BIAS CORRECTION FOR INEQUALITY MEASURES: An application to China and Kenya

Robert Breunig
Australian National University*

Abstract

This article applies an analytical bias correction technique for inequality measures to income data from China and Kenya. We use the coefficient of variation squared and illustrate how the bias is downward for positively skewed distributions. The analytical bias correction technique is then compared to a jackknife estimator in a simulation exercise. The bias will be important, even for moderately large sample sizes.

Keywords: *Coefficient of Variation, Small Sample Bias Correction, Inequality Measurement*

JEL Classification: C13, D31

*Centre for Economic Policy Research, Economics Program, Australian National University, Canberra ACT 0200 AUSTRALIA (02) 6125-2148, Fax: (02) 6125-0087, e-mail: Robert.Breunig@anu.edu.au.

1 INTRODUCTION

Most commonly used inequality measures such as Theil's Entropy indexes, Atkinson's measure, and the coefficient of variation are ratios of random variables and are thus biased in small samples. The expected value of such inequality measures will take the form

$$E\widehat{I} = I_0 + O\left(\frac{1}{n}\right)$$

where I_0 is the true population value of the inequality measure and the bias term of order n^{-1} will become very small in large samples. See Kakwani (1980) or Ravallion (1994) for a review of income inequality estimation. Breunig (2001) discusses the bias problem and demonstrates how the large-n expansion may be used to derive a bias-corrected estimator.

The purpose of this paper is to apply the bias correction technique developed in that paper for the Coefficient of Variation squared (CV^2) to income data from China and Kenya. In this paper, we demonstrate that the bias may be important, even in samples that would normally be considered to be quite large. This question is then explored further in a simple simulation exercise. Furthermore, we illustrate how the degree of bias in the estimated inequality measure is related to the skewness in the distribution of the sample data.

CV^2 has been used as an inequality measure, as has the coefficient of variation (CV) itself. CV^2 gives identical rankings to CV and therefore embodies the same underlying notion of social welfare (See Blackorby and Donaldson (1978) for mapping of inequality measures to implied social welfare functions.) Lehrer and Nerlove (1981), Blackburn and Bloom (1990), and Cancian, Danziger, and Gottschalk (1992) have all used CV^2 as an inequality measure. Formby, Smith, and Zheng (1999) have shown the usefulness of CV^2 for second-order stochastic dominance rankings of normalized distributions.

2 BIAS REDUCTION: APPLICATION

Given a sample of size n: $y_1, \ldots, y_n$; the square of the sample coefficient of variation is expressed as

$$\hat{\theta} = \frac{s^2}{\bar{y}^2} \tag{1}$$

where $\bar{y}$ and s^2 are the sample mean and variance. The bias correction technique proposed by Breunig (2001) is to instead estimate

$$\tilde{\theta} = \hat{\theta} - \widehat{Bias}(\hat{\theta}) \tag{2}$$

where

$$\widehat{Bias}(\hat{\theta}) = \frac{\hat{\theta}^{3/2}}{n}\left[3\hat{\theta}^{1/2} - 2\hat{\gamma}_1\right] \tag{3}$$

and the sample skewness coefficient is calculated as

$$\hat{\gamma}_1 = \frac{1}{n}\sum^{n}\left(y_i - \bar{y}\right)^3 / s^3. \tag{4}$$

The estimator $\tilde{\theta}$ is unbiased upto $O(n^{-1})$. (See Breunig (2001) for details.)

From (3) it is clear that $\tilde{\theta} = \hat{\theta}$ when $3\hat{\theta}^{1/2} = 2\hat{\gamma}_1$. When $\gamma_1 > \frac{3}{2}CV$ in the population, the bias will be negative. Income distributions tend to have large positive skewness, thus inequality estimation using CV^2 will be biased downward under most circumstances. The greater the inequality the greater the degree to which the measure will understate inequality.

We estimate the bias corrected CV^2 for two data sets on income: one from Kenya (1986, Central Bureau of Statistics and the Ministry of National Planning and Development of Kenya) on 2,424 urban households and 1988 Chinese data on 9,009 urban households gathered by a group of six University of California Riverside faculty members along with the Chinese Academy of Social Sciences. For additional information on these data sets see Mwangi wa Githinji (2000) and Khan, et. al. (1991).

Table 1 gives summary statistics for household and per-capita income as well as the estimates of $\hat{\theta}$ and $\tilde{\theta}$. The Gini coefficient is also calculated for reference. Per-capita income is computed by equally dividing household income among all members of the household and thus almost certainly adds a downward bias to inequality. Household

income uses the household as the main unit of analysis, weighting total household income by household size.

[Table 1 here]

It is clear from this table that urban inequality is much larger in Kenya than in China. Comparing the almost-unbiased coefficient of variation squared $\tilde{\theta}$ with the sample CV^2 as it is usually calculated, $\hat{\theta}$, the two data sets give quite different results. For the China data, the bias correction gives almost no change in the squared coefficient of variation. In the Kenya data, however, the usual estimator of CV^2, $\hat{\theta}$, gives an underestimate of inequality. The bias correction give a change of 4.3%, quite substantial for a sample of this size. This is because for the Kenya data the condition $\widehat{\gamma}_1 > \frac{3}{2}\widehat{CV}$ is satisfied. An examination of the nonparametric density functions of the two distributions[1] shows that the Kenya data has a long right-hand tail, signalling both large skewness and greater inequality. The lack of difference between $\hat{\theta}$ the $\tilde{\theta}$ for China is driven by two factors: the China data set is almost four times larger than the Kenyan one so that the Bias $(\hat{\theta})$ in (3) is almost zero– that is $\hat{\theta}$ and hence $\tilde{\theta}$ are both asymptotically unbiased. The second reason is that the China data has a less–skewed distribution– that is the value of $\hat{\gamma}_1$ is small and close to $\frac{3}{2}$CV making the bias of $\hat{\theta}$ in (3) near zero.

3 SIMULATION

Now we turn to the question of how important the bias in the coefficient of variation is for small samples. We attempt to address this question through two simple simulation exercises. First, using the unweighted Kenya data as the "population", we drew simple random samples (with replacement) of size $n = 50$ to $n = 500$, calculated both the sample coefficient of variation squared and the "almost unbiased" estimator of CV^2 for this new sample. In addition, we calculated an alternative bias-corrected estimator, $\widehat{\theta}_{jkn}$, using the leave-one out jackknife estimator calculated in the usual way (see Efron (1982, p. 6)). The jackknife will give similar results to a simple

[1]The graphs of the nonparametric densities as well as the data intself are available from the author.

bootstrap bias correction, however, the jackknife may be more suitable for inequality measures when the sampling is not simple random sampling (See Breunig and Stern (2001).) This exercise was repeated 50000 times. The results are summarized in Figure 1 and Table 2. The first part Figure 1 graphs the bias as a percentage of the true value of the parameter for the three estimators. The second part of Figure 1 graphs the mean squared error (MSE) of the two bias-corrected measures, $\widetilde{\theta}$ and $\widehat{\theta}_{jkn}$, relative to the mean squared error of $\widehat{\theta}$.

As can be seen from the figures, the jackknife estimator has the lowest average bias and the highest mean squared error throughout. The high variability of the estimator is a well-known problem of the jackknife (see Hinkley (1978), for example). Considering the bias, the proposed bias corrected estimator $\widetilde{\theta}$ dominates $\widehat{\theta}$ as it is usually calculated for all sample sizes which were considered. However, unlike the jackknife, the improvement in the bias comes at only a small loss of mean squared error. In cases where bias is the main concern of the practitioner, the bias-corrected estimator is thus preferable. Although the jackknife gives lower bias, the extremely high mean squared error would indicate that one should be suspicious of the results for any given case. $\widetilde{\theta}$ appears to provide a bias reduction that does not come at the expense of too large an increase in variance.

Sample sizes of 300 to 500 are not uncommon in the inequality literature, particularly when making comparisons across different states or regions of a country. These results have important implications for such comparisons. If the distributions across regions are dissimilar, then the degree of bias in the region-specific inequality measures will be quite different. Thus any comparison across regions will be affected. These effects may be quite severe. In the tables we present simulation averages. There were many cases where the bias was much more severe than the average and as a consequence, the bias-correction much larger.

In the second simulation exercise, we repeat the same calculations using the log-normal distribution. The simulated data is generated setting the variance parameter equal to one and the mean equal to $\exp\{0\}$. This gives a small amount of skewness and meets the condition for negative bias: $\gamma_1 > \frac{3}{2}CV$. Simulation results are summarized in Table 3 and Figure 2. We observe similar results to the first simulation.

For very small sample sizes (below 250) the jackknife bias correction does slightly better than $\widetilde{\theta}$ but at the expense of a large increase in mean squared error. For larger sample sizes, the bias-corrected estimator suggested here is equivalent to the jackknife in performance relative to bias but has a lower mean squared error and would thus appear to be superior. Both clearly dominate the usual method of estimating CV^2. The bias correction in this case is much smaller, but it is important to note that most income distributions are more highly skewed than a log-normal distribution and thus the degree of bias will generally be larger than this simulation shows.

4 CONCLUSION

The above findings indicate that when the sample is small or moderately large and the skewness in the distribution is greater than $\frac{3}{2}CV$ that the almost unbiased estimator $\tilde{\theta}$ will be useful for correcting bias in CV^2. In the example from Kenya, the bias correction is over 4% even with a sample of nearly 2500 observations. Usually we consider such sample sizes as immune from small-sample bias problems. The simulations show that the bias-corrected estimator, on average, performs betten than the usual method of calculating CV^2. The jackknife bias-corrected estimator does somewhat better at reducing bias, but much worse in a mean squared error sense. The choice of which technique to use, therefore, is not unambiguous. The increase in variance needs to be weighed against the improvement in bias.

For comparisons across region and over time, the biases illustrated here may be important. In particular, if the degree of skewness is different in the different distributions being compared, then apparent differences in the inequality measure may in fact simply be driven by differing degrees of bias in the measure. An important question which is beyond the scope of this paper is whether the relationship shown here between distribution and bias holds for other inequality measures.

References

[1] Blackburn, M. and D. Bloom (1990) "Changes in the Structure of Family Income Inequality in the U.S. and other Industrialized Nations During the 1980s," Mimeo.

[2] Blackorby, C. and D. Donaldson (1978) "Measures of relative equality and their meaning in terms of social welfare," *Journal of Economic Theory*, Volume 18, pp. 59-80.

[3] Breunig, R. (2001) "An almost unbiased estimator of the coefficient of variation," *Economics Letters*,.Volume 70, pp. 15–19.

[4] Breunig, R. and S. Stern (2001) "Bias correction and variance estimation for inequality measures under complex sampling." mimeo, Australian National University.

[5] Cancian, M., S. Danziger, and P. Gottschalk (1992) "Working Wives and Family Income Inequality among Married Couples," in: S. Danziger and P. Gottschalk, eds., *Uneven Tides: Rising Inequality in America.* Russell Sage Foundation: New York.

[6] Efron, B. (1982) *The Jackknife, the Bootstrap, and Other Resampling Plans*, Philadelphia: Society for Industrial and Applied Mathematics.

[7] Formby, J., W. Smith, and B. Zheng (1999) "The Coefficient of Variation, Stochastic Dominance and Inequality: A new interpretation" *Economics Letters* 62, 319-323.

[8] Githinji, M. (2000), *Ten Millionaires and Ten Million Beggars: A Study of Income Distribution and Development in Kenya.* Aldershott, England: Ashgate Publishing.

[9] Hinkley, D. V. (1978), "Improving the Jackknife with Special Reference to Correlation Estimation," *Biometrika*, Volume 65, number 1, pp. 13-21.

[10] Kakwani, N. (1980), *Income Inequality and Poverty: Methods of Estimation and Policy Applications*. New York: Oxford University Press.

[11] Khan, A. R., K. Griffin, C. Riskin, and Z. Renwei (1991), "Household Income and Its Distribution in China." University of California Riverside, Working paper Series.

[12] Lehrer, E. and M. Nerlove (1981) "The Impact of Female Work on Family Income Distribution in the United States: Black-White Differentials," *Review of Income and Wealth* 27, 423-431.

[13] Maasoumi, E. (1991), guest editor. *Measurement of Welfare and Inequality*. Annals of Econometrics, *Journal of Econometrics*. Volume 50, numbers 1 and 2.

[14] Ravallion, M. (1994), *Poverty Comparisons*. Langhorne, PA: Harwood Academic Publishers.

TABLE 1
Results on Inequality Measures

Unit of income	Kenyan Shillings	Chinese Yuan
Exchange rate in $US (at survey date)	16.04 KS=1.00 US$	4.86 CY=1.00 US$
Survey date	1986	1988
Sample Size	2,424	9,009
HOUSHOLD INCOME	**KENYA**	**CHINA**
Sample statistics		
Mean	46,628.95	6,507.26
Median	16,813	5759.00
Average Household Size	3.55	3.53
Variance	25,022,439,640.12	10,873,357
Standard Deviation	158,184.83	3297.48
Skewness coefficient	20.74	2.83
CV	3.39	0.50674
CV squared	11.51	0.25678
Corrected CV	3.47	0.5068
Corrected CV squared	12.01	0.25684
Gini coefficient	0.645	0.238
PER-CAPITA INCOME	**KENYA**	**CHINA**
Sample statistics		
Mean	12,204.6	1,841.95
Median	7,451.61	1,700.00
Variance	2,403,269,188.45	842,322.83
Standard Deviation	49,023.15	917.78
Skewness coefficient	27.45	3.04
CV	4.02	0.498
CV squared	16.13	0.248
Corrected CV	4.16	0.498
Corrected CV squared	17.28	0.248
Gini coefficient	0.652	0.222

Figure 1a
Bias for Three Estimators of CV^2

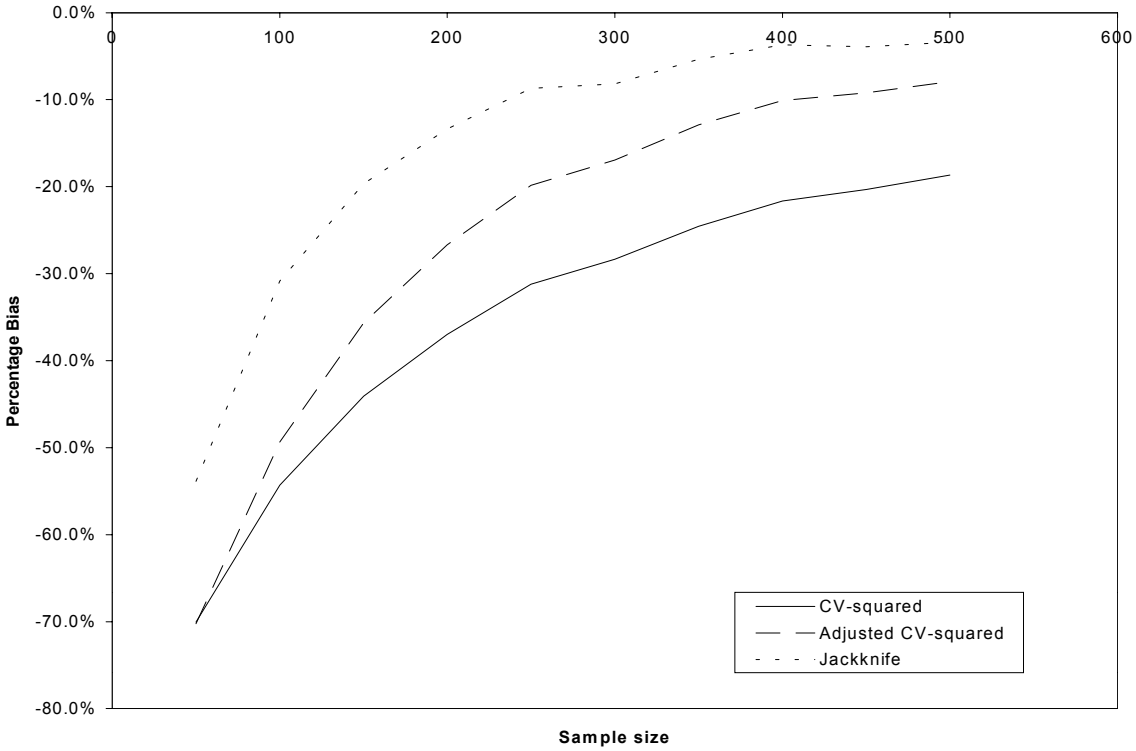

Figure 1b
Relative MSE for Bias-corrected Estimators of CV^2

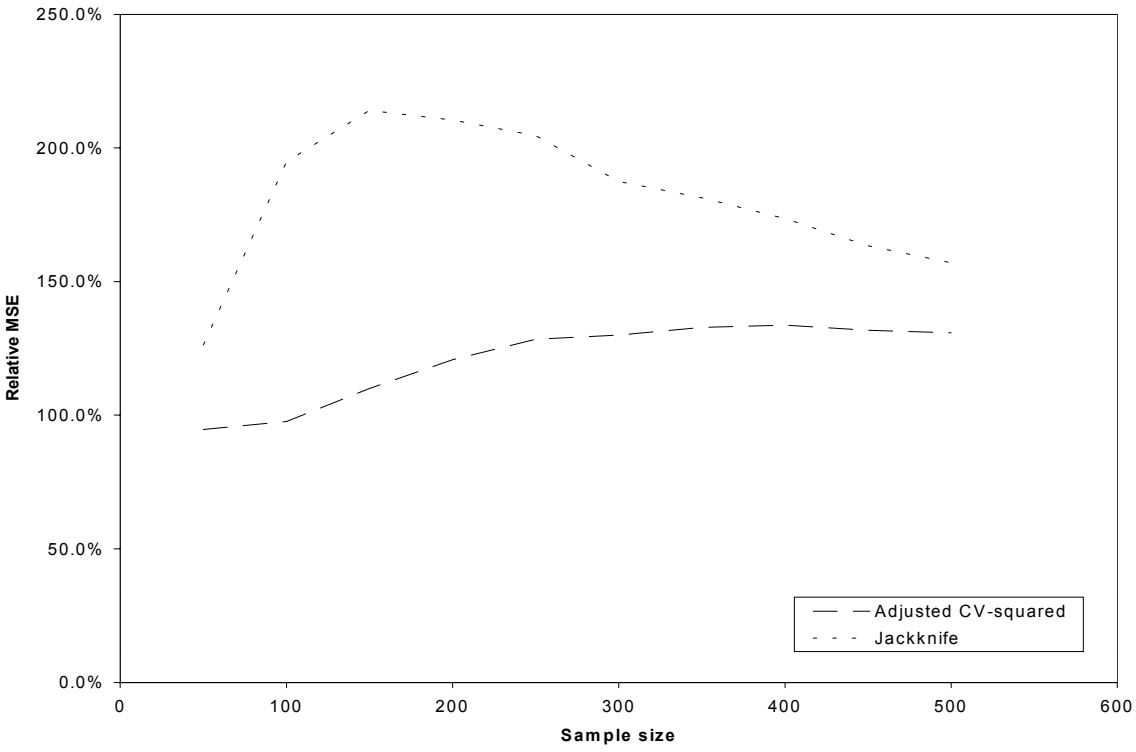

Figure 2a
Bias for Three Estimators of CV^2
(Simulated lognormal data)

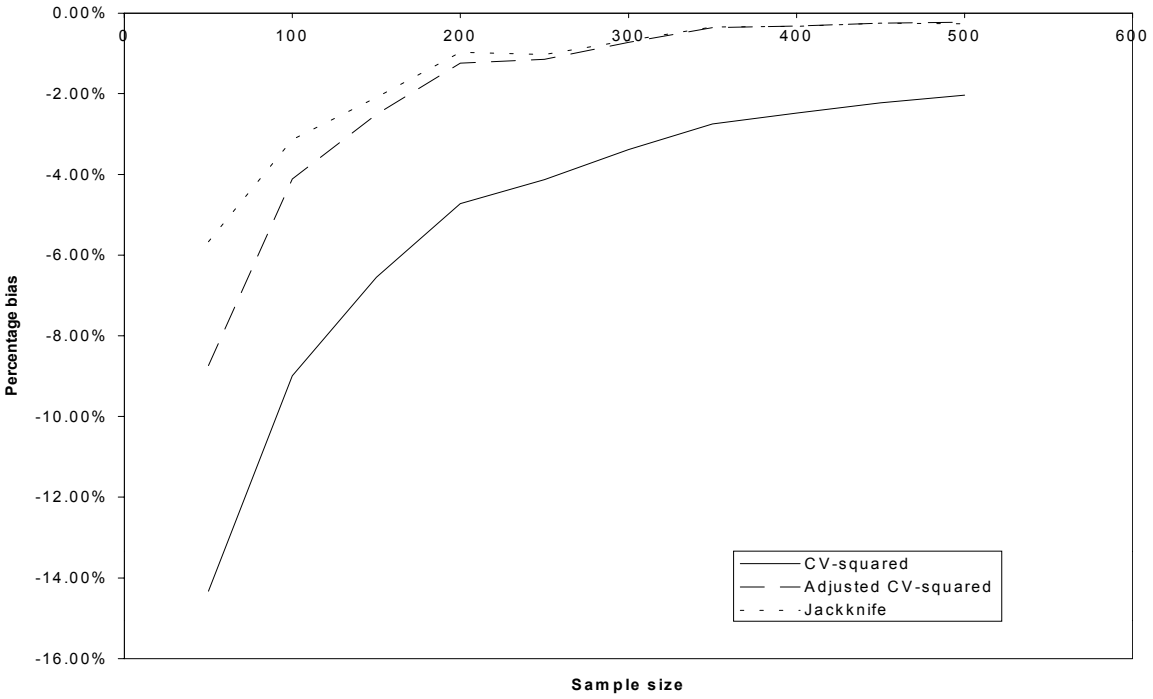

Figure 2b
Relative MSE for Bias-corrected Estimators of CV^2
(Simulated lognormal data)

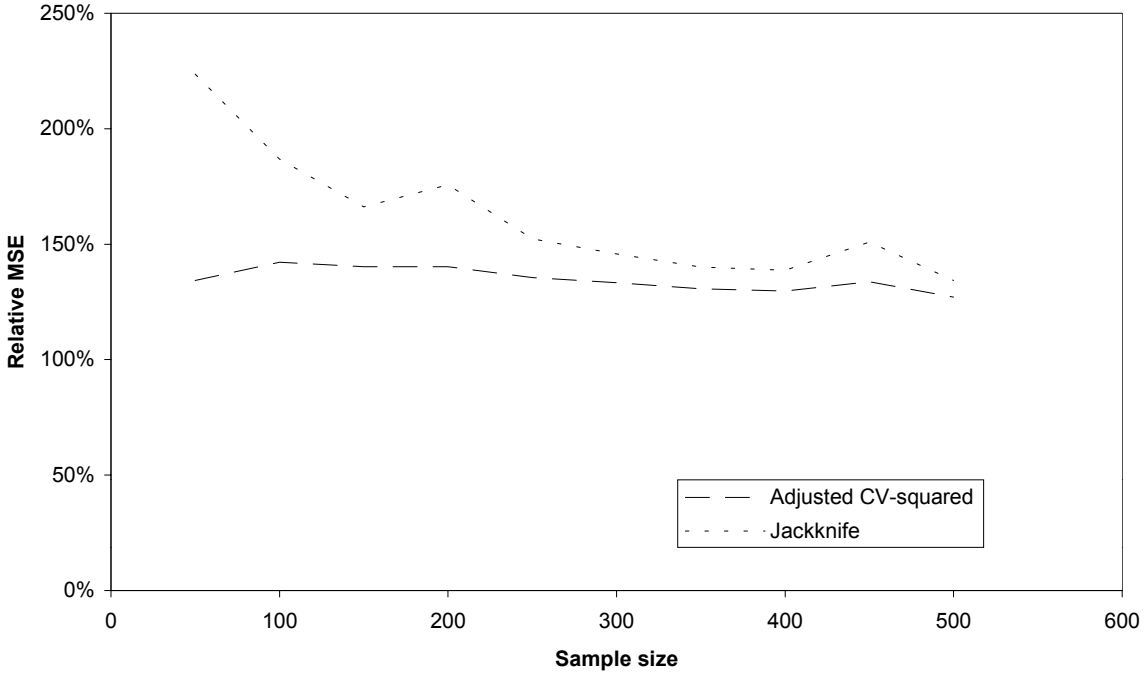

TABLE 2
Average Bias and Mean Squared Error
for Three Estimators of the Coefficient of Variation

	Average Bias			Average Mean Squared Error		
n	$\hat{\theta}$	$\tilde{\theta}$	θ_{jkn}	$\hat{\theta}$	$\tilde{\theta}$	θ_{jkn}
50	-12.95	-12.98	-9.96	209.2	198.2	264.3
100	-10.04	-9.12	-5.70	185.3	180.9	361.1
150	-8.15	-6.60	-3.63	172.8	190.2	369.8
200	-6.84	-4.93	-2.47	163.3	197.2	343.6
250	-5.77	-3.67	-1.60	153.2	196.7	313.3
300	-5.24	-3.13	-1.51	141.0	183.3	264.3
350	-4.54	-2.38	-0.99	131.9	175.2	239.1
400	-4.00	-1.87	-0.68	122.7	164.0	213.1
450	-3.76	-1.70	-0.72	113.6	149.7	185.6
500	-3.45	-1.46	-0.63	106.2	138.9	166.8

Simulation based on 50000 repetitions (population CV-squared is 18.49)

$\hat{\theta}$: CV-squared

$\tilde{\theta}$: Adjusted CV-squared

θ_{jkn}: Jackknife adjusted CV-squared

TABLE 3
Average Bias and Mean Squared Error
for Three Estimators of the Coefficient of Variation
(simulated lognormal data)

	Average Bias			Average Mean Squared Error		
n	$\hat{\theta}$	$\tilde{\theta}$	θ_{jkn}	$\hat{\theta}$	$\tilde{\theta}$	θ_{jkn}
50	-0.246	-0.150	-0.097	0.985	1.323	2.205
100	-0.154	-0.071	-0.054	0.772	1.098	1.442
150	-0.113	-0.043	-0.036	0.662	0.929	1.101
200	-0.081	-0.021	-0.017	0.646	0.905	1.135
250	-0.071	-0.020	-0.018	0.535	0.726	0.815
300	-0.058	-0.013	-0.012	0.493	0.658	0.719
350	-0.047	-0.006	-0.006	0.446	0.582	0.624
400	-0.043	-0.006	-0.006	0.413	0.536	0.573
450	-0.038	-0.004	-0.004	0.439	0.587	0.663
500	-0.035	-0.004	-0.004	0.361	0.459	0.485

Simulation based on 50000 repetitions (population CV-squared is 1.718)

$\hat{\theta}$: CV-squared

$\tilde{\theta}$: Adjusted CV-squared

θ_{jkn}: Jackknife adjusted CV-squared