A Simulation Study of the Distribution of Fay's Successive Difference Replication Variance Estimator

Elizabeth T. Huang and William R. Bell*
Statistical Research Division, U.S. Census Bureau, Washington, DC 20233

Abstract
Small area estimation with area level models requires variance estimates of the direct survey point estimates being modeled. Small area direct variance estimates are likely to be unstable, suggesting modeling the variances to improve them. One aspect of such modeling would be to specify a probability distribution of the variance estimators. Here, we consider this for Fay's successive difference replication variance estimator. More specifically, we examine via simulations whether the variance estimator could be assumed to approximately follow a scaled chi-squared distribution, and if so, with what value of the degrees of freedom? We study these questions for simple random samples of various sizes from various distributions (normal, Poisson, and Bernoulli). The motivation for this study comes from county modeling of ACS (American Community Survey) poverty estimates by the Census Bureau's Small Area Income and Poverty Estimates program, as direct variances of the ACS poverty estimates are produced using Fay's variance estimator.

Key Words: survey variance, degrees of freedom, chi-squared distribution, sample size

1. Introduction

Small area estimation with area level models requires variance estimates of the sampling errors in the direct survey point estimates being modeled. Small area direct sampling variance estimates are likely to be unstable, suggesting modeling the variances to improve them. This might be accomplished by simply fitting a generalized variance function (GVF) to the direct variance estimates to attempt to remove some of the noise in the estimates. The GVF can be thought of as parameterizing the mean function of the direct variance estimates, and if the direct variance estimates are approximately unbiased, then the GVF approximates the true variances. Some authors (e.g., Otto and Bell 1995, Arora and Lahiri 1997, Gershunskaya and Lahiri 2005, You and Chapman 2006) have gone further and specified full probability models for direct survey variance estimates. The latter three papers used models that assumed the direct survey variance estimates were unbiased and followed scaled chi-squared distributions, generally with known degrees of freedom. Otto and Bell modeled direct estimates of sampling error covariance matrices assuming a Wishart distribution (the multivariate version of the chi-squared distribution), and also estimated the degrees of freedom as part of the model fitting.

Suppose the characteristic of interest is the population mean, estimated by the mean of a simple random sample of size n. If the population unit level data are independent and identically normally distributed, then the usual estimate of the variance of the mean is well-known to be distributed proportional to a chi-squared random variable with $n-1$ degrees of freedom. The situation in practice is rarely this simple, however. The sampling may be complex, the unit level data may not be

*Disclaimer: This report is released to inform interested parties of ongoing research and to encourage discussion. The views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

normally distributed, and the point estimator may be more complicated than the sample mean. Also, survey variance estimators are typically more complicated than the usual sample variance estimate, e.g., they may be based on a replication method such as jackknife or bootstrap. Thus, in practice, the assumption that sample variance estimates follow a scaled chi-squared distribution would be an approximation, probably of unknown accuracy. Furthermore, even if the chi-squared approximation were adequate, the appropriate value of the degrees of freedom may be unknown.

There appears to have been little if any study of the accuracy of chi-squared approximations to the distribution of survey variance estimators, of the conditions under which such approximations are reasonably accurate and conditions under which they are not, and of the appropriate value of the degrees of freedom and how this varies over changes in the distribution of the data and changes in the nature of the survey, such as variations in sample size. In this paper we begin a small investigation of some of these issues for Fay's successive difference replication variance estimator (Fay and Train 1995). We examine, via simulations, the distribution of Fay's estimator of the variance of the sample mean under simple random sampling (*srs*) from populations with different distributions of the unit level data—normal, Bernoulli (0-1) with various success probabilities, and Poisson with various occurrence rates. For various sample sizes we examine whether the variance estimator appears distributed approximately proportional to chi-squared and how the degrees of freedom varies with sample size.

A primary motivation for this investigation is the study of models for variance estimates of direct county poverty estimates from the American Community Survey (ACS) that are used in area level models by the Census Bureau's Small Area Income and Poverty Estimates (SAIPE) program. Such models for variance estimates of county estimates of the number of school-age (5-17) children in poverty are explored by Maples, Bell, and Huang (2009). Direct variances of the ACS poverty estimates are produced using Fay's variance estimator. For more discussion of the ACS methodology, including the variance estimation, see U.S. Census Bureau (2009).

Sections 2 and 3 of this paper review Fay's successive difference replication variance estimator and derive some of its properties when applied to estimating the variance of the mean of independent identically distributed (*i.i.d.*) data. Section 4 presents simulation results studying the behavior of this variance estimator for data simulated from normal, Bernoulli, and Poisson populations. We examine bias of the variance estimator, how well its distribution resembles a chi-squared distribution, and the degrees of freedom of a chi-squared approximation. In addition, Section 4 provides such results for samples drawn from an artificial population constructed by pooling 2005 ACS sample data over several counties from the state of Maryland. Section 5 then offers some conclusions.

2. Fay's Successive Difference Replication Variance Estimator

Fay and Train (1995) discuss the successive difference replication variance estimator and its application to CPS (Current Population Survey) Annual Social and Economic Supplement data. Until 2005, CPS data were used as the source of direct state and county poverty estimates that formed the basis of the SAIPE models. U.S. Census Bureau (2009, chapter 12) discusses application of the successive difference replication variance estimator to ACS data. Though the ACS and CPS designs differ substantially, the application of the variance estimator is essentially

similar though based on different numbers of replicates—80 for ACS versus 160 for CPS. Here we review the variance estimator, expressing it in a form that facilitates derivation of some of its properties.

Suppose the quantity being estimated, which we denote by Y_0, is the population mean, and that it is being estimated by the sample mean, $\hat{Y}_0 = \bar{y} = n^{-1}\sum_{i=1}^{n} y_i$, where $y_1, \ldots, y_n$ are the values of the units in the sample of size n. To estimate the variance we form replicate estimates $\hat{Y}_r$ defined by

$$\hat{Y}_r = \frac{1}{n}\sum_{i=1}^{n} f_{ir} y_i = \frac{1}{n}\mathbf{f}_r'\mathbf{y} \qquad r = 1, \ldots, R \tag{1}$$

where the f_{ir} are the replicate factors, $\mathbf{f}_r = (f_{1r}, \ldots, f_{nr})'$, $\mathbf{y} = (y_1, \ldots, y_n)'$, and R is the number of replicates. For now we assume that $R = 4k$ for some k, and that $R \geq n + 2$. The replicate factors are defined by

$$f_{ir} = 1 + \frac{1}{2\sqrt{2}}(a_{i+1,r} - a_{i+2,r}) \qquad i = 1, \ldots, n$$

for $r = 1, \ldots, R$, and where $A = [a_{ij}]$ is an $R \times R$ Hadamard matrix, which has the orthogonality property that $AA' = R \times I_R$ (Plackett and Burman 1946). Note that only rows 2 through $n + 2$ of the Hadamard matrix are used in constructing the f_{ir}; the first row of A is generally a vector of ones and is not used. The variance estimator of $\hat{Y}_0$ is then

$$v_r = \frac{4}{R}(1 - f)\sum_{r=1}^{R}(\hat{Y}_r - \hat{Y}_0)^2 \tag{2}$$

where $f = n/N$ is the sampling fraction and N is the population size.

More generally, $\hat{Y}_0$ could be a survey weighted estimator, $\hat{Y}_0 = \sum_{i=1}^{n} w_i y_i$, and then for the replicate estimates the replicate factors multiply the survey weights, i.e., $\hat{Y}_r = \sum_{i=1}^{n} f_{ir} w_i y_i$. Even more generally, $\hat{Y}_0$ could be a nonlinear function of such survey weighted totals, in which case the replicate estimates would be constructed analogously by multiplying the survey weights in the estimated totals by the replicate factors. Here we examine the simple case of the sample mean ($w_i = n^{-1}$) because we can readily examine some properties of v_r for this case.

From (1) we note that the replicate estimates are linear functions of $\mathbf{y}$. Collecting all these into a vector $\hat{\mathbf{Y}} = (\hat{Y}_1, \ldots, \hat{Y}_R)'$, we can write this as $\hat{\mathbf{Y}} = n^{-1}F\mathbf{y}$, where $F = [\mathbf{f}_1, \ldots, \mathbf{f}_R]'$ is an $R \times n$ matrix. We can then write

$$F = \mathbf{1}_R\mathbf{1}_n' + \frac{1}{2\sqrt{2}}\tilde{A}'D'$$

where $\mathbf{1}_n$ is the $n \times 1$ vector of ones, $\tilde{A}$ consists of rows 2 through $n + 2$ of the Hadamard matrix A, and D is the $n \times (n + 1)$ differencing matrix defined by

$$D = \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -1 \end{bmatrix}.$$

The replicate variance estimator can then be written

$$\begin{aligned} v_r &= \frac{4}{R}(1-f)\left\|\hat{\mathbf{Y}}-\hat{Y}_0\mathbf{1}_R\right\|^2 \\ &= \frac{4}{R}(1-f)n^{-2}\left\|(F-\mathbf{1}_R\mathbf{1}_n')\mathbf{y}\right\|^2 \\ &= \frac{1-f}{2Rn^2}\mathbf{y}'\mathbf{Q}\mathbf{y} \end{aligned} \tag{3}$$

where $\mathbf{Q} = D\tilde{A}\tilde{A}'D'$. From the orthogonality property of the rows of the Hadamard matrix A, it follows that $\tilde{A}\tilde{A}' = R \times I_{n+1}$, so that $\mathbf{Q} = R \times DD'$ and, in this case ($n \leq R-2$), the Hadamard matrix A is not actually needed. Fay and Train (1995) obtained a closely related result to show that a related variance estimator reduces to the successive difference variance estimator (Wolter 1985).

The preceding leaves open the question of what happens when the sample size n exceeds $R-2$, assuming the number of replicates R to be fixed at some value. In this case the matrix $\tilde{A}$ is obtained by repeating rows of the Hadamard matrix A (continuing to leave out the first row of ones), and then $\mathbf{Q}$ does not reduce to $R \times DD'$. In practice at the Census Bureau this repetition of rows of A is not done in the numerical order of the rows, but follows a more involved pattern (Navarro 2001). Also, one additional row of A is omitted from $\tilde{A}$.

The expression (3) can be used to establish some properties of v_r for the case where $y_i \sim i.i.d.\ (\mu, \sigma^2)$. First, note that $E(v_r) = [(1-f)/(2Rn^2)]tr[\mathbf{Q}E(\mathbf{y}\mathbf{y}')] = [(1-f)/(2Rn^2)][\sigma^2 tr(\mathbf{Q}) + \mu^2\mathbf{1}_n'\mathbf{Q}\mathbf{1}_n]$. For the case of $n \leq R-2$, so $\mathbf{Q} = R \times DD'$, one can easily show that $tr(\mathbf{Q}) = 2nR$ and $\mathbf{1}_n'\mathbf{Q}\mathbf{1}_n = 2R$. Then

$$E(v_r) = \frac{1-f}{2Rn^2}[\sigma^2 2nR + 2R\mu^2] = \frac{1-f}{n}\sigma^2 + \frac{1-f}{n^2}\mu^2. \tag{4}$$

Since $\mathrm{Var}(\hat{Y}_0) \equiv \mathrm{Var}(\bar{y}) = (1-f)\sigma^2/n$, we see that v_r is unbiased if and only if $\mu = 0$. The relative bias in v_r is easily shown to be $(\mu^2/\sigma^2)/n$, which depends on μ^2/σ^2 and decreases proportional to n^{-1}. The bias in v_r should thus be small for moderate to large samples (though this result requires $n \leq R-2$).

For the case where $\mu = 0$, and writing $\mathbf{Q} = [q_{ij}]$, one can show that

$$\mathrm{Var}(\mathbf{y}'\mathbf{Q}\mathbf{y}) = 2\sigma^4 tr(\mathbf{Q}^2) + (q_{11}^2 + \cdots + q_{nn}^2)[E(y_i^4) - 3\sigma^4]. \tag{5}$$

From (3), we then have that $\mathrm{Var}(v_r) = [(1-f)/(2Rn^2)]^2\mathrm{Var}(\mathbf{y}'\mathbf{Q}\mathbf{y})$. For the case of $n \leq R-2$, it turns out that this simplifies to

$$\mathrm{Var}(v_r) = \frac{(1-f)^2}{n^3}\left[E(y_i^4) - \frac{\sigma^4}{n}\right]. \tag{6}$$

The relative variance of v_r is then

$$\mathrm{RelVar}(v_r) \equiv \frac{\mathrm{Var}(v_r)}{[E(v_r)]^2} = \frac{1}{n}\left[\frac{E(y_i^4)}{\sigma^4} - \frac{1}{n}\right]. \tag{7}$$

Formulas analogous to (5)–(7) could be developed for the case of $\mu \neq 0$, but these would be complicated and thus not particularly instructive. For the case where $y_i \sim i.i.d.\ N(0, \sigma^2)$, $E(y_i^4) = 3\sigma^4$, so that (7) then simplifies to

$$y_i \sim i.i.d.\ N(0,\sigma^2) \Rightarrow \mathrm{RelVar}(v_r) = \frac{1}{n}\left[3 - \frac{1}{n}\right]. \tag{8}$$

3. A Working Model for the Variance Estimator

As noted earlier, for *i.i.d.* normal data the usual unbiased estimator of σ^2, $s^2 = (n-1)^{-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$ is distributed as $(n-1)s^2/\sigma^2 \sim \chi^2_{n-1}$. This suggests considering whether other variance estimators, such as v_r, might be assumed distributed proportional to χ^2_k for some value of the degrees of freedom k.

In the case of v_r this would need to be an approximation due to a well-known result (e.g., Rao 1973, p. 186), which states that when the y_i are *i.i.d.* $N(0,1)$, $\mathbf{y'Qy} \sim \chi^2_k$ if and only if $\mathbf{Q}$ is idempotent with $k = \text{rank}(\mathbf{Q}) = tr(\mathbf{Q})$. If $\mathbf{Q}$ is instead proportional to an idempotent matrix, or if $\mathbf{Q}$ is idempotent but $\sigma^2 \neq 1$, then $\mathbf{y'Qy}$ would be distributed proportional to χ^2_k. It is easily checked, however, that the matrix $\mathbf{Q}$ used in Section 2 in the definition of v_r is not proportional to an idempotent matrix. In fact, for the case of $n \leq R-2$, $\mathbf{Q} = R \times DD'$, and DD' is a modification of the second difference matrix, which features a band with elements $[-1, 2, -1]$ centered about the diagonal in all but the first and last row. In contrast, $(DD')^2$ is a modification of the fourth difference matrix, which features a diagonal band with elements $[1, -4, 6, -4, 1]$ in all but the first two and last two rows. Hence, v_r cannot be exactly distributed proportional to χ^2_k even for *i.i.d.* $N(0,1)$ data. Still, we can ask whether this would be approximately the case. We use simulations to examine this question in the next section, with both normal and non-normal data (the latter presumably leading to a further degree of approximation.)

If a χ^2_k distribution provides a reasonable approximation, there is still the question of what is the appropriate value of k? Recall that $E(\chi^2_k) = k$ and $\text{Var}(\chi^2_k) = 2k$, so that $\text{RelVar}(\chi^2_k) = 2k/k^2 = 2/k$, and hence $k = 2/\text{ RelVar}(\chi^2_k)$. Thus, given $\text{RelVar}(v_r)$ we can use this result – the "Satterthwaite approximation" (Ames and Webster 1991) – to determine the degrees of freedom for the chi-squared approximation. For $N(0,\sigma^2)$ data, equation (8) shows that we would have approximately (ignoring the n^{-2} term) $k = (2/3)n$, at least up through $n = R-2$. What happens for $n > R-2$ is unclear, though we would presume that k would not exceed the number of replicates R. For non-normal data with mean zero, the behavior of k as n increases could be obtained using the more general equation (7) (again for n up through $R-2$). However, since most of our simulations in the next section involve data with nonzero means, we shall also use these simulations to determine $\text{RelVar}(v_r)$ and thus k.

4. Simulation Study of Fay's Successive Difference Replication Variance Estimator with Samples from Various Populations

To study properties of the successive difference replication variance estimator v_r (of the sample mean, $\bar{y}$), we simulated data from various distributions to create artificial populations, drew a large number (10,000) of simple random samples of various sizes from each population, computed v_r for each sample, and examined the behavior of v_r for a given sample size over the simulations. In all cases the population size was $N = 10,000$. We know the true variance of $\bar{y}$ from a sample of size n is $n^{-1}(1-f)S^2$ where $f = n/N$ is the sampling fraction and $S^2 = (N-1)^{-1}\sum_{i=1}^{N}(y_i - \bar{Y})^2$ is the population variance with $\bar{Y}$ the population mean (Cochran 1977, p. 23). We can compute S^2 from the simulated population (since N is large $S^2 \approx \sigma^2 \equiv \text{Var}(y_i)$), and can thus examine the percent relative bias of v_r, computed as

$$\text{Percent Rel Bias } (v_r) = 100 \times \frac{E(v_r) - \text{ Var}(\bar{y})}{\text{Var}(\bar{y})} \tag{9}$$

where $E(v_r)$ is estimated by averaging v_r over the simulations. We can also estimate $\text{Var}(v_r)$ from the simulations, hence estimate $\text{RelVar}(v_r)$ and $\text{CV}(v_r) = \sqrt{\text{RelVar}(v_r)}$, and from this compute the approximate degrees of freedom $k = 2/\text{RelVar}(v_r)$.

Having determined k, we can also examine how closely the distribution of $k \times v_r/E(v_r)$ matches a χ^2_k distribution. For this let $G(v)$ be the empirical cumulative distribution function (cdf) of $k \times v_r/E(v_r)$ over the simulations, and let $F(v)$ be the corresponding χ^2_k cdf. To judge how well $F(v)$ approximates the true distribution, we use a modified version of the Kolmogorov-Smirnov statistic (Rao 1973, p. 421):

$$\text{K-S} = \sup_v |G(v) - F(v)|. \tag{10}$$

The modification is to omit the normalization by square root of sample size, where in this context the sample size would be the number of simulations (10,000). This normalization would be needed for hypothesis testing, but that is not of interest here since we know from Section 3 that the true distribution of $k \times v_r/E(v_r)$ is not exactly χ^2_k. We are using K-S from (10) merely to get an approximate measure of the difference between the true distribution of $k \times v_r/E(v_r)$ (approximated by $G(v)$ from the large number of simulations) and the approximating χ^2_k distribution. We shall take .10 as a rough criterion value for K-S to indicate when the χ^2_k approximation seems reasonable. So if K-S is substantially less than .10 we would be very satisfied with the χ^2_k approximation, while if K-S were substantially more than .10 we would regard the χ^2_k approximation as inadequate. Values of K-S near .10 are marginal.

In the subsections to follow we report results from simulations using the following population distributions: Normal$(0, 1)$; Bernoulli(p) with $p = 0.1$, 0.25, or 0.5; and Poisson(μ) with $\mu = 0.02$, 0.1, 0.3, or 0.5. We also used some data from the 2005 ACS to create an artificial population from which we drew samples. For all these population distributions we drew samples of various sizes from $n = 2$ to $n = 760$. For each sample size we computed the following quantities from the values of v_r over the simulated samples: (i) K-S from (10), (ii) $\text{CV}(v_r) = \sqrt{\text{RelVar}(v_r)}$, (iii) the degrees of freedom $k = 2/\text{RelVar}(v_r)$, and (iv) Percent Rel Bias (v_r) from (9). We present these results in a set of graphs that plot these statistics against sample size, with one graph for each statistic, and with one set of these four graphs for each population distribution. For the Bernoulli and Poisson cases, each graph contains a different curve for each different parameter value used. We examine the graphs to determine, for each population distribution, for what sample sizes the χ^2_k approximation seems reasonable (judged by comparing K-S to .10), for what sample sizes v_r is approximately unbiased, and how $\text{CV}(v_r)$ and the degrees of freedom k vary with sample size.

Note that the interpretation of k as degrees of freedom is only appropriate for cases where the χ^2_k approximation seems reasonable. Note also that we expect that the degrees of freedom should not exceed the number of replicates (here 80). Recall that, in our application of v_r, two rows of the 80×80 Hadamard matrix A are not used in constructing the replicates, requiring us to repeat rows of A for $n > 78$. This suggests that the expected upper limit on k should probably be 78 not 80. It also raises some interesting questions about the behavior of k as n increases. Does k approach 78 and then stop increasing? Does k increase steadily up to $n = 78$ but then stop increasing or increase more slowly, for $n > 78$? Does k even approach 78 within the sample sizes considered?

In the figures showing the simulation results the four graphs on each page are arranged as follows: The graph of K-S values is in the upper left, the graph of

degrees of freedom is in the upper right, the graph of $\mathrm{CV}(v_r)$ is in the lower left, and the graph of the percent relative bias of v_r is in the lower right.

4.1 Simulation results for the Normal$(0, 1)$ population

For this case and for $n \leq 78$, some of our results of interest follow exactly from results given in Section 2. In particular, since $\mu = 0$ we know from equation (4) that v_r is unbiased for these n. Also, equation (8) shows that $\mathrm{RelVar}(v_r) = n^{-1}(3 - n^{-1})$, so that $\mathrm{CV}(v_r) \approx \sqrt{3/n}$ and

$$k = \frac{2}{\mathrm{RelVar}(v_r)} = \frac{2n}{3 - \frac{1}{n}} \approx \frac{2}{3}n. \tag{11}$$

At $n = 78$ this gives $k \approx 52$. Results for $n > 78$ and for K-S for all n can be obtained from the simulations. Changing the variance would not affect any of the results presented here, so they apply to $N(0, \sigma^2)$ populations not just $N(0, 1)$. The restriction to $\mu = 0$ is important, however, as follows from equation (4).

The graphs of the simulation results with the $N(0, 1)$ data are given by Figure 1. We discuss them in their clockwise order, starting from the plot of K-S values in the upper left. The circles in the plots give the values determined from the simulations for each sample size. In addition, for this case we have added to the plots solid curves showing the corresponding theoretical results for the usual variance estimator for this case, $s^2(1 - f)/n$. This variance estimator is known to be unbiased and distributed proportional to χ^2_{n-1} for all n. Hence, the "curves" in the K-S plot and relative bias plot are just the horizontal axis at 0, and the curve in the degrees of freedom plot is the straight line with intercept -1 and slope 1. The corresponding curve in the CV plot is $\sqrt{2/(n-1)}$. The latter two results provide a comparison of interest with the results for v_r.

The plot of K-S values shows that they are below 0.03 for all sample sizes (2 to 760) considered. Thus, for $N(0, \sigma^2)$ populations, the distribution of $k \times v_r/E(v_r)$ is well approximated by the chi-squared distribution.

The plot of the degrees of freedom for $n \leq 78$ is consistent with the relation (11), and in fact at $n = 78$ the simulation results give $k = 52.9$. For $n > 78$ the degrees of freedom continue to increase, though more slowly and not as regularly. The maximum degrees of freedom is 74.4 for sample size of 760. Compared to the usual variance estimator for this case, the degrees of freedom of v_r is fewer by 1/3 up to $n = 78$, and for larger n is more substantially lower.

The plot of the CVs of v_r starts at 1.13 for $n = 2$, and deceases to 0.31, 0.20, and 0.16 for sample sizes of 30, 75, and 760, respectively. Interestingly, the CV of the usual variance estimator is higher for $n = 2$ (at 1.42), though it decreases more rapidly with n to values of 0.26, 0.16, and 0.05 for sample sizes of 30, 75, and 760, and the CV approaches 0 as $n \to \infty$. In contrast, the simulation results suggest that the CV of v_r will not decline much below 0.16 as n continues to increase.

Since v_r is known to be unbiased for $n \leq 78$, the deviations from 0 over this range of sample sizes in the plot of percent relative biases are due to simulation error. For $n > 78$ the percent relative biases from the simulations are still small, never exceeding 0.5%. So, for normal data with mean 0, in addition to being exactly unbiased for $n \leq 78$, we can conclude that v_r can be regarded as essentially unbiased with $n > 78$.

4.2 Simulation results for the Bernoulli populations

Figure 2 plots the simulation results when v_r was applied to simple random samples drawn from Bernoulli(p) populations for p values of 0.1, 0.25, and 0.5. In these graphs the black solid line is for $p = 0.1$, the green dashed line is for $p = 0.25$, and the red dot-dashed line is for $p = 0.5$. The plot of K-S values shows that a scaled χ^2_k distribution provides a poor approximation to the distribution of v_r for small to moderate sample sizes. The sample sizes needed for K-S values to become smaller than 0.1 are 90, 80, and 78, for $p = 0.1$, 0.25, and 0.5, respectively. The computed values of k are thus not really interpretable as degrees of freedom much below these values of n where the χ^2 approximation is poor.

In the plot of the degrees of freedom we see for $p = 0.1$ and $p = 0.25$ that k mostly increases with n. The increase in k is fairly smooth for $p = 0.1$, with k reaching a maximum value of about 61 at $n = 760$, though it appears k would continue to increase beyond that point. For $p = 0.25$ the increase in k is nearly linear up to $n = 78$, for which k is about 66. For $n > 78$ there is an overall general increase (with some undulations) up to around $n = 400$, followed by a leveling off, with a maximum value for k of around 77. While the behavior of k for $p = 0.25$ is reminiscent of the results for $N(0, 1)$, the behavior of k when $p = 0.5$ is rather odd. It increases faster than n up to a maximum of 157 at $n = 75$, and then decreases, first rapidly, then more slowly. After about $n > 400$, it approximately stabilizes with values of k ranging from around 81 to around 85. We have no concrete explanation for this unusual behavior, but would note again that, in this case, for small values of n the χ^2 approximation is quite poor.

In the CV plot the curve for $p = 0.1$ appears uniformly higher than that for $p = 0.25$, which appears uniformly higher than that for $p = 0.5$, suggesting that the precision of v_r increases with p for a given n. For a given p, the CVs decrease mostly monotonically with increasing n, though with a few points of exception where the CV increases. The most pronounced exceptions occur just after $n = 78$, and are largest for $p = 0.5$. Due to the scale of the CV plot, the undulations there are not so readily apparent as they are in the plot of $k = 2/\ \mathrm{CV}^2$. (These two graphs for $p = 0.5$ may thus appear inconsistent, though they are, in fact, consistent.)

Since $E(y_i) = p \neq 0$, we know from equation (4) that v_r is biased. The plot of the relative biases shows that the bias is large for small n and large p, but is otherwise negligible. Where the bias is not negligible it is positive. Still, even for $p = 0.5$ the bias goes below 5% for $n \geq 20$. For $p = 0.25$ or 0.10 the bias is appreciable only for very small n.

4.3 Simulation results for the Poisson populations

Figure 3 plots the simulation results when v_r was applied to simple random samples drawn from Poisson(μ) populations for μ values of 0.02, 0.1, 0.3, and 0.5. In these graphs the black solid line is for $\mu = 0.02$, the red small dashed line is for $\mu = 0.1$, the blue dot-dashed line is for $\mu = 0.3$, and the green dashed line is for $\mu = 0.5$. The plot of K-S values shows decreasing values with increasing n except for $\mu = 0.02$, for which there is initially (up to $n = 10$) an increase in the K-S values, which is then followed by a sharp decrease. In this case the K-S value does not drop below 0.1 until $n = 300$. For $\mu = 0.1$, 0.3, and 0.5, the K-S values initially decrease more sharply with increasing n, and drop below 0.1 when $n = 50$, 16, and 10, respectively. Thus, the smaller the μ value is, the larger is the sample size needed for the distribution of v_r to be well-approximated by a scaled chi-squared distribution.

The plot of the degrees of freedom shows that, for all values of μ considered, the value of k increases fairly steadily with n, though for $\mu = 0.3$ and $\mu = 0.5$ it appears that k may level off shortly beyond $n = 760$. The maximum values of k are about 26 ($\mu = 0.02$), 52 ($\mu = 0.1$), 65 ($\mu = 0.3$), and 69 ($\mu = 0.5$). Correspondingly, the plot of CV values shows these to initially decrease rapidly with increasing n, then level off, with lower CVs for the higher values of μ.

The plot of the relative biases of v_r shows some large values for $\mu = 0.5$, but only for very small sample sizes. The largest relative bias is 26% for $n = 2$, but it drops below 5% at $n = 15$, and continues to drop towards 0 (though the plotted values aren't 0 due to some simulation error). As would be expected from equation (4), the relative biases are smaller for smaller values of μ.

4.4 Simulations results using ACS 2005 data

The ACS is a nationwide survey designed to provide annual estimates of population and housing characteristics nationally and for states, counties, and other substate areas. SAIPE uses state and county poverty estimates constructed from ACS data as the basis for its state and county poverty models. Here we use some ACS micro data from 2005 to create an artificial population from which we can repeatedly draw samples, create sample based poverty estimates, and apply v_r to estimate the variances of the poverty estimates. We then study the properties of v_r over the samples drawn. Relative to the results just presented which used simulated data, this approach has the advantage of using an artificial population that more accurately reflects the real properties of the ACS poverty data.

ACS has an annual national sample size of about 3,000,000 addresses. Rather than use the full national sample, we wanted a more homogeneous artificial population of the sort that might be present for an actual county. We also wanted to create an artificial population large enough to support drawing samples of a substantial size (we again used 760 as the largest sample size) without this resulting in an unduly large sampling fraction. To achieve these goals we combined data from 19,264 ACS 2005 sample households taken from Maryland's 5 largest county equivalents (Anne Arundel County, Baltimore County, Montgomery County, Prince George's County, and Baltimore city) to define our artificial population.

For the target population parameter of interest we used the mean number per household of 5-17 year-old related (to the head of the household) children in poverty. (This is $\bar{Y}$ if, for each household i in the artificial population that is in poverty, y_i is defined as the number of 5-17 related children, and for all other households y_i is defined as 0.) This characteristic was chosen to provide some comparability with the previous results (which involved variances of sample means) and because multiplying this quantity by the number of households would produce the number of 5-17 related children in poverty, a key characteristic for the SAIPE models. We drew 10,000 simple random samples of households from this population, formed estimates of the target characteristic, and computed variance estimates using v_r. A couple points are worth noting. First, only 26.3% of the households in our artificial population include any related 5-17 year-old children. Second, only about 2% of households in the artificial population both are in poverty and have a related 5-17 year-old.

Figure 4 plots, for various sample sizes, the usual four statistics on v_r, obtained from the samples drawn from the artificial population as just described. These are indicated by the circles on the plots. We also applied the usual variance estimator

defined by $s^2(1-f)/n$. Results for this variance estimator are shown by the solid lines on the plots. The two sets of results are very similar. The usual variance estimator is slightly more precise, as can be seen for the larger sample sizes in the plot of degrees of freedom (though this is hard to spot in the corresponding plot of CVs of the variance estimators.) Because the two sets of results are so similar, we shall not comment further on the results for the usual variance estimator.

The pattern of the plot of the K-S values for v_r is similar to that for the simulated Poisson(0.02) population, but the K-S values are smaller. The K-S value of v_r is 0.13 for $n = 2$, increases to a maximum of 0.46 for $n = 25$, and then decreases sharply with leveling off starting around $n = 200$. The K-S values fall below 0.1 for $n \geq 120$.

The degrees of freedom for the chi-squared approximation increase very nearly linearly with n, but increase slowly, reaching a maximum of only 9.9 for $n = 760$. These values are even lower than those for the simulated Poisson(0.02) population. The CVs show the corresponding rapid decrease with increasing n followed by a leveling off.

The percent relative biases of v_r are negligible except possibly for a few of the smallest sample sizes, and even there only three of the values exceed 4%.

5. Conclusions

We have examined some properties of Fay's successive difference replication variance estimator v_r for estimating the variance of the sample mean from simple random samples. We did this by creating artificial populations simulated from various distributions (normal, Bernoulli, and Poisson), and also by taking sample poverty data for a group of counties from the 2005 ACS to define an artificial population. Our goals were to examine, for these different populations and for various sample sizes n, (1) whether v_r could be regarded as distributed approximately proportional to χ^2_k, (2) how the degrees of freedom k varied with n, and (3) whether v_r had an appreciable bias.

With large samples (say, $n \geq 100$) the simulation results are mostly supportive of using a chi-squared approximation to the distribution of v_r. With smaller samples the results are less clear, being dependent on the form of the distribution of the data and its parameter values. The chi-squared approximation was poorest for the Poisson data with very small mean, the artificial population constructed from ACS poverty data, and the Bernoulli data for all values of the success probability. For normally distributed data with mean zero the chi-squared approximation was good for all sample sizes, and for Poisson data with a substantial mean the approximation was good for all but the very smallest sample sizes.

The degrees of freedom, k, of v_r for the chi-squared approximation (when it seemed reasonable) varied substantially across the various cases considered. While k did generally increase with sample size, it increased at very different rates for various cases. Two things that could be said are (i) k increased more slowly than n (it increased at rate $(2/3)n$ for normal data with mean zero, but at a much slower rate in many other cases), and (ii) k generally remained less than the number of replicates used for v_r (here 80), even as n increased to large values (here up to $n = 760$). An isolated exception to these conclusions occurred for the Bernoulli(0.5) data. Though we have no concrete explanation for this exception, in this case for small n the chi-squared assumption was poor anyway.

Bias of v_r appeared to be of little concern in our results, with a bias of substance

appearing only for a few cases with very small sample sizes (n smaller than 15).

This investigation could be extended in any of four directions: (i) consideration of other variance estimators, (ii) consideration of variance estimators for population characteristics other than the mean, (iii) variance estimation with survey designs other than simple random sampling, and (iv) consideration of some additional population distributions beyond those examined here.

REFERENCES

Ames, Michael H. and Webster, John T. (1991), "On Estimating Approximate Degrees of Freedom," *The American Statistician*, 45, 45–50.

Arora, Vipin and Lahiri, Partha (1997), "On the Superiority of the Bayesian Method Over the BLUP in Small Area Estimation Problems," *Statistica Sinica*, 7, 1053–1063.

Cochran, W. G. (1977) *Sampling Techniques, 3rd ed.*, New York: John Wiley and Sons.

Gershunskaya, Julie B. and Lahiri, Partha (2005), "Variance Estimation for Domains in the U.S. Current Employment Statistics Program," *Proceedings of the American Statistical Association, Survey Research Methods Section*, 3044–3051.

Fay, Robert E. and Train, George F. (1995), "Aspects of Survey and Model-Based Postcensal Estimation of Income and Poverty Characteristics for States and Counties," *Proceedings of the American Statistical Association, Government Statistics Section*, 154–159.

Maples, Jerry J., Bell, William R., and Huang, Elizabeth T. (2009), "Small Area Variance Modeling with Application to County Poverty Estimates from the American Community Survey," *Proceedings of the American Statistical Association, Survey Research Methods Section*, [CD-ROM], Alexandria, VA: American Statistical Association.

Navarro, Alfredo (2001), "2000 American Community Survey (ACS) Comparison County Replicate Factors (ACS-V-01)," internal U.S. Census Bureau memorandum to C. Alexander, Washington, DC, May 23, 2001.

Otto, Mark C. and Bell, William R. (1995), "Sampling Error Modelling of Poverty and Income Statistics for States," *Proceedings of the American Statistical Association, Government Statistics Section*, pp. 160–165.

Plackett, R. L. and Burman, J. Peter (1946), "The Design of Optimal Multifactorial Experiments," *Biometrika*, 33, 305–325.

Rao, C. R. (1973) *Linear Statistical Inference and its Applications, 2nd ed.*, New York: Wiley.

U.S. Census Bureau (2009), "Design and Methodology: American Community Survey," U.S. Government Printing Office, Washington, DC, available at http://www.census.gov/acs/www/Downloads/dm1.pdf.

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

You, Yong and Chapman, Beatrice (2006), "Small Area Estimation Using Area Level Models and Estimated Sampling Variances," *Survey Methodology*, 32, 97–103.

Figure 1. The K−S, Degrees of Freedom, CV and Bias of Fay's variance estimates of Mean−N(0,1)

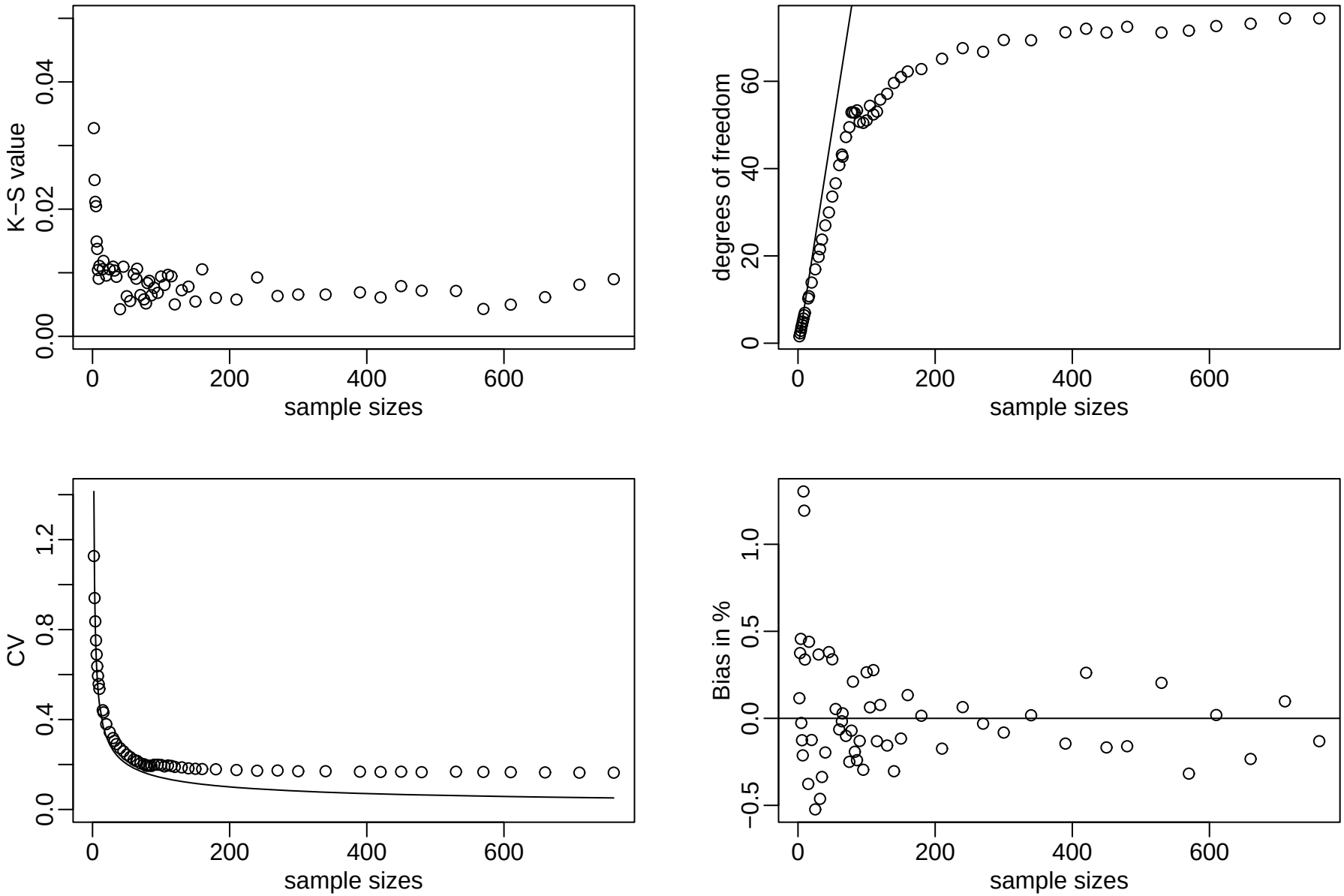

Figure 2. The K−S, Degrees of Freedom, CV and Bias of Fay's variance estimates of Mean−B(p)−p=0.1,0.25,0.5

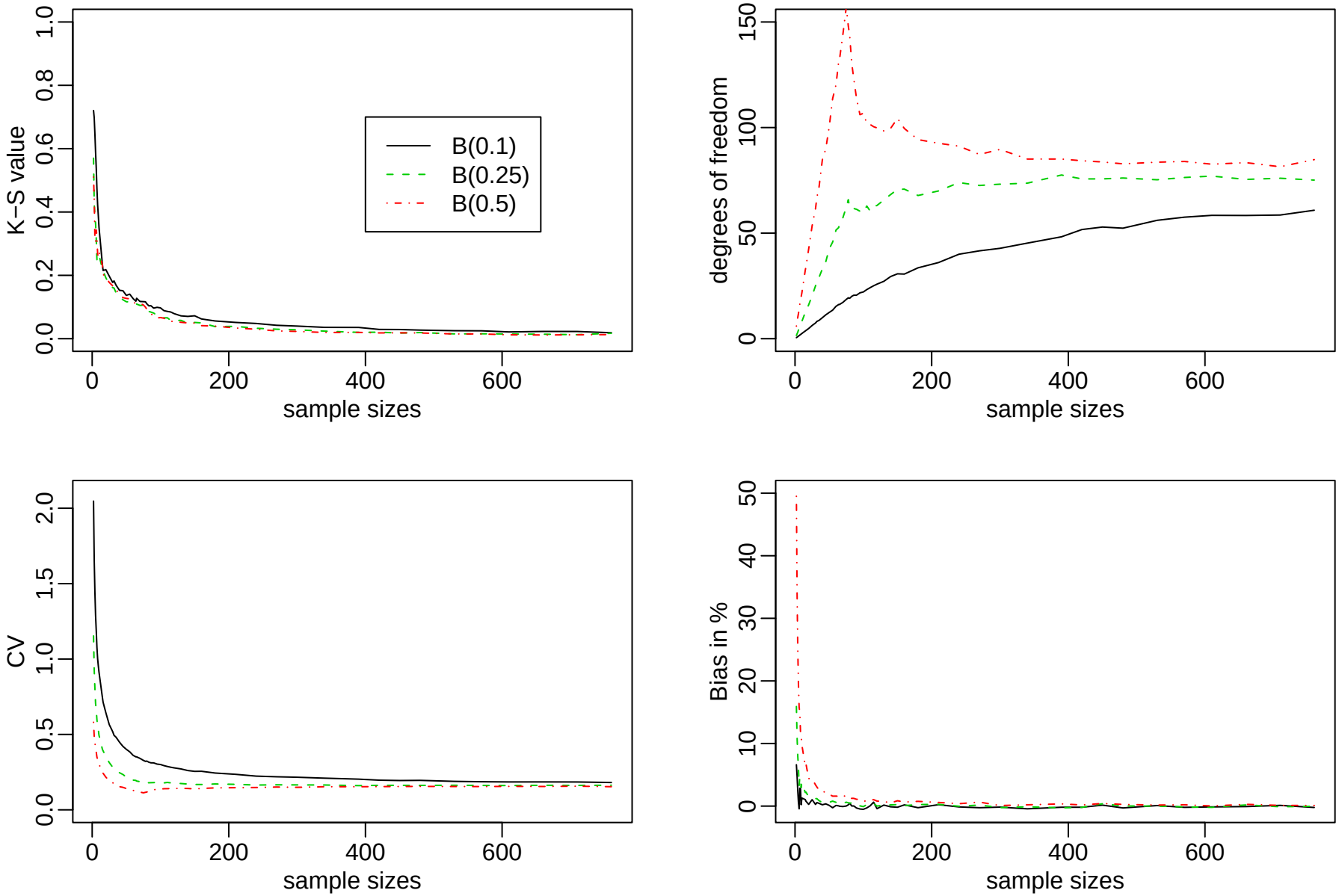

Figure 3. The K−S, Degrees of Freedom, CV and Bias of Fay's variance estimates of Mean−P(mu)−mu=0.02,0.1,0.3,0.5

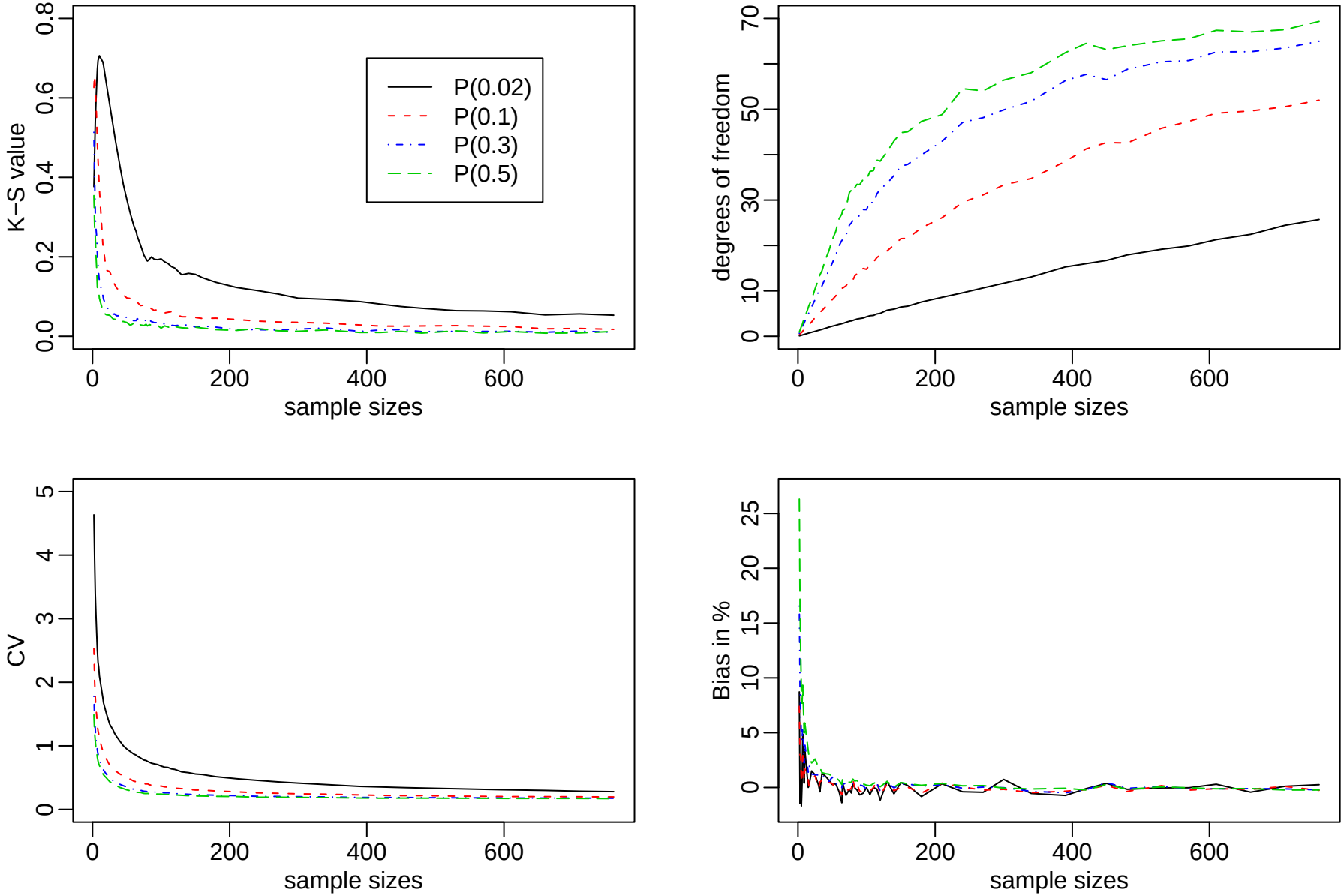

Figure 4. The K−S, Degrees of Freedom, CV and Bias of two variance estimates of mean−ACS05−MD−G1