

DISCOVERY SAMPLING AND SELECTION MODELS

MIKE WEST

Institute of Statistics and Decision Sciences

Duke University

Durham, NC 27708, U.S.A.

ABSTRACT

Various aspects of Bayesian inference in selection and size biased sampling problems are presented, beginning with discussion of general problems of inference in infinite and finite populations subject to selection sampling. Estimation of the size of finite populations and inference about superpopulation distributions when sampling is apparently informative is then developed in two specific problems. The first is a simple example of truncated data analysis, and some details of simulation based Bayesian analysis are presented. The second concerns *discovery* sampling in which units of a finite population are selected with probabilities proportional to some measure of size. A well-known area of application is in the discovery of oil reserves, and some recently published data from this area is analysed here. Solutions to the computational problems arising are developed using iterative simulation methods. Finally, some comments are made on extensions, including multiparameter superpopulations, semi-parametric models and problems of dealing with missing data in discovery sampling.

1. INTRODUCTION

Inference in problems with sampling bias and selection effects often focus on infinite population models for which analysis is relatively straightforward. Bayarri and DeGroot (1987), for example, present the archetype framework in which data y are sampled independently from a distribution with density function $w(y)f(y|\theta)/c(\theta)$. Here $f(y|\theta)$ is an underlying population density function, usually from a standard parametric family with a low dimensional parameter θ , and $w(y)$ is a non-negative weight or selection function that biases sampling relative to the underlying distribution. Truncation to a known set A via the indicator selection function $w(y) = I(y \in A)$ is a common example (Bayarri and DeGroot, 1987). The term $c(\theta)$ is a normalising constant for the density of selected sample values. Inference about θ , which may be analytically straightforward under the original model $f(y|\theta)$, is usually more complicated in the selection model due to the contribution to the likelihood function of terms involving $c(\theta)$. In low dimensional problems, this is not a major issue since analysis will be amenable to standard numerical integration and optimisation techniques. With higher dimensional parameters Monte Carlo methods are needed. There are also

interesting classes of applications in which the random selection sampling structure is extended to include regression of the observed y values on independent variables, leading to higher dimensional parameters through the introduction of regression parameters.

Iterative posterior simulation schemes (Gelfand and Smith, 1990) allow trivial coding of routines to simulate posterior distributions in certain common selection models. General issues of data augmentation (Tanner and Wong, 1987) and of dealing with truncation and censoring (Kuo and Smith, 1992) in simulation analyses are particularly relevant. Suppose the known selection function $w(y) \geq 0$ is a cumulative distribution function (or may be standardised to become one), such as in the common truncation example where $w(y) = 0$ for $y < a$ and $w(y) = 1$ for $y \geq a$ for some fixed threshold a (Bayarri and DeGroot, 1987; Irony, 1992). In such cases, a single observation y can be modelled as arising as follows: a random sequence x_1, x_2, \dots , is generated from $f(\cdot|\theta)$ and values are accepted with probabilities proportional to $w(x_i)$. Then y is the first value accepted. Let k be the number of x_i rejected prior to accepting $y = x_{k+1}$. Assume also that the prior density $p(\theta)$ is conjugate to $f(y|\theta)$ and easily sampled, and that the data distribution $f(\cdot|\theta)$ is also easily sampled for any given θ . Then the joint structure for x_1, \dots, x_k, y, k and θ is such that Gibbs sampling naturally applies to simulate the posterior $p(\theta|y)$. This follows since (a) $p(\theta|x_1, \dots, x_k, y, k) \propto p(\theta)f(y|\theta) \prod_{i=1}^k f(x_i|\theta)$ is of conjugate form and so easily sampled; and (b) given θ , values of x_1, \dots, x_k, y and, implicitly, k , may be directly sampled by drawing sequentially from $f(\cdot|\theta)$ and accepting/rejecting accordingly. Computational requirements in (b) may be eased by directly drawing k from the implied geometric distribution with success probability $c(\theta)$, and then sampling the x_1, \dots, x_k and y from the relevant selection models. If $f(y|\theta)$ is normal and $w(y) = 0(1)$ for $y < a(\geq a)$, for example, then the relevant selection distributions may be sampled directly without using rejection. The scheme (a) and (b) obviously extends directly to a random sample of size n from the selection model, thus providing for easy Monte Carlo approximation to posteriors in infinite population selection models. The key, as usual, lies in the data augmentation — explicitly introducing the latent variables x_i (Tanner and Wong, 1987).

Some common finite population problems are similarly amenable to simulation based analysis. Suppose a finite population of some N units has a characteristic denoted by y , and that the N values of y in the finite population are modelled as a random sample from some superpopulation distribution whose density is $f(y|\theta)$. Again θ is an uncertain set of parameters. In many cases, depending on the nature of the selection mechanism, the infinite population discussion above may be paralleled in the finite population context to provide easy simulation based posterior analysis. As a simple example, consider inference about both N and θ in a special case of the context of Sanathanan (1977). Suppose that we observe only those units with $y \in A$ for some fixed selection set A , setting $w(y) = I(y \in A)$, and that all such units are observed. If there happen to be n such units, we observe data $D \stackrel{\text{def}}{=} \{n, y_1, \dots, y_n\}$ in which all the y_i lie in A , and do not observe $Y \setminus D = \{y_{n+1}, \dots, y_N\}$. Write $c(\theta) = \int_A f(y|\theta)dy$. Assume independent priors for θ and N in which $p(\theta)$ is conjugate to $f(y|\theta)$, and $p(N)$ is Poisson with fixed rate λ — subject to $N > 0$, of course (Note that mixtures of conjugate priors and mixtures of Poissons neatly generalise the framework with no significant impact on computational issues.) It easily follows that (a) $p(\theta|D, Y \setminus D, N)$ is of conjugate form and so easily sampled; (b) given θ and n , the remaining population size $m \stackrel{\text{def}}{=} N - n$ is conditionally independent of the y_i and has the Poisson posterior with rate λ modified to $\lambda(1 - c(\theta))$;

and (c) given N and θ , the latent variables $Y \setminus D$ form a random sample of size $N - n$ from the infinite population selection model with density proportional to $(1 - w(y))f(y|\theta)$, and so may be easily generated. Distributions (a)–(c) provide the structure for iterative sampling of the joint posterior for θ and N . Incidentally, the posterior for the latent variables $Y \setminus D$ is simulated, so that inference about additional population characteristics – any ‘interesting’ functions of the y_i – easily follows.

Similar issues are now discussed and developed more thoroughly in more specific and interesting contexts.

2. TRUNCATED DATA MODEL

A straightforward application of data/parameter augmentation (Tanner and Wong, 1987) arises in a finite population selection sampling problem studied by Irony (1992). An example there supposes an experimenter records nominally $N(\theta, 1)$ observations generated in time according to a Poisson process of rate λ . Here θ and λ are uncertain; the prior assumes independence with normal and gamma margins, $\theta \sim N(m, M)$ and $\lambda \sim G(a, A)$. Subsequently, the investigator learns that a defect in recording equipment has prevented negative data being recorded. In one unit of time, a single observation $x_1 > 0$ is recorded (the analysis developed here trivially extends to more than one observation as may easily be verified). The problem is to make inferences about θ and λ . Data augmentation is obviously appropriate in this endeavour, as follows.

Let n be the number of observations actually generated in the experiment, the $n - 1$ negative and unobserved values denoted by x_2, \dots, x_n , with $x = \{x_1, \dots, x_n\}$ – the order here is irrelevant. We note that, given n and θ , the single positive observation arises as $k = 1$ where k is binomially distributed with n trials and success chance $\Phi(\theta)$ – the standard normal distribution function at θ . Using this information, marginalisation over n easily leads to

$$p(x_1, k = 1 | \theta, \lambda) \propto \lambda \phi(x_1 - \theta) e^{-\lambda \Phi(\theta)},$$

where $\phi(\cdot)$ is the standard normal density function. Augmenting the problem by n and the latent data in x we see that the full joint posterior has the following conditional distributions.

- (a) Conditional on quantities $k = 1$, n and x we obtain the closed-form, conjugate and independent posteriors $(\theta | x, n) \sim N(m^*, M^*)$ and $(\lambda | x, n) \sim G(a + n, A + 1)$ where $M^* = Mn / (Mn + 1)$ and $m^* = (Mm + n\bar{x}) / (Mn + 1)$. Given n and x , values of θ and λ are trivially simulated from this posterior.
- (b) Given θ , λ , $k = 1$ and x_1 , we can write $n = 1 + h$ where h has the Poisson distribution of rate $\Phi(-\theta)\lambda$. This conditional distribution for $(n | \theta, \lambda)$ is trivially simulated.
- (c) Given n , $k = 1$, x_1 , θ and λ , the unobserved (but negative) quantities x_2, \dots, x_n are independent with distribution function $\Phi(x - \theta) / \Phi(-\theta)$ over $x < 0$. Values are trivially simulated using the inverse distribution function $\theta + \Phi^{-1}\{u\Phi(-\theta)\}$ for $0 < u < 1$.

Routine Gibbs sampling concepts now apply directly (Gelfand and Smith, 1990). The joint posterior for $(\theta, \lambda, n, x_2, \dots, x_n | k = 1, x_1)$ may be sampled iteratively by drawing sequentially from distributions specified in (a), (b) and (c), based on some suitable starting values. Illustration appears in Figure 1. Here the priors are specified by $m = 0$, $M = 16$, $a = 0.1$ and $A = 0.2$. The observation is $x_1 = 0.25$, and the likelihood function is concentrated along a ridge spanning values

of λ from small integers values up to many thousands, and concentrated in the θ dimension on negative, single digit integers. The likelihood has a unique mode at $\theta = -3.53$ and $\lambda = 4714$. Posterior density contours appear in Figure 1(a) – the innermost is at a fraction 0.99 of the height at the posterior mode and the outermost at 0.001 times that modal height. Figure 1(b) is a histogram of the simulated values of the sample size n , representing the posterior $p(n|k = 1, x_1)$. Posterior marginal densities for θ and λ computed in the simulation analysis appear as full lines in 1(c) and 1(d), with priors as dotted lines. The former are computed as the usual averages of conditional densities – thus, for example, $p(\theta|x_1) \approx K^{-1} \sum_{n,x} p(\theta|x_1, n, x)$ where K is the simulation sample size and the sum averages over simulated values of n and x ; the summands are normal densities from part (a) above. That the data is almost completely uninformative about λ is clear from the latter figure. The additional dashed line in 1(c) is an ‘exact’ evaluation of the posterior for θ – it is easily verified that the marginal posterior is proportional to $p(\theta)\phi(x_1 - \theta)/(A + \Phi(\theta))^{1+a}$, and simple quadrature serves to normalise the density to produce the ‘exact’ figure. The close correspondence between this and the simulation based curve is a partial check on the adequacy of the Gibbs sampling analysis – which here was based on a Monte Carlo sample of size 2000 with an initial 500 iterations discarded to ‘burn-in’ from the arbitrary initial values. Repeat simulations with different sample sizes and starting values produce closely similar figures.

Figure 1 goes about here

3. DISCOVERY SAMPLING

3.1. SUPERPOPULATION MODEL ANALYSIS WITH KNOWN POPULATION SIZE

Nair and Wang (1989) discuss likelihood inference in finite populations subject to the assumptions of (i) superpopulation models, and (ii) size biased sampling. This development is termed discovery sampling, and applied there in the context of discovery of oil reserves. The structure is as follows.

Univariate, 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 $Y \stackrel{\text{def}}{=} \{y_1, \dots, y_N\}$, and the population size N is assumed known. Values y_i relate to measures of ‘size’ of the units. Sampling from this population is size biased, it being supposed that units having larger y values are more likely to be sampled. In inference about oil reserves, for example, as discussed by Nair and Wang (and other previous authors – see bibliography of Nair and Wang) the units may represent individual oil deposits, or *pools*, in oil rich areas, or *plays*, and the observations will measure physical size of pools — larger pools have larger surface areas and other characteristics that enhance the chance of discovery under investigation of the play. A general framework for size biased sampling supposes that units are selected with probabilities proportional to some non-negative weight function $w(\cdot)$, an increasing function in the case of discovery sampling. A typical assumption is that probability of selection is proportional to a power of size, so $w(y) = y^c$ for some specified constant $c > 0$.

A traditional superpopulation model corresponds essentially to an assumption of exchangeability of the elements of Y , deemed appropriate whatever the population size may be. In this the

population values y_j are assumed conditionally independent with a common density $f(y|\theta)$ depending on an uncertain parameter vector θ . In oil reserve estimation, $f(\cdot)$ may be a log-normal, gamma or similar density determined by specifying the location, scale and/or shape parameters which make up θ . Prior information will often be available about θ – certainly this will be the norm in the oil reserve context. Objectives of analysis include, primarily, inference about characteristics of the finite population, such as predicting the set of values remaining, their total or other summaries. By the way, this will involve inference about the underlying superpopulation parameters θ .

To proceed to explore the model, we follow Nair and Wang by labelling the data so that, without loss of generality, the n observed values are y_1, \dots, y_n , sampled in that order. We also define positive numbers b_1, \dots, b_n by $b_j = w(y_j) + \dots + w(y_n)$ for $j = 1, \dots, n$ – so b_j is the ‘weight’ of units j, \dots, n . Condition on Y — we will ignore conditioning on N and n for notational clarity; note that they are assumed fixed and known, and uninformative about Y or θ (the next section allows uncertainty about N .) It then follows that the chance of selecting the n units indexed $i = 1, \dots, n$, in that order and without replacement, is given by $\prod_{i=1}^n w(y_i)/(t + b_i)$ where $t = \sum_{i=n+1}^N w(y_i)$ is the total weight of the remaining and unobserved units. Write $D = \{y_1, \dots, y_n\}$ and $Y \setminus D = \{y_{n+1}, \dots, y_N\}$ so that $Y = \{D, Y \setminus D\}$. Under the superpopulation structure, we can then deduce the joint density, conditional on N, n and θ , of the observed and unobserved values as

$$p(D, Y \setminus D | \theta) = \frac{N!}{(N - n)!} \left\{ \prod_{i=1}^n \frac{w(y_i)}{t + b_i} \right\} \prod_{j=1}^N f(y_j | \theta). \quad (1)$$

To compute the likelihood for θ given the observed data D , namely $p(D|\theta)$, the next step is to integrate (1) with respect to the unobserved quantities $Y \setminus D$, as in Nair and Wang (1989).[†] However, this produces a likelihood function of considerable complexity, with consequent difficulties in posterior inference. The structure of (1), is much more tractable, and the explicit appearance of the unobserved or latent variables $Y \setminus D$ obviously suggest approaches to numerical analysis using Gibbs, or other, sampling methods. Introduce a prior density $p(\theta)$ and note the following structure. (I) Condition on D and $Y \setminus D$. Then (1) implies the conditional posterior density

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

In common models, a prior $p(\theta)$ that is conjugate to $f(\cdot|\theta)$ implies a conjugate posterior (2) that may be sampled easily.

(II) Condition on D and θ . Now (1) shows that the conditional density for $Y \setminus D$ is just

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

Recall that t depends on $Y \setminus D$ through the definition $t = \sum_{i=n+1}^N w(y_i)$, so that the joint density function here is not a standard form. However, we may introduce additional ‘latent’

[†] Note, incidentally, the limiting case as $N \rightarrow \infty$ with n fixed; then the problem reduces to the more usual infinite population model in which the elements of X are a random sample from the density $p(y|\theta) \propto w(y)f(y|\theta)$. See comments in Section 1.

data or variables to expand the structure and induce conditional distributions that are easily simulated, as follows. To see this, note that, for each i , $(t + b_i)^{-1} = \int_0^\infty e^{-(t+b_i)\phi_i} d\phi_i$ and so

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

Write $\Phi = \{\phi_1, \dots, \phi_n\}$. Then (3) is proportional to the marginal density for $(Y \setminus D | D, \theta)$ from the joint density for $(Y \setminus D, \Phi | D, \theta)$ with the following structure.

(IIa) The conditional density for Φ is

$$p(\Phi | Y \setminus D, D, \theta) \propto \prod_{i=1}^n e^{-(t+b_i)\phi_i},$$

so that the ϕ_i are conditionally independent and exponentially distributed, $(\phi_i | Y \setminus D, D, \theta) \sim \text{Exp}(t + b_i)$; and

(IIb) the conditional density for $Y \setminus D$ is

$$p(Y \setminus D | \Phi, D, \theta) \propto \left\{ \prod_{i=1}^n e^{-t\phi_i} \right\} \prod_{j=n+1}^N f(y_j | \theta).$$

Write $r = \sum_{i=1}^n \phi_i$. Then, using the definition of $t = \sum_{i=n+1}^N w(y_i)$, we have

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

So the latent y_j are conditionally independent and have the common distribution with density proportional to $e^{-r w(y)} f(y | \theta)$.

Iterative posterior sampling exploits the conditional structure laid out in (I) and (II). Assuming that simulation from the component distributions is feasible, the steps are by now familiar to those using Gibbs or other methods:

- (a) Choose initial values of $Y \setminus D$ and compute $t = \sum_{i=n+1}^N w(y_i)$;
- (b) based on the current values of $Y \setminus D$, sample θ from (2);
- (c) based on the current value of t , sample Φ using the independent exponential distributions under (IIa), and save only the value of $r = \sum_{i=1}^n \phi_i$;
- (d) based on the current values of r and θ , draw a random sample of size $N - n$ from the distribution with density (4) to produce a revised $Y \setminus D$;
- (e) proceed to (b), and iterate.

After ‘burning-in’ for some initial iterations (Raftery and Lewis, 1992), subsequent samples will approximate draws from the joint posterior determined by these conditionals, with margins $p(\theta | D)$, $p(Y \setminus D | D)$, $p(t | D)$, etc. Note that both parametric and predictive inference is encompassed here — the simulations leads to approximations to posterior inferences about the superpopulation parameters θ and, coincidentally, about the unknown values $Y \setminus D$ of the remaining population units.

With common models and a conjugate prior for θ , steps (a)–(c) are straightforward. That leaves (d), where the latent or missing data $Y \setminus D$ are generated as random sample with common

density proportional to $e^{-rw(y)}f(y|\theta)$. Recall that the observed data D are assumed sampled from $f(y|\theta)$ with probability proportional to $w(y)$, an increasing function of y , so that the remaining cases $Y \setminus D$ will tend to be smaller. Since $r > 0$ the term $e^{-rw(y)}$ modifies the original density to a form with greater mass on smaller values; the latent data are now effectively a selection sample from an infinite population model with selection probability proportional to this modifying factor. Heuristically, this makes sense: in the iterations, small values of $Y \setminus D$ lead to a small value of the sum of their weights t , which in turn leads to larger values of the exponential quantities ϕ_i and hence a larger value of their sum r ; this implies a selection probability that is biased towards small values of y in generating the sample $Y \setminus D$ for the next iteration. Technically, we still have the problem of simulating these selection samples. Only in very special models will this be direct. For example, if $f(y|\theta)$ is a gamma density, or a finite mixture of gamma densities, and $w(y) = y$, it is easily seen that the selection distribution is also a gamma, or a finite mixture of gammas, hence directly simulated. Otherwise, it is not typically the case that the selection distribution is a standard form, the following general rejection technique may be used. Note that $e^{-rw(y)} = P(x > r|y)$ where x is a random quantity with the conditional distribution $(x|y) \sim Ex(w(y))$. Then the required distribution may be simulated via rejection, each draw from $e^{-rw(y)}f(y|\theta)$ being generated as follows:

- (i) sample y from $f(y|\theta)$ and, independently, $u \sim U(0, 1)$;
- (iii) if $\log(1 - u) > -rw(y)$ reject the value of y and proceed again to (i); otherwise save the value of y and exit.

EXAMPLE.

Oil deposit data from Nair and Wang (1989) are analysed under a log-normal superpopulation model. Data in that paper include observations on various estimated size characteristics of oil pools in the Rimbey-Meadowbrook reef chain of the western Canadian sedimentary basin in central Alberta. The data records cover all pools discovered between 1947, when the first discovery was made, and 1968. There are $n = 23$ pools, and $N = 40$ is assumed by Nair and Wang for their maximum likelihood estimation procedures. For illustration, analysis here focuses on the records of *net pay* per pool, measured in metres in the penultimate column of Table 1 of that reference. Now $f(y|\theta)$ is the density function of the superpopulation distribution defined by $(\log(y)|\theta) \sim N(\mu, \sigma^2)$. Here $\theta = (\mu, \sigma^2)$, and analyses reported below are all based on the reference prior $p(\theta) \propto \sigma^{-2}$. In the area of application, there exists substantial expertise that may be used to explore ranges of informative priors and hence ranges of resulting inferences; the reference prior is simply a benchmark, as usual, with which alternative analyses and other (non-Bayesian) approaches may be compared. Finally, selection bias is modelled via $w(y) = y$.

Analysis assuming $N = 40$ pools in total is partially summarised in Figure 2. The simulation computations are burnt-in for 1000 iterations then the subsequent 5000 samples from the complete joint posterior contribute to the final approximations. The twenty smallest data values in D are indicated as stars on the axis in this frame – the three largest values exceed the upper limit on the axis so do not appear in the frame. The full line in Figure 2(b) is the average of all 5000 such densities – the simulation based estimate of the predictive density for a future pool size (without selection bias) $p(y|D) = \int f(y|\theta)p(\theta|D)d\theta$. This is a Bayes' estimate of the superpopulation density. Figure 2(a) displays just 50 conditional log-normal densities, $f(y|\theta)$ for 50 of the sampled values of θ (equally spaced through the complete 5000 draws) to give an impression of uncertainty about the

function – sampled curves from the posterior for $f(\cdot|\theta)$. For comparison, the dotted line in 2(b) is the log-Student-t predictive density from the reference analysis of the $n = 23$ observations ignoring selection bias; this clearly puts greater mass on overly large values due to the neglect of the bias issue.

Additional analyses support the adequacy of the numerical approximations, different initial values and burn-in sample sizes leading to similar graphs. One additional analysis was done using using $N = 23$; in this case, the ‘correct’ and ‘incorrect’ predictive densities of Figure 2(b) theoretically coincide since the entire population is sampled. The numerical computations lead to graphs that do indeed coincide almost exactly, adding further support to the adequacy of the numerical approximations in this example.

Figure 2 goes about here

3.2. UNKNOWN POPULATION SIZE

Analysis in Nair and Wang (1989) is based throughout on a known population size N . The authors comment on the difficulties inherent in the problem of simultaneously estimating θ and N (even ignoring the difficulties of a precise definition of oil pool so as to imply a fixed and unique value of N). The authors also note that, in the oil reserve context, there is a degree of externally available expert opinion and data-based geographical/geophysical information that could be used to form ranges of informative priors for N and θ . If this is done, then joint inference is feasible. Intuitively, the information about N derivable from the observed data alone is small unless the prior for θ is reasonably informative. To explore this, we need to be able to extend the analysis of the previous section to compute posteriors for N and θ jointly. We may do this by rather neat extension of the simulation based analysis, now described.

Assume a prior distribution for the population size N and, as earlier, that the observed sample size n is uninformative apart from the logical constraint $N \geq n$. Write $m = N - n$ so that the implied prior mass function $p(m)$, ($m = 0, 1, \dots$), is obtained from the original prior for N by simple truncation. Note the implicit additional assumption of independence between m and θ a priori. This may be appropriate in some circumstances, though more generally the framework admits specification of a conditional prior $p(m|\theta)$ – if, for example, prior information relates to expected total oil in the play, then prior dependence between m and θ is implicit and can be incorporated via conditional priors. This is not explored further here. Recall the joint density for D and $Y \setminus D$ in equation (1), and now explicitly include the uncertain remaining population size m in the conditioning so that (1) is denoted $p(D, Y \setminus D|\theta, m)$. Augmenting the conditioning of this joint density with the latent exponential variables Φ introduced in (IIa) of the previous section, recalling $r = \sum_{i=1}^n \phi_i$. This leads to

$$p(D, Y \setminus D|\theta, \Phi, m) \propto \frac{(m+n)!}{m!} \prod_{j=1}^N e^{-r w(y_j)} f(y_j|\theta)$$

where we substitute $m = N - n$ and identify all terms involving m . Thus

$$\begin{aligned} p(m|D, \theta, \Phi) &\propto p(m) p(D|\theta, \Phi, m) \propto p(m) \int p(D, Y \setminus D|\theta, \Phi, m) dy_{n+1} \dots dy_{n+m} \\ &\propto p(m) \frac{(m+n)!}{m!} \gamma(r, \theta)^m \end{aligned} \tag{5}$$

where, for any given r and θ ,

$$\gamma(r, \theta) = \int_0^\infty e^{-rw(y)} f(y|\theta) dy. \quad (6)$$

Simulation from this class of conditional posteriors for m extends the analysis summarised in the previous section, the iterations detailed in points (a)–(e) being modified slightly as follows:

- (a) Choose initial values of the remaining population size m and $Y \setminus D$, and compute the value of $t = \sum_{i=n+1}^{n+m} w(y_i)$;
- (b) based on the current values of m and $Y \setminus D$, sample θ from (2);
- (c) based on the current value of t , sample Φ using the independent exponential distributions under (IIa), and save only the value of $r = \sum_{i=1}^n \phi_i$;
- (d) based on the current values of r and θ , sample m from (5);
- (e) based on the current values of m, r and θ , draw a random sample of size m from the distribution with density (4) to produce a revised $Y \setminus D$;
- (f) proceed to (b), and iterate.

It remains to develop an algorithm to sample the posterior (5) given any specific prior for m . A natural choice assumes an initial truncated Poisson distribution for the total population size N . A Poisson distribution with specified rate λ , and subject to $N > 0$, induces the prior $p(m) \propto \lambda^m / (m + n)!$ over $m = 0, 1, \dots$, so that (5) becomes

$$p(m|D, \theta, \Phi) \propto \{\lambda\gamma(r, \theta)\}^m / m!,$$

and hence m has a conditional posterior that is Poisson with reduced rate $\lambda\gamma(r, \theta)$. Given r and θ , this is trivially simulated once the number $\gamma(r, \theta)$ is evaluated. Unfortunately, for many interesting models, this term will need numerical evaluation, though the required one-dimensional numerical integration in (6) can be easily and efficiently performed – see the following example. Note finally that the Poisson prior here might be generalised to a finite mixture of Poisson distributions of known rates, allowing greater flexibility in the specification of prior information; the resulting analysis may be easily performed with such priors, simply extending the above discussion.

EXAMPLE.

In the log-normal example for the oil reserve data, and with $w(y) = y$, the integral (6) is an evaluation of the moment generating function of a log-normal density which cannot be performed in closed form. The integral can be written as

$$\gamma(r, \theta) = \int_{-\infty}^{\infty} e^{-re^x} \phi((x - \mu)/\sigma) dx / \sigma$$

where $\phi(\cdot)$ is the standard normal density function. 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.

Figure 3 presents summaries of the analysis in which the rate of the prior Poisson distribution for N is $\lambda = 40$. Recall that analysis in Nair and Wang (1989), and that in the previous section here, assumed N fixed at 40, so this prior supports a range of possibilities including this nominal value. Posteriors in Figures 3(c) and 3(d) are analogous to those in 2(a) and 2(b), respectively.

The posterior predictive density $p(y|D)$ – the full line in 3(d) – is similar to that of the earlier analysis, though somewhat less peaked near zero. Uncertainty about the population size N is evident in the histogram of sampled values representing $p(N|D)$ in Figure 3(b). The apparent concordance with the Poisson prior indicates little in the way of additional information about N from the data, as expected. Figure 3(a) presents a corresponding histogram approximation to $p(t|D)$ where $t = \sum_{i=n+1}^{n+m} y_i$, the total of the net-pay values remaining. Alternative approaches to analysis are considerably complex when it comes to evaluating predictive distributions for even very simple function (like t) of uncertain quantities; by comparison, the simulation analysis leads trivially to approximate posteriors in histogram form.

A similar set of graphs appears in Figure 4, but now based on a Poisson prior with $\lambda = 20$. This prior puts decreasing mass on values of N exceeding the observed $n = 23$ so suggesting that sampling has been much more exhaustive than suggested by the previous prior. The resulting posterior in 4(b) suggests this prior conflicts somewhat with the data, posterior mass favouring rather larger values of N though heavily constrained by the prior. Due to the focus on much smaller values of N , the resulting posterior for θ is very much closer to that assuming the n values do in fact exhaust the population – the dashed line in 4(d). Again the concordance between curves in 4(d) helps to validate approximation accuracy in the simulation analysis.

Figures 3 and 4 go about here

3.3. EXTENSIONS

(a) *Multivariate superpopulation*

The data of Nair and Wang (1989) is multivariate, the net-pay observations analysed above representing just one of several size characteristics of oil pools. Those authors develop methods for inference based on the entire multivariate data set in cases when the selection function $w(\cdot)$ depends on only a linear function of the observation per pool. The Bayesian analysis presented above also extends easily to that case, and will be reported in West (1992) with application to the oil pool discovery data.

(b) *Weight function $w(y) = y^c$ for $c > 0$*

Nair and Wang (1989), and other authors, have considered selection functions of the form $w(y) = y^c$, for some positive and known constant c , as alternatives to the linear weight function used in the examples above. Of course, the theory of this paper applies to any weight function whatsoever. Problems of estimating the selection function are outside current scope, however, and would apparently be difficult unless additional prior information about θ and/or c were available. In such a generalised model, estimation of c would allow assessment of the relative support for the special case $c = 0$, of random sampling, and hence provide the means to explore whether or not selection effects are “real”. Other parametric forms of weight function might also be considered.

(c) *Missing data*

Analysis may be extended to problems in which one or more of the ordered values in D are missing. Data may arise with values missing at random, so requiring such extension. Alternatively, individual values, or sets of values, may be chosen for omission from an analysis in order to assess

their influence on inferences when included. For example, values x_2 and x_5 in the net pay data analysed above appear somewhat smaller than perhaps anticipated so early in the sequence, and might be dropped for analysis to be reperformed (though, in this case, the differences between analyses with and without these values are small.)

Suppose just x_k is missing for a given k , $1 \leq k \leq n$ (details extend trivially to cases of more than one missing value.)

Write $Y \setminus y_k = \{Y, D\} - \{y_k\}$. Expanding equation (1) to include the latent variables Φ we can easily deduce

$$p(y_k | Y \setminus y_k, \Phi, \theta) \propto w(y_k) e^{-r_k y_k} f(y_k | \theta)$$

where $r_k = \sum_{i=1}^k \phi_i$. Now the simulation analysis can be augmented by a step that simulates y_k from the conditional posterior with this density at each stage (a) of the iterations. Notice that this sampling involves two weights $w(y)$ and $e^{-r_k y}$. The first simply biases sampling towards larger values than under $f(y|\theta)$ – as in sampling the infinite population model with size bias. The exponential term operates as it does in sampling the remaining values $Y \setminus D$. Note, however, that $r_k < r$, implying the bias towards smaller values is less than in sampling $Y \setminus D$. Also, r_k is an increasing function of k to increase the bias towards smaller values of y_k for larger k .

This extension of the analysis is computationally trivial in the log-normal model with $w(y) = y^c$. It is easily seen that $w(y)f(y|\theta)$ is the density of $(\log(y)|\theta) \sim N(\mu + c\sigma^2, \sigma^2)$, so sampling the missing value uses the same rejection method as in sampling the elements of $Y \setminus D$ but with this corrected underlying log-normal distribution.

(d) *Nonparametric superpopulation models*

Gibbs sampling approaches to data analysis using Dirichlet process mixture models (Escobar and West, 1992) naturally lends itself to the current context. Assume, for example, that the superpopulation distribution is modelled as a mixture of log-normal distributions. The developments in Escobar and West show how simulation based analysis of random samples from such models easily generates algorithms to sample from posterior predictive densities – the Bayesian density estimates. These algorithms may be nested in the iterative computational schemes of the current paper to provide non- or semi-parametric inference about superpopulation distributions. Work in this direction will be reported elsewhere.

ACKNOWLEDGEMENTS

This research was supported in part by the National Science Foundation, grants DMS 89-03842 DMS 90-24793.

BIBLIOGRAPHY

- Bayarri, M.J. and DeGroot, M.H. (1987). Bayesian analysis of selection models. *The Statistician*, **36**, pp. 137-146.
- Escobar, M.D., and West, M. (1992). Bayesian density estimation and inference using mixtures. Invited revision for *J. Amer. Statist. Assoc.*
- Gelfand, A.E., and Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85**, pp. 398-409.
- Irony, T.Z. (1992). Information in sampling rules. *J. Stat. Plan. Inf.* (to appear).
- 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.
- 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.
- Sanathanan, L. (1977). Estimating the size of a truncated sample. *J. Amer. Statist. Assoc.* **72**, pp. 669-672.
- Tanner, M.A., and Wong, W.H. (1987). The calculation of posterior distributions by data augmentation (with discussion). *J. Amer. Statist. Assoc.* **82**, pp. 528-550.
- West, M. (1992). Inference in successive sampling discovery models. ISDS discussion paper #92-A21, Duke University.

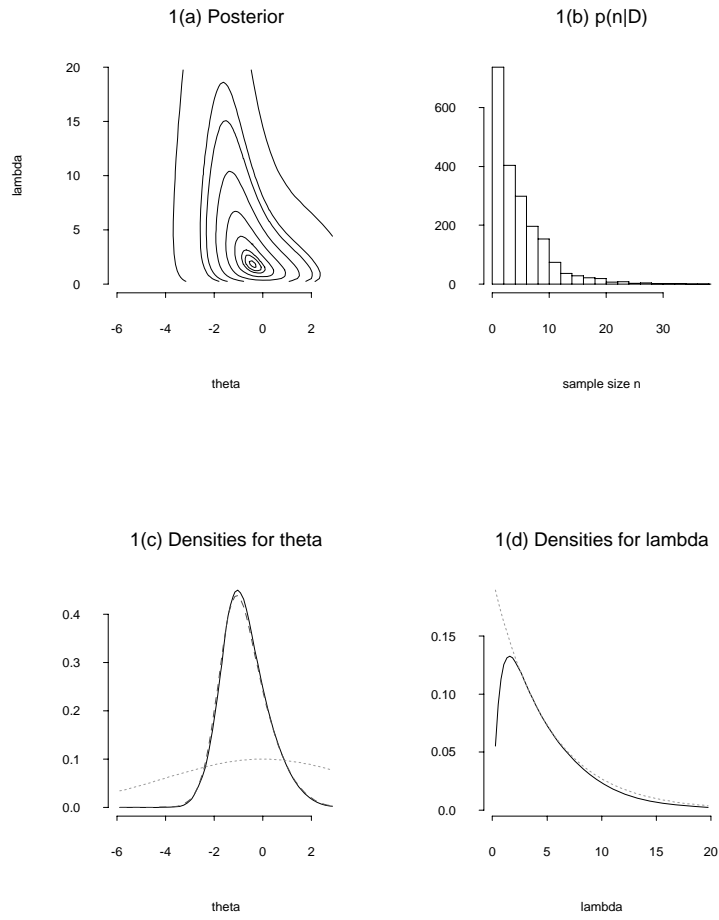


Figure 1. Posteriors in truncated normal data analysis.

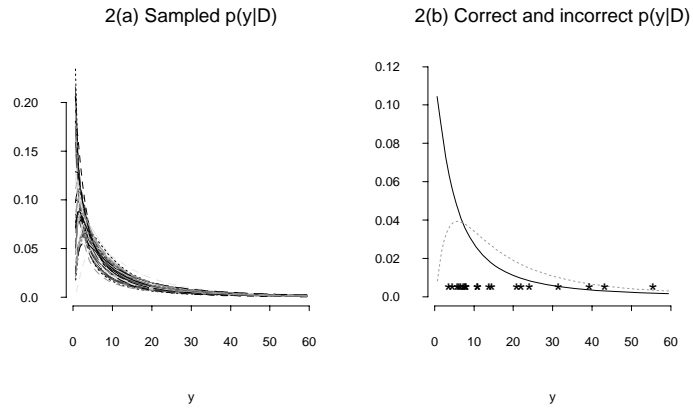


Figure 2. Oil reserve data analysis with $N = 40$.

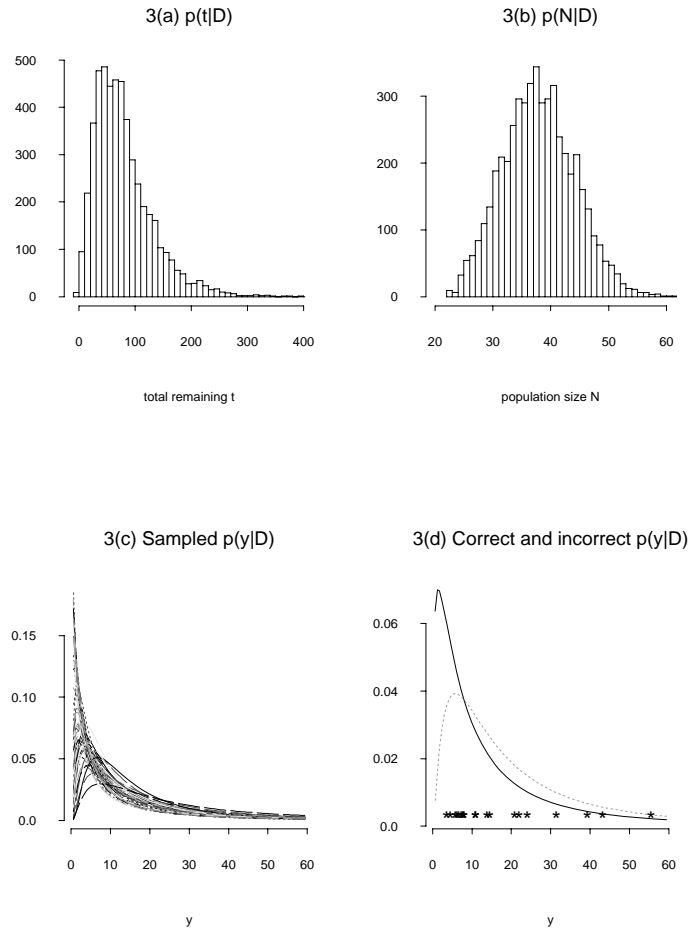


Figure 3. Oil reserve data analysis with $\lambda = 40$.

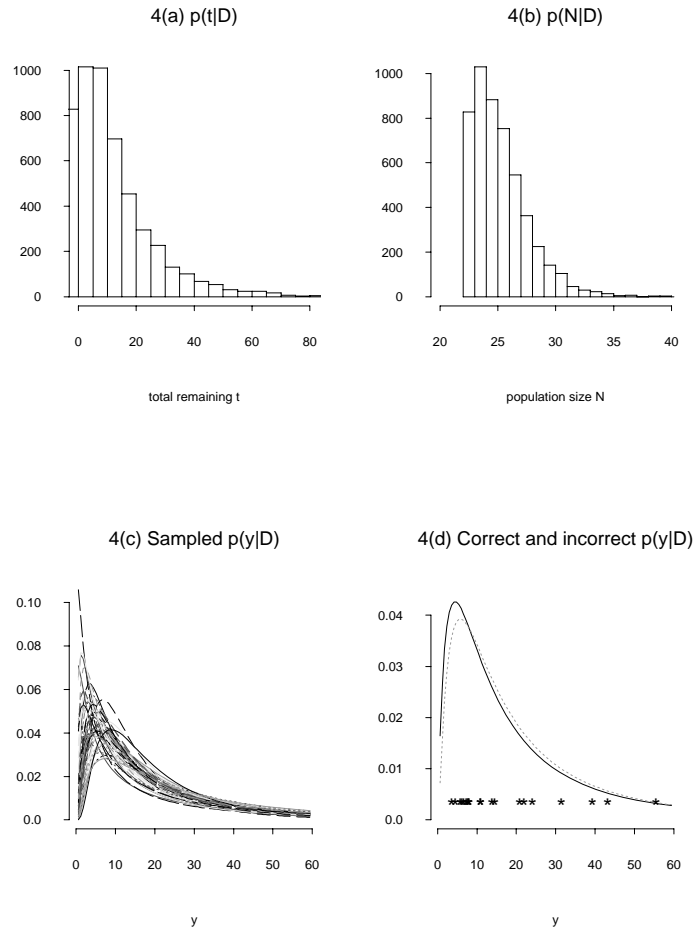


Figure 4. Oil reserve data analysis with $\lambda = 20$.