Maximum Likelihood Estimation under a Successive Sampling Discovery Model

Vijayan N. Nair; Paul C. C. Wang

Technometrics, Vol. 31, No. 4 (Nov., 1989), 423-436.

Stable URL:
http://links.jstor.org/sici?sici=0040-1706%28198911%2931%3A4%3C423%3AMLEUAS%3E2.0.CO%3B2-2

Technometrics is currently published by American Statistical Association.

TECHNOMETRICS, NOVEMBER 1989, VOL. 31, NO. 4

Maximum Likelihood Estimation Under a Successive Sampling Discovery Model

Vijayan N. Nair

AT&T Bell Laboratories
Murray Hill, NJ 07974

Paul C. C. Wang

Department of Mathematics and Statistics
University of Calgary
Calgary, Alberta T2N 1N4
Canada

A common problem encountered in the analysis of discovery data is the size-bias phenomenon in which the larger units tend to be discovered first. One approach to account for this bias is to model the discovery process as sampling successively from a finite population without replacement and with probability proportional to size. We consider in this article a generalized version of this model for analyzing multivariate data with any given measure of size. We assume a superpopulation framework and develop procedures for maximum likelihood estimation of the parameters of the distribution. The use of the EM algorithm for computing the maximum likelihood estimates, associated computational issues, and relationships to regression estimators in survey sampling are discussed. Oil discovery data from the Rimbey–Meadowbrook reef play are used to illustrate the techniques.

KEY WORDS: EM algorithm; Finite population, Parametric inference; Petroleum resource estimation; Size-biased sampling; Superpopulation model.

1. INTRODUCTION

In analyzing discovery data, one frequently has to deal with the size-bias phenomenon, in which the larger units tend to be discovered first. In petroleum resource estimation, it is well-known that measures of size of a pool, such as area and mean formation depth, impact its chance of discovery (Arps and Roberts 1958). Littlewood (1981) considered a software debugging model in which the bugs that make the greatest contribution to the overall failure rate are discovered earlier and so are fixed earlier. Similar situations also arise in testing for design errors in hardware reliability. Any reasonable inference based on the discovered data must take the biased nature of the sample into account; treating the data as a simple random sample from the population of interest can lead to extremely biased predictions.

One of the approaches proposed in the literature to account for this size bias is to model the discovery process as sampling successively from a finite population without replacement and *with probability proportional to size* (Kaufman, Balcer, and Kruyt 1975). See also Cozzolino (1972). We consider in this article a generalized version of this model for analyzing multivariate data with an arbitrary measure of size. Specifically, let $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$, with $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ik})'$ be the values associated with the N units of a finite population that are available for discovery. Let $w(\cdot)$ denote a positive weight function. Then, given $\mathbf{Y}_i = \mathbf{y}_i$ $(i = 1, \ldots, N)$ the probability of observing the *ordered* sample $(i_1, \ldots, i_n)$ under this model is

$$\Pr\{(i_1, \ldots, i_n) \mid \mathbf{Y}_i = \mathbf{y}_i, i = 1, \ldots, N\} = \prod_{j=1}^{n} \frac{w(\mathbf{y}_{i_j})}{\sum_{i=1}^{N} w(\mathbf{y}_i) - \sum_{k=0}^{j-1} w(\mathbf{y}_{i_k})}, \quad (1.1)$$

where $w(y_{i_0}) \equiv 0$. In other words, the sample is obtained by selecting successively without replacement and with probability proportional to $w(\mathbf{y})$ from the finite population of N units.

Note the similarity between (1.1) and the traditional probability-proportional-to-size-without-replacement scheme in survey sampling. In the latter case, however, the size measures associated with all N units of the finite population are known a priori. In (1.1), the probability of selecting a unit is a function of its a priori unknown characteristics, and the characteristics associated with the $N - n$ unobserved units remain unknown even after the sample becomes available.

The initial formulation of this model with $w(y) = y$ (univariate case) was given by Kaufman et al. (1975). The generalization to $w(y) = y^\gamma$ was considered by Bloomfield, Deffeyes, Watson, Benjamini, and Stine (1979), Smith and Ward (1981), and Lee and Wang

(1985). Several methods of estimation have been developed for the univariate problem in the literature. Andreatta and Kaufman (1986), Gordon (1983), Smith and Ward (1981), Wang and Nair (1988), and Bickel, Nair, and Wang (1989) discussed nonparametric estimation procedures. See also Mallows and Nair (1987) for a problem associated with unbiased estimation in such models. Inference for parametric models such as the lognormal distribution was considered by Barouch and Kaufman (1976, 1977), Barouch, Kaufman, and Nelligan (1983), Lee and Wang (1986), and others. Most of these results are based on a superpopulation framework in which the attributes $\mathbf{Y}_i$'s associated with the units of the finite population are assumed to be generated independently and identically according to a distribution function (df) $F(\mathbf{y})$,

This article considers parametric maximum likelihood estimation methods for the (multivariate) biased sampling model in (1.1). We also assume a superpopulation framework, although we view this primarily as a mechanism for doing smooth, parametric estimation of the underlying population. It is assumed that the superpopulation distribution F is known except for a vector of unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)'$, and $f_{\boldsymbol{\theta}}(\mathbf{y})$ denotes the density. We develop procedures for maximum likelihood estimation of $\boldsymbol{\theta}$. In a typical data-analytic approach, one would first estimate the distribution nonparametrically and then use goodness-of-fit procedures (formal or graphical) to choose the appropriate parametric distribution (see Wang and Nair 1988). The lognormal distribution has been found to be useful for modeling oil and gas discovery data. Throughout the article, it is assumed that the weight function $w(\mathbf{y})$ and the finite population size N are known. See the discussion in Section 5, however.

The limiting case, as $N \to \infty$ ($n/N \to 0$), of this biased sampling model was studied by Cox (1969), Patil and Rao (1977, 1978), and Vardi (1982). In this infinite population case, the observations are iid with density

$$g_{\boldsymbol{\theta}}(\mathbf{y}) = \frac{[w(\mathbf{y})f_{\boldsymbol{\theta}}(\mathbf{y})]}{[E_{\boldsymbol{\theta}}w(\mathbf{Y})]}. \tag{1.2}$$

The univariate case with $w(y) = y$ is the familiar length-biased sampling problem. At the other extreme, if $n = N$ ($n/N = 1$), F can be estimated in the usual way based on the N iid observations $\mathbf{Y}_1, \ldots, Y_N$. Since all N units in the finite population have been observed, the sampling design becomes irrelevant, and the inference does not depend on the biased sampling model (1.1). As we shall see in Section 3, the effect of the weight function $w(\mathbf{y})$ in finite populations is in between these two extreme cases.

The article is organized as follows. In Section 2, discovery data from the Rimbey–Meadowbrook reef play and simulation results are used to illustrate the existence and consequences of the size-bias phenomenon. Maximum likelihood estimation under the successive sampling discovery model is considered in Section 3. The likelihood equations under the model are derived, and the EM algorithm for solving the equations and related computational issues are discussed. In Section 4, we consider situations in which the likelihood can be separated into a part that depends on the biased sampling mechanism and another part that is dependent of the bias. The analysis of multivariate data simplifies considerably in such cases. In Section 5, we illustrate the techniques by reanalyzing the oil discovery data from the Rimbey–Meadowbrook reef play.

2. AN EXAMPLE

The Rimbey–Meadowbrook reef chain, located in central Alberta, is one of the most productive plays in the western Canada sedimentary basin. It produces both oil and gas but is predominantly a conventional oil play. The first discovery from this play was made in 1947 and the last significant one in 1984. A detailed description of the geological formation of the Rimbey–Meadowbrook reef chain was given by Stoakes (1980). See also Lee and Wang (1986). Table 1 gives the data for the pools discovered before 1968. This data set was obtained from Energy Resources Conservation Board (1985). The data in Table 1 contain information on the discovery order, as represented by the discovery number, and the following multivariate attributes associated with the discovered pools: *volume*—the volume of oil in place, measured in million cubic meters ($10^6 m^3$); *area*—the area of the pool closure, measured in hectares (ha); *net pay*—the average net-pay thickness of a pool, measured in meters (m); and *depth*—the mean formation depth, measured in meters (m) below sea level. For illustrative purposes, we assume that the superpopulaton distribution of the data

$$\mathbf{Y} = (\textit{volume}, \textit{area}, \textit{net pay}, \textit{depth}) \tag{2.1}$$

associated with each pool is multivariate logonormal with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Table 2 gives the maximum likelihood estimates (MLE's) of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ treating the data in Table 1 as a simple random sample. The estimates corresponding to the *volume* variable are of particular interest, since we want to predict the amount of recoverable oil remaining in the play. From Table 2, the estimate of the mean of log-*volume* is 1.33. Suppose the total number of pools in this play is 40 (see Lee and Wang 1986). Then we can predict the total recoverable oil from the remaining 17 pools as approximately 64 million cubic meters ($17 \times e^{1.33}$).

That this is an overly optimistic prediction is easily

Table 1. Rimbey–Meadowbrook Reef Discovery Data

Discovery number	Field/pool name	Volume ($10^6 m^3$)	Area (ha)	Net pay (m)	Depth (m)
1	Leduc–Woodbend A	61.200	8,812	10.80	1,620.5
2	Redwater	207.000	15,199	31.39	977.8
3	Golden Spike A	49.600	590	135.64	1,728.8
4	Acheson	27.600	1,542	24.12	1,547.8
5	Golden Spike C	.425	158	5.82	1,827.0
6	Wizard Lake A	62.000	1,075	86.13	1,969.0
7	Glen Park A	4.660	173	39.32	1,921.8
8	Leduc–Woodbend B	2.380	751	7.99	1,653.5
9	Bonnie Glen	125.000	3,120	55.44	2,165.6
10	Westerose	31.000	652	72.20	2,204.6
11	St. Albert B	1.750	110	22.00	1,424.9
12	Fairydell–Bon Accord	2.770	405	13.75	1,226.5
13	Homeglen–Rimbey A	14.900	4,563	7.56	2,415.5
14	Leduc–Woodbend F	1.030	81	20.91	1,658.1
15	Yekau Lake A	10.700	250	6.58	1,557.5
16	Morinville A	.091	16	10.97	1,379.2
17	St. Albert A	3.700	101	43.24	1,463.6
18	Morinville B	2.590	323	14.48	1,608.1
19	Sylvan Lake	1.620	987	6.16	2,881.9
20	Morinville C	.615	211	3.51	1,379.8
21	Wizard Lake B	.160	54	4.45	2,108.0
22	Lanaway	.245	65	7.92	2,923.3
23	Yekau Lake B	.040	16	7.32	1,552.7

seen from Figure 1, which shows plots of log-*volume* and log-*area* against the discovery order. The smooth curves, obtained by the *lowess* procedure (Cleveland 1979), illustrate the general declining trend with advancing exploration. It is known in petroleum exploration that the area of a pool significantly affects its probability of discovery. Although *volume* does not impact discovery order directly, it is highly correlated with *area*, thus explaining the declining trend in Figure 1b. These plots indicate clearly that there is a strong size bias associated with the discovery order and that most of the larger pools are likely to have been discovered. Treating the discovered data as a simple random sample will, therefore, result in biased inference. For example, the sample mean of the data will overestimate the underlying population mean and the sample variance will underestimate the population variance.

Table 2. Maximum Likelihood Estimates of the Parameters of the Multivariate Lognormal Distribution for the Rimbey–Meadowbrook Data Under Simple Random Sampling

				$\hat{\rho}_{kj}$			
Variable	γ_k	$\hat{\mu}_k$	$\hat{\sigma}_{kk}$	Volume	Area	Net pay	Depth
Volume	0	1.33 (±.10)	5.45 (±2.70)	1.00	.87	.65	−.09
Area	0	5.92 (±.08)	3.20 (±.93)		1.00	.28	−.01
Net pay	0	2.80 (±.04)	1.00 (±.09)			1.00	−.07
Depth	0	7.46 (±.01)	.06 (±.00)				1.00

NOTE: Standard errors are in parentheses.

This phenomenon is demonstrated very clearly by the results of a simulation study summarized in Figure 2. The study was based on a (univariate) normal superpopulation model with parameters $\mu = 0$ and $\sigma^2 = 1$. We generated a finite population of size $N = 50$ by iid sampling from this superpopulation. A size-biased sample of size $n = 25$ was then selected from this finite population according to (1.1) with $w(y) = e^y$. We simulated 250 such samples, and from each sample we calculated the sample mean ($\overline{Y}$) and sample variance (S_Y^2). We also computed the maximum likelihood estimates $\hat{\mu}$ and $\hat{\sigma}^2$ under the correct biased sampling model. The theory and computation of the MLE's are discussed in Sections 3 and 4. Figure 2a is the boxplots of $\overline{Y}$ and $\hat{\mu}$, and Figure 2b is the boxplots of S_Y^2 and $\hat{\sigma}^2$ from the 250 simulations. It is clear that the sample mean grossly overestimates $\mu = 0$. The median of the 250 simulated values was .51; in fact, all but one of the 250 sample means were positive. Similarly, the sample variance greatly underestimates $\sigma^2 = 1$. On the other hand, the distributions of $\hat{\mu}$ and $\hat{\sigma}^2$ are centered around their respective true values. See Section 3.5 for additional discussion of the MLE's from this simulation study.

It is obvious from the discussion thus far that one must take into account the biased nature of the discovered data in doing inference. Sections 3 and 4 develop maximum likelihood estimation procedures assuming the sampling model (1.1) and a superpopulation framework. The results are illustrated in Section 5 by reanalyzing the Rimbey–Meadowbrook data

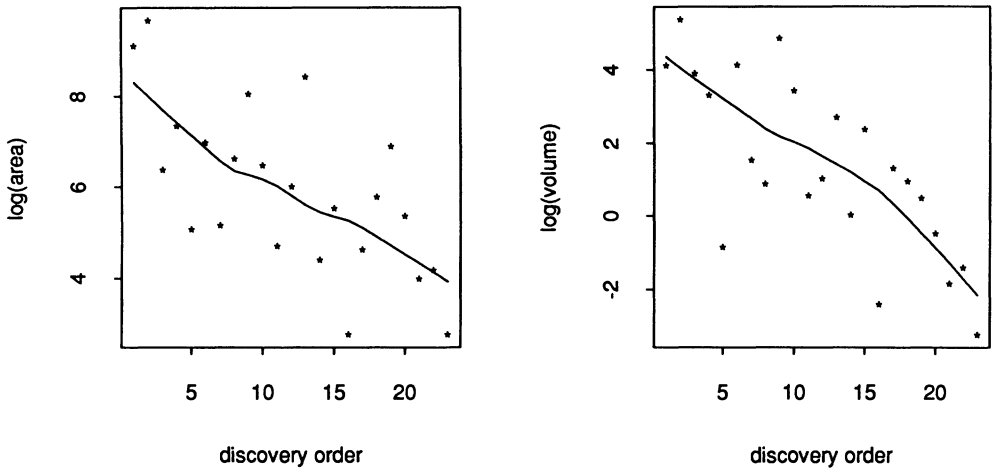

Figure 1. Plots of Log(Area) and Log(Volume) Against Discovery Order. The smooth (lowess) curves show clearly that the larger pools tend to be discovered earlier.

using successive sampling discovery model and the weight function

$$w(\mathbf{y}) = volume^{\gamma_1} \times area^{\gamma_2} \times net\ pay^{\gamma_3} \times depth^{\gamma_4}, \tag{2.2}$$

with $\gamma_1 \equiv 0$. Although this weight function does not depend explicitly on *volume*, inference regarding the *volume* variable is affected by the size bias, since it is correlated with the other variables.

3. MAXIMUM LIKELIHOOD ESTIMATION

3.1 Likelihood Equations

In this section, we derive the likelihood equations arising from the successive sampling discovery model. Let $\mathbf{X}_j$ be the value associated with the jth discovery. We assume the labels are noninformative and so the observed *ordered* sample can be denoted by $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$. We also assume throughout that the *df* F is continuous so that the values associated with the different units of the finite population are distinct with probability 1. Relabel the elements of the finite population so that $\mathbf{X}_j = \mathbf{Y}_j$ $(j = 1, 2, \ldots, n)$. Then, given $\mathbf{Y}_i = \mathbf{y}_i$ $(i = 1, \ldots, N)$, with $\mathbf{y}_i = \mathbf{x}_i$ $(i = 1, \ldots, n)$, the probability of observing the ordered sample $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is

$$\Pr\{(\mathbf{x}_1, \ldots, \mathbf{x}_n) \mid \mathbf{y}_1, \ldots, \mathbf{y}_N\} = \prod_{j=1}^{n} \frac{w(\mathbf{x}_j)}{b_j + w(\mathbf{y}_{n+1}) + \cdots + w(\mathbf{y}_N)}, \tag{3.1}$$

where $b_j = w(\mathbf{x}_j) + \cdots + w(\mathbf{x}_n)$. To get the unconditional distribution of $(\mathbf{X}_1, \ldots, \mathbf{X}_n)$, we have to sum (3.1) over all possible $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$, multiply by the joint density of $(\mathbf{Y}_1, \ldots, \mathbf{Y}_N)$, and integrate over the unobserved values $(\mathbf{Y}_{n+1}, \ldots, \mathbf{Y}_N)$. This gives the joint density of $\mathbf{X}_i = \mathbf{x}_i$ $(i = 1, \ldots, n)$ as

$$\frac{N!}{(N-n)!} \left(\prod_{j=1}^{n} \frac{f_\theta(\mathbf{x}_j) w(\mathbf{x}_j)}{b_j} \right) \times E_\theta \left(\prod_{i=1}^{n} \frac{b_i}{b_i + w(\mathbf{Y}_{n+1}) + \cdots + w(\mathbf{Y}_N)} \right). \tag{3.2}$$

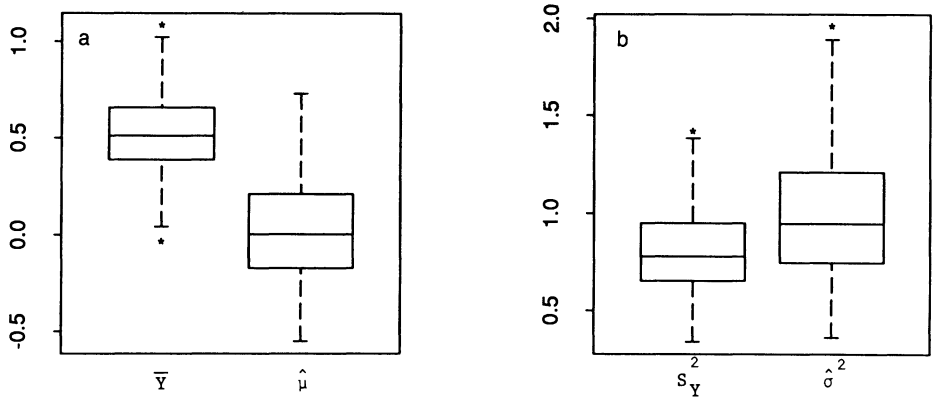

Figure 2. Boxplots of the Estimates of the Parameters μ and σ^2 of a Normal Distribution Under the Size-Biased Sampling Model. The estimates were obtained from a simulation study with $N = 50$, $n = 25$, $w(y) = e^y$, and simulation size 250. The boxplots of the sample mean $(\bar{Y})$ and the MLE $\hat{\mu}$ are given in panel a, and the boxplots of the sample variance (S_Y^2) and the MLE $\hat{\sigma}^2$ are given in panel b.

Let $R = w(\mathbf{Y}_{n+1}) + \cdots + w(\mathbf{Y}_N)$ and $T = \sum_{i=1}^{n} \varepsilon_i/b_i$, where ε_i's are iid standard exponential variates independent of the $\mathbf{Y}_i$'s. Then the expectation term in (3.2) is, in fact,

$$E_{\boldsymbol{\theta}} \prod_{j=1}^{n} \frac{b_j}{b_j + R} = E_R E_T [\exp(-RT)]. \quad (3.3)$$

The second expectation in (3.3) is the Laplace transform of $\mathbf{T}$, a sum of independent exponential random variables.

So let $\phi(t; \boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\{\exp[-tw(\mathbf{Y}_1)]\}$, the Laplace transform of $w(\mathbf{Y}_1)$. By interchanging the expectations in (3.3), the joint density of $\mathbf{X}_1, \ldots, \mathbf{X}_n$ in (3.2) can be written as

$$\frac{N!}{(N-n)!} \prod_{j=1}^{n} \frac{f_{\boldsymbol{\theta}}(\mathbf{x}_j) w(\mathbf{x}_j)}{b_j} \int_0^\infty [\phi(t; \boldsymbol{\theta})]^{N-n} g_n(t)\, dt, \quad (3.4)$$

where $g_n(t)$ is the general gamma or Erlang density (Johnson and Kotz 1970, p. 222) of T,

$$g_n(t) = \sum_{i=1}^{n} c_i [b_i e^{-tb_i}], \quad t > 0, \quad (3.5)$$

and

$$c_i = \prod_{k \neq i} b_k/(b_k - b_i). \quad (3.6)$$

This density is itself a linear combination of exponential densities and equals 0 at the origin since $\sum_{i=1}^{n} c_i b_i = 0$. It depends on the data through the partial sums $b_j = \sum_{i=j}^{n} w(\mathbf{x}_i)$ and is sensitive to the order in which the data are observed.

Define

$$S(\boldsymbol{\theta}) = \int_0^\infty [\phi(t; \boldsymbol{\theta})]^{N-n} g_n(t)\, dt. \quad (3.7)$$

Then the log-likelihood of $\boldsymbol{\theta}$ given the data is

$$\log L(\boldsymbol{\theta}) = \text{constant} + \sum_{j=1}^{n} \log f_{\boldsymbol{\theta}}(\mathbf{x}_j) + \log S(\boldsymbol{\theta}). \quad (3.8)$$

By differentiating (3.8), we get the likelihood equations

$$\sum_{j=1}^{n} \frac{\partial}{\partial \theta_r} \log f_{\boldsymbol{\theta}}(\mathbf{x}_j) + (N-n) \int_0^\infty \left[\frac{\partial}{\partial \theta_r} \log \phi(t; \boldsymbol{\theta})\right] h_{\boldsymbol{\theta}}(t)\, dt = 0 \quad (3.9)$$

for $r = 1, \ldots, m$, where $h_{\boldsymbol{\theta}}(t)$ is the data-dependent density function

$$h_{\boldsymbol{\theta}}(t) = [\phi(t; \boldsymbol{\theta})]^{N-n} g_n(t)/S(\boldsymbol{\theta}). \quad (3.10)$$

Remark 3.1. If $w(\mathbf{y}) \equiv 1$ or if $n = N$, $S(\boldsymbol{\theta})$ is a constant independent of $\boldsymbol{\theta}$ so that the likelihood reduces to the usual simple random sampling model based on n iid observations from the *df* $F_{\boldsymbol{\theta}}$. On the other hand, if we fix n and let $N \to \infty$ (so that $n/N \to 0$), the likelihood tends to the one based on n iid observations from the biased sampling model (1.2).

3.2 EM Algorithm

Neither the log-likelihood (3.8) nor the likelihood equations (3.9) can be expressed explicitly in general. Barouch and Kaufman (1976, 1977) used an asymptotic expansion to approximate the likelihood function in the special univariate lognormal situation with $w(y) = y$. This expansion is involved, and its usage in practice appears difficult. Using the Newton–Raphson procedure to numerically solve the likelihood equations would involve the evaluation of $(m^2 + 3m + 2)/2$ double integrals for *each iteration*—one for the log-likelihood in (3.8), m for the score functions in (3.9), and $m + [m(m-1)]/2$ for the information matrix $I_0(\boldsymbol{\theta})$. Quasi Newton methods using only the first derivatives would be computationally preferable. In this section, we discuss the EM algorithm for computing the MLE's. See also Barouch et al. (1983) for a special case. The application of the EM technique to this problem is particularly appealing because of its interpretability.

Let $h_{\boldsymbol{\theta}}(t)$ be the density given by (3.10). Define another density function

$$k(\mathbf{y} \mid t, \boldsymbol{\theta}) = \exp\{-tw(\mathbf{y})\} f_{\boldsymbol{\theta}}(\mathbf{y})/\phi(t; \boldsymbol{\theta}). \quad (3.11)$$

Under the successive-sampling-discovery model, the conditional distribution of the remaining $\mathbf{Y}_{n+1}, \ldots, \mathbf{Y}_N$, given the data $(\mathbf{x}_1, \ldots \mathbf{x}_n)$, is

$$\begin{aligned} & f_{\boldsymbol{\theta}}(\mathbf{y}_{n+1}, \ldots, \mathbf{y}_N \mid \text{data}) \\ &= \prod_{j=1}^{n} \{b_j/[b_j + w(\mathbf{y}_{n+1}) + \cdots + w(\mathbf{y}_N)]\} \\ &\quad \times \prod_{i=n+1}^{N} f_{\boldsymbol{\theta}}(\mathbf{y}_i)/S(\boldsymbol{\theta}) \\ &= \int_0^\infty \prod_{i=n+1}^{N} e^{-tw(y_i)} f_{\boldsymbol{\theta}}(\mathbf{y}_i) g_n(t)\, dt/S(\boldsymbol{\theta}) \\ &= \int_0^\infty \prod_{i=n+1}^{N} k(\mathbf{y}_i \mid t, \boldsymbol{\theta}) h_{\boldsymbol{\theta}}(t)\, dt. \end{aligned} \quad (3.12)$$

This density is symmetric in its arguments and is a mixture of a product density. From this, the conditional distribution of $\mathbf{Y}_{n+1}$ given the data is obtained as

$$f_{\boldsymbol{\theta}}(\mathbf{y} \mid \text{data}) = \int_0^\infty k(\mathbf{y} \mid t, \boldsymbol{\theta}) h_{\boldsymbol{\theta}}(t)\, dt. \quad (3.13)$$

Note that

$$\frac{\partial}{\partial\theta_r}[\log\phi(t;\boldsymbol{\theta})] = \int\left[\frac{\partial}{\partial\theta_r}\log f_{\boldsymbol{\theta}}(\mathbf{y})\right]\exp(-tw(\mathbf{y})) \times f_{\boldsymbol{\theta}}(\mathbf{y})\,d\mathbf{y}/\phi(t;\boldsymbol{\theta}) = \int\left[\frac{\partial}{\partial\theta_r}\log f_{\boldsymbol{\theta}}(\mathbf{y})\right]k(\mathbf{y}\mid t,\boldsymbol{\theta})\,d\mathbf{y}. \quad (3.14)$$

So the integral in (3.9) can be written as

$$\int_0^\infty\left[\frac{\partial}{\partial\theta_r}\log\phi(t;\boldsymbol{\theta})\right]h_{\boldsymbol{\theta}}(t)\,dt = \int_0^\infty\left\{\int\left[\frac{\partial}{\partial\theta_r}\log f_{\boldsymbol{\theta}}(\mathbf{y})\right]k(\mathbf{y}\mid t,\boldsymbol{\theta})h_{\boldsymbol{\theta}}(t)\,d\mathbf{y}\right\}dt = \int\left[\frac{\partial}{\partial\theta_r}\log f_{\boldsymbol{\theta}}(\mathbf{y})\right]\left(\int_0^\infty k(\mathbf{y}\mid t,\boldsymbol{\theta})h\boldsymbol{\theta}(t)\,dt\right)d\mathbf{y} = E_{\boldsymbol{\theta}}\left(\left[\frac{\partial}{\partial\theta_r}\log f_{\boldsymbol{\theta}}(\mathbf{Y})\right]\,\middle|\,\text{data}\right). \quad (3.15)$$

Therefore, the likelihood equations (3.9) are in fact given by

$$E_{\boldsymbol{\theta}}\left\{\sum_{j=1}^{N}\left[\frac{\partial}{\partial\theta_r}\log f_{\boldsymbol{\theta}}(\mathbf{Y}_j)\right]\,\middle|\,\text{data}\right\} = 0. \quad (3.16)$$

This expression has a nice interpretation; if all of the $\mathbf{Y}_k$'s in the finite population are known, we would just solve $\sum_{k=1}^{N}(\partial/\partial\theta_r)\log f_{\boldsymbol{\theta}}(\mathbf{y}_k) = 0$ to get the maximum likelihood estimates; since some of the $\mathbf{Y}_k$'s are unobservable, we solve instead for its expectation, given the data $\mathbf{x}_1, \ldots, \mathbf{x}_n$. This interpretation is precisely the idea behind the EM algorithm (Dempster, Laird, and Rubin 1977) for incomplete data.

Let

$$Q(\boldsymbol{\theta}\mid\boldsymbol{\theta}^*) = E_{\boldsymbol{\theta}^*}\left\{\left[\sum_{j=1}^{N}\log f_{\boldsymbol{\theta}}(\mathbf{Y}_j)\right]\,\middle|\,\text{data}\right\} = \sum_{j=1}^{N}\log f_{\boldsymbol{\theta}}(\mathbf{x}_j) + (N-n)E_{\boldsymbol{\theta}^*}\{\log f_{\boldsymbol{\theta}}(\mathbf{Z})\mid\text{data}\}, \quad (3.17)$$

where $\mathbf{Z}$ denotes any one of the $(N-n)$ unobserved variables $\mathbf{Y}_{n+1}, \ldots, \mathbf{Y}_N$. The conditional distribution of $\mathbf{Z}$, given the data, is given by (3.13). The EM algorithm for our problem is then given by the following:

1. Start with an initial value $\boldsymbol{\theta}^{(0)}$.
2. Given $\boldsymbol{\theta}^{(p)}$ $(p = 0, 1, \ldots)$, obtain $\boldsymbol{\theta}^{(p+1)}$ as follows:

E Step. Compute $Q(\boldsymbol{\theta}\mid\boldsymbol{\theta}^{(p)})$ in (3.17).

M Step. Set $\boldsymbol{\theta}^{(p+1)}$ to be a value of $\boldsymbol{\theta}$ that maximizes $Q(\boldsymbol{\theta}\mid\boldsymbol{\theta}^{(p)})$ by solving

$$\frac{\partial}{\partial\theta_r}Q(\boldsymbol{\theta}\mid\boldsymbol{\theta}^{(p)}) = \frac{n}{N}\left\{\frac{1}{n}\sum_{j=1}^{n}\frac{\partial}{\partial\theta_r}\log f_{\boldsymbol{\theta}}(\mathbf{x}_j)\right\} + \left(1-\frac{n}{N}\right)E_{\boldsymbol{\theta}^{(p)}}\left\{\left[\frac{\partial}{\partial\theta_r}\log f_{\boldsymbol{\theta}}(\mathbf{Z})\right]\,\middle|\,\text{data}\right\} = 0 \quad (3.18)$$

for $r = 1, \ldots, m$.

3. Repeat the iteration until the log-likelihood (3.9) stops improving.

Under fairly general conditions, the EM iterations $\{\boldsymbol{\theta}^{(p)}\}$ increase the likelihood monotonically and the likelihood converges to a stationery point (Dempster et al. 1977; Wu 1983). If the likelihood is unimodal with only one stationary point, $\hat{\boldsymbol{\theta}}$ will be the unique MLE. Otherwise, we can only determine if it corresponds to a local maximum. In all our applications, however, we repeated the EM algorithm with several starting points, and in all cases the algorithm converged to the same estimates.

Although the EM algorithm can be used with any distribution, the computations can be prohibitive in general. Fortunately, as described by Dempster et al. (1977), the algorithm simplifies considerably when the underlying df F is a member of the regular exponential family

$$f_{\boldsymbol{\theta}}(\mathbf{y}) = b(\mathbf{y})\exp\{\boldsymbol{\theta}'\mathbf{s}(\mathbf{y})\}/a(\boldsymbol{\theta}), \quad (3.19)$$

where $\boldsymbol{\theta}\in\boldsymbol{\Theta}$, an m-dimensional convex set, and $\mathbf{s}(\mathbf{y})$ is an m-dimensional vector of sufficient statistics. In this case, the EM iteration is based on the complete-data sufficient statistic $\mathbf{s}(\mathbf{Y})$ and is given as follows:

E Step. Given $\boldsymbol{\theta}^{(p)}$, compute

$$\mathbf{s}^{(p)} = \frac{n}{N}\bar{\mathbf{s}} + \left(1-\frac{n}{N}\right)E_{\boldsymbol{\theta}^{(p)}}\{\mathbf{s}(\mathbf{Z})\mid\text{data}\}, \quad (3.20)$$

where $\bar{\mathbf{s}} = (1/n)\sum_{j=1}^{n}\mathbf{s}(\mathbf{x}_j)$.

M Step. Obtain $\boldsymbol{\theta}^{(p+1)}$ as the solution in $\boldsymbol{\theta}$ of the equation

$$E_{\boldsymbol{\theta}}\{\mathbf{s}(\mathbf{Y})\} = \mathbf{s}^{(p)},$$

where

$$E_{\boldsymbol{\theta}}\{\mathbf{s}(\mathbf{Y})\} = \left[\frac{\partial}{\partial\theta_1}\log a(\boldsymbol{\theta}), \ldots, \frac{\partial}{\partial\theta_m}\log a(\boldsymbol{\theta})\right]'. \quad (3.21)$$

Note the structure and the intuitive interpretation behind (3.20). The estimator is a convex combination of the observed quantity and the conditional expectation of the unobserved quantity. This is similar to the superpopulation estimators in survey sampling. The relationship between (3.20) and the regression-type estimators in survey sampling will be discussed in Section 4.

Example 3.1. Consider the gamma distribution with density

$$f_{\boldsymbol{\theta}}(y) = \beta^{\alpha} y^{\alpha-1} e^{-\beta y}/\Gamma(\alpha), \qquad Y > 0, \boldsymbol{\theta} = (\alpha, \beta).$$

Let

$$s_1^{(p)} = \frac{n}{N}\left[\frac{1}{n}\sum_{j=1}^{N} x_j\right] + \left(1 - \frac{n}{N}\right) E_{\boldsymbol{\theta}^{(p)}}\{Z \mid \text{data}\}$$

$$s_2^{(p)} = \frac{n}{N}\left[\frac{1}{n}\sum_{j=1}^{N} \log x_j\right] + \left(1 - \frac{n}{N}\right) E_{\boldsymbol{\theta}^{(p)}}\{\log Z \mid \text{data}\}, \tag{3.22}$$

and let $\psi(\alpha) = (d/d\alpha)\log\Gamma(\alpha)$, the digamma function. Then the EM algorithm, given $\alpha^{(p)}$ and $\beta^{(p)}$, is as follows:

1. Determine $\alpha^{(p+1)}$ as the solution of α of

$$\log\alpha - \psi(\alpha) = \log s_1^{(p)} - s_2^{(p)}. \tag{3.23}$$

2. Take $\beta^{(p+1)} = \alpha^{(p+1)}/s_1^{(p)}$.

One iteration of steps 1 and 2 solves the usual likelihood equations in the complete data case. The trigamma function $\psi'(\alpha)$ is needed if Newton's method is used for solving Equation (3.23). An alternative is to use the approximation discussed by Johnson and Kotz (1970, p. 189). Computational issues in evaluating the conditional expectations of Z and $\log Z$ are discussed in Section 3.5.

Example 3.2. Consider the K-dimensional multivariate normal distribution with density

$$f_{\boldsymbol{\theta}}(\mathbf{y}) = (2\pi)^{-K/2}|\boldsymbol{\Sigma}|^{-1/2} \times \exp\{-\tfrac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\boldsymbol{\mu})\}, \qquad \boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Lognormal distributions have been used to model the pool distributions in petroleum exploration. This example can be applied to such cases by transforming the data. See Section 5.

Recall that $\mathbf{x}_i = (x_{i1}, \ldots, x_{iK})'$ $(i = 1, \ldots, n)$ denote the observations in order of appearance, and let $m_k = n^{-1}\sum_{i=1}^{n} x_{ik}$ $(k = 1, \ldots, K)$ and $s_{jk}(\mu_j, \mu_k) = n^{-1}\sum_{i=1}^{n}[(x_{ij} - \mu_j)(x_{ik} - \mu_k)]$ $(j, k = 1, \ldots, K)$. Let σ_{jk} denote the (j, k)th element of $\boldsymbol{\Sigma}$. Then the EM iteration $\boldsymbol{\theta}^{(p)} \to \boldsymbol{\theta}^{(p+1)}$ is given by

$$\mu_k^{(p+1)} = (n/N)m_k + [1 - (n/N)]E_{\boldsymbol{\theta}^{(p)}}\{Z_k \mid \text{data}\}$$

$$\sigma_{jk}^{(p+1)} = (n/N)s_{jk}(\mu_j^{(p+1)}, \mu_k^{(p+1)}) + (1 - (n/N))E_{(\mu^{(p+1)}, \Sigma^{(p)})} \times \{(Z_j - \mu_j^{(p+1)})(Z_k - \mu_k^{(p+1)}) \mid \text{data}\}, \quad j, k = 1, \ldots, K. \tag{3.24}$$

Suppose, for some i, $w(\mathbf{y})$ does not depend on y_i, *and* Y_i is independent of the other Y_j's. Then inference about the marginal distribution of Y_i will be unaffected by the sampling bias. In this case, $E_{\boldsymbol{\theta}^{(p)}}\{Z_1 \mid \text{data}\} = \hat{\mu}_1^{(p)}$ in (3.24), and so the EM iterations will converge to the usual estimates $\hat{\mu}_j = m_i$ and $\hat{\sigma}_{ii} = n^{-1}\sum_{j=1}^{n}(x_{ji} - m_i)^2$, as they should. On the other hand, as long as Y_i is not independent of the other Y_j's, inference about μ_i and σ_{ii} will be affected even if the biased sampling mechanism does not explicitly depend on y_i.

3.3 Interval Estimation

The inverse of the observed Fisher information matrix at $\hat{\boldsymbol{\theta}}$ is often used to estimate the variance–covariance matrix of the MLE. For the general incomplete data problem considered in Dempster et al. (1977), it can be shown under suitable regularity conditions that the (r, s) entry of the observed Fisher information matrix is given by

$$I_{0,rs}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\left\{-\sum_{k=1}^{N}\frac{\partial^2}{\partial\theta_r\partial\theta_s}\log f_{\boldsymbol{\theta}}(\mathbf{Y}_k) \mid \text{data}\right\} - \operatorname{cov}_{\boldsymbol{\theta}}\left\{\sum_{k=1}^{N}\frac{\partial}{\partial\theta_r}\log f_{\boldsymbol{\theta}}(\mathbf{Y}_k), \sum_{j=1}^{N}\frac{\partial}{\partial\theta_s}\log f_{\boldsymbol{\theta}}(\mathbf{Y}_j) \mid \text{data}\right\}. \tag{3.25}$$

Therefore, $I_0(\boldsymbol{\theta})$ is just the difference between the conditional expectation of the complete-data information matrix and the conditional covariance of the complete data-score functions given the data. For the regular exponential family (3.19), let $\mathbf{S} = \sum_{k=1}^{N}\mathbf{s}(\mathbf{Y}_k)$ be the m-dimensional vector of complete data-sufficient statistics. Then the observed Fisher information matrix in (3.25) can be expressed simply as

$$I_0(\boldsymbol{\theta}) = \operatorname{cov}_{\boldsymbol{\theta}}(\mathbf{S}) - \operatorname{cov}_{\boldsymbol{\theta}}(\mathbf{S} \mid \text{data}) \tag{3.26}$$

(Dempster et al. 1977). Computational issues involved in calculating (3.25) and (3.26) are discussed in Section 3.5.

If the MLE's have a limiting normal distribution, we can estimate the standard errors in (3.26) and use the normal approximation to construct confidence intervals. But the usual results on asymptotic nor-

mality of the MLE's are not applicable here since the likelihood in (3.4) is based on data that are neither identically nor independently distributed. We have looked at this problem but have not been able to establish the asymptotic normality of the MLE's. This is an issue that merits further investigation. We conclude this section by summarizing the results of a small simulation study to examine the adequacy of using a normal approximation.

The details of this simulation study are the same as those reported in Section 2. We generated a finite population of size $N = 50$ by iid sampling from a (univariate) lognormal distribution with parameters $\mu = 0$ and $\sigma^2 = 1$. A sample of size $n = 25$ was selected from the finite population according to (1.1) with $w(y) = y$. We computed the MLE's, $\hat{\mu}$ and $\hat{\sigma}^2$, from 250 such simulations.

The boxplots of the simulated $\hat{\mu}$'s and $\hat{\sigma}^2$'s are given in Figure 2. The means (medians) of the 250 simulated values were .02 (.00) and .99 (.95), respectively, so the distributions are centered around the true values of $\mu = 0$ and $\sigma^2 = 1$. To check the adequacy of the normal approximations, we examined the histograms (not shown here) and normal Q–Q plots. Figure 3 is the normal Q–Q plots of the $\hat{\mu}$'s and log-$\hat{\sigma}^2$'s. The log-transformation improved the normal approximation for the $\hat{\sigma}^2$'s. There is a slight asymmetry in the distributions (this was more evident in the histograms), and the fit is not as good near the tails. Overall, the normal approximations seem to be reasonable.

We also computed the estimated standard errors of $\hat{\mu}$ and $\hat{\sigma}$ using (3.26). The averages of these 250 values were .24 and .17, respectively. These were in close agreement with the sample standard deviations of the 250 simulated values of $\hat{\mu}$ and $\hat{\sigma}$—.23 and .18, respectively—so there was no significant bias in using (3.26) to compute the standard errors.

Table 3 compares the nominal and observed levels of the confidence intervals based on the standard errors and the normal approximation. Intervals shown are those based on $\hat{\mu}$ and log $\hat{\sigma}^2$. Results for $\hat{\sigma}^2$ were qualitatively similar to those based on log $\hat{\sigma}^2$. Since the simulations are based on (only) a sample of size 250, the standard error of the observed values in Table 3 is about $\pm .02$. The observed levels are larger than the nominal levels in all cases, suggesting that the tails of the standardized variables are somewhat heavier than a standard normal. The intervals for μ perform slightly better than those for σ^2. Overall, the approximations are not unreasonable. These results, however, are rather preliminary, and further work is needed to examine this issue in more detail.

3.4 Computational Issues

The computation of the likelihood, the likelihood equations, and the information matrix all involve evaluating integrals of the form

$$H(\boldsymbol{\theta}) = \int_0^\infty H(t; \boldsymbol{\theta}) g_n(t)\, dt \qquad (3.27)$$

for some $H(t; \boldsymbol{\theta})$. The function $H(t; \boldsymbol{\theta})$ involves the Laplace transform $\phi(t; \boldsymbol{\theta})$ or its partial derivatives. The main difficulty in computing (3.27) is the accurate evaluation of the density $g_n(t)$, particularly for values of t near 0. Note that

$$g_n(t) = \prod_{j=1}^{n} b_j \frac{t^{n-1}}{(n-1)!} + O(t^n) \qquad (3.28)$$

for t near 0, so the values are close to 0 if n is large. Further, the ratios of the coefficients c_i's in (3.6) can

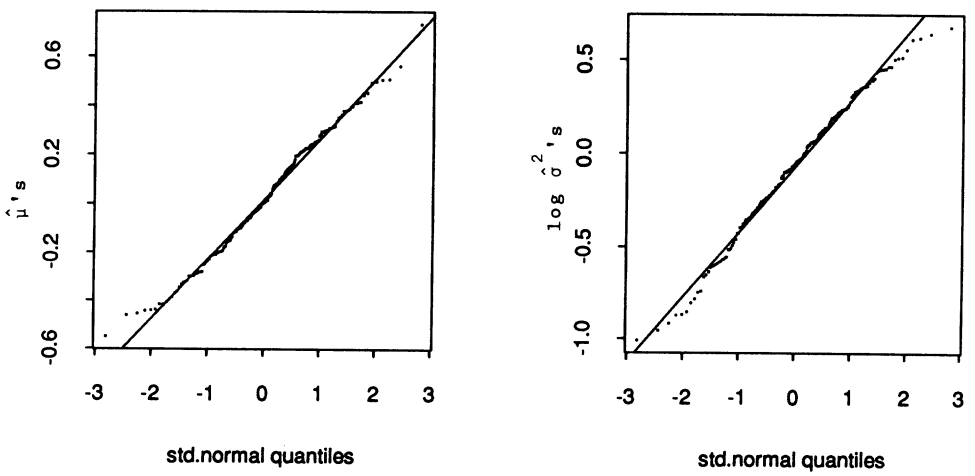

Figure 3. Normal Q–Q Plots of the MLE's $\hat{\mu}$ and Log $\hat{\sigma}^2$. The estimates were obtained from a simulation study (see the description of Fig. 2 for details). The straight lines correspond to the best-fitting normal distributions with parameters equal to the sample mean and standard deviation of the simulated values.

Table 3. Comparisons of Nominal and Observed Levels of the Confidence Intervals for μ and σ^2 for a Lognormal Superpopulation

Nominal level	Observed level	
	$\hat{\mu}$	$\hat{\sigma}^2$
.01	.02	.06
.05	.08	.12
.10	.16	.15

NOTE: The results are based on a simulation study with $N = 50$, $n = 25$, and $w(y) = y$. The intervals for σ^2 were based on log σ^2 as the pivot.

be large, and they alternate in sign so that a large number of cancellations will occur near $t = 0$. Because of round-off errors, Expressions (3.5)–(3.6) are practically useless for computing the density $g_n(t)$ for t near 0. In the special case where $w(\mathbf{y}) \equiv 1$, $b_i = n - i - 1$ and

$$b_i c_i = (-1)^{n-1} n \binom{n-1}{i-1}. \tag{3.29}$$

In this case, it can be shown that

$$S(\boldsymbol{\theta}) = n \int_0^1 z^{N-n}(1-z)^{n-1}\, dz = 1 \Big/ \binom{N}{n}, \tag{3.30}$$

which can be a very small number. These observations suggest that in general we must avoid calculating the partial fractions coefficients in (3.6) in evaluating $g_n(t)$.

Among the different methods available in the literature (Davies 1980; Sheil and O'Muircheartaigh 1977), we found the one based on the inverse Laplace transform (Crump 1976) to perform best. The Laplace transform of $g_n(t)$ is

$$L_n(s) = \prod_{j=1}^{n} \frac{b_j}{b_j + s}, \tag{3.31}$$

and the inverse transform is given by the well-known inversion formula

$$g_n(t) = (e^{Dt}/\pi) \times \int_0^\infty [\mathrm{Re}\{L_n(s)\} \cos wt - \mathrm{Im}\{L_n(s)\} \sin wt]\, dw, \tag{3.32}$$

where $s = D + iw$ and D is any real number such that $L_n(s)$ is analytic for $\mathrm{Re}(s) > D$. Since $g_n(t)$ is of exponential order $-b_n$—that is, $g_n(t) \le Me^{-b_n t}$—this inverse transform can be well approximated on a compact interval by a Fourier series approximation. See Crump (1976) for details. The IMSL (1984) subroutine FLINV can be used for this purpose. Choosing D in (3.32) (called ALPHA in FLINV) appropriately to achieve the desired accuracy can be quite delicate, however. We used an iterative scheme to choose D.

The function $H(t; \boldsymbol{\theta})$ in (3.27) involves the Laplace transform $\phi(t; \boldsymbol{\theta})$ or its partial derivatives. For the gamma and some other distributions, these can be evaluated explicitly. For others, such as the lognormal, they too have to be evaluated numerically. We used the Gauss–Hermite quadrature method to compute this for the lognormal distribution for the Rimbey–Meadowbrook application in Section 5. This particular quadrature method takes advantage of the e^{-x^2} component in the lognormal distribution. For other distributions, different numerical methods should be considered. To compute the integrals of the form (3.27), we used the Gauss–Legendre quadrature method. Specifically, for a given data set, we chose a set of meshpoints carefully within the important domain of the data-dependent density $g_n(t)$. Then we evaluated $H(t; \boldsymbol{\theta})$ and $g_n(t)$ at these points and approximated the integral by a partial sum. The obvious advantage of this strategy is that the $g_n(t_i)$'s need to be evaluated only once for all the different iterations in the EM algorithm. Ideally, we would like to let the meshpoints depend on the value of $\boldsymbol{\theta}$ and the particular function $H(t; \boldsymbol{\theta})$ that is being evaluated. Since $g_n(t)$ is very expensive to evaluate, this is not computationally feasible.

We have only considered numerical integration techniques in our computations. It is also possible to use Monte Carlo integration techniques to compute the integrals involved. We have not investigated this approach.

4. TWO-STAGE ESTIMATION

In the multivariate case, the direct computation of the MLE's can be quite formidable. For example, for the multivariate (log)normal problem in Section 3, one has to solve $K + K(K+1)/2$ iterative equations in (3.24), K for the means and $K(K+1)/2$ for the variance–covariance matrix. In this section, we show that the estimation problem can be simplified considerably under certain conditions.

Suppose that the original problem can be reparameterized (smoothly) from $\boldsymbol{\theta}$ to $(\boldsymbol{\eta}, \boldsymbol{\tau}) \in \boldsymbol{\Omega}_1 \times \boldsymbol{\Omega}_2$ so that the marginal distribution of $W = w(\mathbf{Y})$ depends on $\boldsymbol{\eta} \in \boldsymbol{\Omega}_1$ but not on $\boldsymbol{\tau} \in \boldsymbol{\Omega}_2$. Further, the conditional distribution of $\mathbf{Y}$ given $W = w$ depends on $\boldsymbol{\tau}$ but not on $\boldsymbol{\eta}$. Then W is ancillary for $\boldsymbol{\tau}$ and conditional on $W = w$ and $\mathbf{Y}$ is conditionally sufficient for $\boldsymbol{\tau}$ in the presence of $\boldsymbol{\eta}$ (Cox and Hinkley 1974, p. 35). Since $f_{\boldsymbol{\theta}}(\mathbf{y}) = f_{\boldsymbol{\eta}}(w) f_{\boldsymbol{\tau}}(\mathbf{y} \mid w)$, the log-

likelihood (3.8) is equivalent to

$$\sum_{j=1}^{n} \log f_\tau(\mathbf{y}_j \mid w_j) + \sum_{j=1}^{n} \log f_\eta(w_j) + \log S(\boldsymbol{\eta}), \tag{4.1}$$

where $S(\boldsymbol{\eta})$ given by (3.7) now depends only on $\boldsymbol{\eta}$. So we can estimate $\boldsymbol{\theta}$ by estimating $\boldsymbol{\eta}$ and $\boldsymbol{\tau}$ separately. Estimation of $\boldsymbol{\eta}$ involves the biased likelihood (3.8), but it only depends on the univariate data w_j ($j = 1, \ldots, n$). The estimation of $\boldsymbol{\tau}$ involves only the (conditional) likelihood of $\mathbf{Y}_j$ given $W_j = w_j$ ($j = 1, \ldots, n$). This conditional distribution, however, is unaffected by the biased sampling mechanism involving the w_j's *regardless of what it is*. So we can use the usual techniques to compute the MLE of $\boldsymbol{\tau}$.

When the likelihood can be separated as in (4.1), the information matrix $I_0(\hat{\boldsymbol{\theta}})$ needed for confidence intervals can also be computed more easily. Let $I_0^{(1)}(\boldsymbol{\eta})$ and $I_0^{(2)}(\boldsymbol{\tau})$ be, respectively, the observed information matrices for $\boldsymbol{\eta}$ and $\boldsymbol{\tau}$. We assume that the reparameterization is smooth enough so that $\boldsymbol{\eta}$ and $\boldsymbol{\tau}$ can be differentiated with respect to $\boldsymbol{\theta}$ at least twice. Define

$$\dot{\eta}_r(\boldsymbol{\theta}) = \partial \boldsymbol{\eta}(\boldsymbol{\theta})/\partial \theta_r \tag{4.2}$$

and

$$\dot{\tau}_r(\boldsymbol{\theta}) = \partial \boldsymbol{\tau}(\boldsymbol{\theta})/\partial \theta_r \tag{4.3}$$

for $r = 1, \ldots, m$. Then it can be shown that

$$I_{0,rs}(\hat{\boldsymbol{\theta}}) = \dot{\boldsymbol{\eta}}_r'(\hat{\boldsymbol{\theta}}) I_0^{(1)}(\hat{\boldsymbol{\eta}}) \dot{\boldsymbol{\eta}}_s(\hat{\boldsymbol{\theta}}) + \dot{\boldsymbol{\tau}}_r'(\hat{\boldsymbol{\theta}}) I_0^{(2)}(\hat{\boldsymbol{\tau}}) \dot{\boldsymbol{\tau}}_s(\hat{\boldsymbol{\theta}}). \tag{4.4}$$

As noted before, $I_0^{(2)}(\hat{\boldsymbol{\tau}})$ can be obtained in the usual manner from the (conditional) likelihood of the $\mathbf{Y}_i$'s given the w_i's. It is much easier to compute $I_0^{(1)}(\hat{\boldsymbol{\eta}})$ which involves only the marginal distribution of the w_i's.

Before discussing some applications, note that the conditions necessary for the preceding separate estimation of the parameters are not always satisfied. For example, consider the case in which $\mathbf{Y} = (Y_1 + Y_2)$ and the Y_i's are independent with a gamma distribution with parameters θ_1 and θ_2. Let $w(\mathbf{Y}) = Y_1 + Y_2$. Then W has a gamma distribution with parameter $\eta = \theta_1 + \theta_2$, and the conditional distribution of $\mathbf{Y}$ given $W = w$ is based on the beta distribution with parameter $\boldsymbol{\tau} = (\theta_1, \boldsymbol{\eta} - \theta_1)$. Here $\boldsymbol{\tau}$ depends on $\boldsymbol{\eta}$, so the separate estimation described previously would not work in this case.

Example 4.1: Multivariate Regression. Consider again the multivariate normal Example 3.2 with $\boldsymbol{\theta} = (\boldsymbol{\mu}, \Sigma)$, and suppose that $w(\mathbf{y}) = \exp(\boldsymbol{\gamma}'\mathbf{y})$ with $\boldsymbol{\gamma}$ assumed known. Then $w(\mathbf{Y})$ is lognormal with parameter $\boldsymbol{\eta} = (\mu_w, \sigma_{ww})$, where $\mu_w = \boldsymbol{\gamma}'\boldsymbol{\mu}$ and $\sigma_{ww} = \boldsymbol{\gamma}'\Sigma\boldsymbol{\gamma}$. The conditional distribution of $\mathbf{Y}$ given $w(\mathbf{Y}) = w$ is multivariate normal with parameters $\boldsymbol{\tau} = (\boldsymbol{\nu}, \Gamma)$, where $\boldsymbol{\nu} = \boldsymbol{\mu} + \Sigma\boldsymbol{\gamma}(\log w - \mu_w)/\sigma_{ww}$ and $\Gamma = \Sigma - \Sigma\boldsymbol{\gamma}\boldsymbol{\gamma}'\Sigma/\sigma_{ww}$. The covariance matrix Γ is singular, since the conditional distribution of $\mathbf{Y}$ given $w(\mathbf{Y}) = w$ can be expressed in terms of any $(k - 1)$ of the Y_j's.

We first solve (3.24) for the univariate data w_j's to get the MLE's μ_w and σ_{ww}. For maximum likelihood estimation of $\boldsymbol{\tau}$, we just need to consider the multivariate simple linear regression model

$$\mathbf{Y}_i = \boldsymbol{\alpha} + \boldsymbol{\beta} \log w_i + \boldsymbol{\epsilon}_i, \qquad i = 1, \ldots, n, \tag{4.5}$$

where $\boldsymbol{\alpha} = \boldsymbol{\mu} - \Sigma\boldsymbol{\gamma}\mu_w/\sigma_{ww}$ and $\boldsymbol{\beta} = \Sigma\boldsymbol{\gamma}/\sigma_{ww}$ and the $\boldsymbol{\epsilon}_i$'s are iid multivariate normal with mean 0 and covariance matrix Γ. The maximum likelihood estimate of $\boldsymbol{\tau}$ is obtained by estimating the regression parameters in (4.5) through least squares. Let $m_w = n^{-1} \sum_{i=1}^{n} \log w_i$, and $S_{kw} = n^{-1} \sum_{i=1}^{n} [(x_{ik} - m_k)(\log w_i - m_w)]$, and let S_{ww} be similarly defined. Then, $\hat{\beta}_k = S_{kw}/S_{ww}$, $\hat{\alpha}_k = m_k - \hat{\beta}_k m_w$, and $\hat{\Gamma}_{kj} = S_{kj} - \hat{\beta}_k \hat{\beta}_j S_{ww}$. By reparameterizing, we get the MLE of $\boldsymbol{\theta}$ as

$$\hat{\mu}_k = m_k + \hat{\beta}_k(\hat{\mu}_w - m_w) \tag{4.6}$$

and

$$\hat{\sigma}_{kj} = S_{kj} + \hat{\beta}_k \hat{\beta}_j (\hat{\sigma}_{ww} - S_{ww}). \tag{4.7}$$

Note that $\sum_{j=1}^{K} \gamma_j \hat{\mu}_j = \hat{\mu}_w$, $\sum_{j=1}^{K} \gamma_j \hat{\sigma}_{kj} = \hat{\sigma}_{kw}$, and $\sum_{k=1}^{K} \gamma_k \hat{\sigma}_{kw} = \hat{\sigma}_{ww}$, as they should.

The estimators in (3.24) reduce to these for the special form of $w(\mathbf{y})$ discussed here. Note also the similarity of the estimator $\hat{\mu}_k$ to regression estimators in survey sampling. In fact, if $w_1, \ldots, w_N$ are all known, as would be the case in survey sampling, $\hat{\mu}_w = N^{-1} \sum_{i=1}^{N} \log w_j$, so $\hat{\mu}_k$ does in fact reduce to the regression estimator in survey sampling. The expression for $\hat{\mu}_k$ in (4.6) is more general in that it holds for *any* sampling scheme involving the w_j's.

When $N = n$, $\hat{\mu}_w = m_w$ and $\hat{\sigma}_{ww} = S_{ww}$, so (4.6) and (4.7) reduce to the usual simple random sampling estimators. On the other hand, if $N \to \infty$ with n fixed, it is known (Scheaffer 1972) that the size-biased distribution of $w(\mathbf{Y})$ is itself lognormal with parameters $\mu_{ww} + \sigma_{ww}$ and σ_{ww}. Hence $\hat{\sigma}_{ww} = S_{ww}$ and $\hat{\mu}_w = m_w - S_{ww}$. From this, we see that (4.3) and (4.4) reduce to $\hat{\mu}_k = m_k - S_{kw}$ and $\sigma_{kj} = S_{kj}$, as they should. The estimators can be computed easily in this case.

Example 4.2: Binary Regression. Let $\mathbf{Y} = (Y_1, Y_2)$ with Y_1 a binary random variable, and suppose that the weight function $w(\cdot)$ depends only on y_2. We are interested in estimating $\boldsymbol{\theta}$ or at least certain parameters such as $R = \Pr(Y_1 = 1)$. If the separation (4.1) holds, we can first estimate $\boldsymbol{\eta}$ from the biased

likelihood involving only the Y_{2i}'s. The conditional probability $F_\tau(y) = \Pr(Y_1 = 1 \mid Y_2 = y)$ is free of the selection bias and can be estimated using the usual binary regression techniques. For example, if $F_\tau(y)$ follows a logistic regression model, we can estimate the parameters using the usual logistic MLE routines. The results can then be combined to estimate $\boldsymbol{\theta}$ and the associated parameters of interest such as R.

The preceding situation arises in many fields of application. For example, in petroleum exploration, $Y_1 = 0$ or 1 according to whether or not a wildcat well is dry, and Y_2 denotes the area of the prospect. Then R is the probability of discovering a pool. In the stress-strength problem in reliability, Y_2 denotes the stress applied to a component, and $Y_1 = 0$ or 1 according to whether or not the component failed, which depends on whether or not the stress applied to the unit exceeded its strength. The reliability of the component is given by R. In software debugging, $Y_1 = 0$ or 1 according to whether or not a program contained a bug, and Y_2 represents some appropriate measure of program complexity.

5. APPLICATION TO RIMBEY–MEADOWBROOK DATA

Consider again the discovery data in Section 2 from the Rimbey–Meadowbrook reef play. Assume for illustrative purposes that the superpopulation distribution of the data is multivariate lognormal. See the discussion in Wang and Nair (1988) on data-analytic procedures for identifying an appropriate parametric model, however. Suppose the weight function is given by

$$w(\mathbf{y}) = area^{\gamma_2} \times net\ pay^{\gamma_3} \times depth^{\gamma_4}, \quad (5.1)$$

with $\gamma_1 \equiv 0$ so that $w(y)$ does not depend of *volume*. If we apply a logarithmic transformation to the data, we are exactly in the framework discussed in Example 4.1. Therefore, we can estimate $(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ by first estimating the parameters of the univariate lognormal distribution of $w(\mathbf{Y})$ and then use least squares to estimate $\boldsymbol{\tau}$.

Table 4 gives the MLE's of the means, variances, and the correlation matrix corresponding to $N = 40$ and $\boldsymbol{\gamma} = (0, .84, .82, -2.68)$. Ignoring for the moment how these particular values of N and $\boldsymbol{\gamma}$ were obtained, compare the estimates in Table 4 with those in Table 2. The means of all the variables except *depth* are smaller, and all the variances are larger. This is to be expected since the pools with larger *areas* and *net pays* have already been discovered and the remaining $N - n$ pools are likely to be the smaller ones. The reverse is true for *depth*, which has a negative value of γ. The negative value implies that the shallower pools have been discovered and the remaining ones are likely to be in the deeper horizons of the Rimbey–Meadowbrook reef play. This is consistent with geological information; the play has a northeast–southwest down dip (Stoakes 1980), and the largest discovered pool (Red Water, see Table 1) is located in the shallower northeastern region. The differences in the estimates of the *depth* variable between Tables 2 and 4 are rather small (see also the comparisons in Tables 5 and 6), suggesting that inference for this variable is not very sensitive to the selection bias. Note also that the mean and variance of the *volume* variable have changed, although this variable is not explicitly involved in the weight function (5.1). This is due to its dependence on the other variables. Interestingly enough, all the correlations in Table 3 have also increased in absolute value from those in Table 2.

Consider now the problem of predicting the remaining amount of recoverable oil in the play. Given the observed data, the conditional expectation of the log-*volume* of a remaining pool is -3.55, so the amount of recoverable oil from the remaining $40 - 23 = 17$ pools is (approximately) $17 \times e^{-3.55} = .49$ million cubic meters. This is actually an underestimate since $e^{EZ} < Ee^Z$, the conditional expectation of *volume* of a remaining pool. But the important point here is the comparison between this value and the similarly obtained value in Section 2 under simple random sampling. The predicted value from Table 4 of about .5 million cubic meters is several orders of

Table 4. Maximum Likelihood Estimates of the Parameters of the Multivariate Lognormal Distribution for the Rimbey–Meadowbrook Data Under the Successive Sampling Discovery Model With N = 40

				$\hat{\rho}_{kj}$			
Variable	γ_k	$\hat{\mu}_k$	$\hat{\sigma}_{kk}$	*Volume*	*Area*	*Net pay*	*Depth*
Volume	.0	−.74 (±.75)	11.10 (±3.92)	1.00	.93	.78	−.32
Area	.8	4.46 (±.56)	5.97 (±2.75)		1.00	.55	−.26
Net pay	.8	2.19 (±.31)	1.47 (±1.15)			1.00	−.27
Depth	−2.7	7.55 (±.08)	.08 (±.17)				1.00

NOTE: Standard errors are in parentheses.

Table 5. Comparison of the Estimated/Predicted Values for the Rimbey–Meadowbrook Data for Different Values of the Finite Population Size N [γ = (0, .8, .8, −2.7) in all three cases]

Estimates	N = 35	N = 40	N = 45
$\hat{\mu}_1$—*volume*	−.28	−.74	−1.15
$\hat{\mu}_2$—*area*	4.79	4.46	4.17
$\hat{\mu}_3$—*net pay*	2.33	2.19	2.08
$\hat{\mu}_4$—*depth*	7.52	7.55	7.57
Recoverable oil from the remaining pools	.41	.49	.52

magnitude smaller than the value of 64 million cubic meters obtained from Table 2 under simple random sampling.

Since our purpose is to use the Rimbey–Meadowbrook data to illustrate the results in Sections 3 and 4, we have assumed in our analysis (Table 4) that the values of N and $\boldsymbol{\gamma}$ are known. In petroleum resource applications, these parameters are typically unknown. Our value of $N = 40$ was suggested by an earlier analysis of the Rimbey–Meadowbrook data by Lee and Wang (1986). In Table 4, $\boldsymbol{\gamma}$ was obtained by maximizing the log-likelihood over both $\boldsymbol{\gamma}$ and $\boldsymbol{\theta} = (\boldsymbol{\mu}, \boldsymbol{\Sigma})$ by using a grid-search over the range of γ values. Estimation of N and $\boldsymbol{\gamma}$ is a complex issue, and a comprehensive discussion of the problem is beyond the scope of this article. A few of the relevant points are:

1. We have been concerned here only with the analysis of discovery data. Often, there is additonal geological and geophysical information such as contour plots of prospective drilling targets and area exhaustion maps. Such additional information can suggest possible values of N and other characteristics of the finite population. Most papers on the analysis of oil and gas discovery have assumed the knowledge of such characteristics and have considered only conditional inference (also called anchored estimation).

Table 6. Comparison of the Estimated/Predicted Values for the Rimbey–Meadowbrook Data for Different Weight Functions w(y)

Estimates	δ = .9	δ = 1	δ = 1.1
$\hat{\mu}_1$—*volume*	−.71	−.74	−.77
$\hat{\mu}_2$—*area*	4.48	4.46	4.44
$\hat{\mu}_3$—*net pay*	2.20	2.19	2.19
$\hat{\mu}_4$—*depth*	7.55	7.55	7.55
Recoverable oil from the remaining pools	.53	.49	.46

NOTE: The three cases considered correspond to $w(y)^{\delta}$, where $w(y) = area^{.8} \times net\ pay^{.8} \times depth^{-2.7}$ and δ = .9, 1, and 1.1. The finite population size N = 40 in all three cases.

See, for example, Andreatta and Kaufman (1986). There has been some recent work on simultaneous estimation of N and F from discovery data (Bickel et al. 1989; Gordon 1983; Smith and Ward 1981; Wang and Nair 1988). Note, however, that the actual value of N can vary depending on the precise definition of what (or how small) a pool is. Even when N is variable, the total amount of recoverable oil can be quite stable (see the comparisons in Table 5).

2. In petroleum exploration, the parameter γ in $w(y) = y^{\gamma}$ (univariate case) is sometimes called the coefficient of discoverability; the larger the absolute value of γ, the more efficient the discovery process. Bloomfield et al. (1979) showed how one can estimate γ from mature (exhausted) plays. Simultaneous estimation of N and $\boldsymbol{\gamma}$ (in addition to F) from nonmature plays appears to be a difficult problem.

3. It is worth keeping in mind that in petroleum exploration the actual discovery mechanism is a very complicated process. The successive-sampling discovery model is only a rough approximation of reality. Moreover, one does not rely exclusively on the analysis of discovery data. Information from other sources must also be incorporated into the decision-making process. Often N and $\boldsymbol{\gamma}$ are chosen by examining how the predictions vary as a function of these parameters and choosing the values that lead to conclusions that are consistent with other geological information.

4. The standard errors for $\hat{\boldsymbol{\theta}}$ in Table 4 were obtained by assuming that N and $\boldsymbol{\gamma}$ are known. The referees point out that, since these values are based on estimates, the resulting uncertainty also should be reflected in the standard errors of $\hat{\boldsymbol{\theta}}$. This is difficult to do when the methods used to estimate N and $\boldsymbol{\gamma}$ are not easily quantifiable. The difficulty is somewhat similar to a typical data analysis situation in which a Q–Q plot is used to informally identify an appropriate parametric distribution and the parameters are then estimated. Strictly speaking, one should take into account the uncertainty from estimating the shape of the distribution in constructing confidence intervals for, say, the quantiles of the distribution, and in constructing prediction intervals, but this is rarely done due to the difficulty involved.

Table 5 examines the sensitivity to changes in N of the parameter estimates, $\hat{\mu}$'s, and the predicted amount of remaining recoverable oil for the Rimbey–Meadowbrook data. The values considered are N = 35, 40, and 45. The value of $\boldsymbol{\gamma}$ was fixed at the value in Table 4. We see that $\hat{\mu}_4$ (*depth*) is the least sensitive, whereas $\hat{\mu}_1$ (*volume*) is the most sensitive. Despite this, the predicted amount of recoverable oil

(*volume*) from the remaining pools is remarkably stable. Table 6 considers the same type of sensitivity analysis for changes in γ. The three values considered were $\gamma \times \delta$, where γ was fixed at the value in Table 4 and $\delta = .9$, 1, and 1.1. The population size $N = 40$ in all three cases. The parameter estimates $\hat{\mu}$'s were even more stable in this case with very small changes. The predicted amount of recoverable oil also exhibited the same level of stability, so the results are not very sensitive to small perturbations in N and γ in the neighborhood of the values assumed in Table 4.

6. SUMMARY

The results in this article can be used to analyze multivariate discovery data using a successive-sampling discovery model with a given, arbitrary measure of size. We have illustrated the techniques by applying them to oil discovery data from the Rimbey–Meadowbrook reef play. Other areas of potential application include software debugging and testing for design errors in hardware reliability.

ACKNOWLEDGMENTS

We are grateful to John Chambers, the referees, and the editors for helpful comments.

[*Received October 1987. Revised May 1989.*]

REFERENCES

Andreatta, G., and Kaufman, G. M. (1986), "Estimation of Finite Population Properties When Sampling Is Without Replacement and Proportional to Magnitude," *Journal of the American Statistical Association*, 81, 657–666.

Arps, J. J., and Roberts, T. G. (1958), "Economics of Drilling for Cretaceous Oil Production on the East Flank of the Denver–Julesberg Basin," *Bulletin of the American Association of Petroleum Geologists*, 42, 2549–2566.

Barouch, E., and Kaufman, G. M. (1976), "Probabilistic Modelling of Oil and Gas Discovery," in *Energy: Mathematics and Models*, ed. F. Roberts, Philadelphia: Society for Industrial and Applied Mathematics, pp. 248–260.

——— (1977), "Estimation of Undiscovered Oil and Gas," in *Proceedings of Symposium in Applied Mathematics* (Vol. 21), Providence, RI: American Mathematical Society, pp. 77–91.

Barouch, E., Kaufman, G. M., and Nelligan, J. (1983), "Estimation of Parameters of Oil and Gas Discovery Process Models Using the Expectation-Maximization Algorithm," in *Energy Modelling and Simulation*, eds. A. S. Kydes et al., Amsterdam: North-Holland, pp. 109–117.

Bickel, P. J., Nair, V. N., and Wang, P. C. C. (1989), "Nonparametric Inference Under an Informative Sampling Design," unpublished manuscript.

Bloomfield, P., Deffeyes, K. S., Watson, G. S., Benjamini, Y., and Stine, R. A. (1979), "Volume and Area of Oil Fields and Their Impact on Order of Discovery," Report of the Resource Estimation and Validation Project, Princeton University, Depts. of Statistics and Geology.

Cleveland, W. S. (1979), "Robust Locally Weighted Regression and Smoothing Scatterplots," *Journal of the American Statistical Association*, 74, 829–836.

Cox, D. R. (1969), "Some Sampling Problems in Technology," in *New Developments in Survey Sampling*, eds. N. L. Johnson and H. Smith, Jr., New York: Wiley-Interscience, pp. 506–527.

Cox, D. R., and Hinkley, D. V. (1974), *Theoretical Statistics*, New York: Chapman & Hall.

Cozzolino, J. M. (1972), "Sequential Search for an Unknown Number of Objects of Nonuniform Size," *Operations Research*, 20, 293–308.

Crump, K. S. (1976), "Numerical Inversion of Laplace Transforms Using a Fourier Series Approximation," *Journal of the Association for Computing Machinery*, 23, 89–96.

Davies, R. B. (1980), "The Distribution of a Linear Combination of χ^2 Random Variables," *Applied Statistics*, 29, 323–333.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data Via the E-M Algorithm," *Journal of the Royal Statistical Society*, Ser. B, 39, 1–39.

Energy Resources Conservation Board (1985), *Alberta's Reserves of Crude Oil, Oil Sands, Gas, Natural Gas Liquids, and Sulphur*, Calgary, Alberta: Author.

Gordon, L. (1983), "Estimation for Large Successive Samples With Unknown Inclusion Probabilities," unpublished manuscript.

IMSL (1984), *IMSL User's Manual* (Vol. 2, Edition 9.2), Houston: International Mathematical and Statistical Libraries.

Johnson, N. L., and Kotz, S. (1970), *Continuous Univariate Distributions—1*, Boston: Houghton Mifflin.

Kaufman, G., Balcer, Y., and Kruyt, D. (1975), "A Probabilistic Model of Oil and Gas Discovery," in *Estimating the Volume of Undiscovered Oil and Gas Resources*, ed. J. Haun, Tulsa, OK: American Association of Petroleum Geologists, pp. 113–142.

Lee, P. J., and Wang, P. C. C. (1985), "Prediction of Oil or Gas Pool: Sizes When Discovery Record Is Available," *Journal of the International Association of Mathematical Geology*, 17, 95–113.

——— (1986), "Evaluation of Petroleum Resources From Pool Size Distributions," in *Oil and Gas Assessment—Methods and Applications* (Studies in Geology, No. 21), ed. D. D. Rice, Tulsa, OK: American Association of Petroleum Geologists, pp. 33–42.

Littlewood, B. (1981), "Stochastic Reliability-Growth: A Model for Fault-Removal in Computer-Programs and Hardware-Designs," *IEEE Transactions of Reliability*, 30, 313–320.

Mallows, C. L., and Nair, V. N. (1987), "A Unique Unbiased Estimator With an Interesting Property," *The American Statistician*, 41, 205–206.

Patil, G. P., and Rao, C. R. (1977), "The Weighted Distributions: A Survey of Their Applications," in *Application of Statistics*, ed. P. R. Krishnaiah, Amsterdam: North-Holland, pp. 383–405.

——— (1978), "Weighted Distributions and Size-Biased Sampling With Applications to Wildlife Populations and Human Families," *Biometrics*, 34, 179–189.

Scheaffer, R. L. (1972), "Size-Biased Sampling," *Technometrics*, 14, 635–644.

Sheil, J., and O'Muircheartaigh, I. (1977), "The Distribution of Non-negative Quadratic Forms in Normal Variables," *Applied Statistics*, 26, 92–98.

Smith, J. L., and Ward, G. L. (1981), "Maximum Likelihood Estimates of the Size Distribution of North Sea Oil Fields," *Journal of the International Association of Mathematical Geology*, 13, 399–413.

Stoakes, F. (1980), "Nature and Control of Shale Basin Fill and

Its Effect on Reef Growth and Termination: Upper Devonian Duvernay and Ireton Formations of Alberta, Canada," *Bulletin of Canadian Petroleum Geology*, 28, 345–410.

Vardi, Y. (1982), "Nonparametric Estimation in the Presence of Length Bias," *The Annals of Statistics*, 10, 616–620.

Wang, P. C. C., and Nair, V. N. (1988), "Statistical Analysis of Oil and Gas Discovery Data," in *Quantitative Analysis of Mineral and Energy Resources* (NATO ASI series), eds. C. F. Chung, A. G. Fabbri, and R. Sinding-Larsen, Boston: D. Reidel, pp. 199–214.

Wu, C. F. J. (1983), "On the Convergence Properties of the E-M Algorithm," *The Annals of Statistics*, 11, 95–103.