The Annals of Statistics
1986, Vol. 14, No. 3, 1257–1260

THE EFFICIENCY OF GOOD'S NONPARAMETRIC COVERAGE ESTIMATOR

BY WARREN W. ESTY

Montana State University

The asymptotic efficiency of Good's nonparametric coverage estimator is obtained relative to the best estimator derived under the assumption that all classes are equally likely. Even when that assumption is true, Good's estimator is quite efficient, with an asymptotic relative efficiency of greater than 85% in all cases, and greater than 95% if the expected coverage is less than one-half.

1. Introduction. The coverage, C, of a random sample of size n from a multinomial population is defined to be the sum of the probabilities of the observed classes. Estimating the coverage of a sample from an unknown multinomial distribution is an occupancy problem with applications in ecology (Good and Toulmin, 1956; Engen, 1974; Engen, 1978), vocabulary studies (Efron and Thisted, 1976; McNeil, 1973) and archaeology (American Numismatic Society, 1974; Esty, 1982, 1983). Good (1953) introduced an estimator for C,

$$\tilde{C} = 1 - N_1/n, \tag{1.1}$$

where N_1 denotes the number of classes observed exactly once, which has received much attention (Robbins, 1968; Engen, 1978; Starr, 1979; Chao, 1981; and many others). Esty (1983) found the associated confidence intervals under very general conditions. Numismatists regularly use estimators derived under the hypothesis that all classes are equally likely, which is the hypothesis of the classical occupancy problem (Feller, 1968; Johnson and Kotz, 1977, Section 6.2.1). Carter (1981) compared several of these (Lyon, 1965; Brown 1955/57; Carcassone, 1980; and Mora-Màs, 1981; Schroeck, 1981, has given another one since then) by evaluating them using real data where the classes are varieties of ancient coins, but no theoretical comparison of Good's estimator to other commonly employed estimators has been given. Users of these intervals will ask if the generality of the nonparametric approach is accompanied by a substantial loss in efficiency relative to the methods already in use. The answer is "no"; Good's estimator is remarkably efficient. When all classes are equally likely, the asymptotic efficiency of $\tilde{C}$ relative to the best estimator based on the hypothesis that all classes are equally likely exceeds 85% in all cases and exceeds 90% if the expected coverage is below 76%. An explicit formula for the asymptotic relative efficiency is given by Theorem 3.

2. Discussion and results. Note that the coverage of a sample is not a parameter of the population. Therefore an "estimator" of the coverage is not an

Received January 1985; revised November 1985.
AMS 1980 *subject classifications.* Primary 62G20; secondary 62F12.
Key words and phrases. Coverage, total probability, occupancy problem, unobserved species.

estimator in the usual sense and the "efficiency" of an estimator of C cannot be defined in the usual sense, but it is possible to define the asymptotic relative efficiency of two estimators by comparing the variances of their associated normal limit theorems in the usual manner. Good's nonparametric estimator will be compared to the most restrictive parametric estimator, namely, the estimator based on the equally likely hypothesis, when the equally likely hypothesis is true.

First we need the normal limit law for the estimator based on the equally likely hypothesis. Suppose n balls are distributed at random into k boxes. Let D denote the number of occupied boxes. The asymptotic behavior of D is well known as $n \to \infty$ and $k \to \infty$ such that $n/k \to m > 0$ (Geiringer, 1938; or see Johnson and Kotz, 1977, Chapter 6.1). If k is unknown, D is a sufficient statistic for k (Darroch, 1958). Asymptotically, the estimator for k, Y, is given by the solution of

$$D = Y(1 - e^{-n/Y}). \tag{2.1}$$

Now $C = D/k$, and the corresponding estimator for C is

$$\hat{C} = D/Y. \tag{2.2}$$

The associated normal limit law is:

THEOREM 1. *If all classes are equally likely and* $n \to \infty$ *and* $k \to \infty$ *such that* $n/k \to -\ln(1-c), 0 < c < 1$, *then*

$$n^{1/2}(C - \hat{C}) \to_D N\left(0, \frac{c^2[-(1-c)\ln(1-c)]}{c + (1-c)\ln(1-c)}\right).$$

COMMENT. Note that c has been chosen such that $E(C) \to c$ and $C \to_P c$.

The corresponding result for Good's estimator is:

THEOREM 2. *If all classes are equally likely and* $n \to \infty$ *and* $k \to \infty$ *such that* $n/k \to -\ln(1-c), 0 < c < 1$, *then*

$$n^{1/2}(C - \tilde{C}) \to_D N(0, (1-c)(c - \ln(1-c))).$$

THEOREM 3. *If all classes are equally likely and* $n \to \infty$ *and* $k \to \infty$ *such that* $n/k \to -\ln(1-c), 0 < c < 1$, *then the asymptotic relative efficiency of Good's estimator to the estimator based on the equally likely hypothesis is given by*

$$E = \frac{(-\ln(1-c))c^2}{(c + (1-c)\ln(1-c))(c - \ln(1-c))}.$$

COROLLARY. (a) *As* $c \to 0$, $E \to 1$. (b) *As* $c \to 1$, $E \to 1$. (c) $E > 0.85$ *for all* c, $0 < c < 1$.

For example, if $c = 0.1$, then $E = 0.9913$ and confidence intervals based on Good's estimator are asymptotically only 0.45% longer than those using the equally likely hypothesis. If $c = 0.5$, then $E = 0.9466$ and intervals are 2.8% longer. If $c = 0.9$, $E = 0.8735$ and intervals are 7% longer. The minimum value of E, 0.8516, is attained at $c = 0.978$. Thus Good's estimator is quite efficient.

3. Proofs. Since $D = k - N_0$, where N_0 denotes the number of unoccupied boxes, D is asymptotically normal with mean and variance asymptotic to

$$k(1 - e^{-n/k}) \quad \text{and} \quad ke^{-n/k}\big(1 - e^{-n/k} - (n/k)e^{-n/k}\big), \tag{3.1}$$

respectively (Johnson and Kotz, 1977, page 317, number 3, where an $m = k$ is missing on the right side). Since $C = D/k = 1 - N_0/k$, $E(C) \to c$ and $C \to_P c$. To prove Theorem 1 note that

$$n^{1/2}(C - \hat{C}) = n^{1/2}\left(\frac{D}{k} - \frac{D}{Y}\right) = \frac{D}{Y}n^{1/2}\frac{Y - k}{k}. \tag{3.2}$$

From (2.1), $D/Y \to_P c$, also. Thus in (3.2) it remains to determine the limit law of $n^{1/2}(Y - k)/k$. In (2.1), treating Y as a function of D, $Y' = [1 - e^{-n/Y} - (n/Y)e^{-n/Y}]^{-1}$, by implicit differentiation. Let $d = k(1 - e^{-n/k})$. $Y'(d) \to [c + (1 - c)\ln(1 - c)]^{-1}$. Expanding $Y(D)$ about d in a Taylor series,

$$\begin{aligned} \frac{n^{1/2}}{k}(Y - k) &= \left[1 - e^{-n/k} - \frac{n}{k}e^{-n/k}\right]^{-1}\frac{n^{1/2}}{k}(D - d) \\ &\quad + \frac{n^{1/2}}{k}O\big((D - d)^2\big). \end{aligned} \tag{3.3}$$

Now, $(n^{1/2}/k)(D - d) = (n^{1/2}/k)(D - E(D) + E(D) - d)$. Note that $n^{1/2}(E(D) - d)/k \to 0$. Using (3.1) and $n/k \to -\ln(1 - c)$, (3.3) is asymptotically normal with mean 0 and variance $(-(1 - c)\ln(1 - c))/(c + (1 - c)\ln(1 - c))$. The D/Y factor in (3.2) contributes the extra factor of c^2, and Theorem 1 is proven.

Under the hypotheses, Theorem 2 follows easily from Theorem 4 of Esty (1983), since $E(N_1)/n = (E(N_1)/k)(k/n) \sim (n/k)e^{-n/k}(k/n) \to 1 - c$, and $E(2N_2)/n \sim (n/k)^2e^{-n/k}(k/n) \to (-\ln(1 - c))(1 - c)$.

Theorem 3 follows immediately from Theorems 1 and 2. Corollary (a) is obtained from the Taylor expansion of $\ln(1 - c)$. Corollaries (b) and (c) are straightforward.

4. Conclusion. Good's nonparametric coverage estimator, which is appropriate for a wide variety of multinomial distributions, is remarkably efficient relative to the best coverage estimator developed under the strong hypothesis that all classes are equally likely, even when that hypothesis is true. Thus the advantages of its wider validity serve to recommend the Good estimator for the coverage of a sample even if the classes are approximately equally likely.

REFERENCES

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

BROWN, I. D. (1955 / 57). Some notes on the coinage of Elizabeth I with special reference to her hammered silver. *Brit. J. Numis.* **28** 568–603.

CARCASSONE, C. (1980). Tablés pour l'estimation par la méthod de maximum de vraisemblance du nombre de coins de droit (ou de reverse) ayant dervi à frapper une émission. In *Symposium Numismatico de Barcelona.* Asociación Numismatica Española, Barcelona.

CARTER, G. (1981). Comparison of methods for calculating the total number of dies from die-link statistics. *Statistics and Numismatics* (C. Carcassone and T. Hackens, eds.) 204–213. Council of Europe, Strasbourg.

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. (1978). *Stochastic Abundance Models.* Halsted, New York.

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

ESTY, W. W. (1983). A normal limit law for a nonparametric estimator of the coverage of a random sample. *Ann. Statist.* **11** 905–912.

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

GEIRINGER, H. (1938). On the probability theory of arbitrary linked events. *Ann. Math. Statist.* **9** 260–271.

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.

JOHNSON, N. L. and KOTZ, S. (1977). *Urn Models and Their Application.* Wiley, New York.

LYON, C. S. S. (1965). The estimation of dies employed in a coinage. *Numis. Circ.* **73** 180–181.

MCNEIL, D. R. (1973). Estimating an author's vocabulary. *J. Amer. Statist. Assoc.* **68** 92–96.

MORA-MÀS, F. (1981). Estimation du nombre de coins selon les répetitions dans une trouvaille de monnaies. In *Statistics and Numismatics* (C. Carcassone and T. Hackens, eds.) 173–192. Council of Europe, Strasbourg.

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

SCHROECK, F. E. JR. (1981). Tabulated results on the estimation of the number of dies of a coin. *Numis. Circ.* **89** 37–40.

STARR, N. (1979). Linear estimation of the probability of discovering a new species. *Ann. Statist.* **7** 644–652.

DEPARTMENT OF MATHEMATICAL SCIENCES
MONTANA STATE UNIVERSITY
BOZEMAN, MONTANA 59717