The Annals of Statistics
1983, Vol. 11, No. 3, 905–912

A NORMAL LIMIT LAW FOR A NONPARAMETRIC ESTIMATOR OF THE COVERAGE OF A RANDOM SAMPLE[1]

BY WARREN W. ESTY

Montana State University

The coverage of a multinomial random sample is the sum of the probabilities of the observed classes. A normal limit law is rigorously proved for Good's (1953) coverage estimator. The result is valid under very general conditions and all terms except the coverage itself are observable. Nevertheless the implied confidence intervals are not much wider than those developed under restrictive assumptions such as in the classical occupancy problem. The asymptotic variance is somewhat unexpected. The proof utilizes a method of Holst (1979).

1. Introduction. The coverage of a random sample of size n from a multinomial population with a perhaps countably infinite number of classes is defined to be the sum of the probabilities of the observed classes. Denote the probability that any particular observation belongs to class i by p_i, where $\sum p_i = 1$. Let X_i denote the number of observations of class i and $I_i = 1$ if $X_i \geq 1$ and $I_i = 0$ if $X_i = 0$. Then the coverage, C, is given by

$$C = \sum p_i I_i. \tag{1}$$

$1 - C$ is then equivalent to the probability that the next observation would belong to a new class. The problem is to estimate C given only $\{N_k; k = 1, 2, \cdots\}$ where N_k denotes the number of classes observed exactly k times, and $n = \sum kN_k$.

Good (1953), Good and Toulmin (1956), Harris (1959), Knott (1967), Robbins (1968), Starr (1979), Chao (1981) and Esty (1982) have addressed various aspects of this problem which has been studied in relation to species frequency models, vocabulary word models and artifact preservation models. If the classes are all equally likely, the coverage is the number of observed classes divided by the total number of classes, which gives a relationship with the classical occupancy problem.

The estimation of the number of classes in the population is a related problem (Goodman, 1949, mentions several interesting applications) which requires a parametric model (Fisher, Corbet, and Williams, 1943; McNeil, 1973; Engen, 1974; Efron and Thisted, 1976), for without some restriction on $\{p_k\}$ there could be any number of extremely unlikely classes.

Good (1953) found the estimator

$$C' = 1 - (N_1/n) \tag{2}$$

for the coverage. Note that C is a random variable and not a parameter of the population, so results about C' (Chao, 1981) are insufficient to yield confidence intervals for C. Also, C and C' are dependent (Starr, 1979) so we cannot merely treat the two variables separately and combine results. Therefore the appropriate variable to analyze is $C - C'$.

This paper rigorously proves a normal limit law for $C - C'$ under very general conditions. All the terms except C itself are observable. A corollary gives approximate confidence intervals for the coverage that are easily calculated and compare favorably with

Received October 1982; revised March 1983.
[1] This work supported in part by the MONTS-NSF project ISP-8011449.
AMS 1980 *subject classifications.* Primary 62G15; secondary 60F05 and 62E20.
Key words and phrases. Coverage, total probability, occupancy problem, cataloging problem, unobserved species, urn models.

parametric confidence intervals. The variance is smaller than a previous approximation of it. The proof utilizes a method of Holst (1979).

2. The theorem. In order to obtain a limit theorem, sequences of n's and $\{p_i\}$'s are required, so a subscript, m, is implied but usually suppressed for notational simplicity.

THEOREM 1. *Let* $\langle \{p_{im} : \sum_i p_{im} = 1\}; m = 1, 2, \cdots \rangle$ *and* n_m *be such that*

$$E(N_1/n) \to c_1, \qquad 0 < c_1 < 1 \quad \text{and} \quad E(N_2/n) \to c_2 \geq 0.$$

Then

$$n^{1/2}[C - (1 - (N_1/n))][(N_1/n) + (2N_2/n) - (N_1/n)^2]^{-1/2}$$

converges in distribution to a standard normal.

COROLLARY. *If* n *is large and* N_1/n *is not very near* 0 *or* 1, *then an approximate* $(1 - a)$ *confidence interval for* C *has endpoints*

$$1 - (N_1/n) \pm z_{a/2}([(N_1 + 2N_2)/n - (N_1/n)^2]/n)^{1/2},$$

where $z_{a/2}$ *is the usual constant for a normal confidence interval.*

COMMENTS. For the case $c_1 = 1$ see Esty (1982). The reason for the exclusion of $c_1 = 0$ and $c_1 = 1$ above is in the proof at Theorem 4.

The proof of Theorem 1 uses a method of Holst (1979) in which the characteristic function of $n^{1/2}(C - (1 - (N_1/n)))$ is shown to converge appropriately. The proof is greatly complicated by the facts that the p_i's are not equal and $E(N_1)/n$ does not appear directly in the characteristic function.

3. Proofs. Let $f_{mk}(x) = p_{mk}$ if $x = 0$, $-1/n$ if $x = 1$, and 0 if $x \geq 2$. Recall that n and p_k are functions of m, but we will suppress the m in the following. Unindexed sums will be over all k.

Define $Z_M = \sum_{k \in M} f_k(X_k)$. If M is all k, Z_M defines $Z = (1 - C) - (N_1/n) = C' - C$. We are interested in the limit of $E(\exp(isZn^{1/2}))$.

We need the well-known:

LEMMA 1. *With* $\{X_k\}$ *as in* (1), *for non-negative integers* $\{x_k\}$ *with* $\sum x_k = n$, $P(X_k = x_k; k = 1, 2, \cdots) = P(Y_k = x_k; k = 1, 2, \cdots \mid \sum Y_k = n)$ *where* $\{Y_k\}$ *are independent random variables and* Y_k *is Poisson distributed with mean* np_k.

By Lemma 1, Z_M is distributed as $\sum_M f_k(Y_k) \mid \sum Y_k = n$ and thus $E(\exp(isZ_M n^{1/2})) = E(\exp(is \sum_M f_k(Y_k) n^{1/2}) \mid \sum Y_k = n)$. We also use the following partial inversion formula for characteristic functions due to Bartlett (1938) (see also Holst, 1979).

LEMMA 2. *Let* (U, V) *be a two-dimensional random vector with* U *integer valued. Then*

$$E(\exp(ivV \mid U = n)) = (2\pi P(U = n))^{-1} \int_{-\pi}^{\pi} E(\exp(iu(U - n) + ivV))\, du.$$

Thus $E(\exp(isZ_M n^{1/2}))$ *is*

$$(2\pi P(\sum Y_k = n))^{-1} \int_{-\pi}^{\pi} E[\exp(iu \sum (Y_k - p_k n) + isZ_M n^{1/2})]\, du.$$

Let $t = un^{1/2}$ to obtain

$$(2\pi n^{1/2}P(\textstyle\sum Y_k = n))^{-1} \int_{-\pi n^{1/2}}^{\pi n^{1/2}} E(\exp(it \textstyle\sum (Y_k - np_k)n^{-1/2} + isZ_M n^{1/2}))\, dt.$$

Since $\sum Y_k$ is Poisson distributed with mean n, $(2\pi n)^{1/2}P(\sum Y_k = n) \to 1$. Define

$$H_n(s) = (2\pi)^{-1/2} \int_{-\pi n^{1/2}}^{\pi n^{1/2}} E[\exp(it \textstyle\sum (Y_k - np_k)n^{-1/2} + isZ_M n^{1/2})]\, dt.$$

We are going to evaluate the limit of $H_n(s)$. Three difficulties lie ahead: evaluating the integrand and its limit, relating the limit to $E(N_1)$ and $E(N_2)$, and proving the limit of the integral of the limit.

Unfortunately, if we choose M to be the set of all k in the definition of Z_M, dominated convergence theorems are insufficient to justify the last step, since the integrand has modulus 1. However, if we sum over a set of indicies, M, such that

$$\textstyle\sum_M p_k \to d < 1, \tag{3}$$

we can circumvent the problem at the cost of additional complexity. Note that for any d near 1 there exists such an M because $E(N_1)/n \to c_1 > 0$. Call the complementary set of indicies MC. Thus $H_n(s)$ becomes

$$(2\pi)^{-1/2} \int_{-\pi n^{1/2}}^{\pi n^{1/2}} \textstyle\prod_M E[\exp(it(Y_k - np_k)n^{-1/2} + isf_k(Y_k)n^{1/2})]$$

$$\cdot \textstyle\prod_{MC} E[\exp(it(Y_k - np_k)n^{-1/2})]\, dt.$$

Call the first product $h_{1n}(s, t)$ and the second $h_{2n}(t)$. Now

$$h_{2n}(t) = \exp(-itn \textstyle\sum_{MC} p_k n^{-1/2}) \exp(n \textstyle\sum_{MC} p_k (e^{itn^{-1/2}} - 1)),$$

since $\sum_{MC} Y_k$ is Poisson distributed with mean $n \sum_{MC} p_k$. Because of (3)

$$h_{2n}(t) = \exp(-t^2(1 - d)/2 + o(t^2) + O(t^3 n^{-1/2})).$$

Since $d < 1$, this is integrable and

$$\int_{-\pi n^{1/2}}^{\pi n^{1/2}} |h_{2n}(t)|\, dt \to \int_{-\infty}^{\infty} \exp(-(1 - d)t^2/2)\, dt.$$

Since $|h_{1n}(s, t)| \leq 1$, a generalized Lebesgue dominated convergence theorem (e.g. Rao, 1973, page 136) proves

$$\lim H_n(s) = (2\pi)^{-1/2} \int_{-\infty}^{\infty} \exp(-(1 - d)t^2/2) \lim h_{1n}(s, t)\, dt. \tag{4}$$

The calculation of this limit requires a large number of preliminary results not paralleled in Holst.

Consider the factors of $h_{1n}(s, t)$.

$$\begin{aligned} E(\exp[it(Y_k - np_k)n^{-1/2} + isf_k(Y_k)n^{1/2}]) \\ &= [\exp(-itnp_k n^{-1/2} + isp_k n^{1/2})]\exp(-np_k) \\ &\quad + [\exp(it(1 - np_k)n^{-1/2} - isn^{-1}n^{1/2})]np_k \exp(-np_k) \\ &\quad + \textstyle\sum_{j=2} \exp(it(j - np_k)n^{-1/2})P(Y_k = j) \\ &= \textstyle\sum_{j=0} \exp(it(j - np_k)n^{-1/2})P(Y_k = j) \\ &\quad + [\exp(-itp_k n^{1/2} + isp_k n^{1/2}) - \exp(-itp_k n^{1/2})]\exp(-np_k) \\ &\quad + [\exp(-itp_k n^{1/2})][\exp(itn^{-1/2})][\exp(-isn^{-1/2}) - 1]np_k \exp(-np_k). \end{aligned}$$

Note that the first term is the characteristic function of a Poisson distributed random variable with mean np_k translated to mean 0 and then divided by $n^{1/2}$. Call the terms B_k, C_k, and D_k, respectively.

$$B_k = \exp(-itp_k n^{1/2})[\exp(np_k(e^{itn^{-1/2}} - 1))].$$

$$C_k = \exp(-itp_k n^{1/2})[\exp(isp_k n^{1/2}) - 1]\exp(-np_k). \tag{5}$$

$$D_k = \exp(-itp_k n^{1/2})\exp(itn^{-1/2})[\exp(-isn^{-1/2}) - 1]np_k\exp(-np_k).$$

Denote $C_k + D_k = E_k$. We now need to evaluate the limit of the product $\prod_M (B_k + E_k)$, which is $\lim h_{1n}(s, t)$.

We now list a sequence of substantial results in the remainder of the proof of Theorem 1. Proofs appear in the next section. It is regrettable that so much work is required, but products of sums are complex unless there are stringent conditions satisfied. Theorem 2 is a useful result, but it does not solve the whole problem because the hypotheses are not satisfied for all k.

THEOREM 2. *Let* $\langle\langle B_{mk}\rangle\rangle$, $\langle\langle E_{mk}\rangle\rangle$, *and* $\langle M_m\rangle$ *satisfy (dropping the m)* i) $\prod_M B_k \sim B$, ii) $(\sum_M E_k) - E \to 0$, iii) $B_k \to 1$ *uniformly*, iv) $E_k \to 0$ *uniformly, and there exist constants* D_1 *and* D_2 *such that* v) $\sum_M |B_k - 1| \le D_1$ *and* vi) $\sum_M |E_k| \le D_2$, *then*

$$\prod_M (B_k + E_k) \sim Be^E$$

where B and E may also depend upon m.

LEMMA 3. $h_{1n}(s, t) \sim \exp(-dt^2/2)\exp(\sum_M \exp(-np_k)[(-s^2/2)(p_k + np_k^2) + stp_k])$.

LEMMA 4.

$$\operatorname{Lim} H_n(s) = \exp\{(-s^2/2)[\lim[\textstyle\sum_M p_k e^{-np_k} + \sum_M np_k^2 e^{-np_k}] - (\lim[\sum_M p_k e^{-np_k}])^2]\}.$$

THEOREM 3. *If* $E(N_{1M})/n \to c_3$, *then* $\sum_M p_k e^{-np_k} \to c_3$. *Also if* $E(N_{2M})/n \to c_4$, *then* $\sum_M np_k^2 e^{-np_k} \to c_4$.

Now, from Lemma 4 and Theorem 3 we have immediately

LEMMA 5. *Under the hypotheses of Theorem* 3, *if* $0 < c_3 < 1$

$$n^{1/2}Z_M \to_d N(0, c_3 + c_4 - c_3^2).$$

THEOREM 4. *Under the hypotheses of Theorem* 1,

$$n^{1/2}[C - (1 - (N_1/n))](E(N_1)/n + E(2N_2)/n - [E(N_1)/n]^2)^{-1/2}$$

converges in distribution to a standard normal random variable.

To show that Theorem 1 follows from Theorem 4 we need only show N_1/n and N_2/n converge in probability to c_1 and c_2, respectively, so we may replace the expected values in Theorem 4 with their corresponding observed values. These steps will complete the proof of Theorem 1. The corollary is immediate.

$\operatorname{Var}(N_1/n) \to 0$ implies N_1/n converges to c_1 in probability. Let $Z_i = 1$ if $X_i = 1$ and $Z_i = 0$ otherwise. Then $N_1 = \sum_i Z_i$ and $\operatorname{Var}(N_1/n) \le n^{-2}(\sum_i E(Z_i^2) + \sum_{i \ne j} \operatorname{Cov}(Z_i, Z_j))$. The first term is $E(N_1)/n^2 \to 0$. The second is bounded above by

$$\sum_{i \ne j} p_i p_j (1 - p_i)^{n-2}(1 - p_j)^{n-2}(p_i + p_j)$$

$$\le 4(n-1)^{-1} \sum_i (n-1)p_i^2(1 - p_i)^{n-2}/2 \sum_j p_j(1 - p_j)^{n-2} \le 4(n-1)^{-1} \to 0.$$

A similar proof holds for $\operatorname{Var}(N_2/n) \to 0$.

4. More proofs. The proof of Theorem 2 is straightforward and omitted. For Lemma 3, recall the definitions of B_k, C_k, D_k, and E_k of (5). Lemma 3 evaluates the limit of the product $\prod_M (B_k + E_k)$. Unfortunately, the hypotheses of Theorem 2 are not satisfied as it stands. For instance, neither iii) nor iv) need hold. Therefore we break the product into two factors, each of which is tractable. Let

$$I = \{k : p_k n^{1/2} \le n^{-3/8}\} \quad \text{and} \quad II = \{k : np_k \ge n^{1/8}\} \backslash I. \tag{6}$$

If $k \notin I$ then $p_k > n^{-7/8}$ and $np_k > n^{1/8}$ so $k \in II$. Therefore $I \cup II = \{k\}$. Then

$$\begin{aligned}\prod_M (B_k + E_k) &= \prod_I (B_k + E_k) \prod_{II} (B_k + E_k) \sim \prod_I (B_k + E_k) \prod_{II} B_k \quad \text{by Lemma 6} \\ &\sim \prod_I B_k (\exp \textstyle\sum_I E_k) \prod_{II} B_k \quad \text{by Lemma 7} \\ &\sim \prod_M B_k (\exp \textstyle\sum_M E_k) \quad \text{by Lemma 6.}\end{aligned}$$

The completed proof of Lemma 3 requires the evaluation of these terms. The limits appear as we work our way through the necessary preliminaries.

LEMMA 6. $\prod_{II} (B_k + E_k)/\prod_{II} B_k \to 1$ and $\sum_{II} |E_k| \to 0$.

PROOF. $\sum_{II} |E_k| \le 2 \sum_{II} (e^{-np_k} + np_k e^{-np_k}) \le 2n^{7/8}(1 + n^{1/8})\exp(-n^{1/8}) \to 0$, since the number of indices in II is less than or equal to $n^{7/8}$. Now, for all k,

$$\begin{aligned}B_k &= \exp[-itp_k n^{1/2} + np_k(itn^{-1/2} - (t^2/2n) + O(t^3 n^{-3/2}))] \\ &= \exp[(-t^2/2)p_k + O(t^3 p_k n^{-1/2})]\end{aligned}$$

and B_k is bounded away from 0 so

$$|\ln(B_k + E_k) - \ln B_k| \le |E_k|/(|B_k| - |E_k|),$$

and Lemma 6 follows.

LEMMA 7. *The hypotheses of Theorem* 2 *are satisfied with* $M = I$ *and* B_k *and* E_k *as in* (5).

PROOF. iii), iv), and v) are easily checked. For vi), $E_k = e^{-np_k}\exp(-itp_k n^{1/2})$ $[\exp(isp_k n^{1/2}) - 1 + np_k e^{itn^{-1/2}}(\exp(-isn^{-1/2}) - 1)]$. Since, for $k \in I$, $p_k n^{1/2} \to 0$ uniformly, we have

$$\begin{aligned}E_k &= e^{-np_k}\exp(-itp_k n^{1/2})(isp_k n^{1/2} - (s^2 p_k^2 n/2) + O(s^3 p_k^3 n^{3/2}) \\ &\quad + np_k[1 + itn^{-1/2} - (t^2/2n) + O(t^3 n^{-3/2})][-isn^{-1/2} - (s^2/2n) + O(s^3 n^{-3/2})]) \\ &= e^{-np_k}\exp(-itp_k n^{1/2})[-(s^2/2)(np_k^2 + p_k) + stp_k + O(p_k^3 n^{3/2}) + O(p_k n^{-1/2})].\end{aligned}$$

The second factor tends to one uniformly and will not affect the limit

$$\textstyle\sum_I E_k \sim \sum_I e^{-np_k}(-(s^2/2)(np_k^2 + np_k) + stp_k),$$

using $\sum_I O(p_k^3 n^{3/2}) = \sum_I O(p_k n^{-1/4}) \to 0$, since $p_k^2 \le n^{-7/4}$ on I. This yields the necessary limit. Applying absolute value signs yields Lemma 7.

To finish the proof of Lemma 3, we need only note that, from the definition of B_k in (5), $\lim \prod_M B_k = \exp(-dt^2/2)$.

PROOF OF LEMMA 4. Using the result of Lemma 3 in (4), the integrand is a multiple of a normal probability density and easily integrated. The result is non-degenerate if

$$\lim(\textstyle\sum_M p_k e^{-np_k} + \sum_M np_k^2 e^{-np_k}) - (\lim \sum_M p_k e^{-np_k})^2 \ne 0.$$

PROOF OF THEOREM 3. A result like Theorem 3 might be expected from the Poisson

approximation to the binomial, but this result follows even if the p_k's are not small. Again we need to distinguish the two types of indicies in (6).

$$E(N_1)/n = \sum p_k(1-p_k)^{n-1}. \qquad \sum_{II} p_k(1-p_k)^{n-1} \le \sum_{II} p_k \exp(-p_k(n-1)) \to 0.$$

Thus $\sum_I p_k(1-p_k)^{n-1} \to c_3$. Now

$$\sum_I p_k(1-p_k)^{n-1} \le \sum_I p_k \exp(-p_k(n-1)) \le \exp(\sup_I p_k) \sum_I p_k e^{-np_k}. \tag{7}$$

Also $\sum_I p_k(1-p_k)^{n-1} \ge \sum_I p_k \exp(-p_k(n-1)/(1-p_k))$ since $1 - t \ge \exp(-t/(1-t))$ for $0 \le t < 1$ (Feller, 1968, (12.26)). Note if $p_k = 1$ the result holds anyway.

$$\sum_I p_k(1-p_k)^{n-1} \ge \sum_I p_k \exp(-np_k/(1-\sup_I p_k)) \ge \sum_I p_k \exp((-np_k)(1+2\sup_I p_k))$$

for all large n since $\sup_I p_k \to 0$

$$\ge \exp(-2n(\sup_I p_k)^2) \sum_I p_k \exp(-np_k) \tag{8}$$

and the first factor approaches one by the definition of I. Combining (7) and (8) yields $\sum_I p_k e^{-np_k} \to c_3$. Noting $\sum_{II} p_k e^{-np_k} \to 0$ yields $\sum p_k e^{-np_k} \to c_3$. The second part is similar. To prove Theorem 4, note $Z = Z_M + Z_{MC}$. From the above the limit distributions of Z_M and Z_{MC} follow. Using Lemma 5 of LeCam (1958), the result follows.

In Lemma 5 we see why c_1 cannot be 0 or 1. The variance would be 0 and the scale factor, $n^{1/2}$, incorrect.

5. Applications. The following example is from one of many areas in which the coverage of a sample is of great interest, namely the historical analysis of ancient coin hoards (see the American Numismatic Society bibliography, 1974). Coins are classified by die variety and the completeness of a hoard as measured by its coverage yields information about coinage in antiquity. Since coin hoards cannot be increased in size to improve the accuracy of point estimates, confidence intervals are necessary supplements to the usual point estimates.

The following typical data are from Holst (1981) and are from a hoard of Indo-Greek coins, $N_1 = 156$, $N_2 = 19$, $N_3 = 2$, $N_4 = 1$, and $N_5 = 0$ and $n = 204$. From the corollary to Theorem 1, an approximate 95 per cent confidence interval for C has endpoints $.235 \pm .083$. We return to this example in Section 7.

Harris (1959, page 548), without attempting to obtain a limiting distribution, approximated $E((C - C')^2)$ by

$$E(N_1 + 2N_2)/n^2, \tag{9}$$

which omits the third term in the correct result. In this example the omission would make the confidence interval 61 per cent wider, a serious loss.

6. Comments. C' is not unbiased. A short calculation shows $E(C' - C)$ is approximately $-2E(N_2)/(n(n-1))$. Users may wish to incorporate this factor into their calculations.

It is interesting to note that as $E(N_1)/n \to 1$ the variance tends to that of the "low coverage" result of Esty (1982), where $2N_2/(n-1)$ estimates C and c_1 would be 1. In that context $\mathrm{Var}(n(2N_2/(n-1) - C)) \sim \mathrm{Var}(2N_2) \sim 4E(N_2)$ since N_2 is asymptotically Poisson distributed. Also $E(N_1) \sim n$ and $n - (E(N_1 + 2N_2)) \to 0$. Thus

$$\mathrm{Var}(n(C' - C)) \sim E(N_1 + 2N_2) - E^2(N_1)/n \quad \text{from Theorem 4}$$

$$\sim n - [(n - 2E(N_2))^2/n] \to 4E(N_2).$$

If the third term is omitted as in (9), for small c_1 the variance approximation is far too large (see Esty, 1982, example 1).

7. Comparison with the occupancy problem. If data is actually from a parametric family, we expect the appropriate parametric theory to outperform nonparametric theory. Nevertheless, the intervals from the corollary compare very favorably to the intervals for the case of equally likely classes, the classical occupancy problem.

If all k classes are equally likely, the distribution of the number of observed classes is asymptotically $N(k(1 - e^{-m}), ke^{-2m}(e^m - 1 - m))$ where $m = n/k$ (Weiss, 1958). To find an approximate 95 per cent confidence interval for k, the number of classes, solve for k in

$$N_d = k(1 - e^{-m}) \pm 1.96(ke^{-2m}(e^m - 1 - m))^{1/2}, \tag{10}$$

where N_d denotes the number of distinct classes observed.

To compare intervals, suppose we create the ideal data based on $n = 100$ and $k = 100$. With observed values taken to be their expectations, we would have $N_1 = 37$ (36.8), $N_2 = 18$ (18.4), $N_3 = 6$ (6.13), $N_4 = 1$ (1.5) and $n_5 = 1$ (.5). Thus $N_d = 63$ and $N_0 = 37$. Then $C' = 1 - (37/100) = .63$ and, from (10), $\tilde{k} = 99.2$, which is equivalent to $\tilde{C} = .635$. Note $C = .63$. From the corollary we calculate the approximate 95 per cent confidence interval to be (.479, .781), which is equivalent to $k \in (80.7, 131.5)$. On the other hand, using (10), we find, with a great deal more work, the approximate 95 percent confidence interval (80.8, 128.3), which is only slightly narrower.

Returning to the coin hoard data, if the classes were equally likely, our result would correspond to an estimate of 757 and the interval (559, 1171). Under the restrictive equally likely assumption, using (10), the estimate is 731 and the interval is (526, 1051). The results are comparable. Since numismatists question the validity of the equally likely hypothesis, the concept of coverage is more appropriate for discussing the completeness of a sample than the number of dies, and the use of the result in this paper is to be preferred in numismatic examples, as well as any others in which the equally likely hypothesis is false or in doubt.

REFERENCES

AMERICAN NUMISMATIC SOCIETY (1974). Estimating die and coinage output: Bibliography. Mimeographed report.

BARTLETT, M. S. (1938). The characteristic function of a conditional statistic. *J. London Math. Soc.* **13** 62–67.

CHAO, A. (1981). On estimating the probability of discovering a new species. *Ann. Statist.* **9** 1339–1342.

EFRON, B. and THISTED R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika* **63** 435–447.

ENGEN, S. (1974). On species frequency models. *Biometrika* **61** 263–270.

ESTY, W. W. (1982). Confidence intervals for the coverage of low coverage samples. *Ann. Statist.* **10** 190–196.

FELLER, W. A. (1968). *An Introduction to Probability Theory and its Applications*, Vol. 1, 3rd ed. Wiley, New York.

FISHER, R. A., CORBET, A. S. and WILLIAMS, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* **12** 42–58.

GOOD, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40** 237–264.

GOOD, I. J. and TOULMIN, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika* **43** 45–63.

GOODMAN, L. A. (1949). On the estimation of the number of classes in a population. *Ann. Math. Statist.* **20** 572–579.

HARRIS, B. (1959). Determining bounds on integrals with applications to cataloging problems. *Ann. Math. Statist.* **30** 521–548.

HOLST, L. (1979). A unified approach to limit theorems for urn models. *J. Appl. Probab.* **16** 154–162.

HOLST, L. (1981). Some asymptotic results for incomplete multinomial or Poisson samples. *Scand. J. Statist.* **8** 243–246.

KNOTT, M. (1967). Models for cataloguing problems. *Ann. Math. Statist.* **38** 1255–1260.

LECAM, L. (1958). Un Théorème sur la division d'un intervalle par des points pris au hasard. *Publ. Inst. Statist. Univ. Paris* **7** 7–16.

McNeil, D. R. (1973). Estimating an author's vocabulary. *J. Amer. Statist. Asso.* **68** 92–96.
Rao, C. R. (1973). *Linear Statistical Inference and its Applications*, 2nd ed. Wiley, New York.
Robbins, H. E. (1968). Estimating the total probability of the unobserved outcomes of an experiment. *Ann. Math. Statist.* **39** 256–257.
Starr, N. (1979). Linear estimation of the probability of discovering a new species. *Ann. Statist.* **7** 644–652.
Weiss, I. (1959). Limiting distributions in some occupancy problems. *Ann. Math. Statist.* **29** 878–884.

Department of Mathematical Sciences
Montana State University
Bozeman, Montana 59717