Rank-Sum Tests for Clustered Data

Somnath DATTA and Glen A. SATTEN

The Wilcoxon rank-sum test is widely used to test the equality of two populations, because it makes fewer distributional assumptions than parametric procedures such as the *t*-test. However, the Wilcoxon rank-sum test can be used only if data are independent. When data are clustered, tests based on generalized estimating equations (GEEs) that generalize the *t*-test have been proposed. Here we develop a rank-sum test that can be used when data are clustered. As an application, we use our rank-sum test to develop a nonparametric test of association between a genetic marker and a quantitative trait locus. We also give a rank-sum test for equivalence of three or more populations that generalizes the Kruskal–Wallis test to situations with clustered data. Unlike previous rank tests for clustered data, our proposal is valid when members of the same cluster belong to different groups, or when the correlation between cluster members differs across groups.

KEY WORDS: Association; Clustered data; Kruskal–Wallis test; Quantitative trait; Rank test; Transmission disequilibrium test; Wilcoxon test.

1. INTRODUCTION

The Wilcoxon rank-sum test is an attractive way to compare two groups, and has become a standard procedure among working statisticians. The Wilcoxon test is calculated by pooling observations from the two groups, ranking the pooled observations and then computing the sum of rankings corresponding to observations from one of the groups. The usual assumption for the applicability of Wilcoxon test is that all the observations in the study are independent. However, in many practical situation, there are clusters of correlated observations. Examples of clustered data include repeated measurement of blood pressure from a single individual, responses of litter mates in an experiment using rodents, or body mass index of siblings. Clustered data are typically analyzed using generalized estimating equations (GEEs) to account for correlation between observations to obtain consistent variance estimates. Although model-free, testing hypotheses about two groups using GEEs corresponds to a variance-adjusted *t*-test and is not invariant to monotonic transformations of the data as rank-based procedures are.

In this article we propose rank-sum tests for comparing two groups when data are clustered. We first note that simply averaging the response within clusters and then applying a rank-sum test for independent data may give a test with improper size, because the null hypothesis of equal distribution between groups may be violated if the two groups have different distributions of cluster sizes. Moreover, this simple approach is not available when members of the same cluster may belong to different groups. Additionally, the correlation between cluster members may depend on group membership.

Two broad approaches are possible when constructing a rank test for clustered data. First, one can make assumptions about the nature of the clustering, for example, assuming that cluster members are exchangeable and that the correlation structure within clusters is independent of group. Under these assumptions, Rosner, Glynn, and Lee (2003) recently proposed a rank test for clustered data that in essence stratifies on cluster size to create a rank-sum statistic for clustered data for the case when cluster members necessarily belong to the same group. Here we take a different approach that makes no assumptions on the nature of the clustering, extending an idea for parameter estimation of Hoffman, Sen, and Weinberg (2001) and Williamson, Datta, and Satten (2003) to hypothesis testing. The resulting test is valid in a wide variety of settings, including when members of the same cluster belong to different groups or when the correlation structure depends on group membership.

The rest of the article is organized as follows. In Section 2 we present the notation and the theoretical development of our tests. Section 2.3 contains expressions that generalize our rank-sum test to more than two groups. Section 2.4 contrasts our test to that of Rosner et al. (2003); and Section 2.5 reports results from a simulation study comparing our procedure with the standard Wilcoxon statistic calculated using cluster-averaged response and the Rosner et al. statistic. In Section 3 we show how our approach can be applied in statistical genetics to develop rank-based quantitative trait transmission-disequilibrium tests (qTDTs). Using simulated data, we compare our rank-based qTDT with a *t*-test proposed by Xiong, Krushkal, and Boerwinkle (1998) and apply our test to data on circulating angiotensin-1 converting enzyme (ACE) levels. The main text ends with a discussion in Section 4. The proof of the asymptotic null distribution of our test statistic is deferred to the Appendix.

2. NOTATION AND GENERAL THEORY

Let M denote the number of clusters and let X_{ik} denote the kth observation in the ith cluster, $1 \le k \le n_i$, $1 \le i \le M$, where n_i denotes the number of observations for the ith cluster. Let g_{ik} denote group membership (0 or 1) of the kth observation in the ith cluster and let $\sum_k g_{ik} = n_{i1}$ be the number of members of group 1 in ith cluster. Our data consist of $(\mathbf{X}, \mathbf{g}) := \{X_{ik}, g_{ik} : 1 \le k \le n_i; 1 \le i \le M\}$. To avoid specifying a probability model for cluster sizes and the number of observations from each group in every cluster, we condition on n_i and n_{i1}; that is, we assume that these quantities are fixed. Assume that the g_{ik}, for a given cluster i, are identically distributed. Further assume that although observations from the same cluster may have arbitrary dependence, observations from different clusters are independent. The null hypothesis that we consider is that observations from the two groups follow the same distribution, that is,

$$P(X_{ik} \le x | g_{ik} = 0, n_i, n_{i1}) = P(X_{ik} \le x | g_{ik} = 1, n_i, n_{i1}) = F(x), \quad (1)$$

Somnath Datta is Professor, Department of Statistics, University of Georgia, Athens, GA 30602 (E-mail: *datta@stat.uga.edu*). Glen A. Satten is Mathematical Statistician, Centers for Disease Control and Prevention, Atlanta, GA 30333 (E-mail: *gsatten@cdc.gov*). The first author's research was supported in part by the Centers for Disease Control and Prevention. The authors thank Martin Farrall for generously providing the genetic data analyzed in Section 3.2, and also gratefully acknowledge assistance from Gonçalo Abecasis.

Journal of the American Statistical Association
September 2005, Vol. 100, No. 471, Theory and Methods
DOI 10.1198/016214504000001583

for some (unknown) distribution function, F, for any pair (i, k).

Our proposed statistic is best introduced in terms of the following Monte Carlo test. Suppose that from each cluster we sampled a single individual k_i at random, and then denoted the response X_{ik_i} by X_i^* and the group membership by g_i^*. It is not hard to verify that (X_i^*, g_i^*) are independent $\sim F \times \text{bin}(1, n_{i1}/n_i)$, for $1 \le i \le M$. Using (X_i^*, g_i^*), we could construct a Wilcoxon rank-rum statistic,

$$W^* = \frac{1}{M+1}\sum_{i=1}^{M} g_i^* R_i^*,$$

where R_i^* is the rank of X_i^* among the set $\{X_j^*, 1 \le j \le M\}$. Our proposed test statistic corresponds to averaging W^* over all possible choices of the (X_i^*, g_i^*) values given the observed data. Hence we propose inference on

$$Z = \frac{S - E(S)}{\sqrt{\widehat{\text{var}}(S)}}, \tag{2}$$

where $S = E(W^*|\mathbf{X}, \mathbf{g})$. This averaging is motivated by recent proposals by Hoffman et al. (2001), Rieger, Kaplan, and Weinberg (2001), and Williamson et al. (2003). Note that even when data are clustered, $Z^* := \{W^* - E(W^*)\}/\sqrt{\text{var}(W^*)}$ can be used as a valid test of the null hypothesis (1), because (X_i^*, g_i^*) are independent. However, this test is unappealing for several reasons, including that it is inefficient, using only one observation per cluster, and it depends on the particular observations chosen from each cluster. The averaging approach leading to the test statistic of (2) avoids these difficulties, because it is an explicitly calculable function of all of the data.

The following steps are necessary before (2) can be recommended. First, we must be able to calculate the quantities $S = E(W^*|\mathbf{X}, \mathbf{g})$, $E(S)$, and $\widehat{\text{var}}(S)$. Second, we must establish the (asymptotic) distribution of Z. Finally, we must evaluate the performance of tests based on (2).

2.1 Calculation of Required Quantities

We first calculate the quantities needed to compute (2). To allow for ties in the data, we use the mid-rank (the unweighted average of all possible rankings of an observation). Hence,

$$R_i^* = 1 + \frac{1}{2}\left\{\sum_{j\neq i} I(X_j^* \le X_i^*) + \sum_{j\neq i} I(X_j^* < X_i^*)\right\}. \tag{3}$$

The value of $S = E(W^*|\mathbf{X}, \mathbf{g})$ can be obtained using (3) as follows. We write

$$\begin{aligned}
&E(R_i^* g_i^*|\mathbf{X}, \mathbf{g}) \\
&\quad = E\left[\frac{g_i^*}{2}\left\{\sum_{j\neq i} I(X_j^* \le X_i^*) + \sum_{j\neq i} I(X_j^* < X_i^*)\right\} + g_i^* \middle| \mathbf{X}, \mathbf{g}\right] \\
&\quad = \frac{1}{2}\sum_{j\neq i} E\{g_i^* I(X_j^* \le X_i^*)|\mathbf{X}, \mathbf{g}\} \\
&\qquad + \frac{1}{2}\sum_{j\neq i} E\{g_i^* I(X_j^* < X_i^*)|\mathbf{X}, \mathbf{g}\} + \frac{n_{i1}}{n_i} \\
&\quad = \frac{1}{2}\sum_{j\neq i} E\{g_i^* F_j(X_i^*)|\mathbf{X}, \mathbf{g}\} \\
&\qquad + \frac{1}{2}\sum_{j\neq i} E\{g_i^* F_j(X_i^*-)|\mathbf{X}, \mathbf{g}\} + \frac{n_{i1}}{n_i} \\
&\quad = \frac{1}{n_i}\sum_{j\neq i}\sum_{k=1}^{n_i} g_{ik}\frac{F_j(X_{ik}) + F_j(X_{ik}-)}{2} + \frac{n_{i1}}{n_i},
\end{aligned}$$

where $F_i(x) = (n_i)^{-1}\sum_{k=1}^{n_i} I[X_{ik} \le x]$ is the empirical distribution function of observations from the ith cluster. Therefore,

$$\begin{aligned}
S &= E(W^*|\mathbf{X}, \mathbf{g}) \\
&= \frac{1}{(M+1)} \\
&\qquad \times \sum_{i=1}^{M}\sum_{k=1}^{n_i}\frac{g_{ik}}{n_i}\left[1 + \frac{1}{2}\sum_{j\neq i}\{F_j(X_{ik}) + F_j(X_{ik}-)\}\right].
\end{aligned} \tag{4}$$

Next we note that $E(S) = E(W^*)$. The unconditional expected value of W^* can be obtained by first conditioning on the vector of indicators $\mathbf{g}^* = (g_1^*, \ldots, g_M^*)$, so that

$$\begin{aligned}
E(S) = E(W^*) &= E\{E(W^*|\mathbf{g}^*)\} \\
&= E\left(\frac{1}{2}\sum_{i=1}^{M} g_i^*\right) = \frac{1}{2}\sum_{i=1}^{M}\frac{n_{i1}}{n_i}.
\end{aligned} \tag{5}$$

To calculate $\text{var}(S)$, we use the Hajek projection of S. Let $\mathbf{X}_i, \mathbf{g}_i$ denote the data from cluster i and let $S_i := E\{S|\mathbf{X}_i, \mathbf{g}_i\}$. To facilitate calculation of S_i, note that

$$\begin{aligned}
&E\big[g_{ik}\{F_j(X_{ik}) + F_j(X_{ik}-)\}|\mathbf{X}_i, \mathbf{g}_i\big] \\
&\quad = g_{ik}[F(X_{ik}) + F(X_{ik}-)] \quad \text{for } j \neq i, \\
&E\big[g_{jk}\{F_i(X_{jk}) + F_i(X_{jk}-)\}|\mathbf{X}_i, \mathbf{g}_i\big] \\
&\quad = \frac{n_{j1}}{n_j}\left[2 - \frac{1}{n_i}\sum_{k=1}^{n_i}\{F(X_{ik}) + F(X_{ik}-)\}\right] \quad \text{for } j \neq i,
\end{aligned}$$

and

$$E\big[g_{jk}\{F_h(X_{jk}) + F_h(X_{jk}-)\}|\mathbf{X}_i, \mathbf{g}_i\big] = \frac{n_{j1}}{2n_j} \quad \text{for } i \neq j, i \neq h, j \neq h.$$

After some algebra, we find that

$$S_i = c_i + W_i,$$

where

$$W_i = \frac{1}{2n_i(M+1)}\sum_{k=1}^{n_i}\left\{(M-1)g_{ik} - \sum_{j\neq i}^{M}\frac{n_{j1}}{n_j}\right\} \times \{F(X_{ik}) + F(X_{ik}-)\} \tag{6}$$

and c_i does not depend on $\mathbf{X}_i, \mathbf{g}_i$ and hence will not contribute to $\text{var}(S)$ calculated using S_i. Taking expectations, we get

$$\begin{aligned}
E(W_i) &= \frac{1}{2(M+1)}\left\{(M-1)\frac{n_{i1}}{n_i} - \sum_{j\neq i}^{M}\frac{n_{j1}}{n_j}\right\} \\
&= \frac{M}{2(M+1)}\left\{\frac{n_{i1}}{n_i} - \frac{1}{M}\sum_{j=1}^{M}\frac{n_{j1}}{n_j}\right\}.
\end{aligned}$$

Finally, var(S) is estimated using

$$\widehat{\text{var}}(S) = \sum_{i=1}^{M} \{\widehat{W}_i - E(W_i)\}^2, \tag{7}$$

where $\widehat{W}_i$ is as in (6) with F replaced by its pooled estimate,

$$\widehat{F} = \left(\sum_{i=1}^{M} n_i F_i\right) \Big/ \left(\sum_{i=1}^{M} n_i\right).$$

It is not hard to see that the numerator of (2) reduces to the numerator of the mid-rank–based Wilcoxon rank-sum statistic (Hudgens and Satten 2002) when there is no clustering (i.e., $n_i \equiv 1$), which further reduces to the numerator of the usual Wilcoxon test statistic when no ties are present. However, the variance (7) does not reduce to the usual Wilcoxon variance $M_0 M_1/12(M+1)$, where M_i is the observations in group i, but it is easily shown that (7) is asymptotically equivalent to the usual Wilcoxon variance in the absence of clustering (i.e., the ratio of the two variance estimates converges to 1 in probability). This issue is also addressed in Section 4.

We next illustrate the calculation of our new statistic using the dataset in Table 1, comprising nine observations in three clusters. Note that group membership cuts across clusters. The quantity $\{F_j(X_{ik}) + F_j(X_{ik}-)\}/2$ is the proportion of observations in cluster j that are less than X_{ik} (where we count observations in cluster j that are equal to X_{ik} as having half of their mass less than X_{ik}). Hence for observation 2 (belonging to cluster 1), we find that $\sum_{j\neq 1}\{F_j(X_{12}) + F_j(X_{12}-)\}/2 = \frac{3}{8} + \frac{1}{6} = \frac{13}{24}$, because 3/2 of 4 observations in cluster 2 are counted as less than $X_{ik} = 4$ and 1/2 of 3 observations in cluster 3 are counted as less than 4. Hence, using (4), $S = E(W^*|\mathbf{X}, \mathbf{g}) = \frac{1}{4}\{\frac{(1+13/24)}{2} + \frac{(1+4/3)+(1+3/2)}{4} + \frac{(1+9/8)+(1+2)}{3}\} = \frac{59}{64} \approx .92$. It is easily verified that this value can also be obtained by averaging the $2 \cdot 3 \cdot 4 = 24$ Wilcoxon statistics W^* obtained by selecting one observation at a time from each group, when mid-ranks are used for the tied observations. Further, using (5), $E(S) = \frac{1}{2}(\frac{1}{2} + \frac{1}{2} + \frac{2}{3}) = \frac{5}{6} \approx .83$. Note that $S > E(S)$, indicating a tendency for members of group 1 to have higher X_{ik} values. To calculate $\widehat{\text{var}}(S)$, we calculate $\widehat{W}_i$ for each cluster. For cluster 1, $\widehat{W}_1 = \frac{1}{2\cdot 4}\{-(\frac{2}{4} + \frac{2}{3})\frac{1}{18} + [2 - (\frac{2}{4} + \frac{2}{3})]\frac{7}{18}\} = \frac{14}{432}$ and $E(W_1) = \frac{3}{2\cdot 4}[\frac{1}{2} - \frac{1}{3}(\frac{1}{2} + \frac{1}{2} + \frac{2}{3})] = -\frac{1}{48}$. Similarly, we find that $\widehat{W}_2 = \frac{55}{1{,}728}$, $E(W_2) = -\frac{1}{48}$, $\widehat{W}_3 = \frac{5}{108}$, and $E(W_3) = \frac{1}{24}$. Finally, $\widehat{\text{var}}(S) = (\frac{14}{432} + \frac{1}{48})^2 + (\frac{55}{1{,}728} + \frac{1}{48})^2 + (\frac{5}{108} - \frac{1}{24})^2 = \frac{5{,}603}{995{,}328} \approx .00563$. Using these results, we obtain $(\frac{59}{64} - \frac{5}{6})/\sqrt{.00563} \approx 1.18$ as the final test statistic.

2.2 Asymptotic Distribution of the Rank-Sum Test for Clustered Data

Asymptotic normality under the null hypothesis of the standardized test (2) can be established under the following mild regularity conditions.

Condition 1. $\sum_{i=1}^{M}(n_i/N)^2 \to 0$, as $M \to \infty$, where $N = \sum_{i=1}^{M} n_i$ is the total sample size.

Condition 2. With W_i given by (6),

$$\liminf_{M\to\infty} M^{-1} \sum_{i=1}^{M} \text{var}(W_i) > 0.$$

Recall the definitions of test statistic S, $E(S)$, and $\widehat{\text{var}}(S)$ given by (4), (5), and (7).

Theorem 1. Under Conditions 1 and 2,

$$\frac{S - E(S)}{\sqrt{\widehat{\text{var}}(S)}} \xrightarrow{d} \text{N}(0, 1), \quad \text{as } M \to \infty.$$

Condition 1 is a sample size condition needed for the consistency of the pooled estimate $\widehat{F}$ of the common marginal distribution function. It is satisfied if, for example, the n_i are bounded. Condition 2 is a technical condition needed to ensure that the estimated asymptotic variance $\widehat{\text{var}}(S)$ given by (7) can be used in the standardization of the test statistic. Because $W_i = W_{iM}$ is an average based on potentially dependent variables, it is not possible to give simpler sufficient conditions for Condition 2 in general. However, for the special case when the variables within each cluster are independent or positive dependent, it can be seen by direct calculation (see the App.) that for all large enough M,

$$\text{var}(W_i) \geq .05 \frac{(\alpha_i - \overline{\alpha})^2}{n_i}, \tag{8}$$

where

$$\alpha_i = \frac{n_{i1}}{n_i}, \qquad \overline{\alpha} = M^{-1} \sum_{i=1}^{M} \frac{n_{i1}}{n_i}.$$

Table 1. Synthetic Data to Illustrate Calculation of Our Proposed Test Statistic

ID no.	Cluster (i)	Member (k)	X_{ik}	g_{ik}	$\frac{1}{2}\sum_{j\neq i}\{F_j(X_{ik}) + F_j(X_{ik}-)\}$	$\frac{1}{2}\{\widehat{F}(X_{ik}) + \widehat{F}(X_{ik}-)\}$
1	1	1	1	0	0	$\frac{1}{18}$
2	1	2	4	1	$\frac{13}{24}$	$\frac{7}{18}$
3	2	1	2	0	1	$\frac{3}{18}$
4	2	2	4	0	$\frac{5}{12}$	$\frac{7}{18}$
5	2	3	6	1	$\frac{4}{3}$	$\frac{11}{18}$
6	2	4	7	1	$\frac{3}{2}$	$\frac{7}{9}$
7	3	1	4	1	$\frac{9}{8}$	$\frac{7}{18}$
8	3	2	7	0	$\frac{23}{8}$	$\frac{7}{9}$
9	3	3	8	1	2	$\frac{17}{18}$

Therefore, in this case a sufficient condition for Condition 2 involving just the sample sizes is

$$\liminf_{M\to\infty} M^{-1}\sum_{i=1}^{M}\frac{(\alpha_i-\overline{\alpha})^2}{n_i} > 0.$$

2.3 Extension to *m* Groups

Next we consider an extension of our procedure to the case when individuals within each clusters may belong to one of m possible groups. The null hypothesis to be tested in this context is that the marginal distributions are the same in all of the groups. As before, let g_{ik} denote the group status ($=j$ if it belongs to the jth group; $1 \le j \le m$) of the kth individual in the ith group.

Corresponding to each group $j \in \{1,\ldots,m\}$, define the corresponding rank-sum statistic $S^{(j)}$ obtained by (4), but with g_{ik} replaced by $g_{ik}^{(j)} = I(g_{ik}=j)$. Similarly, $E(S^{(j)})$ under the null hypothesis can be obtained from (5), but with n_{i1} replaced by n_{ij}, the number of group j individuals in cluster i. We propose a statistic that compares $S^{(j)}$ with its expected value $E(S^{(j)})$ under the null hypothesis. Furthermore, let $W_i^{(j)}$ be as in (6), but with the g's replaced by $g^{(j)}$ and n_{i1} replaced by n_{ij}, and define $\widehat{W}_i^{(j)}$ likewise. Calculate the spectral decomposition of $\widehat{\mathbf{V}} := M^{-1}\sum_{i=1}^{M}\{\widehat{\mathbf{W}}_i - E(\mathbf{W}_i)\}\{\widehat{\mathbf{W}}_i - E(\mathbf{W}_i)\}^T = \sum_{j=1}^{m}\widehat{\lambda}_{(j)}^{(M)}\widehat{\mathbf{P}}_j^{(M)}\widehat{\mathbf{P}}_j^{(M)T}$, where $\widehat{\mathbf{W}}_i - E(\mathbf{W}_i)$ is the m-vector with components $\widehat{W}_i^{(j)} - E(W_i^{(j)})$, $\widehat{\lambda}_{(1)}^{(M)} \ge \cdots \ge \widehat{\lambda}_{(m)}^{(M)}$ are the (ordered) eigenvalues of $\widehat{\mathbf{V}}$, and $\widehat{\mathbf{P}}_j^{(M)}$ are the corresponding orthonormal eigenvectors. Let $\widehat{\mathbf{V}}^- = \sum_{j=1}^{m-1}\{\widehat{\lambda}_{(j)}^{(M)}\}^{-1}\widehat{\mathbf{P}}_j^{(M)}\widehat{\mathbf{P}}_j^{(M)T}$. Then one would reject the null hypothesis of equality of marginal distributions across groups for large values of the test statistic

$$T = M^{-1}\{\mathbf{S}-E(\mathbf{S})\}^T\widehat{\mathbf{V}}^-\{\mathbf{S}-E(\mathbf{S})\}, \tag{9}$$

where $\mathbf{S}-E(\mathbf{S})$ is the m-vector with components $S^{(j)}-E(S^{(j)})$. This test can be considered a form of the Kruskal–Wallis test for clustered data. Under appropriate regularity conditions, T will have an asymptotic chi-squared distribution with $m-1$ degrees of freedom under the null hypothesis.

Condition 3. Let $\lambda_{(m-1)}^{(M)}$ denote the second smallest eigenvalue of the matrix $M^{-1}\sum_{i=1}^{M}E\{\mathbf{W}_i - E(\mathbf{W}_i)\}\{\mathbf{W}_i - E(\mathbf{W}_i)\}^T$. Assume that

$$\liminf_{M\to\infty}\lambda_{(m-1)}^{(M)} > 0.$$

This condition ensures that the variance–covariance matrix of $\mathbf{S}-E(\mathbf{S})$ has rank $m-1$. We now state the following asymptotic distribution theorem, proof of which is deferred to the Appendix.

Theorem 2. Under Conditions 1 and 3,

$$T \xrightarrow{d} \chi^2_{m-1}, \quad \text{as } M\to\infty,$$

where T is given by (9).

2.4 Comparison to the Rosner, Glynn, and Lee Test and Other Rank-Sum Tests

Recently, Rosner et al. (hereafter RGL) (2003) proposed a rank test for clustered data for the case where all observations in the same cluster belong to the same group. In addition, the RGL test assumes that cluster members are exchangeable and that the correlation structure within clusters is independent of group. The RGL test stratifies on cluster size; after ranking all observations (ignoring clustering), it compares the rank sum of observations from group 1 having cluster size k to the fraction of group 1 clusters of size k times the total rank sum of observations from clusters of size k. The final statistic sums the comparisons over all cluster sizes k and divides by an appropriate standard deviation.

The nature of the RGL statistic gives an indication of situations where good or poor performance can be expected. Because RGL compares groups at each cluster size, imbalance in the distribution of groups across cluster size strata will result in inefficiency. At the extreme, all data from cluster sizes for which only one group is represented are ignored by the RGL test. Additionally, by scoring clusters by the sum of ranks of cluster members, we can anticipate that RGL will perform best when correlation is weak; as correlation increases, the effective number of independent observations per cluster drops, so that RGL overweights larger clusters. By the same reasoning, our new approach may be less efficient than RGL when correlation is weak, because large clusters are underweighted. We explore these predictions in the context of a simulation study in the next section.

2.4.1 Simulation Study. Here we report a small simulation study to assess the properties of our proposed test statistic, to compare our test with both the RGL tests and the standard Wilcoxon tests that ignores any clustering, and to illustrate a difficulty in using the standard Wilcoxon rank-sum test that treats the cluster as the experimental unit and uses the within-cluster average as the cluster response. We consider the situation where all members of a cluster belong to the same group, which is a requirement when computing the RGL test and when using the within-cluster average response. In Section 3 we present results for our proposed test in a more complex simulation that has both within-cluster correlation and members of the same cluster belonging to different groups.

Let M_0 and M_1 denote the number of clusters whose members are in group 0 and group 1. As $g_{ik} = g_{ik'}$ we drop the second index on g for this section only. We generated data $X_{ik} = \exp(Y_{ik}) + g_i\delta$, where $\mathbf{Y}_i = (Y_{ik},\ldots,Y_{in_i})$ were independent multivariate normal variates with mean $\mathbf{0}$ and exchangeable variance–covariance matrix $(1-\rho_{g_i})\mathbf{I} + \rho_{g_i}\mathbf{1}$, where $\mathbf{I}$ is the $n_i \times n_i$ identity matrix and $\mathbf{1}$ is the $n_i \times n_i$ matrix with all entries equal to 1. Note that when $\rho_0 = \rho_1 = 0$, there is no clustering; that is, data in each group are iid. We chose cluster sizes to allow for the possibility of informative cluster sizes. Clusters with $g_i = 0$ had $n_i = 2$ with probability p_c and cluster size $n_i = 5$ with probability $1-p_c$, whereas clusters with $g_i = 1$ had $n_i = 5$ with probability p_c and cluster size $n_i = 2$ with probability $1-p_c$. Note that for $p_c \ne .5$, the distribution of cluster sizes is different between the two groups (i.e., cluster size is informative). Simulation results for various values of M_0, M_1, p_c, ρ_0, ρ_1, and $\delta = 0$ (null) or $\delta = .5$ (alternative) are given in Table 2.

Table 2. Simulation Results for Size ($\delta = 0$) and Power ($\delta = .5$), for Cluster Average (CA), RGL, New Test (DS), and Standard Wilcoxon (W) Ignoring Cluster Effect

						Size				Power			
Distribution type	M_0	M_1	p_c	ρ_0	ρ_1	*CA*	*RGL**	*DS*	*W*	*CA*	*RGL**	*DS*	*W*
Continuous	10	10	0	0	0	.087		.052	.050	.40		.45	.55
	10	10	.2	0	0	.064	.044	.053	.049	.35	.39	.49	.60
	25	25	.2	0	0	.079	.049	.051	.050	.70	.78	.87	.93
Discrete	10	10	0	0	0	.087		.053	.049	.35		.34	.38
	10	10	.2	0	0	.061	.044	.052	.049	.29	.26	.36	.42
	25	25	.2	0	0	.081	.049	.051	.050	.63	.60	.71	.79
Continuous	25	25	.5	0	0	.049	.049	.051	.050	.57	.94	.91	.95
	25	25	.5	.9	.9	.050	.049	.052	.320	.49	.46	.53	.82
	25	25	.2	.9	−.1	.320	.170	.051	.170	.83	.56	.65	.83

NOTE: The nominal size of all tests is .05.

*The size and power of RGL statistic based on only those simulated datasets for which the RGL test is defined.

To explore the effect of discrete data, for some simulations we replaced X_{ik} by the largest integer less than or equal to X_{ik}. The resulting distribution assigns mass to nonnegative integers and has a 90th percentile between 2 and 3, resulting in heavily tied data. For the RGL test when data are tied, we used the mid-rank in all expressions.

The simulation results in Table 2 illustrate several points. The size of our proposed test was very close to nominal for all simulation conditions. Our new test also performed well even when group membership completely determined cluster size ($p_c = 0$), whereas the RGL test cannot be used for this case. The size of the RGL test was also close to nominal for $p_c > 0$, except when clusters from different groups had different correlation structure ($\rho_0 \neq \rho_1$). We also found that our new test had appropriate size even for heavily tied data, as did the RGL test using mid-ranks. Not surprisingly, power for all tests was lower for discrete data than for continuous data.

For simulations with $p_c = .2$, the power of our new test was higher than that of the RGL test. However, when p_c was increased to .5, the RGL test had higher power. This is because the RGL test can perform poorly when the distribution of cluster sizes differs across groups. Note, however, that even when $p_c = .5$, if ρ_0 and ρ_1 are increased to .9, then the power of the RGL test decreases to below that of our new statistic. When cluster members are very correlated, the RGL statistic overweights large clusters, because it uses a sum of ranks of all cluster members even when the cluster effectively contributes only one observation. Finally, when the correlation between cluster members differs across groups (i.e., when $\rho_0 \neq \rho_1$), only our new test maintained size.

For all simulations having $\rho_0 = \rho_1 = 0$, the standard Wilcoxon test ignoring the clustering is valid. Not surprisingly, the standard Wilcoxon test outperformed both our new test and the RGL test for these simulations. However, for simulations with $\rho_0 \neq 0$ and $\rho_1 \neq 0$, the size of the standard Wilcoxon was far from the nominal .05 level. When cluster size was informative ($p_c \neq .5$), the size of the Wilcoxon test that used the average cluster response was larger than nominal, even as the power was lower than that of our new test. When $p_c = .5$, so that the two groups have the same distribution of cluster sizes, then the Wilcoxon test using the average cluster response was properly sized, but still had lower power than either our new test or the RGL test.

3. TESTING WHEN CLUSTER MEMBERS CAN BELONG TO DIFFERENT GROUPS

In this section we consider the situation where group and cluster memberships do not coincide. As a concrete example, we can consider testing the difference in blood pressure between boys and girls when some study participants are siblings. In this case clusters correspond to sets of siblings, whereas the groups correspond to boys or girls. The restriction that all members of the same cluster belong to the same group would correspond to the requirement that only families composed of only boys or only girls be included in the study. There is no currently available rank-based approach to test group differences when group and cluster memberships do not coincide.

3.1 Testing for Linkage and Association Between a Marker Locus and a Quantitative Trait Locus

Genetic epidemiologists use family-based association tests to determine whether alleles at a marker locus (a locus where genetic variation can be measured but where genetic variability may not be directly related to disease mechanism) are associated (or correlated) with alleles at a locus that does directly affect a trait of interest. Association between alleles at marker and trait loci can arise either because of confounding (called "population stratification" in statistical genetics) or because the marker and trait loci are located in close proximity to each other on the same chromosome. This latter case is called "linkage," and a finding of both association and linkage between alleles at marker and trait loci is an important step in mapping trait loci. Transmission-disequilibrium tests (TDTs) use data from nuclear families and take nonnull values only when both association and linkage exist between marker and trait loci. TDTs for quantitative traits are referred to as qTDTs.

Xiong et al. (1998) (XKB hereafter) proposed a qTDT based on a t-test. Consider a marker locus with two alleles, a and A. We consider data from nuclear families in which at least one parent is heterozygous (i.e., has genotype aA), and compare the trait values X between those children who received the a allele from their heterozygous parent and those children who received the A allele from their heterozygous parent. If the ith family has only one heterozygous parent, then n_i is the number of offspring, X_{ik} is the trait value of the kth offspring, and $g_{ik} = 1$ (0) if the heterozygous parent transmitted the a (A) allele to the kth offspring. If the

ith family has two heterozygous parents, then n_i is *twice* the number of offspring, $\mathbf{X}_i = (X_{i1}, X_{i1}, X_{i2}, X_{i2}, \ldots, X_{io_i}, X_{io_i})$ and $\mathbf{g}_i = (F_{i1}, M_{i1}, F_{i2}, M_{i2}, \ldots, F_{io_i}, M_{io_i})$, where $F_{ik} = 1$ (0) if the father transmitted the a (A) allele to the kth offspring and M_{ik} is defined similarly for mothers. Note that for the heterozygous offspring of two heterozygous parents, we do not actually know whether it was the mother or father who transmitted the a allele, but it is only important that one value of g is 1 and the other is 0.

Given data $\mathbf{X}_i$ and $\mathbf{g}_i$ for M families, XKB (1998) proposed the t-test,

$$T = \frac{(\overline{X}_1 - \overline{X}_0)}{\sqrt{(1/m_1 + 1/m_0)S^2}},$$

where $m_r = \sum_{i=1}^{M} \sum_{k=1}^{n_i} I[g_{ik} = r]$,

$$\overline{X}_r = \frac{1}{m_r} \sum_{i=1}^{M} \sum_{k=1}^{n_i} I[g_{ik} = r] X_{ik} \quad \text{for } r = 0, 1,$$

and

$$S^2 = \frac{1}{(m_1 + m_0 - 2)} \times \sum_{i=1}^{M} \sum_{k=1}^{n_i} \{ g_{ik}(X_{ik} - \overline{X}_1)^2 + (1 - g_{ik})(X_{ik} - \overline{X}_0)^2 \},$$

to compare whether the trait values of offspring that received the a or A alleles differ. XKB showed that such a difference is evidence of association and linkage disequilibrium. We propose replacing the t-test by the expected value of the Wilcoxon rank-sum test averaged over all possible samples of one transmission event per family.

3.2 A Simulation Study for the Quantitative Trait Transmission-Disequilibrium Tests

To compare the performance of our test with the XKB test using simulated data, we simulated data that mimic the inheritance of genetic material and heritable traits. We generated parental genotypes using a binomial distribution with parameter p, the proportion of a alleles. We assumed that each parental allele was equally likely to be transmitted to each offspring (unselected sampling). We generated parental trait values X_{iF} and X_{iM} independently from a distribution F and generated offspring trait values using

$$X_{ik} = \alpha c_{ik} + \beta(X_{iF} + X_{iM}) + \epsilon_{ik}, \qquad k = 1, \ldots, o_i,$$

where c_{ik} and X_{ik} are the number of a alleles and the trait value for the kth offspring, ϵ_{ik} are iid random variables with distribution F, and α and β are constants. A non-0 value of α corresponds to an (additive) effect of genotype at the marker locus on trait values, whereas a non-0 value of β corresponds to familial correlation (possibly due to genetic effects at loci that are not linked to the marker locus).

Each simulated dataset contained data on 50 families. The number of offspring per family followed a uniform distribution on the integers 1, 2, 3, and 4. For each simulation, we generated 100,000 datasets. To determine the effect of the underlying distribution F, we simulated phenotypes (of both parent and offspring) using either a N(0, 1) or a lognormal distribution with log-mean 0 and log-variance 1. Results are shown in Table 3, where we tabulate the size and power for tests with a nominal size of .05.

Table 3. Probability of Rejecting the Null Hypothesis by the qTDTs in a Simulation Experiment

α (additive effect)	β (family effect)	F (error distribution)	Probability of rejection	
			XKB	Rank
0	0	N(0, 1)	.052	.048
0	0	LN(0, 1)	.047	.045
0	.5	LN(0, 1)	.050	.045
.5	0	N(0, 1)	.850	.670
.5	0	LN(0, 1)	.370	.740
.9	.5	LN(0, 1)	.610	.750

When $\alpha = 0$ (no effect of the locus on phenotype), both the XKB and our new test had appropriate size, irrespective of the distribution of phenotypes or the presence of a parental (or random) effect. When phenotypes were normally distributed, the XKB test had higher power to detect departures from $\alpha = 0$ than the rank test. However, when phenotypes were lognormally distributed, the rank test had higher power than the XKB test. Interestingly, a parental (or random) effect decreased the difference in power between the rank and XKB tests; comparing simulations having $\alpha = .5$ and $\beta = 0$ with simulations having $\alpha = .9$ and $\beta = .5$, we see that the power of the rank test is approximately the same, but the power of the XKB test is increased for the simulations having $\beta > 0$.

3.3 Application to Data on Circulating Angiotensin-1–Converting Enzyme Levels

Data on circulating ACE levels in 69 British families were reported by Keavney et al. (1998). ACE levels were standardized separately for men and women; in addition, probands were genotyped at 10 marker loci in the ACE gene. We abstracted data on nuclear families by selecting from each pedigree only those members having no offspring and both parents in the pedigree. Of these nuclear families, we used only those in which both parents were genotyped. Offspring whose circulating ACE levels were not measured were excluded from the analysis. We tested marker locus I/D for linkage to and association with a quantitative trait locus that affects circulating ACE levels. We selected this marker locus because it had the greatest evidence of association in an analysis of these data by Abecasis, Cookson, and Cardon (2000). Of the original 69 pedigrees, we used only 37 nuclear families in the analysis. The joint distribution of the number of offspring and the number of heterozygous parents is given in Table 4. From this table, we see that values of n_i for these data ranged from 1 to 10 (corresponding to twice the number of offspring in a family with two heterozygous parents). Using (2), we obtained $Z = -3.81$ ($p = 1.4 \times 10^{-4}$), giving strong evidence of linkage and association. The sign of the

Table 4. Family Size and Parental Heterozygosity in the ACE Data of Keavney et al. (1998)

Number of heterozygous parents	Number of offspring						
	1	2	3	4	5	6	Total
1	6	9	8	0	0	1	24
2	1	5	5	1	1	0	13
Total	7	14	13	1	1	1	37

statistic indicates that the allele coded 1 corresponds to lower levels of circulating ACE.

4. DISCUSSION

As noted previously, our calculation of the variance of our test statistic does not reduce to the usual variance of the Wilcoxon test statistic in the absence of clustering. An alternative approach to developing a rank-sum test that would reduce to the usual Wilcoxon test statistic in the absence of clustering would be to calculate $\mathrm{var}(S)$ using the subtraction estimator

$$\mathrm{var}(S) = \mathrm{var}(W^*) - E\{\mathrm{var}(W^*|\mathbf{X}, \mathbf{g})\}. \tag{10}$$

Because, given the $\mathbf{g}^*$, W^* is the standard Wilcoxon based on iid data $\mathbf{X}^*$, the classical variance formula of rank-sum test can be used to compute the first term on the right side of (10),

$$\begin{aligned}
&\mathrm{var}(W^*)\\
&= \frac{M}{12(M+1)} E\left(\sum_{i=1}^{M}\left(g_i^* - \overline{g^*}\right)^2\right) + \mathrm{var}\left(\frac{1}{2}\sum_{i=1}^{M} g_i^*\right)\\
&= \frac{M}{12(M+1)}\left[\sum_{i=1}^{M}\frac{n_{i1}}{n_i}\right.\\
&\qquad\left. - \frac{1}{M}\left\{\sum_{i=1}^{M}\frac{n_{i1}}{n_i}\left(1-\frac{n_{i1}}{n_i}\right) + \left(\sum_{i=1}^{M}\frac{n_{i1}}{n_i}\right)^2\right\}\right]\\
&\quad + \frac{1}{4}\sum_{i=1}^{M}\frac{n_{i1}}{n_i}\left(1-\frac{n_{i1}}{n_i}\right),
\end{aligned}$$

assuming no ties. It is easy to see that if there were no clustering (i.e., $n_i \equiv 1$), then this would reduce to the usual variance of Wilcoxon for iid data, and would also equal $\mathrm{var}(S)$, because in this case the second term on the right side of (10) would be 0. In general, the term $E\{\mathrm{var}(W^*|\mathbf{X}, \mathbf{g})\}$ needs to be estimated via projection techniques as before. For example, an estimator of this (without correcting for ties) is given by

$$\frac{1}{M^2}\sum_{i=1}^{M}\left\{\left(\frac{1}{n_i}\sum_{k=1}^{n_i} V_{ik}^2\right) - \left(\frac{1}{n_i}\sum_{k=1}^{n_i} V_{ik}\right)^2\right\},$$

where

$$V_{ik} = g_{ik} + \sum_{j\neq i}\sum_{h=1}^{n_j}\frac{g_{ik} - g_{jh}}{n_j} I(X_{jh} \le X_{ik}).$$

There is, however, a practical difficulty in using this approach. There is no guarantee that the resulting estimator of $\mathrm{var}(S)$ obtained by the subtraction formula (10) will be positive. In fact, using simulated data, we have seen that in some settings (10) is negative in a nonnegligible proportion of samples. This typically occurs when each term on the right side of (10) is large compared to their difference. Further, even for situations when (10) results in a positive variance estimator, we found that the resulting standardized test had worse performance than the test using (7). For these reasons, we did not pursue this approach.

The simulation results the we report in Section 2.4 show that our test can be conservative when the number of clusters is small. In this case it would be desirable to calculate exact p values using a Monte Carlo scheme. When all members of a cluster belong to the same group, this is easily accomplished. A permutation test can be constructed by repeatedly randomly permuting the group memberships of each cluster and then calculating our test statistic. The empirical quantiles of the resulting distribution can be used for hypothesis testing. However, the situation is not so straightforward when group membership can vary within clusters. In this case it appears necessary to make assumptions about the nature of the within-group correlation before a permutation or bootstrap scheme can be suggested. One possibility is to permute the group membership indicators without regard to cluster, which is appropriate if within-group correlations are independent of group membership. An alternative is to stochastically reassign group membership simultaneously to all members of the same group within each cluster. For example, with two groups with independent probability $1/2$ for each cluster, we would reassign all group 0 members to group 1 and all group 1 members to group 0. This scheme is appropriate when the correlation structure within a group is determined by group membership. However, the sensitivity of these permutation schemes to misspecification of the correlation structure is unknown.

The basic idea of calculating the expected value of the rank-sum statistic conditional on sampling one observation from each cluster may be applicable to other tests as well, including linear-rank and signed-rank statistics. We plan to pursue a general theory in subsequent publications. However, preliminary calculation suggests that for linear-rank and signed-rank tests this approach leads to somewhat restrictive conditions on the score-generating function, and that many commonly used scores (e.g., Savage, normal) must be modified to satisfy the conditions required to establish asymptotic normality of the test statistic.

APPENDIX: PROOFS OF ASYMPTOTIC DISTRIBUTION THEOREMS

Consider the probability model conditional on n_{ij}, $i \ge 1$, $j = 0, 1$. Let $U = (M+1)S/\binom{M}{2}$. Then U is very similar to a U-statistic based on independent (but not identically distributed) random vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_M$, with $\mathbf{Y}_i = (X_{i1}, g_{i1}, \ldots, X_{in_i}, g_{in_i})$, $1 \le i \le M$, and a kernel $h = h_{ij}$ of order 2 that depends on M and the n_{ij}. Because the random vectors involved are of different lengths and not identically distributed and the kernel function depends on the sample size, standard theory for U-statistics does not immediately apply. However, we show here that the asymptotic normality of S can be established along similar lines as the standard U-statistic theory.

Proof of Theorem 1

Through careful examination of the reminder term in the first-order Hoeffding decomposition, we will show that

$$U - E(U) = \frac{(M+1)}{\binom{M}{2}}\sum_{i=1}^{M}(W_i - E(W_i)) + o_p\left(\frac{1}{\sqrt{M}}\right), \tag{A.1}$$

as $M \to \infty$. Let

$$r_M = U - E(U) - \frac{(M+1)}{\binom{M}{2}}\sum_{i=1}^{M}(W_i - E(W_i)).$$

It follows from the a slight modification of the representation in equation 5.3.2 of Serfling (1980) (which can be seen to hold for independent but not identically distributed random vectors) that r_M is

in turn a U-statistic based on a centered second-order kernel $H = H_{ij}$ defined by

$$H(\mathbf{y}_i, \mathbf{y}_j) = h(\mathbf{y}_i, \mathbf{y}_j) - Eh(\mathbf{y}_i, \mathbf{Y}_j) - Eh(\mathbf{Y}_i, \mathbf{y}_j) + Eh(\mathbf{Y}_i, \mathbf{Y}_j).$$

Moreover, because $|h(\mathbf{Y}_i, \mathbf{Y}_j)|$ is bounded (uniformly in i, j, and M), so is $|H(\mathbf{Y}_i, \mathbf{Y}_j)|$ by its relationship with h. Therefore, by direct calculation of the second moment of r_M along the line of lemma 5.2.2B of Serfling (1980), it follows that $E(|r_M|^2) = O(M^{-2})$, implying that $r_M = o_P(M^{-1/2})$.

Note that the W_i's are independent and bounded random variables. Therefore, by the Lindeberg central limit theorem for triangular arrays (Billingsley 1986, thm. 27.2),

$$\frac{\sum_{i=1}^M (W_i - E(W_i))}{\sqrt{\sum_{i=1}^M \operatorname{var}(W_i)}} \xrightarrow{d} \mathrm{N}(0, 1), \quad \text{as } M \to \infty, \tag{A.2}$$

provided that

$$\sum_{i=1}^M \operatorname{var}(W_i) \to \infty, \quad \text{as } M \to \infty. \tag{A.3}$$

Clearly, (A.3) follows from condition 2. Moreover, by the law of large numbers for independent and uniformly integrable summands (see, e.g., Fabian and Hannan 1985),

$$M^{-1} \sum_{i=1}^M (W_i - E(W_i))^2 - M^{-1} \sum_{i=1}^M \operatorname{var}(W_i) \xrightarrow{p} 0, \quad \text{as } M \to \infty. \tag{A.4}$$

Next, note that for each x, $\widehat{F}(x)$ is a weighted average of independent summands $F_i(x)$, each of which has expectation $F(x)$, it converges in probability to $F(x)$ under condition 1. Moreover, the convergence holds in sup norm as well, by the usual arguments as in the proof of the Glivenko–Cantelli theorem. Therefore, from (A.4) we get

$$M^{-1} \widehat{\operatorname{var}}(S) - M^{-1} \sum_{i=1}^M \operatorname{var}(W_i) \xrightarrow{p} 0, \quad \text{as } M \to \infty,$$

which further implies that

$$\frac{\sum_{i=1}^M \operatorname{var}(W_i)}{\widehat{\operatorname{var}}(S)} \xrightarrow{p} 1, \quad \text{as } M \to \infty, \tag{A.5}$$

if Condition 2 holds.

Finally, using (A.2) and (A.5) with the representation (A.1), we conclude the proof of the theorem.

Proof of (8)

Assume continuous distributions for simplicity. Then $\int F^2(x)\, dF(x) = 1/3$. Note that $W_i = n_i^{-1} \sum_{k=1}^{n_i} W_{ik}$, say, where

$$EW_{ik} \approx \frac{1}{2}(\alpha_i - \overline{\alpha}),$$

$$EW_{ik}^2 \approx \frac{1}{3}(\alpha_i - 2\alpha_i\overline{\alpha} + \overline{\alpha}^2) \geq \frac{1}{3}(\alpha_i - \overline{\alpha})^2.$$

Therefore, when the W_{ik}'s are independent or positive dependent,

$$\operatorname{var}(W_i) \geq \left(\frac{1}{12} - \epsilon\right)(\alpha_i - \overline{\alpha})^2 n_i^{-1}$$

for any $\epsilon > 0$ and all large enough M. Let $\epsilon = 1/12 - 1/20$ to complete the proof.

Proof of Theorem 2

Consider the spectral decomposition

$$\mathbf{V} := M^{-1} \sum_{i=1}^M E\{\mathbf{W}_i - E(\mathbf{W}_i)\}\{\mathbf{W}_i - E(\mathbf{W}_i)\}^T = \sum_{j=1}^m \lambda_{(j)}^{(M)} \mathbf{P}_j^{(M)} \mathbf{P}_j^{(M)T},$$

where $\lambda_{(1)}^{(M)} \geq \cdots \geq \lambda_{(m-1)}^{(M)} \geq \lambda_{(m)}^{(M)} = 0$ are the eigenvalues and the $\mathbf{P}_j^{(M)}$'s are the corresponding orthonormal eigenvectors of $\mathbf{V}$. Using the Cramer–Wold device and the Lindeberg central limit theorem, we obtain

$$(Z_1, \ldots, Z_{m-1})^T \xrightarrow{d} \mathrm{N}_{m-1}(\mathbf{0}, \mathbf{I}_{m-1}), \tag{A.6}$$

where

$$Z_j = \left\{\lambda_{(j)}^{(M)} M\right\}^{-1/2} \{\mathbf{S} - E(\mathbf{S})\}^T P_j^{(M)}, \qquad 1 \leq j \leq m-1. \tag{A.7}$$

Next, using the law of large numbers and the uniform consistency of $\widehat{F}$, we obtain, as before,

$$\left\| M^{-1} \sum_{i=1}^M \{\widehat{\mathbf{W}}_i - E(\mathbf{W}_i)\}\{\widehat{\mathbf{W}}_i - E(\mathbf{W}_i)\}^T - M^{-1} \sum_{i=1}^M E\{\mathbf{W}_i - E(\mathbf{W}_i)\}\{\mathbf{W}_i - E(\mathbf{W}_i)\}^T \right\| \xrightarrow{P} 0,$$

where $\|\cdot\|$ denotes any norm on $\Re^{m^2}$. Therefore, using Condition 3 and (A.6), we obtain

$$(\widehat{Z}_1, \ldots, \widehat{Z}_{m-1})^T \xrightarrow{d} \mathrm{N}_{m-1}(\mathbf{0}, \mathbf{I}_{m-1}), \tag{A.8}$$

where the $\widehat{Z}_j$'s are given by (A.7) with $\lambda_{(j)}^{(M)}$ and $\mathbf{P}_j^{(M)}$ replaced by $\widehat{\lambda}_{(j)}^{(M)}$ and $\widehat{\mathbf{P}}_j^{(M)}$. We complete the proof by noting that $T = \sum_{j=1}^{m-1} \widehat{Z}_j^2$.

[Received September 2003. Revised August 2004.]

REFERENCES

Abecasis, G. R., Cookson, W. O. C., and Cardon, L. R. (2000), "Pedigree Tests of Transmission Disequilibrium," *European Journal of Human Genetics*, 8, 545–551.

Billingsley, P. (1986), *Probability and Measure* (2nd ed.), New York: Wiley.

Fabian, V., and Hannan, J. (1985), *Introduction to Probability and Mathematical Statistics*, New York: Wiley.

Hoffman, E. B., Sen, P. K., and Weinberg, C. R. (2001), "Within-Cluster Resampling," *Biometrika*, 88, 1121–1134.

Hudgens, M. G., and Satten, G. A. (2002), "Midrank Unification of Rank Tests for Exact, Tied and Censored Data," *Journal of Nonparametric Statistics*, 14, 569–581.

Keavney, B., McKenzie, C. A., Connell, J. M., Julier, C., Ratcliffe, P. J., Sobel, E., Lathrop, M., and Farrell, M. (1998), "Measured Haplotype Analysis of the Angiotensin-I Converting Enzyme Gene," *Human Molecular Genetics*, 7, 1745–1763.

Rieger, R. H., Kaplan, N. L., and Weinberg, C. R. (2001), "Efficient Use of Siblings in Testing for Linkage and Association," *Genetic Epidemiology*, 20, 175–191.

Rosner, B., Glynn, R. J., and Lee, M.-L. T. (2003), "Incorporation of Clustering Effects for the Wilcoxon Rank-Sum Test: A Large-Sample Approach," *Biometrics*, 59, 1089–1098.

Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, New York: Wiley.

Williamson, J. M., Datta, S., and Satten, G. A. (2003), "Marginal Analyses of Clustered Data When Cluster Size Is Informative," *Biometrics*, 59, 36–42.

Xiong, M. M., Krushkal, J., and Boerwinkle, E. (1998), "TDT Statistics for Mapping Quantitative Trait Loci," *Annals of Human Genetics*, 62, 431–452.