A Generalized Good's Nonparametric Coverage Estimator

by

Anne Chao, Shen-Ming Lee and Teh-Chuan Chen

Reprinted from the
CHINESE JOURNAL OF MATHEMATICS
September, 1988, Vol. 16, No. 3

TAIPEI, TAIWAN, REPUBLIC OF CHINA
A Generalized Good's Nonparametric Coverage Estimator

by

ANNE CHAO, SHEN-MING LEE and TEH-CHUAN CHEN

Abstract. A generalized form of Good's estimator (Good, 1953) is proposed to estimate the coverage of a random sample from an unknown multinomial distribution. Some properties of the proposed estimator are derived. This generalized estimator retains all good properties of Good's estimator and has smaller bias if each cell probability is less than 0.5. The results of a simulation study are reported to compare the performance of the proposed method with that of other existing estimators.

1. Introduction

A random sample of size \( n \) is taken with replacement from a population which consists of an unknown number of distinct classes, possibly countably many. Let \( p_i \) be the probability that any observation belongs to the \( i \)th class and \( X_i \) be the number of representatives of the \( i \)th class in the sample, \( \Sigma p_i = 1 \). The sample coverage, \( C \), is defined as the sum of the probabilities of the observed classes, that is,

\[
C = \sum_{i} p_i I[X_i > 0],
\]

where \( I[A] \) is the indicator function of the event \( A \). Note that \( C \) varies with samples and is actually a random variable.

An "estimator" proposed by Good (1953) has been widely used in various fields, see Good and Toulmin (1956), Robbins (1968), Engen (1978), Starr (1979), Chao (1981), Esty (1982, 1983, 1986) and many others. Good's estimator has the following form.

Key Words: sample coverage, occupancy problem, nonparametric estimator, frequency counts, bias-reduction.

Received November 17, 1987.
\[ \hat{C}_1 = 1 - \frac{N_j}{n}, \]

where

\[ N_j = \sum_i I[X_i = j], \quad j = 1, 2, \ldots, n \]

denotes the number of classes with exactly \( j \) representatives in the sample. \( \{N_j\} \) is usually referred to as the frequency counts of the sample. The asymptotic normality and associated confidence interval of Good’s estimator under very general conditions were developed by Esty (1982, 1983).

A closely related problem is to predict \( 1 - \mathcal{C} \), the conditional probability of discovering a new species in one additional observation. Most previous authors mainly focused on estimating the corresponding unconditional probability \( \theta \), where

\[ \theta = E(\mathcal{C}) = 1 - \sum p_i(1 - \hat{p}_i)^* . \]

Estimators based on more additional observations were discussed in Robbins (1968), Starr (1979), Chao (1981) and Clayton and Frees (1987). In this work, we will only consider the estimation problem based on the original sample. Clayton and Frees (1987) recently proposed using the nonparametric maximum likelihood estimator (NMLE), \( \hat{C} \), of \( \theta \) to predict \( \mathcal{C} \), where NMLE \( \hat{C} \) is

\[ \hat{C} = 1 - \sum \hat{p}_i(1 - \hat{p}_i)^*. \]

and \( \hat{p}_i = \frac{X_i}{n} \) is the fraction of times the class \( i \) is observed in the sample. They further recommended a bias-corrected form of NMLE, \( \hat{C}^* \), and

\[ \hat{C}^* = 1 - 2 \sum \hat{p}_i(1 - \hat{p}_i)^*. \]

In classical occupancy problems, it is assumed that the number of classes is finite and all classes are equally likely. Under this restriction, the sample coverage becomes \( D/k \), where \( k \) is the number of classes and
$D$ is the number of distinct classes observed in the sample. Thus our purpose reduces to the problem of estimating the number of classes $k$, which has received much attention in the literature, see Lewontin and Prout (1956), Darroch (1958), Harris (1968) Samuel (1968) Johnson and Kotz (1977, pp. 136-136). In this special case, the "best" estimator $\hat{C}_B$ in the sense of being approximate maximum likelihood and approximate minimum variance unbiased estimators (when the latter exists) under the equally-likely assumption, has the following form:

$$\hat{C}_B = D/k,$$

where $k$ is the unique solution of $D = k(1 - \exp(-n/k))$.

This raises the question that whether the use of Good's estimator is accompanied by a substantial loss in efficiency relative to the best estimator derived under equiprobable assumption. Esty (1986) answered this question and found that Good's estimator is remarkably efficient relative to the best estimator even when the hypothesis is true.

In Section 2, we generalize Good's estimator to a class of new estimators which utilize $N_1, N_2, \cdots, N_n$. Good's estimator thus becomes a special case (i.e. the first order) of the generalized estimators. Some statistical properties of the generalized estimators are obtained. Generally we recommend the estimator using all the frequency counts (i.e. the $n$th order generalized estimator).

A simulation study was carried out to study the performance of various estimators for small and moderate samples. The NMLE may be an attractive estimator for inference based on more additional observations as discussed by Clayton and Frees (1987). However, for our problem, the NMLE will be shown to have some deficiencies if the sample coverage is not sufficiently high. All the comparison results are presented in Section 3.

2. GENERALIZED ESTIMATORS

We first define the $m$th order generalized estimator as
\begin{equation}
\hat{C}_m = 1 - \frac{N_1}{n} + \frac{N_2}{\binom{n}{2}} - \cdots + (-1)^m \frac{N_m}{\binom{n}{m}}, \quad m \leq n.
\end{equation}

If \( \hat{C}_m > 1 \), then this estimator is modified to be equal to 1. For example, this happens to \( \hat{C}_n \) when \( n \) is even, \( N_n = 1 \) and no other frequency counts. It means that all observations belong to the same class. Intuitively, this class is likely to be only observable class and the sample coverage is then approximately equal to 1. When \( m = 1 \), \( \hat{C}_1 \) is exactly Good's estimator using \( N_1 \) only.

Based on frequency counts \( N_1, N_2, \ldots, N_n \), Starr (1979) proved that there is no linear unbiased estimator \( \theta \) based on the frequency counts \( N_1, N_2, \ldots, N_n \) from a sample of size \( n \) and indicated that the claim by Knott (1967) of \( \hat{C}_n \) being an unbiased estimator of \( \theta \) was not correct. Clayton and Frees (1987) further proved that no unbiased estimator of \( \theta \) exists based on a sample of size less than \( n + 1 \).

To see how the generalized estimator was derived, notice that for \( m \leq n \),

\begin{equation}
\theta = 1 - \sum_i p_i (1 - p_i)^n
\end{equation}

\begin{equation}
= 1 + \sum_{j=1}^m \sum_i (-1)^j p_i (1 - p_i)^{n-j} + \sum_i (-1)^{m+1} p_i^{m+1} (1 - p_i)^{n-m},
\end{equation}

and

\begin{equation}
EN_j = \sum_i \binom{n}{j} p_i (1 - p_i)^{n-j}, \quad j = 1, 2, \ldots, n.
\end{equation}

Thus

\begin{equation}
\theta = 1 + \sum_{j=1}^m (-1)^j \frac{EN_j}{\binom{n}{j}} + \sum_i (-1)^{m+1} p_i^{m+1} (1 - p_i)^{n-m}.
\end{equation}

If the last term of (2.4) is ignored and \( EN_j \) is replaced by \( N_j, j = 1, 2, \cdots, m \), we then obtain \( \hat{C}_m \). Actually, the last term is exactly the negative bias of \( \hat{C}_m \) with respect to \( C \) (or to \( \theta \)).

We can also obtain the same class of estimators from a bias-reduction
point of view. Note that Good’s estimator is biased low, and

\[ E(\hat{C}_1 - C) = - \sum p_i (1 - p_i)^{n-1}. \]

From (2.3), the magnitude of this bias can be approximately estimated by

\[ -EN_i\binom{n}{2}/(2) \]

and the corrected estimator will have less absolute bias if

\( p_i < 0.5 \) for all \( i \) as will be shown in Property 2. After correcting this bias, we obtain the second order generalized estimator \( \hat{C}_2 \). However, \( \hat{C}_2 \)
has a positive bias, and the bias is

\[ E(\hat{C}_2 - C) = \sum p_i (1 - p_i)^{n-2} \sim EN_i\binom{n}{3}. \]

Continuing this bias-correcting procedure, we therefore get resulting form.

We have derived some statistical properties about \( \hat{C}_m, m = 1, 2, \ldots, n \), which are presented in the following:

**Property 1:** Under the equiprobable assumption, (i.e., \( p_1 = p_2 = \cdots = p_k = 1/k \)), if \( n \to \infty, k \to \infty \) such that \( n/k \to \lambda, 0 < \lambda < \infty \), then we have for fixed \( m = 1, 2, \ldots, \)

\[ E(\hat{C}_m - C) = O(n^{-m}), \]

\[ E(\hat{C}_m - C)^2 = \lambda e^{-\lambda}\left(1 + \frac{1}{\lambda} - \frac{e^{-\lambda}}{\lambda}\right)n^{-1} + O(n^{-2}), \]

\[ n^{1/2}(\hat{C}_m - C) \xrightarrow{d} N\left(0, \lambda e^{-\lambda}\left(1 + \frac{1}{\lambda} - \frac{e^{-\lambda}}{\lambda}\right)\right). \]

**Property 2:** In a general population, if for all \( i, p_i < 1/2 \), then for any \( n \) and \( k \) greater than 1

\[ |E(\hat{C}_1 - C)| \geq |E(\hat{C}_2 - C)| \geq \cdots \geq |E(\hat{C}_m - C)|. \]

**Property 3:** (This Property shows that how to construct simple asymptotic variance estimators) In a general population, the mean squared error (MSE) of \( \hat{C}_m \) with respect to \( C \) is obtained as

\[ E\left(\hat{C}_m - C\right)^2 = \left(\frac{EN_i}{n^2} + \frac{2EN_i}{n^2} - \frac{EN_i^2}{n^3}\right) + O(n^{-2}), \quad m = 1, 2, \ldots, n. \]
More accurately, we have

\[ E(\hat{C}_1 - C)^2 = \left( \frac{EN_1}{n^2} + \frac{2EN_2}{n^3} - \frac{EN_3}{n^4} \right) \]

\[ + \left( \frac{EN_1}{n^3} + \frac{2EN_2}{n^4} - \frac{12EN_3}{n^5} + \frac{12E(N_1N_2)}{n^6} \right) \]

\[ + \frac{8EN_3^2}{n^4} - \frac{EN_1^2}{n^4} \right) + O(n^{-5}), \]

and

\[ E(\hat{C}_i - C)^2 = E(\hat{C}_1 - C)^2 + \left( \frac{8E(N_1N_2)}{n^4} - \frac{4EN_i^2}{n^4} - \frac{12E(N_1N_3)}{n^6} \right) \]

\[ + O(n^{-5}), \quad i = 2, 3, \cdots, n. \]

Property 4: If \( n \) is large and \( EN_i/n \to \alpha_i \) and \( EN_i/n \to \alpha_i \), \( 0 < \alpha_i < 1 \). Then

\[ n^{1/3}[\hat{C}_n - C]/\left[ N_1/n + 2N_2/n - \left( N_1/n \right)^2 \right]^{1/2} \n \to N(0, 1). \]

Thus an approximate confidence interval for \( C \) based on \( \hat{C}_n \) can be obtained in a usual way.

All the derivation details are omitted here. We only remark that the Property 4 follows directly from the proof of Esty (1983) using a method of Holst (1979).

It is clear from these results that \( \hat{C}_n \) retains all good properties of Good's estimator. For example, it is asymptotically unbiased, consistent, asymptotically normal, and it has the same asymptotic efficiency as Good's estimator relative to the best estimator assuming equiprobable assumption and thus it is as highly efficient as Good's estimator.

Properties 1 and 2 show that in equiprobable case, Good's estimator has bias of order \( O(n^{-1}) \), while \( \hat{C}_n \) has order of \( O(n^{-3}) \); and for general populations, \( \hat{C}_n \) has the smallest magnitude of bias among this class of new estimators as long as \( p_i < 0.5 \) for all \( i \). However, this reduction in bias does not cause any non-negligible increase in \( MSE \) for large sample sizes. For finite sample sizes, the \( MSE \)s of \( \hat{C}_1 \) and \( \hat{C}_n \) are generally comparable based
on our simulation results in next section. Thus, the estimator $\hat{C}_n$ using all frequency counts is suggested for practical use.

3. NUMERICAL COMPARISONS

For small and moderate sample sizes, a simulation study was conducted to evaluate the comparative adequacy of previously mentioned estimation procedures under various distributional assumptions. All the computations were conducted on a CDC Cyber-840 computer using a Fortran program. The uniform random numbers were generated from the uniform multiplicative congruential generator RANF available on CDC computers.

In our study, four classes of distributions were used to model the probability distribution \( \{p_i\} \). These included (a) equiprobable case, with \( p_i = 1/k, \quad i = 1, 2, \ldots, k \), (b) truncated geometric distribution, with \( p_i = (1-p) p_i^{-1}/(1-p^i), \quad i = 1, 2, \ldots, k, \quad 0 < p < 1 \), (c) (normalized) uniform \((0, 1)\) distribution, with \( p_i = p_i / \sum p_i \), where \( p_i^*, p_i^*, \ldots, p_i^* \) are a random sample from a uniform \((0, 1)\) distribution, and (d) (normalized) beta \((1,4)\) distribution, with \( p_i = p_i^* / \sum p_i^* \), where \( p_i^*, p_i^*, \ldots, p_i^* \) are a random sample from beta \((1,4)\) distribution. We remark that the results for other beta distributions are quite similar to those of beta \((1,4)\). Thus they are not reported here.

For each specified distribution, several combinations of values of \( n \) and \( k \) were considered. For each given \( n, k \) and \( \{p_i\} \), we generated 1000 data sets. In equiprobable (non-equiprobable) case, we calculated $\hat{C}_1, \hat{C}_n, \hat{C}_B, \hat{C}, \hat{C}^*$ ($\hat{C}_1, \hat{C}_n, \hat{C}, \hat{C}^*$) and their corresponding biases and MSEs with respect to \( C \) for each data set. Finally, these 1000 biases and MSEs were averaged to give the results of Tables 1–4. Note here $\theta$ is always fixed but \( C \) varies with data sets. We also present the average values of \( C \), which are very close to $\theta$. Our (unreported) numerical results indicated that the relative performance of estimators evaluated with respect to \( C \) and to $\theta$ agree favorably well. Thus we only present here the results with respect to \( C \).

Tables 1–4 show that our proposed estimator $\hat{C}_n$ generally has the smallest bias except the truncated geometric with \( p = .1 \), for which case
Table 1. Comparison of Estimators for Equiprobable Case

<table>
<thead>
<tr>
<th>$k$</th>
<th>$n$</th>
<th>$\theta$</th>
<th>$\bar{C}$</th>
<th>Average Bias</th>
<th>Average RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\hat{C}_1$</td>
<td>$\hat{C}_n$</td>
<td>$\hat{C}_n^*$</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>0.6513</td>
<td>-0.017</td>
<td>-0.002</td>
<td>0.005</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>0.8784</td>
<td>-0.019</td>
<td>-0.002</td>
<td>0.005</td>
</tr>
<tr>
<td>50</td>
<td>10</td>
<td>0.9949</td>
<td>-0.012</td>
<td>-0.002</td>
<td>0.005</td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>0.9999</td>
<td>-0.012</td>
<td>-0.002</td>
<td>0.005</td>
</tr>
</tbody>
</table>

Table 2. Comparison of Estimators for Truncated Geometric Distribution

<table>
<thead>
<tr>
<th>$p$</th>
<th>$k$</th>
<th>$n$</th>
<th>$\theta$</th>
<th>$\bar{C}$</th>
<th>Average Bias</th>
<th>Average RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>$\hat{C}_1$</td>
<td>$\hat{C}_n$</td>
<td>$\hat{C}$</td>
</tr>
<tr>
<td>1</td>
<td>20</td>
<td>0.9779</td>
<td>-0.004</td>
<td>-0.014</td>
<td>-0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>2</td>
<td>20</td>
<td>0.9797</td>
<td>-0.005</td>
<td>-0.015</td>
<td>-0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>0.9954</td>
<td>-0.006</td>
<td>-0.016</td>
<td>-0.004</td>
<td>0.003</td>
</tr>
</tbody>
</table>

Table 3. Comparison of Estimators for Uniform (0, 1) Distribution

<table>
<thead>
<tr>
<th>$k$</th>
<th>$n$</th>
<th>$\theta$</th>
<th>$\bar{C}$</th>
<th>Average Bias</th>
<th>Average RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\hat{C}_1$</td>
<td>$\hat{C}_n$</td>
<td>$\hat{C}$</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>0.6783</td>
<td>-0.010</td>
<td>-0.001</td>
<td>0.004</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>0.8793</td>
<td>-0.013</td>
<td>-0.002</td>
<td>0.004</td>
</tr>
<tr>
<td>50</td>
<td>10</td>
<td>0.9876</td>
<td>-0.016</td>
<td>-0.002</td>
<td>0.004</td>
</tr>
<tr>
<td>100</td>
<td>10</td>
<td>0.9992</td>
<td>-0.017</td>
<td>-0.003</td>
<td>0.004</td>
</tr>
</tbody>
</table>

Note: $\bar{C}$ represents the average value of the estimator, $\hat{C}_1$ and $\hat{C}_n$ represent the bias of the estimator towards the first and last data points, respectively, $\hat{C}$ represents the overall bias, and $\hat{C}^*$ represents the root mean square error of the estimator.
the assumption of Property 2 (for bias-reduction) is not satisfied. Although $\hat{C}_B$ in equiprobable case is nearly unbiased, its bias is larger than that of $\hat{C}_a$. On the other hand, either $\bar{C}$ or $\bar{C}^*$ has the largest bias. Note that $\bar{C}$ or $\bar{C}^*$ works well only in the high coverage cases and it has severe positive bias (and $\text{MSE}$ also) when the sample coverage is low. Intuitively, if a large portion of the probability is distributed over classes with relatively small $p_i$'s, the $\text{NMLE}$ does not work well. For example, in equiprobable case, if the number of classes is large relative to the sample size, most cells are nearly "unobservable" in the sample. However, the cell probability corresponding to empty cell is estimated to be zero in the nonparametric estimation procedure. Hence if the number of empty cells becomes large, the $\text{NMLE}$ is expected to have a substantial bias.

Before we discuss the behavior of root mean squared error ($\text{RMSE}$), notice that for truncated geometric distribution with $p = .1$ and $.5$, for each fixed $n$, there is almost no difference between two biases (or two $\text{RMSEs}$) when $k$ is increased from 20 to 100. This can be explained as follows: for $k = 20$ or 100, $p = .1$ or $.5$, $p = 0$ thus $p_i \approx (1 - p) p^{i-1}$, thus all "observable" cell probabilities are nearly the same for both populations. (Actually, the cell probabilities for $p = .1$ are .9, .099,..., hence there are probably only very few observable cells. Similarly for the case of $p = .5$.) Therefore we would expect that for any population size greater than 20, the performance of any estimator will remain the same as long as the sample size is fixed.

<table>
<thead>
<tr>
<th>$h$</th>
<th>$n$</th>
<th>$e$</th>
<th>$\bar{C}$</th>
<th>Average Bias</th>
<th>Average RMSE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>$\hat{C}_1$</td>
<td>$\hat{C}_2$</td>
<td>$\bar{C}$</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>.7596</td>
<td>.7593</td>
<td>.0366</td>
<td>.0059</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>.9072</td>
<td>.9077</td>
<td>.0061</td>
<td>.0009</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>.9740</td>
<td>.9741</td>
<td>.0006</td>
<td>.0001</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>.9911</td>
<td>.9906</td>
<td>.0001</td>
<td>.0002</td>
</tr>
<tr>
<td>50</td>
<td>10</td>
<td>.2797</td>
<td>.2795</td>
<td>.0253</td>
<td>.0019</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>.4623</td>
<td>.4615</td>
<td>.0128</td>
<td>.0026</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>.7350</td>
<td>.7259</td>
<td>.0023</td>
<td>.0019</td>
</tr>
<tr>
<td>100</td>
<td>100</td>
<td>.8873</td>
<td>.8877</td>
<td>.0030</td>
<td>.0014</td>
</tr>
</tbody>
</table>

Table 4. Comparison of Estimators for Beta (1, 4) Distribution
Now we compare the behavior of the RMSE. In equiprobable cases, as we expected, $\hat{C}_B$ generally has the lowest RMSE. However, the three estimators $\hat{C}_1$, $\hat{C}_n$ and $\hat{C}_B$ produce very close RMSEs. This indicates that even in small and moderate samples, both Good’s and our proposed estimators are remarkably efficient with respect to the best estimator. Note that when the sample coverage is low, both $\bar{C}$ and $\bar{C}^*$ have considerable large values of RMSE.

For truncated geometric distribution, $\bar{C}$ is superior to the others in terms of RMSE except for the case of $p = .9$ and $k = 100$. This is clearly seen because for this distribution, the sample coverages for most cases are quite high especially for small values of $p$. When $p = .9$ and $k = 100$, for which case the sample coverages are usually moderate, $\bar{C}$ performs worst and the other three estimators are comparable.

For uniform and beta distributions, when the sample coverage is very high (i.e., in the case of $k = 10$), $\bar{C}$ again dominates the other estimators with respect to RMSE; when the sample coverage is low or moderate, for which cases $\bar{C}$ and $\bar{C}^*$ usually have non-negligible biases, $\bar{C}$ has the worst RMSE and the other three estimators compare favorably in the values of RMSE.

In summary, the NMLE is preferred in terms of RMSE only for high coverage situations. In other cases, there will be very little differences among the values of RMSE of $\hat{C}_1$, $\hat{C}_n$ and $\hat{C}_B$ under equiprobable assumption, and for general populations, $\hat{C}_1$, $\hat{C}_n$ and $\hat{C}^*$ are usually comparable with respect to REMS. Since in most of the practical cases, $\hat{C}_n$ has the smallest bias, we would recommend the use of $\hat{C}_n$ unless the sample coverage is extremely high.

ACKNOWLEDGEMENTS

The authors thank the referee for reading carefully this manuscript and helpful comments.

REFERENCES


Institute of Statistics
National Tsing Hua University
Hsinchu, Taiwan
Republic of China