On the Assessment of Monte Carlo Error in Simulation-Based Statistical Analyses

Elizabeth KOEHLER, Elizabeth BROWN, and Sebastien J.-P. A. HANEUSE

Statistical experiments, more commonly referred to as Monte Carlo or simulation studies, are used to study the behavior of statistical methods and measures under controlled situations. Whereas recent computing and methodological advances have permitted increased efficiency in the simulation process, known as *variance reduction*, such experiments remain limited by their finite nature and hence are subject to uncertainty; when a simulation is run more than once, different results are obtained. However, virtually no emphasis has been placed on reporting the uncertainty, referred to here as Monte Carlo error, associated with simulation results in the published literature, or on justifying the number of replications used. These deserve broader consideration. Here we present a series of simple and practical methods for estimating Monte Carlo error as well as determining the number of replications required to achieve a desired level of accuracy. The issues and methods are demonstrated with two simple examples, one evaluating operating characteristics of the maximum likelihood estimator for the parameters in logistic regression and the other in the context of using the bootstrap to obtain 95% confidence intervals. The results suggest that in many settings, Monte Carlo error may be more substantial than traditionally thought.

KEY WORDS: Bootstrap; Jackknife; Replication.

1. INTRODUCTION

Statistical experiments, more commonly referred to as Monte Carlo or simulation studies, are used to investigate the behavior of statistical methods and measures under controlled situations. Recent computing advances have led to an increasing popularity of simulation studies as powerful alternatives to formula-based approaches in statistical settings where analytic solutions are unavailable. Such settings include the estimation of standard errors under complex models, power and sample size calculations under various study designs, evaluation of operating characteristics such as bias and relative efficiency for competing methods/estimators, investigation of model misspecification, microsimulation modeling, and characterization of Bayesian posterior distributions.

Typically, a simulation study involves the repeated application of three steps; (i) generating simulated data, (ii) performing some statistical procedure, and (iii) recording the results. After some (necessarily) finite repetition of these steps, a summary statistic is usually calculated. Although flexible and often insightful, Monte Carlo studies are limited by their finite nature, and as such are subject to sampling variability similar to that of any scientific investigation based on a finite sample from a broader population; when a simulation is run more than once, different results are obtained. Here we call this between-simulation variability *Monte Carlo error* (MCE) (e.g., Lee and Young 1999). The extent to which differences occur across simulations depends on the setting of the experiment, as well as on the number of simulated data sets or *replicates*.

The importance of MCE has long been recognized; in the closing statements of their introductory article, Metropolis and Ulam (1949) noted that "in particular, it seems very difficult to estimate in a precise fashion the probability of the error due to the finiteness of the sample. This estimate would be of great practical importance, since it alone would allow us to suit the size of the sample to the desired accuracy." Whereas there is a broad literature on techniques developed to improve efficiency of simulations (e.g. Ripley 1987; Efron and Tibshirani 1993; Gentle 2002; Robert and Casella 2004; Givens and Hoeting 2005), less emphasis has been placed on evaluating and reporting MCE in a broad range of settings. Relevant work in specific settings includes bootstrap estimation of standard errors and confidence intervals (Hall 1986; Booth and Sarkar 1998; Lee and Young 1999), evaluation of integrals (Booth and Caffo 2002; Jank and Booth 2003), and assessment of convergence of Markov chain Monte Carlo schemes in Bayesian inference (Geyer 1992; Roberts 1996; Kosorok 2000; Flegal, Haran, and Jones 2008).

More generally, however, the lack of emphasis in this area may best be reflected by our survey of 2007 issues of *Biometrics*, *Biometrika*, and *JASA*. Although we provide more details later, here we note that of 223 regular articles that reported a simulation study, only 8 provided either a formal justification for the number of replications used or an estimate of MCE. Motivated by this apparent lack of consideration for reporting MCE, in this article we seek to renew attention to MCE. We believe that increased reliance on simulation-based assessment of

Elizabeth Koehler is Biostatistician, Department of Biostatistics, Vanderbilt University, Nashville, TN 37232 (E-mail: *e.koehler@vanderbilt.edu*). Elizabeth Brown is Assistant Professor, Department of Biostatistics, University of Washington, Seattle, WA 98195 (E-mail: *elizab@washington.edu*). Sebastien J.-P. A. Haneuse is Associate Scientific Investigator, Division of Biostatistics, Group Health Center for Health Studies, Seattle, WA 98101 (E-mail: *haneuse.s@ghc.org*). The authors thank the editor, an associate editor, and two referees for their helpful and constructive comments.

 DOI: 10.1198/tast.2009.0030

statistical procedures has made the reporting of MCE more important; therefore, a key goal of this article is to provide simple and practical tools for estimating MCE in a broad range of settings, as well as to provide a means for determining the number of replications required to achieve a prespecified desired level of accuracy.

The remainder of the article is organized as follows. Section 2 outlines some notation, defines MCE, and presents a simple example illustrating that MCE generally may be more substantial than traditionally thought. Section 3 presents the results of a survey conducted to examine the extent to which MCE is considered in the publication of simulation-based results. Section 4 outlines a series of simple and practical numerical and graphical tools for monitoring and quantifying MCE. Section 5 demonstrates the methods as applied to bootstrap-based confidence interval estimation. Finally, Section 6 concludes with a brief discussion.

2. MONTE CARLO ERROR

The assessment of MCE requires consideration of the specific structural and distributional assumptions of the simulation, referred to here as the "design." The details of the design collectively form a data-generating mechanism, denoted by $f_X(\cdot)$, from which it is assumed that we are able to generate independent data samples or replicates, $\{X_1, X_2, \ldots\}$. Given a particular design, let ϕ denote some target quantity of interest and $\hat{\phi}_R$ denote the Monte Carlo estimate of ϕ from a simulation with R replicates.

2.1 Definition

We define *Monte Carlo error* to be the standard deviation of the Monte Carlo estimator, taken across hypothetical repetitions of the simulation, where each simulation is based on the same design and consists of R replications:

$$\mathrm{MCE}(\hat{\phi}_R) = \sqrt{\mathrm{Var}[\hat{\phi}_R]}. \tag{1}$$

We have chosen to characterize error using the standard deviation rather than the variance, because this scale better reflects the scale of investigation.

Traditionally, efforts to decrease uncertainty associated with simulations have focused on developing techniques that yield more efficient simulation schemes, such as importance sampling and conditioning, and are referred to as *variance reduction* (Ripley 1987, Chapter 5). Here we consider a static simulation framework and consider uncertainty specifically related to the choice of simulation sample size, R.

2.2 Illustrative Example

To illustrate MCE, consider a simple example in the context of logistic regression. Suppose that interest lies in the association between a binary exposure X and a binary outcome Y, and assume that the two are related via the logistic regression model

$$\mathrm{logit}\,\mathrm{P}(Y = 1 \mid X) = \beta_0 + \beta_X X. \tag{2}$$

We conducted a simulation consisting of R replicates, each with $N = 100$ individuals with exposure prevalence fixed at $\mathrm{P}(X = 1) = 0.3$, 30 individuals with $X = 1$, and 70 individuals with $X = 0$. Outcomes were generated as Bernoulli random variables based on (2) with $\beta_0 = -1$ and $\beta_X = \log(2)$. Let $\hat{\beta}_X^r$ denote the maximum likelihood estimator (MLE) of β_X based on the rth replicate, $r = 1, \ldots, R$.

We consider three operating characteristics of the MLE: percent bias, coverage of the standard 95% confidence interval, and power to detect an association based on a Wald test. A Monte Carlo estimate of the percent bias for the MLE of β_X is given by

$$\hat{\phi}_R^b = \frac{1}{R}\sum_{r=1}^{R} \frac{\hat{\beta}_X^r - \beta_X}{\beta_X} \times 100. \tag{3}$$

A Monte Carlo estimate of the coverage probability is given by

$$\hat{\phi}_R^c = \frac{1}{R}\sum_{r=1}^{R} I[\hat{\beta}_X^r - 1.96\hat{\mathrm{se}}(\hat{\beta}_X^r) \le \beta_X \le \hat{\beta}_X^r + 1.96\hat{\mathrm{se}}(\hat{\beta}_X^r)], \tag{4}$$

where $I[\cdot]$ is an indicator taking a value of 1 if the argument is true and 0 otherwise, and $\hat{\mathrm{se}}(\cdot)$ denotes the usual standard error estimate from the logistic regression. Finally, a Monte Carlo estimate of the power to detect an association is given by

$$\hat{\phi}_R^p = \frac{1}{R}\sum_{r=1}^{R} I\left[\left|\frac{\hat{\beta}_X^r}{\hat{\mathrm{se}}(\hat{\beta}_X^r)}\right| > 1.96\right]. \tag{5}$$

2.3 Results

Figure 1 provides a summary of the Monte Carlo percent bias calculations for $\hat{\beta}_X$ for five separate simulation runs, each consisting of $R = 10{,}000$ replicates. At any given value of R, the height of the line represents the Monte Carlo estimate of percent bias, $\hat{\phi}_R^b$, had the simulation been stopped at that point. We see that even in this relatively simple and straightforward

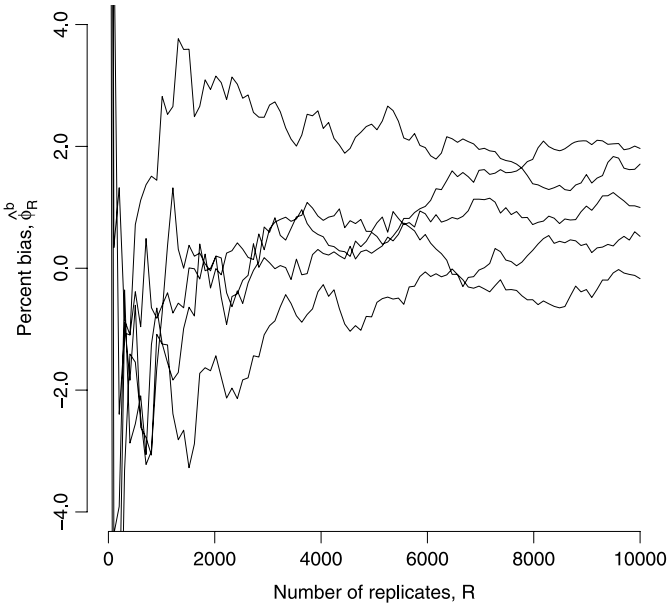

Figure 1. Monte Carlo estimates of percent bias for the MLE $\hat{\beta}_X$, $\hat{\phi}_R^b$, as a function of the number of replicates R, for five simulation runs.

Table 1. Summaries of the Monte Carlo sampling distribution, based on $M = 500{,}000$ simulations, for estimators of three operating characteristics of the log-odds ratio $\hat{\beta}_X$ in the simple logistic regression example of Section 2.2.

Operating characteristic	Number of replications, R	Min.	Max.	Mean	MCE*
Percent bias, $\hat{\phi}_R^b$	100	−34.4	35.3	0.9	6.8
	500	−13.7	15.6	0.9	3.1
	1000	−9.2	12.0	0.9	2.2
	2500	−5.1	7.3	0.9	1.4
	5,000	−4.1	5.6	0.9	1.0
	10,000	−2.3	4.7	0.9	0.7
95% CI coverage, $\hat{\phi}_R^c$	100	84.0	100.0	95.3	2.1
	500	90.4	99.0	95.3	0.9
	1000	91.9	98.1	95.3	0.7
	2500	93.2	97.2	95.3	0.4
	5,000	93.9	96.7	95.3	0.3
	10,000	94.3	96.2	95.3	0.2
Power, $\hat{\phi}_R^p$	100	14.0	55.0	33.0	4.7
	500	24.0	43.4	33.0	2.1
	1000	26.4	39.9	33.0	1.5
	2500	28.7	37.4	33.0	0.9
	5,000	30.0	36.1	33.0	0.7
	10,000	30.8	35.4	33.0	0.5

*Estimated standard deviation of the Monte Carlo sampling distribution, given by (1).

setting, after 10,000 replicates, there is a surprising amount of between-simulation variability in the results; the final point estimates, $\hat{\phi}_R^b$, range between −0.17 and 1.97.

Building on the five simulation runs in Figure 1, we repeated the simulation a total of $M = 500{,}000$ times. For each value of R, we calculated the empirical *Monte Carlo sampling distribution*, based on M experiments, for the estimator of each operating characteristic.

Table 1 provides summary statistics of the three Monte Carlo sampling distributions, including the minimum, maximum, mean, and standard deviation (i.e., MCE). We see that for $R = 1000$, the estimation of percent bias for the MLE $\hat{\beta}_X$ is subject to substantial between-simulation variation; across the M simulations, point estimates $\hat{\phi}_R^b$ range between −9.2% and 12.0%, and the MCE is 2.2%. Although not shown, the central 95% mass of the Monte Carlo sampling distribution is between −3.3% and 5.1%. At $R = 10{,}000$, the minimum and maximum across the M simulations are −2.3% and 4.7%, with MCE decreasing to 0.7%. Practically, this result suggests that ensuring that the central 95% mass of the Monte Carlo sampling distribution for percent bias is within one unit of the overall underlying value of 0.9% (i.e. −0.1%–1.9%) requires that the number of replicates exceed $R = 10{,}000$. For the coverage probability calculations, there is less MCE; Table 1 suggests that around 2500 replications are required to be within one unit of the true value 95% of the time.

We repeated the entire simulation, fixing the marginal exposure probability to equal 0.1 and then 0.5. Although we do not give detailed results here, we found that MCE was greater for $\hat{\phi}_R^b$ when $P(X = 1) = 0.1$ compared to when $P(X = 1) = 0.3$, likely a function of the decreased exposure variation. For example, when $R = 100$, the MCE was 11.1%, and when $R = 1000$, the MCE was 3.5%. In this setting, both $\hat{\phi}_R^c$ and $\hat{\phi}_R^p$ had a comparable MCE to that presented in Table 1, as did all three operating characteristics when $P(X = 1) = 0.5$. Finally, we also repeated the entire simulation, permitting the number exposed to vary across repetitions, setting the number exposed to be a binomial random variable with $P(X = 1) = 0.3$. The results were identical to those given in Table 1, indicating that, at least in this simple setting, introducing additional variability into the marginal exposure distribution across repetitions did not affect the magnitude of the MCE.

3. REPORTING OF SIMULATION STUDIES

The results given in Table 1 serve to illustrate two key points. First, even in simple settings such as logistic regression with a single binary exposure, where simulation-based estimators may be expected to be relatively well behaved, MCE can be substantial. Thus, to obtain accurate Monte Carlo estimates of quantities such as bias and power, we may need to perform a simulation with surprisingly large numbers of replications. Second, the magnitude of MCE, and thus the number of replications required, depends on both the design $f_X(\cdot)$ and the target quantity of interest ϕ. As such, whereas "rules of thumb" are useful in a wide range of settings (e.g. van Belle 2002), it seems unlikely that a single choice for R will provide practical guidance in a broad range of simulation settings. Consequently, for a reader to fully understand and place into context results obtained via a simulation study, the results should be accompanied by some measure of associated uncertainty.

To gauge the extent to which Monte Carlo error is considered and/or reported in the current literature, we conducted a survey of published articles from a nonrandom sample of three statistics journals: *Biometrics*, *Biometrika*, and *JASA*. We considered all regular articles published in 2007, excluding only those for which MCMC was used as part of a single analysis; Bayesian simulation studies, where the entire MCMC process was repeated, were retained. Each article was downloaded electronically, and a search was performed for any of the following terms: "bootstrap," "dataset," "Monte Carlo," "repetition," "replication," "sample," and "simulation." In addition, when indicated by the main article, we also performed the search on supplementary materials available online. Articles for which the search returned a positive result were read in detail to determine whether or not a simulation-based result was reported and, if so, whether or justification for the number of replications and/or estimates of MCE was provided.

Of the 328 regular articles studied, 223 reported the results of a simulation study; only 8 reported estimates of MCE. In a similar survey conducted more than 20 years ago, Hauck and Anderson (1984) found that of the 294 regular articles published in the same three journals in 1981, 63 reported the results of a simulation study, and of those 63, 5 reported some justification for the number of replications. We also recorded the number of replications for each article. Some articles had multiple simulations, for which varying levels of R were used; in such cases we took the largest reported value of R. From Table 2, of the 223 articles reporting a simulation study, 5 did not explicitly report R. For those that did report R, we see wide variability in the number

Table 2. Number of replications associated with simulation studies reported in regular articles published in 2007.

R	*Biometrics*	*Biometrika*	*JASA*	Total
10	3			3
20	1			1
30			1	1
50	4	4		8
100	13	5	13	31
150			1	1
200	4		6	10
250			1	1
300	1		1	2
400	3		2	5
500	9	5	10	24
600	2		1	3
750	1			1
1000	29	22	23	74
1500	1			1
2000	4	2	2	8
3,000			2	2
4,000	1			1
5,000	5		8	13
9,000			1	1
9,999	1			1
10,000	4	6	10	20
20,000			1	1
100,000	1	1	1	3
199,999		1		1
1000,000		1		1
Not reported	4		1	5
Total	91	47	85	223

of replications used. The most common choice was $R = 1000$ (74 articles); only 5 articles used a value of $R > 10{,}000$.

Without the benefit of a reported justification for R, and given the often-complex nature of many recently proposed methods, it is reasonable to assume that the specific choice for many of these articles was driven by time constraints imposed by the computational burden or by the somewhat arbitrary use of round numbers (often multiples of 100 or 1000). Although work continues on improving the efficiency of simulations (e.g. Efron and Tibshirani 1993; Robert and Casella 2004; Givens and Hoeting 2005), in many cases little can be done to substantially reduce the time needed to run even a single iteration, especially as problems to which simulations are applied become increasingly complex. Given the results of the logistic regression example in Section 2.2, however, such simulations may plausibly experience greater MCE than traditionally thought, suggesting that more emphasis should be placed on reporting MCE in the literature.

4. QUANTIFICATION OF MONTE CARLO ERROR

For the example given in Section 2.2, Figure 1 illustrates a simple and effective diagnostic tool for monitoring the simulation as R increases. In addition to plotting the running value of the Monte Carlo estimate, one could provide additional detail on overall uncertainty by augmenting the plot with running standard errors (e.g. Robert and Casella 2004, Chapter 3). An obvious strategy for using this plot to minimize uncertainty is to wait until estimation levels off at some stationary state and then halt the simulation. But this approach seems somewhat subjective and, moreover, does not provide an estimate of MCE itself. In what follows, we outline various tools for quantifying MCE available to practicing statisticians. These tools have been implemented in an R package (R Development Core Team 2007), which is available online at *http://www.r-project.org/*.

4.1 Asymptotics

For a broad range of quantities commonly evaluated using Monte Carlo techniques, one can appeal to asymptotic theory to derive an estimator of the MCE (e.g. Robert and Casella 2004). In brief, suppose that the target quantity has an integral representation given by

$$\phi = \int \phi(x) f_X(x)\, dx.$$

Given a sample of R replicates generated under the design $f_X(\cdot)$, $\mathbf{X} = \{X_1, X_2, \ldots, X_R\}$, a natural Monte Carlo estimate of ϕ is to take the mean of the integrand, evaluated across the replicates:

$$\hat{\phi}_R(\mathbf{X}) \equiv \hat{\phi}_R = \frac{1}{R}\sum_{r=1}^{R} \phi(X_r).$$

By the strong law of large numbers, $\hat{\phi}_R \longrightarrow \mathrm{E}[\phi(X)] = \phi$, as $R \longrightarrow \infty$. Furthermore, under mild regularity conditions, the central limit theorem guarantees that

$$\sqrt{R}(\hat{\phi}_R - \phi) \longrightarrow_d \mathrm{Normal}(0, \sigma_\phi^2), \tag{6}$$

as $R \longrightarrow \infty$, where $\sigma_\phi^2 = \mathrm{E}[(\phi(X) - \phi)^2]$. Thus, from (6), an estimate of the MCE is easily obtained from the replicates themselves as

$$\widehat{\mathrm{MCE}}_{clt}(\hat{\phi}_R) = \frac{\hat{\sigma}_\phi}{\sqrt{R}} = \frac{1}{R}\sqrt{\sum_{r=1}^{R}(\phi(X_r) - \hat{\phi}_R)^2}. \tag{7}$$

Estimation of σ_ϕ^2 is less straightforward for quantities that do not have an integral representation. In some settings, one can appeal to the delta method (e.g. van der Vaart 1998, Chapter 3); however, this may require detailed analytic calculations, and the goal here is to provide simple, practical, and broadly applicable numerical and graphical tools for assessing MCE for a wide range of ϕ. Furthermore, the evaluation of (7) is based on a single simulation of length R, and its accuracy as an estimator of MCE relies on the availability of sufficient replications to get a good initial estimate of ϕ; as can be seen from Figure 1, it may not be obvious when this is the case.

4.2 Resampling-Based Methods

In the setting of evaluating uncertainty associated with bootstrap calculations, Efron (1992) introduced the "jackknife-after-bootstrap" as a means to evaluate the accuracy solely on the basis of the bootstrap replications themselves. This approach can readily be applied in more general Monte Carlo studies as follows.

Suppose that a simulation consists of R replicates, $\mathbf{X} = \{X_1, X_2, \ldots, X_R\}$, from which the Monte Carlo estimate $\hat{\phi}_R(\mathbf{X})$ is evaluated. For each $r = 1, \ldots, R$, evaluate $\hat{\phi}_{R-1}(\mathbf{X}_{(-r)})$, where $\mathbf{X}_{(-r)}$ is the set $\mathbf{X}$ with the rth replicate removed. The jackknife estimate of MCE is given by

$$\widehat{\mathrm{MCE}}_{jack}(\hat{\phi}_R) = \sqrt{\frac{R-1}{R}\sum_{r=1}^{R}\left(\hat{\phi}_{R-1}\left(\mathbf{X}_{(-r)}\right) - \overline{\hat{\phi}_{R-1}(\mathbf{X})}\right)^2}, \tag{8}$$

where

$$\overline{\hat{\phi}_{R-1}(\mathbf{X})} = \frac{1}{R}\sum_{r=1}^{R}\hat{\phi}_{R-1}\left(\mathbf{X}_{(-r)}\right).$$

A natural alternative to this approach is to add a second level of replication and evaluate a bootstrap estimate of the MCE as follows. Given the R simulation replicates $\mathbf{X}$, generate a bootstrap replicate by sampling from $\mathbf{X}$ with replacement, denoted by $\mathbf{X}^*$, and evaluate the statistic of interest, $\hat{\phi}_R(\mathbf{X}^*)$. Repeat this process B times, to give $\hat{\phi}_R(\mathbf{X}_1^*), \ldots, \hat{\phi}_R(\mathbf{X}_B^*)$. An estimate of the MCE is then the standard deviation across the bootstrap statistics

$$\widehat{\mathrm{MCE}}_{boot}(\hat{\phi}_R, B) = \sqrt{\frac{1}{B}\sum_{b=1}^{B}(\hat{\phi}_R(\mathbf{X}_b^*) - \overline{\hat{\phi}_R(\mathbf{X}^*)})^2}, \tag{9}$$

where

$$\overline{\hat{\phi}_R(\mathbf{X}^*)} = \frac{1}{B}\sum_{b=1}^{B}\hat{\phi}_R(\mathbf{X}_b^*).$$

Efron (1992) originally proposed the jackknife specifically to avoid a second level of replication, noting that "... at present, bootstrap-after-bootstrap seems to be too computationally intensive." Although this currently may be of less concern, in the context of evaluating MCE via a second level of replication, as in (9), the approach requires specification of an additional simulation "sample size," B. This raises the potential need to further monitor MCE associated with the MCE estimates (i.e., uncertainty associated with finite B).

4.3 Bootstrap Grouping Prediction Plot

Whereas (8) and (9) provide broadly applicable estimates of the MCE for a given R, it also would be useful, particularly for planning purposes, to be able to specify some desired level of tolerable uncertainty and obtain the corresponding required number of replications. Here we build on both the asymptotic and resampling methods to develop a novel graphical approach for characterizing MCE, as a function of R. We refer to this plot as the bootstrap grouping prediction (BGP) plot and note that as a byproduct, we can easily predict (approximately) the number of replications needed to achieve a given level of accuracy.

Appealing to the asymptotic results of Section 4.1, we note that as $R \longrightarrow \infty$, we expect the magnitude of MCE to be linear in $1/\sqrt{R}$. Furthermore, as $1/\sqrt{R} \longrightarrow 0$, MCE $\longrightarrow 0$. Thus, by estimating the MCE for at least one value of R, we can exploit these facts to numerically characterize the relationship between MCE and $1/\sqrt{R}$. Several reasonable ways forward exist; a simple approach is as follows. Suppose that an initial set of R^i replicates, $\mathbf{X} = \{X_1, \ldots, X_{R^i}\}$, is generated, and choose some sequence $\mathbf{R}^* = \{\mathbf{R}_1^*, \ldots, \mathbf{R}_p^*\}$ of values less than R^i. For each element in $\mathbf{R}^*$, randomly select a subset of R_j^* replicates from the original $\mathbf{X}$ and obtain an estimate of MCE using either (8) or (9), $j = 1, \ldots, p$. Based on these p estimates and with $1/\sqrt{R^*}$ as the predictor, fit a linear regression constrained to pass through the origin (simply by omitting the intercept). The estimated slope, denoted here by $\hat{\beta}^+$, then can be used to "predict" the value of R for any desired level of MCE:

$$R^+ = \left(\frac{\hat{\beta}^+}{\text{target MCE}}\right)^2. \tag{10}$$

We could use a single estimate of the MCE (i.e., with $p = 1$), in which case the extrapolation would rely on the asymptotic result connecting the single estimate with the point (0, 0) on the $(1/\sqrt{R}, \text{MCE})$ plane. In this setting, the calculation for $\hat{\beta}^+$ is trivial; choosing $p = 2$ or 3 remains computationally convenient and will yield a more stable estimate of the slope. Furthermore, to avoid dependence on initial selection of the p subsets, we could bootstrap the entire procedure, say B^+ times, and take the average across the values.

Finally, we note that this approach could be used recursively, especially for settings in which the magnitude of MCE is unknown. For example, depending on time constraints, we could run the simulation with, say, $R = 500$. We then could apply the BGP plot and obtain a crude estimate of R^+ for a desired level of accuracy, then use this value to guide subsequent planning, in particular establishing a trade-off between time considerations and accuracy. Depending on the nature of the problem, repeating the BGP calculation at, say, $R = 5{,}000$ will lead to a more refined estimates of R^+.

4.4 MCE in the Illustrative Example

Returning to the logistic regression example of Section 2.2, we note that the values reported in Table 1 were themselves based on a simulation. Although the corresponding simulation size was deliberately set to be large ($M = 500{,}000$), the values remain subject to uncertainty due to the finiteness of the simulation.

Table 3 provides assessments of MCE associated with two characteristics of the Monte Carlo sampling distribution of the percent bias, $\hat{\phi}_R^b$. Given the large number of replicates, we did not evaluate the jackknife estimator. Furthermore, because the standard deviation does not have a direct integral representation, we evaluated MCE using only the bootstrap-based estimator. From Table 3, we see that in addition to directly quantifying uncertainty, we also could use the results to form interval estimates. For example, in addition to reporting an estimated mean percent bias of 0.89% when $R = 100$, we could (and perhaps should) report a 95% confidence interval of (0.87%, 0.91%).

5. BOOTSTRAP-BASED 95% INTERVAL ESTIMATION

A common application of simulation-based methods is the use of the bootstrap to calculate standard errors and 95% CI estimates when formulas are either unavailable or impractical to implement. Although we establish standard error estimates for

Table 3. Monte Carlo error associated with estimation of characteristics of the Monte Carlo sampling distribution for percent bias, $\hat{\phi}_R^b$, given a simulation of size $M = 500{,}000$.

Characteristic	R	Estimated value	$\widehat{\text{MCE}}_{clt}$	$\widehat{\text{MCE}}_{boot}$ $B = 100$	$B = 200$	$B = 500$
Mean	100	0.8885	0.0097	0.0091	0.0100	0.0098
	500	0.8956	0.0043	0.0039	0.0039	0.0043
	1000	0.9011	0.0031	0.0034	0.0033	0.0031
	2500	0.9052	0.0019	0.0019	0.0020	0.0019
	5,000	0.9048	0.0014	0.0012	0.0012	0.0013
	10,000	0.9051	0.0010	0.0009	0.0009	0.0010
MCE*	100	6.8359	NA	0.0067	0.0068	0.0065
	500	3.0593	NA	0.0029	0.0029	0.0030
	1000	2.1608	NA	0.0023	0.0023	0.0022
	2500	1.3692	NA	0.0013	0.0014	0.0014
	5,000	0.9674	NA	0.0011	0.0010	0.0009
	10,000	0.6848	NA	0.0007	0.0007	0.0007

*Estimated standard deviation of the Monte Carlo sampling distribution.

logistic regression analyses of case-control data (Prentice and Pyke 1979), to illustrate the methods of Section 4, particularly the use of the BGP plot, we consider the use of the bootstrap as a means to obtain 95% CI estimates for MLEs of the odds ratio parameters.

5.1 Ohio Lung Cancer Data

The context that we consider is that of a hypothetical case-control study using lung cancer mortality data from Ohio. For each of 88 counties, population estimates and lung cancer death counts are available by gender, race, age, and year of death; we focus on data from 1988 for individuals age 55–84 years and, for simplicity, stratify age into three groups: 55–64 years, 65–74 years, and 75–84 years. A more detailed description of the data was provided by Waller et al. (1997).

Let A_1 be a binary indicator of whether or not an individual's age is between 65 and 74 years, inclusively, and let A_2 be a binary indicator of whether or not the age is between 75 and 84 years. Furthermore, let X be a binary indicator of gender (0, male; 1, female) and let Z be a binary indicator of race (0, white; 1, nonwhite). Finally, let $Y = 0/1$ be a binary indicator of lung cancer status. We assume the following logistic disease model:

$$\text{logit}(\pi) = \beta_0 + \beta_{A_1} A_1 + \beta_{A_2} A_2 + \beta_X X + \beta_Z Z, \qquad (11)$$

where $\pi = \text{P}(Y = 1 \mid A_1, A_2, X, Z)$. Out of the full data set comprising 2,220,177 individuals with 5,533 lung cancer deaths, we selected 100 cases and 100 controls to form the hypothetical case-control study. The first row of Table 4 provides the MLEs for the odds ratio parameters.

Various bootstrap-based interval estimates have been proposed (Efron and Tibshirani 1993); for simplicity, we consider forming an interval using the 2.5th and 97.5th percentiles of the bootstrap sampling distribution. To obtain these, we sampled $R = 1000$ data sets with replacement from the case-control data and evaluated the MLEs using each data set. The results are given in the second row of Table 4.

5.2 Evaluation of MCE

To evaluate uncertainty in the interval estimate bounds, we calculated the bootstrap-based MCE estimate, given by (9), for both the 2.5th and 97.5th percentiles of the bootstrap sampling distribution. Following the procedure outlined in Section 4.2,

Table 4. Evaluation of MCE for bootstrap-based 95% CI estimates for odds ratio MLEs in a logistic regression analysis of the hypothetical case-control study of Section 5.1.

	Age 65–74 years $\exp\{\beta_{A_1}\}$	Age 75–84 years $\exp\{\beta_{A_2}\}$	Gender $\exp\{\beta_X\}$	Race $\exp\{\beta_Z\}$
Odds ratio				
MLE	2.34	2.92	0.48	2.52
bootstrap-based 95% CI	(1.24, 5.11)	(1.34, 7.14)	(0.25, 0.90)	(1.06, 8.29)
$\widehat{\text{MCE}}_{boot}$				
2.5th percentile	0.033	0.066	0.007	0.037
97.5th percentile	0.207	0.312	0.025	0.604
Monte Carlo 95% CI				
2.5th percentile	(1.18, 1.31)	(1.21, 1.47)	(0.24, 0.26)	(0.99, 1.13)
97.5th percentile	(4.71, 5.52)	(6.52, 7.75)	(0.85, 0.95)	(7.11, 9.47)
R^+				
Target MCE = 0.05*				
2.5th percentile	720	1040	28	561
97.5th percentile	10,577	26,651	226	96,128
Target MCE = 0.005**				
2.5th percentile	72,030	104,048	2824	56,096
97.5th percentile	1,057,715	2,665,144	22,576	9,612,776
Projected MCE for $R = 10{,}000$				
2.5th percentile	0.013	0.016	0.003	0.012
97.5th percentile	0.051	0.082	0.008	0.155

*Width of Monte Carlo 95% CI is approximately 0.2.

**Width of Monte Carlo 95% CI is approximately 0.02.

this required a second level of bootstrap replication; we set $B = 1000$. From Table 4, we see that the 2.5th percentile tended to have a fairly low MCE, whereas the MCE for the 97.5th percentile was consistently higher. Using these MCE estimates, we constructed approximate Monte Carlo 95% CIs for each of the percentiles. For example, we note that whereas an initial estimate of 8.29 was obtained for the upper bound of the 95% CI estimate for the MLE of $\exp\{\beta_Z\}$, substantial uncertainty is associated with having generated only $R = 1000$ bootstrap replications; the corresponding Monte Carlo 95% CI is (7.11, 9.47).

To tie down the MCE, we applied the methods described in Section 4.3 and evaluated the BGP plot for the 2.5th and 97.5th percentile estimates using $\mathbf{R}^* = \{200, 500, 1{,}000\}$. For simplicity, Figure 2 provides results solely for the 75- to 84-year age group and race main effects, $\exp\{\beta_{A_2}\}$ and $\exp\{\beta_Z\}$. Based on these plots, Table 4 also provides the projected number of replications, R^+, required to reduce the percent bias MCE to 0.05 or 0.005 for each of the four 2.5th and 97.5th percentiles.

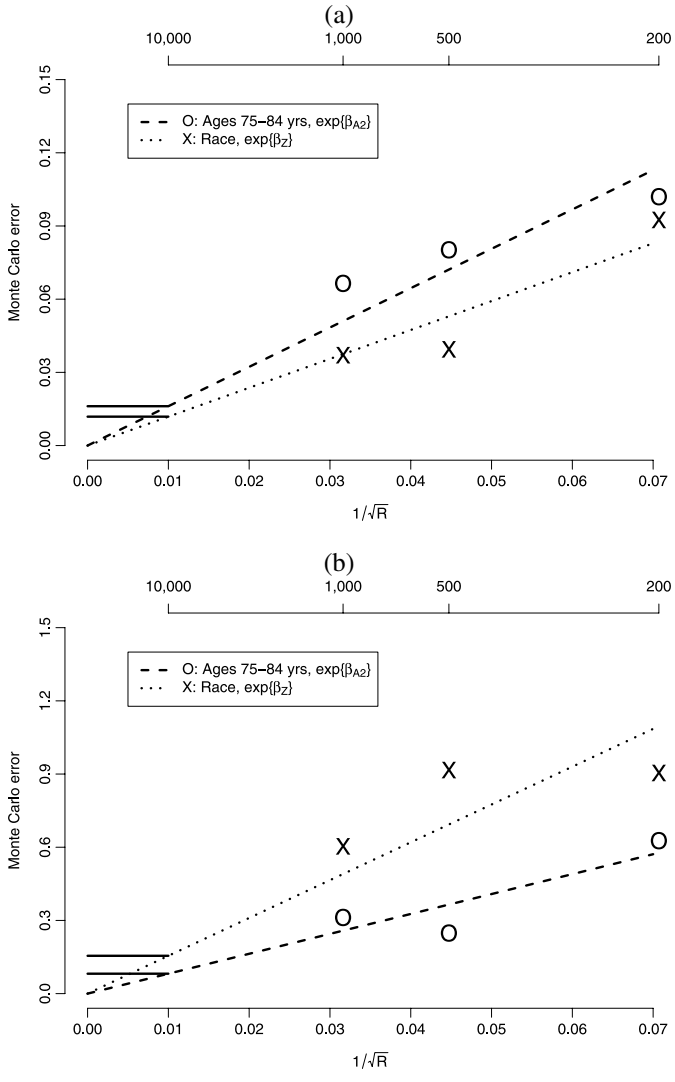

Figure 2. BGP plots, based on $R^* = (200, 500, 1{,}000)$, for the bounds of the bootstrap-based 95% CIs, for two of the odds ratio parameters in the Ohio lung cancer model. Solid lines indicate the projected MCE estimates, based on an anticipated $R = 10{,}000$.

These criteria would reduce uncertainty to levels at which the width of the Monte Carlo 95% CI would be approximately 0.02 and 0.2. Although Table 4 presents both sets of results, given the magnitudes of the respective MLEs, a target MCE of 0.005 might reasonably be used as a threshold for "accurate" estimation of the 2.5th percentile, whereas a target MCE of 0.05 could reasonably be used for the 97.5th percentile.

Overall, we find substantial variability in R^+ across the percentiles and across the odds ratio parameters. For example, accurate Monte Carlo estimation of the 2.5th percentile for the gender main effect odds ratio would require just under $R^+ = 3{,}000$ bootstrap replications; in contrast, accurate estimation for the 97.5th percentile of the race main effect odds ratio would require just under $R^+ = 100{,}000$ bootstrap replications. Although results are not shown, we note that a similar assessment of R^+ on the log-odds ratio scale tends to provide more similar results between the 2.5th and 97.5th percentiles for any given parameter. We have taken the odds ratio scale, because most scientific papers likely would report study results on this scale.

Finally, in most applications there likely will be a practical trade-off between the number of replications and the magnitude of MCE that one is willing to tolerate. From the BGP plots, we can determine a projected MCE for some value of R. For example, Table 4 indicates that if $R = 10{,}000$ bootstrap replications were generated and used as the basis for the bootstrap interval estimates, the projected MCE for the 97.5th percentile of the race main effect would be reduced to 0.155.

6. DISCUSSION

A central role of statisticians is to assess and quantify uncertainty associated with estimation/inference, based on a finite sample from a larger population. In the context of simulation studies, uncertainty associated with a finite sample size (the number of replicates, R) often has been referred to as Monte Carlo error. In this article we have considered three issues relating to MCE. First, the examples presented in Sections 2 and 5 serve to illustrate that MCE may be more substantial than traditionally thought, and that tying down uncertainty to reasonable levels, especially for the relatively complex settings considered in the recent literature, may require running longer simulations than is current common practice (see Table 2). Second, the magnitude of MCE in specific settings likely depends on a range of factors, including the parameter under investigation, the chosen operating characteristic, and the underlying variability in the data. As such, "one-size-fits-all" approaches to MCE may not be reasonable. Third, viewed as statistical or mathematical experiments (Ripley 1987), it could be argued that to aid in the interpretation of results, simulation studies always should be accompanied by some assessment of uncertainty. The literature apparently pays virtually no attention to the reporting of MCE, however. This is in contrast to most scientific studies, in which the reporting of uncertainty (usually in the form of standard errors, p-values, and CIs) is typically insisted on. In overcoming this, reviewers, associate editors, and editors of the statistical literature may need to play a more active role in ensuring the reporting of MCE.

In this work we have focused directly on the standard deviation of the Monte Carlo sampling distribution as a measure of uncertainty associated with R. Other measures of uncertainty have been used as well; a common approach used in previous investigations is to evaluate the coefficient of variation as a measure for determining when to stop a simulation (e.g. Efron and Tibshirani 1993). An important advantage of this measure is that it acknowledges the diminishing returns associated with increasing R. With increasingly powerful computing resources, however, this may be of less concern, and we see an opportunity to add greater emphasis on MCE when designing and reporting statistical experiments. Beyond the uncertainty associated with R, other operating characteristics of a simulation also might be of interest. For example, the assessment of small-sample bias in Monte Carlo estimates may be important in settings where the computational burden is extreme.

The goal of this article has been to cast renewed attention on issues relating to uncertainty in simulations and the reporting thereof. The methods outlined in Section 4 provide practicing statisticians with a range of simple and practical tools for investigating MCE in specific settings. Post hoc MCE calculations are relatively straightforward in a broad range of experimental settings and can provide important insight into uncertainty. The proposed BGP plot also provides a simple approach for determining the number of simulated data sets or replications needed to achieve a desired level of accuracy, and would be particularly useful in planning a simulation study. A broader understanding of the MCE estimators could benefit from future investigation of their operating characteristics. For example, although the bootstrap-based estimator is applicable in a broad range of settings, the required second level of replication (denoted here by B) may quickly become computationally burdensome; thus guidance on how to determine B would be of practical interest.

[Received August 2008. Revised February 2009.]

REFERENCES

Booth, J. G., and Caffo, B. S. (2002), "Unequal Sampling for Monte Carlo EM Algorithms," *Computational Statistics & Data Analysis*, 39 (3), 261–270.

Booth, J. G., and Sarkar, S. (1998), "Monte Carlo Approximation of Bootstrap Variances," *The American Statistician*, 52, 354–357.

Efron, B. (1992), "Jackknife-After-Bootstrap Standard Errors and Influence Functions" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 54, 83–111.

Efron, B., and Tibshirani, R. (1993), *An Introduction to the Bootstrap*, New York: Chapman & Hall.

Flegal, J., Haran, M., and Jones, G. (2008), "Markov Chain Monte Carlo: Can We Trust the Third Significant Figure?" *Statistical Science*, 23 (2), 250–260.

Gentle, J. (2002), *Elements of Computational Statistics*, New York: Springer.

Geyer, C. J. (1992), "Practical Markov Chain Monte Carlo" (with discussion), *Statistical Science*, 7, 473–483.

Givens, G. H., and Hoeting, J. A. (2005), *Computational Statistics*, New Jersey: Wiley.

Hall, P. (1986), "On the Number of Bootstrap Simulations Required to Construct a Confidence Interval," *The Annals of Statistics*, 14, 1453–1462.

Hauck, W. W., and Anderson, S. (1984), "A Survey Regarding the Reporting of Simulation Studies," *The American Statistician*, 38, 214–216.

Jank, W., and Booth, J. (2003), "Efficiency of Monte Carlo EM and Simulated Maximum Likelihood in Two-Stage Hierarchical Models," *Journal of Computational and Graphical Statistics*, 12 (1), 214–229.

Kosorok, M. R. (2000), "Monte Carlo Error Estimation for Multivariate Markov Chains," *Statistics & Probability Letters*, 46 (1), 85–93.

Lee, S. M. S., and Young, G. A. (1999), "The Effect of Monte Carlo Approximation on Coverage Error of Double-Bootstrap Confidence Intervals," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 61, 353–366.

Metropolis, N., and Ulam, S. (1949), "The Monte Carlo Method," *Journal of the American Statistical Association*, 44 (247), 335–341.

Prentice, R. L., and Pyke, R. (1979), "Logistic Disease Incidence Models and Case-Control Studies," *Biometrika*, 66, 403–411.

R Development Core Team (2007), *R: A Language and Environment for Statistical Computing*, Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Ripley, B. (1987), *Stochastic Simulation*, New Jersey: Wiley.

Robert, C., and Casella, G. (2004), *Monte Carlo Statistical Methods* (2nd ed.), New York: Springer.

Roberts, G. (1996), "Markov Chain Concepts Related to Sampling Methods," in *Markov Chain Monte Carlo in Practise*, eds. W. Gilks, S. Richardson, and D. Spiegelhalter, London, U.K.: Chapman & Hall.

van Belle, G. (2002), *Statistical Rules of Thumb*, New Jersey: Wiley.

van der Vaart, A. (1998), *Asymptotic Statistics*, Cambridge, U.K.: Cambridge University Press.

Waller, L. A., Carlin, B. P., Xia, H., and Gelfand, A. E. (1997), "Hierarchical Spatio-Temporal Mapping of Disease Rates," *Journal of the American Statistical Association*, 92, 607–617.