The Korean Communications
in Statistics Vol. 5, No. 3, 1998
pp. 707-722

A Nonparametric Test for Clinical Trial with Low Infection Rate

Mark C. K. Yang[1] and Donguk Kim[2]

Abstract

This paper evaluates a new clinical trial designs for low infection rate disease. This type of sparse disease reaction makes the traditional two sample t-test or Wilcoxon rank-sum test inefficient compared to a new test suggested. The new test, which is based solely on the larger changes, is shown to be more effective than existing method by simulation for small samples. However, this test can be shown to be connected to the locally most powerful rank test under certain practical conditions. This design is motivated in testing the treatment effects in periodontal disease research.

1. Introduction

A typical simple clinical trial experiment is to divide the patients into two groups; one with treatment, and the other serving as controls without treatment or with placebo. The decision of the treatment effect will then be evaluated by some relevant measures pertaining to the disease. In this paper, we consider the situation when the natural infection rate is low and the uninfected individuals will not be affected by the treatment but they can not be identified before treatment. This model is motivated in designing clinical trials for evaluating preventive treatment in periodontal disease. Periodontal disease is a very common disease everywhere in the world. It starts with inflammation at tooth gum tissues, eventually leads to the loss of attachment to the alveolar bone and tooth loss. According to a 1987 NIH report, 95% of the population over age 65 in US are affected by the disease. However, it is a very slow disease. It has been discovered in many demographic study that this disease can be considered as active only in a small proportion of the population at any given time period of 1 to 2 years (see e.g., Lindhe et al. (1989), Papapanou, et al. (1989), and Ismail et al. (1990)). Moreover, the activity of the disease cannot be easily observed. The current "gold standard" in diagnosing the disease is to measure the change of the attachment level by probing. If there is a drastic increase in the average attachment loss compared to the measurement error, then we can

1) Professor, Department of Statistics, University of Florida, Gainesville, Florida 32611, USA.
2) Assistant Professor, Department of Statistics, Sungkyunkwan University, Seoul, 110-745, Korea.

conclude that the disease is or has been active with this person. However, there is still no established method to predict this change by a patient's current condition. Thus we know a patient's activity status only after the fact.

In general, let the measurement pertaining to the disease be continuous, and $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ and $Y_{21}, Y_{22}, \ldots Y_{2n_2}$ denote, respectively, the measurements of n_1 subjects in the control group and those of n_2 subjects in the treatment group. Let $G_i(x)$ denote the distribution functions for $\{Y_{ij}\}_{j=1}^{n_i}$ for $i = 1, 2$. Then, we are testing

$$H_0 : G_1(x) = G_2(x) \quad \text{versus} \quad H_1 : G_1(x) \neq G_2(x), \tag{1.1}$$

or versus some more specific alternative hypotheses H_1.

To test (1.1), we may use the traditional t-test, hoping it is robust against the possible non-normality of the $G_1(x)$ under H_0. Or, we may choose a nonparametric test such as the Wilcoxon rank-sum test to make sure that the probability of making Type I error is protected regardless of the distribution of $G_1(x)$. In the low infection rate model, is the robustness for the t-test hold? And how efficient is the Wilcoxon rank-sum test? These are among many other questions we try to answer.

In the low infection rate model, we let under H_0,

$$G_1(x) = G_2(x) = (1-\pi)F_0(x) + \pi F_1(x), \tag{1.2}$$

where π is the infection rate, $F_0(x)$ and $F_1(x)$ are respectively the distribution functions of an individual who is not inflected and who is inflected without treatment. When the treatment is effective, we assume that the affected patient has a different distribution $F_2(x)$, but the unaffected patients remain to have the same distribution $F_0(x)$, i.e., under H_1, $G_1(x)$ is the same as (1.2), but

$$G_2(x) = (1-\pi)F_0(x) + \pi F_2(x). \tag{1.3}$$

In most situations $F_1(x)$ represents a shift in mean in $F_0(x)$, and $F_2(x)$ is expected to shift the mean of $F_1(x)$ back to $F_0(x)$. Suppose, without loss of generality, that $F_1(x)$ increases the mean. Then it is intuitive that only some of the larger values in both groups should be used for testing, because they are likely to come from the infected persons. The rest of the data are merely noise in evaluating the treatment effect. If the infected individuals can be identified, then to use only their data would be ideal. However, in many situations there is no clear threshold to separate an infected and an uninfected individual. Moreover, some of the infected individual may look uninfected due to the treatment. We assume that the

infection rate has been known with a great accuracy before we design the clinical trial. Let the estimated infection rate be π^*. Then a reasonable choice of the individuals to be included would be the upper $100\pi^*$ %. More specifically, let

$$N = n_1+n_2,\ m_i = [\pi^* n_i+0.5],\ i = 1, 2\ ;\ M = m_1 + m_2,$$

where $[x]$ denote the integer part of x. Define R_{ij} to be the rank of Y_{ij} in the pooled data $\{Y_{ij}\}$; $i = 1, 2$; $j = 1, \ldots, n_i$, and

$$a(x)=\begin{cases} x - (N - M), & \text{if } x>N - M \\ 0, & \text{elsewhere,} \end{cases}$$

$$S_i = \sum_{j=1}^{m_i} a(R_{ij}), \tag{1.4}$$

and H_0 is rejected if S_1 is too large, or equivalently, S_2 is too small. We will refer this test as the quantile rank test, following a similar rank test, the quartile test (see Hajek and Sidak(1967), p. 96), though it is used for a completely different purpose. If $\pi^*=1$ and $F_1(x)$ and $F_2(x)$ have the same distribution except location parameter, then our test coincides with the Wilcoxon rank-sum test.

Two other nonparametric tests seem also reasonable for testing H_0. Boos and Brownie(1986) considered the following model for the existence of non-respondents;

$$H_0 : G_1(x)=G_2(x)=F_0(x),\ \text{versus}\ H_1 : G_2(x)=(1-\pi)F_0(x)+\pi F_0(x-\triangle). \tag{1.5}$$

Apparently, the model assumes that only 100π % individuals on the average will respond to the treatment. Johnson, Verril and Moore(1987) generalize (1.5) to two different F's by letting,

$$H_1 : G_2(x) = (1-\pi)F_0(x) + \pi F_1(x). \tag{1.6}$$

Under the mixed normal alternative,

$$H_0 : G_1(x)=\Phi(x),\ \text{versus}\ H_1 : G_2(x)=(1-\pi)\Phi(x)+\pi\Phi(x-\triangle), \tag{1.7}$$

the test statistic by Johnson et al. based on the approximate scores is

$$\sum_{i=1}^{n_2}[e^{-\Phi^{-1}(u_i)^2/2}e^{\Phi^{-1}(u_i)}-1],\ 0<u_i<1,$$

where $u_i=\dfrac{\gamma_i}{n_1+n_2+1}$, and $\gamma_1\le\gamma_2\le\cdots\le\gamma_{n_2}$ are the ranks of $Y_{21}, Y_{22}, \ldots Y_{2n_2}$ in the

combined sample, and $\Phi(\cdot)$ is the standard normal cumulative distribution function.
It is easy to show that (1.5) and (1.6) are not equivalent to the quantile rank test (1.4). Since they are also nonparametric tests, they are valid to test the null hypothesis. We will include the Johnson et al.'s test in the comparison.

2. Properties of the quantile rank test

Theorem 2.1 Under H_0, the mean and variance of the quantile rank test statistic are

$$E(S_1) = \frac{n_1 M(M+1)}{2N} \quad \text{and,}$$

$$V(S_1) = \frac{n_1 n_2 M(M+1)}{12N^2(N-1)}[2N(2M+1)-3M(M+1)]. \tag{2.1}$$

Proof. See Appendix A. 1.

Note that if $\pi^*=1$, the mean and variance of S_1 coincide with those of the Wilcoxon rank-sum test, respectively. Moreover, the asymptotic normality holds under very mild condition on $G_1(x)$.

To show the optimality of this test, the distribution of the F's have to be specified. If we assume the F's are normally distributed with the same variance, then we will show that (1.4) is connected to the locally most powerful rank test for this model. The normality assumption for the data is quite reasonable for the attachment measurements (see e.g. Gunsolley and Best (1988), Yang et al. (1992), and Namgung and Yang(1994)). The asymptotic power is useful in determining the sample size for clinical trials. We obtain approximate score version of a locally most powerful rank test and a quantile rank test.

Theorem 2.2 Let $Y_{11}, \ldots, Y_{1n_1}$ are random sample from

$$G_1(y) = (1-\pi)\Phi(y) + \pi\Phi(y-\triangle), \tag{2.2}$$

and $Y_{21}, \ldots, Y_{2n_2}$ are random sample from

$$\begin{aligned} G_\theta(y) &= (1-\pi)\Phi(y) + \pi\Phi(y-\triangle') \\ &= (1-\pi)\Phi(y) + \pi\Phi(y-(\triangle-\theta)). \end{aligned} \tag{2.3}$$

The support of $G_\theta(y)$ is contained within the support of $G_1(y)$, and $G_1(y)$ and $G_\theta(y)$ have densities $f_1(y)$ and $f_\theta(y)$, where

$$f_1(y) = (1-\pi)\phi(y) + \pi\phi(y-\triangle),$$

$$f_\theta(y) = (1-\pi)\phi(y) + \pi\phi(y-(\triangle-\theta)),$$

where $\phi(y)$ is the standard normal density function.

Then, the locally most powerful rank test for testing $H_0 : \theta = 0$ versus $H_1 : \theta > 0$ rejects H_0 for large values of

$$\sum_{i=1}^{n_2} E\frac{-(V^{(\gamma_i)}-\triangle)}{\left[1+(\frac{1-\pi}{\pi})\exp^{\frac{1}{2}[-2\wedge V^{(\gamma)}+\wedge^2]}\right]}, \tag{2.4}$$

where $V^{(1)}\leq\cdots\leq V^{(n_1+n_2)}$ is an ordered sample from $G_1(y)$ and $\gamma_1\leq\gamma_2\leq\cdots\leq\gamma_{n_2}$ are ranks of $Y_{21},\ldots,Y_{2n_2}$ in the combined sample.

Proof. See Appendix A. 2.

Now we approximate the locally most powerful rank test by replacing the expectation by the score

$$J(\frac{\gamma_i}{n_1+n_2+1})=\frac{-(G_1^{-1}(\frac{\gamma_i}{n_1+n_2+1})-\triangle)}{\left[1+(\frac{1-\pi}{\pi})\exp^{\frac{1}{2}[-2\wedge G_1^{-1}(\frac{\gamma_i}{n_1+n_2+1})+\wedge^2]}\right]}.$$

Then the approximate locally most powerful rank test statistic is

$$\sum_{i=1}^{n_2} J(\frac{\gamma_i}{n_1+n_2+1}). \tag{2.5}$$

Also if we replace $V^{(\gamma_i)}$ by $S(\gamma_i)$,

$$S(\gamma_i)=\begin{cases}\gamma_i & , \text{ if } \gamma_i \geq N-M\\ 0 & , \text{ otherwise.}\end{cases}$$

Then we get

$$\sum_{i=1}^{n_2}\gamma_i I_{[\gamma_i\geq N-M]},$$

since $(\frac{1-\pi}{\pi})e^{\frac{1}{2}[-2\wedge\gamma_i+\wedge^2]}\rightarrow 0$, for $\gamma_i \geq N-M$.

Hence we reject H_0 for small values of

$$\sum_{i=1}^{n_2}\gamma_i I_{[\gamma_i\geq N-M]}, \tag{2.6}$$

an equivalent version of (1.4).

3. Comparison with other tests

Besides the t, Wilcoxon rank-sum, and Johnson et al.'s tests, another intuitive test should be considered. This last test is to use a threshold to classify a patient as infected or as uninfected and using the proportions of the infected individuals to test the treatment effect. The threshold may be known from pathological viewpoint or have to be artificially determined. In the comparison, we will make the threshold arbitrary and choose the threshold with the highest power. The details are given in Appendix A. 3.

For comparison, we compute the asymptotic power for the Wilcoxon rank-sum test for the alternative we are interested in. We obtain the asymptotic mean and variance for the Wilcoxon rank-sum test statistic W for our model.

Theorem 3.1 Let $Y_{11}, \cdots, Y_{1n_1}$ be a random sample from $G_1(y)$ given by (2.2) and $Y_{21}, \cdots, Y_{2n_2}$ be an independent random sample from $G_2(y)$ given by (2.3). Let $n_1 = n_2 = n$. Then the asymptotic mean and variance for the Wilcoxon rank-sum test for our model are

$$E(W) = \frac{1}{2} n((1+2p_1)n+1), \text{ and}$$
$$V(W) = n^2(p_1 - p_1^2 + (n-1)(p_2 + p_3 - 2p_1^2)), \tag{3.1}$$

where p_1, p_2 and p_3 can be defined as

$$\begin{aligned}
p_1 &= 0.5(1-\pi)^2 + \pi(1-\pi)[\Phi(\triangle'/\sqrt{2}) + \Phi(-\triangle/\sqrt{2})] + \pi^2\Phi((\triangle' - \triangle)/\sqrt{2}),\\
p_2 &= (1-\pi)^3\zeta(0,0) + 2(1-\pi)^2\pi\zeta(0, -\triangle') + (1-\pi)\pi^2\zeta(-\triangle', -\triangle')\\
&\quad + \pi(1-\pi)^2\zeta(\triangle, \triangle) + 2\pi^2(1-\pi)\zeta(\triangle, \triangle - \triangle') + \pi^3\zeta(\triangle - \triangle', \triangle - \triangle'),\\
p_3 &= (1-\pi)^3\zeta(0,0) + 2(1-\pi)^2\pi\zeta(0, \triangle) + (1-\pi)\pi^2\zeta(\triangle, \triangle)\\
&\quad + \pi(1-\pi)^2\zeta(-\triangle', -\triangle') + 2\pi^2(1-\pi)\zeta(-\triangle', \triangle - \triangle') + \pi^3\zeta(\triangle - \triangle', \triangle - \triangle'),
\end{aligned}$$

$$\zeta(x_1, x_2) = \int_{-\infty}^{-x_1/2} \Phi\left(\frac{-x_2 - t}{\sqrt{3/2}}\right)\frac{1}{\sqrt{\pi}} e^{-t^2} dt.$$

Proof. See Appendix A. 4.

Note that under H_0, we get $p_1 = \frac{1}{2}$ and $p_2 = p_3 = \frac{1}{3}$, and the asymptotic mean and variance for the Wilcoxon rank-sum test under the alternative coincide with the mean and variance under the null.

At $\pi = 0.1$, $\triangle = 4$, $\triangle' = 1$ and using Mathematica(1991), we get $p_1 = 0.473802$,

$p_2 = 0.320562$, $p_3 = 0.295$. Hence the asymptotic mean and variance of the Wilcoxon rank-sum test under the alternative (2.3) are

$$E(W) = 0.973802n^2 + \frac{n}{2}, \text{ and}$$
$$V(W) = 0.249314n^2 + 0.166584n^2(n-1). \tag{3.2}$$

For example, at the significance level 0.05, the asymptotic power for the Wilcoxon rank-sum test at $\pi = 0.1$, $\triangle = 4$, $\triangle' = 1$ are obtained to be 0.116 at $n = n_1 = n_2 = 50$, and 0.157 at $n = 100$. Also, at $\pi = 0.1$, $\triangle = 2$, $\triangle' = 1$, those asymptotic powers are 0.088 at $n = 50$, and 0.110 at $n = 100$.

There is always the question of the convergence rate for an asymptotic result. Small sample property and its comparison with other tests were performed with $n_1 = n_2 = 50$, 100, 150, 200, 300, 400 with $F_0(x) \sim N(0,1)$, $F_1(x) \sim N(\triangle, 1)$ and $F_2(x) \sim N(\triangle', 1)$, with $\triangle' = 1$ and $\triangle = 4$ and 2. It can be seen that this simulation should have covered a reasonable range of the sample size and the difference between the three F's. Other sample size are either too small or can be covered by the asymptotic approximation, and for any greater differences between $\triangle$ and $\triangle'$, the results tend to the extremes with very high or low powers.

Tables 1 and 2 give typical simulation results of empirical power. Each of the tabulated values is based on 6,000 simulations at the level of significance 0.05. Table 1 shows empirical powers at $\triangle = 4$ and $\triangle' = 1$ under the alternative, and Table 2 gives those powers at $\triangle = 2$ and $\triangle' = 1$ under the alternative. For each Table, the estimated infection rate π^* is 0.1, 0.15, and 0.2, but the actual infection rate ran from 0.10(0.05)0.25 to cover a large range of possible mis-specification infection rate.

From the simulation, our quantile rank test at $\pi = 0.1$, $\triangle = 4$, $\triangle' = 1$ has the power of 0.612 at $n = n_1 = n_2 = 50$, and 0.862 at $n = 100$. But, the power for the Wilcoxon rank-sum test is 0.117 at $n = 50$, and 0.156 at $n = 100$, which are very close to the asymptotic power from Appendix A. 4. Also, for comparison, at $\pi = 0.1$, $\triangle = 2$, $\triangle' = 1$, the power of the quantile rank test is 0.260 at $n = 50$, and 0.346 at $n = 100$, but that of the Wilcoxon rank-sum test is 0.088 at $n = 50$, and 0.110 at $n = 100$ from simulation. The new test has some advantage over Johnson's et al.'s test for small sample size n, but the power difference diminishes as n becomes large. Also, when π^* is overly estimated, i.e., $\pi^* = 0.2$, Johnson's et al.'s test does better. The overall performance of their test is acceptable compared to the new test, but the new test is much easier to do. We note that the performance of t-test is better than Wilcoxon test in the simulations, and t-test reach the significance level 0.05 with negligible error margins under the null and is robust in the low infection rate model.

Figure 1 gives power curves for the 4 methods with $\pi = 0.1$, $\triangle = 4$, $\triangle' = 1$ in the

alternative, and Figure 2 gives those power curves with $\pi = 0.1$, $\triangle = 2$, $\triangle' = 1$.

4. Example

The Periodontal Disease Research Center (PDRC) and the Fear and Anxiety Clinic (FAC) at the University of Florida was to conduct a clinical trial on the anxiety treatment effect on the peridontal disease. It is well known that disease can be caused by psychological factors (see O'Leary(1990)). Thus, it is not surprising that periodontal disease and other dental disease are found to be related to emotional stress (see e.g., De Marco(1976)). Patients with anxiety disorders will be selected from FAC and examined periodically at the PDRC. One group will be treated by cognitive behavior therapy for 12 months and the other group will served as control. We are interested in the treatment in reduction of attachment loss due to periodontal disease. Due to uncertainty of the distribution of attachment loss in the anxiety disorders group, we feel a nonparametric test is more appropriate. We are interested in the power of this test under various possible treatment effects. Due to low power of the Wilcoxon rank-sum test, we will use the quantile rank test with upper 20% of the data, because the infected rate is expected within the range of 10% to 20%. The final sample size is expected around 100 in both groups.

In a future study of treating periodontal disease patients with Augmentin or clindamycin, the new method analysis should also increase in detecting the treatment effect.

5. Conclusion

The results in the previous sections confirm one intuitive testing strategy in clinical trials: When the infection rate is low and the infected individuals can not be identified in the beginning of the trial, we should use only the extreme values at the end of the trial to form test statistic. A test, (1.4), based on this idea is constructed and shown to be more effective than the common used t or Wilcoxon rank test. It is simpler than and at least as powerful as other nonparametric tests designed for similar purposes. Though the new test depends on the prior knowledge of the infection rate, it is quite robust if this rate is misspecified.

Table 1: Empirical powers at $\triangle = 4$ and $\triangle' = 1$ under the alternative. The value π is the true infection rate and π^* is its estimate used to form the quantile rank test (1.4).

π	$n_1 = n_2$	F	Wilcoxon	Johnson	Quantile Rank		
					$\pi^* = 0.1$	$\pi^* = 0.15$	$\pi^* = 0.2$
0.10	50	0.287	0.117	0.413	0.612	0.456	0.384
	100	0.473	0.156	0.675	0.862	0.736	0.589
	150	0.625	0.189	0.823	0.942	0.843	0.750
	200	0.743	0.235	0.921	0.984	0.928	0.851
	300	0.873	0.305	0.992	0.997	0.985	0.951
	400	0.938	0.359	0.999	1.000	0.998	0.972
0.15	50	0.460	0.181	0.621	0.856	0.723	0.638
	100	0.724	0.260	0.898	0.989	0.955	0.884
	150	0.860	0.346	0.968	0.998	0.985	0.959
	200	0.939	0.409	0.996	1.000	0.999	0.993
	300	0.992	0.530	1.000	1.000	1.000	1.000
	400	0.999	0.626	1.000	1.000	1.000	1.000
0.20	50	0.632	0.261	0.807	0.967	0.903	0.851
	100	0.881	0.406	0.977	0.999	0.997	0.980
	150	0.965	0.531	0.998	1.000	1.000	0.999
	200	0.994	0.636	1.000	1.000	1.000	1.000
	300	1.000	0.780	1.000	1.000	1.000	1.000
	400	1.000	0.870	1.000	1.000	1.000	1.000
0.25	50	0.768	0.371	0.917	0.995	0.981	0.956
	100	0.958	0.577	0.997	1.000	1.000	0.998
	150	0.994	0.730	1.000	1.000	1.000	1.000
	200	0.999	0.835	1.000	1.000	1.000	1.000
	300	1.000	0.936	1.000	1.000	1.000	1.000
	400	1.000	0.974	1.000	1.000	1.000	1.000

Table 2: Empirical powers at $\triangle = 2$ and $\triangle' = 1$ under the alternative. The value π is the true infection rate and π^* is its estimate used to form the quantile rank test (1.4).

π	$n_1 = n_2$	F	Wilcoxon	Johnson	Quantile Rank		
					$\pi^* = 0.1$	$\pi^* = 0.15$	$\pi^* = 0.2$
0.10	50	0.117	0.088	0.189	0.260	0.209	0.187
	100	0.156	0.110	0.284	0.346	0.294	0.251
	150	0.190	0.124	0.351	0.421	0.364	0.343
	200	0.233	0.149	0.440	0.503	0.443	0.409
	300	0.291	0.174	0.588	0.639	0.590	0.514
	400	0.360	0.214	0.700	0.751	0.657	0.571
0.15	50	0.166	0.121	0.275	0.347	0.301	0.271
	100	0.230	0.155	0.431	0.516	0.462	0.409
	150	0.305	0.196	0.549	0.632	0.591	0.559
	200	0.361	0.228	0.656	0.731	0.689	0.649
	300	0.474	0.297	0.838	0.884	0.856	0.797
	400	0.568	0.357	0.906	0.937	0.903	0.843
0.20	50	0.210	0.149	0.365	0.454	0.400	0.376
	100	0.327	0.230	0.580	0.659	0.624	0.569
	150	0.429	0.288	0.719	0.791	0.759	0.749
	200	0.513	0.341	0.822	0.873	0.854	0.832
	300	0.655	0.454	0.943	0.968	0.966	0.941
	400	0.761	0.534	0.978	0.992	0.983	0.969
0.25	50	0.267	0.192	0.461	0.539	0.505	0.479
	100	0.430	0.313	0.704	0.759	0.746	0.708
	150	0.555	0.402	0.841	0.880	0.879	0.874
	200	0.661	0.478	0.914	0.936	0.938	0.930
	300	0.813	0.628	0.983	0.991	0.992	0.986
	400	0.895	0.724	0.994	0.997	0.998	0.997

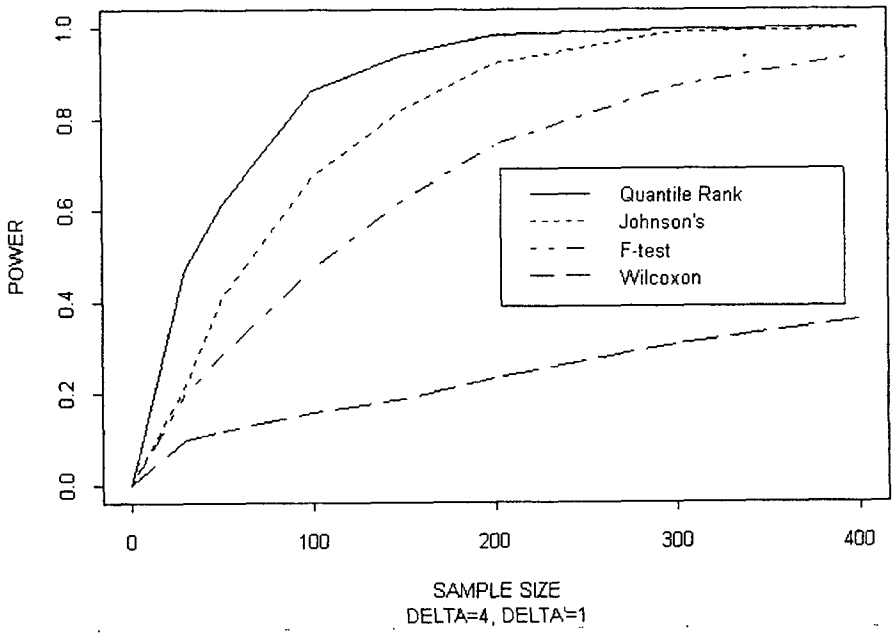

Figure 1. Power curves for the 4 methods with $\pi = 0.1$, $\triangle = 4$, $\triangle' = 1$ in the alternative.

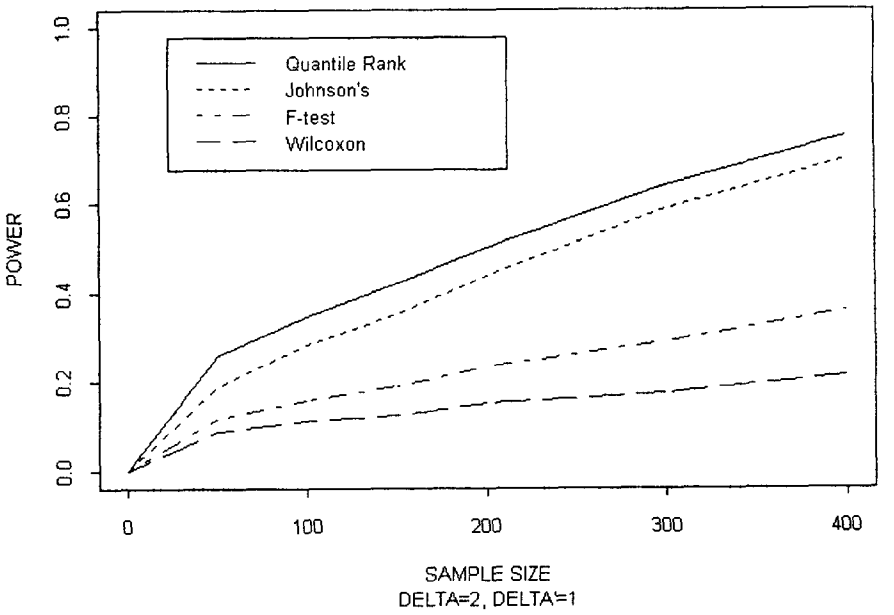

Figure 2. Power curves for the 4 methods with $\pi = 0.1$, $\triangle = 2$, $\triangle' = 1$ in the alternative.

Appendix

In this appendix we provide proofs of Theorem 2.1, Theorem 2.2 and Theorem 3.1 and derivation of the power by classifying a patient as infected or uninfected.

A. 1 Proof of Theorem 2.1

Following Hajek and Sidak(1967), we reindex R_{ij} as a single index R_k, $k=1, 2, \ldots, N$. Then S_1 in (1.4) is equivalent to

$$S_1 = \sum_{k=1}^{N} c_k a(R_k), \text{ with}$$

$$c_k = \begin{cases} 1 & k \leq n_1 \\ 0 & \textit{elsewhere.} \end{cases}$$

$$a(l) = \begin{cases} l & l \leq M \\ 0 & \textit{elsewhere.} \end{cases}$$

According to *Theorem c* in Hajek and Sidak(1967, P.61),

$$E(S_1) = N \cdot \bar{a} \cdot \bar{c}, \text{ and,}$$

$$V(S_1) = \sigma_a{}^2 \sum_{k=1}^{N} (c_i - \bar{c})^2,$$

where $\bar{a}$, $\bar{c}$, and $\sigma_a{}^2$ are the means of $a(l)$, c_k and the variance of $a(l)$ defined in the theorem. Then we get $\bar{a} = \frac{1}{N} \cdot \frac{M(M+1)}{2}$, $\bar{c} = \frac{n_1}{N}$, $\sum_{k=1}^{N} (c_i - \bar{c})^2 = \frac{n_1 n_2}{N}$, and $\sigma_a^2 = \frac{M(M+1)}{12N(N-1)} (2N(2M+1) - 3M(M+1))$. Both $E(S_1)$ and $V(S_1)$ can then be derived. Since we can show that the asymptotic normality holds under the alternative, the asymptotic normality holds under H_0 as a special case.

A. 2 Proof of Theorem 2.2

Refer to the general conditions in Hajek and Sidak(1967, pp70-73).

i) the density

$$d(y,\theta)=(1-\pi)\phi(y)+\pi\phi(y-(\triangle-\theta))$$

is absolute continuous in θ for every y.

ii) the limit

$$\begin{aligned}\dot{d}(y,0) &= \lim_{\theta\to 0}\frac{d(y,\theta)-d(y,0)}{\theta}\\ &= \pi\phi'(y-\triangle)\end{aligned}$$

exists for almost all y.

iii) $$\lim_{\theta\to 0}\int_{-\infty}^{\infty}|\dot{d}(y,\theta)|dy = \int_{-\infty}^{\infty}|-(y-\triangle)\cdot\pi\frac{1}{\sqrt{(2\cdot 3.14159)}}\exp^{\frac{-(y-\triangle)^2}{2}}|dy$$
$$= \pi E|y-\triangle| < \infty$$

holds.

If follows that the locally most powerful rank test is based on the test statistic

$$\sum_{i=1}^{n_1+n_2} c_i E\left\{\frac{\dot{d}(V^{(\gamma_i)},0)}{d(V^{(\gamma_i)},0)}\right\} = \sum_{i=1}^{n_2} E\left\{\frac{\dot{d}(V^{(\gamma_i)},0)}{d(V^{(\gamma_i)},0)}\right\},$$

where $V^{(1)}\leq\cdots\leq V^{(n_1+n_2)}$ is an ordered sample from $G_1(y)$ and $\gamma_1\leq\cdots\leq\gamma_{n_2}$ are the ranks of $Y_{21},\ldots,Y_{2n_2}$ in the combined sample.

The C_i's are 0's for $Y_{11},\ldots,Y_{1n_1}$ and 1's for $Y_{21},\ldots,Y_{2n_2}$.

Now

$$\begin{aligned}\frac{\dot{d}(y,0)}{d(y,0)} &= \frac{\pi\cdot\phi'(y-\triangle)}{(1-\pi)\phi(y)+\pi\phi(y-\triangle)}\\ &= \frac{-(y-\triangle)}{\left[1+\left(\frac{1-\pi}{\pi}\right)e^{\frac{-2\triangle y+\triangle^2}{2}}\right]}.\end{aligned}$$

Hence

$$\sum_{i=1}^{n_2} E\frac{\dot{d}(V^{(\gamma_i)},0)}{d(V^{(\gamma_i)},0)} = \sum_{i=1}^{n_2} E\left\{\frac{-(V^{(\gamma_i)}-\triangle)}{\left[1+\left(\frac{1-\pi}{\pi}\right)e^{\frac{-2\triangle V^{(\gamma)}+\triangle^2}{2}}\right]}\right\}.$$

Then the test with critical region

$$\sum_{i=1}^{n_2} E\left\{ \frac{-(V^{(r_i)}-\triangle)}{\left[1+(\frac{1-\pi}{\pi})e^{\frac{-2\triangle V^{(r)}+\triangle^2}{2}}\right]} \right\} \geq K$$

is the locally most powerful rank test for H_0 against $H_1 : \theta > 0$.

A. 3 Derivation of the power by classifying a patient as infected or uninfected

Let θ be threshold that divides the affected and non-infected, i.e., the subject is considered as infected if and only if $Y_{ij} > \theta$. Then the number of infected subjects in both groups are binomial variables. If we apply the normal approximation to the binomial distribution, it can be shown that the power at significance level α is

$$Power = 1 - \Phi(\frac{t_\alpha - (p_1 - p_2)}{\sqrt{p_1 q_1/n_1 + p_2 q_2/n_2}}),$$

where

$$\begin{aligned} t_\alpha &= z_\alpha\sqrt{pq(1/n_1+1/n_2)}, \text{ with} \\ p_1 &= (1-\pi)(1-\Phi(\theta))+\pi(1-\Phi(\theta-\triangle)), \quad q_1 = 1-p_1, \\ p_2 &= (1-\pi)(1-\Phi(\theta))+\pi(1-\Phi(\theta-\triangle')), \quad q_2 = 1-p_2, \\ p &= \frac{n_1p_1+n_2p_2}{n_1+n_2}, \quad q=1-p, \end{aligned}$$

and $\Phi(x)$ is the standard normal distribution function.

A. 4 Proof of Theorem 3.1

Using Lehmann(1975, pp. 70-71), the asymptotic mean and variance of the Wilcoxon rank-sum test can be written as

$$\begin{aligned} \mu &= n_1n_2p_1+\frac{1}{2}n_2(n_2+1), \\ \sigma^2 &= n_1n_2p_1(1-p_1)+n_1n_2(n_2-1)(p_2-p_1^2)+n_2n_1(n_1-1)(p_3-p_1^2), \end{aligned}$$

where p_1, p_2 and p_3 can be expressed in terms of independent random variables X, X', Y, Y',

$X, X' \sim G_1(x)$, $p_1 = \Pr\{X < Y\}$,
$Y, Y' \sim G_2(x)$, $p_2 = \Pr\{X < Y \text{ and } X < Y'\}$,
X, X', Y, Y', all independent, $p_3 = \Pr\{X < Y \text{ and } X' < Y\}$.

The three probabilities p_1, p_2 and p_3 can be computed by sums of one-dimensional integral for our models, i.e.,

$$
\begin{aligned}
p_1 &= 0.5(1-\pi)^2+\pi(1-\pi)[\Phi(\triangle'/\sqrt{2})+\Phi(-\triangle/\sqrt{2})]+\pi^2\Phi((\triangle'-\triangle)/\sqrt{2}),\\
p_2 &= (1-\pi)^3\zeta(0,0)+2(1-\pi)^2\pi\zeta(0,-\triangle')+(1-\pi)\pi^2\zeta(-\triangle',-\triangle')\\
&\quad +\pi(1-\pi)^2\zeta(\triangle,\triangle)+2\pi^2(1-\pi)\zeta(\triangle,\triangle-\triangle')+\pi^3\zeta(\triangle-\triangle',\triangle-\triangle'),\\
p_3 &= (1-\pi)^3\zeta(0,0)+2(1-\pi)^2\pi\zeta(0,\triangle)+(1-\pi)\pi^2\zeta(\triangle,\triangle)\\
&\quad +\pi(1-\pi)^2\zeta(-\triangle',-\triangle')+2\pi^2(1-\pi)\zeta(-\triangle',\triangle-\triangle')+\pi^3\zeta(\triangle-\triangle',\triangle-\triangle'),\\
\zeta(x_1,x_2) &= \int_{-\infty}^{-x_1/2}\Phi\left(\frac{-x_2-t}{\sqrt{3/2}}\right)\frac{1}{\sqrt{\pi}}e^{-t^2}dt,
\end{aligned}
$$

where $\Phi(x)$ and $\phi(x)$ represent the standard normal distribution and density functions.

Acknowledgements

The authors are grateful to a referee for helpful comments and suggestions.

References

[1] Boos, D. D. and Brownie, C. (1986). Testing for a treatment effect in the presence of nonresponders, *Biometrics*, Vol. 42, 191-197.

[2] De Marco, T. J. (1976). Periodontal emotional stress syndrome, *Journal of Periodontology*, Vol. 47, 67-68.

[3] Good, P. I. (1979). Detection of a treatment effect when not all experimental subjects will respond to treatment, *Biometrics*, Vol. 35, 483-489.

[4] Gunsolley, J.C. and Best, A.M. (1988). Change in attachment level, *Journal of Periodontology*, Vol. 59, 450-456.

[5] Hajek, J. and Sidak, Z. (1967). *Theory of Rank Tests*, Academic Press, New York.

[6] Ismail, A.I., Morrison, E.C., Burt, B.A., Caffesse, R.G. and Kavanagh, M.T. (1990). Natural history of periodontal disease in adults: Findings from the Tecumseh periodontal disease study, 1959-87, *Journal of Dental Research*, Vol. 69, 430-435.

[7] Johnson, R.A., Verrill, S. and Moore, D.H. (1987). Two sample rank test for detecting changes that occur in a small proportion of the treated population, *Biometrics*, Vol. 43, 641-655.

[8] Lehmann, E. L. (1975). *Nonparametrics : Statistical Methods Based on Ranks*,