ON THE ESTIMATION OF THE NUMBER OF CLASSES IN A POPULATION[1]

By Leo A. Goodman

Princeton University

1. Summary. This paper deals with the following problem: 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 K of classes which subdivide the population on the basis of the sample results and our knowledge of the population size.

There is exactly one real valued statistic S which is an unbiased estimate of K when the sample size n is not less than the maximum number q of elements contained in any class. The restriction placed upon q is unimportant for many practical problems where either there is a reasonably low bound for q or those classes containing more than n elements are known. An unbiased estimate does not exist when there is no such knowledge.

Since the unbiased estimate can be very unreasonable, modifications of S are considered. The statistic

$$T' = \begin{cases} S' = N - \dfrac{N(n-1)}{n(n-1)}\, x_2, & \text{if } S' \geq \sum_{i=1}^{n} x_i, \\ \sum_{i=1}^{n} x_i, & \text{if } S' < \sum_{i=1}^{n} x_i, \end{cases}$$

where x_i is the number of classes containing i elements in the sample, is the most suitable estimate, in comparison with three other statistics, for a hypothetical population.

The case where each element in the population has an equal and independent chance of coming into the sample is used as a model for some sampling procedures and also as an approximation to the case of random sampling.

2. Introduction.

The problem discussed may be described in terms of colored balls in an urn. How should we estimate the number of colors present in the urn on the basis of both the sample which gives the number of, say, white balls, red balls, etc., and our knowledge of the total number of balls in the urn:

The following practical cases illustrate some of the ways in which this problem presents itself:

(1) A company has received a large number of requests for a free sample of its product. It is known that the same people often send more than one request.

[1] Prepared in connection with research sponsored by the Office of Naval Research.

From a sample of the requests we wish to estimate how many different people have sent requests.[2]

(2) The Social Security Board possesses a large collection of Social Security cards. It is known that some people obtain different cards when they change jobs. From a sample of the cards it is desired to estimate how many different people have Social Security cards.[3]

(3) A person who sells durable commodities anticipates opening a store which is to be located at a highway intersection. He would like to know how many different automobiles pass through the intersection in a given time period. The total number of automobiles may be easily observed but some probably pass through more than once. This type of inquiry is also useful to advertising agencies which must decide the most efficient location for billboards.

(4) The State Unemployment Compensation Board possesses a large list of the people receiving unemployment benefits. It is desired to estimate the total number of families benefiting from the insurance program on the basis of a random sample of the people named on the list.

(5) The number of words in a book may be easily estimated and a sample can be taken. The problem of estimating the number of different words in a book is another analogue of the general problem.[4]

3. Results and derivations. In order to show that an unbiased estimate of the number of classes in a population exists when the sample size n is not less than the maximum number q of elements contained in any class, we need prove the following two statements:

LEMMA 1. *Suppose we have K classes of N similar elements with n_1 elements in class 1, n_2 elements in class 2, $\cdots$, n_K elements in class K. The class of an element is readily identifiable when the element is examined. Let*

$$q = \max (n_s).$$

Suppose a random sample is drawn without replacement. If x_i is the number of classes containing i elements in the sample, and K_j is the number of classes containing j elements in the population, then

$$E(x_i) = \sum_{j=i}^{q} \Pr(i \mid j, N, n)K_j ,$$

where $Pr(i \mid j, N, n)$ shall henceforth be an abbreviation of

$$\frac{C_i^j \, C_{n-i}^{N-j}}{C_n^N} .$$

[2] Submitted by Charles Callard to Question and Answers, *The American Statistician*, Vol. 3, No. 1, p. 23.

[3] Mentioned to the author by Dr. J. Stevens Stock of Opinion Research Corporation.

[4] Mentioned in letter to the author from Frederick Mosteller of Harvard University.

PROOF. Let y_s be the number of elements appearing in the sample from the s-th class. The statement is proved by considering $E(x_i) = \sum_{s=1}^{K} E(\delta_{iy_s})$, where

$$\delta_{iy_s} = \begin{cases} 1, & \text{if } y_s = i, \\ 0, & \text{if } y_s \neq i. \end{cases}$$

LEMMA 2. *Let*

$$a^{(t)} = \begin{cases} a(a-1)(a-2) \cdots (a-t+1), & \text{for } t > 0, \\ 1 & , \text{ for } t = 0. \end{cases}$$

If

$$a_i = 1 - (-1)^i \frac{[N-n+i-1]^{(i)}}{n^{(i)}},$$

then

$$\sum_{i=1}^{j} A_i \Pr(i \mid j, N, n) = 1.^{5}$$

This result follows directly from the fact that

$$\sum_{i=0}^{j} (-1)^i C_i^j [N-n+i-1]^{(j-1)} = 0, \text{for } j \geqq 1.$$

The following theorem may be proved directly by the preceding lemmas:

THEOREM 1. *Suppose a sample of n elements is drawn without replacement from a population of size N which is subdivided into K classes. Let*

$$A_i = 1 - (-1)^i \frac{[N-n+i-1]^{(i)}}{n^{(i)}}.$$

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.

THEOREM 2. *There is at most one real valued statistic which is an unbiased estimate of the number of classes in a population.*[6]

PROOF. Let us order the points of the sample space in the following manner: Letting x_i be the number of classes containing i elements in the sample, order the sample points by increasing values of x_n; for equal values of x_n, order the points by increasing values of x_{n-1}; for equal values of x_{n-1}, order the points

[5] The author is indebted to Professor Frederick F. Stephan of Princeton University for a statement leading to a simplification of the original result.

[6] This statement was mentioned to the author by M. P. Peisakoff of Princeton University.

by increasing values of x_{n-2} ; $\cdots$; for equal values of x_3, order the points by increasing x_2. Let

$$x_1 = n - \sum_{j=2}^{n} jx_j .$$

To prove the theorem, we must show that to each 0_i there corresponds a unique value $S(i)$, which must be the value of our estimate when 0_i is observed, in order that the statistic be unbiased. To each

$$0_i = [x_1(i), x_2(i), x_3(i), \cdots , x_n(i)],$$

let us associate the population

$$P_i = \left[N - \sum_{j=2}^{n} jx_j(i), x_2(i), x_3(i) \cdots , x_n(i) \right].$$

If P_1 is the underlying population, then 0_i for all $i > 1$ will occur with a probability of zero. Since there are N classes in P_1, the value of the statistic must be $S(1) = N$ whenever 0_1 is observed in order that the estimate be unbiased. The theorem may now be proved by induction.

Since all the P_i used in the proof of Theorem 2 satisfied the condition that the maximum number q of elements contained in any class be not more than the sample size n, the statistic S is the only real valued statistic which is an unbiased estimate when $q \leq n$.

When the restriction that $q \leq n$ is removed, it is useless to search for an unbiased estimate since we have

THEOREM 3. *There does not exist an unbiased estimate of the number of classes subdividing a population when it is not known whether the maximum number q of elements contained in any class is not more than the sample size n.*

By the preceding theorems it is clear that if an unbiased estimate exists it must equal S. However, S is generally not unbiased when $n < q$.

THEOREM 4. *Suppose the statistics S_1, S_2, $\cdots$, S_n are the solutions of the system of linear equations*

$$x_i = \sum_{j=1}^{n} \Pr(i \mid j, N, n)S_j , \text{for } i = 1, 2, \cdots , n,$$

where x_i is the number of classes containing i elements in a sample of size n from a population of N elements. If K_j is the number of classes containing j elements in the population, then $E(S_j) = K_j$, for $j = 1, 2, \cdots , n$ when n is not less than the maximum number q of elements contained in any class.

PROOF. We observe that the statement is certainly true for $j = q + 1, q + 2, \cdots , n$, since

$$E(S_j) = K_j = 0, \quad \text{for} \quad j = q + 1, q + 2, \cdots , n.$$

The statement is also true for $j = q$, since

$$E(S_q) = E(x_q) \frac{N^{(q)}}{n^{(q)}} = K_q .$$

To prove that $E(S_j) = K_j$, for any $j < q$, we assume it to be true for all $i > j$, whereupon its truth for j follows.

By Theorem 2, and 3, it is clear that $\sum_{j=1}^{n} S_j = S$. Since

$$\sum_{j=1}^{q} jK_j = N,$$

it seems reasonable to ask whether the values of the estimates $S_1, S_2, \cdots, S_n$ are in agreement with the known value of the size of the population. The unbiased estimate of K can be shown to be internally consistent by

THEOREM 5. *Suppose a sample of size n is drawn without replacement from a population of N elements which is divided into classes. If x_i is the number of classes containing i elements in the sample, and if the linear equations*

$$x_i = \sum_{j=1}^{n} \Pr(i \mid j, N, n)S_j,$$

are solved simultaneously for S_j, then

$$\sum_{j=1}^{n} jS_j = N.$$

The theorem follows readily from the fact that

$$\sum_{i=1}^{j} i \Pr(i \mid j, N, n) = n \frac{j}{N} \text{ and } \sum_{i=1}^{n} ix_i = n.$$

The variance of S may now be calculated by means of the formula

$$\sigma_S^2 = \sum_{i,j=1}^{n} A_i A_j u_{ij} = \sum_{i,j=1}^{n} A_i A_j \left\{ \sum_{s,t=1}^{q} m_{st}(i, j)K_s K_t + \sum_{s=1}^{q} [m_s(i, j) - m_{ss}(i, j)]K_s \right\},$$

where u_{ij} is the covariance between x_i and x_j, $m_{st}(i, j)$ is the covariance between δ_{iy_r} and δ_{jy_h} when $r \neq h$, $n_r = s$ and $n_h = t$, and $m_s(i, j)$ is the covariance between δ_{iy_r} and δ_{jy_r} when $n_r = s$.

Since the statistic S can be very unreasonable, we consider other possible estimates of K. The statistic

$$S' = N - \frac{N^{(2)}}{n^{(2)}} x_2$$

may be shown to be a modification of S which replaces the number x_i of classes containing $i > 2$ elements in the sample by an additional ix_i classes, each containing only one element. Since the values of K_i for $i > 2$ are relatively small in the practical problems of Section 2, S' might be used as an estimate.

Another statistic which may be used to estimate K is

$$S'' = \frac{N}{n} \sum_{i=1}^{n} x_i.$$

This statistic may be shown to overestimate K whenever $q \neq 1$. The estimate

$$S'' = \sum_{i=1}^{n} x_i$$

underestimates K when $n \leq N - m$ where m is the least number of elements contained in any class.

4. Binomial sampling. Let us suppose that each element from a population of N elements has an equal and independent chance $p = 1/r$ of entering the sample s. In this case, the size of the sample obtained is a random variable η which is binomially distributed with mean Np. If a large random sample of n elements is drawn without replacement from a large population of size N, then the results when interpreted in terms of binomial samples where $p = 1/r = n/N$ are a good approximation to the results obtained by the usual model. Binomial sampling may be considered a model of the case where one attempts to obtain the sampling ratio $p = 1/r$ by drawing simultaneously an uncounted sample of elements which is estimated as being of the appropriate size.

In the case of binomial sampling, the statistic

$$B = \sum_{i=1}^{N} B_i x_i, \quad \text{where} \quad B_i = 1 - (1 - r)^i$$

may be shown to be an unbiased estimate of the number of classes in a population from which binomial samples are drawn.

Let us now consider the statistic which corresponds to S' for the case of binomial sampling; i.e.,

$$B' = N - r^2 x_2 .$$

It may be shown that

$$E(B') = K_1 + K_2 + \sum_{j=3}^{q} [j - C_2^j (1 - p)^{j-2}] K_j .$$

Hence, the statistic B' will underestimate K whenever

$$p < 1 - \left(\frac{2}{j}\right)^{1/j-2}, \quad \text{for} \quad j = 3, 4, \cdots, q.$$

Since

$$1 - \left(\frac{2}{j}\right)^{1/j-2}$$

is a decreasing function of j for $j > 2$, when $p > \frac{1}{3}$, B' overestimates, and when

$$p < 1 - \left(\frac{2}{p}\right)^{1/q-2},$$

B' underestimates the value of K. When p is such that

$$1 - \left(\frac{2}{q}\right)^{1/q-2} \leq p \leq \tfrac{1}{3},$$

the expected value of B' is brought closer to K by underweighting some K_j and overweighting others.

5. A hypothetical population.[7] Suppose we draw a random sample of 1000 elements without replacement from a population of 10,000 elements where

$$K_1 = 9225, \qquad K_2 = 336, \qquad K_3 = 33, \qquad K_4 = 1.$$

Hence, $K = 9595$. By means of Table 1, let us now compare on the basis of binomial sampling the estimates which have been presented in the preceding sections. Since N and n are large, these results are a good approximation to the case of random sampling without replacement.

TABLE 1

Estimate	*Expected value*	*Bias*	$\sqrt{\textit{Mean Square Error}}$
S	9595	0	347
S'	9570	−25	207
S''	9959	364	490
S'''	996	−8599	8600

It is clear that the best estimates of the number of classes in this particular population are S or S', since S has the least bias, $E(S) - K$, and S' has the least mean square error, $E(S' - K)^2$. One might argue that both S and S' are the statistics which are capable of giving nonsensical estimates. However, we may decide to modify S or S' in order to always get reasonable estimates by using the statistics

$$T = \begin{cases} S, & \text{if } N \geq S \geq \sum_{i=1}^{n} x_i, \\ N, & \text{if } S > N, \\ \sum_{i=1}^{n} x_i, & \text{if } S < \sum_{i=1}^{n} x_i \end{cases}$$

$$T' = \begin{cases} S', & \text{if } S' \geq \sum_{i=1}^{n} x_i, \\ \sum_{i=1}^{n} x_i, & \text{if } S' < \sum_{i=1}^{n} x_i. \end{cases}$$

[7] Other examples have been investigated by Frederick Mosteller in Questions and Answers, *The American Statistician,* Vol. 3, No. 3, p. 12.

Although these modified statistics T and T' are not unbiased, they have the desirable property that

$$MSE(T) \leq MSE(S), \quad \text{and} \quad MSE(T') \leq MSE(S').$$

Since this hypothetical population is a plausible one for the practical problems of Section 2, the modified statistics T or T' seem, therefore, to be "best" for estimating the number of classes for these problems, where the "best" statistic is defined as the one which never gives unreasonable estimates and has the least mean square error.

The author wishes to express his appreciation to Professor John W. Tukey whose suggestions were very helpful.