SAMPLING PHYSICIAN PRACTICES FOR A NATIONAL VASECTOMY SURVEY

Paul M. Gargiullo, Lisa M. Koonin, Centers for Disease Control, and Shelby Marston-Ainley, Libby Antarsh, Association for Voluntary Surgical Contraception
Paul M. Gargiullo, Division of Reproductive Health, K-21, Centers for Disease Control, Atlanta, GA 30333

KEY WORDS: Single-Stage Cluster Sampling, Probability Proportional to Size, PPS, Subgroup Estimation.

INTRODUCTION

The Association for Voluntary Surgical Contraception conducted a national vasectomy survey, with technical assistance from the Centers for Disease Control, to estimate the total number of vasectomies performed in the United States in 1991, the number of physician practices that performed vasectomies, the average number of physicians per practice, the average number of vasectomies performed per physician (by specialty), occlusion method, and type of anesthesia used. The survey was restricted to urologists, general surgeons, and family practitioners because these specialists include almost all the practitioners believed to perform vasectomies. The survey included only nonfederal physicians and medical residents involved in patient care (i.e., office-based and hospital-based practices).

Previous surveys of vasectomy incidence considered individual physicians as sample units (Kendrick et al., 1987; Orr et al., 1985). However, managers of medical practices told us that it is easier to report the number of vasectomies performed per practice than per physician within a practice. By sampling physician practices, data can be captured for one or more physicians, for the cost of a single respondent contact. Furthermore, practices are less likely than individual physicians to have frequent changes of address and telephone numbers. Therefore, the sample units for this survey were practices, which are defined as individual and group private practices, medical school departments, and divisions of health maintenance organizations. A sample of practices was drawn by way of a sample of physicians listed in the American Medical Association's (AMA) Physician Master File, the most comprehensive and current list available of physicians and medical residents nationwide. The file contains name, address, historical data, medical specialty, and other details of current professional activities. It is periodically updated through mail-in questionnaires, and information from hospitals, medical societies, state licensing agencies, and other sources.

This article deals with the sample design and theory for estimating the total number of vasectomies performed in 1991. We conclude by computing sample sizes using data from a pilot survey. The survey is ongoing, and results should be available by fall of 1992.

MEASURES OF SIZE AND THE SELECTION RULE

Since a practice is a "cluster" of physicians, sampling of practices can be thought of as single-stage cluster sampling. The information collected applies to the entire cluster rather than to individual physicians (second-stage units). In single-stage cluster sampling, if the variable of interest is positively correlated with some measure of size (MOS) for the clusters, then it is statistically efficient to select clusters with probability proportional to MOS. The result is a probability-proportional-to-size (PPS) cluster sample. A frequently used MOS is the number of second-stage units within the clusters. We believe that the number of vasectomies performed in a practice is at least weakly correlated with the number of physicians in the practice. Therefore, we sampled practices with probability proportional to the number of physicians in the practices. The i^{th} practice has an MOS designated by M_{hi}, which is the number of stratum h physicians (e.g., urologists in the northeast census region) who consider practice i to be their primary practice. M_{hi} includes all stratum h physicians in practice i, whether or not they perform vasectomies. (Although it would be efficient to restrict sampling to practices that perform vasectomies, we can not know in advance which practices these

are.) One or more strata may be represented in a single practice. For example, both family practitioners and general surgeons may work in a large group practice. Then, the value of M_{hi} is different for each specialty in the group practice. In fact, we considered the different specialties within a group practice to represent different practices in a statistical sense.

We selected a sample of practices from within each stratum so that selection probabilities were proportional to the M_{hi}. This is accomplished by the following selection rule. Physician records were selected with equal probability and with replacement from within each of 12 strata of the file -- four census regions by the three specialties (Table 1). A practice was considered selected when a physician in that practice was selected in a single draw from the file. This allowed us to compute probabilities of selection for all sample units. For practice i, stratum h, the "single draw" probability of selection was

$$z_{hi} = M_{hi} / M_{ho}$$

where

$$M_{ho} = \sum_{i=1}^{N_h} M_{hi} \quad , \quad \sum_{i=1}^{N_h} z_{hi} = 1$$

and N_h the number of practices in the population of stratum h. Thus, the single-draw probability of selection for a practice in stratum h is proportional to the number of stratum-h physicians working at that practice as their primary practice. In sampling with replacement, a practice may appear in the sample in more than one draw.

A basic assumption of this sample design is that a practice of size M_{hi} has exactly M_{hi} physicians listed in the file. The names of the physicians in the practice do not have to match the names in the file, but the *number* of physicians in the practice must match the *number* of physician names in the file that are associated with the practice. We determined the size of each sampled practice M_{hi} by inquiring of the practice rather than by checking the file. The M_{ho} was obtained by summing listings of physicians within strata of the file. Alternatively, the M_{hi} could be computed strictly from the file, however, it would difficult to identify physicians in the same practice and group these physicians by practice to draw a PPS sample.

Although the M_{ho} are computed by tallying names within strata in the file, not all of these names can be associated with active practices because of retirement, death, or an erroneous listing. We refer to these listings as "dead" listings. Although the entire file could be cleaned of dead listings prior to sample selection, such listings could be identified in the sample. We considered a listing dead after extensive efforts to contact the practice failed. We handled the problem of dead listings as a problem in subgroup estimation. We used formulas for subgroup estimation in any event, because vasectomies were not performed in all practices, and those that did perform them comprised the subgroup of interest. Sampled practices that did not perform vasectomies were assigned a zero annual vasectomy count. Likewise, a dead listing found in the sample was considered to be a "practice" of size $M_{hi} = 1$, with a zero annual vasectomy count. This avoided statistical bias due to dead listings, but at the cost of increased variance.

POPULATION TOTAL

We now define the population or actual number of vasectomies performed in 1991.

- Y = the population total,
- h = the stratum indicator $(1,2,\ldots,L)$,
- Y_h = the population total within stratum h,
- N_h = the number of practices in the population in stratum h,
- N_h' = the number of practices in which vasectomies were performed in the population in stratum h, and
- y_{hi} = the response variable (number of vasectomies) associated with practice i in stratum h, where $i=1,2,\ldots,N_h$ or N'_h.

Practices in which vasectomies were performed are considered to be a subgroup of the total population of

practices ($N_h' \le N_h$). If a practice does not perform vasectomies, then $y_{hi} = 0$ for that practice. We define the population total as

$$Y = \sum_{h=1}^{L} Y_h = \sum_{h=1}^{L} \sum_{i=1}^{N_h} y_{hi} = \sum_{h=1}^{L} \sum_{i=1}^{N_h'} y_{hi} \qquad (1)$$

ESTIMATE OF TOTAL

The derivation of estimators for single-stage cluster sampling with PPS can be found in Cochran (1977) and Kish (1965). Modifications of these estimators are presented here to account for subgroup estimation (Kish, 1965; Durbin, 1958). If n_h practices are selected from stratum h, then an unbiased estimate of Y is

$$\hat{Y}_{ppz} = \sum_{h=1}^{L} \hat{Y}_{h_{ppz}} = \sum_{h=1}^{L} \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{y_{hi}}{z_{hi}} = \sum_{h=1}^{L} \frac{1}{n_h} \sum_{i=1}^{n_{hi}'} \frac{y_{hi}}{z_{hi}} \qquad (2)$$

where

$\hat{Y}_{h_{ppz}}$ = the estimated number of vasectomies performed in stratum h,

n_h = the number of practices sampled, and

n_h' = the number of sampled practices that performed vasectomies ($n_h' \le n_h$).

VARIANCE OF $\hat{Y}_{ppz}$

From the standard formulas for stratified PPS sampling, we have

$$V(\hat{Y}_{ppz}) = \sum_{h=1}^{L} V(\hat{Y}_{h_{ppz}}) = \sum_{h=1}^{L} \frac{1}{n_h} \sum_{i=1}^{N_h} z_{hi} \left(\frac{y_{hi}}{z_{hi}} - Y_h \right)^2 \qquad (3)$$

where

$y_{hi} = 0$ if no vasectomies were performed in practice i, stratum h.

Next we partition the sum of the squares of equation 3 for practices that performed vasectomies (N_h') and practices that did not ($N_h - N_h'$) :

$$\sum_{i=1}^{N_h} z_{hi} \left(\frac{y_{hi}}{z_{hi}} - Y_h \right)^2 = \sum_{i=1}^{N_h'} z_{hi} \left(\frac{y_{hi}}{z_{hi}} - Y_h \right)^2 + \sum_{i=1}^{N_h - N_h'} z_{hi} \left(\frac{0}{z_{hi}} - Y_h \right)^2 \qquad (4)$$

Now, since $\sum_{i=1}^{N_h'} z_{hi} + \sum_{i=1}^{N_h - N_h'} z_{hi} = 1$, we can derive the following expression:

$$\sum_{i=1}^{N_h - N_h'} z_{hi} \left(\frac{0}{z_{hi}} - Y_h \right)^2 = \sum_{i=1}^{N_h - N_h'} z_{hi} Y_h^2 = \left(1 - \sum_{i=1}^{N_h'} z_{hi} \right) Y_h^2 \qquad (5)$$

This leads to equation 6, which is an alternate expression for the variance that involves only those practices in the subgroup in which vasectomies were performed:

$$V(\hat{Y}_{ppz}) = \sum_{h=1}^{L} \frac{1}{n_h} \left[\sum_{i=1}^{N_h'} z_{hi} \left(\frac{y_{hi}}{z_{hi}} - Y_h \right)^2 + \left(1 - \sum_{i=1}^{N_h'} z_{hi} \right) Y_h^2 \right] \qquad (6)$$

The variance in equation 6 has two components. The first component reflects variability due to lack of correlation between a practice's response variable y_{hi} (number of vasectomies performed) and the practice's MOS M_{hi} (number of specialists in the practice). When correlation is perfect, which is ideal for PPS sampling, this component vanishes. The second component reflects the amount of uncertainty in finding practices in which vasectomies were performed. This component is large when most vasectomies were performed in either (a) a few practices or (b) small practices (i.e. z_{hi} is small). Few or small practices are less likely to be picked up in a PPS sample than numerous or large practices. This component becomes zero if all practices performed vasectomies.

ESTIMATION OF VARIANCE

Using the standard variance estimate for stratified PPS sampling with replacement, we have the following unbiased variance estimate:

$$\upsilon(\hat{Y}_{ppz}) = \sum_{h=1}^{L} \upsilon(\hat{Y}_{h_{ppz}}) = \sum_{h=1}^{L} \frac{\sum_{i=1}^{n_h} \left(\frac{y_{hi}}{z_{hi}} - \hat{Y}_{h_{ppz}} \right)^2}{n_h(n_h-1)} \quad (7)$$

where

$y_{hi} = 0$ if no vasectomies were performed in practice i, stratum h.

As we did in equation 6, we partition the sum of the squares from equation 7 into components for practices that performed vasectomies and practices that did not:

$$\sum_{i=1}^{n_h} \left(\frac{y_{hi}}{z_{hi}} - \hat{Y}_{h_{ppz}} \right)^2 = \sum_{i=1}^{n_h'} \left(\frac{y_{hi}}{z_{hi}} - \hat{Y}_{h_{ppz}} \right)^2 + \sum_{i=1}^{n_h - n_h'} \left(\frac{0}{z_{hi}} - \hat{Y}_{h_{ppz}} \right)^2 \quad (8)$$

where vasectomies were performed in n_h' of the n_h sampled practices.

Equation 8 leads to the following alternate variance estimator, which contains the sum of the squares for only those practices in which vasectomies were performed.

$$\upsilon(\hat{Y}_{ppz}) = \sum_{h=1}^{L} \frac{1}{n_h(n_h-1)} \times \left[\sum_{i=1}^{n_h'} \left(\frac{y_{hi}}{z_{hi}} - \hat{Y}_{h_{ppz}} \right)^2 + (n_h - n_h') \hat{Y}^2_{h_{ppz}} \right] \quad (9)$$

The first component of variability in equation 9 reflects lack of correlation between y_{hi} and z_{hi}, and the second component reflects the uncertainty in finding practices that perform vasectomies. The latter component becomes zero when all sampled practices perform vasectomies that is, when $n_h = n_h'$.

SAMPLE SIZE ESTIMATION AND OPTIMUM ALLOCATION

In this section we use traditional methods to estimate required sample size under two different constraints: (1) minimize the variance of the estimate under constant survey costs and (2) minimize the cost for constant variance. First we define a simple cost function for the survey.

$$C = c_0 + \sum_{h=1}^{L} n_h c_h \quad (10)$$

where

C = total survey cost (fixed plus variable costs),

c_0 = fixed costs (including personnel salaries and benefits, transportation, and other overhead expenses),

n_h = number of practices to be sampled from stratum h, and

c_h = cost per practice sampled from stratum h (e.g., mailing and telephone costs, printing and stationery costs, royalties and fees for use of the file, and mailing and data entry costs).

Given the large number of listings on the file (82,805), it is highly improbable that more than one physician would be selected from any given practice. Thus, for sample size estimation, the number of practices sampled from each stratum (n_h) can be equated with the number of listings selected from the file. The total sample size is given by

$$n = \sum_{h=1}^{L} n_h$$

The optimum allocation of n among the L strata is the set of n_h that minimizes the variance of the estimated total number of vasectomies $V(\hat{Y}_{ppz})$ subject to the constraint that

$$C - c_0 = \sum_{h=1}^{L} n_h c_h$$

We used Lagrange's method of undetermined multipliers to determine the n_h and the multiplier λ that minimizes the following expression for the variance and cost constraint:

$$G = \sum_{h=1}^{L} \frac{1}{n_h} \sum_{i=1}^{N_h} z_{hi} \left(\frac{y_{hi}}{z_{hi}} - Y_h \right)^2 + \lambda \left(\sum_{h=1}^{L} n_h c_h - C + c_0 \right) \quad (11)$$

For simplicity we used the unpartitioned equation 3 for $V(\hat{Y}_{ppz})$. In this form, $y_{hi} = 0$ for practices in which vasectomies were not performed.

We have L partial derivatives of the form

$$\frac{\partial G}{\partial n_h} = -\frac{\sum_{i=1}^{N_h} z_{hi}\left(\frac{y_{hi}}{z_{hi}} - Y_h\right)^2}{n_h^2} + \lambda c_h \qquad (12)$$

Setting all partial derivatives to zero and rearranging, we have

$$n_h = \sqrt{\frac{\sum_{i=1}^{N_h} z_{hi}\left(\frac{y_{hi}}{z_{hi}} - Y_h\right)^2}{\lambda c_h}} \qquad (13)$$

We eliminate the undetermined multiplier λ by dividing equation 13 by the sum of the n_h to obtain the following L allocation fractions:

$$\frac{n_h}{n} = \frac{\sqrt{\frac{1}{c_h}\sum_{i=1}^{N_h} z_{hi}\left(\frac{y_{hi}}{z_{hi}} - Y_h\right)^2}}{\sum_{h=1}^{L}\sqrt{\frac{1}{c_h}\sum_{i=1}^{N_h} z_{hi}\left(\frac{y_{hi}}{z_{hi}} - Y_h\right)^2}} \qquad (14)$$

The optimum allocation favors strata with either high variability or low cost per unit c_h.

To compute the total sample size (n) that minimizes the variance $V(\hat{Y}_{ppz})$ for a constant cost C, we first rearrange the allocation fractions (equation 14) so that only n_h is on the left side; then we substitute these n_h into the cost function (equation 10) to obtain

$$n = \frac{(C - c_0)\sum_{h=1}^{L}\sqrt{\frac{1}{c_h}\sum_{i=1}^{N_h} z_{hi}\left(\frac{y_{hi}}{z_{hi}} - Y_h\right)^2}}{\sum_{h=1}^{L}\sqrt{c_h\sum_{i=1}^{N_h} z_{hi}\left(\frac{y_{hi}}{z_{hi}} - Y_h\right)^2}} \qquad (15)$$

To compute the total sample size (n) that minimizes the total cost C for a constant variance $V(\hat{Y}_{ppz})$, we again rearrange equation 14 and substitute the optimum set of n_h into the formula for the variance (equation 3) to obtain

$$n = \frac{1}{V(\hat{Y}_{ppz})}\left(\sum_{h=1}^{L}\sqrt{\frac{1}{c_h}\sum_{i=1}^{N_h} z_{hi}\left(\frac{y_{hi}}{z_{hi}} - Y_h\right)^2}\right) \times \left(\sum_{h=1}^{L}\sqrt{c_h\sum_{i=1}^{N_h} z_{hi}\left(\frac{y_{hi}}{z_{hi}} - Y_h\right)^2}\right) \qquad (16)$$

Once n is found, the optimum allocation fractions (equation 14) can be used to determine the optimum set of n_h to sample from the strata. Equation 17 below suggests an estimate that can be "plugged into" the sample allocation formulas in equations 14 through 16. This estimate can be computed by using data from a pilot survey or from prior surveys.

$$\mathbf{E}\left[\frac{\sum_{i=1}^{n_h}\left(\frac{y_{hi}}{z_{hi}} - \hat{Y}_{h_{ppz}}\right)^2}{n_h - 1}\right] = \sum_{i=1}^{N_h} z_{hi}\left(\frac{y_{hi}}{z_{hi}} - Y_h\right)^2 \qquad (17)$$

SAMPLE SIZE COMPUTATIONS FOR VASECTOMY SURVEY

Sample size calculations and allocation among strata were performed to minimize $V(\hat{Y}_{ppz})$ while holding survey cost constant. Appropriate values were used in equation 15 in which C was the total survey cost ($62,000), and c_o was fixed survey costs ($54,538). The variable survey costs were $7,462 ($C - c_o$). Recall that the constant cost constraint requires that $C - c_o = \sum_{h=1}^{L} n_h c_h$. The average cost per practice or c_h was estimated at $4.90 for urology, $4.29 for general surgery, and $3.71 for family practice. General surgery and family practices were pre-screened by telephone to determine if vasectomies were performed in them, and if so, they were mailed questionnaires. Since most urologists perform vasectomies, these practices were first contacted by mail. Due to a low rate of return, follow-up telephone costs for urology strata were higher than anticipated.

The only quantities remaining for calculating sample size in equation 15 are $\sum_{i=1}^{N_h} z_{hi}\left(\frac{y_{hi}}{z_{hi}} - Y_h\right)^2$, which can be estimated using equation 17. Equation 17 was estimated for each medical specialty from a pilot survey of 90 practices. The estimates were much larger for the urology strata than for family practice and general surgery strata. We computed a sample size of n = 1,684 practices. Table 2

shows the allocation fractions n_h/n computed by using equation 14. Multiplying the total sample size by the allocation fractions yielded the sample sizes shown in Table 3.

CONCLUSION

We have described a design that we believe maximizes the precision that can be obtained from a survey with a very limited budget. To accomplish this, certain compromises had to be made. The file was not cleaned of dead listings before sampling, but such listings were handled during estimation, at the cost of increased variance. Also, the weights assigned to practices were approximate. The measures of size (number of physicians) were believed to be up to date, since they were obtained by interview. However, the actual selection probabilities reflect measures of size that were only as current as the file. We believe that this discrepancy is minor and nonsystematic, therefore, the effect is again to increase the variance.

REFERENCES

Cochran, W.G. (1977). Sampling Techniques. 3rd ed. John Wiley & Sons, New York.

Durbin, J. (1958). Sampling theory for estimators based on fewer individuals than the number selected. Bull. Int. Stat. Inst. 36(3):113-119.

Kendrick, J.S., Gonzales, B., Huber, D.H., Grubb, G.S., and Rubin, G.L. (1987). Complications of vasectomies in the United States. J. Fam. Pract. 25(3):245-248.

Kish, L. (1965). Survey Sampling. John Wiley & Sons, New York.

Orr, M.T., Forrest, J.D., Johnson, J.H., and Tolman, D.L. (1985). The provision of sterilization services by private physicians. Fam. Plann. Perspect. 17:216-220.

Table 1. Number of listings from AMA Physician Master File (M_{h0}).

	Census Regions			
Type of Practice	Northeast	Midwest	South	West
Family Practice	7,408	13,076	15,149	10,277
General Surgery	7,730	6,520	8,994	5,148
Urologists	2,035	1,800	2,985	1,683

Table 2. Sample allocation fractions (n_h/n).

	Census Regions			
Type of Practice	Northeast	Midwest	South	West
Family Practice	0.03100	0.05472	0.06339	0.04300
General Surgery	0.10674	0.09003	0.12419	0.07108
Urologists	0.09952	0.08803	0.14598	0.08231

Table 3. Sample sizes (n_h).

	Census Regions			
Type of Practice	Northeast	Midwest	South	West
Family Practice	52	92	107	72
General Surgery	180	152	209	120
Urologists	168	148	246	139