ENRICHING ONE SAMPLE WHILE IMPROVING ANOTHER: LINKING DIFFERENTIALLY STRATIFIED SAMPLES OF DOCUMENTS FILED BY EXEMPT ORGANIZATIONS

James M. Harte and Cecelia H. Hilgert, Internal Revenue Service
James M. Harte, Statistics of Income, P.O. Box 2608, Washington DC 20013-2608

KEY WORDS: Permanent Random Number

Introduction

Tax-exempt organizations are in the nonprofit sector, as economists call it. This sector is increasingly important having doubled in the last 15 years. In testimony before the Congress (June 1993), the Commisioner of the Internal Revenue (IRS) noted that the public charity part of the nonprofit sector had revenues in excess of $400 billion for 1990, representing 7.4 percent of the gross domestic product. Some of the organizations also conduct for-profit business activities. Linkage of the data from profit-making activities with the nonprofit side is important for policy analysis from the tax-exempt organization point of view and from the point of view of their profit-seeking competitors. This paper is about such a linkage.

The Statistics of Income (SOI) Division of the IRS conducts sample surveys of administrative documents filed by individuals and organizations under tax law, namely the Internal Revenue Code (IRC) [See reference 1]. These include documents filed by tax-exempt organizations engaging in charitable, educational, religious and other nonprofit activities. Tax-exempt activities are usually reported on Form 990. The SOI Form 990 sample represents six of the twenty-five types of nonprofit organizations filing the document. Another sample is of Form 990T, the tax return covering business activities unrelated to the tax-exempt purpose. This SOI sample covers any type of tax-exempt organization. (See references 2 and 3 for more information about nonprofit organizations.)

Starting with the samples covering 1993 activities (documents filed in 1994 and 1995), the Form 990 SOI sample is to be enriched. Form 990T data are to be added to the Form 990 sample record if a T return was filed. These T returns were to come from the separate T sample and be an integral part of it, if possible. This requirement seemed analagous to feeding the multitude with a few loaves and fishes as the following comparison shows (estimates for 1992):

	Form 990	Form 990T
Population:	273,000	42,000
Sample:	20,000	5,000
Organization:	6 Types	Any

We wished to preserve the broad representative character of the T sample without increasing the target sample size unduly. The extra T's needed for the match might crowd out other T's needed to represent the rest of the broad T population. Also wished were a sample selection and an estimation method which made best use of all sample units. Excluding the extra T's from the estimates seemed unsatisfactory. Including them as a self-representing sample stratum might not be much better. We had no data to judge whether these wishes were feasible. A feasibility study matching the two 1988 samples was conducted. The study used a special method of estimation available for SOI samples.

For the 1993 sample, a method of selection based on cross classifying organizations by information on each document was adopted. From these joint strata samples will be selected. Additional returns needed to enrich the Form 990 database will be secured through a higher sampling rate in the joint stratum whenever the T sampling rate based on income alone is too low. *For estimation there are no additional returns since no distinction is made in calculating the estimates.* The rest of the paper discusses the feasibility study, presents the 1993 design, and concludes with evaluation, notes and references.

Section 1 - The 1988 Feasibility Study

Strategy

The joint distribution of organizations with documents in either sample was estimated. Using the joint distribution, the 1988 T sample was hypothetically augmented with the extra returns needed to complete the match with the Form 990 sample. This augmented sample was compared with the actual sample, providing a point of view to initially discuss and assess the feasibility issues of sample size, representativeness and sampling error.

Form 990 Sample

The sample population is divided into two parts: (1) returns from organizations exempt under subsection 501(c)(3) of the Internal Revenue Code; and (2) returns from those exempt under subsections 501(c)(4) through (c)(9). (See the Exhibit 1 for a description of these organizations.) Returns with less than $25,001, the threshhold for required filing, are excluded. Each part is stratified by size of assets, but the size categories differ. These size categories, in use since 1988, are part of the detail of Table 1. More recently, smaller organizations may file a short Form 990EZ. For sampling purposes no distinction is made between the the short form and the longer form.

Form 990T Sample

The 1988 sample of this income tax return was stratified based on the size of net unrelated business income. All returns with income $100,000 were selected, while samples were drawn from four strata with smaller amounts. Starting with the 1992 sample, gross unrelated business income replaced net income as the stratifier. Table 1 from the 1988 study uses the newer gross income classes. Estimation was based, of course, on the 1988 net income stratification.

Sample Selection

Selection of documents from each stratum is actually based on selection of the organizations filing the documents. This is true of SOI samples, generally. (See reference 4 for background.) Each organization has an account number, the Employer Identification Number (EIN). From the EIN, a function is computed, yielding an integer with many digits. The low order digits are pseudo-random numbers. Each organization is assigned a permanent random number from the low order digits. The decision to include an organization's document in the sample from a stratum depends on whether its random number is less than some constant associated with the stratum. A hypothetical example: suppose the last three digits of the integer function are used as random numbers. Then, a sampling rule might be "include the document if the random number is less than 100. Otherwise, do not include it." This rule would be expected to yield a 10 percent sample for the stratum. Furthermore, if the document was included under the this rule for a 10 percent sample, then any other document filed by the organization in a different sample would be included whenever the sampling rate is 10 percent or more.

Estimation and Matching

Define AB as the event that documents from the same organization have been included in stratum A for sample 1 and stratum B for sample 2, respectively. Stipulate that pr[A] =< pr[B], which are the individual probabilities of inclusion in each sample. The inclusion probability for both documents is denoted by pr[AB], and pr[B/A] denotes the conditional probability that B will be included given that A has been included. Generally,

$$pr[AB] = pr[A]^* Pr[B/A]$$

but, in SOI sampling $pr[B/A] = 1$

Consequently, $pr[AB] = pr[A]$

This means that estimates from matching at the sample level may be derived from the inverse of the probability of selection in one of the original SOI samples. In practice, data files from the 1988 samples of Form 990 and Form 990T were compared. When a match on the EIN was detected a joint document was created. Each one of the pair had a weight from the original sample. In making estimates for Table 1, the higher of these two weights was applied. This method was considered reasonable for producing counts but not for financial estimates.

Joint Population

Table 1 classifies the population of organizations by asset size from Form 990 and by gross income size from Form 990T. It may be seen that 55 percent of the Form 990T population was from organizations represented by the SOI Form 990 sample: 15 percent from c(3) organizations and 40 percent from the c(4) through c(9) types of organizations.

The Augmented Sample

To the joint population classes the sampling rates for Form 990 were applied *whenever they were larger than the actual T rates.* This notionally produced a larger sample in which the match with the 990 sample was complete. To the actual sample of about 7,000 an additional 1,500 returns would have been required to augment the sample and complete the match with the Form 990 sample. About 2,800 of the required matching T's were in the actual sample. A comparison of the augmented and actual sample is as follows:

	Augmented Sample	Actual Sample	Additional Sample
501(c)(3)	1,823	1,141	682
501(c)(4)-(9)	2,500	1,652	848
All other	4,201	4,201	0
Totals	8,524	6,994	1,530

These results gave us a basis to believe that our objectives were feasible. If the sample had been absorbed rather than augmented, then the additional sample of 1,530 returns would have crowded out other returns but the all other returns category totaled 4,201.

Section 2 - SOI 1993 Sample Design

First, the Form 990 sample was designed, since the T sample had to conform to it. Population projections to 1993 were made as usual for the sample strata. For (c)(3) returns a sample of 11,500 was allocated among the strata, while 8,500 returns were allocated to the c(4) through (c)(9) returns. Formerly, equal allocations were made between these two groups.

Next an initial version of the T sample was made as if there were no requirement to serve the other sample. Population projections were made to gross income strata which had been established for the 1992 project. A sample of 5,000 was allocated paralleling the allocation made for 1992. This determined minimum sampling rates.

The 1993 population projections for the T returns and for Form 990 returns were the estimated marginal totals for the joint distribution of the population represented by one or both of the SOI samples. The internal cells of the joint distribution were prorated from the marginals following Table 1 from the 1988 feasibility study. To each asset-income stratum the higher of the T rate or the Form 990 rate was applied yielding an augmented sample of about 6,500 returns. This was scaled back to 5,500 returns by lowering and flattening the rates for income classes $60,000 under $300,000. (It was required to take all returns with $300,000 or more into the sample.) The final sampling rates for Form 990 and Form 990T are given in Table 2 . For joint strata the larger of the two rates will be applied. Table 3 contains the projected population counts for Form 990T, while Table 4 contains the projected allocation of the sample to the joint strata.

Section 3 - Evaluation

The allocation of the 1993 sample is compared with the 1992 allocation:

	1992	1993
Total	5,000	5,500
SOI 990 filed	3,500	4,500
No SOI 990 filed	1,500	1,000

The 1,000 returns with No SOI 990 filed includes an estimated 500 returns with gross income of $300,000 or more, the take-all category. This motivated the increase in the sample target from 5,000 to 5,500 returns. The domain estimates for organizations filing both returns will be improved because of the increased sample size and the deeper stratification. The estimates for the domain of returns not covered by the SOI sample will have larger sampling error because of the samller sample size. Neither of these domains is of great interest. Most estimates should be improved. The general representativeness of the T sample was preserved.

Effective use will be made of all returns because of joint stratification, assignment of permanent random numbers to organizations, and the flexible threshold rule for deciding inclusion in the sample.

Acknowledgments

The authors would like to thank Dan Skelly for his technical comments on this paper, Wendy Alvey and Beth Kilss for their review and assistance in the presentation of the material at San Francisco, and Paul Arnsberger for his technical assistance with the paper. Earlier, Elizabeth Nelson and Stephanie Hughes had worked on the 1988 feasibility study.

Notes and References

[1] For general information on Statistics of Income programs, see Scheuren, Fritz, and Petska, Tom, "Turning Administrative Systems into Information Systems," *Journal of Official Statistics,* Vol. 9, No. 1, 1993 Statistics Sweden; pp. 109-119.

[2] For information on the Statistics of Income studies on Forms 990, 990-PF, and 990-T, see *Statistics of Income—Compendium of Studies of Tax-Exempt Organizations, 1974-87, and Compendium of Studies of Tax-Exempt Organizations, 1987-92, Vol. 2.*

[3] See Hilgert, Cecelia, and Arnsberger, Paul, "Charities and Other Tax-Exempt Organizations, 1989," *Statistics of Income Bulletin, Winter 1993-94,* Volume 13, Number 3 (forthcoming) and also Meckstroth, Alicia, "Private Foundations and Charitable Trusts, 1990," *Statistics of Income Bulletin, Winter 1993-94*, Volume 13, Number 3 (forthcoming).

[4] See Harte, James, M., "Some Mathematical and Statistical Aspects of the Transformed Taxpayer Identification Number: A Sample Selection Tool Used at IRS," *1986 American Statistical Association Proceedings, Section on Survey Research Methods,* pp. 603-608.

Exhibit 1. Selected Types of Tax-Exempt Organizations, by Internal Revenue Code Section

Code section	Description of organization	Type of activities
501(c)(3)	Religious, charitable, educational, scientific or testing for public safety	Activities of nature implied by class or organization
501(c)(4)	Civic leagues, social welfare organizations, and local associations of employees	Promotion of community welfare, charitable, educational and recreational activities
501(c)(5)	Labor, agricultural and horticultural organizations	Educational or instructive, the purpose being to improve conditions of work, and to improve products and efficiency
501(c)(6)	Business leagues, chambers of commerce, real estate boards, etc.	Improvement of business conditions of one or more lines of business
501(c)(7)	Social and recreational clubs	Pleasure, recreational, and social activities
501(c)(8)	Fraternal beneficiary societies and associations	Lodge providing for payment of life, sickness accident or other benefits to members
501(c)(9)	Voluntary employees' beneficiary associations	Provides for payment of life, sickness, accident or other benefits to members

Table 1. Estimated SOI 1988 Form 990T Population by Gross Income (Form 990T) and Assets (Form 990)

Form 990 Assets ($000)	Form 990T Gross Business Income ($): under 1,000	1,000 under 20,000	20,000 under 60,000	60,000 under 150,000	150,000 under 300,000	300,000 or more	Totals
IRC 501(c)(3)							
Under 250	0	989	304	243	0	0	1,536
250 under 500	0	345	60	20	0	20	445
500 under 1,000	0	307	164	61	41	0	573
1,000 under 2,500	20	324	176	82	27	9	638
2,500 under 5,000	0	102	132	125	31	41	431
5,000 under 10,000	0	114	69	70	18	78	349
10,000 or more	0	210	264	249	228	397	1,348
Subtotals	20	2,391	1,169	850	345	545	5,320
IRC 501(c)(4) - (9):							
Under 125	0	3,374	1,277	173	43	0	4,867
125 under 400	0	3,007	641	359	121	29	4,157
400 under 1,000	0	1,064	604	321	170	82	2,241
1,000 under 2,500	0	500	434	404	181	79	1,598
2,500 under 10,000	3	171	136	250	183	194	937
10,000 or more	0	11	10	28	25	163	237
Subtotals	3	8,127	3,102	1,535	723	547	14,037
Totals							
990T and 990	23	10,518	4,271	2,385	1,068	1,092	19,357
990T, no 990	21	8,799	3,834	2,055	770	634	16,114
990T total	44	19,317	8,105	4,440	1,838	1,726	35,471

Table 2. Sampling Rates: SOI 1993 Form 990 or Form 990EZ, and Form 990T

Form 990 or Form 990EZ IRC 501(c)(3) Assets ($000)	Rate	Form 990 or Form 990EZ IRC 501(c)(4) (9) Assets ($000)	Rate	Form 990T All sections Income ($)	Rate
Under 250	0.007	Under 125	0.019	Under 1,000	0.000
250 under 500	0.007	125 under 400	0.045	1,000 under 20,000	0.014
500 under 1,000	0.014	400 under 1,000	0.091	20,000 under 60,000	0.045
1,000 under 2,500	0.045	1,000 under 2,500	0.200	60,000 under 150,000	0.147
2,500 under 5,000	0.073	2,500 under 10,000	0.400	150,000 under 300,000	0.147
5,000 under 10,000	0.147	10,000 or more	1.000	300,000 or more	1.000
10,000 or more	1.000				

Table 3. Projected SOI 1993 Form 990T Population by Gross Income (Form 990T) and Assets (Form 990)

Form 990 Assets ($000)	Form 990T Gross Business Income ($): under 1,000	1,000 under 20,000	20,000 under 60,000	60,000 under 150,000	150,000 under 300,000	300,000 or more	Totals
IRC 501(c)(3)							
Under 250	0	1,474	453	362	0	0	2,289
250 under 500	0	429	75	25	0	25	554
500 under 1,000	0	408	218	81	54	0	761
1,000 under 2,500	0	437	238	111	36	12	834
2,500 under 5,000	0	131	170	161	40	53	555
5,000 under 10,000	0	143	86	88	23	98	438
10,000 or more	0	259	326	308	282	491	1,666
Subtotals	0	3,281	1,566	1,136	435	679	7,097
IRC 501(c)(4) - (9):							
Under 125	0	3,984	1,508	204	51	0	5,747
125 under 400	0	3,350	714	400	135	32	4,631
400 under 1,000	0	1,191	676	359	190	92	2,508
1,000 under 2,500	0	646	561	522	234	102	2,065
2,500 under 10,000	0	245	195	358	262	278	1,338
10,000 or more	0	17	16	44	39	256	372
Subtotals	0	9,433	3,670	1,887	911	760	16,661
Totals							
990T and 990	0	9,692	3,996	2,195	1,193	1,251	18,327
990T, no 990	9,869	5,987	2,166	1,179	454	563	20,218
990T total	9,869	18,701	7,402	4,202	1,800	2,002	43,976

Table 4. Projected SOI 1993 Form 990T Sample by Gross Income (Form 990T) and Assets (Form 990)

Form 990 Assets ($000)	Form 990T Gross Business Income ($): under 1,000	1,000 under 20,000	20,000 under 60,000	60,000 under 150,000	150,000 under 300,000	300,000 or more	Totals
IRC 501(c)(3)							
Under 250	0	21	20	53	0	0	94
250 under 500	0	6	3	4	0	25	38
500 under 1,000	0	6	10	12	8	0	36
1,000 under 2,500	0	20	11	16	5	12	64
2,500 under 5,000	0	10	12	24	6	53	105
5,000 under 10,000	0	21	13	13	3	98	148
10,000 or more	0	259	326	308	282	491	1,666
Subtotals	0	343	395	430	304	679	2,151
IRC 501(c)(4) - (9):							
Under 125	0	76	68	30	7	0	181
125 under 400	0	151	32	59	20	32	294
400 under 1,000	0	108	62	53	28	92	343
1,000 under 2,500	0	129	112	104	47	102	494
2,500 under 10,000	0	98	78	143	105	278	702
10,000 or more	0	17	16	44	39	256	372
Subtotals	0	579	368	433	246	760	2,386
Totals							
990T and 990	0	838	694	741	528	1,251	4,052
990T, no 990	0	84	97	173	67	563	984
990T total	0	1,006	860	1,036	617	2,002	5,521