This greatest lower bound can also be obtained by using the Bhattacharyya bound (3.2) without applying a limiting operation:

$$\frac{p_0 q_0}{n} = \frac{1}{\sigma^2_{S_1 - S_2/2! + \cdots - (-1)^n S_n/n!}},$$

where the S's are defined (3.1) as the divided differences corresponding to ordinary differences with interval h (h being chosen sufficiently small that all p's fall between 0 and 1).

REFERENCES

[1] D. G. CHAPMAN AND H. ROBBINS, "Minimum variance estimation without regularity assumptions," *Annals of Math. Stat.*, Vol. 22 (1951), pp. 581–586.

[2] E. L. LEHMANN, "Notes on the Theory of Estimation," mimeographed notes.

ON THE ANALYSIS OF SAMPLES FROM k LISTS[1]

BY LEO A. GOODMAN

The University of Chicago

1. Introduction and summary. Suppose we have k lists of names, no name appearing more than once in each list. We are interested in estimating the following parameters: (a) the number of names occurring in common in pairs, triples, $\cdots$, of lists; (b) the number of names occurring in 1, 2, $\cdots$, k lists. This note presents unbiased estimators for these parameters when a random sample is drawn from each list. It is also observed that the estimators presented are the only real-valued statistics which are unbiased estimators of the parameters, and hence must be the minimum variance unbiased estimators. This yields another example in which "insufficient" statistics have been used to obtain minimum variance unbiased estimators.

These unbiased estimators may at times give unreasonable estimates. In such cases, it is suggested that the statistics be modified so that the nearest reasonable estimate is used. Although this procedure introduces some bias, it usually reduces the mean square error.

This problem arises when we are interested in tracing the interrelations of agencies through the individual members. The problem also arises in the work of H. H. Fussler and J. M. Dawson of the University Library, University of Chicago, who are interested in comparing the acquisitions of various libraries. For special problems other sampling schemes may be more economical or more efficient than taking a sample from each list. Professor F. F. Stephan of Princeton University pointed out to the author that, in the special case of the "library problem," the Book Catalog and author cards used by many libraries provide a convenient means of drawing matched samples. (There is a brief discussion

[1] This work was prepared in connection with research supported by the Office of Naval Research.

of this kind of sampling problem on page 571 of [1].) A sampling scheme based on the last digit or two of the serial number of the cards could be used to search each library reference file for the same list of books. Special provision must be made for accessions made outside the sampling period and for books not covered by the Library of Congress cards. The analysis presented herein deals with the case in which (either for good, bad, or no reasons) a random sample has been drawn from each list.

The restriction that no name appear more than once in each list may be weakened to obtain somewhat more general results.

The problem discussed in this paper was brought to the author's attention by Professor W. Allen Wallis of the University of Chicago.

2. Results.

THEOREM 1. *Given k lists of names, let d_{12} names occur in common in lists 1 and 2, d_{13} names occur in common in lists 1 and 3, $\cdots$, $d_{[t]}$ names occur in common in lists $[t]$ (where $[t]$ is some subset containing at least two of the integers $1, 2, \cdots, k$), $\cdots$, $d_{123\ldots k}$ names occur in all lists. Suppose a random sample of $n_i = N_i/g_i$ names is drawn from list i, which contains N_i names, for $i = 1, 2, \cdots, k$. If e_{12} names occur in common in the samples from lists 1 and 2, e_{13} names occur in common in the samples from lists 1 and 3, $\cdots$, $e_{[t]}$ names occur in common in the samples from lists $[t]$, $\cdots$, $e_{12\ldots k}$ names occur in all samples, then an unbiased estimator of $d_{[t]}$ is*

$$d^1_{[t]} = \prod_i g_i e_{[t]},$$

where the product is taken over all values of i appearing in $[t]$.

The proof is based on the fact that $e_{[t]} = \sum \delta_{j[t]}$, where

$$\delta_{j[t]} = \begin{cases} 1, \text{ if name } j \text{ appears in all the samples from lists } [t] \\ 0, \text{ otherwise,} \end{cases}$$

and the summation is taken over all names.

THEOREM 2. *An unbiased estimator of the number of names occurring in ν lists is*

$$\sum_{i=0}^{k-\nu} (-1)^i C_\nu^{\nu+i} d^1(\nu + i),$$

where $d^1(\nu + i) = \sum d^1_{[t]}$ and the summation is taken over all $[t]$ containing $\nu + i$ integers. Also, an unbiased estimator of the number of names occurring in at least ν lists is

$$\sum_{i=0}^{k-\nu} (-1)^i C_{\nu-1}^{\nu+i-1} d^1(\nu + i).$$

The proof of these results follows from Theorem 1 and some combinatorics.

THEOREM 3. *Let F be a real-valued function of the parameters $d_{12}, d_{13}, \cdots, d_{[t]}, \cdots, d_{12\ldots k}$. Then there can be at most one real-valued function S of the sample results $e_{12}, e_{13}, \cdots, e_{[t]}, \cdots, e_{12\ldots k}$, such that $E\{S\} = F$, for all values of the parameters.*

PROOF. Let $2^k - k - 1 = M$. Suppose we order the M subsets $\{[t]\}$. To simplify notation we shall designate $d_{[t]}$ and $e_{[t]}$ by d_i and e_i, respectively, where $i = i([t])$ is the rank of the ordered subset $[t]$. The sample space consists of a subset $\{[e_1, e_2, \cdots, e_M]\}$ of M-dimensional Euclidean space. Let us order this subset by increasing values of e_M; for equal values of e_M, we order the vectors by increasing values of $e_{M-1}, \cdots$, for equal values of e_2, we order the vectors by increasing values of e_1. Hence, we may describe the sample space as a sequence $O_1 = [e_1(1), e_2(1), \cdots, e_M(1)]$, $O_2 = [e_1(2), e_2(2), \cdots, e_M(2)], \cdots$, where O_1 is the smallest ordered vector, O_2 is the next smallest, etc. To each sample point O_j let correspond the parameter point $P_j = [d_1(j), d_2(j), \cdots, d_M(j)]$, where $d_i(j) = e_i(j)$. Let $\Pr\{O_i ; P_j\}$ be the probability of obtaining sample point O_i when P_j is the true parameter point. Then it is easy to see that $\Pr\{O_i ; P_j\} = 0$ for $i > j$ and $\Pr\{O_i ; P_i\} > 0$. Hence, any unbiased estimate $S(O_i)$ of a function $F(P)$, defined on the parameter space P, must be such that

$$\sum_{i=1}^{j} S(O_i) \Pr\{O_i ; P_j\} = F(P_j)$$

for $j = 1, 2, 3, \cdots$. This necessary condition insures the uniqueness of $S(O_i)$, since $S(O_i)$ must satisfy the recursion relation associated with the necessary condition.

In order to calculate the variance of these statistics, we again consider the estimators in terms of δ's. We then see that the variance of $d'[t]$ is

$$\sigma^2_{d'_{[t]}} \sim d_{[t]} \prod_i g_i ,$$

where the product is taken over all values of i appearing in $[t]$, which permits the calculation of standard errors for the estimators. Similar results may be obtained for the other estimators presented.

By Theorem 3 we see that if one wishes to have unbiased estimators, then using the results of Theorems 1 and 2 is the best possible move. That is, the statistics described in those theorems are the only unbiased estimators of the parameters, and hence must be minimum variance unbiased estimators. The reader may have observed that $e_{[t]}$ is not a sufficient statistic for $d_{[t]}$. We see, therefore, that minimum variance unbiased estimators have been obtained using statistics which are not sufficient.

REFERENCE

[1] FREDERICK F. STEPHAN, "Practical problems of sampling procedure," *Am. Sociol. Rev.*, Vol., 1 (1936), pp. 569-580.