



A Poisson model for coverage problems with an
application in genomic research

By CHANGXUAN MAO and BRUCE G. LINDSAY

Technical Report #01-06-01

2001

Center for Likelihood Studies
DEPARTMENT OF STATISTICS
THE PENNSYLVANIA STATE UNIVERSITY
UNIVERSITY PARK, PA 16802

A Poisson model for coverage problems with an application in genomic research

BY C. MAO AND B. G. LINDSAY

*Department of Statistics, Pennsylvania State University,
326 Thomas Building, University Park, PA 16802-2111, U.S.A.
cmao@stat.psu.edu, bgl@psu.edu*

SUMMARY

Suppose a population has infinitely many individuals and is partitioned into N disjoint classes. For any k , the abundance- k coverage of a random sample from the population is defined to be the sum of the proportions of the classes that contribute exactly k individuals in the sample. The sample coverage is the total proportions of the classes that contribute at least one individual in the sample. The asymptotic distribution for the abundance- k coverage is developed under a Poisson model. A new derivation of the well-known Turing's estimators is presented. It shows that Turing's estimators are sensible if N is large enough. As an application, a gene classification issue in genomic research is addressed. Since Turing's approach is method of moment estimation, maximum likelihood estimation is presented as an alternative approach for the coverage problem. Finally, we show that any Turing-type estimator is asymptotic fully efficient among a class of estimators satisfying the regularity conditions defined by Tierney and Lambert.

Some Key words: Abundance- k coverage; sample coverage; number of species; Poisson process.

1. INTRODUCTION

Suppose there is a population composed of infinitely many individuals. The population has been partitioned into N disjoint classes. A random sample is taken from the population. For any k , the *abundance- k coverage* of the random sample is defined to be the sum of the proportions of the classes that contribute exactly k individuals in the sample. The *sample coverage* is the sum of the proportions of the classes that contribute at least one individual in the sample. The well-known coverage estimators due to Turing were discussed in Good (1953, 1956) as well as some related issues, such as estimation of the number of classes and estimation of the probability of discovering a new class. Darroch and Ratcliff (1980) and Chao (1992, 1993) developed estimators for N based on the sample coverage. There are many papers addressing the probability that a new class is observed when the sample is enlarged, for example, see Robbins (1968), Starr (1979), Chao (1981), and Clayton and Frees (1987).

The authors were motivated to address inference about coverage by certain problems arising in genomic studies. Scientists are interested in categorizing genes by abundance, for example, see Cantor and Smith (1999). One issue of

interest is the estimation and confidence inference about the total proportions of the abundant genes and the rare genes in a library. The only available confidence inference is about the sample coverage studied in Esty (1982, 1983). The analysis in Esty (1982, 1983) is quite complicated, which encourages us to consider modeling the coverage problem by other means.

It has been recognized that Turing's estimators have wide validity. The original derivation of Turing's estimators given in Good (1953) is from a Bayesian point of view. This paper develops a simple frequentist model in which the coverage problem can be readily analyzed. In particular, it provides a framework to assess Turing-type estimators as well as other natural model-based competitors. In addition to providing estimators, such a model allows direct and easy development of confidence statements both asymptotically and by bootstrapping.

Our results are derived by treating the sampling model as a superposition of N independent Poisson processes. The rates of these Poisson processes are assumed to be a random sample from a latent distribution with finite variance. Such a representation brings much simplicity in modeling. A mixture model arises when the sampling is stopped at any fixed time. The limiting distribution for the abundance- k coverage is developed under the mixture model.

In Section 2, we build up a Poisson model for the coverage problem and establish a central limit theorem from which the limiting distribution of the abundance- k coverage can be readily derived. In Section 3, the estimation of the mean abundance- k coverage and prediction of the abundance- k coverage are addressed together. The limiting distributions about the estimation and prediction are derived from the central limit theorem developed in Section 2. In Section 4, a genomic data set is studied as an application, in which it shows that our framework facilitates the analysis of any specific Turing-type estimator. In the last section, alternative approaches and efficiency of Turing-type estimators are investigated. From the mixture model point of view, Turing-type estimators are moment estimators, maximum likelihood estimation provides an alternative approach. Inference about the abundance- k coverage can be easily obtained from bootstrapping. By applying a general result in Tierney and Lambert (1984) about asymptotic efficiency of functionals of mixture distributions, it turns out that Turing-type estimators are asymptotically fully efficient among a class of estimators that satisfy some regularity conditions given in Tierney and Lambert (1984).

2. THE COVERAGE PROBLEM

Let the classes be indexed by $1, 2, \dots, N$. The number of individuals from the i th class in the sample is assumed to follow a homogeneous Poisson process, denoted by $X_i(t)$, with rate ξ_i , $i = 1, 2, \dots, N$. The ξ_i 's are assumed to be a sample from a latent distribution $\tilde{Q}(\xi)$, which is assumed to have finite variance. The $X_i(t)$'s are assumed to be independent. Let $\lambda_i(t)$ be the mean

value function of $X_i(t)$. Then

$$\lambda_i(t) = \xi_i t, \quad i = 1, 2 \dots N. \quad (1)$$

The $\lambda_i(t)$'s are a sample arising from a latent distribution $Q_t(\lambda)$, where

$$Q_t(\lambda) = \tilde{Q}(\lambda/t). \quad (2)$$

So for any finite t , the latent distribution $Q_t(\lambda)$ also has finite variance. Let π_i be the true proportion of i th class defined by

$$\pi_i = \frac{\xi_i}{\sum_{j=1}^N \xi_j} = \frac{\lambda_i(t)}{\sum_{j=1}^N \lambda_j(t)}, \quad i = 1, 2 \dots N. \quad (3)$$

Each $X_i(t)$ is a Poisson random variable with mean $\lambda_i(t)$. Let Ω be the sample space of a Poisson distribution, which is the set of all nonnegative integers. The $X_i(t)$'s constitute a *Poisson sample* for any fixed time t . Let $s(t)$ be the number of individuals observed at time t , called the *sample size*. For any k in Ω , let $n_k(t)$ be the number of classes that contribute exactly k individuals up to time t , called *abundance- k frequency*, and let $n(t)$ be the number of distinct classes observed in the sample up to time t , which can be written as $N - n_0(t)$, that is,

$$n_k(t) = \sum_{i=1}^N I(X_i(t) = k), \quad n(t) = \sum_{k=1}^{+\infty} n_k(t) \text{ and} \quad (4)$$

$$s(t) = \sum_{i=1}^N X_i(t). \quad (5)$$

Conditioning on the sample size $s(t)$, the $X_i(t)$'s constitute a *multinomial sample*. That is, they arise from a multinomial distribution with index N and cell probabilities π_i 's. Coverage models are usually based on this multinomial framework, for example, see Good (1953) and Esty (1982, 1983). The Poisson framework used here seems to be both new and useful.

For any k in Ω , let $A(k, t)$ be the total proportion of those classes that contribute exactly k individuals up to time t . We will call this *abundance- k coverage*, that is,

$$A(k, t) = \sum_{j=1}^N \pi_j I(X_j(t) = k), \quad (6)$$

where $I(E)$ is the indicator function of the event E .

The abundance-0 coverage is the total proportion of those classes that have not been observed through time t . The *sample coverage* at time t is the total proportion of those classes actually seen up to time t , denoted by $C(t)$, that is,

$$C(t) = \sum_{j=1}^N \pi_j I(X_j(t) > 0) = 1 - A(0, t). \quad (7)$$

We can express $A(k, t)$ and $C(t)$ in terms of the $\lambda_i(t)$ instead of the π_i 's as

$$A(k, t) = \frac{\sum_{j=1}^N \lambda_j(t) I(X_j(t) = k)}{\sum_{j=1}^N \lambda_j(t)} \quad \text{and} \quad (8)$$

$$C(t) = \frac{\sum_{j=1}^N \lambda_j(t) I(X_j(t) > 0)}{\sum_{j=1}^N \lambda_j(t)}. \quad (9)$$

Under such a representation, $A(k, t)$ and $C(t)$ can be regarded as random variables which are functions of the observed $X_i(t)$'s and the unobserved $\lambda_i(t)$'s. The original definitions of $A(k, t)$ and $C(t)$ contain constrained fixed parameters and random variables, which makes inference much more difficult. The key idea in this paper is to utilize the new representation to derive some results about the estimation and confidence inference of the coverages.

We consider a sampling framework in which observations have been made up to a fixed time t . Therefore, we will omit t in later discussions. That is, $s(t)$ will be written as s , $n_k(t)$ as n_k , $n(t)$ as n , $A(k, t)$ as $A(k)$, $C(t)$ as C , Q_t as Q , $X_i(t)$ as X_i , and $\lambda_i(t)$ as λ_i . To emphasize that λ is a (latent) random variable, we will use the notation Λ .

Let Λ be a random variable from the latent distribution $Q(\lambda)$. Let X be a Poisson random variable with conditional mean $\Lambda = \lambda$. Let $f(x; \lambda)$ be the density of X given $\Lambda = \lambda$ with respect to the counting measure $\omega(x)$. Let $f(x; Q)$ be the mixture density of X with respect to $\omega(x)$. Hence

$$f(x; Q) = Ef(x; \Lambda) = \int f(x; \lambda) dQ(\lambda), \quad x \in \Omega. \quad (10)$$

From our present point of view, X_1, X_2, \dots, X_N , constitute a random sample from (10). The joint density of $\{X_i\}_{i=1}^N$ is given by

$$\prod_{i=1}^N f(x_i; Q) = \prod_{j \in \Omega} f(j; Q)^{n_j}.$$

The joint density of the abundance frequencies $\{n_j\}_{j \in \Omega}$ is given by

$$f(n_0, n_1, \dots; Q) = \frac{N}{\prod_{j \in \Omega} n_j!} \prod_{j \in \Omega} f(j; Q)^{n_j}. \quad (11)$$

That is, $\{n_j\}_{j \in \Omega}$ follows a multinomial distribution with index N and probabilities $f(j; Q)$, $j \in \Omega$. The following result is clear:

$$En_k = Nf(k; Q), \quad k \in \Omega. \quad (12)$$

We note that model (10) could also arise in sampling from non-homogeneous Poisson processes provided the mean value function at time t can be considered

as a random sample from Q . Our asymptotic results will be based on the number of classes N becoming infinite rather than letting t go to infinity. One unusual feature of this is that the number of classes N is not known since n_0 is not observed. So the application of the asymptotic results requires special care.

Our analysis depends on a general central limit theorem result for which we need to define the following random variables. For all j , $j = 1, 2, \dots, N$, let

$$Y_j = \Lambda_j I(X_j = k), \quad (13)$$

$$Z_j = \Lambda_j, \quad (14)$$

$$U_j = (k+1)I(X_j = k+1), \quad (15)$$

$$V_j = \sum_{k \in \Omega} (k+1)I(X_j = k+1) \text{ and} \quad (16)$$

$$\mathbf{W}_j = (Y_j, Z_j, U_j, V_j)^T. \quad (17)$$

Now set

$$\bar{Y} = N^{-1} \sum_{j=1}^N Y_j \quad (18)$$

$$\bar{Z} = N^{-1} \sum_{j=1}^N Z_j, \quad (19)$$

$$\bar{U} = N^{-1} \sum_{j=1}^N U_j, \quad (20)$$

$$\bar{V} = N^{-1} \sum_{j=1}^N V_j \text{ and} \quad (21)$$

$$\bar{\mathbf{W}} = (\bar{Y}, \bar{Z}, \bar{U}, \bar{V})^T. \quad (22)$$

The mean and asymptotic covariance matrix will be determined by the following terms. For all k in Ω , set

$$\alpha_k = (k+1)f(k+1; Q) \quad \text{and} \quad \beta_k = (k+1)\alpha_{k+1}. \quad (23)$$

Then set

$$\alpha = \sum_{j \in \Omega} \alpha_j \quad \text{and} \quad \beta = \sum_{j \in \Omega} \beta_j. \quad (24)$$

Finally, define

$$\boldsymbol{\mu} = (\alpha_k, \alpha, \alpha_k, \alpha)^T \text{ and} \quad (25)$$

$$\Sigma = \begin{pmatrix} \beta_k & \beta_k & 0 & k\alpha_k \\ \beta_k & \beta & \beta_k & \beta \\ 0 & \beta_k & (k+1)\alpha_k & (k+1)\alpha_k \\ k\alpha_k & \beta & (k+1)\alpha_k & \alpha + \beta \end{pmatrix} - \boldsymbol{\mu}\boldsymbol{\mu}^T. \quad (26)$$

Our main results are based on the following theorem.

Theorem 1 Let $\mathbf{W} = \mathbf{W}_1$. The mean vector of \mathbf{W} is $\boldsymbol{\mu}$ and Σ is the variance-covariance matrix for \mathbf{W} . As N goes to infinity, we have

$$\sqrt{N}(\bar{\mathbf{W}} - \boldsymbol{\mu}) \xrightarrow{d} N(0, \Sigma). \quad (27)$$

The proof is straightforward and is given in the appendix. The convergence of the series used to define α and β in (24) is guaranteed by the finite variance assumption of Q .

As a corollary, we can provide asymptotic normality results for the abundance- k coverage random variable $A(k)$ and the sample coverage C . For k in Ω , define

$$A_e(k) = \alpha_k/\alpha \text{ and } \sigma_k^2 = \alpha^{-4}[\beta_k\alpha^2 - 2\alpha\alpha_k\beta_k + \beta\alpha_k^2]. \quad (28)$$

Additionally, set

$$C_e = 1 - A_e(0) \text{ and } \sigma^2 = \sigma_0^2. \quad (29)$$

The following proposition describes the limiting distribution of $A(k)$ (and C).

Proposition 2 For all k in Ω , as N goes to infinity, we have

$$\sqrt{N} \frac{A(k) - A_e(k)}{\sigma_k} \xrightarrow{d} N(0, 1). \quad (30)$$

Additionally,

$$\sqrt{N} \frac{C - C_e}{\sigma} \xrightarrow{d} N(0, 1). \quad (31)$$

The proof is given in the appendix.

Proposition 2 can be stated informally as follows:

$$A(k) \sim N(A_e(k), N^{-1}\sigma_k^2) \text{ and } C \sim N(C_e, N^{-1}\sigma^2) \text{ approximately.} \quad (32)$$

3. SEMI-PARAMETRIC INFERENCE

In the present framework, one might consider estimating the parameter $A_e(k)$ (and C_e), which represents the mean abundance- k coverage (mean sample coverage), or estimating the random quantity $A(k)$ (and C). The latter might be termed a prediction problem rather than an estimation one. Both points of view will be considered here.

We consider first the estimation of $A_e(k)$, $N^{-1}\sigma_k^2$, C_e and $N^{-1}\sigma^2$. For all k in Ω , $f(k; Q)$ can be estimated by n_k/N , denoted by $\hat{f}(k; Q)$. Essentially these estimators are moment estimators since a mixture density can be represented as moments of a measure, see Lindsay (1995). One can estimate α_k , β_k , α and β

in (23) and (24) using the $\widehat{f}(x; Q)$'s. The estimators are denoted by $\widehat{\alpha}_k$, $\widehat{\beta}_k$, $\widehat{\alpha}$ and $\widehat{\beta}$ respectively. When $\widehat{\alpha}_k$, $\widehat{\beta}_k$, $\widehat{\alpha}$ and $\widehat{\beta}$ replace α_k , β_k , α and β respectively in formulas (28) and (29) for $A_e(k)$, σ_k^2 , C_e and σ^2 , we obtain estimators for them, denoted by $\widehat{A}_e(k)$, $\widehat{\sigma}_k^2$, \widehat{C}_e and $\widehat{\sigma}^2$ respectively. All of these estimators are consistent, see the appendix. Set

$$m = \sum_{j \in \Omega} (j+1)(j+2)n_{j+2}. \quad (33)$$

We can write these estimates as follows:

$$\widehat{A}_e(k) = \frac{(k+1)n_{k+1}}{s}, \quad (34)$$

$$\widehat{C}_e = 1 - \frac{n_1}{s}, \quad (35)$$

$$N^{-1}\widehat{\sigma}_k^2 = \frac{(k+1)^2(k+2)}{s^4} \left[\frac{n_{k+2}s^2}{k+1} - 2n_{k+1}n_{k+2}s + \frac{mn_{k+1}^2}{k+2} \right] \text{ and} \quad (36)$$

$$N^{-1}\widehat{\sigma}^2 = \frac{1}{s^4} [2n_2s^2 - 4n_1n_2s + mn_1^2]. \quad (37)$$

Note that $\widehat{A}_e(k)$, \widehat{C}_e , $N^{-1}\widehat{\sigma}_k^2$ and $N^{-1}\widehat{\sigma}^2$ do not depend on the unknown constant N . Formulas in (34) and (35) are identical to Turing's estimators for $A(k)$ and C given in Good (1953).

We need to consider the variance of $\widehat{A}_e(k)$ (and \widehat{C}_e) when it is used as an estimator for $A_e(k)$ (and C_e). Define the variance functions:

$$\rho_k^2 = \alpha^{-4}(k+1)\alpha_k\alpha^2 + \alpha_k^2[\beta - (2k+1)\alpha] \text{ and} \quad (38)$$

$$\rho^2 = \rho_0^2 = \alpha^{-4}[\alpha_0\alpha^2 + \alpha_0^2(\beta - \alpha)]. \quad (39)$$

The following proposition is a corollary of **Theorem 1**.

Proposition 3 (Estimation) For all k in Ω , as N goes to infinity, we have

$$\sqrt{N} \frac{\widehat{A}_e(k) - A_e(k)}{\rho_k} \xrightarrow{d} N(0, 1). \quad (40)$$

Additionally,

$$\sqrt{N} \frac{\widehat{C}_e - C_e}{\rho} \xrightarrow{d} N(0, 1). \quad (41)$$

The proof is given in the appendix.

We next turn to the prediction problem. While $\widehat{A}_e(k)$ is devised as an estimator for $A_e(k)$ and \widehat{C}_e as an estimator for C , $\widehat{A}_e(k)$ can also be regarded as

a predictor for $A(k)$ and \widehat{C}_e as a predictor for C . Define the following variance functions for prediction:

$$\delta_k^2 = \alpha^{-4}[\alpha^2(\alpha_k + k\alpha_k + \beta_k) - \alpha\alpha_k^2] \text{ and} \quad (42)$$

$$\delta^2 = \delta_0^2 = \alpha^{-4}[\alpha^2(\alpha_0 + \beta_0) - \alpha\alpha_0^2]. \quad (43)$$

The following proposition is also a corollary of **Theorem 1**

Proposition 4 (Prediction) *For all k in Ω , as N goes to infinity, we have*

$$\sqrt{N} \frac{A(k) - \widehat{A}_e(k)}{\delta_k} \xrightarrow{d} N(0, 1). \quad (44)$$

Additionally,

$$\sqrt{N} \frac{C - \widehat{C}_e}{\delta} \xrightarrow{d} N(0, 1). \quad (45)$$

The proof is given in the appendix.

One can also estimate ρ_k^2 , ρ^2 , δ_k^2 and δ^2 by using $\hat{f}(x; Q)$. These estimators are denoted by $\hat{\rho}_k^2$, $\hat{\rho}^2$, $\hat{\delta}_k^2$ and $\hat{\delta}^2$ respectively. We can also write the variance estimates as:

$$N^{-1} \hat{\rho}_k^2 = \frac{(k+1)^2 n_{k+1}}{s^4} \{s^2 + n_{k+1}[m - (2k+1)s]\}, \quad (46)$$

$$N^{-1} \hat{\rho}^2 = \frac{n_1}{s^4} [s^2 + n_1(m - 3s)], \quad (47)$$

$$N^{-1} \hat{\delta}_k^2 = \frac{k+1}{s^3} \{s[(k+1)n_{k+1} + (k+2)n_{k+2}] - n_{k+1}^2\} \text{ and} \quad (48)$$

$$N^{-1} \hat{\delta}^2 = \frac{(n_1 + 2n_2)s - n_1^2}{s^3}. \quad (49)$$

We have the consistency results about the variance estimators.

Proposition 5 *For all k in Ω , as N goes to infinity, we have*

$$\hat{\rho}_k^2 \xrightarrow{P} \rho_k^2 \text{ and } \hat{\delta}_k^2 \xrightarrow{P} \delta_k^2. \quad (50)$$

Additionally,

$$\hat{\rho}^2 \xrightarrow{P} \rho^2 \text{ and } \hat{\delta}^2 \xrightarrow{P} \delta^2. \quad (51)$$

The result in (49) was previously given in Esty (1983) in the multinomial model. All the results for the estimation problem and abundance- k prediction problem are new. One assumption in Esty (1983) requires n_1/s converges to some constant, which is clearly guaranteed by the existence of $A_e(0)$. These

results are semi-parametric in the sense that the latent distribution Q is treated as completely unknown.

Proposition 5 paves the way for the construction of asymptotic estimation confidence interval for $A_e(k)$ (and C_e) and asymptotic prediction confidence interval for $A(k)$ (and C).

4. AN APPLICATION

Now we return to the gene categorization problem. A prepared cDNA library typically consists of 10^6 clones. Each clone represents one copy of cDNA from a gene. The cDNA copy numbers of genes, that is, the expression levels, may differ by $10^3 \sim 10^4$ fold. Genes in a library are categorized into abundant and rare ones according to their expression levels. Since the actual gene expression levels are unknown, one strategy is to sequence a sample of clones from a cDNA library. The “single-pass” cDNA sequences are called expressed sequence tags (ESTs). The ESTs are clustered into unique genes. The number of ESTs from a unique gene in the sample is regarded as an indicator of the expression level of that gene. Suppose there are N genes indexed by $1, 2 \dots N$, where N is unknown. Let X_i be the number of ESTs from i th gene. Suppose scientists set a demarcation d and categorize genes according to the following rule:

If $X_i \leq d$, then the i th gene is rare, else it is abundant.

The issue of interest is the total proportions of the rare and abundant genes in the library. Note that all unobserved genes have to be assumed to be rare. When d equals 0, one can apply the sample coverage results directly. Now we consider the general case. Let A^* be the coverage of all rare genes defined by

$$A^* = \sum_{i=1}^N \pi_i I(X_i \leq d). \quad (52)$$

Note that the coverage of all abundant genes is $1 - A^*$. We may write

$$A^* = \sum_{j=0}^d A(j), \quad A_e^* = \sum_{j=0}^d A_e(j) \quad \text{and} \quad \widehat{A}_e^* = \sum_{j=0}^d \widehat{A}_e(j). \quad (53)$$

It is clear from our earlier derivations that A^* is a normal random variable asymptotically. \widehat{A}_e^* is a Turing-type moment estimator for A_e^* and predictor for A^* . Asymptotically, \widehat{A}_e^* and $A^* - \widehat{A}_e^*$ are normal random variables because of the following proposition. First, we define

$$\alpha^* = \sum_{j \leq d} \alpha_j, \quad \beta^* = \sum_{j \leq d} \beta_j, \quad \theta^* = \sum_{j \leq d} r \alpha_j = \beta^* - \beta_d, \quad (54)$$

$$\rho^{*2} = \alpha^{-4} [\alpha^2 (\alpha^* + \theta^*) + (\beta - \alpha) \alpha^{*2} - 2\alpha \alpha^* \theta^*] \quad \text{and} \quad (55)$$

$$\delta^{*2} = \alpha^{-4} [\alpha^2 (\alpha^* + \beta_d) - \alpha \alpha^{*2}]. \quad (56)$$

Our key result for the gene categorization is the following proposition, which is a corollary of another central limit theorem given in the appendix.

Proposition 6 *As N goes to infinity, we have*

$$\sqrt{N} \frac{\widehat{A}_e^* - A_e^*}{\rho^*} \xrightarrow{d} N(0, 1), \quad (57)$$

$$\sqrt{N} \frac{A^* - \widehat{A}_e^*}{\delta^*} \xrightarrow{d} N(0, 1). \quad (58)$$

There are consistent plug-in estimators for ρ^{*2} and δ^{*2} , denoted by $\widehat{\rho}^{*2}$ and $\widehat{\delta}^{*2}$. The unknown constant N is cancelled out in $N^{-1}\widehat{\rho}^{*2}$ and $N^{-1}\widehat{\delta}^{*2}$. The consistency of variance estimators allows us to construct confidence intervals as follows.

The estimation confidence interval for A_e^ and prediction confidence interval for A^* are given by*

$$\widehat{A}_e^* \pm Z_{\gamma/2} \sqrt{N^{-1}\widehat{\rho}^{*2}}, \quad \widehat{A}_e^* \pm Z_{\gamma/2} \sqrt{N^{-1}\widehat{\delta}^{*2}}, \quad (59)$$

respectively, where Z_γ is the γ upper quantile of the standard normal distribution.

One tomato flower cDNA library is studied here. It was made from 0 ~ 3 mm buds of tomato flowers. Totally 2586 ESTs were generated from the library. The nonzero EST abundance frequencies are given in the following table.

j	0	1	2	3	4	5	6	7	8
n_j	?	1434	253	71	33	11	6	2	3
j	9	10	11	12	13	14	16	23	27
n_j	1	2	2	1	1	1	1	1	1

Data source: TIGR Tomato Gene Index. Library Identifier: T1526.

Reference: Quackenbush et al (2000).

©1995-2000 The Institute for Genomic Research. All rights reserved.

The predicted abundance-0 coverage is 0.555 with prediction standard error 0.013. The 95% prediction confidence interval is (0.529, 0.580). That is, an estimated 55% of the clones in the library are from those genes that are not identified by the sample. If we set d to be one, then the predictors for A^* and C^* are 0.750 and 0.250 respectively with prediction standard error 0.012. The 95% prediction confidence intervals for A^* and C^* are (0.726, 0.774) and (0.227, 0.273) respectively.

5. DISCUSSION

The Poisson representation of the coverage problem suggests alternative approaches in which Q is modeled. Inference about the coverage problem becomes

easy when the estimators for the two parameters of interest Q and N are obtained. For example, bootstrap re-sampling can be readily applied. Maximum likelihood estimation can be an alternative approach, for example, see Ord and Whitmore (1986) and Norris and Pollock (1995, 1996, 1998). The authors are investigating a conditional maximum likelihood approach, in which estimation of Q can be separated from N and estimation of N will be done given the estimator for Q . Such a conditional approach removes the confounding of Q and N in (11) and facilitates the analysis and computation. Estimation of Q itself provides direct information about the distribution of gene abundance, and could be used to provide empirical Bayesian estimation of the true abundance λ_i of the i th gene.

The mixture framework also simplifies evaluation of the efficiency of Turing-type estimators. It was shown by Esty (1986) under the multinomial framework that Turing's sample coverage estimator is quite efficient when it is compared to the mle in the case that all classes have the same abundance. The Poisson representation allows us to investigate the efficiency of all Turing-type estimators for the general case. Note that all mean abundance- k coverage or their summations over some index set \mathcal{A} are functionals of the mixture density, which can be expressed in the ratio form as follows:

$$\frac{\sum_{j \in \mathcal{A}} (j+1) f(j+1; Q)}{\sum_{j \in \Omega} (j+1) f(j+1; Q)}. \quad (60)$$

Turing-type estimators simply estimate $f(j; Q)$ by the sample proportion n_j/N . Although N is unknown, it is cancelled out due to the ratio structure of the functionals. Tierney and Lambert (1984) studied the asymptotic efficiency of estimators for functionals of a mixture density. Note that in the non-parametric mixture of Poisson densities, the corollary in Section 3 of Tierney and Lambert (1984) can be applied. So we conclude that any Turing's estimator is asymptotically fully efficient among all estimators that satisfy their regularity conditions. Tierney and Lambert (1984) conjectured that the non-parametric mixture mle is efficient. In their setting, n_0 and hence N are known. So the conjecture does not apply to the mle here. Although the mle should give less variable estimators in finite samples, the simple computation of Turing-type estimators and their confidence intervals make them more attractive than mle for coverage estimation.

ACKNOWLEDGMENTS

Thanks for the Institute for Genomic Research for the tomato EST data and Dr. Claude dePamphilis and Ms. Liying Cui for their help in preparation of and advice on the tomato EST data. Professor Lindsay's research was supported by National Science Foundation grant DMS-9870193.

APPENDIX

First, we calculate the necessary moments of the defined random variables.

$$\begin{aligned}
EY &= (k+1)f(k+1; Q) = \alpha_k. \\
EZ &= \sum_{k \in \Omega} (k+1)f(k+1; Q) = \alpha. \\
EU &= (k+1)f(k+1; Q) = \alpha_k. \\
EV &= \sum_{k \in \Omega} (k+1)f(k+1; Q) = \alpha.
\end{aligned}$$

$$\begin{aligned}
EY^2 &= (k+1)(k+2)f(k+2; Q) = \beta_k. \\
EZ^2 &= \sum_{k \in \Omega} (k+1)(k+2)f(k+2; Q) = \beta. \\
EU^2 &= (k+1)^2 f(k+1; Q) = (k+1)\alpha_k. \\
EV^2 &= \sum_{k \in \Omega} (k+1)^2 f(k+1; Q) = \alpha + \beta. \\
EYZ &= (k+1)(k+2)f(k+2; Q) = \beta_k. \\
EYU &= 0. \\
EYV &= k(k+1)f(k+1; Q) = k\alpha_k. \\
EZU &= (k+1)(k+2)f(k+2; Q) = \beta_k. \\
EZV &= \sum_{k \in \Omega} (k+1)(k+2)f(k+2; Q) = \beta. \\
EUV &= (k+1)^2 f(k+1; Q) = (k+1)\alpha_k.
\end{aligned}$$

Note that the variance of $Q(\lambda)$ is assumed to be finite, which implies α and β are finite. For all k in Ω , α_k and β_k are finite.

Since

$$\forall x \in \Omega, \quad EI(X = x) = f(x; Q) < +\infty, \quad (61)$$

by Khinchine's WLLN, we have consistency of $\hat{f}(x; Q)$ as follows:

$$\hat{f}(x; Q) = \frac{n_x}{N} = \frac{1}{N} \sum_{i=1}^N I(X_i = x) \xrightarrow{P} f(x; Q). \quad (62)$$

The consistency of $\hat{\alpha}_k, \hat{\beta}_k$ follow this directly. Define

$$\bar{Y}^2 = \frac{1}{N} \sum_{j=1}^N Y_j^2 \text{ and } \bar{Z}^2 = \frac{1}{N} \sum_{j=1}^N Z_j^2. \quad (63)$$

Then

$$\hat{\alpha} = \bar{Z} \text{ and } \hat{\beta} = \bar{Z}^2. \quad (64)$$

Also by Khinchine's WLLN, we obtain the consistency of $\hat{\alpha}$ and $\hat{\beta}$.

It is clear now that $\boldsymbol{\mu}$ is the mean vector of \mathbf{W} and Σ is the variance-covariance matrix of \mathbf{W} . By the CLT, we have

$$\sqrt{N}(\bar{\mathbf{W}} - \boldsymbol{\mu}) \xrightarrow{d} N(0, \Sigma). \quad (65)$$

Theorem 1 has been shown. Next we consider the propositions.

$$A_e(k) = \frac{EY}{EZ} = \frac{EU}{EV}. \quad (66)$$

$$A(k) = \frac{\sum_{j=1}^N \Lambda_j I(X_j = k)}{\sum_{j=1}^N \Lambda_j} = \frac{\sum_{j=1}^N Y_j}{\sum_{j=1}^N Z_j} = \frac{\bar{Y}}{\bar{Z}}. \quad (67)$$

$$\hat{A}_e(k) = \frac{(k+1)n_{k+1}}{s} = \frac{\bar{U}}{\bar{V}}. \quad (68)$$

Set

$$\mathbf{R} = \left(\frac{1}{\bar{Z}}, -\frac{EY}{\bar{Z}EZ}, 0, 0 \right)^T, \quad (69)$$

$$\mathbf{S} = \left(0, 0, \frac{1}{\bar{V}}, -\frac{EU}{\bar{V}EV} \right)^T, \quad (70)$$

$$\mathbf{T} = \mathbf{R} - \mathbf{S} = \left(\frac{1}{\bar{Z}}, -\frac{EY}{\bar{Z}EZ}, -\frac{1}{\bar{V}}, \frac{EU}{\bar{V}EV} \right)^T, \quad (71)$$

$$\mathbf{R}_e = \left(\frac{1}{EZ}, -\frac{EY}{(EZ)^2}, 0, 0 \right)^T, \quad (72)$$

$$\mathbf{S}_e = \alpha^{-2}(0, 0, \alpha, -\alpha_k)^T \text{ and} \quad (73)$$

$$\mathbf{T}_e = \mathbf{R}_e - \mathbf{S}_e = \alpha^{-2}(\alpha, -\alpha_k, -\alpha, \alpha_k)^T. \quad (74)$$

Then

$$\sigma_k^2 = \mathbf{R}_e^T \Sigma \mathbf{S}_e, \quad (75)$$

$$\rho_k^2 = \mathbf{S}_e^T \Sigma \mathbf{S}_e, \quad (76)$$

$$\delta_k^2 = \mathbf{T}_e^T \Sigma \mathbf{T}_e, \quad (77)$$

$$\sqrt{N}[A(k) - A_e(k)] = \mathbf{R}^T (\bar{\mathbf{W}} - \boldsymbol{\mu}), \quad (78)$$

$$\sqrt{N}[\hat{A}_e(k) - A_e(k)] = \mathbf{S}^T (\bar{\mathbf{W}} - \boldsymbol{\mu}) \text{ and} \quad (79)$$

$$\sqrt{N}[A(k) - \hat{A}_e(k)] = \mathbf{T}^T (\bar{\mathbf{W}} - \boldsymbol{\mu}). \quad (80)$$

Because \bar{Z} and \bar{V} go to EZ and EV respectively in probability as N goes to infinity, we have

$$\mathbf{R} \xrightarrow{P} \mathbf{R}_e, \quad (81)$$

$$\mathbf{S} \xrightarrow{P} \mathbf{S}_e \text{ and} \quad (82)$$

$$\mathbf{T} \xrightarrow{P} \mathbf{T}_e. \quad (83)$$

Then

$$\sqrt{N}[A(k) - A_e(k)] \xrightarrow{d} N(0, \sigma_k^2), \quad (84)$$

$$\sqrt{N}[\hat{A}_e(k) - A_e(k)] \xrightarrow{d} N(0, \rho_k^2) \text{ and} \quad (85)$$

$$\sqrt{N}[A(k) - \hat{A}_e(k)] \xrightarrow{d} N(0, \delta_k^2). \quad (86)$$

$\hat{A}_e(k)$ is a consistent estimator for $A_e(k)$. The consistency of $\hat{\sigma}_k^2$, $\hat{\rho}_k^2$ and $\hat{\delta}_k^2$ can be easily derived since σ_k^2 , ρ_k^2 and δ_k^2 are functions of $f(x; Q)$, α_k , β_k , α and β .

The results about C , C_e and \hat{C}_e follow the results about $A(0)$, $A_e(0)$ and $\hat{A}_e(0)$ directly.

Finally, we develop another center limit theorem which is used in the application of the gene categorization. For $j = 1, 2, \dots, N$, let

$$Y_j^* = \Lambda_j I(X_j \leq d) \text{ and} \quad (87)$$

$$U_j^* = \sum_{k \leq d} (k+1) I(X_j = k+1). \quad (88)$$

Using Z_j and V_j from (14), (16), set

$$\mathbf{W}_j^* = (Y_j^*, Z_j, U_j^*, V_j)^T. \quad (89)$$

Then let

$$\bar{Y}^* = N^{-1} \sum_{j=1}^N Y_j^* \text{ and } \bar{U}^* = N^{-1} \sum_{j=1}^N U_j^*. \quad (90)$$

Using \bar{Z} and \bar{V} from (19) and (21), set

$$\bar{\mathbf{W}}^* = (\bar{Y}^*, \bar{Z}, \bar{U}^*, \bar{V})^T. \quad (91)$$

Define

$$\boldsymbol{\mu}^* = (\alpha^*, \alpha, \alpha^*, \alpha)^T \text{ and} \quad (92)$$

$$\Sigma^* = \begin{pmatrix} \beta^* & \beta^* & \theta^* & \theta^* \\ \beta^* & \beta & \beta^* & \beta \\ \theta^* & \beta^* & \alpha^* + \theta^* & \alpha^* + \theta^* \\ \theta^* & \beta & \alpha^* + \theta^* & \alpha + \beta \end{pmatrix} - \boldsymbol{\mu}^* \boldsymbol{\mu}^{*T}. \quad (93)$$

Theorem 7 Let \mathbf{W}^* be \mathbf{W}_1^* . The mean vector of \mathbf{W}^* is $\boldsymbol{\mu}^*$ and Σ^* is the variance-covariance matrix for \mathbf{W}^* . As N goes to infinity, we have

$$\sqrt{N}(\bar{\mathbf{W}}^* - \boldsymbol{\mu}^*) \xrightarrow{d} N(0, \Sigma^*). \quad (94)$$

The proof is some simple calculations of moments. Then **Proposition 6** follows directly because

$$\rho^{*2} = (0, 0, \alpha, -\alpha^*)\Sigma^*(0, 0, \alpha, -\alpha^*)^T \text{ and} \quad (95)$$

$$\delta^{*2} = (\alpha, -\alpha^*, -\alpha, \alpha^*)\Sigma^*(\alpha, -\alpha^*, -\alpha, \alpha^*)^T. \quad (96)$$

REFERENCES

- CANTOR, C. R. & SMITH, C. L. (1999). Genomics: the sciences and technology behind the Human Genome Project. New York: Wiley.
- CHAO, A. (1981). On Estimating the Probability of Discovering a New Species. *Ann. Statist.* **9**, 1339–1342.
- CHAO, A. & LEE, S. M. (1992). Estimating the Number of Classes Via Sample Coverage. *J. A. Statist. Assoc.* **87**, 210–217.
- CHAO, A., MA, M. C. & YANG, M. C. K. (1993). Stopping Rules and Estimation for Recapture Debugging With Unequal Failure Rates. *Biometrika* **80**,193–201.
- CLAYTON, K. M. & FREES, W. E. (1987). Nonparametric estimation of the probability of discovering a new species. *J.A.Statist.Assoc.* **82**, 305–311.
- DARROCH, J. N. & RATCLIFF, D. (1980). A Note on Capture-recapture Estimation. *Biometrics* **36**,149–153.
- 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.
- ESTY, W. W. (1986). The efficiency of Good's nonparametric coverage estimator. *Ann. Statist.* **14**, 1257–1260.
- GOOD, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika.* **40**, 237-264.
- GOOD, I. J. & TOULMIN, G. H. (1956). The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika.* **43**,45-63.
- LINDSAY, B. (1995). Mixture Models: Theory, geometry and applications. NSF-CBMS Regional Conference Series in Probability and Statistics. **5**.
- NORRIS, L. J. I. & POLLOCK, K. H. (1995). A Capture-recapture Model With Heterogeneity and Behavioural Response. *Envir. Ecolog. Statist.* **2**,305–313.
- NORRIS, L. J. I. & POLLOCK, K. H. (1996). Nonparametric MLE Under Two Closed Capture-recapture Models With Heterogeneity. *Biometrics.* **52**, 639–649.
- NORRIS, L. J. I. & POLLOCK, K. H. (1998). Non-parametric MLE for Poisson Species Abundance Models Allowing for Heterogeneity Between Species. *Envir. Ecolog. Statist.* **5**, 391–402.
- ORD, J. K. & WHITMORE, G. A. (1986). The Poisson-inverse Gaussian Distribution As a Model for Species Abundance. *Comm. Statist, Part A* **15**, 853–871.
- QUACKENBUSH, J., CHO, J., LEE, D., LIANG, F., HOLT, I., KARAMYCHEVA, S., PARVIZI, B., PERTEA, G., SULTANA, R. & WHITE, J., (2000). The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species. *Nulceic Acids Res.* **29**, 159-164.

- ROBBINS, E. H. (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. Statistics.* **7**, 644-652.
- TIERNEY, L. & LAMBERT, D. (1984). Asymptotic efficiency of estimators of functionals of mixed distributions. *Ann. Statist.* **12**, 1380-1387.