International Journal of Clinical Trials
Zhu X. Int J Clin Trials. 2021 Aug;8(3):184-195
http://www.ijclinicaltrials.com
pISSN2349-3240 | eISSN 2349-3259
DOI: https://dx.doi.org/10.18203/2349-3259.ijct20212840

Original Research Article

Sample size calculation for Mann-Whitney U test with five methods

Xiaoping Zhu*

Department of Biostatistics and Data Management, Regeneron Pharmaceuticals, New Jersey, USA

Received: 28 February 2021
Accepted: 08 April 2021

***Correspondence:**
Dr. Xiaoping Zhu,
E-mail: xiaoping.zhu@regeneron.com

ABSTRACT

Background: Precise sample size estimation plays a vital role in the planning of a study specifically for medical treatment expenses that are expensive and studies that are of high risk.
Methods: Among a variety of sample size calculation methods for the nonparametric Mann-Whitney U test, five potential methods are selected for evaluation in this article. The evaluation of method performance is based on the results obtained from high precision Monte Carlo simulations.
Results: The sample size deviations (from the simulation ones) are performance indicators. The sum of the squared deviations over all scenarios is used as the criterion for ranking the five methods. For power comparisons, the percentage errors (relative to the simulated powers) are used. The effect size and target power both have large impacts on the minimum required sample sizes.
Conclusions: Based on the ranking criterion, Shieh's method has the best performance. Noether's method always overestimates the minimum required sample sizes but not too severe.

Keywords: Mann-Whitney U test, Nonparametric, Sample size, Power, Monte Carlo simulation

INTRODUCTION

Sample size calculation plays an important role in the planning phase in any research area such as agriculture, medical science, economics, and other fields of research. Researchers would like to minimize experiment costs by using a minimal required sample size to detect a practical meaningful effect with a certain power. In testing the difference of two independent samples for continuous data, apparently, the most popular method is the two-sample t test. This test is more powerful and has the optimal power among all unbiased tests.[1] Although the normality will be met for large samples (usually 30 observations or more) by the central limit theorem, in reality it is difficult to meet the strict assumptions of normality and equal variances. Also, data distribution is rarely exact normal in practice. Thus, the result of the unpaired t test is unreliable especially for small sample sizes, heavy-tailed, severely skewed distributions, or outliers (i.e., abnormal extreme values). On the other hand, for categorical (i.e., ordinal and nominal) data, t tests is not appropriate. For these situations, the nonparametric rank-based tests are much preferred. In this paper, we will study rank-based nonparametric tests for two independent samples from the same distribution but with a location shift for the second sample.

In testing whether two independent samples come from the same distribution, nonparametric statistical tests are very useful. These methods do not require any specific form for the sampling distribution and do not make normality assumption. Oftentimes, we prefer median (instead of mean) for location shift distribution which is strongly skewed (either to the right or to the left), asymmetric (e.g., exponential distribution), long-tailed (e.g., double-exponential distribution), heavy-tailed (e.g., t(3) distribution). In addition, the rank-based nonparametric methods do not use the actual values of

the observations; instead, they are based on the rank (place in order) of each observation. Thus, the results of inference are not sensitive to outliers.

Moreover, there is no sample size requirement for the exact Mann-Whitney U test (a.k.a. Wilcoxon-Mann-Whitney test, or WMW test for short) to be valid. However, for the asymptotic WMW test to be valid, Siegel and Castellan recommended the following: m = 3 or 4 and n > 12; m > 4, and n > 10, where m and n are the sample sizes for the two groups, respectively.[2] Therefore, it is primarily useful for data with small sample sizes (<20) and extremely small sample sizes (i.e., m, n in the range of 2 to 6). Finally, the WMW test is also widely applied for the ordinal data.

More sophisticated nonparametric tests such as WMW test has been developed in many years ago. Neither distributional assumption nor normality assumption is employed for the data and hence is preferable. In this paper, we focus on five rank-based approaches to estimate power and sample size for two-sample using either linear rank tests or asymptotic WMW tests. These five methods are Lehmann, Noether, Wang et al, Shieh et al., and Doll and Klein.[3-7] The method of Doll and Klein is the two-sample linear rank tests approach and the other approaches use different approximations for the WMW test statistic. Details of the five methods are described in the section "methods" below. The aim of this paper is to assess their advantages and disadvantages about which approach is the most reliable and likely to estimate the minimum sample size needed in achieving power at a given level of significance for researches. More importantly, no research to date has compared the asymptotic WMW tests together with a linear rank test on a common ground.

Current paper is organized as follows. In the section "methods", the modelling of the local shift is introduced (see the subsection A). Five rank-based nonparametric methods for sample size calculation are described in the subsection B before an extensive simulation study is presented in the subsection C. The findings are illustrated by simulating results of the five methods in the section "results". In the section "discussion", a brief discussion of the finding is discussed in the use of these five methods. We close with a short conclusion in the section "conclusion". In this paper, we only consider the 2 independent samples (2-sided) on continuous data for normal location shift distribution. In practice, there are, of course, other test problems such as 1-sided test or matched pair sample test that might have some specific applications which we will not cover.

METHODS

This section provides a brief description of the five methods for the power and sample size calculation based on the rank-based tests on continuous data. Doll and Klein's method for power calculation uses the generalized linear rank test statistics based on the score generating functions. The other four methods use the traditional normal approximation for the power calculation based on the asymptotic WMW test statistic for large samples. Our investigation only focuses on these conventional methods because they all belong to the linear rank tests.

For most practical applications, the assumption of the approximate normality of the Wilcoxon test statistics is sufficiently accurate and adequate for comparing two samples when the sample sizes of the two independent samples are large. In fact, the basic assumption is the same for all the methods investigated.

Models of location shift

In what follows, we begin with the data assuming the distributions of two groups have the same shape and only differ by a location shift. Suppose $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are two independent random samples with continuous cumulative distribution functions F_X and F_Y, respectively, where, the values of X and Y are mutually independent and identically distributed (i.i.d.). The location shift modelling assumes the same shape except by alocation shift θ with probability density functions (pdf) f_X and f_Y. The location shift θ for the case of exponential and normal distributions of two samples is shown in (Figure 1).

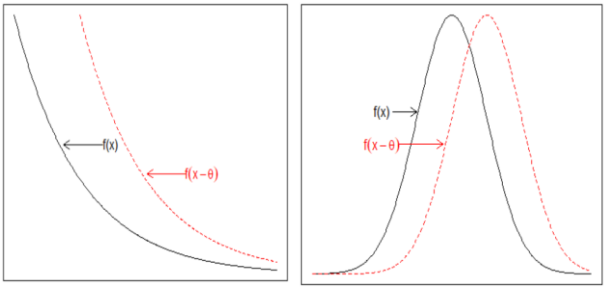

Figure 1: Location shift model for the same shape and spread by a shift of θ, A) exponential distributions; B) normal distributions.

According to the nature of the data and to simplify the complicated procedure, without loss of generality, some further assumptions are made as follows: the samples sizes m and n are the same (i.e., m=n) to maximize the power of the hypothesis test.[8] Happ et al also recommended a balance design for the WMW test.[9] Both F_X and F_Y must have the continuous distributions of the same shape and spread. B¨urkner et al showed that symmetric, continuous distributions under a location shift model that is optimal for the WMW test.[10] The location shift ($\theta \geq 0$) model follows a normal distribution, that is, $F_X(x) = F_Y(x-\theta), \forall x, y$. By using the normal model, we can pursue a unified approach that can be used of midranks in the formulas for the test statistics.[11,12]

Five methods for power and sample size calculation

This section provides a brief description of the five methods for the power and sample size calculation based on the rank-based tests on continuous data. Doll and Klein's method for power calculation uses the generalized linear rank test statistics based on the score generating functions. The other four methods use the traditional normal approximation for the power calculation based on the asymptotic WMW test statistic for large samples. Our investigation only focuses on these conventional methods because they all belong to the linear rank tests.

For most practical applications, the assumption of the approximate normality of the Wilcoxon test statistics is sufficiently accurate and adequate for comparing two samples when the sample sizes of the two independent samples are large. In fact, the basic assumption is the same for all the methods investigated.

In practical applications, it is of interest to investigate whether a shift (or difference) in location has occurred between two independent samples after conducting an experiment. Let $X_1, \ldots, X_m$ be stochastically independent, identically distributed (i.i.d.) with continuous cumulative distribution function (cdf) F_X and $Y_1, \ldots, Y_n$ be stochastically i.i.d. with continuous cdf F_Y. We consider independent samples such that X_i and Y_j are stochastically independent for $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$. Further, assuming the distributions of two samples have the same shape but differ by a shift in location, that is, $F_X(x) = F_Y(x - \theta)$, where θ is a location shift. A two-sided hypothesis of θ can be stated as

$$H_0: \theta = 0 \text{ vs. } H_a: \theta \neq 0$$

The Wilcoxon's rank sum test is frequently used for the above two independent samples test. This test is a nonparametric test based on the ranks of the individual observations rather than on their actual values. The Man-Whiney test statistic (a.k.a. WMW statistic) is given by

$$W = \sum_{i=1}^{m}\sum_{j=1}^{n} \psi(y_j - y_i) \ldots . \text{eq. } 1$$

Where $\psi(y_j - x_i) = 1$ if $y_j - x_i > 0$ and 0 otherwise. The assumption of the two independent samples, the hypothesis of θ, and the WMW statistic will be used in describing the five methods below unless stated otherwise.

Lehmann method

Lehmann first introduced the asymptotic power and sample size estimation based on the Wilcoxon rank sum test. Given two i.i.d. samples X and Y, the hypothesis is to test if θ = 0 for such two samples. For large m and n, the WMW test statistic's asymptotic normality in Equation (1) can be modified under the alternative hypothesis H_a. Thus, if the null hypothesis ($H_0: \theta = 0$) is rejected, the power for the WMW test against the alterative hypothesis ($H_a: \theta \neq 0$) can be approximated by

$$\text{Power} = 1 - \beta \approx \Phi\left(\sqrt{\frac{12mn}{N+1}}\theta f^*(0) - z_{\alpha/2}\right) \text{ eq. } 2$$

where $f^*(0)$ is the density of the distribution of difference of the two groups (X and Y) evaluated at zero and $f^*(0) = E[f(X_1)]$.

For $r = m/n$, the sample size can be solved from Equation 2.

$$n \approx \frac{(z_{\alpha/2} + z_\beta)^2(1+r)}{12\theta^2(f^*(0))^2 r} \text{ and } m = rn \ldots . . \text{eq. } 3$$

Noether method

Noether proposed a sample size determination method based on the Mann-Whitney statistic U for the two-sample Wilcoxon rank sum test given the two i.i.d. samples of X and Y. For large m and n, the assumption of the approximate normality of the test statistics is used for Equation (1) at the α asymptotic significance level. In addition, assuming the variance of the WMW statistic is the same under both hypotheses of H_0 and H_a. Noether provided the approximate power of the WMW test under H_a that θ ≠ 0 for $m = rn$:

$$\text{Power} = 1 - \beta \approx \Phi\left(\sqrt{\frac{12rn^2}{n + rn + 1}}(p_1 - 0.5) - z_{\alpha/2}\right)$$

$$\approx \Phi\left(\sqrt{\frac{12rn}{r+1}}(p_1 - 0.5) - z_{\alpha/2}\right) \text{ eq. } 4$$

Where the last approximation in Equation (4) is obtained by ignoring the "+1" in the denominator and $p_1 = P(X_1 < Y_1) = \int F_X dF_Y$.

Hence, by solving the above equation, the sample size required to achieve the target power can be obtained as

$$n \approx \frac{(z_{\alpha/2} + z_\beta)^2(1+r)}{12(p_1 - 0.5)^2 r} \text{ eq. } 5$$

Zhao et al. further generalized Noether's method so that it can be applied to continuous data as well as ordinal data (with or without ties).[13]

Wang et al method

Wang et al. derived an explicit sample size formula for Wilcoxon's rank sum test. They assumed two independent random samples $of\ X$ and Yas described in the subsection "models of location shift" follow the same continuous distribution except a location shift θ. The hypotheses to be tested are $H_0: \theta = 0$ versus $H_a: \theta \neq 0$. The test statistic W is the sum of ranks in Y group, where ranks are based on the combined observation with rank 1 to $m + n$. For large m and n, the test statistic W can be assumed as asymptotically standard normal (Z^*)

$$Z^* = \frac{W - \mu_W}{\sigma_W} \text{ eq. } 6$$

Where;

$$\mu_W = \frac{n(n+1)}{2} + mnp_1$$

$$\sigma_W^2 = mnp_1(1-p_1) + mn(m-1)(p_2 - p_1^2) + mn(n-1)(p_3 - p_1^2)$$

$$p_1 = P(y_1 \geq x_1),\ p_2 = P(y_1 \geq x_1 \text{ and } y_1 \geq x_2),\ p_3 = P(y_1 \geq x_1 \text{ and } y_2 \geq x_1)$$

For $r = m/n$, the power of the test can be approximate by

$$\text{Power} = 1 - \beta = \Phi\left(\frac{z_{\alpha/2}\sqrt{r(r+1)/12} + \sqrt{n}r(1/2 - p_1)}{\sqrt{r^2(p_1 - p_1^2) + r(p_3 - p_1^2)}}\right) \text{ eq. } 7$$

Thus, by solving the above equation, the sample size needed to achieve the target power can be obtained as

$$m = rn, n = \frac{\left(z_{\alpha/2}\sqrt{r(r+1)/12} + z_\beta\sqrt{r^2(p_2 - p_1^2) + r(p_3 - p_1^2)}\right)^2}{r^2(1/2 - p_1)^2} \text{ eq. } 8$$

Readers can consult their paper for more details.[5]

Shieh et al method

Shieh et al introduced explicit sample size and power formulas for the WMW test. The derivation of sample size and power formulas is based on the asymptotic normal distribution of the WMW statistic with an exact variance large-sample method. They assumed two independent random samples of X and Y as described in the subsection "models of location shift". The hypotheses to be tested are $H_0: \theta = 0$ versus $H_a: \theta \neq 0$. The Mann-Whitney form of the MWM statistic is define as

$$W = \sum_{i=1}^{m}\sum_{j=1}^{n} \psi(y_j - x_i), \text{ eq. } 9$$

Where $\psi(y_j - x_i) = 1$ if $y_j - x_i > 0$ and 0 otherwise.

They found that as n and m tend to infinity, $(W - \mu)/\sigma$ tends to the standard normal distribution (i.e., $N(0,1)$) with $\mu = mnp_1$ and $\sigma^2 = mnp_1(1-p_1) + mn(n-1)(p_2 - p_1^2) + mn(m-1)(p_3 - p_1^2)$, where $p_1 = P(X_1 < Y_1) = \int F_Y dF_X, p_2 = P(X_1 < Y_1 \cap X_1 < Y_2) = \int (1 - F_X)^2\, dF_Y$ and $p_3 = P(X_1 < Y_1 \cap X_2 < Y_1) = \int {F_Y}^2 dF_X$.

Finally, the power of the test can be approximated by

$$\text{Power } = 1 - \beta \approx \Phi\left(\frac{\mu - \mu_0 - z_0\sigma_o}{\sigma}\right). \text{eq. } 10$$

where $\mu_0 = \frac{nm}{2}$ and $\sigma_0^2 = \frac{mn(N+1)}{12}$. The sample size needed to achieve the target power can be obtained by solving Equation (10) using numerical methods such as bisection method or Newton-Raphson method. Alternatively, sample size can be obtained by a linear search method due to power increases monotonically as sample size increases. Readers can consult their paper for more details.[6]

Doll and Klein method

Doll and Klein introduced a new sample size analysis method for arbitrary linear rank tests for location shifts of continuous distributions.[7] The WMW test is a special case of the linear rank tests. The method is based on linear rank tests' asymptotic normality, while mean and variance can be expressed by score generating functions for large m and n.

Finally, for $m = n$, the power of the test can be expressed as

$$\text{Power} = 1 - \beta \approx 1 - \Phi\left(z_\alpha/2 - \sqrt{\frac{n}{2}}\delta\sigma\frac{\int_0^1 \phi(u)\phi(u;f_0)du}{\sqrt{\int_0^1 (\phi(u) - \overline{\phi})^2 du}}\right) + \Phi\left(-z_\alpha/2 - \sqrt{\frac{n}{2}}\delta\sigma\frac{\int_0^1 \phi(u)\phi(u;f_0)du}{\sqrt{\int_0^1 (\phi(u) - \overline{\phi})^2 du}}\right) \text{ eq. } 11$$

The sample size needed to achieve the target power can be obtained by solving Equation (11) using numerical methods (e.g., bisection method or Newton-Raphson method). Because power increase monotonically as sample size increases, a straightforward linear search can also be used to obtain the sample size. For more detailed information, readers can consult their paper.[7]

Formulas for the five methods

Using the five methods for the asymptotic WMW test described above, we list systematically their simplified explicit formulas of the power and sample size calculation in (Table 1) corresponding to both equal and unequal sample sizes with practical assumptions.

However, as noted, it is not always possible to come with a closed-form formula for the sample size (e.g., Shieh and Doll methods described above). Generally, an iterative numerical procedure can be applied to obtain the sample size whenever the explicit formula is not available. The flowchart of the iterative numerical procedure is illustrated in (Figure 2) for sample size determination when a specific power formula is given. In addition, the notations in (Table 1) have been adjusted for the ease of understanding, comparison, and applications.

Table 1: Formulas of power and sample size calculations for five methods using WMW test.

Methods	Sample sizes (n, m) and power $(1-\beta)$ assumed a normal location shift model	
	$\mathbf{r \neq 1}$ **(or** $\mathbf{n \neq m}$**)**	$\mathbf{r = 1}$ **(or** $\mathbf{n = m}$**)**
Lehmann (1975)	$1-\beta \approx \Phi\left[\sqrt{12nB}\,\theta\frac{1}{2\sqrt{\pi}} - z_{\alpha/2}\right]$ $n \approx \frac{\pi(z_{\alpha/2} + z_{\beta})^2}{3B\theta^2}$ and $m = rn$ where $B = r/(1+r)$; $\pi = 3.14159$	$1-\beta \approx \Phi\left[\sqrt{6n}\,\theta\frac{1}{2\sqrt{\pi}} - z_{\alpha/2}\right]$ $n \approx \frac{2\pi(z_{\alpha/2} + z_{\beta})^2}{3\theta^2}$ and $m = n$
Noether (1987)	$1-\beta \approx \Phi\left[\sqrt{12nB}\,(p_1 - 0.5) - z_{\alpha/2}\right]$ $n \approx \frac{(z_{\alpha/2} + z_{\beta})^2}{12B(p_1 - 0.5)^2}$ and $m = rn$ where $p_1 = P(Y_1 \geq X_1) = 0.5 + \frac{\theta}{2\sigma\sqrt{\pi}}$ $\sigma = \sigma_X = \sigma_Y = 1$	$1-\beta \approx \Phi\left[\sqrt{6n}\,(p_1 - 0.5) - z_{\alpha/2}\right]$ $n \approx \frac{(z_{\alpha/2} + z_{\beta})^2}{6(p_1 - 0.5)^2}$ and $m = n$
Wang et al (2003)	$1-\beta \approx \Phi\left\{\frac{1}{G}\left[\sqrt{\frac{12nB}{r}}\,(p_1 - 0.5) - z_{\alpha/2}\right]\right\}$ $n \approx \frac{(z_{\alpha/2} + z_{\beta}G)^2 r}{12B(p_1 - 0.5)^2}$ and $m = rn$ where $G = \sqrt{B\,[(p_2 - {p_1}^2)/r + (p_3 - {p_1}^2)]}$ $p_2 = P(Y_1 \geq X_1 \text{ and } Y_1 \geq X_2)$; $p_3 = P(Y_1 \geq X_1 \text{ and } Y_2 \geq X_1)$	$1-\beta \approx \Phi\left\{\frac{\sqrt{6n}(p_1 - 0.5) - z_{\alpha/2}}{\hat{G}}\right\}$ $n \approx \frac{(z_{\alpha/2} + z_{\beta}\hat{G})^2}{6(p_1 - 0.5)^2}$ and $m = n$ where $\hat{G} = \sqrt{0.5[(p_2 - {p_1}^2) + (p_3 - {p_1}^2)]}$
Shieh et al (2006)	$1-\beta \approx \Phi\left(\frac{\mu - \mu_0 - z_{\alpha/2}\,\mu_0}{\sigma}\right)$ $n \rightarrow$ rely on the power formula above with numerically algorithm; $m = rn$ Where $\mu_0 = rn^2/2$; $\sigma_0 = \sqrt{rn^2(n + rn + 1)/12}$ $\mu = rn^2 p_1$ $\sigma_a = \sqrt{rn^2 + (1-n)S_2 + (rn-1)S_3}$ $S_1 = p_1 - p_1^2$; $S_2 = p_2 - p_1^2$; $S_3 = p_3 - p_1^2$	$1-\beta \approx \Phi\left(\frac{\mu - \mu_0 - z_{\alpha/2}\,\mu_0}{\sigma_a}\right)$ $n \rightarrow$ rely on the power formula above with numerically algorithm; $m = n$ Where $\mu_0 = n^2/2$; $\sigma_0 = \sqrt{n^2(2n+1)/12}$ $\mu = n^2 p_1$ $\sigma_a = \sqrt{n^2[S_1 + (1-n)(S_2 + S_3)]}$
Doll and Klein (2019)	$1-\beta \approx 2 - \Phi(z_{\alpha/2} - D) - \Phi(z_{\alpha/2} + D)$ $n \rightarrow$ rely on the power formula above with numerically algorithm; $m = rn$ where: $D = \sqrt{n/2}\,\theta\,(0.2821/\sqrt{1/12})$	$1-\beta \approx 2 - \Phi(z_{\alpha/2} - \hat{D}) - \Phi(z_{\alpha/2} + \hat{D})$ $n \rightarrow$ rely on the power formula above with numerically algorithm; $m = n$ where: $\hat{D} = \sqrt{n/2}\,\theta$

α and β are fixed probabilities of the type-I and type-II error rates, respectively.
$z_{\alpha/2}$and z_{β}denote the upper $\alpha/2$ th and βth quantile of the standard normal distribution

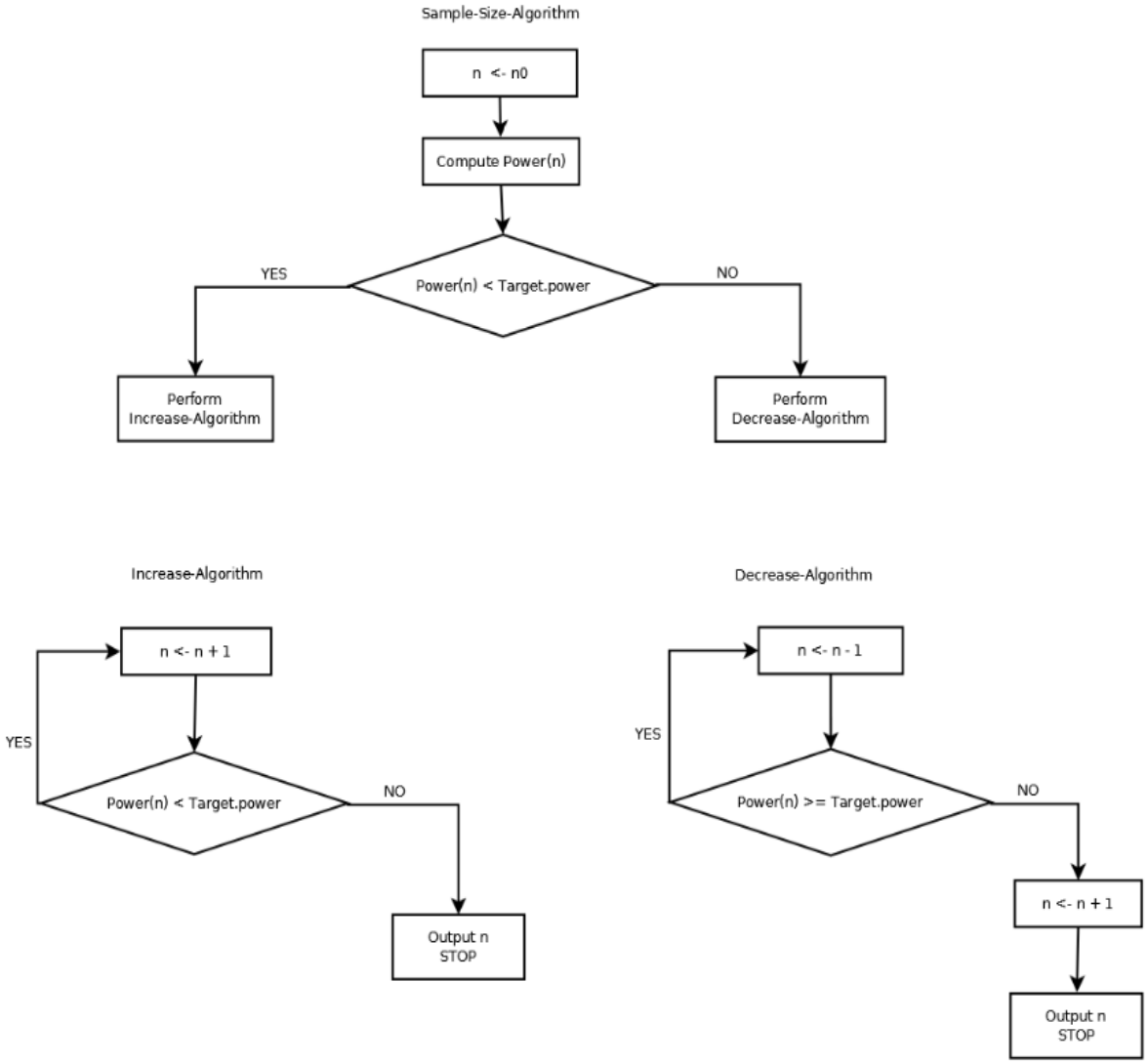

Figure 2: An iterative numerical process to determine the required sample size for a given target power.

Simulation study

To illustrate the applications of the five methods for sample size and power calculations for the WMW test described in the subsection "five methods for power and sample size calculation", a simulation study was performed to study the properties of these five methods.

The aim of the simulation study primarily focuses on two aspects: evaluate the differences of the sample sizes obtained from the formulas and compare them to those from the simulation under various powers and shifts. Assess the accuracy of the approximate powers obtained from the five methods and compare to the almost true powers from the simulation when the sample sizes are given. In current paper, simulation technique is used to compute sample sizes and powers. The simulation is implemented in R software using the Wilcox test function from R stats package.[14] The simulation results are served as a reference (i.e., ground truth) for comparing the five methods under investigation.

Simulation study design

We performed the simulation study over a range of settings designed to consider the real-life circumstances frequently encountered in clinical research. As described in the subsection "models of location shift", the corresponding two independent and identical normal distributions of X and Y for continuous data were generated. That is, $X_1, \ldots, X_m \sim N(0,1)$ and $Y_1, \ldots, Y_n \sim N(\theta, 1)$ are independent continuous random samples, and $F_Y(x) = F_X(x - \theta)$ where $F_X(x)$ is the standard normal distribution so $p_1 = \Phi(\theta/\sqrt{2})$ and $p_2 = p_3 = E[\{\Phi(Z + \theta)\}^2]$ where $Z \sim N(0,1)$. Without loss of generality, let the variances of X and Y be 1 (i.e., $\sigma_X = \sigma_Y = 1$). In nonparametric analysis, θ (or p_1) is served as the effect size. The seven effect sizes (θ) considered in

this study are $0.2, 05, 0.8, 1.0, 1.2, 1.5$ and 2.0. Four target powers are investigated (i.e., $80\%, 85\%, 90\%,$ and 95%). In this study, we paid more attention to the larger effect sizes (i.e., $\theta \geq 1.0$) for studying the cases of the smaller sample sizes. The purpose is to explore the behaviors of the five methods under study. For all simulation scenarios, we only used 2-sided tests and a significance level of $\alpha = 0.05$ with equal sample sizes (i.e., m=n), without loss of generality. In addition, the asymptotic approximate powers for these five methods are compared with the almost true power from the simulation to study the accuracy and reliability among these methods. In the simulation, the power of the WMW test was estimated by counting the number of cases when the hypothesis H_0 was rejected and dividing it by the total number of simulations ($N = 10^6$). Flowchart for the power calculation using a Monte Carlo simulation is shown in (Figure 3).

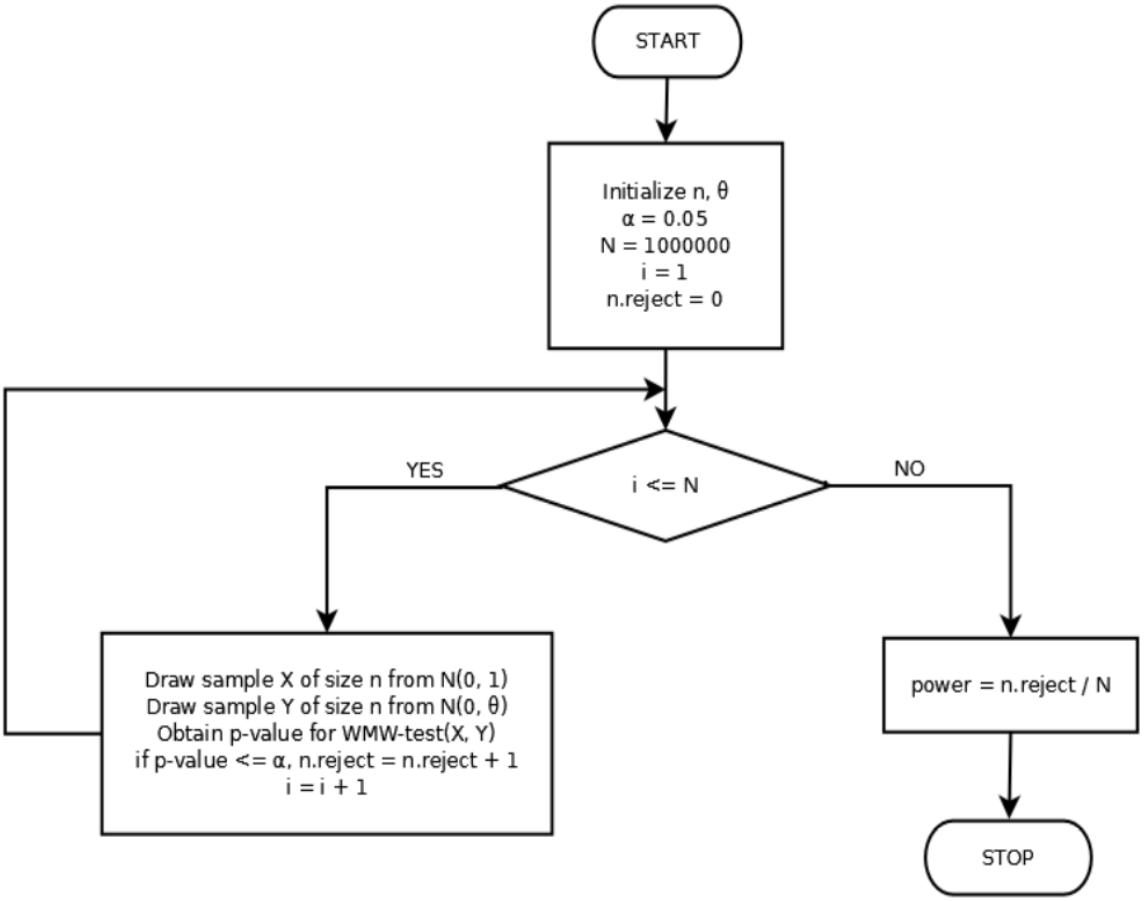

Figure 3: The power computation through a Monte Carlo simulation.

Number of simulations

The simulations were carried out with 10^6 replications for each of the 28 scenarios. The 10^6 simulations were used to make the Monte Carlo simulation error small enough. In general, simulation error is approximately proportional to $1/\sqrt{10^6}$=0.001. In other words, the maximum Monte Carlo standard error of the simulated power is less than 0.001 in each scenario in this study. To justify the necessary use of 10^6 simulations for high accuracy, four numbers of simulations (i.e., 10^3, 10^4, 10^5, and 10^6) are usedto show the corresponding accuracy for the case of 90% power. The powers for n=89 and n=90 are considered. A 95% confidence interval (CI) of the simulated power was constructed by a nonparametric bootstrapping method for each of sample size and number of simulation combinations. The variation of confidence intervals for different number of simulations is clearly seen in (Figure 4). The length of a confidence interval implies the corresponding precision. Shorter confidence interval means higher precision as in the case of $N = 10^6$. As depicted in (Figure 4) n = 90 is the required sample size of a target power of 90% but not n=89 when using 10^6 simulations. Performing 10^6 simulations takes a lot of time even for a sample size of 90. It takes longer time (e.g., a few hours) for larger sample sizes. Therefore, it is impractical to use in the case of a lot of scenarios needed to be considered. However, in studying the various properties of a method or methods, comparing methods, or verifying the validity of methods, simulation technique is a good and perhaps the only method to use for researchers.

RESULTS

The results of the simulation will be presented in 2 separate parts on sample size and power calculation for the five methods for the 28 settings described in the subsection "simulation study". To simplify labelling and reference, let $n_s(pwr_s)$ denote the sample size (power) obtained from the simulation and $n_L(pwr_L)$, $n_N(pwr_N)$, $n_W(pwr_W)$, $n_{Sh}(pwr_{Sh})$, and $n_D(pwr_D)$ denote the corresponding five methods (i.e., Lehmann, Noether, Wang, Shieh, and Doll) for sample size (power) calculation using asymptotic normal approximation. For convenience, the sample sizes or powers obtained from the five methods are referred to as "formula-based methods" hereafter.

Required sample sizes

The sample sizes obtained from the Monte Carlo simulations and the formula-based methods for the equal sample sizes ($m = n$) case are shown in (Table 2). The formula based sample sizes as mentioned through terms ($n_{(.)}$ denoted for n_L, n_N, n_W, n_{sh}, and n_D) were calculated from the formulas of the five methods and the simulation sample size (n_s) is obtained by applying the general sample size calculation algorithm (Table 1). The sample size deviations ($d_{(.)}$) between each of methods and n_s(i.e., the almost true sample sample) are also calculated. The formula-based sample sizes are in good agreement with that of simulated sample size whenever a small deviation is observed. In general, the deviation ($n_{(.)}$ -n_s) of either 1 or -1 is small enough to be regarded as no difference in practice. For small effect size ($\theta = 0.2$), as expected, the required sample sizes are large and the deviations from n_s across the powers for all methods except Noether's method are large as well. Noether's method has very small deviation values (0, 2, 0, -1) corresponding to the four target powers (80%, 85%, 90%, and 95%) while Wang's method has larger deviation values (-10, -6, -16, and 13) for each target power. When $\theta = 0.5$, the deviations of the sample sizes from n_s are similar for the methods of Lehmann (-1, -1, -1, -1) and Shieh (0, -1, -1, 0) compared with the other three methods. Also, Noether's method has a slightly large positive deviations (1, 2, 2, 3) but Wang's method (-2, -1, -3, -2) has negative deviation which is in contrary to Noether's for the target powers. Overall, for θ greater than or equal to 0.5, the sample sizes gradually declined especially it is dramatically reduced for θ changing from 0.2 to 0.5. For illustrative purpose, the method ranking referred to the sample size calculated to each method is provided by the difference between $n_{(.)}$ and n_s. As a result, (Table 3) provides both the method rankings and the sum of the squared deviations (SSD), where SSD for each method is defined as:

$$\text{SSD} = \sum_{i=1}^{i=28} d_i^2 \quad \text{eq. 12}$$

The SSD is the criterion used to rank performances of the five methods. Smaller SSD implies better performance. Thus, the smallest SSD has rank 1 and the second smallest SSD has rank 2, and so on. In addition, average ranks will be assigned in the case of tie. Results of ranking are shown in (Table 3) for the four target powers. In the power of 85% case, there is a tie situation (i.e., the SSD values of Lehmann and Shieh method are the same) and average rank of 1.5 is assigned to both methods. As can be seen in (Table 3), Shieh's method has the first rank for power of 80%, 85%, and 90% and has the second rank for power of 95%. Lehmann's method has the first rank for powers of 85% and 95%, the second rank for power of 90%, and the third rank for power of 80%. Wang's method has the fifth rank due to its larger deviations from the referenced sample size for all powers. It appears that the resulting sample sizes are not accurate enough. To demonstrate the precision of the sample size calculations for WMW tests visually, box plots of the overall deviations regardless of effect size (θ) and target power are shown in (Figure 5). As a graphical illustration, it is easy to see that the ordering of the precision for the sample size calculation is Shieh >Lehmann >Doll >Wang >Noether method. Overall, Shieh's method is consistently superior to the other four methods and Noether's method has a tendency to overestimate the sample sizes.

Statistical power

In this section, we show the estimation of powers in the simulation and the formula-based for the five methods across all 28 stimulation settings. The target powers are used for the sample size calculation based on 1 million simulations for each setting. By the simulation sample size (n_s), the simulated (true) power (pwr_s) is estimated by a Monte Carlo method and the formula-based power is determined by the formulas of the five methods in (Table 1). The algorithm of the simulated powers using the standardized WMW test statistics was described in the subsection "simulation study". The results are given in (Table 4). Also, the percentage errors ($e_{(.)}$) between the simulated powers and each of the formula-based powers as defined by Shieh are presented in (Table 4), where the $e_{(.)}$ is defined as;

$$\text{Percentage error} = e_{(.)} = \frac{pwr_{(.)} - pwr_s}{pwr_s} \quad \text{eq. 13}$$

Where $(.)$ denotes one of the methods (Lehmann, Noether, Wang, Doll,or Shieh).[3-6] When θ is small (θ= 0.2), the formula-based powers of the five methods are very close to the simulated power for a given large sample size (Table 4). The percentage errors of the five methods are very small and unimportant for small θ, where the e_L, e_{Sh}, and e_D corresponding Lehmann, Shieh, and Doll methods have small positive values especially the e_L is very small; Noether's method e_N has a small negative value; the error is either negative or positive for Wang's method (where e_W is negative for $n_s = 414$ or

554 and is positive for $n_s = 472$ or 686). However, when θ is median or large (θ= 0.5 or θ= 0.8), we observe that the powers of the five methods are gradually far away from the simulated powers and the values of $e_{(.)}$ increases as well. Moreover, when θ is extremely large, the powers of the five methods compared with the simulated powers are considerably inconsistent and give a large positive or negative error, which the values of error ($e_{(.)}$ = ~±4%, ~±5%, ~±8%, and ~±14%) across all the methods of effect size θof 1, 1.2, 1.5, and 2, respectively. Overall, an observation that is revealed by (Table 4) is that, in all settings, it was a very constant phenomenon that the estimated powers for Noether's method was found to be less than the corresponding simulated powers with negative e_N whilethe other methods (Lehmann, Shieh, and Doll) are consistently larger than the simulated power with positive errors.

Table 2: Results of sample size from the simulation study. The simulated sample size (n_s), the formula-based sample size ($n_{(.)}$), and the deviation $d_{(.)} = n_{(.)} - n_s$are presented.

θ	Target power	Simulation Sample size n_s	Formula-based Sample Sizes [$n_{(.)}$] and Deviation [$d_{(.)} = n_{(.)} - n_s$]									
			Lehmann		Noether		Wang		Shieh		Doll	
			n_L	d_L	n_N	d_N	n_W	d_W	n_{Sh}	d_{Sh}	n_D	d_D
0.2	80%	414	412	-2	414	0	404	-10	417	3	411	-3
0.2	85%	472	471	-1	474	2	466	-6	474	2	471	-1
0.2	90%	554	551	-3	554	0	538	-16	553	-1	551	-3
0.2	95%	686	681	-5	685	-1	699	13	681	-5	681	-5
0.5	80%	68	67	-1	69	1	66	-2	68	0	66	-2
0.5	85%	77	76	-1	79	2	76	-1	76	-1	76	-1
0.5	90%	90	89	-1	92	2	87	-3	89	-1	89	-1
0.5	95%	111	110	-1	114	3	109	-2	110	-1	109	-2
0.8	80%	28	27	-1	29	1	27	-1	28	0	26	-2
0.8	85%	31	30	-1	33	2	31	0	31	0	30	-1
0.8	90%	36	35	-1	39	3	35	-1	35	-1	35	-1
0.8	95%	45	44	-1	48	3	42	-3	44	-1	43	-2
1	80%	18	17	-1	20	2	18	0	18	0	17	-1
1	85%	21	20	-1	23	2	19	-2	20	-1	19	-2
1	90%	24	23	-1	26	2	23	-1	23	-1	23	-1
1	95%	29	28	-1	32	3	27	-2	28	-1	28	-1
1.2	80%	14	12	-2	15	1	13	-1	13	-1	12	-2
1.2	85%	15	14	-1	17	2	14	-1	15	0	14	-1
1.2	90%	17	16	-1	20	3	16	-1	17	0	16	-1
1.2	95%	21	20	-1	24	3	19	-2	19	-2	19	-2
1.5	80%	10	8	-2	11	1	9	-1	9	-1	8	-2
1.5	85%	10	9	-1	12	2	10	0	10	0	9	-1
1.5	90%	12	11	-1	14	2	11	-1	11	-1	10	-2
1.5	95%	14	13	-1	18	4	12	-2	13	-1	13	-1
2	80%	6	5	-1	8	2	6	0	6	0	5	-1
2	85%	7	6	-1	9	2	6	-1	6	-1	5	-2
2	90%	8	6	-2	10	2	6	-2	7	-1	6	-2
2	95%	9	8	-1	13	4	7	-2	8	-1	7	-2

Table 3: Method rankings: the sum-squared deviations (SSDs) corresponding to each method by the four target powers are shown in parentheses across all the θs.

Method	Target powers			
	80%	85%	90%	95%
Lehmann	3 (16)	1.5 (7)	2 (18)	1 (31)
Noether	2 (12)	4 (28)	4 (34)	4 (69)
Wang	5 (107)	5 (43)	5 (273)	5 (198)
Shieh	1 (11)	1.5 (7)	1 (6)	2 (34)
Doll	4 (27)	3 (13)	3 (21)	3 (43)

Table 4: Results of powers estimation from the simulation study. The simulated power (pwr_s), the formula-based power ($pwr_{(.)}$), and the percent error ($e_{(.)}$) are presented.

θ	Target Power	Simulated		Formula-based powers ($pwr_{(.)}$) and error (%)									
				Lehmann		Noether		Wang		Shieh		Doll	
		n_s	pwr_S	pwr_L	e_L	pwr_N	e_N	pwr_W	e_W	pwr_{Sh}	e_{Sh}	pwr_D	e_D
0.2	0.8	414	0.8018	0.8024	0.079	0.8003	-0.187	0.7946	-0.89	0.8036	0.225	0.8029	0.139
0.2	0.85	472	0.8509	0.851	0.012	0.8491	-0.218	0.8528	0.215	0.8511	0.02	0.8514	0.056
0.2	0.9	554	0.9014	0.9017	0.039	0.9001	-0.142	0.8964	-0.55	0.9024	0.113	0.902	0.068
0.2	0.95	686	0.9511	0.9514	0.032	0.9503	-0.083	0.9513	0.019	0.9545	0.365	0.9515	0.046
0.5	0.8	68	0.8067	0.8102	0.435	0.797	-1.207	0.8058	-0.11	0.8028	-0.476	0.813	0.785
0.5	0.85	77	0.8521	0.8559	0.437	0.8437	-0.992	0.8507	-0.17	0.8589	0.792	0.8581	0.698
0.5	0.9	90	0.9023	0.9047	0.261	0.8945	-0.869	0.9052	0.321	0.9082	0.651	0.9062	0.432
0.5	0.95	111	0.9516	0.9527	0.113	0.9458	-0.612	0.961	0.983	0.9582	0.685	0.9535	0.198
0.8	0.8	28	0.817	0.8262	1.128	0.7928	-2.958	0.8233	0.769	0.8123	-0.582	0.8328	1.929
0.8	0.85	31	0.8556	0.8629	0.854	0.8318	-2.778	0.8595	0.457	0.8604	0.567	0.8682	1.476
0.8	0.9	36	0.9029	0.9089	0.669	0.8826	-2.243	0.9115	0.961	0.9126	1.084	0.9126	1.078
0.8	0.95	45	0.9555	0.958	0.265	0.9406	-1.557	0.9613	0.615	0.9637	0.858	0.9598	0.454
1	0.8	18	0.803	0.8243	2.653	0.7718	-3.887	0.8183	1.914	0.8124	1.179	0.8344	3.915
1	0.85	21	0.8674	0.8789	1.325	0.8318	-4.104	0.8705	0.357	0.8662	-0.141	0.8862	2.167
1	0.9	24	0.9083	0.9178	1.05	0.8776	-3.377	0.9187	1.146	0.9151	0.752	0.923	1.617
1	0.95	29	0.9538	0.9581	0.453	0.9296	-2.537	0.9659	1.271	0.9588	0.526	0.9609	0.742
1.2	0.8	14	0.8446	0.8618	2.038	0.7902	-6.439	0.8601	1.827	0.8466	0.232	0.8734	3.407
1.2	0.85	15	0.8606	0.8848	2.808	0.8171	-5.056	0.8892	3.32	0.8823	2.52	0.8946	3.952
1.2	0.9	17	0.9084	0.9207	1.356	0.862	-5.104	0.9278	2.141	0.9142	0.643	0.9277	2.129
1.2	0.95	21	0.958	0.9637	0.6	0.9235	-3.594	0.9729	1.563	0.9684	1.085	0.9671	0.956
1.5	0.8	10	0.853	0.8923	4.603	0.7865	-7.796	0.892	4.571	0.8624	1.099	0.9062	6.237
1.5	0.85	10	0.853	0.8923	4.603	0.7865	-7.796	0.8792	3.069	0.8686	1.833	0.9062	6.237
1.5	0.9	12	0.919	0.9404	2.326	0.8548	-6.987	0.9551	3.928	0.9478	3.137	0.9485	3.211
1.5	0.95	14	0.9605	0.9679	0.771	0.903	-5.984	0.9764	1.654	0.9735	1.359	0.9725	1.246
2	0.8	6	0.8285	0.9019	8.858	0.715	-13.69	0.8891	7.313	0.8194	-1.099	0.923	11.401
2	0.85	7	0.8862	0.9421	6.302	0.7796	-12.03	0.9551	7.776	0.9146	3.2	0.9551	7.772
2	0.9	8	0.9491	0.9665	1.84	0.8313	-12.41	0.9873	4.032	0.9604	1.189	0.9743	2.663
2	0.95	9	0.962	0.981	1.981	0.8721	-9.344	0.9941	3.338	0.9892	2.829	0.9856	2.456

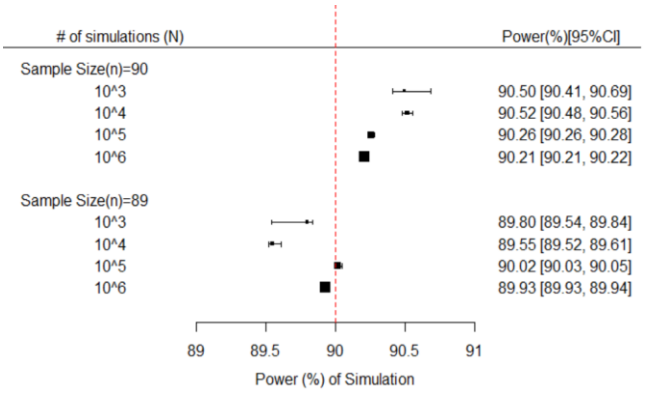

Figure 4: Power of four numbers of simulations with 95% CIs using bootstrap method.

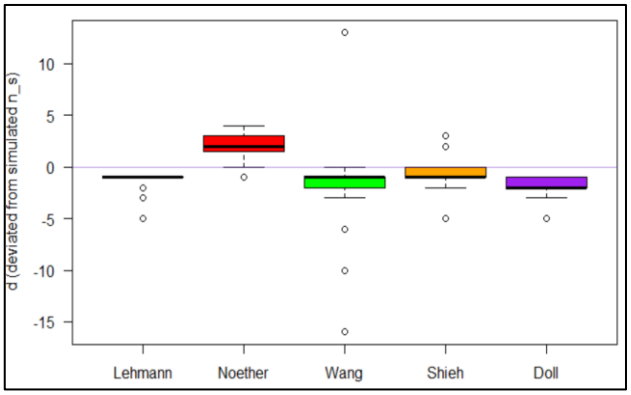

Figure 5: Box plot of the sample size deviations from the simulation for the five methods.

DISCUSSION

The sample size calculation is a fundamental aspect in clinical research to detect a clinically relevant effect size. It reflects several things to be considered when calculating the required sample size, including the minimal clinically relevant effect size (θ), hypothesis testing framework, data distribution, variance of the outcome, the significance level (α) and target power (1-β) of the 1-sided or 2-sided test.

In this paper, the five nonparametric methods commonly used in the sample size determination for the asymptotic Mann-Whitney U test are evaluated and investigated for their performance through the simulation study. There are strengths of our research in comparing the five methods. That is, the application of a large number of simulations (1 million) and the comprehensive considerations of 28 settings. Consequently, we will be able to conclude with sufficient confidence for the choice of those methods when calculating sample size and power for the WMW test on continuous data in most applications. However, the simulation in this paper is not without limitations. We focused only on continuous data with a normal location shift. As described in the section "introduction", the WMW test can be used to analyze continuous data that are not normally distribution. For example, Shieh and Wang derived explicit formulas for the sample size and power for non-normal distributions (i.e., uniform, exponential, and double exponential). The formulas for both Lehmann's and Noether's methods avoid the evaluation of p_2 and p_3. Lehmann proposed the density of F distribution be evaluated at zero, $f^*(0)$, which the quantity of $f^*(0)$ are specified as $f^*(0) = 1, 1/4, \text{and } 1/2$ for uniform $(-1/2, 1/2)$, double exponential (0, 1) and exponential (1) distribution, respectively.

Doll's method for sample size is suitable for various distributions of continuous data and similar to the method proposed by Mollan for computing power of the exact WMW test.[16] Nevertheless, Doll's method resolved the computation burden of Van de Wiel's method which is generally very time consuming and only works for sample sizes up to 40 (i.e., 20 in each of the two groups). Further, for simplicity, in the evaluation of the five methods for the asymptotic WMW test, the formulas (Table 1) for equal sample size in comparing two groups is considered. In fact, the formulas for unequal sample size of two groups are also provided in the table. Note that Shieh's method gives different power estimates when the values of m and n are interchanges (e.g., m=12, n=6 versus m=6, n=12) for a fixed effect size and p_2 and p_3 are unequal for non-symmetric distributions.[17] Furthermore, there may be cases that a 1-sided test is more appropriate than a 2-sided test. In such cases, $z_{\alpha/2}$ in the 2-sided test formulas can simply be replaced by z_α. In addition, the value of significance level (α) (i.e., the type I error rate, probability of reject H_0 when H_0 is true) is not restricted but needs to be specified in advance.

In practical applications, α is often set to 5% or 1%, however, the formulas are still valid for any other value of α. Finally, we will provide the implementation of these methods in popular sample size calculation software packages. Shieh method is very popular and more precise in calculating sample size or power for the WMW test, but the computation is slightly more involved because the exact variance of W statistic is used. Fortunately, the method has been implemented in the wmwpow R package. The function shiehpwr can be used to compute the power for uniform, normal, exponential, or double exponential data. Moreover, due to the simplicity and good performance, both Lehmann's and Noether's methods has been implemented in commercial software nQuery Advisor 6.0 and East 5. In practice, in addition to the five methods of the sample size calculation for the asymptotic WMW test, other methods can also be applied, for example, Al-Sunduqchi method (implemented in NCSS-PASS software) using the familiar standard two-sample t test sample size formula with simple adjustment factors of $1, 2/3, \text{and } \pi/3$ for uniform, double exponential, and normal data, respectively.[15] However, as always, there are still debates on which is the most appropriate method of sample size calculation to detect the reasonable effect size in clinical research.

CONCLUSION

Five potential sample size calculation methods for Mann-Whitney U test are evaluated based on high precision simulation results. Among the five methods, Shieh's method has the best performance. Lehmann's method is very stable and has second best performance. Doll's method can be used for any linear rank tests which Mann-Whitney U test is a special case. It has good performance too. Noether's method consistently and slightly overestimates the required sample size. If no financial constraints, Neother's method can be regarded as a method with an extra margin of safety. In other words, it is a conservative method. Wang's method does not perform well compared to the other four methods.

Funding: No funding sources
Conflict of interest: None declared
Ethical approval: Not required

REFERENCES

1. Lehmann EL. Testing statistical hypotheses. New York: Chapman and Hall; 1959:10-369.
2. Siegel S, and Castellan NJ. Nonparametric statistics for the behavioural sciences. New York: MeGraw-Hill Inc; 1988:45-85.
3. Lehmann EL. Nonparametric: statistical methods based on ranks. New Jersey: Prentice Hall; 1975:87-98.
4. Noether GE. Sample size determination for some common nonparametric tests. J Amn Stat Assoc. 1987;82(398):645-7.

5. Wang H, Chen B, Chow S-C. Sample size determination based on rank tests in clinical trials. J Biopharma Stat. 2003;13(4):735-51.
6. Shieh G, Jan S, Randles RH. On power and sample size determinations for the Wilcoxon–Mann–Whitney test. J Nonparametric Stat. 2006;18(1):33-3.
7. Doll M, Klein I. Sample size analysis for two-sample linear rank tests. fau discussion papers in economics Friedrich-Alexander-Universität (FAU) Erlangen-Nürnberg, Institute for Economics. Nürnberg. 2006;218-35.
8. Dette H, Brien O, Timothy E. Efficient experimental design for the Behrens-fisher problem with application to bioassay. Am Stat. 2003;58(2):138-43.
9. Happ M, Bathke AC, Brunner E. Optimal sample size planning for the Wilcoxon-Mann-Whitney test. Stat Med. 2019;38(3):363-75.
10. Bürkner PC, Doebler P, Holling H. Optimal design of the Wilcoxon-Mann-Whitney-test. Biom J. 2017;59(1):25-40.
11. Ruymgaart FH. A unified approach to the asymptotic distribution theory of certain midrank statistics. In: Statistique non Parametrique Asymptotique. USA: Springer; 1980: 1-18.
12. Akritas MG, Arnold SF, Brunner E. Nonparametric hypotheses and rank statistics for unbalanced factorial designs. J Am Stat Assoc. 1997; 92(437): 258-65.
13. Zhao YD, Rahardja D, Qu Y. Sample size calculation for the Wilcoxon-Mann-Whitney test adjusting for ties. Stat Med. 2008;27(3):462-8.
14. A language and environment for statistical computing. Available at: https://www.R-project.org/. Accessed on 20 August 2020.
15. Van De WMA. Exact non-null distributions of rank statistics, communications in statistics. Simul Comput. 2001; 30:1011-29.
16. Mollan KR, Trumble IM, Reifeis SA, Ferrer O, Bay CP, Baldoni PL, et al. Exact power of the rank-sum test for a continuous variable. ArXiv. 2019;1901-7.
17. Al-Sunduqchi MS. Determining the appropriate sample size for inferences based on the Wilcoxon statistics. Available at: https://arxiv.org/pdf/1805.12249.pdf. Accessed on 20 August 2020.

Cite this article as: Zhu X. Sample size calculation for Mann-Whitney U test with five methods. Int J Clin Trials 2021;8(3):184-95.