Journal of Official Statistics, Vol. 24, No. 1, 2008, pp. 21–51

Applications of the Bayesian Bootstrap in Finite Population Inference

Murray Aitkin[1]

This article extends the Bayesian bootstrap analysis, applied to studies of single finite population survey data in Hartley and Rao (1968), Ericson (1969), and Rubin (1981), to regression models for numerically-valued response variables in stratified and clustered samples.

This extension provides an alternative Bayesian analysis of complex finite population survey data which for some applications requires only standard statistical modeling software to implement.

Key words: Survey sampling; stratification; clustering; Dirichlet prior.

1. Introduction

The Bayesian bootstrap approach to inference in finite population survey sampling was described by Hartley and Rao (1968), Ericson (1969), and Rubin (1981). This approach does not use a parametric model assumption for the distribution of a numerically-valued response variable Y in the finite population of size N. Instead, the key to the approach is the *multinomial model* for the finite population, when tabulated by the distinct values $Y_1 < \cdots < Y_J < \cdots < Y_D$ which the variable Y can take. As these values are always measured with finite precision, denoted by δ, the possible values of Y form an equally spaced discrete grid of values Y_J with step-length δ, with counts N_J and proportions $p_J = N_J/N$ at Y_J. Population parameters like the mean and variance can be expressed as functions of the proportions p_J

$$\mu = \sum_J p_J Y_J$$

$$\sigma^2 = \sum_J p_J (Y_J - \mu)^2$$

A simple random sample from the population can be correspondingly expressed through the sample counts n_J at Y_J (most of these will be zero). If the sample size n is small

[1] University of Melbourne, Department of Psychology, Melbourne Vic 3010, Australia. Email: murray.aitkin@unimelb.edu

Acknowledgments: We appreciate support for this work from Gary Phillips. This work has been funded partly by the Australian Research Council under grant number DP0559684 and partly by Federal funds from the U.S. Department of Education, National Center for Education Statistics, under contract number RN95127001. The content of this publication does not necessarily reflect the views or policies of the U.S. Department of Education, National Center for Education Statistics, nor does mention of trade names, commercial products or organizations imply endorsement by the U.S. Government. Tom Chadwick contributed to part of this work.

compared to the population size N, (the alternative case is considered later), so that sampling with replacement accurately approximates sampling without replacement, the multinomial probability of the sample counts n_J is

$$\Pr(n_1, \ldots, n_D) = m(n; p_1, \ldots, p_D) = \frac{n!}{\prod_{J=1}^{D} n_j!} \prod_{J=1}^{D} p_J^{n_J}$$

Given the sample counts, the likelihood is the *multinomial* likelihood

$$L(p_1, \ldots, p_D) = \prod_{J=1}^{D} p_J^{n_J}$$

where the term $n!/\prod_{J=1}^{D} n_j!$ is a known constant and can be omitted from the likelihood.

Formally, we need to know the smallest and largest values of Y, and the number of distinct values D in the population, to be able to compute this likelihood, but for any unobserved values of Y_J the corresponding n_J is zero, so the likelihood can be reexpressed in terms of the p_j for only the observed Y_J. Thus the p_J for the unobserved Y_J do not contribute to the likelihood, and so these Y_J do not need to be known unless the prior gives them nonzero weight.

Maximizing the multinomial likelihood over the p_J for a fixed μ gives the "empirical" profile likelihood in μ (Owen 1988), discussed extensively in Owen (2001). A fully Bayesian analysis follows from the specification of a prior for the p_J; a very convenient choice is the natural conjugate Dirichlet prior, used by Hartley and Rao (1968), Ericson (1969), and Rubin (1981), which has density

$$\pi(p_1, \ldots, p_D) = C(a_1, \ldots, a_D) \prod_{J=1}^{D} p_J^{a_J - 1}$$

over the D-dimensional simplex $p_J > 0, \sum_{J=1}^{D} p_J = 1$, where $C(a_1, \ldots, a_D)$ is the normalizing constant

$$C(a_1, \ldots, a_D) = \frac{\Gamma\left(\sum_1^D a_J\right)}{\prod_1^D \Gamma(a_J)}$$

The posterior distribution is again Dirichlet

$$\pi(p_1, \ldots, p_D | y) = C(n_1 + a_1, \ldots, n_D + a_D) \prod_{1}^{D} p_J^{n_J + a_J - 1}$$

Priors and posteriors for functions of the p_J follow automatically from the Dirichlet prior for the p_J: no additional prior specifications are necessary (for example, for the mean μ).

The Dirichlet is a special case of the Dirichlet process prior (Ferguson 1973); this was used by Binder (1982) for the more general case of finite population values which are arbitrary real numbers. However for most real sampled populations with fixed

measurement or recording precision, the simpler equally spaced grid of population values is sufficient, and we restrict consideration to this case. The multinomial/Dirichlet model and prior have been proposed recently as the fundamental nonparametric distribution and prior model, in the work of Gutierrez-Pena and Walker (2005, 2007).

Since even in large samples many of the positive values of n_J will be 1 or a small integer, the choice of prior is more important than usual in parsimonious parametric models. The effective information provided by the prior is easily seen from the form of the posterior: the sample counts n_J are augmented by the prior "weights" a_J. The "total prior weight" $a = \sum_J a_J$ augments the total sample weight $n = \sum_J n_J$.

Ericson (1969) considered the proper prior with $a_J = \epsilon_J$ with $\epsilon = \sum_J \epsilon_J$ "small", of the order of 1. He showed that many standard survey sampling results followed as limiting cases as $\epsilon \rightarrow 0$, though he expressed reservations about the properties of such an unrealistically "rough" prior.

Rubin (1981) introduced the term *Bayesian bootstrap* for posterior inference with the improper Haldane prior with $a_J = 0 \forall J$. This prior is used by Gutierrez-Pena and Walker (2005, 2007).

It leaves the total sample weight unchanged, but has the curious property that for any value Y_J not observed in the sample, the posterior distribution for the corresponding p_J has a nonintegrable spike at the zero value of these p_J. This is equivalent to assigning zero prior probability to these unobserved values. The computation of the posterior distribution can then be restricted to the d observed distinct sample values y_J rather than the D distinct population values, a great saving. This saving is shared with Owen's empirical likelihood: the construction of the empirical profile likelihood depends only on the observed sample values and their sample frequencies.

The term "Bayesian bootstrap" comes from the analogy with the frequentist bootstrap, which resamples from the observed sample. The Bayesian bootstrap also uses only the observed sample, but it resamples from the *posterior distribution* of the *probabilities* attached to each observed value, rather than from the values themselves.

Rubin (1981, pp. 133–134) highlighted the difficulty with the Haldane prior approach:

". . . First, is it reasonable to use a model specification that effectively assumes all possible distinct values of [Y] have been observed?"

". . . Second, even assuming all distinct values of [Y] have been observed, is it reasonable to assume a priori independent parameters, constrained only to sum to 1, for these values? If two values of [Y] are "close," isn't it often realistic to assume that the associated probabilities of their occurrence should be similar? Shouldn't the parameters be smoothed in some way?"

Banks (1988) took up these criticisms by developing a smoothing of the Dirichlet posterior: given the Haldane prior, he proposed generating a random value of p_J for each observed Y_J, and then spreading it uniformly over this Y_J and all unobserved values to the left of this Y_J down to the next observed value. In this way the posterior mass was spread over the whole sample range from $y_{(1)}$ to $y_{(n)}$, though in an *ad hoc* way.

An apparently unreasonable model specification could be expected to perform poorly. We demonstrate the contrary with a simulation study of several methods. This evaluates the frequentist performance of the Bayesian bootstrap relative to other frequentist procedures, following the precept that to be useful, Bayes procedures need to be

well calibrated in the frequentist sense (Rubin 1987, p. 62). In large samples from parsimonious regular parametric models giving normal likelihoods, it is easily shown that with flat priors, $100(1-\alpha)\%$ credible intervals or regions are identical to $100(1-\alpha)\%$ likelihood-based confidence intervals or regions, and so have repeated-sampling confidence coverage of $100(1-\alpha)\%$. However the multinomial/Dirichlet model is nonparsimonious and the sample size at each sample support point is very small, so this result may not apply to credible intervals from the Bayesian bootstrap for derived parameters.

We first give a brief example.

1.1. Example – Income Population

The following simple random sample of $n = 40$ values of family income Y (in hundreds of 1962 dollars) at the birth of a child come from the StatLab boy population (Hodges, Krech, and Crutchfield 1975) of $N = 648$ families with a boy baby:

Family incomes for random sample of 40 families (hundreds of dollars)													
26	35	38	39	42	46	47	47	47	52	53	55	55	56
58	60	60	60	60	60	65	65	67	67	69	70	71	72
75	77	80	81	85	93	96	104	104	107	119	120		

The sample mean is $\bar{y} = 67.1$ and the (unbiased) variance is $s^2 = 500.87$. Figure 1 shows the full income population as a maximum resolution histogram. The classical large-sample 95% confidence interval for the mean is $\bar{y} \pm 1.96s/\sqrt{n}$, which is [60.1, 74.0]; this is nearly identical to the t-interval [60.0, 74.2] assuming a normal distribution for income. The design-based interval using the finite population correction of $(1 - 40/648) = 0.938$ gives the slightly shorter interval [60.6, 73.6]. If income could be assumed to be normally distributed, the equal-tailed 95% confidence interval based on the χ^2_{39} distribution for the income variance would be [336.1, 825.9].

For the Bayesian bootstrap analysis we tabulate the sample by the distinct values of Y. We first make an analogous notational change: since we use only the d ordered distinct sample values, we will denote them by y_j with sample frequencies n_j.

j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
y_j	26	35	38	39	42	46	47	52	53	55	56	58	60	65	67	69
n_j	1	1	1	1	1	1	3	1	1	2	1	1	5	2	2	1
j	17	18	19	20	21	22	23	24	25	26	27	28	29	30		
y_j	70	71	72	75	77	80	81	85	93	96	104	107	119	120		
n_j	1	1	1	1	1	1	1	1	1	1	2	1	1	1		

Simulation of the p_j, and therefore of any *marginal* function of the p_j, from the Dirichlet posterior with the Haldane prior is particularly simple: for a single simulation we generate

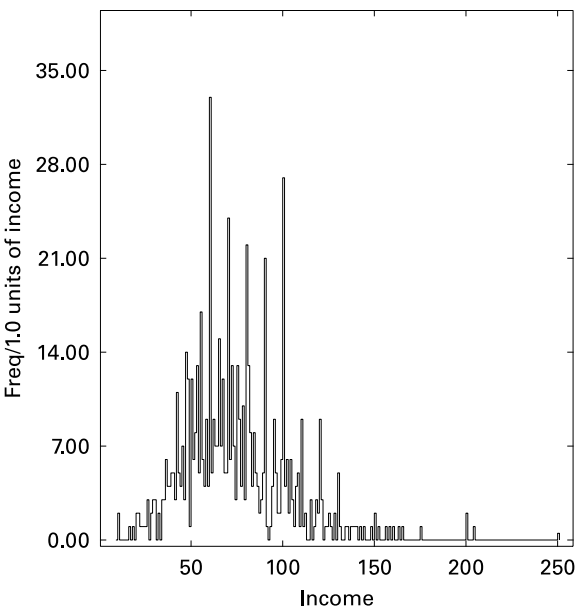

Fig. 1. Income histogram

$d = 30$ independent $U(0,1)$ values U_j, transform them to d independent gamma variables G_j with parameters 1 and n_J, and then define $p_j = G_j / \sum_j G_j$. Repeating the simulations M times gives M simulated values $p_j^{[m]}$ of the p_j, and hence M simulated values

$$\mu^{[m]} = \sum_j p_j^{[m]} y_j$$

from the marginal posterior distribution of μ.

We show in Figure 2 the posterior cdf of the mean μ from a simulation of size $M = 10,000$, and in Figure 3 a kernel density estimate using a bandwidth of 1.0, together with the simulated values.

(Approximate) percentiles of the posterior distribution can be read directly from Figure 2 (or from the list of ordered values). The posterior density in Figure 3 has only very mild skew. The sample median is 67.0, and the sample mean is 67.1. The 95% equal-tailed credible interval is [60.6, 74.2]; it is slightly shorter than the t-interval and slightly asymmetric.

Simulation is not restricted to the mean μ – we can simulate *any* parametric function of the p_J – the variance or standard deviation and higher moments are just as simple.

Figures 4 shows the joint marginal scatter of the $M = 10,000$ values of μ and σ for the income sample, and Figure 5 shows the joint posterior scatter of the standardized third and fourth cumulants of the income distribution. The point (0,0) is in the extreme edge of the point scatter in Figure 5: there is no question that the population is both skewed and heavy-tailed.

The 95% equal-tailed credible interval for the variance is [308.3, 708.0], substantially different from the normal-based interval, and the corresponding interval for the standard deviation is [17.6, 26.6].

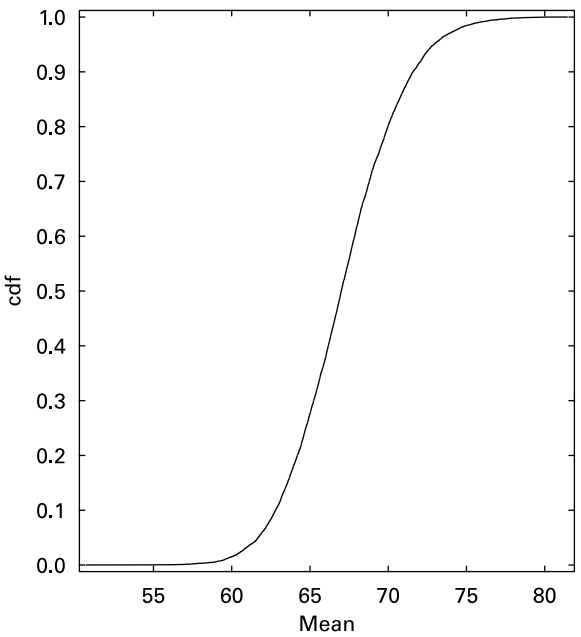

Fig. 2. Posterior cdf, income mean

1.2. Simulation Study

In the study we compared coverage of intervals for the mean based on the Bayesian bootstrap central credible intervals using the Haldane prior, with confidence intervals based on the gamma and normal distributions (the latter closely equivalent to the survey sampling intervals), and with (frequentist) bootstrap percentile intervals. As noted in the

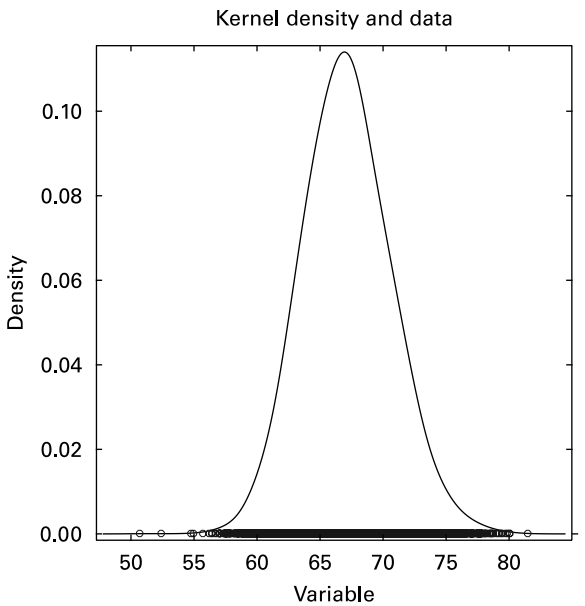

Fig. 3. Posterior density, income mean

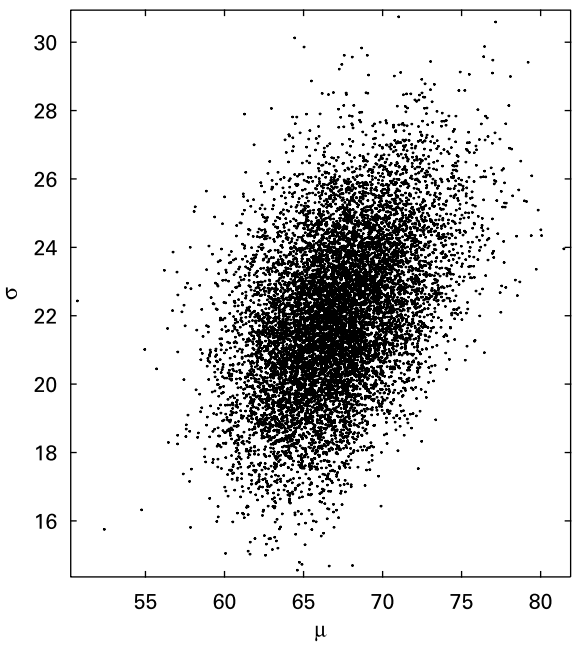

Fig. 4. Joint posterior, income mu and sigma

Introduction, we evaluate the frequentist performance of the Bayesian bootstrap relative to other frequentist procedures.

We drew 1,000 random samples of size 40 from the StatLab boy population of $N = 648$, and constructed 80%, 90%, and 95% confidence intervals for the mean

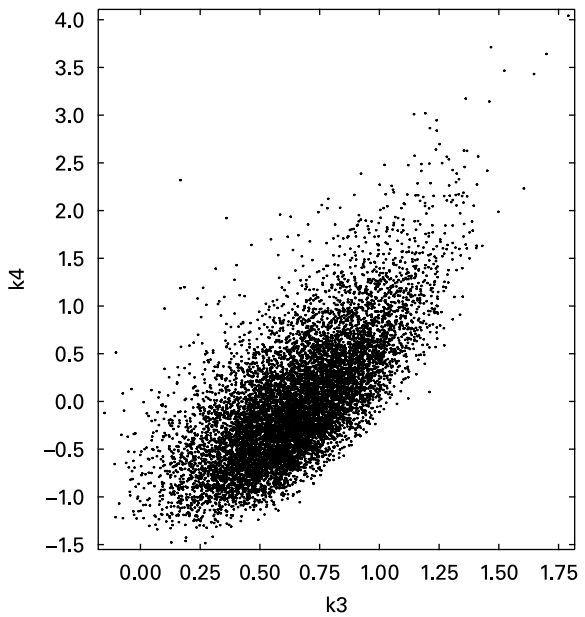

Fig. 5. Joint posterior, income k3 and k4

μ based on a normal income distribution (using t percentage points), and corresponding intervals based on a gamma income distribution with scale parameter r, from the sample mean and its estimated standard error $\bar{y}/\sqrt{n\hat{r}}$ using the MLE $\hat{r}$, and using t rather than the normal percentage points. (This choice provides an approximate "small-sample" adjustment to the asymptotic normal percentage points which also allows a direct comparison with the normal income distribution intervals – see below.) The StatLab income population is chosen for the simulation as it is a real one, and also has features which may be expected in other real populations, like irregularity, rounding and preference for multiples of 5 and 10.

For each sample we drew 10,000 bootstrap samples (sampling with replacement) and constructed the empirical 80%, 90%, and 95% bootstrap percentile intervals (equal-tailed) for the mean, and drew 10,000 Bayesian bootstrap samples from the posterior distribution of the mean based on the Haldane prior. We also extended the support to the full range of the observed sample, and used an Ericson-type Dirichlet prior with parameters $1/l$, where l is the number of support points in the grid from $y_{(1)}$ to $y_{(n)}$. The last prior gives an equivalent prior weight of 1 compared to the sample weight of 40. From these samples we constructed the equal-tailed 80%, 90%, and 95% credible intervals for the mean.

We give in Table 1 below, in the first panel the average length of the intervals across the 1,000 samples, and in the second panel the actual coverage (c) of the intervals, as well as the proportion of left (lnc) and right (rnc) noncoverage.

Apart from the Ericson prior, the table shows a consistent pattern: the interval lengths decrease slightly across the columns, and the coverages decrease slightly. The intervals based on the Ericson prior on the extended observed support behave qualitatively differently: compared with the Haldane prior, interval lengths increase but coverages *decrease*. Larger prior weight on the unobserved values accentuates this effect (results not shown): the effect of further increasing prior weight on unobserved values of Y is to further decrease the coverage and increase the length of the credible intervals.

Table 1. Average interval length and coverage, $n = 40$

Coefficient	Normal	Gamma	Boot	Haldane	Ericson
80%	11.90	11.68	11.54	11.26	11.32
90%	15.38	15.09	14.81	14.57	14.63
95%	18.46	18.12	17.64	17.49	17.55
c	0.799	0.789	0.777	0.770	0.767
80% lnc	0.079	0.085	0.092	0.097	0.112
rnc	0.122	0.126	0.131	0.133	0.121
c	0.890	0.892	0.879	0.876	0.872
90% lnc	0.038	0.039	0.046	0.047	0.056
rnc	0.072	0.069	0.075	0.077	0.072
c	0.945	0.941	0.933	0.927	0.925
95% lnc	0.017	0.020	0.023	0.030	0.034
rnc	0.038	0.039	0.044	0.043	0.041

Two criticisms can be made of the Ericson prior:

- it biases the posterior mean towards the sample median, which is inappropriate since the sample income distribution is clearly skewed;
- it cannot be assigned until the data are observed, so it is a post-data prior.

These criticisms illustrate the difficulty of the Dirichlet approach if it requires prior assignment to unobserved values – how is this to be done? However, this comparison shows that even a prior weight of 1, over a conservative range, compared to the sample weight of 40 results in poorer coverage and longer intervals than the Haldane prior. We now consider the comparison of the Haldane prior intervals with those from the other methods.

The pattern of shorter interval lengths with reduced coverage makes it difficult to compare the methods directly – would a method with lower coverage but shorter intervals than another method have the same, better or worse coverage if the interval length were increased to match that of the other method? We address this question by modeling the coverage probabilities using a probit analysis with interval length as an explanatory variable; the interval methods are an explanatory factor which is tested for significance in the analysis. We regress the probits of the coverage probabilities against the interval lengths for the sample size of 1,000, with method (5 levels) and nominal coverage (3 levels) as explanatory factors.

This analysis shows that coverage probability is very strongly determined by interval length; method and nominal coverage show no significant variation once interval length is included; the simple interval length model has a goodness-of-fit χ^2 value of 2.24 with 13 degrees of freedom.

Thus the methods are equivalent in coverage after adjustment for interval length: the survey sampling, model-based gamma, frequentist bootstrap and Haldane prior interval methods perform equally well in coverage, though the survey sampling and model-based gamma methods have the closest to nominal coverage.

The apparently "unreasonable" Haldane prior provides the best set of credible intervals: apparently more reasonable priors which do not exclude unobserved values perform less well than the Haldane prior.

1.3. Extensions of the Bayesian Bootstrap

The above discussion of the Bayesian bootstrap analysis is limited to simple random sampling and moment parameters of the multinomial distribution of *Y*. In the remainder of this article we extend the Bayesian bootstrap approach in several directions. Section 2 extends the one-sample approach to sampling without replacement, using the approach of Hoadley (1969). Section 3 discusses regression models and gives an example supported by a simulation study. Section 4 extends the implicit multinomial model to multiple subpopulations. This provides the analysis for stratified sampling in Section 5 and for cluster sampling in Section 6. Section 7 discusses a complex example of regression in a stratified and clustered sample. Section 8 has discussion and conclusions; it will be clear from the extensions that the Bayesian bootstrap approach can handle survey designs of considerable complexity.

2. Sampling Without Replacement

The multinomial likelihood construction in Section 1 is based on the assumption that the sample size n is small compared with the population size N. When this is not so, we need a more careful construction of the likelihood. The probability that a sample of size n contains n_J of the N_J values of Y_J in the population is now the hypergeometric

$$\Pr(n_1, \ldots, n_D) = \left[\prod_{J=1}^{D} \binom{N_J}{n_J}\right] / \binom{N}{n}$$

Given the sample values $n_1, \ldots, n_D$, the likelihood in the parameters $N_1, \ldots, N_D$ is the *hypergeometric likelihood*

$$L(N_1, \ldots, N_D) = \prod_{J=1}^{D} \binom{N_J}{n_J}$$

where the known constant denominator is omitted, and the N_J must be larger than or equal to n_J. Since $\binom{N}{0} = 1$, the zero counts can again be omitted from the likelihood, which can be expressed in terms of only the observed sample counts n_j:

$$L(N_1, \ldots, N_d) = \prod_{j=1}^{d} \binom{N_j}{n_j}$$

Thus the sample is again uninformative about the population counts at unobserved values of Y. (Note that if all $n_j = 1$, $L(N_1, \ldots, N_d) = N_1 \cdot N_2 \cdots N_d$, for the observed values $y_1, \ldots, y_d$. This *is* informative about these values of N_j.)

The population counts are not free parameters: they must satisfy $N_j \geq n_j$. Write $N_J^* = N_J - n_J$; we take the $N_J^* \geq 0$ to be the unknown parameters, with $\sum_{J=1}^{D} N_J^* = N - n = N^*$. The likelihood in the N_j^* is

$$L(N_1^*, \ldots, N_d^*) = \prod_{j=1}^{d} \binom{N_j^* + n_j}{n_j}$$

For this form of likelihood there is no simple conjugate prior distribution for the N_J^*. Following Hoadley (1969), we embed the model in *two* levels of prior distribution.

Conditional on category proportions p_j, we treat the d unobserved population counts N_J^* as drawn from a multinomial distribution

$$m(N^*; p_1, \ldots, p_d) = \frac{N^*!}{\prod_{j=1}^{d} N_j^*!} \prod_{j=1}^{d} p_j^{N_j^*}$$

in which the probabilities p_j, conditional on the observed sample sizes n_j, have the Dirichlet distribution of Section 1:

$$\pi(p_1, \ldots, p_d | n_1, \ldots, n_d) = \frac{\Gamma(n)}{\prod_{j=1}^{d} \Gamma(n_j)} \prod_{j=1}^{d} p_j^{n_j - 1}$$

Integrating out the p_j gives a *compound multinomial distribution* as the posterior distribution of the N_j^* given the n_j:

$$\begin{aligned}
&\Pr[N_1^*, \ldots, N_d^* | n_1, \ldots, n_d] \\
&\quad = \int \cdots \int \Pr[N_1^*, \ldots, N_d^* | p_1, \ldots, p_d] \cdot \Pr[p_1, \ldots, p_d | n_1, \ldots, n_d] \mathrm{d}p_1 \ldots \mathrm{d}p_d \\
&\quad = \int \cdots \int \frac{N^*!}{\prod_{j=1}^{d} N_j^*!} \prod_{j=1}^{d} p_j^{N_j^*} \cdot \frac{\Gamma(n)}{\prod_{j=1}^{d} \Gamma(n_j)} \prod_{j=1}^{d} p_j^{n_j - 1} \mathrm{d}p_1 \ldots \mathrm{d}p_d \\
&\quad = c \cdot \frac{N^*!}{\prod_{j=1}^{d} N_j^*!} \prod_{j=1}^{d} \frac{\Gamma(N_j^* + n_j)}{\Gamma(N^* + n)}
\end{aligned}$$

This distribution does not lend itself to direct simulation, but its integral derivation provides a very simple indirect sampling formulation for simulation of the mean μ (or other parameters):

- generate M values $p_j^{[m]}$ of the p_j as in Section 1;
- from these, generate M values $N_j^{*[m]}$ from the multinomial distributions $m(N^*; p_1^{[m]}, \ldots, p_d^{[m]})$
- calculate the M values $\mu^{[m]} = \sum_{j=1}^{d} (N_j^{*[m]} + n_j) Y_j / \sum_{j=1}^{d} (N_j^{*[m]} + n_j)$.

This approach avoids completely the awkward form of the posterior in the N_j^*, and requires only the additional multinomial simulation step. An alternative approach to the unobserved N_j, not used in this article, is the Polya urn model (Ghosh and Meeden 1997, p. 42).

2.1. Simulation Study

We replicate part of the simulation study in Section 1, with the same sample size from the StatLab population, to compare the posterior distributions of the mean for sampling with and without replacement. We restrict the study to just the two posteriors based on the same Haldane prior-based posterior for the p_j. Results are given in Table 2.

The intervals based on the hypergeometric likelihood are shorter, but have lower coverage, than those based on the multinomial likelihood. Adjusting again for interval length, the coverages are equivalent – the deviance for the single interval length model is 0.046 on 4 df. Recognizing the finiteness of the population does not bring increased precision in inference about its mean; the sample fraction of 40/648 = 0.062 is not sufficiently large to give the theoretical improvement.

Table 2. Average credible interval length and coverage, $n = 40$

Coefficient	With rep	Without rep
80%	11.34	10.99
90%	14.67	14.21
95%	17.61	17.06
c	0.756	0.741
80% lnc	0.126	0.134
rnc	0.118	0.125
c	0.858	0.846
90% lnc	0.069	0.077
rnc	0.073	0.077
c	0.923	0.913
95% lnc	0.036	0.041
rnc	0.041	0.046

3. Regression Models

We want to relate an outcome variable Y to an explanatory variable X through a regression. We use an example from Royall and Cumberland (1981) for illustration.

Royall and Cumberland (1981) discussed a finite population of 393 short-stay hospitals for which data were available on the number of patients Y discharged in one year and the number of hospital beds X in that year. The data came from the NCHS Hospital Discharge Survey, a national sample of short-stay hospitals with fewer than 1,000 beds (Herson 1976). We treat this as the population for this study. A simple random sample of size $n = 32$ is shown in Figure 6 and the values of Y and X are given below.

Number of patients Y and hospital beds X								
Y	1,076	577	1,258	134	795	1,219	486	1,095
X	260	128	474	118	261	400	154	400
Y	1,040	297	22	625	955	1,948	1,084	57
X	418	74	20	192	228	461	247	10
Y	828	487	795	1,326	2,031	2,089	518	695
X	145	159	261	584	509	712	103	200
Y	247	635	1,231	609	337	490	389	479
X	57	185	374	265	145	244	110	180

We are interested in the total number of short-stay patients across the population, and we assume a simple proportionality of the number of such patients in each hospital to the number of beds in that hospital. We know from administrative records the number of beds X in each hospital and hence the total number of beds T_X in all

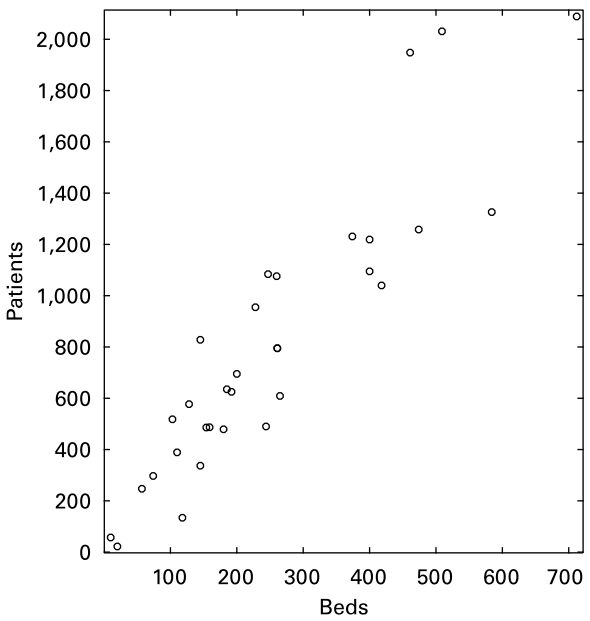

Fig. 6. Patients and beds for hospital sample

the hospitals ($T_X = 107,956$), and draw an SRS of hospitals of size n, recording in each hospital the number of short-stay patients and the number of beds. From the sample data we want to estimate the total number T_Y of short-stay patients in all the hospitals.

Figure 6 shows that the variance of Y is clearly increasing with X, so the *ratio estimator* is an appropriate choice in the survey sampling approach. The ratio estimator is

$$\hat{T}_Y = \frac{\sum_i y_i}{\sum_i x_i} \cdot T_X = \frac{\bar{y}}{\bar{x}} T_X = \hat{B} T_X$$

where $\hat{B} = \sum_i y_i / \sum_i x_i$. From a model-based viewpoint, this estimator would be optimal (in the weighted least squares sense) under a model in which Y has mean BX and variance $\sigma^2 X$.

3.1. Design-Based Approach

We introduce an additional notation, by indexing the population values by $I^* = 1, \ldots, N$, and define the population *ratio regression coefficient* by

$$B = \frac{\sum_I Y_{I^*}}{\sum_I X_{I^*}} = \frac{T_Y}{T_X} = \frac{\mu_Y}{\mu_X}$$

and the population total T_Y estimate by $\hat{T}_Y = \hat{B}T_X$. Let Z_{I^*} be the sample selection indicators, with $Z_{I^*} = 1$ if population member I^* is included in the sample, and $Z_{I^*} = 0$ otherwise. We have

$$\hat{B} = \frac{\sum_{I^*} Y_{I^*} Z_{I^*}}{\sum_{I^*} X_{I^*} Z_{I^*}}$$

The sampling distribution of this ratio is complicated by the appearance of the Z_{I^*} in both numerator and denominator. Exact results are not available, but and approximate variance for $\hat{B}$ is $(1 - n/N)s_e^2/(n\mu_X^2)$ (Lohr 1999, p. 68), where

$$s_e^2 = \sum_i \frac{(y_i - \hat{B}x_i)^2}{n-1}$$

and this can be used to construct approximate confidence intervals for B, and hence for T_Y. An alternative robust sandwich variance estimate is obtained by replacing the normal model variance estimate $\sigma^2 / \sum x_i$ by

$$\mathrm{Var}[\hat{B}] = \frac{\sum_i \mathrm{Var}[Y_i]}{\left(\sum_i x_i\right)^2} \doteq \frac{\sum_i (y_i - \hat{B}x_i)^2}{\left(\sum_i x_i\right)^2}$$

where the variance model is not assumed to be correct.

3.2. *Bayesian Bootstrap Approach*

We follow the same approach as in the simple mean model. The population consists of N pairs Y_{I^*}, X_{I^*}. We tabulate them conceptually into the D *distinct* pairs (Y_J, X_J) with frequency N_J. The probability that a randomly drawn sample value gives the pair (Y_J, X_J) is $p_J = N_J/N$. Our interest is not in the p_J but in the ratio regression coefficient

$$B = \frac{\sum_J p_J Y_J}{\sum_J p_J X_J}$$

We draw a random sample of size n (we assume with replacement) and obtain counts n_J for the distinct values. The likelihood of the sample is as before (omitting the known constant)

$$L(\mathbf{p}) = \prod_J p_J^{n_J}$$

We use the Haldane Dirichlet D(**0**) prior with $a_J = 0$ for all J, giving the Dirichlet posterior D(**n**), now defined on the d distinct values in the observed support:

$$\pi(p_1, \ldots, p_d|y) = \frac{\Gamma(n)}{\prod_j \Gamma(n_j)} \prod_j p_j^{n_j - 1}$$

We draw $M = 10,000$ random values $p_j^{(m)}$ of the p_j for the observed support, and compute the 10,000 values of $B^{(m)}$. The 95% central credible interval for B is computed from the 250th and 9750th ordered values of $B^{(m)}$.

For the hospital example, the 10,000 values of B generated from the posterior distribution of the p_j give a median of 3.200 and a central 95% credible interval of [2.917, 3.515] (the population value is 2.966). The corresponding credible interval for T_Y is [314,908, 379,465]. The true value is 320,159. The posterior cdf and density estimate for B are shown in Figures 7 and 8; those for T_Y are just rescaled.

Note that any other function of the p_J could be simulated in the same way; for example, if it were clear that the variance of Y was proportional to X^2 rather than X, while the mean was linear in X, the estimate of B^* would be

$$\hat{B}^* = \sum_i \left(\frac{y_i / x_i}{n} \right)$$

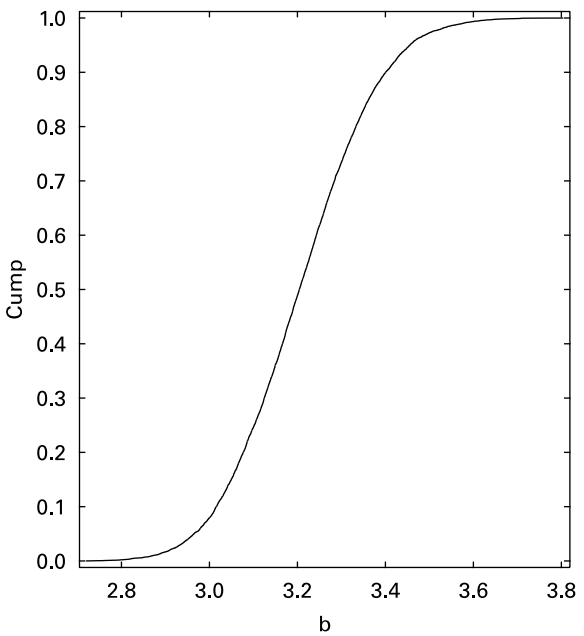

Fig. 7. Posterior cdf of ratio regression coefficient B

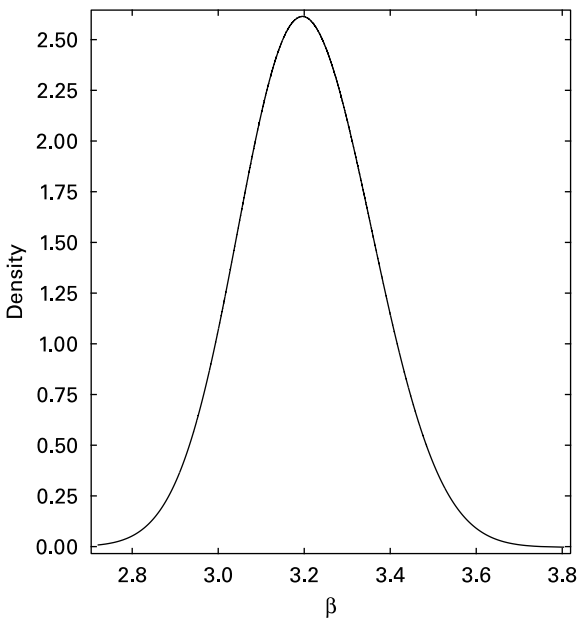

Fig. 8. Posterior density of ratio regression coefficient B

implying a population definition of

$$B^* = \sum_J \frac{p_J Y_J}{X_J}$$

whose posterior distribution could be simulated at the same time as that of B.

3.3. Simulation Study

We used the hospital population in a simulation study of the performance of the survey sampling estimate and approximate confidence interval, the confidence interval from the normal model using the information-based standard error, the bootstrap confidence interval and the credible interval using the Haldane prior.

We generated 1,000 random samples of size $n = 50$ from the hospital population, and for each sample constructed the 80%, 90%, and 95% confidence intervals for the population ratio regression coefficient by five methods:

- the ML estimate assuming a normal model with variance proportional to X (the ML estimate is identical to the ratio estimate) and with standard error from the normal model information matrix;
- the ML estimate assuming a normal model with variance proportional to X but with *robust* "sandwich" standard error from the normal model information matrix and the squared residuals;
- the sample survey ratio estimate, with approximate standard error from the sampling distribution of the Z_{I^*};

- the bootstrap central percentile interval from 10,000 bootstrap samples;
- the central credible interval from 10,000 samples from the posterior distribution of B.

These are given in Table 3.

Apart from the normal intervals with the information-based standard error, all the methods performed very similarly, with slight under-coverage at the higher confidence levels. Relative coverage is not related to interval length in this example, and the only significant effect in the probit analysis of coverage proportions is the lower coverage of the first normal method. The information-based standard error appears to be too small, probably a consequence of the variance of Y not being proportional to X.

3.4. Ancillary Information

It might appear that the Bayesian analysis could be strengthened by using the additional information in the hospital bed population. We know the *marginal* proportions of hospitals with exactly X beds. If the marginal *sample* proportions departed from these, then it would appear that we should adjust the posterior distribution of T_Y (which was just that for B scaled by T_X) by scaling-up the predictions for each X_J by the actual population proportions at these X_J. However, the scaling by T_X already achieves this (since T_X incorporates these population proportions), so the marginal proportions of hospitals at each bed number cannot provide more information. This result follows also from incomplete data theory (Little and Rubin 1987): if the unobserved Y are "missing at random" (that is, selection into the sample is not based on Y) then the full information about B is contained in the observed sample pairs (y_j,x_j), and the additional observed X_J provide no further information about B, and hence about T_Y.

3.5. Sampling without Replacement

The analysis and simulations above assume that sampling is with replacement. We may simply adapt the hypergeometric analysis in Section 2 to the regression model. As before,

Table 3. Average interval length and coverage, $n = 50$

Coefficient	Normal	Sandwich	Survey	Boot	Haldane
80%	0.2814	0.3337	0.3362	0.3300	0.3187
90%	0.3632	0.4307	0.4339	0.4232	0.4109
95%	0.4353	0.5162	0.5200	0.5036	0.4919
c	0.724	0.794	0.802	0.796	0.784
80% nlc	0.149	0.113	0.117	0.118	0.122
nrc	0.127	0.093	0.081	0.086	0.094
c	0.835	0.886	0.880	0.882	0.879
90% nlc	0.092	0.064	0.075	0.070	0.076
nrc	0.073	0.050	0.045	0.048	0.055
c	0.904	0.943	0.937	0.937	0.931
95% nlc	0.050	0.033	0.044	0.040	0.042
nrc	0.046	0.024	0.019	0.023	0.027

the population consists of N pairs Y_{I^*}, X_{I^*}. We tabulate them conceptually into the D *distinct* pairs (Y_J,X_J) with frequency N_J. The probability that a randomly drawn sample value gives the pair (Y_J,X_J) is $p_J = N_J/N$. The ratio regression coefficient is now expressed as

$$B = \frac{\sum_J N_J Y_J}{\sum_J N_J X_J}$$

We draw a random sample of size n, now without replacement, and obtain counts n_J for the distinct values. Given the sample values $n_1, \ldots, n_D$, the likelihood in the parameters $N_1, \ldots, N_D$ is the hypergeometric likelihood

$$L(N_1, \ldots, N_D) = \prod_{J=1}^{D} \binom{N_J}{n_J}$$

which is expressed in terms of only the observed sample counts n_j:

$$L(N_1, \ldots, N_d) = \prod_{j=1}^{d} \binom{N_j}{n_j}$$

As before, write $N_J^* = N_J - n_J$, with $\sum_{J=1}^{D} N_J^* = N - n = N^*$. The likelihood in the N_j^* is

$$L(N_1^*, \ldots, N_d^*) = \prod_{j=1}^{d} \binom{N_j^* + n_j}{n_j}$$

Conditional on category proportions p_j, we treat the d unobserved population counts N_J^* as drawn from a multinomial distribution

$$m(N^*; p_1, \ldots, p_d) = \frac{N^*!}{\prod_{j=1}^{d} N_j^*!} \prod_{j=1}^{d} p_j^{N_j^*}$$

in which the probabilities p_j, conditional on the observed sample sizes n_j, have the Dirichlet distribution of Section 1:

$$\pi(p_1, \ldots, p_d | n_1, \ldots, n_d) = \frac{\Gamma(n)}{\prod_{j=1}^{d} \Gamma(n_j)} \prod_{j=1}^{d} p_j^{n_j - 1}$$

Integrating out the p_j gives a *compound multinomial distribution* as the posterior distribution of the N_j^* given the n_j:

$$\begin{aligned}
&\Pr[N_1^*, \ldots, N_d^* | n_1, \ldots, n_d] \\
&= \int \cdots \int \Pr[N_1^*, \ldots, N_d^* | p_1, \ldots, p_d] \cdot \Pr[p_1, \ldots, p_d | n_1, \ldots, n_d] \mathrm{d}p_1 \ldots \mathrm{d}p_d \\
&= \int \cdots \int \frac{N^*!}{\prod_{j=1}^d N_j^*!} \prod_{j=1}^d p_j^{N_j^*} \cdot \frac{\Gamma(n)}{\prod_{j=1}^d \Gamma(n_j)} \prod_{j=1}^d p_j^{n_j - 1} \mathrm{d}p_1 \ldots \mathrm{d}p_d \\
&= c \cdot \frac{N^*!}{\prod_{j=1}^d N_j^*!} \prod_{j=1}^d \frac{\Gamma(N_j^* + n_j)}{\Gamma(N^* + n)}
\end{aligned}$$

We use the simple indirect sampling formulation for simulation of the ratio regression coefficient:

- generate M values $p_j^{[m]}$ of the p_j as in Section 1;
- from these, generate M values $N_j^{*[m]}$ from the multinomial distributions $m(N^*; p_1^{[m]}, \ldots, p_d^{[m]})$
- calculate the M values $B^{[m]} = \sum_{j=1}^d (N_j^{*[m]} + n_j) Y_j / \sum_{j=1}^d (N_j^{*[m]} + n_j) X_j$

3.6. Simulation Study

We replicate part of the simulation study in Section 3.3, but with sample size 40 from the hospitals population, to compare the posterior distributions of the ratio regression coefficient for sampling with and without replacement. We restrict the study to just the two posteriors based on the same Haldane prior-based posterior for the p_j. Results are given in Table 4.

The hypergeometric intervals are shorter, but have lower coverage, as in the single-sample case. The probit analysis of actual coverage with length, method and nominal

Table 4. Average credible interval length and coverage, $n = 40$

Coefficient	With rep	Without rep
80%	0.3551	0.3362
90%	0.4578	0.4337
95%	0.5481	0.5192
c	0.769	0.743
80% lnc	0.125	0.136
rnc	0.106	0.121
c	0.867	0.850
90% lnc	0.071	0.081
rnc	0.062	0.069
c	0.929	0.911
95% lnc	0.036	0.046
rnc	0.035	0.043

coverage as explanatory variables shows that the simple interval length model is sufficient to describe the results: the deviance of the single interval length model is 0.104 on 4 df, so there is no improvement in coverage from the hypergeometric likelihood. The theoretical gain in precision is again not visible with a sampling fraction of 0.10.

4. More General Regression Models

The approach in this section can be readily extended to general regression models. For complex models we adopt the "working model" language of Valliant, Dorfman, and Royall (2000, p. 50), in which the "working" probability model leads to an optimal estimator under the model, which is then used without the working model being assumed to hold.

Suppose that a working model has $\mathrm{E}[Y|\mathbf{x}] = \mathbf{B}'\mathbf{x}$, $\mathrm{Var}[Y] = \sigma^2$, leading to the usual least squares estimate $(X'X)^{-1}X'\mathbf{y}$ of $\mathbf{B}$. This can be immediately treated in the same Bayesian way, expressed by definition as $\mathbf{B} = (X'PX)^{-1}X'P\mathbf{y}$, where P is a diagonal weight matrix of the population proportions at each support point in the $(Y,\mathbf{x})$ space. We simulate M values $p_j^{[m]}$ from the posterior Dirichlet distribution of the p_j, giving the M values $\mathbf{B}^{[m]} = (X'P^{[m]}X)^{-1}X'P^{[m]}\mathbf{y}$. This will in general require M matrix inversions of the weighted SSP matrix. Nonconstant variance models can be easily incorporated.

The ability to use standard software for the Dirichlet analysis with an additional weight vector greatly extends the generality of the Bayesian bootstrap analysis. We illustrate this with the analysis of a complex example in Section 7.

5. The Multinomial Model for Multiple Populations

Consider a population of size N which is made up of S subpopulations indexed by $s = 1, \ldots, S$, with N_s members and proportion $p_s = N_s/N$ in subpopulation s. A response variable of interest Y takes values in the full population. As for the case of a single population, we conceptually tabulate the full population by the *distinct* values $Y_1 < \cdots < Y_J < \cdots < Y_D$. In subpopulation s the proportions of the subpopulation at the values Y_J are denoted by $p_{sJ} = N_{sJ}/N_s$ where N_{sJ} is the number of members at Y_J in subpopulation s. We do not assume that the proportions p_{sJ} are related across subpopulations: the set of multinomial distributions is completely general.

The subpopulation means and variances of Y are

$$\mu_s = \sum_J p_{sJ}Y_J, \qquad \sigma_s^2 = \sum_J p_{sJ}(Y_J - \mu_s)^2$$

We draw a random sample of size n_s from the sth subpopulation, with total sample size $n = \sum_s n_s$, and obtain n_{sJ} sample values at Y_J in the sth subpopulation. The *sample fraction* π_s is the proportion n_s/N_s drawn from the sth subpopulation.

The subpopulation sample means and variances are

$$\bar{y}_s = \sum_J \frac{n_{sJ}Y_J}{n_s}, \qquad s_s^2 = \sum_J \frac{n_{sJ}(Y_J - \bar{y}_s)^2}{n_s - 1}$$

We consider first the case when the sample fractions are small, so that sampling with and without replacement are equivalent. Large sample fractions require the hypergeometric likelihood rather than the multinomial; the approach of Section 2 can be adapted to the more general case here.

The overall population mean is

$$\mu = \sum_{s=1}^{S} \frac{N_s \mu_s}{N} = \sum_{s=1}^{S} p_s \mu_s$$

and the overall sample mean is

$$\bar{y} = \sum_{s=1}^{S} \frac{n_s \bar{y}_s}{n}$$

These are conveniently summarized in Table 5 below.

We make use of this structure for two different types of sampling.

6. Complex Sample Designs

6.1. Stratified Sampling

Stratified sampling is designed to reduce variability in estimation due to known population heterogeneity – the population is made up of homogenous subpopulations with substantial differences in mean and/or variance among them. If some of the subpopulations are small, a simple random sample may miss them completely, or give only small subsamples from them. Stratified sampling sometimes *oversamples* small strata to give comparable sample sizes from all strata – the assessment of strata differences is most precise, for a fixed total sample size, when the strata sample sizes are proportional to the strata variances (and so are equal if the strata variances are equal).

We now identify the s label with *stratum* (usually denoted by h). For a single response variable Y, we wish to estimate the stratum means μ_s and the overall population mean, allowing for the different sampling fractions $p_s = n_s/N_s$ in the different strata. For Bayesian inference about the individual μ_s, we proceed as for the single population mean in Section 1. We draw M values $p_{sJ}^{[m]}$ from the posterior Dirichlet distribution of the p_{sJ} on the observed support in stratum s using the Haldane prior, and map these into M values

$$\mu_s^{[m]} = \sum p_{sJ}^{[m]} Y_J$$

Table 5. Subpopulation structure

	Subpopulation	Proportion	Sample fraction	Mean	Variance
Population	s	$p_s = N_s/N$		μ_s	σ_s^2
Sample	s	n_s/n	$\pi_s = n_s/N_s$	$\bar{y}_s$	s_s^2

of the posterior distribution of μ_s. Then for posterior inference about $\mu = \sum_s p_s \mu_s$, we simply combine the M simulated values of μ_s:

$$\mu^{[m]} = \sum_s p_s \mu_s^{[m]}$$

to give M values from the posterior distribution of μ.

We postpone an example to Section 7.

6.2. *Cluster Sampling*

Cluster sampling has a similar formal structure to stratification, but the population parameters of interest are different. Cluster sampling is a form of two-stage sampling, in which the population is divided into clusters which are defined by geographic contiguity or other similarities, which make units sampled within the same cluster more homogeneous than those sampled from different clusters. Clustering frequently reduces sampling costs compared with simple random sampling.

The two-stage design selects clusters at random according to a sample design, and samples units within clusters according to a second sample design (sometimes a full sample of all units in the clusters).

The analysis in cluster sampling allows for the greater homogeneity *within* clusters than that *among* clusters, and this is naturally represented through *variances*.

We now change notation, using the subscript c to represent cluster identification; the design has C clusters. We adapt Table 5 to represent clustering in Table 6:

The overall population mean is $\mu = \sum_c p_c \mu_c$, and the overall population variance is

$$\begin{aligned}
\sigma^2 &= \sum_c \sum_J \frac{N_{cJ}(Y_J - \mu)^2}{N} \\
&= \sum_c \sum_J \frac{p_{cJ} N_c (Y_J - \mu_c + \mu_c - \mu)^2}{N} \\
&= \sum_c \frac{N_c \left[\sigma_c^2 + \sum_j p_{cJ}(\mu_c - \mu)^2 \right]}{N} \\
&= \sum_c p_c [\sigma_c^2 + (\mu_c - \mu)^2] = \sigma_W^2 + \sigma_A^2
\end{aligned}$$

Table 6. Subpopulation structure

	Cluster	Proportion	Sample fraction	Mean	Variance
Population	c	$p_c = N_c/N$		μ_c	σ_c^2
Cluster	c	n_c/n	$\pi_c = n_c/N_c$	$\bar{y}_c$	s_c^2

where N_{cJ} is the number of cluster c values of Y equal to Y_J, $p_{cJ} = N_{cJ}/N_c$,

$$\sigma_W^2 = \sum_c p_c \sigma_c^2$$

is the (average) *pooled within-cluster variance* and

$$\sigma_A^2 = \sum_c p_c (\mu_c - \mu)^2$$

is the *among-cluster variance*.

The posterior distributions of both these *variance components* can be simulated directly from their definitions in terms of the cluster means, variances and proportions. Denote the sample data from cluster c by y_{cj}, $j = 1, \ldots, n_c$. We assume for simplicity of notation that all the observations in a cluster are distinct, though tabulation for the distinct values is in general needed as for a single population. The cluster population proportion at Y_{cj} is p_{cj}, and given the sample data, we simulate M values $p_{cj}^{[m]}$ from the posterior Dirichlet distribution of the p_{cj} as in Section 1. From these we compute the M values $\mu_c^{[m]}$, $\sigma_c^{2[m]}$, $\mu_c^{[m]}$, $\sigma_A^{2[m]}$ and $\sigma_W^{2[m]}$ from their definitions above.

An important point here is that the cluster sizes N_c in the population do not need to be known for this analysis, nor the total population size N: only the proportions p_c are used, and these are based on the sample proportions at each observed value.

We illustrate with a small example from Box and Tiao (1973, p. 246), a designed experiment in which five samples were randomly chosen from six batches of raw material, and a single laboratory determination made (of the yield of dyestuff in grams of standard color) on each sample. This example is artificial for population survey sampling, but our aim is to show how variance components are estimated.

Box and Tiao give the details of the Bayesian treatment of the normal variance component model, and the joint posterior distribution of the "among-batch" and "within-batch" variance components. In Box and Tiao's Bayesian analysis the within-batch variances are assumed to be the same across batches; we relax this assumption.

The data are given in Table 7, with the batch mean and (unbiased) variance. For each batch c, we generate $M = 10,000$ random values $p_{jc}^{[m]}$ of the p_{jc} for the observed values y_{jc} in that batch, and substitute them in the various means and variances for each c, and the variance components. The posterior distributions for the batch means and variances are shown in Figures 9 and 10.

Table 7. Dyestuff data

Batch	1	2	3	4	5	6
	1,545	1,540	1,595	1,445	1,595	1,520
	1,440	1,555	1,550	1,440	1,630	1,455
	1,440	1,490	1,605	1,595	1,515	1,450
	1,520	1,560	1,510	1,465	1,635	1,480
	1,580	1,495	1,560	1,545	1,625	1,445
Mean	1,505.0	1,528.0	1,564.0	1,498.0	1,600.0	1,470.0
Variance	3,975.0	1,107.5	1,442.5	4,720.0	2,500.0	962.5

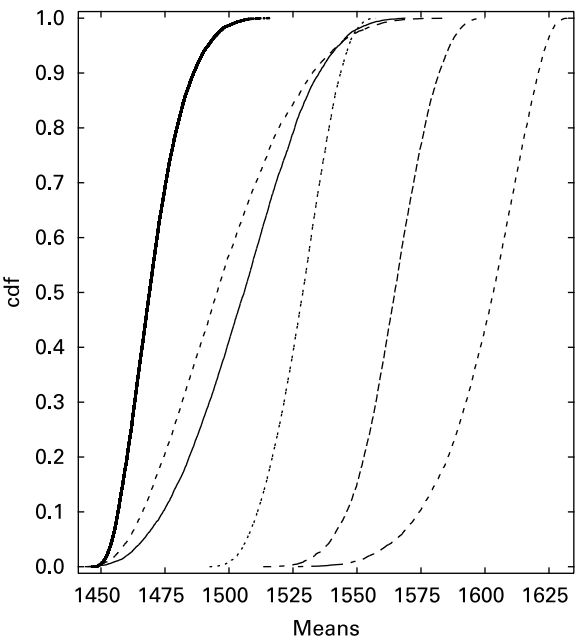

Fig. 9. Posteriors, batch means

They are very diffuse, a consequence of the small sample size in each batch. (The batch can be identified in the figures by matching the sample mean to the posterior median.) The sample means differ considerably, and there is some sample variance heterogeneity – the largest variance ratio between batches is 4.90. The posterior

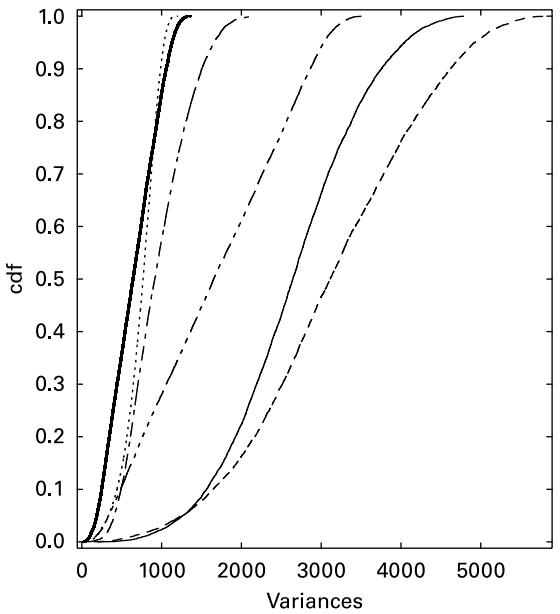

Fig. 10. Posteriors, batch variances

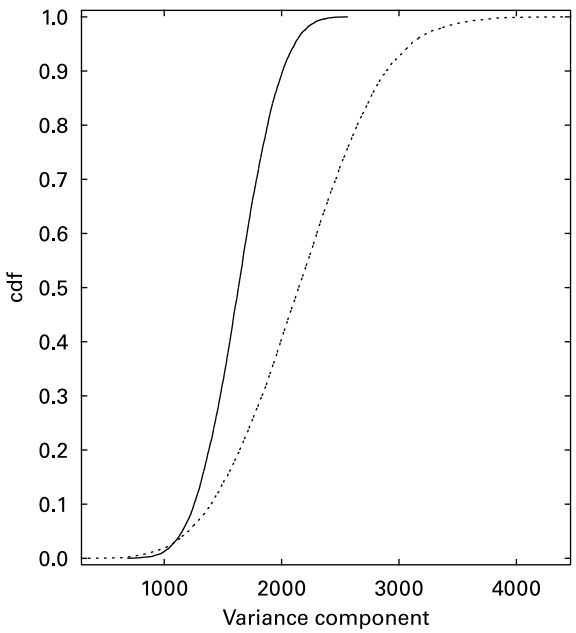

Fig. 11. Posteriors, batch variance components

distributions of the among-batch and pooled within-batch variance components are shown in Figure 11.

6.3. Shrinkage Estimation

The variance components are widely used in *small-area estimation*, in which the estimation of an area mean is improved by "borrowing strength" from the other area means through their variation as measured by the among-area variance component. In fully (normal) model-based inference a *shrinkage estimator* of an area mean may be superior to the simple area sample mean, if the area sample size is small. In the normal two-level model

$$y_{jc}|\mu_c \sim N(\mu_c, \sigma_c^2)$$
$$\mu_c \sim N(\mu, \sigma_A^2)$$

it follows immediately that

$$\bar{y}_c|\mu_c \sim N(\mu_c, \sigma_c^2/n_c)$$
$$\mu_c|\bar{y}_c \sim N(\mu + w_c(\bar{y}_c - \mu),\ \sigma_A^2(1 - w_c))$$

where

$$w_c = \frac{n_c\theta_c}{1 + n_c\theta_c}, \quad \theta_c = \frac{\sigma_A^2}{\sigma_c^2}$$

The posterior mean of μ_c can be written

$$\tilde{\mu}_c = w_c \bar{y}_c + (1 - w_c)\mu$$

This is widely used as a *shrinkage estimator* of the area mean. The difficulty with this estimator in frequentist theory (apart from the assumption of normality) is how to specify correctly its variability; in the fully normal model-based analysis *empirical Bayes* estimators are widely used, with ML estimates replacing the unknown variance component parameters and overall mean, but the variability of the resulting shrinkage estimator is very difficult to establish; further, the posterior *variance* of the μ_c is widely ignored.

The same Dirichlet posterior analysis provides the inference about the μ_c, correctly adjusted for parameter uncertainty. We first substitute the simulated values above into the variance component ratio and the means, using $\mu^{[m]} = \sum p_c \mu_c^{[m]}$, giving

$$\theta_c^{[m]} = \frac{\sigma_A^{2[m]}}{\sigma_c^{2[m]}}$$

$$w_c^{[m]} = \frac{n_c \theta^{[m]}}{1 + n_c \theta^{[m]}}$$

$$\tilde{\mu}_c^{[m]} = w_c^{[m]} \mu_c^{[m]} + (1 - w_c^{[m]})\mu^{[m]}$$

$$\widetilde{\mathrm{Var}}_c^{[m]} = \sigma_A^{2[m]}(1 - w_c^{[m]})$$

where $\widetilde{\mathrm{Var}}_c$ is the posterior variance of μ_c. Then for each m we draw the random value $\mu_c^{[m]}$ from $N(\widetilde{\mu_c^{[m]}}, \widetilde{\mathrm{Var}_c^{[m]}})$. These values allow for *all* the uncertainty in the parameters, *and* for the variance of the posterior distribution of μ_c.

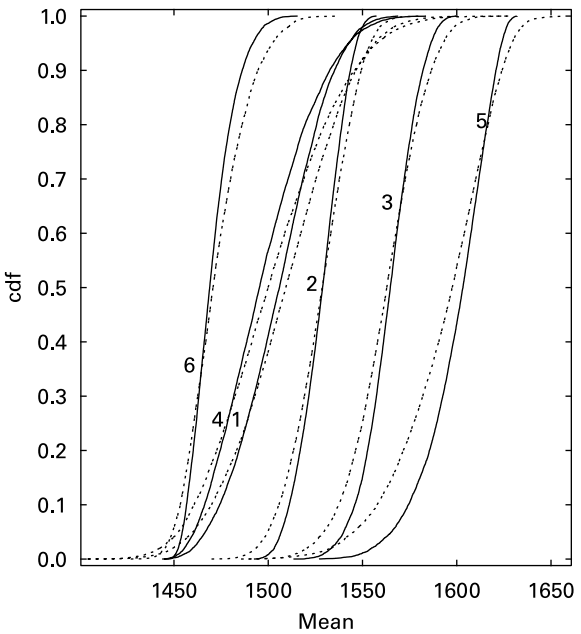

Fig. 12. Posterior and fixed effect means, all batches

This is one of the great strengths of the Bayesian analysis: the simulation variability in the population proportions p_j is *propagated* throughout the subsequent functions of these parameters. We illustrate with the Box and Tiao example.

Figure 12 shows the posterior distributions (as cdfs), for each batch, of the "fixed effect" mean μ_c of the batch (without using the batch random effect distribution – solid curves) and of the batch random effect (dashed curves), derived as described above from $M = 10,000$ samples.

Surprisingly, the random effect posterior distributions are more diffuse than those for the "fixed effects" – it appears that incorporating the additional information has *decreased*, rather than *increased*, the precision of inference!

There are two reasons for this result. First, we have not assumed that the batch variances are homogeneous. In a frequentist analysis, this assumption is made routinely: without it, the batch variance for each batch has four degrees of freedom instead of the 24 of the pooled within-batch variance. As a result, *all* of the batch means, variances and variance component ratios are based on very small samples and are very imprecise. The additional information provided by the among-batch variance component (which is itself based on only five degrees of freedom) does not overcome this imprecision.

Second, we are not assuming a normal (or any other) parametric distribution for the dyestuff yield, and so the multinomial distributional model in each batch is based on only five values, which are unrelated across batches.

The homogeneity assumption is important for inference about the batch means: if we use the *average batch variance* instead of the individual batch variances, the random effect batch posteriors (not shown) are *more precise* than the fixed effect posteriors, and also show shrinkage towards the overall population mean. This is the usual conclusion from empirical Bayes analyses, but its validity may depend strongly on the homogeneity of variance assumption.

7. A Complex Example

We conclude this article with an analysis of the Labor Force Survey data from Valliant, Dorfman, and Royall (2000, Appendix B.5). The sample of 478 individuals is stratified in three strata and clustered in 115 clusters within strata, with an average of four individuals per cluster. We illustrate the general approach with a main-effect regression of weekly wage on sex, age and hours worked, allowing for the stratification and clustering.

Since the (stratum, cluster) cells hold only four cases each on average, we assume a constant variance of wage across these cells. We comment on this assumption below.

We index the data by (i,c,s) for person i in cluster c and stratum s, and write y_{ics} for the weekly wage of person i in cluster c and stratum s, a_c and b_s for the random cluster and fixed stratum effects, and $\mathbf{x}_i$ for the explanatory variables on person i. We adopt the working model

$$E[y_{ics}|a_c] = \beta'\mathbf{x}_i + a_c + b_s$$

$$\text{Var}[y_{ics}|a_c] = \sigma^2$$

$$E[a_c] = 0$$

$$\text{Var}[a_c] = \sigma_A^2$$

$$\text{Cov}[y_{ics}, y_{i'c's'}] = \delta(c,c')\sigma_A^2$$

where $\delta(c, c') = 1$ if $c = c'$ and zero otherwise. These are the usual assumptions of the two-level cluster random effect model with fixed stratum effects. If in addition the cluster and wage variables were assumed to be normally distributed, the optimal estimates of β and the stratum effects would be the ML estimates from the two-level normal variance component model. These can be expressed as generalized least squares estimates: writing $\mathbf{b}$ for the vector of stratum effects and $Z = [X, B]$ for the design matrix of explanatory variables and stratum effects, the MLEs are the solutions of

$$Z'VZ[\boldsymbol{\beta}, \mathbf{b}] = Z'V\mathbf{y}$$

where V is the block-diagonal covariance matrix of the observations. This solution can be obtained from any standard two-level maximum likelihood model program. We adopt the ML estimators of fixed effects and variance components as defining the population parameters for the Bayesian bootstrap analysis, but without the assumption of normality.

We express the posterior distributions of the population parameters – regression coefficients, stratum effects and variance components – in terms of the posterior Dirichlet distributions of the probabilities at each sample point in each cluster within stratum, based on the number (2–5) of persons in each cluster. The full simulation procedure is surprisingly simple:

- from the observations within each cluster, construct the Dirichlet posterior, with the Haldane prior, of the probabilities p_{jc} on the observed support y_{jc} within that cluster;
- draw M values $p_{jc}^{[m]}$ from the posteriors of the p_{jc};
- using the $p_{jc}^{[m]}$ as explicit *weights* for each observation y_{jc} in cluster c, carry out M weighted ML fits of the y_{cj} to the explanatory variables, to obtain parameter estimates $\beta^{[m]}$, $\mathbf{b}^{[m]}$, $\sigma^{2[m]}$ and $\sigma_A^{2[m]}$.

These M values provide the required posterior distributions of the parameters.

For the conventional two-level normal model analysis we assume the wage variance is constant. For the main effect model of age, sex, hours worked and stratum, the ML estimates and standard errors (omitting those for strata) are given below.

Maximum likelihood estimates and standard errors, wage example

	Intercept	Age	Sex	Hours	Sigma	Sigma_A
MLE	−3.20	1.988	−107.8	7.061	136.7	76.76
SE	7.95	0.125	3.46	0.154		

The Bayesian bootstrap analysis with $M = 10,000$ gave posterior distributions for all the parameters which were indistinguishable from normal, apart from slightly heavier tails for the among-cluster variance component. The posterior means and standard deviations for the parameters are shown below.

Posterior means and standard deviations, wage example

	Intercept	Age	Sex	Hours	Sigma	Sigma_A
Mean	127.3	1.792	−115.2	7.514	137.2	77.01
SD	32.1	0.412	10.6	0.497	4.03	8.28

The striking difference in intercepts is of no real consequence since wages were not centered for the analysis. The variance component estimates are very close; the other parameter estimates are less close but similar, as might be expected from a weighted analysis, but their precisions are very different – the posterior SDs are 3–4 times as large as the SEs. The reason for this is very clear – a graph of cluster sample variance against mean shows that the wage variance increases with mean, so the constant variance model gives a compromise variance which misrepresents the nature of the variability. The Bayes analysis gives a variance estimate which is almost the same, but the effect of the weighting is similar to that of "sandwiching" the variance estimates in a frequentist analysis which allows for variance heterogeneity: the model uncertainty, in *both* distributional *and* variance terms, is allowed for by the Dirichlet distributions in each cluster.

The Bayesian bootstrap analysis required about 4 hours of time for the 10,000 draws, on a laptop with 1.6 GHz processor and 1.24 GB RAM, running the GLIM4 Gaussian quadrature macro (Aitkin 1999) for a two-level normal GLMM weighted by the Dirichlets generated by sequential reads through the clusters. This time could be substantially reduced with more efficient model fitting and simulation.

We emphasize that *any* package which can both simulate gamma random variables and fit two-level models could be used for this analysis; we do not give a specific package code.

8. Discussion

The Bayesian bootstrap approach to finite population analysis is quite general. It accepts the survey sampling axiom of not assuming a full parametric model for the population form of the response variable, but it nevertheless provides full information about the defined population parameters through the multinomial likelihood and the noninformative Haldane Dirichlet prior. The prior restricts analysis to the observed support, in the same way as maximum empirical likelihood and the frequentist bootstrap. This approach, like parametric model-based approaches, does not use the sampling distributions of the sample selection indicators, but unlike parametric model approaches, does not require the examination or validation of the parametric model by residual examination.

It is surprising that the Bayesian bootstrap analysis is equivalent to a series of M weighted maximum likelihood analyses with randomly varying weights; the variation among the M resulting parameter realizations provides the full posterior distribution of these parameters, in a (model) distribution-free way, since it depends only on the primitive multinomial model, which is not a model assumption which can be contradicted by the data.

The ability to use standard software packages (provided these allow for weights) is a particularly useful feature of the Bayesian bootstrap analysis. Running 10,000 regression analyses may appear computationally intensive, especially for GLMMs, but since the weights are varying only randomly, the ML estimates can be used as starting values for each analysis, and so convergence of each of the M model-fitting steps can be faster than for the ML analysis itself.

This analysis has some similarities to the empirical likelihood analysis of Owen (2001), but the latter depends on asymptotic frequentist likelihood theory for the calibration of empirical likelihood confidence intervals and regions. As noted in the examples in this article, the credible intervals tend to under-cover in small samples (as do likelihood-based confidence intervals in some models), but this calibration is based only on the equivalence in larger samples of credible and likelihood-based confidence intervals; the credible intervals always have their usual Bayesian interpretation.

The disadvantages of the Bayesian bootstrap approach are shared with survey sampling analysis, empirical likelihood and bootstrapping: without an explicit response probability model there is no optimal choice for population or regression parameters. The decision to use (say) the ratio estimator is not based on data properties, except in so far as the form of the variance function can be assessed from data plotting or residual examination. Different choices of the power of X parameter in the variance function lead to different estimators, but there is no obvious way of choosing which is more appropriate, since the multinomial likelihood is a function only of the population proportions at each support point, and not explicitly of the variance parameter.

Different variance parameters provide different population regression coefficient definitions, for all of which the multinomial likelihood and Dirichlet posterior provide credible intervals for the regression coefficients, but we are unable to compare these coefficients (and the implicit variance functions which justify them) through the model likelihoods, since they depend on the same unconstrained multinomial parameters.

A more refined examination of the variance form (for example, the choice of the most appropriate value of the power parameter of X) requires an explicit parametric model for the response, or else a constrained multinomial likelihood, to provide different likelihoods for different models which can then be compared. A similar problem occurs with the variance homogeneity assumption: this is a *model constraint* on the multinomial probabilities in each batch which is difficult to implement in a Bayesian framework.

Thus the Bayesian bootstrap approach is not completely general, but within its limitations optimal Bayes procedures are available and fruitful, and readily computable with general software. Further investigation of model comparison methods and constrained multinomial probabilities is required for a fully general analysis.

9. References

Aitkin, M. (1999). A General Maximum Likelihood Analysis of Variance Components in Generalized Linear Models. Biometrics, 55, 117–128.

Banks, D. (1988). Histospline Smoothing the Bayesian Bootstrap. Biometrika, 75, 673–684.

Binder, D.A. (1982). Non-parametric Bayesian Models for Samples from Finite Populations. Journal of the Royal Statistical Society, Series B, 44, 388–393.

Box, G.E.P. and Tiao, G.C. (1973). Bayesian Inference in Statistical Analysis. Reading, MA: Addison-Wesley.

Ericson, W.A. (1969). Subjective Bayesian Models in Sampling Finite Populations (with Discussion). Journal of the Royal Statistical Society, Series B, 31, 195–233.

Ferguson, T.S. (1973). A Bayesian Analysis of Some Nonparametric Problems. Annals of Statistics, 1, 209–230.

Ghosh, M. and Meeden, G. (1997). Bayesian Methods for Finite Population Sampling. London: Chapman and Hall.

Gutierrez-Pena, E. and Walker, S.G. (2005). Statistical Decision Problems and Bayesian Nonparametric Methods. International Statistical Review, 73, 309–330.

Hartley, H.O. and Rao, J.N.S. (1968). A New Estimation Theory for Sample Surveys. Biometrika, 55, 547–557.

Herson, J. (1976). An Investigation of Relative Efficiency of Least Squares Prediction to Conventional Probability Sampling Plans. Journal of the American Statistical Association, 71, 700–703.

Hoadley, B. (1969). The Compound Multinomial Distribution and Bayesian Analysis of Categorical Data from Finite Populations. Journal of the American Statistical Association, 64, 216–229.

Hodges, J.L., Krech, D., and Crutchfield, R.S. (1975). StatLab: An Empirical Introduction to Statistics. New York: McGraw-Hill.

Little, R.J. and Rubin, D.B. (1987). Statistical Analysis with Missing Data. New York: Wiley.

Lohr, S.L. (1999). Sampling: Design and Analysis. Pacific Grove: Duxbury.

Owen, A.B. (1988). Empirical Likelihood Ratio Confidence Intervals for a Single Functional. Biometrika, 75, 237–249.

Owen, A.B. (2001). Empirical Likelihood. CRC, Boca Raton: Chapman and Hall.

Royall, R.M. and Cumberland, W.G. (1981). An Empirical Study of the Ratio Estimator and Estimators of Its Variance (with Discussion). Journal of the American Statistical Association, 76, 66–88.

Rubin, D.B. (1981). The Bayesian Bootstrap. Annals of Statistics, 9, 130–134.

Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

Valliant, R., Dorfman, A.H., and Royall, R.M. (2000). Finite Population Sampling and Inference: A Prediction Approach. New York: Wiley.

Walker, S.G. and Gutierrez-Pena, E. (2007). Bayesian Parametric Inference in a Nonparametric Framework. Test, 16, 188–197.

Received January 2006
Revised July 2007