FEASIBILITY STUDY OF THE USE OF CHROMY'S ALGORITHM IN POISSON-SAMPLE SELECTION FOR THE ANNUAL SURVEY OF MANUFACTURES

Laura Zayatz and Richard Sigman*
U.S. Census Bureau

1.0. Introduction

Every five years, when new universe data become available from the Census of Manufactures, the Census Bureau selects a new sample for the Annual Survey of Manufactures (ASM). A probability of selection is assigned to each unit, and then Poisson sampling is used to select the sample (Waite and Cole, 1980). This paper describes a feasibility study we conducted to see if Chromy's algorithm could be used to calculate the probabilities of selection so that constraints on variances of product-class, industry, and selected state-level estimates are satisfied.

2.0. Chromy's Algorithm

Prior to selecting a Poisson sample, the survey designer assigns to each sample unit a probability of its being selected. This is analogous to sample allocation in stratified sampling, in which prior to sample selection the survey designer assigns a sample size to each stratum. The sample-allocation problem has been studied extensively. Consequently, we reviewed some of the recent theoretical results for sample allocation to suggest ways to determine Poisson-sampling probabilities. This section describes Chromy's algorithm for sample allocation in stratified sampling. The next section discusses the application of Chromy's algorithm to determining selection probabilities for Poisson sampling.

To discuss sample allocation for stratified sampling, we need some notation. A stratified sample of H strata is to be selected in order to collect data for I items. We assume that the variance of the estimator for item i, denoted V_i', has the form

$$V_i' = V_i + V_{i0} = \Sigma_h V_{ih}^2/n_h + V_{i0}, \quad (1)$$

where n_h is the stratum sample size for stratum h, and the V_{i0} do not depend on n_h. This variance formula describes a rather large class of estimators. For example, it describes direct expansion estimators, difference estimators, regression estimators, and ratio estimators.

When $I=1$, the survey designer can (1) minimize variance for fixed total sample size or (2) minimize total sample size for fixed variance. In either case the allocation formula is

$$n_h \propto V_{1h},$$

where "$\propto$" means "proportional to". If the total sample size is fixed at level n_f, then

$$n_h = V_{1h} n_f / \Sigma_k V_{1k};$$

and if the variance is fixed at level V_1^*, then

$$n_h = V_{1h} \Sigma_k V_{1k} / (V_1^* - V_{10}). \quad (2)$$

When $I>1$, there are two general approaches to allocating a stratified sample. One approach considers the weighted sum $A = \Sigma A_i V_i'$ and then (1) minimizes A for fixed total sample size or (2) minimizes total sample size for fixed A. Then

$$n_h \propto (\Sigma_i A_i V_{ih}^2)^{1/2}.$$

The second approach involves non-linear programming. The survey designer specifies an upper bound, V_i^*, for each variance and then minimizes the total sample size subject to $V_i' \le V_i^*$ for all i. Bethel (1989) explains how it follows from the Kuhn-Tucker theorem that there exist λ_i such that

$$n_h = (\Sigma_i \lambda_i V_{ih}^2)^{1/2} \quad (3)$$

is the desired allocation. Chromy (1987) describes the following algorithm for solving for the λ_i:

STEP 1: For all i set $\lambda_i = 1$ for the first iteration.
STEP 2: Calculate the n_h using (3).
STEP 3: Calculate the $V_i = V_i' - V_{i0}$ using (1).
STEP 4: Calculate revised λ_i, denoted λ_i', using the updating equation

$$\lambda_i' = \lambda_i [V_i / (V_i^* - V_{i0})]^2 .$$

Steps 2 through 4 are then repeated over and over again with λ_i' each time replacing λ_i. The minimum-sample-size solution is obtained when $V_i' \le V_i^*$ and $\lambda_i(V_i' - V_i^*) = 0$ for all i. (When $I=1$, allocation (2) is obtained on the second iteration.)

A result derived by Causey (1983) can be used to define a stopping rule for Chromy's algorithm. Causey's result is that

$$\Delta = \Sigma \lambda_i | V_i' - V_i^* | \quad (4)$$

is an approximate upper bound on the distance (in terms of the total sample size) that a solution is away from the minimum-sample-size solution. In Appendix A, we derive an alternative stopping rule. The alternative rule assumes that the sample sizes converge to the minimum sample size as a negative-exponential function of the iteration number.

3.0. Poisson Sampling

The Census Bureau's Technical Paper 24 (Ogus and Clark, 1971) describes in detail the sample design and the selection and estimation procedures for the Annual Survey of Manufactures. Much of this section was taken from Technical Paper 24.

3.1. Sample Selection Procedure

In the selection of a new ASM sample, each establishment is sampled independently of the selection or nonselection of every other establishment and the probability of selection varies from

establishment to establishment. This sampling procedure is called Poisson sampling.

With Poisson sampling, a linear unbiased estimate of a total Y is given by

$$\hat{Y} = \sum_{S} Y_h W_h ,$$

where S is the set of units selected, Y_h is the current value of unit h, and W_h, the sampling weight, is the reciprocal of the probability of selection (p_h) of unit h. The variance of an estimated total $\hat{Y}$ is

$$V(\hat{Y}) = \sum_{h=1}^{N} Y_h^2 \frac{q_h}{p_h} = \sum_{h=1}^{N} (W_h - 1) Y_h^2 ,$$

where $q_h = 1 - p_h$ and N is the size of the population. The total number of units selected, n, is a variable. The expected sample size, $E(n)$, is

$$E(n) = \sum_{h=1}^{N} p_h ,$$

and the variance of the sample size, $V(n)$, is

$$V(n) = \sum_{h=1}^{N} p_h q_h .$$

The Census Bureau assigns a new probability of selection for the ASM to each unit when new universe data become available from the Census of Manufactures. They then select a new ASM sample. The pertinent question is "How should we determine the probabilities of selection?" Before describing various methods of determining selection probabilities, we must first discuss the ASM difference estimator.

3.2. Difference Estimator for Population Total

The ASM uses a difference estimator that takes advantage of the correlations between current data and census totals. This improves the reliability of the annual totals.

3.2.1. Formula for the Difference Estimator

Most of the estimates in the ASM are developed by the difference estimator formula

$$\hat{Y}_{DIFF} = \hat{D} + X ,$$

where $\hat{D} = (\hat{Y} - \hat{X})$ is the sample estimate of the change from the last census and X is the total from that census.

3.2.2. Variance Formulas

ASM publications provide various breakdowns of manufacturers' shipments. One breakdown is by the industry of the responding establishment. Another breakdown is by the product class of the manufactured merchandise. The variance of the difference estimator for category i of breakdown t is

$$V(\hat{Y}_{DIFF}^{(t,i)}) = V(\hat{D}^{(t,i)}) = \sum_{h=1}^{N} (\frac{1}{p_h} - 1) D_{tih}^2 ,$$

where $D_{tih} = (Y_{th} - X_{th})$ is the difference in shipments by establishment h in category i of breakdown t between the sample year and the census year.

Since direct measures of the differences, D_{tih}, are unavailable for the entire population, they are predicted for use in sample selection from (1) shipments reported by the companies in the most recent census and (2) estimated regression coefficients that describe the historical relationships between shipments and year-to-year shipment differences. In particular, the prediction equation is

$$\hat{D}_{tih}^2 = \hat{\beta}_{ti} X_{tih}^2$$

where X_{tih} is the shipments in the census year by establishment h in category i of breakdown t. An estimated regression coefficient, $\hat{\beta}_{ti}$, is calculated for each category i and breakdown t using the formula

$$\hat{\beta}_{ti} = \frac{\sum_{h=1}^{n} W_h x_{tih}^2 d_{tih}^2}{\sum_{h=1}^{n} W_h x_{tih}^4}$$

where

x_{tih} = the base year's value of shipments for establishment h and category i of breakdown t, and

d_{tih} = the year-to-year shipment difference for establishment h and category i of breakdown t.

If x_{tih} is zero, then x_{tih} is assigned the census year's value of shipments.

The estimated regression coefficients are reviewed, and some, if deemed appropriate, are modified. Technical Paper 24 describes the review and modification process.

The estimated regression coefficients permit the calculation of the predicted variance,

$$\hat{V}(\hat{Y}_{DIFF}^{(t,i)}) = \hat{V}(\hat{D}^{(t,i)}) = \sum_{h=1}^{N} \hat{\beta}_{ti} X_{tih}^2 (\frac{1}{p_h} - 1) ,$$

which is used in the sample selection process.

After the sample is selected and sample data become available, the unbiased estimated variance,

$$v(\hat{Y}_{DIFF}^{(t,i)}) = v(D^{(t,i)}) = \sum_{h=1}^{n} (\frac{1}{p_h} - 1) \frac{D_{tih}^2}{p_h} ,$$

is calculated and appears in the published tables.

3.3. Determination of Selection Probabilities

3.3.1. Univariate Case

If we wish to constrain the variance of the difference estimator for a single category i of breakdown t, that is

$$V(\hat{D}^{(t,i)}) = \sum_{h=1}^{N} (\frac{1}{p_h} - 1) D_{tih}^2 \quad < \quad V_{ti}^* ,$$

then we assign probabilities of selection so that

$$p_h \propto \hat{D}_{tih} .$$

From Section 2, we have

$$p_h = \hat{D}_{tih} \frac{\sum_k \hat{D}_{tik}}{V_{ti}^* + \sum_k \hat{D}_{tik}^2} . \qquad (5)$$

3.3.2. Multivariate Case

For the multivariate case, it is unlikely that (5) can be satisfied for all t and all i for every establishment. One approach to the multivariate case is to take

$$p_h \propto \hat{D}_h = \sqrt{\sum_{ti} \hat{D}_{tih}^2} .$$

D_h is referred to as a measure of size for establishment h. The probabilities are then calculated as a function of the budgeted expected sample size, that is,

$$p_h = \frac{E(n) \hat{D}_h}{\sum_{h=1}^{N} \hat{D}_h} .$$

This is the approach that was used to select a new ASM sample based on 1987 census data.

A different approach is to use Chromy's algorithm, as described in Section 2, to calculate the probabilities of selection. Each establishment is treated as a stratum, and the same method of calculating the stratum samples sizes, the n_h's, is used to calculate the probabilities, the p_h's. Replacing the notation in Section 2 with the notation we have been using for this particular problem, the four steps outlined in Section 2 for carrying out Chromy's algorithm are

STEP 1: For all t and all i, set $\lambda_{ti} = 1$ for the first iteration.

STEP 2: Calculate the $p_h = \sqrt{\sum_{ti} \lambda_{ti} \hat{D}_{tih}^2}$.

STEP 3: Calculate the $\hat{V}(\hat{D}^{(t,i)}) = \sum_{h=1}^{N} \hat{D}_{tih}^2 (\frac{1}{p_h} - 1)$.

STEP 4: Calculate the revised λ_{ti} , denoted λ_{ti}' , using the updating equation

$$\lambda_{ti}' = \lambda_{ti} \left[\frac{\hat{V}(\hat{D}^{(t,i)}) + \sum_{h=1}^{N} \hat{D}_{tih}^2}{V_{ti}^* + \sum_{h=1}^{N} \hat{D}_{tih}^2} \right]^2 .$$

Steps 2 through 4 may be repeated (with λ_{ti}' each time replacing λ_{ti}) until convergence or until a defined stopping point is reached. Convergence occurs when

$$\sum_{h=1}^{N} \hat{D}_{tih}^2 (\frac{1}{p_h} - 1) \leq V_{ti}^*$$

and

$$\lambda_t \left(\sum_{h=1}^{N} \hat{D}_{tih}^2 \left(\frac{1}{p_h} - 1\right) - V_{ti}^*\right) = 0$$

for all t and i.

Stopping rules for Chromy's algorithm are described in Section 2 and in Appendix A. We used Chromy's algorithm instead of the algorithm by Causey (1983) or Bethel (1989) because it was the easiest to program. The ability of a SAS implementation of Chromy's algorithm to handle a very large number of strata (i.e., establishments for the ASM problem) was especially appealing. (The implementation of Bethel's algorithm by Mergerson (1988) contains several arrays indexed by strata and for the ASM problem would have exceeded available memory.) In addition, Bethel (1989) observed that based on several comparisons Chromy's algorithm appears to converge faster than Bethel's algorithm.

4.0. Results of Testing with 1987 Census of Manufactures Data

We conducted eight test runs of Chromy's algorithm with 1987 Census of Manufactures data. Runs 1 through 4 executed Chromy's algorithm as it is described in Section 3. The first run involved non-certainty establishments and an original set of constraints on variances of product class estimates. The second run involved non-certainty establishments and a revised set of constraints on variances of product class estimates. The third run was a continuation of the second run, but we added in a set of constraints on variances of industry estimates. The fourth run involved non-certainty and certainty establishments and product class and industry constraints.

The first three runs completed successfully but required a large amount of CPU time (26, 56, and 69 hours) and elapsed time (31, 74, and 94 hours). For these three runs, we stopped the iterations of Chromy's algorithm when we were within 500 of the minimum expected sample size (as indicated by equation (4)). Very few constraints were not satisfied upon completion of these runs. We stopped the fourth run before we were within 500 of the minimum expected sample size because convergence was extremely slow.

Because of the algorithm's slow convergence, we contacted Dr. Chromy and asked him if he knew of ways to speed up the algorithm's convergence. Dr. Chromy suggested a slight change in the way the lambdas are updated in each iteration. This change would allow some lambdas to converge to zero very quickly and would allow them to again become positive if necessary. (See Appendix A.) We made this change and performed four more test runs (Runs 5, 6, 7, and 8). The fifth run involved non-certainty establishments and the product class and industry constraints as in Run 3. The sixth run involved non-certainty and certainty establishments and product class and industry constraints. The seventh run involved non-certainty establishments with constraints on product class, industry, and selected state-level estimates. The eighth run was a continuation of the seventh.

All four of these runs completed successfully, and the reduction in CPU and elapsed time was large. Runs 5, 6, and 7 were stopped when we were within 500 of the minimum expected sample size (as indicated by equation (4)). Run 8 was stopped when we were within 50 of the minimum expected sample size. Very few constraints were not satisfied upon completion of these runs, and resulting probabilities were adjusted so that all specified constraints were satisfied. Zayatz and Sigman (1993), provides additional details and discussion on runs. In comparing Runs 1 through 4 with Runs 5 through 8, we observed that Runs 1 through 4 converged from above -- that is, subsequent iterations produced smaller and smaller expected sample sizes -- whereas, Runs 5 through 9 converged from below -- subsequent runs produced larger and larger expected sample sizes.

5.0. Conclusions and Recommendations

The test results discussed in Section 4 show that it is feasible to use Chromy's algorithm to assign probabilities of selection so that constraints on the variances of product-class, industry, and selected state-level estimates are satisfied.

In the 1987 ASM redesign, the Census Bureau assigned probabilities and selected the new sample in a two-stage process. The first stage ensured that variance constraints on product class estimates were met, and the second stage ensured that variance constraints on industry estimates were met.

Chromy's algorithm offers the benefit of minimizing the expected sample size by enforcing variance constraints on product-class, industry, and selected state-level estimates **simultaneously**. We recommended that the algorithm (in its modified version found in Appendix B) be used to assign probabilities of selection in the 1992 ASM redesign.

6.0. References

1. Bethel, J., "Sample Allocation in Multivariate Surveys," Survey Methodology, Vol. 15, 1989, pp.

47-57.
2. Causey, B.D., "Computational Aspects of Optimal Allocation in Multivariate Stratified Sampling," SIAM Journal of Scientific and Statistical Computing, Vol. 4, 1983, pp. 322-329.
3. Chromy, J., Design Optimization with Multiple Objectives, "Proceedings of the Survey Research Methods Section, American Statistical Association, 1987, pp. 194-199.
4. Dee, T. and Dorinski, S. (1993), "1994 ASM Sample Design Research," unpublished manuscript, Research and Methodology Staff, Industry Division, Bureau of the Census, Washington, DC 20233.
5. Dorinski, S. (November 17, 1992), "SAS Files For ASM Sample Design Research," unpublished manuscript, Research and Methodology Staff, Industry Division, Bureau of the Census, Washington, DC 20233
6. Dorinski, S. (December 14, 1992) "Updated SAS Files For ASM Sample Design Research," unpublished manuscript, Research and Methodology Staff, Industry Division, Bureau of the Census, Washington, DC 20233.
7. Mergerson, J.W., "ALLOC.P: A Multivariate Allocation Program," The American Statistician, Vol. 42, 1988, p. 85.
8. Ogus, J. and Clark, D. (1971), "Technical Paper 24: The Annual Survey of Manufactures: A Report on Methodology," U.S. Department of Commerce, Bureau of the Census, Washington, DC.
9. Waite, P. J. and Cole, S. J. (1980), "Selection of a New Sample Plan for the Annual Survey of Manufactures," Proceedings of the Social Statistics Section, American Statistical Association, pp. 307-311.
10. Zayatz, L. and Sigman, R. (1993), "Feasibility Study of the Use of Chromy's Algorithm in Poisson-Sample Selection for the Annual Survey of Manufacturers," ESMD Report Series ESMD-9305, August 1993, Economic Statistical Methods Division, Bureau of the Census, Washington, DC 20233.

APPENDIX A
MODIFIED CHROMY'S ALGORITHM

STEP 0: For all t and all i, calculate

$$a_{ti} = \text{"univariate } \lambda_{ti}\text{"} = \left[\frac{\sum_k \hat{D}_{tik}}{V_{ti}^* + \sum_k \hat{D}_{tik}^2} \right]^2 .$$

STEP 1': For all t and all i set

b_{ti} = scaling factor = 1,

and

$$\lambda_{ti} = a_{ti} b_{ti} .$$

STEP 2 and STEP 3 remain unchanged.

STEP 4': Calculate the revised scaling factors b_{ti}:

$$b_{ti}' = \begin{cases} b_{ti} c_{ti}^2 & \lambda_{ti} > 0 \\ 1 & \hat{V}(\hat{D}^{(t,i)}) > V_{ti}^* \\ & \lambda_{ti} = 0 \end{cases}$$

where

$$c_{ti} = \frac{\sum_{h=1}^{N} \frac{\hat{D}_{tih}^2}{p_h}}{V_{ti}^* + \sum_{h=1}^{N} \hat{D}_{tih}^2} .$$

(If $\lambda_{ti} = 0$ and $\hat{V}(\hat{D}^{(t,i)}) \le V_{ti}^*$, it is not necessary to update b_{ti}.) and revised λ_{ti} :

$$\lambda_{ti}' = \begin{cases} 0 & \lambda_{ti} = 0 \\ & \hat{V}(\hat{D}^{(t,i)}) \le V_{ti}^* \\ 0 & \lambda_{ti} > 0 \\ & b_{ti} \le \epsilon \\ & \hat{V}(\hat{D}^{(t,i)}) \le V_{ti}^* \\ a_{ti} b_{ti}' & \lambda_{ti} > 0 \\ & b_{ti} > \epsilon \\ a_{ti} b_{ti}' = a_{ti} & \lambda_{ti} = 0 \end{cases} .$$

We used $\epsilon = 0.005$.

APPENDIX B
DERIVATION OF ADJUSTMENT FORMULA

The computer time needed to determine selection probabilities with Chromy's algorithm can be reduced by stopping the algorithm before it has completely converged. If this is done, then one or more constraints from the set of all constraints may not be satisfied. Our solution to this problem is to add a post-processing step that slightly inflates some of the selection probabilities so that all constraints are satisfied. Probabilities associated with establishments involved only in constraints that were satisfied will be slightly decreased in this process. The remainder of

this appendix derives a formula for this adjustment.

The variance constraints are

$$\sum_h \hat{D}_{tih}^2 (\frac{1}{p_h} - 1) \leq V_{ti}^*; \; t=1,2,3; \; i=1,2,\ldots,I_t$$

where

t indexes the type of estimate i.e. type of shipments breakdown,

i indexes the category within a breakdown,

h indexes the establishment,

$\hat{D}_{tih}^2$ = the predicted squared difference between the census year and the sample year for shipments by establishment h in category i of breakdown t,

p_h = the selection probability for establishment h, and

V_{ti}^* = the target variance for category i of breakdown t.

Given a set of selection probabilities, let

R_{ti} = the ratio of the predicted design variance to the target variance for category i of breakdown t. Then

$$\sum_h \hat{D}_{tih}^2 (\frac{1}{p_h} - 1) = R_{ti} V_{ti}^*$$

or

$$\sum_h \hat{D}_{tih}^2 \frac{(\frac{1}{p_h} - 1)}{R_{ti}} = V_{ti}^*.$$

Let

$$r_{hti} = \begin{cases} R_{ti} \text{ if establishment } h \text{ has shipments in category } i \text{ of breakdown } t \\ 1 \text{ otherwise (so } \hat{D}_{tih} = 0). \end{cases}$$

Then

$$\sum_h \hat{D}_{tih}^2 \frac{(\frac{1}{p_h} - 1)}{r_{hti}} = V_{ti}^*.$$

If the constraint for category i of breakdown t is not satisfied, then $r_{hti} > 1$ for $\hat{D}_{tih}^2 > 0$. Hence, all constraints will be satisfied if

$$\sum_h \hat{D}_{tih}^2 (\frac{1}{p_h'} - 1) \leq \sum_h \hat{D}_{tih}^2 \frac{(\frac{1}{p_h} - 1)}{r_{hti}} = V_{ti}^*;$$

that is

$$(\frac{1}{p_h'} - 1) \leq \frac{(\frac{1}{p_h} - 1)}{r_{hti}}.$$

Solving for p_h' gives

$$p_h' \geq \frac{r_{hti}}{r_{hti} + \frac{1}{p_h} - 1}. \qquad (B\text{-}1)$$

The right-hand side of (B-1) is a decreasing function of r_{hti}. Thus, (B-1) is satisfied by

$$p_h' = \frac{\max_{t,i} r_{hti}}{\max_{t,i} r_{hti} + \frac{1}{p_h} - 1}. \qquad (B\text{-}2)$$

It also follows from this result that if $p_h < 1$ then $p_h' < 1$.

We have written a SAS program that takes as input the probabilities resulting from the application of Chromy's algorithm and the ratios of expected coefficients of variation to target coefficients of variation for product-class, industry, and selected state-level estimates. The program then adjusts the probabilities according to (B-2) so that all constraints are satisfied.

---

* 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 relect those of the Census Bureau.