A noninformative Bayesian approach to finite population sampling using auxiliary variables

Glen Meeden
University of Minnesota

http://www.stat.umn.edu/glen/talks

Joint work with
Radu Lazar, Jeremy Strief
and David Nelson

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 auxiliary variables to the characteristic of interest.

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

Our approach is Bayesian and objective but we will not explicitly assume a model and so there will be no model selection!

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 $s \subset \{1, 2, \ldots, N\}$ using some sampling design Δ, usually SRSWOR.

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

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

How to relate the "unseen" to the "seen?

The Bayesian Way

Ericson (1969) JRSSB

Need joint prior distribution for the population

$$P\Big(y_1, y_2, \ldots, y_N\Big)$$

After observing sample must find

$$P\Big(y_j \,:\, j \notin s \,\Big|\, y_i \,:\, i \in s\Big)$$

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

$$E\Big(\mu(\mathbf{y}) \,\Big|\, y_i \; i \in s\Big) = \overline{y}_s$$

and

$$Var\Big(\mu(\mathbf{y}) \,\Big|\, y_i \; i \in s\Big) = \Big(1 - \frac{n}{N}\Big) \frac{v_s}{n} \frac{n-1}{n+1}$$

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

Noninformative Bayesian justification for some design based procedures.

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

Stepwise Bayes proves admissibility.

Instead of Polya sampling why not iid?

Relation to bootstrap

Assume SRSWOR is the sampling design and $N = kn$ for some integer k.

Given a sample s a good guess for the population is just k copies of $y(s)$.

Assuming our guess is the "truth" we can take repeated samples of size n from this "population" to get an estimate of variance for estimator.

Gross (1980) and Booth, Bulter and Hall (1994)

The Polya posterior uses the sample to construct many possible copies of the population to get an estimate of variance.

The bootstrap uses the sample to construct a single guess for the population and considers repeated draws from it to get an estimate of 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 seen

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

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 λ_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 λ, the vector of λ_j's, has approximately the uniform distribution on the $n-1$ dimensional simplex $\sum_{j \in s} \lambda_j = 1$.

Easy to Simulate from this distribution.

Easy to add constraints to this space. If $\mu(\mathbf{x})$ is know 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 $\mathbf{y}$ 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.

cor(PCS,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 3 strata and 2 auxiliary variables.

Size	The x_1's	The x_2's	The errors
300	gamma(10,1)	gamma(2,1)	normal$(0, 1)$
200	gamma(15,1)	gamma(7,1)	normal$(0, 1.5^2)$
400	gamma(5,1)	gamma(3,1)	normal$(0, 3.5^2)$

The true model

stratum 1: $y_i = 1 + x_{1i}x_{2i} + \epsilon_i$
stratum 2: $y_i = 3 + x_{1i} + x_{1i}x_{2i} + \epsilon_i$
stratum 3: $y_i = 2 + x_{2i} + x_{1i}x_{2i} + \epsilon_i$

We assumed that the population median of x_1 and the population mean of x_2 were known.

Simulation results

We generated 500 random samples selecting 25 units from the first stratum, 20 units from the second stratum and 35 units from the third stratum.

	Ave. value	Ave. abs err	Ave. low bd	Ave. len	Freq of cov
Mean	48.0	4.82	36.4	23.1	1.000
Strat	43.4	2.07	38.2	10.4	0.942
CnstPp	43.4	1.52	40.2	6.75	0.936

If just information about auxiliary means is available then the empirical likelihood based methods of Chen and Sitter (1999) and Zhong and Rao (2000) could be used. But what means?

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 the 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, $\overline{y}_s$. 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 population mean of x.

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

This allows us to estimate the strata means and strata sizes using the proportion of units in the second sample that fall within each stratum for each simulated copy of the entire population.

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
bigsmp	14.99	29.02	23.42
CPP	14.99	29.12	23.42
errCPP	1.09	1.35	

Weights and Standard Theory

Weights usually come from the sampling design. A unit's weight indicates how many units of the population it represents.

Taylor series argument for estimating the variance of estimators of complicated functions.

Weights are often adjusted; examples are raking and calibration.

Standard theory is sometimes obscure when it comes to variance estimation.

Why not be a Bayesian?

Bayesian Weights

Recall λ is the vector of proportions of units in the sample for a completed simulated copy of the population. Let

$$p = E_{CPP}(\lambda)$$

then

$$W = Np = \{Np_i : i \in s\}$$

is a set of Bayesian weights for the sample.

Note cannot arise in a full Bayesian analysis. Happens here because the CPP assumes that only the values that appear in the sample can occur in the population.

Why should a Bayesian care about W?

More on Bayesian Weights

A sophisticated Bayesian probably will not care. But in public use files where doing simulation is to hard naive users want weights attached to units.

The Bayesian weights will incorporate the same kinds of information that are used in the design based approach.

If the range of the weights is not to large then using them in the the usual frequentist Taylor series approach to variance estimation can be thought of as an approximation to a full blown CPP analysis.

Recall, the Horvitz-Thompson estimator is not used in practice when the range of the weights gets to large.

Can the CPP protect against really bad samples? Maybe.

A fun read

Statistical Information and Likelihood
A Collection of Critical Essays
by Dr. D. Basu

J. K. Ghosh, Editor (1988)

Basu gives an elegant argument for the Bayesian approach to finite population sampling. He demonstrates that for most designs the posterior does not depend on the design.

So the one area in statistics where prior information is often used the standard methods cannot be given a standard Bayesian justification!

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.