Objective Stepwise Bayes Weights in Survey Sampling

Glen Meeden
University of Minnesota

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

Population information is used to select a design and to adjust weights after the sample is observed.

Objectivity, i.e. good frequentist properties, under the selected design is important.

For a Bayesian the information is summarized in their prior distribution. It is both easy and impossible to implement.

A stepwise Bayes approach can be easier to use and be objective.

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

y_i is value of single characteristic for unit i.

$\mathbf{y} = (y_1, \ldots, y_N) \in \Theta \subseteq \mathcal{R}^N$

Let Δ denote the sampling design which is used to select a sample $s \subset \{1, 2, \ldots, N\}$ of size n

The basic problem in sample survey is how to relate the information in the sample

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

to

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

We will be interested in estimating a population total or mean or median. Let

$t(y) = \sum_{i=1}^{N} y_i = N\mu(\mathbf{y})$

When Δ is simple random sampling without replacement (SRSWOR) the usual estimator of $t(y)$ is $N\overline{y}_s = N\sum_{i\in s} y_i/n$

An unbiased estimator of its variance is

$$N^2\Big(1-\frac{n}{N}\Big)\frac{v_s}{n}$$

where

$$v_s = \sum_{i\in s}(y_i-\overline{y}_s)^2/(n-1)$$

is the sample variance.

Design based weights

Weights usually come from the sampling design.

If π_i is the probability that unit i is included in the sample then $w_i = 1/\pi_i$ is the weight assigned to that unit.

A sampled unit's weight represents how many units of the population it represents.

Under SRSWOR $\pi_i = n/N$ and so $w_i = N/n$

The Horvitz-Thompson Estimator

For a general design Δ an unbiased estimator of $t(y)$ is

$$\delta_{HT}(y(s)) = \sum_{i \in s} w_i y_i$$

Weights are often adjusted; examples are raking and calibration

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

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

Auxiliary Variable

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

Sometimes all the x_i's are known and in other cases only x(s) is known along with $\mu(\mathbf{x})$, the population mean of $\mathbf{x}$.

In the second case either the ratio estimator or the regression estimator can be used.

These estimators work best when the relationship between the y_i's and x_i's follow certain linear models.

The Bayesian Way

Ericson (JRSSB (1969)) and Basu (See Ghosh (1988))

Need a joint prior distribution for the population

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

After observing the sample one 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"

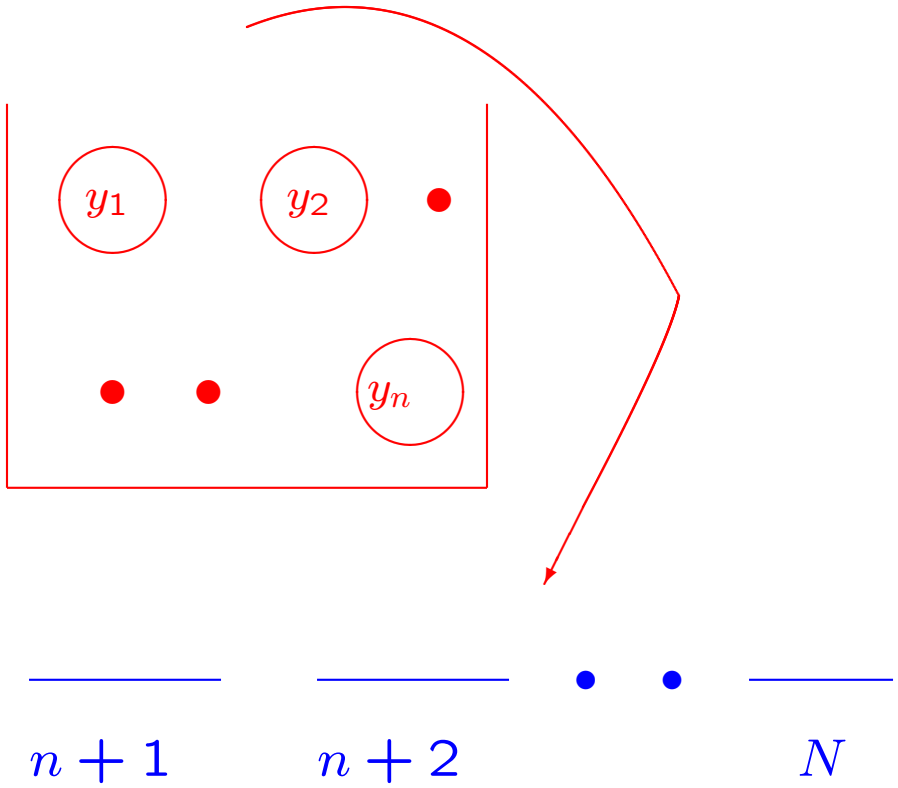
Simulate from the posterior to get completed copies of the entire population.

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 Polya posterior



Under the Polya Posterior

It is easy to generate a simulated complete copy of the population (SCCP) using the Polya posterior

A SCCP will only contain values that appear in the sample.

If $p_i = (\#$ of times y_i appears in a SCCP$)/N$ then $E(p_i) = 1/n$.

Inferences can be made by generating many SCCP's.

The total weight assigned to the sample values in the initial urn is n.

Under the Polya Posterior

The Polya posterior makes sense when little prior information is available about the population, i.e. for a Bayesian their beliefs about the y_i's are exchangeable. In such a case a frequentist would use SRSWOR.

Under this posterior distribution one finds

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

and

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

Not just a TTD

It yields a non-informative stepwise Bayes justification for some standard design based procedures by proving their admissibility.

Joshi (1966)
Ghosh and Meeden (1997)
Lo (1988) Annals and Rubin (1981) Annals
Nelson and Meeden (2006) JSPI – Median
Lazar, Meeden and Nelson (2008) Survey Methodology
Strief and Meeden (2013) Survey Methodology
Geyer and Meeden (2013) Bayesian Analysis

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

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 simulated full copy it calculates the parameter of interest. It uses these values to get a point estimate and an estimate of its variance.

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 the distribution of $\mathbf{p} = (p_1, \ldots, p_n)$ under the Polya posterior is approximately the uniform distribution on the $n-1$ dimensional simplex $\sum_{j \in s} p_j = 1$.

So there are two ways to simulate from the Polya posterior; the exact and the approximate.

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 over the subset of the simplex defined by

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

Harder to simulate values from this restricted space.

The Constrained Polya Posterior (CPP)

For situations where the regression estimator would be used the point and interval estimators 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. Meeden (1995) showed that in a variety of populations the CPP did on the average 10% better.

The CPP can incorporate constraints involving the median of $\mathbf{x}$. More generally it can incorporate linear equality and inequality constraints, for several auxiliary variables.

We will now argue that the CPP yields weights for the units in the sample in a natural way and that these weights yield inferences with good design properties.

Stepwise Bayes Weights

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

$$w_j = NE_{CPP}(p_j)$$

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.

A "best guess" constructed population

Given a sample $y(s)$ along with its weights 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.

Easier to simulate from the WDP than the CPP.

Posterior mean and variance of $\mu(y)$ under the WDP

It is easy to see that under the WDP

$$E\Big(\sum_{i=1}^{n} p_i y_i\Big) = \sum_{i=1}^{n} E(p_i) y_i = \sum_{i=1}^{n} \frac{w_i}{N} y_i = \bar{y}_{bw}$$

$$V\Big(\sum_{i=1}^{n} p_i y_i\Big) = \frac{1}{n+1} \sigma^2_{bw}$$

where $\bar{y}_{bw}$ and σ^2_{bw} 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 this "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. The usual recommended approach is to assume that the sampling was done with replacement, even when it was the not. This yields the estimate

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

where

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

Note that when the design is simple random sampling and $N = nk$ then $\gamma_{dw} = 0$ and this estimate almost agrees with the WDP estimate. It is off by the factor $(n-1)/(n+1)$.

A simulation example

For a population of size 2,000 the x variable was a random sample from a gamma(5) distribution + 20.

In population A the distribution of $y_i|x_i$ was normal with mean $5x_i$ and standard deviation 20 which yielded $\rho_{y,x} = 0.49$.

In population B 400 was added to each y_i .

The sampling design was pps based on x.

The Horvitz-Thompson (HT) estimator should work well for population A.

It will be compared to the WDP estimator which assumes $\mu(x)$ is known.

Simulation results

For A $t(y) = 249,044$. Results based on 500 samples of size 50. The HTRW estimator is the HT estimator renormalized so that the weights sum to $N = 2,000$. The nominal coverage for each method is 0.95

Population	Method	Ave. abs err	Ave. len	Freq of coverage
A	HT	4,628	21,898	0.940
B	HT	8,965	43,914	0.960
A& B	WDP	4,706	24,381	0.960
A	HTRW	5,051	21,897	0.896
B	HTRW	5,051	43,919	0.998

Why is the 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

$$\sum_{i \in s} w_i y_i' = \sum_{i \in s} w_i y_i + 400 \sum_{i \in s} w_i$$

For the HT estimator the second term in the above equation is adding additional variablity. For 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 $2x_i$

Another example

$N = 2000$

The X_i's are iid gamma(5)

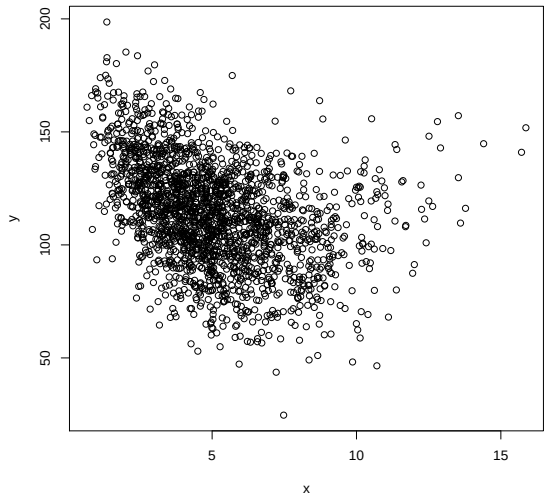
$Y_i|x_i = 100 + (x_i - 8)^2 + Z_i$

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.

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

We took 500 samples under 4 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
wtdirch 227546.9 4190.6 216032.0 23029.7 0.958

Results for estimating the median = 114.12

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
wtdirch 227471.1 4919.2 216117.8 22706.6  0.936

Results for estimating the median = 114.12

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
wtdirch 227721.2 4420.6   216270.5 22901.4  0.950

Results for estimating the median = 114.12

wtdirch 113.558 2.694 106.464 14.265 0.952
```

WDP is slightly less efficient then SRS

y Dependent (Range(y) $\rightarrow [1, 2]$)

```
Results for estimating the total = 227923.0

method  ptest        abser    length    freqcov
freqstr 231,590.0    5,229.0  21,170.8   0.892
freqreg 231,424.4    5,143.4  21,127.9   0.902
wtdirch 231,139.1    4,967.6  22,867.0   0.938
```

WDP seems to give some slight protection against bias in sampling design but can only do so much. If Range(y) $\rightarrow [1, 4]$ the WD intervals only cover 86% of the time while the freqstr intervals cover just 80% of the time.

Stratification and estimating the median

When only a few observations are taken from each strata, say 2, finding a good confidence interval for the population median can be hard.

Suppose we have L strata where N_j is the size of stratum j.

Applying the Polya posterior independently within each stratum means that each of the two sampled units in stratum j should get weight LN_j/N, since then $2\sum_{j=1}^{L} LN_j/N = 2L$, the total sample size.

Will compare the WDP with these weights to the standard estimate.

The population

$L = 20$ and N_j's $\sim$ iid Poisson(100).

The strata means $\sim$ iid Normal(150,σ^2) with either $\sigma = 10$ or $\sigma = 20$.

The strata standard deviations $\sim$ iid from a gamma distribution with scale parameter one and shape parameter $\gamma\sigma$ with either $\gamma = 0.10$ or $\gamma = 0.25$.

We took 500 samples and compared the two methods.

Simulation Results

Method	Ave. value	Ave. err	Ave. len	Freq of coverage
sigma=10 and gamma=0.10				
Stand	148.40	2.37	8.30	0.808
WDP	148.39	2.22	12.20	0.950
sigma=10 and gamma=0.25				
Stand	144.28	5.70	20.59	0.834
WDP	144.18	5.41	28.3	0.950
sigma=20 and gamma=0.10				
Stand	152.75	3.02	10.52	0.828
WDP	152.61	2.78	22.88	0.996
sigma=20 and gamma=0.25				
Stand	155.94	6.72	23.17	0.826
WDP	155.89	6.35	34.96	0.962

Comments

The point estimator based on the WDP seems to do just a bit better. but the confidence intervals produced by WDP are clearly superior.

Results when $L = 40$ are similar.

Too long intervals for WDP happen when the strata means vary widely and the strata variances tend to be relatively small.

When the sample size was increased to four units per stratum the difference between the two methods is not so dramatic but the story remains much the same.

More on Weights

The frequentist theory of weights is a mess? Gelman (2007)

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. (Rao and Wu (2010))

Stepwise Bayes weights incorporate the same kinds of information that are used in the design based approach.

If the range of the Stepwise Bayes weights is not too large then one can use them in the the usual frequentist Taylor series approach to variance estimation. (Strief (2007))

Concluding Remarks

Estimators based on the CPP are consistent. Geyer and Meeden (2013).

The CPP and WDP have the advantages of the Bayesian approach but are objective.

Will work when prior information involves linear equality and inequality constraints on population quantities of auxiliary variables and yields estimates of population quantities other than the mean.

More work needs to be done for more complicated designs. Meeden (1999) considered cluster sampling.

Computations were done using the R package **polyapost** available on CRAN. Simulating complete copies of the population becomes harder for more complicated designs.