BUREAU OF THE CENSUS
STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR–91/08

ESTIMATION OF THE PERCENT OF UNIQUE POPULATION ELEMENTS ON A MICRODATA FILE USING THE SAMPLE

by

Laura Voshell Zayatz
Statistical Research Division
U.S. Bureau of the Census
Washington, DC 20233 U.S.A.

Report issued: August 14, 1991

Estimation of the Percent of Unique Population Elements on a Microdata File Using the Sample

Laura Voshell Zayatz

ABSTRACT

National statistical agencies publicly release information about a nation's population that has been collected under a pledge of confidentiality. 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 to the same combination of variables on some other data base containing identifiers and thus link the element to its microdata record. The percent of unique population elements on a microdata file can be regarded as one component of a measure of disclosure risk. In this paper, two methods of estimating the percent of unique population elements on a sample microdata file using information from that sample are presented and evaluated. A third method of estimation was discussed by Willenborg, Mokken, and Pannekoek (1990) and is reviewed here.

KEY WORDS: Microdata, Disclosure Avoidance, Unique Population Element

I. 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.

An element in a population is called "unique" if that element possesses a combination of characteristics which distinguishes it from all other elements in that population. 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 <u>key variables</u> (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:

$$\text{Risk} = 1 - (1 - f_s f_u)^n$$

where n is the sample size, f_s is the fraction of population elements for

which the intruder knows the values of the key variables, and f_u is the fraction of population elements which have a unique combination of the key variables. Anyone wishing to use this measure to assess the disclosure risk of a sample microdata file must estimate f_u using only information from the sample. The estimation of f_u 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. One method, presented in Section II, uses subsampling and a relationship between the subsample and the sample which is also present between the sample and the population as a basis for estimation. The second method involves separating the records in the sample into groups of all records with the same combination of key variables. These groups are called equivalence classes. The technique, presented in Section III, uses the distribution of the sizes of the equivalence classes in the sample as a basis for estimation.

A third method of estimating the percent of unique population elements on a microdata file using the sample is discussed by Willenborg, Mokken, and Pannekoek (1990) and is reviewed here. This method involves the distribution of the sizes of all possible equivalence classes and is discussed in Section IV.

The three methods of estimating the percent of unique population elements on a microdata file 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 V. An abbreviated version of this paper will be presented at the 1991 Annual Meetings of the American Statistical Association and will appear in the proceedings of those meetings (Zayatz 1991).

II. Procedure Using Subsampling

A. Background

One 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. The use of such an approximation when the sampling fraction is relatively large is justified by the following work.

Using records from Population Data Set #3 which is described in detail in the Appendix, 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. This plot contains 190 points, and these points represent data sets and subsets of all different sizes and different ratios of sizes. See Figure 1.

This entire process was repeated using sets of records with 15 categorical

variables from Population Data Set #9. The origins of all Population Data Sets as well as the categorical breakdowns of their variables are described in the Appendix. See Figure 2.

We see from these graphs 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. This fact leads to an estimation of the percent of records which are unique in the sample that are truly unique in the population.

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).

B. The Procedure

We begin the estimation procedure by taking a subsample from the sample microdata set using the same sampling fraction that was used to take the sample from the population. We then list all records which are unique in the sample. We also list all records which are unique in the subsample. Comparing the two lists, we can find the percent of records which are unique in the subsample that are also unique in the sample. From work explained above, we know that this percent can be used to approximate the percent of records which are unique in the sample that are truly unique in the population. We estimate the number of records in the sample that are truly unique in the population to be this percent of the number of records which are unique in the sample. This estimate may be multiplied by 100 and divided by the number of elements in the sample to obtain an estimate of the percent unique population elements in the sample.

C. An Example

For our population, we will use a data set of 56372 records with 15 categorical variables from Population Data Set #9. 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 then make one list of all records which are unique in the sample and another list of all records which are unique in the subsample. We find the intersection of the two lists and count the records which are unique in 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 the true percent of unique population elements in the sample is 39.073%.

D. Effects of Sampling Fraction

A good property of any method which uses a sample to estimate a population parameter is an increased accuracy of the estimate as the sampling fraction increases. Here we present an example which conveys the tendency of the estimate to be more accurate as the sampling fraction increases. Similar work was conducted on other populations, and the same trend was found.

The estimate of the percent of unique population elements on a microdata file produced by this method will increase in accuracy as the percent of records which are unique in the subsample that are also unique in the sample becomes closer to the percent of records which are unique in the sample that are truly unique in the population. Using the data set of 56372 records with 15 variables from Population Data Set #9, we calculated the absolute values of the differences between these two percents for sampling fractions of 0.1, 0.2, ..., 0.9. See Table 1 where this information and the actual estimates are displayed. Recall that the true percent of unique population elements in the sample is 39.073%.

As seen in the table, the absolute value of the differences between the two percents tended to decrease as the sampling fraction increased. This caused an increase in the accuracy of the estimate as the sampling fraction increased. As stated before, this trend was seen in other populations as well and thus suggests that the accuracy of the estimate of the percent of unique population elements in a sample increases as the sampling fraction increases using this method.

If this procedure is used with a sampling fraction of 1, the sample will actually be the entire population. The subsample will be the entire sample, and it will also be the entire population. Thus

$$f = 1 \Rightarrow N = n_1 = n_2.$$

Because the subsample is the sample,

$$u_1 = u_2,$$

and 100% of the records which were unique in the subsample will also be unique in the sample. Therefore, the estimate of the percent of records which are unique in the sample that are truly unique in the population will be

$$p_1 = 100 * u_1 / u_2 = 100\%.$$

The estimate of the number of unique population elements which are in the sample will be

$$u_s = u_1 * p_1 / 100 = u_1.$$

Finally, the estimate of the percent of unique population elements in the

sample will be

$$p_2 = 100 * u_s / n_1 = 100 * u_1 / n_1.$$

This estimate is exactly equal to the true percent of unique population elements in the sample.

III. Procedure Using Equivalence Classes

A. 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. The use of such an approximation when the sampling fraction is relatively large is justified by the following work.

Using the data set of 56372 records and 15 variables from Population Data Set #9, we grouped the records into equivalence classes and counted the number of equivalence classes of each size. We also calculated the percent of equivalence classes that were of each size. The results are shown in Table 2. 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 counted the number of equivalence classes of each size and calculated the percent of equivalence classes that were of each size in the sample. The results are shown in Table 3.

It can be seen from these two tables 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, 77.8% of all equivalence classes in the original data set are of size 1, and 83.8% of all equivalence classes in the sample are of size 1. For our purposes, 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.

This technique of estimating the percent of unique population elements in a sample involves two probabilities. The first is the probability that a given equivalence class in the population is of a certain size. This probability is equal to the percent of equivalence classes in the population which are of that size divided by 100. As discussed above, we may estimate the probability that a given equivalence class in the population is of a given size as the percent of equivalence classes in the sample which are of that size divided by 100. For example, using the data set of 56372 records described above, the actual probability that a given equivalence class in the population is of size 1 is 0.778. Using the sample, we would estimate this probability as 0.838. The probability that an equivalence class in the population is of size C will be denoted Prob (C_p) and will be estimated by the percent of equivalence classes in the sample that are of size C divided by 100.

Also involved in the estimation procedure is the probability that one and only one element from an equivalence class of a given size in the population will

be chosen in the sample. In other words, we make use of the probability that an equivalence class of a given size in the population will be represented by an equivalence class of size 1 in the sample. Let

N = the size of the population,
n = the size of the sample,
C_p = the size of the equivalence class in the population,
$Prob\ (\ 1_s \mid C_p)$ = 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.

Then 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 is

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

where

$$\binom{x}{y} = \frac{x!}{y!(x-y)!}$$

and we define

$$\binom{x}{y} = 0 \quad \text{if} \quad y > x.$$

The probability that an equivalence class is of size C in the population and is represented by an equivalence class of size 1 in the sample is the product of the two probabilities we have discussed.

$$Prob\ (\ 1_s \cap C_p\ ) = Prob\ (\ C_p\ ) * Prob\ (\ 1_s \mid C_p\ )$$

Recall that the first probability in the product can be estimated for all values of C, and the second probability in the product can be calculated exactly for all values of C using the formula given above. This product is involved in the estimation of the number of records which are unique in the sample that are truly unique in the population. This estimate in turn leads to an estimation of the percent of unique population elements in the sample.

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).

B. The Procedure

We begin by estimating the probability that a record is unique in the population given that it is unique in the sample. By Bayes' rule,

$$Prob\ (\ 1_p \mid 1_s\ ) = \frac{Prob\ (\ 1_p \cap 1_s\ )}{Prob\ (\ 1_s\ )} = \frac{Prob\ (\ 1_p\ ) * Prob\ (\ 1_s \mid 1_p\ )}{\sum_C Prob\ (\ C_p\ ) * Prob\ (\ 1_s \mid C_p\ )}$$

We estimate $Prob\ (\ 1_p \mid 1_s\ )$ using our estimates of $Prob\ (\ C_p\ )$ and our calculations of $Prob\ (\ 1_s \mid C_p\ )$ for all C. We then multiply this estimate of probability with 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 unique population elements in the sample.

C. An Example

We will again use the data set of 56372 records from Population Data Set #9. 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 | C_p) using the formula given above and, using our sample, estimating Prob (C_p) for all class sizes C. See Table 4.

Note that we need not calculate and estimate these values for classes of size greater than 20 because, according to the sample, Prob (C_p) for C > 20 is approximately 0, hence the product Prob (C_p) * Prob (1_s | C_p) is approximately 0 for C > 20. We now estimate

$$\text{Prob } (1_p) * \text{Prob } (1_s \mid 1_p) = 0.838 * 0.167 = 0.140$$

and

$$\sum_C \text{Prob } (C_p) * \text{Prob } (1_s \mid C_p) = 0.191$$

Thus our estimate of the probability that a record which is unique in the sample is truly unique in the population is

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

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

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

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 just 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%.

D. Effects of Sampling Fraction

Using this method of estimation, we would hope to find an increase in the accuracy of the estimate as the sampling fraction increases. Here we present an example which conveys the tendency of the estimate to be more accurate as the sampling fraction increases. Similar work with this method was conducted on other populations, and the same trend was found.

The estimate of the number of unique population elements in a sample produced by this method will increase in accuracy as the percent of equivalence classes

of each given size in the sample approaches the percent of equivalence classes of that size in the population. Using the 56372 records with 15 variables from Population Data Set #9, we calculated these percents for sampling fractions of 0.1, 0.2, ..., 0.9 and for the population. See Table 5.

Note that as the sampling fraction increases, the percent of equivalence classes of each given size in the sample tends to grow closer to the percent of equivalence classes of that size in the population. This causes a tendency in the estimate of the percent of unique population elements in a sample to become more accurate as the sampling fraction increases. See Table 6.

This trend was seen in other populations as well and thus suggests that the accuracy of the estimate of the percent of unique population elements in a sample increases as the sampling fraction increases using this method.

If this procedure is used with a sampling fraction of 1, then

$$N = n$$

and

$$\binom{N-C}{n-1} = \binom{N-C}{N-1} = 0 \text{ for } C > 1.$$

Thus

$$\text{Prob}\ (\ 1_s \mid C_p\ ) = \frac{\binom{C}{1}\binom{N-C}{n-1}}{\binom{N}{n}} = 0 \text{ for } C > 1,$$

$$\sum_C \text{Prob}\ (\ C_p\ ) * \text{Prob}\ (\ 1_s \mid C_p\ ) = \text{Prob}\ (\ 1_s\ ) * \text{Prob}\ (\ 1_s \mid 1_p\ ),$$

and

$$\text{Prob}\ (\ 1_p \mid 1_s\ ) = \frac{\text{Prob}\ (\ 1_p\ ) * \text{Prob}\ (1_s \mid 1_p)}{\sum_C \text{Prob}\ (C_p\ ) * \text{Prob}\ (\ 1_s \mid C_p\ )} = 1.$$

The estimate of the probability that a record which is unique in the sample is truly unique in the population is 1. Therefore the estimate of the number of unique population elements which are present in the sample will be

$$u_s = u_1 * \text{Prob}\ (\ 1_p \mid 1_s\ ) = u_1.$$

Thus, the estimate of the percent of unique population elements in the sample will be

$$p_2 = 100 * u_s / n = 100 * u_1 / n.$$

This estimate is exactly equal to the true percent of unique population elements in the sample.

IV. Procedure Described in (Willenborg, Mokken, and Pannekoek 1990)

A. Background

Willenborg, Mokken, and Pannekoek (1990) describe a method of estimating the

percent of unique elements in a population using information from a sample of that population. We are interested in the percent of unique population elements present in a sample of a population as would be found on a microdata file, but for the case of simple random samples, the two percents are approximately equal. The estimation procedure will not be fully described but can be found in (Willenborg, Mokken, and Pannekoek 1990). It involves the distribution of the sizes of all potential equivalence classes. A potential equivalence class is represented by any possible combination of the key variables.

B. Limitations

In order to calculate the estimate of the percent of unique population elements, one must obtain and use in calculations the number of records in each potential equivalence class. This becomes impossible if the number of key variables is large and the number of categories of each variable is large. For example, a data set with 10 variables having 10 categories each has 10 billion potential equivalence classes. Thus the method of estimation is limited to cases where the number of possible combinations of the key variables is not extremely large.

C. Assumptions

The frequency (or size) of a potential equivalence class is the number of population elements possessing that class's combination of key variables. For an arbitrary class frequency Y, it is assumed that

$$Y \mid \mu = m \sim \text{Poisson}(m)$$

and

$$\mu \sim \text{gamma}(\alpha, \beta).$$

This implies that the marginal distribution of Y is negative binomial. This assumption may not be valid. Using the class frequencies of a set of 87959 records with 6 variables from Population Data Set #3, we calculated the estimates of the parameters corresponding to the negative binomial distribution and then tested the hypothesis that the frequencies followed a negative binomial distribution with those parameters. The Kolmogorov Goodness of Fit test rejected the hypothesis at the 0.01 level.

D. Example of Performance

The technique was used to estimate the percent of unique population elements using a sampling fraction of 1/6 and a set of 87959 records with 6 variables from Population Data Set #3. The percent of unique population elements was 0.380%. The procedure yielded an estimate of 0.011%. More examples of performance and a comparison of performances between this and the two previously described methods of estimation are presented in Section V.

E. Effects of Sampling Fraction

The percent of unique population elements in that same set of 87959 records from Population Data Set #3 was estimated by applying this method to samples containing from 5% to 95% of the records in the data set. As stated before, the percent of unique population elements was 0.380%. The estimates ranged from 0.0112% to 0.0119%. See Figure 3.

Note that the estimates did not change much by using different sized samples. One would expect that an estimate given by a 95% sample would be much better than an estimate given by a 5% sample. This, however, is not the case. The reason is that the estimate mainly involves the variance, scaled according to

sample size, of the frequencies of all potential equivalence classes in the sample. This scaled variance does not change much from a 5% sample to a 95% when the number of potential equivalence classes is large and the frequencies of most potential equivalence classes are 0. This same phenomenon is seen in Figure 4.

Here we took a one in six sample of the set of 87959 records with 6 variables from Population Data Set #3. We estimated the percent of unique population elements several times using this sample as if it were a sample from different sized populations. Using the sample as if it were a 5% sample of a population, the estimate of the percent of unique population elements was 0.011309%. Using the sample as if it were a 95% sample of a population, the estimate of the percent of unique population elements was 0.011313%. These two estimates are almost equal because the scaled variances of the frequencies of potential equivalence classes are almost equal.

If this procedure is used with a sampling fraction of 1, the estimate of the percent of unique population elements will, in most cases, not equal the true percent of unique population elements. This can be seen in Figure 3 where 100% of the population is in the sample and the estimate of the percent of unique population elements is 0.0112% while the true percent is 0.3797%. This is because this method attempts to fit the frequencies of potential equivalence classes to a distribution using information from the sample. It then uses this distribution to find the expected number of equivalence classes of size one in the population. An equivalence class of size one in the population constitutes a unique population element. Even if the sample is actually the entire population, it is unlikely that the distribution fitted to the class frequencies will have an expected value of the number of equivalence classes of size one equal to the true number of unique population elements. Thus the estimate of the percent of unique population elements will not be the true percent of unique population elements.

V. Performance

We estimated the percent of unique population elements in simple random samples of several different populations using the three methods described above. Sampling fractions of 1/6 and 1/100 were used. In Tables 7 and 8, we provide the number of population elements, the number of key variables, the true percent of unique population elements in the sample, and the three estimates of this percent for each population. Note that the method discussed in (Willenborg, Mokken, and Pannekoek 1990) could not be applied to populations with a large number of possible combinations of key variables. More detailed descriptions of the population data sets are given in the Appendix.

The method using subsampling and the method using equivalence classes seemed to perform at about the same level. When a sampling fraction of 1/6 was used, these two 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. This phenomenon can be seen in Table 5. The method discussed in (Willenborg, Mokken, and Pannekoek 1990) consistently under-estimated the percent of unique population elements in a sample and could not be used for four of the nine data sets because of the large number of potential equivalence classes. When a sampling fraction of 1/100 was used, none of the three methods provided good estimates of the percent of unique population elements in a sample.

VI. 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. A third method discussed in (Willenborg, Mokken, and Pannekoek 1990) was reviewed. Examples of performance have been provided.

The two methods of estimation that involve subsampling and the distribution of the sizes of the equivalence classes in the sample 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 of National Death Index records, and the two methods may be used as part of the process of investigating the disclosure risk of the file.

VII. 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.

VIII. 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 Number of Unique Population Elements Using a Sample," Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington, D.C., to appear.

Table 1

Sampling Fraction	% of Records which were Unique in Subsample that were also Unique in Sample	% of Records which were Unique in Sample that were Truly Unique in Population	Abs. Value Difference in Percents	Estimate
0.1	71.4	59.3	12.1	46.241%
0.2	73.3	66.4	6.9	43.185%
0.3	78.1	75.0	3.1	40.676%
0.4	82.0	79.9	2.1	40.286%
0.5	80.9	83.9	3.0	37.593%
0.6	88.7	87.7	1.0	39.635%
0.7	91.6	91.0	0.6	39.280%
0.8	94.5	94.2	0.3	39.280%
0.9	94.7	97.2	2.5	38.028%

Table 2

Equivalence Classes in the Data Set

Class Size	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	22026	77.8	22026	77.8
2	2954	10.4	24980	88.2
3	1090	3.8	26070	92.1
4	560	2.0	26630	94.0
5	354	1.3	26984	95.3
6	223	0.8	27207	96.1
7	173	0.6	27380	96.7
8	109	0.4	27489	97.1
9	106	0.4	27595	97.4
10	87	0.3	27682	97.7
11	64	0.2	27746	98.0
12	53	0.2	27799	98.2
13	54	0.2	27853	98.4
14	48	0.2	27901	98.5
15	26	0.1	27927	98.6
16	37	0.1	27964	98.7
17	25	0.1	27989	98.8
18	14	0.0	28003	98.9
19	21	0.1	28024	99.0
20	16	0.1	28040	99.0
21	18	0.1	28058	99.1
22	12	0.0	28070	99.1
23	23	0.1	28093	99.2
24	18	0.1	28111	99.3
25	15	0.1	28126	99.3
26	11	0.0	28137	99.4
27	9	0.0	28146	99.4
28	7	0.0	28153	99.4
29	7	0.0	28160	99.4
30	9	0.0	28169	99.5
31	8	0.0	28177	99.5
32	12	0.0	28189	99.5
33	5	0.0	28194	99.6
34	7	0.0	28201	99.6
35	6	0.0	28207	99.6
36	8	0.0	28215	99.6
37	7	0.0	28222	99.7
38	3	0.0	28225	99.7
39	4	0.0	28229	99.7
40	3	0.0	28232	99.7
41	6	0.0	28238	99.7
42	5	0.0	28243	99.7
43	2	0.0	28245	99.7
44	1	0.0	28246	99.7
45	4	0.0	28250	99.8
46	6	0.0	28256	99.8
47	3	0.0	28259	99.8
48	3	0.0	28262	99.8
49	1	0.0	28263	99.8
50	2	0.0	28265	99.8
51	2	0.0	28267	99.8
52	3	0.0	28270	99.8
53	3	0.0	28273	99.8
54	1	0.0	28274	99.8
55	4	0.0	28278	99.9
56	1	0.0	28279	99.9
57	2	0.0	28281	99.9

Table 2, continued

58	2	0.0	28283	99.9
59	1	0.0	28284	99.9
60	4	0.0	28288	99.9
61	3	0.0	28291	99.9
62	2	0.0	28293	99.9
64	4	0.0	28297	99.9
68	1	0.0	28298	99.9
69	2	0.0	28300	99.9
70	2	0.0	28302	99.9
72	2	0.0	28304	99.9
75	1	0.0	28305	99.9
76	1	0.0	28306	100.0
77	1	0.0	28307	100.0
78	1	0.0	28308	100.0
79	1	0.0	28309	100.0
80	2	0.0	28311	100.0
86	1	0.0	28312	100.0
87	1	0.0	28313	100.0
88	2	0.0	28315	100.0
101	1	0.0	28316	100.0
103	1	0.0	28317	100.0
121	1	0.0	28318	100.0
141	1	0.0	28319	100.0
298	1	0.0	28320	100.0

Table 3

Equivalence Classes in the Subset

Class Size	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	5563	83.8	5563	83.8
2	591	8.9	6154	92.8
3	171	2.6	6325	95.3
4	97	1.5	6422	96.8
5	54	0.8	6476	97.6
6	44	0.7	6520	98.3
7	29	0.4	6549	98.7
8	23	0.3	6572	99.1
9	10	0.2	6582	99.2
10	10	0.2	6592	99.4
11	10	0.2	6602	99.5
12	12	0.2	6614	99.7
13	5	0.1	6619	99.8
14	5	0.1	6624	99.8
15	3	0.0	6627	99.9
16	1	0.0	6628	99.9
17	3	0.0	6631	99.9
18	1	0.0	6632	100.0
19	1	0.0	6633	100.0
22	1	0.0	6634	100.0
66	1	0.0	6635	100.0

Table 4

Class Size C	Calculation of Prob (1_s | 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 5

Class Size	Sampling Fraction 10%	Sampling Fraction 20%	Sampling Fraction 30%	Sampling Fraction 40%	Sampling Fraction 50%
C	Percent of Classes	Percent of Classes	Percent of Classes	Percent of Classes	Percent of Classes
1	85.5	83.1	81.6	80.9	80.1
2	8.4	8.8	9.3	9.2	9.5
3	2.2	3.0	3.1	3.6	3.5
4	1.7	1.7	1.7	1.7	1.8
5	0.6	0.9	1.0	1.2	1.2
6	0.5	0.5	0.7	0.6	0.8
7	0.5	0.6	0.6	0.5	0.4
8	0.3	0.5	0.3	0.4	0.3
9	0.1	0.1	0.2	0.3	0.4
10	0.0	0.2	0.3	0.3	0.3
11	0.1	0.1	0.2	0.2	0.2
12	0.0	0.1	0.3	0.2	0.2
13	0.0	0.2	0.0	0.1	0.2
14	0.0	0.1	0.1	0.1	0.1
15	0.0	0.0	0.1	0.1	0.1
16	0.0	0.0	0.1	0.1	0.1
17	0.0	0.0	0.1	0.1	0.0
18	0.0	0.0	0.0	0.0	0.1
19	0.0	0.0	0.1	0.0	0.0
20	0.0	0.0	0.0	0.1	0.1
21	0.0	0.0	0.1	0.0	0.1
22	0.0	0.0	0.0	0.1	0.1
23	0.0	0.0	0.0	0.0	0.0
24	0.0	0.0	0.0	0.0	0.0
25	0.0	0.0	0.0	0.0	0.0
26	0.0	0.0	0.0	0.0	0.0
27	0.0	0.0	0.0	0.0	0.0
28	0.0	0.0	0.0	0.0	0.0
29	0.0	0.0	0.0	0.0	0.0
30	0.0	0.0	0.0	0.0	0.0

Table 5, continued

Class Size	Sampling Fraction 60%	Sampling Fraction 70%	Sampling Fraction 80%	Sampling Fraction 90%	Population
C	Percent of Classes	Percent of Classes	Percent of Classes	Percent of Classes	Percent of Classes
1	79.6	79.0	78.6	78.1	77.8
2	9.6	9.9	10.0	10.3	10.4
3	3.7	3.8	3.8	3.8	3.8
4	1.8	1.8	2.0	2.0	2.0
5	1.0	1.1	1.1	1.1	1.3
6	1.0	0.9	0.7	0.8	0.8
7	0.5	0.6	0.7	0.6	0.6
8	0.4	0.4	0.4	0.4	0.4
9	0.3	0.3	0.3	0.4	0.4
10	0.2	0.3	0.3	0.3	0.3
11	0.2	0.2	0.2	0.2	0.2
12	0.2	0.1	0.2	0.2	0.2
13	0.1	0.2	0.2	0.2	0.2
14	0.1	0.1	0.1	0.1	0.2
15	0.1	0.1	0.1	0.1	0.1
16	0.1	0.1	0.1	0.1	0.1
17	0.0	0.1	0.1	0.1	0.1
18	0.1	0.1	0.1	0.1	0.0
19	0.1	0.0	0.1	0.1	0.1
20	0.1	0.1	0.1	0.1	0.1
21	0.1	0.0	0.1	0.0	0.1
22	0.1	0.0	0.0	0.1	0.0
23	0.0	0.1	0.0	0.1	0.1
24	0.0	0.0	0.0	0.0	0.1
25	0.1	0.1	0.0	0.0	0.1
26	0.1	0.0	0.0	0.0	0.0
27	0.0	0.0	0.1	0.0	0.0
28	0.0	0.1	0.1	0.0	0.0
29	0.0	0.0	0.0	0.0	0.0
30	0.0	0.0	0.0	0.0	0.0

Table 6

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

Sampling Fraction	Estimate of Percent of Unique Population Elements in Sample
0.1	46.754%
0.2	42.521%
0.3	40.662%
0.4	40.277%
0.5	39.706%
0.6	39.633%
0.7	39.275%
0.8	39.300%
0.9	39.062%

Table 7

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	Estimate Using Willenborg's Method
#1	67685	4	0.194	0.223	0.228	0.056
#2	116504	5	1.548	2.018	1.786	0.236
#3	87959	6	0.380	0.434	0.368	0.011
#4	117290	7	3.479	3.728	3.346	0.154
#5	117458	8	4.837	5.303	4.862	0.195
#6	10321	9	15.531	17.876	16.878	NA
#7	87959	10	8.936	10.355	10.434	NA
#8	10000	11	84.690	90.300	90.890	NA
#9	87959	15	35.139	39.117	39.611	NA

Table 8

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	Estimate Using Willenborg's Method
#1	67685	4	0.194	2.275	1.862	0.365
#2	116504	5	1.548	2.992	4.747	0.202
#3	87959	6	0.380	0.958	1.227	0.013
#4	117290	7	3.479	12.959	11.502	0.150
#5	117458	8	4.837	20.433	13.834	0.170
#6	10321	9	15.531	73.636	54.084	NA
#7	87959	10	8.936	32.085	33.624	NA
#8	10000	11	84.690	100.000	100.000	NA
#9	87959	15	35.139	78.522	78.590	NA

Figure 1

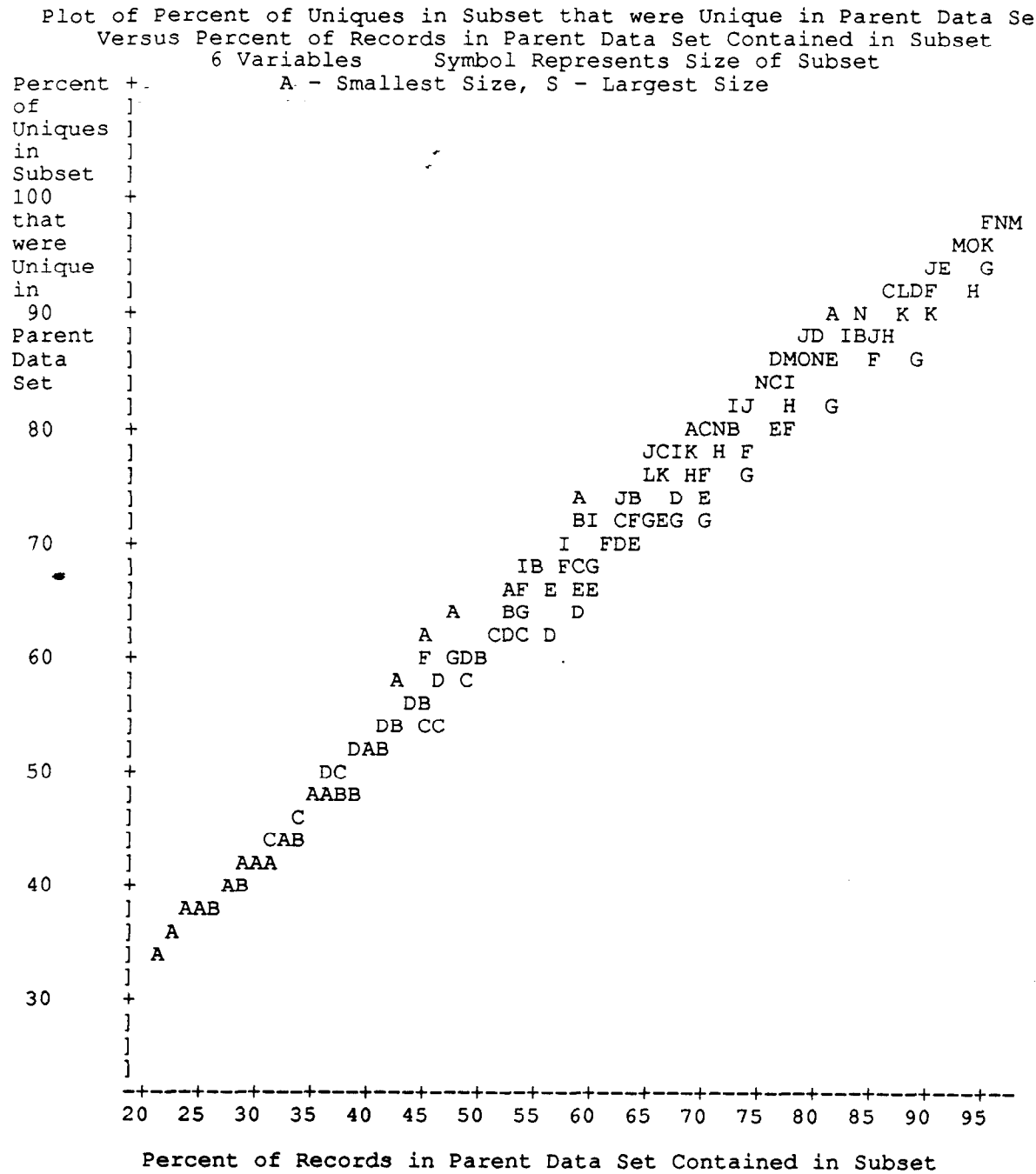

Plot of Percent of Uniques in Subset that were Unique in Parent Data Set
Versus Percent of Records in Parent Data Set Contained in Subset
6 Variables Symbol Represents Size of Subset
A - Smallest Size, S - Largest Size
Percent of Uniques in Subset that were Unique in Parent Data Set
Percent of Records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

Figure 2

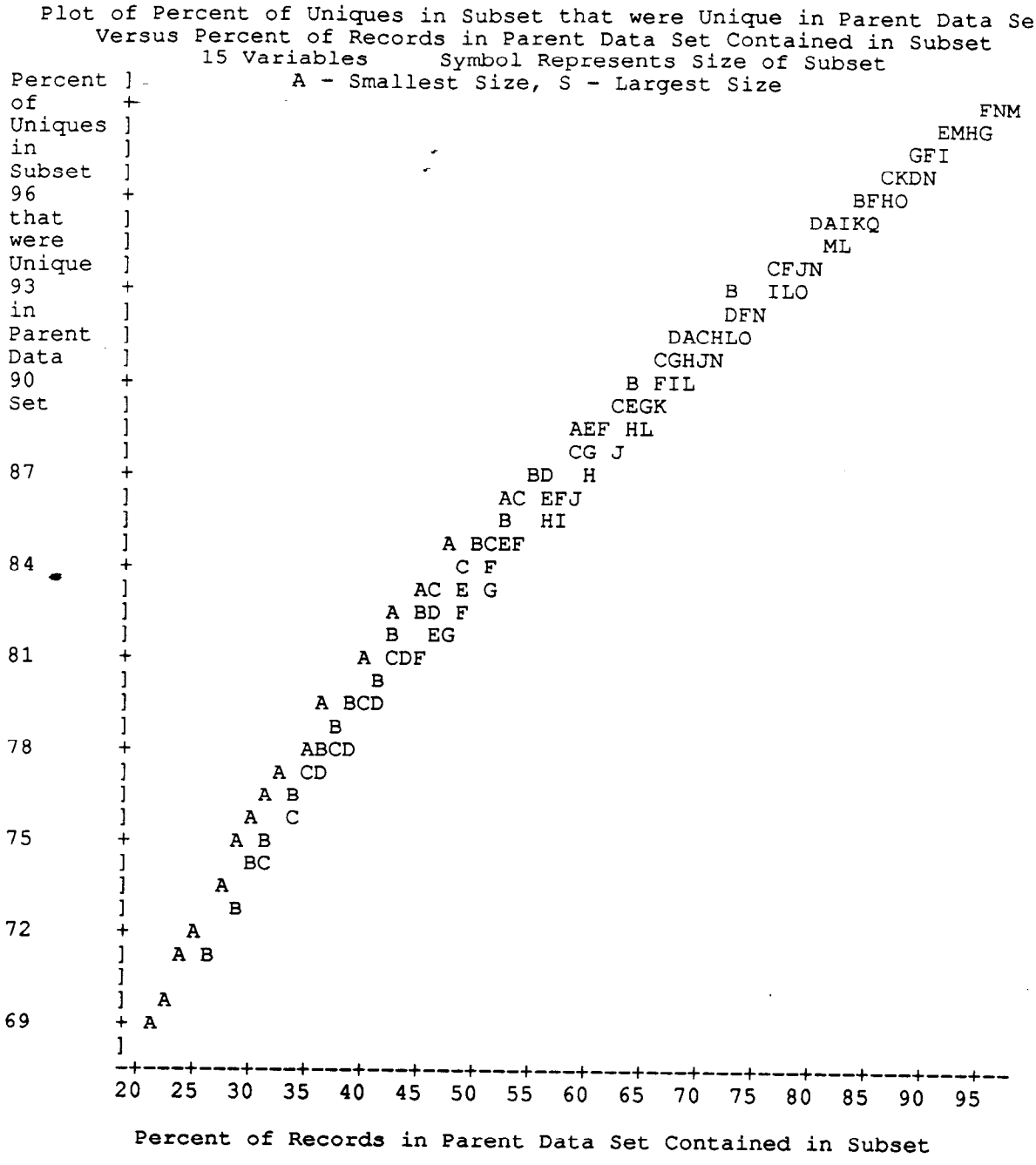

Plot of Percent of Uniques in Subset that were Unique in Parent Data Set
Versus Percent of Records in Parent Data Set Contained in Subset
15 Variables Symbol Represents Size of Subset
A - Smallest Size, S - Largest Size
Percent of Uniques in Subset 96 that were Unique 93 in Parent Data 90 Set
87
84
81
78
75
72
69
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

Figure 3

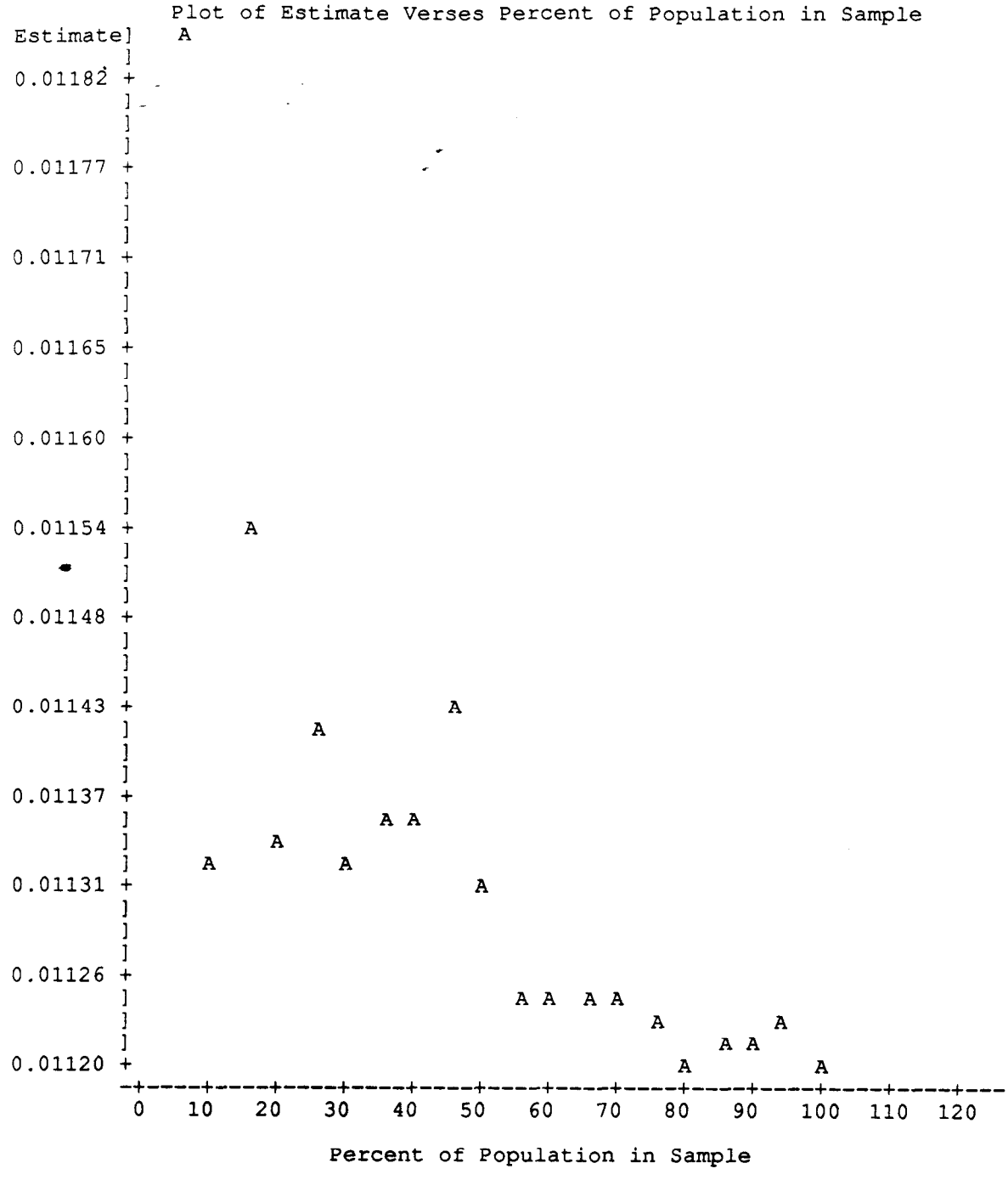
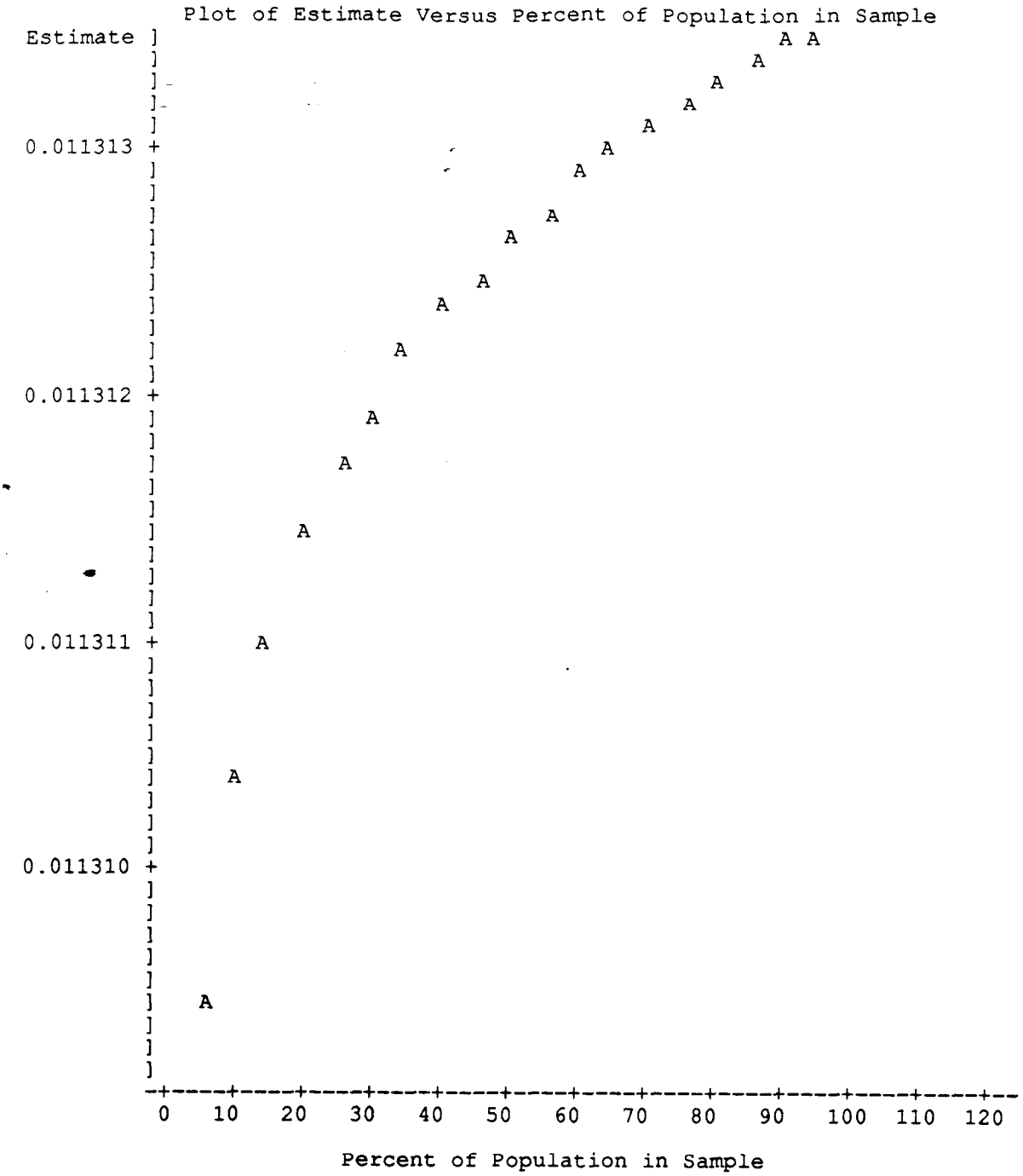

Plot of Estimate Verses Percent of Population in Sample
Estimate
0.01182
0.01177
0.01171
0.01165
0.01160
0.01154
0.01148
0.01143
0.01137
0.01131
0.01126
0.01120
0 10 20 30 40 50 60 70 80 90 100 110 120
Percent of Population in Sample

Figure 4


Plot of Estimate Versus Percent of Population in Sample
Estimate
0.011313
0.011312
0.011311
0.011310
0 10 20 30 40 50 60 70 80 90 100 110 120
Percent of Population in Sample

APPENDIX

POPULATION DATA SET #1
Data from the 1980 Decennial Census obtained from the Population Division of the Bureau of the Census

1. Persons

   Number of Person Records from Household (0 - 31)

2. Tenure

a. NA
b. Owner Occupied
c. Renter with Cash Rent
d. Renter with No Cash Rent

3. Household Type

a. NA
b. Married-couple Family Household
c. Family Household with Male Householder, No Wife Present
d. Family Household with Female Householder, No Husband Present

4. Household Income

   Household Income rounded to the nearest $10,000
   Bottom Code: -$10,000
   Top Code: $80,000

POPULATION DATA SET #2
Data from the January, 1987 Current Population Survey obtained from the Data User Services Division of the Bureau of the Census

1. Ethnicity

a. Mexican American
b. Chicano
c. Mexican / Mexicano
d. Puerto Rican
e. Cuban
f. Central or South American
g. Other Spanish
h. Don't Know
i. NA
j. All Other

2. Age

Age rounded to the nearest year

3. Marital Status

a. Married, Civilian, Spouse Present
b. Married, Armed Forces, Spouse Present
c. Married, Spouse Absent but Not Separated
d. Widowed
e. Divorced
f. Separated
g. Never Married

4. Sex

a. Male
b. Female

5. War

a. Vietnam
b. Korea
c. World War II
d. World War I
e. Other Service

POPULATION DATA SET #3
Data from the 1980 Decennial Census obtained from the Population Division of the Bureau of the Census

1. Tenure

a. NA
b. Owner Occupied
c. Renter with Cash Rent
d. Renter with No Cash Rent

2. Household Type

a. Everyone in Household Related
b. At Least Two but Not All Persons in Household Related
c. Single Person Household
d. Otherwise

3. Race

a. Class One, White Husband, White Wife
b. Class One, White Husband, Black Wife
c. Class One, White Husband, Indian Wife
d. Class One, White Husband, Asian / Pacific Islander Wife
e. Class One, Black Husband, White Wife
f. Class One, Black Husband, Black Wife
g. Class One, Black Husband, Indian Wife
h. Class One, Black Husband, Asian / Pacific Islander Wife
i. Class One, Indian Husband, White Wife
j. Class One, Indian Husband, Black Wife
k. Class One, Indian Husband, Indian Wife
l. Class One, Indian Husband, Asian / Pacific Islander Wife
m. Class One, Asian / Pacific Islander Husband, White Wife
n. Class One, Asian / Pacific Islander Husband, Black Wife
o. Class One, Asian / Pacific Islander Husband, Indian Wife
p. Class One, Asian / Pacific Islander Husband, Asian / Pacific Islander Wife
q. Class Two, Male Householder, White
r. Class Two, Female Householder, White
s. Class Two, Male Householder, Black
t. Class Two, Female Householder, Black
u. Class Two, Male Householder, Indian
v. Class Two, Female Householder, Indian
w. Class Two, Male Householder, Asian / Pacific Islander
x. Class Two, Female Householder, Asian / Pacific Islander
y. Class Three, White
z. Class Three, Black
aa. Class Three, Indian
bb. Class Three, Asian / Pacific Islander
cc. Otherwise

4. Ethnicity

a. Class One, Both Spouses Spanish
b. Class One, Male Spouse Spanish
c. Class One, Female Spouse Spanish
d. Class Two, Male Householder Spanish
e. Class Two, Female Householder Spanish
f. Class Three, Spanish
g. Otherwise

POPULATION DATA SET #3, continued

5. Children

a. NA
b. Householder with Own Children Under 6
c. Householder with Own Children Ages 6 - 17
d. Householder with Own Children, Some Under 6 and Some 6 - 17
e. Householder without children

6. Marital Status

a. Now Married
b. Widowed
c. Divorced
d. Separated
e. Never Married

POPULATION DATA SET #4
Data from the November, 1986 Current Population Survey obtained from the Data User Services Division of the Bureau of the Census

1. Ethnicity

a. Mexican American
b. Chicano
c. Mexican / Mexicano
d. Puerto Rican
e. Cuban
f. Central or South American
g. Other Spanish
h. Don't Know
i. NA
j. All Other

2. Age

Age rounded to the nearest ten years

3. Marital Status

a. Married, Civilian, Spouse Present
b. Married, Armed Forces, Spouse Present
c. Married, Spouse Absent but Not Separated
d. Widowed
e. Divorced
f. Separated
g. Never Married

4. Sex

a. Male
b. Female

5. War

a. Vietnam
b. Korea
c. World War II
d. World War I
e. Other Service

POPULATION DATA SET #4, continued

6. Highest Grade

a. None
b. One Year Elementary
c. Two Years Elementary
d. Three Years Elementary
e. Four Years Elementary
f. Five Years Elementary
g. Six Years Elementary
h. Seven Years Elementary
i. Eight Years Elementary
j. One Year High School
k. Two Years High School
l. Three Years High School
m. Four Years High School
n. One Year College
o. Two Years College
p. Three Years College
q. Four Years College
r. Five Years College
s. Six + Years College

7. Completed

a. Yes
b. No

POPULATION DATA SET #5
Data from the October, 1986 Current Population Survey obtained from the Data User Services Division of the Bureau of the Census

1. Ethnicity

a. Mexican American
b. Chicano
c. Mexican / Mexicano
d. Puerto Rican
e. Cuban
f. Central or South American
g. Other Spanish
h. Don't Know
i. NA
j. All Other

2. Age

Age rounded to the nearest ten years

3. Marital Status

a. Married, Civilian, Spouse Present
b. Married, Armed Forces, Spouse Present
c. Married, Spouse Absent but Not Separated
d. Widowed
e. Divorced
f. Separated
g. Never Married

4. Sex

a. Male
b. Female

5. War

a. Vietnam
b. Korea
c. World War II
d. World War I
e. Other Service

POPULATION DATA SET #5, continued

6. Highest Grade

a. None
b. One Year Elementary
c. Two Years Elementary
d. Three Years Elementary
e. Four Years Elementary
f. Five Years Elementary
g. Six Years Elementary
h. Seven Years Elementary
i. Eight Years Elementary
j. One Year High School
k. Two Years High School
l. Three Years High School
m. Four Years High School
n. One Year College
o. Two Years College
p. Three Years College
q. Four Years College
r. Five Years College
s. Six + Years College

7. Completed

a. Yes
b. No

8. Race

a. White
b. Black
c. Other

POPULATION DATA SET #6
Data from the 1980 Decennial Census obtained from the Population Division of the Bureau of the Census

1. Sex

a. Male
b. Female

2. Age

Age rounded to the nearest 10
Top Code: 90

3. Marital Status

a. Now Married
b. Widowed
c. Divorced
d. Separated
e. Single or NA

4. Race

a. White
b. Black
c. American Indian, Eskimo, Aleut
d. Japanese
e. Chinese
f. Filipino
g. Korean
h. Asian Indian
i. Vietnamese
j. Hawaiian
k. Other Asian and Pacific Islander
l. Spanish
m. Other

5. Spanish Origin

a. NA
b. Mexican
c. Puerto Rican
d. Cuban
e. Other

6. Highest Grade Attended

a. Never Attended School, or Attended Only Nursery School or Kindergarten
b. First through Eighth Grades
c. Ninth through Twelfth Grades
d. College

7. Disability

a. NA
b. With a Work Disability
c. No Work Disability

POPULATION DATA SET #6, continued

8. Labor

a. NA
b. Civilian, At Work
c. Civilian, With Job but Not At Work
d. Civilian, Unemployed
e. Armed Forces, At Work
f. Armed Forces, With Job but Not At Work
g. Not in Labor Force

9. Income from All Sources

Income rounded to the nearest $10,000
Bottom Code: -$10,000
Top Code: $80,000

POPULATION DATA SET #7
Data from the 1980 Decennial Census obtained from the Population Division of the Bureau of the Census

1. Tenure

a. NA
b. Owner Occupied
c. Renter with Cash Rent
d. Renter with No Cash Rent

2. Household Type

a. Everyone in Household Related
b. At Least Two but Not All Persons in Household Related
c. Single Person Household
d. Otherwise

3. Race

a. Class One, White Husband, White Wife
b. Class One, White Husband, Black Wife
c. Class One, White Husband, Indian Wife
d. Class One, White Husband, Asian / Pacific Islander Wife
e. Class One, Black Husband, White Wife
f. Class One, Black Husband, Black Wife
g. Class One, Black Husband, Indian Wife
h. Class One, Black Husband, Asian / Pacific Islander Wife
i. Class One, Indian Husband, White Wife
j. Class One, Indian Husband, Black Wife
k. Class One, Indian Husband, Indian Wife
l. Class One, Indian Husband, Asian / Pacific Islander Wife
m. Class One, Asian / Pacific Islander Husband, White Wife
n. Class One, Asian / Pacific Islander Husband, Black Wife
o. Class One, Asian / Pacific Islander Husband, Indian Wife
p. Class One, Asian / Pacific Islander Husband, Asian / Pacific Islander Wife
q. Class Two, Male Householder, White
r. Class Two, Female Householder, White
s. Class Two, Male Householder, Black
t. Class Two, Female Householder, Black
u. Class Two, Male Householder, Indian
v. Class Two, Female Householder, Indian
w. Class Two, Male Householder, Asian / Pacific Islander
x. Class Two, Female Householder, Asian / Pacific Islander
y. Class Three, White
z. Class Three, Black
aa. Class Three, Indian
bb. Class Three, Asian / Pacific Islander
cc. Otherwise

4. Ethnicity

a. Class One, Both Spouses Spanish
b. Class One, Male Spouse Spanish
c. Class One, Female Spouse Spanish
d. Class Two, Male Householder Spanish
e. Class Two, Female Householder Spanish
f. Class Three, Spanish
g. Otherwise

POPULATION DATA SET #7, continued

5. Children

a. NA
b. Householder with Own Children Under 6
c. Householder with Own Children Ages 6 - 17
d. Householder with Own Children, Some Under 6 and Some 6 - 17
e. Householder without children

6. Marital Status

a. Now Married
b. Widowed
c. Divorced
d. Separated
e. Never Married

7. Payment (Rent or Mortgage Plus Utilities, Tax, Insurance, Etc.)

a. = 0
b. < 50
c. < 75
d. < 100
e. < 125
f. < 150
g. < 175
h. < 200
i. < 250
j. < 300
k. < 400
l. < 500
m. < 600
n. < 700
o. < 800
p. < 900
q. < 1000
r. ≥ 1000

8. Employment / Unemployment

a. Class One, Both Spouses Unemployed
b. Class One, Husband Unemployed, Wife Employed
c. Class One, Husband Unemployed, Wife Not in Labor Force
d. Class One, Husband Employed, Wife Unemployed
e. Class One, Husband Not in Labor Force, Wife Unemployed
f. Class One, Both Spouses Not in Labor Force
g. Class One, Husband Not in Labor Force, Wife Employed
h. Class One, Husband Employed, Wife Not in Labor Force
i. Class One, Both Spouses Employed
j. Class Two, Male Householder Unemployed
k. Class Two, Male Householder Not in Labor Force
l. Class Two, Male Householder Employed
m. Class Two, Female Householder Unemployed
n. Class Two, Female Householder Not in Labor Force
o. Class Two, Female Householder Employed
p. Class Three, Unemployed
q. Class Three, Not in Labor Force
r. Class Three, Employed
s. Other

POPULATION DATA SET #7, continued

9. Veteran Status

a. Class One, Husband Veteran
b. Class One, Wife Veteran
c. Class One, Both Spouses Veterans
d. Class Two, at Least One Male in Household is Veteran
e. Class Two, at Least One Female in Household is Veteran
f. Class Two, at Least One Male and at Least One Female are Veterans
g. Class Three, Veteran
h. Otherwise

10. Disability

a. Class One, Husband Disabled
b. Class One, Wife Disabled
c. Class One, Both Spouses Disabled
d. Class Two, Male Householder Disabled
e. Class Two, Female Householder Disabled
f. Class Three, Disabled
g. Otherwise

POPULATION DATA SET #8
Data from the 1980 Decennial Census obtained from the Population Division of the Bureau of the Census

1. Percent Non-Hispanic White rounded to the nearest 10%
2. Percent Age 65+ rounded to the nearest 10%
3. Percent 4+ Years of College rounded to the nearest 10%
4. Percent Families Below Poverty rounded to the nearest 10%
5. Percent Female Families with Children rounded to the nearest 10%
6. Percent Age 14-18 in School rounded to the nearest 10%
7. Percent Males in Labor Force rounded to the nearest 10%
8. Percent Males Unemployed rounded to the nearest 10%
9. Median Age of Housing Unit rounded to the nearest 10 years
10. Median Value rounded to the nearest $10,000
11. Median Rent rounded to the nearest $100

POPULATION DATA SET #9
Data from the 1980 Decennial Census obtained from the Population Division of the Bureau of the Census

1. Tenure

a. NA
b. Owner Occupied
c. Renter with Cash Rent
d. Renter with No Cash Rent

2. Household Type

a. Everyone in Household Related
b. At Least Two but Not All Persons in Household Related
c. Single Person Household
d. Otherwise

3. Race

a. Class One, White Husband, White Wife
b. Class One, White Husband, Black Wife
c. Class One, White Husband, Indian Wife
d. Class One, White Husband, Asian / Pacific Islander Wife
e. Class One, Black Husband, White Wife
f. Class One, Black Husband, Black Wife
g. Class One, Black Husband, Indian Wife
h. Class One, Black Husband, Asian / Pacific Islander Wife
i. Class One, Indian Husband, White Wife
j. Class One, Indian Husband, Black Wife
k. Class One, Indian Husband, Indian Wife
l. Class One, Indian Husband, Asian / Pacific Islander Wife
m. Class One, Asian / Pacific Islander Husband, White Wife
n. Class One, Asian / Pacific Islander Husband, Black Wife
o. Class One, Asian / Pacific Islander Husband, Indian Wife
p. Class One, Asian / Pacific Islander Husband, Asian / Pacific Islander Wife
q. Class Two, Male Householder, White
r. Class Two, Female Householder, White
s. Class Two, Male Householder, Black
t. Class Two, Female Householder, Black
u. Class Two, Male Householder, Indian
v. Class Two, Female Householder, Indian
w. Class Two, Male Householder, Asian / Pacific Islander
x. Class Two, Female Householder, Asian / Pacific Islander
y. Class Three, White
z. Class Three, Black
aa. Class Three, Indian
bb. Class Three, Asian / Pacific Islander
cc. Otherwise

4. Ethnicity

a. Class One, Both Spouses Spanish
b. Class One, Male Spouse Spanish
c. Class One, Female Spouse Spanish
d. Class Two, Male Householder Spanish
e. Class Two, Female Householder Spanish
f. Class Three, Spanish
g. Otherwise

POPULATION DATA SET #9, continued

5. Children

a. NA
b. Householder with Own Children Under 6
c. Householder with Own Children Ages 6 - 17
d. Householder with Own Children, Some Under 6 and Some 6 - 17
e. Householder without children

6. Marital Status

a. Now Married
b. Widowed
c. Divorced
d. Separated
e. Never Married

7. Payment (Rent or Mortgage Plus Utilities, Tax, Insurance, Etc.)

a. = 0
b. < 50
c. < 75
d. < 100
e. < 125
f. < 150
g. < 175
h. < 200
i. < 250
j. < 300
k. < 400
l. < 500
m. < 600
n. < 700
o. < 800
p. < 900
q. < 1000
r. ≥ 1000

8. Employment / Unemployment

a. Class One, Both Spouses Unemployed
b. Class One, Husband Unemployed, Wife Employed
c. Class One, Husband Unemployed, Wife Not in Labor Force
d. Class One, Husband Employed, Wife Unemployed
e. Class One, Husband Not in Labor Force, Wife Unemployed
f. Class One, Both Spouses Not in Labor Force
g. Class One, Husband Not in Labor Force, Wife Employed
h. Class One, Husband Employed, Wife Not in Labor Force
i. Class One, Both Spouses Employed
j. Class Two, Male Householder Unemployed
k. Class Two, Male Householder Not in Labor Force
l. Class Two, Male Householder Employed
m. Class Two, Female Householder Unemployed
n. Class Two, Female Householder Not in Labor Force
o. Class Two, Female Householder Employed
p. Class Three, Unemployed
q. Class Three, Not in Labor Force
r. Class Three, Employed
s. Other

POPULATION DATA SET #9, continued

9. Veteran Status

a. Class One, Husband Veteran
b. Class One, Wife Veteran
c. Class One, Both Spouses Veterans
d. Class Two, at Least One Male in Household is Veteran
e. Class Two, at Least One Female in Household is Veteran
f. Class Two, at Least One Male and at Least One Female are Veterans
g. Class Three, Veteran
h. Otherwise

10. Disability

a. Class One, Husband Disabled
b. Class One, Wife Disabled
c. Class One, Both Spouses Disabled
d. Class Two, Male Householder Disabled
e. Class Two, Female Householder Disabled
f. Class Three, Disabled
g. Otherwise

11. Household Class

a. Householder has Spouse Present
b. Householder has No Spouse Present, Living with One or More Other Persons
c. Single Person Household

12. Household Income

a. $\leq$ 0
b. < 1000
c. < 3000
d. < 5000
e. < 7000
f. < 9000
g. < 11000
h. < 13000
i. < 15000
j. $\geq$ 15000

13. Social Security

a. = 0
b. < 500
c. < 1000
d. < 1500
e. < 2000
f. < 2500
g. $\geq$ 2500

14. Public Assistance

a. = 0
b. < 500
c. < 1000
d. < 1500
e. < 2000
f. < 2500
g. $\geq$ 2500

POPULATION DATA SET #9, continued

15. Other Income

a. = 0
b. < 500
c. < 1000
d. < 1500
e. < 2000
f. < 2500
g. < 5000
h. < 10000
i. < 15000
j. ≥ 15000