ACCURACY OF DIFFERENCE METHOD FOR APPROXIMATING SAMPLING ERRORS FOR PROPORTION ESTIMATES FROM A COMPLEX SAMPLE

Iris M. Shimizu, National Center for Health Statistics

1. Introduction

The sampling error for a proportion statistic P based on a complex sample is frequently derived by using the difference of rel-variances for the numerator and denominator of the proportion to approximate the rel-variance of P when both numerator and denominator are derived from the same survey. The results of this approximation method are accurate for simple random samples; however, little is known about the accuracy when the method is used on complex sample data.

In practice the method is used because for complex samples neither covariances of aggregate estimates nor variances of proportion estimates are easily accessible or the available resources do not permit computing the variance estimates for every desired statistic. In fact the very reasons that lead to use of this approximation method have also made it difficult or impossible to evaluate its accuracy.

A computer program developed by the National Center for Health Statistics (NCHS) to compute variances by using the balanced half-sample repeated replication (BRR) technique ([2],[3]) does permit such evaluation. The program was designed specifically for use on data from complex samples. It produces as by-products the rel-variances of proportion estimates when it produces the rel-variances of the aggregate estimates that are used in the numerator and denominator of the proportion estimates.

The NCHS BRR program was used to obtain the sampling errors for statistics from the 1978 sample-based National Reporting System for Family Planning Services (NRSFPS) conducted by NCHS. We used the output for these sampling errors to investigate the accuracy of results obtained from the difference method for approximating sampling errors of proportion statistics.

On the basis of our study, it appears that because of an excessive sample size in NRSFPS the approximation method gives reasonably accurate results except possibly in tests of hypotheses. However, results would be less than satisfactory for sample sizes of 20,000 or less in the same survey.

This paper describes the investigation and the detailed findings.

2. NRSFPS Survey Design

The NRSFPS collected data about the utilization of family planning clinics in the United States and selected territories. Prior to mid 1977, data were collected on all patients at all family planning visits at all clinics which voluntarily participated in the system. In mid 1977, the NRSFPS was converted to a multistage probability sample survey.

The first stage of the NRSFPS sample consisted of a stratified sample of clinics offering medical family planning services. The second stage consisted of systematic random samples of those visits to sampled clinics in which the patient received one or more medical services pertaining to a family planning method or infertility treatment. For each sampled visit, clinic personnel collected the required information from the patient and/or the clinic medical records. These data were then sent to Informatics, Inc. of Rockville, Maryland, the contractor for the survey.

The sample data were weighted to produce both State and national estimates. Variances for the U.S. estimates were computed by using the NCHS BRR computer program.

3. Notation and Study Description

In the discussion that follows, let

$P = N/D$ = a proportion estimate

N = numerator estimate

D = denominator estimate

X = estimate

$S(X)$ = standard error of X

$RSE(X) = S(X)/X$ = the relative standard error of X

$V^2(X) = [RSE(X)]^2$ = the rel-variance of X.

Estimates for the sampling errors of proportion estimates can be derived through the use of an approximation for the rel-variance of P. That approximation can be written as

$$V^2(P) \doteq V^2(N) + V^2(D) - 2\ Cov(N,D)/ND.[1] \quad (1)$$

It can be proven that when both the numerator and the denominator of a proportion estimate are derived from the same simple random sample, then

$$Cov(N,D)/ND = V^2(D) \quad (2)$$

so that equation (1) becomes

$$V^2(P) = V^2(N) - V^2(D). \quad (3)$$

The relationship (3) between the rel-variances has not been proven true in complex samples, however. Yet, in practice, the relationship is frequently used to approximate the rel-variances of proportion estimates based on data from complex samples without any evaluation of its accuracy for the data set used. We were able to evaluate the accuracy of the approximations in the NRSFPS data set.

From NRSFPS the proportion estimates for which rel-variances were available included proportions of visits made by females having less than 12 years, 12 years, and more than 12 years of education, the proportions of female patients having each of the three levels of education, and the proportions of female patients using pills, IUD's, diaphragms, foam, natural methods, sterilization, other methods, or no method of contraception prior to their visit to a family planning clinic. These statistics were for four age groups (less than 20 years, 20-24 years, 25-29 years, and 30 or more years) and for three race groups (white, black, and other). These estimates were based on the U.S. portion (excluding the territories) of the NRSFPS sample for 1978. In that year there were about 1,200 sample clinics that were in scope for the NRSFPS in the U.S. of which about 1,000 responded with total samples of about 280,000 visits by females and about 140,000 female patients.

For estimated proportion P = N/D we defined

$$V_2{}^2(P) = V^2(N) - V^2(D) \qquad (4)$$

and then computed V_2 for each P in the data set using the values of $V^2(N)$ and $V^2(D)$ which were each based on the 1978 NRSFPS and computed by using the BRR technique. The $V_2{}^2(P)$ is the approximation to the rel-variance of P expressed in equation (3). We then computed $V_2(P)$ and $S_2(P)$, the corresponding approximations to RSE(P) and S(P).

For simplicity in the remainder of this paper, attention is primarily focused upon the results of this method of approximation for the RSE's and standard errors, since in practice RSE's and standard errors, instead of rel-variances, are used to define reliable statistics and significant differences. We defined the values of RSE(P) and S(P) which were calculated directly by the BRR computer program from the 1978 NRSFPS data as the correct values and compared $V_2(P)$ and $S_2(P)$ against those values.

We first consider the frequency with which the approximations were relatively close to the correct values being approximated. For this part of the investigation we calculated the ratio

$$R(P) = V_2(P)/RSE(P) = S_2(P)/S(P) \qquad (5)$$

for each proportion estimate. These ratio values were then plotted in scatter diagrams and the results summarized in Table 1. A value of R(P) less than one indicates under-approximation to the correct value. When values of R(P) are close to one, the approximations to RSE(P) and S(P) may be considered reasonable. For discussion purposes, approximations $V_2(P)$ and $S_2(P)$ are arbitrarily defined here to be close to the correct values RSE(P) and S(P) when the associated R value falls between 0.8 and 1.2, that is, when the approximations are within 20 percent of the correct values. For example, a ratio of R(P) = 0.80 means that the approximation is $V_2(P) = 0.24$ when the correct value is $V(P) = 0.30$. It is noted, however, that in some circumstances, a difference of 20 percent between the correct and the approximated values could be unacceptable.

We also investigate the magnitude of differences that may exist between the approximations and the corresponding correct values for both RSE's and standard errors, i.e., the differences

$$RSE(P) - V_2(P) \quad \text{and} \quad S(P) - S_2(P).$$

These results are summarized in Table 2.

4. Findings

The following discussion pertains to only the statistics included in this study because no attempt was made in this study to include a probability sample of all proportion statistics derived from the 1978 NRSPFS.

Since $V_2(P)/RSE(P) = S_2(P)/S(P)$, statements of $V_2(P)$ relative to RSE(P) apply also to the ratio of $S_2(P)$ to S(P) and the accuracy of $S_2(P)$ relative to S(P). The percent distribution for the ratios is presented in Table 1 for several variables. As can be seen, the approximations exeeded the correct values as often as not and not quite one half of them lie between 0.8 and 1.2. That means the correct values are underestimated about as often as they are overestimated by the approximation method and the approximations do not fall within 20 percent of the correct values quite as often as they fall outside that range. In addition, only 68 percent of the approximations fall within 40 percent of the correct values.

In Table 1 the likelihood with which approximations V_2 are relatively close to the correct RSE values varies with the attribute for which the proportion statistics are derived and inversely with domain size. The approximations fall within 20 percent of the correct values 20 percent of the time for estimated proportions of patients having different levels of education and 59 percent of the time for estimated proportions of patients using various contraceptive methods prior to their visit to a family planning service site. There is no complete explanation for the variation by attribute but it is known that domain size played a part in the observed variation.

The decrease in likelihood of an approximation being close to the correct value as the domain size increased could have been expected to some extent on the basis of an earlier evaluation of the NRSFPS variances.([4],[5]) There it was noted that for visit statistics the average design effects for standard errors increased from 2.6 for domains of less than one million to 7.8 for domains of three million or more visits. Similarly for patient statistics, the average design effects increased from 2.2 to 4.0 for the same domain sizes in the patient population. A smaller design effect implies that the correct sampling errors suffer less deviation from the corresponding errors that would have been obtained had a simple random sample been used instead of a complex sample. Thus it follows that when design effects are smaller, the estimates of RSE and standard errors based on the difference approximation (3) should be more likely to be close to the correct values since the difference approximation gives the correct value when simple random sampling is used.

It can also be seen in Table 1 that the smaller the estimated proportion P, the more likely it is for the corresponding approximation V_2 to be relatively close to the correct RSE value. The likelihood ranges from a high of 0.71 when P is less than 0.10 to a low of 0.16 when P is 0.40 or more. For the statistics included in this study the approximations are not likely to be relatively close to the correct RSE values when P exceeds 0.10. Indeed when P is 0.40 or more, the approximations differ by at least 40 percent of the correct values about two-thirds of the time.

Correspondingly, it is also seen in Table 1 that the likelihood of relatively close approximation to correct RSE values increases from 0.15 when the RSE is less than 3 percent to 0.82 when RSE is 20 percent or more. Relatively close approximation occurs 78 percent or more of the time when the correct RSE values are 10 percent or more.

While approximations V_2 and S_2 have been considered in the prior discussion as relatively close whenever they were within 20 percent of the corresponding correct RSE and S(P) values, approximations that are within 20 percent of the correct values may still lead to erroneous results in tests of hypotheses. In the simple t-tests, use of under-approximations for standard errors could result in rejection of a hypothesis that would not be rejected if the correct values were used. Conversely use of over-approximations for standard errors could prevent rejection of a hypothesis that would have been rejected if the correct values were used. These two types of errors resulting from use of approximations for the standard errors appear almost equally likely to occur since, according to Table 1, the ratios of approximations to the correct values are almost symmetrically distributed about 1 except where the approximations are from 20 to 40 percent away from the correct values.

For a simple illustration of the effect which use of approximations in place of correct values for standard errors may have on test results, consider the t-test of the hypothesis H_0: $P'' = P'$. The correct value of the test statistics is formulated as

$$t = \frac{P'' - P'}{\sqrt{S^2(P'') + S^2(P')}} . \qquad (6)$$

Suppose that $S_2(P) = 0.9\ S(P)$ for both P' and P'' and $S_2(P)$ is substituted for S(P) in (6). Then the approximated test statistic becomes $t_2 = 1.1 \cdot t$. If the critical value is Z = 2.0 and if the correct value lies between the ratio 2.0/1.1 = 1.8 and 2.0, then the use of the approximations would result in rejection of H_0 whereas use of correct values would lead to the opposite conclusion. If, on the other hand, $S_2(P) = 1.1\ S(P)$ for both P' and P'' and if the correct value of the test statistic was between 2.0 and 2.2, then use of the approximations would prevent rejection of H_0 whereas use of correct values would lead to rejection.

If one is not testing hypotheses and the difference between correct and approximate values is negligible, then the approximations may be quite satisfactory despite poor showings in relative accuracy. Hence, we now consider absolute differences between the correct and the approximate values for RSE's and standard errors.

In Table 2, it is seen that on the average the absolute difference between approximate and correct values for RSE(P) and S(P) are indeed small for the statistics from the 1978 NRSFPS which are included in this study. The average differences are less than 2 percentage points for RSE's and less than one percentage point for standard errors, regardless of attribute class or P value. The differences for individual statistics, however, range as high as 0.09 for RSE's and as high as 0.02 for standard errors. The differences generally decreased as P increased.

The generally small approximation errors are probably due to the large sample size (280,000 visits). Both the correct and approximated values of the sampling errors are functions of the square root of the sample size n, thus the error in the approximation is also a function of the sample size. Hence, the smallest sample size required to produce a given approximation error E on the average can be computed. For the 1978 NRSFPS, the minimum sample size required to give a maximum error for RSE's can be formulated as

$$n = 280{,}000 \times (0.014/E)^2 \qquad (6)$$

and the corresponding formula for S(P) is

$$n = 280{,}000 \times \left[\frac{S(P) - S_2(P)}{E} \right]^2 . \qquad (7)$$

When the maximum error desired is E = 0.05 percentage points, the minimum sample sizes needed are 22,000 for RSE's and 20,000 for S(P) when P = 0.40.

The effect on approximation errors caused by reducing the sample size can be seen in Table 2. There the errors in approximations are not negligible when the sample size is less than 20,000.

We also considered the effects which the approximation method has on reliability. The RSE is frequently used in defining reliability for statistics. If, as is commonly done, statistics having RSE less than 30 percent are defined as reliable, use of approximation V_2 in place of the correct RSE to define reliability would result in error for only three out of the 280 cases considered from the 1978 NRSFPS. All three of these occured when P was less than 0.10 and the correct RSE was 27-32 percent; in other words the statistics involved were small and were borderline cases with regard to reliability.

5. Conclusion

The approximation of rel-variances for proportion statistics derived by taking the differences of the rel-variances for the numerator and the denominator in the proportion is convenient when the latter rel-variances are available and both numerator and denominator are derived from the same sample. However the approximation method is with some error when the data come from complex samples.

For the statistics used in our investigation, the approximations appear to be sufficiently accurate on the average except possibly when used in tests of hypotheses. Accuracy, however, depends on sample size. For the statistics observed, sample sizes in excess of 20,000 units were required to obtain a reasonable level of accuracy. Also, approximations were most likely to be relatively close to the correct values for smaller P values and for larger RSE values.

The accuracy of the approximations also appears to vary inversely with survey design effects. This suggests that results from use of the difference method for approximating variances for P statistics in other surveys may differ from those observed in this limited study, especially if design effects differ from those observed in the NRSFPS survey.

References

[1] Hansen, Morris H.; Hurwitz, William N.; and Madow, William G. (1953). Sample Survey Methods and Theory, Vol I. John Wiley & Sons.

[2] Jones, Gretchen (1977). "HES Variance and Crosstabulation Program; Version 2." Unpublished document. National Center for Health Statistics. Hyattsville, Maryland.

[3] McCarthy, Philip J. (1969). "National Center for Health Statistics: Pseudoreplication: Further Evaluation and Application of the Balanced Half-sample Technique." Vital and Health Statistics. PHS PUB. No. 1000-Series 2-No. 31. Public Health Service. Washington. U.S. Government Printing Office.

[4] Shimizu, Iris M. (1980). "Variances for NRSFPS Visit Statistics." Unpublished document. National Center for Health Statistics.

[5] Shimizu, Iris M. (1980). "Variances for 1978 NRSFPS Patient Statistics." Unpublished document. National Center for Health Statistics.

Appendix

Let $\hat{P} = \hat{X} / \hat{Y}$ be a proportion estimate where $\hat{X}$ and $\hat{Y}$ are estimated from the same simple random sample. Then the approximation to the rel-variance of $\hat{P}$ is:

$$V^2(\hat{P}) = V^2(\hat{X}) - V^2(\hat{Y}). \tag{A1}$$

Walt R. Simmons, formerly of NCHS, originally proved this identity. However, to the author's knowledge, the proof for it is not published. Hence, we outline the proof here.

Proof: Define

$$X = \sum_{i=1}^{N} X_i \quad \text{and} \quad Y = \sum_{i=1}^{N} Y_i, \tag{A2}$$

N = total number of units in the universe,

n = sample size from the universe,

$$X_i = \begin{cases} 1 \text{ if the i-th sample unit has both X and Y attributes} \\ 0 \text{ otherwise} \end{cases}$$

and

$$Y_i = \begin{cases} 1 \text{ if the i-th sample unit has the Y attribute} \\ 0 \text{ otherwise} \end{cases}$$

In a simple random sample, the aggregate estimates for the attribute variables are:

$$\hat{X} = \frac{N}{n} \sum_{i=1}^{N} X_i\delta_i \quad \text{and} \quad \hat{Y} = \frac{N}{n} \sum_{i=1}^{N} Y_i\delta_i \tag{A3}$$

where

$$\delta_i = \begin{cases} 1 \text{ if the i-th unit is in the sample} \\ 0 \text{ otherwise.} \end{cases}$$

The Taylor series approximation to $V^2(\hat{P})$ is

$$V^2(\hat{P}) = V^2(\hat{X}) + V^2(\hat{Y}) - 2\,\mathrm{Cov}(\hat{X},\hat{Y})/XY. \tag{A4}$$

It is sufficient to show that

$$V^2(\hat{Y}) = \mathrm{Cov}(\hat{X},\hat{Y})/XY. \tag{A5}$$

$$\text{Now } \mathrm{Cov}(\hat{X},\hat{Y}) = E(\hat{X}\ \hat{Y}) - E(\hat{X})E(\hat{Y}). \tag{A6}$$

It can be shown for a simple random sample that

$$E(\hat{X}\ \hat{Y}) = \frac{N}{n} X \left[1 + \frac{n-1}{N-1}(Y - 1)\right]. \tag{A7}$$

Hence, the Cov $(\hat{X},\hat{Y})$ can be written as

$$\frac{\mathrm{Cov}(\hat{X},\hat{Y})}{X\ Y} = \frac{1}{Y}\left[\frac{N}{n}\left[\ 1 + \frac{n-1}{N-1}(Y - 1)\right] - Y\right]. \tag{A8}$$

For the left side of the equation (A5),

$$V^2(\hat{Y}) = [E(\hat{Y}^2) - E^2(\hat{Y})]/Y^2. \tag{A9}$$

But it can be shown for a simple random sample that

$$E(\hat{Y}^2) = \frac{N}{n}\ Y \left[1 + \frac{n-1}{N-1}(Y - 1)\right]. \tag{A10}$$

Hence,

$$V^2(\hat{Y}) = \frac{\frac{N}{n}\left[1 + \frac{n-1}{N-1}(Y - 1)\right] - Y}{Y}. \tag{A11}$$

The right side of (A11) is identical to the right side of (A8).

TABLE 1: Percent Distribution for Ratios $V_2(P)/RSE(P)$ in 1978 National Reporting System for Family Planning Services: U.S.

	Number of Proportion Statistics Studied	$V_2(P)/RSE(P)$							
		0.8- 1.2	0.6- 1.4	<0.6	0.6- 0.8	0.8- 1.0	1.0- 1.2	1.2- 1.4	1.4<
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
					Percent				
Total	280	49*	68*	17	6	26	22	13	15
Attribute Class									
Visits by Education	60	48	72	15	8	23	25	15	13
Patients by Education	60	20	43	30	3	8	12	20	27
Patients by prior method	160	59	76	13	6	34	25	10	11
Domain Size for Statistics									
Less than one million	201	53	73	15	7	29	23	12	12
1 - 3 million	62	40	58	23	3	18	23	15	19
3 million or more	17	29	47	24	-	24	6	18	29
Estimated Proportion P									
Less than 10%	128	71	90	5	7	43	28	12	5
10 - 29%	53	41	65	19	4	21	19	21	17
30 - 39%	49	29	49	14	1	12	16	18	37
40% or more	50	16	34	42	8	8	8	10	24
Correct RSE of P									
Less than 3%	75	15	36	29	4	8	7	17	35
3 - 4%	71	45	69	21	10	24	21	14	10
5 - 9%	63	49	71	14	10	21	29	13	14
10 - 19%	49	78	98	2	6	51	27	14	-
20% or more	21	82	95	-	5	45	36	9	5

*These percents may not equal sum of percents in columns (5) - (8) due to rounding of figures.

TABLE 2: Average Absolute Differences Between the Correct and the Approximated Sampling Errors and the Range of those Differences in the 1978 National Reporting System for Family Planning Services: U.S.

Variable	Number of Proportion Statistics Studied	$\lvert RSE(P) - V_2(P) \rvert$		$\lvert S(P) - S_2(P) \rvert$		Range of Actual Values for RSE(P)
		Average	Range	Average	Range	
	(1)	(2)	(3)	(4)	(5)	(6)
				Percentage points		
Total	280	1.40	0.02- 9.20	0.32	0.00- 2.25	0.72- 48.14
Attribute Class						
Visits by Education	60	1.03	0.02- 3.90	0.34	0.01- 2.09	0.80- 12.23
Patients by Education	60	1.84	0.10- 5.70	0.63	0.04- 1.80	0.92- 16.07
Patients by prior method	160	1.37	0.02- 9.20	0.19	0.00- 2.25	0.72- 48.14
Estimated Proportion P						
Less than 10%	128	1.21	0.07- 9.20	0.05	0.00- 0.57	2.69- 48.14
10 - 29%	53	1.65	0.06- 7.23	0.36	0.02- 1.69	1.22- 16.07
30 - 39%	49	1.73	0.10- 7.10	0.62	0.08- 2.25	0.71- 12.35
40% or more	50	1.30	0.02- 5.00	0.66	0.01- 2.20	0.72- 05.07
Sample Size						
4,000 visits	280	11.64	0.16-76.50	2.66	0.02-18.71	5.99-400.33
20,000 visits	280	5.21	0.07-34.20	1.19	0.00- 8.37	2.68-179.03
50,000 visits	280	3.29	0.05-21.64	0.75	0.00- 5.29	1.69-113.23
100,000 visits	280	2.32	0.03-15.30	0.53	0.00- 3.74	1.19- 80.07