THE USE OF UNEQUAL PROBABILITY SAMPLING TO REDUCE RESPONDENT BURDEN

Crop
Reporting
Board

Economics, Statistics,
and Cooperatives Service

THE USE OF UNEQUAL PROBABILITY SAMPLING TO REDUCE RESPONDENT BURDEN

By

Robert D. Tortora
Keith N. Crank

Mathematical Statisticians
Sample Survey Research Branch
Statistical Research Division
Economics, Statistics, and Cooperatives Service
U.S. Department of Agriculture

Washington, D.C. April, 1978

Table of Contents

Page

Introduction 1

Current Surveys 2

Methodology 4

Methods of Computing Expected Respondent Burden 8

Creation of a Population 16

Comparison of Estimators 17

Summary 24

Appendix A, Derivation of the PPS Estimators 25

INTRODUCTION

The reduction of respondent burden is a major problem faced by any organization that repeatedly contacts the public to obtain information. The Office of Management and Budget [3] has emphasized this as a goal of ~ny Federal Agency that collects statistics. Also, refusal rates are increasing. For some major national agricultural surveys refusal rates are approaching 15 percent with some states exceeding 25 percent. Particularly bothersome is the fact that more and more of the refusals are coming from the large farming operations that can account for a major percentage of a characteristic to be estimated. Therefore any procedure to reduce the burden placed on these large operators may help reverse this trend toward refusing.

One way to reduce respondent burden is to make fewer contacts. Fewer contacts may come about by conducting fewer surveys or reducing sample sizes, i.e., reducing overall burden, or by simply not contacting the same sampling units too often, i.e., reducing the individual burden. When estimates of characteristics have to be made for the same population over many different subject matter areas, the individual burden placed on units that have a large 'size' for one or more subject matter areas can be substantial. In fact, if they are asked to participate in a survey for which they have a small 'size', this invitation may precipitate future refusals on all surveys.

Presently various methods are used to reduce this individual burden. For nonprobability surveys sampling units that have no positive data associated with it for that survey item, i.e., a control value of zero, or units that have been contacted many times in the past are given a zero selection probability. However, as more surveys are placed on a probability basis these techniques violate the principle that every sampling unit must have a known positive probability of selection. This paper studies a method proposed by Tortora [7] based on probability proportional to size (PPS) sampling using 'burden'as an inverse measure of size.

We assume that we are interested in estimating population totals for different subject matter areas and the technique of rotation sampling is not permitted. We compare the proposed PPS sampling scheme with the current method used by the Economics, Statistics, and Cooperatives Service (ESCS). Two issues are examined. First, it is shown that the burden of those units that already have a large burden can be reduced substantially. Second, various PPS estimators are shown to be nearly as efficient as the current estimator used by ESCS.

CURRENT SURVEYS

The ESCS estimates crop acreage and livestock numbers as well as many other items on a state and national level. The area frame is the primary frame for probability acreage estimates of major crops. A list frame is used in each state, either alone or in conjunction with the area frame, to improve the estimates of minor crop acreage or livestock items and provide more geographical detail. In the paper we restrict ourselves to estimates made using a list frame for surveys conducted over a one year period.

For each subject matter area the list frame is stratified and a simple random sample(srs) of frame units (names) is drawn from each stratum. Table 1 gives the surveys of interest for this study. These are probability surveys currently employed in South Dakota and the strata are those sampled less than 100 percent.

TABLE 1: South Dakota probability surveys, stratum boundaries, population sizes, and sample sizes.

	Strata						
	10	20	30	40	50	60	70
Surveys							
Cattle on Feed							
stratum boundary values	1-100	101-300					
population size	6558	1224					
sample size	2200	571					
Cattle							
stratum boundary values	0	1-74	75-124	125-224	225-349	350-549	530-999
population size	8929	15494	6005	4964	1934	975	319
sample size	318	342	214	300	268	162	64
Sheep							
stratum boundary values	1-99	100-399	400-999	1000-2999	3000+		
population size	3376	1186	240	75	9		
sample size	2008	650	174	45	5		
Hog							
stratum boundary values	0	1-74	75-149	150-249	250-399		
population size	26019	8009	2772	1117	486		
sample size	553	500	346	213	97		
Dairy							
stratum boundary values	1-24	25-49	50-149				
population size	4762	1891	729				
sample size	1018	505	390				
Chicken							
stratum boundary values	1-149	150-349	350-1599				
population size	5071	1186	428				
sample size	962	307	226				

A sample for a Cattle on Feed survey is drawn once and contacted four times. Samples selected for Cattle and Sheep surveys are contacted twice during the year. Two independent samples are selected for Hog surveys each year, with each sample being contacted twice. Samples for Dairy and Chicken surveys are drawn once a year and contacted monthly.

For the surveys shown in Table 1 comparisons of coefficients of variation (C.V.) will be made between the PPS estimator and current ESCS estimator of population total. The sample sizes given in Table 1 are used for the comparisons. Note that comparisons of C.V.'s for extreme operator (E.O.)*/ strata will not be made since in those strata we assumed the current methodology would want to be retained.

METHODOLOGY

The proposed methodology is to assign different probabilities to the names on the list according to the size of their burdens relative to others in the same stratum. The selection would be done in such a manner that those with larger burdens would have lower probabilities of selection and those with smaller burdens would have higher probabilities of selection. This should result in individual burdens being closer to a common mean.

The method for assigning probabilities is based on the previous number of contacts. This method requires that the population be divided into classes according to size of burden. These classes are then ordered from smallest burden to largest burden. Probabilities are given in the following formula:

$$P_i = c^{1-i}/(M_1 + M_2\ c^{-1} + M_3\ c^{-2} + \ldots\ M_d c^{1-d}) \qquad (1)$$

*/ Extreme operator denotes those operations that are large in one or more specie of livestock and fall into a stratum where with probability one each sampling unit is contacted at least once per year.

where i is the class number, d is the number of classes, M_i is the number in class i, and c is a constant (> 1). The value of c can be varied to produce a larger or smaller spread in the probabilities (if c = 1 then all of the probabilities are equal). The method assumes an unstratified design. For stratified designs selection probabilities must be computed for each stratum.

Before dividing the population into classes, a burden must be assigned to each name on the list. There are a number of ways of doing this, but only two were used in this study. The simpler of the two methods was to assign a burden to a name according to the number of times that name was contacted in the previous year for the seven surveys in the study. Thus, if a name was contacted twice for the hog survey, twelve times for the dairy survey and not at all for the other surveys his burden would be fourteen. (In addition, if it were known that a particular name has been previously interviewed in an area sample then the burden from the area sample could be included as part of the name's total burden. However, that information was not available for this study.)

The second method of assigning burden assumes that not all surveys are equal in terms of the burden they place on farmers. This method uses a response burden index [1] for each survey. This index is based not only on the number of contacts in a year, but also on the length of the questionnaire and the period of time for which information must be remembered. This means that surveys which only required information for the previous month would have a lower index than those for which six months of information is required. After calculating a response burden index (RBI) for each survey the burden assigned to a name is simply the sum of the indices of those surveys for which the name was selected.

It should be noted that both methods of assigning burden require a knowledge of which names were selected for the seven surveys for the previous year. Unfortunately this information is not readily available under the present system. For this reason the study was done using expected number of contacts and expected RBI. Using expected values instead of actual values also has the advantage of allowing a direct comparison of the two sampling methods (for a given individual or group of individuals) in terms of change in burden.

Even after the burdens are assigned it is still not possible to assign probabilities. In equation (1) there are three parameters, the constant (c), the number of classes (d), and the number of units in each class (M_i), which must be known before probabilities can be calculated. Unfortunately the effects of these parameters are not independent and at present there is no analytic method for assigning or evaluating these parameters. To prevent needless calculations as a result of too many classes the following subjective criterion was used in assigning classes: not more than one class would contain less than 5 percent of the population.

The choice of c was considered to be fairly important. If c is close to one, the sampling plan is very similar to a simple random sample. This prevents a possible large increase in the variance, but it also prevents much improvement in response burden. Three values of c, namely 1.1, 1.25, and 1.5 were used in the study.

The values of burden were rounded to integers and placed into groups. Then the smaller groups were combined with other groups of similar burden to satisfy the 5 percent rule. Finally the groups that now existed were combined to form five classes. Although the choice of five classes was aribitrary, it was necessary to choose some value which could be used consistently from survey to survey for all three values of c and for both methods of assigning burden. This allows different "spreads" in the probabilities to be compared. Using c = 1.1, the largest probability is less than one and one half times the smallest probability. For c = 1.25, this ratio is 2.4 and for c = 1.5, it is 5.1.

In order to compare the two sampling methods in terms of relative efficiency, it is necessary to know the survey design and what estimators are being used. We use the stratum boundary values given in Table 1. For the present methodology the estimator used (within each stratum) is $\hat{Y}_i = \sum_{j=1}^{n_i} \frac{N_i}{n_i} y_{ji}$, where $\hat{Y}_i$ is the estimate for the ith stratum, n_i and N_i are the sample size and population size, respectively, for the ith stratum, and y_{ji} is the reported value for the jth individual in the ith stratum. The within stratum estimator for the proposed sampling plan is

$$\hat{Y}_i = \sum_{j=1}^{n_i} \frac{y_{ji} - k_i}{n_i P_{ji}} + N_i k_i \qquad (2)$$

where $\hat{Y}_i$, y_{ij}, n_i and N_i are the same as above, P_{ji} is the probability of selecting the jth individual in stratum i and k_i is a preassigned constant chosen to reduce the variance of the estimate. We compare 4 PPS estimators. The first, denoted $\hat{Y}_{usl}$, is the usual stratifed PPS estimator of a population total. The second, denoted $\hat{Y}_{opt}$, is the PPS estimator with k_i chosen to minimize variance within each stratum. In this sense it is the "best" PPS estimator. The third, denoted $\hat{Y}_{con}$, uses a k_i based on the control data on the frame. The fourth estimator, denoted $\hat{Y}_{str}$ uses a k_i equal to the lower stratum boundary value. For a more detailed explanation of these modified PPS estimators, see Appendix A.

METHODS OF COMPUTING EXPECTED RESPONDENT BURDEN

In order to evaluate the procedure proposed, it was decided to measure the reduction in expected burden. This allows easier and more consistent comparisons between the methods of computing expected burden based on the number of contacts versus a RBI. These expected burdens were computed in the following manner. The list frame from South Dakota was obtained. Each name on the frame was assigned four expected burdens by multiplying its inclusion probability by the number of contacts or the RBI for each survey of interest. For purposes of illustration we omit the required subscripts for each stratum which have their own inclusion probabilities. The following formula was used for each name on the list:

$$B_{ik} = \sum_{j=1}^{7} m_{ij} \Pi_{jk} \quad , \quad i=1, 2; \quad k=1, 2,$$

where

Π_{j1} is the probability of inclusion on the jth survey under the current stratified srs scheme, (the inclusion probability is sample size times one over population size).

Π_{j2} is the probability of inclusion on the jth survey under the PPS stratified scheme, (the inclusion probability is sample size times selection probability computed from equation (1)),

m_{1j} is the number of times the jth survey is conducted in a year, and

m_{2j} is the RBI for the jth survey.

B_{11} is the expected burden based on expected number of contacts under the current methodology. B_{12} is the expected burden based on expected number of contacts using the PPS scheme. B_{21} is the expected burden based on expected

RBI using the current methodology. B_{22} is expected burden based on expected RBI using the proposed methodology. The upper limit on the summation sign is 7 instead of 6 (Table 1) because the control data for Sheep on Feed was also used in assigning burden even though all names with positive data are sampled.

Since the extreme operators (E.O.'s) are contacted most often under our current survey procedures we would, ideally, attempt to reduce their burden over all surveys for which they are not E.O. Reducing the burden of the E.O.'s has the potential of lowering their refusal rate. So, a small decrease in precision caused by the PPS estimator(s) could be more than offset by a decrease in nonsampling error caused by more E.O. survey participation. Therefore, the major question to be answered is "how much is the expected burden of the large or extreme operators reduced?"

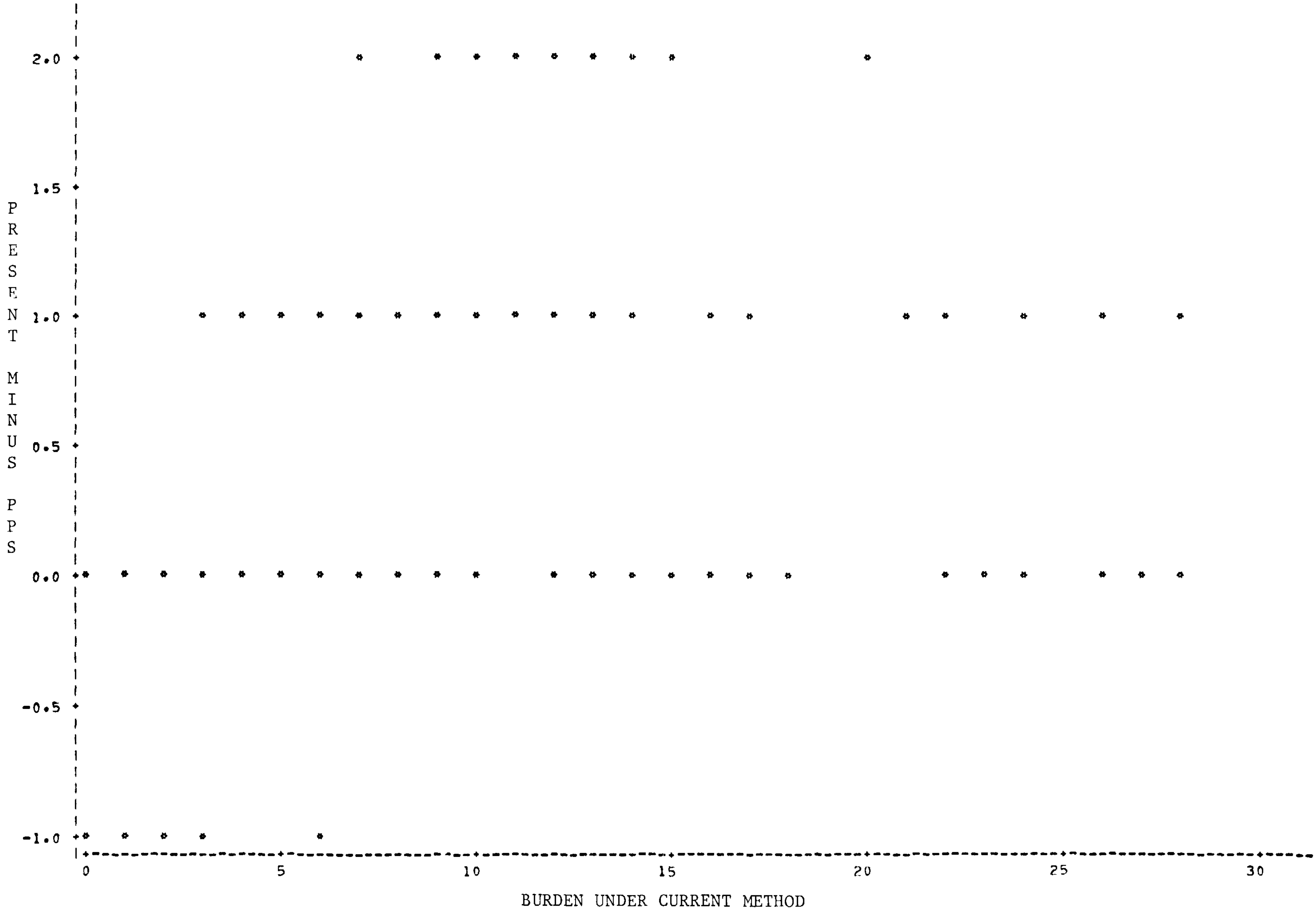
Table 2 gives the definition of E.O.'s by specie of livestock and poultry.

TABLE 2: Extreme Operator Definitions

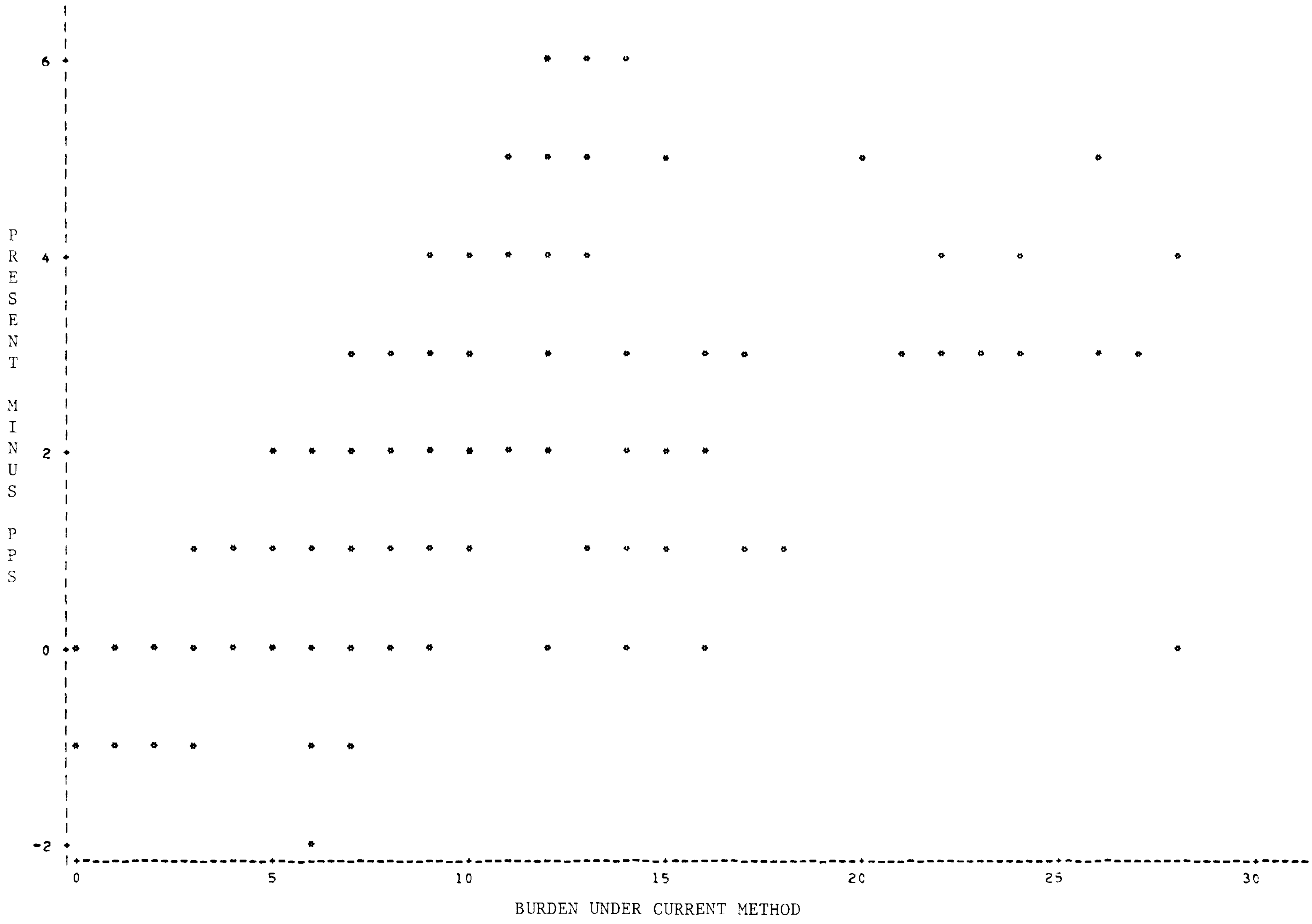
Survey	Minimum number to qualify as Extreme Operator
Cattle on Feed	400
Cattle	1000
Sheep	1000
Sheep on Feed	1000
Hog	400
Dairy	150
Chicken	1600

Three graphs on the following pages show the changes in expected burden (contacts) using values of c of 1.1, 1.25, and 1.5. The expected burden under the current method is plotted on the x-axis; and the difference between this and the expected burden under the PPS method are plotted on the y-axis. The first graph (for c = 1.1) shows a maximum increase of 1 in the burden (shown as -1 on the y-axis), but the maximum decrease is also small. As c becomes larger in the next two graphs the maximum increase and decrease in burden also become larger, but the maximum decrease is much greater than the maximum increase. Also the trend toward a positive slope is pronounced for larger values of c. Thus a c of 1.5 can be seen to be much better than a c of either 1.1 or 1.25. The same situation occurs for expected burden based on RBI. Therefore the rest of the paper will refer only to results obtained using c = 1.5.

Graph 1: Plot of Difference Between Expected Burden Under Present Sampling System and Expected Burden Under Proposed PPS System for c = 1.1.



Graph 2: Plot of Difference Between Expected Burden Under Present Sampling System and Expected Burden Under Proposed PPS System for c = 1.25.



Graph 3: Plot of Difference Between Expected Burden Under Present Sampling System and Expected Burden Under Proposed PPS System for c = 1.5.

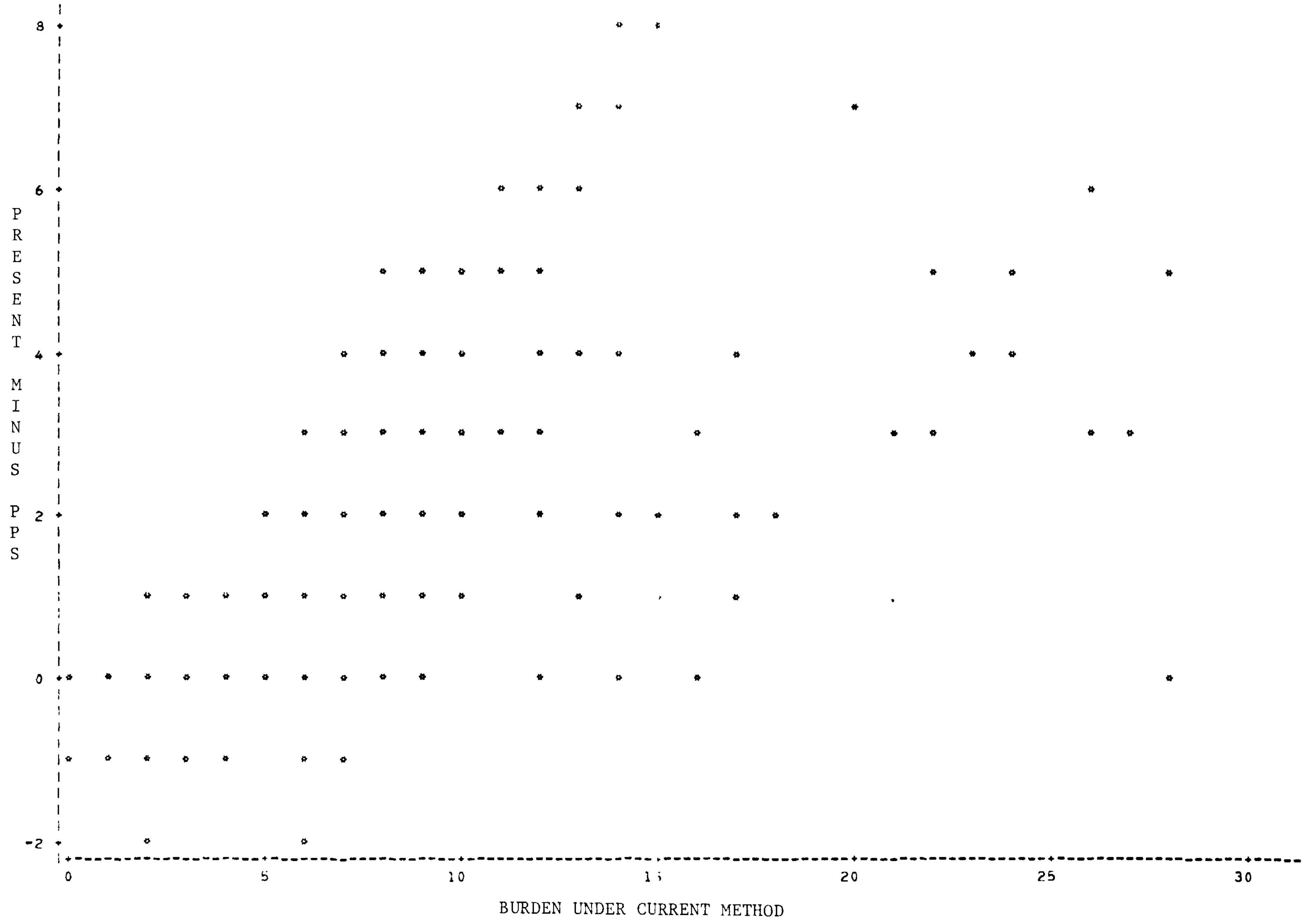


Table 3 compares the average expected burden based on number of contacts for the PPS scheme and current scheme for selected groups in the frame.

TABLE 3: Average expected burden, PPS versus current, based on number of contacts, c = 1.5.

Group	Number of Operators	Current Expected Burden	PPS Expected Burden	Percent Change in Burden
Cattle on Feed EO's	177	6.040	5.513	-8.7
Cattle EO's	193	4.093	3.754	-8.3
Sheep EO's	83	5.266	4.544	-13.7
Sheep on Feed EO's	25	7.150	5.912	-17.3
Hog EO's	415	5.167	4.472	-13.4
Dairy EO's	12	15.295	14.358	-6.1
Chicken EO's	95	16.326	14.722	-9.8
Non-EO's	37666	1.803	1.837	+1.9
Current expected burden of more than 10	231	13.447	10.100	-24.9
Current expected burden of more than 15	42	21.184	17.924	-15.4
Current expected burden of more than 20	26	23.998	20.058	-16.4
Current expected burden of at most 10	38271	1.813	1.836	+1.2

Table 3 indicates that: 1) The average expected burden for EO's decreased from 6 to 17 percent, with an increase for non-EO's of less than 2 percent. 2) For those farmers and ranchers with large expected burden under the present scheme, the average decrease in burden ranged from 15 to 25 percent, depending on the size of burden that was considered large. The average increase for the other farmers was only slightly more than 1 percent.

Table 4 compares the average expected burden based on the RBI for the PPS and current schemes for these same groups.

TABLE 4: Average expected burden, PPS versus current, based on Respondent Burden Index

Group	Number of Operators	Current Expected Burden	PPS Expected Burden	Percent Change in Burden
Cattle on Feed EO's	177	6.439	5.644	-12.4
Cattle EO'S	193	5.152	4.275	-17.0
Sheep EO's	83	5.643	4.494	-20.4
Sheep on Feed EO's	25	8.184	6.320	-22.8
Hog EO's	415	10.574	9.161	-13.4
Dairy EO's	12	11.039	9.667	-12.4
Chicken EO's	45	14.050	12.400	-11.7
Non-EO's	37666	1.699	1.736	+2.2
Current expected burden of more than 10	149	17.894	16.007	-10.6
Current expected burden of more than 15	98	21.166	19.296	-8.8
Current expected burden of more than 20	43	25.427	23.116	-9.1
Current expected burden of at most 10	38353	1.431	1.493	+4.3

Examination of Table 4 shows that: 1) The average expected burden for EO's decreased from almost 12 to almost 23 percent, with an increase for non-EO's of slightly more than 2 percent. 2) For those farmers and ranchers with large expected burden under the present scheme, the average decrease in burden was approximately 10 percent. The average increase for the other farmers was 4 percent.

A note of caution. Comparisons between the two methods of computing expected burdens are very difficult to make since one would be comparing number of contacts with an index number that includes as one of its components the number of contacts. The main purpose of Tables 3 and 4 is to show that the PPS scheme does indeed decrease the burden of large operators.

CREATION OF A POPULATION

Although the change in expected burden was encouraging, this alone is not sufficient reason for accepting the new method of sampling. A comparison must also be made of the variances of the estimators used in the two different plans. When this was done using the control data, it was discovered that the C.V.'s of total numbers of hogs or cattle under both methods were very small. However, there is also the problem that even though this PPS type estimator works well when the actual bounds on the data are known and used, it may not work nearly as well on unknown data which could differ considerably from the available control data. Thus, it was necessary in some way to create "real" data.

Survey data from a list frame in South Dakota was acquired for hogs and cattle. Sample correlation between reported and control data were computed by strata and showed no significant difference from zero. This made the creation of "real" data easier because it was possible to completely randomize the assignments within each stratum. Outliers were found for each stratum and only these operators were assigned their reported values from the survey.

Frequency distributions were obtained from the total hogs reported in the survey for each stratum. For small values of reported hogs an interval of 25 hogs was used. This interval was increased to 50 and then 100 as the reported number of hogs increased and the number of reports decreased. A uniform distribution was assumed within each interval.

Using the frequencies, corresponding percentages of the names on the list in each stratum were assigned randomly to each interval. This random assignment was done on the computer using a uniform random number generator. Then using another random number within each interval each name was assigned a "reported" number of livestock (for hogs and cattle). This program was run separately for hogs and cattle using the parameters and stratum boundaries for that specie. In this way each name on the list was given a value for hogs and for cattle to be used in comparing the two methods of sampling and estimation.

COMPARISON OF ESTIMATORS

Before beginning the comparisons, some general remarks concerning the efficiency of PPS estimators are in order. Rao [6] has shown that, when using the area frame, PPS estimation can be less efficient than the estimator based on simple random sampling. This inefficiency of the PPS estimation occurs whenever the condition (Raj [5])

$$\sum_{i=1}^{N} (1/P_i - N)Y_i^2/P_i > 0 \qquad (3)$$

is violated. In words, equation (3) requires that $1/P_i$ and Y_i^2/P_i be positively correlated. However, it is difficult to measure the relationship specified in equation (3). Even if we could measure the magnitude by which equation (3) differs from zero we still would not have a feel for the relative efficiency of the PPS estimator.

All the comparisons made here assume with replacement sampling for the calculation of variances. This was done for ease of computation of the variance of the PPS estimators. So the conclusions hold for without replacement sampling to the extent that finite population correction factors for simple random without replacement and PPS without replacement are equal.

Comparison of design effects (DEFF) and C.V.'s are made. We use the stratum boundary values specified in Table 1 and define the DEFF for each stratum as the ratio of the standard error of the PPS estimator to the standard error of the srs estimator within stratum. The DEFF for the entire population is defined as the ratio of the standard error of the PPS stratified estimator of population total to the standard error of the stratified srs estimator of population total, $\hat{Y}_{srs}$.

Comparisons are made for the different PPS estimators suggested in the section Methodology. These estimators are compared with the usual estimator of a population total based on a stratified design. Two situations are examined. First, when the control data is used for population values, and second, when the generated data is used for population values. The former situation is less realistic and we examine it primarily to indicate the potential of the PPS estimators.

The comparison for control data is made only for selection probabilities computed from a burden of expected number of contacts. Table 5 presents these comparisons for the estimates of total number of hogs and cattle.

TABLE 5: Comparison of Estimators Using Control Data and Selection Probabilities Computed from Expected Number of Contacts.

Estimator	Hogs		Cattle	
	DEFF	C.V.	DEFF	C.V.
$\hat{Y}_{srs}$	1.00	0.9	1.00	0.7
$\hat{Y}_{usl}$	1.92	1.8	2.18	1.4
$\hat{Y}_{opt}$	1.09	1.0	1.08	0.7
$\hat{Y}_{str}$	1.25	1.2	1.35	0.9

Table 5 shows that the PPS estimators are nearly as efficient as the srs estimator of population total. However, since the results are based on control data the C.V.'s are unrealistically low. Therefore, we compare in more detail the estimators using generated data for both methods of computing burden.

A key comparison concerns the efficiency of Y_{opt} with $\hat{Y}_{usl}$, $\hat{Y}_{con}$, and $\hat{Y}_{str}$. $\hat{Y}_{opt}$ is the "best" PPS estimator. However, it could never be used in practice since it assumes knowledge of data for the entire population. Therefore, $\hat{Y}_{opt}$ is a lower bound for the PPS estimators of concern in terms of efficiency.

Examination of Tables 6 and 7 shows that the PPS estimators using a $k_i > 0$ in equation (2) are still very efficient estimators. In fact $\hat{Y}_{con}$ and $\hat{Y}_{str}$ are as efficient (C.V.'s) as the best PPS estimator $\hat{Y}_{opt}$. All the PPS estimators compare favorably with the srs estimator.

Tables 8 and 9 compare the estimators when expected RBI is used as basis for computing selection probabilities. They repeat the results obtained for the PPS estimators based on expected number of contacts. However, these estimators based on expected RBI have a higher level of C.V. when compared to the current estimator.

TABLE 6 : DEFF and C.V.'s Using Generated Hog Data Selection Probabilities Based on Expected Number of Contacts.

	Stratum 10		Stratum 20		Stratum 30		Stratum 40		Stratum 50		Total	
Estimator	DEFF	C.V.	DEFF	C.V.	DEFF	C.V.	DEFF	C.V.	DEFF	C.V.	DEFF	C.V.
$\hat{Y}$	1.00	27.2	1.00	7.5	1.00	5.6	1.00	5.6	1.00	8.6	1.00	4.9
$\hat{Y}_{usl}$	1.08	29.3	1.12	8.4	1.20	6.8	1.24	7.0	1.26	10.8	1.11	5.4
$\hat{Y}_{opt}$	1.07	29.2	1.09	8.2	1.12	6.3	1.10	6.2	1.10	9.4	1.08	5.3
$\hat{Y}_{con}$	1.08	29.3	1.09	8.2	1.12	6.3	1.11	6.2	1.10	9.4	1.09	5.3
$\hat{Y}_{str}$	1.08	29.3	1.12	8.4	1.12	6.3	1.10	6.2	1.10	9.4	1.09	5.3

TABLE 7 : DEFF and C.V.'s Using Generated Cattle Data Selection Probabilities Based on Expected Number of Contacts.

Estimator	Stratum 10		Stratum 20		Stratum 30		Stratum 40		Stratum 50		Stratum 60		Stratum 70		Total	
	DEFF	C.V.	DEFF	C.V.	DEFF	C.V.	DEFF	C.V.	DEFF	C.V.	DEFF	C.V.	DEFF	C.V.	DEFF	C.V.
$\hat{Y}$	1.00	19.5	1.00	6.0	1.0	5.4	1.00	4.2	1.00	4.3	1.00	5.7	1.00	7.8	1.00	2.3
Y_{usl}	1.04	20.2	1.15	6.8	1.23	6.6	1.22	5.1	1.22	5.2	1.19	6.8	1.28	10.0	1.17	2.7
$\hat{Y}_{opt}$	1.03	20.1	1.08	6.4	1.08	5.8	1.08	4.6	1.07	4.6	1.06	6.0	1.11	8.7	1.07	2.5
$\hat{Y}_{con}$	1.04	20.2	1.08	6.5	1.08	5.8	1.08	4.6	1.07	4.6	1.06	6.1	1.12	8.8	1.07	2.5
$\hat{Y}_{str}$	1.04	20.2	1.15	6.8	1.09	5.8	1.09	4.6	1.08	4.6	1.06	6.1	1.11	8.7	1.09	2.5

TABLE 8 : DEFF and C.V.'s Using Generated Hog Data. Probabilities Based on Expected RBI.

Estimator	Stratum 10		Stratum 20		Stratum 30		Stratum 40		Stratum 50		Total	
	DEFF	C.V.	DEFF	C.V.	DEFF	C.V.	DEFF	C.V.	DEFF	C.V.	DEFF	C.V.
$\hat{Y}$	1.00	27.2	1.00	7.5	1.00	5.6	1.00	5.6	1.00	8.6	1.00	4.9
$\hat{Y}_{usl}$	1.27	34.5	1.24	9.3	1.24	7.0	1.37	7.7	1.46	12.6	1.28	6.2
$\hat{Y}_{opt}$	1.26	34.4	1.20	9.0	1.13	6.4	1.15	6.5	1.22	10.4	1.23	6.0
$\hat{Y}_{con}$	1.27	35.5	1.20	9.0	1.13	6.4	1.17	6.6	1.22	10.4	1.24	6.0
$\hat{Y}_{str}$	1.27	35.5	1.24	9.3	1.14	6.4	1.16	6.6	1.22	10.4	1.24	6.1

TABLE 9 : DEFF and C.V.'s Using Generated Cattle Data probabilities Based on Expected RBI.

Estimator	Stratum 10		Stratum 20		Stratum 30		Stratum 40		Stratum 50		Stratum 60		Stratum 70		Total	
	DEFF	C.V.	DEFF	C.V.	DEFF	C.V.	DEFF	C.V.	DEFF	C.V.	DEFF	C.V.	DEFF	C.V.	DEFF	C.V.
$\hat{Y}$	1.00	19.5	1.00	6.0	1.00	5.4	1.00	4.2	1.00	4.3	1.00	5.7	1.00	7.8	1.00	2.3
$\hat{Y}_{usl}$	1.08	21.0	1.16	6.9	1.25	6.7	1.38	5.8	1.23	5.2	1.24	7.1	2.25	17.6	1.20	2.8
$\hat{Y}_{opt}$	1.08	20.9	1.09	6.5	1.10	5.9	1.20	5.1	1.08	4.6	1.07	6.1	1.48	11.6	1.13	2.6
$\hat{Y}_{con}$	1.08	21.0	1.09	6.5	1.10	5.9	1.20	5.1	1.08	4.6	1.07	6.1	1.44	11.7	1.13	2.6
$\hat{Y}_{str}$	1.08	21.0	1.15	6.9	1.10	5.9	1.21	5.1	1.08	4.6	1.08	6.2	1.49	11.7	1.15	2.6

SUMMARY

This paper has shown that it is possible to reduce the expected burden of larger farm operators at practically no loss in sampling efficiency by using a probability of selection computed inversely to the operator burden. Burden was computed using expected number of contacts and using an expected response burden index. Using expected contacts the burden of E.O.'s dropped by 6 to 17 percent depending on the characteristic to be estimated. When the expected burden computed by the index was used the burden dropped by about 12 to about 23 percent.

The variance of four different PPS estimators was compared to the variance of the stratified srs estimator of population total for both methods of computing expected burden. Under both methods the PPS estimators were nearly as efficient. The PPS estimators based on expected number of contacts showed the smallest decrease in precision with DEFF's ranging from 1.07 to 1.17.

APPENDIX A

Derivation of the PPS Estimators

Although it is not needed to determine the change in burden, the estimator is required for an overall evaluation of the sampling plan. Much time and effort was put into a search for a feasible estimator, (feasible meaning having a relatively small variance which is easily estimated). The estimator used for comparison, while possibly not the "best" estimator, is felt to be adequate for comparing the proposed methodology with the present methodology, because of its simplicity and its improvement over the usual PPS estimator. For the de ivations we drop the subscripts to denote strata. However to motivate the derivations we discuss properties of the estimator by strata.

Naturally the first estimator looked at was the usual PPS estimator,

$$\hat{Y}_1 = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{P_i}$$

with variance

$$V(\hat{Y}_1) = \frac{1}{n} \left(\sum_{i=1}^{N} \frac{Y_i^2}{P_i} - Y^2 \right).$$

Using control data for hogs and cattle the C.V.'s combined over all strata were twice as large for the PPS estimator as for the srs estimator. This meant that any gains in burden achieved through unequal probabilities would be lost to the increase in sample sizes made necessary by the large C.V.'s.

One point that was noted for the hog data (and which was later noted in the cattle data) was that the ratio of the PPS variance to the srs variance increased as the strata increase, i.e., as the number of livestock increased. One possible explanation for this was that the data was shifted away from the origin. Instead of having a linear relationship of the form $Y_i = YP_i$, the actual relationship more closely resembled $Y_i = XP_i + K$, for some constant K, where $X = Y - NK$. This led to the estimator

$$\hat{Y}_2 = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i \cdot K}{P_i} + NK$$

which is unbiased.

The following page shows three graphs which help to show why the estimator $\hat{Y}_2$ is preferred to the regular PPS estimator. Figure 1 shows the ideal situation for using the regular PPS estimator, $\hat{Y}_1$. All sample points lie on the line $Y_i = YP_i$ so that $Y = \frac{Y_i}{P_i}$. Figure two shows what happens if we use $\hat{Y}_1$ as an estimator when the points lie on the line $Y_i = (Y - KN)\ P_i + K$. As the sample points deviate from the intersection of the two lines $(\frac{1}{N}, \frac{Y}{N})$, the variance increases. By shifting the Y-axis (Figure 3) the variance is decreased and the resulting estimator $\hat{Y}_2$ is as efficient as $\hat{Y}_1$ was for Figure 1.

Before giving the variance of $\hat{Y}_2$, it will be useful to know a little bit about K. First of all, for any given population with given probabilities of selection (not all equal) there is a unique value of K (call it K_{opt}) which minimizes the variance of $\hat{Y}_2$. The formula for K_{opt} is

$$K_{opt} = \frac{\sum_{i=1}^{N} \frac{Y_i}{P_i} - NY}{\sum_{i=1}^{N} \frac{1}{P_i} - N^2}$$

It is easily seen that the variance of $\hat{Y}_2$ is

$$V(\hat{Y}_2) = \frac{1}{n} \sum_{i=1}^{N} P_i \left(\frac{Y_i - K}{P_i} + KN - Y\right)^2.$$

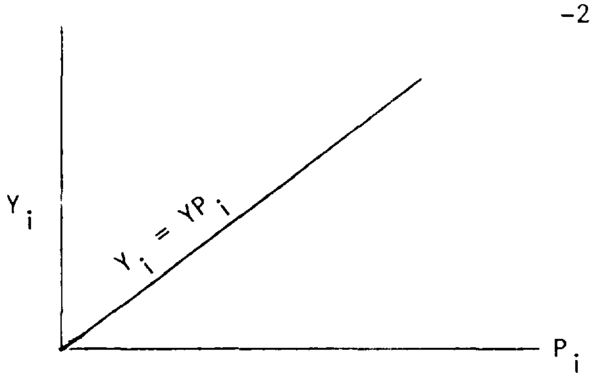


Figure 1: PPS Estimator $\hat{Y} = \frac{Y_i}{P_i}$

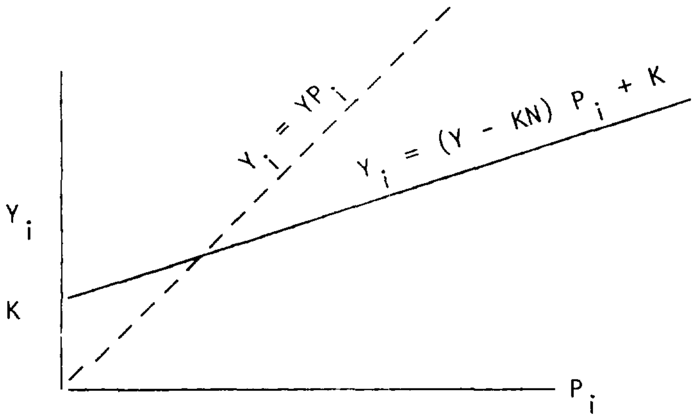


Figure 2: PPS Estimator $\hat{Y} = \frac{Y_i}{P_i}$

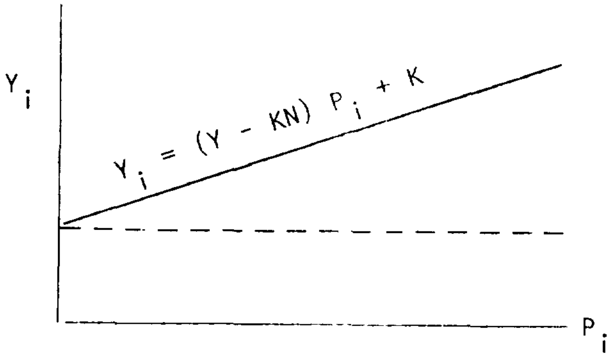


Figure 3: Improved Estimator: $\hat{Y} = \frac{Y_i - K}{P_i} + KN$

Taking the derivative of $V(\hat{Y}_2)$ with respect to K and setting it equal to zero produces, after some rearrangement,

$$\sum_{i=1}^{N} \frac{Y_i}{P_i} - NY = (\sum_{i=1}^{N} \frac{1}{P_i} - N^2)K$$

Solving for K gives K_{opt} which minimizes $V(\hat{Y}_2)$. The denominator is non-negative and is zero if and only if $P_i = \frac{1}{N}$ for all i. In this case the numerator is also zero and the expression is undefined. (This is merely a restatement of the well known fact that adding or subtracting a constant to all values in a simple random sample does not change the variance, i.e., for all values of K, the variances are the same.) Secondly, any value of K which lies in the interval $[0, 2K_{opt}]$ ($[2K_{opt}, 0]$ if $K_{opt} < 0$) produces an estimator whose variance is at least as small as that for the usual PPS estimator. This can be derived explicitly by solving the inequality $V(\hat{Y}_1) - V(\hat{Y}_2) \geq 0$, or it can be found geometrically be referring back to Figure 2. By rotating the dashed line around the point of intersection with the solid line, the variance can be decreased until it reaches a minimum when the two lines are identical. Continuing to rotate past the solid line it is possible to go an equal distance on the other side before the variance of $\hat{Y}_2$ becomes, once again, as large as the variance of $\hat{Y}_1$. (See Figure 4.) For this situation the distance of the two dashed lines from the solid line are equal for a given value of P_i. Thus their variances are equal.

Both of these facts show up in the variance formula:

$$V(\hat{Y}_2) = \frac{1}{n} (\sum_{i=1}^{N} \frac{(Y_i - K)^2}{P_i} - (Y - NK)^2)$$

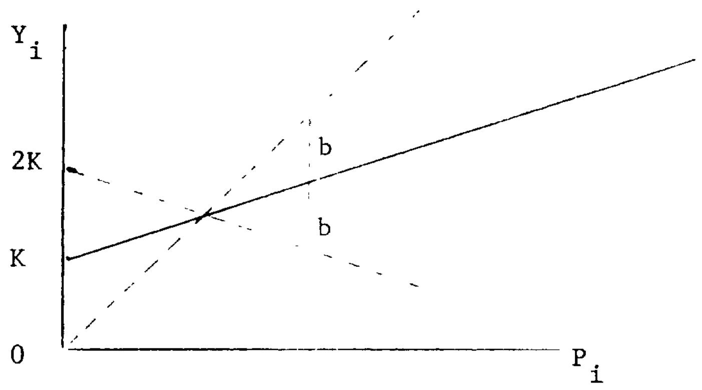


Figure 4: Region for which estimator $\hat{Y}_2$ has smaller variance than $\hat{Y}_1$.

which reduces to

$$V(\hat{Y}_2) = V(\hat{Y}_1) + \frac{KD}{n}(K - 2K_{opt}) \text{ where } D = \sum_{i=1}^{N} \frac{1}{P_i} - N^2.$$

Unfortunately K_{opt} is not known unless the population values are known. Thus its use at first glance seems impractical. However, there are two possibilities concerning its use. First there is the possibility of estimating K_{opt}. However, if this is done using the same data that is used for estimating Y, the results are biased. Although this in itself is not reason to disregard it, its mean square error appears to show no improvement over the variance of the usual PPS estimator and it is much more difficult to compute. Also no gains are achieved by trying to remove the bias. The second possibility, which is supported by the research so far, is to preassign a value of K. This will be an improvement over the usual PPS estimator as long as the assigned value of K is between zero and twice K_{opt}.

Three methods of assigning K seem plausible. Since control data is available for all names, these can be used to determine the optimal K for the control data ($\hat{Y}_{con}$). Also, since the smaller strata have the smaller variance ratios it might be possible to merely shift the strata so that they started at zero. In other words K would be the lower bound of the strata ($\hat{Y}_{str}$). Finally, once a survey is finished, that data could be used to estimate K for the next survey. None of these is necessarily best and no recommendation is made as to which should be used.

Other estimators were studied, including some for without replacement sampling [2], [4]; however, it was felt that an examination of all of these was a research project in itself. The estimator $\hat{Y}_2$ was felt to be sufficiently good (and it is fairly simple) to compare the two sampling plans. If the proposed methodology is accepted, then more time can be spent in examining the different estimators.

REFERENCES

1. Arends, W. (1976), "Sample Selection System", Proceedings of SRS National Conference, USDA, pp. 135 - 141.

2. Chaudhuri, Arijit, "Some Properties of Estimators Based on Sampling Schemes with Varying Probabilities", Australian Journal of Statistics, 1975, Vol 17, pp. 22 - 28.

3. Graham, B. (1976), "Long Range Plans", Proceedings of SRS National Conference, USDA, pp. 17 - 27

4. Hartley, H. O., and Rao, J. N. K., "Sampling with Unequal Probabilities and Without Replacement", Annals of Mathematical Statistics, 1962, Vol 33, pp. 350 - 374.

5. Raj, D. (1968), Sampling Theory, McGraw Hill, New York

6. Rao, J. N. K. (1966), "Alternative Estimators in PPS Sampling for Multiple Characteristics", Sankhya, Series A, 29, pp. 47 - 60.

7. Tortora, R. D. (1977), "Reducing Respondent Burden for Repeated Samples", Agricultural Economics Research, 30, 2, pp. 41-44.