Two-sample rank tests under complex sampling

Thomas Lumley and Alastair Scott

Department of Statistics,

University of Auckland,

Auckland, NZ

t.lumley@auckland.ac.nz

June 20, 2012

Author's Footnote:

Thomas Lumley is Professor of Biostatistics and Alastair Scott is Professor of Statistics, Department of Statistics, University of Auckland. This research was funded in part by a grant from the Marsden Fund to the University of Auckland. We thank J.N.K. Rao for helpful advice on known weak-convergence results in survey sampling, and Jianqing Wang for access to a preprint of Wang (2012).

Abstract

Rank tests are widely used for exploratory and formal inference in the health and social sciences. With the increasing use of data from complex survey samples in medical research, there is increasing demand for versions of rank tests that account for the sampling design. In the absence of design-based rank tests, naive unweighted rank tests are being used in survey analyses even by researchers who otherwise use inferential methods appropriate for the sampling design. We propose a general approach to constructing design-based rank tests when comparing groups within a complex sample and when using a national survey as a reference distribution, and illustrate both scenarios with examples. We show that the tests have asymptotically correct level and that the relative power of different rank tests is not greatly affected by complex sampling.

Keywords:multistage sampling; cluster sampling; health surveys; weak convergence; Wilcoxon test

1. INTRODUCTION

Despite limitations such as non-transitivity (Brown and Hettmansperger 2002) and difficulty of interpretation, rank-based tests are widely used by researchers in the social and health sciences. Data from complex multistage survey designs are increasingly important in these areas, with more and more large studies publishing public-use data. The extension of rank tests to data from complex samples would be valuable to researchers who wish to do the same analyses with data from, say, NHANES (National Center for Health Statistics 1981) or the British Household Panel Survey (Taylor 2010) as they would do with data from a cohort or cross-sectional sample. In this paper we give a general and computationally simple approach to design-based rank tests with complex sampling. The term 'design-based' here implies two criteria to be satisfied, at least asymptotically: that the test has the specified level when the population null hypothesis is true, and that it

tests the same population null hypothesis regardless of the sampling scheme.

In the absence of correct design-based methods, rank tests, especially the Wilcoxon rank-sum test, are currently being used on data from complex surveys by simply ignoring the sampling scheme, even in papers that correctly use sampling weights in other aspects of the analysis such as fitting regression models or estimating summary statistics. For example, Knovich et al. (2008) examined the relationship between serum copper and anaemia in the NHANES II sample, using design-based logistic regression for their primary analyses, but unweighted Wilcoxon rank-sum tests for comparing serum copper between non-anaemic and anaemic groups. Lamprecht et al. (2011) used the Wilcoxon rank-sum test to compare continuous variables between groups in the Burden of Obstructive Lung Disease study, a multi-country complex survey that typically does account for the sampling in their analyses (Buist et al. 2005). Hailpern et al. (2007) examined the association between kidney disease and cognitive function, using data from NHANES III. They accounted for the NHANES III sampling scheme when fitting logistic regression models and estimating proportions, but used unweighted Wilcoxon rank-sum tests for unadjusted comparisons between groups.

Similar issues arise with environmental health research using the extensive NHANES blood analysis dataset as a control sample. Herrick et al. (2011) compared serum PCB levels in teachers from PCB-containing schools to the NHANES 2003–4 sample, using the Wilcoxon rank-sum test. Feng et al. (2011) compared dioxin and furan levels in people living near wood-treatment plants to the NHANES 2003–4 sample, also using the Wilcoxon rank-sum test. Castorina et al. (2010) compared pesticide, herbicide, and fungicide residues in blood samples from pregnant women in the Salinas Valley to the NHANES 1999–2002 sample using the Wilcoxon rank-sum test and quantile regression.

In all these examples a simple design-based version of Wilcoxon rank-sum or quantile test would have been preferable, and would likely have been used if it had been readily available. There has been little previous literature directly on this topic. Rosner et al. (2003) described an extension of the Wilcoxon test to clustered data, where the clusters were assumed to be nested within the groups being compared. Rosner et al. (2006) removed this assumption to allow clusters to have members from both groups. Both papers used a permutation approach, obtaining a sampling distribution by randomization of the group assignments. Jung and Jeong (2003) described weighted logrank tests for censored clustered data, again assuming that each cluster belongs to a single group. They used a sandwich-type variance estimator and theory developed for the Cox proportional hazard model. Datta and Satten (2005) also addressed clustered data, using an approach based on thinning the data to a single observation per cluster and then averaging to restore the information thus lost. None of these tests is explicitly design-based, and none of them allows unequal sampling probabilities.

Recently, Natarajan et al. (2010) described an extension of the Wilcoxon rank-sum test to complex samples, based on fitting a proportional odds regression model to the data, and using the score test, which is known to be asymptotically equivalent to the Wilcoxon test under iid sampling. Their proposal was limited to ordinal categorical data: neither the theoretical nor the computational approach generalize immediately to continuous data, where the underlying proportional odds model would have as many parameters as observations. The approach also does not generalize to other rank tests; the score tests in other cumulative link models for ordinal data do not reduce to rank tests in the same way.

Complex sampling degrades some of the attractive theoretical properties of the rank tests. The lack of exchangeability under complex sampling means that rank tests are not distribution-free in small samples, and the need to reweight observations to the population

complicates the elegant optimality properties of rank tests under location-shift alternatives. These limitations should not affect most uses of rank tests in complex samples, where the sample size is usually large enough for inference based on the central limit theorem, and where it is unusual to know *a priori* that two groups differ only by a location shift.

In Section 2 we introduce the test statistics and describe the estimation approach and the computations needed, both for comparing groups within a complex survey and for using a complex survey as a reference distribution. In Section 3 we use simulation to study the level and power of design-based versions of some popular rank tests. In Section 4 we illustrate the impact of the design-based tests on two examples from the NHANES series. In the Discussion we point out some of the ways in which a design-based rank test has different aims from the clustered rank tests described by previous authors.

2. CONSTRUCTION OF THE RANK TEST

2.1 Comparing groups within a survey

In this section we describe the test statistic and give an outline of the underlying theoretical justification. Details of the asymptotics are given in the Appendix.

We are testing the null hypothesis that a real-valued random variable Y is independent of a grouping variable G (initially taking just two values, 0 or 1) against the alternative that Y is stochastically ordered by G. Suppose that we have data from a sample of n units, drawn from a finite population of N units in which the ith population unit has values (Y_i, G_i). We shall assume that the finite population values, $\{(Y_i, G_i);\ i = 1, \ldots, N\}$, are generated independently from some joint distribution with marginal distribution function F_Y. (Note that the asymptotic properties of the test below depend very weakly on the assumption that the finite population values are independent if N is large — see Graubard and Korn

(2002). All we really need is that the finite-population rank statistics defined in equation 1 has expected value zero when Y is independent of G.)

First consider the finite-population quantity that will be estimated by the sample rank test. Let $\mathbb{F}_N(y) = \frac{1}{N}\sum_{i=1}^{N} I(Y_i \le y)$ denote the empirical finite-population distribution function of Y. The scaled finite population mid-ranks, $R_1, R_2, \ldots, R_N$, are defined by setting

$$R_i = \frac{1}{N}\sum_{j=1}^{N}\left[I(Y_j < Y_i) + 0.5 \times I(Y_j = Y_i)\right] = \frac{1}{2}\left[\mathbb{F}_N(Y_i) + \mathbb{F}_N(Y_i-)\right].$$

The use of mid-ranks allows the test to be defined for discrete as well as continuous Y(Conover 1973; Hudgens and Satten 2002). We define the finite-population rank test statistic, T_N, as the difference in the mean of $g(R_i)$ between the groups for a suitable function $g()$. For example, the Wilcoxon test uses $g(R_i) = R_i$, the normal-scores test uses $g(R_i) = \Phi^{-1}(R_i)$, and Mood's test for the median uses $g(R_i) = I(R_i > 1/2)$. Thus

$$T_N = \frac{1}{M_0}\sum_{\{i:G_i=0\}} g(R_i) - \frac{1}{M_1}\sum_{\{i:G_i=1\}} g(R_i), \tag{1}$$

where $M_\ell = \sum_1^N I(G_i = \ell)$ is the number of finite population units in group ℓ $(\ell = 0, 1)$.

We draw a sample, s, of n units from the finite population using some probability sampling method with sampling probabilities π_i, and corresponding sampling weights $w_i = 1/\pi_i$, and we observe the values of Y_i and G_i for the sampled units. In large surveys the sampling probabilities and weights will often include adjustments for non-response, frame errors, and other imperfections; in two-phase designs the probabilities and weights may not be entirely pre-specified. The estimated population mid-ranks $\widehat{R}_i$ are defined as

$$\widehat{R}_i = \frac{1}{\widehat{N}}\sum_{j\in s}\left[w_j I(Y_j < Y_i) + 0.5 w_j I(Y_j = Y_i)\right] = \frac{1}{2}\left[\widehat{F}_n(Y_i) + \widehat{F}_n(Y_i-)\right].$$

where $\widehat{N} = \sum_{j \in s} w_j$ is the estimated population size and

$$\widehat{F}_n(y) = \frac{1}{\widehat{N}} \sum_{j \in s} w_j I(Y_j \leq y)$$

is the Hájek estimator of $\mathbb{F}_N(y)$ (and hence a consistent estimator of the finite-population and super-population distribution functions, $\mathbb{F}_N(y)$ and $F_Y(y)$ respectively). Note that these are not the same as the sample ranks unless we have a self-weighting design. We can now define $\widehat{T}_n$, the sample version of the rank test statistic, as the estimator of the finite-population quantity T_N in equation 1:

$$\widehat{T}_n = \frac{1}{\widehat{M}_0} \sum_{i \in s_0} w_i g(\widehat{R}_i) - \frac{1}{\widehat{M}_1} \sum_{i \in s_1} w_i g(\widehat{R}_i), \tag{2}$$

where, for $\ell = 0$ or 1, $s_\ell = \{i \in s : G_i = \ell\}$ and $\widehat{M}_\ell = \sum_{i \in s_\ell} w_i$ is the Horvitz-Thompson estimator of M_ℓ.

If $\widehat{R}_i$ were a fixed quantity associated with the ith unit in the realised finite population, then $\widehat{T}_n$ would be the difference between two estimated domain means and inference based on it would be straightforward. Inference is complicated by the fact that the value of $\widehat{R}_i$ is not fixed for the finite population but depends on the values of the other units drawn in the sample as well as on the sampling design. If we write $U_i = \frac{1}{2}[F_Y(Y_i) + F_Y(Y_i-)]$, replacing the estimate $\widehat{F}_n$ in the definition of $\widehat{R}_i$ with the superpopulation quantity F_Y, then the U_is do not depend on the sample values or the sampling design in any way. (They are also independent and identically distributed in the superpopulation under the assumptions above, although we will not make use of this in our derivations.) The classical proof that U_i can be substituted for R_i with no effect on the asymptotic distribution of the full finite population statistic T_N relies heavily on exchangeability(Hájek and Šidák 1967, Chapter 5) and does not carry over to $\widehat{R}_i$ under complex sampling. However, an alternative approach using the functional delta method and the weak convergence of $\sqrt{N}\,(\mathbb{F}_N(y) - F_Y(y))$ to a

Brownian Bridge (Pyke and Shorack 1968; van der Vaart and Wellner 1996) can be adapted for complex samples. We adopt this approach here and show in the appendix that, under suitable conditions, $\widehat{T}_n$ defined in 2 has the same asymptotic null distribution as

$$\widetilde{T}_n = \frac{1}{\widehat{M}_0} \sum_{i \in s_0} w_i g(U_i) - \frac{1}{\widehat{M}_1} \sum_{i \in s_1} w_i g(U_i).$$

More precisely, under conditions described in the Appendix, $\sqrt{n}\left(\widehat{T}_n - \widetilde{T}_n\right) \xrightarrow{p} 0$ when the null hypothesis that Y is independent of the grouping variable G in the superpopulation is true.

To make use of this equivalence, we need an estimate of $\widetilde{\sigma}^2$, the asymptotic variance of $\sqrt{n}\,\widetilde{T}_n$. If we knew the values of U_i for the sampled units, $\widetilde{T}_n$ would be the difference between two estimated domain means and all computer packages with procedures for analyzing survey data have software for producing such a variance estimate, $\widetilde{V}(\mathbf{U_s})$ say. We show in the appendix that $\widetilde{V}(\widehat{\mathbf{R}}_\mathbf{s})$, the value we get by replacing U_i with $\widehat{R}_i$ for $i \in s$ is also a consistent estimator of $\widetilde{\sigma}^2$ and hence, under H_0, of $var\{\widehat{T}_n\}$. This allows standard survey programs to be used. In particular, a design-based rank test can be performed as a weighted t-test on the transformed estimated ranks $g(\widehat{R}_i)$. Computation of $\widehat{R}_i$ is straightforward and weighted t-tests under complex sampling are now available in most general-purpose statistical software. We have provided an implementation for R, the `svyranktest()` function in the survey package(Lumley 2011). This implementation uses a t reference distribution rather than the asymptotic Normal distribution, with degrees of freedom defined as $C - H$, where C is the number of primary sampling units and H is the number of strata, an adjustment that is widely used for inference in survey statistics. (Note that this can make a substantial difference even in big surveys since the degrees of freedom may be small even when the sample size is large. For example, the public-use data sets from the current continuous NHANES surveys have 14 strata for each two-year period, with two primary

sampling units per stratum, giving only 14 degrees of freedom per two years in a sample size of approximately 10,000.)

The assumptions necessary to establish the equivalence of $\widehat{T}_n$ and $\widetilde{T}_n$ amount to a strong form of the central limit theorem for the sequence of sampling designs. Most importantly, we assume that the sequence of designs is such that $\sqrt{n}\begin{pmatrix} \widehat{F}_{0n} - F_{0Y} \\ \widehat{F}_{1n} - F_{1Y} \end{pmatrix}$, where $F_{\ell Y}(y)$ denotes the conditional distribution function of Y given $G = \ell$ in the superpopulation and $\widehat{F}_{\ell n}(y) = \frac{1}{\widehat{M}_\ell}\sum_{i \in s_\ell} w_i I(Y_i \leq y)$ its sample estimate ($\ell = 0, 1$), converges weakly to a bivariate Gaussian process as n, $N \to \infty$. This last assumption is true under mild conditions for simple random sampling and stratified random sampling with a fixed number of strata. It has been established for a reasonably broad class of single-stage designs in Breslow and Wellner (2007), Cardot and Josserand (2011), and Wang (2012). These results have been extended to multi-stage designs, under the assumption that design effects of estimated differences of step functions are bounded, in Lumley (2012). The assumption of bounded design effects, while almost certainly stronger than necessary, is an intuitively natural one in practice since sampling designs are typically chosen to minimize the design effects of important variables subject to a fixed cost. A very large design effect would indicate that the design is not an appropriate one for the variables under study.

For some rank tests it is more common simply to define the test statistic as the sum of $g(R_i)$ over one of the groups. For example, the Wilcoxon test statistic is often defined as the sum of the ranks over the smaller group. This definition is a linear function of the definition in equation 1, with coefficients that depend only on the group sizes M_0 and M_1. Under simple random sampling or independent sampling from a superpopulation, if the null hypothesis is true, $\widehat{M}_0$ and $\widehat{M}_1$ are independent of the ranks and it is standard to condition on them, so the two definitions give the same test. Under complex sampling

the group sizes are typically not independent of the ranks unless the groups were used to stratify the sampling, so the two definitions of the rank test are not equivalent. We have chosen equation 1 as estimating a more readily interpretable population quantity; this form was also the one used by Hájek and Šidák (1967).

A test based on the sample ranks, as distinct from the estimated population ranks $\widehat{R}_i$, would also be possible, but would not be design-based, in the sense that the mapping of values of Y to sample ranks would depend on the sampling scheme, and so the alternatives against which the test was consistent would also depend on the sampling scheme. For example, it would be possible for the rank test statistic in the population to be positive, but for the sample rank-test statistic to be negative with arbitrarily high probability.

2.2 Comparing a targeted sample to a population survey

The NHANES series of surveys includes a wide range of assays performed on blood samples, giving population distributions for nutrients, environmental pollutants, disease biomarkers, and other variables. Researchers often wish to compare the distribution of blood measurements in a targeted sample (a case series or a cohort) to national reference values from NHANES, and as we illustrated in the Introduction, often use rank tests for these comparisons.

When the targeted sample is a well-defined probability sample from a population that is not the NHANES population, the targeted sample and NHANES can be treated as two strata of a single stratified sample from the combined population. Examples would include comparisons across time and comparisons between countries. The targeted sample could also be a well-defined probability sample from a subset of the US, eg comparing data from a state survey with national data. In this scenario we can treat the data as a dual-frame survey(Lohr and Rao 2000; Metcalf and Scott 2009). That is, NHANES is a

sample from the US population, and the targeted sample is a sample from some subset of the US population. Metcalf and Scott (2009) described a large class of estimators for dual-frame surveys that use the original design weights for the non-overlapping subsets of the two surveys and rescale the weights to prevent double-counting for population in the overlap of the two sampling frames. For example, if data from a California survey were being compared to NHANES it would be necessary to decide how to apportion the weight for California between the Californian survey and the Californian subset of NHANES. A simple and reasonably efficient choice is to apportion the weight in proportion to the sample size the two surveys have for California. The two surveys would then be treated as two strata in a combined data set with the adjusted weights.

More commonly, as in the references we cited above, a reference distribution from NHANES is being compared to a small targeted sample that was not drawn according to any probability mechanism. In this situation we can still model the data as coming from a dual-frame survey, but one in which the sampling frame for the targeted sample is just the sample itself. Since the overlap is a negligible fraction of the national sampling frame for NHANES, we propose to use the NHANES sampling weights without modification. We use weights $w_i = 1$ for the targeted sample, reflecting the fact that they are members of the national sampling frame but need not be sampled in a way that makes them representative of any larger subset of the population. Again, the two samples are then treated as strata in a combined data set.

3. SIMULATIONS

3.1 Large sample with few clusters

NHANES, like a number of other large public studies, has a sampling design with a moderate number of strata and only two, large, clusters per stratum. The usual approximation to degrees of freedom for estimating standard errors is $C - H$, where C is number of PSUs in the sample and H is the number of strata. As we have noted above, his can be quite small even in very large surveys. For example, NHANES II had $n = 20,322$ participants completing clinical exams but only 32 degrees of freedom. For this reason, it is important to evaluate the performance of the rank tests in situations where the sample size is large but the number of PSUs is relatively small.

[Figure 1 about here.]

We simulated a population of size 100,000 with $Y \sim N(0,1)$ and $G \sim Bernoulli(1/3)$. Since the rank tests are invariant to monotone transformations there is no loss of generality in using a Normal distribution for Y. Strata were defined by quantiles of $YG + \epsilon$, where $\epsilon \sim N(0, \sigma^2)$, with the first 5 strata containing 10% each of the population, the next 9 strata containing 5%, four containing 2% and 2 containing 1%. Clusters of size 100 were defined by quantiles of $Y + \eta$, with $\eta \sim N(0, \tau^2)$, and two clusters per stratum were sampled. Each simulation scenario was replicated 10,000 times. Figure 1 shows quantile-quantile plots of $-\log_{10}(p\text{-values})$ for the Wilcoxon test, the median test, the normal-scores test,, and the t-test, for simulations with $\sigma = \tau = 5$. An unweighted t-test or Wilcoxon test has a median z-statistic of approximately 3.5 under this sampling design, so the sampling is strongly biased.

As Figure 1 shows, using a t reference distribution with the design degrees of freedom gives a test with close to nominal size; in fact 6.0–6.5% of p-values were below 0.05. Using a

Normal reference distribution is moderately anticonservative, with about 8% of p-values below 0.05. Increasing the number of clusters to three per stratum, giving 40 design degrees of freedom, improves performance, with 5.5–6.0% of p-values below 0.05 using the t_{40} reference distribution and 6.0–6.5% using the Normal.

Repeating the simulation with $\tau = \sigma = 1$, which gives much stronger sampling bias and a median unweighted z-statistic of about 19, the test using a t-reference distributions were conservative for α near 0.1, but anti-conservative for very small α; the rank tests still had similar performance to the t-test.

These simulations confirm that the rank tests, while not exact, have acceptable control of Type I error even under strongly biased sampling with a small number of clusters, and that the performance is improved by using a t distribution rather than a Normal distribution to compute p-values.

3.2 Tail sampling and power

Under independent sampling, rank tests are locally optimal for certain location-shift alternatives. For example, the Wilcoxon test is locally optimal for location shift in a logistic distribution, the median test for location shift in a double exponential (Laplace) distribution, and the normal-scores test for location shift in a Normal distribution. The situation is more complicated under complex sampling, because the sampling affects the distributions of the variables, and because the estimated group sizes $\widehat{M_1}$ and $\widehat{M_2}$ are typically not independent of the ranks unless sampling is stratified on group. We carried out a simulation to investigate the impact on power of random group size and of oversampling the tails of the outcome distribution.

Group membership for a population of 10,000 was simulated as Bernoulli(0.3). Variables

X, Y, and Z were simulated from independent Normal, double exponential, and logistic distributions with unit variance and with mean $G/5$. Sampling probabilities were either equal for all units or oversampled in the tails, with $\pi_i \propto |X_i + Y_i + Z_i| + 2$.

The sample size was 500, either sampled in a single stratum or with sampling stratified on G to give 150 units with $G = 1$ and 350 with $G = 0$. Table 1 shows the efficiencies of the Wilcoxon rank-sum test, Mood's test for the median, and the normal-scores test, relative to a design-based t-test. The relative efficiency was estimated by the ratio of squared z-statistics, based on 10,000 replications.

The results in Table 1 show that the relative efficiencies of the tests are affected both by the random group size when sampling is not stratified on group, and by the oversampling of tails. Both of these factors tended to attenuate the difference in efficiency between tests, though the pattern of differences remained consistent. The equivalence of the normal-scores test and t-test for Normal data persisted in all four scenarios and the patterns of higher and lower efficiency tended to remain the same. These results suggest that criteria for choosing a rank test under complex sampling should be similar to the criteria used in standard situations when data are sampled independently.

[Table 1 about here.]

4. EXAMPLES

4.1 Serum copper and anemia in NHANES II

To show the potential impact of the sampling design on inference we repeat an analysis from Knovich et al. (2008). The authors conjectured that copper deficiency would explain some cases of anemia, and compared serum copper concentrations in people with and without anemia using data from NHANES II. They reported unweighted median serum copper

concentrations in anemic and non-anemic subjects as 1260 and 1190 μg/dl, respectively, and described this as statistically significant using an unweighted Wilcoxon test. We compute the unweighted Wilcoxon p-value to be 1.3×10^{-7}. The weighted estimates of the population median are 1200 and 1160μg/dl, noticeably lower than the unweighted estimates, and the design-based Wilcoxon p-value is 0.011. The Wilcoxon test still reports a statistically significant difference, but the p-value is much larger and a test for difference in medians gives a p-value of only 0.079.

Three factors are responsible for the inflated significance of the unweighted Wilcoxon test. First, ignoring the weights gives a slightly larger difference between the distributions of serum copper. Second, ignoring the clustering in NHANES II overstates the precision of the comparison. Finally, the NHANES design with 64 sampling units and 32 strata has low design degrees of freedom, so a t reference distribution is more appropriate than the Normal distribution used in the Wilcoxon test.

The main results of Knovich et al. (2008) do not come from the Wilcoxon test, but from a logistic regression model that did account for the clustering in the NHANES II design, and their conclusions of a U-shaped relationship between serum copper and anemia are still supported by the analysis.

4.2 Comparing to NHANES III

The Heart and Estrogen/Progestin Study (HERS) was a randomized trial of estrogen and progestin supplementation in post-menopausal women(Hulley et al. 1998). Some data from this trial have been made available in Vittinghoff et al. (2004), and we will use them to illustrate a comparison between a targeted sample and a national survey. We compared HDL cholesterol, systolic and diastolic blood pressure, and self-rated global health for 2763 women in HERS and the 6695 women over age 50 among the 18162 NHANES III

participants with examination and lab data, who represent a subpopulation of 29 million women. HDL cholesterol and blood pressure are continuous measurements, but self-rated health is a five-level discrete variable with levels from "Excellent" to "Poor".

As HERS was a trial in women who had existing coronary disease we would expect the participants to have worse health than the population as a whole. HERS participants had lower HDL cholesterol, lower blood pressure, and worse self-rated global health than the NHANES women over 50. For systolic blood pressure the design-based Wilcoxon test had a t-statistic of -2.5 and p-value of 0.013, but ignoring the design gave a z-statistic of -9.4 and p-value of 10^{-21}. For diastolic blood pressure the design-based Wilcoxon test had a t-statistic of -11.9 and ignoring the design gave a z-statistic of -10.3. For HDL cholesterol the design-based test had $t = -10.6$ and ignoring the design gave -13.8. For self-rated health the design-based test had a t-statistic of -4.9 and ignoring the design gave a z-statistic of -5.5.

The Wilcoxon test for the discrete self-rated health variable is equivalent to a score test in a proportional-odds model, so we also computed the Wald test and the Rao-Scott likelihood ratio test(Rao and Scott 1984) under the proportional-odds model. The Wald z-statistic was -3.34. The Rao-Scott χ^2 statistic was 11.95; taking the appropriately-signed square root gives a z-statistic of -3.45. The large difference between the score test and the other two tests in this example appears to be because the proportional-odds assumption is far from true for these data.

The impact of the sampling design on the Wilcoxon test is sometimes large and is not consistent from variable to variable, so ignoring the design makes results difficult to interpret.

5. DISCUSSION

We have shown how to construct design-based analogs of the Wilcoxon rank-sum test, Mood's test for quantiles, and other two-sample rank tests. Although they are asymptotic rather than exact tests, the actual significance levels are close to their nominal values in simulated data sets representative of population surveys. The tests are easy to implement and do not require substantially more computational resources than a design-based t-test would. The impact on p-values of taking the sampling design into account can be large, but is quite variable, so using a design-based test is preferable to *ad hoc* strategies for interpreting standard rank-test results.

Datta and Satten (2005) and Rosner et al. (2003) both discuss the issue of informative cluster size, i.e., the possibility that the distribution of Y could be systematically different in large and small clusters. This is an important problem in both cluster-randomized trials and in observational studies of repeated measurements. It is a less serious problem in design-based inference. In our design-based setting the clusters are an aspect of the sampling plan, not of the scientific question, and the goal is to estimate the same population quantity for any sampling plan. There is no difficulty in deciding how much weight should be given to large versus small clusters, or between-cluster versus within-cluster contrasts: they should be weighted so as to reproduce the population totals.

As an illustration of this issue, consider the example in Datta and Satten (2005, Section 3.3) of linkage and association between presence of a gene variant and circulating levels of a biomarker (ACE: angiotensin converting enzyme). The clusters in this example are 37 nuclear families, extracted from a sample of 69 previously sampled pedigrees. A design-based test would require sampling probabilities for each family and would estimate whether, in the population, ACE levels are higher in people who carry the variant. In this example

Datta and Satten's clustered test is a test for linkage and association, so it is immune to confounding by population structure, which could bias the design-based test. On the other hand, in a setting where clusters are merely a feature of the sampling plan, the removal of cluster-level association performed by Datta and Satten's test would be undesirable.

Some extensions of our approach are straightforward. Handling more than two groups, as in the Kruskal–Wallis test, simply involves replacing the weighted t-test by a weighted one-way ANOVA on $g(S_i)$. We have not explicitly considered non-monotone transformations of ranks, such as in the Ansari–Bradley test for scale differences, but theory and computations for these tests follow using similar arguments to those in section 2. One-sample tests such as the Wilcoxon signed-rank test should also be straightforward; we did not consider them because there appears to be little demand. A more-speculative direction for future research is extending the median test to provide tests and confidence intervals for design-based quantile regression.

REFERENCES

Breslow, N. E. and Wellner, J. A. (2007), "Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression," *Scand J Statist*, 34, 86–102.

Brown, B. M. and Hettmansperger, T. P. (2002), "Kruskal-Wallis, multiple comparisons and Efron dice." *ANZ Journal of Statistics*, 44, 427–38.

Buist, A. S., Vollmer, W. M., Sullivan, S. D., Weiss, K. B., Lee, T. A., Menezes, A. M. B., Crapo, R. O., Jensen, R. L., and Burney, P. G. J. (2005), "The Burden of Obstructive Lung Disease Initiative (BOLD): Rationale and Design," *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 2, 277–83.

Cardot, H. and Josserand, E. (2011), "Horvitz–Thompson estimators for functional data: asymptotic confidence bands and optimal allocation for stratified sampling," *Biometrika*, 98, 107–118.

Castorina, R., Bradman, A., Fenster, L., Barr, D. B., Bravo, R., Vedar, M. G., Harnly, M. E., McKone, T. E., Eisen, E. A., and Eskenazi, B. (2010), "Comparison of Current-Use Pesticide and Other Toxicant Urinary Metabolite Levels among Pregnant Women in the CHAMACOS Cohort and NHANES," *Environmental Health Perspectives*, 11.

Conover, W. J. (1973), "Rank Tests for One Sample, Two Samples, and k samples Without the Assumption of a Continuous Distribution Function," *The Annals of Statistics*, 1, pp. 1105–1125.

Datta, S. and Satten, G. (2005), "Rank-sum tests for clustered data," *Journal of the American Statistical Association*, 100, 908–15.

Feng, L., Wu, C., Tam, L., Sutherland, A., Clark, J., and Rosenfeld, P. (2011), "Dioxin furan blood lipid and attic dust concentrations in populations living near four wood treatment facilities in the United States," *Journal of Environmental Health*, 73, 34–47.

Graubard, B. I. and Korn, E. L. (2002), "Inference for Superpopulation Parameters using sample surveys," *Statistical Science*, 17, 73–96.

Hailpern, S. M., Melamed, M. L., Cohen, H. W., and Hostetter, T. H. (2007), "Moderate Chronic Kidney Disease and Cognitive Function in Adults 20 to 59 Years of Age: Third National Health and Nutrition Examination Survey (NHANES III)," *Journal of the American Society of Nephrology*, 18, 2205–13.

Hájek, J. and Šidák, Z. (1967), *Theory of Rank Tests*, Prague: Academic Press.

Herrick, R. F., Meeker, J. D., and Altshul, L. (2011), "Serum PCB levels and congener profiles among teachers in PCB-containing schools: a pilot study," *Environmental Health*, 10.

Hudgens, M. and Satten, G. (2002), "Midrank unification of rank tests for exact, tied, and censored data," *Journal of Nonparametric Statistics*, 14, 569–581.

Hulley, S., Grady, D., Furberg, C., Herrington, D., Riggs, B., and Vittinghoff, E. (1998), "Randomized trial of estrogen plus progestin for secondary prevention of coronary heart disease in postmenopausal women. Heart and Estrogen/progestin Replacement Study (HERS) Research Group." *JAMA*, 280, 605–13.

Jung, S.-H. and Jeong, J.-H. (2003), "Rank tests for clustered survival data," *Lifetime Data Analysis*, 9, 21–33.

Knovich, M. A., Il'yasova, D., Ivanova, A., and Molnár, I. (2008), "The association between serum copper and anaemia in the adult Second National Health and Nutrition Examination Survey (NHANES II) population," *British Journal of Nutrition*, 99, 1226–9.

Kosorok, M. R. (2008), *Introduction to Empirical Processes and Semiparametric Inference*, Springer.

Lamprecht, B., McBurnie, M. A., Vollmer, W. M., Gudmundsson, G., Welte, T., Nizankowska-Mogilnicka, E., Studnicka, M., Bateman, E., Anto, J. M., Burney, P., Mannino, D. M., Buist, S. A., and the BOLD Collaborative Research Group (2011), "COPD in Never Smokers : Results From the Population-Based Burden of Obstructive Lung Disease Study," *Chest*, 139, 752–63.

Lohr, S. L. and Rao, J. N. K. (2000), "Inference from dual-frame surveys," *Journal of the American Statistical Association*, 94, 271–80.

Lumley, T. (2011), *survey: analysis of complex survey samples*, R package version 3.25.

Lumley, T. (2012), "An empirical-process central limit theorem for complex sampling under bounds on the design effect," Tech. rep., Department of Statistics,University of Auckland.

Metcalf, P. A. and Scott, A. J. (2009), "Using multiple frames in health surveys." *Statistics in Medicine*, 28, 1512–1523.

Natarajan, S., Lipsitz, S. R., Sinha, D., and Fitzmaurice, G. (2010), "An Extension of the Wilcoxon Rank-Sum Test for Complex Sample Survey Data," in *Program of the 2010 Joint Statistical Meetings*, abstract # 308217.

National Center for Health Statistics (1981), *Plan and operation of the second National Health and Nutrition Examination Survey, 1976–1980.*, no. 15 in Series 1, Programs and collection procedures, National Center for Health Statistics.

Pyke, R. and Shorack, G. R. (1968), "Weak convergence of a two-sample empirical process and a new approach to Chernoff–Savage theorems," *Annals of Mathematical Statistics*, 39, 755–771.

Rao, J. N. K. and Scott, A. J. (1984), "On Chi-Squared Tests for Multiway Contingency Tables with Cell Proportions Estimated from Survey Data," *The Annals of Statistics*, 12, pp. 46–60.

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, pp. 1089–1098.

Rosner, B., Glynn, R. J., and Lee, M.-L. T. (2006), "Extension of the rank sum test for clustered data: two-group comparisons with group membership defined at the subunit level." *Biometrics*, 62, 1251–9.

Taylor, M. F. (ed.) (2010), *British Household Panel Survey User Manual*, vol. A, Colchester: University of Essex.

van der Vaart, A. W. and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, Springer.

Vittinghoff, E., Glidden, D. V., Shiboski, S. C., and McCulloch, C. E. (2004), *Regression methods in Biostatistics*, New York: Springer.

Wang, J. C. (2012), "Sample distribution function based goodness-of-fit test for complex surveys," *Computational Statistics and Data Analysis*, 56, 664–679.

APPENDIX A. ASYMPTOTICS

Some conditions on the sequence of populations and samples are necessary for the rank tests to be valid. We will use a fairly simple set of conditions; weaker assumptions are almost certainly possible.

We have a sequence of finite populations indexed by ν with values, $\{(Y_i, G_i);\ i = 1, \ldots, N_\nu\}$, generated independently from some joint distribution with marginal distribution function F_Y. A sample of size n_ν is drawn from the νth population using some well-defined probability sampling scheme. We suppose that $n_\nu, N_\nu \to \infty$ with lim sup $n_\nu/N_\nu < 1$ as $\nu \to \infty$. (To avoid the notation getting too cumbersome, we shall omit the subscript ν from here on.) We shall assume throughout that the sequence of sampling designs supports a central limit theorem for Horvitz-Thompson estimators. More specifically, assume that:

A0 If $\widehat{\mu}_{\mathrm{HT}}$ is the Horvitz-Thompson estimator of the mean, μ say, of any variable with finite fourth moment and $\widehat{\sigma}^2_{\mathrm{HT}}$ is the Horvitz-Thompson estimator of $var\left\{\sqrt{n}\widehat{\mu}_{\mathrm{HT}}\right\}$, then $Z_n = \sqrt{n}\ \left(\widehat{\mu}_{\mathrm{HT}} - \mu\right)/\widehat{\sigma}_{\mathrm{HT}} \xrightarrow{d} N(0,1)$ as $\nu \to \infty$.

Conditions under which this assumption is valid are discussed in Section 1.3 of Fuller (2009), for example.

As in Section 2.1, let $F_{\ell Y}(y)$ denote the conditional distribution function of Y given $G = \ell$ and $\widehat{F}_{\ell n}(y) = \frac{1}{\widehat{M}_\ell}\sum_{i\in s_\ell} w_i I(Y_i \le y)$ its sample estimator ($\ell = 0, 1$). Our derivation depends on the following assumption:

A1 The sequence $\sqrt{n}\begin{pmatrix} \widehat{F}_{0n} - F_{0Y} \\ \widehat{F}_{1n} - F_{1Y} \end{pmatrix}$ converges weakly to a bivariate Gaussian process as $\nu \to \infty$.

Let $D_Y(y) = F_{0Y}(y) - F_{1Y}(y)$ and $\widehat{D}_n(y) = \widehat{F}_{0n}(y) - \widehat{F}_{1n}(y)$. Note that $D_Y(y) \equiv 0$ under the

null hypothesis that $F_{0Y} = F_{1Y}$. Define the midrank function, $U_Y(y)$, by setting $U_Y(y) = \frac{1}{2}[F_Y(y) + F_Y(y^-)]$ and let $\widehat{R}_n$ denote its sample estimator, $\widehat{R}_n(y) = \frac{1}{2}[\widehat{F}_n(y) + \widehat{F}_n(y^-)]$. We need the following preliminary result:

Lemma 1. Under assumptions A0 and A1, $\sqrt{n}\begin{pmatrix} \widehat{D}_n - D_Y \\ \widehat{R}_n - U_Y \end{pmatrix}$ converges weakly to a bivariate Gaussian process, say $Z(y) = \begin{pmatrix} Z_1(y) \\ Z_2(y) \end{pmatrix}$.

Proof: This follows directly from A0 and A1 on noting that $\widehat{R}_n - U_Y$ is a function of $\widehat{F}_n - F_Y$ which can be written in the form $\widehat{F}_n - F_Y = p_0(\widehat{F}_{0n} - F_{0Y}) + p_1(\widehat{F}_{1n} - F_{1Y}) + (\widehat{p}_0 - p_0)D_Y + \epsilon_n$ where p_ℓ denotes the probability that $G = \ell$ in the superpopulation, $\widehat{p}_\ell = \sum_{i \in s_\ell} w_i / \sum_{i \in s} w_i$ is its sample estimator ($\ell = 0, 1$)and $\epsilon_n = (\widehat{p}_0 - p_0)(\widehat{D}_n - D_Y)$. □

Our first result requires the following additional assumption:

A2 $g()$ is differentiable on $(0, 1)$, with derivative bounded and continuous on closed subintervals of $(0, 1)$.

Result 1 If assumptions A0, A1, and A2 hold, then:
(a) $\sqrt{n}(\widetilde{T}_n - \delta_Y) \xrightarrow{d} N(0, \tilde{\sigma}^2)$, where $\delta_Y = \int g(U_Y)\, dD_Y$;
(b) If $H_0 : F_{0Y} \equiv F_{1Y}$ is true, then $\delta_Y = 0$ and $\sqrt{n}(\widehat{T}_n - \widetilde{T}_n) \xrightarrow{p} 0$.

Proof: We can write $\widehat{T}_n$ in the form

$$\widehat{T}_n = \int g(\widehat{R}_n)\, d\widehat{F}_{0n} - \int g(\widehat{R}_n)\, d\widehat{F}_{1n} = \int g(\widehat{R}_n)\, d\widehat{D}_n$$

It follows from Lemmas 12.2 and 12.3 of (Kosorok 2008) that $(D, R) \mapsto \phi(D, R) = \int g(R)\, dD$ is Hadamard differentiable under A2 with Hadamard derivative

$$\phi'(\alpha, \beta) = \int g(R) d\alpha + \int \beta g'(R) dD.$$

It then follows from Lemma 1 and the functional delta method (Kosorok 2008, Theorem 12.1) that

$$\sqrt{n}(\widehat{T}_n - \delta_Y) \overset{d}{\rightsquigarrow} \int g(U_Y) dZ_1 + \int Z_2 g'(U_Y) dD_Y,$$

where $\delta_Y = \int g(U_Y)\, dD_Y$. Similarly we can write $\widetilde{T}_n$ in the form

$$\widetilde{T}_n = \int g(U_Y)\, d\widehat{D}_n$$

and use the same argument to show that

$$\sqrt{n}(\widetilde{T}_n - \delta_Y) \overset{d}{\rightsquigarrow} \int g(U_Y) dZ_1 \sim N(0, \tilde{\sigma}^2).$$

It also follows that

$$\sqrt{n}\left(\widehat{T}_n - \widetilde{T}_n\right) \overset{d}{\rightsquigarrow} \int Z_2 g'(U_Y) dD_Y,$$

which is normally distributed with mean zero and finite variance. If H_0 is true, then $D_Y = 0$ so that $\delta_Y = \int g(U_Y)\, dD_Y = 0$ and $\sqrt{n}\left(\widehat{T}_n - \widetilde{T}_n\right) \xrightarrow{p} 0$. □

Assumption A2 rules out several statistics of interest, most notably Mood's test for the median and similar tests for quantiles. To cover such cases, we can replace A2 by the alternative assumptions:

A2a $g(r)$ is the indicator function of a subinterval (a, b) of $(0, 1)$.

Now, however, we need to make an additional assumption of absolute continuity:

A3 Y has a bounded density, f_Y say, with respect to Lebesgue measure in the superpopulation.

Result 2 If assumptions A0, A1, A2a, and A3 hold, then:
(a) $\sqrt{n}(\widetilde{T}_n - \delta_Y) \xrightarrow{d} N(0, \tilde{\sigma}^2)$, where here $\delta_Y = D_Y(U_Y^{-1}(b)) - D_Y(U_Y^{-1}(a))$;
(b) If $H_0 : F_{0Y} \equiv F_{1Y}$ is true, then $\delta_Y = 0$ and $\sqrt{n}(\widehat{T}_n - \widetilde{T}_n) \xrightarrow{p} 0$.

Proof: Here we can write $\widehat{T}_n$ in the form

$$\widehat{T}_n = \widehat{F}_{0n}(\widehat{R}_n^{-1}(b)) - \widehat{F}_{0n}(\widehat{R}_n^{-1}(a)) - \widehat{F}_{1n}(\widehat{R}_n^{-1}(b)) + \widehat{F}_{1n}(\widehat{R}_n^{-1}(a)) = \widehat{D}_n(\widehat{R}_n^{-1}(b)) - \widehat{D}_n(\widehat{R}_n^{-1}(a)).$$

Let $d(y) = D'(y) = f_0(y) - f_1(y)$, where $f_\ell(y)$ is the conditional density of Y given $G = \ell$. Note that $U_Y = F_Y$ under A3 and that $d(y) = 0$ under H_0. By assumption A3, and Lemmas 12.2 and 12.8 of Kosorok (2008), the map $(D, R) \mapsto \phi(D, R) = D(R^{-1})$ is Hadamard differentiable at the superpopulation values (D_Y, R_Y), with derivative

$$\phi'(\alpha, \beta) = \alpha\left(R^{-1}\right) + \frac{d(R^{-1})}{f_Y(R^{-1})}\beta(R^{-1}).$$

It follows that

$$\sqrt{n}(\widehat{T}_n - \delta_Y) \overset{d}{\rightsquigarrow} Z_1\left(U_Y^{-1}(b)\right) - Z_1\left(U_Y^{-1}(a)\right) + k(b)Z_2\left(U^{-1}(b)\right) - k(a)Z_2\left(U^{-1}(a)\right),$$

where $k(p) = \frac{d(U^{-1}(p))}{f_Y(U^{-1}(p))}$.

Similarly we can write $\widetilde{T}_n$ in the form

$$\widetilde{T}_n = \widehat{D}_n(U_Y^{-1}(b)) - \widehat{D}_n(U_Y^{-1}(a))$$

which, from Lemma 1, leads directly to

$$\sqrt{n}(\widetilde{T}_n - \delta_Y) \overset{d}{\rightsquigarrow} Z_1\left(U_Y^{-1}(b)\right) - Z_1\left(U_Y^{-1}(a)\right),$$

and hence that $\sqrt{n}(\widetilde{T}_n - \delta_Y)$ is asymptotically Normal. It also follows that

$$\sqrt{n}\left(\widehat{T}_n - \widetilde{T}_n\right) \overset{d}{\to} k(b)Z_2\left(U^{-1}(b)\right) - k(a)Z_2\left(U^{-1}(a)\right).$$

If H_0 is true, then $d(y)$, and hence k(p), is identically zero and $\sqrt{n}\left(\widehat{T}_n - \widetilde{T}_n\right) \overset{p}{\to} 0$.

□

To use these results, we need to be able to estimate the asymptotic variance $\tilde{\sigma}^2$ of $\sqrt{n}\,\widetilde{T}_n$. Now $\widetilde{T}_n$ is just the difference between two estimated domain means and all packages for

the analysis of survey data will produce an estimate, say $\widetilde{V}_n(\mathbf{U}_s)$ where $\mathbf{U}_s = \{U_i; i \in s\}$, of the variance of $\sqrt{n}\,\widetilde{T}_n$. Thus, if the U_is were known for the sample units, we could use the standardized quantity, $\widetilde{Z}_n = \sqrt{\frac{n}{\widetilde{V}_n}}\,\widetilde{T}_n$ as a test statistic. Under assumption A0, $\widetilde{Z}_n$ will be asymptotically normal with mean zero under H_0 provided that the fourth moment of $g(U_i)$ is finite (which is true for all standard choices of $g(.)$) and $p_\ell = P(G = \ell) > 0$ for $\ell = 0$ or 1 since $\widetilde{V}_n$ is an algebraic function of Horvitz-Thompson variance estimators. Since we do not know U_i, we substitute the estimated midrank $\widehat{R}_i$ and use $\widehat{Z}_n = \sqrt{\frac{n}{\widehat{V}_n}}\,\widehat{T}_n$ with $\widehat{V}_n = \widetilde{V}_n(\widehat{\mathbf{R}}_s)$ as our test statistic.

Result 3 Under the assumptions of either Result 1 or Result 2,

$$\widehat{Z}_n = \sqrt{\frac{n}{\widehat{V}_n}}\,\widehat{T}_n \xrightarrow{d} N(0,1)$$

as $\nu \to \infty$ if H_0 is true.

Proof: This follows directly from Lemma 1.

List of Figures

1 Quantile–quantile plot of 10,000 realizations of $-\log_{10}(p-\text{values})$ for three rank tests and the t-test, in a stratified cluster sample with 20 strata and 40 clusters of size 100. Grey dots indicate a Normal reference distribution for the test statistic; black circles indicate a t distribution with 20 degrees of freedom. 30

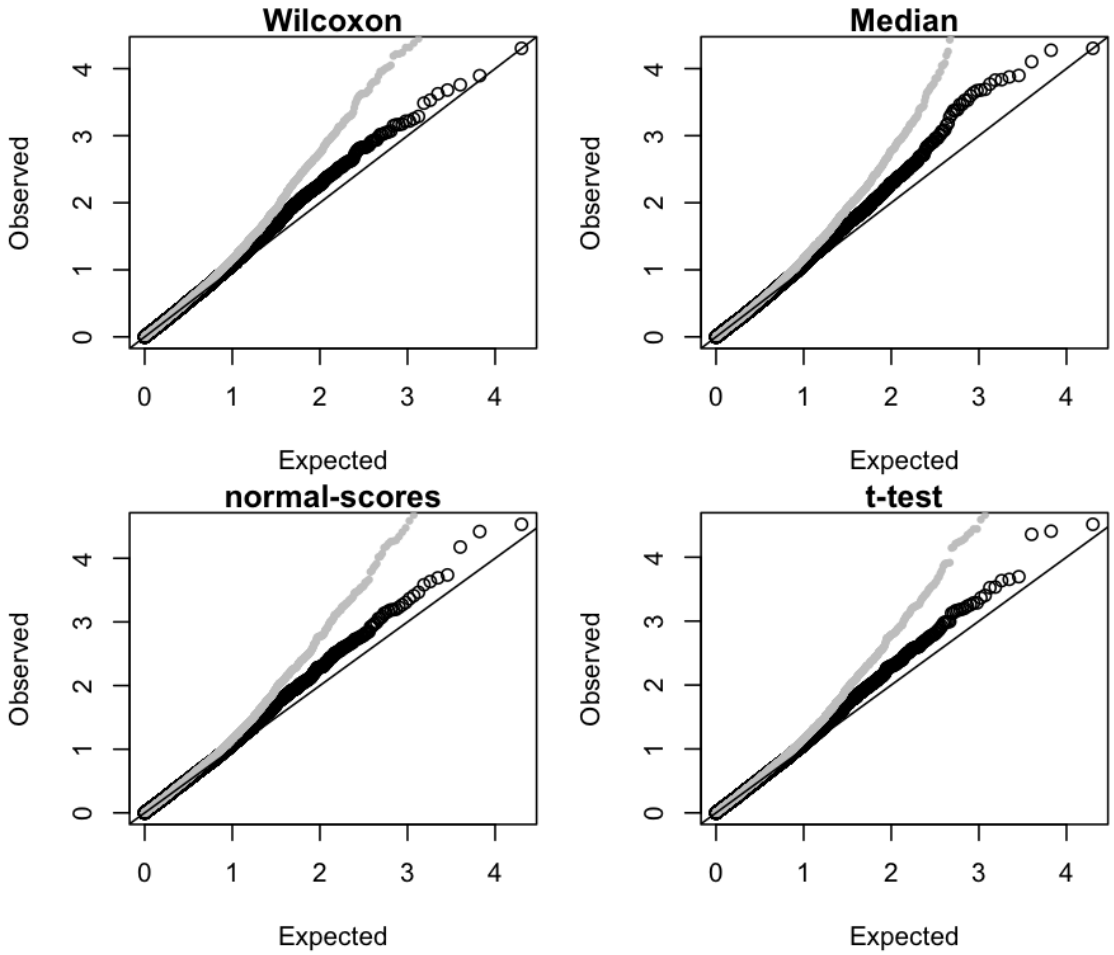

Figure 1: Quantile–quantile plot of 10,000 realizations of $-\log_{10}(p-\text{values})$ for three rank tests and the t-test, in a stratified cluster sample with 20 strata and 40 clusters of size 100. Grey dots indicate a Normal reference distribution for the test statistic; black circles indicate a t distribution with 20 degrees of freedom.

List of Tables

1 Estimated relative efficiencies relative to a (weighted) t-test for three rank tests and three superpopulation outcome distributions, under four sampling designs. N is a Normal distribution, E is a double exponential (Laplace) distribution, L is a logistic distribution. Fixed/Random M means the group sizes are fixed or random in the sampling design, equal/unequal π means the sampling is with equal probability or with oversampling of the tails of the distributions. Based on 10,000 replications, samples of size 500 from a population of size 10,000 32

Table 1: Estimated relative efficiencies relative to a (weighted) t-test for three rank tests and three superpopulation outcome distributions, under four sampling designs. N is a Normal distribution, E is a double exponential (Laplace) distribution, L is a logistic distribution. Fixed/Random M means the group sizes are fixed or random in the sampling design, equal/unequal π means the sampling is with equal probability or with oversampling of the tails of the distributions. Based on 10,000 replications, samples of size 500 from a population of size 10,000

	Wilcoxon			Median			Normal-scores		
	N	E	L	N	E	L	N	E	L
Random M, unequal π	0.91	1.41	1.03	0.71	1.62	0.85	1.00	1.23	1.01
Fixed M, unequal π	0.95	1.62	1.10	0.78	1.98	0.97	0.99	1.33	1.03
Random M, equal π	0.89	1.45	0.99	0.43	1.77	0.70	1.00	1.23	1.00
Fixed M, equal π	0.93	1.65	1.07	0.49	2.23	0.79	1.00	1.31	1.02