AN APPLICATION OF GOODMAN'S METHOD FOR ESTIMATING CLASSES IN A POPULATION

Barry I. Graubard and Robert J. Casady
National Center for Health Statistics

INTRODUCTION

A problem that confronts survey statisticians is how to estimate the number of different persons represented by a listing of records, where more than one record belongs to the same person. We want to accomplish this task by utilizing a simple random sample of records selected without replacement. Goodman[1] stated this problem as follows: "Suppose a population of known size N is subdivided into an unknown number of mutually exclusive classes. It is assumed that the class in which an element is contained may be determined, but that the classes are not ordered. Let us draw a random sample of n elements without replacement from the population. The problem is to estimate the total number of K classes which subdivide the population on the basis of sample results and knowledge of population size." Our population consists of a collection of records (these are our population elements) each of which correspond to only one person. The set of records corresponding to a person represent a class. We will assume that each record contains items of information such as name, age, sex, race, and mailing address of the person to whom the record belongs. This information will be used in a matching algorithm to determine which records in the sample belong to the same person. Goodman presented an unbiased estimator (we will call $\hat{K}_G$) of the number of classes, provided that the sample size is not less than the maximum number of elements in any single class. Also, he proved that this estimator was the unique unbiased estimator. He proposed three additional biased estimators and showed, using a hypothetical population, that one of the biased estimators (we will call $\hat{K}_{G2}$) had a smaller mean squared error (MSE) than $\hat{K}_G$.

In this paper, we will present another derivation of $\hat{K}_G$ and propose another biased estimator ($\hat{K}_{A2}$) which is a natural consequence of our derivation of $\hat{K}_G$. $\hat{K}_{A2}$ has a larger MSE than $\hat{K}_{G2}$ with respect to a hypothetical population provided by Goodman. However, $\hat{K}_{A2}$ is easier to obtain than $\hat{K}_{G2}$ when a matching algorithm is required to identify which pairs of elements are in the same class. Finally, we will consider estimators of K when our elements in the population are first divided up into clusters and a simple random sample of clusters are selected without replacement. An example is given of how this procedure might be used to increase the efficiency of the estimation of K.

DERIVATION OF UNBIASED ESTIMATOR

Goodman stated and proved the following:

Theorem: Suppose a sample of n elements is drawn with replacement from a population of size N which is subdivided into K classes. Let

$$A_i = 1-(-1)^i \frac{\binom{N-n+i-1}{N-n-1}}{\binom{n}{n-i}}, \quad i=1,\ldots,n \quad .$$

If there are x_i classes containing i elements in the sample then

$$E\left[\sum_{i=1}^{n} A_i x_i\right] = K$$

provided that n is not less than the maximum number q of elements contained in any class in the population.

Thus,

$$\hat{K}_G = \sum_{i=1}^{n} \left[1-(-1)^i \frac{\binom{N-n+i-1}{N-n-1}}{\binom{n}{n-i}}\right] x_i \qquad (1)$$

is the unique unbiased estimator (as Goodman proved).

Let us consider another approach which results in the same estimator $\hat{K}_G$. Suppose that we consider all $\binom{N}{j}$ subsets of the population which are size $j=1,2,\ldots,q$. If we set

$$\delta_{ji} = \begin{cases} 1 & \text{if all elements in the } i^{th} \text{ subset of size } j \text{ belong to the same class} \\ 0 & \text{otherwise} \end{cases}$$

then

$$K = \sum_{j=1}^{q} (-1)^{j-1} \sum_{i=1}^{\binom{N}{j}} \delta_{ji}, \text{ by an application of}$$

addition rule for the union of a collection of sets. If we set the random variable

$$\alpha_{ji} = \begin{cases} 1 & \text{if the sample contains the } i^{th} \text{ subset of size } j \\ 0 & \text{otherwise} \end{cases}$$

then

$$\hat{K}_A = \sum_{j=1}^{n} (-1)^{j-1} \frac{\binom{N}{n}}{\binom{N-j}{n-j}} \sum_{i=1}^{\binom{N}{j}} \delta_{ji} \alpha_{ji} \qquad (2)$$

is an unbiased estimator of K since $E(\alpha_{ji}) = \frac{\binom{N-j}{n-j}}{\binom{N}{n}}$ for $j=1,\ldots,q$ and $i=1,\ldots,\binom{N}{j}$,

provided that $n \geq q$.

It can be shown that $\hat{K}_A = \hat{K}_G$. This derivation is a simpler one than the one given by Goodman because it makes use of the well-known addition formula for the union of sets as opposed to a more complicated combinatorial identity that

Goodman uses. We will assume for the rest of the paper that $n \geq q$.

DERIVATION OF BIASED ESTIMATORS

Let K_i denote the number of classes containing i elements in the population so that $K_1+\ldots+K_q = K$. Since the statistic $\hat{K}_G$ can have a large variance, when there are classes in the population with more than two elements, it is reasonable to consider biased estimators for K that may have smaller MSE than $\hat{K}_G$. Goodman proposed several biased estimators of K. The one that is discussed here (we will denote as $\hat{K}_{G2}$) had a smaller MSE than $\hat{K}_G$ for a hypothetical population which Goodman constructed. The estimator

$$\hat{K}_{G2} = N - \frac{N(N-1)}{n(n-1)} x_2 \tag{3}$$

is very similar to another biased estimator

$$\hat{K}_{A2} = N - \frac{N(N-1)}{n(n-1)} \sum_{i=1}^{\binom{N}{2}} \delta_{2i}\, a_{2i} \ . \tag{4}$$

Note that $\hat{K}_{A2}$ is just the first two terms of $\hat{K}_A$. $\hat{K}_{G2}$ requires somewhat different information than $\hat{K}_{A2}$. In order to calculate $\hat{K}_{G2}$, one needs to identify and count the classes that have exactly two of their elements in the sample, whereas in order to use $\hat{K}_{A2}$ one needs to count the number of times that pairs of elements belong to the same class. The difference between these two estimates lies in the type of matching algorithm that is needed to identify when two elements (records) belong to the same class (person). The estimator $\hat{K}_{G2}$ requires the matching algorithm to be able to link all the elements that belong to the same class and the estimator $\hat{K}_{A2}$ requires the matching algorithm to only make determinations about whether each pair of elements belong to the same class. When each element in the population does not have a unique identifier which determines the the class it belongs (as is the usual case in practice) then constructing a reliable matching algorithm to link together all elements in samples that are in the same class can be quite difficult. Thus, for some problems it may be easier to apply $\hat{K}_{A2}$ than $\hat{K}_{G2}$.

The MSE's of $\hat{K}_{G2}$ and $\hat{K}_{A2}$ were compared analytically, assuming that the elements in the sample were selected using a binomial sampling scheme. Binomial sampling is a reasonable approximation to simple random sampling without replacement and it simplifies the computations, enabling us to compare the MSE's of $\hat{K}_{G2}$ and $\hat{K}_{A2}$. If each element in the population has equal and independent chance p of entering the sample, where the size of the sample is a random variable which is binomially distributed with mean Np, then we say we used binomial sampling to select the sample.

Denote the estimators $\hat{K}_{G2}$ and $\hat{K}_{A2}$ under binomial sampling by $\hat{B}_{G2}$ and $\hat{B}_{A2}$, respectively, then

$$\hat{B}_{G2} = N - \frac{1}{p^2} x_2 \tag{5}$$

and

$$\hat{B}_{A2} = N - \frac{1}{p^2} \sum_{i=1}^{\binom{N}{2}} \delta_{2i} \ . \tag{6}$$

Note that $\hat{B}_{G2} = \hat{B}_{A2}$ (as well as $\hat{K}_{G2} = \hat{K}_{A2}$) when $K_i = 0$ for $i=3,\ldots,q$.

$$\text{The MSE } (\hat{B}_{G2}) = \sum_{j=2}^{q} \binom{j}{2} \frac{(1-p)^{j-2}}{p^2} K_j - \sum_{j=2}^{q} \binom{j}{2} (1-p)^{2(j-2)} K_j + \left[\sum_{i=3}^{q} [1-i+\binom{i}{2} (1-p)^{i-2}] K_i \right]^2 \tag{7}$$

$$\text{and MSE } (\hat{B}_{A2}) = \frac{(1-p)(1+p)}{p^2} \sum_{i=2}^{q} \binom{i}{2} K_i + 3! \frac{(1-p)}{p} \sum_{i=3}^{q} \binom{i}{3} K_i + \left[\sum_{i=3}^{q} \binom{i-1}{2} K_i \right]^2 \tag{8}$$

$$\text{where Bias } (\hat{B}_{G2}) = \sum_{i=3}^{q} [(1-i)+ \binom{i}{2} (1-p)^{i-2}] K_i$$

$$\text{and Bias } (\hat{B}_{A2}) = \sum_{i=3}^{q} \binom{i-1}{2} K_i \ .$$

It can easily be seen that if p is small (i.e. $(1-p) \approx 1$) then Bias $(\hat{B}_{G2}) \approx$ Bias $(\hat{B}_{A2})$ and further the $|$ Bias $(\hat{B}_{G2})$ $| \leq$ Bias $(\hat{B}_{A2})$ for all $0 < p \leq 1$. It should be also be noted that the Bias $(\hat{B}_{A2}) \geq 0$ so that $\hat{B}_{A2}$ always overestimates K if $K_i > 0$ for any $i=3,\ldots,q$ and the Var $(\hat{B}_{G2}) \leq$ Var $(\hat{B}_{A2})$ which implies that MSE $(\hat{B}_{G2}) \leq$ MSE $(\hat{B}_{A2})$.

Using Goodman's hypothetical population of 10,000 elements with $K_1 = 9225$, $K_2 = 336$, $K_3 = 33$, and $K_4 = 1$ and selecting a binomial sample with $p = 1/10$, we obtain an empirical comparison of $\hat{B}_{A2}$ and $\hat{B}_{G2}$ with the $\sqrt{\text{MSE } (\hat{B}_{A2})} = 214$ and $\sqrt{\text{MSE } (\hat{B}_{G2})} = 207$. This population is typical of populations one sees in practice. Thus, $\hat{B}_{A2}$ and $\hat{B}_{G2}$ can have MSE's that are close in magnitude.

Goodman calculated that the $\sqrt{\text{MSE } (\hat{K}_G)} = 347$, and it is easy to see the great savings one can achieve by using biased estimators such as $\hat{B}_{2A}$ and $\hat{B}_{2G}$. However, if we had a population with a greater frequency of classes containing three, four and more elements per class then the variances and biases of $\hat{B}_{A2}$ and $\hat{B}_{G2}$ will become larger. The next section describes a method that could reduce the MSE's for estimators like $\hat{K}_{A2}$ and $\hat{K}_{G2}$.

DERIVATION OF ESTIMATORS FOR CLUSTER SAMPLING

In this section we will suggest grouping our elements in the population into clusters in such a way so that as the frequency of classes with elements in more than two clusters is smaller than the frequency of classes with more than two elements (i.e., in the previous sections we used clusters with one element) and that most of the clusters represent just one class. Next, we will consider estimators of K that are functions of a sample of clusters.

We will formalize this new sampling procedure as follows: Suppose that we divide a population of N elements into L clusters (not all necessarily the same size). The clusters are constructed so that they have a relatively small number of elements. The strategy for determining the clusters should be such that most of the elements in each cluster come from the same class and that the number of classes with elements in more than two clusters is small. Next we select a simple random sample of clusters without replacement. Each cluster that is selected into the sample has its elements grouped into the classes to which they belong. This step should be relatively simple since the cluster sizes are small. The estimator of K which is a simple extension of $\hat{K}_G$ is:

$$\hat{C}_G = \sum_{i=1}^{\ell} \left[1-(-1)^i \frac{\binom{L-\ell+i-1}{L-\ell-1}}{\binom{\ell}{\ell-i}}\right] z_i \qquad (9)$$

where z_i are the number of classes containing elements in i clusters in the sample. Again, $\hat{C}_G$ is an unbiased estimator for K provided that ℓ is not less than the maximum number w of clusters in which any class in the population contains elements. Another form for $\hat{C}_G$, which parallels $\hat{K}_A$, is

$$\hat{C}_A = \sum_{j=1}^{\ell} (-1)^{j-1} \frac{\binom{L}{\ell}}{\binom{L-j}{\ell-j}} \sum_{i=1}^{L} \gamma_{ji}\, \beta_{ji}, \qquad (10)$$

where γ_{ji} is the number of classes that have elements in all the clusters of the i^{th} subset of j clusters and

$$\beta_{ji} = \begin{cases} 1 \text{ if all the clusters in the } i^{th} \text{ subset of } j \text{ clusters are selected into the sample} \\ 0 \text{ otherwise} \end{cases}$$

Biased estimators can be constructed using the first two terms of $\hat{C}_G$ and $\hat{C}_A$ in the same way as before when $\hat{K}_{G2}$ and $\hat{K}_{A2}$ were formed. These biased estimators are

$$\hat{C}_{G2} = \frac{L}{\ell}\left[\sum_{i=1}^{L} \gamma_{1i}\, \beta_{1i}\right] - \frac{L(L-1)}{\ell(\ell-1)} z_2 \text{ and} \qquad (11)$$

$$\hat{C}_{A2} = \frac{L}{\ell}\left[\sum_{i=1}^{L} \gamma_{1i}\, \beta_{1i}\right] - \frac{L(L-1)}{\ell(\ell-1)}\left[\sum_{i=1}^{\binom{L}{2}} \gamma_{2i}\, \beta_{2i}\right] \qquad (12)$$

Under binomial sampling of clusters, the MSE's for $\hat{C}_{A2}$ and $\hat{C}_{G2}$ do not simplify enough to be able to make a comparison between them or with the MSE's of $\hat{B}_{A1}$ and $\hat{B}_{G2}$. However, if we further simplify the problem by supposing that the elements in the population are divided into clusters where no cluster has elements from more than one class then the MSE $(\hat{C}_{G2}) \leq$ MSE $(\hat{B}_{G2})$. This is a straightforward result of equation (7) and it indicates how one can reduce the MSE of the estimators for K by clustering the elements.

In the last section, we will describe one way to form clusters of elements that will, hopefully, increase the efficiency of the estimation of K.

A PROPOSED APPLICATION

A real life problem in which we plan to apply the estimators in this paper can be described as follows: The National Center for Health Statistics (NCHS) has a file of State pharmacist records where each record consists of information provided to NCHS by a pharmacist at the time of the renewal of his State pharmacy license. Therefore, the number of records a pharmacist has in the file is the same as the number of States in which he is licensed. Each record contains a name and a mailing address of a pharmacist; and for pharmacists that provided NCHS with the requested information there is age, race, sex, year of graduation and other information about his practice of pharmacy.

We want to estimate the number of pharmacists (classes) represented by our listing of records (elements) by selecting a sample of records. In order to identify the records in the sample which belong to the same pharmacist, we plan to use a matching algorithm that compares items of information on each pair of records in the sample and makes a determination whether the records correspond to the same pharmacist or not. Of course, a pharmacist may be licensed under different names or have different mailing addresses Thus, there can be matching errors made by the algorithm but we will ignore these errors for reasons of simplicity.

One way to select a sample of records that may increase the efficiency of our estimation is to select clusters of records where the clusters contain most of the records for one pharmacist. A proposed way to divide our population into clusters is by grouping together records with the same name on them. Thus, pharmacists with more than one license under different names will have licenses in more than one cluster and pharmacists with the same name will have their records in the same cluster. We feel that for most of the clusters the records in each cluster will belong to just one pharmacist and that few pharmacists will have records in more than one cluster. If this is true, we should be able to decrease the MSE of our estimators for the total number of pharmacists by selecting a sample of clusters of records.

The generalization of Goodman's method for estimating the number of classes in a population seems to have theoretical promise for problems where good auxiliary information is available for determining the clusters of population elements. However, applications of this methodology needs to be conducted in order to determine empirically its gains in efficiency over not clustering the elements.

[1]Leo A. Goodman, "On the Estimation of Classes in a Population", Ann. Math. Statist. (1949) pps 72-79.