

A noninformative Bayesian approach to finite population  
sampling using auxiliary variable

Glen Meeden  
University of Minnesota

Joint work with  
Radu Lazar and David Nelson

<http://www.stat.umn.edu/~glen/talks>

## Basu and finite population sampling

Statistical Information and Likelihood  
A Collection of Critical Essays  
by Dr. D. Basu  
J. K. Ghosh, Editor (1988)

Gives an elegant argument for the Bayesian approach to finite population sampling.

The one area in statistics where prior information is often used but traditionally in a non-Bayesian manner. Why?

## Auxiliary variables and survey sampling

Auxiliary variables often contain information about the population.

In standard theory one needs to assume a model that relates the characteristic of interest to the auxiliary variables.

For example, in stratified populations should you use a single regression model for the whole population or a different model in each stratum?

Our approach will be objective and Bayesian but we will not need to assume a model.

## Some Notation

$\mathcal{U}$  is a finite population with  $N$  units.

$y_i$  is the characteristic of interest for unit  $i$ .

$x_i$  is an auxiliary variable for unit  $i$ .

We observe a sample  $s \subset \{1, 2, \dots, N\}$  using some sampling design  $\Delta$ , usually SRSWOR.

$y(s) = \{y_i : i \in s\}$  are the “seen”

$y(s') = \{y_j : j \notin s\}$  are the “unseen”

How to relate the “unseen” to the “seen”?

## The Bayesian Way

Ericson (1969) JRSSB

Need joint prior distribution for the population

$$P(y_1, y_2, \dots, y_N)$$

After observing sample must find

$$P(y_j : j \notin s \mid y_i : i \in s)$$

the conditional distribution of the unseen given the seen.

Simulate from the posterior to get completed copies of the entire population. For each of the simulated copies compute the parameter of interest . Use these computed values to find point and interval estimates of the parameter.

## The Polya Posterior

Imagine a MC is using SRSWOR to select units one at a time from a population. After selecting  $n$  of the  $N$  units she asks you to use the **seen** to estimate the population mean.

How should we relate the **seen** to the **unseen**?

Select a unit at random from the **unseen**. Select a second unit at random from the **seen** and assign its value to the selected **unseen** unit and place both with the **seen**.

Repeat this process using the  $N - n - 1$  **unseen** and the  $n + 1$  **“seen”**.

Repeat until all the **unseen** have been assigned a simulated value.

## Under Polya Posterior

$$E(\mu(\mathbf{y}) \mid y_i \ i \in s) = \bar{y}_s$$

and

$$Var(\mu(\mathbf{y}) \mid y_i \ i \in s) = \left(1 - \frac{n}{N}\right) \frac{v_s}{n} \frac{n-1}{n+1}$$

where  $\bar{y}_s$  and  $v_s$  are the sample mean and sample variance and  $\mu(\mathbf{y})$  is the population mean.

Ghosh and Meeden (1997), Lo (1988) Annals and Rubin (1981) Annals , Nelson and Meeden (2006) JSPI and Magnussen and Kohl (2002) Forest Science

## Relation to bootstrap

Assume SRSWOR and  $N = kn$  for some integer  $k$ . Given a sample  $s$  a good guess for the population is just  $k$  copies of  $y(s)$ .

The bootstrap assumes the guess is the “truth” and takes many repeated samples of size  $n$  from the guess. For each resample it calculates the estimate and uses these values to get an estimate of variance. Gross (1980) and Booth, Butler and Hall (1994)

The Polya posterior uses the sample to construct many possible different guesses for the population. For each simulate full copy it calculates the parameter of interest and uses these values to get an estimate and an estimate of its variance.

## Auxiliary Variable

$x_i$  is value of an auxiliary variable for unit  $i$ .

Assume  $\mu(\mathbf{x})$ , the population mean of  $\mathbf{x}$  is known and we observe  $y_i$  and  $x_i$  for all the units in the sample.

How should the Polya posterior incorporate knowing  $\mu(\mathbf{x})$ ?

Just do restricted Polya sampling using the

$$\text{seen} = \{(y_i, x_i) : i \in s\}$$

in such a way that every simulated copy of the population satisfies the constraint on  $\mu(\mathbf{x})$ .

## Can constraints be satisfied?

Suppose  $N = 10$ ,  $n = 2$  and we know

$$\mu(\mathbf{x}) = 1.47$$

If

$$\{x_i : i \in s\} = \{0, 1.2\}$$

then there are no simulated copies of the population which satisfy the constraint.

If

$$\{x_i : i \in s\} = \{0, 2\}$$

then again there are no simulated copies of the population which satisfy the constraint.

## Approximating the Polya posterior

Under the Polya posterior the only values appearing in a simulated copy of the entire population are those that appeared in the sample.

For a  $j \in s$  let  $\lambda_j$  be the proportion of units in a completed simulated copy which take on the value  $y_j$ . If  $n/N$  is small then under the Polya posterior  $\lambda$ , the vector of  $\lambda_j$ 's has approximately the uniform distribution on the  $n-1$  dimensional simplex  $\sum_{j \in s} \lambda_j = 1$ . Easy to simulate complete copies.

Easy to add constraints to this space. If  $\mu(\mathbf{x})$  is known then we are restricted to

$$\sum_{j \in s} x_j \lambda_j = \mu(\mathbf{x})$$

Harder to simulate in restricted problem.

## The Constrained Polya Posterior (CPP)

For situations where the regression estimator would be used the point and interval estimator of the CPP behave almost the same.

Chen and Qin (1993) *Biometrika* considered a point estimator of the median of  $y$  assuming  $\mu(\mathbf{x})$  is known. In a variety of populations the CPP did on the average 10% better.

The CPP can incorporate constraints involving the median of  $\mathbf{x}$ .

The CPP can incorporate linear inequality constraints, for example  $\mu(\mathbf{x})$  is known to lie in an interval.

## An Example

A population of 2500 veterans. They are classified by gender (F and M) and health status (Good, Average and Poor).

The characteristic of interest is PCS, a measure of overall quality of life.

The auxiliary variable is age and its population mean is known.

$$\text{cor}(\text{PCS}, \text{age}) = -0.22$$

Strata and Sample Sizes

	Good	Average	Poor
F	353(20)	155(10)	117(10)
M	890(30)	493(20)	492(10)

## The Results

Results for estimating PCS using 200 random samples.

Strata is the usual stratified estimator which assume the strata sizes are known.

The CPP estimator assumes the row and column totals of the strata sizes are know along with the average age of the individuals in the population.

Meth	A est	A aber	A lwbd	A len	F cov
Mean	37.23	1.04	34.91	4.65	0.938
Strata	36.65	0.93	34.32	4.65	0.948
CPP	36.64	0.93	34.34	4.61	0.958

## A toy stratified population

We constructed a population with 2 strata and 2 auxiliary variables.

Strat.	Size	The $x_{1i}$ 's	The $x_{2i}$ 's	The errors
1	600	gamma(20,1)	gamma(18,1)	normal(0,7 <sup>2</sup> )
2	400	gamma(7,1)	gamma(12,1)	normal(0,8 <sup>2</sup> )

$$\text{stratum 1: } y_i = 37 + 1.2x_{1i} + \sqrt{x_{2i}} + e_i$$

$$\text{stratum 2: } y_i = 40 + 3x_{1i} + 0.5\sqrt{x_{2i}} + e_i$$

Assume pop median of  $x_1$  and pop mean of  $x_2$  are known.

Also we have constraints from the known strata sizes.

$$\text{cor}(y, x_1) = 0.196 \text{ and } \text{cor}(y, x_2) = 0.526$$

## Simulation results

We took 500 samples of size 60 from the population.

The true population total = 76297.1

sampling weights	Est method	Ave. value	Ave. abs err	Ave. len	Freq. of cov
srswor	postStrat	76263.9	1184.7	5664.9	0.944
	ConstPp	76296.3	846.8	4037.9	0.932
gamma(5)+5	postStrat	76387.8	1103.3	5554.4	0.964
	ConstPp	76412.0	822.6	4058.0	0.956
$X_2$	postStrat	76473.3	1125.6	5529.4	0.942
	ConstPp	76478.7	835.4	4087.3	0.94
$X_1$	postStrat	78224.3	2079.4	5713.3	0.726
	ConstPp	76454.3	1001.0	4162.5	0.900

## Simulation results

If you “know” the model you can do better.

With just info about aux. means one can use empirical likelihood methods of Chen and Sitter (1999) or Zhong and Rao (2000).

Easy to estimate other population parameters.

The true population median = 71.09

	Ave. value	Ave. abs err	Ave. len	Freq. of cov
srswor	70.65	1.39	6.55	0.93
gamma(5) + 5	70.66	1.29	6.62	0.928
$X_2$	70.61	1.41	7.07	0.924
$X_1$	70.80	1.31	6.36	0.898

## Minnesota Population Center

The center is a leading developer and disseminator of demographic data.

For example, it creates decade by decade micro copies of the USA population so that researchers can study time related questions.

Since survey questions and definitions change over time presenting data in a consistent fashion can be a problem.

<http://www.pop.umn.edu/>

## A Simple Problem

Suppose we have a large random sample from a population with sample mean,  $\bar{y}_s$ . The population consists of two strata whose sizes are unknown. In addition the large sample contains no strata information.

Suppose we have a much, much smaller random sample where we learn the  $y$  values, the stratum membership and the value of the auxiliary variable  $x$  for each unit in the sample.

The population mean of  $x$  is assumed to be known.

How can we combine this information to get a good estimate of the strata means?

## A Solution

Use the second sample and the CPP to generate complete simulated copies of the population which satisfy two constraints.

One constraint comes from knowing the mean of  $x$ .

The other forces the mean of every simulated population to agree with  $\bar{y}_s$ .

This allows us to estimate the strata means and strata sizes using the units within each strata in the second sample.

## An Example

In Stratum 1

$x_i$ 's iid gamma(5);  $y_i|x_i$  ind Norm( $10 + x_i, 5^2$ )

In Stratum 2

$x_i$ 's iid gamma(7);  $y_i|x_i$  ind Norm( $8 + 3x_i, 15^2$ )

40% of the population belongs to Stratum 1.

The two sample sizes were 1000 and 40.

500 pairs of random samples were taken.

	str1	str2	pop
truemeans	15	29	23.4
big Samp	14.99	29.02	23.42
CPP	14.99	29.12	23.42
errCPP	1.09	1.35	

## Concluding Remarks

- Computations were done using the R package **polyapost** available in CRAN.
- Can estimate population quantities other than the mean.
- Will work when prior information involves linear equality and inequality constraints on population quantities.
- The CPP has the advantages of the Bayesian approach but only uses the kinds of prior information that are usually available.
- No need to select a model.