Published in final edited form as:

Int Stat Rev. 2015 August ; 83(2): 228–238. doi:10.1111/insr.12087.

Assessing Variability of Complex Descriptive Statistics in Monte Carlo Studies using Resampling Methods

Dennis D. Boos and Jason A. Osborne

Department of Statistics, North Carolina State University, Raleigh, NC 27695-8203

Dennis D. Boos: boos@stat.ncsu.edu

Summary

Good statistical practice dictates that summaries in Monte Carlo studies should always be accompanied by standard errors. Those standard errors are easy to provide for summaries that are sample means over the replications of the Monte Carlo output: for example, bias estimates, power estimates for tests, and mean squared error estimates. But often more complex summaries are of interest: medians (often displayed in boxplots), sample variances, ratios of sample variances, and non-normality measures like skewness and kurtosis. In principle standard errors for most of these latter summaries may be derived from the Delta Method, but that extra step is often a barrier for standard errors to be provided. Here we highlight the simplicity of using the jackknife and bootstrap to compute these standard errors, even when the summaries are somewhat complicated.

Key words and phrases

Bootstrap; jackknife; coefficient of variation; delta method; influence curve; standard errors; variability of ratios

1 Introduction

Due to the complexity of modern statistical methods, Monte Carlo studies are often a key tool for evaluating their effectiveness. However, perusal of recent statistical journals suggests that these “experiments” are not evaluated or summarized with the same rigor that we statisticians ask of our scientific friends in other fields. Specific examples are given at the end of this section. Why is this so? We think that at least partially the reason is that deriving and computing the required standard errors is an extra step that can be burdensome for complex summaries. The main point of this paper is to demonstrate the ease with which these standard errors can be provided by the jackknife or bootstrap. (Note: For us, the term “standard error” refers to any estimate of the standard deviation of a sampling distribution.)

Typically a Monte Carlo study is used to study estimators and associated variance estimators, statistical tests, and confidence intervals. Many of the summaries displayed in tables or graphs are simply means of the Monte Carlo output, and standard errors are easily obtained. For example, suppose $(\hat{\theta}_1, V_1), \dots, (\hat{\theta}_N, V_N)$ are results from computing an estimator and associated variance estimator from N simulated samples. In general, we call this raw Monte Carlo output the “Monte Carlo result matrix $\mathbf{X}$.” The Monte Carlo standard error of the bias estimator $\bar{\hat{\theta}} - \theta_0$ is $S_N / \sqrt{N}$, where $\bar{\hat{\theta}}$ and S_N are the sample mean and

standard deviation of the $\hat{\theta}_i$, respectively, and θ_0 is the true parameter value. Similarly, the Monte Carlo estimate of the mean squared error is the average of the $(\hat{\theta}_i - \theta_0)^2$, the estimate of $E(V)$ is the average of the V_i , and an estimate of the coverage of the confidence interval $\hat{\theta} \pm 1.96 \sqrt{\hat{V}}$ is the average of the 0–1 variable obtained by checking whether each of the N intervals contain θ_0 .

But even in this simple situation with $\mathbf{X} = (\hat{\theta}_1, V_1), \dots, (\hat{\theta}_N, V_N)$, more complicated summaries may arise. Suppose that there is interest in assessing the approximate normality of the estimator, say by the classical Pearson measures of skewness and kurtosis: $\text{Skew} = E\{X - E(X)\}^3 / [E\{X - E(X)\}^2]^{3/2}$ and $\text{Kurt} = E\{X - E(X)\}^4 / [E\{X - E(X)\}^2]^2$, taking values 0 and 3, respectively, for normally distributed X . Here the Pearson skewness estimator is

$\hat{\text{Skew}} = N^{-1} \sum_i (\hat{\theta}_i - \bar{\theta})^3 / S_N^3$, but how to obtain a standard error? The usual Delta Method (e.g., Ver Hoef, 2012) calculus for the asymptotic variance is tedious, but may be found in Cramér (1946, p. 357),

$$\text{AVar}(\hat{\text{Skew}}) = \frac{4\mu_2^2\mu_6 - 12\mu_2\mu_3\mu_5 - 24\mu_2^3\mu_4 + 9\mu_3^2\mu_4 + 35\mu_2^2\mu_3^2 + 36\mu_2^5}{4\mu_2^5 N},$$

where μ_k is the k th central moment. Then estimates of the sample moments,

$N^{-1} \sum_i (\hat{\theta}_i - \bar{\theta})^k$, could be inserted to get a standard error. The effort to derive this Monte Carlo standard error correctly (without using the Cramér reference) is beyond what most people are willing to expend. But the jackknife or bootstrap standard error is immediate with a simple R function call (as illustrated in Sections 3 and 4). In fact, it is no more difficult to obtain a jackknife or bootstrap standard error for $\hat{\text{Skew}}$ than it is for the bias estimate $\bar{\theta} - \theta_0$.

A second example is related to the variance estimator V associated with $\hat{\theta}$. Often the mean $\bar{V}$ of the V_i is reported along with S_N^2 ; if they are close, one may feel confident that V is approximately unbiased for $\text{Var}(\hat{\theta})$ (estimated by S_N^2). Even with associated Monte Carlo standard errors given, however, the correlation between $\bar{V}$ and S_N^2 makes it difficult to assess whether these estimates are significantly different. We argue in Section 2 that the ratio $\bar{V}/S_N^2$ is a better summary (easy to compare visually with 1), and coupled with a standard error makes inference easy. However, computing the Delta Method standard error for $\bar{V}/S_N^2$ is not so easy, albeit easier than for the skewness estimator.

Our underlying premise is that any table or plot of Monte Carlo estimates should include a summary of standard errors for each different type of estimate displayed. However, except for the sample mean, computation of the required standard errors can be burdensome and distract from the main focus of research. Our goal then is to show that jackknife and bootstrap standard errors are so simple and effective for use in Monte Carlo studies that they are worth considering as part of almost any analysis of simulations. An additional benefit of

having standard errors readily available is that choosing the Monte Carlo replication size N can be facilitated by calculating the standard errors in preliminary runs.

How widely applicable are the jackknife and bootstrap standard errors? To answer this, we make a distinction between 1) The Monte Carlo output X and 2) The summaries of this output.

1. The Monte Carlo output X is made up of N rows, typically quantities like estimators, estimated variances of estimators, test statistics, etc.; each row is computed from an independently generated data set. However, these quantities may be anything calculated from the generated data and parameters: for example, they might be nonregular estimators like the sample extremes or results from model selection. Their sampling distribution has nothing to do with the applicability of the jackknife and bootstrap.
2. For standard errors, the jackknife and bootstrap are applied to summaries of X , that is, functions of X , say $g(X)$, like sample means and variances and higher sample moments and regular functions of these. But robust summaries like medians or trimmed means are also possible. The summaries may be quite complex, but they should be regular in the sense that they are asymptotically normal as the Monte Carlo sample size N goes to infinity. In choosing between the jackknife and bootstrap, there are two issues:
  - a. Scope of applicability. For regular summaries like differentiable functions of the sample moments of X , either the jackknife or bootstrap may be used. (See, for example, Shao and Tu, 1995, Theorems 2.1 and 3.8.) However, the jackknife variance estimator is not consistent for percentiles, but the bootstrap is. So whenever a summary involves percentiles, the bootstrap should be used. For example, the median is an integral part of the boxplot, and thus to give a standard error for the boxplot center based on Monte Carlo output, the bootstrap should be used. Or when estimating the upper percentile of a test statistic, the bootstrap should be used.
  - b. Computing time. The jackknife computations, explained in Sections 3 and 4, are basically to drop out rows of the Monte Carlo output resulting in N “resamples,” and then to compute the summary of interest for each. For example, we have found jackknife calculations in R for output from $N = 1,000$ generated samples are essentially instantaneous, whereas those for $N = 10,000$ samples can take up to a minute or two when handling a number of summaries at the same time. The bootstrap computations are based on drawing B simple random samples with replacement from the rows of the Monte Carlo output, and calculating the summary of interest for each resample. If one uses $B = 1,000$, then the computations for the bootstrap are comparable to the jackknife when $N = 1,000$, but the bootstrap will be faster than the jackknife when $N = 10,000$.

For most applications, both the jackknife and bootstrap are applicable, and there is essentially no human cost once the programming is understood and little additional

computer cost. Typically with N at least 100, the jackknife and bootstrap give exactly the same standard errors to several decimals, as we shall find in the Tables of Section 2.

To illustrate further the need for standard errors and the results of this paper, we looked through recent issues of this journal. Because *ISR* is a review and expository journal, it has few Monte Carlo studies. However, we found two from the April 2014 issue that help us make our point.

In Figure 1 of Niebuhr and Kreiss (2014), we find side-by-side boxplots of a bootstrap method of estimating 95th percentiles of estimated autocovariances and one based on asymptotic normal approximations. In addition they plot the true 95th percentile as a dotted line, making for an easy visual comparison. However, they say “Figure 1 shows that the bootstrap procedure outperforms the normal approximation in this situation.” Clearly the center of the boxplot of bootstrap percentiles is much closer to the true percentile than that of the normal approximation, but as statisticians we would challenge a biologist making such a statement without at least giving standard errors. Connecting to the above discussion of applicability of the jackknife and bootstrap, the fact that the boxplots are based on estimated percentiles is irrelevant, it could have been boxplots of any random quantity computed from the simulated autoregressive processes. However, the key quantity underlying their superiority claim is the center of the boxplot, i.e., the median, and this means that only the bootstrap and not the jackknife could be used here for standard errors.

A second example is from Zhu (2014). In his Table 1 we find a bias estimator and two adjoining columns labeled $se_e(\cdot)$ for “empirical standard error” (that we called S_N in our example above) and $se_m(\cdot)$ for “average model-based standard error” (the average of $\sqrt{\hat{V}_i}$ in our example). Since here $N = 1000$, the $se_e(\cdot)$ could have been divided by $\sqrt{1000}$ to give a Monte Carlo standard error for the bias estimates, but there are no standard errors for the “standard error” columns. The author states “Moreover, for all the estimation methods, the empirical and model-based standard errors are very close, ...” We admit that they are indeed very close, but it still would be good practice to give some Monte Carlo standard errors. In fact, as we argue in Section 2, an alternative display could have replaced those two standard error columns by their ratio plus a Monte Carlo standard error from the jackknife or bootstrap to more easily assess their closeness.

A brief outline of the paper is as follows. Section 2 introduces notation and reports on a small Monte Carlo study to illustrate the main points of the paper. Section 3 reviews the basics of the jackknife and bootstrap, and Section 4 gives the detailed explanation of how to use the jackknife and bootstrap to obtain Monte Carlo standard errors. R code is sprinkled throughout in order to be clear about practical use.

2 Notation and Motivating Example

To keep notation and ideas clear, we use the term “Monte Carlo standard error” and denote it by SE_{MC} for any estimated standard deviation of a Monte Carlo summary computed from a sample of N vectors, each vector computed from a computer-generated random sample. Precise language and notation are important because Monte Carlo studies involve two kinds

of samples, the original N samples or datasets, and then the resulting sample of N vectors of estimates computed from those datasets that we call the Monte Carlo result matrix $\mathbf{X}$. SE_{MC} is based on $\mathbf{X}$.

To illustrate, we use a small Monte Carlo study of the sample coefficient of variation

$\hat{CV} = s/\bar{Y}$ for a sample $Y_1, \dots, Y_n$, where $\bar{Y} = n^{-1} \sum_{i=1}^n Y_i$ and $s^2 = (n-1)^{-1} \sum_{i=1}^n (Y_i - \bar{Y})^2$. For a second estimator, we use the bias-corrected estimator of Bao (2009), CV_{BC} , obtained by plugging $\hat{CV}$, $\hat{Skew}$, and $\hat{Kurt}$ into his large sample bias expression

$$E(\hat{CV}) - CV \approx (CV/n) \{ CV^2 - 1/4 - CV(Skew)/2 - (Kurt - 3)/8 \},$$

and subtracting the resulting bias estimator from $\hat{CV}$. For a variance estimator for either, we use estimates plugged into the asymptotic variance expression

$$AVar(\hat{CV}) = (CV^2/n) \{ CV^2 - CV(Skew) + (Kurt - 1)/4 \},$$

which may be found in Serfling (1980, p. 137).

We generated $N = 4,000$ samples of size $n = 20$ and $n = 50$ from normal, gamma (shape=2), and exponential populations with mean adjusted to have $CV=0.5$ for each, and then calculated $\hat{\theta}_1 = \hat{CV}$, $\hat{\theta}_2 = \hat{CV}_{BC}$, and $\hat{V}_1 = \hat{V}_2 = A\hat{V}_{Ar}$ for each. Thus, the resulting Monte Carlo result matrix $\mathbf{X}$ is $4,000 \times 3$. To illustrate, the first three rows of $\mathbf{X}$ from the normal population are

	est	est.bc	est.avar
1	0.6340277	0.6239584	0.026301120
2	0.5049589	0.5016624	0.006976995
3	0.4766496	0.4743313	0.004432886

The columns in Table 1 are simply the estimated biases of $\hat{CV}$ and CV_{BC} (columns 1 and 2), the sample variances (times n) of the 4,000 $\hat{CV}$ and CV_{BC} (columns 3 and 4), the respective estimated mean square errors (times n , columns 5 and 6), and the average of the estimated asymptotic variance (times n , labeled $\overline{nA\hat{V}_{Ar}}$, column 7). For each entry, a Monte Carlo standard error was calculated by the jackknife and bootstrap, as explained in Section 4, and the mean of each column of standard errors is included in the last row of the table. Although the jackknife and bootstrap standard errors are a little different, their respective means are identical for the number of digits displayed.

We might be interested in comparing columns 1 and 2, 3 and 4, 5 and 6, 3 and 7, and 4 and 7. For bias, it is fairly easy to see that the bias-corrected estimator has significantly less bias in most cases, but that even for the worst case for $n = 20$, $\hat{CV}$ has only $(0.029/0.5)100 =$

5.8% bias. Moving to the estimated variances in columns 3 and 4, $\hat{C}\hat{V}$ appears to have lower variance than CV_{BC} in a number of cases, but bounding the standard error of the difference by $(0.006^2 + 0.006^2)^{1/2} = 0.01$ (since the estimators are highly positively correlated) is too crude to declare statistical significance in all cases. Clearly, we could estimate the standard deviation of all the relevant paired differences and carry out appropriate inference, but it is not easy to display the elements needed for the analysis. Moreover, it is visually difficult to make all the pairwise comparisons. A similar comment could be made for comparing mean square errors and for comparing average estimated asymptotic variance (column 7) to Monte Carlo estimated variances (columns 6 and 7).

As an alternative display, consider Table 2 that keeps the first two columns of Table 1 but adds three of the relevant ratios of entries from Table 1: $\overline{MSE}$ Ratio is the ratio of the estimated mean squared error of $\hat{C}\hat{V}$ to that of CV_{BC} , and $\overline{A\hat{V}ar}/S_N^2$ in the last two columns are ratios of $\overline{A\hat{V}ar}$ to the Monte Carlo estimated variances S_N^2 for $\hat{C}\hat{V}$ and CV_{BC} , respectively. Now the comparisons are all immediately obvious, and inference is trivial to carry out by comparing the ratios to 1.0 along with the standard error averages at the bottom of each column. For example, CV_{BC} has slightly improved mean squared error compared to $\hat{C}\hat{V}$ for the normal distribution but much worse mean squared error for the gamma and exponential distributions. The last two columns show that the estimated asymptotic variance badly under-estimates the true variance for both estimators for the gamma and exponential population. Part of the reason for this under-estimation is the use of $\hat{S}_{skew}$ and $\hat{K}_{urt}$, which are biased low, in the estimated asymptotic variance expression. In other work, we have seen that the jackknife estimate of variance is much better for estimating the variance of $\hat{C}\hat{V}$ than merely estimating the asymptotic variance directly.

Thus for making comparisons, we prefer the ratios in Table 2, but note that Table 1 has an advantage if one is interested in the actual values of the variances and mean squared errors and not just their comparisons. Note also that without use of the jackknife or bootstrap, the standard errors of the ratios are much more challenging to derive than those in Table 1.

The Monte Carlo standard errors calculated for Tables 1 and 2 for the jackknife and bootstrap are essentially the same, and identical for the number of digits reported in the last rows of those tables. That makes sense because $N = 4,000$ is a very large sample in terms of convergence of these standard errors to the true standard deviations. With smaller Monte Carlo replication size, say $N = 100$, there will be some differences between them, but not much. Thus, we believe that the choice between using the jackknife or bootstrap for Monte Carlo standard errors should be based mainly on personal preference. Now we proceed to briefly review the jackknife and bootstrap and then illustrate their use in calculating the standard errors of Tables 1 and Table 2.

3 Jackknife and Bootstrap Basics

The jackknife was introduced by Quenouille (1949, 1956) as a general method to remove bias from estimators, and Tukey (1958) suggested its use for variance and interval estimation. The bootstrap was introduced by Efron (1979), and has been very popular for

estimating both bias and variance as well as other aspects of a sampling distribution. Here we briefly describe each method, starting with the jackknife.

For an estimator $\hat{\theta}$ based on an iid sample $Y_1, \dots, Y_n$, let $\hat{\theta}_{[i]}$ be the “leave-1-out” estimator obtained by computing $\hat{\theta}$ with Y_i deleted from the sample. We denote the average of these “leave-1-out” estimators by $\bar{\theta}_1 = n^{-1} \sum_{i=1}^n \hat{\theta}_{[i]}$ and define the pseudo-values by

$$\hat{\theta}_{ps,i} = n\hat{\theta} - (n-1)\hat{\theta}_{[i]}. \quad (1)$$

The average of these pseudo-values is the bias-adjusted jackknife estimator

$$\hat{\theta}_j = \frac{1}{n} \sum_{i=1}^n \hat{\theta}_{ps,i} = n\hat{\theta} - (n-1)\bar{\theta}_1 = \hat{\theta} - (n-1)(\bar{\theta}_1 - \hat{\theta}).$$

The jackknife variance estimator for $\hat{\theta}$ or $\hat{\theta}_j$ is

$$\hat{V}_j = \frac{n-1}{n} \sum_{i=1}^n (\hat{\theta}_{[i]} - \bar{\theta}_1)^2 = \frac{1}{n} \frac{1}{n-1} \sum_{i=1}^n (\hat{\theta}_{ps,i} - \hat{\theta}_j)^2. \quad (2)$$

The last expression is in the form of a sample variance of the $\hat{\theta}_{ps,i}$ divided by n , much like the estimated variance of the sample mean. In fact, when $\hat{\theta} = Y$, then $\hat{\theta}_{ps,i} = Y_i$ and the jackknife reproduces results for the sample mean. More generally, suppose that $\hat{\theta}$ has the typical approximation-by-averages representation

$$\hat{\theta} - \theta = \frac{1}{n} \sum_{i=1}^n IC(Y_i, \theta) + R_n, \quad (3)$$

where $\sqrt{n}R_n \xrightarrow{p} 0$ as $n \rightarrow \infty$, θ is the probability limit of $\hat{\theta}$, and $IC(Y_i, \theta)$ is the influence curve (Hampel, 1974). The influence curve was introduced in the context of robust estimation, but here we are not thinking of robust estimation in particular, but using it as the commonly accepted name for the components of the approximating average. If $E\{IC(Y_1, \theta)\} = 0$ and $E\{IC(Y_1, \theta)\}^2 = \sigma_{IC}^2$ exists, then by the Central Limit Theorem, $\hat{\theta}$ is asymptotically normal with parameters $(\theta, \sigma_{IC}^2/n)$. The Delta Method, and more generally the Influence Curve Method, estimate σ_{IC}^2/n by a sample variance of $\hat{IC}(Y_i, \hat{\theta})$, where $\hat{IC}$ means that the influence curve has been estimated in addition to replacing θ by $\hat{\theta}$ (or more simply by plugging estimates into σ_{IC}^2/n). Using (3), one can show that $\hat{\theta}_{ps,i} - \hat{\theta} \approx \hat{IC}(Y_i, \hat{\theta})$ and thus the jackknife variance estimate $\hat{V}_j$ in (2) is just an approximate version of the Delta and Influence Curve Methods. An appealing feature of $\hat{V}_j$ is that you do not need to know calculus or $IC(Y_i, \theta)$ in order to calculate $\hat{V}_j$. You just need a computer program.

To illustrate the calculations, consider the data in Table 3 on the biomass from small plots at the end of one year for three types of planting methods for a new variety of centipede grass (from seed, sprigs, or a combination of seed and sprigs). There is interest in comparing the means of the three planting methods as well as the standard deviations, but it might make

more sense to compare the relative standard deviations, i.e., the coefficients of variation (CV). Using seed is much cheaper than using sprigs, but the relative standard deviation 1.38 appears high compared to the 0.54 for sprigs or to the 0.86 for a combination of sprigs and seeds. We have added $SE_J = (V_J)^{1/2}$ in parentheses, and now explain their computation in R.

To get the standard errors in Table 3, we first give R code for the jackknife variance estimate adapted from the Appendix of Efron and Tibshirani (1993):

```
jack.var <- function(x, theta, ...){
  n <- length(x)
  u <- rep(0, n)
  for(i in 1:n) {
    u[i] <- theta(x[-i], ...) # leave-1-out estimators
  }
  jack.var <- ((n - 1)/n) * sum((u - mean(u))^2)
  return(jack.var)
}
```

For those less familiar with R, the ... in the first line of code allows extra arguments of the function `theta` to be included, as in `jack.var(x, theta=mean, trim=.10)` to obtain $\hat{V}_J$ for the 10% trimmed mean. Putting the first row of Table 3 in data vector `y`, the call producing $\hat{V}_J$ is `jack.var(y, theta=cv)`, where `cv=function(y){sd(y)/mean(y)}`. Repeating for rows two and three and taking square roots gives standard errors 0.34, 0.16, and 0.24, respectively.

The (nonparametric) bootstrap is perhaps easier to explain than the jackknife. From the sample $Y_1, \dots, Y_n$, we draw a random sample of size n with replacement, say $Y_1^*, \dots, Y_n^*$, and compute the estimator from this resample, say θ^* . To illustrate with the first sample in Table 3, a resample might look like 2, 79, 79, 11, 11, 2, 15, 89, 2, where repeats occur naturally because of sampling with replacement. Another way to view the resample is that it is an iid sample obtained from an estimate of the distribution function of the data. We independently repeat this process a total of B times resulting in a sample of estimators $\hat{\theta}_1^*, \dots, \hat{\theta}_B^*$. Then the bootstrap variance estimate is the sample variance of this sample of estimators,

$$\hat{V}_B = \frac{1}{B-1} \sum_{i=1}^B \left(\hat{\theta}_i^* - \bar{\hat{\theta}}^* \right)^2. \quad (4)$$

The approximation-by-averages representation (3) may be used to prove consistency of the bootstrap, but it is more intuitive to view $\hat{V}_B$ as an analogue to the definition of the variance of $\hat{\theta}$. To calculate $\hat{V}_B$ in R, we use

```
boot.var <- function(x,B,theta, ...){
  n <- length(x)
```



```

    bootsam <- matrix(sample(x,size=n*B,replace=T),nrow=B,ncol=n)
    thetastar <- apply(bootsam,1,theta,...)
    boot.var <- var(thetastar)
    return(boot.var)
}

```

Applying this function to the first row of Table 3 is simple,

```

set.seed(284)
boot.var(y,B=1000,theta=cv)

```

but there is the slight added complexity of specifying B and setting the random number seed in order to be able to replicate the calculation. Repeating the call for the second and third samples and taking square roots leads to bootstrap standard errors 0.32, 0.14, and 0.20. How large is the effect of using $B = 1,000$ instead of $B = \infty$? Rerunning the calculations starting with seed=285 yields (0.33, 0.14, 0.21), not much different. Efron and Tibshirani (1993, Section 6.4) give calculations suggesting that $B = 200$ is adequate for bootstrap standard errors in most situations, but because the computing cost is so low here and generally for Monte Carlo standard errors, we recommend $B = 1,000$ as a default value. Comparing to the jackknife standard errors (0.34, 0.16, 0.24) in Table 3, there is a suggestion that the jackknife has slightly larger standard errors on average (see Efron and Stein, 1981, for some general results on the upward bias of $\hat{V}_j$).

Introductions to the bootstrap and jackknife and comparisons may be found in Efron (1982), Efron and Tibshirani (1993), and Boos and Stefanski (2013). Clearly there can be differences in their estimated variances in small samples, but the key point for our use of the 12 bootstrap and jackknife to obtain Monte Carlo standard errors is that they are very likely to be identical when rounded to the appropriate number of digits. Their implementation in R is equally easy, as seen here for simple univariate samples and to be seen in the next section for multivariate samples (e.g., for a Monte Carlo result matrix $\mathbf{X}$ with multiple columns). The bootstrap has a slightly more complicated call but has a slightly wider scope (e.g., can handle Monte Carlo estimates of percentiles such as the median found in boxplots) and flexibility in terms of the number of computations (the resample size B is not tied to the number of Monte Carlo replications N).

4 Jackknife and Bootstrap Calculations in R for Monte Carlo Summaries

Now we turn our attention back to the central focus of the paper, Monte Carlo standard errors like those reported in Tables 1 and 2, and denoted SE_{MC} . R code will be given in terms of `jack.var`, but the bootstrap calls via `boot.var` are exactly the same except for adding the argument B . Recall that the Monte Carlo result matrix $\mathbf{X}$ used to create those tables is the $N \times 3$ matrix with each row of the form $((\hat{C}\hat{V}, \hat{C}\hat{V}_{BC}, A\hat{V}_{ar}))$ calculated from a single computer-generated data set $Y_1, \dots, Y_n$.

Actually, there are 6 different situations reported in those tables with 6 different matrices $\mathbf{X}$ corresponding to sampling from $N(2,1)$ populations, $\text{gamma}(2,1)+0.828$ populations, and standard exponential populations, respectively, at $n = 20$ and $n = 50$. Note that each population has $CV = 0.5$, but the results in the tables depend mainly on the skewness and kurtosis values for those populations, $(\text{Skew}, \text{Kurt}) = (0,3)$ for the normal population, $(\text{Skew}, \text{Kurt}) = (\sqrt{2}, 6)$ for the gamma ($\alpha = 2$) population, and $(\text{Skew}, \text{Kurt}) = (2,9)$ for the exponential population. In the following, $\mathbf{X}$ will refer to one of those 6 matrices.

For each sampling situation in Table 1, the entries are simply sample means and variances. Thus, the R code using `jack.var` from Section 3 to obtain jackknife variance estimates for columns 1–4 and 7 are

```
var.J1 <- jack.var(x=X[,1],theta=mean)
var.J2 <- jack.var(x=X[,2],theta=mean)
var.J3 <- jack.var(x=X[,1],theta=var)
var.J4 <- jack.var(x=X[,2],theta=var)
var.J7 <- jack.var(x=X[,3],theta=mean)
```

and SE_{MC} values follow by taking square roots (and multiplying by n when the summary is multiplied by n). For the mean squared errors, we need to add several columns to $\mathbf{X}$ consisting of squared deviations from the true parameter value 0.5, and then mimic the above with `theta=mean`. For all the cases based on means, we do not need the jackknife because the exact same result is obtained by `var(X[,i])/N`. Also, because the sample variance has a well-known asymptotic variance in terms of second and fourth central moments, it is easy to get standard errors for an S_N^2 without using the jackknife. However, for other summaries it can be more challenging to find and estimate the asymptotic variance. For example, in this study of the coefficient of variation we might also be interested in the approximate normality of $\hat{C}_V$ and $\hat{C}_{V_{BC}}$, and assess this with estimates of the Pearson skewness and kurtosis measures defined in the Introduction, given here in R code by

```
skew <- function(x){mean((x-mean(x))^3)/(mean((x-mean(x))^2)^(1.5))}
kurt <- function(x){mean((x-mean(x))^4)/(mean((x-mean(x))^2)^2)}
```

Then to get the jackknife variance estimates for these Monte Carlo summaries, the R code is

```
var.J.skew <- jack.var(x=X[,1],theta=skew)
var.J.kurt <- jack.var(x=X[,1],theta=kurt)
```

The point is that the jackknife computation is completely trivial compared to the Delta Method described in the Introduction.

For statistics that are functions of a vector like a ratio of two sample variances, the jackknife computation is slightly more complicated. R code to compute the ratio of Monte Carlo variances (not displayed in Table 2) and its jackknife variance estimate is

```
ratio.var <- function(index,xdata){var(xdata[index,1])/var(xdata[index,2])}
var.J <- jack.var(1:N,theta=ratio.var,xdata=X[,c(1,2)])
```

The salient feature here is that `jack.var` needs to receive “data” `1:N` so that it can drop out one row at a time to compute the leave-1-out estimators. Thus the function `ratio.var` also needs to receive data `1:N`, and the real data in the first two columns of X come in as the additional argument `xdata=X[,c(1,2)]`. Similarly, the bootstrap call here is

```
var.B <- boot.var(1:N,B=1000,theta=ratio.var,xdata=X[,c(1,2)])
```

To get the jackknife estimate of variance for the ratio of mean squared errors in Table 2, the code is

```
ratio.mse <- function(index,xdata,true){
  mean((xdata[index,1]-true)^2)/mean((xdata[index,2]-true)^2)}
var.J <- jack.var(1:N,theta=ratio.mse,xdata=X[,c(1,2)],true=0.5)
```

Once standard errors for every table entry have been computed, it is easy to report a summary of these standard errors. In Tables 1 and 2 we have given the mean of the standard errors in each column. Other summaries such as the maximum or median or range might also be used. Such summaries could also be placed in the table caption. There should be freedom to make the standard error summaries informative but not overly obtrusive. In fact, we prefer not to include a separate standard error for each entry because that distracts from the main message of the table. Our working principle is to aim at giving enough information about the variability of entries, but not too much.

5 Summary

We believe that summaries of Monte Carlo results should include standard errors whenever possible. In many situations, these summaries are just sample means, and the results of this paper are unnecessary. However, for any other kind of summary such as a sample variance, percentile, or ratios of quantities, the Delta Method step required is burdensome enough to keep researchers from providing Monte Carlo standard errors. Thus, we think that jackknife or bootstrap standard errors should be routinely used for all summaries, simple and complex, because they are so easy to use. An added benefit of having these standard errors readily available is that when planning Monte Carlo studies, the standard errors computed from preliminary runs can be used for choosing the Monte Carlo replication size N .

ACKNOWLEDGEMENTS

This work was supported by NSF grant DMS-0906421 and NIH grant P01 CA142538-01.

References

- Boos, DD.; Stefanski, LA. Essential Statistical Inference: Theory and Methods. New York: Springer; 2013.
- Bao Y. Finite-Sample Moments of the Coefficient of Variation. *Econometric Theory*. 2009; 25:291–297.
- Cramér, H. Mathematical Methods of Statistics. Princeton: Princeton University Press; 1946.
- Efron B. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*. 1979; 7:1–26.
- Efron B, Stein C. The Jackknife Estimate of Variance. *The Annals of Statistics*. 1981; 9:586–596.
- Efron, B. The Jackknife, the Bootstrap, and Other Resampling Plans. New Philadelphia: Society for Industrial and Applied Mathematics; 1982.
- Efron, B.; Tibshirani, RJ. An Introduction to the Bootstrap. New York: Chapman & Hall; 1993.
- Gelman A, Pasarica C, Doddhia R. Let's practice what we preach: turning tables into graphs. *American Statistician*. 2002; 56:121–130.
- Hampel FR. The Influence Curve and Its Role in Robust Estimation. *Journal of the American Statistical Association*. 1968; 69:567–582.
- Miller RG. Jackknifing Variances. *The Annals of Mathematical Statistics*. 1974; 39:383–393.
- Niebuhr T, Kreiss J-P. Asymptotics for Autocovariances and Integrated Periodograms for Linear Processes at Lower Frequencies. *International Statistical Review*. 82:123–140.
- Quenouille MH. Approximate Use of Correlation in Time Series. *Journal of the Royal Statistical Society, Series B*. 1949; 11:18–84.
- Quenouille MH. Notes on Bias in Estimation. *Biometrika*. 1956; 43:353–360.
- Serfling, RJ. Approximation Theorems of Mathematical Statistics. New York: Wiley; 1980.
- Shao, J.; Tu, D. The Jackknife and Bootstrap. New York: Springer; 1996.
- Tukey J. Bias and Confidence in not Quite Large Samples (abstract). *The Annals of Mathematical Statistics*. 1958; 29:614–614.
- Ver Hoef JM. Who Invented the Delta Method? *The American Statistician*. 2012; 66:124–127.
- Zhu H. Non-parametric Analysis of gap Times for Multiple Event Data: An Overview. *International Statistical Review*. 2014; 82:106–122.

Table 1

Monte Carlo estimated bias, variance (times n), and mean squared error (times n) of $\hat{CV}$ = sample coefficient of variation and of the reduced-bias version CV_{BC} , and the average estimated asymptotic variance (times n). Based on $N = 4,000$ samples of size $n = 20$ and $n = 50$ from normal, gamma ($\alpha = 2$), and exponential populations, each with true $CV = 0.5$.

Population	$n = 20$					
	Bias		nS_N^2		$n\hat{MSE}$	
	$\hat{CV}$	CV_{BC}	$\hat{CV}$	CV_{BC}	$\hat{CV}$	$\overline{nA\hat{Var}}$
normal	0.002	0.000	0.21	0.20	0.21	0.19
gamma ($\alpha = 2$)	-0.013	-0.004	0.17	0.20	0.18	0.10
exponential	-0.029	-0.015	0.23	0.28	0.25	0.11
	$n = 50$					
	normal	0.001	0.000	0.19	0.19	0.19
	gamma ($\alpha = 2$)	-0.006	-0.001	0.18	0.20	0.15
	exponential	-0.012	-0.004	0.26	0.30	0.19
	Average SE_{MC}	0.001	0.001	0.006	0.006	0.005

“Average SE_{MC} ” is the mean of the 6 standard errors (not displayed) in each column. Jackknife and bootstrap give the same averages to 3 decimals.

Table 2

Monte Carlo estimated bias of $\hat{CV}$ = sample coefficient of variation and of the reduced-bias version CV_{BC} , ratio of the estimated mean squared error of $\hat{CV}$ to that of CV_{BC} , and ratios of the mean of the asymptotic variance estimate to the estimated variances. Based on $N = 4,000$ samples of size $n = 20$ and $n = 50$ from normal, gamma ($\alpha = 2$), and exponential populations, each with true $CV = 0.5$.

Population	$n = 20$					
	Bias		MSE Ratio		$\widehat{AVar}/S_N^2$	
	$\hat{CV}$	CV_{BC}	$\hat{CV}/CV_{BC}$	$\hat{CV}$	CV_{BC}	CV_{BC}
normal	0.002	0.000	1.07	0.92	0.98	0.98
gamma ($\alpha = 2$)	-0.013	-0.004	0.87	0.61	0.52	0.52
exponential	-0.029	-0.015	0.88	0.47	0.39	0.39
$n = 50$						
normal	0.001	0.000	1.02	0.99	1.01	1.01
gamma ($\alpha = 2$)	-0.006	-0.001	0.91	0.80	0.72	0.72
exponential	-0.012	-0.004	0.89	0.70	0.61	0.61
Average SE_{MC}	0.001	0.001	0.005	0.02	0.02	0.02

„Average SE_{MC} ” is the mean of the 6 standard errors (not displayed) in each column from the jackknife (and bootstrap).

Table 3

Biomass of a New Variety of Centipede

	$\bar{y}$										s	$\hat{C}V(SE_J)$
Seed:	1	2	79	5	17	11	2	15	85	24.1	33.4	1.38 (0.34)
Sprig:	37	60	48	14	76	23				43.0	23.2	0.54 (0.16)
Combo:	3	61	7	5	27	25	35	17		22.5	19.4	0.86 (0.24)