Published in final edited form as:

Biometrics. 2011 June ; 67(2): 646–653. doi:10.1111/j.1541-0420.2010.01488.x.

Power and Sample Size Estimation for the Clustered Wilcoxon Test

Bernard Rosner and Robert J. Glynn

Channing Laboratory, Harvard Medical School, 181 Longwood Avenue, Boston, MA 02115

Bernard Rosner: bernard.rosner@channing.harvard.edu

Summary

The Wilcoxon rank sum test is widely used for two-group comparisons of non-normal data. An assumption of this test is independence of sampling units both within and between groups which will be violated in the clustered data setting such as in ophthalmological clinical trials, where the unit of randomization is the subject, but the unit of analysis is the individual eye. For this purpose, we have proposed the clustered Wilcoxon test to account for clustering among multiple subunits within the same cluster (Rosner, Glynn and Lee, *Biometrics* 2003, 2006). However, power estimation is needed to plan studies that use this analytic approach. We have recently published methods for estimating power and sample size for the ordinary Wilcoxon rank sum test (Rosner and Glynn, *Biometrics*, 2009). In this paper we present extensions of this approach to estimate power for the clustered Wilcoxon test. Simulation studies show good agreement between estimated and empirical power. These methods are illustrated with examples from randomized trials in ophthalmology. Enhanced power is achieved with use of the subunit as the unit of analysis instead of the cluster using the ordinary Wilcoxon rank sum test.

Keywords

Clustered data; Ophthalmology; Rank sum test

1. Introduction and Notation

The Wilcoxon rank sum test or equivalently the Mann-Whitney U Test (henceforth referred to as the Wilcoxon Mann-Whitney (WMW) test) is a frequently used nonparametric test to compare the location parameter of two distributions. An assumption of this test is independence of sample points, both between and within samples. This will be violated in the setting of clustered data where there is independence between clusters but not between multiple subunits within the same cluster. For the sake of specificity, we consider data from the Sorbinil Retinopathy Trial (SRT) (Sorbinil Retinopathy Trial, 1990). The SRT was conducted among 497 type I diabetic patients who had little or no evidence of retinopathy at baseline. Patients were randomized to either Sorbinil, an aldose-reductase inhibitor or placebo, and were seen at 1 year and then every 9 months up to 48 months after randomization. In addition, all subjects had a final visit at the end of the trial (max = 56 months). Sixteen of the patients provided no follow-up. The primary endpoint of the trial was based on change in retinopathy severity level from baseline to the maximum follow-up

Correspondence to: Bernard Rosner, bernard.rosner@channing.harvard.edu.

Supplementary Materials: SAS macros to implement the methods in this paper for both balanced and unbalanced designs are available at the *Biometrics* website (<http://www.biometrics.tibs.org>) as well as the following web site (<https://sites.google.com/a/channing.harvard.edu/bernardrosner/home>).

(i.e., severity level at maximum follow-up minus severity level at baseline) for each patient based on a patient-level score. To obtain this patient-level score, each eye was graded into a 9 category ordinal scale (ETDRS scale). A patient-level grade at each time was obtained from a cross-classification of eye-specific grades in the right and left eyes resulting in a 40-step ordinal classification scheme. Person-level change was obtained as the difference in this score over 2 time points. In a previous paper (Rosner and Glynn, 2009) we considered power of the Wilcoxon rank sum test based on changes in this person-specific score. In this paper, we consider power of the clustered Wilcoxon test based on change scores in individual eyes.

Table 1 shows the cross-tabulation of change scores for right eyes $\times$ left eyes for the 243 subjects in the placebo group. The change scores for individual eyes ranged from -2 (improvement) to $+4$ (worsening) with the distribution of change scores somewhat right skewed. 122 of the subjects (50%) had the same change score in both eyes, 103 (42%) had change scores which differed by 1 unit and 18 (7%) differed by 2+ units. The weighted kappa with quadratic weights was 0.48 indicating substantial, but not perfect agreement between change scores for fellow eyes.

In this setting, we let X_{ij} = change in ETDRS grade for the j^{th} subunit (eye) from the i^{th} cluster (subject), $i = 1, \dots, m$; $j = 1, \dots, g$ in the control group and Y_{kl} = change in ETDRS grade for the l^{th} subunit (eye) from the k^{th} cluster (subject), $k = 1, \dots, n$; $l = 1, \dots, g$ in the active treatment group. We wish to test the hypothesis:

$$H_0: F_X = F_Y \text{ vs. } H_1: F_Y(y) = F_X(y - \delta), \text{ where } \delta \neq 0. \quad (1)$$

Let $R_{ij} = \text{Rank}(X_{ij})$ for observations in the X group and $R_{m+k,l} = \text{Rank}(Y_{kl})$ for observations in the Y group where ranks are determined from gN scores over all subunits among the $N = m + n$ clusters in the control and treatment groups combined. We use the clustered Wilcoxon statistic as a test statistic:

$$W_{c,\text{obs}} = \sum_{i=1}^m \sum_{j=1}^g R_{ij}.$$

To evaluate statistical significance, we consider the randomization distribution corresponding to $W_{c,\text{obs}}$ given by

$$W_c = \sum_{i=1}^N \delta_i R_{i+}, \quad (2)$$

where $R_{i+} = \sum_{j=1}^g R_{ij}$, $\delta_i = 1$ if $i \in I$, $\delta_i = 0$ if $i \notin I$, and I = the set of m unique integers from $\{1, \dots, N\}$ corresponding to the m clusters which hypothetically could have been assigned to the control group out of the N clusters in a randomization test design. We have previously shown that the test statistic

$$Z_c = \{W_c - E(W_c)\} / \{\widehat{\text{var}}(W_c)\}^{1/2} \quad (3)$$

is asymptotically $N(0, 1)$ as $N \rightarrow \infty$, provided that $\lim_{N \rightarrow \infty} m/N = \xi$, where $0 < \xi < 1$ (Rosner, et al, 2003),

$$E(W_c) = gm(gN+1)/2, \text{ and}$$

$$\widehat{\text{var}}(W_c) = [mn/\{N(N-1)\}] \sum_{i=1}^N \{R_{i+} - g(1+gN)/2\}^2.$$

We refer to this test as the clustered Wilcoxon test for balanced designs. In this paper, we consider the power of the test procedure in finite samples and also extend it to unbalanced designs where g is variable for different clusters. We anticipate an increase in power when using the clustered Wilcoxon test based on change scores in individual eyes vs. the ordinary Wilcoxon test based on person-specific change scores. In Section 2.1, we present power formulas for the balanced case followed by the unbalanced case in Section 2.2 using a new type of shift alternative hypothesis referred to as the probit shift alternative. In Section 2.3, we compare power of the Mann-Whitney U Test under the general location shift alternative vs. the probit shift alternative for the extreme value location-shift family. Section 3 presents simulation results for both designs. Section 4 presents an application to the design of future ophthalmologic trials based on data from the Sorbinil Retinopathy Trial, compares power to alternate simpler designs based on the subject as the unit of analysis, and considers the planning of a new treatment trial for retinitis pigmentosa patients where the unit of analysis is the eye, but the number of eyes available for analysis varies for individual patients.

2. Methods

2.1 Balanced Design

It is difficult to compute power under the shift alternative in equation 1, since F_X is almost always unknown and the meaning of δ will be different for each F_X . Instead, we consider a different type of shift alternative. For this purpose, we let $H_X(X_{ij}) = \Phi^{-1}\{F_X(X_{ij})\}$ = underlying probit for the j^{th} subunit from the i^{th} X cluster, $i = 1, \dots, m$; $j = 1, \dots, g$; $H_Y(Y_{kl}) = \Phi^{-1}\{F_Y(Y_{kl})\}$, $k = 1, \dots, n$; $l = 1, \dots, g$. In addition, we let

$$\begin{aligned} H_X(X_{ij}) &= H_X(X_i) + e_X(X_{ij}), \quad i=1, \dots, m; \quad j=1, \dots, g, \\ H_Y(Y_{kl}) &= H_Y(Y_k) + e_Y(Y_{kl}), \quad k=1, \dots, n; \quad l=1, \dots, g, \end{aligned} \quad (4)$$

where $H_X(X_i) \sim N(0, \rho)$, $H_Y(Y_k) \sim N(\mu, \rho)$, $e_X(X_{ij}) \sim N(0, 1 - \rho)$, $e_Y(Y_{kl}) \sim N(0, 1 - \rho)$ and $H_X(X_i)$, $H_Y(Y_k)$, $e_X(X_{ij})$, $e_Y(Y_{kl})$ are mutually independent.

Suppose X, Y correspond to random variables obtained from a control group and an active treatment group, respectively, and let $\{X_1, \dots, X_m\}$, $\{Y_1, \dots, Y_n\}$ be realizations of X, Y . We will assume that X and Y are continuous random variables with no ties. Let Y_c be the counterfactual random variable obtained if each member of the active treatment group received the control treatment. By definition, $H_{Y_c} = H_X$. We consider the class of shift alternatives given by

$$H_0: H_X = H_Y = H_{Y_c} \text{ vs. } H_1: H_Y = H_{Y_c} + \mu, \text{ where } \mu \neq 0. \quad (5)$$

It follows from (4) that

$$\text{corr}\{H_X(X_{ij1}), H_X(X_{ij2})\} = \text{corr}\{H_Y(Y_{kl1}), H_Y(Y_{kl2})\} = \rho, H_X(X_{ij}) \sim N(0, 1) \text{ and } H_Y(Y_{kl}) \sim N(\mu, 1).$$

The rationale for using the probit transformation to define the shift alternative in (5) is that

- a. all continuous distributions (F_X) have a corresponding underlying probit transformation (H_X)
- b. it is natural to define location shifts between distributions in terms of a normally distributed scale, which by definition is satisfied by the probit transformation.

Indeed, analysis of variance models are often used on transformed scales (such as (4)) to achieve approximate normality so that location shifts between distributions are more interpretable. Furthermore, based on (5), μ , can be interpreted as an effect size on the probit scale, which is very useful for sample size and power calculations.

To illustrate the meaning of μ , we plot the probability mass function (P_X) of right eye change scores in the control group of the Sorbinil Retinopathy Trial from Table 1 and hypothetical probit scores (P_Y) in the active treatment group with $\mu = 0.2, 0.5, 0.8$, where $F_Y(y) = \Phi[\Phi^{-1}\{F_X(y)\} + \mu]$, $P_X(y) = F_X(y) - F_X(y-1)$, $y = -1, 0, \dots, 4$, $F_X(-2) = 0$, and similarly for $P_Y(y)$.

As μ increases, there is a smooth transition from increasing probability mass for change scores of 0 (no change) or -1 (improvement) and decreasing probability mass for changes scores of ≥ 1 (worsening).

It will be easier to compute power under the alternative hypothesis in (5) if the test procedure in (3) is reformulated in terms of θ_c , where $\theta_c = \Pr(X_{ij} < Y_{kl})$. Indeed, the Wilcoxon rank sum test statistic is consistent for the hypothesis $H_0: \theta_c = 1/2$ vs. $H_1: \theta_c \neq 1/2$ in the sense that the power of this test $\rightarrow 1$ as $m, n \rightarrow \infty$ (Hollander and Wolfe, 1999). Furthermore, some authors emphasize the importance of θ_c due to its ease of interpretability compared with the location shift parameter δ in equation 1 (Wolfe and Hogg, 1971; Browne, 2010). Specifically, we consider the test statistic

$$Z_c^* = (\hat{\theta}_c - 1/2) / \{\text{var}_0(\hat{\theta}_c)\}^{1/2}, \quad (6)$$

where

$$\hat{\theta}_c = \sum_{i=1}^m \sum_{j=1}^g \sum_{k=1}^n \sum_{l=1}^g U(X_{ij} - X_{kl}) / (mng^2),$$

and $U(a) = 1$ if $a < 0$; $= 1/2$ if $a = 0$, and $= 0$ if $a > 0$.

It is easy to show that $E(\hat{\theta}_c) = \theta_c$ and $\theta_c = \Phi(\mu / \sqrt{2})$. The asymptotic power of the test procedure in (6) can be expressed in the form:

$$\text{Power} = \Phi[|\theta_c - 1/2|/\{\text{var}_1(\widehat{\theta}_c)\}^{1/2} - z_{1-\alpha/2}\{\text{var}_0(\widehat{\theta}_c)/\text{var}_1(\widehat{\theta}_c)\}^{1/2}] + \Phi[-|\theta_c - 1/2|/\{\text{var}_1(\widehat{\theta}_c)\}^{1/2} - z_{1-\alpha/2}\{\text{var}_0(\widehat{\theta}_c)/\text{var}_1(\widehat{\theta}_c)\}^{1/2}], \quad (7)$$

where $z_p = p^{\text{th}}$ percentile of a $N(0, 1)$ distribution, and $\text{var}_0(\widehat{\theta}_c)$, $\text{var}_1(\widehat{\theta}_c)$ = variance of $\widehat{\theta}_c$ under H_0 and H_1 , respectively.

To compute the power in (7) we need to determine $\text{var}_0(\widehat{\theta}_c)$ and $\text{var}_1(\widehat{\theta}_c)$.

Based on the results in the Web Appendix, we have:

$$\begin{aligned} \text{var}_1(\widehat{\theta}_c) &= [\theta_c(1 - \theta_c) + 2(g - 1)[\Phi_2\{\Phi^{-1}(\theta_c), \Phi^{-1}(\theta_c), (1+\rho)/2\} - \theta_c^2] + (g - 1)^2[\Phi_2\{\Phi^{-1}(\theta_c), \Phi^{-1}(\theta_c), \rho\} - \theta_c^2] + (m+n-2)g[\Phi_2\{\Phi^{-1}(\theta_c), \Phi^{-1}(\theta_c), 1/2\} - \theta_c^2] + (g - 1)[\Phi_2\{\Phi^{-1}(\theta_c), \Phi^{-1}(\theta_c), \rho/2\} - \theta_c^2]] / (mng^2), \end{aligned} \quad (8)$$

where $\Phi_2(c_1, c_2, \rho) = \Pr\left[Z_1 \leq c_1 \text{ and } Z_2 \leq c_2 \mid (Z_1, Z_2) \sim N\left\{\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right\}\right]$. Note that $\Phi_2(c_1, c_2, \rho)$ can be computed using the PROBBNRM function of SAS Version 9.0 (SAS, 2006) or the binormal function of Stata (Stata, 2009).

Note that under H_0 , $\theta_c = 1/2$ and equation 8 reduces to

$$\text{var}_0(\widehat{\theta}_c) = [1/4 + 2(g-1)\sin^{-1}\{(1+\rho)/2\}/(2\pi) + (g-1)^2\sin^{-1}(\rho)/(2\pi) + (m+n-2)g\{1/12 + (g-1)\sin^{-1}(\rho/2)/(2\pi)\}]/(mng^2). \quad (9)$$

Upon combining (7), (8) and (9) we obtain the power of the clustered Wilcoxon test for balanced designs.

2.1.1 Special Cases—Note that if $g = 1$, then equation 8 reduces to the power of the ordinary Wilcoxon rank sum test (Rosner and Glynn, *Biometrics*, 2009, equation 16) given by

$$\text{var}(\widehat{\theta}) = [\theta(1-\theta) + (m+n-2)[\Phi_2\{\Phi^{-1}(\theta), \Phi^{-1}(\theta), 1/2\} - \theta^2]]/(mn).$$

Furthermore, if $g = 1$, then equation 9 reduces to the ordinary Mann-Whitney variance under H_0 given by

$$\text{var}_0(\widehat{\theta}) = (m+n+1)/(12mn).$$

If $g > 1$ and $\rho = 0$, then equation 9 reduces to

$$\text{var}_0(\widehat{\theta}) = (m^* + n^* + 1)/(12m^*n^*),$$

where $m^* = mg$ and $n^* = ng$.

2.2 Unbalanced Design

We now consider power for the clustered Wilcoxon statistic for an unbalanced design. We consider the test statistic in equation 20 of (Rosner, Glynn & Lee, 2006) given by

$$Z_{c,s} = (\widehat{\theta}_{c,s} - 1/2) / \{\widehat{\text{var}}(\widehat{\theta}_{c,s})\}^{1/2}, \quad (10)$$

where

$$\widehat{\theta}_{c,s} = \sum_{g=1}^{g_{\max}} w^{(g)} \widehat{\theta}_c^{(g)} / \sum_{g=1}^{g_{\max}} w^{(g)},$$

and

$$\begin{aligned} \text{var}(\widehat{\theta}_{c,s}) &= 1 / \sum_{g=1}^{g_{\max}} w^{(g)}, \\ 1/w^{(g)} &= \frac{1}{N^{(g)}\{N^{(g)}-1\}\{g^4 m^{(g)} n^{(g)}\}} \sum_{i=1}^{N^{(g)}} [R_{i+}^{(g)} - g\{1 + gN^{(g)}\}/2]^2, \end{aligned} \quad (11)$$

where $m^{(g)}$ and $n^{(g)}$ are the number of controls and cases among clusters of size g , $N^{(g)} = m^{(g)} + n^{(g)}$, and $R_{ij}^{(g)}$ = rank of the j^{th} subunit from the i^{th} cluster of size g among all subunits

for all clusters of sizes g and $R_{i+}^{(g)} = \sum_{j=1}^g R_{ij}^{(g)}$.

Therefore, similar to equation 7, we have:

$$\text{Power}_{c,s} = \Phi \left[\frac{|\theta_c - 1/2| / \{\text{var}_1(\widehat{\theta}_{c,s})\}^{1/2} - z_{1-\alpha/2} \{\text{var}_0(\widehat{\theta}_{c,s}) / \text{var}_1(\widehat{\theta}_{c,s})\}^{1/2}}{\sqrt{1 + \frac{1}{2} \left(\frac{1}{\text{var}_1(\widehat{\theta}_{c,s})} - \frac{1}{\text{var}_0(\widehat{\theta}_{c,s})} \right) \left(\frac{1}{\text{var}_1(\widehat{\theta}_{c,s})} - \frac{1}{\text{var}_0(\widehat{\theta}_{c,s})} \right)^{-1}}}} \right], \quad (12)$$

where

$$\begin{aligned} \text{var}_0(\widehat{\theta}_{c,s}) &= 1 / \sum_{g=1}^{g_{\max}} w_0^{(g)}, \\ 1/w_0^{(g)} &= [1/4 + 2(g-1)\sin^{-1}\{(1+\rho)/2\}/(2\pi) + (g-1)^2\sin^{-1}(\rho)/(2\pi) + \{m^{(g)} + n^{(g)} - 2\}g\{1/12 + (g-1)\sin^{-1}(\rho/2)/(2\pi)\}]/\{m^{(g)}n^{(g)}g^2\}, \\ \text{var}_1(\widehat{\theta}_{c,s}) &= \sum_{g=1}^{g_{\max}} [\{w_0^{(g)}\}^2 / w_1^{(g)}] / \left\{ \sum_{g=1}^{g_{\max}} w_0^{(g)} \right\}^2, \\ 1/w_1^{(g)} &= [\theta_c(1-\theta_c) + 2(g-1)[\Phi_2\{\Phi^{-1}(\theta_c), \Phi^{-1}(\theta_c), (1+\rho)/2\} - \theta_c^2] + (g-1)^2[\Phi_2\{\Phi^{-1}(\theta_c), \Phi^{-1}(\theta_c), \rho\} - \theta_c^2] + \{m^{(g)} + n^{(g)} - 2\}g[\Phi_2\{\Phi^{-1}(\theta_c), \Phi^{-1}(\theta_c), \rho\} - \theta_c^2]]. \end{aligned}$$

To use equation 12, we need to know $m^{(g)}, n^{(g)}, g = 1, \dots, g_{\max}$. If this information is not readily available, then we could assume that the overall cluster size distribution is fixed, and the proportion of active treated subjects is the same for all g , whereby

$\lambda^{(g)} = N^{(g)}/N$, $m^{(g)}/n^{(g)}$ is constant for all g . Thus,

$$m^{(g)} = (m_+/N) Ns \text{ constant for all } = m_+ \lambda^{(g)} \text{ and } n^{(g)} = n_+ \lambda^{(g)}, g = 1, \dots, g_{\max}, m_+ = \sum m^{(g)}, n_+ = \sum n^{(g)}.$$

2.3 Comparison of Power for the Mann-Whitney U Test under the General Location Shift Alternative in (1) vs. the Probit Shift Alternative in (5)

The extreme value location shift family occurs frequently in the context of parametric survival analysis based on the Weibull distribution (Lawless, 2003). One advantage of this family is that one can explicitly obtain the variance of the Mann-Whitney U statistic under both H_0 and H_1 for the general location shift alternative in (1). Thus, we used this family to compare power estimates for the Mann-Whitney U test under the location shift alternative in (1) vs. the probit shift alternative in (5).

For simplicity, we consider the ordinary Mann-Whitney U test in the absence of clustering. A crucial element in estimating power for the Mann-Whitney U test is determination of $\text{var}_1(\widehat{\theta})$, where $\theta = \Pr(X_i < Y_k)$, $\widehat{\theta} = U/(mn)$, m = number of control subjects and n = number of active treatment subjects. Lehmann (1998) has provided a general formula for $\text{var}(U)$ and equivalently $\text{var}(\widehat{\theta})$ given by

$$\text{var}(\widehat{\theta}) = \{\theta(1-\theta) + (n-1)(p_2 - \theta^2) + (m-1)(p_3 - \theta^2)\} / (mn) \quad (13)$$

where

$$\begin{aligned} p_2 &= \Pr(X_i < Y_{k_1} \text{ and } X_i < Y_{k_2}), i=1, \dots, m; k_1 \neq k_2=1, \dots, n; \\ p_3 &= \Pr(X_{i_1} < Y_k \text{ and } X_{i_2} < Y_k), i_1 \neq i_2=1, \dots, m; k=1, \dots, n. \end{aligned}$$

Under the general location shift alternative in (1), we have

$$\begin{aligned}
 p_2 &= \int_{x=-\infty}^{\infty} f_X(x) \{1 - F_X(x - c)\}^2 dx \\
 p_3 &= \int_{x=-\infty}^{\infty} f_X(x - c) F_X^2(x) dx.
 \end{aligned}
 \tag{14}$$

We consider the specific case of the extreme value distribution location-scale family, where

$$\begin{aligned}
 f_X(x) &= \exp\{x/b - \exp(x/b)\}/b, \quad S_X(x) = \exp\{-\exp(x/b)\} \\
 f_Y(x) &= \exp[(x - c)/b - \exp\{(x - c)/b\}]/b, \quad S_Y(x) = \exp[-\exp\{(x - c)/b\}],
 \end{aligned}$$

$S_X = 1 - F_X(x)$, and similarly for S_Y .

We can interpret c/b as an effect size denoted by d . From (14), it can be shown that

$$\begin{aligned}
 \theta &= \exp(d) / \{1 + \exp(d)\} \\
 p_2 &= \exp(d) / \{2 + \exp(d)\} \\
 p_3 &= 2\exp(2d) / [\{1 + \exp(d)\}\{1 + 2\exp(d)\}],
 \end{aligned}$$

and

$$\text{var}(\hat{\theta}) = \left[\frac{\exp(d)}{\{1 + \exp(d)\}^2} + \frac{(n - 1)\exp(d)}{[2 + \exp(d)]\{1 + \exp(d)\}^2} + \frac{(m - 1)\exp(2d)}{[1 + 2\exp(d)]\{1 + \exp(d)\}^2} \right] / (mn). \tag{15}$$

Similarly, based on Rosner and Glynn (2009), under the probit alternative:

$$\text{var}_{\text{probit}}(\hat{\theta}) = \left[\frac{\exp(d)}{\{1 + \exp(d)\}^2} + (m + n - 2) \left[\Phi_2 \left[\Phi^{-1} \left\{ \frac{\exp(d)}{1 + \exp(d)} \right\}, \Phi^{-1} \left\{ \frac{\exp(d)}{1 + \exp(d)} \right\}, 1/2 \right] - \left\{ \frac{\exp(d)}{1 + \exp(d)} \right\}^2 \right] \right] / (mn). \tag{16}$$

The power of the Mann-Whitney U test is given by

$$\text{Power} = \Phi[|\theta - 1/2| / \{\text{var}_1(\hat{\theta})\}^{1/2} - z_{1-\alpha/2} \{\text{var}_0(\hat{\theta}) / \text{var}_1(\hat{\theta})\}^{1/2}] + \Phi[-|\theta - 1/2| / \{\text{var}_1(\hat{\theta})\}^{1/2} - z_{1-\alpha/2} \{\text{var}_0(\hat{\theta}) / \text{var}_1(\hat{\theta})\}^{1/2}]. \tag{17}$$

Note that for both (1) and (5), under H_0 , $\text{var}_0(\hat{\theta}) = (m + n + 1)/(12mn)$.

We have evaluated the power of the Mann-Whitney U test using (17), where $\text{var}(\hat{\theta})$ is obtained under the general shift alternative in (15) and the probit shift alternative in (16), for $d = 0.25, 0.50, 0.75, 1.0$ and $(m, n) = (50, 50), (30, 70)$ and $(70, 30)$. The results are given in Table 2.

We see that for $m = n = 50$ the power is virtually the same using either (15) or (16). For $(m, n) = (70, 30)$ or $(30, 70)$, the difference in power under the 2 alternatives is very small, generally $\leq 1\%$, which is negligible in practice.

Of course, in most cases, it is impossible to specify a functional form for F in (1). Nevertheless, the results in Table 2 give some reassurance that the power under the probit shift alternative in (5) gives a reasonable approximation to the power under the general shift alternative in (1) for the same value of θ .

Note that under the general location shift alternative in (1),

$$\theta = \Pr(X < Y) = \int_{x=-\infty}^{\infty} f_X(x) \{1 - F_X(x - c)\} dx.$$

Thus, for a given F , there is a 1 – 1 correspondence between the location shift c and the parameter θ . However, the probabilities p_2 and p_3 in (14) are not strictly a function of θ and in general are not the same.

Under the probit shift alternative in (5), the probabilities p_2 and p_3 are the same and can be uniquely determined as a function of θ given by

$$p_2 = p_3 = \Phi_2[\Phi^{-1}(\theta), \Phi^{-1}(\theta), 1/2].$$

Thus, the variance of $\hat{\theta}$ under H_1 can be determined as a function of θ .

3. Simulation Studies

We performed simulation studies to assess the validity of the power formulas in equation 7 and 12 in the setting of a balanced and unbalanced design, respectively. For the balanced design, we considered the case of 40 clusters with g subunits per cluster, where $g = 2, 3$ or 4. Twenty of the clusters were randomly assigned to be unexposed (X clusters) and 20 to be exposed (Y clusters). A nominal type I error of 0.05 was used. If H_{ij} = underlying probit for the j^{th} subunit from the i^{th} cluster, then

$$H_{ij} = H_i + e_{ij}, i = 1, \dots, 40; j = 1, 2 \quad (18)$$

where

$H_i \sim N(0, \rho)$ for unexposed clusters, $H_i \sim N(0.8, \rho)$ for exposed clusters $e_{ij} \sim N(0, 1 - \rho)$ and H_i is independent of e_{ij} .

4000 simulated datasets were obtained.

We computed the empirical power given by $\text{Prob}(|Z_c| > 1.96)$ where Z_c is obtained from equation 3 and compared it with the theoretical power given by equation 7. Simulations were performed for each combination of $g = 2, 3, 4$ and $\rho = 0.2, 0.5, 0.8$.

For unbalanced designs, we considered the case of 40 clusters with 20 clusters of size 2 and 20 clusters of size 4. Within each cluster size, 10 of the clusters were randomly assigned to be unexposed (X clusters) and 10 to be exposed (Y clusters). Four thousand simulated datasets were generated as in equation 18. For each simulated dataset we computed the test statistic $Z_{c,s}$ in equation 10 and the empirical power given by $\Pr(|Z_{c,s}| > 1.96)$ and compared

it with the mean theoretical power in equation 12. The simulation results for the balanced and unbalanced designs are presented in Table 3.

We see that the theoretical power tends to be slightly larger than the empirical power, although the difference is small, generally at most 2%. As expected, the power increases as the number of subunits per cluster (g) increases and decreases as the intraclass correlation (ρ) increases. Overall, the results indicate that the power formulas in equation 7 and 12 are adequate for designs with at least 40 clusters.

4. Examples

4.1 Clinical trial of treatment for diabetic retinopathy among patients with type I diabetes

We present an example of the methodology using data from the Sorbinil Retinopathy Study as described in the introduction. In Table 4, we present the power of the clustered Wilcoxon test based on eye-specific change scores for specific combinations of $(m, n) = (100, 100), \dots, (500, 500)$ and effect sizes $\mu = 0.1, \dots, 0.5$ and compare these with the power of the ordinary Wilcoxon test (Rosner and Glynn, 2009) based on person-specific change scores.

It is clear that there are substantial increases in power from considering eye-specific change scores vs. person-specific change scores, especially for low to moderate effect sizes ($\mu = 0.1 - 0.3$).

4.2 Evaluation of Docosahexaenoic Acid in Patients with Retinitis Pigmentosa Receiving Vitamin A Treatment

A prior randomized clinical trial showed that patients treated with vitamin A palmitate (15,000 IU/day) had on average a slower rate of decline of retinal function as monitored by the electroretinogram (ERG) (Berson, et al, 1993). A second study was subsequently performed assessing the effect of docosahexaenoic acid (DHA) in slowing the rate of field loss among patients with retinitis pigmentosa (RP) taking 15,000 IU/day of vitamin A (Berson, et al, 2004a). In brief, 208 RP patients were randomized to either 1200 mg of DHA plus 15,000 IU/day of vitamin A or control fatty acids + 15,000 IU/day of vitamin A and were followed for 4 years. The primary endpoint was change in central visual field measured by the Humphrey Field analyzer 30-2 program.

The overall results of the study were not significant. However, significant differences were found when patients were stratified by whether they were on vitamin A therapy prior to entry with a significant beneficial treatment effect among patients not taking vitamin A prior to entry ($p = 0.01$, p -interaction = 0.002) (Berson, et al, 2004b). Because the significant interaction effect was not expected, it is worth considering the necessary sample size to replicate these findings in a new study among patients not on vitamin A prior to entry.

Furthermore, it was recognized that the distribution of 4 year change in Humphrey 30-2 field was non-normal and both baseline field and change in field over time can vary substantially in fellow eyes for an individual patient. In addition, some patients were only evaluable on 1 of 2 eyes due to invalid field data. Hence, the clustered Wilcoxon test with an unbalanced design is an appropriate method of analysis for this study.

Thus, we consider power calculations based on equation 12 of this paper.

We first estimated $\hat{\theta}_c$ from the previous study, separately for patients contributing 1 or 2 eyes to the analysis (subsequently referred to as 1-eyed and 2-eyed patients) and then combined these estimates based on equation 10 to yield an overall $\hat{\theta} = 0.669$. We then estimated ρ_X from $\hat{\rho}_X = \text{corr}\{H_X(X_{i1}), H_X(X_{i2})\}$ and ρ_Y from $\hat{\rho}_Y = \text{corr}\{H_Y(Y_{k1}), H_Y(Y_{k2})\}$,

converted them to the Fisher's Z scale using $\hat{Z}_X = 0.5 \ln\{(1 + \hat{\rho}_X)/(1 - \hat{\rho}_X)\}$, $\hat{Z}_Y = 0.5 \ln\{(1 + \hat{\rho}_Y)/(1 - \hat{\rho}_Y)\}$, combined them using inverse-variance weighting to obtain $\hat{Z} = \{(m^{(2)} - 3)\hat{Z}_X + (n^{(2)} - 3)\hat{Z}_Y\}/(m^{(2)} + n^{(2)} - 6)$ and then estimated ρ from $\hat{\rho} = \{\exp(2\hat{Z}) - 1\}/\{\exp(2\hat{Z}) + 1\} = 0.706$.

We propose a new study with 50 patients in each group of whom 5 are 1-eyed patients and 45 are 2-eyed patients. We then estimate $w_0^{(g)}$ and $w_1^{(g)}$, $g = 1, 2$ from equation 12 and used them to calculate $\text{var}_0(\hat{\theta}_{c,s})$ and $\text{var}_1(\hat{\theta}_{c,s})$ and thereby $\text{Power}_{c,s}$ in equation 12.

The power is approximately 90% which should be adequate for the proposed new study.

5. Discussion

We have presented methods of power calculation for the clustered Wilcoxon test which is a generalization of the ordinary Wilcoxon rank sum test, using the subunit within a cluster as the unit of analysis and accounting for correlation between multiple subunits within the same cluster. Traditionally, power calculations for non-normally distributed outcome data have dichotomized the outcome and used standard methods for power estimation with binary outcomes and extensions of these methods to account for within-cluster correlation (Donner and Klar, 2000). This will result in a loss of power relative to using the entire scale in a nonparametric analysis (Rosner and Glynn, 2009). Other methods of adjusting standard rank tests for clustering have been proposed (Datta and Satten, 2005; Larocque, 2005). These methods are similar to the clustered Wilcoxon test in equation 6 for balanced designs. They differ from the methods for unbalanced designs in equation 10 according to the weight given to clusters of different size. Another approach, primarily applicable to the balanced design, is to summarize subunit-specific scores into a single score for a given cluster and perform power calculations using approaches for non-clustered ranked data from Lehmann (1998) or Rosner and Glynn (2009). However, as we have shown in Table 4, this can result in a loss of power relative to using subunit-specific scores within the context of a rank-based inference approach.

One practical advantage of our approach is that it is not necessary to specify a parametric family for the distribution of scores for controls, which is usually unknown. Instead, only an estimate of θ and of the correlation between probit scores for multiple subunits within the same cluster is required. We have shown, at least for the extreme value location-scale family, that power based on probit shift alternatives is similar to power based on location shift alternatives, for the same value of θ . One limitation is that the power estimates are only applicable for clustered data where all subunits in a cluster belong to the same group.

Acknowledgments

This work was supported by NIH grant EY12269 from the National Eye Institute. We acknowledge programming support from Marion McPhee and Rong Chen.

References

- Berson EL, Rosner B, Sandberg MA, et al. A randomized trial of vitamin A and vitamin E supplementation for retinitis pigmentosa. *Archives of Ophthalmology*. 1993; 111:761–772. [PubMed: 8512476]
- Berson EL, Rosner B, Sandberg MA, et al. Clinical trial of docosahexaenoic acid in patients with retinitis pigmentosa receiving vitamin A treatment. *Archives of Ophthalmology*. 2004a; 122:1297–1305. [PubMed: 15364708]

- Berson EL, Rosner B, Sandberg MA, et al. Further evaluation of docosahexaenoic acid in patients with retinitis pigmentosa receiving vitamin A treatment. *Archives of Ophthalmology*. 2004b; 122:1306–1314. [PubMed: 15364709]
- Browne RH. The t-test p value and its relationship to the effect size and $P(X > Y)$. *The American Statistician*. 2010; 64:30–33.
- Datta S, Satten GA. Rank sum tests for clustered data. *Journal of the American Statistical Association*. 2005; 100:908–915.
- Donner, A.; Klar, N. *Design and Analysis of Cluster Randomization Trials in Health Research*. London: Arnold; 2000.
- Hollander, M.; Wolfe, D. *Nonparametric Statistical Methods*. 2nd. New York: John Wiley & Sons, Inc.; 1999.
- Larocque, D. The Wilcoxon signed-rank test for cluster correlated data. In: Pierre, Duchesne; Bruno, Remillard, editors. *Statistical Modeling and Analysis for Complex Data Problems*. New York: Springer; 2005. p. 309–323.
- Lawless, J. *Statistical Models and Methods for Lifetime Data*. 2nd. New York: John Wiley & Sons, Inc.; 2003.
- Lehmann, E. *Nonparametrics: Statistical Methods Based on Ranks* (rev ed). New York: John Wiley & Sons, Inc.; 1998.
- Rosner B, Glynn RJ, Lee MLT. Incorporation of clustering effects for the Wilcoxon rank sum test: a large sample approach. *Biometrics*. 2003; 59:1089–1098. [PubMed: 14969489]
- Rosner B, Glynn RJ, Lee MLT. Extension of the rank sum test for clustered data: two-group comparisons with group membership defined at the subunit level. *Biometrics*. 2006; 62:1251–1259. [PubMed: 17156300]
- Rosner B, Glynn RJ. Power and sample size estimation for the Wilcoxon rank sum test with application to comparison of C statistics from alternative prediction models. *Biometrics*. 2009; 65:188–197. [PubMed: 18510654]
- SAS Institute. *SAS Version 9.1.3*. Cary, N. C.: 2006.
- Sorbinil Retinopathy Trial Research Group. A randomized trial of Sorbinil, an aldose reductase inhibitor, in diabetic retinopathy. *Archives of Ophthalmology*. 1990; 108:1234–1244. [PubMed: 2119168]
- Stata Corp. *Stata Statistical Software: Release 11*. College Station, TX: 2009.
- Wolfe DA, Hogg RV. On constructing statistics and reported data. *American Statistician*. 1971; 25:27–30.

Figure 1. Illustration of Probit Shift Alternative for the Sorbinil Retinopathy Trial Placebo Group Data

Table 1
Relationship Between Change Scores in Fellow Eyes, in the Placebo Group of the Sorbinil Retinopathy Trial.

Right Eye	Left Eye							Total
	-2	-1	0	1	2	3	4	
-2	0	0	0	0	0	0	0	0
-1	0	4	6	0	0	0	0	10
0	0	11	82	30	0	1	0	124
1	1	3	30	30	7	1	0	72
2	0	0	7	10	3	5	0	25
3	0	0	2	2	3	3	0	10
4	0	1	0	0	0	1	0	2
Total	1	19	127	72	13	11	0	243

Power of the Wilcoxon Rank Sum Test Under the Location Shift Alternative and the Probit Shift Alternative for the Extreme-Value Location-Scale Family of Distributions

Table 2

<i>m</i>		<i>n</i>	<i>d</i> [*]	Power	
				Location Shift	Probit Shift
50	50	0.25		0.186	0.186
		0.50		0.562	0.562
		0.75		0.890	0.891
70	30	1.0		0.990	0.991
		0.25		0.167	0.163
		0.50		0.490	0.489
30	70	0.75		0.816	0.829
		1.0		0.967	0.975
		0.25		0.159	0.163
		0.50		0.489	0.489
		0.75		0.840	0.829
		1.0		0.982	0.975

* *d* = effect size = *c*/*b*

Clustered Wilcoxon Test, simulation results based on 4000 simulations, nominal type I error = 0.05, for each combination of m , n , g and ρ .

Table 3

		ρ			
m	n	g	0.2	0.5	0.8
20	20	2	power*	0.887	0.804
				0.871	0.796
			power**		0.716
20	20	3	power*	0.955	0.856
				0.950	0.850
			power**		0.743
20	20	4	power*	0.979	0.881
				0.973	0.861
			power**		0.738
10	10	2	power ⁺	0.948	0.843
10	10	4	power ⁺⁺	0.928	0.827

* based on the theoretical power in equation 7

** based on the empirical power given by $\Pr(|Z_C| \geq 1.96)$

⁺ based on the theoretical power in equation 12

⁺⁺ based on the empirical power given by $\Pr(|Z_{C,s}| \geq 1.96)$

Table 4
Power of the Sorbinil Retinopathy Trial Based on Eye and Person Specific Change Scores.

<i>m</i>	<i>n</i>	unit of analysis	μ				
			0.1	0.2	0.3	0.4	0.5
100	100	eye	0.13	0.38	0.70	0.91	0.99
		person	0.10	0.26	0.51	0.75	0.91
200	200	eye	0.22	0.65	0.94	> 0.99	> 0.99
		person	0.16	0.47	0.81	0.97	> 0.99
300	300	eye	0.30	0.82	0.99	> 0.99	> 0.99
		person	0.21	0.63	0.94	> 0.99	> 0.99
400	400	eye	0.38	0.91	> 0.99	> 0.99	> 0.99
		person	0.26	0.76	0.98	> 0.99	> 0.99
500	500	eye	0.46	0.96	> 0.99	> 0.99	> 0.99
		person	0.32	0.85	> 0.99	> 0.99	> 0.99

Table 5

Power Calculation for a Proposed Study of the Effect of DHA Supplementation on Change in Visual Field Over 4 Years Among RP Patients Who Initiate Vitamin A Supplementation of 15,000 IU/day at Randomization

study	variable	treatment group [*]	Number of eyes available for analysis (g)		
			1	2	total
Previous Study	$m^{(g)}$	1	3	31	34
	$n^{(g)}$	2	2	27	29
	$\hat{\theta}^{(g)}$		0.167	0.699	0.669
	$\hat{\rho}$	1	—	0.575	
New Study		2		0.814	
		overall		0.706	
	$m^{(g)}$	1	5	45	50
	$n^{(g)}$	2	5	45	50
	$\text{var}_0\{\hat{\theta}^{(g)}\}$		0.0367	0.00315	0.002896 [†]
	$w_0^{(g)}$		27.3	318.0	
	$\text{var}_1\{\hat{\theta}^{(g)}\}$		0.0317	0.00269	0.002478 [‡]
	$w_1^{(g)}$		31.5	372.0	
	Power _{c,s}			0.899	

^{*} Treatment group 1 = control group taking 15,000 IU/day of Vitamin A + placebo Treatment group 2 = active group taking 15,000 IU/day of Vitamin A and 1200 mg/day of DHA.

[†] $\text{var}_0(\hat{\theta}_{c,s})$ in equation 12.

[‡] $\text{var}_1(\hat{\theta}_{c,s})$ in equation 12.