COMPARISON OF VARIANCE ESTIMATION TECHNIQUES FOR A COMPLEX SAMPLE SURVEY

Donna Brogan and Beverly Martinez, Emory University

OBJECTIVE AND MAIN RESULT

This paper empirically compares two methods for estimating variances of point estimates in complex sample surveys, the Taylor Series expansion and the independent random groups technique (sometimes called the replicate technique). The use of three independent replicates is compared to the second order Taylor Series expansion. Comparison of these two methods gives approximately equivalent estimates of the variance 27% of the time. We conjecture the agreement to be low because of the low degrees of freedom for the replicates, i.e. 2 df, even though the sample size in each replicate is moderately large.

DESCRIPTION OF COMPLEX SAMPLE SURVEY

A complex sample survey of housing units (HU's) in six counties in south Georgia was conducted in 1981 to assess the prevalence of hypertension and its treatment status. The sample was designed originally as four independent replicates (random groups) of 240 HU's each, with each replicate balanced across the six counties, which served as six strata. First stage sampling within each county was at the enumeration district (ED) level, with second stage sampling of each ED resulting in a segment of about 24 HU's. About 12 HU's per segment were selected for the sample. An interviewer visited each selected sample HU, enumerated all persons living there, and asked all enumerated adults 18 years or older to participate in a personal health interview, including measurement of blood pressure, height and weight. Interviewers were trained for three days on blood pressure measurement and interviewing techniques and were supervised by one person. Data editing, coding, entry, management and analysis were done at a central statistical office.

Field work was completed for all segments in three of the four replicates, providing only three complete replicates to use in the replicate technique. In the fourth replicate some segments were completed, some partially completed, and some never assigned to interviewers. Considering all completed segments from all four replicates, the enumeration stage response rate was 92.1% and the interview stage response rate was 92.9%, for an overall response rate of 85.6% (.921 x .929). The sample of all completed segments was weighted according to probability of selection, adjusted for undercoverage in listing, and adjusted for nonresponse at the HU and individual level by a series of steps, including poststratification to the 1980 Census of Population and Housing. Further information on sample design, field work, weighting procedures and results of the survey is in Brogan et al (1983).

METHODOLOGY

Taylor Series

The Taylor Series technique for variance estimation used subjects from the three completed replicates plus subjects from the completed segments in the fourth replicate, all appropriately weighted, adjusted and poststratified so as to make inference to all adults in the six county area. Poststratification was done separately to a four-county area (the experimental counties) and to a two-county area (the control counties) because of a planned longitudinal study of the effectiveness of a community intervention in the experimental counties. Each point estimate, its standard error and design effect (DEFF) were obtained from the software program SESUDAAN, developed at Research Triangle Institute (RTI). All stages of sampling below the ED level were ignored. The sample size for this method was 653 adults for the four-county area and 604 adults for the two-county area, for a total sample size of 1257.

Replicate Technique (Independent Random Groups)

The replicate technique used only the three completed replicates, for a somewhat smaller total sample size of 1144 adults. Sample size for replicates 1, 2, and 3 in experimental and control counties, respectively, was (191, 196, 208) and (193, 164, 202). Each replicate was weighted appropriately so as to make inference to adults in the six-county area. Poststratification to a limited number (8) of age-race-sex cells was carried out separately for each replicate since one replicate seemed to have somewhat higher underlisting and nonresponse. The 24 poststratification weights (8 for each replicate) for the four-county area ranged from 1.025 to 2.143. The 24 poststratification weights for the two-county area ranged from 0.76 to 3.03.

Population Parameters

We selected five population parameters to estimate for the purpose of comparing the two techniques, three proportions and two means. The proportions are P_1, the proportion of adults who are normotensive[1]; P_2, the proportion who are hypertensive and on antihypertensive medication and controlled[1]; and P_3, the proportion who have an elevated blood pressure.[1] These three proportions sum to one. The two population means are diastolic blood pressure (DBP) and systolic blood pressure (SBP), estimated separately for those on and not on antihypertensive medication. We estimated each parameter for the following domains: all adults, whites, blacks, men, women, white men, white women, black men and black women.

Comparison Criterion

It is not fair to compare the two variance estimation techniques to each other via the two standard errors because the Taylor Series technique had a slightly larger total sample size than the replicate technique. Thus, we calculated the design effect $(DEFF)^2$ for each technique and then compared the two techniques by taking the square root of the ratio of the replicate DEFF to the Taylor Series DEFF. This comparison index is denoted by I in Tables 2 through 4.

RESULTS

Tables 2 to 4 give the results of many of the calculations leading to the index I. For example, in line 1 of Table 2, for the estimation of the percentage of adults in the four experimental counties who are normotensive, the three separate replicates yield point estimates of 73.8, 75.7 and 78.2%. Using all three replicates pooled

together, the point estimate is 75.9%, based on 585 subjects. Obtaining the sample variance of the three point estimates with 2 df, and then dividing by 3 (number of replicates), the standard error of the point estimate 75.9% is estimated as 1.27%. DEFF in this instance is estimated to be 0.52. Using the Taylor Series technique the point estimate of the parameter is 75.1%, based on 653 subjects. The standard error is estimated as 2.13%, with a DEFF of 1.60. The index I, square root of the ratio of the two DEFF's, is 0.57.

For each parameter in the three tables, the weighted point estimates obtained from the two approaches are similar to each other, as expected. However, the estimates of the variance are quite different. The two DEFF columns in each table indicate that results from the replicate technique are more variable across the parameters. The Taylor Series technique yields DEFF values which cluster around 1.00, ranging from 0.52 to 2.51, whereas the replicate technique yields DEFF values as low as 0.002 and as high as 5.48. The low DEFF values of the replicate technique are obtained when the estimates from the three independent replicates are quite close to each other, while the high DEFF values are due to the high variability among the estimates of the same parameter from the three replicates.

For our purposes, we consider an index between 0.80 and 1.20 as indicative of moderate agreement between the two estimates of variability. Seventeen out of 63 (27%) of the comparison indices are within this range. For estimating proportions, 9 out of 27 comparison indices (33%) are within moderate agreement.

This empirical study does not indicate in general which technique gives a higher estimate of variance. In Table 2 for proportions, the Taylor Series technique gives a higher DEFF in 19 out of 27 (or 70%) instances. For means, the replicate technique gives higher estimates of DEFF 61% of the time in Table 3 but only 11% of the time in Table 4. In any case, the number of disagreements is too many to be ignored.

CONCLUSION

Based on this empirical study, we conclude that the Taylor Series technique is better than the independent replicate technique when the latter is based only on three replicates.

Footnotes

[1] Elevated blood pressure is defined as Diastolic Blood Pressure greater than or equal to 90 mm Hg. Normotensive is nonelevated blood pressure and not on antihypertensive medication. Controlled is on antihypertensive medication and blood pressure not elevated. Needing treatment also means elevated.

[2] DEFF is the ratio of the estimated variance using the complex survey design to the estimated variance assuming that the obtained sample resulted from simple random sampling of adults in Georgia.

References

1. Brogan, Donna, Beverly Martinez, Madelyn Chenault, Neil Shulman, Baldeo K. Mishra, Katherine Otto and Cecile L. Cate. "Prevalence, Treatment and Control Status of Hypertension in Rural Middle Georgia." Unpublished manuscript.
2. Shah, B.V. "SESUDAAN: Standard Errors Program for Computing of Standardized Rates from Sample Survey Data." Report #RTI/5250/00-01S. Research Triangle Institute, Research Triangle Park, N.C. April, 1981.

Acknowledgements

This research upon which this publication is based was performed pursuant to Grant No. HL 24004 and Contract No. NO1-HV-92192 with the National Heart, Lung, and Blood Institute, NIH. Acknowledgements are expressed to staff at Fort Valley State College, Emory University School of Medicine, and Georgia Department of Human Resources for contributing to the implementation of the sample survey, to staff at Centers for Disease Control for consultation on variance computation, and to Robert Vogel for reviewing the manuscript.

Table 1

Symbols Used in Tables 2 thru 4

A. Independent Replicate Technique

1. $\hat{\theta}_i$ = point estimate of the population parameter obtained from the ith complete replicate, i = 1, 2, 3
2. $\hat{\theta}$ = point estimate of the population parameter obtained from the three completed replicates pooled together
3. n = total sample size (all three replicates added together)
4. $\hat{\theta}^2$ = weighted estimate of the population variance, assuming simple random sampling

$$= \sum_{i=1}^{n} W_i(X_i - \bar{X}_W)^2/(n-1) \qquad \text{where } \bar{X}_W = \sum_{i=1}^{n} W_i X_i / \sum_{i=1}^{n} W_i$$

5. S.E. $(\hat{\theta})$ = standard error of the estimate

$$= \sqrt{\sum_{i=1}^{3} (\hat{\theta}_i - \hat{\theta})^2/6}$$

6. DEFF = design effect

$$= \frac{\left[\text{S.E.}(\hat{\theta})\right]^2}{\hat{\sigma}^2/n} \text{ for means}$$

$$= \frac{\left[\text{S.E.}(\hat{\theta})\right]^2}{(\hat{\theta})(100-\hat{\theta})/n} \text{ for percentages}$$

B. Taylor Series Technique

1. $\hat{\theta}$ = Point estimate of the population parameter obtained from three completed replicates plus completed segments from the fourth replicate
2. n = total sample size (all three replicates added together plus part of the fourth replicate)
3. S.E. $(\hat{\theta})$ = standard error of the estimate calculated by SESUDAAN
4. DEFF = design effect calculated by SESUDAAN

C. Comparison index

$$I = \frac{\text{DEFF (replicate)}}{\text{DEFF (Taylor)}}$$

Parameter (θ)	INDEPENDENT REPLICATE TECHNIQUE							TAYLOR SERIES TECHNIQUE				
	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}$	n	S.E.($\hat{\theta}$)	DEFF	$\hat{\theta}$	n	S.E.($\hat{\theta}$)	DEFF	I
1. % all adults normotensive	73.8	75.7	78.2	75.9	585	1.27	0.52	75.1	653	2.13	1.60	0.57
2. % all whites "	74.0	76.1	83.8	78.0	342	2.98	1.77	76.9	369	3.22	2.16	0.91
3. % all blacks "	73.7	75.1	71.4	73.4	243	1.08	0.14	73.0	284	3.30	1.57	0.30
4. % all men "	75.4	74.1	81.1	76.9	253	2.15	0.66	75.6	285	3.07	1.46	0.67
5. % all women "	72.5	77.0	75.6	75.0	332	1.33	0.31	74.7	368	2.40	1.12	0.53
6. % all white men "	73.1	72.2	79.7	75.0	156	2.36	0.47	73.4	169	4.64	1.86	0.50
7. % all white women "	74.8	79.8	87.6	80.7	186	3.72	1.66	80.2	200	3.04	1.16	1.20
8. % all black men "	78.4	76.7	83.2	79.4	97	1.95	0.22	78.7	116	3.95	1.08	0.45
9. % all black women "	70.1	73.9	62.6	68.8	146	3.32	0.75	68.7	168	4.63	1.68	0.67
10. % all adults controlled on med.	13.2	10.5	12.0	11.9	585	0.78	0.34	12.5	653	1.47	1.29	0.51
11. % all whites "	14.3	7.8	9.6	10.6	342	1.94	1.36	10.7	369	1.96	1.49	0.96
12. % all blacks "	11.9	13.7	14.9	13.5	243	0.87	0.16	14.7	284	2.48	1.39	0.34
13. % all men "	11.8	4.6	10.3	8.9	253	2.19	1.50	8.5	285	1.56	0.89	1.30
14. % all women "	14.3	15.5	13.4	14.4	332	0.61	1.00	15.9	368	2.27	1.41	0.84
15. % all white men "	14.0	1.6	11.5	9.1	156	3.79	2.70	8.6	169	2.77	1.64	1.28
16. % all white women "	14.5	13.6	7.8	12.0	186	2.10	0.78	12.6	200	2.21	0.88	0.94
17. % all black men "	9.0	8.6	8.7	8.8	97	0.12	0.002	8.2	116	1.83	0.52	0.06
18. % all black women "	14.2	17.6	19.5	17.1	146	1.55	0.25	19.5	168	3.98	1.69	0.38
19. % all adults needing treatment	13.0	13.8	9.9	12.2	585	1.19	0.77	12.4	653	1.38	1.15	0.82
20. % all whites " "	11.8	16.1	6.6	11.5	342	2.75	2.53	12.4	369	2.15	1.58	1.27
21. % all blacks " "	14.4	11.1	13.8	13.1	243	1.01	0.22	12.4	284	1.87	0.92	0.49
22. % all men " "	12.8	21.3	8.5	14.2	253	3.76	2.94	15.9	285	2.75	1.61	1.35
23. % all women " "	13.1	7.5	11.0	10.5	332	1.63	0.94	9.4	368	1.27	0.70	1.16
24. % all white men " "	12.8	26.2	8.8	16.0	156	5.26	3.21	18.0	169	3.94	1.78	1.34
25 % all white women " "	10.8	6.7	4.6	7.3	186	1.82	0.91	7.2	200	1.92	1.11	0.91
26. % all black men " "	12.6	14.7	8.2	11.8	97	1.92	0.34	13.1	116	3.48	1.24	0.52
27. % all black women " "	15.8	8.4	17.9	14.0	146	2.88	1.01	11.8	168	2.14	0.74	1.17

Table 2. Results of the Replication and Taylor Series Techniques to Estimating Variances of Estimates of the Percentage Distribution of Hypertension by Race and by Sex for Adults in Four Rural Counties in Georgia, 1981

Parameter Θ	INDEPENDENT REPLICATE TECHNIQUE								TAYLOR SERIES TECHNIQUE				
	$\hat{\Theta}_1$	$\hat{\Theta}_2$	$\hat{\Theta}_3$	$\hat{\Theta}$	n	$\hat{\sigma}^2$	S.E.($\hat{\Theta}$)	DEFF	$\hat{\Theta}$	n	S.E.($\hat{\Theta}$)	DEFF	I
Mean DBP of													
1. all adults on medication	81.70	84.64	80.25	82.12	206	215.92	1.29	1.59	82.09	239	0.99	1.11	1.20
2. whites " "	82.68	82.24	79.17	81.59	110	215.82	1.12	0.63	81.71	122	1.55	1.41	0.67
3. blacks " "	80.18	86.19	81.14	82.67	96	215.49	1.87	1.55	82.45	117	1.33	0.98	1.26
4. men " "	83.69	88.54	81.46	84.33	71	161.88	2.10	1.93	84.40	82	1.41	1.06	1.35
5. women " "	80.28	82.44	79.51	80.77	135	244.05	0.88	0.43	80.76	157	1.19	0.93	0.68
6. white men " "	86.16	87.85	83.54	85.57	42	164.47	1.27	0.41	85.45	47	2.17	1.32	0.56
7. white women " "	79.88	79.61	73.47	78.40	68	234.20	2.16	1.35	78.73	75	1.83	1.14	1.09
8. black men " "	79.05	89.03	77.16	82.53	29	152.72	3.72	2.63	82.82	35	1.90	0.98	1.64
9. black women " "	80.84	84.72	82.30	82.73	67	243.78	1.13	0.35	82.30	82	1.54	0.79	0.67
Mean SBP of													
1. all adults on medication	138.27	145.88	133.70	139.25	206	474.59	3.55	5.48	138.84	239	1.86	1.74	1.77
2. whites " "	141.15	145.89	136.98	141.29	110	366.50	2.57	1.98	139.92	122	2.08	1.38	1.20
3. blacks " "	133.78	145.88	130.99	137.16	96	576.82	4.57	3.48	137.80	117	2.98	1.85	1.37
4. men " "	137.53	143.63	137.29	139.14	71	411.94	2.09	0.75	138.56	82	2.56	1.37	0.74
5. women " "	138.79	147.00	131.50	139.32	135	512.79	4.48	5.28	139.00	157	2.16	1.40	1.94
6. white men " "	138.07	148.36	138.05	140.07	42	438.29	3.58	1.23	138.98	47	3.68	1.51	0.90
7. white women " "	143.64	144.73	135.59	142.26	68	307.02	2.96	1.94	140.66	75	2.17	1.00	1.39
8. black men " "	136.52	140.22	135.72	137.78	29	370.45	1.40	0.15	137.94	35	2.76	0.78	0.44
9. black women " "	132.18	148.83	129.61	136.89	67	669.56	6.02	3.63	137.74	82	3.68	1.71	1.46

Table 3. Results of the Replication and Taylor Series Techniques to Estimating Variances of Estimates of Mean Blood Pressure by Race and by Sex for Adults on High Blood Pressure Medication in Six Rural Counties in Georgia, 1981

Parameter (θ)	INDEPENDENT REPLICATE TECHNIQUE								TAYLOR SERIES TECHNIQUE				
	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$	$\hat{\theta}$	n	$\hat{\sigma}^2$	S.E.($\hat{\theta}$)	DEFF	$\hat{\theta}$	n	S.E.($\hat{\theta}$)	DEFF	I
Mean DBP of													
1. all adults Not on medication	75.68	76.75	76.12	76.19	939	145.73	0.31	0.62	75.94	1018	0.50	1.73	0.60
2. whites " "	73.98	76.71	75.66	75.50	539	140.58	2.43	2.43	75.57	573	0.61	1.50	1.27
3. blacks " "	77.65	76.80	76.73	77.06	400	150.94	0.30	0.23	76.41	445	0.80	1.85	0.35
4. men " "	77.76	80.01	78.71	78.85	441	143.31	0.65	1.31	78.84	486	0.69	1.59	0.91
5. women " "	73.80	73.64	73.76	73.73	498	135.40	0.05	0.01	73.22	532	0.64	1.60	0.08
6. white men " "	76.30	80.67	79.19	78.81	266	152.40	1.28	2.88	78.91	284	0.91	1.55	1.36
7. white women " "	71.74	72.67	72.54	72.34	273	108.82	0.29	0.21	72.39	289	0.58	0.88	0.49
8. black men " "	79.59	79.04	78.10	78.90	175	131.18	0.44	0.25	78.75	202	1.10	1.82	0.37
9. black women " "	76.03	74.84	75.43	75.45	225	162.79	0.34	0.16	74.28	243	1.12	1.86	0.29
Mean SBP of													
1. all adults Not on medication	122.19	122.43	123.11	122.58	939	289.58	0.28	0.25	122.21	1018	0.68	1.66	0.39
2. whites " "	123.10	122.35	122.58	122.66	539	301.41	0.22	0.09	122.17	573	0.89	1.53	0.24
3. blacks " "	121.12	122.54	123.82	122.48	400	273.79	0.78	0.89	122.27	445	1.24	2.51	0.60
4. men " "	126.02	128.29	127.05	127.14	441	259.64	0.66	0.73	126.55	486	0.76	1.10	0.81
5. women " "	118.72	116.85	119.52	118.37	498	279.78	0.79	1.11	118.15	532	0.87	1.43	0.88
6. white men " "	127.40	127.99	127.08	127.50	266	282.87	0.27	0.07	126.97	284	0.98	1.00	0.26
7. white women " "	118.93	116.61	118.59	118.04	273	275.33	0.72	0.52	117.57	289	1.07	1.21	0.66
8. black men " "	124.27	128.72	127.00	126.66	175	228.24	1.30	1.29	126.01	202	1.28	1.47	0.82
9. black women " "	118.50	117.16	120.80	118.79	225	284.98	1.06	0.89	118.89	243	1.47	1.81	0.70

Table 4. Results of the Replication and Taylor Series Techniques to Estimating Variances of Estimates of Mean Blood Pressure by Race and by Sex for Adults not on High Blood Pressure Medication in Six Rural Counties in Georgia, 1981