



Moment-based nonparametric estimators for the  
number of classes in a population

By CHANGXUAN MAO and BRUCE G. LINDSAY

Technical Report #01-06-15

2001

---

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

# Moment-based nonparametric estimators for the number of classes in a population

Chang Xuan Mao and Bruce G. Lindsay <sup>1</sup>

Chang Xuan Mao (Corresponding author)  
Graduate student  
Interdepartmental Group in Biostatistics  
University of California, Berkeley  
367 Evans Hall  
Berkeley, CA 94720-3860  
Email:cmao@stat.berkeley.edu

Bruce G. Lindsay  
Distinguished Professor of Statistics  
Department of Statistics  
Pennsylvania State University  
326 Thomas Building  
University Park, PA 16802-2111.

---

<sup>1</sup>Professor Lindsay's research was supported by National Science Foundation grant DMS-980193. The authors thank Drs. Anne Chao, David Hunter, Francesca Chiaromonte, John Bunge and Robert Colwell for their insightful comments.

# Moment-based nonparametric estimators for the number of classes in a population

Chang Xuan Mao and Bruce G. Lindsay

August 27, 2001

## **Abstract**

Suppose a closed population has infinitely many individuals and is partitioned into unknown  $N$  disjoint classes. A random sample is drawn from the population, which may be a Poisson sample or a binomial sample. The issue of interest is the estimation of  $N$ . A new nonparametric method is presented in this paper when the population is heterogeneous, which means that the classes vary in abundance. A simple graphical diagnostic is developed to detect the existence of heterogeneity. An estimator sequence is developed for  $N$ , which is based on moment representation of mixture densities and approximation of the total mass of a measure on the positive half of the real line through its higher moments. A bootstrap confidence inference methodology is provided. As illustrations of the procedure, the number of expressed genes in a cDNA library is estimated from a tomato EST dataset and

the population size of rabbits is estimated from a capture-recapture dataset.

KEY WORDS: Number of species; Population size; Multinomial sample; Capture-recapture; Heterogeneity; Poisson mixture; Binomial mixture; Moment estimator; Mixture identifiability.

## 1 INTRODUCTION

Suppose there are an unknown number  $N$  of distinct classes in a closed population made up of infinitely many individuals. The classes are indexed by  $1, 2, \dots, N$ . A random sample is drawn from the population. The issue of interest is the estimation of  $N$ . Such an estimation problem has interesting practical applications in many different fields, see Bunge and Fitzpatrick (1993). There are two kinds of samples: *Poisson samples* and *binomial samples*. The *multinomial sampling* model is also often used by statisticians. It can be obtained from the Poisson model by conditioning.

Suppose that the number of individuals from each class is available. Let  $x_i$  be the number of individuals from the  $i$ th class that are present in the sample,  $i = 1, 2, \dots, N$ . Each  $x_i$  is assumed to follow a Poisson distribution with mean  $\lambda_i$  and the  $x_i$ 's are independent. The  $x_i$ 's constitute a *Poisson sample*. Let  $s$  be the total number of individuals in the sample, that is,

$$s = \sum_{i=1}^N x_i.$$

Conditioning on  $s$ , the  $x_i$ 's follow a multinomial distribution with

index  $N$  and cell probabilities as follows:

$$\pi_i = \frac{\lambda_i}{\sum_{i=1}^N \lambda_i}, \quad i = 1, 2 \dots N.$$

Thus the Poisson sample becomes a *multinomial sample* by conditioning.

Suppose only presence or absence information for each class is available at  $M$  independent observing occasions. Each class is present or absent independently at each occasion. Let  $y_{ij}$  be 1 if the  $i$ th class is present at the  $j$ th occasion, else  $y_{ij}$  be 0, and let  $\pi_{ij}$  be the probability that the  $i$ th class is present at the the  $j$ th occasion,  $i = 1, 2 \dots N$ ,  $j = 1, 2 \dots M$ . The  $y_{ij}$ 's constitute a *multiple Bernoulli sample*. We further assume that each class has the same probability to be present at all occasions, that is,

$$\pi_{ij} = \pi_i, \quad j = 1, 2 \dots M, i = 1, 2 \dots N.$$

For each  $i$ , let  $\lambda_i$  be the odds of  $\pi_i$ , and  $x_i$  be the number of occasions when the  $i$ th class is present in the sample, that is,

$$\lambda_i = \pi_i / (1 - \pi_i), \quad x_i = \sum_{j=1}^M y_{ij}, \quad i = 1, 2 \dots N.$$

Then  $x_i$  is a binomial random variable with index  $M$  and odds  $\lambda_i$ . The  $x_i$ 's are independent given the  $\lambda_i$ 's and constitute a *binomial sample*.

Under the so-called *homogeneity assumption*, which says that the  $\lambda_i$ 's (or  $\pi_i$ 's) are identical over  $i$ , there is a vast literature concerning the estimation of  $N$ , for example, see Harris (1959) and Darroch and Ratcliff (1980). However, the homogeneity assumption is unlikely to hold in real populations. When the homogeneity assumption is

violated, the  $\lambda_i$ 's (or  $\pi_i$ 's) are often assumed to arise as a random sample from a *latent distribution*. This is often called the *heterogeneous case*. For example, the binomial mixture model was considered by Burnham and Overton (1978, 1979), Chao (1989) and Mingoti and Meeden (1992). The Poisson mixture model was used by Efron and Tibshirani (1976), Ord and Whitmore (1986) and Zelterman (1988). More recently, nonparametric maximum likelihood estimation was investigated by Norris and Pollock (1995, 1996, 1998). There are other approaches dealing with the multinomial sample, such as the nonparametric models in Chao (1984) and Chao and Lee (1992), the parametric models in Kalinin (1965) and McNeil (1973), and the Bayesian models in Lewins and Joanes (1984) and Keener, Rothman and Starr (1987).

The authors are working with biologists on a floral genome project. One interesting and important question is how many genes are expressed in a specific tissue when the expressed sequence tags (ESTs) are generated by sequencing sampled clones in a cDNA library made from the target tissue (Cantor and Smith 1999). In this project, some well-known traditional nonparametric methods were applied, but the simulation results suggested that the systematic bias of most of traditional nonparametric estimators was too large. The large bias is not surprising since gene expression levels are highly differentiated while most of traditional nonparametric estimators work well for the case in which all genes have the same expression level. The authors were thereby motivated to develop new approaches by which the systematic

bias can be reduced greatly.

The mixture models for both the Poisson sample and the binomial sample are introduced in Section 2. The two mixture models share the same mathematical structure, which allows us to develop estimators for the Poisson sample and binomial sample in a new unified framework. A simple log ratio plot is developed in Section 3, and it is shown to provide an easy way to detect heterogeneity. An estimator based on linear extrapolation from the log ratio plot is defined.

The mixture densities can be represented by moments of a measure on the positive half of the real line  $\mathcal{R}^+$ . In Section 4, it will be shown that the total mass of a finite measure on  $\mathcal{R}^+$  can be approximated by a sequence of lower bounds. Each lower bound is a function of the successive higher order moments of the measure. The moment representation of the mixture densities and the mathematical results about approximating the total mass through higher order moments naturally motivate the development of a sequence of estimators for  $N$ . The estimators in the sequence can be regarded as generalizations of the linear extrapolation estimator. The methodology is nonparametric in that the latent distribution is not modeled directly. Properties of confidence inference and bootstrap confidence statements are provided at the end of Section 4.

Two examples are studied in Section 5. The first example concerns a genomic problem: the number of expressed genes in a cDNA library needs to be estimated by a sequenced sample from the library. The second example is the famous live-trapping dataset of cotton-tail

rabbits. Discussion of the properties of the proposed method and its generalizations are presented in Section 6.

## 2 MIXTURE MODELS

We first present some basic facts about the Poisson distribution, the binomial distribution and their mixtures. Let  $\lambda$  be the mean of a Poisson distribution or the odds of a binomial distribution. Both the Poisson and binomial distributions are *power series distributions*, which mathematically can be represented as densities with respect to some  $\sigma$ -finite measure  $d\omega(x)$  on the set of nonnegative integers in the following form,

$$f(x; \lambda) = c(\lambda)\lambda^x, \quad x \in \Omega, \quad \lambda \in \mathcal{R}^+.$$

Here,  $\Omega$  is the *sample space* and  $c(\lambda)$  is the *normalization function*. Let  $I(E)$  be the indicator for the event  $E$ . The Poisson and binomial representations arise as follows:

$$\Omega = \begin{cases} \{0, 1, \dots\} & \text{(Poisson)} \\ \{0, 1, \dots, M\} & \text{(binomial)} \end{cases},$$

$$c(\lambda) = \left[ \int \lambda^x d\omega(x) \right]^{-1} = \begin{cases} e^{-\lambda} & \text{(Poisson)} \\ (1 + \lambda)^{-M} & \text{(binomial)} \end{cases},$$

$$d\omega(x) = \begin{cases} \frac{I(x \in \Omega)}{x!} & \text{(Poisson)} \\ \frac{M! I(x \in \Omega)}{(M-x)! x!} & \text{(binomial)} \end{cases}.$$

Let  $h(x)$  be  $d\omega(x)$ . Then

$$f(x; \lambda) = c(\lambda)h(x)\lambda^x, \tag{1}$$

which is the density with respect to counting measure on the set of nonnegative integers.

To create a mixture model, we assume that  $\lambda$  arises from a latent distribution  $Q(\lambda)$  and that  $x$  has a Poisson or binomial distribution conditioning on  $\lambda$ . The conditional density  $f(x; \lambda)$  is most naturally called a *component density*. The marginal density of  $x$ , called the *mixture distribution*, is denoted by  $f(x; Q)$ . It can be written as

$$f(x; Q) = \int f(x; \lambda) dQ(\lambda), \quad \forall x \in \Omega. \quad (2)$$

We next form mixture models for both the Poisson sample and the binomial sample. Let  $x_i$  be called the *frequency* of the  $i$ th class in the sample. The  $x_i$ 's are assumed to arise as a random sample from the mixture density  $f(x; Q)$ . Let

$$n_j = \sum_{i=1}^N I(x_i = j), \quad j \in \Omega, \quad \text{and} \quad n = \sum_{\{j>0\}} n_j.$$

Here,  $n_j$  is the number of classes that occur with frequency  $j$  in the sample. The  $n_j$ 's are called *frequency counts*. Note that  $n_0$  is not observed and  $n$  is the number of distinct classes observed in the sample. The joint density of the frequencies  $x_i$ 's can be written as

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

Therefore, the joint density of the frequency counts  $\{n_j : j \in \Omega\}$  is given by

$$\frac{N!}{\prod_{j \in \Omega} n_j!} \prod_{j \in \Omega} f(j; Q)^{n_j}. \quad (3)$$

That is,  $\{n_j : j \in \Omega\}$  has a multinomial distribution with index  $N$  and cell probabilities  $f(j; Q)$ 's. In ( 3 ), we have given the “complete” data likelihood for  $\{n_j : j \in \Omega\}$ . Since  $n_0$  is not observed, the observed data likelihood is

$$\frac{N!}{(N-n)! \prod_{j \in \Omega_0} n_j!} [f(0; Q)]^{N-n_0} \prod_{j \in \Omega_0} f(j; Q)^{n_j}, \quad (4)$$

where  $\Omega_0 = \Omega - \{0\}$ . Note that the observed likelihood in ( 4 ) can be further factored into a binomial density for  $n$  times a multinomial density for  $\{n_j : j \in \Omega_0\}$  given  $n$ . The density of  $\{n_j : j \in \Omega_0\}$  given  $n$  can be written as

$$\frac{n!}{\prod_{j \in \Omega_0} n_j!} \prod_{j \in \Omega_0} \left[ \frac{f(j; Q)}{1 - f(0; Q)} \right]^{n_j}. \quad (5)$$

The density of  $n$  is given by

$$\frac{N!}{n!(N-n)!} f(0; Q)^{N-n} [1 - f(0; Q)]^n. \quad (6)$$

Comparing ( 5 ) and ( 6 ), we can see that ( 6 ) contains all the information about  $N$  in the sample.

Because  $N = n + n_0$ , the problem of inferring  $N$  can be treated as one of inferring the latent variable  $n_0$  via the set of observed variables  $\{n_j : j \in \Omega_0\}$ . Note that  $En_0 = Nf(0; Q)$ . So one could also think of the problem as one of estimating the parameter  $Nf(0; Q)$  and adding it to  $n$ , which estimates  $N(1 - f(0; Q))$ .

In this paper, we will develop a graphical diagnostic for heterogeneity and an estimation procedure for  $N$ , which do not require direct modeling of the latent distribution  $Q$ .

### 3 HETEROGENEITY DETECTION

Graphical diagnostics are often used in the literature of mixture models, see Titterington, Smith and Makov (1985), Lindsay and Roeder (1992) and Lindsay (1995). They are easy to use and have good power for detecting heterogeneity. A simple log ratio plot is developed in this section. It is also used to define an estimator for  $N$  by linear extrapolation.

#### 3.1 The log ratio plot

In the two mixture models constructed in Section 2, the component densities form an exponential family. For any fixed  $\lambda_0$ , define the *ratio function* by

$$R(x) = f(x; Q)/f(x; \lambda_0) \text{ where } \lambda_0 \in \mathcal{R}^+.$$

**Proposition 1** *The function  $\log R(x)$  is convex in  $x$  and it is linear if and only if  $Q$  is degenerate.*

Although the proof is simple, readers are referred to Lindsay and Roeder (1992) and Lindsay (1995).

To adapt this plot to the case in which  $n_0$  and  $N$  are unknown, we define a translated version of  $\log R(x)$  as follows,

$$H(x) = \log[c(\lambda_0)NR(x)] = \log Nf(x; Q) - \log h(x) - x \log \lambda_0. \quad (7)$$

Note the following result can be easily obtained from **Proposition 1**.

**Proposition 2** *The log ratio function  $H(x)$  is convex in  $x$  and it is linear if and only if  $Q$  is degenerate.*

We next develop an empirical version of  $H(x)$ . For every  $x$ ,  $Nf(x; Q)$  has an unbiased estimator  $n_x$ , which can be seen from (3). Plugging this estimator into (7) yields an estimator  $\hat{H}(x)$  for  $H(x)$  of the form:

$$\hat{H}(x) = \log n_x - \log h(x) - x \log \lambda_0.$$

We will leave  $\hat{H}(x)$  undefined if  $n_x = 0$ . The plot of  $\{(x, \hat{H}(x)) : x \in \Omega_0\}$ , called the *log ratio plot*, therefore has the following interpretation: when  $N$  is sufficiently large, the log ratio plot should be approximately linear in the homogeneous case and should be strictly convex in the heterogeneous case. In the latter case, methods based on the homogeneity assumption will tend to be conservative. If neither linearity nor strict convexity shows, the model itself should be questioned. Examples of these plots are given in Section 5.

Finally, note that since  $n_0$  is not observed, and  $N = n + n_0$ , the estimation of  $N$  can be viewed as an extrapolation problem, where we extrapolate the value of  $\hat{H}(0)$  from the set of values  $\{\hat{H}(x) : x \in \Omega_0\}$ .

### 3.2 Linear extrapolation

We present a simple extrapolation in this sub-section. Since  $H(x)$  is convex in  $x$ , the point  $(0, H(0))$  should be above the straight line through the points  $(1, H(1))$  and  $(2, H(2))$  when  $Q$  is not degenerate, or on the line if  $Q$  is degenerate. Suppose the line is  $ax + b$  with a slope  $a$  and an intercept  $b$ , then we can obtain  $a$  and  $b$  via the equations:

$$H(1) = a + b, \quad H(2) = 2a + b.$$

So

$$H(0) \geq 2H(1) - H(2). \quad (8)$$

When we replace  $H(x)$  with  $\hat{H}(x)$ , an empirical version of the inequality ( 8 ) is obtained as follows:

$$\hat{H}(0) \geq 2\hat{H}(1) - \hat{H}(2),$$

which yields,

$$n_0 = n_0/h(0) \geq \frac{h(2)n_1^2}{[h(1)]^2n_2}, \quad (9)$$

since  $h(0) = 1$  in both models. The right-hand side of the inequality ( 9 ) can be regarded as a lower bound predictor for  $n_0$ . It arises from a linear extrapolation of the log ratio plot. We therefore define a lower bound estimator  $\hat{N}_{linear}$  for  $N$  by

$$\hat{N}_{linear} = n + \frac{h(2)n_1^2}{[h(1)]^2n_2} = \begin{cases} n + \frac{n_1^2}{2n_2} & \text{(Poisson)} \\ n + \frac{(M-1)n_1^2}{2Mn_2} & \text{(binomial)} \end{cases}.$$

The estimator  $\hat{N}_{linear}$  for a Poisson sample was obtained in Chao (1984) and  $\hat{N}_{linear}$  for a binomial sample was obtained in Chao (1989). The linear extrapolation is equivalent to the theory behind Chao (1989). In a homogeneity model,  $\hat{N}_{linear}$  is tight for  $N$ . In a heterogeneous population, the estimator  $\hat{N}_{linear}$  is only a lower bound for  $N$ . Our next task is to create more sophisticated extrapolation methods.

## 4 THE ESTIMATION PROCEDURE

In the Poisson or binomial model, the sample space is a lattice (finite in the binomial case and infinite in the Poisson case). Such a structure implies that the Poisson mixture and the binomial mixture densities have moment representations.

### 4.1 Moment representation

Considering the mixture models introduced in Section 2, we first define a new measure  $\nu$  by re-weighting the latent distribution  $Q$ . Let

$$d\nu(\lambda) = c(\lambda)dQ(\lambda). \tag{10}$$

We can invert the relationship in ( 10 ) to obtain  $Q$  from  $\nu$  as well.

$$dQ(\lambda) = [c(\lambda)]^{-1}d\nu(\lambda). \tag{11}$$

Thus the re-weighting is a bijection. Note that the re-weighted measure  $\nu$  and the latent distribution  $Q$  have the same support points.

Let  $\mu(x)$  be the  $x$ th moment of  $\nu$ , that is,

$$\mu(x) = \int \lambda^x d\nu(\lambda), \quad x = 0, 1, \dots$$

From ( 1 ) and ( 2 ), the mixture density with respect to counting measure can be represented in terms of the moment sequence  $\mathcal{M} = \{\mu(x)\}_{x \in \Omega}$  of  $\nu$ :

$$f(x; Q) = h(x)\mu(x), \quad x \in \Omega,$$

which yields the moment sequence of  $\nu$  as

$$\mu(x) = f(x; Q)/h(x), \quad x \in \Omega. \quad (12)$$

Note that the Poisson model generates an infinite moment sequence but in the binomial model the moment sequence terminates at  $\mu(M)$ .

## 4.2 Measure and its moments

The moment structure of the mixture density can be exploited to characterize  $\mu(0)$ , and hence  $f(0; Q)$ , in terms of the remaining higher moments  $\mu(x)$ ,  $x \in \Omega_0$ , thus providing a more general extrapolation tool. To do so, we will develop some results about a measure and its moment sequence.

Let  $\mathcal{M}$  be any real number sequence  $\{\mu(x)\}_{x=0}^{+\infty}$ . For any nonnegative integer  $p$ , the *Hankelian matrices* of  $\mathcal{M}$ ,  $H_p(\mathcal{M})$  and  $\bar{H}_p(\mathcal{M})$ , are defined as follows:

$$H_p(\mathcal{M}) = \begin{bmatrix} \mu(0) & \mu(1) & \cdots & \mu(p) \\ \mu(1) & \mu(2) & \cdots & \mu(p+1) \\ \vdots & \vdots & \ddots & \vdots \\ \mu(p) & \mu(p+1) & \cdots & \mu(2p) \end{bmatrix},$$

$$\bar{H}_p(\mathcal{M}) = \begin{bmatrix} \mu(1) & \mu(2) & \cdots & \mu(p+1) \\ \mu(2) & \mu(3) & \cdots & \mu(p+2) \\ \vdots & \vdots & \ddots & \vdots \\ \mu(p+1) & \mu(p+2) & \cdots & \mu(2p+1) \end{bmatrix}.$$

The corresponding determinants are called *Hankelian determinants*.

The following two theorems will play a significant role in our procedure.

**Theorem 3** (*Stieltjes Moment Theorem*) *A sequence  $\mathcal{M}$  is a moment sequence of some finite measure  $\nu(\lambda)$  on  $\mathcal{R}^+$  with infinitely many support points if and only if the Hankelian determinants  $|H_p(\mathcal{M})|$  and  $|\bar{H}_p(\mathcal{M})|$  are positive for each nonnegative integer  $p$ .*

**Theorem 4** *Let  $T$  be a natural number. A sequence  $\mathcal{M}$  is a moment sequence of some finite measure on  $\mathcal{R}^+$  with exactly  $T$  support points if and only if  $|H_p(\mathcal{M})|$  and  $|\bar{H}_p(\mathcal{M})|$  are positive for each nonnegative integer when  $p < T$ , and  $\text{rank}(H_p(\mathcal{M})) = T$  and  $\text{rank}(\bar{H}_p(\mathcal{M})) = T$  for each  $p$  when  $p \geq T$ .*

These theorems are results in the Stieltjes moment problem, which is a special case of the Hamburger moment problem. Readers are referred to Widder (1941), Lindsay (1989) and Gyires (1998). Lindsay (1989, 1995) addressed the moment problem in mixture models. For more detailed description about the moments, Hankelian matrices and their properties, see Karlin (1968) and Dette and Studden (1997).

Let  $\nu(\lambda)$  be a finite positive measure on  $\mathcal{R}^+$  with  $T$  (possibly infinite) support points. Let  $\mu(x)$  be the  $x$ th moment of  $\nu(\lambda)$  and let  $\mathcal{M}$  be the moment sequence of  $\nu$ . Let  $\mathcal{M}(y)$  be the moment sequence where all terms are identical to terms in  $\mathcal{M}$  but the first term is replaced by an unknown real number  $y$ . Let  $H_p$  and  $\bar{H}_p$  be the Hankelian matrices of  $\mathcal{M}$  and  $H_p(y)$  and  $\bar{H}_p(y)$  be the Hankelian matrices of  $\mathcal{M}(y)$ . Let  $\Delta_p$  be the set of real numbers  $y$  where  $|H_p(y)|$  is nonnegative and let

$\mu_p$  be the infimum of the set  $\Delta_p$ . That is,

$$\begin{aligned}\mathcal{M}(y) &= \{y, \mu(1), \mu(2), \dots\} \text{ and,} \\ H_p &= H_p(\mathcal{M}), \quad \bar{H}_p = \bar{H}_p(\mathcal{M}), \\ H_p(y) &= H_p(\mathcal{M}(y)), \quad \bar{H}_p(y) = \bar{H}_p(\mathcal{M}(y)), \\ \Delta_p &= \{y \in \mathcal{R} : |H_p(y)| \geq 0\} \\ \mu_p &= \inf \Delta_p.\end{aligned}$$

The determinant  $|H_p(y)|$  is linear in  $y$ . If the coefficient of  $y$  is not zero, then it is clear that  $\mu_p$  is the single root of the following equation:

$$|H_p(y)| = 0.$$

We will say that  $\mu_p$  is the *p-th order lower bound* for  $\mu(0)$ . This terminology is justified by the following results.

**Proposition 5** *For all  $p$  in  $\mathcal{N}$ ,  $\mu(0)$  is in  $\Delta_p$  and so  $\mu_p \leq \mu(0)$ .*

The proposition comes from definitions and **Theorem 3** directly.

The elements of  $\Delta_p$  are further described by the following proposition.

**Proposition 6** *If  $y$  is in  $\Delta_p$ , then it is nonnegative. Also,  $\sup \Delta_p = +\infty$ .*

The proof is given in the appendix. This proposition implies that there does not exist a corresponding upper bound theory for  $\mu(0)$ .

The next proposition shows that the lower bounds are monotonically increasing. Define two sub-matrices of the Hankelian matrices

$H_p(\mathcal{M})$  of a real number sequence  $\mathcal{M}$  as follows:

$$\mathbf{b}_p(\mathcal{M}) = (\mu(1), \mu(2), \dots, \mu(p))' \text{ and}$$

$$A_p(\mathcal{M}) = \begin{bmatrix} \mu(2) & \mu(3) & \cdots & \mu(p+1) \\ \mu(3) & \mu(4) & \cdots & \mu(p+2) \\ \vdots & \vdots & \ddots & \vdots \\ \mu(p+1) & \mu(p+2) & \cdots & \mu(2p) \end{bmatrix}.$$

For notational simplicity, set

$$A_p = A_p(\mathcal{M}), \quad \mathbf{b}_p = \mathbf{b}_p(\mathcal{M}),$$

where  $\mathcal{M}$  is the moment sequence of  $\nu$ .

**Proposition 7** *Let  $\mu_0 = 0$ . For each natural number  $p$ , we have*

$$\begin{aligned} \mu_p &= \mathbf{b}_p' [A_p]^{-1} \mathbf{b}_p \text{ and} \\ \mu_p &= \mu_{p-1} + \beta_{p-1}, \end{aligned} \tag{13}$$

where

$$\beta_0 = \mu_1, \quad \beta_{p-1} = \frac{|\bar{H}_{p-1}|^2}{|A_{p-1}| |A_p|} > 0, \quad p \geq 2. \tag{14}$$

When  $T$  is finite, we only consider  $p \leq T$ .

The proof is given in the appendix. We will call  $\beta_p$  the  $p$ th order *bias-correction*.

When the measure  $\nu$  has finitely many support points, the largest lower bound is identical to the total mass, that is

$$\mu_T = \mu(0),$$

because  $|H_T| = 0$ , which can be seen from **Theorem 4**.

**Remark 1** *Note that for the binomial sample, the sequence  $\{\mu(x)\}_{x=0}^M$  terminates at  $\mu(M)$ . The lower bound sequence therefore terminates at  $\mu_{[M/2]}$ , where  $[M/2]$  represents the largest integer no greater than  $M/2$ . Then for the binomial sample, if the latent distribution has more than  $[M/2]$  support points, then there is a positive difference between the total mass  $\mu(0)$  and the largest lower bound  $\mu_{[M/2]}$ . If  $T$  is no greater than  $[M/2]$ , then  $\mu_T$  equals  $\mu(0)$ .*

The sequence  $\{\mu_p\}_{p=0}^{+\infty}$  is strictly increasing and is bounded above by  $\mu(0)$  when  $T$  is infinite. Therefore, it converges to some real number no greater than  $\mu(0)$ , denoted by  $\mu_\infty$ , that is,

$$\mu_\infty = \lim_{p \rightarrow +\infty} \mu_p \leq \mu(0).$$

The following proposition is about the sequence  $\mathcal{M}(\mu_\infty)$ , which is derived from the sequence  $\mathcal{M}$  with  $\mu(0)$  being replaced by  $\mu_\infty$ .

**Proposition 8** *The sequence  $\mathcal{M}(\mu_\infty)$  is a a moment sequence of some finite measure on  $\mathcal{R}^+$  with infinitely many support points.*

The proof is given in the appendix.

The next proposition gives a sufficient condition for  $\mu_\infty$  to equal  $\mu(0)$ .

**Proposition 9** *If the moment generating function of  $\nu$  exists, then*

$$\mu_\infty = \mu(0).$$

The proposition is proved in the appendix. Note that  $\mu_T$  equals  $\mu(0)$  whenever  $T$  is finite or infinite.

**Remark 2** *For the Poisson sample, if the latent distribution  $Q$  has finitely many support points, then the total mass  $\mu(0)$  is always identical to the largest lower bound. When the latent distribution  $Q$  has infinitely many support points, then  $\mu_\infty$  equals  $\mu(0)$  because the moment generating function of  $\nu$  exists. In fact, we have*

$$\forall t < 1, \quad \int e^{t\lambda} d\nu(\lambda) = \int e^{(t-1)\lambda} dQ(\lambda) < +\infty.$$

We can also represent the difference between  $\mu_T$  and  $\mu_p$  in terms of the bias-corrections as follows:

**Proposition 10** *Let  $\beta_T = 0$ . Then*

$$\mu_T - \mu_p = \sum_{k=p}^T \beta_k, \quad p = 0, 1, \dots, T.$$

The proof is simple. Due to **Proposition 7**, we have

$$\mu_p + \sum_{k=p}^m \beta_k = \mu_{m+1}, \quad \forall m \geq p.$$

The proposition is established by letting  $m$  go to  $T$ .

Note that when  $T$  is infinite, the last proposition yields two simple results as follows:

$$\lim_{p \rightarrow +\infty} \beta_p = 0, \quad \lim_{p \rightarrow +\infty} \sum_{k=p}^{\infty} \beta_k = 0.$$

Hence, the bias-corrections must go to zero as  $p$  increases.

**Remark 3** *In the preceding theory, it is important that  $\nu$  is allowed no mass at zero, as otherwise we could add arbitrary mass there without changing any of the moments except for  $\mu(0)$ . Thus the total mass can not be characterized by the set of higher moments  $\{\mu(x) : x \in \Omega_0\}$ .*

### 4.3 The estimator sequence

We next create an empirical version of the moment sequence. In both the Poisson sample and the binomial sample, all but a finite number of  $n_x$ 's are zero. In a typical sample of interest, the nonzero  $n_x$ 's correspond to values of  $x$  near zero. In fact, if we had, for example,  $n_1 = n_2 = 0$ , then we would have high confidence in predicting that  $n_0 = 0$ .

Define  $S$  to be the largest value of  $p$  such that

$$n_x \neq 0, \quad x = 1, 2, \dots, p.$$

Let

$$\hat{\mu}(x) = \frac{n_x}{N h(x)}, \quad x = 0, 1, \dots, S. \quad (15)$$

Note that going from in ( 12 ) to ( 15 ), the mixture density  $f(x; Q)$  has been replaced by its “estimator”  $n_x/N$ . The term “estimator” here is somewhat unusual because there is an unknown constant in the expression.

**Remark 4** *Since the frequency counts are used as moment estimators, the estimation becomes less reliable when the frequency counts are too small. In ( 15 ), we only require that  $n_x > 0$ , which is the*

weakest requirement because we cannot use zero to estimate a positive quantity. Additionally, we may require that  $n_x$  is larger than some positive integer, say 5, in order to have more reliable moment estimators.

The moment estimator  $\hat{\mu}(x)$  is an empirical version of  $\mu(x)$ , and is consistent because of the law of large numbers:

$$\hat{\mu}(x) \xrightarrow{P} \mu(x) \text{ as } N \longrightarrow +\infty.$$

We can now construct an empirical version of the moment sequence  $\mathcal{M}$  by defining

$$\widehat{\mathcal{M}} = \{\hat{\mu}(0), \hat{\mu}(1), \dots, \hat{\mu}(S)\},$$

as well as the empirical moment matrices:

$$\hat{\mathbf{b}}_p = \mathbf{b}_p(\widehat{\mathcal{M}}) \text{ and } \widehat{A}_p = A_p(\widehat{\mathcal{M}}).$$

When  $A_p$  is positive definite, we have

$$P\{\widehat{A}_p \text{ is positive definite}\} \longrightarrow 1,$$

as  $N$  goes to infinity.

When  $\widehat{A}_p$  is positive definite, based on ( 13 ), we define

$$\hat{\mu}_p = \hat{\mathbf{b}}_p^T \widehat{A}_p^{-1} \hat{\mathbf{b}}_p,$$

which can be regarded an empirical version of  $\mu_p$ , and a lower bound estimator for  $\mu(0)$ . That is, asymptotically

$$\hat{\mu}_p \xrightarrow{P} \mu_p \leq \mu(0). \tag{16}$$

From ( 16 ), it is reasonable to regard  $N\hat{\mu}_p$  as a lower bound predictor for  $n_0$ . Note that from ( 15 )  $N$  is cancelled out in  $N\hat{\mu}_p$ .

Let  $\hat{\mu}_0 = 0$  and define

$$\hat{N}_p = n + N\hat{\mu}_p. \quad (17)$$

We call  $\hat{N}_p$  the *p*th order moment estimator for  $N$ . The 0th order estimator is the naive estimator  $n$ . The estimators can be regarded as generalizations of the linear extrapolation estimator because

$$N\hat{\mu}_1 = \frac{h(2)n_1^2}{[h(1)]^2n_2} \text{ and } \hat{N}_1 = n + N\hat{\mu}_1 = \hat{N}_{linear}.$$

Although we have developed estimators for  $N$  directly, one could also develop the methods from the likelihood using moment-based nuisance parameter estimation. There are two parameters involved in the procedure: the parameter of interest  $N$  and the nuisance parameter  $f(0;Q)$ . Let  $\theta$  be the corresponding odds, that is,  $\theta = f(0;Q)/[1 - f(0;Q)]$ . The density of  $n$  given in ( 6 ) can be expressed in terms of  $\theta$ , denoted by  $L(N;\theta)$  and written as

$$L(N;\theta) = \frac{N!}{n!(N-n)!} \left( \frac{\theta}{\theta+1} \right)^{N-n} \left( \frac{1}{\theta+1} \right)^n.$$

When  $\theta$  is known,  $L(N;\theta)$  can be used as a likelihood to infer  $N$ . From Lindsay and Roeder (1987), the maximum likelihood estimator for  $N$ , is given by

$$\hat{N} = n + n\theta. \quad (18)$$

Comparing ( 17 ) and ( 18 ), we see that the sequence  $\{N\hat{\mu}_p/n : p = 1, 2, \dots, [S/2]\}$  can be regarded as an estimator sequence for  $\theta$ . If we

substitute an estimator  $\hat{\theta}$  for  $\theta$  into ( 18 ), we obtain an estimator for  $N$  of the type called pseudo-likelihood estimators by Gong and Samaniego (1981). Note that since  $N$  is an integer, technically, we should use  $[\hat{N}_p]$  as an estimator for  $N$  (Lindsay and Roeder 1987).

**Remark 5** *It should also be noted that even though one could estimate lower bounds for  $\theta$  by other methods, one cannot improve on the asymptotic efficiency of the moment method in the nonparametric setting. See Mao and Lindsay (2001) for more on this point.*

**Remark 6** *If  $T$  is infinite, then no consistent nonparametric estimator for  $N$  can be constructed via just a finite number of the  $n_x$ 's. And in a Poisson sample, it is difficult to use infinitely many  $n_x$ 's because all but a finitely many are zero at any fixed sample size.*

**Remark 7** *If  $T$  is finite, then  $\hat{\mu}_T$  is consistent for  $\mu_T$  and hence for  $\mu(0)$ . Thus  $\hat{N}_T$  is tight for a latent distribution with  $T$  or fewer support points. Otherwise, it displays a systematic bias.*

**Remark 8** *When  $T$  is finite and  $S \geq 2T$ , we can construct a moment estimator for the latent distribution  $Q$ . First, we estimate the moments of  $\nu$  as follows:*

$$\tilde{\mu}(0) = \frac{N \hat{\mu}_T}{\hat{N}_T}, \quad \tilde{\mu}(x) = \frac{n_x}{\hat{N}_T h(x)}, \quad x = 1, 2 \dots T.$$

*Then we construct a moment-based estimator  $\hat{\nu}$  for the re-weighted measure  $\nu$ . Lindsay (1989) provides a detailed discussion of method of moment estimation of a latent distribution. Finally, we can obtain  $\hat{Q}$  by re-weighting  $\hat{\nu}$  through ( 11 ). From this point of view, one could*

also say that  $\widehat{N}_p$  is the moment estimator of  $N$  based on a  $p$ -point mixture model.

**Remark 9** *The determination of the number of support points  $T$  is a one-sided inference problem in the sense that one can not construct a two-sided confidence interval for the number of support points given a sample of finite size but one can construct a lower confidence limit, see Donoho (1988).*

#### 4.4 Fractional undercount

It is clear from ( 17 ) that the moment-based estimator  $\widehat{N}_p$  estimates  $N_p$ , where  $N_p = N(1 - f(0; Q)) + N\mu_p$  and so the ratio

$$1 - \frac{N_p}{N} = f(0; Q) - \mu_p$$

represents the *fractional undercount*. That is, if the fractional undercount is 0.25, we are estimating  $0.75N$  instead of  $N$ . Another way to measure the undercount is to ask what fraction of the unseen classes,  $n_0 \approx Nf(0; Q)$ , will never be counted. This can be assessed by the *fractional unseen undercount*,  $1 - \mu_p/f(0; Q)$ .

The only time that the undercounts will be zero is if the true latent distribution  $Q$  has  $T$  support points and  $p \geq T$ . In order to demonstrate the potential magnitude of the fractional undercounts in continuous mixtures, we consider some classic mixture models.

We start with the gamma-mixed Poisson distribution, which results in the negative binomial distribution. That is,  $f(x; \lambda)$  is Poisson

( $\lambda$ ) and  $\lambda \sim \text{gamma}(\alpha, \beta)$ . The mixture density is

$$f(x; Q) = \binom{\alpha + x - 1}{\alpha - 1} \left( \frac{\beta}{\beta + 1} \right)^x \frac{1}{(\beta + 1)^\alpha}, \quad \alpha > 0, \quad \beta > 0.$$

For  $\alpha = 1$  and  $\beta = 0.0526$ , we have  $f(0; Q) = 0.95$ . This represents a situation where 95% of the classes are not observed. The fractional undercounts for  $p = 1$  to  $p = 9$  are 47.5%, 31.7%, 23.7%, 19.0%, 15.8%, 13.6%, 11.9%, 10.6% and 9.5%. Thus we cannot reduce undercount to 10% without using  $p = 9$ . On the other hand, for  $\alpha = 2$  and  $\beta = 1.7778$ , we have  $f(0; Q) = 0.36$ , so we expect to see 64% of the classes. In this case, the fractional undercounts are 12.0%, 6.0%, 3.6%, 2.4%, 1.7%, 1.3% and 1.0% for  $p = 1$  to  $p = 7$ . Here undercounting is not such a severe problem.

As a second example, we consider the beta-binomial family, in which  $f(x; \lambda)$  is  $\text{binomial}(M, \lambda/(1+\lambda))$  and  $\pi = \lambda/(1+\lambda) \sim \text{beta}(\alpha, \beta)$ . The mixture density is given by

$$f(x; Q) = \frac{M!}{x!(M-x)!} \frac{(\alpha + \beta)^M}{(M + \alpha + \beta)^M} \frac{\binom{M-x+\beta}{\alpha} \binom{M-x+\beta}{\beta}}{\binom{M-x+\beta}{\alpha} \binom{M-x+\beta}{\beta}}, \quad \alpha > 0, \quad \beta > 0.$$

If we set  $M = 5$ ,  $\alpha = 0.1$  and  $\beta = 0.1$ , then  $f(0; Q) = 41.4\%$ . We then have the fractional undercounts 38.5%, 37.9% for  $p = 1$  and  $p = 2$  respectively. In this case, the fractional unseen undercounts are 93.1% and 91.6%, indicating that we have an extreme poor estimate of the number of unseen classes. This case is rather extreme in that the beta distribution puts a great deal of mass on the classes with a small probability of being viewed because the distribution on  $\pi$  is U-shaped.

A more realistic setting might give mean detectability  $\alpha/(\alpha + \beta) = 0.25$ , with standard deviation  $\{\alpha\beta/[(\alpha + \beta)^2(\alpha + \beta + 1)]\}^{1/2} = 0.18$

to the distribution on  $\pi$ . This results in  $\alpha = 1.2$  and  $\beta = 3.6$ , and a unimodal distribution, which is skewed to the right. In this case, for  $M = 5$ , we have  $f(0; Q) = 5.7\%$  so that a large fraction of classes will be seen and the fractional undercounts are 2.3% and 1.6% for  $p = 1$  and  $p = 2$  respectively.

## 4.5 Confidence inference and bootstrap

The confidence inference for the estimator  $\widehat{N}_p$  requires the description of its distribution. Let

$$\mathbf{f}_p = (f(0; Q), f(1; Q) \dots, f(2p; Q))'$$

and

$$\begin{aligned} \Sigma_p &= \text{diag}(\mathbf{f}_p) - \mathbf{f}_p \mathbf{f}_p', & \xi_p &= \xi_p(\mathbf{f}_p) = 1 - f(0; Q) + \mu_p, \\ v_p &= \frac{\partial \xi_p(\mathbf{f}_p)}{\partial \mathbf{f}_p}, & w_p &= \frac{\partial \log \xi_p(\mathbf{f}_p)}{\partial \mathbf{f}_p}, \\ \sigma_p^2 &= v_p' \Sigma_p v_p, & \delta_p^2 &= w_p' \Sigma_p w_p. \end{aligned}$$

Then we have

$$(N \sigma_p^2)^{-1/2} (\widehat{N}_p - N_p) \xrightarrow{d} N(0, 1), \quad (19)$$

and

$$(N / \delta_p^2)^{1/2} \log(\widehat{N}_p / N_p) \xrightarrow{d} N(0, 1) \quad (20)$$

as  $N$  goes to infinity. Similar central limit theorems hold for all moment-based nonparametric estimators, see Burnham and Overton (1978, 1979).

However, it is difficult to express  $\sigma_p^2$  and  $\delta_p^2$  explicitly for each  $p$  except for  $p = 1$ . Thus a bootstrap re-sampling procedure is presented here to obtain confidence statements. Since the frequency counts  $n_j$ 's arise from a multinomial distribution with index  $N$  and cell probabilities  $f(x; Q)$ ,  $x \in \Omega$ , and the estimator sequence only depends on the observed frequency counts, it is reasonable to re-sample from the multinomial distribution. Let  $\hat{N}$  be an estimator for the index  $N$ . The simple estimators for the cell probabilities are given by

$$\hat{f}(x; Q) = \frac{n_x}{\hat{N}}, \quad x \in \Omega.$$

We may use a collapsed multinomial distribution by combining the higher order frequency counts. For example, since we only use the first  $S + 1$  frequency counts, we may bootstrap from the following multinomial distribution with index  $\hat{N}$  and probabilities.

$$\hat{f}(x; Q) = \frac{n_x}{\hat{N}}, \quad x = 0, 1, \dots, S, \quad \hat{f}(S+; Q) = \hat{N}^{-1} \sum_{\{x>S\}} n_x.$$

The bootstrap procedure has two steps.

- (1). Generate a random vector  $(n_0^*, n_1^*, n_2^*, \dots, n_{S+}^*)'$  from the estimated multinomial distribution and compute  $\hat{N}^*$  from  $n_1^*, n_2^*, \dots, n_{S+}^*$ .
- (2). Repeat (1)  $B$  times and obtain  $\hat{N}^{*1}, \hat{N}^{*2}, \dots, \hat{N}^{*B}$ .

Let

$$G(x) = B^{-1} \sum_{b=1}^B I(\hat{N}^{*b} \leq x) \text{ and } G^{-1}(u) = \inf\{x : G(x) \geq u\}.$$

Then for any  $\alpha$  with  $0 < \alpha < 0.5$ , a  $(1 - \alpha)$  two-sided bootstrap confidence interval is given by  $(G^{-1}(\alpha/2), G^{-1}(1 - \alpha/2))$  and a  $(1 - \alpha)$  one-sided bootstrap confidence interval is given by  $(G^{-1}(\alpha), +\infty)$ .

We note that the bootstrap intervals should not be interpreted as confidence intervals for the true  $N$  except for the unlikely situation in which  $T$ , the number of support points in  $Q$ , is known and finite. They are rather intervals for the corresponding lower bounds. As such, the lower limits do represent consecutive lower limits for  $N$ , but the upper limits are only valid for the lower bounds  $N_p$ .

It should not be surprising that the bootstrap yields a wide confidence interval because the confidence interval necessarily becomes wider and wider as  $N$  increases. In fact, the confidence interval width can shrink at the usual rate  $1/\sqrt{N}$  only on the log scale and the confidence intervals for  $N$  have width proportional to  $\sqrt{N}$ , see ( 19 ) and ( 20 ).

**Remark 10** *There are model-based bootstrap procedures in which  $Q$  is modeled parametrically or nonparametrically. There exists a moment-based estimator for the latent distribution as seen in **Remark 8**. Nonparametric maximum likelihood estimation of the latent distribution is another nonparametric approach, see Norris and Pollock (1995, 1996, 1998). The present authors are investigating alternative approaches to obtain moment estimators and nonparametric maximum likelihood estimators for the latent distribution  $Q$ . In there, estimation of  $Q$  can be separated from  $N$  and then  $N$  is estimated given the nuisance parameter  $\theta$ , which is a functional of the latent distribution.*

## 4.6 Selection rules

If one wishes to select a single  $\widehat{N}_p$  from the lower bound estimators, the following considerations are relevant.

The moment-based estimator  $\widehat{N}_p$  does not estimate  $N$  directly but rather  $N_p$ , where  $0 \leq N_p \leq N$ . Thus as an estimator for  $N$ , it has bias of the form  $N(\mu(0) - \mu_p)$ . The estimator  $\widehat{N}_p$  has an asymptotic variance of the magnitude  $O(N)$ . Thus for fixed  $p$ , the mean squared error of  $\widehat{N}_p$ ,  $E(\widehat{N}_p - N)^2$ , is eventually dominated by the bias-squared term. On the other hand, as a function of  $p$ , for fixed  $N$ , the bias of  $\widehat{N}_p$  decreases in  $p$  while the variance increases. Thus one can anticipate that the optimal choice for  $p$  will depend on  $N$  and  $Q$ , and will be increasing as a function of  $N$ .

In the face of this complexity, we therefore propose two simple rules of selection. For the first, let  $K$  be the largest  $p$  such that

$$\widehat{N}_1 \leq \widehat{N}_2 \leq \dots \leq \widehat{N}_p,$$

and set that  $\widehat{N}_{max} = \widehat{N}_K$ . The logic here is that as long as  $\widehat{N}_p$  is increasing, we seem to be removing bias. We also note that this is the largest  $p$  that makes the empirical moment sequence consistent with a theoretical moment sequence.

A second strategy attacks the bias-variance tradeoff more directly. Consider the construction of lower 95% confidence limits  $\{\widehat{L}_1, \widehat{L}_2, \dots, \widehat{L}_{[S/2]}\}$  for the true values of  $\{N_1, N_2, \dots, N_{[S/2]}\}$ . Since the lower confidence limits for  $N_p$  is also a conservative limit for  $N$ , the largest  $\widehat{L}_p$  in the sequence is the best lower limit for  $N$ . Let  $\widehat{N}_{LL}$  be  $\widehat{N}_p$  for this value of  $p$ .

This choice balances the decreasing bias of  $\widehat{N}_p$  against the increasing variance.

## 5 EXAMPLES

### 5.1 The tomato EST data

A prepared cDNA library typically consists of  $10^6$  clones. Each clone represents one copy of cDNA from a gene. Usually a cDNA library is highly heterogeneous in the sense that the numbers of cDNA copies of each gene, that is, the expression levels, may differ by  $10^3 \sim 10^4$  fold. When sequencing a sample of clones from a cDNA library, the “single-pass” cDNA sequences are called expressed sequence tags (ESTs). The ESTs from the sample are clustered into groups, each corresponding to a distinct gene and the data are used to make inference about  $N$ , the unknown number of distinct genes in the library. Let  $x_i$  be the number of ESTs from the  $i$ th gene,  $i = 1, 2, \dots, N$ . We study one tomato flower cDNA library, which was made from  $0 \sim 3$  mm buds of tomato flowers. A sample of 2586 ESTs was generated from the library. We list the nonzero EST frequency counts here:  $n_1 = 1434$ ,  $n_2 = 253$ ,  $n_3 = 71$ ,  $n_4 = 33$ ,  $n_5 = 11$ ,  $n_6 = 6$ ,  $n_7 = 2$ ,  $n_8 = 3$ ,  $n_9 = 1$ ,  $n_{10} = n_{11} = 2$ , and  $n_{12} = n_{13} = n_{14} = n_{16} = n_{23} = n_{27} = 1$ . The original data are from the Tomato Gene Index at the Institute of Genomic Research (TIGR). The library identifier is T1526. For more information about the Gene Index at TIGR, the readers are referred to Quackenbush et al. (2000).

It is reasonable to assume the data approximate a Poisson sample. Since the estimation will rely on the frequency counts with  $x$  near zero, we only plot  $(x, \widehat{H}(x))$  relative to first 6 frequency counts in **Figure 1**, in which the log ratio plot shows a clear convexity. The lower bound estimates are  $\widehat{N}_1 = 5889$ ,  $\widehat{N}_2 = 7024$ ,  $\widehat{N}_3 = 7300$ ,  $\widehat{N}_4 = 7430$ ,  $\widehat{N}_5 = 8037$ ,  $\widehat{N}_6 = 7414$  and  $\widehat{N}_7 = 7437$ . However, since the last four estimators depend very small frequency counts, they are less reliable. if we follow the first selection rule, the estimate for  $N$  is  $\widehat{N}_5 = 8037$ . If we follow **Remark 4** and the first selection rule, then the estimate for  $N$  is  $\widehat{N}_3 = 7300$  because  $n_7 = 2 < 5$  and only the first six frequency counts are used to construct the estimator sequence.

When the estimate  $\widehat{N}_3 = 7300$  is used and 500 bootstrap re-samples are generated, the two-sided 95% confidence interval for  $N_3$  is (5902, 24245). The upper-end is noninformative because it is known that the total number of tomato genes is about 23000. The one-sided 95% confidence interval for  $N_3$  is (6103,  $+\infty$ ). When the estimate  $\widehat{N}_5 = 8037$  is used, the two-sided 95% confidence interval  $N_5$  is (6001, 31277). The upper-end is noninformative again. The one-sided 95% confidence interval for  $N_5$  is (6236,  $+\infty$ ), representing an improvement in confidence for  $N$  itself over using  $\widehat{N}_3$ .

## 5.2 The cotton-tail rabbit data

The estimation of population size via capture-recapture experiments is mathematically equivalent to the estimation of the number of classes via a multiple Bernoulli sample. If each individual is assumed

to have the same probability to be captured on all occasions, the multiple Bernoulli sample becomes a binomial sample. The heterogeneity model in capture-recapture studies allows the capture probabilities to vary among individuals. The methodology developed in this paper for a binomial sample can be applied to the heterogeneity model in capture-recapture studies. The following example is from such a study.

Edwards and Eberhardt (1967) performed a live-trapping study on a closed population of wild cotton-tail rabbits with known population size  $N = 135$ . The experiment was performed for 18 consecutive nights and the presence/absence of particular rabbits was recorded. The nonzero observed frequency counts are  $n_1 = 43$ ,  $n_2 = 16$ ,  $n_3 = 8$ ,  $n_4 = 6$ ,  $n_6 = 2$  and  $n_7 = 1$ . Here, for example, the frequency count  $n_1 = 43$  means that 43 rabbits were caught exactly once in 18 nights. This famous dataset has received attention from many authors. The Schnabel-type estimates were reported in Edwards and Eberhardt (1967). The Schnabel estimate is 96. The Schumacher and Eschmeyer estimate is 97. Edwards and Eberhardt (1967) fit a geometric series and obtained the estimates 136 and 164 using two different procedures. Burnham and Overton (1978, 1979) used a third-order jackknife estimate, 159, with a 95% confidence interval (116, 202), and an interpolated jackknife estimate, 142, with a 95% confidence interval (112, 172). The estimate in Chao (1984) is 134, with a bootstrap confidence interval (97, 168). Chao (1987) used a refined version of the estimator in Chao (1984), which yields an estimate, 136, with a 95% confidence interval (87, 185).

**Figure 2** shows that the log ratio plot is convex. This means a mixture binomial model may be appropriate. The moment lower bound estimate sequence is  $\hat{N}_1 = 131$  and  $\hat{N}_2 = 144$ , and the maximum estimate is  $\hat{N}_2 = 144$ . The smallest frequency count used in  $\hat{N}_2$  equals six.

When  $\hat{N}_2 = 144$  is used, the (500 re-samples) bootstrap 95% two-sided confidence interval for  $N_2$  is (103, 546). The one-sided 95% confidence interval for  $N_2$  is (106,  $+\infty$ ). If we use  $\hat{N}_1 = 131$ , then the 95% two-sided confidence interval for  $N_1$  is (101, 202) and one-sided 95% confidence interval for  $N_1$  is (107,  $+\infty$ ) from 500 re-samples. Thus the lower variation of  $\hat{N}_1$  gives some improvement for inference on  $N$ .

Comparing of the preceding estimates with the ones considered here is rendered more difficult by two considerations:

- (1) As all point estimates are based on lower order frequency counts, they must have some systematic bias in the mixture model, bias that varies from method to method.
- (2) The confidence limits themselves represent different methodologies. One way to make a fair comparison of variability would be to use the same procedure to assess variability of all estimates, a subject which is still under investigation.

## 6 DISCUSSION

The detection of heterogeneity is important in the estimation of the number of classes  $N$ . An asymptotic  $\chi^2$  test statistic was introduced in

a binomial sample by Burnham and Overton (1978). However, many estimation methods often assume either homogeneity or heterogeneity without providing a model-checking procedure, for example, see Chao (1984). The diagnostic proposed in this paper displays heterogeneity in an easy way without requiring model fitting. Convexity is a strongly graphical property and easily assessed by the simple log ratio plot. The authors are investigating more delicate graphical diagnostics and a dispersion score test.

The proposed moment estimator sequence has the advantage of having an explicit expression for its bias under minimal assumptions. The combinations of observed frequency counts used to define the estimators are based on the properties of the latent distribution only through the mixture density. The total bias is expressed as the summation of a sequence of bias-corrections, which can be estimated by functions of observed frequency counts. If the lower order bias terms are very large and can be well estimated, then the significant improvement can be made. All nonparametric estimators use some function of the observed frequency counts as an predictor for  $n_0$ . Unfortunately, higher order frequency counts usually will introduce large variance into the estimator. On the other hand, using only lower order frequency counts may significantly under-estimate the true value. Unfortunately, there is no upper bound theory that enables one to trap the true value of  $N$  from above.

The estimation of the number of shared classes in two populations is a more difficult task. Only Chao et al. (2000) generalized

sample-coverage based methods to address a two-population problem. However, the approach in this paper can be conveniently generalized to estimate the total number of classes of multiple populations, which automatically yields estimators for the number of shared classes of two populations. The generalization will be reported in a sequel paper.

Finally, the moment results in this paper can be rephrased in terms of the identifiability of the parameters in the observed data sampling models ( 4 ), ( 5 ) and ( 6 ). That is, the multinomial distribution of  $\{n_x : x \in \Omega_0\}$  given  $n$  can be used to consistently estimate (as  $n \rightarrow +\infty$ ) the cell-probabilities  $f(j; Q)/[1 - f(0; Q)]$  for  $j$  in  $\Omega_0$ . However, if these cell probabilities do not determine  $f(0; Q)$  ( or  $\theta$  ) uniquely, then there is no hope that the binomial distribution of  $n$  will uniquely determine  $N$ . The question as to whether and how the cell probabilities constrain  $f(0; Q)$  can be deduced from the moment results in this paper.

## APPENDIX: PROOFS

In all the proofs, for a matrix  $A$  we will use the notation  $A > 0$  to represent that  $A$  is positive definite and  $A \geq 0$  to represent that  $A$  is nonnegative definite.

### Proof of Proposition 6

Since  $H_p(y)$  has entries identical to those of  $H_p$  except for the entry in the left-upper corner, the condition that  $|H_p(y)| \geq 0$  and the nonnegative definiteness of  $H_p$ , which can be obtained from **Theorem**

3 and **Theorem 4**, ensure that  $H_p(y) \geq 0$  by simple algebra theorems for testing nonnegative definiteness. Hence we have  $y \geq 0$  due to the nonnegative definiteness of  $H_p(y)$ .

## Proof of Proposition 7

One result about the determinant of the sum of two matrices and two results about the inverse and determinant of a partitioned matrix are used in the proof. For convenience, we describe them first, see Schott (1997).

Let  $A$  be a matrix,  $\mathbf{b}$  be a vector and  $c$  be a scalar, such that  $A$  and  $A + c\mathbf{b}\mathbf{b}'$  are nonsingular. Then

$$|A + c\mathbf{b}\mathbf{b}'| = |A|(1 + c\mathbf{b}'A^{-1}\mathbf{b}).$$

Let

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \quad B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}.$$

Then

$$|A_{22}| \neq 0 \implies |A| = |A_{22}| |(A_{11} - A_{12}A_{22}^{-1}A_{21})|,$$

$$|A_{11}| \neq 0 \implies |A| = |A_{11}| |(A_{22} - A_{21}A_{11}^{-1}A_{12})|.$$

Suppose  $A$ ,  $A_{11}$  and  $A_{22}$  are nonsingular, and  $B$  is the inverse of  $A$ .

Then

$$B_{11} = (A_{11} - A_{12}A_{22}^{-1}A_{21})^{-1} = A_{11}^{-1} + A_{11}^{-1}A_{12}B_{22}A_{21}A_{11}^{-1},$$

$$B_{22} = (A_{22} - A_{21}A_{11}^{-1}A_{12})^{-1} = A_{22}^{-1} + A_{22}^{-1}A_{21}B_{11}A_{12}A_{22}^{-1},$$

$$B_{12} = -A_{11}^{-1}A_{12}B_{22},$$

$$B_{21} = -A_{22}^{-1}A_{21}B_{11}.$$

We first consider the case  $T = +\infty$ . We have

$$H_p > 0 \implies A_p > 0.$$

Let

$$B_p = A_p^{-1} \text{ and } \mathbf{c}_p = (\mu(p+2), \mu(p+3), \dots, \mu(2p+1))'.$$

Then

$$A_{p+1} = \begin{pmatrix} A_p & \mathbf{c}_p \\ \mathbf{c}'_p & \mu(2p+2) \end{pmatrix}, \quad \bar{H}_p = \begin{pmatrix} \mathbf{b}_p & A_p \\ \mu(p+1) & \mathbf{c}'_p \end{pmatrix},$$

and for all  $y$  in  $\mathcal{R}$ ,

$$H_p(y) = \begin{pmatrix} y & \mathbf{b}'_p \\ \mathbf{b}_p & A_p \end{pmatrix}.$$

$$|A_{p+1}| = |A_p|(\mu(2p+2) - \mathbf{c}'_p A_p^{-1} \mathbf{c}_p) > 0.$$

$$|\bar{H}_p| = (-1)^p |A_p|(\mu(p+1) - \mathbf{b}'_p A_p^{-1} \mathbf{b}_p) > 0.$$

For all  $y$  in  $\Delta_p$ , we obtain

$$|H_p(y)| = |A_p|(y - \mathbf{b}'_p A_p^{-1} \mathbf{b}_p).$$

We can conclude that

$$\mu_p = \mathbf{b}'_p A_p^{-1} \mathbf{b}_p.$$

Let

$$B_{p+1} = \begin{pmatrix} E_p & \mathbf{d}_p \\ \mathbf{d}'_p & g_p \end{pmatrix},$$

where  $E_p \in \mathcal{R}^{p \times p}$ ,  $\mathbf{d}_p \in \mathcal{R}^p$  and  $g_p \in \mathcal{R}$ . Then

$$E_p = (A_p - \mathbf{c}_p[\mu(2p+2)]^{-1} \mathbf{c}'_p)^{-1} = A_p^{-1} + A_p^{-1} \mathbf{c}_p g_p \mathbf{c}'_p A_p^{-1},$$

$$g_p = (\mu(2p+2) - \mathbf{c}'_p A_p^{-1} \mathbf{c}_p)^{-1} = |A_p| |A_{p+1}|^{-1},$$

$$\mathbf{d}_p = -A_p^{-1} \mathbf{c}_p g_p.$$

$$\begin{aligned}
\mu_{p+1} &= \mathbf{b}'_{p+1} A_{p+1}^{-1} \mathbf{b}_{p+1} \\
&= \mathbf{b}'_{p+1} B_{p+1} \mathbf{b}_{p+1} \\
&= (\mathbf{b}'_p, \mu(p+1))' \begin{pmatrix} E_p & \mathbf{d}_p \\ \mathbf{d}'_p & g_p \end{pmatrix} \begin{pmatrix} \mathbf{b}_p \\ \mu(p+1) \end{pmatrix} \\
&= \mathbf{b}'_p E_p \mathbf{b}_p + 2\mu(p+1) \mathbf{b}'_p \mathbf{d}_p + g_p [\mu(p+1)]^2 \\
&= \mathbf{b}'_p (A_p^{-1} + A_p^{-1} \mathbf{c}_p g_p \mathbf{c}'_p A_p^{-1}) \mathbf{b}_p \\
&\quad + 2\mu(p+1) \mathbf{b}'_p (-A_p^{-1} \mathbf{c}_p g_p) + g_p [\mu(p+1)]^2 \\
&= \mathbf{b}'_p A_p^{-1} \mathbf{b}_p + g_p [\mu(p+1) - \mathbf{b}'_p A_p^{-1} \mathbf{c}_p]^2 \\
&= \mu_p + \frac{|\bar{H}_p|^2}{|(-1)^p A_p|^2 |A_{p+1}|} A_p \\
&= \mu_p + \frac{|\bar{H}_p|^2}{|A_p| |A_{p+1}|}.
\end{aligned}$$

Using  $\beta_p$  defined in ( 14 ), we get

$$\beta_p > 0 \text{ and } \mu_{p+1} = \mu_p + \beta_p > \mu_p.$$

We next consider the case in which  $T$  is finite. From **Theorem 4**, the expression for  $\mu_p$  holds for  $p < T$ . We only need to show that  $A_T > 0$  such that  $\mu_T$  is well defined. Then it is clear that the relation between successive lower bounds holds. Consider a re-weighted measure  $\nu^+(\lambda)$  defined by

$$d\nu^+(\lambda) = \lambda^2 d\nu(\lambda).$$

It is clear that  $\nu^+(\lambda)$  also has exactly  $T$  support points. Consider the moment sequence  $\mathcal{M}^+$  of the measure  $\nu^+(\lambda)$  from  $\nu(\lambda)$ , defined by

$$\mathcal{M}^+ = \{\mu^+(x) : \mu^+(x) = \int \lambda^x d\nu^+(\lambda), x = 0, 1, \dots\}.$$

Then

$$\mu^+(x) = \mu(x+2) \implies A_T = H_{p-1}(\mathcal{M}^+) > 0.$$

## Proof of Proposition 8

$\mathcal{M}(\mu_\infty)$  is a strictly increasing sequence. From the definitions and **Proposition 6**, we have

$$\forall p \in \mathcal{N}, \quad \mu_p < \mu_\infty \implies |H_p(\mu_\infty)| > 0 \implies H_p(\mu_\infty) > 0.$$

From **Theorem 3**,

$$0 < \bar{H}_p = \bar{H}_p(\mu_\infty).$$

Then **Theorem 3** ensures that the proposition holds.

## Proof of Proposition 9

Suppose there exists  $\delta$  such that

$$\int e^{t\lambda} d\nu(\lambda) < +\infty, \quad t < \delta.$$

It is clear that

$$\int e^{t\lambda} \lambda d\nu(\lambda) < +\infty, \quad t < \delta/2.$$

Let  $\nu_*(\lambda)$  be one measure determined by  $\mathcal{M}(\mu_\infty)$ . Let  $d\nu^+(\lambda) = \lambda d\nu(\lambda)$  and  $d\nu_*^+(\lambda) = \lambda d\nu_*^+(\lambda)$ . Let  $\mu^+(x)$  be the  $x$ th moment of  $\nu^+(\lambda)$ . Then

$$\mu^+(x) = \mu(x+1), \quad \forall x = 0, 1, \dots$$

$\mu^+(x)$  is also the  $x$ th moment of  $\nu_*^+(\lambda)$ . The moment generating function of  $\nu^+(\lambda)$  exists because

$$\int e^{t\lambda} d\nu^+(\lambda) = \int e^{t\lambda} \lambda d\nu(\lambda) < +\infty, \quad t < \delta/2.$$

Thus  $\{\mu^+(x)\}_{x=0}^{+\infty}$ , that is,  $\mathcal{M} - \{\mu(0)\}$ , uniquely determines  $\nu^+(\lambda)$ .

So we have

$$\nu^+(\lambda) = \nu_*^+(\lambda).$$

The proposition holds because

$$\mu_\infty = \int \lambda^{-1} d\nu_*^+(\lambda) = \int \lambda^{-1} d\nu^+(\lambda) = \mu(0).$$

## REFERENCES

- Bunge, J. and Fitzpatrick, M. (1993). “Estimating the Number of Species: A Review”. *Journal of the American Statistical Association*, 88:364-373.
- Burnham, K. P. and Overton, W. S. (1978). “Robust Estimation of Population Size When Capture Probabilities Vary Among Animals”. *Biometrika*, 65:625-634.
- (1979). “Estimation of the Size of a Closed Population When Capture Probabilities Vary Among Animals”. *Ecology*, 60:927-936.
- Cantor, C. R. and Smith, C. L. (1999). *Genomics : the science and technology behind the Human Genome Project*. New York: Wiley.
- Chao, A. (1984). “Nonparametric Estimation of the Number of Classes in a Population”. *Scandinavian Journal of Statistics*, 11:265-270.
- (1989). “Estimating Population Size for Sparse Data in Capture-recapture Experiments”. *Biometrics*, 45:427-438.
- Chao, A. and Lee, S. M. (1992). “Estimating the Number of Classes Via Sample Coverage”. *Journal of the American Statistical Association*, 87:210-217.

- Chao, A. and Hwang, W. H. and Chen, Y. C. and Kuo, C. Y. (2000). “Estimating the number of shared species in two communities”. *Statistica Sinica*, 10:227-246.
- Darroch, J. N. and Ratcliff, D. (1980). “A Note on Capture-recapture Estimation”. *Biometrics*, 36:149-153.
- Dette, H. and Studden, W. J. (1997). *The theory of canonical moments with applications in statistics, probability, and analysis*, New York: Wiley.
- Donoho, D. (1988). “One-sided inference about functionals of a density”. *Annals of Statistics*, 16:1390-1420.
- Edwards, W. and Eberhardt, L. L. (1967). “Estimating cottontail abundance from live-trapping data”. *Journal of Wildlife management*, 31:87-96.
- Efron, B. and Tibshirani, R. (1976). “Estimating the number of unseen species: how many words did Shakespeare know?”. *Biometrika*, 63:435-447.
- Gong, G. and Samaniego, F. J. (1981). “Pseudo Maximum Likelihood Estimation: Theory and Applications.” *Annals of Statistics*, 9:861-869.
- Gyires, B. (1998). *Linear Approximation in Convex Metric Spaces and the Application in the Mixture Theory of Probability Theory*, Word Scientific Publishing Co. Pte. Ltd.
- Harris, B. (1959). “Determining bounds on integrals with applications to cataloging problems”. *Annals of Mathematical Statistics*, 30:521-528.

- Kalinin, V. M. (1965). "Functionals related to the Poisson distribution, and the statistical structure of a text". *Proceedings of the Steklov Institute of Mathematics*, 79:6-19.
- Karlin, S. (1968). *Total positivity*, Stanford University Press, Vol. 1.
- Keener, R. and Rothman, E. and Starr, N. (1987). "Distributions on partitions". *Annals of Statistics*, 15:1466-1481.
- Lewins, W. A. and Joanes, D. N. (1984). "Bayesian estimation of the number of species". *Biometrics*, 40:323-328.
- Lindsay, B. G. (1989). "Moment Matrices: Applications in Mixtures". *Annals of Statistics*, 17:722-740.
- (1995). *Mixture Models: Theory, geometry and applications*, NSF-CBMS Regional Conference Series in Probability and Statistics.
- Lindsay, B. G. and Roeder, K. (1987). "A Unified Treatment of Integer Parameter Models". *Journal of the American Statistical Association*, 82:758-764.
- (1992). "Residual diagnostics for mixture models". *Journal of the American Statistical Association*, 87:785-794.
- Mao, C. and Lindsay, B. G. (2001). "A Poisson model for coverage problems with an application in genomic research". *Submitted to Biometrika*.
- McNeil, D. (1973). "Estimating an author's vocabulary". *Journal of the American Statistical Association*, 68:92-96.
- Mingoti, S. A. and Meeden, G. (1992). "Estimating the total number of distinct species using presence and absence data". *Biometrics*, 48:863-875.

- Norris, J. L. I. and Pollock, K. H. (1995). “A Capture-recapture Model With Heterogeneity and Behavioural Response”. *Environmental and Ecological Statistics*, 2: 305-313.
- (1996). “Nonparametric MLE Under Two Closed Capture-recapture Models With Heterogeneity”. *Biometrics*, 52:639-649.
- (1998). “Nonparametric MLE for Poisson Species Abundance Models Allowing for Heterogeneity Between Species”. *Environmental and Ecological Statistics*, 5:391–402.
- Ord, J. K. and Whitmore, G. A. (1986). “The Poisson-inverse Gaussian Distribution As a Model for Species Abundance”. *Communications in Statistics, Part A – Theory and Methods*, 15:853-871.
- Quackenbush, J., Cho, J., Lee, D., Liang, F., Holt, I., Karamycheva, S. Parvizi, B., Pertea, G., Sultana, R. and White, J. (2000). “The TIGR Gene Indices: analysis of gene transcript sequences in highly sampled eukaryotic species”. *Nucleic Acids Research*, 29:159-164.
- Schott, J. R. (1997). *Matrix Analysis for Statistics*, New York: Wiley.
- Titterton, D. M., Smith, A. F. and Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*, Chichester: Wiley.
- Widder, D. V. (1941). *The Laplace transformation*, Princeton University Press.
- Zelterman, D. (1988). “Robust estimation in truncated discrete distributions with application to capture-recapture experiments”. *Journal of statistical planning and inference*, 18:225-237.

Figure 1: The log ratio plot (EST data,  $\lambda_0 = 1$ )

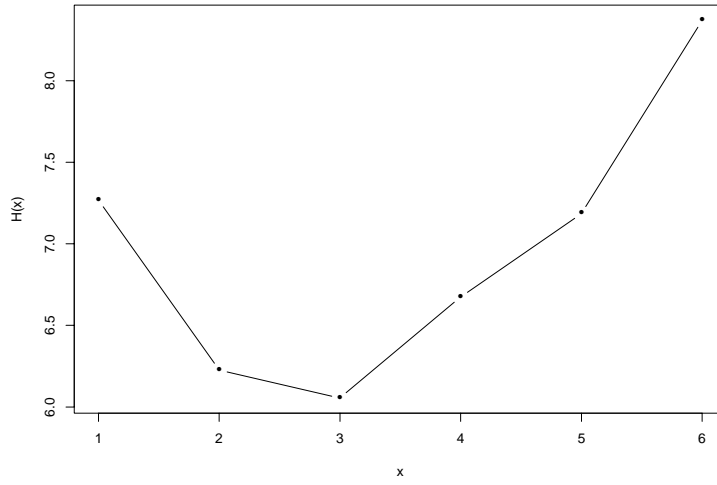


Figure 2: The log ratio plot (rabbit data,  $\lambda_0 = 1/9$ )

