Paper 209-31

Exploiting the Link Between the Wilcoxon-Mann-Whitney Test and a Simple Odds Statistic

Ralph G. O'Brien, Cleveland Clinic Foundation, Cleveland, OH
John Castelloe, SAS Institute Inc., Cary, NC

ABSTRACT

Over a quarter-century ago, Alan Agresti (Biometrics, 1980) proposed using the generalized odds ratio (genOR) to summarize the association between two ordinal variables. Unfortunately, genOR is still largely unknown, even though it is elegantly straightforward and fills key voids in the working statistician's toolbox. An extension of it, a statistic we are calling "WMWodds," is an ideal effect-size measure for properly interpreting and reporting results based on the common Wilcoxon-Mann-Whitney (WMW) two-group test. In addition, the distribution theory suggests a sound and general way to perform sample-size analyses for the WMW test, an assertion strongly supported by Monte Carlo results. These matters are developed through realistic examples using SAS,® including a downloadable macro. This is an interim report of work still progressing; see www.bio.ri.ccf.org/robrien/WMWodds for later communications.

KEY WORDS: Wilcoxon rank-sum test, Mann-Whitney U test, effect size, generalized odds ratio, power, sample-size analysis

INTRODUCTION

How should you interpret results from the Wilcoxon-Mann-Whitney (WMW) two-group test? Some software, including PROC NPAR1WAY in SAS/STAT,® provides the mean ranks for each group, but these have no interpretation outside the study, and their expected difference increases as the sample size increases. Even worthy statistics books (and knowledgeable statisticians!) state that the WMW test compares the two medians, but this is only true in the rarest of cases in which the population distributions of the two groups are merely shifted versions of each other (i.e., differing only in location, and not shape or scale). In fact, a WMW statistic can have a p-value near 0.00 even when the two groups have *identical* sample medians.

Here, we present a simple and highly useful way to better understand, interpret, and present results from the Wilcoxon-Mann-Whitney test by estimating a simple odds parameter and computing its confidence interval. The methodology also appears to provide a sound approximation for computing statistical powers for the WMW test.

A SIMPLE ODDS PARAMETER FOR THE WILCOXON-MANN-WHITNEY (WMW) TEST

In an all-fictitious example, including the disease and experimental treatment, suppose Dr. Uri Ologist is testing the effectiveness of SHS67 in treating martinsarebia, a very painful chronic bladder disease that affects men 8 times more than women and has no effective treatment. Specifically, Dr. Ologist compared SHS67 to placebo in a balanced, randomized, double-blind study of males only. The primary outcome was self-reported overall improvement after four weeks on treatment, assessed using a seven-point Likert scale:

Compared to when I started this study, my condition is:

much worse	-3
worse	-2
slightly worse	-1
the same	0
slightly better	+1
better	+2
much better	+3

Suppose the counts for this trial are:

	Response							
	-3	-2	-1	0	+1	+2	+3	
Placebo	27	23	35	28	23	8	3	147
SHS67	11	34	30	27	19	15	12	148

Note first that the group sample medians are the same: slightly worse (-1). But this does not imply that the treatments produced identical outcomes. The Wilcoxon-Mann-Whitney test provides a better comparison, summarized in this snippet of output from the SAS/STAT procedure NPAR1WAY:

```
Group          N    Mean Score
============================
placebo      147         137.9
SHS67        148         158.1

  Wilcoxon Two-Sample Test
============================
Z                       -2.06
Two-Sided Pr > |Z|     0.0391
```

The higher mean rank for SHS67 (158 vs. 138) along with the 0.04 p-value supports the conclusion that the SHS67 patients reported better outcomes. But how much better?

Consider the following way to characterize the Mann-Whitney-Wilcoxon test. Let Y_1 and Y_2 be observations drawn independently from two distributions. Then the null hypothesis being tested is:

$$H_0 : \pi = 0.5, \quad \text{where}$$

$$\pi = \text{Prob}(Y_1 < Y_2) + \frac{1}{2}\text{Prob}(Y_1 = Y_2)$$

with the alternative hypothesis conforming to the sidedness of the test:

$$H_1 : \begin{cases} \pi > 0.5, & \text{upper 1-sided} \\ \pi < 0.5, & \text{lower 1-sided} \\ \pi \neq 0.5, & \text{2-sided} \end{cases}$$

In other words, H_0 asserts that it is equally likely for Y_1 to be less than Y_2 or greater than Y_2. This essential form of the hypothesis illustrates that the WMW test has nothing directly to do with means, median, or even the shapes for the distributions.

Now, converting π to an odds measure, we define

$$WMW_{\text{odds}} = \frac{\pi}{1-\pi}.$$

For example, if $WMW_{\text{odds}} = 2.0$, the odds are 2:1 that Y_1 is less than Y_2, splitting ties evenly.

Expressed in terms of WMW_{odds}, the hypothesis for the WMW test is:

$$H_0 : WMW_{\text{odds}} = 1$$

$$H_1 : \begin{cases} WMW_{\text{odds}} > 1, & \text{upper 1-sided} \\ WMW_{\text{odds}} < 1, & \text{lower 1-sided} \\ WMW_{\text{odds}} \neq 1, & \text{2-sided} \end{cases}$$

Estimating WMW_{odds} only involves counting properly, and it provides a clear way to quantify how much the two distributions differ in the manner examined by the WMW test. For the martinsarebia data, there are 10628 Y_1, Y_2 pairs having $Y_1 < Y_2$ (concordant), and 7650 pairs having $Y_1 > Y_2$ (discordant), and 3478 pairs having $Y_1 = Y_2$ (ties). Thus $\widehat{WMW}_{\text{odds}}$ = (10628 + 3478/2)/(7650 + 3478/2) = 1.32. In other words, for each pair randomly assigned to different groups, the one who took SHS67 was about 30% more likely to report a better outcome. The effect is not large, but it is hardly trivial either.

Of course, this 1.32 value is only an estimate. What is the sampling distribution of $\widehat{WMW}_{\text{odds}}$ or, say, $\ln(\widehat{WMW}_{\text{odds}})$? By knowing that, we can then answer: What is its 95% confidence interval? How "significantly" different is $\widehat{WMW}_{\text{odds}} = 1.32$ from 1.0?

THE SAMPLING DISTRIBUTION OF WMW_{ODDS}

A measure quite similar to WMW_{odds} is Agresti's (1980) generalized odds ratio:

$$genOR = \frac{\text{Prob}(Y_1 < Y_2)}{\text{Prob}(Y_1 > Y_2)}.$$

When used in a 2x2 table, it is identical to the usual odds ratio, hence the term "generalized odds ratio." But it fails to be a suitable effect size measure for the WMW test, because it ignores the ties rather than split them evenly, in effect "overstating" the group difference: $genOR \geq WMW_{\text{odds}}$, with equality holding when there are no ties. For the martinsarebia data, $\widehat{genOR}$ = 9985/5894 = 1.39 versus 1.32 for $\widehat{WMW}_{\text{odds}}$, perhaps a minor difference to Dr. Ologist, but we could have shown other examples with substantially greater discrepancies.

Fortunately, Agresti's formulas for $genOR$ can be readily applied to WMW_{odds}. $genOR$ is a simple transformation of the Goodman-Kruskall gamma (γ) statistic,

$$genOR = \frac{1+\gamma}{1-\gamma}$$

Accordingly, Agresti took known results for γ, in particular the asymptotic expression for $SE(\hat{\gamma})$, and applied the delta method to derive large-sample expressions for $genOR$, $\text{SE}(\widehat{genOR})$, $\ln(genOR)$, and $\text{SE}(\ln(\widehat{genOR}))$. The asymptotic ("in distribution") Normality of $\ln(\widehat{genOR})$ provides confidence limits and p-values for $genOR$.

We can extend $genOR$ to WMW_{odds} by modifying the data in a way that leaves $\widehat{WMW}_{\text{odds}}$ unaffected but destroys all ties between the two groups. Thus, $\widehat{genOR} = \widehat{WMW}_{\text{odds}}$ and Agresti's results apply immediately.

Using the martinsarebia example, one can see that the trick is rather simple. Recall the original table of counts:

	Response							
	-3	-2	-1	0	+1	+2	+3	
Placebo	27	23	35	28	23	8	3	147
SHS67	11	34	30	27	19	15	12	148

Now add and subtract some arbitrarily small quantity, $\epsilon > 0$, from each of the values in the second group, as so:

	Response									
	-3.1	-3.0	-2.9	-2.1	-2.0	-1.9	-1.1	-1.0	-0.9	...
Placebo	-	27	-	-	23	-	-	35	-	...
SHS67	5.5	-	5.5	17	-	17	15	-	15	...

Here, $\epsilon = 0.1$, but it can be any positive value small enough to preserve the ordering of the counts across the categories.

This 2 x 21 table of counts produces the same $\widehat{WMW}_{\text{odds}}$ = 1.32 value as we obtained from the original 2 $\times$ 7 table. On the other hand, $\widehat{genOR}$ is reduced from 1.39 to 1.32. Using Agresti's formulas for $genOR$ on the modified table yields a 95% confidence interval for WMW_{odds} of [1.01, 1.72], with p = 0.040 (corresponding to Z = 2.059). For comparison, the 95% confidence interval for $genOR$ is [1.02, 1.90], with p = 0.035 (corresponding to Z = 2.11). Since there are 3478 ties, $W\widehat{MW}_{\text{odds}}$ is closer to 1.0, and its corresponding Z statistic is smaller. Recall that the Wilcoxon-Mann-Whitney results were Z = 2.06 and p = 0.039.

Operating like odds ratios and hazard ratios, $W\widehat{MW}_{\text{odds}}$ and its confidence interval give us a simple effect size that can be compared with those obtained from any other ordinal (including continuous) responses from this or any other study. These can be effective displayed in a single forest plot (See Lewis and Clarke, 2001; free download http://bmj.bmjjournals.com/cgi/reprint/322/7300/1479.), as sketched here in Figure 1.

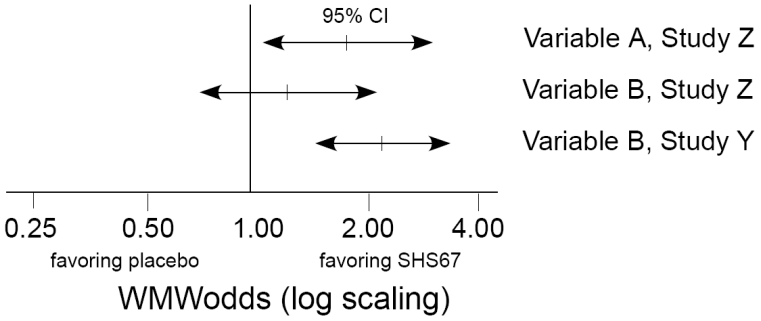

Figure 1. Forest plot

The SAS macro that performed these calculations is downloadable from www.bio.ri.ccf.org/robrien/WMWodds. Currently (January 2006), the call is:

```
%macro WMWodds
 (   label=,       /* a short single line of text                                */
     method=,      /* WMWODDS (default; aligns with WMW test for 2 group GENOR */
                   /*   (Agresti's Generalized Odds Ratio (ignores ties))       */
     dataset=,     /* SAS dataset name                                          */
     outcome=,     /* the outcome variable, ordinal or continuous, required     */
     group=,       /* the predictor variable, two groups or ordinal, required   */
     FlipGroups=,  /* YES, to flip group ordering, optional                     */
     weight=,      /* weight variable for PROC FREQ, optional, default = 1      */
     alphaCI=,     /* alpha value for conf. int., optional, default = 0.05      */
     print=        /* NO, to suppress FREQ output                               */
```

```
);
```

The martinsarebia example above used:

```
data martinsarebia;
  do SHS67 = 0 to 1;
    do MartinLikert = -3 to +3;
      input count @@;
      output;
      end;
    end;
datalines;
27 23 35 28 23  8  3
11 34 30 27 19 15 12
;
run;

%WMWodds ( label= GenOR analysis of Likert scores for placebo vs. SHS67,
           method = GenOR,
           dataset=martinsarebia,
           group=SHS67,
           outcome=MartinLikert,
           weight = count,
           alphaCI=0.05,
           print = no)

%WMWodds (label= WMWodds analysis of Likert scores for placebo vs. SHS67,
          method = WMWodds,
          dataset=martinsarebia,
          group=SHS67,
          outcome=MartinLikert,
          weight = count,
          alphaCI=0.05,
          print = no)
```

For the method=WMWodds run, the output contained:

```
WMWodds analysis:
-------------------------------
WMWodds analysis of SLikert scores for placebo vs. SHS67
Group variable: SHS67
Outcome variable: MartinLikert

WMWodds: 1.317
SE(WMWodds):  0.18

Confidence limits for WMWodds (alpha: 0.05)
  Based on Goodman-Krusal gamma limits: 1.012   1.732
  Based on log(WMWodds) -> Normal: 1.008   1.721
```

```
Testing Ho: WMWodds = 1, two tailed
  Z based on G-K gamma: 2.046   p: 0.041
  SE(WMWodds) under H0: 0.134
  Z based on log(WMWodds) -> Normal: 2.059   p: 0.040
```

Note that two sets of statistics are given here, one obtained from transforming the values obtained from PROC FREQ's computations for the Goodman-Kruskal gamma and the other obtained by working more directly on $\ln(\widehat{WMW}_{\text{odds}})$ and taking it to be Normally distributed. Formulas for $\widehat{WMW}_{\text{odds}}$, $\text{SE}(\widehat{WMW}_{\text{odds}})$, $\ln(\widehat{WMW}_{\text{odds}})$, and $\text{SE}(\ln(\widehat{WMW}_{\text{odds}}))$ are merely the sample analogues of those for the population values given in the "Computational Formulas for WMW_{odds}" section (page 7).

POWER ANALYSIS EXAMPLE

Dr. Ologist is encouraged by the favorable results ($p < 0.05$) obtained in his first study treating martinsarebia with SHS67 (see above), but being a savvy and prudent clinical scientist, he understands that the chance is much greater than 0.05 that this is a false positive (Lee and Zelen, 2000). He therefore plans to retest the same hypothesis in a second study and will use alpha = 0.01, not 0.05. In addition, instead of using the seven-point Likert scale described above, he decides to use a visual analog scale, shown in Figure 2.

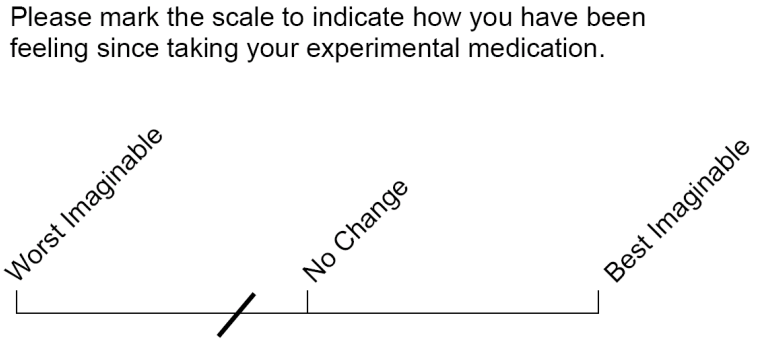

Figure 2. Visual analog scale

Using a scale from 0.00 to 1.00 in .01 increments, this subject would have scored 0.38. The scale is essentially continuous but there will still be ties, unless the sample size is quite small.

The design will again be balanced design and the VAS scores in the two groups will be compared using the Wilcoxon-Mann-Whitney test augmented with the WMW_{odds} estimate and confidence interval. What sample size is required to obtain 0.90 power?

What kind of group difference does Dr. Ologist envision now? His collaborating statistician, Dr. Dowtin Thomas, explains that such data may conform well to a beta distribution, and he first shows Dr. Ologist several possibilities for the placebo group. Ologist likes the shape of the beta(2.5, 2.0) distribution, which characterizes the natural human bias to report improvements in well-being, especially early in a clinical trial, no matter what group one is assigned to. Dr. Thomas suggests that a beta(3.0, 1.9) distribution be conjectured for the SHS67 group, in particular because this produces a population value for WMW_{odds} of 1.37, in line with the [1.01, 1.72] confidence interval from the previous study. Dr. Ologist accepts this scenario, which is illustrated in Figure 3.

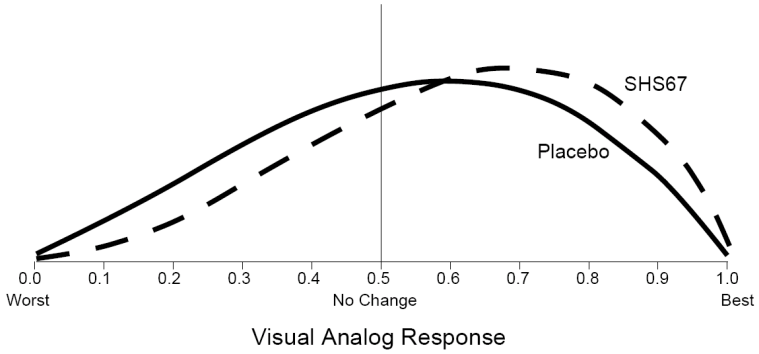

Figure 3. Scenario for placebo and SHS67 distributions

Applying the theory outlined in the "Computational Formulas for WMW_{odds}" section (page 7), the required total sample size is 600 (300 per group).

Suppose instead that Dr. Ologist was still going to use the Likert scale he used before, and he envisioned that the beta(2.5, 2.0) and beta(3.0, 1.9) distributions were the "latent" distributions that populated two 7-category ordinal distributions with intervals (0, 1/7] = -3, (1/7, 2/7] = -2, ... (6/7, 1) = +3. In this case, achieving 0.90 power (alpha = 0.05, 2-sided) would require 620 subjects total, only a 3.3% increase.

COMPUTATIONAL FORMULAS FOR WMW_{ODDS}

Note: This section can be skipped without misunderstanding the essentials. Equations for the sample statistics are not presented here, because they are isomorphic to those given for the power computations.

We have already mentioned (in the "The Sampling Distribution of WMW_{odds}" section (page 3)) that the computations for $\widehat{WMW}_{\text{odds}}$ can be performed by applying Agresti's (1980) results for $genOR$ to a contingency table modified to split all ties evenly. Identically, one can just augment Agresti's (1980) equations with correction terms for the ties. It is these extended equations that are given in this section.

While the power computation method is most directly applicable to the test using $\widehat{WMW}_{\text{odds}}$ (being derived from its asymptotic distribution, after all), it also appears to serve as a sound power approximation for the traditional flavors of the WMW test (such as those implemented in PROC NPAR1WAY in SAS/STAT). This is not surprising considering the simple relationship between WMW_{odds} and the traditional WMW test, with the former based on the odds parameter constructed from the latter.

Let Y_1 and Y_2 be independent observations from any two distributions that we wish to compare using a WMW test. For purposes of deriving the asymptotic distribution of $\widehat{WMW}_{\text{odds}}$ (and consequently the power computation as well), these distributions must be formulated as ordered categorical ("ordinal") distributions. In addition, all of the conditional probabilities (of response given group membership) must specified, along with the usual power analysis ingredients such as alpha, sample size per group, and sidedness of test.

If a distribution is not continuous, it can be discretized using a large number of categories with negligible loss of accuracy. Our suggested discretization method is to break each non-ordinal distribution into b categories (where the choice of b depends upon computational feasibility and desired accuracy), with breakpoints evenly spaced on the probability scale. That is, each bin contains an equal probability $1/b$ for that distribution. Then the breakpoints across both distributions are pooled to form a collection of C bins (heretofore called "categories"), and the probabilities of bin membership for each distribution are re-calculated. The motivation for this method of binning is to avoid degenerate representations of the distributions (i.e., small handfuls of large probabilities among mostly empty bins), as may be caused by something like an evenly spaced grid across raw values rather than probabilities.

After the discretization process just mentioned above, we now have two ordinal distributions, each with a set of probabilities across a common set of C ordered categories. For simplicity of notation, we assume (without loss of generality) the response values to be $1, \ldots, C$. Represent the conditional probabilities as

$$\tilde{p}_{ij} = \text{Prob}\left(Y_i = j \mid \text{group} = i\right), i \in \{1, 2\} \quad \text{and} \quad j \in \{1, \ldots, C\}$$

and the group allocation weights as

$$w_i = \frac{n_i}{N} = \text{Prob}\left(\text{group} = i\right), \quad i \in \{1, 2\}$$

The joint probabilities can then be calculated simply as

$$p_{ij} = \text{Prob}\left(\text{group} = i, Y_i = j\right) = w_i \tilde{p}_{ij}, i \in \{1, 2\} \quad \text{and} \quad j \in \{1, \ldots, C\}$$

The next step in the power computation is to compute the probabilities that a randomly chosen pair of observations from the two groups is concordant, discordant, or tied. It is useful to define these probabilities as functions of the terms Rs_{ij} and Rd_{ij}, defined as follows, where Y is a random observation drawn from the joint distribution across groups and categories:

$$\begin{aligned} Rs_{ij} &= \text{Prob}\left(Y \text{ is concordant with cell}(i,j)\right) + \frac{1}{2}\text{Prob}\left(Y \text{ is tied with cell}(i,j)\right) \\ &= \text{Prob}\left((\text{group} < i \text{ and } Y < j) \text{ or } (\text{group} > i \text{ and } Y > j)\right) + \\ &\quad \frac{1}{2}\text{Prob}\left(\text{group} \neq i \text{ and } Y = j\right) \\ &= \sum_{g=1}^{2}\sum_{c=1}^{C} w_g \tilde{p}_{gc}\left[\text{I}_{(g-i)(c-j)>0} + \frac{1}{2}\text{I}_{g\neq i, c=j}\right] \end{aligned}$$

and

$$\begin{aligned} Rd_{ij} &= \text{Prob}\left(Y \text{ is discordant with cell}(i,j)\right) + \frac{1}{2}\text{Prob}\left(Y \text{ is tied with cell}(i,j)\right) \\ &= \text{Prob}\left((\text{group} < i \text{ and } Y > j) \text{ or } (\text{group} > i \text{ and } Y < j)\right) + \\ &\quad \frac{1}{2}\text{Prob}\left(\text{group} \neq i \text{ and } Y = j\right) \\ &= \sum_{g=1}^{2}\sum_{c=1}^{C} w_g \tilde{p}_{gc}\left[\text{I}_{(g-i)(c-j)<0} + \frac{1}{2}\text{I}_{g\neq i, c=j}\right] \end{aligned}$$

For an independent random draw Y_1, Y_2 from the two distributions, we have

$$\begin{aligned} P_c &= \text{Prob}\left(Y_1, Y_2 \text{ concordant}\right) + \frac{1}{2}\text{Prob}\left(Y_1, Y_2 \text{ tied}\right) \\ &= \sum_{i=1}^{2}\sum_{j=1}^{C} w_i \tilde{p}_{ij} Rs_{ij} \end{aligned}$$

and

$$\begin{aligned} P_d &= \text{Prob}\left(Y_1, Y_2 \text{ discordant}\right) + \frac{1}{2}\text{Prob}\left(Y_1, Y_2 \text{ tied}\right) \\ &= \sum_{i=1}^{2}\sum_{j=1}^{C} w_i \tilde{p}_{ij} Rd_{ij} \end{aligned}$$

Then

$$WMW_{\text{odds}} = \frac{P_c}{P_d}$$

Proceeding to compute the theoretical standard error associated with WMW_{odds} (that is, the population analogue to the sample standard error), we have

$$\text{SE}(WMW_{\text{odds}}) = \frac{2}{P_d}\left[\sum_{i=1}^{2}\sum_{j=1}^{C} w_i \tilde{p}_{ij} \left(WMW_{\text{odds}} Rd_{ij} - Rs_{ij}\right)^2 / N\right]^{\frac{1}{2}}$$

Converting to the log scale using the delta method,

$$\text{SE}(\ln(WMW_{\text{odds}})) = \frac{\text{SE}(WMW_{\text{odds}})}{WMW_{\text{odds}}}$$

The next step is to produce a "smoothed" version of the $2 \times C$ cell probabilities that conforms to the null hypothesis $WMW_{\text{odds}} = 1$ (in other words, independence in the $2 \times C$ contingency table of probabilities). Let $\text{SE}_{H_0}(\ln(WMW_{\text{odds}}))$ denote the theoretical standard error of $\ln(WMW_{\text{odds}})$ assuming H_0.

Finally we have all of the terms needed to compute the power, using the noncentral Chi-square and normal distributions:

$$\text{power} = \begin{cases} P\left(Z \geq \frac{\text{SE}_{H_0}(\ln(WMW_{\text{odds}}))}{\text{SE}(\ln(WMW_{\text{odds}}))} z_{1-\alpha} - \delta^{\star} N^{\frac{1}{2}}\right), & \text{upper 1-sided} \\ P\left(Z \leq \frac{\text{SE}_{H_0}(\ln(WMW_{\text{odds}}))}{\text{SE}(\ln(WMW_{\text{odds}}))} z_{\alpha} - \delta^{\star} N^{\frac{1}{2}}\right), & \text{lower 1-sided} \\ P\left(\chi^2(1, (\delta^{\star})^2 N) \geq \left[\frac{\text{SE}_{H_0}(\ln(WMW_{\text{odds}}))}{\text{SE}(\ln(WMW_{\text{odds}}))}\right]^2 \chi^2_{1-\alpha}(1)\right), & \text{2-sided} \end{cases}$$

where

$$\delta^{\star} = \frac{\ln(WMW_{\text{odds}})}{N^{\frac{1}{2}} \text{SE}(\ln(WMW_{\text{odds}}))}$$

is the primary noncentrality, i.e., the "effect size" that quantifies how much the two conjectured distributions differ. Z is a standard normal random variable, $\chi^2(df, nc)$ is a noncentral χ^2 random variable with degrees of freedom df and noncentrality nc, and N is the total sample size.

MONTE CARLO STUDIES OF ACCURACY

How accurate is the WMW_{odds}-based power approximation when applied to the standard WMW test using the normal-based Z statistic? Focusing on the power neighborhood of 0.9, we use two sets of scenarios, one comparing a variety of different beta distributions (Table 1) and another comparing ordered categorical distributions (Table 2).

The standard beta(p, q) distribution is handy for specifying possible parent distributions, because it takes on very different shapes simply by changing its two parameters. The first five cases in Table 1 took Y_1 to be beta(2, 3), which has a bell-skewed desity with some moderate right skewness. For Y_2, we used beta($2 + a, 3 - a$), where a = 0.2 to 1.0 by 0.2, thus ending with beta(3, 2), which is moderately skewed left. The second five cases took Y_1 to beta(0.5, 1), which drops quickly after 0.00 and then levels off. For Y_2, we used beta($0.5 + a, 1 - a$), a = 0.1 to 0.5 by 0.1, ending with beta(1, 0.5), which rises gradually after 0.00 and then rises rapidly towards 1.0.

Table 1. Comparison of nominal and simulated powers with beta distributions

Parent 1	Parent 2	Total Sample Size	Nominal Power	Simulated Power	95 % Simulation 95% C.I.
beta(2, 3)	beta(2.2, 2.8)	1106	0.900	0.897	(0.883, 0.910)
beta(2, 3)	beta(2.4, 2.6)	282	0.900	0.898	(0.884, 0.911)
beta(2, 3)	beta(2.6, 2.4)	128	0.901	0.892	(0.878, 0.906)
beta(2, 3)	beta(2.8, 2.2)	74	0.904	0.904	(0.891, 0.916)
beta(2, 3)	beta(3.0, 2.0)	48	0.903	0.895	(0.882, 0.908)
beta(0.5, 1)	beta(0.60,0.90)	796	0.900	0.899	(0.886, 0.912)
beta(0.5, 1)	beta(0.70,0.80)	214	0.902	0.903	(0.890, 0.916)
beta(0.5, 1)	beta(0.80,0.70)	100	0.903	0.910	(0.898, 0.923)
beta(0.5, 1)	beta(0.90,0.60)	58	0.903	0.898	(0.885, 0.912)
beta(0.5, 1)	beta(1.00,0.50)	38	0.905	0.898	(0.885, 0.911)

The nominal powers in Table 1 were computed using the formulas given in the "Computational Formulas for WMW_{odds}" section (page 7) and used a balanced design, α = 0.05, sample sizes necessary to achieve powers around 0.9, and 1000 bins per distribution during discretization. The mean and 95% confidence interval for simulated power are based on 2000 Monte Carlo simulations, which a standard error about of about 0.0067 for estimating powers near 0.90. Note that all 10 nominal powers fall within the 95% confidence intervals.

Table 2 gives the corresponding results when the beta distributions described above are segmented into seven ordered categories, with cut-points at 1/7, 2/7, etc. For example, the ordered categorical distribution converted from the beta(2, 3) looks like this:

X	-3	-2	-1	0	+1	+2	+3
Prob(X)	0.100	0.223	0.250	0.213	0.140	0.063	0.010

Table 2 shows the nominal and simulated powers using the same sample sizes as in Table 1, again using a balanced design, α = 0.05 and 2000 Monte Carlo simulations. Again, all 10 nominal powers fell within the 95% confidence intervals.

Table 2. Comparison of nominal and simulated powers with 7-point Likert scales derived from beta distributions

Latent Parent 1	Latent Parent 2	Total Sample Size	Nominal Power	Simulated Power	95 % Simulation 95% C.I.
beta(2, 3)	beta(2.2, 2.8)	1106	0.889	0.893	(0.879, 0.907)
beta(2, 3)	beta(2.4, 2.6)	282	0.889	0.888	(0.874, 0.902)
beta(2, 3)	beta(2.6, 2.4)	128	0.891	0.887	(0.873, 0.901)
beta(2, 3)	beta(2.8, 2.2)	74	0.894	0.892	(0.879, 0.906)
beta(2, 3)	beta(3.0, 2.0)	48	0.893	0.899	(0.886, 0.912)
beta(0.5, 1)	beta(0.60, 0.90)	796	0.858	0.854	(0.838, 0.869)
beta(0.5, 1)	beta(0.70, 0.80)	214	0.870	0.884	(0.869, 0.898)
beta(0.5, 1)	beta(0.80, 0.70)	100	0.877	0.873	(0.858, 0.887)
beta(0.5, 1)	beta(0.90, 0.60)	74	0.881	0.879	(0.864, 0.893)
beta(0.5, 1)	beta(1.00, 0.50)	38	0.884	0.884	(0.870, 0.898)

The perceptive reader will note that there is little loss of power when the continuous beta variates are partitioned into the seven categories. The worst case occurred with beta(0.5, 1.0) versus beta(0.6, 0.9) which had 0.90 power when the data were kept continuous and 0.86 power when the data were categorized. Worse cases could be defined, such as beta(0.2, 3.0) versus beta(0.3, 2.8), which makes 59% of the (Y_1, Y_2) pairs tied in just first category $(X < 1/7)$. What is impressive, however, is that the nominal and simulated powers tracked so closely.

Though not reported here, similar Monte Carlo simulations were performed to assess the empirical Type I error rates. Those results were uniformly encouraging, even when the two distributions had markely different shapes, yet WMW_{odds} = 1.0.

We have also examined the empirical Type I error rates and powers for the the exact WMW test (as performed using the EXACT WILCOXON statement in PROC NPAR1WAY). Although we used only 200 trials in each simulation case, the results obtained were completely encouraging.

We have yet to compare these values with the methods of Kolassa (1995) and Whitehead (1993), both of which are tailored to scenarios conforming to the proportional odds model. The method of Hilton and Mehta (1993) is excluded from comparison because it is designed for the exact conditional form of the WMW test, whereas our scope here is limited to unconditional tests. The bootstrap method of Collins and and Hamilton (1988) is aimed at the classic shift alternative, a model that has limited appeal. We see no reason why our approximate approach will not work satisfactorily with either proportional odds or shift alternatives, but, of course, methods tailored specifically to them should work better.

CONCLUSION

We have demonstrated the utility of the WMW_{odds} parameter in providing interpretable estimates in the nonparametric comparison of two distributions with ordered values. The tests and confidence intervals formulated using WMW_{odds} are a viable alternative to the traditional WMW analysis, which lacks meaningful and generalizable summary statistics. In addition, because of the similarity between the hypotheses of the two tests (both splitting ties evenly), the asymptotic distribution of WMW_{odds} provides a promising power approximation not only for its own test statistic, but also for the WMW test.

It should also be noted that $genOR$ and WMW_{odds} also extend readily to G $>$ 2 ordered groups, a topic not explored here.

This paper is a preliminary communication of work in progress, and later communications will be posted to www.bio.ri.ccf.org/robrien/WMWodds.

Feedback is welcome.

REFERENCES

Agresti, A. (1980), "Generalized Odds Ratios for Ordinal Data," *Biometrics*, 36, 59–67.

Collings, B.J. and Hamilton, M.A. (1988), "Estimating the Power of the Two-Sample Wilcoxon Test for Location Shift," Biometrics, 44, 847–860.

Hilton, J.F. and Mehta, C.R. (1993), "Power and Sample Size Calculations for Exact Conditional Tests with Ordered Categorical Data," *Biometrics*, 49, 609–616.

Kolassa, J.E. (1995), "A Comparison of Size and Power Calculations for the Wilcoxon Statistic for Ordered Categorical Data," *Statistics in Medicine*, 14, 1577–1581.

Lee, S.J. and Zelen, M. (2000), "Clinical Trials and Sample Size Considerations: Another Perspective," *Statistical Science*, 15, 95–100.

Lewis, S and Clarke, M. (2001), "Forest plots: trying to see the wood and the trees," *BMJ*, 322, 1479–80.

Whitehead, J. (1993), "Sample Size Calculations for Ordered Categorical Data," *Statistics in Medicine*, 12, 2257–2271.

CONTACT INFORMATION

Ralph O'Brien, Cleveland Clinic Foundation
Department of Quantitative Health Sciences, WB-4

Cleveland, OH 44195
Email: obrien.ralph@gmail.com

John Castelloe, SAS Institute Inc.,
SAS Campus Drive, Cary, NC 27513
Email: john.castelloe@sas.com

Version 1.0