Estimating Variance Components for a Two-Stage Design with Second-Stage Strata Nested within PSUs

Jun Liu, Vincent Iannacchione, Jill Kavee, Research Triangle Institute
Jun Liu, RTI, P.O.Box 12194, RTP, NC, 27709, junliu@rti.org

Key Words: Stratified Two-stage PPS Design, Variance Components, Cost/Variance Optimal Sample Allocation.

ABSTRACT

The sampling design for the 1998 DoD Survey of Health Related Behaviors Among Military Personnel is a two-stage design where primary sampling units (PSUs) are selected with probabilities proportional to size (PPS) and the second-stage strata are nested within the PSUs. We derived formulas for estimating the design-consistent variance components associated with this design and then used data from the 1995 survey to estimate between-PSU and within-PSU variance components. In this paper, we discuss the utility of the formulas for use in a cost/variance constrained optimal sample allocation.

1. INTRODUCTION

The two-stage sample design where the first-stage sampling units (PSUs) are selected with probability proportional to size (PPS) without replacement and the second-stage strata are nested within PSUs is a commonly used sample design in large scale surveys. In most of the situations when the first-stage sampling fraction is small, we can assume a PPS sample design with replacement and calculate the variances for the estimates accordingly. For that reason, variance formulas for multi-stage PPS sample designs that are found in many standard text books, for example, Hansen et al (1953) and Cochran (1977), typically ignore stratification at the second-stage. However, variance structure and the decomposition of the variance components and their estimation can be of interest by themselves. In this paper, we derive the formulas for the variance components and discuss how they can be estimated. We then apply the results in a sample allocation problem.

2. VARIANCE DECOMPOSITION

We consider stratified two-stage sample designs where the second-stage strata are nested within the first-stage units (PSUs). The first-stage sampling frame is stratified into H first-stage strata, indexed by h. The SSUs (second-stage units) are stratified into J second-stage strata, indexed by j. The PSUs are selected with probability proportional to size (PPS); a random sample of SSUs is selected independently within each second-stage stratum within each PSU. Because the sampling method of the second-stage units does not affect the variance formula, we will present the result with general designs.

When the total number of second-stage units M_d are known for the d-th domain, p_d, the proportion of a certain attribute of the domain d population can be estimated using the following linear estimator,

$$\hat{p}_d = \bar{y}_d = \frac{1}{M_d}\hat{y}_d = \frac{1}{M_d}\sum_{h=1}^{H}\hat{y}_{dh} \tag{1}$$

where $\hat{y}_{dh}$ is the Horvitz-Thompson estimator of the total in the d-th domain D_d and h^{th} first-stage stratum, given by

$$\hat{y}_{dh} = \sum_{i=1}^{n_h}\frac{\hat{y}_{dhi}}{\pi_{hi}} = \frac{1}{n_h}\sum_{i=1}^{n_h}\frac{\hat{y}_{dhi}}{z_{hi}}. \tag{2}$$

Here, π_{hi} is the inclusion probability for the i^{th} PSU in the first-stage stratum h. The single-draw selection probability for the same PSU is z_{hi}. The domain total for the i^{th} PSU in the h^{th} first-stage stratum can be estimated as

$$\hat{y}_{dhi} = \sum_{j\in D_d} M_{hij}\bar{y}_{hij} = \sum_{j\in D_d}\frac{M_{hij}}{m_{hij}}\sum_{k=1}^{m_{hij}} y_{hijk} \tag{3}$$

where,

m_{hij} is being the sample size in the j^{th} second-stage stratum within the i^{th} PSU of the h^{th} first-stage stratum, and

M_{hij} is the population total for the j^{th} second-stage stratum within the i^{th} PSU of the h^{th} first-stage stratum.

In the above, we also define

$$M_{dhi} = \sum_{j\in D_d} M_{hij},\ M_{dh} = \sum_{i=1}^{N_h} M_{dhi},\ \text{and, } M_d = \sum_{h=1}^{H} M_{dh}.$$

It can be proven that the variance of the estimated proportion from the domain d, p_d, can be written as

$$r(\bar{y}_d) = \frac{1}{M_d^2}\sum_{h=1}^{H}\frac{1}{n_h}\left\{\sum_{i=1}^{N_h} z_{hi}\left(\frac{Y_{dhi}}{z_{hi}} - Y_{dh}\right)^2 + \sum_{j\in D_d}\sum_{i=1}^{N_h}\frac{Var(\hat{y}_{hij})}{z_{hi}}\right.$$

$$= \frac{1}{M_d^2}\sum_{h=1}^{H}\left\{\sum_{i=1}^{N_h} z_{hi}\left(\frac{Y_{dhi}}{z_{hi}} - Y_{dh}\right)^2\right\}$$
$$+ \frac{1}{M_d^2}\sum_{h=1}^{H}\frac{1}{n_h}\left\{\sum_{j\in D_d}\sum_{i=1}^{N_h}\frac{Var(\hat{y}_{hij})}{z_{hi}}\right\}$$

$$= Var_{PSU}(\bar{y}_d) + Var_{SSU}(\bar{y}_d). \tag{4}$$

If the SSUs are drawn by stratified simple random sampling, then

$$Var_{SSU}(\bar{y}_d) = \frac{1}{M_d^2}\sum_{h=1}^{H}\frac{1}{N_h}\left\{\sum_{j\in D_d}\sum_{i=1}^{N_h}\frac{M_{hij}^2(1-f_{hij})}{z_{hi}}\frac{S_{hij}^2}{m_{hij}}\right\}$$
$$= \frac{1}{M_d^2}\sum_{h=1}^{H}\sum_{j\in D_d}\sum_{i=1}^{N_h}\frac{M_{hij}^2(1-f_{hij})S_{hij}^2}{\pi_{hi}m_{hij}}.$$

Since the sample size for the j^{th} second-stage stratum, within the i^{th} PSU and the h^{th} first-stage stratum is given by

$$m_{hij} = \frac{f_{hj}M_{hij}}{\pi_{hi}} = \frac{m_{hj}M_{hij}}{M_{hj}\pi_{hi}},$$

we have

$$Var_{SSU}(\bar{y}_d) = \frac{1}{M_d^2}\sum_{h=1}^{H}\sum_{j\in D_d}\sum_{i=1}^{N}\frac{M_{hik}M_{hj}(1-f_{hij})S_{hij}^2}{m_{hj}}. \tag{5}$$

Here,

- S_{hij}^2 is the population variance of the j^{th} second-stage stratum within the i^{th} PSU of the h^{th} first-stage stratum;
- m_{hj} is the number of sampled individuals in the j^{th} second-stage stratum within the h^{th} first-stage stratum;
- M_{hij} is the total number of individuals in the j^{th} second-stage stratum within the i^{th} PSU of the h^{th} first-stage stratum;
- M_{hj} is the total number of individuals in the j^{th} second-stage stratum within the h^{th} first-stage stratum; and
- M_d is the population size of the domain d.

3. ESTIMATING VARIANCE COMPONENTS

To facilitate the estimation of the variance components, we recast (5) in the following form:

$$\hat{V}ar(\bar{y}d) = \frac{1}{M_d^2}\sum_{h}^{H}\left\{\frac{\hat{\sigma}_{b,dh}^2}{n_h} + \sum_{k\in D_d}\frac{\hat{\sigma}_{w,dhk}^2}{m_{hk}}\right\},$$

then

$$\hat{\sigma}_{w,dhk}^2 = \sum_{i=1}^{n_h}\frac{M_{hk}M_{hik}}{\pi_{hi}}(1-f_{hik})s_{2hik}^2,$$

where

$$(1-f_{hik})s_{2hik}^2 = \frac{(M_{hik} - M_{hik})}{M_{hik}}\frac{M_{hik}}{(M_{hik}-1)}\hat{p}_{hik}\hat{q}_{hik}$$

and

$$\hat{\sigma}_{b,dh}^2 = \frac{1}{(N_h-1)}\sum_{i}^{n_h}\left(\frac{\hat{y}_{dki}}{z_{hi}} - \hat{Y}_{dh}\right)^2 - \sum_{k\in D_d}\sum_{i}^{n_h}\frac{M_{hk}M_{hik}}{z_{hi}}\frac{(1-f_{hik})s_{2hik}^2}{M_{hk}}.$$

If we write

$$\hat{V}ar(\bar{y}d) = \sum_{h}^{H}\left\{\frac{\hat{\sigma}_{PSU,dh}}{n_h} + \frac{\hat{\sigma}_{SSU,dh}}{N_h\bar{m}_n}\right\}$$

then,

$$\hat{\sigma}_{SSU,dh}^2 = \sum_{k\in D_d}\sum_{i}^{n_h}\left(\frac{M_{hk}M_{hik}}{M_d^2}\right)\frac{(1-f_{hik})s_{2hik}^2}{\pi_{hi}\theta_{rk}} = \frac{1}{M_d^2}\sum_{k\in D_d}\sum_{i}^{n_h}\frac{(1-f_{hik})s_{2hik}^2}{\pi_{hi}\theta_{rk}},$$

and

$$\sum_{k}\theta_{rk} = 1.$$

4. APPLICATION IN SAMPLE ALLOCATION

The sample allocation problem can be stated in terms of determining the number of installations and active-duty members to include in the sample such that the precision requirements set for the survey are met for the least cost. That is, the sample sizes determined by the sampling design are a balance between satisfying analytical requirements of the survey and the fiscal constraints imposed on the survey.

The sample design for 1998 DoD Survey of Health Related Behaviors Among Military Personnel (Iannacchione, at el. 1998) is a stratified two-stage design with the second-stage stratification nested within the first-stage units (PSUs). The first-stage sampling frame was stratified into eight first-stage strata, indexed by h. The SSUs (second-stage units) were stratified into 12 second-stage strata, indexed by j. The PSUs were selected with probability proportional to size (PPS); a simple random sample (SRS) of SSUs was selected independently within each second-stage stratum within each PSU.

When the total number of active-duty members M_d are known for the d-th domain, p_d, the proportion of a certain attribute of the domain d population can be estimated using the following linear estimator ,

$$\hat{p}_d = \bar{y}_d = \frac{1}{M_d}\hat{y}_d = \frac{1}{M_d}\sum_{h=1}^{8}\hat{y}_{dh}$$

where $\hat{y}_{dh}$ is the Horvitz-Thompson estimator of the total in the d-th domain and h^{th} first-stage stratum, given by

$$M_{dhi} = \sum_{j\in D_d} M_{hij},\ M_{dh} = \sum_{i=1}^{N_h} M_{dhi},\ \text{and, } M_d = \sum_{h=1}^{H} M_{dh}\,.$$

We set up a nonlinear optimization problem using the Kuhn-Tucker conditions (Chong and Zak, 1996) to search for the optimal sample size and allocation. For a design like the 1998 DoD Survey, the variance of the estimated proportion from domain d can be expressed as in (4).

As one can see, the variance formula depends on the first- and second-stage sample size, n_h and m_{hj}, respectively. We can also formulate the cost function for the survey in terms of n_h and m_{hj} as well:

$$C = C_0 + \sum_{h=1}^{8}\left\{c_{1h}n_h + \sum_{j=1}^{12} c_{2hj}m_{hj}\right\} \qquad (6)$$

where C_0 is the fixed cost and is assumed zero for the optimization purpose. Parameters c_{1h} and c_{2hk} are the variable cost associated with adding an additional PSU and SSU, respectively.

If we denote the precision requirement for the sample proportion from the d^{th} domain as V_d, the sample allocation problem can then be formulated as minimizing the cost function (4) subject to the following constraints:

$$\text{Var}(\hat{p}_d) \le V_d, \quad d=1,2,...15, \qquad (7)$$

and,

$$n_h \ge 0,\quad m_{hj} \ge 0,\quad \text{for } h=1,2,...,8, \text{ and } j = 1,2,...,12. \qquad (8)$$

The variance constraints are given in the form of the variance components of (4). The variance components were estimated from data collected in the 1995 DoD Survey. To provide stable estimates, three groups of outcomes were used in the estimation (***Table 2***). The variance components used in the variance constraints were calculated by averaging the estimated variance components of the outcome categories within each outcome group. Negative estimates were converted to zero. The domains on which constraints were imposed are given in ***Table 3***. The variance components estimated using the 1995 allocation and the 1998 allocation are also compared in this exhibit.

In addition to the constraints in (4) and (5), we imposed the practical limitations that are listed in ***Table 4***. For example, we set an upper limit on the number of SSUs (active-duty members) to be selected from an installation so that the group sessions would not become unmanageable. The realized sample allocation from the constrained optimization is given in ***Table 5***.

5. CONCLUSIONS

A design consistent variance component formula is derived. Its utility is demonstrated through a sample allocation problem. Other areas of application include assessing design effects, etc.. More research is planned to study the new variance formula to have a better understanding of the design consistent formula and the formula that assumes simple random sampling at both stages.

Table 2. Outcome Groups Used in the Calculation of Variance Constraints for the Sample Allocation

Outcome Group	Outcome Category
Drug Use	Marijuana Use
	Any Drug Except Marijuana
	Any Drug Use
Tobacco Use	Any Smoking in Past 30 Days
	Heavy Smoking in Past 30 Days
	Smokeless Tobacco Use (Males Only)
	Percent Attempted to Quit Smoking
Alcohol Use	Percent of Abstainers
	Percent of Infrequent to Light Drinkers
	Percent of Moderate Drinkers
	Percent of Moderate to Heavy Drinkers
	Percent of Any Drinking Versus Abstainers
	Percent With Serious Consequences Due to Alcohol
	Percent With Productivity Loss Due to Alcohol
	Percent With Alcohol Dependence Symptoms

Table 3. Design Constraints used in the Allocation

Design Constraints			Target	Achieved
Constraints on the Number of PSUs				
Min # of PSUs per Stratum >=			2	2.0
Total # of PSUs <=			65	58.5
Max # of PSUs per Service <=			18	15.8
Max # of PSUs for Army OCONUS <=			6	6.0
Max # of PSUs for Navy OCONUS <=			6	6.0
Max # of PSUs fpr Marine OCONUS <=			2	2.0
Max # of PSUs fpr Air Force OCONUS <=			4	4.0
Min # of PSUs per Service >=			12	13.5
Constraints on the Number of SSUs				
Max Total SSUs <=			18,000	18,000.0
Min SSUs per Cell >=	Male		2	12.5
	Female		1	1.7
Max SSUs per Cell <=	Male		1,300	1,017.8
	Female		300	300.0
Min # of DoD female SSUs >=			4000	4000.0
Min # of SSUs per PSU >=			250	275.0
Max # of SSU per PSU <=	Army	CONUS	300	300.0
		OCONUS	350	350.0
	Navy	CONUS	300	275.0
		OCONUS	350	350.0
	Marine	CONUS	300	281.1
		OCONUS	350	350.0
	Air Force	CONUS	300	300.0
		OCONUS	350	350.0

6. REFERENCES

Chong, E.K.P., and S.H. Zak (1996). *An Introduction to Optimization.* John Wiley & Sons, New York.

Cochran, W.G. (1977). *Sampling Techniques.* John Wiley & Sons, New York.

Hansen, W.H., Hurwitz, W.W., and Madow, W.G. (1953). *Sampling Survey Methods and Theory.* John Wiley and Son, New York.

Iannacchione, V.G., Liu, J., Kavee, J.D., Crump, C.J. (1998). 1998 DoD Survey of Health Related Behaviors among Miliary Personnel, Prepared for the Office of the Assistant Secretary of Defense (Health Affairs). RTI Report 7034-01-FR.

Table 3. Variance Constraints Used in the Sample Allocation

		Alcohol			Drug			Smoking		
		WWD95	WWD98	Reduction	WWD95	WWD98	Reduction	WWD95	WWD98	Reduction
Service	Army	8.57	6.77	21.03%	10.74	8.76	18.40%	8.25	6.63	19.56%
	Navy	10.38	9.98	3.80%	6.89	6.50	5.68%	11.80	11.40	3.38%
	Marine Corps	10.34	9.13	11.74%	11.45	10.02	12.51%	9.37	8.27	11.74%
	Air Force	8.27	7.59	8.24%	4.98	4.65	6.66%	8.39	7.73	7.80%
Rank	E1-E3	5.78	4.85	16.10%				5.68	4.65	18.14%
	E4-E6	5.23	4.69	10.34%				5.45	4.99	8.42%
	E7-E9	5.83	5.33	8.61%				6.87	6.22	9.42%
	W1-W5	25.23	21.15	16.19%				10.74	9.15	14.86%
	O1-O3	12.74	9.46	25.76%	7.25	5.03	30.55%	11.55	8.77	24.05%
	O4-O10	18.17	13.80	24.05%	6.04	5.63	6.77%	10.55	8.74	17.10%
Service X Gender	DoD, Male	4.81	4.28	10.88%				4.64	4.19	9.66%
	Army, Female	12.16	8.14	33.10%				16.55	10.77	34.92%
	Navy, Female	13.97	11.93	14.59%				32.12	27.37	14.77%
	Marine,	15.55	12.04	22.58%				22.57	17.47	22.56%
	Air Force,	19.31	16.13	16.49%				17.13	14.16	17.34%

Table 5. Rounded Sample Allocation for the First- and Second-Stage Sample Size

		Army		Navy		Marine Corps		Air Force		DoD	
		CONUS	OCONUS	CONUS	OCON/Afl*	CONUS	OCONUS	CONUS	OCONUS		
PSUs per Cost Stratum		10	6	10	6	12	2	10	4		60
Males	E1 - E3	300	246	272	265	879	209	295	147		
	E4 - E6	616	501	625	608	1018	239	1001	499		
	E7 - E9	588	472	508	485	275	65	512	255		
	W1 - W5	168	143	39	37	100	13				
	O1 - O3	177	145	194	192	177	42	228	113		
	O4 - O10	282	230	194	166	184	40	189	96		
Females	E1 - E3	214	68	200	113	157	32	192	81		
	E4 - E6	266	145	256	154	288	53	300	143		
	E7 - E9	123	90	91	52	67	5	94	34		
	W1 - W5	19	8	10	2	24	2				
	O1 - O3	101	30	100	21	37	4	91	21		
	O4 - O10	80	30	80	11	24	3	89	16		
Summary											
PSUs / SSUs per Service		16	5,042	16	4,675	14	3,937	14	4,396	60	18,050
Total SSUs per Stratum		2,934	2,108	2,569	2,106	3,230	707	2,991	1,405		18,050
Average SSUs per PSU		293	351	257	351	269	354	299	351		316
Total Females per Stratum		803	371	737	353	597	99	766	295		4,021
Total Males per Stratum		2,131	1,737	1,832	1,753	2,633	608	2,225	1,110		14,029
Females / Males per		1,174	3,868	1,090	3,585	696	3,241	1,061	3,335	4,021	14,029
Percent of Females / Males		23.3%	76.7%	23.3%	76.7%	17.7%	82.3%	24.1%	75.9%	22.3%	77.7%
Total Officers / Enlisted		1,413	3,629	1,046	3,629	650	3,287	843	3,553	3,952	14,098
Percent of Officer / Enlisted		28.0%	72.0%	22.4%	77.6%	16.5%	83.5%	19.2%	80.8%	21.9%	78.1%

* OCONUS and Afloat Personnel