RANK TESTS WITH DATA FROM A COMPLEX SURVEY

THOMAS LUMLEY AND ALASTAIR SCOTT*

ABSTRACT. Rank tests are widely used for exploratory and formal inference in the health and social sciences. With the widespread use of data from complex survey samples in medical and social research, there is increasing demand for versions of rank tests that account 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: Complex sampling; Multistage sampling; Rank Tests; Sample surveys.

1. INTRODUCTION

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 or the British Household Panel Survey 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 means that two criteria should be satisfied: that the same population null hypothesis is tested regardless of the sampling scheme and that the test has the specified level, at least asymptotically, when this hypothesis is true.

In the absence of design-based methods, rank tests, especially the Wilcoxon rank-sum test, are currently 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. To give some idea of the extent of this use, GoogleScholar lists almost 2,500 papers with "NHANES" and "Wilcoxon" in the abstract. 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. Similarly, the Wilcoxon rank-sum test was used by Leece(2000) to compare households with fixed- and floating-rate mortgages using data from the British Household Panel Survey. 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.

Recently, Natarajan et al (2012) developed 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 random sampling. Their method is 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.

2. Construction of the rank test

2.1. **Comparing groups within a survey.** 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), where Y_i is real-valued and $G_i \in \{0, 1\}$ is a grouping variable. 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. We want to test the null hypothesis that Y is independent of G against the alternative that Y is stochastically ordered by G. In other words, we want to test $H_0 : F_{0Y}(y) \equiv F_{1Y}(y)$ where $F_{\ell Y}(y)$ denotes the conditional distribution function of Y given $G = \ell$ $(\ell = 0, 1)$.

First consider the finite-population quantity that will be estimated by the sample rank test. Let

$$\mathbb{F}_N(y) = \frac{1}{N} \sum_{i=1}^{N} I(Y_i \leq y)$$

denote the empirical finite-population distribution function of Y. The scaled finite population ranks, $R_1, \ldots, R_N$, are defined by setting $R_i = \mathbb{F}_N(Y_i)$. 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 $M_\ell = \sum_{i=1}^{N} I(G_i = \ell)$ is the number of finite population units in group ℓ.

For simplicity, we shall assume that Y has a continuous distribution. However, when g is a continuous function, the test can be extended to include discrete Y by replacing

R_i with the mid-rank, $R_i^* = \{\mathbb{F}_N(y) + \mathbb{F}_N(y^-)\}/2$ with $\mathbb{F}_N(y^-) = \sum_{j=1}^N I(Y_j < y)/N$, as in Conover (1973) or Hudgens & Satten (2002).

We draw a sample, s, of n units from the finite population using some probability sampling design with selection 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. The estimated population ranks $\widehat{R}_i$ are defined as $\widehat{R}_i = \widehat{F}_n(Y_i)$ where

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

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

If $\widehat{R}_i$ was 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 through the weights. If we set $U_i = F_Y(Y_i)$, 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. The classical proof that U_i can be substituted for R_i with no effect on the asymptotic null distribution of the full finite population statistic T_N relies heavily on exchangeability (see Hajek & Sidak, 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 $N^{1/2}\{\mathbb{F}_N(y) - F_Y(y)\}$ to a Brownian Bridge (van der Vaart & Wellner, 1996) can be adapted for complex samples. We adopt this approach here and show 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, if we set $\widehat{D}_n(y) = \widehat{F}_{0n}(y) - \widehat{F}_{1n}(y)$ where $\widehat{F}_{\ell n}(y) = \widehat{M}_\ell^{-1}\sum_{i\in s_\ell} w_i I(Y_i \le y)$ is the estimator of $F_{\ell Y}$ ($\ell = 0$ or 1), then we can write

$$n^{1/2}\left(\widehat{T}_n - \widetilde{T}_n\right) = n^{1/2}\int \left\{g(\widehat{F}_n) - g(F_Y)\right\} d\widehat{D}_n(y).$$

Let $\delta_Y = \int g(y)\, dD_Y(y)$. Then, under suitable conditions, $n^{1/2}(\widehat{T}_n - \delta_Y)$ and $n^{1/2}(\widetilde{T}_n - \delta_Y)$ are asymptotically normal with mean zero. Moreover, if $H_0 : F_{0Y} \equiv F_{1Y}$ is true, then $\delta_Y = 0$ and $n^{1/2}(\widehat{T}_n - \widetilde{T}_n) \to 0$ in probability. Details are given in Lumley & Scott (2013).

Now $\widetilde{T}_n$ is simply a difference between two estimated domain means. Methods for computing an estimate of its variance, $\widetilde{V}_n(U)$ say, are standard and now available routinely in most general-purpose statistical software. The asymptotic equivalence of $\widehat{T}_n$ and $\widetilde{T}_n$ means that we can use replace U_i by $\widehat{R}_i$ without affecting the asymptotic null distribution. Thus, a design-based rank test can be based on the test statistic $\widehat{Z}_n = \widehat{T}_n / \widehat{V}_n^{1/2}$, where $\widehat{V}_n = \widetilde{V}_n(\widehat{R})$. We have provided an implementation in the `svyranktest` function in the R package `survey`. 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. 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.

2.2. **Comparing a targeted sample to a population survey.** Another common use of rank tests is to compare measurements from a targeted sample to reference values obtained from a large national survey. The NHANES series of surveys, which includes a wide range of assays performed on blood samples giving population distributions for nutrients, environmental pollutants, disease biomarkers, and other variables, is particularly useful for such comparisons..

When the targeted sample is a well-defined probability sample from a population that is distinct from that sampled in the big survey, the targeted sample and main sample 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 main population, as when comparing data from a state survey with national data. In this scenario we can treat the data as a dual-frame survey (Lohr & Rao, 2000). Metcalf & 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 data from a nation–wide survey, it would be necessary to decide how to apportion the weight for California between the Californian survey and the Californian subset of the national survey. 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, a small targeted sample that was not drawn according to any probability mechanism is being compared to a reference distribution obtained from a large survey. 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 will be a negligible fraction of the larger sampling frame , we propose to use the sampling weights from the larger survey without modification. We use weights $w_i = 1$ for the targeted sample, reflecting the fact that they are members of the main 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. Example: Serum copper and anaemia 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 anaemia, and compared serum copper concentrations in people with and without anaemia using data from NHANES II. They reported unweighted median serum copper concentrations in anemic and non-anaemic 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 overstates the precision of the comparison. Finally, the survey 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 naive Wilcoxon test.

We should note that 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 design, and their conclusions of a U-shaped relationship between serum copper and anemia are still supported by the analysis.

References

[1] Conover, W. J. (1973). Rank tests for one sample, two samples, and k samples without the assumption of a continuous distribution function. *Annals of Statistics* 1, 1105-1125.

[2] Hajek, J. & Sidak, Z. (1967). *Theory of Rank Tests.* Prague: Academic Press.

[3] Hudgens, M. & Satten, G. (2002). Midrank unification of rank tests for exact, tied, and censored data. *Journal of Nonparametric Statistics* 14, 569-581.

[4] Knovich, M. A., IlYasova, D., Ivanova, A. & Molnar, 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.

[5] Korn E.L. & Graubard B.I. (1999) *Analysis of Health Surveys.* John Wiley and Sons.

[6] Leece, D. (2000). Household choice of fixed versus floating rate debt. *Oxford Bulletin of Economics & Statistics* 62, 61-82

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

[8] Lumley, T. (2013). "survey: analysis of complex survey samples". R package version 3.29-3. URL: http://cran.r-project.org/package=survey.

[9] Lumley, T. & Scott, A.J. (2013). Two-sample rank tests under complex sampling *Biometrika*, **100**, to appear.

[10] Metcalf, P. A. & Scott, A. J. (2009). Using multiple frames in health surveys. *Statistics in Medicine* 1512–1523.

[11] Natarajan, S., Lipsitz, S. R., Sinha, D. & Fitzmaurice, G. (2012). An extension of the Wilcoxon rank-sum test for complex sample survey data. *Applied Statistics* 61, 653-664

[12] van der Vaart, A. W. & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* New York: Springer-Verlag.

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

Department of Statistics, University of Auckland, Auckland, NZ

E-mail address: a.scott@auckland.ac.nz