulas yield markedly different results.

The objectives of this article are to provide a clear understanding and demonstration of various competing approaches to power and sample size determinations for the Wilcoxon–Mann–Whitney test and to offer useful and well-supported recommendations on the most reliable approach for use by researchers. In the process, we also hope to account for some inconsistent findings in the literature. We investigate the exact variance formula for power and sample size calculations based on the asymptotic normal distribution of the Wilcoxon–Mann–Whitney statistic. Accordingly, closed form expressions are provided for the prominent cases studied in the literature: uniform, normal, double exponential and exponential distributions. Numerical analysis is performed to assess the accuracy of the exact variance large-sample method and to explore the impact of various modifications of the suggested asymptotic formula for several location shifts and sample allocation schemes under the four prescribed distributions.

The rest of this paper is organized as follows. Section 2 describes the important details of power and sample size considerations for the Wilcoxon–Mann–Whitney test. In section 3, the estimated sample sizes are provided and a Monte Carlo simulation study is reported that compares the finite-sample performance of both exact variance and approximate methods. Finally, the important implications of the results are summarized in the last section.

2. Power and sample size determinations

Let $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ be two independent samples from continuous cumulative distribution functions F and G, respectively. We restrict our attention to the simple but important location shift situation where F and G are of the same shape and $G(x) = F(x - \theta)$

for all x and $\theta \geq 0$. We wish to test the hypothesis that the two samples have come from the same population against the alternative that G is stochastically larger than F. That is, we wish to test H_0: $\theta = 0$ versus H_1: $\theta > 0$. The Mann–Whitney form of the Wilcoxon–Mann–Whitney statistic is defined as follows.

$$W = \sum_{i=1}^{m} \sum_{j=1}^{n} \varphi(Y_j - X_i),$$

where $\varphi(Y_j - X_i) = 1$ if $Y_j - X_i > 0$ and 0 otherwise. Obviously, the formulation requires the computation of mn differences. Although the exact null distribution of W is available, the asymptotic null distribution is commonly used to provide critical values. Assume m/N approaches c as the sample sizes m and n tend to infinity, where $N = m + n$ and $0 < c < 1$. It follows that $(W - \mu_0)/\sigma_0$ has a limiting standard normal distribution under H_0: $\theta = 0$ where

$$\mu_0 = \frac{mn}{2} \quad \text{and} \quad \sigma_0^2 = \frac{mn(N+1)}{12}. \tag{1}$$

See ref. [8, Theorem 3.2.4] for a detailed proof. Suppose that the sample sizes are large enough so that the critical value can be determined from the standard normal approximation rather than the exact null distribution. Then, the test is carried out by rejecting H_0: $\theta = 0$ if the standardized value $(W - \mu_0)/\sigma_0$ is greater than z_α, where α is the specified significance level and z_α is the $100(1-\alpha)$th percentile of the standard normal distribution.

For the purpose of power calculation, it is crucial to derive the distribution of the ranks when F is continuous and $\theta > 0$. This is considerably more complicated than the exact distribution of W for the case of $\theta = 0$ (i.e. $F = G$). However, useful results can be obtained from the normal approximation to the power. The approximation corresponds to the fact that the distribution of $(W - \mu)/\sigma$ tends to the standard normal distribution as m and n tend to infinity [1, section 2.3] where

$$\mu = mnp_1, \quad \sigma^2 = mnp_1(1-p_1) + mn(n-1)(p_2 - p_1^2) + mn(m-1)(p_3 - p_1^2), \tag{2}$$

$p_1 = P(X_1 < Y_1) = \int F \, \mathrm{d}G$, $p_2 = P(X_1 < Y_1 \cap X_1 < Y_2) = \int (1-G)^2 \, \mathrm{d}F$ and $p_3 = P(X_1 < Y_1 \cap X_2 < Y_1) = \int F^2 \, \mathrm{d}G$. When the distributions are symmetric, it can be shown that $p_2 = p_3$. Hence, if the hypothesis $H_0 : \theta = 0$ is rejected with specified significance level α, then the power of the test against any fixed alternative $\theta > 0$ is approximated by the probability

$$P\{W > \mu_0 + z_\alpha \sigma_0\} \doteq \Phi\left(\frac{\mu - \mu_0 - z_\alpha \sigma_0}{\sigma}\right), \tag{3}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. Application of formula 3 involves the values of p_1, p_2 and p_3. In principle, the approximate power can be computed for any alternative. However, there are some important cases typically of great interest. For illustrative purposes, explicit analytical forms are presented for $G(x) = F(x - \theta)$ where $F(x)$ has the following continuous distributions.

(i) Uniform $(-1/2, 1/2)$:

$$p_1 = \frac{1}{2} + \theta\left(1 - \frac{\theta}{2}\right) \quad \text{and} \quad p_2 = p_3 = \frac{1}{3} + \theta - \frac{\theta^3}{3} \ \text{ for } \theta \leq 1.$$

(ii) Standard normal $N(0, 1)$:

$$p_1 = \Phi\left(\frac{\theta}{\sqrt{2}}\right) \quad \text{and} \quad p_2 = p_3 = E[\{\Phi(Z + \theta)\}^2], \ \text{where } Z \sim N(0,1).$$

(iii) Double exponential (0, 1):

$$p_1 = 1 - \frac{1}{2}\left(1 + \frac{\theta}{2}\right)\mathrm{e}^{-\theta} \quad \text{and} \quad p_2 = p_3 = 1 - \left(\frac{7}{12} + \frac{\theta}{2}\right)\mathrm{e}^{-\theta} - \frac{1}{12}\mathrm{e}^{-2\theta}.$$

(iv) Exponential (1):

$$p_1 = 1 - \frac{1}{2}\mathrm{e}^{-\theta}, \quad p_2 = 1 - \frac{2}{3}\mathrm{e}^{-\theta} \quad \text{and} \quad p_3 = 1 - \mathrm{e}^{-\theta} + \frac{1}{3}\mathrm{e}^{-2\theta}.$$

These expressions are employed to examine the accuracy of the exact variance method in the subsequent section.

Note that the computations of p_2 and p_3 are typically more involved than that of p_1. These can be avoided by the following two alternative approximations by Lehmann [1] and Noether [5]. They are motivated by the fact that when θ is sufficiently close to zero, the asymptotic variance σ^2 of W can be approximated well by the null variance σ_0^2. Consequently, application of the two methods is confined to the models with small location shifts. The modification of the power function by Lehmann [1, equation 2.29] is

$$P\{W > \mu_0 + z_\alpha \sigma_0\} \doteq \Phi\left\{\sqrt{\frac{12mn}{N+1}}\theta f^*(0) - z_\alpha\right\}, \tag{4}$$

where $f^*(0)$ is the density of F^* evaluated at zero and F^* denotes the distribution of the difference of two independent variables, each with distribution F. It can be shown that $f^*(0) = E[f(X_1)]$. Specifically, $f^*(0) = 1, 1/(2\sqrt{\pi}), 1/4$ and $1/2$ for uniform $(-1/2, 1/2)$, standard normal, double exponential $(0, 1)$ and exponential (1) distributions, respectively. Noether [5] considered an alternative approximation:

$$P\{W > \mu_0 + z_\alpha \sigma_0\} \doteq \Phi\left\{\sqrt{\frac{12mn}{N}}\left(p_1 - \frac{1}{2}\right) - z_\alpha\right\}. \tag{5}$$

Along the same line of providing simplified power approximation, the lower and upper bounds for the variance of the Wilcoxon–Mann–Whitney statistic furnish the basis for two other potential approaches. It was shown in Birnbaum and Klose [7] that $\sigma_{\mathrm{L}}^2 \le \sigma^2 \le \sigma_{\mathrm{U}}^2$, where

$$\sigma_L^2 = mn\left\{\frac{[N + 1 + 2\sqrt{(m-1)(m-n)(2p_1-1)^3}]}{3} - [mp_1^2 + n(1-p_1)^2 + p_1(1-p_1)]\right\}$$

$$\text{if } \frac{n-1}{m-1} \le 2(1-p_1),$$

$$= mn\left\{\left[\frac{4(1-p_1)}{3}\right]\sqrt{2(m-1)(n-1)(1-p_1)} - (m+n-2)(1-p_1)^2 + p_1(1-p_1)\right\}$$

$$\text{if } 2(1-p_1) < \frac{n-1}{m-1} \le \frac{1}{2(1-p_1)},$$

$$= mn\left\{\frac{[N + 1 + 2\sqrt{(n-1)(n-m)(2p_1-1)^3}]}{3} - [m(1-p_1)^2 + np_1^2 + p_1(1-p_1)]\right\}$$

$$\text{if } \frac{1}{2(1-p_1)} < \frac{n-1}{m-1},$$

and

$$\sigma_{\mathrm{U}}^2 = mn\left\{v\left[\frac{k}{3} - (1-p_1)^2\right] + u\left[\frac{-2k}{3} + 1 - p_1^2\right] + \frac{k}{3} - p_1(1-p_1)\right\},$$

with $u = \min(m, n)$, $v = \max(m, n)$ and $k = 1 - (2p_1 - 1)^{3/2}$. Therefore, the corresponding power approximations based on the general lower and upper bounds are

$$P\{W > \mu_0 + z_\alpha\sigma_0\} \doteq \Phi\left(\frac{\mu - \mu_0 - z_\alpha\sigma_0}{\sigma_{\mathrm{L}}}\right) \tag{6}$$

and

$$P\{W > \mu_0 + z_\alpha\sigma_0\} \doteq \Phi\left(\frac{\mu - \mu_0 - z_\alpha\sigma_0}{\sigma_{\mathrm{U}}}\right) \tag{7}$$

respectively. It is easily seen that the application of formulas (4)–(7) circumvent the evaluation of p_2 and p_3 and only necessitate specifying the quantity $f^*(0)$ and/or the probability p_1. The impact of these simplifications and the accuracy of the exact variance large-sample method (3) are investigated in the subsequent numerical examinations.

There are also many situations when it is desirable to determine the sample size needed to adequately achieve a given power. All the aforementioned power formulas (3)–(7) can be modified to derive sample size estimates required for testing a specified alternative hypothesis with significance level α and power $1 - \beta$. To be specific, the sample size estimate based on the exact variance method (3) is the minimum N which satisfies $(\mu - \mu_0) \geq (z_\alpha\sigma_0 + z_\beta\sigma)$ and $m/N \doteq c$. Similarly, one can perform the sample size calculations for the other four formulas (4)–(7).

3. Numerical study

To illustrate the application of various competing approaches to sample size and power determinations presented in the preceding section, we begin by considering sample size requirements to detect a specified location shift for a given power at the significance level $\alpha = 0.05$. Then we examine their accuracy for power computations through a Monte Carlo simulation study.

Table 1. Sample sizes required to attain power levels 0.90 and 0.95 for uniform shift alternatives.

	0.90				0.95			
Power $\Delta = \theta/\sigma$	0.3	0.5	1.0	1.5	0.3	0.5	1.0	1.5
Equal group sizes ($m = n$)								
Exact variance	410	154	42	20	516	192	52	24
Lehmann	382	140	36	18	482	176	46	22
Noether	416	160	48	26	526	202	60	32
Lower bound	410	154	42	20	516	192	52	24
Upper bound	428	166	48	24	542	210	60	28
Average bound	418	160	44	22	530	202	56	26
Unequal group sizes ($m = n/3$ and $m = 3n$)								
Exact variance	548	204	56	28	688	256	68	32
Lehmann	512	184	48	24	644	232	60	28
Noether	556	216	64	36	704	272	80	44
Lower bound	496	180	52	24	612	224	60	28
Upper bound	640	256	76	36	820	328	96	48
Average bound	568	220	64	32	720	280	80	40

Table 2. Sample sizes required to attain power levels 0.90 and 0.95 for normal shift alternatives.

	0.90				0.95			
Power $\Delta = \theta/\sigma$	0.3	0.5	1.0	1.5	0.3	0.5	1.0	1.5
Equal group sizes ($m = n$)								
Exact variance	400	146	38	18	504	182	46	22
Lehmann	400	146	38	18	506	184	48	22
Noether	406	150	44	24	512	190	54	30
Lower bound	400	144	38	18	502	180	46	20
Upper bound	416	156	42	20	526	196	54	26
Average bound	408	150	40	20	514	188	50	24
Unequal group sizes ($m = n/3$ and $m = 3n$)								
Exact variance	532	192	52	24	672	244	60	28
Lehmann	536	196	52	24	676	244	64	28
Noether	540	200	60	32	684	252	72	40
Lower bound	480	172	44	24	596	208	52	24
Upper bound	620	240	68	32	800	308	88	40
Average bound	552	208	56	28	700	260	72	32

First, we summarize the estimated sample sizes corresponding to both equal and unequal group sizes ($m = n$, $m = n/3$ and $m = 3n$) for uniform, normal, double exponential and exponential shift alternatives in tables 1–4, respectively. The power levels are set at 0.90 and 0.95 and a total of four selected values of location shift in terms of $\Delta = \theta/\sigma$ are considered: 0.3, 0.5, 1.0 and 1.5. For example, when $\Delta = 0.3$, the distribution is uniform, and group sizes are balanced, the exact variance method gives the sample size estimate 410 in order to achieve power level 0.90. In the same situation, the sample size estimates of the other four approximations defined in (4)–(7), denoted by Lehmann, Noether, lower bound and upper bound, are 382, 416, 410 and 428, respectively. As expected, the results in the tables reveal the general relation that sample sizes increase and decrease with increasing power and shift, respectively. Comparatively, the equal-group-sizes design yields a smaller total of the sample sizes than the other two unbalanced designs. This shows that a balanced design is the optimal sample allocation scheme to maximize the overall power. Analytically, it can be shown that $\sigma_L^2 = \sigma^2$ for uniform distribution with equal group sizes ($m = n$). Hence, in this case the exact variance

Table 3. Sample sizes required to attain power levels 0.90 and 0.95 for double exponential shift alternatives.

	0.90				0.95			
Power $\Delta = \theta/\sigma$	0.3	0.5	1.0	1.5	0.3	0.5	1.0	1.5
Equal group sizes ($m = n$)								
Exact variance	264	100	30	16	332	124	36	20
Lehmann	256	94	24	12	322	118	30	14
Noether	268	104	34	22	338	132	44	26
Lower bound	262	98	28	16	328	122	34	18
Upper bound	276	108	34	18	348	136	40	22
Average bound	268	102	30	16	338	128	38	20
Unequal group sizes ($m = n/3$ and $m = 3n$)								
Exact variance	352	132	40	20	444	168	48	24
Lehmann	340	124	32	16	432	156	40	20
Noether	356	140	48	28	452	176	60	36
Lower bound	312	116	36	20	384	140	40	24
Upper bound	416	168	52	28	536	216	68	36
Average bound	368	144	44	24	464	180	56	28

Table 4. Sample sizes required to attain power levels 0.90 and 0.95 for exponential shift alternatives.

	0.90				0.95			
Power $\Delta = \theta/\sigma$	0.3	0.5	1.0	1.5	0.3	0.5	1.0	1.5
Equal group sizes ($m = n$)								
Exact variance	166	70	26	16	210	88	30	18
Lehmann	128	48	14	6	162	60	16	8
Noether	170	74	30	20	216	94	38	24
Lower bound	164	68	24	14	206	86	28	16
Upper bound	176	76	28	16	224	96	34	20
Average bound	170	72	26	16	214	90	32	18
Unequal group sizes ($m = n/3$)								
Exact variance	240	104	40	24	308	132	48	28
Lehmann	172	64	20	8	216	80	24	12
Noether	228	100	40	28	288	128	52	32
Lower bound	196	80	28	20	240	100	36	20
Upper bound	272	120	44	24	348	156	56	32
Average bound	236	100	36	24	296	128	44	28
Unequal group sizes ($m = 3n$)								
Exact variance	204	84	28	20	252	100	36	20
Lehmann	172	64	20	8	216	80	24	12
Noether	228	100	40	28	288	128	52	32
Lower bound	196	80	28	20	240	100	36	20
Upper bound	272	120	44	24	348	156	56	32
Average bound	236	100	36	24	296	128	44	28

method and the lower bound method give the identical sample sizes as shown in table 1. It is clear that the exact variance method gives identical sample size estimates for the two unbalanced designs with respective sample ratio $m/n = 1/3$ and 3 under the symmetric shift alternatives in tables 1–3. This is due to the fact that $p_2 = p_3$, provided the distribution is symmetric. In addition, all four approximate formulas give the same sample sizes for the two unbalanced designs throughout the four different distributions. Such phenomena shall continue to exist between unbalanced designs with reverse sample ratios in other continuous settings. It is worthwhile to note that the ordering of sample size estimates is consistently upper bound > Noether > exact variance method. The methods of Lehmann and Lower bound tend to produce small sample sizes without a conclusive order between the two approaches. Lehmann's approximation may seriously underestimate the sample size and can therefore be misleading in some cases. As suggested by a referee, we also evaluate the 'average bound' approach that employs the simple mean $(\sigma_L^2 + \sigma_U^2)/2$ for the exact variance σ^2. It appears that the resulting sample sizes are not accurate as expected.

In order to identify the most reliable method, we proceed to evaluate the performance of these formulas in terms of the discrepancy between their nominal power and the estimated actual power, where they all use the same sample size. The sample sizes of the exact variance method presented in tables 1–4 at the power value of 0.90 are utilized as the basis in the simulation study. Accordingly, the nominal (approximate) powers are computed for all six approaches and the results are presented in tables 5–8 for the four shift models, respectively. The nominal values of the exact variance method are slightly larger than the pre-specified nominal level 0.90, whereas those associated with the other approaches vary around 0.90. Estimates of actual power for the given sample sizes and model configurations are then computed through Monte Carlo simulation using 10,000 replicate data sets. For each replicate, m and n observations are generated from the selected distributions F and G, respectively. Then the standardized Wilcoxon–Mann–Whitney test statistic $(W - \mu_0)/\sigma_0$ is computed and the

Table 5. Nominal power and simulated power at specified sample size for uniform shift alternatives.

$\Delta = \theta/\sigma$		0.3		0.5		1.0		1.5
Sample sizes (m, n)		(205, 205)		(77, 77)		(21,21)		(10, 10)
Simulated power		0.8985		0.8995		0.9032		0.9086
Nominal power (percentage error)								
Exact variance	0.9003	(0.20)	0.9018	(0.25)	0.9084	(0.57)	0.9181	(1.05)
Lehmann	0.9175	(2.12)	0.9261	(2.96)	0.9403	(4.11)	0.9483	(4.37)
Noether	0.8963	(−0.24)	0.8913	(−0.91)	0.8703	(−3.64)	0.8372	(−7.86)
Lower bound	0.9003	(0.20)	0.9018	(0.25)	0.9084	(0.57)	0.9181	(1.05)
Upper bound	0.8899	(−0.96)	0.8829	(−1.85)	0.8697	(−3.71)	0.8627	(−5.05)
Average bound	0.8950	(−0.39)	0.8920	(−0.83)	0.8878	(−1.70)	0.8879	(−2.27)
Sample sizes (m, n)		(137, 411)		(51, 153)		(14, 42)		(7, 21)
Simulated power		0.9000		0.8967		0.9020		0.9262
Nominal power (percentage error)								
Exact variance	0.9011	(0.12)	0.9005	(0.42)	0.9102	(0.91)	0.9362	(1.08)
Lehmann	0.9182	(2.03)	0.9251	(3.16)	0.9414	(4.37)	0.9584	(3.48)
Noether	0.8970	(−0.34)	0.8896	(−0.79)	0.8703	(−3.51)	0.8527	(−7.94)
Lower bound	0.9273	(3.03)	0.9329	(4.03)	0.9428	(4.52)	0.9596	(3.61)
Upper bound	0.8616	(−4.27)	0.8443	(−5.84)	0.8258	(−8.44)	0.8343	(−9.93)
Average bound	0.8910	(−0.99)	0.8822	(−1.62)	0.8729	(−3.23)	0.8850	(−4.45)
Sample sizes (m, n)		(411, 137)		(153,51)		(42, 14)		(21, 7)
Simulated power		0.8998		0.9005		0.9075		0.9252
Nominal power (percentage error)								
Exact variance	0.9011	(0.14)	0.9005	(0.00)	0.9102	(0.30)	0.9362	(1.19)
Lehmann	0.9182	(2.05)	0.9251	(2.73)	0.9414	(3.74)	0.9584	(3.59)
Noether	0.8970	(−0.32)	0.8896	(−1.21)	0.8703	(−4.10)	0.8527	(−7.84)
Lower bound	0.9273	(3.05)	0.9329	(3.59)	0.9428	(3.89)	0.9596	(3.72)
Upper bound	0.8616	(−4.25)	0.8443	(−6.24)	0.8258	(−9.00)	0.8343	(−9.83)
Average bound	0.8910	(−0.97)	0.8822	(−2.03)	0.8729	(−3.81)	0.8850	(−4.35)

Table 6. Nominal power and simulated power at specified sample size for normal shift alternatives.

$\Delta = \theta/\sigma$		0.3		0.5		1.0		1.5
Sample sizes (m, n)		(200, 200)		(73, 73)		(19, 19)		(9, 9)
Simulated power		0.8994		0.8986		0.9081		0.9012
Nominal power (percentage error)								
Exact variance	0.9005	(0.12)	0.9033	(0.52)	0.9096	(0.16)	0.9173	(1.78)
Lehmann	0.9003	(0.10)	0.9027	(0.45)	0.9079	(−0.02)	0.9165	(1.69)
Noether	0.8971	(−0.26)	0.8937	(−0.54)	0.8716	(−4.02)	0.8335	(−7.51)
Lower bound	0.9011	(0.19)	0.9048	(0.69)	0.9142	(0.67)	0.9248	(2.62)
Upper bound	0.8906	(−0.98)	0.8853	(−1.48)	0.8736	(−3.80)	0.8670	(−3.80)
Average bound	0.8957	(−0.41)	0.8948	(−0.42)	0.8926	(−1.71)	0.8933	(−0.88)
Sample sizes (m, n)		(133, 399)		(48, 144)		(13, 39)		(6, 18)
Simulated power		0.8986		0.8981		0.9105		0.9110
Nominal power (percentage error)								
Exact variance	0.9000	(0.16)	0.9001	(0.22)	0.9185	(0.88)	0.9220	(1.21)
Lehmann	0.8998	(0.13)	0.8996	(0.16)	0.9158	(0.59)	0.9195	(0.93)
Noether	0.8964	(−0.25)	0.8901	(−0.90)	0.8790	(−3.46)	0.8335	(−8.51)
Lower bound	0.9271	(3.17)	0.9342	(4.02)	0.9528	(4.64)	0.9535	(4.66)
Upper bound	0.8607	(−4.22)	0.8443	(−5.99)	0.8372	(−8.05)	0.8223	(−9.73)
Average bound	0.8904	(−0.91)	0.8827	(−1.71)	0.8846	(−2.84)	0.8739	(−4.07)
Sample sizes (m, n)		(399, 133)		(144, 48)		(39, 13)		(18, 6)
Simulated power		0.9008		0.8941		0.9122		0.9121
Nominal power (percentage error)								
Exact variance	0.9000	(−0.09)	0.9001	(0.67)	0.9185	(0.69)	0.9220	(1.09)
Lehmann	0.8998	(−0.11)	0.8996	(0.61)	0.9158	(0.40)	0.9195	(0.81)
Noether	0.8964	(−0.49)	0.8901	(−0.45)	0.8790	(−3.64)	0.8335	(−8.62)
Lower bound	0.9271	(2.92)	0.9342	(4.48)	0.9528	(4.45)	0.9535	(4.54)
Upper bound	0.8607	(−4.45)	0.8443	(−5.57)	0.8372	(−8.22)	0.8223	(−9.84)
Average bound	0.8904	(−1.15)	0.8827	(−1.27)	0.8846	(−3.02)	0.8739	(−4.18)

Table 7. Nominal power and simulated power at specified sample size for double exponential shift alternatives.

$\Delta = \theta/\sigma$		0.3		0.5		1.0		1.5
Sample sizes $(m,\ n)$		(132, 132)		(50, 50)		(15, 15)		(8, 8)
Simulated power		0.8996		0.8984		0.8976		0.9053
Nominal power (percentage error)								
Exact variance	0.9015	(0.21)	0.9028	(0.49)	0.9122	(1.63)	0.9174	(1.33)
Lehmann	0.9090	(1.04)	0.9195	(2.35)	0.9510	(5.95)	0.9726	(7.43)
Noether	0.8975	(−0.24)	0.8917	(−0.75)	0.8707	(−2.99)	0.8324	(−8.06)
Lower bound	0.9036	(0.45)	0.9079	(1.06)	0.9259	(3.15)	0.9384	(3.66)
Upper bound	0.8900	(−1.07)	0.8837	(−1.64)	0.8806	(−1.89)	0.8785	(−2.96)
Average bound	0.8967	(−0.33)	0.8953	(−0.34)	0.9018	(0.46)	0.9060	(0.07)
Sample sizes (m, n)		(88, 264)		(33, 99)		(10, 30)		(5, 15)
Simulated power		0.9032		0.9052		0.9046		0.8819
Nominal power (percentage error)								
Exact variance	0.9018	(−0.16)	0.9007	(−0.49)	0.9145	(1.09)	0.9008	(2.14)
Lehmann	0.9092	(0.66)	0.9178	(1.39)	0.9524	(5.28)	0.9661	(9.55)
Noether	0.8975	(−0.63)	0.8890	(−1.79)	0.8707	(−3.74)	0.8108	(−8.07)
Lower bound	0.9328	(3.28)	0.9393	(3.77)	0.9558	(5.66)	0.9478	(7.47)
Upper bound	0.8575	(−5.06)	0.8412	(−7.08)	0.8306	(−7.59)	0.8109	(−8.05)
Average bound	0.8907	(−1.38)	0.8825	(−2.51)	0.8848	(−2.18)	0.8633	(−2.11)
Sample sizes $(m,\ n)$		(264, 88)		(99, 33)		(30, 10)		(15, 5)
Simulated power		0.9087		0.8970		0.9036		0.8883
Nominal power (percentage error)								
Exact variance	0.9018	(−0.76)	0.9007	(0.42)	0.9145	(1.20)	0.9008	(1.41)
Lehmann	0.9092	(0.06)	0.9178	(2.31)	0.9524	(5.40)	0.9661	(8.76)
Noether	0.8975	(−1.23)	0.8890	(−0.89)	0.8707	(−3.64)	0.8108	(−8.73)
Lower bound	0.9328	(2.65)	0.9393	(4.71)	0.9558	(5.77)	0.9478	(6.70)
Upper bound	0.8575	(−5.64)	0.8412	(−6.23)	0.8360	(−7.48)	0.8109	(−8.71)
Average bound	0.8907	(−1.98)	0.8825	(−1.62)	0.8848	(−2.08)	0.8633	(−2.81)

simulated power is the proportion of the 10,000 replicates whose values exceed the critical value z_α. Consequently, the adequacy of the power approximations is determined by the difference between their nominal power and the simulated power. The results of both simulated power and percentage error are presented in tables 5–8 where percentage error is defined as

$$\text{Percentage error} = \frac{\text{nominal power} - \text{simulated power}}{\text{simulated power}} \times 100\%.$$

Examination of these tables shows that the performance of the exact variance approach is excellent over the whole range of conditions that were considered. The accuracy of the four approximate methods varies considerably with the structures of the shift alternatives. It is easily seen that the nominal powers of Lehmann's [1] approach tend to be higher than the simulated powers while Noether's [5] method shows the opposite pattern, namely that the nominal powers are generally lower than the simulated powers. Specifically, the approximation of Lehmann [1] gives satisfactory results for $\Delta = 0.3$ and 0.5 and is surprisingly accurate for all normal shift alternatives. In contrast, its percentage errors for exponential shift models are too large to be acceptable. On the other hand, the simplified formula of Noether [5] is reliable for $\Delta \leq 0.5$ throughout the simulation except in the last unbalanced setting of the exponential shift models. For the two methods based on the general lower and upper bounds of the non-null variance, the results suggest that both are extremely vulnerable to error when the sample sizes are unbalanced. Note that the upper bound approach frequently gives the largest errors among the formulas. The only situations where both variance bound methods maintain a reasonable magnitude of errors are with small location shifts and balanced designs. Unsurprisingly, the percentage errors of the average bound method are always between those

Table 8. Nominal power and simulated power at specified sample size for exponential shift alternatives.

$\Delta = \theta/\sigma$	0.3		0.5		1.0		1.5	
Sample sizes (m, n)		(83, 83)		(35, 35)		(13, 13)		(8, 8)
Simulated power		0.9038		0.8974		0.9104		0.9259
Nominal power (percentage error)								
Exact variance	0.9001	(−0.41)	0.9015	(0.45)	0.9211	(1.17)	0.9375	(1.25)
Lehmann	0.9547	(5.63)	0.9746	(8.60)	0.9964	(9.45)	0.9997	(7.97)
Noether	0.8938	(−1.10)	0.8861	(−1.26)	0.8742	(−3.98)	0.8523	(−7.95)
Lower bound	0.9036	(−0.03)	0.9092	(1.31)	0.9393	(3.17)	0.9625	(3.96)
Upper bound	0.8855	(−2.03)	0.8797	(−1.97)	0.8918	(−2.04)	0.9077	(−1.97)
Average bound	0.8943	(−1.06)	0.8938	(−0.41)	0.9141	(0.41)	0.9338	(0.85)
Sample sizes (m, n)		(60, 180)		(26, 78)		(10, 30)		(6, 18)
Simulated power		0.9012		0.9000		0.9214		0.9159
Nominal power (percentage error)								
Exact variance	0.9003	(−0.10)	0.9029	(0.32)	0.9261	(0.52)	0.9376	(2.37)
Lehmann	0.9666	(7.26)	0.9847	(9.41)	0.9988	(8.40)	0.9999	(9.17)
Noether	0.9141	(1.43)	0.9138	(1.53)	0.9121	(−1.01)	0.8868	(−3.18)
Lowerbound	0.9502	(5.64)	0.9638	(7.08)	0.9846	(6.86)	0.9922	(8.33)
Upper bound	0.8714	(−3.30)	0.8681	(−3.54)	0.8882	(−3.60)	0.9004	(−1.69)
Average bound	0.9074	(0.69)	0.9102	(1.13)	0.9332	(1.28)	0.9456	(3.24)
Sample sizes (m, n)		(153, 51)		(63, 21)		(21, 7)		(15, 5)
Simulated power		0.9056		0.9108		0.9006		0.9525
Nominal power (percentage error)								
Exact variance	0.9042	(−0.15)	0.9094	(−0.15)	0.9095	(0.99)	0.9690	(1.74)
Lehmann	0.9407	(3.88)	0.9618	(5.60)	0.9879	(9.70)	0.9995	(4.93)
Noether	0.8711	(−3.81)	0.8554	(−6.08)	0.8062	(−10.49)	0.8317	(−12.68)
Lower bound	0.9156	(1.11)	0.9175	(0.73)	0.9104	(1.09)	0.9697	(1.81)
Upper bound	0.8257	(−8.83)	0.8064	(−11.46)	0.7768	(−13.75)	0.8425	(−11.55)
Average bound	0.8634	(−4.66)	0.8504	(−6.64)	0.8256	(−8.33)	0.8948	(−6.06)

of the two procedures based on the lower and upper bounds. However, the performance of the average bound method is not as satisfactory as anticipated. Overall, the exact variance method has a clear advantage over the approximate counterparts.

4. Conclusion

In this article, we discuss power and sample size calculations for the Wilcoxon–Mann–Whitney test within the framework of location shift models. The asymptotic normal property is employed to obtain an approximation to the power function of the Wilcoxon–Mann–Whitney statistic. Extensive numerical study of power and sample size determinations is conducted to evaluate the impact of certain simplifications made by various competing approximations. A major criterion for the selection of an appropriate method is how well the nominal power matches the actual power corresponding to a given sample size and model configuration. The modifications of the exact mean and variance made in previously proposed large-sample approximations appear to affect and degrade the accuracy of sample size and power calculations in significant and distinctive ways. The application of these approximations should be restricted to shift models with relatively small departures from the null hypothesis. The findings in our numerical study suggest that the methods of Lehmann [1] and Noether [5] may be quite inaccurate for location shifts $\Delta = \theta/\sigma > 0.5$. In contrast to previous results in the literature, we find that the use of Noether's approximation for medium and large shift alternatives is problematic.

The sample sizes produced by the exact variance method are not only based on a more realistic mean and variance structure than the other approximate approaches, they also give smaller power discrepancies than the other approximations with the same sample size. Although computation is slightly more involved when using the exact variance formula, the extra complexity is outweighed by its superiority in accuracy. The normal approximation to the power of the Wilcoxon–Mann–Whitney test can be further improved through an Edgeworth expansion. However, the empirical results of this study show a simple normal approximation is adequate for most purposes. We feel that the method advocated in this paper is a practical way to investigate power and sample size selection, important issues that are often overlooked in discussions of the Wilcoxon–Mann–Whitney test.

References

[1] Lehmann, E.L., 1998, *Nonparametrics: Statistical Methods Based on Ranks* (Upper Saddle River, New Jersey: Prentice Hall).
[2] Milton, R.C., 1970, *Rank Order Probabilities: Two-Sample Normal Shift Alternatives* (New York: Wiley).
[3] Haynam, G.E. and Govindarajulu, Z., 1966, Exact power of Mann–Whitney test for exponential and rectangular alternatives. *Annals of Mathematical Statistics*, **37**, 945–953.
[4] Ramsey, F.L., 1971, Small sample power functions for nonparametric estimates of location in the double exponential family. *Journal of the American Statistical Association*, **66**, 149–151.
[5] Noether, G.E., 1987, Sample size determination for some common nonparametric tests. *Journal of the American Statistical Association*, **82**, 645–647.
[6] Vollandt, R. and Horn, M., 1997, Evaluation of Noether's method of sample size determination for the Wilcoxon–Mann–Whitney test. *Biometrical Journal*, **39**, 823–829.
[7] Birnbaum, Z.W. and Klose, O.M., 1957, Bounds for the variance of the Mann–Whitney statistic. *Annals of Mathematical Statistics*, **28**, 933–945.
[8] Hettmansperger, T.P., 1984, *Statistical Inference Based on Ranks* (New York: Wiley).