Not to be cited without permission of the authors

Canadian Atlantic Fisheries Scientific Advisory Committee

CAFSAC Research Document 81/ 74

A comparison of four estimators of the variance of the mean from a two dimensional systematic sample

by

S.J. Smith
Marine Fish Division
Bedford Institute of Oceanography
PO Box 1006
Dartmouth, N.S. B2Y 4A2

and

K.S. Naidu
Research and Resource Services
Northwest Atlantic Fisheries
Centre
PO Box 5667
St. John's, NFLD A1C 5X1

Abstract

The problem of estimating the variance of the mean from a single systematic sample of a spatially autocorrelated population is investigated by simulation methods. Using a Gaussian Markov stationary autocorrelated model the behaviour of four contenders for the estimator of this variance was studied under various conditions. It was found that a sample based form of the unconditional expected variance of systematic sampling was the most efficient of the four. Further, it was found that using a jackknife autocorrelation estimator instead of the natural autocorrelation estimator did not improve the efficiency of this variance estimator.

Résumé

Des méthodes de simulation ont servi à analyser le problème de l'estimation de la variance de la moyenne à partir d'un échantillon systématique unique d'une population à autocorrélation spatiale. Nous faisons appel à un modèle gaussien stationnaire de Markov pour étudier, dans diverses conditions, le comportement de quatre estimateurs possibles de variance. On a constaté que le plus efficace des quatre estimateurs était une forme, fondée sur l'échantillon, de la variance anticipée inconditionnelle de l'échantillonnage systématique. On a trouvé en outre que l'emploi d'un estimateur d'autocorrélation dit "jackknife" au lieu d'un estimateur naturel n'améliorait pas l'efficacité de cet estimateur.

1. INTRODUCTION

Sample surveys of commercially important fish and invertebrate species are often carried out to gather information on temporal changes in population size. This information is used in conjuction with commercially based indices, e.g. catch per unit effort, to provide conservation and management advice. Due to the nature of the use of the results of such surveys it is important that the survey design used be as optimum a design as possible with respect to providing precise estimates of the mean numbers or weights of animals present. In general these types of surveys are very difficult to design because the only available "a priori" information is a rough idea of the range or boundaries of the population being studied. In the case of sedentary benthic invertebrates, it may be assumed that spatial autocorrelation is present, that is samples taken close together in space are more likely to be similar than samples taken further apart. Often where surveys cover a large area, stratified random designs are employed usually with some arbitrary ranges of depths used as the stratifying variable. The presence of spatial autocorrelation in these types of surveys is usually ignored; temporal changes can affect the spatial patterns in mobile species, and the large distances between samples can reduce the effect of autocorrelation on the variance estimates. For small scale surveys of sedentary animals where samples may be taken very close together the degree of autocorrelation can have a greater effect.

In recent issues of Biometrika there have been a number of papers dealing with the problem of finding an optimum sampling design for autocorrelated populations in one (Blight, 1973) and two dimensions (Bellhouse 1977, Martin 1979). In these cases it was determined from theoretical comparisons that a type of systematic sampling scheme provided the optimum design with respect to minimum variance.

In addition to the above studies systematic sampling schemes have recently received some attention in the fisheries literature (Fieldler (1978), Venrick (1978), and Lenarz and Adams (1980)). These studies compare systematic, stratified and random schemes in one dimension by simulation and/or empirical methods. They conclude that systematic schemes are to be preferred on the basis of increased precision of the estimate of the population mean.

One drawback of this sampling method is the lack of a general estimator of the variance of the mean. Heilbron (1978) using simulation techniques studied this problem when sampling an autocorrelated population in one dimension. He concluded that a sample based approximation of the unconditional expected value, derived by Cochran (1946) and discussed further by Quenouille (1949), performed well enough with respect to a quadratic loss criterion to be useful in practice. We propose to extend some of the techniques used by Heilbron to the two dimensional case and again by simulation, study the behaviour of four possible estimators of the variance. The familiar superpopulation model (Cochran 1946, Quenouille 1949 and others) is used along with a Gaussian stationary serial correlation model. The

The problem of estimating the variance and the model are explained in §2 while the method of study and the estimators examined are discussed in §3.

2. THE MODEL

Envisage the area to be sampled divided up into a rectangular lattice with N_1 rows and N_2 columns so that there are $N = N_1 \times N_2$ sample units of equal size available. A rectangular shaped sample area may not be attainable in practice but the infinite number of permutations of non-rectangular shapes would be unmanageable in a study of this sort. The size of the sample units would be determined by the type of sampling gear used in the survey, i.e. in the benthic marine environment the gear could be dredges or bottom trawling nets which would be towed over a standard distance by a vessel.

Further, associated with each sample unit is an unknown value y_{ij} which is to be observed, where i and j denote the row and column indices respectively, $i = 1, \ldots, N_1$ and $j = 1, \ldots, N_2$. After Quenouille (1949) and Bellhouse (1977) we assume that these N observations (y_{ij}) are to be considered a finite population which is in turn a sample from an infinite superpopulation which has the following characteristics,

$$E[y_{ij}] = \mu,$$
$$E[(y_{iy} - \mu)^2] = \sigma^2 \qquad (1)$$
$$E[(y_{ij} - \mu)(y_{i+u,j+v} - \mu)] = \sigma^2 \rho_{u,v}.$$

The quantity $\rho_{u,v}$ represents the autocorrelation between sample units of distance u and v units apart.

Further, we wish to sample this finite population by employing a systematic scheme in the following manner. We first rewrite the row and columns indices of the rectangular lattice as $N_1 = n_1 K_1$ and $N_2 = n_2 K_2$ respectively, where K_1, K_2, n_1 and n_2 are all integers. Then from the following ranges 1, ... , K_1 and 1, ... , K_2 randomly select integers i' and j' with probabilities $1/K_1$ and $1/K_2$. The sample will consist of those units identified by the n_1 x n_2 combinations of the row indices i', i' + K_1, i' + $2K_1$, ... , i' + $(n_1 - 1)K_1$ and column indices j', j' + K_2, j' + $2K_2$, ... , j' + $(n_2 - 1)K_2$. The sample will be structured as a rectangular lattice in space, and when taken in this manner is called and aligned sample. Alternatively the i' and j' could have been chosen such that only the rows, or columns, are aligned or neither are aligned. Bellhouse (1977) has shown that the aligned sample is less efficient than the case where the alignment is in one direction only or where there is no alignment but no conclusions could be drawn as to the relative efficiency between the latter two types. We prefer to restrict our attention to aligned samples because we have found them to be easier and more efficient to implement especially where vessel time and associated costs are concerned.

From the sample we are interested in estimating the finite population mean $\bar{Y}$ by the sample mean,

$$\bar{y}_{\ell,s} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} y_{ij} \quad , \tag{2}$$

where ℓ and s denote the particular values of i' and j' chosen to obtain the sample. The sampling variance of $\bar{y}_{\ell,s}$ is defined by:

$$V_{sy} = \frac{1}{K_1 K_2} \sum_{\ell=1}^{K_1} \sum_{s=1}^{K_2} (\bar{y}_{\ell,s} - \bar{Y})^2 \quad . \tag{3}$$

The quantities required for equation 3 can only be obtained if the finite population was completely enumerated. Since we have only one value of $\bar{y}_{\ell,s}$ for any one systematic sample we cannot estimate the variance by directly using the definition in equation 3. As noted in the introduction therein lies the problem with this sampling method. In §3 we will discuss some possible ways of estimating this variance.

In our simulation study we assumed that the population values followed a Gaussian stationary autocorrelation model with $\rho_{u,v} = \rho_1^u \rho_2^v$.

3. ESTIMATORS AND METHOD OF STUDY

To compare the variances obtained from various types of sampling designs with respect to sampling an autocorrelated population in two dimensions, Quenouille (1949) derived the unconditional expected value

for each variance. The expected value for the variance of the mean of a systematic sample (EVS) therein derived is as follows:

$$EVS = \frac{\sigma^2}{n_1 n_2}\left(\frac{K_1K_2-1}{K_1K_2}\right)\left\{1 - \frac{1}{K_1K_2n_1n_2(K_1K_2-1))} \times \sum_{u,v\epsilon s}\sum (K_1n_1-|u|)\right.$$

$$\left. \times (K_2n_2-|v|)\,\rho_{u,v} + \frac{K_1K_2}{(n_1n_2(K_1K_2-1))} \times \sum_{u,v\epsilon s'}\sum (n_1-|u|)\right. \qquad (4)$$

$$\left. \times (n_2-|v|)\,\rho_{K_1u,K_2v}\right\} .$$

In this equation σ^2, K_1, K_2, n_1 and n_2 are as defined previously while $\rho_{u,v}$ refers to the autocorrelation between sample units of distance u,v units apart in the finite population and ρ_{K_1u,K_2v} is the autocorrelation between units contained in a single systematic sample. The first double summation is carried out over a region s where $|u| < K_1n_1$, $|v| < K_2n_2$ and excludes $u = v = o$. The second summation exists over the region s' where $|u| < n_1$, $|v| < n_2$ and again $u = v = o$ is excluded.

This equation was used as the basis for three of the estimates studied here by replacing the parameters σ^2, $\rho_{u,v}$ and ρ_{K_1u,K_2v} with sample based estimates analogous to the approach taken by Heilbron (1978).

The estimators compared in the simulation study are defined as follows:

(1) The EVSY: The equation in (4) with σ^2 replaced by the sample variance and $\rho_{K_1u,K_2v} = \rho_1^{K_1u}\ \rho_2^{K_2v}$ estimated by

$$\hat{\rho}_{K_1u,K_2v} = \frac{\sum_{i=1}^{n_1-u}\sum_{j=1}^{n_2-v}(Y_{ij} - \bar{Y}')(Y_{i+u,j+v} - \bar{Y}'')}{\left\{\sum_{i=1}^{n_1-u}\sum_{j=1}^{n_2-v}(Y_{ij} - \bar{Y}')^2 \times \sum_{i=1}^{n_1-u}\sum_{j=1}^{n_2-v}(Y_{i+u,i+v} - \bar{Y}'')^2\right\}^{\frac{1}{2}}} \quad (5)$$

where $\bar{Y}' = \sum_{i=1}^{n_1-u}\sum_{j=1}^{n_2-v} Y_{ij}/(n_1-u)(n_2-v)$,

$$\bar{Y}'' = \sum_{i=1}^{n_1-u}\sum_{j=1}^{n_2-v} Y_{i+u,j+v}/(n_1-u)(n_2-v)$$

The values (u,v) are set to (1,0) when estimating ρ^{K_1u} and (0,1) for $\rho_2^{K_2v}$.

(2) EVSJ: Again we use equation (4) and replace σ^2 with the sample variance but the autocorrelation estimator is a jackknifed version of (5) $\rho^{J}_{K_1u,K_2v}$, defined as:

$$\hat{\rho}^{J}_{K_1u,K_2v} = 2\hat{\rho}_{K_1u,K_2v} - \tfrac{1}{2}(\hat{\rho}'_{K_1u,K_2v} + \hat{\rho}''_{K_1u,K_2v}) \quad , \qquad (6)$$

where $\hat{\rho}_{K_1u,K_2v}$ and $\hat{\rho}_{K_1u,K_2v}$ are computed by using (5) on the first and second halves of the data set, respectively, divided in half along the jth direction for $\hat{\rho}_1^{K_1}$ and along the ith direction for $\hat{\rho}_2^{K_2}$.

(3) EVSO = $\frac{\hat{\sigma}^2}{n_1 n_2}\left(\frac{K_1K_2-1}{K_1K_2}\right)$, that is using the expected value form but ignoring the autocorrelation.

(4) SUB4: This is strictly an ad hoc estimate which was designed to imitate equation (3) but at the sample level. The sample is subdivided into four pseudo-systematic samples and the mean is calculated for each sample. The original sample mean is used as an estimate of the population mean and the pseudo sample means are used in place of the $\bar{\bar{y}}_{\ell,s}$ values. From preliminary tests it was observed that this estimator was biased by a factor of approximately 4 and therefore the simulations were carried with equation (3) divided by 4.0.

The simulations were carried out by generating N pseudo-random numbers into a $N_1 \times N_2$ array replicated $500/K_1K_2$ times for each set of parameters studied. These numbers were generated using the IMSL (International and Mathematics and Statistics Libraries) subroutine GGNML (IMSL, 1980). It was found that there existed a residual serial correlation between the numbers generated but this was eliminated by "shuffling" rows and columns of the generated array.

The desired autocorrelation structure was obtained by first applying the following transformation along the rows of the data matrix,

$$Y'_{ij} = (1 - \rho_1^2)^{\frac{1}{2}} Y_{ij} + \rho_1 Y_{i,j-1} \quad (j=2, \ldots, N_{2;\forall i})$$

Then the columns were treated in an analogous manner by applying,

$$Y''_{ij} = (1 - \rho_2^2)^{\frac{1}{2}} Y'_{ij} + \rho_2 Y'_{i-1,j} \quad (i=2, \ldots, N_{i;\forall j})$$

The following 4 sets of runs were carried out:

A: $N_1 = N_2 = 40$, $K_1 = K_2 = 2$

B: $N_1 = N_2 = 40$, $K_1 = K_2 = 5$

C: $N_1 = N_2 = 80$, $K_1 = K_2 = 2$

D: $N_1 = N_2 = 80$, $K_1 = K_2 = 5$

In each case the simulations were run for the following three sets of autocorrelation; (1) $\rho_1^{K_1 u} = \rho_2^{K_2 v} = 0.2$, (2) $\rho_1^{K_1 u} = \rho_1^{K_2 v} = 0.4$ and (3) $\rho_1^{K_1 u} = \rho_2^{K_2 v} = 0.8$. The autocorrelation values were set at the sample level since it is at this level that the researcher will detect them. In the results any particular simulation run is identified by a letter and a number, e.g.

A1: $N_1 = N_2 = 40$, $K_1 = K_2 = 2$ and $\rho_1^{K_1 u} = \rho_2^{K_2 v} = 0.2$

Additional runs were carried out for the A series for $\rho_1^{K_1 u} = 0.8$, $\rho_2^{K_2 v} = 0.2$ and $\rho_1^{K_1 u} = 0.8$, $\rho_2^{K_2 v} = 0.4$. The results of these experiments were in accordance with those from the cases listed above and therefore will not be reported here.

Since the autocorrelations were set to be greater than zero, autocorrelation estimates encountered by the simulation program which were less than zero were set equal to zero and a record was kept of the number of occurrences.

The estimators are compared by the criterion of relative effeciency (R.E.) defined here as the ratio of the mean squared error of EVS to the mean squared error of the variance estimator (T, say), that is:

$$R.E.\ (T) = \sum_{i=1}^{t} (EVS_i - VSY)^2 / \sum_{i=1}^{t} (T_i - VSY)^2, \qquad (7)$$

where t is the total number of simulations carried out for any particular case.

Use of the above criterion differs from the approach taken by Heilbron (1978) where the conditional expectation of VSY was used in place of EVS in (7). This conditional expectation was derived under the assumption that the observations resulted from an arbritary Gaussian distribution and then specialized to the Markov serial correlation model. We preferred to use the unconditional expected value because being distribution-free it would facilitate comparisons between our results and those obtained if alternate distributions or models were studied.

4. RESULTS AND DISCUSSION

The results from the relative efficiency comparison are presented in Figure 1 and 2. These results can be summarized as follows; R.E. (EVSY) $\geq$ R.E. (EVSJ) $>$ R.E. (EVSO) $>$ R.E. (SUB4) . The estimated relative bias for each estimator, i.e. R.B.($\bar{T}$) = [($\bar{T}$-VSY)/VSY]x100, is plotted in Figures 3 and 4. Here the results are not as obvious as the relative efficiency results but the following trends can be seen. The |R.B. (EVSO)| was always greater in value than those obtained for the other estimators while |R.B.(SUB4)| behaved erratically. In accordance with the relative efficiency results |R.B. (EVSY)| was always less than or equal to |R.B. (EVSJ)|.

Intuitively one would expect, for the model assumed in this study, that those estimators which take the autocorrelation into account would be more efficient and in general less biased than those that did not. However, we did not expect to see the estimator EVSY perform as well as or better than EVSJ. Recall that the autocorrelation estimator in EVSJ was the jackknife estimator which is expected to be less biased than the natural estimator (5) used in EVSY. Comparing the estimated average values obtained from the simulation experiment for the two types of autocorrelation estimators we see that this is indeed true (Table 1). However from Table 2 where the variances of the estimators and the percentage of times that the estimates were less than zero are compared it is obvious that the jackknife estimator is also a less precise estimator. Therefore the advantage to be gained by using the less biased jackknife estimator is lost due the associated larger variance.

The main point to be made then is that Heilbron's (1978) findings for the one dimensional case can now be extended to two dimensions. In addition our results show that for the two dimensional case, use of the jackknife autorcorrelation estimator instead of the less complicated natural autocorrelation estimator (5) does not improve the relative efficiency of the sample-base unconditional expected variance estimator.

We would like to thank J. McGlade, R.K. Misra and D. Rivard for their very helpful comments on an earlier draft.

REFERENCES

Bellhouse, D.R. (1977). Some optimal designs for sampling in two dimensions. Biometrika 64, 605-611.

Blight, B.J.N. (1973). Sampling from an autocorrelated finite population. Biometrika 60, 375-385.

Cochran, W.G. (1946). Relative accuracy of systematic and stratified random from a certain class of populations. Ann. Math. Statist. 17, 164-77.

Fieldler, P.C. (1978). The precision of simulated transect surveys of northern anchovy, Engraulis mordax, school groups. Fishery Bulletin 76, 679-685.

Heilbron, D.C. (1978). Comparison of estimators of the variance of systematic sampling. Biometrika 65, 429-33.

IMSL (1980). International Mathematical and Statistical Libraries, Inc. Houston, Tex. USA.

Lenarz, W.H. and Adams, P.B. (1980). Some statistical considerations of the design of trawl surveys for rockfish (Scorpaenidae). Fishery Bulletin 78, 659-74.

Martin, R.J. (1979). A subclass of lattice processes applied to a problem in planar sampling. Biometrika 66, 209-17.

Quenouille, M.H. (1949). Problems in plane sampling. Ann. Math. Statist. 20, 355-75.

Venrick, E.L. (1978). Systematic sampling in a planktonic ecosystem. Fishery Bulletin 76, 617-27.

Table 1. Results of the Simulation Study: Comparing Average values obtained from the autocorrelation estimators. Relative Bias (R.B.)=(estimator-parameter)/parameter, expressed as a percentage .

Simulation Set No.	Expected Value	Avg $(\hat{\rho}_1^{K_1 u})$ (R.B.)	Avg $(\hat{\rho}_1^{J})$ (R.B.)	Avg $(\hat{\rho}_2^{K_2 v})$ (R.B.)	Avg $(\hat{\rho}_2^{J})$ (R.B.)
A 1	0.2	0.1969 (-1.55)	0.2023 (1.15)	0.1896 (-5.20)	0.1945 (-2.75)
A 2	0.4	0.3913 (-2.19)	0.4012 (0.30)	0.3830 (-4.25)	0.3916 (-2.10)
A 3	0.8	0.7587 (-5.16)	0.7867 (-1.66)	0.7521 (-5.99)	0.7772 (-2.85)
B 1	0.2	0.1580 (-21.00)	0.1720 (-14.00)	0.1380 (-13.00)	0.1540 (-23.00)
B 2	0.4	0.3276 (-19.00)	0.3696 (-7.60)	0.3087 (-22.83)	0.3465 (-13.38)
B 3	0.8	0.6185 (-22.69)	0.7048 (-11.90)	0.5887 (-26.41)	0.6439 (-19.51)
C 1	0.2	0.1984 (-0.70)	0.1999 (-0.05)	0.1927 (-1.40)	0.1985 (-0.75)
C 2	0.4	0.3969 (-0.78)	0.3997 (-0.08)	0.3946 (-1.35)	0.3969 (-0.78)
C 3	0.8	0.7855 (-1.77)	0.7961 (-0.45)	0.7839 (-2.01)	0.7926 (-0.93)
D 1	0.2	0.1950 (-2.50)	0.2040 (2.00)	0.1920 (-4.00)	0.1980 (-1.00)
D 2	0.4	0.3842 (-3.95)	0.4014 (0.35)	0.3827 (-4.33)	0.3955 (-1.13)
D 3	0.8	0.7296 (-8.80)	0.7696 (-3.80)	0.7262 (-9.23)	0.7631 (-4.61)

Table 2. Results of the Simulation Study: Comparison of the variances ($\times 10^{-2}$) of autocorrelation estimators and percentage of times that estimates were less than zero (% <0.)

Simulation Set No.	Var $(\hat{\rho}_1^{K_1 u})$	Var $(\hat{\rho}_2^{J})$	Var $(\hat{\rho}_2^{K_2 v})$	Var $(\hat{\rho}_2^{J})$
	% <0.	% <0.	% <0.	% <0.
A 1	0.2567	0.2803	0.2572	0.2693
	0.0	0.08	0.0	0.6
A 2	0.2792	0.3136	0.2757	0.2879
	0.0	0.0	0.0	0.0
A 3	0.3205	0.4759	0.3187	0.4159
	0.0	0.0	0.0	0.0
B 1	1.7398	1.9675	1.7330	2.6350
	11.6	25.6	15.0	33.8
B 2	1.8063	2.5267	1.8798	2.5604
	2.20	9.4	1.4	12.20
B 3	1.8474	3.4288	3.2550	4.1549
	0.0	1.8	0.2	3.0
C 1	0.0630	0.0643	0.0715	0.0719
	0.0	0.0	0.0	0.0
C 2	0.0738	0.0782	0.0776	0.0788
	0.0	0.0	0.0	0.0
C 3	0.1037	0.1341	0.1012	0.1162
	0.0	0.0	0.0	0.0
D 1	0.4980	0.5760	0.4327	0.4744
	0.8	3.4	3.4	3.4
D 2	0.5462	0.6572	0.4573	0.4807
	0.0	0.0	0.0	0.0
D 3	0.4418	0.6268	0.2974	0.4516
	0.0	0.0	0.0	0.0

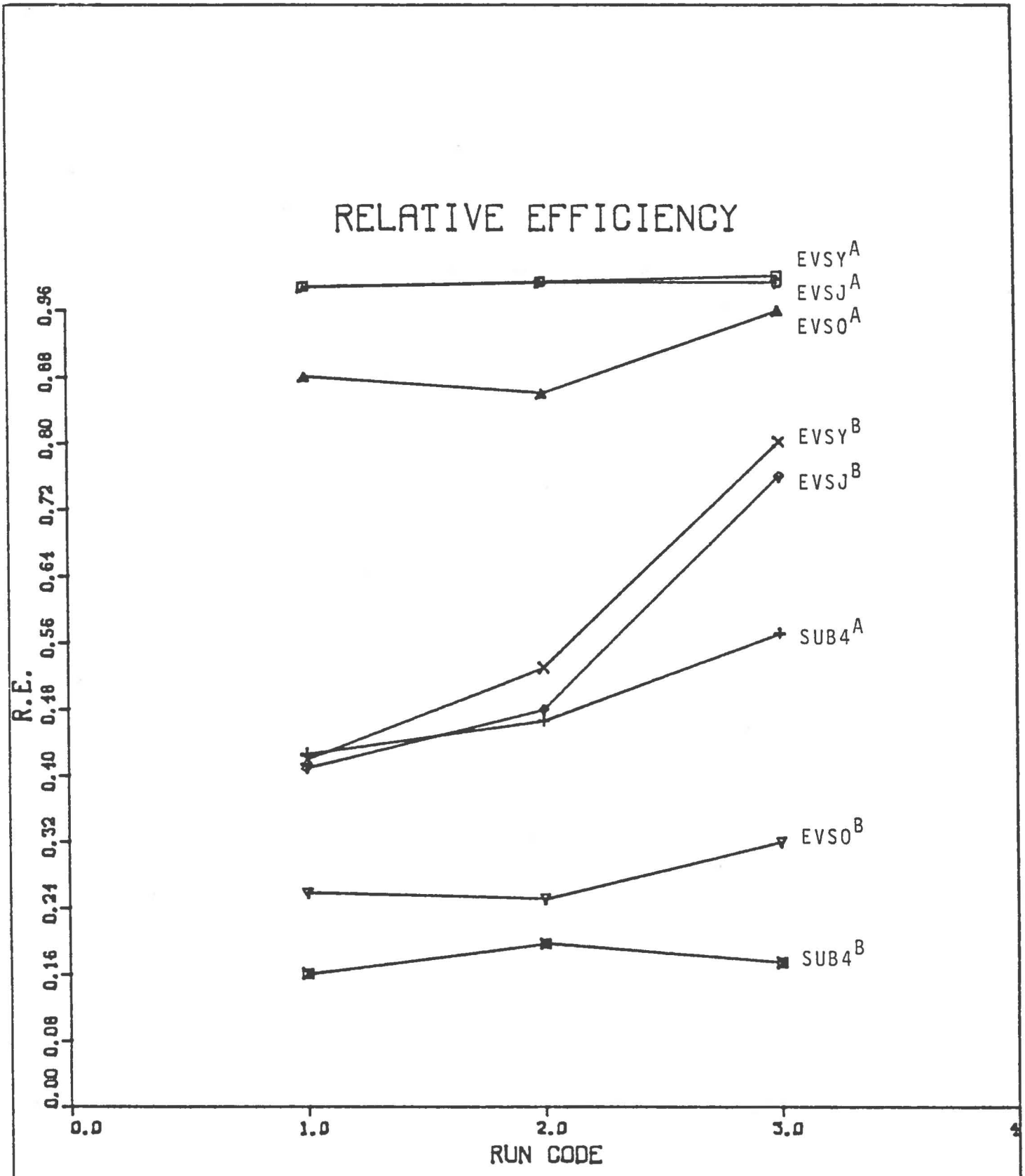

Figure 1. Relative efficiency results from the simulation study Sets A and B. See text for explanation of estimators.

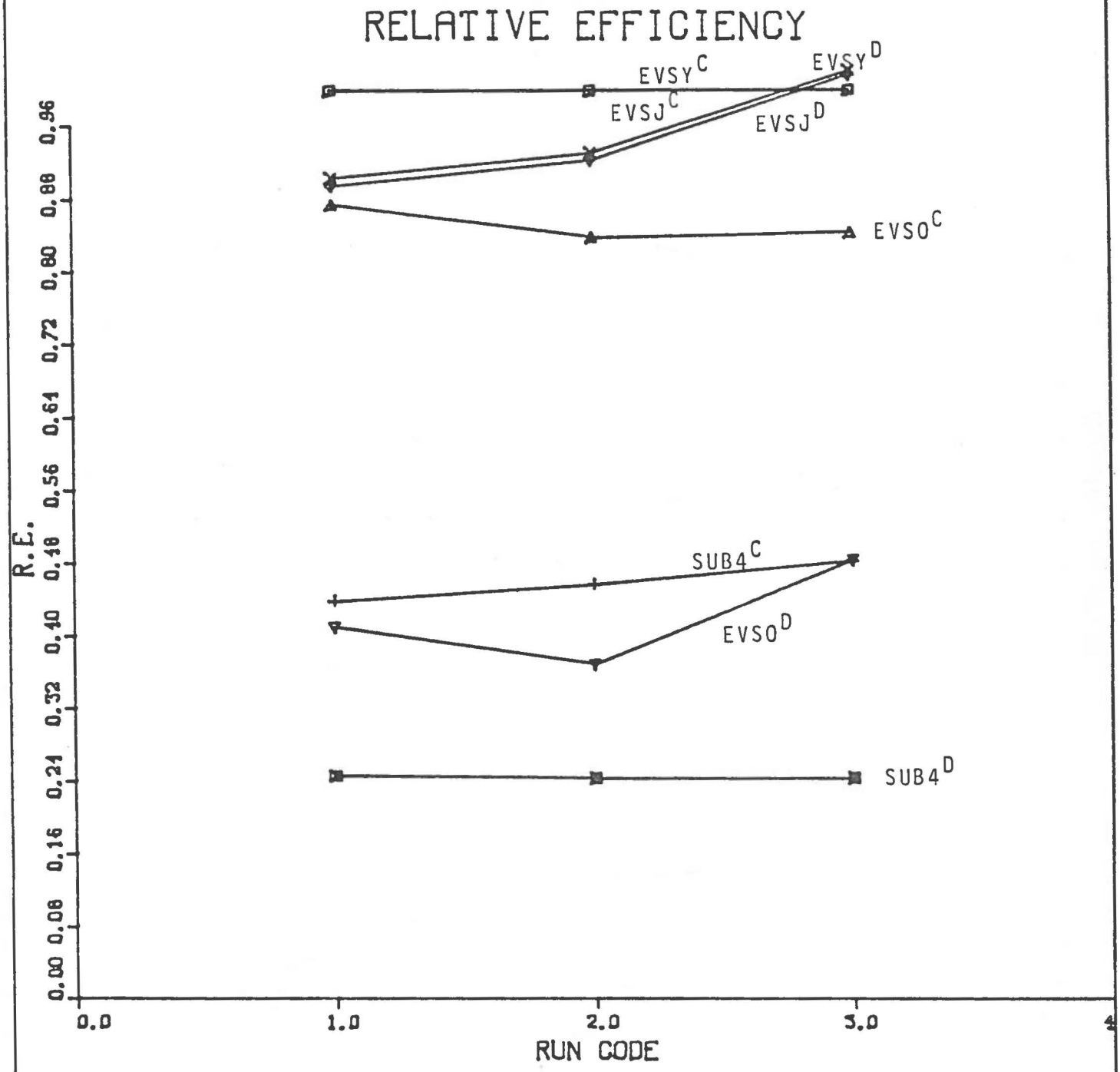

Figure 2. Relative efficiency results from the simulation study Sets C and D. See text for explanation of estimators.

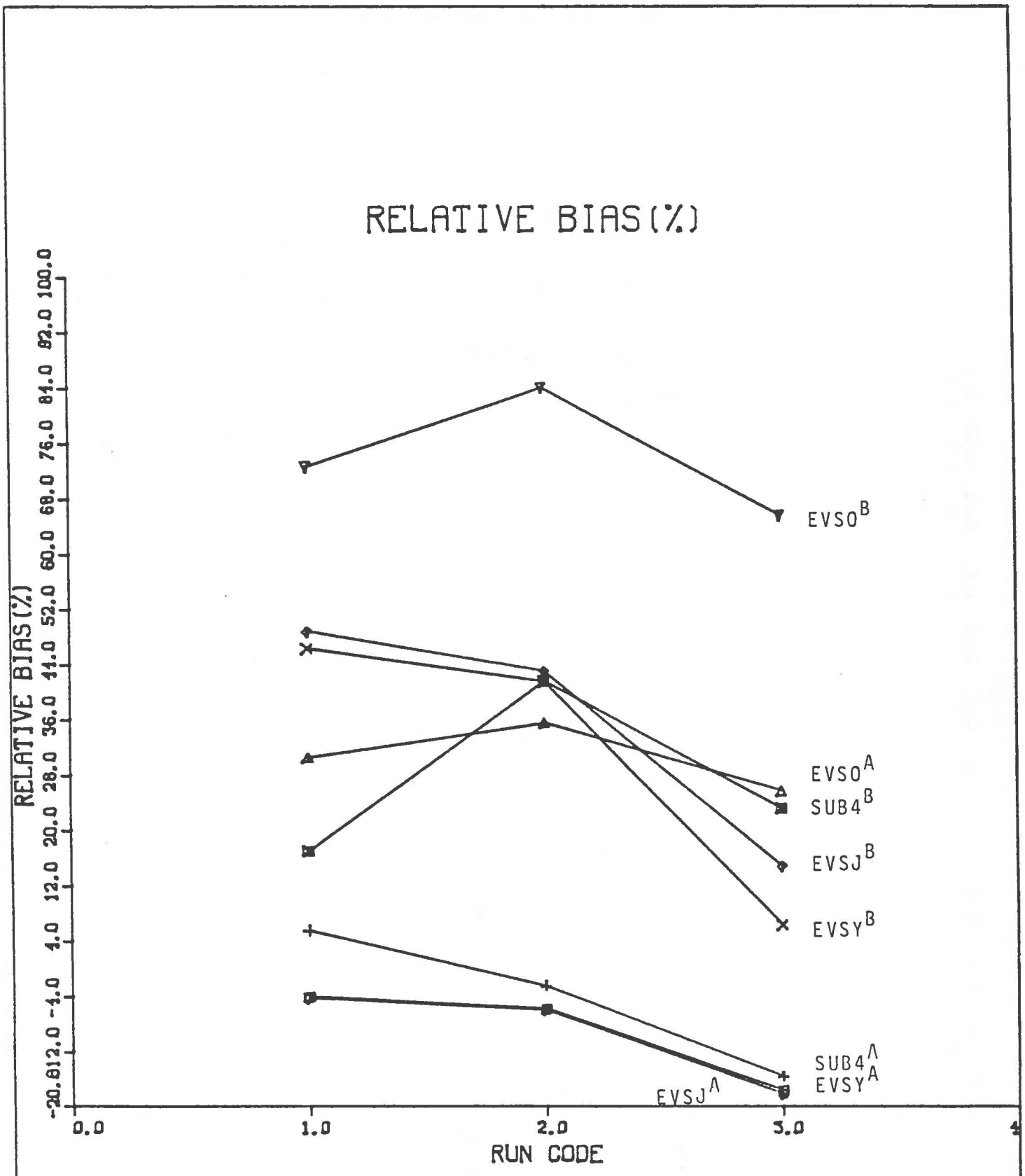

Figure 2. Percentage bias results from the simulation study Sets A and B. See text for explanation of estimators.

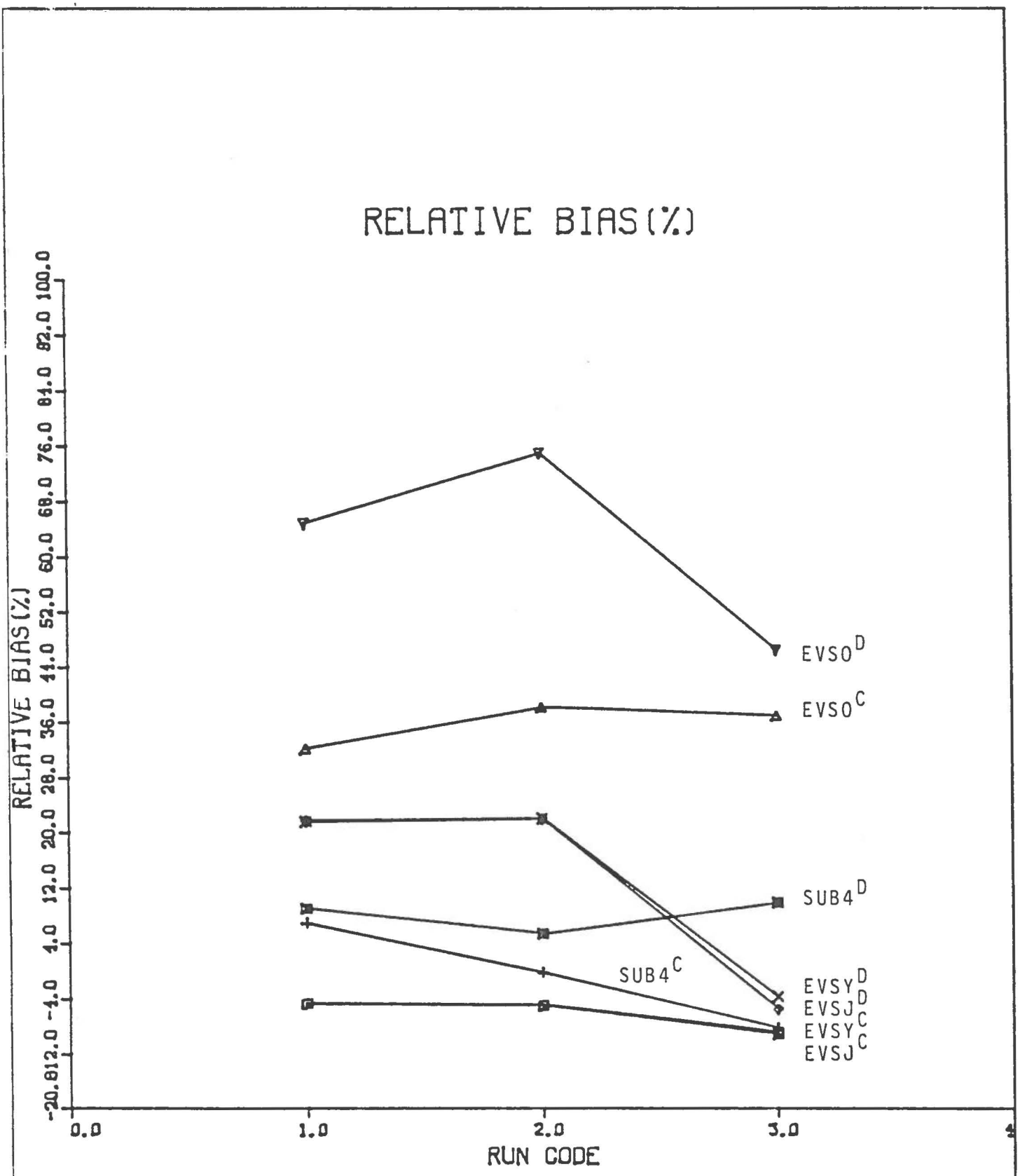

Figure 4. Percentage bias results from the simulation study Sets C and D. See text for explanation of estimators.