

Objective Bayesian weights
in finite population sampling

Glen Meeden
University of Minnesota

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

Joint work with Jeremy Strief

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?

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.

The posterior does not depend on the design.

Using the posterior

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 approximately point and interval estimates of the parameter of interest.

Can we find posteriors that have good design based properties?

The Approximate Polya Posterior

Suppose our beliefs about the unseen given the seen are exchangeable and $n \ll N$.

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 . Then under the approximate Polya posterior λ , (the vector of λ_j 's) has the uniform distribution on the $n - 1$ dimensional simplex $\sum_{j \in s} \lambda_j = 1$. Then

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

and

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

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

Not just a TTD

Easy to simulate from this distribution.

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

Lazar, Meeden and Nelson (2008) Survey Methodology

Stepwise Bayes proves admissibility.

Note that on the average for each $i \in s$ the value y_i appears N/n times.

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

Frequentist weights

For $i \in s$ let w_i be its associated weight which is the reciprocal of its inclusion probability.

Assume $\sum_{i \in s} w_i = N$. (Why not?)

Then the usual estimate of the population mean is

$$\bar{y}_w = \sum_{i \in s} \frac{w_i}{N} y_i$$

with an estimate of variance given by

$$\hat{V}_f(\bar{y}_w) = \frac{1}{n(n-1)} \sum_{i \in s} n \left(\frac{w_i}{N} y_i - \bar{y}_w \right)^2$$

This estimate assumes the sampling was done with replacement even when it was not.

Constructing a Good Guess

Assume the w_i 's for $i \in s$ are integers. Then given the sample a good guess for the population is the one that contains w_i copies of y_i for $i \in s$.

The mean of this constructed population is \bar{y}_w and let σ^2 denote its variance. Then

$$\widehat{V}_f(\bar{y}_w) = \frac{N-1}{N} \frac{\sigma^2}{n-1} + \frac{1}{n-1} \sum_{i \in s} \frac{w_i}{N} y_i^2 \left(n \frac{w_i}{N} - 1 \right)$$

The second term will be small if the w_i 's do not vary to much. It is 0 under SRSWOR where $w_i = N/n$.

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 approximate Polya posterior incorporate knowing $\mu(\mathbf{x})$?

Use the uniform distribution of the subset of the simplex defined by

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

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.

The CPP can incorporate linear constraints on several auxiliary variables.

Bayesian Weights

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

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

then

$$W = N\gamma = \{N\gamma_i : i \in s\}$$

is a set of Bayesian weights for the sample.

Note this cannot arise in a full Bayesian analysis. It 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 Bayesian weights is not to large then one can use them in the the usual frequentist Taylor series approach to variance estimation.

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

A Bayesian Way to Use Weights

Given some w_i 's for $\{i \in s\}$ let $\alpha_i = n(w_i/N)$.

Let λ_i be the proportion of units in a simulated copy of the population which take on the value y_i . Let λ have the Dirichlet($\alpha_1, \dots, \alpha_n$) distribution. Then

$$E(\mu(\mathbf{y}) \mid y_i \ i \in s) = \bar{y}_w = \sum_{i \in s} \frac{w_i}{N} y_i$$

and

$$\text{Var}(\mu(\mathbf{y}) \mid y_i \ i \in s) = \frac{N-1}{N} \frac{\sigma^2}{n+1}$$

where σ^2 is the variance of the constructed population created by using the weights.

When $w_i = N/n$ this is just the approximate Polya posterior.

Call this the Weighted Dirichlet (WDP) posterior.

Comparing CPP and WDP

The Bayesian weights depend on the $\{x_i : i \in s\}$ and the population mean of x but not on the design. They are the number of units in the population that each member of the sample represents.

The WDP is a looser version of the CPP. Under WDP simulated copies of the population only satisfy the population mean constraint for X on the average while under CPP each simulated copy must satisfy the constraint.

For small sample sizes this can be a good thing because the CPP interval estimates can sometimes be a bit too short and undercover.

Under WDP the posterior variance of the population mean will close to

$$\widehat{V}_f(\bar{y}_w) = \frac{N-1}{N} \frac{\sigma^2}{n-1} + \frac{1}{n-1} \sum_{i \in s} \frac{w_i}{N} y_i^2 \left(n \frac{w_i}{N} - 1 \right)$$

An example

$$N = 2000$$

The X_i 's are iid gamma(5)

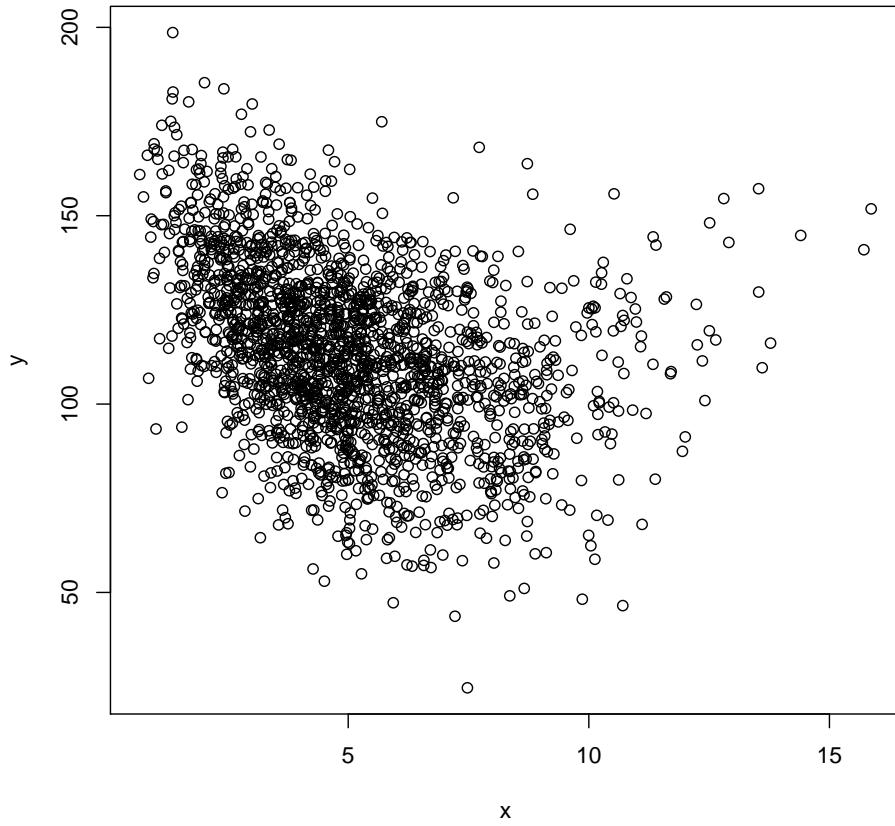
$$Y_i = 100 + (X_i - 8)^2 + Z_i$$

Where the Z_i 's are iid normal(0,20²)

The total of the y_i 's is 227,923.0

The median of the y_i 's is 114.12

The plot



More on the Example

$n = 60$ and we assume all the x_i 's in the population are known.

We form 3 post-strata using $x_{[20]}$ and $x_{[40]}$ the twentieth and fortieth largest members of the sample.

We consider the post-stratified estimator and the regression estimator.

CPP use the constraints from the post-stratification and the population mean of x .

We took 500 samples under 3 different sampling plans.

SRS without replacement

ave min and max of CPP wts 0.658 1.58

Results for estimating the total = 227923.0

| method | pctest | abserr | lowbd | length | freqcov |
|---------|----------|--------|----------|---------|---------|
| freqstr | 227856.1 | 4165.0 | 217190.1 | 21332.1 | 0.950 |
| freqreg | 227602.1 | 4302.7 | 216951.9 | 21300.3 | 0.944 |
| cnstpp | 227546.9 | 4190.6 | 217592.7 | 19909.7 | 0.932 |
| wtdirch | 227546.9 | 4190.6 | 216032.0 | 23029.7 | 0.958 |

Results for estimating the median = 114.12

| | | | | | |
|---------|--------|-------|---------|--------|-------|
| cnstpp | 113.31 | 2.731 | 106.843 | 13.219 | 0.936 |
| wtdirch | 113.33 | 2.675 | 106.205 | 14.554 | 0.956 |

PPS proportional to x

ave min and max of CPP wts 0.374 3.024

Results for estimating the total = 227923.0

| method | pctest | abserr | lowbd | length | freqcov |
|---------|----------|--------|----------|---------|---------|
| freqstr | 225295.8 | 5228.9 | 213791.6 | 23008.4 | 0.916 |
| freqreg | 224207.2 | 5611.2 | 213317.1 | 21780.3 | 0.878 |
| cnstpp | 227471.1 | 4919.2 | 216973.2 | 20947.3 | 0.894 |
| wtdirch | 227471.1 | 4919.2 | 216117.8 | 22706.6 | 0.936 |

Results for estimating the median = 114.12

| | | | | | |
|---------|---------|-------|---------|--------|-------|
| cstpp | 113.643 | 2.733 | 106.803 | 13.816 | 0.944 |
| wtdirch | 113.587 | 2.734 | 106.486 | 14.273 | 0.95 |

PPS proportional to iid gamma(5) + 5

ave min and max of CPP wts 0.651 1.583

Results for estimating the total = 227923.0

| method | pctest | abserr | lowbd | length | freqcov |
|---------|----------|--------|----------|---------|---------|
| freqstr | 227976.5 | 4371.2 | 217349.4 | 21254.1 | 0.938 |
| freqreg | 227715.5 | 4462.2 | 217062.5 | 21305.9 | 0.934 |
| cnstpp | 227721.2 | 4420.6 | 217719.6 | 19924.4 | 0.908 |
| wtdirch | 227721.2 | 4420.6 | 216270.5 | 22901.4 | 0.95 |

Results for estimating the median = 114.12

| | | | | | |
|---------|---------|-------|---------|--------|-------|
| cstpp | 113.549 | 2.707 | 107.044 | 12.988 | 0.926 |
| wtdirch | 113.558 | 2.694 | 106.464 | 14.265 | 0.952 |

Slightly less efficient than SRS

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 and WDP have the advantages of the Bayesian approach but only uses the kinds of prior information that are usually available.
- CPP yields Bayesian weights that can either be used in a Bayesian manner in WDP or in standard frequentist formulas.