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

Stuart Lipsitz
Brigham and Women's Hospital
Boston, MA

Outline

- Background for Complex Sample survey
- Extension of Wilcoxon Rank-Sum for Complex Survey Data
- Example

Complex Sample Surveys

- Complex survey sampling is often used to sample a fraction of a large finite population.

- In general, each sampling unit has a different probability of being selected into the sample.

- For generalizability to popultion, both design and the probability of being must be incorporated into the analysis.

- analyses of ready availability of public-use data from large population-based complex sample surveys has led to: **newly discovered important associations between risk factors and disease**

- Many seminal papers published in leading medical journals have used such complex sample survey data.
- **Paper:** Epidemic of obesity in UK children
  Journal: The Lancet (Reilly and Dorosty, 1999)
  Survey: Health Survey for England (HSE)
- **Paper:** Adolescent Overweight and Future Adult Coronary Heart Disease
  Journal: New England Journal of Medicine (Bibbins-Domingo et al., 2007)
  Survey: US National Health and Nutrition Examination Surveys (NHANES)
- A search of PubMed (National Library of Medicine) abstracts using the word "NHANES" yielded 7699 articles in the last 5 years
- And NHANES is just one of at least 100 complex surveys.

- Usually, reporting of regression analyses is the main goal, but initial summaries in terms of bivariate analyses are regularly reported in ‘Table 1’ in a medical paper.

- Wilcoxon rank-sum test is one of the most frequently used statistical tests for comparing an ordinal outcomes between two groups, and are often used in ‘Table 1’.

- Unfortunately, no simple extension of the Wilcoxon rank sum test has been proposed for complex survey data.

- The mutli-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.

- Extensions of the rank-sum tests have been proposed for clustered data (Jung and Kang, 2001; Rosner, Glynn, and Lee, 2003), without stratification or unequal selection probabilities.

- With independent subjects,

  Wilcoxon rank-sum test statistic=score test statistic for a group effect from a proportional-odds cumulative logistic regression model (McCullagh, 1989; Agresti, 2002)

- Using this framework, for complex survey data,

  1. we propose formulating a similar proportional-odds cumulative logistic regression model for the ordinal variable

  2. using an estimating equations score statistic for no group effect as an extension of the Wilcoxon test.

MEPS DATA

- Example: 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.
- Designed to produce national estimates of the health care use, expenditures, sources of payment, and insurance coverage of the United States civilian noninstitutionalized population.

- MEPS is a stratified, multistage probability cluster sample.
- 203 geographical regions form the strata .
- Two or three clusters (area segments) were sampled within each stratum.

- By design, each subject in the population has a known probability π_i of being sampled
- Over-sampled

  -Hispanics, African-Americans,

  -adults with functional impairments,

  -children with limitations in activities

  -individuals predicted to incur high levels of medical expenditures

  -low income individuals.

- Each subject in sample has known weight '$w_i = 1/\pi_i$'

- Because of the complex sampling frame utilized in these surveys, must use design-based analyses that incorporate the weighting, stratification, and clustering variables.

- We analyze data from 25,388 subjects who participated in the Household Component of the MEPS.

- Goal: See if people with and without health insurance differ in the ordinal variables

- Eduction (1=no degree, 2=ged, 3=high school diploma,4=bachelor's degree, 5=master's degree, 6=doctorate degree)

- Income (1=Poor, 2=Near-poor, 3=Low income, 4=Middle income, 5=High income)

- Perceived health status (1=Excellent, 2=Very Good, 3=Good, 4=Fair, 5=Poor)
- BMI
  - 1=underweight, BMI < 18.5 kg/m^2
  - 2=normal, BMI: 18.5 to 24.9 kg/m^2
  - 3=overweight, BMI = 25.0 to 29.9 kg/m^2
  - 4= obese, BMI > 30.0 kg/m^2
- Want to use Wilcoxon test, but incorporate the weighting, stratification, and clustering variables.
- Table 1 show fake data from 25 typical subjects, including strata, cluster, and weights

Table 1. Example (Fake) Data on 25 subjects from MEPS study

Subject	Strata	Cluster	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

Weights rescaled so that their sum=population=226,043,351
Weight for indiviudal= # of people in popultation one person represents.

Table 1 Example (Column Percent, Ignoring Design)

Variable	Levels	Health Insurance No	Health Insurance Yes	Wilcoxon X^2(P-value)
Education				959.81(< .0001)
	No Degree	31.3	17.9	
	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				1933.38(< .0001)
	Poor	21.0	7.8	
	Near-poor	7.4	3.1	
	Low	22.5	10.7	
	Middle	30.8	31.0	
	High	18.3	47.5	
Perceived Health Status				0.03(0.864)
	Excellent	26.0	25.8	
	Very Good	31.4	34.6	
	Good	30.6	26.6	
	Fair	9.4	9.5	
	Poor	2.6	3.5	
BMI				0.52(0.472)
	Under	2.7	2.0	
	Normal	38.7	37.7	
	Over	34.8	35.7	
	Obese	23.8	24.6	

Aside: Estimated $38,929,595/226,043,351 = 17.2\%$ US Citizen's without health insurance in 2002.

$95\% CI : (16.4\%, 18.0\%)$

Wilcoxon Rank-Sum Test = Score test from Proportional odds model

- First, consider typical sampling scheme of n indedendent subjects ($i = 1, ..., n$)

- Ordinal discrete random variable, Y_i
- Without loss of generality, assume Y_i takes on positive integer values $j = 1, 2, , ..., J$.

- Form J indicator random variables Y_{ij}, where

  $Y_{ij} = 1$ if subject i has response j

  $Y_{ij} = 0$ if otherwise.

- Goal; Determine if this ordinal outcome differs across two groups

- 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.

- Denote the probability of response j given x_i as

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

- Multinomial probability mass function for subject i equals

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

- Proportional odds model can be written as

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

- γ_{ij} is a 'cumulative probability'.

- Since

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

- Likelihood for subject i can be rewritten as

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

- Our main interest is in testing for no group effect, i.e.,

$$H_0: \beta = 0 \ .$$

- Under this null hypothesis the distribution of the ordinal variable is identical in the two groups.

- The Wilcoxon rank-sum test statistic can be shown to equals score test statistic for testing $\beta = 0$. (McCullagh, 1980)

- Briefly discuss score test

General Score test

- General form of a score test statistic for testing $\beta = 0$.

$$X^2 = \mathbf{U}(\widehat{\boldsymbol{\theta}}_0, 0)' \{\text{Var}[\mathbf{U}(\boldsymbol{\theta}, \beta)]\}^{-1}_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_0, \beta=0} \mathbf{U}(\widehat{\boldsymbol{\theta}}_0, 0),$$

where

- $\mathbf{U}(\widehat{\boldsymbol{\theta}}, \beta)$ is the score vector

- $\widehat{\boldsymbol{\theta}}_0$ is MLE of $\boldsymbol{\theta}$ under $H_0 : \beta = 0$
- $\mathbf{U}(\widehat{\boldsymbol{\theta}}_0, 0)$ is the score vector evaluated at $(\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_0, \beta = 0)$

- $\{Var[\mathbf{U}(\boldsymbol{\theta}, \beta)]\}_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_0, \beta=0}$ is the variance of $\mathbf{U}(\boldsymbol{\theta}, \beta)$ evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_0, \beta = 0$.

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

Proportional Odds Model

- For the proportional odds model (McCullagh, 1980), the only non-zero component of $\mathbf{U}(\widehat{\boldsymbol{\theta}}_0, 0)$ is

$$U_0 = \sum_{i=1}^{n} x_i \sum_{j=1}^{J} S_j(Y_{ij} - \hat{p}_j),$$

where

$$\hat{p}_j = n^{-1} \sum_{i=1}^{n} Y_{ij}$$

is the proportion of subject with response level j, regardless of group, and

$$S_j = \hat{p}_j/2 + \sum_{k=1}^{j-1} \hat{p}_k$$

which is the 'ridit' score.

- Note that $n \cdot S_j$ equals the average rank for a subject with response level j,
- This test statistic is identical to the Wilcoxon rank-sum statistic, which sums, for group $x_i = 1$

(average rank in category j $\times$ the number of subjects in category j)

across all categories

- 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 proportional odds models in the complex sample survey setting,
- without having to develop new theory for ranks in complex survey data.

Extension of the Wilcoxon Rank-Sum Test for complex survey data

- First, we discuss weighted estimating equations (WEE) for estimating $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}})$ in complex surveys.
- In complex sample surveys, target popultation is usually thought to be of finite size N, where N is often so large that for practical purposes the population is infinite.
- Assume the sample is still of size n (out of population N)
- To indicate which n subjects are sampled from 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 $i = 1, ..., N$,

- with $\Sigma_{i=1}^{N} \delta_i = n$.
- Depending on the sampling design, some of the δ_i could be correlated (e.g., for two subjects within the same cluster).

- As before, let π_i equal the (known by design) probability of being selected into the survey.
- Depending on the sampling design, π_i may depend on the outcome of interest, the independent variables, or additional variables (screening variables, for example) not in the model of interest.
- For a simple random sample (SRS), $\pi_i = n/N$ is a constant.
- Assume that the proportional odds model holds for all subjects in the population
- To obtain a consistent estimate of $(\boldsymbol{\theta}, \beta)$, one can use a weighted estimating equation, which is the solution to

$$\mathbf{U}_{wee}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}}) = \mathbf{0}$$

where

$$\mathbf{U}_{wee}(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}}) = \sum_{i=1}^{N} \frac{\delta_i}{\pi_i} \sum_{j=1}^{J} \frac{d}{d(\boldsymbol{\theta}, \beta)} [y_{ij} \log(\gamma_{ij} - \gamma_{i,j-1})]$$

- Here, the 'weights' are $w_i = \frac{\delta_i}{\pi_i}$. ($w_i = \frac{1}{\pi_i}$ if sampled $\delta_i = 1$).
- weighted likelihood score equations under (GEE) working 'independence' of subjects.

Properties of WEE

- $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}})$ has an asymptotic multivariate normal distribution with mean $(\boldsymbol{\theta}, \beta)$ and sandwich covariance matrix

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

- Note, $\{\mathrm{Var}[\mathbf{U}_{wee}(\boldsymbol{\theta}, \beta)]\}$ depends on the sample design (stratification and clustering).

- Empirically, $\mathrm{Var}[(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}})]$ is estimated via 'sandwich variance estimator' found in sample survey programs in SAS, Sudaan, R, and Stata.

Estimating Equations Score test

- We apply an estimating equations score test statistic (Rotnitzky and Jewell, 1990) for the null hypothesis of H_0:$\beta = 0$, in the proportional odds model.

- Here, let $\widehat{\boldsymbol{\theta}}_0$ denote the WEE estimate of $\boldsymbol{\theta}$ under the null hypothesis that $\beta = 0$.

- Similar to the usual score test, the estimating equations score test statistic for H_0:$\beta = 0$ is

  $$X^2 = \mathbf{U}_{wee}(\widehat{\boldsymbol{\theta}}_0, 0)'\{\text{Var}[\mathbf{U}_{wee}(\boldsymbol{\theta}, \beta)]\}^{-1}_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_0,\beta=0}\mathbf{U}_{wee}(\widehat{\boldsymbol{\theta}}_0, 0),$$

  where the form of $\mathbf{U}_{wee}(\widehat{\boldsymbol{\theta}}_0, 0)$ and $\{\text{Var}[\mathbf{U}_{wee}(\boldsymbol{\theta}, \beta)]\}_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_0,\beta=0}$ are both dervied under the alternative, but evaluated at $(\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_0, \beta = 0)$.

- In particular, sandwich

  $$\text{Var}[\mathbf{U}_{wee}(\boldsymbol{\theta}, \beta)]\}_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}_0,\beta=0} =$$

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

- Using central limit theorem for complex surveys (Binder, 1983), X^2 asymptotically chi-square distribution 1 degree-of-freedom under null

- although the definition of 'asymptotic' is sometimes non-standard if the finite population size N is small.

- Similar to the score test for non-complex survey data, the only non-zero component of $\mathbf{U}(\widehat{\boldsymbol{\theta}}_0, 0)$ is

$$U_0 = \sum_{i=1}^{N} w_i \sum_{j=1}^{J} x_i S_j (Y_{ij} - \hat{p}_j),$$

  where

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

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

$$S_j = \hat{p}_j/2 + \sum_{k=1}^{j-1} \hat{p}_k \ .$$

  which is a weighted 'ridit' score.

- 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 printed out, and requires a simple two step procedure.

Application: MEPS study

- Goal: See if people with and without health insurance differ in the ordinal variables

- Eduction (1=no degree, 2=ged, 3=high school diploma,4=bachelor's degree, 5=master's degree, 6=doctorate degree)

- Income (1=Poor, 2=Near-poor, 3=Low income, 4=Middle income, 5=High income)

- Perceived health status (1=Excellent, 2=Very Good, 3=Good, 4=Fair, 5=Poor)
- BMI
  - 1=underweight, BMI < 18.5 kg/m^2
  - 2=normal, BMI: 18.5 to 24.9 kg/m^2
  - 3=overweight, BMI: 25.0 to 29.9 kg/m^2
  - 4= obese, BMI > 30.0 kg/m^2

Results (weighted proportions)

Variable	Levels	Health Insurance No	Health Insurance Yes	Ignoring Design Wilcoxon (Propotional-odds) X^2(P-value)	Complex-survey Propotional-odds X^2(P-value)
Education				959.81(< .0001)	448.41(< .0001)
	No Degree	31.3	17.9		
	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				1933.38(< .0001)	982.84(< .0001)
	Poor	21.0	7.8		
	Near-poor	7.4	3.1		
	Low	22.5	10.7		
	Middle	30.8	31.0		
	High	18.3	47.5		
Perceived Health Status				0.03(0.864)	1.03(0.31)
	Excellent	26.0	25.8		
	Very Good	31.4	34.6		
	Good	30.6	26.6		
	Fair	9.4	9.5		
	Poor	2.6	3.5		
BMI				0.52(0.472)	3.36(0.067)
	Under	2.7	2.0		
	Normal	38.7	37.7		
	Over	34.8	35.7		
	Obese	23.8	24.6		

Results

- X^2 quite different depending whether design taken into account

- For education and income, X^2 taking the design into account almost half the size, albeit all are very significant.

- On the other hand, for Perceived Health Status and BMI, we see that the opposite is true, taking the design into account gives much larger X^2.

- for BMI, X^2 is borderline significant (P=0.067) using design, whereas not close to significance without design (p=0.472).

- The results of the analyses of the MEPS data indicate that failure to incorporate the design in the analysis can potentially yield misleading inferences about the associations.

Conclusion

- In summary, we have proposed an extension of the Wilcoxon Rank-Sum test to complex survey data.

- The approach is not ad hoc, but is based on the connection betwen the Wilcoxon rank-Sum test and the Proportional odds score test for a group effect.

- Based on estimating equations score statistic, no need to develop complicated probabiliy theory for ranks.

- Could 'extend' in other directions like adjusting for covariates, missing data.

- Will it work for continuous outcomes (instead ordinal) ?

- I think you can go through the theory to show that the test will be chi-square 1 under null, except it might test the edge of computing power.

- If BMI was continuous for example, with no ties, we have 25,000 intercepts in the proportional odds model, and, at least in sas, you run out of 'computer memory'