AUGMENTING A SAMPLE TO SATISFY SUBPOPULATION RELIABILITY REQUIREMENTS

Promod Chandhok, Rachel Weinstein, and Carol Gunlicks, Bureau of the Census[1]
Promod Chandhok, Washington, DC 20233

I. Introduction

The Current Population Survey (CPS) was established in 1940 to provide estimates of labor force characteristics. Since that time, expansion in the sample and changes in the design have enabled the CPS to provide monthly labor force estimates for the nation and the 11 largest states. In addition, annual average estimates for the remaining states and selected metropolitan areas are produced. In this paper, two of four design options are investigated that would produce monthly labor force estimates for all states and the District of Columbia (D.C.).

In the current CPS sample design, independent samples are selected from each state using a stratified multi-stage cluster design. One nonself-representing (NSR) primary sampling unit (PSU) is selected from each stratum with probability of selection proportional to the population size of the PSU. PSUs with a large population are in the sample with certainty and are designated self-representing (SR) PSUs.

A maximum 8% coefficient of variation (CV) on the monthly estimate of number unemployed is maintained for the eleven largest states, 10% monthly CV for the Los Angeles and New York City areas, and 8% CV on the annual estimates of unemployed for the remaining states and D.C. The 1990 CPS sample (D_{90}) will be selected using the current design but will use 1990 census data.

The Bureau of Labor Statistics is interested in producing monthly state estimates for all the states and D.C. The eleven largest states and the substate areas of California and New York must maintain current reliability. Estimates for the remaining states and D.C. must achieve a maximum monthly CV of 10% on the estimate of unemployed. The sample design used to obtain these estimates is referred to as CPS Two-Phase (D_2). We plan to select samples at the same time for both D_{90} and D_2 in 1994. The CPS Two-Phase design would be implemented approximately two years after D_{90}. More details about CPS expansion are in Tupek, Waite and Cahoon (1990).

Four methods of expanding the sample are being explored. The methods are:

1. Stratify and select a sample independently for both D_{90} and D_2. (Benchmark option)
2. Independently stratify and select PSUs to supplement D_{90} in order to obtain D_2. (Independent Supplement option)
3. Stratify for D_{90} and D_2 independently and use controlled selection to simultaneously select PSUs for both samples. (Controlled Selection option)
4. Assign multiple workloads to each strata using the D_{90} stratification. (Multiple workload option)

The controlled selection option is discussed in Ernst (1990). The multiple workload option is currently being researched and the results will be reported elsewhere.

One can compare the independent supplement option to the benchmark option in several ways. We compare assuming the same sample size for the two approaches. This implies that the within-PSU variances of estimates based on the two approaches are equal. Therefore, in comparing the two approaches we only look at the between-PSU variances. Although the within-PSU variance is the same for the two approaches this may represent a substantial portion of the total variance and therefore the total variance is also considered. Cost is also an important consideration in choosing a particular sample design. It is expensive to add a new PSU to the sample because of the start-up costs involved. Therefore, in addition to low total variance, a goal of the new sample design is to maximize the overlap of PSUs between the new design and the previous design.

Research for all methods uses 1980 data to simulate stratifications for D_{90} and D_2. In the simulations, 1985 projected data was used to estimate workloads and sampling intervals.

Section II of this paper contains notation used throughout, while section III describes the benchmark approach and section IV describes an independent supplement option. Section V evaluates these two approaches with a comparison of the results of the simulations.

II. Notation

Since samples are selected independently for each state, all notation applies to the state level if not specified. Let

H be the number of strata (including SR PSUs) in the original stratification (D_{90}),

N_h the number of PSUs in stratum h, h = 1, 2, ..., H,

X_{hi} the measure of size (MOS) of PSU i in stratum h,

$$X_{h.} = \sum_{i=1}^{N_h} X_{hi} \quad \text{, stratum MOS,}$$

$$X_{..} = \sum_{h} \sum_{i} X_{hi} = \sum_{h} X_{h.} \quad \text{, total MOS,}$$

Y_{hi} the number of units with a given characteristic in the i-th PSU of stratum h,

$Y_{h.}$ the number of units with a given characteristic in stratum h,

K the number of strata in the stratification for the independent supplement,

N_k the number of PSUs in stratum k, k = 1, 2, .., K
X_{ki} the MOS of PSU i in stratum k, i = 1, 2, .., N_k,
$X_{k.}$, $X_{..}$, Y_{ki} and $Y_{k.}$ be defined similarly as for the original stratification, with k strata instead of h,

$\hat{Y}$ the estimate of Y based on the independent supplement approach,

$V_B(\hat{Y})$ the between-PSU variance of the estimate of Y based on the independent supplement approach,

$V_W(\hat{Y})$ the within-PSU variance of the estimate of Y based on the independent supplement approach and

SI the sampling interval used to meet the reliability requirements for D_2 in the benchmark approach.

A prime (') indicates the notation applies to the benchmark option.

III. The Benchmark Approach

The following approach to achieving the increased reliability required by D_2 will be used as a benchmark to compare the various methods of obtaining increased reliability. The benchmark (or independent sample) method entails selecting an independently stratified sample. The stratification is obtained from an algorithm that clusters PSUs to minimize the between-PSU variance for one PSU per stratum designs. For a description of the clustering algorithm used see Kostanich, Judkins, Singh and Schautz (1981).

The benchmark option is evaluated by computing the between-PSU variance of the number of unemployed and the number of persons in the civilian labor force. For each state the variances are of the form

$$V_B(\hat{Y}') = \sum_{h=1}^{H'} \sum_{i} X_{hi}\, X_{h.} \left(\frac{Y_{hi}}{X_{hi}} - \frac{Y_{h.}}{X_{h.}} \right)^2 .$$

The formula assumes that a census is undertaken, however this data is obtained from only a sample. Using sample data in the formula produces sampling bias, which is generally positive, and should be taken into account. Since our main intent is to compare the two options, the bias is ignored. Table 1 provides state between-PSU variances for the estimated number of unemployed and civilian labor force for only 29 states. The other states were omitted for the following reasons. For the eleven largest states the reliability requirements are already met in D_{90} and hence additional sample is not needed in these states. The remaining states either consisted entirely of SR PSUs (in which case the between-PSU variance is zero) or had data file problems.

The approximate within-PSU variance is

$$V_w(\hat{Y}') = (SI)\,(CNP16+)\,(P)\,(Q)\,(DE)$$

where

CNP16+ = number of civilian noninstitutional persons 16 and older (state total, projected to implementation date),

P = proportion with a given characteristic,

Q = 1 - P and

DE = design effect for the given characteristic and within-PSU sample design.

The advantage of using this approach for D_{90} and D_2 is that both designs are optimal in terms of variance. They both achieve an optimal stratification based on a given CV. The disadvantages are that cost is traded for the advantage of the variance and the 1990 design is not guaranteed to be a subset of the Two-Phase design. It is possible that PSUs in sample in D_{90} will be dropped two years later when D_2 is implemented. This will be costly especially if these PSUs were not in sample before D_{90}. Also, PSUs may be dropped in D_{90} and then added again when the Two-Phase occurs. Cost estimates involved have not been obtained, however it may be possible to find out the expected number of PSUs that will be dropped in D_2.

IV. Independent Supplement Approach

In this approach the D_{90} sample is augmented with an independently selected supplement. First, the additional sample size is determined so that the size of the D_{90} sample plus that of the supplement equals the benchmark sample, i.e.,

$$\frac{1}{SI} = \frac{1}{SI_1} + \frac{1}{SI_2} \qquad (1)$$

or,

$$SI_2 = SI\; SI_1 (SI_1 - SI)^{-1}$$

where SI_1 is the sampling interval used to select the D_{90} sample and SI_2 the sampling interval needed for the supplement. SI_2 together with a target workload of 50 for an NSR PSU helps to determine the number of strata needed for the supplement. Then, the clustering program is used to produce a stratification for which the between-PSU variance is minimum. This between-PSU variance is calculated assuming one PSU is selected with probability proportional to size from each stratum. The between-PSU variance formula for the supplement is the same as for the benchmark option.

An estimator of Y based on the two portions, i.e., the D_{90} sample and the supplement, is given by

$$\hat{Y} = SI\left(\sum_{h=1}^{H} \hat{Y}_{h.} + \sum_{k=1}^{K} \hat{Y}_{k.} \right)$$

$$= SI \sum_{r=1}^{2} \frac{1}{SI_r}\, \hat{Y}_r = \sum_{r} \frac{SI}{SI_r}\, \hat{Y}_r$$

where

$\hat{Y}_1 = SI_1 \sum_{h=1}^{H} \hat{Y}_{h.}$ = estimate of Y based on the D_{90} sample, and

$\hat{Y}_2 = SI_2 \sum_{k=1}^{K} \hat{Y}_k$ = estimate of Y based on the

$$\hat{Y}_1 = SI_1 \sum_{h=1}^{H} \hat{Y}_h.$$ = estimate of Y based on the D_{90} sample, and

$$\hat{Y}_2 = SI_2 \sum_{k=1}^{K} \hat{Y}_k$$ = estimate of Y based on the supplement.

Hence the weight, w_r, assigned to the r-th portion of the combined sample is

$w_r = (SI/SI_r)$, r = 1,2.

These weights are proportional to corresponding sample sizes. The within-PSU variance is given by

$$V_w(\hat{Y}) = \sum_r \left(\frac{SI}{SI_r}\right)^2 V_w(\hat{Y}_r),$$

where

$$V_w(\hat{Y}_r) = (SI_r)(CNP16+)P\,Q(DE).$$

$V_w(\hat{Y}_r)$ denotes the approximate within-PSU variance of the estimate of Y based on the r-th portion of the combined sample (r = 1, 2). Thus,

$$V_w(\hat{Y}) = (SI)^2\ (CNP16+)\ (P)\ (Q) * (DE) \sum_r \frac{1}{SI_r^2} SI_r$$

$$= (SI)^2\ (CNP16+)\ (P)\ (Q) * (DE) \sum_r \frac{1}{SI_r}$$

$$= (SI)\ (CNP16+)\ (P)\ (Q)\ (DE)$$

using equation (1). Hence, the approximate within-PSU variance for this option is the same as that for the benchmark option (given in table 3). The between-PSU variance of this option is given by

$$V_B(\hat{Y}) = \left(\frac{SI}{SI_1}\right)^2 V_B(\hat{Y}_1) + \left(\frac{SI}{SI_2}\right)^2 V_B(\hat{Y}_2)$$

where $V_B(\hat{Y}_r)$ denotes the between-PSU variance based on the r-th portion of the combined sample (r = 1, 2). This $\hat{Y}$ may not meet the stricter reliability requirements. However, our main goal was to compare, using between-PSU variance as the criterion, different options for sample expansion. Therefore, we choose weights such that the within-PSU variance for this option is the same as for other options.

Table 2 provides between-PSU variances for estimates of number unemployed in 1980 (UE80) and number in civilian labor force in 1980 (CLF80) based on the D_{90} sample, the supplement and D_2.

V. Comparison

The independent supplement approach has at least one advantage over the benchmark approach, the D_{90} sample is a subset of D_2. This would mean a minimum increase in field costs going from D_{90} to D_2. In the benchmark approach, D_{90} is not a subset of D_2. Thus some D_{90} PSU's could be dropped and more new PSU's may be added than with the independent supplement method. This may initially increase the non-sampling error since the quality of data collected from a currently sampled PSU is generally better than from a new PSU. In a sample of five states it was seen, using the technique for estimating expected overlap in Ernst (1986), that 92% of D_{90} PSUs are expected to be selected in D_2.

The independent supplement option provides a D_2 which may not be optimal whereas the benchmark option gives an optimal D_2. Both options may add PSU's dropped in the implementation of D_{90}. If this loss of 1980 design PSUs is of particular concern, the method of controlled selection (Ernst, 1990) is able to maximize overlap with the 1980 PSUs.

Table 4 provides ratios of variances using the independent supplement to variances using the benchmark approach, for each state and each of the two evaluation variables. In the heading of the last two columns, the average stands for the average over the two evaluation variables. For example, the fifth column is the average of the first and third columns. The second to last row contains the average of ratios, over states, for each evaluation variable. The last row gives the ratio of the variances for the sum of the state totals for each evaluation variable. It is seen that in terms of between-PSU variance, the independent supplement approach is not viable. But in terms of total variance this approach is competitive if the cost of dropping D_{90} PSUs is non-trivial.

References

Ernst, Lawrence R. (1986) "Maximizing the Overlap Between Surveys When Information is Incomplete," European Journal of Operational Research, 192-200.

Ernst, Lawrence R. (1990) "Simultaneous Selection of Primary Sampling Units for Two Designs," Proceeding of the Section on Survey Research Methods, American Statistical Association, to appear.

Kostanich, D., Judkins, D., Singh, R., and Schautz, M. (1981),"Modification of Friedman-Rubin's Clustering Algorithm for Use in Stratified PPS Sampling," Proceedings of the Section on Survey Research Methods, American Statistical Association, 285-290.

Raj, Des (1968) Sampling Theory, New York, McGraw-Hill.

Tupek, A., Waite, P., and Cahoon, L. (1990), "Sample Expansion Plans for the Current Population Survey," Proceedings of the Section on Survey Research Methods, American Statistical Association, to appear.

Acknowledgments

The authors would like to thank Lynn Weidman and Dave Chapman of Statistical Research Division, U.S. Bureau of the Census for review of this paper. Thanks are due to Todd Williams for some programming support.

[1] This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the authors and do not necessarily reflect those of the Census Bureau.

Table 1: 1980 Between-PSU Variances Using the Benchmark Approach

(in millions)

STATE	VARIANCE OF NUMBER UNEMPLOYED	VARIANCE OF NUMBER IN CIVILIAN LABOR FORCE
Alabama	1.2826	23.4348
Arizona	0.4380	36.3781
Arkansas	0.4425	27.5571
Georgia	1.6340	88.3356
Idaho	0.1672	1.1575
Iowa	0.8660	7.1823
Kansas	0.6980	19.0554
Kentucky	1.3429	24.8051
Louisiana	1.3136	19.4688
Maryland	0.8815	9.4802
Minnesota	5.7273	33.5941
Mississippi	0.3062	19.8983
Missouri	1.9238	77.8580
Montana	0.2479	5.3641
Nebraska	0.1689	3.8869
Nevada	0.0442	14.5978
New Mexico	0.2691	10.6502
North Dakota	0.0922	1.3108
Oklahoma	0.7990	33.0730
Oregon	1.0697	15.1268
South Carolina	0.0601	2.9771
South Dakota	0.1590	1.6067
Tennessee	1.5847	63.0653
Utah	0.1076	7.1377
Virginia	1.2076	58.8836
Washington	5.7959	36.7214
West Virginia	0.0556	3.0979
Wisconsin	1.5401	53.9748
Wyoming	0.0077	0.3736

Table 2: 1980 Between-PSU Variances Using the Independent Supplement Approach

(in millions)

STATE	VARIANCE OF NUMBER UNEMPLOYED			VARIANCE OF NUMBER IN CIVILIAN LABOR FORCE		
	D 90	SUPPLEMENT	D 2	D 90	SUPPLEMENT	D 2
Alabama	7.5202	3.1793	2.3159	166.9928	55.4677	45.6915
Arizona	2.1873	1.8006	1.0084	186.5696	67.4025	52.8102
Arkansas	1.2271	1.3231	0.6516	47.5917	41.0424	22.0761
Georgia	4.1668	2.7760	1.7300	315.7452	206.6706	130.1582
Idaho	0.6496	0.8574	0.3931	12.2360	26.5802	10.6294
Iowa	2.2321	2.0863	1.0784	49.5031	47.6836	24.2932
Kansas	1.8244	1.0852	0.6994	89.5828	65.5016	37.9888
Kentucky	10.3380	4.6272	3.4271	192.3196	78.6116	61.3564
Louisiana	5.8990	3.2324	2.1708	428.0979	62.7176	105.3753
Maryland	2.2043	4.5218	1.5124	130.3982	409.0451	111.5665
Minnesota	9.9163	8.9621	4.7086	163.2228	155.9539	79.8964
Mississippi	1.8633	2.2470	1.0371	85.1082	35.9218	29.3023
Missouri	7.3237	7.7476	3.7744	277.1723	366.6369	158.4331
Montana	1.5702	1.0578	0.6462	8.5584	13.6191	5.6786
Nebraska	0.5506	0.3398	0.2170	26.6008	18.9286	11.1913
Nevada	0.0841	0.2615	0.1103	32.9193	71.0834	31.4939
New Mexico	0.8971	0.4140	0.2957	44.7675	70.4513	31.9095
North Dakota	0.3295	0.3560	0.1713	8.1631	11.2092	4.8277
Oklahoma	3.2813	1.7031	1.2508	188.4204	156.5274	86.3334
Oregon	6.5895	4.0309	2.5041	151.9108	160.0811	82.1852
South Carolina	2.9005	1.1588	0.9557	42.6806	14.5693	13.3469
South Dakota	0.4455	0.3688	0.2045	9.8163	8.3541	4.5607
Tennessee	5.4810	5.8062	2.8523	195.9452	149.7502	85.1381
Utah	0.6457	0.3826	0.2411	71.3772	32.2430	23.0292
Virginia	6.3770	5.5179	2.9679	476.8672	524.6721	250.8063
Washington	17.6610	13.6121	7.7154	329.5929	111.6519	93.6220
West Virginia	1.1429	0.1189	0.2461	25.5502	1.2576	5.0398
Wisconsin	41.4520	6.5697	11.6613	173.5238	186.8491	90.2626
Wyoming	0.1375	0.1393	0.0706	9.7838	5.6817	3.6573

Table 3: 1980 Within-PSU Variances for Two-Phase

(in millions)

STATE	VARIANCE OF NUMBER UNEMPLOYED	VARIANCE OF NUMBER IN CIVILIAN LABOR FORCE
Alabama	92.4125	465.4400
Arizona	84.1473	364.0277
Arkansas	40.8795	179.8064
Georgia	255.6976	1050.1737
Idaho	7.1712	31.0839
Iowa	68.0172	271.6088
Kansas	46.7830	181.2626
Kentucky	104.3241	458.8515
Louisiana	136.1796	598.9595
Maryland	176.8181	643.9977
Minnesota	158.7792	522.1312
Mississippi	42.5799	199.7374
Missouri	180.9153	806.1789
Montana	5.0581	21.0899
Nebraska	19.6891	78.6564
Nevada	10.6160	34.8745
New Mexico	12.0781	57.3242
North Dakota	3.4344	13.0751
Oklahoma	79.2581	330.2077
Oregon	73.9523	286.5478
South Carolina	71.5775	323.2127
South Dakota	3.4771	13.2323
Tennessee	173.4630	762.7459
Utah	18.0444	71.0296
Virginia	238.5792	966.1812
Washington	157.3513	655.3480
West Virginia	22.5509	120.0103
Wisconsin	185.7463	719.8619
Wyoming	2.8522	9.5602

Table 4: Ratios of the Variance Using the Independent Supplement Approach to the Variance Using the Benchmark Approach

STATE	UNEMPLOYED		CIVILIAN LABOR FORCE		AVERAGE	
	BETWEEN-PSU VARIANCE RATIO	TOTAL VARIANCE RATIO	BETWEEN-PSU VARIANCE RATIO	TOTAL VARIANCE RATIO	BETWEEN-PSU VARIANCE RATIO	TOTAL VARIANCE RATIO
Alabama	1.8057	1.0110	1.9497	1.0455	1.8777	1.0283
Arizona	2.3025	1.0067	1.4517	1.0410	1.8771	1.0239
Arkansas	1.4728	1.0051	0.8011	0.9736	1.1369	0.9893
Georgia	1.0588	1.0004	1.4735	1.0367	1.2661	1.0186
Idaho	2.3509	1.0308	9.1834	1.2938	5.7671	1.1623
Iowa	1.2453	1.0031	3.3824	1.0614	2.3138	1.0322
Kansas	1.0020	1.0000	1.9936	1.0945	1.4978	1.0473
Kentucky	2.5520	1.0197	2.4735	1.0756	2.5128	1.0476
Louisiana	1.6525	1.0062	5.4125	1.1389	3.5325	1.0726
Maryland	1.7158	1.0036	11.7683	1.1562	6.7421	1.0799
Minnesota	0.8221	0.9938	2.3783	1.0833	1.6002	1.0386
Mississippi	3.3872	1.0170	1.4726	1.0428	2.4299	1.0299
Missouri	1.9620	1.0101	2.0349	1.0911	1.9984	1.0506
Montana	2.6063	1.0751	1.0586	1.0119	1.8325	1.0435
Nebraska	1.2846	1.0024	2.8792	1.0885	2.0819	1.0455
Nevada	2.4943	1.0062	2.1574	1.3415	2.3259	1.1739
New Mexico	1.0989	1.0022	2.9962	1.3128	2.0475	1.1575
North Dakota	1.8572	1.0224	3.6831	1.2445	2.7702	1.1334
Oklahoma	1.5655	1.0056	2.6104	1.1466	2.0879	1.0761
Oregon	2.3409	1.0191	5.4331	1.2223	3.8870	1.1207
South Carolina	15.9049	1.0125	4.4832	1.0318	10.1940	1.0221
South Dakota	1.2867	1.0125	2.8385	1.1991	2.0626	1.1058
Tennessee	1.7999	1.0072	1.3500	1.0267	1.5749	1.0170
Utah	2.2398	1.0074	3.2264	1.2033	2.7331	1.1053
Virginia	2.4577	1.0073	4.2594	1.1872	3.3585	1.0973
Washington	1.3312	1.0118	2.5495	1.0822	1.9404	1.0470
West Virginia	4.4249	1.0084	1.6268	1.0158	3.0259	1.0121
Wisconsin	7.5718	1.0540	1.6723	1.0469	4.6221	1.0505
Wyoming	9.1555	1.0220	9.7901	1.3306	9.4728	1.1763
Mean of Ratios	2.8534	1.0132	3.3927	1.1250	3.1231	1.0691
Ratios of Variances for Sum of State Totals	1.9727	1.0105	5.0858	1.1287	3.4793	1.0696