Inference in Successive Sampling Discovery Models

MIKE WEST

Institute of Statistics and Decision Sciences
Duke University
Durham
North Carolina 27708-0251, USA.

Abstract

A variety of practical problems of finite population inference can be addressed in the framework of successive sampling discovery models – population units are assumed drawn from a superpopulation distribution and then successively sampled according to a specified 'size-biased' selection mechanism. Formal statistical analysis of discovery data under such models is technically challenging, as exemplified by the likelihood analyses of Nair and Wang (1989). Assessment of uncertainties about superpopulation parameters and, more critically, appropriate forms of predictive inference for the unsampled units in the finite population, are open issues that are addressed here from a Bayesian perspective. Motivated by the likelihood analysis of Nair and Wang (1989), we develop a formal Bayesian approach to analysis in the same class of models; we show how simulation methods provide for the computation of required posterior and predictive distributions of relevance. We further develop model extensions to cover problems of uncertainty about finite population sizes, uncertainty about sample selection mechanisms, and other practical issues. Several analyses of the oil reserve data of Nair and Wang (1989) are used for illustration.

1 Introduction

In a recent article, Nair and Wang (1989) describe analysis of a *successive sampling discovery model* characterised by two key features:

- finite population inference assuming a superpopulation distribution for characteristics of the finite population;
- data drawn from the finite population subject to size biased sampling.

In the context of a multivariate normal superpopulation model and specific functional forms of size biasing, Nair and Wang (1989) develop extended EM algorithms for maximum likelihood estimation of superpopulation parameters, and various methods of extrapolation to inference about characteristics of the finite population being sampled. Discovery of oil reserves is their central, very interesting example.

The current paper concerns Bayesian inference in these and more general contexts. Bayesian inference in standard parametric models of infinite populations subject to selection effects (e.g.

Bayarri and DeGroot, 1987), or generally subject to non-random sampling, has until recently been hampered by computational difficulties. As described and illustrated in Kuo and Smith (1992), and West (1994), however, Markov Chain Monte Carlo methods now allow what is often trivial coding of routines to simulate posterior distributions in common models subject to truncation, censoring and selection effects. Some finite population problems are similarly amenable to simulation based analysis, as indicated in West (1994); that paper includes a simple treatment of a univariate successive sampling model. This current paper provides a full and extensive treatment of the multivariate discovery models of Nair and Wang (1989), with various practical generalisations, and illustrations. We show how posterior and, critically, predictive inferences can be generated using simulation and data augmentation, directly addressing the issues of uncertainty assessment in estimation and prediction. We further develop analyses to provide inference on finite population sizes, assessments of sensitivity to hypothesised sampling bias mechanisms, and to answer various interesting predictive questions.

The underlying statistical framework is as follows. Multivariate, non-negative observations are obtained sequentially in time. It is assumed that the data values are characteristics of observational units drawn without replacement from a finite population; the population of values is denoted by $Y \stackrel{\text{def}}{=} \{y_1, \ldots, y_N\}$, each y_i being a $p-$vector for some fixed $p \geq 1$. The population size N is usually assumed assumed known. Values y_i relate to measures of 'size' of the units. From Y, a sample of specified size $n \leq N$ is successively drawn, without replacement. Sampling is supposed to be biased; the selection probability for an unsampled unit i depends, at each stage in time, on both y_i and the characteristics of the other remaining units.

In the oil discovery application of Nair and Wang (1989), for example, the four-dimensional observations are estimates of surface area, volume, net-pay and depth of oil pools in an oil rich area, or oil play (see also the bibliography of Nair and Wang). As the observations measure the physical sizes and locations of pools, their values are naturally supposed to be informative about the discovery process. Larger pools have larger surface areas and other characteristics that enhance the chance of discovery under investigation of the play; deeper pools are more difficult to find. A general framework for biased sampling supposes that units are selected with probabilities proportional to some non-negative weight or selection function, $w(y)$.

The superpopulation structure assumes an underlying distribution, assumed to have a density $f(y|\theta)$, from which the values in Y are randomly sampled. Here θ is a collection of parameters characterising the superpopulation model. $f(y|\theta)$ is most simply interpretable as a prior distribution for the y_i, though non-Bayesians assume a global frequency-based interpretation (e.g. the distribution of oil pool 'sizes' worldwide). Analysis is aimed, primarily, at inference about characteristics of the finite population, such as predicting the set of values remaining, their total or other summaries. Secondarily, this will involve inference about the underlying superpopulation parameters θ.

Section 2 develops the Bayesian analysis of this general model, describing posterior simulation via Gibbs sampling. Section 3 discusses details of multivariate log-normal superpopulation models with log-linear size biasing. A first analysis of oil reserve data from Nair and Wang (1989) appears in Section 4, providing comparisons with the likelihood analysis and illustrating Bayesian inference. Section 5 concerns uncertainty about the population size N and extends analysis to incorporate estimation of N. This is illustrated in some further analyses Section 6, where we also focus on predictive inference and the changes in inference as data are successively recorded. Section 7 discusses

uncertainty about weight functions and describes inference incorporating prior distributions over weight functions; further analysis of the oil reserve data is summarised in this context.

2 Posterior distributions and their simulation

By way of notation:

- Label the data so that, without loss of generality, the n observed values are $y_1, \ldots, y_n$, sampled in that order. Write D for the observed data and U for the unobserved values, so that $Y = \{D, U\}$ with $D = \{y_1, \ldots, y_n\}$ and $U = \{y_{n+1}, \ldots, y_N\}$.
- Let $b_j = w(y_j) + \ldots + w(y_n)$ for $j = 1, \ldots, n$. Then $b_j > 0$ is the 'weight' of the observed units $j, \ldots, n$.
- Let $t(U) = \sum_{i=n+1}^{N} w(y_i)$ be the total weight of the remaining, unobserved units U; note the dependence on the unobserved units U.

The sampling structure is summarised as follows. First, the N elements of Y are randomly drawn from $f(y|\theta)$. Next, units $i = 1, \ldots, n$ are selected in that order; conditional on Y, and having sampled units $1, \ldots, i-1$, the chance of selecting unit i as the next observation is $w(y_i)/(t(U)+b_i)$. Putting these pieces together gives the joint density

$$p(Y|\theta) \equiv p(D, U|\theta) = \frac{N!}{(N-n)!}\{\prod_{i=1}^{n} \frac{w(y_i)}{(t(U)+b_i)}\} \prod_{j=1}^{N} f(y_j|\theta), \tag{1}$$

where the factorials count subsets of size n. (Notation here ignores dependence on N and n for clarity.) Hence the density of the observed data D is

$$p(D|\theta) = \frac{N!}{(N-n)!} \int \cdots \int \{\prod_{i=1}^{n} \frac{w(y_i)}{(t(U)+b_i)}\}\{\prod_{j=1}^{N} f(y_j|\theta)\} \prod_{j=n+1}^{N} dy_j.$$

Having observed D, this equation gives the likelihood function for θ. Evaluation of the likelihood at any point θ involves $N-n$ nested $p-$dimensional integrations, a daunting task in general. At this point, Nair and Wang (1989) embark on development of iterative EM solution of the likelihood equations. With common forms of superpopulation density, this, and other standard approaches, may produce adequate approximations to maximum likelihood estimates but estimating associated uncertainties is very difficult. Appropriate assessment of the implications of parameter uncertainties on predictive inferences about unobserved features of the finite population is similarly extremely difficult. Direct Bayesian analysis using traditional approximations is essentially impossible. The structure of (1) is much more tractable, and suggests that simulation of the posterior density $p(\theta|D)$ can be performed by iteratively simulating the conditional posteriors $p(\theta|Y) \equiv p(\theta|U, D)$ and $p(U|\theta, D)$ – the standard Gibbs sampling paradigm. These two densities are detailed, based on an assumed prior density $p(\theta)$.

A. For known U, (1) implies

$$p(\theta|U, D) \propto p(\theta) \prod_{j=1}^{N} f(y_j|\theta). \tag{2}$$

In common models, a prior $p(\theta)$ that is conjugate to $f(.|\theta)$ implies a conjugate posterior that may be directly sampled and evaluated.

B. For known D and θ, (1) implies

$$p(U|\theta, D) \propto \{\prod_{i=1}^{n}(t(U)+b_i)^{-1}\} \prod_{j=n+1}^{N} f(y_j|\theta).$$

As the elements of U appear in complicated ways through $t(U)$ here, this joint density, in $N-n$ dimensions, is not easy to work with. The data augmentation concept helps out, however, to induce conditional distributions that are easily simulated, as follows. Note that, for each i, $(t(U)+b_i)^{-1} = \int_0^\infty \mathrm{e}^{-(t(U)+b_i)\phi_i} d\phi_i$ and so

$$p(U|\theta, D) \propto \{\prod_{i=1}^{n} \int_0^\infty \mathrm{e}^{-(t(U)+b_i)\phi_i} d\phi_i\} \prod_{j=n+1}^{N} f(y_j|\theta). \tag{3}$$

With $\Phi = \{\phi_1, \ldots, \phi_n\}$, (3) is the marginal density for U from a joint density for $(U, \Phi|\theta, D)$ with the following defining conditionals.

- $p(\Phi|\theta, U, D) \propto \prod_{i=1}^{n} \mathrm{e}^{-(t(U)+b_i)\phi_i}$, hence the ϕ_i are conditionally independent exponentials, $(\phi_i|\theta, U, D) \sim Ex(t(U)+b_i)$. Note incidentally that Φ is conditionally independent of θ here.
- $p(U|\Phi, \theta, D) \propto \{\prod_{i=1}^{n} \mathrm{e}^{-t(U)\phi_i}\} \prod_{j=n+1}^{N} f(y_j|\theta)$. Defining $r = \sum_{i=1}^{n} \phi_i$, we then have

$$p(U|\Phi, \theta, D) \propto \prod_{j=n+1}^{N} \mathrm{e}^{-r\,w(y_j)} f(y_j|\theta). \tag{4}$$

Hence the elements of U are (conditionally) a random sample of size $N-n$ with common density proportional to $\mathrm{e}^{-r\,w(y)} f(y|\theta)$. Note incidentally that U is conditionally independent of D here.

Analysis via Gibbs sampling involves iteratively simulating the conditional posteriors in A and B, as follows.

(a) Choose an initial value of U and compute $t(U) = \sum_{i=n+1}^{N} w(y_i)$.
(b) Sample θ from (2) conditional on the current U.
(c) Similarly sample the exponentials Φ conditional on the current U; save only the summary $r = \sum_{i=1}^{n} \phi_i$.
(d) Draw U from (4) conditional on current values of r and θ.
(e) Return to (b), and iterate.

This iterative scheme determines a Markov Chain in $\{\theta, U\}$ space whose stationary distribution is the joint posterior $p(\theta, U|D)$. Following some 'burn-in' iterations, successive draws are saved as the basis of summary inferences; write $\{\theta_k, U_k\}_{k=1}^{K}$ for these samples. In addition to the raw posterior samples, the known conditional distributions may be used in producing final approximations to posterior quantities. Note also that draws are implicitly made from the posteriors for characteristics of the finite population – any functions of the elements of U. Suppose $size(y)$ represents an "interesting" summary measure of "size" of population units, such as the net volume of an oil pool with characteristics y. Then, for example, the total size of the remaining units is $s(U) \stackrel{\text{def}}{=} \sum_{i=n+1}^{N} size(y_i)$, and the posterior for this total may be directly approximated by a histogram

(or a smoothed version of it) of the implied posterior sample $\{s(U_k)\}_{k=1}^{K}$. Forecasts of the sizes of future cases, in order, are similarly available; thus, for example, we are able to evaluate predictions for units remaining in an hypothetical future discovery process. This is illustrated below.

Sampling the elements of U in (4) requires comment. The U values are a random sample with common density proportional to $\mathrm{e}^{-rw(y)}f(y|\theta)$. This makes sense – the observed data D are sampled from $f(y|\theta)$ with probability proportional to $w(y)$, an increasing function of y, so that the remaining cases U will tend to be smaller as determined under this modified density. Since $r > 0$ the term $\mathrm{e}^{-rw(y)}$ is decreasing, so the modified density function for sampling the unobserved units concentrates on smaller values than does $f(y|\theta)$. As a result, the elements of U are a regular selection sample (Bayarri and DeGroot, 1987) – drawn from the infinite population distribution but with selection probabilities proportional to $\mathrm{e}^{-rw(y)}$. Technically, we still have the problem of simulating these selection samples. Though $f(y|\theta)$ will typically be easily simulated, this is unlikely to be true of the selection distribution in practical models. The selection structure provides for sampling via rejection. Note simply that $\mathrm{e}^{-rw(y)} = P(x > r|y)$ where, given the vector y, x is exponentially distributed, $(x|y) \sim Ex(w(y))$; then (4) is sampled as follows:

(i) Draw y from $f(y|\theta)$ and, independently, $u \sim U(0,1)$.

(ii) If $\log(1-u) > -rw(y)$ reject y and return to (i); otherwise save y and stop.

Repeating this $N-n$ times generates the required latent data $U = \{y_{n+1}, \ldots, y_N\}$.

3 Log-normal model with log-linear weighting

Let x be the $p-$vector whose elements are the natural logs of the corresponding elements of y. Suppose $x \sim N(\mu, \Sigma)$, a $p-$variate normal distribution whose mean vector μ and variance matrix Σ determine the parameter $\theta = \{\mu, \Sigma\}$. Then $f(y|\theta)$ is multivariate log-normal, as in Nair and Wang (1989). Further following these, and previous authors, assume the weighting of sampled units is log-linear with $w(y) = \mathrm{e}^{a'x}$ for some specified $p-$vector a of non-negative elements; thus $w(y)$ is a weighted geometric mean of the elements of y.

In this special model, conjugate priors for θ lie in the normal-inverse Wishart class (Press, 1985, section 7.1.6). General results based on proper priors in this class are straightforward to derive. For benchmark analysis, the usual reference prior $p(\theta) \propto |\Sigma|^{-(p+1)/2}$ is appropriate. Under this prior, which is used in data analyses below, the relevant conditional posterior in equation (2) is as follows. Write $\bar{x}$ and S for the sample mean vector and sum of squares matrix of the x_i, $(i = 1, \ldots, N)$, so that $\bar{x} = \sum_{i=1}^{N} x_i/N$ and $S = \sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})'$. Then under the posterior (2), μ given Σ is normal, $N(\bar{x}, \Sigma/N)$, and Σ has the inverse Wishart distribution $W^{-1}(S, p, N+p)$ (in the notation of Press, 1985; see section 7.1.6). This provides for simulation of θ at point A in Section 2. Standard methods may be used to generate normal deviates, and inverse Wishart variates are obtained by inverting Wishart draws; the latter may be (and are below) generated using the technique of Odell and Fieveson (1966).

At point B, the main issue is simulation of the latent items U in equation (4). Generally, sampling via rejection as outlined can be implemented, though could be computationally expensive since, at each step of the Gibbs iterations, the $N-n$ sampled $p-$vectors are obtained at the expense of rejecting possibly many more. Sometimes this direct rejection method will be necessary. However, with the special log-linear weight function, some efficiency can be gained by mapping the rejection

steps down from p to just one dimension. To see this, write $z = a'x$ so that $w(y) = \mathrm{e}^z$. Since the rejection step only involves z, it makes sense to restrict the candidate draws to the one-dimensional z alone. Now, under $(x|\theta) \sim N(\mu, \Sigma)$, we have a marginal distribution $(z|\theta) \sim N(a'\mu, q)$ with $q = a'\Sigma a$, and a set of conditional distributions $(x|z, \theta)$ which are singular normal,

$$(x|z, \theta) \sim N(\mu + A(z - a'\mu), \Sigma - AA'q), \tag{5}$$

with $A = \Sigma a/q$. The corresponding conditionals $(y|z, \theta)$ are the resulting multivariate log-normal distributions. Now, we want to sample the distribution whose density is proportional to

$$\mathrm{e}^{-r\,w(y)}\, f(y|\theta) \;=\; \int_{-\infty}^{\infty} \mathrm{e}^{-r\mathrm{e}^z}\, p(y|z, \theta) p(z|\theta) dz.$$

This is simulated as follows:

- Sample z from density proportional to $\mathrm{e}^{-r\mathrm{e}^z} p(z|\theta)$ where $(z|\theta)$ is the above univariate normal; this is done via rejection as noted in section 2. Once a sampled value is accepted, use it at the next step.
- Given z, draw x from the multivariate normal in (5), and transform to y via antilogs.

In the second step, x is drawn from the singular normal distribution $p(x|z, \theta)$. A direct way to do this is to use the singular variance matrix in terms of its singular value decomposition $\Sigma - AA'q = LDDL'$ where L is the matrix whose columns are eigenvectors of $\Sigma - AA'q$, and D is the diagonal matrix of square roots of the corresponding eigenvalues. One diagonal element in D is zero due to the singularity; let E represent the $p \times (p-1)$ matrix obtained by deleting the corresponding column of zeroes from LD. Then x is generated as $x = \mu + A(z - a'\mu) + E\epsilon$ where ϵ is a $p-1$ vector of independent standard normal deviates.

Note that this dimensionality reduction process may be effected in any model whose weight function $w(y)$ depends only on a linear function of the logged elements of y.

4 A first analysis of oil reserve data

Oil deposit data in Table 1 of Nair and Wang (1989) give estimated size characteristics of oil pools in an oil play in the Rimbey-Meadowbrook reef chain located in central Alberta, the discoveries having been made during 1947 and 1968. The data include estimated volume, surface area, net pay and depth for each of the $n = 23$ pools discovered during that period; so $p = 4$ and y is the $4-$vector of these measurements in stated order, and x is the $4-$vector of logged values, assumed drawn from the $N(\mu, \Sigma)$ superpopulation prior subject to size biasing. Nair and Wang analyse this data using $w(y) = \exp(a'x)$ with $a' = (0, 0.84, 0.82, -2.68)$; we use this weight function here in preliminary analysis of this data. We also adopt the reference prior for μ and Σ, though note that in the area of application, there exists substantial expertise that should be used to explore ranges of informative priors, and hence to the derived ranges of posterior inferences. We use the reference prior as a benchmark, as usual, with which inferences under alternative priors and from non-Bayesian approaches may be compared; we are particularly interested in comparing results with the likelihood approach of Nair and Wang.

This first analysis assumes $N = 40$ pools to compare with Nair and Wang (1989). The Gibbs sampling procedure is easily implemented. This requires initial values for the $N - n = 17$ unobserved

vectors in U. Repeat simulations using various starting values have verified the insensitivity to these values of the posterior inferences reported here and in subsequent analyses in later sections; widely differing starting values led to insignificant differences in reported posterior inferences, supporting the assumption of convergence of the Gibbs sampling iterations. The simulation computations were burnt-in for 2,000 iterations (Raftery and Lewis, 1992); values of μ, Σ and U sampled in these initial iterations are not used in final inferences. 45,000 further iterations of the Gibbs sampling scheme were performed, and every fifth set of values of μ, Σ and U were saved to provide an approximate sample of size $K = 9,000$ from the joint posterior $p(\mu, \Sigma, U|D)$. These computations, with various subsidiary calculations and data manipulation, take under five minutes on a DECstation 5000/200 in RISC Fortan code written by the author. The output data are trivially summarised.

To provide comparison with the likelihood analysis, point estimates of μ and Σ are quoted; these are the Monte Carlo approximations to $E(\mu|D)$ and $E(\Sigma|D)$ computed by averaging the 9,000 simulated conditional means, $E(\mu|D, U_k)$ and $E(\Sigma|D, U_k)$. Similar computations provide the approximate posterior standard deviations quoted for μ. The estimates for μ, $\pm$one posterior standard deviation in each case, appear below together with the MLEs and corresponding approximate standard deviations from Nair and Wang (1989). The elements are, in order, the means for Volume, Area, Net pay and Depth, respectively.

$$\begin{array}{ll} E(\mu'|D): & -1.21 \pm 0.68,\ \ 4.15 \pm 0.49,\ \ 2.06 \pm 0.24,\ \ 7.57 \pm 0.05 \\ \text{MLE:} & -0.74 \pm 0.75,\ \ 4.46 \pm 0.56,\ \ 2.19 \pm 0.31,\ \ 7.55 \pm 0.08 \end{array}$$

It is notable that the MLEs overestimate the first three elements – the average volume, area and net pay of oil pools, all measuring physical size – relative to the posterior mean. This corresponds to negative skewness in the corresponding univariate marginal posterior densities. Posterior uncertainties as measured by the quoted standard deviations are smaller than those associated with the MLEs. As we shall note below, however, the likelihood analysis quite radically underestimates variation as described by Σ so that Bayesian predictive distributions–estimates of the superpopulation density–reflect much greater uncertainty than the likelihood analysis might suggest. This comment is based on the posterior estimates of diagonal elements of $E(\Sigma|D)$, given in the first column of the table below; the corresponding MLEs are in parentheses. The remaining columns give posterior estimates of correlations ρ_{ik} derived directly from $E(\Sigma|D)$, again with MLEs in parentheses.

Σ_{kk}	ρ_{1k}	ρ_{2k}	ρ_{3k}	ρ_{4k}
18.38 (11.10)	1.00 (1.00)	0.94 (0.93)	0.80 (0.78)	−0.36 (−0.32)
9.72 (5.97)		1.00 (1.00)	0.59 (0.55)	−0.30 (−0.26)
2.35 (1.47)			1.00 (1.00)	−0.31 (−0.27)
0.15 (0.08)				1.00 (1.00)

In view of the posterior means for the Σ_{kk}, the variability in the superpopulation is apparently quite seriously underestimated by the MLEs. The posterior estimates of correlations are, however, in close agreement with the MLEs. Further study of the posterior distribution may be based on the sampled values, or on more efficient approximations to marginal posterior densities and other features. There is, however, little intrinsic interest in parameter estimation. Predictions about the remaining population units are the overriding issue – the superpopulation parameters θ serve mainly to structure the problem of inference about, in this case, the remaining oil pools. Consider, for example, inference about the *total volume* of oil remaining in the current reserve, namely

$t = \sum_{i=n+1}^{N} y_{i1}$, where y_{i1} is the first (log volume) element of the vector y_i. Each simulated draw from $p(U|D)$ provides a simulated value from $p(t|D)$ simply by summing the appropriate elements of the sampled y_i. An histogram of the 9,000 sampled values in this analysis is unimodal with mode near $t \approx 0.5$ million cubic metres, close to the rough MLE calculation in Nair and Wang (1989). The density $p(t|D)$ is, however, highly skewed over larger totals, well supporting values up to three or four million cubic metres (and with a general shape similar to the third frame in Figure 7 below, which is based on an extended analysis described below). Any single point estimate of total potential volume is clearly misleading, and the MLE radically so; a full appreciation of appropriate values for, and uncertainties about, t (or other quantities) is difficult without access to the full distribution $p(t|D)$. Further discussion, with graphical summaries of predictive distributions, is given below in extended analyses.

5 Uncertainty about N

Assume now that we admit uncertainty about the finite population size N. In many contexts, substantial prior information may be available and can be incorporated in analysis through possible priors, or classes or priors, over N. It may be that N and θ are plausibly related *a priorii*. In the oil pool discovery problem, θ includes parameters describing mean pool sizes and pool locations, so that elicitation of oil experts' prior expectations of mean pool sizes and total oil content of the play will naturally relate to functions of N and θ together, inducing dependencies in the prior. Clearly prior form and structure are heavily context dependent. Write $m = N - n$ for the number of items remaining undiscovered. In this paper, we restrict attention to independent priors $p(m, \theta) = p(m)p(\theta)$, for specific illustration of what is clearly a more general theory. Note that this prior is implicitly dependent on n, and may be derived from an initial prior for N and θ given n. Generally, this may depend on n in ways other that through the simple logical constraint that $m \geq 0$, allowing for the possibilities that observing the sample size n may be informative about n and/or θ other than just via this constraint.

As in section 2, the joint density for D and U in equation (1) combines with $p(m, \theta)$ to give a joint posterior for m, θ and U. The conditional posteriors for θ and U given m are just as derived in that section with $N = n + m$ fixed. Now that analysis is extended to include learning about m. Again introducing the useful latent variables Φ, recalling $r = \sum_{i=1}^{n} \phi_i$, we see that

$$\begin{aligned} p(m|\theta, \Phi, D) \propto & p(m)\, p(D|m, \theta, \Phi) \propto p(m) \int p(D, U|m, \theta, \Phi) dy_{n+1} \ldots dy_{n+m} \\ \propto & p(m)\gamma(r, \theta)^m (m+n)!/m! \end{aligned} \tag{6}$$

where, for fixed r and θ,

$$\gamma(r, \theta) = \int_0^\infty \mathrm{e}^{-rw(y)} f(y|\theta) dy. \tag{7}$$

Note that $0 < \gamma(r, \theta) < 1$.

The Gibbs sampling analysis of section 2 now extends to incorporate m. Referring to the iterations outlined in items (a)–(e) of that section, note simply that we can insert a step to simulate a value of m from (6) following the drawing of Φ at point (b). Otherwise, the analysis is modified only by adding an initial value of m to seed the burn-in iterations.

The two technical steps in sampling (6) involve evaluating the required integral (7), and then performing the required simulation. The integral will not typically be analytically manageable, so numerical methods are needed for that step. In the application of the next section the structure is such that simple and efficient Gauss Hermite quadrature is a natural method, and trivially implemented. In other models, other approaches might be preferable. The simulation step is also heavily model dependent – whether the resulting discrete posterior (6) is easily simulated depends most heavily on the form of the prior $p(m)$. Given an algorithm to evaluate the density function $p(m)$, we can of course evaluate (6) across a finite range of values of m and hence normalise, thus approximately determining the posterior for simulation.

One special class of priors permitting easy analysis for m is as follows. Suppose the initial prior for N given n is a discrete mixture of some h Poisson distributions truncated so that $N \geq n$; take density function $p(N) \propto \sum_{i=1}^{h} \pi_i \lambda_i^N \mathrm{e}^{-\lambda_i}/N!$ for $N = n, n+1, \ldots$, and assume N is independent of θ in the prior. The specified rates $\lambda_i > 0$, and probabilities π_i may depend on both n and θ. Under the induced prior for m, equation (6) leads easily to

$$p(m|\theta, \Phi, D) = \sum_{i=1}^{h} \pi_i^* \{\lambda_i \gamma(r,\theta)\}^m \mathrm{e}^{-\lambda_i \gamma(r,\theta)}/m!, \qquad (m \geq 0)$$

where $\pi_i^* \propto \pi_i \lambda_i^n \mathrm{e}^{-\lambda_i(1-\gamma(r,\theta))}$, subject to unit sum. This is a mixture of Poisson posteriors in which the term $\gamma(r,\theta) < 1$ acts to appropriately decrease the Poisson means from the initial values λ_i. This conditional posterior for m is easily computable and trivially simulated once $\gamma(r,\theta)$ is computed.

6 Analysis of oil reserve data incorporating uncertainty about N

6.1 Preliminaries

In the oil reserve data analysis in Section 4, and with $w(y) = \exp(a'y)$, the integral (7) is an evaluation of the moment generating function of a multivariate log-normal density which cannot be performed in closed form. Drawing on the univariate reduction in Section 3, we can reduce (7) to

$$\gamma(r,\theta) = \int_{-\infty}^{\infty} \exp(-r\mathrm{e}^z) p(z|\theta) dz$$

where the density $p(z|\theta)$ is normal, $(z|\theta) \sim N(a'\mu, a'\Sigma a)$. Note also that the existing Gibbs sampling algorithm already evaluates the moments $a'\mu$ and $a'\Sigma a$ each step. As a result, simple Gauss-Hermite quadrature is an efficient and accurate method of numerical integration in this case. In the examples summarised below, nine-point quadrature is applied to approximately evaluate $\gamma(r,\theta)$ in (7).

Consider now priors for N. Nair and Wang (1989) discuss the sensitivity to the assumed value for N of their MLE calculations and rough predictions, examining differences between analyses based on $N = 35$, 40 and 45. They also discuss issues of available prior information about N in the oil reserve context; such prior information exists based on historical explorations and might be incorporated to determine classes of appropriate priors for N. They conclude that, for predicting total volume of oil in undiscovered pools and other "interesting" quantities, results differ little from those based on N fixed across the 35–45 range. One key reason is that the remaining pools are

indicated as likely quite small, so that adding a few more to the tail will not unduly affect predictions of total volume remaining. Inferences about the superpopulation parameters θ are, however, much more sensitive to N, even radically so. To the extent that the superpopulation model is introduced mainly to provide structure and a framework for prediction about "what is left?", this is not a major concern. However, inference for N may be a real issue in other applications. Here we summarise analyses based on a fairly diffuse prior over the identified range–that displayed in the first frame of Figure 2. This prior is an equally weighted mixture of eight Poissons with means at 25(5)60; this is close to uniform over the range 30–50, but puts appreciable mass on smaller and larger values. Whether this prior makes scientific sense is not an issue here; the issue is whether or not constraining N to the 35–45 range as in previous analyses masks features of the likelihood function of interest and possible importance.

The iterative simulation procedure is easily implemented, and very fast so as to enable a great deal of experimentation with different models, data configurations, starting values and simulation sample sizes. Such explorations have confirmed the adequacy of the reported approximations to posterior distributions below. In each case, posterior reconstructions are based on simulation sample sizes of 9000.

6.2 Analysis of the first 18 discoveries

A first analysis considers only the first $n = 18$ discoveries. Figure 1 displays histograms of the marginal posteriors for the four elements of the mean μ of the superpopulation normal distribution. Superimposed are the posteriors arising from analysis ignoring possible biases in sampling and also the finite population structure; these are just student T densities from the reference analysis of the 18 observed values as a normal random sample. Typical, and here quite dramatic, over-estimation of means is apparent for log volume, log area and, to a lesser extent, log net pay in the "incorrect" analysis; correspondingly, mean log depth is underestimated, again as expected. Also evident is the (expected) substantial over-precision of the direct random sampling analysis. Figure 2 displays the mixture prior and the histogram of posterior draws for the population size N. The posterior mass on values up to 60 or 70 is appreciable.

Predictive inferences of interest relate to characteristics of undiscovered oil pools and, more specifically, features of pools likely to be discovered early in the continued exploration process. For example, assuming a continued discovery process with the same selection mechanism, what are the likely characteristics of the next pool to be discovered? And the next? At each stage of the simulation analysis, a sampled value of N and the set $U = \{y_{n+1}, \ldots, y_N\}$ is obtained. Recall that the elements of U represent draws from the conditional posterior of the remaining $N - n$ units but are not ordered to account for the future discovery process. To obtain an appropriate ordering, simply sample from the finite set U without replacement and according to the selection weight function. For example, a draw from the predictive distribution for the *next* pool to be discovered is obtained by choosing y_{n+j} with selection probability proportional to $w(y_{n+j})$, $(j = 1, \ldots, N - n)$. Removing the selected case from U leaves $N - n - 1$ items which can be further sampled this way to successively produce draws from the predictive distributions for the second, third, and subsequent discoveries assuming a continuation of the discovery sampling process.

Focus on just the volume of oil in remaining pools. Following the above procedure, the first elements of the sampled $4-$vectors at each stage provide draws from the predictives distributions for

volume of undiscovered pools and future discoveries. Figures 3 presents the predictive histogram for log oil volume of the next pool, that numbered 19 in the continuing discovery process. For reference, the X marked on the axis indicates the actual value eventually realised, the four Os indicate the log volumes of the final four future pools. This out-of-sample prediction seems, in the light of eventual observation, adequate. Similar computations lead to predictions for the 23rd pool in the discovery sequence having observed just 18, also appearing in Figure 3; here again future discoveries 19–23 are marked on the axis, with the actual log volume of the 23rd discovery indicated by X. Note that the eventually observed value is rather low, down in the left hand tail of this predictive density. It appears that, at the time of the 18th discovery, predictions about the continuing process on this basis would have rather optimistic. Similar inferences can be derived for other features of undiscovered items. For example, the log oil volume measures in each of the vectors in each sampled set U can be used to calculate the implied total oil volume remaining; the final frame in Figure 3 displays the corresponding predictive histogram in this analysis, for log total remaining. Notice that these predictions all incorporate the uncertainty about the number of pools remaining as each simulation step involves a value of N drawn from $p(N|D)$. Figure 4 displays plots corresponding to those in Figure 3, but now transformed to the actual oil volume scale rather than the log scale. Notice, in particular, the heavy tails of these densities.

6.3 Analysis of all 23 discoveries

A second analysis assumes availability of all $n = 23$ oil pools recorded, as in Section 4 but now under the mixture prior on N. Figure 5 displays margins of the posterior for μ based on these 23 observations; these are to be compared with those in Figure 1 based on only the first 18 observations. Superimposed again are the "incorrect" posterior T distributions from reference analysis of the 23 observations as if they were a normal random sample. Also, the X labels on the axes indicate the maximum likelihood values for the elements of μ obtained by Nair and Wang (1989). One contrast with Figure 1 is that posteriors for mean log volume, area and net pay are shifted to lower values, and for log depth to (slightly) higher values, with the effects most marked on the volume and area characteristics. This is induced by the fact that the last few observed pools 19–23 are really quite small relative to those up to discovery 18, perhaps quite surprisingly so in retrospect; this recalls the earlier comment about optimistic predictions of oil volume remaining at discovery 18. In addition, the posteriors are apparently rather more diffuse than those based on just the first 18 discoveries. Figure 6(a) indicates some reason why this is so; the corresponding posterior for the population size N is wider spread than at the 18th discovery, giving much greater mass to lower values down to $N = 23$ while continuing to appreciably support values up to 60–70. This greater diffuseness, or uncertainty, about the population size is reflected in the increased spread in the posterior for μ. What has happened here is that the last few discoveries represent much smaller pools than predicted after the 18th, so that it now appears much more likely that the population is close to exhaustion based on these few cases; the posterior mass and mode in the region of $N \approx 30$ in Figure 6(a) is induced by these cases, with the upper concentration of mass and the second mode near $N \approx 55$ based largely on the earlier discoveries.

Updated predictions about the future of the discovery process and oil volume remaining appear in Figure 7, to be compared with similar predictions made after 18 discoveries in Figure 4. The circles on the axes represent the actual volumes of the 22nd and 23rd discoveries. Notice, in

particular, the shift to much lower values in predicting total oil volume remaining; the final five discoveries are heavily influential in inducing a shift from the density appreciably supporting up to 40 or 50 million cubic metres after 18 discoveries, to the range up to just 5 or 6 million cubic metres after 23 discoveries.

7 Uncertainty about sampling biases

Some investigation of sensitivity to the assumed form $w(y)$ of the selection function is often desirable. Though we do not address the issues associated with estimating $w(y)$ here, it is of interest to explore deviations away from the assumed form and to try to account for data based deficiencies in the selection function within the analysis. In the oil reserve data example, early discovery of very small pools or late discovery of very large pools are potentially very influential observations that may distort inferences about θ and further predictions; if allowance can be made for such events by suitably modifying the selection mechanisms, then such distortions of inferences can be ameliorated. Additionally, information about deviations away from the assumed selection mechanism can be fed-back to adapt in future inferences. A rather simple mechanism for doing this is developed here.

Modify the basic selection mechanism as follows; assume that a population unit i, having characteristics y_i, is observed with probability proportional to $\beta_i w(y_i)$ where $w(.)$ is the nominal weight function used so far, and the β_i are positive quantities representing possible deviations away from the nominal form. Suppose further that the β_i are 'random effects' drawn as a random sample from some specified prior $p(\beta_i)$, $i = 1, \ldots, N$. With these assumptions, there is no notion of predicting deviations away from the nominal selection mechanism; rather we have a model which allows such deviations (to degrees determined by $p(\beta_i)$) and hence provides an approach, via posterior inference for the weight multipliers β_i, for post-data assessment of such deviations. Additionally, estimation of, and uncertainties about, the β_i will feed through to predictive inference for the unobserved segment of the population.

In particular, assume the β_i are randomly sampled from a gamma prior, $\beta_i \sim G(\alpha, \alpha)$ with $p(\beta_i) \propto \beta_i^{\alpha-1}\exp(-\alpha\beta_i)$ for some $\alpha > 0$. Then $E(\beta_i) = 1$, and so $w(y)$ is the prior expected value for the selection function. Likely degrees of variation away from this expectation are determined by the hyperparameter α, assumed specified. If we choose α large, the β_i will not deviate much from unity, so that sensitivity analysis is really impossible. We proceed to analysis in the context of the general model in Section 6. In the development of Section 2 (extended to include learning on N too), include now the weight vector $\beta = \{\beta_i, \ldots, \beta_N\}$ in the conditioning of all posterior distributions; the analysis described there applies with the simple modification that the weights $w(y_i)$ are replaced by $\beta_i w(y_i)$ throughout. Thus, given a value for β, the posterior $p(\theta, U, N|D, \beta)$ may be sampled as already described for the special case $\beta_i = 1$. The modification to include β implies that

- $b_j = \sum_{i=j}^{n} \beta_i w(y_i)$ for $j = 1, \ldots, n$, and
- $t(U) \equiv t(U, \beta) = \sum_{i=n+1}^{N} \beta_i w(y_i)$;

note the original definitions apply if $\beta_i = 1$ for each i.

Hence, Gibbs sampling can proceed if each iteration is augmented by a step to simulate β from an appropriate conditional posterior distribution. The appropriate distribution is developed

by noting, from the development after equation (3), that

$$p(D, U, \Phi | N, \theta, \beta) \propto \prod_{i=1}^{n} \beta_i w(y_i) \mathrm{e}^{-\phi_i (t(U)+b_i)}$$

as a function of β. This easily reduces to

$$p(D, U, \Phi | N, \theta, \beta) \propto \left(\prod_{i=1}^{n} \beta_i \mathrm{e}^{-\beta_i r_i w(y_i)} \right) \prod_{i=n+1}^{N} \mathrm{e}^{-\beta_i r w(y_i)}$$

where, for $i = 1, \ldots, n$, $r_i = \sum_{j=1}^{i} \phi_j$, and $r \equiv r_n$ as in the original analysis of Section 2. It is now clear that the independent gamma priors for the β_i are conjugate to this conditional likelihood function; as a result, the β_i are conditionally independent with $p(\beta_i | D, U, N, \Phi, \theta) \equiv p(\beta_i | D, U, \Phi)$ given by

$$(\beta_i | D, U, N, \Phi, \theta) \sim \begin{cases} G(\alpha + 1, \alpha + r_i w(y_i)), & \text{for } i = 1, \ldots, n; \\ G(\alpha, \alpha + r w(y_i)), & \text{for } i = n+1, \ldots, N. \end{cases}$$

These gamma posteriors are easily simulated; at each Gibbs iterate therefore, values of the β_i may be sampled to include their estimation and to account for the uncertainty about the weight multipliers. Notice that the conditionals for weight modifiers of undiscovered units are shifted downwards relative to the prior; the means, for example, are $E(\beta_i | D, U, N, \theta, \Phi) = \alpha/(\alpha + r w(y_i))$ for $i > n$; this is reasonable, as larger undiscovered pools have larger baseline weights $w(y_i)$ and so, as they are not yet discovered, the expectation is that the weight modifier is acting to decrease discovery chances.

A further analysis of the full $n = 23$ oil pools incorporates this extension with a gamma prior based on $\alpha = 5$. The analysis of Section 6.3 is modified only in this respect. Figure 6(b) displays the resulting posterior for N. Comparison with Figure 6(a) indicates a marked shift to smaller values, while retaining the bimodality. This is consistent with weight modifiers that are smaller on the larger pools discovered late in the process–a smaller population size would have led to any such larger pools discovered earlier, and their later occurrence suggests lower selection probabilities. To explore inferences about the weight modifiers, Figure 8 displays partial summary information. The simulation analysis provides samples from the posterior for β. For each j, the sampled β_j values are log transformed – the logged values can be expected to have a posterior closer to symmetry than the untransformed value, and the prior for the logged values are close to symmetry – and approximate posterior means and standard deviations computed. The first frame in Figure 8 gives error bars plots for the β_j in discovery order; the error bars represent one standard deviation either side of the mean. For reference, the horizontal dashed lines represent 0 ± 0.45, essentially the prior mean with one prior standard deviation error bars on the log beta scale. Notice that discoveries 17 and 18, particularly, and discovery 20 to a lesser extent, have posteriors suggestive of rather lower weights than the baseline weight function provides. This identifies these pools as rather larger than expected this late in the discovery process, confirming the earlier suspicions about the over-optimistic predictions based on the first 18 pools alone. The second frame in Figure 8 provides a similar plot against the log linear quantities $a' y_j$ appearing in the baseline weight function. This plot, together with additional possible displays against individual component elements of each y_j,

can be useful in indentifying possible functions relationships between the weight modifiers and the y_j, perhaps leading to alternative functional forms for weight functions.

Acknowledgements

Partial support was provided by the National Science Foundation under grant DMS 90-24793.

References

Bayarri, M.J., and DeGroot, M.H. (1987). Bayesian analysis of selection models. *The Statistician*, **36**, pp. 137-146.

Kuo, L., and Smith, A.F.M. (1992). Bayesian computations in survival models via the Gibbs sampler. In *Survival Analysis: State of the Art*, J.P. Klein and P.K. Goel (eds.). Kluwer.

Nair, V.J., and Wang, P.C.C. (1989). Maximum likelihood estimation under a successive sampling discovery model. *Technometrics* **31**, pp. 423-436.

Odell, P.L., and Fieveson, A.H. (1966). A numerical procedure to generate a sample covariance matrix. *J. Amer. Statist. Assoc.* **61**, pp. 199-203.

Press, S.J. (1985). *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference,* Krieger, California.

Raftery, A., and Lewis, S. (1992). How many iterations in the Gibbs sampler? In *Bayesian Statistics IV*, J.O. Berger, J.M. Bernardo, A.P. Dawid and A.F.M. Smith (eds.). Oxford University Press.

West, M. (1994). Discovery sampling and selection models. In *Decision Theory and Related Topics IV*, J.O. Berger and S.S. Gupta (eds.). Springer Verlag: New York.

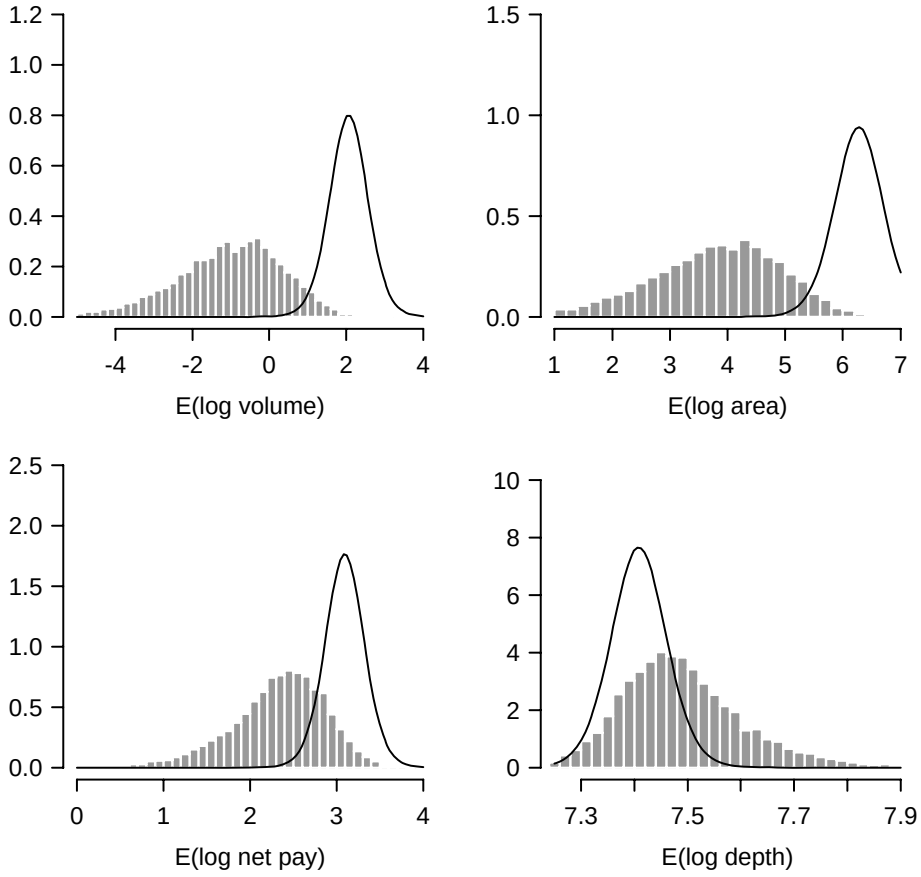

Figure 1. Histograms are approximate margins of $p(\mu|D)$ in analysis based on the first $n = 18$ oil pools. The curves are margins of the reference posterior from analysis of this data as a normal random sample.

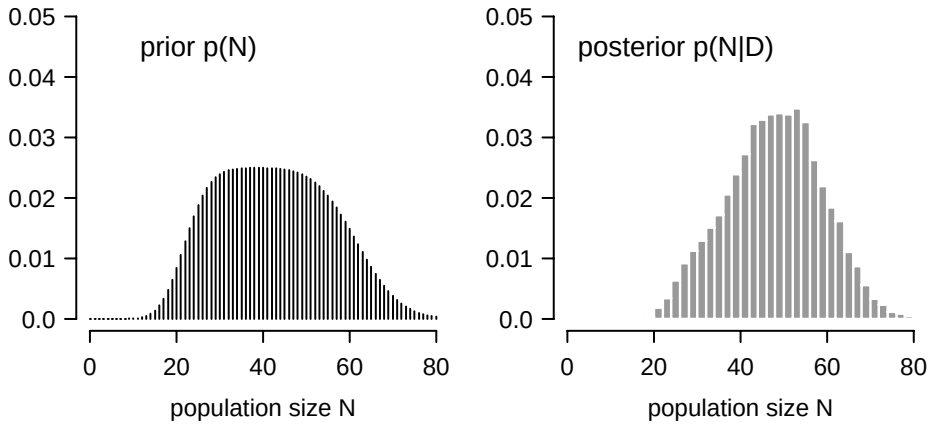

Figure 2. Prior and approximate posterior for population size N in analysis based on the first $n = 18$ oil pools.

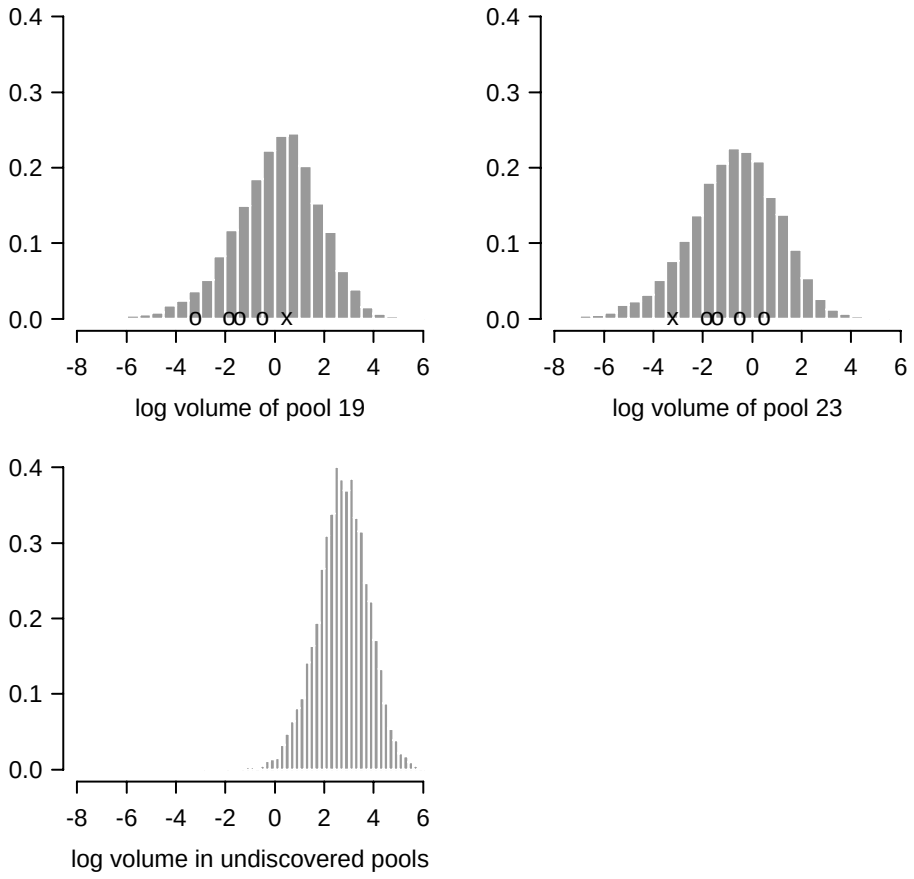

Figure 3. Approximate predictive densities for log oil volume in further pools in analysis based on the first $n = 18$ oil pools. The X labels in the first two frames indicate the actual log volume of pools 19 and 23 respectively, and the circles indicate the log volumes of the other four pools in the last five discovered.

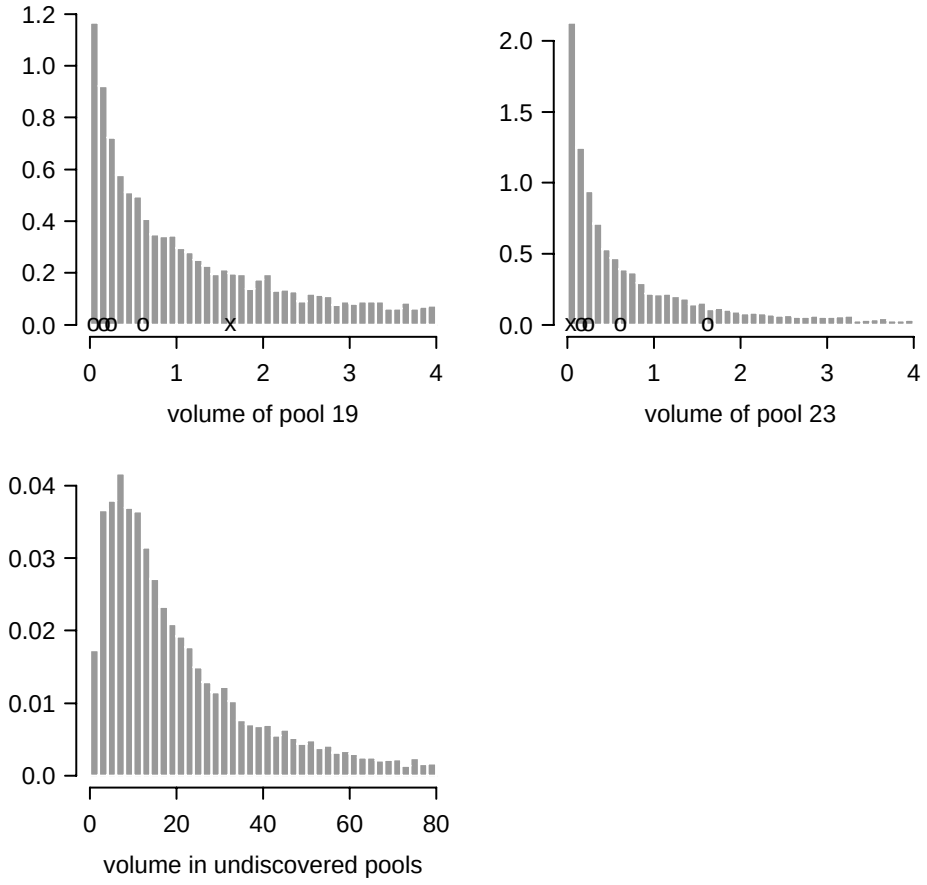

Figure 4. Approximate predictive densities for oil volume in further pools in analysis based on the first $n = 18$ oil pools. The X labels in the first two frames indicate the actual volume of pools 19 and 23 respectively, and the circles indicate the volumes of the other four pools in the last five discovered.

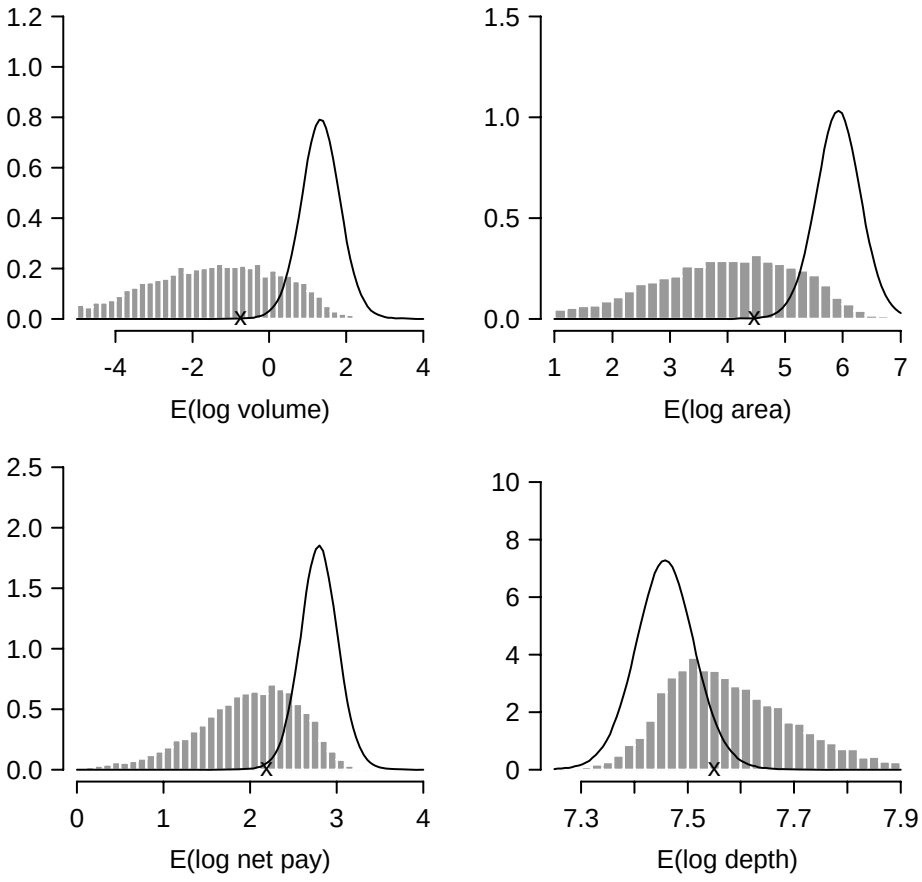

Figure 5. Histograms are approximate margins of $p(\mu|D)$ in analysis based on the full $n = 23$ available oil pools. The curves are margins of the reference posterior from analysis of this data as a normal random sample. The X labels indicate maximum likelihood estimates of these four parameters.

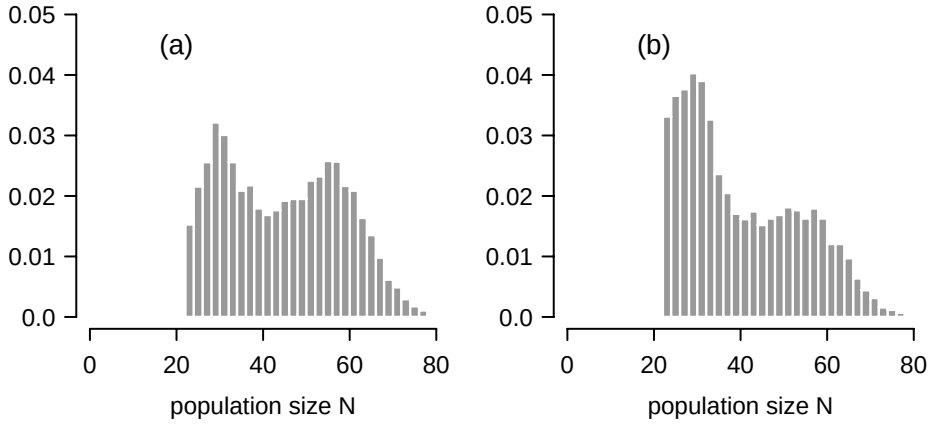

Figure 6. Approximate posteriors for population size N in analyses of the full $n = 23$ oil pools. Frame (a) is from original analysis, frame (b) is from analysis extended to include uncertainty about the selection mechanism via the weight modifiers β.

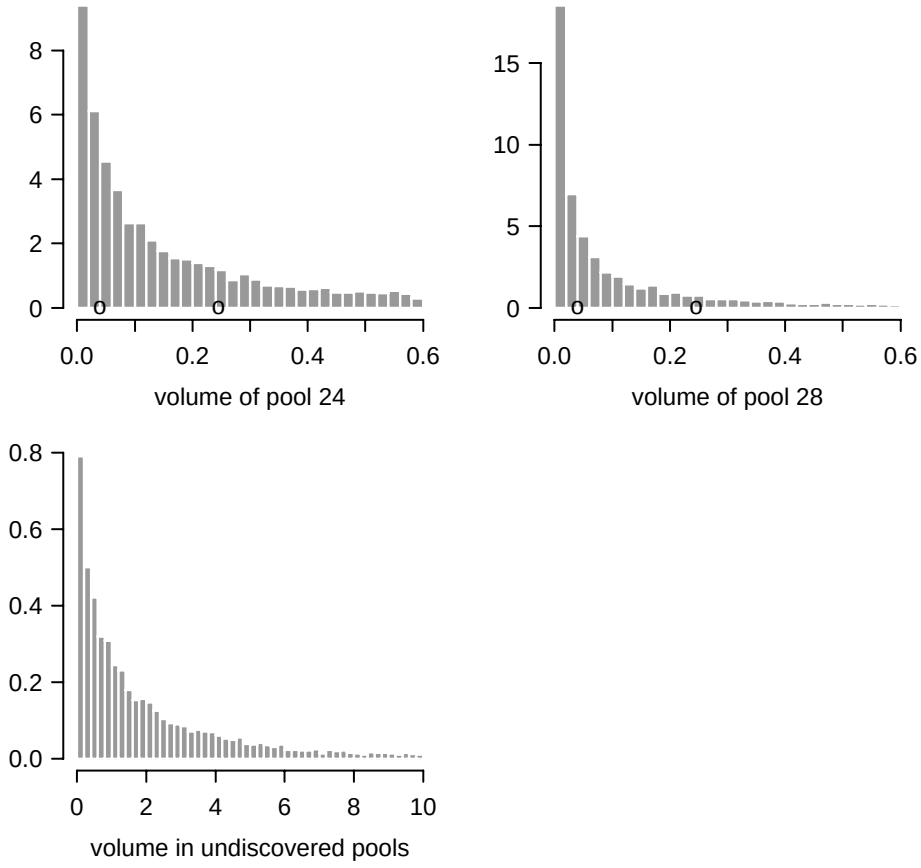

Figure 7. Approximate predictive densities for oil volume in further pools in analysis based on the full $n = 23$ available oil pools. The circles indicate the volumes of pools 22 and 23, for reference.

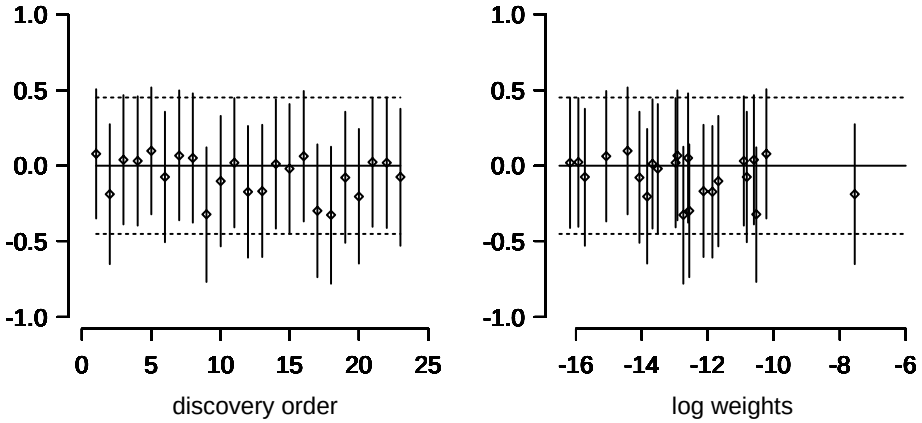

Figure 8. Approximate posterior means and one standard deviation error bars for the log weight modifiers, $\log(\beta_j), (j = 1, \ldots, 23)$, *in analysis of the full* $n = 23$ *available oil pools. The first frame graphs these versus discovery order, the second against the linear function* $a'y_j$ *of the nominal selection weight function. The dashed lines represent approximate one standard deviation intervals from the priors for the* β_j, *symmetrically located about the approximate prior mean of zero.*