Journal of Official Statistics
Vol. 6. No. 3, 1990, pp. 223–239

Fay's Method for Variance Estimation

David R. Judkins[1]

Abstract: The standard balanced repeated replication (BRR) method of estimating variances involves dividing the sample in each stratum into half-samples, selecting a balanced set of half samples across all strata, re-computing the statistic of interest (generally nonlinear) on each selected half-sample, and taking the mean square difference of among the replicate estimates as the variance estimate. One problem that occasionally arises is that one or more replicate estimates will be undefined due to division by zero. This is particularly common when ratio estimation has been used with very small cell sizes. Robert Fay suggested a solution to this problem several years ago: Instead of increasing the weights of one half sample by 100% and decreasing the weights of the other half sample to zero, he recommended perturbing the weights by $\pm x\%$. In this article, his suggestion is evaluated with simulation techniques. It is shown to be useful when variance estimates are needed for both smooth and nonsmooth statistics or when there are very few degrees of freedom available for variance estimation. The paper also discusses further modifications to the technique that are useful for variance estimation when only one PSU is selected per stratum.

Key words: Balanced half-samples; BRR; jackknife: Taylor linearization; Monte Carlo study.

1. Introduction

The variance estimation technique of balanced half-samples or balanced repeated replications (BRR) is well known. Formalized by McCarthy (1966) with roots at the U.S. Census Bureau in the 1950s it has long been used as one of a number of techniques to estimate the variance of nonlinear statistics from complex designs. (For a detailed and accessible discussion, see Wolter (1985.) In fact, it is often used for linear statistics from simple designs because of the simplicity of the system once it has been set up for complicated situations.

The advantages of this system relative to Taylor linearization and the jackknife have long been debated. Recently, Kovar (1985), and Hansen and Tepping (1985), reported on simulations which compared the jackknife, BRR, Taylor linearization, and the bootstrap. Rao and Wu (1985) have also studied the estimators theoretically. For ratio estimates, the results seem pretty clear that the jackknife and Taylor linearization are nearly equivalent, and that both of them are superior to BRR or the bootstrap. Contrary findings are given in Andersson, Forsman, and Wretman (1987), but they inadvertently compared an unstratified

[1] Senior Statistician with Westat, Inc., 1650 Research Blvd., Rockville, MD 20850, U.S.A. This work was supported in part by the National Institute of Dental Research as part of work by Westat on the Epidemiologic Survey of Oral Health in Adults.

jackknife with versions of Taylor linearization that did reflect the stratification (Judkins 1989). Thus, arguments between the methods generally revolve around which method is easier and cheaper to implement. The resampling methods (BRR, jackknife, and bootstrap) are generally easier to program but require lots of CPU time to run. The replicate weights required for unsophisticated users to use the method also require substantial increases in record lengths. This issue is best resolved by examining the complexity of the estimation procedures, the variety of estimates to be produced, and who will produce them. For extremely complex estimation procedures that involve the use of weights that have been iteratively adjusted, the resampling methods definitely come out ahead by virtue of programming ease. Where, on the other hand, the estimation procedures are fairly simple, the range of items for which variances are needed is fairly limited, and where sampling statisticians are available to assist in all variance calculations, the Taylor method is a good choice. Given the facts thus far, it would seem that the jackknife is generally the best choice for complex designs with complex estimation procedures.

More recently, however, Kovar, Rao, and Wu (1988) have shown that the jackknife provides much worse estimates of variances for medians than does BRR. (Taylor linearization cannot be used since the median is not smooth.) Presumably their results are also true for other non-smooth statistics. One solution for the median is to use the jackknife method to compute the variance on an estimated percentage of 50% and then use Woodruf's method (1952). However, Kovar, Rao, and Wu (1988, p. 42) show that this is also not a good choice when the population was "stratified by a concomitant variable [that is] highly correlated with [the variable of interest]." A statistician is then faced with a difficult choice if a procedure is needed that is good both for ratios and for medians. It is in this context that a modification to BRR suggested by Robert Fay is extremely useful. Fay's method (Dippo, Fay, and Morganstein 1984, sec. 4) is a compromise between BRR and the jackknife. His technique was motivated by the observation that the standard half-sample variance estimator occasionally runs into problems estimating the variance of ratios because the denominators are zero for some replicates. Less drastic but also a problem, some ratio replicates can be extremely large because of near-zero denominators. This is caused by the fact that when half the sample is zero weighted and half is double weighted, less common groups disappear more frequently than in the full sample. (The jackknife avoids these problems by only dropping one observation at a time.) Even where estimation of ratios is not an analytic objective, estimation procedures are frequently used that involve ratio-weighting. Ideally in the case of ratio-weighting, a half sample estimate should have all stages of weighting performed on it in exactly the same manner as they were performed on the full sample. If cells for nonresponse adjustment or post-stratification have small cell sizes, however, dropping out half the sample can result in cell sizes smaller than would have been tolerated in the full sample. (For the full sample weighting, the cells would have been collapsed.) The question then was how to maintain the structure of the balanced half samples but maintain the full cell sizes. He solved the puzzle by suggesting that the weights be perturbed more gently in each half sample.

The standard BRR perturbs the weights rather sharply. All weights are either multiplied by two or by zero. Then the mean

square error of the replicate estimates from the full sample estimate is used as an estimate of the variance. Fay's idea was to use the weights of 0.5 and 1.5 instead of 0 and 2 for the half samples within each stratum, or more generally, weights of k and $2 - k$ for $0 \leqslant k < 1$. (Such weights were in fact used for the 1984 panel of the U.S. Census Bureau's Survey of Income and Program Participation. They have also been used in the Epidemiologic Survey of Oral Health in Adults). Of course, the mean square error of the replicate estimates from the full sample estimate becomes much smaller when the weights are perturbed more gently. Fay noted that for linear statistics, the mean square error is too small by a factor of $(1 - k)^2$. He thus suggested that the mean square error be multiplied by $1/(1 - k)^2$ to obtain a reasonable estimate of variance. Heuristically, it makes sense that gentler perturbation should lead to fewer extreme replicate estimates, thereby yielding lower kurtosis of the replicates and hence better stability of the variance estimator.

Formulaically, Fay's method may be compared with BRR as follows. For simplicity, assume that there are two sample units in each of H strata. (Later in the paper, this simplifying assumption will be dropped.) Let $\tilde{w} = (w_{11}, w_{12}, \ldots, w_{H1}, w_{H2})$ be the vector of unbiased weights for the sample units. These unbiased weights are usually taken to be inverse selection probabilities. The final weights involve nonresponse adjustment and iterative post-stratification. Thus, the final weight for a case depends on many of the other weights in the sample, possibly all of them. To reflect this, let $w_{Fhi} = f_{hi}(\tilde{w})$ be the final weight for the ith case in the hth stratum. For BRR, the tth replicate weight is taken to be

$$w_{Fhi}^{(t)} = f_{hi}(\tilde{w}^{(t)}), \quad \text{where}$$

$$\tilde{w}^{(t)} = ((1 + d_{th})w_{11}, (1 - d_{th})w_{12}, \ldots$$

$$\ldots, (1 + d_{th})w_{H1}, (1 - d_{th})w_{H2}),$$

and d_{th} is the element in the tth row and hth column of a matrix of ones and negative ones with orthogonal columns and orthogonal rows. The problem mentioned before is that sometimes $f_{hi}(\tilde{w}^{(t)})$ is undefined even when $f_{hi}(\tilde{w})$ is defined. Fay's method solves this by taking

$$\tilde{w}^{(t)} = ((1 + d_{th}(1 - k))w_{11},$$

$$(1 - d_{th}(1 - k))w_{12}, \ldots$$

$$\ldots, (1 + d_{th}(1 - k))w_{H1},$$

$$(1 - d_{th}(1 - k))w_{H2}),$$

where $0 \leqslant k < 1$. In this paper, $100(1 - k)$ will be referred to as the perturbation factor. The perturbation factor for BRR is 100% ($k = 0$). Fay originally suggested a perturbation factor of 50%. In this paper, several other perturbation factors are also evaluated.

Let $\tilde{w}_F$ be the vector of w_{Fhi} and $\tilde{w}_F^{(t)}$ be the vector of $w_{Fhi}^{(t)}$. Let $\zeta(\tilde{w}_F, \tilde{x})$ be some statistic of interest such as a population total, ratio of domain means, a regression coefficient, or a log-odds ratio. The BRR estimate of variance is

$$(1/T)\Sigma_t[\zeta(\tilde{w}_F^{(t)}, \tilde{x}) - \zeta(\tilde{w}_F, \tilde{x})]^2,$$

where T is the number of replicates. Fay's estimate of variance is

$$[1/(1 - k)^2](1/T)\,\Sigma_t[\zeta(\tilde{w}_F^{(t)}, \tilde{x}) - \zeta(\tilde{w}_F, \tilde{x})]^2,$$

where the replicate weights are formed as indicated above.

As an example, suppose there are two strata, each with two PSUs. For simplicity of illustration, assume that $w_{hi} \equiv 1000$ for all four PSUs and that there is no ratio-adjustment. Then $\{\tilde{w}_F^{(t)}\} = \{(2, 0, 2, 0),$

(2, 0, 0, 2)} is one possible balanced set of replicates for BRR. For Fay's method with 50% perturbation, the corresponding set of replicates is $\{\tilde{w}_F^{(i)}\} = \{(1.5, 0.5, 1.5, 0.5), (1.5, 0.5, 0.5, 1.5)\}$. Consider the case where we are interested in the ratio of two variables and the bivariate distribution on the sample is $\tilde{x} = ((0, 0), (1, 2), (1, 3), (0, 0))$. The full sample estimate of the ratio of the first variable to the second is 2000/5000 = 0.4. The first BRR replicate estimate of the ratio is 2000/6000 = 0.333, but the second BRR replicate estimate is undefined. With Fay's method, the first replicate estimate is $2000/5500 = 0.\overline{36}$ and the second replicate estimate is 1000/2500 = 0.4. So Fay's estimate of variance is

$$\frac{1}{2}\left(\frac{1}{1 - 0.5}\right)^2 [(0.\overline{36} - 0.4)^2 + (0.4 - 0.4)^2] = 0.002645.$$

Practically, Fay's method is just as easy to implement as BRR. After the replicate weights have been created, the only change from current procedure is in the application of the multiplier. For example, if $k = 2/3$, then all variance estimates produced with old software need to be multiplied by 9 (standard error estimates by 3). Making it even easier, software has been developed at Westat which will implement Fay's method (or most other common resampling methods) when replicate weights are fed to it. From the standpoint of computer resources, Fay's method, BRR, and the jackknife (two PSUs per stratum, only dropping one) are all identical.

Extending the generality of the method, the two units within a stratum may be (1) ultimate units, (2) first stage units in a multistage design, (3) groups of ultimate units, or (4) groups of first stage units. In the second case, all second stage units within the same first stage unit will receive the same perturbation. The third case is useful when more than two ultimate units were selected per stratum. Similarly, the fourth case is useful when more than two PSUs have been selected per stratum. In the third or fourth case, it is necessary to collapse the ultimate or primary units within each stratum into just two groups per stratum. Rust (1986) discusses the implications of such collapsing.

Given the intuitive appeal of Fay's method and the fact that it does not require any more CPU time than the other resampling methods, it seemed important to determine the bias and stability properties of the method. In section 2, some simple properties of Fay's method are discussed. The method is shown to have desirable properties for estimating the variance of linear statistics. Theoretical discussion of the method's properties for nonlinear statistics is much more difficult, and none is presented. Instead, a Monte Carlo simulation study was done that investigated the method's properties for ratios, regression coefficients, and medians. The methodology for the study is discussed in Section 4. The results of this study were very favorable and are presented in Section 5. In Section 6, some extensions of Fay's method are presented that are useful for estimating variances when only one PSU has been selected per stratum. Section 6 closes with a summary and directions for further research.

2. Equivalency for Linear Statistics

Fay's method is identical to BRR for linear statistics. Since BRR is itself equivalent, for linear statistics, to the textbook variance estimate for the Horvitz-Thompson estimate from a design where two units are selected with probability proportionate to size and with replacement (Wolter 1985,

p. 123), Fay's method is shown to reduce to a variance estimator that is widely agreed to be optimal.

Fay's variance estimator in this case is

$$\begin{aligned}
\text{estvar} &= (1/T)[1/(1-k)]^2 \\
&\quad \times \Sigma_t\{\Sigma_h\{[1+d_{th}(1-k)]w_{h1}x_{h1} \\
&\quad + [(1-d_{th}(1-k)]w_{h2}x_{h2}\} \\
&\quad - \Sigma_h(w_{h1}x_{h1}+w_{h2}x_{h2})\}^2 \\
&= (1/T)[1/(1-k)]^2 \\
&\quad \times \Sigma_t\{\Sigma_h[d_{th}(1-k)w_{h1}x_{h1} \\
&\quad - d_{th}(1-k)w_{h2}x_{h2}]\}^2 \\
&= (1/T)\Sigma_t[\Sigma_h(d_{th}w_{h1}x_{h1} \\
&\quad - d_{th}w_{h2}x_{h2})]^2 \\
&= (1/T)\Sigma_t\{\Sigma_h[(1+d_{th})w_{h1}x_{h1} \\
&\quad + (1-d_{th})w_{h2}x_{h2}] \\
&\quad - \Sigma_h(w_{h1}x_{h1}+w_{h2}x_{h2})\}^2
\end{aligned}$$

which is the same as the BRR variance estimate.

3. Methodology for Nonlinear Statistics

Investigation of the bias and stability properties of Fay's method relied upon empirical investigation. Conceivably, these properties could be studied for smooth statistics theoretically by taking all fourth partial derivatives with respect to all stratum half-sample totals, but such an approach was not pursued. The empirical study was patterned after an earlier Monte Carlo study by Hansen and Tepping (1985) that was also the basis for Kovar (1985) and for Kovar, Rao, and Wu (1988). Hansen and Tepping created three main artificial populations with a common structure. They then varied the parameters to create a larger set of populations. In this study, only the first of their main populations was used. The common structure was 32 strata with two independent and identically distributed observations, (x_{h1}, y_{h1}) and (x_{h2}, y_{h2}), from a bivariate normal population within each stratum. The correlation, ρ, between x and y was assumed to be constant over all strata. The coefficient of variation (cv) per stratum for x was constant at 10% over all strata. The cv per stratum for y was roughly uniform at 24%. (The overall cv's for x and y were 1.8% and 5.6%, respectively.) The means of x and y were allowed to change across strata. Table 1 gives the stratum weight, means of x and y and their standard errors by stratum for population number 1. To create variations within this main population, the correlation (ρ) is varied from 0.8 to 0.2, the standard error of x is varied by a uniform factor of 1, 5, 10, or 15, and the standard error of y is varied by a uniform factor of 1, 5, or 10. Some of the resulting variations would be rare in practice and not to be recommended. Nonetheless, such variations are sometimes of interest. As will be seen later, significant differences between the methods occur only when the sample is being pushed hard (i.e., when variances are large).

Some study was also made of a population with less than 32 strata. The results are not presented, but they were similar to the results with very high stratum level variances. One exception is that as the number of strata decreases, the jackknife becomes more like BRR. The advantages that Fay's method has over BRR then apply to the comparison with the jackknife as well.

The sample units generating (x, y) may be ultimate units, PSUs, or groups of either,

Table 1. Parameters for population 1

Stratum (h)	W_h	μ_{xh}	μ_{yh}	σ_{xh}	σ_{yh}
1	0.042	100	90	10.0	25
2	0.042	95	75	9.5	24
3	0.042	90	70	9.0	22
4	0.039	98	75	9.8	22
5	0.039	93	70	9.3	20
6	0.037	98	75	9.8	24
7	0.037	96	75	9.6	23
8	0.037	94	75	9.4	22
9	0.037	92	70	9.2	24
10	0.034	96	75	9.6	23
11	0.034	94	70	9.4	20
12	0.034	92	70	9.2	22
13	0.034	90	70	9.0	22
14	0.031	96	75	9.6	25
15	0.031	94	70	9.4	20
16	0.031	92	70	9.2	18
17	0.031	90	70	9.0	19
18	0.031	88	70	8.8	20
19	0.031	86	65	8.6	20
20	0.031	84	60	8.4	18
21	0.031	82	60	8.2	16
22	0.031	80	60	8.0	20
23	0.028	90	70	9.0	22
24	0.028	85	65	8.5	18
25	0.028	80	60	8.0	20
26	0.025	90	70	9.0	20
27	0.025	85	60	8.5	18
28	0.025	80	50	8.0	15
29	0.025	75	50	7.5	14
30	0.020	75	50	7.5	16
31	0.016	75	45	7.5	14
32	0.013	75	45	7.5	12
Overall	1.000	89.744	68.245	1.646	3.830

as discussed above. Since the normality assumption is made, the simulation is probably most appropriate for either the case of two ultimate units per stratum with a normal variable such as height or the case of two PSUs per stratum with at least 30 second stage units per PSU. With this latter assumption, the PSU totals will be approximately normal from the central limit theorem. Of course, the most common survey variables are categorical in nature (counts of units in various categories). So it would be rare to have negative PSU totals in practise. Thus, a more realistic distribution would be one that approximates the normal over most of its range but is never allowed to be negative.

However, since the earlier authors did not force the simulated PSU totals to be nonnegative, PSU totals were allowed to be negative in this study so as to have better comparability across studies.

The version of the jackknife that was simulated is described by Kovar (1985) as the half jackknife. All previous work indicates little reason to investigate the full jackknife since the properties of the two are very similar except for use of CPU time; the full version requires twice as much (for nonlinear statistics). Formulaically, the only difference from BRR is in the definition of the matrix (d_{th}). For BRR, this matrix is taken to be an orthogonal (32 × 32) matrix of positive and negative ones. For the half jackknife (hereafter referred to as simply "the jackknife"), (d_{th}) is taken to be the identity matrix of order 32, with the assumption that the order of two units within each stratum is not significant in any way.

The pseudo-random numbers were generated by the standard congruential operator available in a package called GAUSS. When different methods were being tested on the same population parameters, a single seed and multiplier were used for the algorithm so that the methods were compared on the same population. When the population parameters changed, new seeds and occasionally new multipliers were used.

From the simulation, a number of evaluative statistics were generated for each underlying statistic and variance estimation method. It was first necessary to define a standard against which variance estimates could be evaluated. The standard was taken to be the mean square error of the underlying statistic across all replicates

$$M(\zeta) = \Sigma_r(\zeta_r - \bar{\zeta})^2/R + (\bar{\zeta} - \zeta_p)^2,$$

where ζ_r is the value of the underlying statistic on the r-th simulation, $\bar{\zeta}$ is the average of the ζ_r across all R simulations, and ζ_p is the value of the underlying statistic for the entire population.

The values of ζ_p were calculated for the ratio, the regression coefficient and the median as follows:

$$R_p = \frac{\mu_y}{\mu_x} = \frac{\Sigma_h W_h \mu_{yh}}{\Sigma_h W_h \mu_{xh}}$$

$$\beta_p = \frac{\Sigma_h W_h[\rho\sigma_{yh}\sigma_{xh} + (\mu_{yh} - \mu_y)(\mu_{xh} - \mu_x)]}{\Sigma_h W_h[\sigma_{xh}^2 + (\mu_{xh} - \mu_x)^2]}$$

m_p = median of the distribution function $\Sigma_h W_h \Phi[(y - \mu_{yh})/\sigma_{yh}]$, where Φ is the standard normal distribution function.

These definitions are plain enough for the ratio and the regression coefficient, but some explanation is needed for the definition of the population median since it is different from that chosen by Kovar, Rao, and Wu (1988). Those authors defined the population median implicitly as the lowest stratum mean with cumulative weight of 0.5 or higher. In this paper, it was assumed that the median of interest is the median of the random variable that results from selecting one of the 32 strata with probability proportionate to the stratum weights W_h and then selecting one (normally distributed) observation from the selected stratum. (This approximates the population median if the population of each stratum is assumed to be very large but finite and proportional to W_h.) This random variable has a mixed normal distribution. Note that with this assumption, the median need not coincide with one of the stratum means.

The values of ζ_r were calculated for the ratio, the regression coefficient and the

median as follows:

$$\hat{R}_r = \frac{\hat{\mu}_{ry}}{\hat{\mu}_{rx}}$$

$$= \frac{(\Sigma_h W_h \hat{\mu}_{ryh})}{(\Sigma_h W_h \hat{\mu}_{rxh})},$$

($\hat{\mu}_{ryh}$ is the average of the two observations from the stratum)

$$\hat{\beta}_r = \frac{\Sigma_h W_h \hat{\mu}_{ryh} \hat{\mu}_{rxh} - \mu_y \mu_x}{\Sigma_h W_h \hat{\mu}_{rxh} \hat{\mu}_{rxh} - \mu_x \mu_x}$$

$\hat{m}_r$ = Result of linear interpolation to the cumulative density 0.5078125 between the PSU means with cumulative density just below and just above that point.

The cumulative density of 0.5078125 was chosen as the interpolation point from the formula $0.5 + 1/(4H)$. The motivation for choosing this rather odd point stems from the fact that there are an even number of observations (64). If there were no weights, the natural choice would be to average the middle two PSU estimates. Interpolation to 0.5 usually results in taking the estimate from the 32nd PSU instead of averaging the 32nd and 33rd PSUs. Empirically, this continuity adjustment was found to improve the bias of the sample median. (Note that Kovar, Rao, and Wu (1988) used the estimate from the first PSU with cumulative density to equal or exceed 0.5.)

For each variance estimation method, the average, $\bar{v}$ and the mean square error, $M(\hat{v}, M(\zeta))$, across all simulations of the estimated variances from $M(\zeta)$ was calculated

$$\bar{v} = \Sigma_r \hat{v}_r / R$$

$$M(\hat{v}, M(\zeta)) = \Sigma_r [\hat{v}_r - M(\zeta)]^2 / R.$$

The ratio of $\bar{v}$ to $M(\zeta)$ is denoted as the "bias" and the ratio of $\sqrt{M(\hat{v}, M(\zeta))}$ to $M(\zeta)$ is denoted as the "stability." These definitions are consonant with the definitions of the earlier authors.

One caution noted by the earlier authors is that this sort of Monte Carlo design yields much smaller variances on $\bar{v}$ than on $M(\zeta)$. To correct for this disparity in accuracy, it is necessary to use a very large number of replicates to estimate $M(\zeta)$. For this study, 1000 replicates were used to estimate $M(\zeta)$ for the ratio and for the median, and 5000 for the regression coefficient. For estimating $\bar{v}$ for the ratio and for the median, the same sets of 1000 replicates were used as had been used to estimate $M(\zeta)$. For the regression coefficient, 200 simulations were used to estimate $\bar{v}$.

4. Results

Table 2 summarizes the results for the ratio. Consistent with earlier work, the jackknife is found to be generally better than BRR for estimating the variance on a ratio. The jackknife has somewhat better bias, much better stability, and marginally better confidence interval coverage. The superiority of the jackknife is particularly evident when there is high variance in the denominator at the stratum level. In this situation, BRR is practically useless. (The value of stability greater than 900 or 90,000% is not a typographical error.) This is, of course, exactly the type of situation that motivated Fay's method. So it is reassuring to see that Fay's method is indeed considerably better than BRR for this case. In fact, as k approaches unity, Fay's method appears to converge to the jackknife and thus to linearization.

Other observations from Table 2:

- For less extreme cases where the ratio is better behaved, there are no significant differences between the methods. In fact, the better behaved situation is probably the rule rather than the

Table 2. Estimating the variance of the ratio of two normal variables from a stratified sample – comparison of bias and stability for BRR, Fay's method and the jackknife

Coefficient of variation per stratum (and overall) %			Overall cv on ratio %	Method or perturbation factor %	Variance estimates		Observed error rates for nominal 90% confidence intervals %		
					Bias	Stability			
Y (numerator)	*X* (denominator)	ρ					Left	Right	Total
24	10	0.8	3.0	BRR	0.99	0.30	6.2	4.8	11.0
(5.6)	(1.8)			50	0.99	0.30	6.2	4.8	11.0
				1	0.99	0.30	6.4	4.8	11.2
				Jackknife	0.99	0.30	6.4	4.8	11.2
		0.5	3.5	BRR	1.01	0.29	5.7	4.3	10.0
				50	1.01	0.29	5.6	4.4	10.0
				1	1.01	0.29	5.5	4.3	9.8
				Jackknife	1.01	0.29	5.4	4.3	9.7
		0.2	4.0	BRR	0.96	0.29	5.6	5.3	10.9
				50	0.96	0.29	5.6	5.2	10.8
				1	0.96	0.29	5.6	5.3	10.9
				Jackknife	0.96	0.29	5.6	5.3	10.9
120	50	0.8	14.9	BRR	1.11	0.46	5.1	3.4	8.5
(28.1)	(9.2)			50	1.09	0.42	5.3	3.7	9.0
				1	1.09	0.40	5.5	3.8	9.3
				Jackknife	1.09	0.40	5.5	3.7	9.2
		0.5	17.1	BRR	1.08	0.38	4.8	4.4	9.2
				50	1.07	0.36	5.0	4.5	9.5
				1	1.07	0.36	5.3	4.5	9.8
				Jackknife	1.07	0.36	5.3	4.4	9.7
		0.2	18.8	BRR	1.11	0.39	4.5	5.7	10.2
				50	1.10	0.38	4.6	5.8	10.4
				1	1.10	0.38	4.6	5.8	10.4
				Jackknife	1.10	0.38	4.6	5.9	10.5

Table 2 (Cont.). Estimating the variance of the ratio of two normal variables from a stratified sample – comparison of bias and stability for BRR, Fay's method and the jackknife

Coefficient of variation per stratum (and overall) %			Overall cv on ratio %	Method or perturbation factor %	Variance estimates		Observed error rates for nominal 90% confidence intervals %		
					Bias	Stability	Left	Right	Total
Y (numerator)	X (denominator)	ρ							
24 (5.6)	100 (18.3)	0.8	10.8	BRR	1.26	1.81	0.9	8.3	9.2
				50	1.07	0.94	1.4	8.7	10.1
				1	1.02	0.81	1.3	9.3	10.6
				Jackknife	1.03	0.79	1.3	9.3	10.6
		0.5	12.4	BRR	1.27	1.30	1.6	6.2	7.8
				50	1.08	0.86	2.2	6.4	8.6
				1	1.03	0.75	2.3	6.1	8.4
				Jackknife	1.04	0.77	2.5	6.3	8.8
		0.2	13.4	BRR	1.28	1.16	0.6	8.2	7.8
				50	1.12	0.86	1.3	8.4	9.7
				1	1.08	0.84	1.1	8.5	9.6
				Jackknife	1.09	0.84	1.0	8.4	9.4
240 (56.1)	150 (27.5)	0.2	44.0	BRR	35.22	909.99	1.7	5.4	7.1
				50	1.17	1.40	2.9	6.1	9.0
				1	1.08	0.96	3.1	6.1	9.2
				Jackknife	1.09	0.96	2.7	6.1	8.8

Table 3. Estimating the variance of the regression of one normal variable on another from a stratified sample – comparison of bias and stability for BRR, Fay's method and the jackknife

Coefficient of variation per stratum (and overall) %			Overall cv on β %	Method or perturbation factor %	Variance estimates		Observed error rates for nominal 90% confidence intervals %		
					Bias	Stability			
Y (dependent)	*X* (independent)	ρ					Left	Right	Total
24	100	0.8	11.8	BRR	1.10	0.53	6.0	8.0	14.0
(5.6)	(18.3)			50	0.98	0.46	6.5	9.0	15.5
				1	0.94	0.44	7.0	9.0	16.0
				Jackknife	0.96	0.46	7.0	8.5	15.5
		0.5	25.0	BRR	1.07	0.50	5.5	6.5	12.0
				50	0.96	0.43	6.5	7.0	13.5
				1	0.93	0.43	6.5	7.5	14.0
				Jackknife	0.95	0.45	6.5	7.5	14.0
		0.2	69.0	BRR	1.15	0.65	7.0	7.0	14.0
				50	1.00	0.53	8.5	8.5	17.0
				1	0.96	0.50	9.0	8.5	17.5
				Jackknife	0.97	0.50	8.5	8.0	16.5
		−0.7	17.2	BRR	1.12	0.52	5.0	5.0	10.0
				50	1.00	0.45	5.5	5.5	11.0
				1	0.96	0.44	7.0	6.0	13.0
				Jackknife	0.99	0.48	6.5	6.0	17.5

Table 4. Estimating the variance of the median of one normal variable on another from a stratified sample – comparison of bias and stability for BRR, Fay's method and the jackknife

Coefficient of variation per stratum (and overall) %	Overall cv on median %	Method or perturbation factor %	Variance estimates		Observed error rates for nominal 90% confidence intervals %		
			Bias	Stability	Left	Right	Total
Y							
24	4.9	BRR	1.17	0.69	5.5	6.6	12.1
(5.6)		50	1.39	1.19	6.7	6.7	13.4
		1	2.16	4.69	17.3	14.7	32.0
		Jackknife	1.75	2.42	10.6	8.8	19.4
120	24.1	BRR	1.16	.68	6.0	4.8	10.8
(28.1)		50	1.35	1.15	7.3	5.7	13.0
		1	1.93	3.99	15.6	18.2	33.8
		Jackknife	1.72	2.50	9.9	9.6	19.5
240	48.7	BRR	1.12	.67	5.6	7.0	12.6
(56.1)		50	1.28	1.16	7.3	7.3	14.6
		1	1.99	5.61	18.7	17.7	36.4
		Jackknife	1.61	2.56	10.3	10.8	21.1

Table 5. Estimating the variance ratio of two normal variables and of the median of one normal variable from a stratified sample – comparison of bias, stability, and confidence interval coverage for BRR, Fay's method with various perturbation factors and the jackknife

Method or Perturbation Factor %	Evaluation for Ratio					Evaluation for Median				
	Bias	Stability	Observed error rates for nominal 90% confidence intervals %			Bias	Stability	Observed error rates for nominal 90% confidence intervals %		
			Left	Right	Total			Left	Right	Total
BRR	1.26	1.81	0.9	8.3	9.2	1.17	0.69	5.5	6.6	12.1
90	1.20	1.45	1.0	8.4	9.4	1.22	0.76	5.8	6.2	12.0
80	1.16	1.25	1.0	8.6	9.6	1.25	0.84	5.8	5.9	11.7
70	1.12	1.11	1.1	8.7	9.8	1.29	0.93	5.9	5.9	11.8
50	1.07	0.94	1.4	8.7	10.1	1.39	1.19	6.7	6.7	13.4
1	1.02	0.81	1.3	9.3	10.6	2.16	4.69	17.3	14.7	32.0
Jackknife	1.03	0.79	1.3	9.3	10.6	1.75	2.42	10.6	8.8	19.4

Note: Population variant corresponds to that used for the seventh set of numbers in Table 2 and first set of numbers in Table 4.

exception. For most common problems, all of the estimators work very well.

- Although no situations with fewer than 32 strata are reported in this paper, the results with fewer strata are similar to the results with higher variances.
- One-sided confidence interval coverage is quite poor for all the methods when the denominator has a greater cv than the numerator. This is caused by a high correlation between the estimated ratio and its estimated standard error, for this unusual situation.
- The bias estimates for the different methods are highly correlated with each other. Thus the relative biases between them are very accurate. The absolute level of the bias is less reliable because of instability in $M(\zeta)$. The same is true of the stability estimates. Similarly, the coverage properties of the confidence intervals may be accurately compared between the methods on the same population variant. The absolute level of coverage is less accurate.

Table 3 summarizes the results for the regression coefficient. Simulations for the regression coefficient were limited to the most interesting population variants from Table 2. Consistent with earlier research, there seems to be a tendency for BRR to overestimate variance and for the jackknife to underestimate it. Fay's method with a perturbation factor of 50% strikes a compromise between BRR and the jackknife that has smaller bias than either of them. Stability for Fay's method appears comparable to that for the jackknife. Confidence intervals for all the methods tend to be too optimistic. Among the methods, BRR appears to be marginally better than Fay's method or the jackknife in this respect.

Table 4 summarizes the results for the median. Results are almost completely turned around from those for the ratio. As is well known, the jackknife does not perform at all well for the median. Bias is high, stability is very poor, and nominal 90% confidence intervals do not cover the population median anywhere near 90% of the time. BRR is biased less seriously, has much better stability, and has confidence intervals that are pretty good. Again, Fay's method with a perturbation factor of 50% strikes a compromise between these extremes. On the other hand, it is important to note that for very small perturbation factors ($k = 0.99$), Fay's method is actually worse than the jackknife.

Having observed that Fay's method seems to be a reasonable compromise between BRR and the jackknife for the ratio, the regression coefficient, and for the median, the critical question becomes: "What perturbation factor should be used?" Table 5 attempts to shed some light on this question. It presents results for one of the more interesting population variants with BRR, the jackknife, and Fay's method with five different perturbation factors. As can be seen, perturbation factors in the range of 50 to 70% appear to be the most robust.

5. Adaptation for Collapsed Strata

Although BRR and Fay's method were developed for the case where two PSUs are selected with replacement per stratum, they are easily adapted to other situations. The case of one PSU per stratum is discussed below in some detail. The case of more then two PSUs per stratum is discussed in Wolter (1985) and in Rust (1986). The collapsed stratum variance estimator proposed here is very similar to that proposed in Hansen, Hurwitz, and Madow (1953, Vol. II, pp. 218–222). To assist comparisons, the nota-

tion has been kept nearly consistent. This type of variance estimator was evaluated by Shapiro, Singh, and Bateman (1980).

Suppose that the strata are collapsed into G groups with L_g strata in the gth group. Assume the L_g strata in each group are divided as evenly as possible into two half samples. If L_g is odd, then let the first half sample have K_g strata and the second half sample have M_g strata, where $M_g - K_g = 1$ and $M_g + K_g = L_g$. If L_g is even, we assume that each half sample has $M_g = K_g = L_g/2$ strata. Assume that there exists a variable which is known at the stratum level prior to data collection and is thought to be highly correlated with the characteristics to be observed in the survey. Let A_{gh} be the value of this variable for the hth half sample in the gth group. (In demographic surveys, A_{gh} is typically the total population in the half sample. Where a half sample consists of multiple strata, A_{gh} is formed by summing the stratum totals.) Let $A_g = A_{g1} + A_{g2}$.

The replicate weights for unit i in half sample 1 and unit j in half sample 2 are

$$\left[1 + \frac{L_g}{\sqrt{K_g M_g}} d_{tg}(1 - k) \frac{A_{g2}}{A_g}\right] W_{g1i}$$

and

$$\left[1 - \frac{L_g}{\sqrt{K_g M_g}} d_{tg}\,(1 - k) \frac{A_{g1}}{A_g}\right] W_{g2j},$$

where W_{g1i} and W_{g2j} are the unbiased weights. Note in particular that the adjustment factor for the first half sample involves the "size" (A_{g2}) of the second half sample. A caution: if k is picked too small, these weights can be negative. Also, note that if L_g is even and the half samples are equal in size, then these replicate weights reduce to the ordinary replicate weights for Fay's method:

$$[1 + d_{tg}(1 - k)]W_{g1i}$$

and

$$[1 - d_{tg}(1 - k)]W_{g2j}.$$

By definition, the variance estimate for a linear statistic is

$$\hat{v} = \frac{1}{(1 - k)^2} \frac{1}{T}$$

$$\times \sum_{t=1}^{T} \left\{ \sum_{g=1}^{G} \left[\left(1 + \frac{L_g}{\sqrt{K_g M_g}} d_{tg}(1 - k) \frac{A_{g2}}{A_g}\right) x_{g1} \right.\right.$$

$$\left.\left. + \left(1 - \frac{L_g}{\sqrt{K_g M_g}} d_{tg}(1 - k) \frac{A_{g1}}{A_g}\right) x_{g2} - x_g \right] \right\}^2$$

where T is the number of balanced replicates, x_{g1} and x_{g2} are unbiased estimates of the linear statistic for half-samples 1 and 2 in the gth group, and $x_g = x_{g1} + x_{g2}$.

Using the balancing property of the d_{tg}, it may be shown that

$$\hat{v} = \sum_{g=1}^{G} \frac{L_g^2}{K_g M_g} \left(\frac{A_{g2}}{A_g} x_{g1} - \frac{A_{g1}}{A_g} x_{g2} \right)^2.$$

The expected value of this variance estimator is

$$E(\hat{v}) = \sum_{g=1}^{G} \frac{L_g^2 A_{g1}^2 A_{g2}^2}{K_g M_g A_g^2}$$

$$\times \left[\left(\frac{X_{g1}}{A_{g1}} - \frac{X_{g2}}{A_{g2}} \right)^2 + \frac{\text{Var } x_{g1}}{A_{g1}^2} \right.$$

$$\left. + \frac{\text{Var } x_{g2}}{A_{g2}^2} \right]$$

where $X_{gi} = Ex_{gi}$.

The first term inside the brackets is usually very large so that the estimator is conservative. Obviously, the better the prior knowledge represented by A_{gi}, the smaller the bias in the variance estimate will be. Note that it is important for A_{g1} and A_{g2} to be roughly equal. (This is fairly easy to achieve for an even number of strata by sorting the strata on

A_{gi} and then grouping two by two.) If they are exactly equal, then the expected value reduces to

$$E(\hat{v}|A_{g1} = A_{g2}\ \forall\ g) = \sum_{g=1}^{G} \frac{L_g^2}{K_g M_g} \times \tfrac{1}{4}[(X_{g1} - X_{g2})^2 + \text{Var}\, x_{g1} + \text{Var}\, x_{g2}].$$

If, furthermore, the L_g are all even, then this simplifies to

$$E(\hat{v}|A_{g1} = A_{g2} \text{ and } L_g \text{ even } \forall\ g) = \sum_{g=1}^{G} [(X_{g1} - X_{g2})^2 + \text{Var}\, x_{g1} + \text{Var}\, x_{g2}] = \sum_{g=1}^{G} (X_{g1} - X_{g2})^2 + \text{Var}\, x.$$

The advantage to the special adjustment for odd L_g becomes apparent if the original strata are equal in size rather than the half samples. In this case, the size of each half sample is proportional to the number of original strata it contains. Thus,

$$E(\hat{v}|A_{g1} = K_g A_g/L_g \text{ and } A_{g2} = M_g A_g/L_g) = \sum_{g=1}^{G} \left[\left(\sqrt{\frac{M_g}{K_g}}\, X_{g1} - \sqrt{\frac{K_g}{M_g}}\, X_{g2}\right)^2 + \frac{M_g}{K_g}\text{Var}\, x_{g1} + \frac{K_g}{M_g}\text{Var}\, x_{g2}\right].$$

This value is intuitively appealing since the first term in brackets is likely to be closer to zero than $(X_{g1} - X_{g2})^2$ and since we may expect very roughly that

$$\frac{\text{Var}\, x_{g1}}{K_g} \approx \frac{\text{Var}\, x_{g2}}{M_g} \approx \frac{\text{Var}\, x_g}{L_g}$$

so that the sum of the second and third terms within the brackets may be approximated by

$$\frac{M_g\ \text{Var}\, x_g}{L_g} + \frac{K_g\ \text{Var}\, x_g}{L_g} = \frac{L_g}{L_g}\text{Var}\, x_g = \text{Var}\, x_g.$$

Deciding how many collapsed groups to form is a difficult question beyond the scope of this paper. (Rust and Kalton (1987) discuss this question in detail for the simplified case where $A_{gi} = 1$). One broad point to keep in mind is that the degrees of freedom increase with increase in G. Increasing the degrees of freedom improves the stability of the variance estimator, but it is easier to equalize A_{g1} and A_{g2}, and thus reduce the bias, with smaller G.

6. Conclusions and Future Study

Summing up from the work in this paper, Kovar (1985) and Kovar, Rao, and Wu (1988), there is no best resampling method for estimation of variances. For surveys where medians and other non-smooth statistics are not of interest, the jackknife is probably the best method. For surveys where estimation of such non-smooth statistics is important, BRR is a good choice provided that variance estimates for all domains have substantial degrees of freedom. If non-smooth statistics are of interest and there are domains where it is impossible to obtain adequate degrees of freedom, Fay's modification is the best method developed thus far provided that the perturbation factor is chosen in the range of 50 to 70%. Obviously, it would be interesting to extend this work to more populations, other statistics, and more values of the perturbation factor. Another interesting direction would be to synthesize Fay's method with the recent modification to BRR for the case

of more than two PSUs per stratum by Wu and Rao (1989). Most recently, Fay (1989) has shown how his method can be used when PSUs are selected without replacement.

7. References

Andersson, C., Forsman, G., and Wretman, J. (1987). Estimating the Variance of a Complex Statistic: A Monte Carlo Study of Some Approximate Techniques. Journal of Official Statistics, 3, 251–265.

Dippo, C.S., Fay, R.E., and Morganstein, D.H. (1984). Computing Variances from Complex Samples with Replicate Weights. Proceedings of the American Statistical Association, Section on Survey Research Methods, 489–494.

Fay, R. E. (1989). Theory and Application of Replicate Weighting for Variance Calculations. Proceedings of the American Statistical Association, Section on Survey Research Methods, forthcoming.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1963). Sample Survey Methods and Theory, Vol. 2, 218–222. New York: Wiley.

Hansen, M.H. and Tepping, B.J. (1985). Estimation of Variance in NAEP. Unpublished manuscript.

Judkins, D.R. (1989). Letter to the Editor. Journal of Official Statistics, 5, 293–294.

Kovar, J. (1985). Variance Estimation of Nonlinear Statistics in Stratified Samples. Methodology Branch Working Paper No. 85-052E, Statistics Canada.

Kovar, J.G., Rao, J.N.K., and Wu, C.F.J. (1988). Bootstrap and Other Methods to Measure Errors in Survey Estimates. The Canadian Journal of Statistics, 16 (Supplement), 25–46.

McCarthy, P.J. (1966). Replication: An Approach to the Analysis of Data from Complex Surveys. Vital and Health Statistics, ser. 2, no. 14, National Center for Health Statistics. Washington, D.C: U.S. Government Printing Office.

Rao, J.N.K. and Wu, C.F.J. (1985). Inference from Sample Surveys: Second Order Analysis of Three Methods for Nonlinear Statistics. Journal of the American Statistical Association, 80, 620–630.

Rust, K. (1986). Efficient Replicated Variance Estimation. Proceedings of the American Statistical Association, Section on Survey Research Methods, 81–87.

Rust, K. and Kalton, G. (1987). Strategies for Collapsing Strata for Variance Estimation. Journal of Official Statistics, 3, 69–81.

Shapiro, G.M., Singh, R.P., and Bateman, D. (1980). Empirical Research Involving an Alternative Variance Estimator to the Collapsed Stratum Variance Estimator. Proceedings of the American Statistical Association, Section on Survey Research Methods, 793–798.

Wolter, K.M. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.

Woodruff, R.S. (1952). Confidence Intervals for Medians and Other Position Measures. Journal of the American Statistical Association, 47, 635–646.

Wu, C.F.J. and Rao, J.N.K. (1989). Pseudo-replication Methods for Complex Survey Data. Proceedings of the American Statistical Association, Section on Survey Research Methods, forthcoming.

Received May 1988
Revised February 1990