The Annals of Statistics
1982, Vol. 10, No. 1. 190–196

CONFIDENCE INTERVALS FOR THE COVERAGE OF LOW COVERAGE SAMPLES

By Warren W. Esty

Montana State University

The coverage of a random sample from a multinomial population is defined to be the sum of the probabilities of the observed classes. The problem is to estimate the coverage of a random sample given only the number of classes observed exactly once, twice, etc. This problem is related to the problem of estimating the number of classes in the population. Non-parametric confidence intervals are given when the coverage is low such that a Poisson approximation holds. These intervals are related to a coverage estimator of Good (1953).

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

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

This concept makes sense even if the number of classes is countably infinite. Also, if each class is equally likely, i.e. $p_i = 1/s$ for each $i = 1, 2, \cdots, s$, then $C = N_d/s$, where $N_d = \sum_{k=1} N_k$ denotes the number of distinct classes observed, and statements about C can be converted into statements about s. Thus there is a relationship with the "unobserved species" and "author's vocabulary" problems which develop estimators for s.

The problem is to estimate C given $\{N_k;\ k = 1, 2, \cdots\}$ where N_k denotes the number of classes observed exactly k times and $N = \sum kN_k$. Good (1953) found the estimator

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

for the coverage. The following results will not improve upon this point estimator, but rather will address the question of a limiting distribution and confidence intervals for C.

Robbins (1968) proved an exact relationship for $E(C)$ similar to (2). Harris (1959) obtained an approximation,

$$E\{(C - C'')^2\} \doteq E(N_1 + 2N_2)/N^2,$$

under different conditions. Under (3) it is not an improvement on Robbin's "universal inequality" for the variance of $C - C''$. With either result, if the observed number of duplicates is small, the lower confidence limit may be 0, which is trivial, and the lack of a limiting distribution leaves the calculation of the confidence interval in doubt. The following result handles both problems. An example comparing these confidence intervals follows the theorems.

Let $D = \sum_2 N_k$ denote the number of classes observed at least twice. If all the p_i's are sufficiently small, then D is approximately Poisson in distribution. This fact can be used to create confidence intervals for the coverage when few duplicates are observed. It will be shown that, under certain conditions, $(N-1)C/2 \to_P E(D) > 0$. Then the usual confidence

Received December, 1979; revised May, 1981.
AMS 1970 *Subject Classification.* Primary 62G15.
Key words and phrases. Coverage, occupancy problem, unobserved species, total probability.

interval for $E(D)$ based on the observation of D yields a confidence interval for C. Furthermore, the result improves upon previously obtained confidence intervals if the conditions hold.

The above formulation assumes sampling with replacement. If each class, i, $i \le s < \infty$, when s is the number of classes in the population, is represented in the population by a finite number of elements, M_i, and a random sample is drawn without replacement by selection of each of the $\sum_{i=1}^{s} M_i$ members with probability p, related results are obtained.

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

THEOREM 1. *Let* $\{p_{in}, \sum_i p_{in} = 1; i = 1, 2, \cdots\}$ *and* $\{N_n\}$ *be such that*

$$N \max p_i \to 0 \qquad \text{and} \qquad N(N-1) \textstyle\sum_i p_i^2/2 \to m > 0. \tag{3}$$

Then

(i) *D and N_2 are asymptotically Poisson distributed with mean m, and $\sum_{k=3} kN_k$ converges in probability to* 0,

and

(ii) $(N-1)C/2 \to_p m$.

COROLLARY 1. *If Np_i is small for all i, then D is approximately Poisson distributed with mean $m' = N(N-1) \sum p_i^2/2$ and $P(D = 0 \mid N)$ is approximately $e^{-m'}$.*

COROLLARY 2. *Let d and n_1 denote the observed values of D and N_1. If n_1 is nearly N and much larger than d, an estimator for the coverage, C, is given by*

$$C' = 2d/(N-1). \tag{4}$$

Furthermore if (a, b) is a $(1-\alpha)$ confidence interval for the mean of a Poisson random variable based on a single observation d, then an approximate confidence interval for C of size $(1-\alpha)$ is given by

$$2a/(N-1) \le C \le 2b/(N-1). \tag{5}$$

COROLLARY 3. *If all s classes are equally likely ($p_i = 1/s$ for all i), and if n_1 is nearly N and much greater than d, an estimator, s', for s is given by*

$$s' = n_d(N-1)/2d, \tag{6}$$

where n_d denotes the number of distinct classes observed. Furthermore, an approximate confidence interval for s of size $(1-\alpha)$ is given by

$$n_d(N-1)/2b \le s \le n_d(N-1)/2a \tag{7}$$

where a and b are as in (5).

COMMENTS. Although N and $N-1$ are asymptotic, the proof (see equation (14)) suggests that $N-1$ is the appropriate factor. In relatively small sample problems, such as Example 3, it performs better.

By (i) the results hold with N_2 replacing D. As an approximation, however, D is preferable since the asymptotically negligible term of (14) in C corresponds approximately to the term of (13) in D that is not in N_2.

The estimator of (4) is in essential agreement with Good's result (2), since, using Theorem 1 (i),

$$1 - (N_1/N) = \textstyle\sum_{k=2} kN_k/N \sim 2 \sum_{k=2} N_k/(N-1) = 2D/(N-1).$$

Unfortunately, the linear combination of N_k's in a calculation paralleling (13) which best approximates (14) is $\sum_{k=2} kN_k/2$, which is not necessarily integer valued and does not satisfy the limit law that integer valued linear combinations with coefficient 1 on N_2 satisfy. Good's result would suggest using $\sum_{k=2} (k-1)N_k$. It and $\sum_{k=2} N_k = D$ have expectations differing from (14) by the same amount, the former overestimating it and the latter underestimating it. I have opted to use the latter because of the type of application in Example 1 where the randomness assumption is sometimes violated by groups of observations in the same class. In that case D gives less weight to the extra duplicates.

In the equally likely case of Corollary 3, a computation shows that the maximum likelihood estimator of s (Good, 1950, page 73) is asymptotic to s' of (6). Also, $C = n_d/s$ so that an estimate or limit on C corresponds to an estimate or limit on s.

The second theorem pertains to a different formulation of the problem: suppose each class, i, is represented in a population by a finite number of elements, M_i, and a random sample is drawn without replacement from the population of $\sum_{i=1}^{s} M_i$ members by selecting each element with probability p. Then the sample size, N, is itself random. By making p small and s large, results paralleling those of Theorem 1 may be obtained. If the M_i's are not all very large, the effect of sampling without replacement alters the result somewhat. In this context let

$$C = \sum_{i=1}^{s} M_i Y_i / \sum_{i=1}^{s} M_i,$$

where Y_i is as in (1).

THEOREM 2. *Let $\{M_i\}$ be a fixed sequence of positive integers (not necessarily large) and let $s \to \infty$ and $p \to 0$ such that*

$$p \sum_{i=1}^{s} M_i \to \infty \tag{8}$$

$$p \max_{i \le s} M_i \to 0, \tag{9}$$

and

$$p^2 \sum_{i=1}^{s} M_i(M_i - 1)/2 \to m > 0. \tag{10}$$

Then

(i) *D and N_2 are asymptotically Poisson distributed with mean m, and $\sum_{k=3} kN_k$ converges to 0 in probability,*

and

(ii) $N(C-p)/2 \to_p m$.

Furthermore, if $M_i = M$ for all i,

(iii) $\{M/(M-1)\}NC/2 \to_p m$.

REMARKS. The results of Theorem 2 can be used to give point estimates and confidence intervals for C. Also, (iii) gives us a feeling for how large the M_i's should be to be able to disregard their effect if they are not precisely known. One obvious corollary is that if M_i is large for all i and the observed value of N is large and n_1 is much larger than n_2, then the results of the previous corollaries hold. The accuracy of the approximations is, however, diminished by using N for $E(N)$ and disregarding $\{M_i\}$. Note that in (ii) and (iii) and the associated corollaries, N is the proper factor and not $N-1$ as in Theorem 1.

This sampling approach has a further generalization. Suppose that the M_i's are themselves i.i.d. positive integer-valued random variables, but that otherwise the context of Theorem 2 is maintained. The conclusions would hold if the hypotheses held with probability one.

THEOREM 3. *The hypotheses to Theorem 2 hold with probability one if $p \to 0$, $p^2 s \to m' > 0$, $E(M^3)$ exists and M is not trivially always one.*

3. Examples.

EXAMPLE 1. Eddy (1967), in a hoard of ancient coins, found among 662 coins of the emperor Gordian III (244–249 AD) only two pairs struck from the same dies. The die varieties observed differ so minutely that numismatists are satisfied that they do not form a collection of differing varieties. Assuming, then, that the sample was random, this implies that a huge number of dies were employed in producing the coins. Numismatists would like a confidence interval for C. Theorem 3 is appropriate since the dies produced independent, identically distributed random numbers of coins. It is known that each M_i is on the order of 10,000 (Sellwood, 1963). Therefore we cite (iii) to justify using $NC/2 \doteq E(D)$. The corollary parallel to (4), with N in place of $N - 1$, as in all corollaries to Theorems 2 and 3, yields $C = 4/662 = .00604$ which is the same as (2). A bound on the variance from Robbins's "universal inequality" would be $1/(N + 1)$ and no non-trivial lower confidence limit for C would be possible since $(1/663)^{1/2} = .039$ is much too large. Harris's approximation (page 548) $E\{(C - C'')^2\} \doteq E(N_1 + 2N_2)/N^2$ is not smaller and inappropriate under (3). But (5) gives non-trivial justified intervals for any desired confidence. For instance, a 95 per cent confidence interval of the form $C_0 \leq C$ is given by $.701/662 = .00106 = C_0 \leq C$. Numismatists examining this data find the coverage surprisingly low and are therefore interested in a confidence interval of the form $C \leq C_1$. Such a 95 per cent confidence interval is $C \leq C_1 = 12.6/662 = .0190$. This result also differs substantially from any obtained from the naive incorrect application of a normal limit law. If it is assumed that $p_i = 1/s$ for all i, confidence intervals for s have been obtained under assumptions other than (3). Usually a sample contains many duplicate observations and a normal limit law can be obtained as, say $N/s \rightarrow k > 0$ ((3) implies $k = 0$). Results based on a normal limit law (for example, Darroch, 1958; or see Seber, 1973, Section 4.1.2) require a normal limit law for D which is not reasonable unless D is large. Of course, if D is large and yet the present hypotheses hold, the Poisson distribution is well approximated by a normal distribution and the two results coincide.

EXAMPLE 2. The calculation of $P(D = 0 \mid N)$ when not all p_i's are the same is the "generalized birthday problem." Gail et al (1979) noted an application to cancer research of the case when s is large and N is moderate. They obtained Corollary 1 and calculated $P(D = 0 \mid N = 40)$ when $p_i = 1/10{,}000$ for each i. The approximation, .924964, differs from the true value, .924869, by one digit in the fourth decimal place, accurate enough for most purposes.

EXAMPLE 3. Suppose a sample of size 23 from s (unknown) equally likely classes yields no duplicates. Even in this extreme case (7) can be used to obtain a confidence interval for s (although not with an upper limit) and also a point estimator. The estimate from (6) would be $s = \infty$, which is not likely to be useful. This reflects the fact that the probability of no duplications is increasing in s. In that case a 50 per cent confidence point may be of some use. Choosing b such that $(0, b)$ is a 50 per cent confidence interval yields $2b = 1.386$ and the point estimate $s_{.50} = 365.08$. This is the reverse of the birthday problem. An approximate 95 per cent confidence interval of the form $s \geq s_0$ is given by $s \geq 84.5$. The true 96 per cent confidence interval is $s \geq 92$. This and other examples show that the lower confidence bounds for s and upper bounds for C tend to be conservative.

4. Proofs of Theorems.

PROOF OF THEOREM 1. Let X_i denote the number of observations of class i in the sample of size N. For notational simplicity the subscripts, n, on X_i, N, and p_i are suppressed. Under (3)

$$\sum_{j=k} jP(X_i = j) \sim kP(X_i = k) \tag{11}$$

uniformly in i, since

$$(k+1)P(X_i = k+1) = N(N-1)\cdots(N-k)p_i^{k+1}(1-p_i)^{N-(k+1)}/k!$$
$$= \{(N-k)p_i/(1-p_i)\}P(X_i = k).$$

By (3) the first factor can be made less than ε for all k uniformly in i for n sufficiently large. Summing from $k+1$ to infinity yields (11).

Now

$$E(\textstyle\sum_{k=3} kN_k) = \sum_i \sum_{k=3} kP(X_i = k) \sim 3 \sum_i P(X_i = 3) \tag{12}$$
$$\leq (N-2)\max p_i \textstyle\sum_i N(N-1)p_i^2(1-p_i)^{N-3}/2 \to 0$$

by (3). Since $\sum_{k=3} kN_k$ is nonnegative, by (12) it converges in probability to 0, proving part of (i).

For the other half of (i), let $D_i = 1$ if $X_i \geq 2$ and $D_i = 0$ if $X_i \leq 1$. Then $D = \sum D_i$, and

$$E(D) = \textstyle\sum E(D_i) = \sum P(X_i \geq 2) = \sum P(X_i = 2) + \sum P(X_i \geq 3) \tag{13}$$
$$\sim N(N-1)p_i^2(1-p_i)^{N-2}/2 \sim N(N-1)p_i^2/2 \to m,$$

where we have used (11) and (3) in the last three steps.

Note that the variables D_i are dependent. That N_2 and D converge in distribution to the Poisson distribution with mean m will follow (Sevastyanov, 1972) if we show the r-dimensional joint probabilities are uniformly asymptotic to the corresponding product of the marginals.

$$P(X_{i_1} = X_{i_2} = \cdots = X_{i_r} = 2)/\{P(X_{i_1} = 2)\cdots P(X_{i_r} = 2)\}$$
$$= \frac{N!(1-p_{i_1}-p_{i_2}-\cdots-p_{i_r})^{N-2r}}{(N-2r)!\{N(N-1)\}^r(1-p_{i_1})^{N-2}\cdots(1-p_{i_r})^{N-2}} \to 1$$

uniformly, since $N\max p_i \to 0$ by hypothesis. We did not need to separate out Sevastyanov's "rare" sets. This proves the result for N_2. The result for D follows from (12).

Recall the definition of C given by (1).

$$E\{(N-1)C/2\} = \textstyle\sum (N-1)p_iE(Y_i)/2$$
$$= \textstyle\sum N(N-1)p_i^2(1-p_i)^{N-1}/2 + \sum (N-1)p_iP(X_i \geq 2)/2 \tag{14}$$
$$\sim N(N-1)p_i^2/2 \to m.$$

Also

$$\mathrm{Var}\{(N-1)C/2\} \leq \textstyle\sum \mathrm{Var}\{(N-1)p_iY_i/2\}$$

since the covariances are negative, and the right hand side is bounded above by

$$\textstyle\sum N(N-1)^2p_i^3/4 \leq (\max Np_i) \sum N(N-1)p_i^2/4 \to 0.$$

Thus $(N-1)C/2 \to_P m$, and Theorem 1 is proven.

The condition that n_1 is nearly N in Corollary 2 is required since $E(N_1) = \Sigma P(X_i = 1) \sim \Sigma Np_i = N \to \infty$. Since $E(D)$ is finite, d is required to be much less than N. Since the sample mean is an estimator for the mean of a Poisson distribution, and since D is approximately Poisson distributed, combining (i) and (ii) we obtain (4) and (5). If, in addition, all s classes are equally likely, $C = n_d/s$. Solving for s in (4) and (5) yields Corollary 3.

The role of the "low coverage" assumption in this paper is to compel a function of C, (ii), to converge in probability to a constant. It is important to recall that C is not a parameter of any distribution, but rather a random variable. Thus the estimation of C by a random variable has sources of error in each random variable. The restrictive assumptions reduce that to one source of error, which is handled by (i). Presumably the coverage of a moderate coverage sample could be made to conform to a normal limit law, but it is not clear how to do so. Normal limit laws were discussed further in Example 1.

PROOF OF THEOREM 2. The proof of Theorem 2 is similar to that of Theorem 1, but the differences are worth noting. In the context of Theorem 2,

$$E(D) = E(\sum D_i) = \sum P(X_i \geq 2) \sim \sum M_i(M_i - 1)p^2(1-p)^{M_i-2}/2$$
$$\sim \sum M_i(M_i - 1)p^2/2 \to m$$

by (9) and (10). Since the D_i's are independent and $P(D_i = 1) \to 0$ uniformly in i, D is asymptotically Poisson with mean m. This proves (i).

Since N is binomially distributed and $E(N) = p \Sigma M_i \to \infty$,

(15) $$\{N - E(N)\}/E(N) \to_P 0.$$

Combining

(16) $$E(C-p)E(N)/2 \to m$$

and

(17) $$\mathrm{Var}\{(C-p)E(N)/2\} \to 0$$

yields

(18) $$(C-p)E(N)/2 \to_P m.$$

This and (15) yield

(19) $$(C-p)N/2 \to_P m.$$

This is (ii). To prove (16),

$$(C-p)E(N) = \{(\sum M_i Y_i/\sum M_i) - p\} p \sum M_i = p \sum M_i Y_i - p^2 \sum M_i$$
$$= p \sum (M_i - 1)Y_i + p \sum Y_i - p^2 \sum M_i,$$

and

$$E(p \sum Y_i - p^2 \sum M_i) \sim p \sum M_i p(1-p)^{M_i-1} - p^2 \sum M_i \to 0$$

and

$$E\{p \sum (M_i - 1)Y_i\} \sim \sum (M_i - 1)M_i p^2(1-p)^{M_i-1} \to m.$$

To prove (17),

$$\mathrm{Var}\{(C-p)E(N)\} = E^2(N)\mathrm{Var}\, C \leq p^2(\sum M_i)^2 \sum M_i^2(pM_i)/(\sum M_i)^2$$
$$\leq p \max M_i \sum M_i^2 p^2 \to 0$$

By (9) and (10). Then (19) follows from (18) and (15) by adding and subtracting $E(N)$.

Suppose, for (iii), that $M_i = M$ for all i. Then

$$E(N)E(C) \sim \sum M^2p^2 = \{M/(M-1)\} \sum M(M-1)p^2 \to 2Mm/(M-1).$$

PROOF OF THEOREM 3. Note that the existence of $E(M^3)$ implies the existence of $E(M^2)$ and $E(M)$. The first two conditions imply $ps \to \infty$, so (8) holds with probability 1. Because $p^2 s \to m' > 0$ and because $E\{M(M-1)\}$ exists and is not zero, (10) is implied. To obtain (9), note first that

$$1 \geq P(p \max_{1\leq s} M_i < \varepsilon) = F^s(\varepsilon/p) = [1 - \{1 - F(\varepsilon/p)\}]^s \geq 1 - s(1 - F(\varepsilon/p)).$$

Since $p^2 s \to m'$, $\varepsilon/p \sim \varepsilon' s^{1/2}$ and since $s \to \infty$ we need only that $x^2\{1 - F(\varepsilon' x)\} \to 0$. Recall that if $E(M^3)$ exists then $\int x^2\{1 - F(x)\}\, dx$ exists, implying $x^2\{1 - F(x)\} \to 0$, which suffices for (9). Theorem 3 is proven.

REFERENCES

DARROCH, J. N. (1958). The multiple-recapture census. I: Estimation of a closed population. *Biometrika* **45** 343–359.

EDDY, SAMUEL K. (1967). *The Minting of Antoniani A. D.* 238–249 *and the Smyrna Hoard.* Numismatic Notes and Monographs no. 156, The American Numismatic Society, New York.

GAIL, M. H., WEISS, G. H., MANTEL, N., and O'BRIEN, S. J. (1979). A solution to the generalized birthday problem with application to allozyme screening for cell culture contamination. *J. Appl. Probab.* **16** 242–251.

GOOD, I. J. (1950). *Probability and the Weighing of Evidence.* Griffin, London.

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

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

ROBBINS, HERBERT E. (1968). Estimating the total probability of the unobserved outcomes of an experiment. *Ann. Math. Statist.* **39** 256–257.

SEBER, G. A. F. (1973). *The Estimation of Animal Abundance.* Hafñer Press, New York.

SEVASTYANOV, B. A. (1972). Poisson limit law for a scheme of sums of dependent random variables. *Theor. Probab. Appl.* **17** 695–699.

SELLWOOD, D. (1963). Some experiments in Greek minting technique. *Numismatic Chronicle* **3** 217–231.

DEPARTMENT OF MATHEMATICS
COLLEGE OF LETTERS AND SCIENCE
MONTANA STATE UNIVERSITY
BOZEMAN, MT 59717