Sample Size Analysis for Two-Sample Linear Rank Tests

Monika Doll
University of Erlangen-Nürnberg

Ingo Klein
University of Erlangen-Nürnberg

Sample Size Analysis for Two-Sample Linear Rank Tests

Monika Doll[a,], Ingo Klein[a,]

[a]*University of Erlangen-Nürnberg, Lange Gasse 20, 90403 Nürnberg, Germany*

April 3, 2018

Abstract

Sample size analysis is a key part of the planning phase of any research. So far, however, limited literature focusses on sample size analysis methods for two-sample linear rank tests, although these methods have optimal properties at different distributions. This paper provides a new sample size analysis method for linear rank tests for location shift alternatives based on score generating functions. Results show a slightly anti-conservative behavior, no severe risk of an occuring circular argument at small to moderate variances of the population's distribution, and good performance compared to alternate sample size analysis methods for the most well-known linear rank test, the Wilcoxon-Mann-Whitney test.

Keywords: Sample Size Analysis, Linear Rank Test, Score Generating Function, Circular Argument

Email addresses: `monika.doll@fau.de` (Monika Doll), `ingo.klein@fau.de` (Ingo Klein)

1. Introduction

Prior to conducting an experiment, a research's planning phase includes an a priori power analysis, respectively sample size analysis. A minimum sample size is required to ensure a reasonable effect size with fixed probabilities for the error of the first (α) and second (β) kind. Accordingly, depending on the population's distribution, a suitable statistical method to test a hypothesis has to be determined in advance. For examining two samples' mean differences, one of the most often applied tests is the so-called t-test. This test is uniformly most powerful unbiased for normally distributed populations at one-sided alternatives, thus it maximizes the power among all unbiased tests in this situation (Lehmann (1959)). The usage of nonparametric tests, to which linear rank tests belong, is often recommended as they provide considerable advantages in efficiency when the assumption of normality is in doubt (Lehmann (1975)). Thus, the Wilcoxon-Mann-Whitney (WMW) test is asymptotically optimal or locally optimal in the set of all linear rank tests at a Logistic distribution (Mann and Whitney (1947), Wilcoxon (1945)) and Mood's Median test is asymptotically optimal at a double exponential or Laplace distribution (Mood (1954), Hájek et al. (1999)). Moreover, in case of a normal distribution the van der Waerden test is an asymptotically optimal linear rank test (van der Waerden (1952), Gibbons (1971)). Consequently, depending on the population's distribution, one test will be preferred over another, while linear rank tests provide (asymptotically) optimal properties not only when the assumption of normality is in doubt. However, although extensive literature exists that is concered with sample size analysis for parametric tests, only limited literature can be found regarding sample size analysis for linear rank tests. Herein, the most often considered linear rank test is the WMW test. This is not surprising as it is probably the most well-known nonparametric test. However, as other linear rank tests are preferrable in case a population is not logistically distributed and few literature exists providing sample size analysis methods for general linear rank tests, enhanced attention on this issue is needed. As sample size analysis is of central importance for experimental research planning, this paper introduces a new sample size analysis method for general linear rank tests on location shifts. The proposed method is based on linear rank tests' asymptotic normality, while mean and variance can be expressed by score generating functions (Hájek et al. (1999)). Comparisons to the most common alternate tests on normal

2

distributions for the WMW statistic display its good performance. As however this method is based on linear rank test statistics' asymptotical properties, the risk of an underlying circular argument is apparent. This means, that for a sufficiently good approximation of a test statistic's exact distribution, a specific sample size is needed. Nonetheless, at an a priori power analysis, a specific sample size will be determined. Thus, by making use of a statistic's asymptotical properties, the risk exists to calculate a minimum required sample size at which the sample size analysis method is not operating sufficiently well. Therefore, to check for the risk of an apparent circular argument, sample sizes were determined using simulated or exact null-distributions of the regarded test statistics. Sample sizes provided by the method based on the test statistics' asymptotic distributions compared to those using their exact or simulated distributions are not found to differ severly at distributions with small to moderate variance.

This paper is structured as follows: Section 2 gives a short literature review on sample size analysis methods for two-sample linear rank tests on location shifts. Section 3 presents the asymptotic distribution of two-sample linear rank tests based on score generating functions and introduces the new sample size analysis method. Section 4 displays the results of this method for different distributions and discusses the findings. In Section 5, a comparison of the sample size analysis method based on score generating functions to alternative methods for the WMW test is shown. Section 6 summarizes and provides concluding remarks.

2. Literature Review

This section provides a short literature review on sample size analysis for two-sample linear rank tests for location shifts of continuous distributions. Methods that will be used in Section 5 for comparison will be presented in more detail. Most research is done referring the WMW test, as it probably is the most often applied linear rank test. However, almost no literature on sample size analysis methods considers alternate linear rank tests. Thus, first, this literature review's focus will be on sample size analysis methods for the WMW test, while afterwards attention will be given to methods for alternate or general linear rank tests for location shifts.

The Wilcoxon-Mann-Whitney test (Wilcoxon (1945), Mann and Whitney (1947)) has the statistic

$$S_W = \sum_{i=1}^{n} a_{n,m}(R_i)$$

with scores $a_{n,m}(i) = i$, $i = 1, 2, \ldots, n + m$ and R_i being the respective scores of X_i in the combined sample of X, Y, while $X_1, \ldots, X_m$ and $Y_1, \ldots Y_n$ are independent random samples. Sample size analyses can be based on a test statistic's exact distribution or its asymptotic distribution. The WMW test statistic's exact distribution requires computationally intensive total enumeration. Hilton and Mehta (1993) introduced a network algorithm based on a conservative lower bound of the critical value to generate conditional power followed by applying an iterative process to determine the required sample size. Their algorithm is however restricted to ordered categorical data and feasable only for small sample sizes. To solve the computational burden, a half interval search algorithm was described by Wan et al. (2009), and an asymptotic approach of Hilton and Mehta (1993)'s algorithm by Rabbee et al. (2003). A method based on probability generating functions was proposed by van de Wiel (2000), which has feasable computational time for sample sizes smaller than 40. However, all of the previously mentioned methods have the disadvantage of being computationally and time intensive, which can be avoided by using a test statistic's asymptotical properties. Al-Sunduqchi (1990) proposed to multiply the minimum required sample size for the parametric t-test by 1.156, which is reasoned by the WMW test's lowest bound of asymptotic relative efficiency to the t-test being 0.864 (Lehmann (1975)). Assuming the WMW test statistic to be approximately normal, Noether (1987) suggested calculating the minimum required sample size by

$$n = \frac{(\lambda_{1-\alpha} + \lambda_{1-\beta})^2}{6(p'' - 0.5)^2},$$

with $p'' = P(X < Y)$. The parameter λ denotes the quantile of the standard normal distribution. The central assumption of Noether (1987) is that alternatives do not differ much from the null hypothesis, such that the test statistic's variance under the alternative can be assumed to equal the statistic's variance under the null hypothesis of no location shift for moderate sample sizes. By not applying this assumption, an estimate of the statistic's variance under the alternative is required. Birnbaum and Klose (1957) suggested to use the WMW

test statistic's variance's lower, respectively upper bound for calculating the required sample size. Inspired by Noether (1987)'s as well as Birnbaum and Klose (1957)'s suggestions, Vollandt and Horn (1997) proposed an exact large sample method for calculating the WMW test statistic's variance's upper bound via solving the following inequalities numerically

for $3/8 < p < 1/2$:

$$\frac{1/2 - p - \lambda_{1-\alpha}[(2n+1)/(12n^2)]^{1/2}}{([17n^2 - 20n + 6]/[12(2n-1)^3])^{1/2}} \geq \lambda_{1-\beta},$$

for $0 < p \leq 3/8$:

$$\frac{1/2 - p - \lambda_{1-\alpha}[(2n+1)/12n^2]^{1/2}}{[(n\delta_1 + \delta_2)/(3n^2)]^{1/2}} \geq \lambda_{1-\beta},$$

with $p = P(X > Y)$, $\delta_1 = 1/2 - 6\delta^2 + (2\delta)^{3/2}$, $\delta_2 = 1/4 + 3\delta^2 - (2\delta)^{3/2}$, and $\delta = 1/2 - p$. While Birnbaum and Klose (1957)'s upper and lower bounds of the WMW test statistic's variance is obtained in terms of p, Vollandt and Horn (1997)'s upper bound depends on the variance's maximum for the two ranges of p. Moreover, their suggestion is valid for any alternative. Based on the assumption of the test statistics' asymptotic normality, Lehmann (1953) proposed to solve numerically

$$\beta = 1 - \Phi\left(\frac{\lambda_\alpha \frac{mn(N+1)}{12} - mn(p_1 - 0.5)}{\sqrt{\sigma_W^2}}\right),$$

with

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

while

$$p_1 = P(X_1 < Y_1),$$

$$p_2 = P(X_1 < Y_1 \text{ and } X_1 < Y_2),$$

and

$$p_3 = P(X_1 < Y_1 \text{ and } X_2 < Y_1).$$

Here, X_1, X_2 are independently distributed with continuous distribution F and Y_1, Y_2 are independently distributed with continuous distribution G. Wang et al. (2003) proposed to estimate the variance pointwise using pilot data. Shieh et al. (2006) provided an exact variance large-sample method. Whenever pilot data is available, bootstrapping techniques can

be applied (Efron (1979)). Therefore, Hamilton and Collings (1991) avoided the problem of assuming normality and presented a bootstrapping method for estimating the required sample size for nonparametric tests for location shifts. Comparing their method to the one of Noether (1987) in a simulation study displayed similar results for both methods. Based on Hamilton and Collings (1991)'s suggestion, Chakraborti et al. (2006) proposed further methods for sample size analysis for the WMW test combining bootstrapping and the formula of Noether (1987) by estimating p'' via $\hat{p}'' = W/mn$, with W being the WMW statistic estimated from the bootstrapped samples. As another method they proposed to estimate p'' using linearly smoothed empirical cumulative distribution functions (ecdf) of two pilot samples via $\widehat{p_x} = \int F_X(x)dF_X(x-\delta)$ with $F_X(x)$ being the piecewise linear ecdf's from the X-sample and $F_X(x-\delta)$ being the piecewise linear ecdf from the Y-sample. Their results showed that compared to the bootstrapping method of Hamilton and Collings (1991) at different underlying population distributions, their proposals work faster and are superior to Noether (1987)'s suggestion. Moreover, they found that Noether (1987)'s formula provides larger minimum required sample sizes than those caluclated using the exact method of van de Wiel (2000), by comparing them for sample sizes smaller than $n = m = 20$. Taking cost considerations into account, Guo (2012) proposed a sample size formula based on Chakraborti et al. (2006)'s proposal on estimating p'' of Noether (1987)'s formula. Likewise relying on the use of pilot data, Divine et al. (2010) presented a so-called 'exemplary dataset method' for calculating the required minimum sample size for the WMW test statistic. This method is similar to the one presented by Zhao et al. (2008), whose method can be applied at continuous as well as ordered categorical data by accounting for ties while assuming equal variances under the null hypothesis and the alternative hypothesis. For continuous distributions, Zhao et al. (2008)'s suggestion equals Noether (1987)'s proposal.

Up to now, the focus was on sample size analysis methods for the WMW test. However, limited literature could be found regarding samples size analysis for alternate linear rank tests on location shifts. Chakraborti et al. (2006) displayed a general sample size analysis formula for linear rank tests for location shifts along the lines of Noether (1987)

$$n = \frac{[\lambda_{1-\alpha} + \lambda_{1-\beta}]^2}{(\mu_\delta - 0.5)^2}(\int_0^1 J^2(u)du - [\int_0^1 J(u)du]^2)$$

6

with

$$\mu_\delta = \int_{-\infty}^{\infty} J[0.5F_x(t) + 0.5F_x(t - \delta)]dF_x(t),$$

where $J(\cdot)$ being the usual score function and $m = n$. This 'Noether-like' formula considers the underlying distribution through the mean functional μ_δ (see Gibbons and Chakraborti (2003)).

To summarize, the available literature on sample size analysis for linear rank tests on location shifts is limited and further approaches are needed. Therefore, this paper proposes a new method to satisfy this need. This new sample size analysis method, which is based on score generating functions, will be introduced in the next section.

3. Asymptotic Sample Size Analysis Method

3.1. Asymptotic Distribution of Two-Sample Linear Rank Statistics

This section will show that linear rank test statistics based on score generating functions are asymptotically normally distributed.

Let $X_1, \ldots, X_n$ be stochastically independent, identically distributed (iid) with cumulative distribution function (cdf) F and $Y_1, \ldots, Y_m$ be iid with cdf G. We consider independent samples such that X_i and Y_j are stochastically independent for $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, m$. R_i is the rank of X_i in the combined sample for $i = 1, 2, \ldots, n$. Then

$$S_a = \sum_{i=1}^{n} a_{n,m}(R_i)$$

with scores $a_{n,m}(i)$, $i = 1, 2, \ldots, n + m$, defines a two-sample linear rank test statistic. It can be used to test $H_0 : F = G$ against one-sided alternatives concerning the stochastic ordering of F and G or the two-sided alternative of F and G being different. F and G shall be continuous such that no ties appear. These tests have a long tradition in statistics and belong to the standard tools of applied statisticians. As mentioned in the previous section, the probably most famous two-sample linear rank test statistic is the WMW statistic with scores $a_{n,m}(i) = i$. This statistic counts the amount of X_i being smaller or equal to Y_j for $j = 1, 2, \ldots, n$. Other choices for the scores $a_{n,m}(i)$ lead to the van der Waerden (van der Waerden (1952)) or Mood's Median statistic (Mood (1954)). Under the null hypothesis the ranks are

independent of the population's distribution F. However, already the null distribution of S_a requires combinatorics, while the computational effort increases exponentially with n, as a higher-dimensional integral has to be solved (see e.g. Haynam and Govindarajulu (1966)). Thus, as already mentioned in the previous section, sample size analysis proposals using a test statistic's exact distribution have the disadvantage of only working sufficiently fast and computationally feasable for sample sizes smaller than 40, as of today. Hence, practically feasable solutions should use a linear rank test statistic's asymptotic properties instead.

Therefore, let f_0 be a density with mean 0 and variance σ^2, f_0' has to exist, and the Fisher-information $I(f_0)$ will be assumed to be finite. We consider two independent samples of size n and m from f_0. $X_1, X_2, \ldots, X_n$ are iid with $f_0(x - \delta\sigma)$ and $Y_1, Y_2, \ldots, Y_m$ are iid with $f_0(x)$. $R_1, R_2, \ldots, R_n$ are the ranks of $X_1, X_2, \ldots, X_n$ in the combined sample. δ will be called the effect size being tested.

Set

$$\varphi(u; f_0) = -\frac{f_0'(F_0^{-1}(u))}{f_0(F_0^{-1}(u))}, \quad u \in [0, 1].$$

Then, we consider a linear rank test with scores $a_{n,m}$ to test a right-sided alternative

$$H_0 : \delta \leq 0 \quad \text{against} \quad H_A : \delta > 0,$$

while the corresponding test statistic is

$$S_{n,m} = \sum_{i=1}^{n} a_{n,m}(R_i) - n\frac{1}{N+1}\sum_{i=1}^{N} a_{n,m}(i).$$

If there is a square integrable function φ with

$$\lim_{N \to \infty} \int_0^1 (a_N(1 + [uN]) - \varphi(u))^2 du = 0$$

while $[uN]$ denotes the largest integer not exceeding $N = n + m$, with $\overline{\varphi} = \int_0^1 \varphi(u) du$, then $S_{n,m}$ is asymptotically normal

1. under the null hypothesis with mean 0 and variance

$$\sigma_{n,m}^2 = \frac{nm}{n+m} \int_0^1 (\varphi(u) - \overline{\varphi})^2 du$$

   or

$$Var(S_{n,m}) = \frac{nm}{(n+m-1)}\frac{1}{n+m}\sum_{i=1}^{N}(a_{n,m}(i) - \overline{a}_{n,m})^2$$

2. under the alternative with mean

$$\mu_{n,m}(\delta) = \delta\sigma\frac{nm}{n+m}\int_0^1 \varphi(u)\varphi(u;f_0)du$$

and the same variance as under the null hypothesis (see Hájek et al. (1999)).

3.2. Sample Size Analysis Method based on Score Generating Functions

The asymptotical normal distribution of $S_{n,m}$ under the existence of a function φ can be used to determine the minimum sample size.

Theorem 3.1. *Let α and β be fixed sizes of the probabilities of the errors of the first and the second kind, δ a fixed effect size, and σ^2 the variance of the underlying density f_0. In case of equal sample sizes, for a decision that the null hypothesis is correct with error probability α or the alternative holds with error probability β, a two-sample linear rank test with score generating function φ*

1. *for $H_0: \ \delta \leq 0$ against $H_A: \ \delta > 0$ requires a minimum sample size of*

$$n = 2\left(\frac{\lambda_{1-\alpha} - \lambda_\beta}{\delta\sigma}\right)^2 \frac{\int_0^1 (\varphi(u) - \overline{\varphi})^2 du}{\left(\int_0^1 \varphi(u)\varphi(u;f_0)du\right)^2}.$$

2. *For $H_0: \ \delta = 0$ against $H_A: \ \delta \neq 0$ the minimum sample size n is given by the solution of*

$$
\begin{aligned}
\beta \ = \ & \Phi\left(\lambda_{1-\alpha/2} - \sqrt{\frac{n}{2}}\delta\sigma\frac{\int_0^1 \varphi(u)\varphi(u;f_0)du}{\sqrt{\int_0^1 (\varphi(u) - \overline{\varphi})^2 du}}\right) \\
& + \Phi\left(\lambda_{1-\alpha/2} + \sqrt{\frac{n}{2}}\delta\sigma\frac{\int_0^1 \varphi(u)\varphi(u;f_0)du}{\sqrt{\int_0^1 (\varphi(u) - \overline{\varphi})^2 du}}\right) - 1.
\end{aligned}
$$

Proof:

1. From the asymptotical results we get the asymptotical power function

$$\pi(\delta) \approx 1 - \Phi\left(\lambda_{1-\alpha} - \frac{\mu_{n,m}}{\sigma_{n,m}}\right).$$

9

Inserting the asymptotical moments gives

$$\pi(\delta) \approx 1 - \Phi\left(\lambda_{1-\alpha} - \sqrt{\frac{nm}{n+m}}\delta\sigma\frac{\int_0^1 \varphi(u)\varphi(u;f_0)du}{\sqrt{\int_0^1 (\varphi(u)-\overline{\varphi})^2 du}}\right).$$

For a known density f_0, fixing α, β, and δ leads to an equation for the sample sizes n, m by solution of

$$\beta = \Phi\left(\lambda_{1-\alpha} - \sqrt{\frac{nm}{n+m}}\delta\sigma\frac{\int_0^1 \varphi(u)\varphi(u;f_0)du}{\sqrt{\int_0^1 (\varphi(u)-\overline{\varphi})^2 du}}\right).$$

In the special situation of equal sample sizes $n = m$ we get

$$\Phi^{-1}(\beta) = \lambda_{1-\alpha} - \sqrt{\frac{n}{2}}\delta\sigma\frac{\int_0^1 \varphi(u)\varphi(u;f_0)du}{\sqrt{\int_0^1 (\varphi(u)-\overline{\varphi})^2 du}}.$$

Solving with respect to sample size leads to

$$n = \left(\sqrt{2}\frac{\lambda_{1-\alpha}-\lambda_\beta}{\delta\sigma}\frac{\sqrt{\int_0^1 (\varphi(u)-\overline{\varphi})^2 du}}{\int_0^1 \varphi(u)\varphi(u;f_0)du}\right)^2.$$

2. For the two-sided alternative we get the asymptotical power function

$$\pi(\delta) \approx 1 - \Phi\left(\lambda_{1-\alpha/2} - \sqrt{\frac{n}{2}}\delta\sigma\frac{\int_0^1 \varphi(u)\varphi(u;f_0)du}{\sqrt{\int_0^1 (\varphi(u)-\overline{\varphi})^2 du}}\right)$$

$$+ \Phi\left(-\lambda_{1-\alpha/2} - \sqrt{\frac{n}{2}}\delta\sigma\frac{\int_0^1 \varphi(u)\varphi(u;f_0)du}{\sqrt{\int_0^1 (\varphi(u)-\overline{\varphi})^2 du}}\right).$$

Due to the symmetry of the standard normal distribution the result follows immediately.

$\square$

The following examples demonstrate the introduced sample size analysis method for the WMW statistic in case of a normal distribution and a $t(3)$ distribution.

Example 3.1. *Let f_0 be the standard normal density with $\varphi(u;f_0) = \Phi^{-1}(u)$, $u \in [0,1]$ and Wilcoxon scores generating function $\varphi(u) = u - 1/2$, $u \in [0,1]$. With*

$$\int_0^1 (\varphi(u)) - \overline{\varphi})^2 du = \int_0^1 (u-1/2)^2 du = 1/12$$

and numerical integration we get

$$\int_0^1 \varphi(u; f_0)\varphi(u)du = 0.2821.$$

By setting $\alpha = \beta = 0.05$, $\delta = 0.8$, and due to $\sigma^2 = 1$ we get

$$n = \left(\sqrt{2}\frac{2 \cdot 1.645}{0.8 \cdot 1}\frac{\sqrt{1/12}}{0.2821}\right)^2 = 35.4.$$

Thus, the minimum required sample size is 36. For the two-sided alternative n has to be calculated numerically. Solving

$$0.05 = \Phi\left(1.96 - \sqrt{\frac{n}{2}}0.8\frac{0.2821}{\sqrt{1/12}}\right) + \Phi\left(1.96 + \sqrt{\frac{n}{2}}0.8\frac{0.2821}{\sqrt{1/12}}\right) - 1$$

leads to $n = 42.53$. Thus, the minimum required sample size is 43.

Example 3.2. *Let f_0 be the $t(k)$ density with variance $\sigma^2 = k/(k-2)$ for $k > 2$. Then it holds*

$$-\frac{d\ln f_0(x)}{dx} = \frac{k+1}{k}\frac{x}{1+x^2/k}, \quad x \in \mathbb{R}$$

such that

$$\varphi(u; f_0) = \frac{k+1}{k}\frac{t^{-1}(u; k)}{1+t^{-1}(u; k)^2/k}, \quad u \in [0, 1]$$

with the quantile function $t^{-1}(.; k)$. Consider the special case $k = 3$. Again we consider the WMW test. Numerical integration gives

$$\int_0^1 \varphi(u; f_0)\varphi(u)du = 0.2297.$$

By setting $\alpha = \beta = 0.05$, $\delta = 0.8$, and due to $\sigma^2 = 3$ we get

$$n = \left(\sqrt{2}\frac{2 \cdot 1.645}{0.8\sqrt{3}}\frac{\sqrt{1/12}}{0.2297}\right)^2 = 17.80.$$

Thus, the minimum required sample size is 18. For the two-sided alternative, solving

$$0.05 = \Phi\left(1.96 - \sqrt{\frac{n}{2}}0.8\sqrt{3}\frac{0.2297}{\sqrt{1/12}}\right) + \Phi\left(1.96 + \sqrt{\frac{n}{2}}0.8\sqrt{3}\frac{0.2297}{\sqrt{1/12}}\right) - 1$$

leads to $n = 21.38$. Thus, the minimum required sample size is 22.

This means that the WMW test needs a remarkably smaller sample size for a $t(3)$ distribution than for the normal distribution to cover a decision for fixed α, β, and δ.

Furthermore, for a known density f_0, we can consider the corresponding locally optimal linear rank test with scores $a_{n,m}(i) = E(\varphi(U^i; f_0))$ or the asymptotically optimal rank test with scores $a_{n,m}(i) = \varphi(i/(2n+1); f_0)$, with U^i being ranks of indendent random variables, each uniformly distributed over $(0, 1)$. The following corollary shows the sample size analysis in this case.

Corollary 3.1. *Let α and β be fixed probabilities of the error of the first and the second kind, δ be a fixed effect size and σ^2 the variance of the underlying density f_0. In case of equal sample sizes a two-sample linear rank test with generating function $\varphi(.; f_0)$*

1. *for $H_0 : \delta \leq 0$ against $H_A : \delta > 0$ needs at least a sample size of*

$$n = 2 \left(\frac{\lambda_{1-\alpha} - \lambda_\beta}{\delta\sigma} \right)^2 \frac{1}{\left(\int_0^1 \varphi(u; f_0)^2 du \right)}$$

   for a decision that the null hypothesis is correct with error probability α or the alternative holds with error probability β.

2. *For $H_0 : \delta = 0$ against $H_A : \delta \neq 0$ the necessary sample size n is given by the solution of*

$$\begin{aligned} \beta &= \Phi \left(\lambda_{1-\alpha/2} - \sqrt{\frac{n}{2}}\delta\sigma\sqrt{\int_0^1 \varphi(u; f_0)^2 du} \right) \\ &+ \Phi \left(\lambda_{1-\alpha/2} + \sqrt{\frac{n}{2}}\delta\sigma\sqrt{\int_0^1 \varphi(u; f_0)^2 du} \right) - 1 \end{aligned}$$

   for a decision that the null hypothesis is correct with error probability α or the alternative holds with error probability β.

This corollary can e.g. be applied to the van der Waerden test statistic for a normal distribution. The corresponding two-sample test is asymptotically optimal.

Example 3.3. *Let f_0 be the standard normal density with $\varphi(u; f_0) = \Phi^{-1}(u)$, $u \in [0, 1]$ and $a_{n,m}(i) = \Phi^{-1}(i/(n+m-1))$, $i = 1, 2, \ldots, n$, the van der Waerden scores with generating function $\varphi(u) = \varphi(u; f_0)$, $u \in [0, 1]$. It is*

$$\int_0^1 \varphi(u; f_0)^2 du = \int_0^1 (\Phi^{-1}(u))^2 du = \sigma^2 = 1.$$

12

By setting $\alpha = \beta = 0.05$, $\delta = 0.8$, and due to $\sigma^2 = 1$ we get

$$n = \left(\sqrt{2} \frac{2 \cdot 1.645}{0.8} \right)^2 = 33.83.$$

Thus, the minimum required sample size is 34. For the two-sided alternative we have to solve

$$0.05 = \Phi \left(1.96 - \sqrt{\frac{n}{2}} 0.8 \right) + \Phi \left(1.96 + \sqrt{\frac{n}{2}} 0.8 \right) - 1$$

numerically for n. This results in $n = 40.61$. Thus, the minimum required sample size is 41.

This proposed sample size analysis method is suitable for arbitrary linear rank tests for location shift alternatives on various distributions.

4. Sample Size Analysis for Two-Sample Linear Rank Tests

This section presents the results of the new sample size analysis method at different conditions. The most often applied linear rank statistics will be discussed. To these belong the WMW, Mood's Median, and the van der Waerden statistic. The corresponding two-sample linear rank tests are asymptotically optimal for the Logistic, the Laplace, and the normal distribution. To provide results for a situation where none of these tests is optimal, the $t(3)$ distribution was also used for comparison next to the ones mentioned previously. However, whether asymptotical results hold depends on the sample size. Therefore, a circular argument could occure at determining the necessary sample size using a test statistic's asymptotical properties. Thus, we examined whether a test statistic's asymptotical properties hold for the calculated sample sizes by comparing these sample sizes to those calculated using a test statistic's exact distribution. The exact power analysis was performed by simulating the distribution under the null (if necessary) as well as the alternative hypothesis. The null distribution of Mood's Median test is given by the hypergeometric distribution. For the WMW test's null distribution a recurrence relation holds, while under the alternative a higher dimensional integral has to be solved. Thus, at analyzing the 12 conditions (three test statistics at four different distributions), we applied as effect sizes δ the values proposed by Cohen (1969) for small ($\delta = 0.2$), medium ($\delta = 0.5$), and large ($\delta = 0.8$) effects. Additionally, we considered extremly large effect sizes ($\delta = 1.2$, $\delta = 1.6$), which provide very small minimum required sample sizes to receive detailed insights into whether the risk of an

13

occuring circular argument is apparent. The probability for the error of the first kind has been fixed at $\alpha = 5\%$. For the probabilities of the second kind (β) we applied 5%, 10%, and 20%.

In the following, the results of the simulations at all conditions will be displayed separately for one-sided and two-sided alternatives. Moreover, n_e refers to those sample sizes that were calculated using a test statistic's exact distribution and n_a refers to those sample sizes that were calculated using the new sample size analysis method based on linear rank test statistics' asymptotical properties. For the normal distribution with $\sigma^2 = 1$ the generating function is $\varphi(u; f_0) = \Phi^{-1}(u)$, $u \in (0, 1)$. For the Logistic distribution with variance $\sigma^2 = \pi^2/3$ it is $\varphi(u; f_0) = u - 1/2$ for $u \in (0, 1)$. For the Laplace distribution with variance $\sigma^2 = 2$ the generating function is $\varphi(u; f_0) = \text{sign}(u - 1/2)$. Finally, the variance of a t distribution with three degrees of freedom is $\sigma^2 = 3$ and

$$\varphi(u; f_0) = \frac{4}{3} \frac{t^{-1}(u; 3)}{1 + t^{-1}(u; 3)^2/3}, \quad u \in [0, 1]$$

with quantile function $t^{-1}(.; 3)$.

4.1. Wilcoxon-Mann-Whitney statistic

For the WMW statistic the scores generating function is $\varphi(u) = u - 1/2$ for $u \in (0, 1)$. Therefore,

$$\int_0^1 (\varphi(u) - \overline{\varphi})^2 du = 0.0833.$$

The sample size analysis results using the WMW statistic for the one-sided alternatives are shown in Table 1, while for the two-sided alternatives the results are presented in Table 2.

Results displayed in Table 1 and Table 2 show that the new method operates relatively precise. Overall, a slightly anti-conservative behavior can be observed, which means that the sample size analysis method introduced in Section 3.2 slightly underestimates the minimum required sample size to ensure an effect size with fixed probabilities for the errors of the first and second kind. The calculated sample size is heavily influenced by the chosen size of the effect. Thus, the higher the effect size, the smaller the minimum required sample size. Moreover, the higher the fixed probability of the error of the second kind (β), the smaller the

14

Table 1: Sample size analysis results using the Wilcoxon-Mann-Whitney statistic for one-sided alternatives

		Normal		Logistic		Laplace		$t(3)$	
δ	β	n_e	n_a	n_e	n_a	n_e	n_a	n_e	n_a
0.2	0.05	568	567	495	494	368	361	287	285
	0.10	450	449	395	391	293	286	228	226
	0.20	324	324	283	282	210	207	162	163
0.5	0.05	92	91	81	79	62	58	48	46
	0.10	72	72	63	63	51	46	39	37
	0.20	53	52	47	46	37	33	29	27
0.8	0.05	37	36	33	31	27	23	21	18
	0.10	29	29	26	25	22	18	17	15
	0.20	22	21	19	18	17	13	13	11
1.2	0.05	17	16	16	14	15	11	12	8
	0.10	14	13	13	11	12	8	9	7
	0.20	10	9	9	8	9	6	7	5
1.6	0.05	10	9	9	8	9	6	7	5
	0.10	9	8	9	7	7	5	6	4
	0.20	7	6	6	5	6	4	5	3

minimum required sample size. Regarding the deviation between n_a to n_e no severe difference is observable for the one- and two-sided alternatives. However, the deviation is heavily influenced by the population's distribution's variance. This means that for distributions with a higher variance (e.g. the Laplace and the $t(3)$ distribution) the deviation between n_a to n_e is larger than for distributions with a small to moderate variance (e.g. the normal and the Logistic distribution). Moreover, as the effect size increases, thus the calculated minimum required sample size decreases, the deviation becomes larger. For the WMW test statistic, the results show that over both alternatives, for the normal and the Logistic distribution the deviation of n_a to n_e has a maximum of 1.34% at $\delta = 0.2$, 2.75% at $\delta = 0.5$, and 6.45% at $\delta = 0.8$. For the extremely large effect sizes, at which the minimum required

Table 2: Sample size analysis results using the Wilcoxon-Mann-Whitney statistic for two-sided alternatives

		Normal		Logistic		Laplace		$t(3)$	
δ	β	n_e	n_a	n_e	n_a	n_e	n_a	n_e	n_a
0.2	0.05	683	681	604	596	438	434	347	343
	0.10	555	551	479	480	357	351	280	277
	0.20	410	411	368	358	268	262	210	207
0.5	0.05	112	109	97	95	76	70	59	55
	0.10	89	89	79	77	62	57	47	45
	0.20	66	66	59	58	47	42	36	34
0.8	0.05	45	43	39	38	33	28	25	22
	0.10	36	35	33	30	27	22	20	18
	0.20	27	26	24	23	20	17	16	13
1.2	0.05	20	19	19	17	17	13	14	10
	0.10	17	16	16	14	14	10	11	8
	0.20	13	12	11	10	11	8	8	6
1.6	0.05	13	11	11	10	11	7	10	6
	0.10	11	9	10	8	10	6	8	5
	0.20	8	7	8	6	8	5	7	4

sample size is small, the maximum deviation is 18.18% at $\delta = 1.2$ and 33.33% at $\delta = 1.6$. Over both alternatives, for those distributions with larger variances, thus the Laplace and $t(3)$ distribution, the maximum deviation of n_a to n_e is 2.45% at $\delta = 0.2$, 12.12% at $\delta = 0.5$, and 30.77% at $\delta = 0.8$. For the extremely large effect sizes the maximum deviation is 50% at $\delta = 1.2$ and 75% at $\delta = 1.6$.

4.2. Mood's Median statistic

Mood's Median test is known to have poor relative efficiency in comparison to alternate linear rank tests (Büning and Trenkler (1994), Mood (1954)), has low power at small sample sizes, and is even suggested to be 'retired from general use' (Freidlin and Gastwirth (2000)). Despite this, this test is one of the most well-known linear rank tests, is presented in many sta-

tistical textbooks, and included in most statistical software packages (Freidlin and Gastwirth (2000)). Therefore, we decided to present the new sample size analysis method's performance for this test, although we are aware of its slow convergence rate to the normal distribution (Klein (1978)).

For Mood's Median statistic the scores generating function is $\varphi(u) = \text{sign}(u - 1/2)$, for $u \in (0, 1)$. Therefore,

$$\int_0^1 (\varphi(u) - \overline{\varphi})^2 du = 1.$$

The sample size analysis results for Mood's Median test for the one-sided alternatives are shown in Table 3 and for the two-sided alternatives in Table 4.

Table 3: Sample size analysis results using Mood's Median statistic for one-sided alternatives

		Normal		Logistic		Laplace		$t(3)$	
δ	β	n_e	n_a	n_e	n_a	n_e	n_a	n_e	n_a
0.2	0.05	836	850	645	658	290	271	325	334
	0.10	662	673	509	521	229	215	255	265
	0.20	474	486	366	376	164	155	180	191
0.5	0.05	127	136	98	106	53	44	50	54
	0.10	100	108	75	84	40	35	38	43
	0.20	70	78	52	61	29	25	25	31
0.8	0.05	47	54	37	42	22	17	18	21
	0.10	36	43	28	33	18	14	14	17
	0.20	24	31	18	24	9	10	8	12
1.2	0.05	19	24	14	19	11	8	7	10
	0.10	14	19	9	15	7	6	6	8
	0.20	8	14	6	11	4	5	3	6
1.6	0.05	9	14	7	11	6	5	4	6
	0.10	6	11	5	9	4	4	3	5
	0.20	4	8	3	6	3	3	3	3

Table 4: Sample size analysis results using Mood's Median statistic for two-sided alternatives

		Normal		Logistic		Laplace		$t(3)$	
δ	β	n_e	n_a	n_e	n_a	n_e	n_a	n_e	n_a
0.2	0.05	1008	1021	779	790	345	325	392	401
	0.10	812	826	625	638	282	263	312	325
	0.20	607	617	466	478	212	197	233	243
0.5	0.05	157	164	121	127	66	52	61	65
	0.10	124	133	95	103	52	43	48	52
	0.20	88	99	68	77	37	32	34	39
0.8	0.05	59	64	45	50	29	21	24	26
	0.10	45	52	36	40	24	17	20	21
	0.20	32	39	26	30	16	13	13	16
1.2	0.05	24	29	20	22	14	10	11	12
	0.10	20	23	15	18	11	8	9	10
	0.20	14	18	9	14	7	6	5	7
1.6	0.05	13	16	11	13	9	6	6	7
	0.10	10	13	8	10	7	5	5	6
	0.20	7	10	5	8	5	4	4	4

Table 3 and Table 4 show a good performance of the new sample size analysis method for the one-sided as well as for the two-sided alternatives, especially considering Mood's Median test's slow convergence rate to normality (Klein (1978)). Overall, except for the Laplace distribution, Mood's Median test requires a higher minimum sample size to cover a decision that the null hypothesis is correct with error probability α or the alternative holds with error probability β than does the WMW test. This can be explained by its smaller relative efficiency in comparison to the WMW test (see e.g. Freidlin and Gastwirth (2000)) and its slower convergence rate to the normal distribution (Klein (1978)). In general, a conservative behavior can be observed. The only exception are the sample size analysis results for the Laplace distribution, while here the smallest minimum sample size is required for this test to

ensure an effect. Deviations between n_a and n_e are larger for Mood's Median test than for the WMW test. Overall, the deviations are larger the higher the population's distribution's variance. Thus, over both alternatives, for the normal and the Logistic distribution n_a to n_e has a maximum deviation of 2.66% at $\delta = 0.2$, 14.75% at $\delta = 0.5$, and 22.58% at $\delta = 0.8$. For the extremely large effect sizes, at which the minimum required sample size is small, the maximum deviation is 45.45% at $\delta = 1.2$ and 50% at $\delta = 1.6$. For the distributions with higher variance, thus, the Laplace and the $t(3)$ distribution, the deviations are larger. For the $t(3)$ distribution a conservative behavior, thus an overestimation of the minimum required sample size can be observed, which is in line with results for the normal and the Logistic distribution. Over both alternatives, the maximum deviation between n_a and n_e is 5.75% at $\delta = 0.2$, 19.35% at $\delta = 0.5$, 33.33% at $\delta = 0.8$, 50% at $\delta = 1.2$, and 40% at $\delta = 1.6$. For the Laplace distribution, for which this test is the asymptotically optimal one, an anti-conservative behavior, thus an underestimation of the minimum required sample size can be observed. In this case, over both alternatives, the maximum deviation between n_a and n_e is 7.61% at $\delta = 0.2$, 26.92% at $\delta = 0.5$, and 41.17% at $\delta = 0.8$. For the extremely large effect sizes the maximum deviation is 37.5% at $\delta = 1.2$, and 50% at $\delta = 1.6$. These results show, that despite the slow convergence rate, the new sample size analysis also can be applied for Mood's Median test without any adjustments. To achieve a better approximation, a continuity correction could be implemented for the sample size analysis for Mood's Median test.

4.3. Van der Waerden statistic

For the van der Waerden statistic the scores generating function is $\varphi(u) = \Phi^{-1}(u)$, for $u \in (0,1)$. Therefore,

$$\int_0^1 (\varphi(u) - \overline{\varphi})^2 du = \int_0^1 \Phi^{-1}(u)^2 du = 1.$$

The sample size analysis results for the van der Waerden test for the one-sided alternatives are shown in Table 5 and for the two-sided ones in Table 6.

Results presented in Table 5 and Table 6 for the van der Waerden test show a slightly anti-conservative behavior, displayed by a slight underestimation of the required sample size.

Table 5: Sample size analysis results using the van der Waerden statistic for one-sided alternatives

δ	β	Normal		Logistic		Laplace		$t(3)$	
		n_e	n_a	n_e	n_a	n_e	n_a	n_e	n_a
0.2	0.05	540	542	519	517	428	425	331	331
	0.10	431	429	409	409	339	337	262	262
	0.20	310	310	295	296	242	243	190	189
0.5	0.05	89	87	85	83	71	68	55	53
	0.10	69	69	67	66	57	54	43	42
	0.20	52	50	49	48	42	39	31	31
0.8	0.05	35	34	34	33	30	27	23	21
	0.10	28	27	27	26	25	22	18	17
	0.20	21	20	20	19	18	16	14	12
1.2	0.05	17	16	16	15	15	12	12	10
	0.10	14	12	13	12	12	10	10	8
	0.20	10	9	10	9	9	7	7	6
1.6	0.05	10	9	10	9	10	7	8	6
	0.10	8	7	8	7	8	6	7	5
	0.20	6	5	6	5	6	4	5	3

The general behavior is similar to the one of the WMW test, while results are found to be more precise. Thus, for the van der Waerden test statistic, the results diplay that over both alternatives, for the normal and the Logistic distribution n_a to n_e has a maximum deviation of 1.08% at $\delta = 0.2$, 2.47% at $\delta = 0.5$, and 8% at $\delta = 0.8$. For the extremely large effect sizes, at which the minimum required sample size is small, the maximum deviation is 18.18% at $\delta = 1.2$ and 33.33% at $\delta = 1.6$. Over both alternatives, for those distributions with larger variances, thus the Laplace and the $t(3)$ distribution, the maximum deviation of n_a to n_e is 1.69% at $\delta = 0.2$, 7.69% at $\delta = 0.5$, and 16.67% at $\delta = 0.8$. For the extremely large effect sizes the maximum deviation is 42.86% at $\delta = 1.2$ and 75% at $\delta = 1.6$.

Table 6: Sample size analysis results using the van der Waerden statistic for two-sided alternatives

δ	β	Normal		Logistic		Laplace		$t(3)$	
		n_e	n_a	n_e	n_a	n_e	n_a	n_e	n_a
0.2	0.05	657	650	622	621	519	511	397	397
	0.10	530	526	501	502	420	413	325	321
	0.20	395	393	377	375	309	309	242	240
0.5	0.05	104	104	103	100	87	82	65	64
	0.10	87	85	83	81	72	67	52	52
	0.20	64	63	61	60	52	50	39	39
0.8	0.05	43	41	42	39	37	32	28	25
	0.10	35	33	32	32	30	26	22	21
	0.20	27	25	24	24	21	20	17	15
1.2	0.05	20	19	19	18	19	15	15	12
	0.10	16	15	16	14	15	12	12	9
	0.20	13	11	13	11	12	9	10	7
1.6	0.05	12	11	12	10	12	8	10	7
	0.10	10	9	10	8	10	7	8	6
	0.20	8	7	8	6	8	5	7	4

Overall, the new sample size analysis method based on score generating functions shows to perform well for the one-sided as well as for the two-sided alternatives as shown in Tables 1 to 6. In general, a higher sample size is required to ensure a fixed effect with given probabilities for the errors of the first and second kind the larger the distribution's variance. For those distributions, for which one of the here compared linear rank tests is the asymptotically optimal one, the respective test presents the lowest minimum required sample size. Thus, at normal distributions, the van der Waerden test displays the lowest minimum required sample size for the one- as well as for the two-sided alternatives, at the Logistic distribution the WMW test, and at the Laplace distribution Mood's Median test. At the $t(3)$ distribution, the new sample size analysis method also provides a good performance.

However, a slightly anti-conservative behavior was observable using the WMW as well as using the van der Waerden statistic. This is displayed by a slight underestimation of the minimum required sample size compared to those sample sizes calculated using a test statistic's exact distribution. In contrast to this, in general, a rather conservative behavior was observable at using Mood's Median statistic, which overestimated the minimum required sample size. Therefore, the results are in line with the respective test statistics' convergence rates to the normal distribution (Klein (1978)). Thus, smallest deviations between n_a to n_e are found for the van der Waerden test and the largest ones for Mood's Median test. Already Mann and Whitney (1947) stated the WMW test statistic's approximation to normality to be sufficient for sample sizes up from $n = m = 8$. Thus, the rather precise results of the new method for the conditions with small to moderate distributions' variances, even at extremely high effect sizes, lead to the interpretation that the risk of an occuring circular argument seems to be not severe. However, at those conditions with extremely large effect sizes for the distributions with higher variance the risk could be apparent. In these cases, e.g. for the WMW test, exact sample size methods are computationally feasible and thus could be applied. Moreover, as the limitation of the method introduced in Section 3.2 is a slight underestimation of the required minimum sample size using the WMW and the van der Waerden test statistic, one possibility to avoid the occurence of underestimating the minimum required sample size in the extreme cases could be to multiply the results by the factor 1.75. This is, as for the most extreme case, thus for the t(3) distribution with the extreme effect size of 1.6, the maximum deviation found for either test statistic was 75%. In addition to that, regardless the size of the effect, another possibility could be to add 2 (4) observations to the calculated required minimum sample size at distributions with low (high) variance to avoid the slight underestimation in almost 90% of all considered cases.

Besides focussing on the new method's performance at different conditions, it is of interest to compare its performance to the one of alternate sample size analysis methods. This will be done in the next section.

5. Performance Comparison of Sample Size Analysis Methods for the Wilcoxon-Mann-Whitney Test

In this section we compare the new sample size analysis method to common alternate methods. The comparison was done for the WMW test, as to our knowledge the sample size analysis method provided by Chakraborti et al. (2006) is so far the only one suitable for arbitrary linear rank tests on location shifts. Therefore, using Wilcoxon scores, next to the method of Chakraborti et al. (2006) (Chak.), the most common sample size analysis methods for the WMW test will be used for the performance comparison. Those sample size analysis methods are the ones presented by Al-Sunduqchi (1990) (Al-S.), Vollandt and Horn (1997) (V&H), Noether (1987) (Noeth.), and Lehmann (1975) (Leh.) (see Section 2). Moreover, Table 7 displays the sample size analysis results using the test statistic's exact distribution (n_e), as well as the new method's results (n_a). The performance comparison was conducted for one-sided alternatives on a normal distribution (compare Table 1).

Table 7 demonstrates that the sample size analysis method presented in Section 3.2 performes well in comparison to alternate methods for the WMW test. In contrast to its slight underestimation, most of the common alternate methods overestimate the minimum required sample size. As expected, the highest overestimation of minimum required sample sizes, especially for small effect sizes, can be observed for the method proposed by (Al-Sunduqchi, 1990), which is based on the asymptotic relative efficiency of the WMW test to the t-test, as well as the method proposed by Vollandt and Horn (1997), which is based on the WMW test statistic's variance's upper bound. Not surprisingly, using WMW scores, the 'Noether-like' method of Chakraborti et al. (2006) and the method proposed by Noether (1987) provide the same results. Furthermore, the method based on score generating functions introduced in Section 3.2 (n_a) is similar to the one proposed by Lehmann (1975). In addition to that, adding 2 observations to the minimum required sample size n_a calculated using the new method still provides more precise results than the alternate methods that overestimate the minimum required sample size.

However, in general, for sample size analyses for linear rank tests at an underlying normal distribution, it is recommended to apply the van der Waerden test, as it is the asymptotically optimal one. Referring to Table 5, it is observable that this test requires a minimum sample

Table 7: Comparison between different sample size analysis methods on normal distribution for the Wilcoxon-Mann-Whitney statistic for one-sided alternatives

δ	β	n_e	n_a	Al-S.	V&H	Noeth.	Chak.	Leh.
0.2	0.05	568	567	627	589	571	571	567
	0.10	450	449	496	465	452	452	449
	0.20	324	324	358	334	326	326	325
0.5	0.05	92	91	101	98	95	95	91
	0.10	74	72	80	78	75	75	73
	0.20	53	52	59	56	54	54	53
0.8	0.05	37	36	40	40	40	40	36
	0.10	29	29	32	32	32	32	29
	0.20	22	21	24	23	23	23	22
1.2	0.05	17	16	18	19	20	20	16
	0.10	14	13	15	15	16	16	13
	0.20	10	9	11	12	12	12	10
1.6	0.05	10	9	11	11	14	14	9
	0.10	9	8	9	9	11	11	8
	0.20	7	6	6	7	8	8	7

size to ensure a fixed effect with fixed probabilities of the errors of first and second kind that is lower than the minimum sample sizes needed by all compared methods for the WMW test (see Table 7).

6. Conclusion

A new sample size analysis method for arbitrary linear rank tests for location shifts of continuous distributions was introduced. This method is based on score generating functions and Section 3.1 showed its asymptotic normality. In Section 4 its good performance at different distributions was demonstrated, while in general a slightly anti-conservative behavior was observable. Moreover, the risk of an occuring circular argument was examined

by comparing the new method's results to sample sizes calculated using a test statistic's exact distribution. Focussing on those conditions at extremely large effect sizes of 1.2 and 1.6 suggests the risk to not be severe when the distribution's variance is small to moderate. Caution has to be taken when the underlying distribution's variance is high. Furthermore, at a performance comparison for the Wilcoxon-Mann-Whitney test on normal distributions it could be shown that most alternate sample size analysis methods overestimate the minimum required sample size, while the method introduced in Section 3.2 demonstrates to be more precise than these alternate methods. However, as at different population's distributions different linear rank tests are the preferred ones by being (asymptotically) optimal, sample size analysis methods for linear rank tests besides the Wilcoxon-Mann-Whitney tests are needed. Thus, the here introduced method provides sample size analysis for arbitrary linear rank tests and also demonstrated its performance for Mood's Median test and the van der Waerden test. In addition to providing a sample size analysis method for arbitrary linear rank tests, the method proposed in Section 3.2 has the big advantages of being easy to implement and not being computationally and time intensive. As sample size analysis is of central importance for any experimental research's planning phase and limited literature is yet available providing methods for sample size analysis for linear rank tests besides the Wilcoxon-Mann-Whitney test, this paper proposes a new method to account for this need.

Bibliography

Al-Sunduqchi, M. S. (1990). *Determining the appropriate sample size for inferences based on the Wilcoxon statistics: Unpublished Ph.D. dissertation.* Department of Statistics, University of Wyoming, Laramie, USA.

Birnbaum, Z. W. and Klose, O. M. (1957). Bounds for the variance of the mann-whitney statistic. *The Annals of Mathematical Statistics*, 28(4):933–945.

Büning, H. and Trenkler, G. (1994). *Nichtparametrische statistische Methoden.* de Gruyter, Berlin.

Chakraborti, S., Hong, B., and van de Wiel, M. A. (2006). A note on sample size determination for a nonparametric test of location. *Technometrics*, 48(1):88–94.

Cohen, J. (1969). *Statistical power analysis for the behavioural sciences.* Academic Press, New York, USA.

Divine, G., Kapke, A., Havstad, S., and Joseph, C. (2010). Exemplary dataset sample size calculation for wilcoxon-mann-whitney tests. *Statistics in Medicine*, 29(1):108–115.

Efron, B. (1979). Bootstrap methods: another look at the jacknife. *The Annals of Statistics*, 7(1):1–26.

Freidlin, B. and Gastwirth, J. L. (2000). Should the median test be retired from general use? *The American Statistician*, 54(3):161–164.

Gibbons, J. D. (1971). *Nonparametric statistical inference.* McGraw-Hill, Tokyo, Japan.

Gibbons, J. D. and Chakraborti, S. (2003). *Nonparametric statistical inference.* Marcel Dekker, Inc., New York and Basel, fourth edition, revised and expanded edition.

Guo, J.-H. (2012). Optimal sample size planning for the wilcoxon-mann-whitney and van elteren tests under cost constraints. *Journal of Applied Statistics*, 39(10):2153–2164.

Hájek, J., Sidák, Z., and Sen, P. K. (1999). *Theory of Rank Tests.* Academic Press, London, England.

Hamilton, M. A. and Collings, B. J. (1991). Determining the appropriate sample size for nonparametric tests for location shifts. *Technometrics*, 33(3):327–337.

Haynam, G. E. and Govindarajulu, Z. (1966). Exact power of mann-whitney test for exponential and rectangular alternatives. *The Annals of Mathematical Statistics*, 37(4):945–953.

Hilton, J. F. and Mehta, C. R. (1993). Power and sample size calculations for exact conditional tests with ordered categorical data. *Biometrics*, 49(2):609–616.

Klein, I. (1978). *Konvergenzgeschwindigkeit von linearen Rangordnungsstatistiken mit lokal trennscharfen (optimalen) Scores.* Diplom-Arbeit, Kiel.

Lehmann, E. L. (1953). The power of rank tests. *The Annals of Mathematical Statistics*, 24(1):23–43.

Lehmann, E. L. (1959). *Testing statistical hypotheses.* John Wiley & Sons, Inc., New York, USA.

Lehmann, E. L. (1975). *Nonparametrics: Statistical Methods based on Ranks.* McGraw-Hill, New York, USA.

Mann, H. B. and Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60.

Mood, A. M. (1954). On the asymptotic efficiency of certain nonparametric two-sample tests. *The Annals of Mathematical Statistics*, 25(3):514–522.

Noether, G. E. (1987). Sample size determination for some common nonparametric tests. *Journal of the American Statistical Association*, 82(398):645–647.

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

Shieh, G., Show-Li, J., and Randles, R. H. (2006). On power and sample size determinations for the wilcoxon-mann-whitney test. *Nonparametric Statistics*, 18(1):33–43.

van de Wiel, M. A. (2000). *Exact distributions of distributionsfree test statistics: PhD-thesis.* Eindhoven University of Technology, Netherlands.

van der Waerden, B. L. (1952). Order tests for the two-sample problem and their power. *Indagationes Mathematicae*, 14:453–458.

Vollandt, R. and Horn, M. (1997). Evaluation of noether's method of sample size determination for the wilcoxon-mann-whitney test. *Biometrical Journal*, 39(7):823–829.

Wan, W.-M., Wu, C.-H., Tseng, Y.-M., and Wang, M.-J. (2009). An improved algorithm for sample size determination of ordinal response by two groups. *Communications in Statistics - Simulation and Computation*, 38(10):2235–2242.

Wang, H., Chen, B., and Chow, S.-C. (2003). Sample size determination based on rank tests in clinical trials. *Journal of Biopharmaceutical Statistics*, 13(3):735–751.

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

Zhao, Y. D., Rahardja, D., and Qu, Y. (2008). Sample size calculation for the wilcoxon-mann-whitney test adjusting for ties. *Statistics in Medicine*, 27(3):462–468.