ESTIMATION OF THE NUMBER OF UNIQUE POPULATION ELEMENTS USING A SAMPLE

Laura Voshell Zayatz*, Bureau of the Census
Commerce/Census/SRD/3227-4 Washington, DC 20233

KEY WORDS: Microdata, Disclosure Avoidance, Unique Population Element

1. INTRODUCTION

National statistical agencies publicly release information about a nation's population that has been collected under a pledge of confidentiality. One method of releasing information is in the form of microdata files which consist of respondent level records containing characteristics of a sample of the elements (individuals or households) in a certain population. There are no obvious identifiers of respondents such as name or address on microdata files, and any agencies that release microdata must try to ensure that no intruders are able to link a respondent to its record on a microdata file. Any such linking would be a disclosure of confidential information.

A population element which has a unique combination of characteristics and is represented in a sample of microdata where those characteristics appear as categorical variables is at risk of disclosure. An intruder could match the element's unique combination of variables on the microdata file to the same combination of variables on some other data base containing identifiers. Because the element is unique, a one-to-one match could be obtained. Thus the intruder could link a unique respondent to its record. The categorical variables which the intruder might use for this purpose will be termed key variables (Bethlehem, Keller, and Pannekoek 1990; Greenberg 1990).

There is no set definition of the "disclosure risk" of a microdata file, however, it makes sense that the definition should involve the percent of population elements represented in that file which have a unique combination of key variables. Willenborg, Mokken, and Pannekoek (1990) suggest a measure of the disclosure risk of a microdata file which involves the percent of population uniques on the file. Anyone wishing to use this measure to assess the disclosure risk of a sample microdata file must estimate the percent of population uniques on the file using only information from the sample. The estimation of the percent of population uniques on the file is difficult because a sample record which is unique compared to all other records in the sample may or may not be truly unique in the population.

In this paper, two methods of estimating the percent of unique population elements on a sample microdata file using information from the sample are presented and evaluated. The first method, presented in Section 2, involves subsampling and the second method, presented in Section 3, involves equivalence class structure.

The two techniques were applied to simple random samples of several different population data sets. The estimates and the true percents of unique population elements on the sample files are presented in Section 4. Further observations and appendices omitted from this paper due to space limitations are contained in the report by Zayatz (1991), from which this paper is an extract.

2. PROCEDURE USING SUBSAMPLING

2.1 Background

The first method of estimating the percent of unique population elements on a microdata file involves taking a subsample from the sample microdata set using the same sampling fraction that was used to obtain the sample from the population. As we stated before, some records on the microdata sample are unique with respect to all other records on the sample but are not unique in the population. Likewise, there will be some records in the subsample which are unique with respect to all other records in the subsample but which are not unique with respect to all other records in the sample. The percent of the records which are unique in the subsample that are also unique in the sample can be used to approximate the percent of records which are unique in the sample that are truly unique in the population. This is seen in Figure 1.

Using records from the 1980 Decennial Census, we created several different sized data sets containing the same 6 categorical variables and took many different sized subsets from each one. We then plotted the percent of unique records in each subset that were also unique in the parent data set versus the percent of records in the parent data set contained in the subset.

We see from this graph that the actual sizes of the data sets and subsets did not play much of a role in determining the percent of records which were unique in the subset that were also unique in the parent data set. It was the ratio of the sizes of the subset and parent data set that determined this percent.

Our only assumption concerning the sample data sets which we use when performing this estimation procedure is that they contain real-life data. The phenomenon described above may not occur in simulated data sets with odd equivalence class structures (Greenberg and Zayatz 1991).

2.2 The Procedure

We begin the estimation procedure by taking a subsample from the sample microdata set using the same sampling fraction that that was used to take the sample from the population. We then find the percent of records which are unique in the subsample that are also unique in the sample. This percent is used to approximate the percent of records which are unique in the sample that are truly unique in the population. This enables us to estimate the number of records in the sample that are truly unique in the population, and finally this estimate may be multiplied by 100 and divided by the number of elements in the sample to obtain an estimate of the percent of unique population elements in the sample.

2.3 An Example

For our population, we will use a data set of 56372 records with 15 categorical variables from the 1980 Decennial Census. The true percent of unique population elements in our sample is 39.073%. Let

N = 56372 be the population size,
n_1 = 9383 be the sample size,
$f = n_1 / N = 9383 / 56372 = 0.166$ be the sampling fraction.

We begin by taking a subsample of the sample using the sampling fraction f = 0.166. We count the number of records which are unique in the subsample that are also unique in the sample. Let

$n_2 = |n_1 * f| = |9383 * 0.166| = 1562$ be the subsample size where $|x|$ denotes the nearest integer to x,
u_1 = 5563 be the number of records which are unique in the sample,
u_2 = 1263 be the number of records which are unique in the subsample,
u_i = 921 be the number of records which are unique in the subsample that are also unique in the sample.

We now calculate the percent of the records which are unique in the subsample that are also unique in the sample. Let

$$p_1 = 100 * u_i / u_2 = 100 * 921 / 1263 = 72.922\%$$

be this percent. This percent is used as an estimate of the percent of records which are unique in the sample that are truly unique in the population. The estimate of the number of records in the sample that are unique in the population is now calculated. Let

$$u_s = |u_1 * p_1 / 100| = |5563 * 72.922 / 100| = 4057$$

be this estimate. Finally, we calculate the estimate of the percent of unique population elements in the sample. Let

$$p_2 = 100 * u_s / n_1 = 100 * 4057 / 9383 = 43.238\%$$

be this estimate. The procedure is completed. Recall that the true percent of unique population elements in the sample is 39.073%.

3. PROCEDURE USING EQUIVALENCE CLASSES

3.1 Background

This method of estimating the percent of unique population elements on a sample microdata file involves dividing the records in the sample into groups of all records possessing the same combination of key variables. These groups are called equivalence classes (Greenberg and Voshell 1990). The number of records in each group is the size of that equivalence class. The percent of all equivalence classes in the sample that are of a given size can be used to approximate the percent of all equivalence classes in the population that are of that size. This is seen in Table 1.

Using the data set of 56372 records and 15 variables, we grouped the records into equivalence classes and calculated the percent of equivalence classes that were of each size. Note that the percent of unique population elements (100% * 22026 / 56372 = 39.1%) is not equal to the percent of equivalence classes in the population that are of size one (100% * 22026 / 28320 = 77.8%). It is important to keep in mind that these are two different percentages.
We then took a simple random sample of 9383 records from the data set, and again calculated the percent of equivalence classes that were of each size in the sample.

It can be seen from this table that the percent of equivalence classes that are of any given size in the sample can be used as a rough approximation of the percent of equivalence classes of that same size in the original data set (which for this purpose simulates our population). For example, for the percent of equivalence classes of size 1, 83.8% can be used to approximate 77.8% This same procedure was carried out on several other different data sets and subsets of different sizes, and the same phenomenon was noticed.

3.2 The Procedure

This second estimation technique begins by estimating the proportion of sample uniques that are truly unique in the population. Let

N = the size of the population,
n = the size of the sample,
Prob(C_p) = the probability that a given equivalence class in the population is of size C,
Prob (1_s | C_p) = the probability that an equivalence class of size C in the population will be represented by an equivalence class of size 1 in the sample,
Prob (1_p | 1_s) = the probability that an equivalence class of size 1 in the sample (a sample unique) was chosen from an equivalence class of size 1 in the population (a population unique).

We begin by estimating Prob (1_p | 1_s). By Bayes' Rule,

$$\text{Prob}(1_p|1s) = \frac{\text{Prob }(1_p \cap 1_s)}{\text{Prob }(1s)} = \frac{\text{Prob}(1_p)*\text{Prob}(1_s|1p)}{\sum_C \text{Prob}(Cp)*\text{Prob}(1_s|Cp)}$$

Recall that Prob (C_p), the probability that a given equivalence class in the population is of size C, can be estimated by the probability that a given equivalence class in the sample is of size C. Using the hypergeometric distribution, we can calculate

$$\text{Prob }(1_s \mid C_p) = \frac{\binom{C}{1}\binom{N-C}{n-1}}{\binom{N}{n}}$$

Thus we can estimate Prob (1_p | 1_s). We then multiply this estimate by the number of unique records in the sample to obtain an estimate of the number of records which are unique in the sample that are truly unique in the population. This estimate may be multiplied by 100 and divided by the number of elements in the sample to obtain an estimate of the percent of unique population elements in the sample.

3.3 An Example

We will again use the data set of 56372 records from the 1980 Decennial Census. Let

N = 56372 be the population size,
n = 9383 be the sample size,
$f = n / N = 9383 / 56372 = 0.166$ be the sampling fraction,
u_1 = 5563 be the number of records which are unique in the sample.

We begin by calculating Prob ($1_s \mid C_p$) using the formula given above and, using our sample, estimating Prob (C_p) for all class sizes C. See Table 2. We now estimate the probability that a record which is unique in the sample is truly unique in the population.

$$\text{Prob}(1_p|1_s)=\frac{\text{Prob }(1_p)*\text{Prob}(1_s \mid 1_p)}{\sum\limits_{C} \text{Prob}(C_p)*\text{Prob}(1_s \mid C_p)} \doteq \frac{0.140}{0.191}=0.732$$

This probability estimate is now used to estimate the number of records in the sample which are unique in the population. Let

$$u_s = |u_1 * \text{Prob}(\ 1_p \mid 1_s\ )| = |5563 * 0.732| \approx 4071$$

be this estimate. Finally, we calculate the estimate of the percent of unique population elements in the sample. Let

$$p_2 = 100 * u_s / n = 100 * 4071 / 9383 = 43.387\%$$

be this estimate. The procedure is completed. This estimate is slightly higher than the estimate obtained by the method of subsampling (43.238%). Recall that the true percent of unique population elements in the sample is 39.073%.

4. PERFORMANCE

We estimated the percent of unique population elements in simple random samples of several different populations using the two methods described above. Sampling fractions of 1/6 and 1/100 were used. In Tables 3 and 4, we provide the number of population elements, the number of key variables, the true percent of unique population elements in the sample, and the estimates of this percent for each population.

The methods seemed to perform at about the same level. When a sampling fraction of 1/6 was used, the methods estimated the percent of unique population elements in a sample fairly well, with a tendency to over-estimate. This apparent upward bias, in both cases, is caused by the fact that the percent of equivalence classes in the sample that are of size one is usually slightly higher than the percent of equivalence classes in the population that are of size one. When a sampling fraction of 1/100 was used, the methods did not provide good estimates of the percent of unique population elements in a sample.

The estimates of the number of unique population elements in a sample produced by these two methods will increase in accuracy as the sampling fraction increases. See Table 5. If the procedures are used with a sampling fraction of 1, the estimates will be exactly equal to the true percent of unique population elements in the sample.

5. CONCLUSION

As was stated earlier, a national statistical agency can regard the percent of unique population elements on a microdata file as one part of a measure of the disclosure risk of that file. In this report, we have presented two methods of estimating the percent of unique population elements in a sample microdata file. Examples of performance have been provided.

The two methods are currently being used to investigate how an increase in geographic detail would affect the percent of unique population elements on a microdata file from the Survey of Income and Program Participation. The Microdata Review Panel at the Census Bureau is currently reviewing a proposal to release a microdata file containing National Death Index information and information from the Current Population Survey, and the two methods have been used as part of the process of investigating the disclosure risk of the file.

Acknowledgement

The author wishes to thank Brian Greenberg for his valuable discussions on the topics in this paper and for his comments on earlier versions. She also thanks Alfredo Navarro and William Winkler for their comments.

* This paper reports the general results of research undertaken by Census Bureau staff. The views expressed are attributable to the author and do not necessarily reflect those of the Census Bureau.

REFERENCES

Bethlehem, J. G., Keller, W. J., and Pannekoek, J. (1990), "Disclosure Control of Microdata," Journal of the American Statistical Association, Vol. 85, pp. 38-45.

Greenberg, B. (1990) "Disclosure Avoidance Research at the Census Bureau," Proceedings of the Bureau of the Census Sixth Annual Research Conference, Bureau of the Census, Washington, D.C., pp. 144-166.

Greenberg, B. and Voshell, L. (1990), "Relating Risk of Disclosure for Microdata and Geographic Area Size," Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington, D.C., pp. 450-455.

Greenberg, B. and Zayatz, L. (1991), "Releasing Public Use Microdata Files at the U.S. Bureau of the Census," Special Issue of Statistica Neerlandica on Statistical Disclosure Avoidance, The Netherlands, to appear.

Willenborg, L. C. R. J., Mokken, R. J., and Pannekoek, J. (1990), "Microdata and Disclosure Risks," Proceedings of the Bureau of the Census Sixth Annual Research Conference, Bureau of the Census, Washington, D.C., pp. 167-180.

Zayatz, L. (1991), "Estimation of the Percent of Unique Population Elements on a Microdata File Using the Sample," Statistical Research Division Report Series, Census/SRD/RR-91/08, Bureau of the Census, Statistical Research Division, Washington, D.C.

Figure 1

Plot of Percent of Uniques in Subset that were Unique in Parent Data Set
Versus Percent of Records in Parent Data Set Contained in Subset
Symbol Represents Size of Subset
A - Smallest Size, S - Largest Size

```
100       +
Percent   ]                                                 FNM
of        ]                                              MOK
Uniques   ]                                              JE  G
in        ]                                          CLDF  H
 90       +                                        A N  K K
Subset    ]                                      JD IBJH
that      ]                                    DMONE  F  G
were      ]                                    NCI
also      ]                                  IJ  H  G
 80       +                                 ACNB  EF
Unique    ]                               JCIK H F
in        ]                             LK HF  G
Parent    ]                            A  JB  D E
Data      ]                            BI CFGEG G
 70       +                            I  FDE
Set       ]                         IB FCG
          ]                        AF E EE
          ]                    A   BG   D
          ]                  A    CDC D
 60       +                    F GDB
          ]                 A  D C
          ]                  DB
          ]                DB CC
          ]              DAB
 50       +              DC
          ]            AABB
          ]           C
          ]         CAB
          ]       AAA
 40       +       AB
          ]  AAB
          ] A
          ]A
          ]
 30       +
          ]
          ]
          ]
         -+---+---+---+---+---+---+---+---+---+---+---+---+---+---+---+--
          20  25  30  35  40  45  50  55  60  65  70  75  80  85  90  95

         Percent of Records in Parent Data Set Contained in Subset
```

Table 1

Equivalence Classes

	Data Set		Subset	
Class Size	Frequency	Percent	Frequency	Percent
1	22026	77.8	5563	83.8
2	2954	10.4	591	8.9
3	1090	3.8	171	2.6
4	560	2.0	97	1.5
5	354	1.3	54	0.8
6	223	0.8	44	0.7
7	173	0.6	29	0.4
8	109	0.4	23	0.3
9	106	0.4	10	0.2
10	87	0.3	10	0.2
11	64	0.2	10	0.2
12	53	0.2	12	0.2
13	54	0.2	5	0.1
14	48	0.2	5	0.1
15	26	0.1	3	0.0
16	37	0.1	1	0.0
17	25	0.1	3	0.0
18	14	0.0	1	0.0
19	21	0.1	1	0.0
20	16	0.1	0	0.0
21	18	0.1	0	0.0
22	12	0.0	1	0.0

Table 2

Class Size C	Calculation of Prob ($1_s \mid C_p$)	Estimate of Prob (C_p)
1	0.167	0.838
2	0.278	0.089
3	0.347	0.026
4	0.386	0.015
5	0.402	0.008
6	0.402	0.007
7	0.391	0.004
8	0.372	0.003
9	0.349	0.002
10	0.323	0.002
11	0.296	0.002
12	0.269	0.002
13	0.243	0.001
14	0.218	0.001
15	0.195	0.000
16	0.173	0.000
17	0.153	0.000
18	0.135	0.000
19	0.119	0.000
20	0.104	0.000

Table 3

Sampling Fraction = F = 1/6

Pop. Data Set	No. of Population Elements	No. of Variables	% of Unique Population Elements in Sample	Estimate Using Subsampling Method	Estimate Using Eq. Class Method
#1	67685	4	0.194	0.223	0.228
#2	116504	5	1.548	2.018	1.786
#3	87959	6	0.380	0.434	0.368
#4	117290	7	3.479	3.728	3.346
#5	117458	8	4.837	5.303	4.862
#6	10321	9	15.531	17.876	16.878
#7	87959	10	8.936	10.355	10.434
#8	10000	11	84.690	90.300	90.890
#9	87959	15	35.139	39.117	39.611

Table 4

Sampling Fraction = F = 1/100

Pop. Data Set	No. of Population Elements	No. of Variables	% of Unique Population Elements in Sample	Estimate Using Subsampling Method	Estimate Using Eq. Class Method
#1	67685	4	0.194	2.275	1.862
#2	116504	5	1.548	2.992	4.747
#3	87959	6	0.380	0.958	1.227
#4	117290	7	3.479	12.959	11.502
#5	117458	8	4.837	20.433	13.834
#6	10321	9	15.531	73.636	54.084
#7	87959	10	8.936	32.085	33.624
#8	10000	11	84.690	100.000	100.000
#9	87959	15	35.139	78.522	78.590

Table 5

True Percent of Unique Population Elements in the Sample: 39.073%

Sampling Fraction	Subsampling Estimate	Equivalence Class Estimate
0.1	46.241%	46.754%
0.2	43.185%	42.521%
0.3	40.676%	40.662%
0.4	40.286%	40.277%
0.5	37.593%	39.706%
0.6	39.635%	39.633%
0.7	39.280%	39.275%
0.8	39.280%	39.300%
0.9	38.028%	39.062%
1.0	39.073%	39.073%