Published in final edited form as:
J R Stat Soc Ser C Appl Stat. 2012 August ; 61(4): 653–664. doi:10.1111/j.1467-9876.2011.01028.x.

An extension of the Wilcoxon Rank-Sum test for complex sample survey data

SUNDAR NATARAJAN,
VA New York Harbor Healthcare System, New York, NY, U.S.A

STUART R. LIPSITZ,
Harvard Medical School, Boston MA, U.S.A

GARRETT M. FITZMAURICE,
Harvard Medical School, Boston MA, U.S.A

DEBAJYOTI SINHA,
The Florida State University, Talahassee, FL, U.S.A

JOSEPH G. IBRAHIM,
University of North Carolina, Chapel Hill NC, U.S.A

JENNIFER HAAS, and
Harvard Medical School, Boston MA, U.S.A

WALID GELLAD
University of Pittsburgh, Pittsburgh PA, U.S.A

Summary

In complex survey sampling, a fraction of a finite population is sampled. Often, the survey is conducted so that each subject in the population has a different probability of being selected into the sample. Further, many complex surveys involve stratification and clustering. For generalizability of the sample to the finite population, these features of the design are usually incorporated in the analysis. While the Wilcoxon rank sum test is commonly used to compare an ordinal variable in bivariate analyses, no simple extension of the Wilcoxon rank sum test has been proposed for complex survey data. With multinomial sampling of independent subjects, the Wilcoxon rank-sum test statistic equals the score test statistic for the group effect from a proportional odds cumulative logistic regression model for an ordinal outcome. Using this regression framework, for complex survey data, we formulate a similar proportional odds cumulative logistic regression model for the ordinal variable, and use an estimating equations score statistic for no group effect as an extension of the Wilcoxon test. The proposed method is applied to a complex survey designed to produce national estimates of the health care use, expenditures, sources of payment, and insurance coverage.

Keywords

Cumulative logistic model; Medical Expenditure Panel Survey; Proportional odds model; Score statistic; Weighted estimating equations

1 Introduction

The Wilcoxon rank-sum test is a frequently used statistical test to compare an ordinal outcome between two groups of subjects. Even in cases where regression analyses are subsequently performed, initial summaries in terms of bivariate analyses are regularly reported at the beginning of the results section of published papers. In this paper, we propose

an extension of the Wilcoxon rank-sum test to complex survey data. In complex survey sampling, a fraction of a finite population is sampled, while accounting for its size and characteristics. Based on certain subject characteristics (for example, age, race, gender), some individuals may be over or under sampled. Thus, individuals in the population may have different probabilities of being selected into the sample. Further, the sampling design can have multiple stages of stratification and clustering. Thus, in general, the design for complex sample surveys often includes stratification, clustering, and different selection probabilities. Although alternative approaches have been proposed for analyzing complex survey data (Chambers and Skinner, 2003), for generalizability of the sample to the finite population (Korn and Graubard, 1999), in this paper we incorporate the design in the analysis, including sampling weights (derived from the probability of selection into the survey), strata and/or cluster variables.

Extensions of rank-sum tests have been proposed for clustered data (Jung and Kang, 2001; Rosner, Glynn, and Lee, 2003; Datta and Satten, 2005), which would arise from a random sample of independent clusters. In these proposed tests, the variance of the usual Wilcoxon test is adjusted for clustering. The multi-stage sampling design, with different probabilities of selection, has been the roadblock in developing a general extension of the Wilcoxon test procedure to complex surveys.

The ready availability of public-use data from large population-based complex sample surveys has led to the calculation of population estimates of frequency of disease (incidence and prevalence) and to associations between risk factors and disease. Many seminal papers published in journals such as the British Medical Journal, the Lancet, the New England Journal of Medicine and the Journal of the American Medical Association have been based on such complex survey data. In the United States, examples of such complex surveys include the National Health and Nutrition Examination Surveys (NHANES), Behavioral Risk Factor Surveillance System (BRFSS), National Health Care Surveys (NHCS), the Nationwide Inpatient Sample (NIS), the Hospital Consumer Assessment of Healthcare Providers and Systems (HCAHPS) survey, and the Medical Expenditure Panel Survey (MEPS). In the United Kingdom, examples are the Annual Population Survey (APS), British Social Attitudes (BSA), Family Resources Survey (FRS), Health Survey for England (HSE), and the Scottish Health Survey (SHeS). For example, the paper ‘Epidemic of obesity in UK children’ (Reilly and Dorosty, 1999), using HSE data, was published in *The Lancet* and the paper ‘Adolescent Overweight and Future Adult Coronary Heart Disease’ (Bibbins-Domingo et al., 2007), using NHANES data, was published in the New England Journal of Medicine. Further, a search of PubMed (National Library of Medicine) abstracts using the words “NHANES” yielded 7699 articles in the last 5 years; NHANES is just one of at least a hundred national complex surveys conducted in different countries around the world. Despite the huge increase in the use of such complex sample surveys, there has not been a simple proposed extension of the Wilcoxon test for comparing an ordinal outcome between two groups of subjects in complex survey data.

Our motivating example is the Medical Expenditure Panel Survey (MEPS; Cohen, 2003) for the year 2002, conducted by the United States National Center for Health Statistics, Centers for Disease Control and Prevention. The 2002 cross-sectional survey was designed to produce national and regional estimates of the health care use, expenditures, sources of payment, and insurance coverage of the United States civilian non-institutionalized population. MEPS is a stratified, multistage probability cluster sample. We analyze data from the 25,388 subjects who participated in the Household Component of MEPS. In the design of the study, the population was first stratified into 203 geographical regions, defined by census and state regions, metropolitan status, and so-ciodemographic measures. Within each stratum, the geographical region was subdivided into area segments, which are

composed of counties or groups of contiguous counties. Two or three clusters (area segments) were sampled within each stratum. In 161 of the strata, there were 2 clusters; in the remaining 42 strata, there were 3 clusters. On average, a typical cluster contained 59 subjects, with a range of 2 to 316 subjects. Although the clusters within strata were sampled without replacement, for purposes of analysis we can assume they were sampled with replacement, since the fraction of clusters sampled within each stratum is much less than 1%. The study over-sampled Hispanics, African-Americans, adults with functional impairments, children with limitations in activities, individuals predicted to incur high levels of medical expenditures, and low income individuals.

For this MEPS study, we explored if patients with and without health insurance differ in the following ordinal variables: education level, income (defined as percent of poverty line), perceived health status, and body mass index (BMI). Thus, for these cross-sectional data, the 'group' is health insurance (yes, no), and the ordinal variables are education level (no degree, high school graduate equivalency degree–ged, high school diploma, bachelor's degree, master's degree, doctorate degree), income (poor, near-poor, low income, middle income, high income), perceived health status (excellent, very good, good, fair, poor), and BMI (underweight, BMI < 18.5 kg/m^2; normal, BMI 18.5 to 24.9 kg/m^2; overweight, BMI 25.0 to 29.9 kg/m^2; obese, BMI > 30.0 kg/m^2). Table 1 show individual-level data from 25 typical subjects, including strata, cluster, and weights (note that these data are typical, not true data so that subjects cannot be identified).

Motivated by the need to develop an analog for complex sample survey data, we propose an extension of the Wilcoxon rank-sum test to the MEPS. Since subjects in complex surveys are not independent, the assumptions needed for applying the Wilcoxon rank-sum test do not hold. With a multinomial sample of independent subjects, the Wilcoxon rank-sum test statistic is shown to equal the score test statistic for no effect of a single dichotomous covariate in a proportional odds cumulative logistic regression model (McCullagh, 1980) for the ordinal outcome. Using this regression framework, the Wilcoxon test is then extended to general complex sample surveys. With complex survey data, we propose formulating a proportional odds cumulative logistic regression model for an ordinal outcome, with the group as a dichotomous covariate. Weighted estimating equations (WEE; Shah et al., 2001; Binder, 1983; Pfefferman, 1993) are used to account for the weighting and sampling design. We propose use of an estimating equations score test (Rao et al.,1998) for no group effect for the proportional odds model; this score test reduces to the standard Wilcoxon test if the design is a multinomial sample. The test can be obtained with minimal programming in common statistical software such as SAS Proc Surveylogistic.

In Section 2, we introduce some notation and discuss the score test for a simple multinomial sample. In Section 3, we describe weighted estimating equations (WEE) for complex survey data, and our proposed score test statistic based on WEE. Finally, in Section 4, we present the results of analyses of the MEPS data to illustrate the proposed method.

2 Proportional Odds Model and the Wilcoxon Rank-Sum Test

In this section, we describe the proportional odds model and the Wilcoxon rank-sum test for a multinomial sample of n independent subjects, $i = 1, 2, \ldots, n$. In the next section, we extend these results to complex survey sampling. We assume the outcome Y_i is an ordinal discrete random variable which can take on positive integer values $j = 1, 2, , \ldots, J$. Since the outcome has J levels, we can form J indicator random variables Y_{ij}, where $Y_{ij} = 1$ if subject i has response j and $Y_{ij} = 0$ if otherwise. Our goal is to determine if the distribution of this ordinal outcome differs across two groups. Thus we form a dichotomous covariate x_i, where

$x_i = 1$ if subject i is in group 1 and $x_i = 0$ if subject i is in group 2. Then, we denote the probability of response j given x_i as

$$p_{ij}=pr(Y_i=j|x_i)=pr(Y_{ij}=1|x_i),$$

and the multinomial probability mass function for subject i equals

$$f(y_{i1}, y_{i2}, \ldots, y_{iJ})=\prod_{j=1}^{J} p_{ij}^{y_{ij}}. \quad (1)$$

The proportional odds model can be written as

$$\gamma_{ij}=pr(Y_i \leq j|x_i, \theta, \beta)=\frac{\exp(\theta_j - x_i\beta)}{1+\exp(\theta_j - x_i\beta)}, \quad (2)$$

where γ_{ij} is a 'cumulative probability' and $\boldsymbol{\theta}' = [\theta_1, \ldots, \theta_{J-1}]$ is the vector of cumulative intercepts. Since

$$\begin{aligned} p_{ij}&=pr(Y_i=j|x_i) \\ &=pr(Y_i \leq j|x_i, \theta, \beta)-pr(Y_i \leq j-1|x_i, \theta, \beta) \quad (3) \\ &=\gamma_{ij}-\gamma_{i,j-1}, \end{aligned}$$

with $\gamma_{iJ} = 1$ and $\gamma_{i0} = 0$, the contribution to the likelihood for subject i can be rewritten as

$$L_i(\theta, \beta)=\prod_{j=1}^{J}[\gamma_{ij}-\gamma_{i,j-1}]^{y_{ij}}. \quad (4)$$

Our main interest is in testing for no group effect in (2), i.e.,

$$H_0; \beta = 0.$$

Under this null hypothesis the distribution of the ordinal variable is identical in the two groups. As we briefly describe here, the Wilcoxon rank-sum test statistic equals the score test statistic for testing $\beta = 0$. Next, consider the general form of a score test statistic for testing $\beta = 0$. If $\hat{\boldsymbol{\theta}}_0$ is the maximum likelihood estimate of $\boldsymbol{\theta}$ under the null hypothesis that $\beta = 0$, then the score test statistic for testing the null hypothesis has the general form

$$X^2=\mathbf{U}(\widehat{\theta_0}, 0)' \{\mathrm{Var}[\mathbf{U}(\theta, \beta)]\}^{-1}_{\theta=\widehat{\theta_0}, \beta=0} \mathbf{U}(\widehat{\theta_0}, 0), \quad (5)$$

where $\mathbf{U}(\boldsymbol{\theta}, \beta)$ is the score (first derivative) vector of the log of the likelihood in (4) with respect to $\boldsymbol{\varphi} = (\boldsymbol{\theta}', \beta)'$; $\mathbf{U}(\hat{\boldsymbol{\theta}}_0, 0)$ is the score vector evaluated at $(\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_0, \beta = 0)$; and $\{\mathrm{Var}[\mathbf{U}(\boldsymbol{\theta}, \beta)]\}_{\theta=\hat{\theta}_0 \beta=0}$ is the variance of $\mathbf{U}(\boldsymbol{\theta}, \beta)$ evaluated at $(\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_0, \beta = 0)$. Note, for regular likelihood problems such as this when subjects are independent,

$$\mathrm{Var}[\mathbf{U}(\theta,\beta)]=E[\mathbf{U}(\theta,\beta)\mathbf{U}(\theta,\beta)']=-E\left[\frac{d}{d\boldsymbol{\varphi}}\mathbf{U}(\theta,\beta)'\right].$$

Under the null hypothesis, X^2 in (5) has an asymptotic chi-square distribution with 1 degree-of-freedom.

For the proportional odds model (McCullagh and Nelder, 1989), the general form of the score vector equals

$$\mathbf{U}(\theta,\beta)=\sum_{i=1}^{n}\sum_{j=1}^{J}\frac{dp_{ij}}{d\boldsymbol{\varphi}}(y_{ij}-p_{ij})/p_{ij}.$$

The only non-zero component of $\mathbf{U}(\hat{\boldsymbol{\theta}}_0, 0)$ equals

$$U_0=\sum_{i=1}^{n}\sum_{j=1}^{J}\left[\frac{dp_{ij}}{d\beta}\right]_{\theta=\widehat{\theta}_0,\beta=0}(y_{ij}-\widehat{p}_j)/\widehat{p}_j=\sum_{i=1}^{n}\sum_{j=1}^{J}\left[\frac{d(\gamma_{ij}-\gamma_{i,j-1})}{d\beta}\right]_{\theta=\widehat{\theta}_0,\beta=0}(y_{ij}-\widehat{p}_j)/\widehat{p}_j. \quad (6)$$

where $\widehat{p}_j=n^{-1}\sum_{i=1}^{n}Y_{ij}$ is the level j proportion (regardless of group). Since γ_{ij} is a logistic regression model, using results for ordinary logistic regression,

$$\frac{d(\gamma_{ij}-\gamma_{i,j-1})}{d\beta}=x_i[\gamma_{ij}(1-\gamma_{ij})-\gamma_{i,j-1}(1-\gamma_{i,j-1})].$$

Thus, the non-zero component of $\mathbf{U}(\hat{\boldsymbol{\theta}}_0, 0)$ can be written as

$$U_0=\sum_{i=1}^{n}x_i\sum_{j=1}^{J}S_j(y_{ij}-\widehat{p}_j), \quad (7)$$

where

$$S_j=\widehat{p}_j^{-1}[\gamma_{ij}(1-\gamma_{ij})-\gamma_{i,j-1}(1-\gamma_{i,j-1})]_{\theta=\widehat{\theta}_0,\beta=0}=\frac{\widehat{\gamma}_j(1-\widehat{\gamma}_j)-\widehat{\gamma}_{j-1}(1-\widehat{\gamma}_{j-1})}{\widehat{\gamma}_j-\widehat{\gamma}_{j-1}},$$

and $\widehat{\gamma}_j=\sum_{k=1}^{j}\widehat{p}_k$ is the cumulative proportion (regardless of group). Straightforward algebra shows that $S_j = 1 - (\hat{\gamma}_j + \hat{\gamma}_{j-1})$, and that (7) is proportional to

$$U_0=\sum_{i=1}^{n}x_i\sum_{j=1}^{J}.5(\widehat{\gamma}_j+\widehat{\gamma}_{j-1})(y_{ij}-\widehat{p}_j), \quad (8)$$

where

$$.5(\widehat{\gamma}_j+\widehat{\gamma}_{j-1})$$

is the 'ridit' score (Bross, 1958). We further show in the Appendix that (8) is proportional to

$$U_0=\sum_{i=1}^{n}x_i\sum_{j=1}^{J}R_j(y_{ij}-\widehat{p}_j)=\sum_{j=1}^{J}R_j\sum_{i=1}^{n}y_{ij}x_i-\sum_{j=1}^{J}R_j\widehat{p}_j\sum_{i=1}^{n}x_i, \quad (9)$$

where

$$R_j=n(\widehat{\gamma}_j+\widehat{\gamma}_{j-1})/2+0.5$$

is the average rank or 'midrank' for subjects in category j. Subjects with $x_i = 0$ do not contribute to (9). For subjects in group $x_i = 1$, this statistic multiplies the average ranks in category j (R_j) by the number of subjects in category j ($\sum_{i=1}^{n}x_iy_{ij}$), and then sums across all categories. Further, the sum of the average ranks in group $x_i = 1$ under the null is subtracted from the observed ranks. Thus, the form of the score statistic in (9) is equivalent to the Wilcoxon rank-sum test statistic.

By formulating the Wilcoxon test statistic in terms of a score test statistic from the proportional odds model, one can apply theory developed for estimating equations score tests to the proportional odds models in the complex sample survey setting, without having to develop new theory for ranks in complex survey data. As discussed in Agresti (2010), the Wicoxon rank-sum test statistic can be written either in terms of the ridits as in (8) or the midranks as in (9). In the next section, we discuss the estimating equations score test for complex survey data in terms of the ridits.

3 Extension of the Wilcoxon Rank-Sum Test for Complex Survey Data

To develop our extension of the Wilcoxon rank-sum test, we first discuss weighted estimating equations for estimating ($\boldsymbol{\theta}$, $\boldsymbol{\beta}$) in complex sample surveys. For complex sample surveys, the target population is usually thought to be of finite size N. We assume the sample is still of size n. To indicate which n subjects are sampled from the population of N subjects, we define the indicator random variable

$$\delta_i=\begin{cases}1 & \text{if subject } i \text{ is selected into sample}\\ 0 & \text{if subject } i \text{ is not selected into sample}\end{cases},$$

for $i = 1, \ldots, N$, where $\sum_{i=1}^{N}\delta_i=n$. Depending on the sampling design, some of the δ_i could be correlated (e.g., for two subjects within the same cluster). We let $\boldsymbol{\pi}_i$ denote the probability of subject i being selected into the survey, which is typically specified in the design of the study. Depending on the sampling design, $\boldsymbol{\pi}_i$ may depend on the outcome of interest, the independent variables, or additional variables (screening variables, for example) not in the model of interest. In particular, $\boldsymbol{\pi}_i = pr(\delta_i = 1|y_i, \mathbf{x}_i, \mathbf{s}_i)$, where $\mathbf{s}_i$ is a vector of additional variables. We assume that the proportional odds model holds for subjects in the

population, and the distribution of the ordinal outcome for a subject in the population follows a multinomial distribution as in (4).

To obtain a consistent estimate of ($\boldsymbol{\theta}$, β), one can use a weighted estimating equation, which is the solution to $\mathbf{U}_{wee}(\hat{\boldsymbol{\theta}}, \hat{\beta}) = \mathbf{0}$, where

$$\mathbf{U}_{wee}(\widehat{\theta},\widehat{\beta})=\sum_{i=1}^{N}\frac{\delta_i}{\pi_i}\sum_{j=1}^{J}\left[\frac{dp_{ij}}{d\boldsymbol{\varphi}}\right]_{\theta=\widehat{\theta},\beta=\widehat{\beta}}(y_{ij}-\widehat{p}_{ij})/\widehat{p}_{ij}, \quad (10)$$

and $\boldsymbol{\varphi} = (\boldsymbol{\theta}', \beta)'$. Here, the 'weights' are $w_i=\frac{\delta_i}{\pi_i}$ $(w_i=\frac{1}{\pi_i}\text{if sampled}\,\delta_i=1)$. Note, also, these are weighted likelihood score equations under a working 'independence' assumption for the N subjects (disregarding any clustering).

Using a first order Taylor series expansion and a suitable central limit theorem for sample survey data (Binder, 1983), ($\hat{\boldsymbol{\theta}}$, $\hat{\beta}$) has an asymptotic multivariate normal distribution with mean ($\boldsymbol{\theta}$, β) and covariance matrix

$$V_{\theta,\beta}=\mathrm{Var}[(\widehat{\theta},\widehat{\beta})]=\left[E\left(\frac{d\mathbf{U}_{wee}(\theta,\beta)}{d\boldsymbol{\varphi}}\right)\right]^{-1}\mathrm{Var}[\mathbf{U}_{wee}(\theta,\beta)]\left[E\left(\frac{d\mathbf{U}_{wee}(\theta,\beta)}{d\boldsymbol{\varphi}}\right)\right]^{-1}, \quad (11)$$

Note, Var[$\mathbf{U}_{wee}(\boldsymbol{\theta}, \beta)$] depends on the sample design (stratification, clustering, sampling with or without replacement) as well as the finite population correction factor. Empirically, (11) is estimated via the 'sandwich variance estimator'. For multionomial and ordinal logistic regression, sample survey programs in SAS, Sudaan, R, and Stata can be used to calculate this variance.

Next, we apply an estimating equations score test statistic (Rao et al., 1998) for the null hypothesis, H_0:$\beta = 0$, in the proportional odds model. Simlar to the previous section, we let $\hat{\boldsymbol{\theta}}_0$ denote the WEE estimate of $\boldsymbol{\theta}$ under the null hypothesis that $\beta = 0$. Then, similar to the usual score test, the estimating equations score test statistic for H_0:$\beta = 0$ is

$$X^2=\mathbf{U}_{wee}(\widehat{\theta}_0,0)'\left[\{\mathrm{Var}[\mathbf{U}_{wee}(\theta,\beta)]\}_{\theta=\widehat{\theta}_0,\beta=0}\right]^{-1}\mathbf{U}_{wee}(\widehat{\theta}_0,0), \quad (12)$$

where the form of $\mathbf{U}_{wee}(\hat{\boldsymbol{\theta}}_0, 0)$ and $\{\mathrm{Var}[\mathbf{U}_{wee}(\boldsymbol{\theta}, \beta)]\}_{\theta=\hat{\theta}_0,\beta=0}$ are both derived under the alternative, but evaluated at ($\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}_0$, $\beta = 0$). **In particular,**

$$\{\mathrm{Var}[\mathbf{U}_{wee}(\theta,\beta)]\}_{\theta=\widehat{\theta}_0,\beta=0}=\left[E\left(\frac{d\mathbf{U}_{wee}(\theta,\beta)}{d\boldsymbol{\varphi}}\right)\right]_{\theta=\widehat{\theta}_0,\beta=0}\{\mathrm{Var}[(\widehat{\theta},\widehat{\beta})]\}_{\theta=\widehat{\theta}_0,\beta=0}\left[E\left(\frac{d\mathbf{U}_{wee}(\theta,\beta)}{d\boldsymbol{\varphi}}\right)\right]_{\theta=\widehat{\theta}_0,\beta=0}, \quad (13)$$

where

$$\left[E\left(\frac{d\mathbf{U}_{wee}(\theta,\beta)}{d\boldsymbol{\varphi}}\right)\right]_{\theta=\widehat{\theta}_0,\beta=0}=\sum_{i=1}^{N}\frac{\delta_i}{\pi_i}\sum_{j=1}^{J}\widehat{p}_j^{-1}\left[\frac{dp_{ij}}{d\boldsymbol{\varphi}}\right]_{\theta=\widehat{\theta}_0,\beta=0}\left[\frac{dp_{ij}}{d\boldsymbol{\varphi}}\right]'_{\theta=\widehat{\theta}_0,\beta=0}. \quad (14)$$

is the negative of the information matrix obtained if one ignores the complex survey design and assumes all subjects are independent with weights w_i. The central limit theorem can be

used to show that asymptotically X^2 has a chi-square distribution with 1 degree-of-freedom under the null (Rao et al., 1998), although the definition of 'asymptotic' is sometimes non-standard in complex sample surveys. Finite sample approximations for the distribution of (13) are given in Rao et al. (1998). For example, for a stratified, cluster design, if we let S =number of strata and C =number of clusters, then Rao et al. (1998) propose approximating the distribution of (12) with an F-distribution with 1 and $f = C - S$ degrees-of-freedom. For the MEPS data we analyze, since $f = 448 - 203 = 245$, the F- and chi-square approximations are practically identical, so we use the chi-square approximation in the following section.

Similar to the score test for non-complex survey data, the only non-zero component of $\mathbf{U}_{wee}(\hat{\boldsymbol{\theta}}_0, 0)$ can be written as

$$U_{wee,0} = \sum_{i=1}^{N} w_i \sum_{j=1}^{J} \left[\frac{dp_{ij}}{d\beta} \right]_{\theta=\hat{\theta}_0,\beta=0} (y_{ij} - \hat{p}_j)/\hat{p}_j = \sum_{i=1}^{N} w_i x_i \sum_{j=1}^{J} .5(\hat{\gamma}_j + \hat{\gamma}_{j-1})(Y_{ij} - \hat{p}_j), \quad (15)$$

where

$$\hat{p}_j = \frac{\sum_{i=1}^{N} w_i Y_{ij}}{\sum_{i=1}^{N} w_i}$$

is the weighted proportion of subjects with response level j, regardless of group,

$$\hat{\gamma}_j = \sum_{k=1}^{j} \hat{p}_k$$

is the cumulative weighted proportion (regardless of group), and

$$.5(\hat{\gamma}_j + \hat{\gamma}_{j-1})$$

is the weighted ridit. Subjects with $x_i = 0$ do not contribute to (15). Then, the estimating equations score statistic has the same form, in terms of the ridits, as the usual proportional odds statistic from an multinomial sample, and can be considered an extension of the usual Wilcoxon rank-sum test to complex survey data.

Most sample survey programs allow fitting of the proportional odds model for ordinal data from complex sample surveys. However, the estimating equations score statistic is not directly available, and requires a two step procedure. First, one fits the proportional odds model under the null $H_0: \beta = 0$ to get $\hat{\boldsymbol{\theta}}_0$. Then, one fits a proportional odds model under the alternative $H_0: \beta \neq 0$ with starting values $(\hat{\boldsymbol{\theta}}_0, 0)$, but instead of iterating until convergence, perform 0 iterations. From this fit, we can get $\mathbf{U}_{wee}(\hat{\boldsymbol{\theta}}_0, 0)$ as well as $\{\text{Var}[(\hat{\boldsymbol{\theta}}, \hat{\beta})]\}_{\theta=\hat{\theta}_0,\beta=0}$ in (13). Finally, one ignores the design and assumes all subjects are independent with weights w_i; under these assumptions, we fit a proportional odds model under the alternative $H_0: \beta \neq 0$ with starting values $(\hat{\boldsymbol{\theta}}_0, 0)$, and perform 0 iterations; this gives us (14). The SAS Proc Surveylogistic code for the example analyzed in Section 4 can be found at http://www.blackwellpublishing.com/rss/SeriesC1.htm. As an alternative to the score statistic, one can also use a Wald statistic (estimate of β divided by its estimated standard error) to test

H_0:$\beta = 0$. The Wald and score statistics have identical large sample properties under the null, but have different properties under the alternative, with the score statistic typically being more powerful (Hauck and Donner, 1977).

4 Application: MEPS Study

In this section, we present results for analyses of data from the MEPS example discussed in the Introduction. There are 203 strata in this study, and 2 or 3 clusters were sampled without replacement within each stratum. Although the sampling was performed without replacement in each stratum, the total (population) number of clusters within each stratum was so large that the finite population correction factor can be ignored. The weights used in the analysis are the (Horvitz-Thompson) survey weights provided by MEPS, so that the weights sum to the population total. These weights account for unit nonresponse. Data on 25 subjects from this dataset are given in Table 1. Our goal is to explore bivariate analyses between health insurance status (yes, no) and the ordinal variables: education level (no degree, high school graduate equivalency degree–ged, high school diploma, bachelor's degree, master's degree, doctorate degree), income (poor, near-poor, low income, middle income, high income), and perceived health status (excellent, very good, good, fair, poor), and BMI (underweight, normal, overweight, and obese). In this dataset, 5401 (21.2%) of 25388 subjects did not have health insurance, although the weighted percentage of subjects without health insurance was 17.2%. Table 2 gives the weighted column percentages given the subject has or does not have health insurance, as well as the usual Wilcoxon test (the proportional odds score test) ignoring the complex survey design (including stratification, clustering, and weighting), and our proposed proportional odds score test which takes the complex survey design into account (and uses a chi-square approximation with 1 degrees-of-freedom).

We see from Table 2 that, as expected, the usual Wilcoxon test (proportional odds score test) ignoring the complex survey design and the proportional odds score test taking the design into account are quite different in value. For education and income, the proportional odds score test statistics taking the design into account are almost half the size of those that do not take the design into account, albeit all are very significant. On the other hand, for perceived health status and BMI, we see that the opposite is true; the test statistics taking the design into account are much larger than those that do not. In fact, for BMI, the test statistic taking the design into account is borderline significant (P=0.070), whereas the test statistics not taking the design into account is far from significant (p=0.472). (The SAS Proc Surveylogistic code for the example is given in an Appendix posted on the Web). We note here that if one used the Wald statistic (which is printed out directly in SAS Proc Surveylogistic) instead of the score statistic, one obtains P-values that are very similar to the score statistic: education level ($P < .0001$), income ($P < .0001$), perceived health status ($P = 0.30$), and BMI ($P = 0.067$); however, this very close correspondence between the Wald and score test cannot be expected in general. The results of analyses of the MEPS data indicate that failure to incorporate the design in the analysis can potentially yield misleading inferences about the associations.

From Table 2, we see that patients without health insurance are less educated and poorer, and have slightly lower BMI. There does not appear to be any difference in perceived health status between patients with or without health insurance. With the large number of subjects sampled in complex surveys, we usually have high power to detect small differences. We see this is the case for BMI, in that patients with health insurance are approximately 2% more likely (in absolute terms) to be overweight or obese.

In order to explore which of the design factors (stratification, clustering, weighting) has the largest impact on the P-values, we re-calculated the proposed test statistic under three scenarios. In the first scenario, we ignored the strata, and just assumed that the design is a cluster sampling design in which we randomly sampled 448 clusters from the population; we also included the weights in this analysis. In the second scenario, we ignored the clusters, and just assumed that the design is a stratified sampling design; we again also included the weights in the analysis. Finally, in the third scenario, we kept stratification and clustering as in the design, but we ignored the weights (set all weights equal to 1). Overall, the average of the weights is 8903.6, with standard deviation 5446.2, and range 386.9 to 49958.0. Across the 203 strata, the mean of the weights range from 2441.6 to 15711.4, and the standard deviations range from 1436.0 to 8750.0. Given this variability in the weights across strata, we might expect the weights to play an important role in calculating the test statistics. The results under these three scenarios are given in Table 3.

In general, stratification tends to decrease the variance, and we see that the values of the test statistics tend to be smaller (due to the larger variance) when stratification is ignored. Clustering tends to increases the variance, and we see that the values of the test statistics tend to be larger (due to the smaller variance) when clustering is ignored. Finally, ignoring the weighting does not yield a clear pattern in that the value of two of the test statistics (for education and income) are larger, and the value of two of the test statistics (for health and BMI) are smaller. Based on the pattern of results in Table 3, for this particular example it appears that the relative importance of stratification, clustering and weighting differs for the four ordinal variables. For example, for perceived health status and BMI, clustering and weighting are the most important design factors to take account of in the analysis; stratification has little impact on the test statistics. However, for income, stratification and clustering appear to be the most important design factors. This differential relative importance of the three design factors for the ordinal outcomes explains in part why the test statistics taking the design into account can be either larger or smaller than those that do not.

5 Conclusion

In summary, we propose an extension of the Wilcoxon rank-sum test to complex survey data. The approach is not ad hoc, but is based on the connection between the Wilcoxon rank-sum test and the proportional odds score test for the group effect. Our proposed test statistic uses an estimating equations score statistic (Rao et al., 1998) for no group effect in a proportional odds logistic regression model.

By formulating the test statistic in terms of a score test statistic from the proportional odds model for complex survey data, one can apply theory developed for estimating equations score tests without having to develop new theory for ranks in complex survey data. The huge increase in use of population-based complex sample surveys has led to analyses that have been published in leading medical journals, yet no simple approach has been developed to test for association between group and an ordinal categorical variable. This paper provides such an approach that can be used for any complex survey design.

One issue that may arise is the maximum number of outcome categories (J) allowed for our proposed test to be approximately chi-square with one degree-of-freedom under the null. Considering the data as arising from a ($2 \times J$) contingency table (Reynolds, 1984), for multinomial samples there should be at least $5 \cdot 2 \cdot J$ observations in the dataset, e.g. $n \geq 5 \cdot 2 \cdot J$. If we let $DEFF$ represent the design effect for a complex sample survey (Korn and Graubard, 1999), where $DEFF$ represents the ratio of the variance of (15) under the given design to a multinomial sample, then $n \geq DEFF \cdot 10 \cdot J$, or, equivalently, $J \leq n/(10 \cdot DEFF)$. Typically, design effects for complex surveys are less than 2 (Heeringa and Liu, 2006),

giving the rule-of-thumb that our proposed method can be applied provided $J \leq n/20$. For a typical large complex survey dataset, with say, 10,000 subjects, this means we could have an ordered categorical variable with 500 levels.

Based on this paper, in future, researchers analyzing complex samples could use our approach to formulate a non-parametric test with sample survey data. The same approach could be used to formulate Wilcoxon-type test statistics in other directions, such as adjusting for covariates and missing data.

Acknowledgments

We are grateful for the support provided by grants from the United States National Institutes of Health.

References

Agresti, A. Analysis of Ordinal Categorical Data. 2. New York: Wiley; 2010.

Chambers, RL.; Skinner, CJ. Analysis of Survey Data. New York: Wiley; 2003.

Conover, WJ. Practical Nonparametric Statistics. 3. New York: Wiley; 1998.

Binder D. On the variance of asymptotically normal estimators from complex surveys. International Statistical Review. 1983; 51:279–292.

Bross IDJ. How to use ridit analysis. Biometrics. 1958; 14:18–38.

Cohen SB. Design strategies and innovations in the Medical Expenditure Panel Survey. Medical Care. 2003; 41:5–12.

Datta S, Satten GA. Rank-sum tests for clustered data. Journal of the American Statistical Association. 2005; 100:908–915.

Hauck WW, Donner A. Wald's test as applied to hypotheses in logit analysis. Journal of the American Statistical Association. 1977; 72:851–853.

Heeringa SG, Liu J. Complex sample design effects and inference for mental health survey data. International Journal of Methods in Psychiatric Research. 2006; 7:56–65.

Jung S, Kang S. Tests for 2 × K contingency tables with clustered ordered categorical data. 2001; 20:785–794.

Bibbins-Domingo K, Coxson P, Pletcher MJ, Lightwood J, Goldman L. Adolescent overweight and future adult coronary heart disease. N Engl J Med. 2007; 357:2371–2379. [PubMed: 18057339]

Korn, EL.; Graubard, BI. Analysis of Health Surveys. New York: John Wiley; 1999.

McCullagh P. Regression models for ordinal data (with discussion). JRSS B. 1980; 42:109–142.

McCullagh, P.; Nelder, JA. Generalized Linear Models. 2. London: Chapman and Hall; 1989.

Pfeffermann D. The role of sampling weights when modeling survey data. International Statistical Review. 1993; 61:317–337.

Rao JNK, Scott AJ, Skinner CJ. Quasi-score tests with survey data. Statistica Sinica. 1998; 8:1059–1070.

Reilly JJ, Dorosty AR. Epidemic of obesity in UK children. The Lancet. 1999; 354:1874–1875.

Reynolds, HT. Analysis of Nominal Data. 2. Beverly Hills, Calif: Sage Publications; 1984.

Rosner B, Glynn RJ, Lee ML. Incorporation of clustering effects for the Wilcoxon rank sum test: a large-sample approach. Biometrics. 2003; 59:1089–1098. [PubMed: 14969489]

Shah, BV.; Barnwell, BG., et al. SUDAAN User's Manual. Release 8.0. Research Triangle Park, NC: Research Triangle Institute; 2001.

Appendix–Derivation of Wilcoxon test from proportional odds score test

In (8), we can further rewrite $(\hat{\gamma}_j + \hat{\gamma}_{j-1})$ in terms of the 'average rank' for category j. First, note that $n\hat{\gamma}_{j-1}$ is the number of subjects with outcome less than j; also, let $n_j = \sum_{i=1}^{n} Y_{ij} = n\widehat{p}_j$ denote the number of subjects with outcome j. Then, the average rank for the n_j subjects in

category j equals the average of the numbers $n\hat{\gamma}_{j-1} + 1$ to $n\hat{\gamma}_{j-1} + n_j$. Then, using the formula for sums given in Conover (1998), the average rank for the n_j subjects in category j equals

$$\begin{aligned} R_j &= n_j^{-1} \sum_{m=n\widehat{\gamma}_{j-1}+1}^{n\widehat{\gamma}_{j-1}+n_j} m \\ &= n_j^{-1}(n\widehat{\gamma}_{j-1}+n_j+n\widehat{\gamma}_{j-1}+1)n_j/2 \\ &= (n\widehat{\gamma}_{j-1}+n_j+n\widehat{\gamma}_{j-1}+1)/2 \\ &= (n\widehat{\gamma}_{j-1}+n\widehat{p}_j+n\widehat{\gamma}_{j-1}+1)/2 \\ &= n(\widehat{\gamma}_j+\widehat{\gamma}_{j-1})/2+0.5, \end{aligned} \quad (16)$$

since $n\hat{\gamma}_{j-1} + n\hat{p}_j = n\hat{\gamma}_j$. Then, in terms of the average rank R_j, we can rewrite $(\hat{\gamma}_j + \hat{\gamma}_{j-1})$ as

$$(\widehat{\gamma}_j+\widehat{\gamma}_{j-1}) = 2R_j/n - 1/n. \quad (17)$$

Then, inserting (17) in (8),

$$U_0 = \sum_{i=1}^{n} x_i \sum_{j=1}^{J} .5(2R_j/n - 1/n)(y_{ij} - \widehat{p}_j) = 1/n \sum_{i=1}^{n} x_i \sum_{j=1}^{J} R_j(y_{ij} - \widehat{p}_j).$$

Since the factor $1/n$ does not affect the score statistic in (5), without loss of generality, we write the numerator of the score statistic as

$$U_0 = \sum_{i=1}^{n} x_i \sum_{j=1}^{J} R_j(y_{ij} - \widehat{p}_j) = \sum_{j=1}^{J} R_j \sum_{i=1}^{n} y_{ij}x_i - \sum_{j=1}^{J} R_j\widehat{p}_j \sum_{i=1}^{n} x_i.$$

Table 1

Example Data on 25 subjects from MEPS study

Subject	Stratum	Cluster	Sampling Weight	Health Insurance	Eduction	Income	perceived health status	BMI
1	1	1	7080.48	yes	Bachelor's	Middle	Good	normal
2	1	2	4714.22	yes	No Degree	High	Good	normal
3	2	2	6925.06	yes	High School	High	Excellent	obese
4	3	2	9358.85	yes	No Degree	High	Very Good	over
5	4	1	6081.79	no	No Degree	Middle	Good	normal
6	4	2	3728.20	no	High School	Poor	Very Good	normal
7	5	1	4056.79	no	High School	Middle	Good	over
8	6	1	5936.66	yes	Master's	High	Excellent	over
9	7	2	2871.62	no	Bachelor's	High	Good	normal
10	8	2	2671.22	yes	Doctorate	High	Very Good	obese
11	9	1	5101.48	yes	High School	Middle	Very Good	normal
12	10	1	3569.07	yes	High School	Poor	Poor	over
13	11	1	4751.75	yes	High School	Poor	Excellent	over
14	12	1	9790.85	yes	GED	Middle	Very Good	over
15	13	1	7168.04	yes	GED	High	Excellent	over
16	14	2	5762.49	yes	No Degree	High	Excellent	over
17	15	1	7382.55	yes	High School	Middle	Excellent	normal
18	15	1	10140.54	no	No Degree	Middle	Excellent	under
19	16	1	4952.08	yes	High School	High	Good	normal
20	17	1	6989.89	no	No Degree	High	Excellent	over
21	18	1	2649.72	yes	GED	High	Very Good	obese
22	19	2	3363.35	yes	High School	High	Very Good	under
23	20	2	5425.54	yes	High School	Middle	Fair	normal
24	21	2	9417.92	no	High School	Low	Excellent	over
25	22	1	2017.34	no	No Degree	Middle	Very Good	obese

J R Stat Soc Ser C Appl Stat. Author manuscript; available in PMC 2013 July 31.

Table 2

Weighted column percentages and Wilcoxon test statistics for MEPS data.

		Health Insurance			
Variable	**Levels**	**No**	**Yes**	**Ignoring Design Wilcoxon (Propotional-odds) X^2(P-value)**	**Complex-survey Propotional-odds X^2(P-value)**
Education	No Degree	31.3	17.9	959.81(< .0001)	355.10(< .0001)
	GED	7.3	4.2		
	High School	49.3	49.5		
	Bachelor's	9.7	18.8		
	Master's	2.0	7.7		
	Doctorate	0.5	2.0		
Income	Poor	21.0	7.8	1933.38(< .0001)	626.24(< .0001)
	Near-poor	7.4	3.1		
	Low	22.5	10.7		
	Middle	30.8	31.0		
	High	18.3	47.5		
Perceived Health Status	Excellent	26.0	25.8	0.03(0.864)	1.06(0.30)
	Very Good	31.4	34.6		
	Good	30.6	26.6		
	Fair	9.4	9.5		
	Poor	2.6	3.5		
BMI	Under	2.7	2.0	0.52(0.472)	3.28(0.070)
	Normal	38.7	37.7		
	Over	34.8	35.7		
	Obese	23.8	24.6		

Table 3

Wilcoxon test statistics for MEPS data accounting for different features of the survey design.

Variable	Scenario	X^2	P-value
Education	Ignoring Design	959.81	< .0001
	Ignoring Strata	267.28	< .0001
	Ignoring Clusters	686.69	< .0001
	Ignoring Weights	467.91	< .0001
	Incorporating Design	355.10	< .0001
Income	Ignoring Design	1933.38	< .0001
	Ignoring Strata	400.79	< .0001
	Ignoring Clusters	1347.41	< .0001
	Ignoring Weights	632.16	< .0001
	Incorporating Design	626.24	< .0001
Perceived Health Status	Ignoring Design	0.03	0.864
	Ignoring Strata	0.99	0.319
	Ignoring Clusters	1.33	0.249
	Ignoring Weights	0.03	0.872
	Incorporating Design	1.06	0.301
BMI	Ignoring Design	0.52	0.472
	Ignoring Strata	3.28	0.070
	Ignoring Clusters	3.89	0.048
	Ignoring Weights	0.40	0.525
	Incorporating Design	3.28	0.070