Variance Estimation for Ordered Categories

Steven Pedlow[1], Yongyi Wang[1], Kirk Wolter[1]
NORC at the University of Chicago[1]

Abstract

The National Longitudinal Survey of Youth, sponsored by the U.S. Department of Labor, examines youth transitions from school to work. Approximately 9,000 youths born in the years 1980 through 1984 were interviewed in 1997 and are still being interviewed annually. Job classification information has been collected and the sum of the top five job categories of youths for various subgroups has been reported. We consider this top five variable to be a measure of concentration; i.e., how concentrated the youths are within the top five job categories. Since the ranking of job categories is subject to sampling error, calculating variances for the sum of the top five is an interesting problem. The variance in the measure of concentration is not equivalent to the variance of being in the top five job categories actually observed. This paper compares several different methods of variance estimation for this measure of concentration, with the conclusions that Taylor series methods cannot be used, and that the Jackknife Repeated Replication method is superior to the Balanced Half-Sampling method.

Keywords: Measures of Concentration, Taylor Series, Jackknife, JRR, Balanced Repeated Replication, BRR, Balanced Half Samples, BHS

1. Introduction

The National Longitudinal Survey of Youth (NLSY97) is the latest in a series of surveys sponsored by the Bureau of Labor Statistics (BLS) at the U.S. Department of Labor to examine issues surrounding youth entry into the work force and subsequent transitions in and out of the work force. The NLSY97 is following a cohort of approximately 9,000 youths who completed an interview in 1997 (the base year). These youths were between 12 and 16 years of age as of December 31, 1996, and are being interviewed annually using a mix of some core questions asked annually and varying subject modules. The tenth round of interviews will be done over the course of the 2006-2007 school year.

The NLSY97 is not a simple random sample. The NLSY97 sample design involved the selection of two independent area-probability samples: 1) a cross-sectional sample designed to represent the various segments of the eligible population in their proper population proportions, and 2) a supplemental sample designed to produce, in the most statistically efficient way, the required oversamples of Hispanic and non-Hispanic, black youths.

Both the cross-sectional and supplemental samples were selected by standard area-probability sampling methods. Sampling was in three essential stages: primary sampling units (PSUs) consisting mainly of Census metropolitan statistical areas (MSAs) or single counties, segments consisting of single census blocks or clusters of neighboring blocks, and housing units (HUs). All eligible youths in each household were then selected for interviewing and testing.

Section 2 talks more about the measures of concentration that we studied. Section 3 sets forth the methods of variance estimation we used in our work that correct for the clustered sample design, and Section 4 contains the main results. We close with a brief summary in Section 5.

2. Measures of Concentration

The February 18, 2004 issue of BLS News (Bureau of Labor Statistics, 2004) concerns the employment of teenagers during the school year and summer. Tables presented in the issue contain estimated percentages of youths with jobs, but they omit estimates of the corresponding standard errors.

While some tables contain straightforward percentages, Tables 2 and 3 of the BLS News issue involve novel statistics on the concentration of youths in the top job categories. These tables list the top five job categories by age and gender. BLS Table 2 lists summer and school year job categories separately for enrolled youths. Exhibit 1 below shows a part of Table 2 from the BLS News issue. Exhibit 1 shows that among enrolled 17 year-old males, the most common summer job category was cook. This paper's focus is on the numbers on the bottom row. Comparing these two groups, the females are more concentrated (38.1 percent vs. 32.8 percent) in their top five job categories, which is due to their strong concentration of cashiers.

Exhibit 1. Excerpt from Table 2 of BLS News issue.

Males– Age 17		Females– Age 18	
SUMMER		SCHOOL YEAR	
	Pct.		Pct.
1. Cooks	8.6	1. Cashiers	15.6
2. Cashiers	8.3	2. Sales (Other)	7.0
3. Stock/Bagger	7.5	3. Waitresses	6.8
4. Sales (Other)	4.6	4. Office Clerks	4.6
5. Gardeners	3.9	5. Receptionists	4.1
TOTAL Top 5	**32.8**	TOTAL Top 5	**38.1**

BLS Table 3 lists the top five job categories for non-enrolled youths by high school graduation/dropout status. Each job category estimate, of course, has a standard error that can be straightforward to calculate. However, these tables also report the percentage of youths concentrated within the top five job categories (for each subgroup separately), which we define as a "measure of concentration." This variable answers the question, "How concentrated are youths within the top 5 job categories?"

These top five job categories are different for each subgroup and the ranks are themselves subject to sampling error. Standard errors for this concentration estimate would not be correct by simply calculating the standard error of the sum of the five binary categories (as fixed categories) that happened to be the top five for this particular (full) sample. A different sampling of youths across the United States could result in different job categories making the Top 5. This paper examines statistics about ordered categories and addresses the novel problems of variance estimation they represent by comparing four different methods of variance estimation.

3. Variance Estimation Techniques

Standard statistical packages such as SAS (SAS Institute, 1999) cannot calculate correct standard errors for a complex survey design such as for NLSY97 without special care (SAS has recently released special procedures that use Taylor series methods). These packages assume in their default standard error calculations that the data come from a simple random sample. Such standard errors tend to be too small for a cluster-sampling design. For illustration, we have included in our analyses the incorrect standard errors under the assumption of simple random sampling (calculated with SAS).

The two most popular classes in the calculation of correct variance estimates for complex survey designs are Taylor series (TS) methods and replication-based methods.

3.1 Taylor Series Method

TS methods work by approximating the estimator of the population parameter of interest by a linear function of the observations. These approximations rely on the validity of Taylor (or binomial) series expansions. An estimator of the variance of the approximation is then used as an estimator of the variance of the estimator itself. In this project, we used the software package SUDAAN (Research Triangle Institute, 2004) for our TS variance calculations.

The standard error of these job category concentrations cannot be estimated validly with the TS method because the method must falsely treat the job categories as fixed. Since the calculations cannot reflect the variability in the ranking of job categories, the variance calculations will be too small.

3.2 Replication-Based Methods

We explored two different replication-based methods in this research: balanced half-samples (BHS) and jackknife (J) estimation. Replication-based methods compute multiple estimates in a systematic way, and use the variability in these estimates to estimate the variance of the full-sample estimator. BHS was originally conceived for use when two primary sampling units (PSUs) are selected from each stratum. A half-sample is then one primary sampling unit from each stratum (with double weight). BHS uses an orthogonal set of half-samples as specified by a Hadamard matrix (Hall, 1967). The variability of the half-sample estimates is taken as an estimate of the variance of the full-sample estimator. In a similar fashion, the J method creates a series of replicate estimates by removing only one primary sampling unit from only one stratum at a time (doubling the weight for the other primary sampling unit). Complete definition of the TS, BHS, and J estimates appears in Wolter (1985).

The NLSY97 was not designed to have two primary sampling units per stratum. NLSY97 has certainty PSUs which act as their own strata, but all other PSUs were selected together. The sampling was done with systematic sampling (with probabilities proportional to size) on a sorted file, which means the PSUs are an approximately stratified sample of PSUs, but there were no discrete strata used, and there was no attempt to select exactly two PSUs from each stratum. Instead, we create pseudo-strata and pseudo-PSUs. To do this, we take the first-stage sampling units in order of selection and collapse two together into a pseudo-stratum. The first of these two PSUs is one pseudo-PSU while the second is the other. It is important to note that the NLSY97 has 19 certainty PSUs in the cross-sectional sample, and 17 certainty PSUs in the supplemental sample. Since these certainty PSUs don't involve sampling until the segment-level, we paired the segments within certainty PSUs and call each pair a pseudo-stratum and each segment within a pair a pseudo-PSU. Non-certainty PSUs are considered pseudo-PSUs, so two consecutive non-certainty PSUs comprise one pseudo-stratum. In all, we formed 323 pseudo-strata: 211 in the cross-sectional sample (171 related to certainty PSUs, 40 to non-certainty PSUs) and 112 in the supplemental sample (71 related to certainty PSUs, 41 to non-certainty PSUs).

We used the software package WESVAR (Brick et al., 2000) to carry out all of the replication-based estimates. We created 336 replicate weights for the BHS method using a 336 x 336 Hadamard matrix. The variability of the BHS estimates is the variance estimate. For the J method, we did not compute 2*323 = 646 replicate weights for the J method. For storage reasons, we combined three pseudo-strata together to make 323/3 = 108 pseudo-strata. We then created 2*108 = 216 replicate weights for the J method. The variance of the J estimates is the variance estimate divided by (k-1), where k is the number of jackknife estimates.

In creating replicate weights for the BHS and Jackknife methods, the standard process is to adjust the final sampling weight for each particular replicate. Essentially, we set to zero the weights for cases in the removed pseudo-PSU. The other pseudo-PSU within that stratum then has its weight doubled since it is now representing both pseudo-PSUs in the pseudo-stratum. The difference between the BHS and Jackknife methods is that this happens in all pseudo-strata for the BHS method, but happens in only one pseudo-stratum per Jackknife replicate.

However, the above method is a simplification of the true impact removing the pseudo-PSU(s) would have. Many steps in the weighting process would proceed differently (e.g., different non-response and other adjustments) if only those cases in the replicate were in the sample. Theoretically then, it may be superior to re-run the weighting algorithm separately for each replicate. We will refer to this as replicate re-weighting. We would expect the standard errors after replicate re-weighting to be larger because of the extra variability accounted for. As part of our work, we explored replicate re-weighting by comparing J and BHS standard errors with and without replicate re-weighting. We saw very little difference in the replicate re-weighted estimates, so this work was not presented, and is not shown here.

Replication methods can be used to estimate the variance of measures of concentration because the top five job categories can be determined (and summed) separately for each replicate. The separate replicate estimates can then be combined in the standard way to get standard error estimates.

3.3 A Simulation-Based Estimation Method

In addition to these two standard methods for these unconventional variables, we also used a new simulation method to estimate the standard errors for the job category concentrations in BLS Tables 2 and 3. In order to perform the simulation, we first calculated a standard error for all 301 job categories in the school year and all 311 summer job categories (BLS Table 2), and all 243 BLS Table 3 job categories (only the top five job categories are shown in BLS Tables 2 and 3) within each subgroup (e.g., enrolled 17-year-old males who had a summer job, non-enrolled 18-year-old female dropouts, etc.). We used SUDAAN to calculate these standard errors.

We then used the estimates and these standard errors to conduct our simulation, which had three steps (for each of the subgroups):

Step 1. For each of 1000 random draws, draw a realized percentage for each job category from a normal distribution with the BLS Table 2 or 3 estimate as the mean and the SUDAAN-calculated standard error as the standard deviation.

Step 2. Sort the job categories by their realized percentages and determine the top five.

Step 3. Add together the realized percentages for the top five for this random draw's estimate of the job category concentration.

This process resulted in 1000 estimates for the job category concentration. The 95-percent confidence interval then has as its lower bound the 25th smallest estimate, while its upper bound is the 976th smallest estimate (25th largest). For comparison with the BHS and Jackknife estimates, we translated this interval into a normal-theory estimate and standard error. To determine the estimate, we simply took the midpoint of the interval (the mean of the 25th smallest and 976th smallest estimates). To determine the standard error estimate, we assumed that this confidence interval was the mean plus or minus 1.96 standard, and divided the interval length by 2*1.96 = 3.92.

4. Results

4.1 Results for Means

We performed BHS, J, and Simulation calculations for the job category concentrations in BLS Tables 2 and 3 of the BLS News. As explained above, the only way to estimate the standard error for this sum by the Taylor series method would be to treat the top five categories as fixed. Note that for the BHS, Jackknife, and simulation method, the top five categories can change by replicate or simulate. We determine the top five job categories for each replicate or simulate and take the sum of them as the estimated concentration for that replicate or simulate. Exhibits 2 and 3 in this paper show the estimates and standard errors under all four of these methods. Conveniently, Exhibit 2 of our paper matches to BLS Table 2 from the BLS News. Similarly, Exhibit 3 of our paper matches to BLS Table 3 from the BLS News. The first three columns indicate the subgroup, the fourth gives the estimate reported by SAS. The next eight columns then give the estimates and standard errors for the four variance estimation methods.

Looking only at the original estimates column, two trends are clear. First, males are less concentrated than females of the same age. Second, with the exception of male dropouts, older youths are less concentrated than younger youths.

Comparing the different estimates produced by the variance estimation techniques, the Taylor series and Jackknife estimates are very close to the replicated estimate column, but the BHS and simulation estimates are biased upwards. For enrolled youths, Exhibit 2 shows that these biases increase for the older youths. The bias is 1 percent or less for 17-year-olds, 1-2 percent for 18-year-olds, and 2 percent to 5 percent for 19-year-olds. For non-enrolled youths, Exhibit 3 shows that the bias is also between 1 percent and 5 percent, but the pattern is less clear. What seems surprising is the close agreement between the BHS and simulation. In Exhibit 2, the difference between the two estimates is 0.1 percent or less for seven of the twelve estimates with a maximum difference of 0.5 percent; in Exhibit 3, the difference is 0.1 percent or less for five of the eight estimates with a maximum difference of 0.8 percent.

To explain the bias in the BHS method and the simulation, recall that for Jackknife, BHS, and the simulation, the percentage for each category will change and thus the top five could change. We hypothesize that when the fourth and/or fifth ranked job category are higher for a particular replicate/simulate, they will remain in the top five and make the concentration higher. However, when the fourth and/or fifth ranked job category are lower for a particular replicate/simulate, they might be replaced and thus the concentration will not go lower as much as it will go higher – this causes a bias. The bias is larger when the sixth and seventh, etc. top job categories are similar in size to the fourth and fifth categories. This tends to happen when the youths are less concentrated (smaller sums of the top five job categories). We further hypothesize that the Jackknife method does not suffer from this bias because each replicate has subtler changes (only one variance stratum is affected at a time) and so the top five job categories are more stable from replicate to replicate.

4.2 Results for Standard Errors

With regard to the standard errors, there is wide disagreement between the methods. The Jackknife estimates are always the highest with the BHS estimates always being the smallest. We had expected the Taylor series method estimates to be smaller because they fix the top five job categories, thus missing one component of the variance. However, the BHS estimates are actually smaller than the Taylor series method estimates. The simulation estimates were second-largest for seven of the twelve subgroups, but smaller than Taylor series estimates for the other five. We hypothesize that the same effect that results in the upward bias in estimates mentioned above also results in a downward bias for the standard errors by preventing the estimated concentration from being as far below the observed concentration as it should. In other words, the concentration parameter can be larger by the appropriate amount, but the amount it can be below the observed estimate is constrained by the replacement of job categories in the top five, which incorrectly shrinks the standard error estimate. This implies that the standard errors for the BHS and simulation methods are too small, and that the best standard error estimates are provided by the Jackknife method.

5. Summary

For measures of concentration, the Jackknife seems clearly superior for the concentration parameter. It is a well-known result that the Jackknife is incorrect for medians, so we wondered if the concentration parameter, which does have a relationship with order statistics, would have flawed Jackknife variance estimates as well. However, it is the BHS and Simulation methods that seem flawed for this concentration parameter. Both methods provide percentage estimates that are biased upwards while the standard error estimates are biased downwards. An alternative simulation method that might be worth exploration is the bootstrap method.

References

Brick, J. M., Morganstein, D., and Valliant, R. (2000). Analysis of Complex Sample Data Using Replication. http://www.westat.com/wesvar/techpapers/ACS-Replication.pdf

Bureau of Labor Statistics (2004). Employment of Teenagers During the School Year and Summer. BLS News, February 18, 2004. USDL 04-217. United States Department of Labor, Washington, D.C. Also Available at: http://www.bls.gov/nls/nlsy97r5.pdf

Hall, Jr., M. (1967). Combinatorial Theory. Waltham, Massachusetts: Blaisdell.

Research Triangle Institute (2004). SUDAAN Language Manual, Release 9.0. Research Triangle Park, NC: Research Triangle Institute.

SAS Institute, Inc. (1999). SAS/STAT User's Guide, Version 8. Cary, NC: SAS Institute Inc.

Wolter, K.M. (1985). *Introduction to Variance Estimation*, Springer-Verlag, New York.

Exhibit 2. Top Five Job categories of Enrolled Youths with an Employee Job During the 2000-01 School Year and the Following Summer (Percent Employed).

Job	Sex	Age	Original Estimates	Taylor Series		Jackknife		BHS		Simulation	
				Estimate	S.E	Estimate	S.E	Estimate	S.E	Estimate	S.E
School year	Male	17	36.1	36.1	2.55	36.1	3.32	37.0	2.39	37.1	2.84
Summer	Male	17	30.9	30.9	2.23	30.9	3.01	31.7	2.05	31.9	2.49
School year	Female	17	42.8	42.8	2.48	42.8	4.26	43.3	2.41	43.2	2.95
Summer	Female	17	36.4	36.4	2.45	36.4	3.14	37.1	2.19	37.4	2.45
School year	Male	18	26.0	26.0	2.55	26.0	2.92	27.3	2.21	27.3	2.52
Summer	Male	18	21.8	21.8	2.28	21.8	2.81	23.2	1.91	23.2	2.21
School year	Female	18	36.5	36.5	3.09	36.6	4.04	38.1	2.88	38.0	3.35
Summer	Female	18	35.1	35.1	3.07	35.2	4.29	36.5	2.84	36.5	3.20
School year	Male	19	21.6	21.6	3.73	21.6	5.91	25.3	2.65	25.7	3.16
Summer	Male	19	18.1	18.1	2.98	18.1	5.46	22.7	2.34	23.1	2.62
School year	Female	19	32.8	32.8	3.16	32.8	3.95	35.6	2.51	35.5	3.52
Summer	Female	19	30.3	30.3	3.10	30.3	4.52	32.4	2.66	32.9	3.17

Exhibit 3. Top Five Job categories of Non-Enrolled Youths with an Employee Job During the 2000-01 School Year and the Following Summer (Percent Employed).

Sex	Age	High school graduation status	Original Estimates	Taylor Series		Jackknife		BHS		Simulation	
				Estimate	S.E	Estimate	S.E	Estimate	S.E	Estimate	S.E
Male	18	dropouts	23.8	23.8	4.00	23.9	6.01	28.9	3.18	28.8	3.72
Male	18	graduates	20.7	20.7	2.51	20.7	3.34	23.9	2.00	23.9	2.38
Female	18	dropouts	47.9	47.9	5.41	48.0	6.64	52.1	4.69	51.3	6.04
Female	18	graduates	36.5	36.5	3.18	36.5	4.48	37.5	2.78	37.9	3.66
Male	19	dropouts	29.9	29.9	4.27	29.9	5.99	33.3	3.29	33.3	3.95
Male	19	graduates	15.8	15.8	2.05	15.8	2.91	18.0	1.49	18.3	1.74
Female	19	dropouts	41.2	41.2	4.91	41.2	7.58	45.8	5.09	45.7	6.07
Female	19	graduates	26.6	26.6	2.12	26.6	3.17	28.5	2.10	28.4	2.40