REFERENCES

ABC, Ann Arbor, MI: Consortium Software.
ANOVA-TM, Dearborn, MI: Advanced Systems and Designs.
Crisp, San Francisco: Crunch Software.
Lotus 1-2-3, Cambridge, MA: Lotus Development.
NWA Statpak, Portland, OR: Northwest Analytical.
SAS/PC, Cary, NC: SAS Institute.
Searle, S. R. (1971), *Linear Models*, New York: John Wiley.
SPSS/PC+, Chicago, IL: SPSS.
Statgraphics, Princeton, NJ: Statistical Graphics.
Statplan, Glastonbury, CT: Futures Group.

Computer Selection of Size-Biased Samples

DAVID R. FOX*

This article describes a method for producing size-biased probability samples as originally proposed by Hanurav (1967) and Vijayan (1968). The complexity of the procedure has led to the development of microcomputer software that greatly facilitates the production of sampling plans as well as the computation of population estimates.

KEY WORDS: πps sampling; Size-biased sampling; Survey sampling.

1. INTRODUCTION

Whilst procedures such as simple random sampling, systematic sampling, stratified sampling, cluster sampling, and other survey sampling techniques are well understood by most statisticians, the same is perhaps not true of unequal probability sampling. Cochran (1977) devoted one chapter to the topic of subsampling with units of unequal sizes, although a far more comprehensive treatment was provided by Brewer and Hanif (1983).

In unequal probability sampling we generally assume that we have available some auxiliary measure x (typically a measure of size) on each of the N population units and that x is moderately to strongly correlated with y, the variable of interest (in particular we assume that x and y have the "ratio property"; i.e., $^{x}i/y_i$ is approximately constant for all i).

Our aim is to obtain a sample drawn *without* replacement such that the probability that the ith population unit is included in the sample is proportional to x_i. For example, in estimating the sales at a number of shopping centers we may already have information on the total floor space at each location. If one can reasonably assume that floor space (x) and sales (y) possess the aforementioned ratio property, then we may expect to improve the precision with which the total sales (Y) can be estimated by using our knowledge of x.

The use of auxiliary information in survey sampling is commonplace. Two standard estimators of a population total are provided by (a) the ratio estimator and (b) the regression estimator. Ordinarily these estimators are associated with random samples (i.e., equal probability sampling).

The use of unequal probabilities in sampling was first suggested by Hansen and Hurwitz (1943), who considered the sampling with replacement case. In a later development, Horvitz and Thompson (1952) provided an unbiased estimator of Y in the sampling without replacement case. Sen (1953) and Yates and Grundy (1953) independently derived a conditionally unbiased estimator of the variance of the Horvitz–Thompson estimator for samples of fixed size n. In the years that have followed, a plethora of papers have emerged detailing the mechanics of obtaining a probability proportional to size (or πps) sample in both the sampling with and without replacement cases. Brewer and Hanif (1983) presented a catalog of approximately 50 of these procedures. Unfortunately, many of the procedures are unnecessarily restrictive in that either sampling is with replacement (which can lead to inefficiencies) and/or the sample size is limited to 2. A general πpswor (inclusion probability proportional to size, sampling without replacement) procedure applicable for any n was described by Vijayan (1968) and represented a generalization of earlier work for the $n = 2$ case by Hanurav (1967). The scheme has thus come to be known as the Hanurav–Vijayan method. In the following sections I give some basic results for πps sampling and describe the Hanurav–Vijayan method and the associated computer software.

2. πps SAMPLING—SOME BASIC RESULTS

In the following, assume that we have a population of N units for which a measure of size, x, is available for each. From this population a probability sample is taken. We denote by π_i the probability that unit i is included in the sample and by π_{ij} the probability that units i and j appear together in our sample.

Let $\{S\}$ denote the (unordered) sample—that is, the set of *distinct* units in the sample, and let m_s be the effective sample size (cardinality of S). We define a *fixed effective sample size* design as one in which m_s is constant. If this constant is n, then we denote the resulting sample as FES(n).

For any FES(n) design the following identities hold:

$$\sum_{i=1}^{N} \pi_i = n, \qquad \sum_{j \neq i}^{N} \pi_{ij} = (n-1)\pi_i,$$

*David R. Fox is Lecturer, School of Mathematics and Statistics, Curtin University of Technology, Perth, Western Australia, Australia.

and

$$\sum_{i<j}^{N}\sum \pi_{ij} = \frac{n(n-1)}{2}.$$

The quantity of interest is Y, the population total of the y's; that is,

$$Y = \sum_{i=1}^{N} Y_i. \tag{2.1}$$

When sampling is without replacement the standard estimator of Y is the Horvitz–Thompson estimator $\hat{Y}_{ht}$:

$$\hat{Y}_{ht} = \sum_{i=1}^{n} \frac{Y_i}{\pi_i}. \tag{2.2}$$

The Sen–Yates–Grundy (SYG) estimator of the variance of $\hat{Y}_{ht}$ is

$$\underset{\text{SYG}}{\text{var}}[\hat{Y}_{ht}] = \sum_{\substack{i,j=1\\ j>i}}^{n}\sum^{n} \frac{\pi_i\pi_j - \pi_{ij}}{\pi_{ij}}\left[\frac{y_i}{\pi_i} - \frac{y_j}{\pi_j}\right]^2. \tag{2.3}$$

Note that one potential problem with Equation (2.3) is that it can assume negative values, although this situation is avoided if the sampling procedure ensures that $\pi_{ij} < \pi_i\pi_j$.

3. THE HANURAV–VIJAYAN PROCEDURE

Without loss of generality assume that $x_1 \le x_2 \le \cdots \le x_N$. Vijayan originally gave his sampling procedure in two parts. In the first case it was assumed that $x_{N-n+1} = x_N$; the more general case $x_{N-n+1} < x_N$ was treated separately. I shall review only the latter case.

Vijayan's procedure satisfies the following three basic requirements of a πps sampling scheme:

1. $\pi_i = np_i = n(x_i/X)$.
2. Each sample contains n distinct units.
3. $\pi_{ij} > 0$.

In addition, experience has shown that the following two properties are usually satisfied by the Hanurav–Vijayan procedure:

4. $\pi_{ij} \le \pi_i\pi_j$, which ensures nonnegativity for $\text{var}_{\text{SYG}}[\hat{Y}_{ht}]$.
5. $\pi_{ij}/\pi_i\pi_j > \beta$, β not too close to 0, which ensures stability of $\text{var}_{\text{SYG}}[\hat{Y}_{ht}]$.

Chandhuri and Voss (1988) noted, however, that for the aforementioned general case the nonnegativity property of requirement 4 is not guaranteed. Also note by virtue of requirement 1, values of the auxiliary variable x must be such that

$$\max_i x_i \le \frac{1}{n}\sum_{i=1}^{N} x_i = \frac{X}{n}.$$

Sampling Procedure

Step 1. Choose one of the integers $1, 2, \ldots, n$ with probabilities $\theta_1, \theta_2, \ldots, \theta_n$, where

$$\theta_i = n(p_{N-n+i+1} - p_{N-n+i})\frac{S + ip_{N-n+1}}{S}, \tag{3.1}$$

$p_j = x_j/X$, $S = \sum_{j=1}^{N-n} p_j$, and by definition $p_{N+1} = 1/n$ (this choice ensures that $\sum_{i=1}^{n}\theta_i = 1$).

Step 2. If the integer at Step 1 is i, then the last $(n-i)$ population units form part of the sample and the remaining i units are selected in accordance with Steps 3–6.

Step 3. Define new normed measures of size p_j^* for the $(N-n+i)$ population units as yet unselected, where

$$p_j^* = \frac{p_j}{S + ip_{N-n+1}}, \quad j \le N-n+1$$

$$= \frac{p_{N-n+1}}{S + ip_{N-n+1}}, \quad N-n+1 < j \le N-n+i.$$

Step 4. Select the first of the remaining i units from the first $(N-n+1)$ population units with probabilities equal to $a_j(1)$, where $a_1(1) = np_1^*$,

$$a_j(1) = np_j^*\prod_{k=1}^{j-1}[1-(i-1)P_k], \quad j = 2, \ldots, N-n+1,$$

and $P_k = x_k/(x_{i+1} + x_{i+2} + \cdots + x_N)$.

Step 5. Let j_1 be the population unit selected at Step 4. Then select the second unit from the (j_1+1)th up to the $(N-n+2)$nd with probabilities equal to $a_j(2, j_1)$, where

$$a_{j_1+1}(2, j_1) = (i-1)p_{j_1+1}^*$$

and

$$a_j(2, j_1) = (i-1)p_j^*\prod_{k=j_1+1}^{j-1}[1-(i-2)P_k],$$

$$j = j_1+2, \ldots, N-n+2.$$

Step 6. Repeat Step 5 until the last sample unit is selected. In general, if the $(l-1)$st of the i units remaining to be selected at the end of Step 3 is chosen to be the j_{l-1}th population unit, then select the lth unit from the $(j_{l-1}+1)$st up to the $(N-n+l)$th with probabilities equal to $a_j(l, j_{l-1})$, where

$$a_{j_{l-1}+1}(l, j_{l-1}) = (n-l+1)p_{j_{l-1}+1}^*$$

and

$$a_j(l, j_{l-1}) = (n-l+1)p_j^*\prod_{k=j_{l-1}+1}^{j-1}[1-(i-l)P_k],$$

$$j = j_{l-1}+1, \ldots, N-n+l.$$

Inclusion Probabilities

The inclusion probabilities as given by Vijayan are

$$\pi_i = np_i \tag{3.2}$$

and

$$\pi_{ij} = \sum_{r=1}^{n}\theta_r K_{ij}^{(r)},$$

where

$$K_{ij}^{(r)} = 1, \qquad N-n+r < i \le N-1$$

$$= \frac{rp_{N-n+1}}{S + rp_{N-n+1}},$$

$$N - n < i \le N - n + r, j > N - n + r$$

$$= \frac{rp_i}{S + rp_{N-n+1}}, \quad 1 \le i \le N - n, j > N - n + r$$

$$= \pi_{ij}^{(r)}, \qquad j \le N - n + r,$$

and

$$\pi_{ij}^{(r)} = \frac{r(r-1)}{2} \prod_{k=1}^{i-1} (1 - P_k) P_i p_j. \tag{3.3}$$

4. COMPUTER SOFTWARE

Clearly, the sampling procedure outlined in Section 3 is tedious and, for all but the smallest problem, would be too time-consuming to attempt by hand. The mechanics of the Hanurav–Vijayan procedure have been programmed for IBM and compatible computers. Features of this software are described next.

The UPS Sample Generator

The system provided goes far beyond the execution of the rather tedious calculations required by the Hanurav–Vijayan method. Incorporated into the software is a data-base and file-management system that greatly facilitates the task of keeping track of several survey plans. The salient features of the program include the following:

1. efficient data-base routines for sorting, merging, and appending files
2. full screen editing capability
3. a completely interactive structure, requiring little or no training
4. data entry from either keyboard or file (will read from delimited ASCII files produced by dBaseIII, Wordstar, Wordperfect, Multiplan, Lotus 1-2-3, Minitab, SPSS-X, SAS, GAUSS, GLIM, and others)
5. full set of file utilities, including creation of data-base files, file deletion—either individual or bulk (e.g., deletion of all files created between two specified dates), file backup and restoration, data editing, and file searches and manipulation.

I illustrate the use of this software using data on corn production taken from Battese, Harter, and Fuller (1988). These data report the actual number of hectares of corn planted in 12 Iowa counties and the corresponding number of pixels obtained from a Landsat image. In the present context, the Landsat data are the auxiliary variable (x) and we are interested in estimating the total number of hectares of corn (Y) for the 12 counties (one normally does not have available any prior information on the y's).

Invoking the UPS software brings up the following main menu:

1. Set/Modify environment
2. Create file and input X-data
3. Produce a sampling plan
4. Input Y-data and compute population estimates
5. Edit information
6. Print information
7. File management tools
9. Quit

Enter selection →

Note that the Landsat data had previously been entered into the data base. To produce a sampling plan we select option 3 of the menu. After identifying the data-base file name, a brief description of the file and its present status is displayed:

FILENAME	*FILE DESCRIPTION*	*CREATED*	*N(pop)*	*N(samp)*	*YDATA?*
corn	Data taken from JASA—Landsat corn data	06/14/88	36	0	.F.

Apart from displaying the file description and the date the file was created, the output indicates that the population size is 36 and that no sample has yet been constructed for these data. The .F. appearing under the column headed YDATA? is a true/false variable to indicate whether the survey data have been entered into the data base. In this case they have not. On this particular occasion we requested that a sample of size $n = 6$ be selected from the 36 population units. The following output shows which units were selected:

CASE #	*X*
2	209.000
4	432.000
15	459.000
29	343.000
30	342.000
31	294.000

At this stage one would normally go out into the field and survey the chosen population units. Once the sample (y) data have been collected, they are then entered into the data base:

$X2$	= 209.000	Y	*96.320*
$X4$	= 432.000	Y	*185.350*
$X15$	= 459.000	Y	*206.390*
$X29$	= 343.000	Y	*109.910*
$X30$	= 342.000	Y	*122.660*
$X31$	= 294.000	Y	*104.210*

The Y-data shown in italics were entered manually from the keyboard.

Finally, after some calculation the following population estimates are produced and displayed:

	Yhat	*Standard error*
Hanurav–Vijayan	4,217.81	247.4806
Ratio estimate	4,230.925	266.5357
Regression est.	4,105.565	282.5032

Using the published data we are able to determine that the actual total of the y's is 4,363.4.

The variance estimate for the Hanurav–Vijayan procedure is computed using Equation (2.3), whereas simple random sampling formulas are used for the variance of the ratio and regression estimators. The latter two estimates are provided for comparison only and do not reflect the unequal proba-

bility nature of the sample. We see from the displayed output that the estimates are reasonably close, with the Hanurav–Vijayan estimate having the smallest standard error.

[*Received June 1988. Revised March 1989.*]

REFERENCES

Battese, G. E., Harter, R. M., and Fuller, W. A. (1988), "An Error-Components Model for Prediction of County Crop Areas Using Survey and Satellite Data," *Journal of the American Statistical Association*, 83, 28–36.

Brewer, K. R. W., and Hanif, M. (1983), *Sampling With Unequal Probabilities* (Lecture Notes in Statistics), New York: Springer-Verlag.

Chandhuri, A., and Vos, J. W. E. (1988), *Unified Theory and Strategies of Survey Sampling*, Amsterdam: North-Holland.

Cochran, W. G. (1977), *Sampling Techniques*, New York: John Wiley.

Hansen, M. H., and Hurwitz, W. N. (1943), "On the Theory of Sampling From a Finite Population," *The Annals of Mathematical Statistics*, 14, 333–362.

Hanurav, T. V. (1967), "Optimum Utilization of Auxiliary Information: πps Sampling of Two Units From a Stratum," *Journal of the Royal Statistical Society*, Ser. B, 29, 374–391.

Horvitz, D. G., and Thompson, D. J. (1952), "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 77, 89–96.

Sen, A. R. (1953), "On the Estimate of Variance in Sampling With Varying Probability," *Journal of the Indian Society of Agricultural Statistics*, 7, 57–69.

Vijayan, K. (1968), "An Exact πps Sampling Scheme: Generalization of a Method of Hanurav," *Journal of the Royal Statistical Society*, Ser. B, 30, 556–566.

Yates, F., and Grundy, P. M. (1953), "Selection Without Replacement From Within Strata With Probability Proportional to Size," *Journal of the Royal Statistical Society*, Ser. B, 15, 253–261.

Statistical Computing Software Reviews

KENNETH BERK, Section Editor

This section is similar in organization to a book-review section in other journals; however, software of interest to statisticians is the subject of review here. Emphasis is on software for microcomputers. Programs that operate only in larger mainframe computers will seldom receive review. Normally, producers of programs make a copy of their product available to the section editor, who then selects one or more persons to test the product and prepare a review.

Producers of computer software who wish to have their product reviewed are invited to contact Section Editor Kenneth Berk, Department of Mathematics, 313 Stevenson Hall, Illinois State University, Normal, Illinois 61761.

Findings and opinions expressed in every review are solely those of the author. They should not be construed as reflecting endorsement of the product, or opinions held, by the American Statistical Association, nor is any warranty implied about any product reviewed.

|STAT (Version 5.3)

Available from Gary Perlman, Department of Computer and Information Science, Ohio State University, Columbus, OH 43210. $20 for UNIX C source and on-line documents; $15 for MS–DOS executables and on-line documents; $10 for 100-page manual.

1. Introduction

|STAT is a collection of 29 data-management and data-analysis programs for MS–DOS and UNIX. First introduced in 1980 in a UNIX version, the programs were converted for use on MS–DOS machines in 1985. This review focuses on the MS–DOS implementation.

The programs are inexpensive and require minimal resources. They will run on a machine with one diskette drive and 96K of memory, and they do not require or use graphics hardware. They may be distributed freely provided there is no material gain to the person making the copies, mass distribution (such as electronic bulletin boards) is not used, and the information in the flyer accompanying |STAT accompanies the copies. Copying for use in classrooms is expressly suggested.

|STAT is neither menu-driven nor command-driven. It follows the UNIX philosophy for module design; that is, each module should do only a particular task. This allows users to combine the executions of modules to perform a desired analysis. No software that a user already has (such as an editor) is duplicated. This is a tremendous strength in its overall design. |STAT uses only ASCII files for input to and output from all of its modules. This means that any results of any module can form the input to any other module. Readers who have struggled with traditional packages should recognize what a powerful concept this is. SAS and other major packages provide some of their output in proprietary system data sets, other output in listing files, and some tools for reading listings. Multiple programming techniques must be used to handle the variety of output formats, with the result that most users just work with the results that can be made available using standard techniques. In |STAT, all results are always in a consistent format, available for additional processing.

|STAT modules are executed from the MS–DOS command prompt or MS–DOS batch files. Results are displayed on the terminal or put in files, at the user's request. In effect, MS–DOS batch-file language is the command language for |STAT. Simple batch files created to execute |STAT analyses could contain only |STAT module invocations, but any MS–DOS batch command can be used. The examples in the |STAT documentation use several of the MS–DOS batch commands. |STAT makes full use of the standard input and standard output files in MS–DOS as well as full use of MS–DOS pipes. Much power and flexibility is achieved by relying on the tools the operating system provides. Users do not have to learn special |STAT conventions for printing, creating files, and editing. Standard MS–DOS utilities are used for each.

There are two major disadvantages to relying so heavily on the facilities of the operating system, and they are far outweighed by the advantages achieved in power and flexibility. The first disadvantage is that many users do not know the MS–DOS conventions for pipes and standard input and output. They do not create batch files, do not use the MS–DOS copy command, and do not