CORRECTING THE BIAS IN THE RANGE OF A STATISTIC ACROSS SMALL AREAS

David R. Judkins, Westat, Inc. and Jun Liu, Research Triangle Institute
David Judkins, Westat, Inc., 1650 Research Blvd., Rockville, MD 20850

Key Words: Range Estimator, Small Area Estimation, Design-based, Empirical Bayes, Bayes

Introduction

Interest in state estimates has increased over the past year with the passage of block grants to the states replacing certain welfare programs. Currently two federal demographic surveys are producing estimates for every state and more may follow. The Current Population Survey (CPS) has been producing state estimates since the late 1970s, originally in response to the CETA act which required accurate state-specific labor force data to distribute federal funds. The National Immunization Survey has been producing state estimates since 1984, in response to President Clinton's immunization initiative. Other state specific surveys are under consideration or partial implementation (such as the BRFSS). As these surveys are analyzed, it is frequently overlooked that the properties of an ensemble of statistics can be as important as the properties of each specific state estimator. Important exceptions to this general neglect include Louis (1984), Spjøtvoll and Thomsen (1987), Lahiri (1990), and Ghosh (1992). Ghosh and Rao (1994) nicely summarized this earlier work in Section 7.2 of their paper. It is important to consider the properties of the ensemble since states at one extreme are punished (through unfavorable publicity or lack of additional funds) while states at the other extreme are rewarded. Some statistics that depend on the ensemble include the rank ordering, the minimum, the maximum, the range, and percentiles. In this paper, we focus on the range. If there is a large range, then there will be calls for action to address the range and debates on the reasons for the disparity among the states. Much the same thing can happen in an industrial plant where some output from different sites is tested for quality in some way. Managers of sites with low quality ratings get punished and those with high ratings get rewarded.

It would thus seem that although an argument can be made for publishing the best possible estimate for each state considered individually, it might be better to compromise the quality of individual state estimates in order to improve the properties of the ensemble. In particular, we argue that ensemble estimates should be published that are expected to have the correct range across the states or other domains of analytic interest. Concern over the range was first voiced to one of the authors in the early 1980s by Barbara Bailar, then an Associate Director at the U. S. Bureau of the Census. When sample sizes are large for each state, the range of the sample means is approximately correct for the range of the true parameters. However, when the sample sizes per state are small and the natural variability among the states is small, the range in the sample means can be seriously positively biased. Beverly Causey developed an unpublished proof that this was true, but a solution to the problem alluded statisticians at the Bureau.

Some years later, it is clear that the solution lies in constrained hierarchical Bayesian or constrained empirical Bayesian methods, as coined by Ghosh (1992). Pure design-based methods do not admit a solution to this problem. In this paper, we first review the basic results that the design-based estimate of the dispersion across states can be severely positively biased, that Bayes and Empirical Bayes estimates of the dispersion under standard loss functions can be severely negatively biased and how the constrained Bayes methods can strike the appropriate compromise between these opposing methods. We reason that if the dispersion is over- or under-estimated, then the same must be true for the range. In this review, we focus on the problem of estimating state means of normal characteristics since the results are easier to derive. Binary characteristics are of more interest for most demographic surveys, but there do not appear to be closed form expressions for the bias of the Bayes and Empirical Bayes procedures for binary characteristics. We then give the results from a small simulation study for a binary characteristic. Although we applied the constrained empirical Bayes method to obtain a solution to the original CPS problem, there was not enough space to show it.

Review of Theory

Let $\theta = (\theta_1, \cdots, \theta_L)$ be the vector of true state means for the characteristic of interest at a particular point in time. Examples would include state per capita incomes, unemployment rates, poverty rates,

immunization rates, substance abuse rates, and so on. Assume that the process giving rise to these true means at different points in time is a stochastic process. While it is impossible to verify whether the process is deterministic or stochastic, viewing the process as stochastic and making some further assumptions about the distribution governing the process allows us to use Bayesian methods to make stronger inferences about the state means than is possible through design-based methods. It also allows us to make inferences about the characteristics of any ensemble of estimates for θ; something that is impossible with design-based methods.

Assume that the process giving rise to the true state means is independent across the states with common mean μ and variance σ^2. If interest focuses on a binary characteristic, it is necessary to make the additional restriction that $\sigma^2 < \mu(1-\mu)$. Assume that a sample of size n_i was drawn from the i-th state. Let $\hat{\theta}_i^D$ be a design-unbiased estimate of θ_i. Let $\hat{\theta}_\cdot^D$ be the average of the design-unbiased estimates across the states. Then the expected dispersion of the design-based estimates with respect to both the sample design and the model is

$$E_M E_D \frac{1}{L-1}\sum_{i=1}^{L}\left(\hat{\theta}_i^D - \hat{\theta}_\cdot^D\right)^2 = \frac{1}{L}\sum_{i=1}^{L} E_M \mathrm{Var}_D\left(\hat{\theta}_i^D \middle| \theta_i\right) + \sigma^2,$$

where E_M denotes expectation with respect to the model and E_D denotes expectation with respect to the design. Note that the first term on the right is the expected measurement variance while the second term is the true process variance. This may be rewritten as

$$E_M E_D S_D^2 = \varphi^2 + \sigma^2.$$

Since the measurement variance is strictly positive, the sample means are expected to be more dispersed than the true state means. With higher dispersion, it is clear that the range of the estimated state means will be positively biased. Of course, if the measurement variance is negligible, then the bias in the range will also be negligible.

If the characteristic is binary, then this expected dispersion of the design-based estimates is

$$E_M E_D S_D^2 = \frac{\mu(1-\mu) - \sigma^2}{\tilde{n}} + \sigma^2,$$

where $\tilde{n}$ is the harmonic mean of the state sample sizes. So the relative bias of design-based estimate of the dispersion across the states is

$$\mathrm{RB}(S_D^2) = \frac{\frac{\mu(1-\mu)}{\sigma^2} - 1}{\tilde{n}} = \frac{1-\rho}{\tilde{n}\rho},$$

where $\rho = \sigma^2 / (\mu(1-\mu))$ is the intrastate correlation for the characteristic.

Table 1 shows examples of the magnitude of overestimation for different intrastate correlations and state sample sizes. This bias is trivial when the intrastate correlation is high and the state sample sizes are high. For small intrastate correlation and small state sample sizes, the bias can be very large. Generally, intrastate correlations tend to be quite small for large geographic classes. For example, the intrastate correlation (at the state level) was just 0.008 for the percent of the total population age 16 and older that was employed in 1994. The intrastate correlation for the unemployment rate in 1993 was just 0.003. This suggests that the dispersion across the states will not be reasonably estimated by design-based estimators based on fewer than several thousand interviews per state. When high measurement variance is present (due to small sample sizes in the states), then many turn to model-based or Bayesian estimation procedures to estimate the state means. This is the field known as small area estimation.

Table 1. Relative Bias in Design-based Estimates of the Dispersion across States

State Sample Size	Intrastate correlation						
	0.001	0.005	0.01	0.025	0.05	0.1	0.125
30	3330%	663%	330%	130%	63%	30%	23%
40	2498%	498%	248%	98%	48%	23%	18%
50	1998%	398%	198%	78%	38%	18%	14%
100	999%	199%	99%	39%	19%	9%	7%
200	500%	100%	50%	20%	10%	5%	4%
400	250%	50%	25%	10%	5%	2%	2%
800	125%	25%	12%	5%	2%	1%	1%
1600	62%	12%	6%	2%	1%	1%	0%
3200	31%	6%	3%	1%	1%	0%	0%
6400	16%	3%	2%	1%	0%	0%	0%

However, Louis (1984) noticed that Bayesian estimates of the state means compress the variation too much. We may infer from this that the range of the Bayesian estimates is negatively biased. Louis (1984) dealt with the simple case where the distribution of the state means is normal with known variance and, given a set of realized state means,

each state sample is a simple random sample from a normal distribution with known variance. In this case, the conditional distribution of $\hat{\theta}_i^D$ given θ_i is $N(\theta_i, \psi_i^2)$, the prior distribution for θ_i is $N(\mu, \sigma^2)$, μ, $\{\psi_i^2\}$ and σ^2 are all fixed and known, and the loss function is $\sum_{i=1}^{L}(\hat{\theta}_i - \theta_i)^2$. It is well known that for this model, the standard Bayes estimate of θ_i is

$$\hat{\theta}_i^B = (1-\gamma_i)\mu + \gamma_i \hat{\theta}_i^D$$

where

$$\gamma_i = \sigma^2 / (\psi_i^2 + \sigma^2).$$

Assume further that the state measurement variances are all equal. In this case, $\varphi^2 \equiv \psi_i^2$ and $\gamma \equiv \gamma_i$. Louis (1984) showed that under these conditions, the actual dispersion in the estimates is

$$S_B^2 = \gamma^2 S_D^2.$$

Combining this result with the earlier result, we have that

$$E_M E_D S_B^2 = \gamma^2(\varphi^2 + \sigma^2) = \gamma\sigma^2.$$

Using the new symbol γ, we note that the expected dispersion of the design-based estimates is

$$E_M E_D S_D^2 = \varphi^2 + \sigma^2 = \frac{\sigma^2}{\gamma}.$$

So the design-based and Bayesian methods both mis-estimate the expected dispersion across the states by a factor of γ. The design based estimate of the dispersion is too high by a factor of $1/\gamma$, while the Bayesian method is too low by a factor of γ. Louis (1984) then proposed a constrained Bayesian estimator with the correct expected dispersion across the states.

Spjøtvoll and Thomsen (1987) did similar work for a binary outcome variable. They developed an estimator by using design-based estimates of the measurement variance, and then subtracting this from the total observed weighted dispersion among the states to estimate the true variance among the states. They then used the components of variance to adjust the state estimates. Lahiri (1990) followed a similar approach but with some slight changes was able to develop an estimator with nice consistency properties. Ghosh (1992) developed a more general approach to the entire problem that relaxes some of Lahiri's assumptions.

Lahiri's estimators work fine if a prior for the state means can be found such that the posterior expected values of the state means can be expressed as a linear function of the state sample means. When the characteristic of interest is binary, such linearity in the posterior estimates can not usually be achieved. The only admissible priors are those that yield state means in the range of 0 to 1. The most common priors for the means of binary variables are the beta distribution, the logit-normal distribution, the probit-normal distribution, and the truncated normal distribution. The logit normal is the most popular since it allows the specification of fixed and random effects on the same scale (Zeger and Karim (1991), Breslow and Clayton (1993), McCulloch (1997), among others). For these priors, the posterior expected state means can only be found by numerically intensive iterative methods. It appears that a combination of these methods with Ghosh's adjustment to correct the dispersion would constitute a very promising line of research. Unfortunately, these combinations would be very difficult to program and apply to many applications. We thus thought it useful to apply Spjøtvoll and Thomsen and Lahiri's simple estimators to some simulated populations where the assumptions do not apply exactly to determine if these simple methods yield useful improvements over the design-based estimates.

Spjøtvoll-Thomsen Estimator

There appears to be an error in Spjøtvoll and Thomsen's estimator of the process variance. They estimate the true process variance as

$$\hat{\sigma}_{ST}^2 = \max\left\{0, \frac{\sum \frac{n_i}{n}(\bar{y}_i - \bar{\bar{y}})^2 - k\bar{\bar{y}}(1-\bar{\bar{y}})}{T-k}\right\}.$$

In every application we tried, this resulted in an estimated zero process variance. We think that they perhaps meant to write:

$$\hat{\sigma}^2_{ST} = \max\left\{0, \frac{T\Sigma\frac{n_i}{n}(\bar{y}_i - \bar{\bar{y}})^2 - k\bar{\bar{y}}(1-\bar{\bar{y}})}{T-k}\right\}.$$

The latter gives reasonable results, but we used the estimator of σ^2 suggested by Lahiri as described in the next section. Spjøtvoll and Thomsen then estimated the measurement variance as

$$\hat{\varphi}^2_i = \frac{\bar{\bar{y}}(1-\bar{\bar{y}}) - \hat{\sigma}^2_{ST}}{n_i}.$$

They then define a compositing factor of

$$\alpha_i = \sqrt{\frac{\hat{\sigma}^2_{ST}}{\hat{\varphi}^2_i + \hat{\sigma}^2_{ST}}}$$

and an estimator for the i-th state of

$$\hat{\theta}^{ST}_i = \alpha_i \bar{y}_i + (1-\alpha_i)\bar{\bar{y}},$$

where $\bar{\bar{y}} = \frac{1}{n}\sum_i n_i \bar{y}_i$.

Lahiri Estimator

Rather than working with the design-based estimates, Lahiri first derives empirical Bayes estimates of the state means under the linearity assumption and then adjusts the empirical Bayes estimates by expanding their dispersion. Although Lahiri calls these estimates, "adjusted empirical Bayes estimates," the label of Ghosh seems more appropriate. Hence, we call them "constrained empirical Bayes (CEB) estimates." Lahiri defines τ^2 as the expected variance given a single observation. Thus,

$$\varphi^2_i = \tau^2 / n_i.$$

He then estimates

$$\hat{\tau}^2 = \frac{1}{n-L}\sum_i n_i \bar{y}_i(1-\bar{y}_i),$$

$$\hat{\sigma}^2 = \max\left\{0, \left[\frac{\Sigma n_i(\bar{y}_i - \bar{\bar{y}})^2}{L-3} - \hat{\tau}^2\right]\frac{(L-1)n}{n^2 - \Sigma n_i^2}\right\}$$

with,

$$\hat{\gamma}_i = \hat{\sigma}^2/(\hat{\tau}/n_i + \hat{\sigma}^2), \quad \hat{\mu} = \Sigma\hat{\gamma}_i\bar{y}_i/\Sigma\hat{\gamma}_i,$$

and

$$\hat{\theta}^{EB}_i = \hat{\gamma}_i\bar{y}_i + (1-\hat{\gamma}_i)\hat{\mu}, \quad \hat{\theta}^{EB}_\cdot = \frac{1}{L}\Sigma\hat{\theta}^{EB}_i,$$

and

$$F = \begin{cases} \sqrt{1 + \dfrac{\hat{\tau}^2(L-1)\Sigma\dfrac{\hat{\gamma}_i}{n_i}}{L\Sigma(\hat{\theta}^{EB}_i - \hat{\theta}^{EB}_\cdot)^2}} & \text{if } \Sigma(\hat{\theta}^{EB}_i - \hat{\theta}^{EB}_\cdot)^2 > 0 \\ 1 & \text{otherwise.} \end{cases}$$

The constrained empirical Bayes estimates are then given by

$$\hat{\theta}^{CEB}_i = \hat{\theta}^{EB}_\cdot + (\hat{\theta}^{EB}_i - \hat{\theta}^{EB}_\cdot)F.$$

Note that since the factor F is greater than one, the constrained empirical Bayes estimate for a particular state will be further distant from the average of the empirical Bayes estimates than the empirical Bayes estimate for that same state.

It may be demonstrated that for equal sample sizes in the states, $F \approx \sqrt{\gamma_i^{-1}}$ for all i and from this that Lahiri's and Spjøtvoll and Thomsen's estimators are very similar for this special case.

Simulation Studies

Although our interest focused on the bias in the range estimator, our discussions above mainly rely on the results on the dispersion estimators. We also pointed out that it is not very clear how Lahiri's "posterior linearity" assumption can be verified in practical situations. To gather some empirical experience about the performance of the estimators we have discussed, a simulation study was carried out to

- Examine the "Posterior linearity" assumption;
- Verify conclusions on the Range estimator;
- Compare performance of competing estimators.

We used a truncated-normal model to generate true state means for our simulation. We knew that the performance of the range estimators would depend upon the state sample sizes and on the intrastate correlation. In order to simplify the presentation of the results, we restricted our attention to a scenario where every state has the same sample size. Across the two-dimensional space defined by sample size

and intrastate correlation, we reduced the size of the simulation study by examining behavior of the estimators on two lines, the first defined by a fixed sample size of 30 observations per state, the second by a fixed intrastate correlation of 0.005. These values were chosen as reasonable and instructive. A total of 500 superpopulations were examined along each line.

For each superpopulation model, we generated 100 populations and then drew 200 samples from each population for a total of 20,000 samples per superpopulation.

The results of holding the state sample size fixed and varying the intrastate correlation are shown on the left-hand side of Exhibit 1. As one can see from the relative bias in the estimated range, the design based estimator of the range was always biased upward, and the standard empirical Bayes estimator (EB) estimator of the range was always biased downward. The constrained empirical Bayes estimator (CEB) of Lahiri and the Spjøtvoll -Thomsen estimator (ST) did well until ρ was very small. When ρ was very small, the relative biases and the relative root mean squared errors of all estimators under consideration became unacceptably large. This suggests that a state sample size of 30 is too small unless there is good prior evidence that the intrastate correlation is at least 0.015, a fairly large value. Another interesting observation is that CEB and ST were almost indistinguishable for most of the range of ρ, although ST was much simpler and easier to compute. When ρ was small, ST, which became more heavily influenced by the overall mean estimator $\bar{y}$ in its formula, behaved more like the design based estimator DB.

The results of holding the intrastate correlation fixed at 0.005 and varying the state sample size are shown on the right-hand side of Exhibit 1. As expected, the bias of the estimators became larger as the sample size shrank. CEB and ST were clearly less biased than the design-based and empirical Bayes estimators.

Conclusions

The range across the states of design-based estimates can be seriously positively biased when either the state sample sizes are small or the intrastate correlation is small. Such biases can affect policy discussions. Constrained empirical Bayes estimators have been proposed that can substantially reduce the bias. In the case of the original CPS problem described in the introduction, application of the CEB method reduced the range in the state response biases from 41 points to 21 points, clearly a critical adjustment for discussion of state allocations based on CPS statistics. Incorporating Ghosh's (1992) adjustments to PQL estimators (Breslow and Clayton, 1993) and to Gibbs sampling (Zeger and Karim, 1991) looks like a promising avenue for further research.

References

Breslow, N. E. and Clayton, D. C. (1993). "Approximate Inference in Generalized Linear Mixed Models." Journal of the America Statistical Association, 88, 9-25.

Ghosh, M. (1992). "Constrained Bayes Estimation with Applications." Journal of the America Statistical Association, 87, pp. 533-540.

Ghosh, M. and Rao, J. N. K. (1994). Small Area Estimation: An Appraisal." Statistical Science, Vol. 9, pp55-93.

Lahiri, P. (1990). "Adjusted Bayes and Empirical Bayes Estimation in Finite Population Sampling." Sankhyã: The Indian Journal of Statistics, Volume 52, Series B, Pt. 1, pp. 50-66.

Louis, T. A. (1984). "Estimating a Population of Parameter Values Using Bayes and Empirical Bayes Methods." Journal of the America Statistical Association, 79, pp. 393-398.

McCulloch, C. E. (1997). "Maximum Likelihood Algorithms for Generalized Linear Models." Journal of the America Statistical Association, 93, pp. 162-170.

Spjøtvoll, E. and Thomsen, I. (1987). "Applications of some Empirical Bayes Methods to Small Area Statistics." Bulletin of the International Statistical Institute, V 2, pp. 435-449.

Zeger, S. L. and Karim, M. R. (1991). "Generalized Linear Models with Random Effects: A Gibbs Sampling Approach." Journal of the America Statistical Association, 86, pp79-86.

Exhibit 1. Relative Bias and Relative Root MSE of Range Estimators as Functions of Intrastate Correlation ρ and State Sample Size n_i

(Range of the 50 state estimates of a binary outcome. Bias are truncated at -60% and root MSEs are truncated at 100%)