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.

There is a 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 a 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.

The posterior does not depend on the design.

Bayesian inference is “easy”

Simulate from the posterior to get completed copies of the entire population.

For the parameter of interest compute its value for each simulated complete copy of the population. (No need to treat estimating a mean or a median as different problems.)

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 $n << N$.

For a $j \in s$ let p_j be the proportion of units in a completed simulated copy which take on the value y_j. Then then under the approximate Polya posterior $\mathbf{p} = (p_1, \ldots, p_n)$ has the uniform distribution on the $n-1$ dimensional simplex $\sum_{j \in s} p_j = 1$. When estimating the population mean, $\mu(\mathbf{y})$,

$$E\Big(\mu(\mathbf{y}) \,\Big|\, y_i \; i \in s\Big) = \bar{y}_s \quad \text{and} \quad Var\Big(\mu(\mathbf{y}) \,\Big|\, y_i \; i \in s\Big) = \frac{v_s}{n} \frac{n-1}{n+1}$$

where $\bar{y}_s$ and v_s are the sample mean and sample variance.

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. It uses these values to get a point 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 approximate Polya posterior incorporate knowing $\mu(\mathbf{x})$?

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

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

Harder to simulate in the 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.

The CPP can incorporate linear constraints on several auxiliary variables.

Bayesian weights

For $j = 1, \ldots, n$ let

$$w_j = N E_{CPP}(p_j)$$

where the expectation is taken with respect to the CPP. Note $\sum_{j=1}^{n} w_j = N$.

These weights depend only on the observed values of the auxiliary variables and the known population constraints and have the usual interpretation.

Note such weights 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.

They do not depend explicitly on the sampling design. In many problems however they can be used in frequentist formulas just like design based weights.

The "best guess" constructed population

Consider the constructed population where the number of units in the population of type (y_i, x_i) is w_i for $i = 1, \ldots, n$.

This then is our best guess for the unknown population and

$$\bar{y}_{bw} = \sum_{i=1}^{n} \frac{w_i}{N} y_i \quad \text{and} \quad \sigma^2_{bw} = \sum_{i=1}^{n} \frac{w_i}{N} (y_i - \bar{y}_{bw})^2$$

are the mean and variance of this constructed population.

Next we will give an alternative way to think about this "best guess" population.

The Weighted Dirichlet posterior (WDP)

Given our weights, the w_j's, consider the the Dirichlet distribution over the simplex defined by the vector $(nw_1/N, \ldots, nw_n/N)$ as an alternative posterior distribution for $p = (p_1, \ldots, p_n)$. It can be used to generate complete simulated copies of the population.

We call this posterior the weighted Dirichlet posterior (WDP).

Note the WDP is a looser version of the CPP. Under the CPP every complete copy of the population will satisfy the constraints. Under the WDP, only the average of all the simulated populations will satisfy the constraints.

Posterior mean and variance of WDP

It is easy to see that under the WDP

$$E\left(\sum_{i=1}^{n} p_i y_i\right) = \sum_{i=1}^{n} \mu_i y_i = \bar{y}_{bw}$$

$$V\left(\sum_{i=1}^{n} p_i y_i\right) = \frac{1}{n+1}\sigma_{bw}^2$$

where $\bar{y}_{bw}$ and σ_{bw}^2 are the mean and variance of our "best guess" constructed population.

The CPP and WDP will have the same point estimate of the population mean but the posterior variance of WDP will be larger.

Design based weights

Let the w_i's be the inverse of the inclusion probabilities.

Let $W = \sum_{i=1}^{n} w_i$. Recall W need not equal N.

The mean and variance of the "best guess" population is

$$\bar{y}_{dw} = \sum_{i=1}^{n} \frac{w_i}{W} y_i \quad \text{and} \quad \sigma^2_{dw} = \sum_{i=1}^{n} \frac{w_i}{W} (y_i - \bar{y}_{dw})^2$$

Design based estimate of variance of $\bar{y}_{dw}$

This is a hard problem. Assuming the sampling was done with replacement, even when it was the not, the usual recommended estimate is

$$\begin{aligned}\widehat{V}_d(\bar{y}_{dw}) &= \frac{1}{n(n-1)}\sum_{i=1}^{n}\left(n\frac{w_i}{W}y_i - \bar{y}_{dw}\right)^2 \\ &= \frac{\sigma^2_{dw} + \gamma_{dw}}{n-1}\end{aligned}$$

where

$$\gamma_{dw} = \sum_{i=1}^{n}\frac{w_i}{W}y_i^2\left(n\frac{w_i}{W} - 1\right)$$

Note that when the design is simple random sampling and $N = nk$ then $\gamma_{dw} = 0$ and this estimate agrees with the WDP estimate.

A simulation example

The x variable was a random sample from a lognormal population.

In population A the distribution of $y_i|x_i$ was normal with mean $2x_i$ and standard deviation $5\sqrt{x_i}$ which yielded $\rho_{y,x} = 0.49$.

Population B was constructed in the same way except each y_i had an additional 400 added to it.

The sampling design was pps using x in both cases. The Horvitz-Thompson estimator should work well for population A. It will be compared to the CPP and WDP estimators.

Simulation results

Comparing the HT estimator of the population total to those based on the CPP and the WDP which assume the population mean of x is known. Results are based on 300 samples of size 50 drawn using pps(x). The nominal coverage for each method is 0.95

Population	Method	Ave. abs err	Ave. len	Freq of coverage
A	HT	4434.8	22008.5	0.933
B	HT	8738.8	45027.4	0.970
A	CPP	4435.9	21180.1	0.920
B	CPP	4250.8	21375.0	0.950
A	WDP	4435.9	24172.7	0.953
B	WDP	4250.8	25424.7	0.977

Why are the CPP and WDP estimators more robust?

Let y_i denote a unit's value in population A and y_i' its corresponding value in population B. Then

$$\sum_{i=1}^{50} w_i y_i' = \sum_{i=1}^{50} w_i y_i + 400 \sum_{i=1}^{50} w_i$$

For the HT estimator the second term in the above equation is adding additional variablity. In population B calculations show that the term γ_{dw} is positive and can be quite large. (It tends to be small and negative in population A.)

γ_{dw} is accounting for the extra variablity in the HT estimator in population B which results from that fact that here $y_i \propto 2x_i + 400$ and not x_i

Another example

$N = 2000$

The X_i's are iid gamma(5)

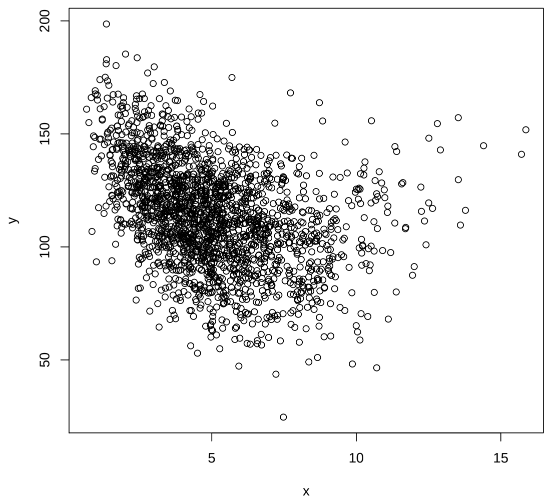
$Y_i|x_i = 100 + (x_i - 8)^2 + Z_i$

Where the Z_i's are iid normal(0,20^2)

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. The 1st stratum is all the units in the population $\leq x_{[20]}$. The 2nd all the population units between $x_{[20]}$ and $\leq x_{[40]}$.

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	ptest	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	ptest	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	0.950

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	ptest	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	0.950

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 then SRS

Consistency of CPP in a special case

Let x be a single auxiliary variable which takes on just finitely many values $x_1, \ldots, x_k$.

Let λ_i be the true population proportion of x_i.

The CPP assumes that the population mean of x, say $\mu_x = \sum_{i=1}^{k} \lambda_i x_i$ is known.

The sampling design is such that within each x_i strata the inclusion probabilities are the same for each unit.

Let n_i be the number of units in the sample with value x_i. If the sampling design is such $n_i/n \asymp \lambda_i$ for large n then the CPP estimator of the population mean is consistent.

More on Weights

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

If the range of the Bayesian weights is not too large then one can use them in the the usual frequentist Taylor series approach to variance estimation.

A good set of weights is one which yields a good "best guess" for the population. The weights need not depend on the selection probabilities.

Ignoring this fact creates (I believe) many difficulties for the standard theory.

Concluding Remarks

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.

Can estimate population quantities other than the mean.

Will work when prior information involves linear equality and inequality constraints on population quantities.

Computations were done using the R package **polyapost** available in CRAN.