ALTERNATIVE SAMPLING STRATEGIES FROM THE NEW BLS BUSINESS ESTABLISHMENT LIST

Principe, Jacqueline and Sommers, John Paul
Jacqueline Principe, Bureau of Labor Statistics, Washington, D.C. 20212

Key Words: Sampling Unit, Stratified, Costs

Survey

The Current Employment Statisitcs Survey (CES) produced by the Bureau of Labor Statisitics (BLS) is the nation's primary source of data on total nonagricultural employment within various industries and areas of the country. The CES is based upon a sample of approximately 380,000 reporting units that are collected monthly to provide monthly estimates for over 250 areas within the country, state level estimates and national estimates. Along with other economic statistics, CES produces estimates of total employment by industry using the Standard Industrial Classification(SIC) codes that are published. Detail ranges from major industry division of one-digit SIC's down to more specialized classifications within industry of four-digit level SIC's. Estimates are made by summing estimates for a set of primary estimation cells. (For national estimates these are generally strata with 4-digit SICs.) Monthly estimates for each primary cell are made using a link relative estimator defined as follows:

Let EST_t be the estimate for a primary cell for month t, then $EST_t = R_{t,t-1} * EST_{t-1}$

where $R_{t,t-1}$ is the ratio of the total sample employment in month t to the total sample employment in month t-1 for all sample units reporting data for both months.

Frame

Samples are drawn from a frame developed from Unemployment Insurance (UI) reports submitted by almost all nonagricultural employers to the individual states. Since employers are required by law to submit regular reports of monthly employment, the UI serves as a frame as well as a benchmark used to measure error of CES estimates. As reports for the previous time period are finalized, estimates for that time period are compared with the actual values from the updated frame for the past time period. At the same time these "benchmark" estimates are used to correct past estimates and update current estimates by replacing past values of EST_t with actual values. This means that the CES estimates cannot build large amounts of error by continually linking month to month estimates for indefinite time periods. Because of the benchmarking and the large sample available, the CES has a history of producing very good estimates at the national level.

Occasionally weaknesses in the frame have hurt the effectiveness of the CES estimator. Reporters to the UI frame were only required to report to Washington on an annual basis for data collected quarterly. By the time data was processed, the frame was badly out-of-date. More importantly, employers reported data on a reporting unit basis, which could represent more than one worksite. This created several problems:

1. geographic subunits could be for areas other than those reported ,
2. the same could be true for industrial classes and
3. it was impossible to monitor the sample because the entities which made up the sample units could change, allowing non comparable units from month to month used to measure employment change. To alleviate these problems, employers produced separate worksite reports, on a voluntary basis, if they had more than 50 employees in secondary sites in a particular industrial category.

Frame Improvement

Beginning in the first quarter of 1989, BLS began to improve reporting with the addition of the Business Establishment List (BEL) to supplement the Universe Data Base (UDB). Starting in 1989, data was to be reported quarterly to BLS, rather than annually as before, allowing the frame to be available on a more timely basis. Secondly, firms with multiple sites were asked to report separately if they had more than 10 employees in secondary sites. Furthermore, if they did separate reports, it was done by individual worksite, the lowest level of detail. Samples could be selected and tracked from month to month for comparative establishments and be placed in the proper geographic areas.

This new level of detail gives BLS a new level of sample options and decisions much like those available in household surveys. Now the CES can select individual worksites, much as households could be selected, or blocks of worksites, companies, effectively leaving the CES with the option of selecting from various forms of cluster

sample. As with household surveys, to determine the most effective method one needs to understand the amount of variations within clusters and the relative costs and feasibility of collecting worksites with companies (clusters). As with household surveys, one would expect a simple random sample of individual worksites to be statistically more efficient than cluster sampling. However, costs of data collection could result in cluster sampling being the preferred method.

Study Overview

This paper describes initial efforts to determine statistical properties for a variety of sampling methods using the BEL. Other studies must complement this study to determine costs and methods to collect and track clustered data from companies versus costs to collect data from individual worksites. In this study we calculate the design effects due to stratification and cluster sampling under various schemes (equal costs for each worksite) and compare relative standard errors for the sampling plans with fixed and marginal collection costs. The study uses the full universe of data from the first 13 states with full BEL breakdown in the first quarter of 1990. They are: Minnesota, Colorado, Kansas, South Dakota, Utah, Wyoming, Virginia, West Virginia, Wisconsin, Iowa, Montana, North Dakota, Oregon. Implementation of multisite reporting was phased into production over a period of time and is now almost complete. Six two digit SIC's were selected to cover continuum of average number of sites per employer. The industry classes are: 15, general building contractors; 20, manufacturing of food and kindred products; 38, instruments and related products; 42, trucking and warehousing; 54, food stores; 60, depository institutions. SIC 42, trucking and warehousing, is an example of low multiple site industry and SIC 54, food stores, is an example of a high multiple site industry.

The sample size is the total current CES sample size in the above states and industries. Data for three months from these sample units were used to calculate relative standard errors for a two month change. Since the frame is built state by state, multiple site companies are defined as companies reporting within each state and UI number. Multiple site companies, treated as clusters, and individual sites were the sample units.

Comparisons Made

Comparisons are made for the following sample plans (all size stratifications were done separately within each industry),

1. Simple random sampling of worksite units with 1 to 5 size class strata, where stratification is by employment size in the worksite.
2. Simple random sampling of worksite units with 2 to 6 strata which are put together in a two step process. The first stage is to break the universe into single and multiple unit companies, the second is to further stratify each group into 1 to 3 strata by size of worksite.
3. Cluster sampling after worksites are separated into 1 to 5 size class strata and then grouped by UI number within each strata. For each stratification, clusters of grouped worksites, each containing M_i sites, are randomly selected, then one, two, three, half, and all the units within the selected clusters are sampled. (m_i= 1, 2, 3, 1/2M_i, M_i,) (m_i= (M_i+1)/2 if M_i is odd)
4. Cluster sampling of grouped worksites where stratification of 1 to 5 size strata is by total employment in each cluster of M_i worksites. All worksites within the randomly selected clusters are selected. (m_i=M_i)
5. Cluster sampling of grouped worksites with stratification as in (4) above and one, two, three, or half the units are selected from multiunit clusters. (m_i= 1, 2, 3, 1/2M_i)
6. Samples are selected from 2 to 6 strata which are put together in a two step process. The first stage is to separate the universe into single units and grouped multiple site units. Each group is further stratified into 1 to 3 strata by employment size of cluster. For each stratification, clusters of M_i sites are randomly selected then one, two, three, half, and all the units within multiunit companies are selected. (m_i= 1, 2, 3, 1/2M_i, M_i) (M_i=1 for single worksite companies)
7. The same as (6) above except stratification is done by average size of worksite within a company rather than grouped multiunit size.

For each of the sampling plans, various cost structures are assumed:

1. Cost of collecting all units within a company is \$1 without regard to the number of worksites collected,
2. Cost of a multiple unit company is \$2, \$3, \$4, or \$5,
3. Cost of collecting the first worksite within a cluster, or company, is \$1 and each additional unit has a fixed marginal cost: with fixed marginal costs of \$0.10, \$0.25, or \$0.50.

Total costs are assumed to be $7,352, one dollar for each unit of the current sample; N=7,352.

Optimal boundaries for the strata were determined by the cum√f method based on the population distribution of the stratification variable in the first month. The standard ratio estimator was used and the relative standard error of the estimates was calculated using the estimate of change over a two month period. Variances for each strata were calculated using the following formula from Cochran:

$$v(\hat{Y}_R) = \frac{N^2(1-f)}{n}\frac{\sum(Y_i-\hat{R}\hat{X}_i)^2}{n-1}$$

$$+\frac{N}{n}\sum_{i}^{n}\frac{M_i^2(1-\frac{m_i}{M_i})s^2_{d'2i}}{m_i} \quad (1)$$

where;

N= total number of units,

M_i= number of worksites in company i,

m_i= number of worksites selected from company i,

$\hat{R} = \overline{Y}/\overline{X}$ s.t.

$\overline{X}$ is the employment population mean of the first month,

$\overline{Y}$ is the employment population mean at the end of the second month,

$s^2_{d'2i} = \frac{1}{M_i-1}\sum_{i}^{M_i}[(y_{ij}-Rx_{ij})-(\overline{Y}_i-R\overline{X}_i)]^2$ is the within multiple unit variance.

The second term of equation (1), that calculates the within cluster variance, was assumed to be zero when simple random sampling was used or all sites from a company were selected. Allocation was done using standard optimal allocation given variable costs as shown in Cochran.

Design Effects

In this section we compare the design effects of a variety of sampling plans. Design effect here is defined as the variance of the employment estimate due to alternate sampling strategies divided by the variance due to simple random sampling from a population of the same size. We are using three strata to compare effects because observed relative standard errors showed little gain for more strata and operationally such plans would be preferable for BLS. To calculate these effects we calculated the variances under each plan for a sample which contain 7,352 worksites. Table 1 below gives the design effect ratio for each stratification and sampling plan:

Table 1. Design Effects For Comparison Of Alternative Sample Designs

	Collection method:					
Stratified by:	single wkste	group $m_i=M_i$	group $m_i=1$	group $m_i=2$	group $m_i=3$	group, $m_i=1/2M_i$
1. worksite, using SRS	1.00					
2. single/multi, then by wksite using SRS	0.95					
3. worksite then group multis, cluster sampling		1.68	4.04	2.68	2.19	1.39
4.,5. grouped multis using cluster sampling		1.69	4.12	2.88	2.73	5.55
6. sing/multi then by grped multis, cluster sampling		1.55	4.60	3.13	3.13	6.17
7.single/multi then by ave. size of wksite		1.80	4.76	4.34	4.93	4.72

As shown above, the only design that is an improvement in variance to worksite SRS is the two stage process of first stratifying on the condition of the unit being single or a member of a grouped unit and then size stratification of worksite within those two strata. One reason for this may be that there are a total of six strata being used instead of three. (These six strata only represent three size classes, whereas simple random sampling with only four size class strata (not listed) is still more efficient and easier to do than stratification by single and multiple sites with three size classes, making SRS by worksite still a more preferable design). All of the cluster sampling procedures have design effects much greater than 1 and appear to be less desirable methods, when costs are not considered. This is due to a great amount of variation of worksite employment size within companies and little intra company correlation. Although lack of intra company correlation may increase variances, efficiency may be gained when sampling a given size of a whole multiple worksite company and when costs are only marginal for the collection of

the additional worksites' data that a central office may already have.

Numerical Results

Simple random sampling by worksite is statistically more efficient than cluster sampling to estimate employment for states and industrial classes when costs are equal by worksite. However, costs of data collection could make cluster sampling the preferred method of application. For each sample design, relative standard errors were calculated using the variances of stratified ratio estimates in the following formula:

$$\text{relative standard error} = \sqrt{\frac{\sum_{h=1}^{L} v(\hat{Y}_R)_h}{(\sum_{h=1}^{L} T_h)^2}} \qquad (2)$$

where; $v(\hat{Y}_R)$ = is as seen in formula (1),

$T_h = \sum N_h \mu_h$ is the total employment at the end of the two month period.

Table 2 was constructed to show the effects of cost on the sampling plans for three strata across SIC's.

Under many cost scenarios, collection of entire companies as a sample unit reduces the relative standard error from that of worksite sampling. Subsampling within clusters has no benefit under these cost structures. Simple random worksite sampling for three strata and the two step process of stratifying on the condition that the unit belongs to a single or multiple worksite report before size stratification produces relative standard errors of 0.172% and 0.169% respectively. When summing up multiple units and stratifying according to company size(plans 4, 5), cluster sampling relative standard errors are lower than the best possible case of worksite sampling (plan 2), 0.169%, for costs up to three times the cost of single site selection and for marginal costs of $1 for the first site plus $0.25 for each additional multiple site. The most realistic cases are those with marginal costs, where the cost is the same as a single worksite plus a smaller amount for each additional site within the company. These could be computer costs to output, collect, edit, and process the additional data, or dollars in personnel time to solicit the employers' additional units and centralize reporting. Similar results are seen at the SIC level. Table 3, below breaks down relative standard errors by SIC so that industrial classes can be compared for different marginal costs and for the variety of sampling plans.

Table 2. Relative Standard Errors For Comparison Of Sample Designs When Costs Considered (in percent)

		Sampling Costs for Grouped Multiple Units							
Sample Plan	**Units Selected↓**	**$1+.00**	**$1+.10**	**$1+.25**	**$1+.50**	**$2**	**$3**	**$4**	**$5**
1.	M_i=1	.172							
2.	M_i=1	.169							
3.	m_i=M_i	.147	.150	.154	.160	.194	.214	.232	.249
4.,5.	m_i=M_i	.130	.142	.158	.183	.144	.157	.171	.184
	m_i=1								
	m_i=2	.267	.268	.271	.276				
	m_i=3	.223	.226	.232	.245				
	m_i=1/2M_i	.169	.183	.217	.281				
6.	m_i=M_i	.129	.142	.157	.179	.144	.156	.168	.178
	m_i=1								
	m_i=2	.266	.269	.276	.281				
	m_i=3	.223	.228	.235	.254				
	m_i=1/2M_i	.168	.185	.223	.290				
7.	m_i=M_i	.134	.146	.169	.199	.149	.167	.183	.197
	m_i=1								
	m_i=2	.271	.277	.288	.307				
	m_i=3	.227	.238	.256	.295				
	m_i=1/2M_i	.168	.183	.213	.266				

Table 3: Relative Standard Errors For Comparison of SIC's By Cost and Design For 3 Strata (in percent)

Stratification by:	Cost	SIC:15	20	38	42	54	60
1. worksite, SRS	1.00	.841	.307	.367	.482	.324	.311
2. Single/multi SRS by worksite	1.00	.813	.300	.338	.476	.319	.303
3. worksite then group multiunits $m_i=M_i$	1, .00	.772	.254	.326	.429	.249	.248
	1, .10	.780	.260	.331	.435	.257	.255
	1, .25	.791	.268	.337	.443	.270	.266
	1, .50	.809	.282	.348	.457	.289	.282
4. grouped worksite units, cluster sample, $m_i=M_i$	1, .00	.683	.222	.278	.391	.223	.209
	1, .10	.726	.242	.304	.425	.244	.244
	1, .25	.778	.267	.336	.470	.271	.317
	1, .50	.861	.299	.382	.530	.330	.404
6. single/ multiple site units, cluster sample, $m_i=M_i$	1, .00	.688	.209	.260	.398	.221	.208
	1, .10	.733	.234	.285	.428	.242	.255
	1, .25	.779	.260	.315	.462	.277	.318
	1, .50	.854	.302	.363	.513	.312	.392

As shown above, SIC's, that have a high percentage of multiple site companies, such as SIC 54, food stores, and SIC 60, depository institutions, have reduced relative standard errors of approximately 31% and 33% when comparing cluster sampling grouped (UI) units to SRS of worksite units for equal costs. However, SIC's with smaller proportions of multiple site employers, such as SIC 15, general building contractors, show less reductions in relative standard error of about 19% . Industries with a smaller proportion of multiple worksite employers appear to be less affected by higher marginal costs, simply because employers have fewer additional worksites.

It is very likely that clustering would be shown to be even more beneficial in this study if data for the whole country was available. The thirteen states used are relatively small, and therefore less likely to have many multiple site companies across SIC's, whereas states such as California, Texas, and New York may have many.

Summary

In this study several sampling designs were compared for a variety of cost constraints using the new BLS BEL. If all costs were equal in the collection of employment data from business worksites the most statistically efficient sampling plan would be; separating single and multiple site companies into two strata, further stratify the two strata into three worksite employment size strata and randomly sample them by worksite, if three size class strata are used. However, there are costs involved in data collection. When costs are considered, cluster sampling using the grouped worksite as the sampling unit becomes the most efficient design, especially in industries with a high rate of multiple sites. In this study we experimented with several costs that could be possible, because the actual costs were not available to us. This is only part of the analysis. Before a final decision can be made we must : 1. Measure the actual costs of sampling multiple units; 2. Determine how the larger firms will react to this extra burden; 3. Devise methods to collect multiple worksite data which suit the respondent and BLS systems; and 4. Assess the effects of grouped sampling on small area estimates. Studies using the Business Establishment List are in early stages, much more research can be done to study effective methods of estimating the employed population for the nation and for smaller areas, such as states and counties.

My thanks to Stephen Woodruff, mathematical statistician at BLS, for his many helpful comments.

John Sommers is currently at the Agency For Health Care Policy Research, Rockville, MD 20852.

References

Bureau of Labor Statistics Bulletin 2285, *BLS Handbook of Methods* (1988) Chapter 2, Employment, Hours, and Earnings from the Establishment Survey, 13-27.

Cochran, William G. (1977), *Sampling Techniques*, John Wiley and Sons, New York, third edition.