Methodology Evaluation of a Survey of High School Students in Iowa

Variance Estimation in a One-Per-Stratum Design

Lu Lu and Michael D. Larsen

Center for Survey Statistics and Methodology
Department of Statistics at Iowa State University

June 19th, 2007

Outline

1. Iowa's State Board of Education (ISBE) employment preparation (EP) survey
2. Variance estimation for individual strata in a one-per-stratum design
    - Collapsing strata followed by synthetic variance redistribution (CSSV)
    - (Restricted) generalized variance functions ((R)GVF)
3. Simulation studies
    - A limited example
    - The ISBE EP survey

The ISBE EP Survey

- The purpose of the survey is to study the availability of EP courses and the degree to which students in Iowa's public high schools enroll in those courses
- Estimators are designed for the average numbers of EP courses taken by public high school students for the State of Iowa and populations of small, medium, and large school districts

A Stratified Multi-Stage Design

- Stratification was by district size (small, medium, large) and 12 Area Education Agencies (AEAs)
- Large districts were included with certainty; medium and small districts were sampled by probability proportional to size sampling without replacement
- All schools in selected districts were included
- Students from ninth or twelfth grade and general or special educational groups were selected by simple random sampling

Establishment Characteristics

- Administrative data at school or establishment level
- Geographical and size-related variability
- Size distributions are highly skewed
- Potentially similar to businesses or hospitals by county

One Sample Unit In Some Strata

- Large districts
  - all schools are sampled
- Medium districts
  - 7 AEAs have 2 schools sampled
  - 5 AEAs have **1 school** sampled
- Small districts
  - 7 AEAs have 2 schools sampled
  - 5 AEAs have **1 school** sampled

Ratio Estimator

$$\hat{t}_{st,ra} = \hat{t}_{st,\pi} \frac{N_{st}}{\hat{N}_{st,\pi}}$$

- t_{st} = No. of EP classes taken in a stratum
- Aggregate $\hat{t}_{st,ra}$ for size, AEA and state estimates

One Per Stratum And Variance Estimation

- Issue: with one PSU per stratum (small or medium districts within AEAs), we cannot directly estimate variance at the stratum level
- Strategies:
  1. Collapse and redistribute
  2. Generalized variance functions

Collapsing Strata Synthetic Variance Estimation of Stratum Variance

- Arrange strata in a non-increasing sequence based on total enrollment size and then collapse strata with one PSU per stratum into pairs or groups sequentially
- Estimate variance of a group consisting of L_g strata by

$$\hat{v}\left(\hat{t}_{coll}^{(g)}\right) = \frac{L_g}{L_g - 1} \sum_{k=1}^{L_g} (\hat{t}_k^{(g)} - \frac{\sum_{k=1}^{L_g} \hat{t}_k^{(g)}}{L_g})^2$$

- Assume that strata in the same group are homogeneous in terms of within strata variation
- The ratio of variances of two strata within a group is approximately the ratio of squared total enrollment sizes
- Variance of a stratum could be obtained through redistributing the group variance proportional to squared total enrollment size

Generalized Variance Function Estimation of Stratum Variance

- Model the relationship between relative variances and expectations of the total estimators for individual strata
- Predict the variance in a stratum from the estimated total through the estimated function

- A traditional GVF model (Valliant, 1987)
  - $V_T^2 = \alpha + \frac{\beta}{T}$
  - could produce negative predictions of variance
- A restricted generalized variance function (Wolter, 1985):

$$V_T^2 = \beta \left(\frac{1}{T} - \frac{1}{N} \right)$$

- The unknown parameter β can be estimated using iteratively reweighted least squares estimation or maximum likelihood estimation algorithms

GVF Procedure for Medium or Small Strata

1. Estimate totals in all strata
2. Estimate variances in strata with two PSUs
3. Fit the RGVF to the variance and total estimates from strata with two PSUs
4. Predict variances based on estimated totals for strata with one PSU sampled

Simulation

- Population Setup:

$$
\begin{aligned}
y_{h,ij} &\sim \text{Poisson}\left(\lambda_{h,i}\right) \\
\lambda_{h,i} &= 0.1h + \tau_{h,i} \\
\tau_{h,i} &\sim \text{Uniform}\left(5, 10\right)
\end{aligned}
$$

  - Strata: $h = 1, \cdots, H = 50$
  - Clusters: $i = 1, \cdots, I = 20$
  - Units: $j = 1, \cdots, N_{h,i}$, $N_{h,i} \sim \text{Uniform}\left(30, 80\right)$

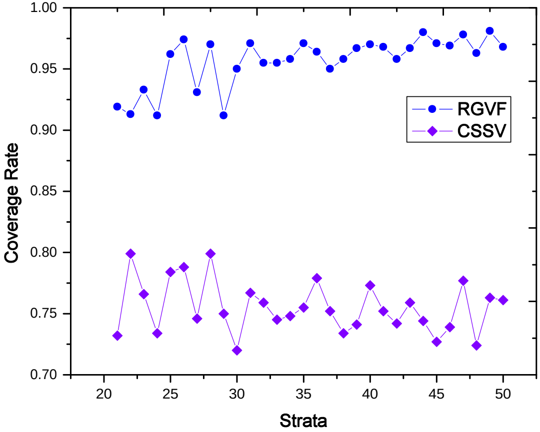
- Sampling designs compared:
  - $m = 10, 20, 30$ strata with two PSUs and $50 - m = 40, 30, 20$ strata with one PSU
  - PSUs were selected by SRS or PPS
  - $n_{h,i} = 5$ elementary units were sampled by SRS
- Variance estimation
  - CSSV
  - Restricted GVF

- Results from the Monte Carlo study for designs with $m = 20$ strata have two PSUs sampled per stratum
- The ratio of variance estimate relative to true variance

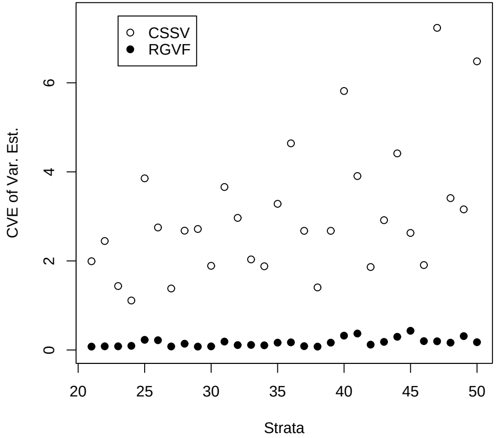
	CSSV	RGVF
PPS	1.10	1.04
SRS	1.27	1.15

- RGVF produces smaller variance estimates than CSSV for a group of strata

- The coverage rate of confidence intervals under PPS design



- The coefficient of variation of variance estimates under PPS design



- CSSV could overestimate the variance on a large scale with a substantial probability
- RGVF outperforms CSSV in terms of
  - Smaller variance estimates for a group of strata
  - A higher coverage rate of confidence intervals
  - Consistently smaller coefficients of variation of variance estimates for individual strata
- Increasing the degrees of freedom for fitting the RGVF model does improve the predictions of variance in terms of a higher coverage rate of confidence intervals and more stable performance

Target Population

- The population database of EP courses taken by twelfth grade students in Iowa's public high schools was created through simulation
- The numbers of EP courses taken by students in a school were generated as independent Poisson random variables with a rate for the school
- The Poisson rates were generated independently from a random effects model with main effects due to school size and AEA

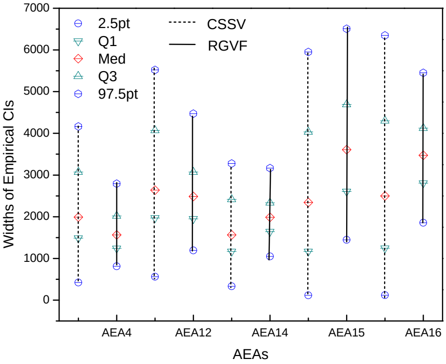
- The coefficients of variation of total estimates

Aggregation	CSSV	RGVF
State	3.35	3.25
Medium	6.05	5.89
Small	5.20	4.91

- Number of confidence intervals covering true totals for strata of medium districts with one PSU sampled

Variance	Area Education Agencies				
Method	4	12	14	15	16
CSSV	983	883	968	751	838
RGVF	996	939	1000	834	991

- Empirical percentiles of the widths of confidence intervals for strata of medium districts with one PSU sampled: RGVF is less variable



- The RGVF method outperforms the CSSV method in terms of producing
    - smaller coefficients of variation of total estimates for a group of strata
    - a higher coverage rate of confidence intervals and consistently more stable performance for individual strata and the group as a whole

Summary and Discussion

- The ISBE EP survey motivated the examination of variance estimation methods for designs with one-per-stratum selection of PSUs
- Traditional collapsing strata estimator is widely applied for estimating the variance of a total for a group of strata
- When a variance estimate is needed for an individual stratum, using a generalized variance function and choosing a reasonable estimate based on some model diagnostics might be helpful
- Negative predictions could be prevented by adding some restrictions to a generalized variance function

- Our simulation studies indicate that a restricted GVF estimator could improve a CSSV estimator by producing consistently smaller coefficients of variation of total estimates for a group of strata, a higher coverage rate of confidence intervals and more stability of performance for individual strata and higher levels of aggregations
- Future study will be focused on small area estimation using hierarchical Bayesian predictive methods and making use of auxiliary information to improve estimation efficiency

References

1. COCHRAN, W.G. (1977). *Sampling Techniques*, third edition. New York: J. Wiley
2. HANSEN, M.H., HURMITZ W.N., and MADOW W.G. (1953). *Sampling survey methods and theory*, Vol I, 399-401, and Vol II, 218-222. New York: Wiley
3. VALLIANT, R. (1987). Generalized variance functions in stratified two-stage sampling, *Journal of the American Statistical Association*, 82, 499-508
4. WOLTER, K.M. (1985). *Introduction to Variance Estimation.* Springer Series in Statistics

Acknowledgments

- Thanks to ISBE for partial support of this work and bringing this survey to CSSM at ISU
- The authors wish to acknowledge the involvement of Barb Guy, Pat Sitlington, and Alan Frank from the offices of State Board and Dianne Anderson and Jan Larson from SRS in CSSM at ISU