A Guide for Simulation Studies in Statistics

04 November 2009

1 Introduction

One of the goals of statistical research is to find new methods (estimators, tests, confidence sets, etc.) that are better than those in current use. Over the last thirty years, the pursuit of purely mathematical tools to show that a new method is better or the best has been increasingly difficult and sometimes fruitless. In contrast, during that same time period, our ability to confidently write computer code to implement complicated statistical procedures has improved greatly. As a result, the simulation or Monte Carlo experiment has become the most important tool in statistical research. However, statisticians have not been practicing in their own research the same methodology of experimental design and analysis that they promote in classes and consulting with researchers in other disciplines. A simulation experiment is an experiment that should be designed and analyzed according to statistical standards. The purpose of this article is to begin to codify the standards of this methodology.

The goals of a simulation study are commonly the computation of:

1. the bias, variance, and MSE of an estimator
2. the size and power of a hypothesis test
3. the size (length, area, volume) and coverage of a confidence set

for the comparison of methods. Other questions that may be addressed include: Do the standard errors accurate reflect the accuracy of an estimator? For what sample size does the asymptotic distribution apply?

We denote $t(\mathbf{X})$ as the result of statistical procedure: an estimator, a test statistic or indicator variable arising from it, where $\mathbf{X}$ denotes the data. Our mathematical skills allow us to derive the distribution of $t(\mathbf{X})$ only in a limited number of cases. In other situations, we may be able to compute moments to address bias or compare MSE. The power of the simulation experiment is that if we can generate samples or *replicates* $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \ldots, \mathbf{X}^{(N)}$, independently from the same distribution as $\mathbf{X}$, then we can estimate any population parameter of the distribution of $t(\mathbf{X})$ (quantiles, mean (bias), variance, etc.) using sample means.

Consider the case of comparing two estimators. For an estimator $t_m(\mathbf{X})$ of a parameter vector θ based on the data vector $\mathbf{X}$, the bias in this estimator is given by

$$E\left[t_m(\mathbf{X})\right] - \theta^*$$

where θ^* denotes the true value of the parameter vector. The Monte Carlo method for estimating such a quantity is to construct *iid* samples of the data $\mathbf{X}^{(k)}, k = 1, \ldots, N$ from the same distribution as $\mathbf{X}$ and use the sample mean as an estimate of the bias, a population quantity:

$$\frac{1}{N}\sum_{k=1}^{N} \mathbf{t}_m(\mathbf{X}^{(k)}) - \theta^*.$$

To compare two estimators, t_ℓ and t_m on the basis of expected absolute errors, $E\{|t(\mathbf{X}) - \theta^*|\}$, we might estimate these quantities with different samples, that is, estimate $E\{|t_\ell(\mathbf{X}) - \theta^*|\}$ with the sample mean

$$\frac{1}{N}\sum_{j=1}^{N} \left[|t_\ell(\mathbf{X}^{(j)}) - \theta^*|\right] \tag{1}$$

and $E\{|t_m(X) - \theta^*|\}$ similarly with a different sample

$$\frac{1}{N}\sum_{k=1}^{N} \left[|t_m(\mathbf{X}^{(k)}) - \theta^*|.\right] \tag{2}$$

But since estimators $t_\ell(\mathbf{X})$ and $t_m(\mathbf{X})$ are commonly highly positively correlated, using the same data $\mathbf{X}$ would lead to sample means of differences with smaller variance. This fact motivates *pairing* or *blocking*, and suggests estimation of the difference

$$E\{|t_\ell(\mathbf{X}) - \theta^*|\} - E\{|t_m(\mathbf{X}) - \theta^*|\},$$

with the sample mean of differences

$$\frac{1}{N}\sum_{k=1}^{N} \left[|t_\ell(\mathbf{X}^{(k)}) - \theta^*| - |t_m(\mathbf{X}^{(k)}) - \theta^*|\right]. \tag{3}$$

Since the gains from pairing can be substantial, a *paired* analysis or *blocking* by dataset should be employed as a matter of routine; any deviation requires justification.

For this discussion, we will use as an example a simulation study by Swallow & Monahan (1984) on variance component estimators in the presence of imbalance. (It's safer to point out the shortcomings of one's own work.) In the next section, we lay out the usual analyses. Then we'll talk about computational issues, finishing up with a section on displaying the information.

2 Design and Analysis

Consider the problem addressed in Swallow & Monahan in examining the performance of variance component estimators. The problem is the simple one-way random effects model,

$$Y_{ij} = \mu + \alpha_i + e_{ij}, i = 1, \ldots, a; j = 1, \ldots, n_i,$$

where the random effects α_i are *iid* Normal$(0, \sigma_a^2)$ and the errors e_{ij} are *iid* Normal$(0, \sigma_e^2)$. The parameters of the statistical problem are the mean μ, random effects variance σ_a^2, and the error variance σ_e^2. The issue of interest to the authors of this study was the effect of balanced/imbalanced sample sizes of the group: n_i. For this problem, $\mathbf{X}$ represents the vector of data Y_{ij} having a multivariate normal distribution with its covariance structure a function of the two variances σ_a^2, σ_e^2. The different estimators to be compared (ANOVA, ML, REML) are represented by $\mathbf{t}_m(\mathbf{X})$, and the parameter vector is $\theta = (\mu, \sigma_a^2, \sigma_e^2)^T$. In this problem, the focus is on estimation, so that $\mathbf{t}(\mathbf{X})$ is a vector of estimators; at times, however, we may reduce to a scalar quantity $t_m(\mathbf{X})$.

One of the first questions to be addressed is whether an estimator $t_m(\mathbf{X})$ is biased, say, in this case, REML estimates. The natural estimator of the bias (yes, of an estimator) is the sample mean

$$\frac{1}{N}\sum_{k=1}^{N} t_m(\mathbf{X}^{(k)}) - \theta^* = \bar{t}_m - \theta^*,$$

whose accuracy would be assessed with the standard error of the mean

$$\sqrt{\frac{1}{N(N-1)}\sum_{k=1}^{N}\left[t_m(\mathbf{X}^{(k)}) - \bar{t}_m\right]^2}.$$

To say that estimator t_ℓ has less bias than estimator t_m would lead to the difference of sample means $\bar{t}_\ell - \bar{t}_m$ as the estimate of its expectation and inference from the paired t-test. Proper analysis of simulation results requires standard statistical practice: estimates with statements of accuracy, and test statistics with critical values and/or p-values. (Discussion of the assumption of normality is postponed until Section 3.2.) This statement that estimator t_ℓ has less bias than estimator t_m holds only for the distribution of $\mathbf{X}$ that was generated, that is, only for one particular value of the parameters that characterize its distribution: $\mu, \sigma_a^2, \sigma_e^2$, and the pattern of sample sizes $(n_1, \ldots, n_a)$.

Consequently, the design of the simulation experiment requires the identification of the factors that affect the distribution of the data. Often location and scale invariance can reduce the number of effective factors. In this situation, since all of the estimators to be considered are location and scale invariant, we can set $\mu = 0$ and $\sigma_e^2 = 1$ and interpret now σ_a^2 as the ratio of variances, without any loss of generality. But the pattern $P = (n_1, \ldots, n_a)$ of sample sizes remains as the second design factor of this experiment.

To illustrate the potential of a well-designed simulation experiment, consider comparing estimators $t_m(\mathbf{X})$ in term of MSE across different values of variance ratio σ_a^2, say $\sigma_{(i)}^2, i = 1, \ldots, I$, and different patterns $P_j, j = 1, \ldots, J$. Denoting $\mathbf{X}^{(i,j,k)}$ as the kth replicate dataset generated with variance ratio i and pattern j, then taking the sample quantity as the response,

$$Y_{m,i,j,k} = \left[t_m(\mathbf{X}^{(i,j,k)}) - \theta^*\right]^2, \tag{4}$$

opens the door for the full panoply of analysis for an experiment with four factors

- methods m, e.g. ANOVA or ML
- settings i of the variance ratio σ_a^2, e.g. $\sigma_{(1)}^2 = .1$
- sample size pattern j, e.g. $P_1 = (3, 5, 7)$
- replicate k, a common random effect.

To compare method 1 and method 2 in terms of MSE, we would use $\overline{Y}_{1\ldots} - \overline{Y}_{2\ldots}$ with its interpretation driven by its expectation. Its expectation is the difference in MSE averaged over the design space (variance ratio i, pattern j). The accuracy of this difference of means would require the full analysis of a model for the data, which may include interactions and random effects.

3 Implementation issues

3.1 Factors and levels

In this simulation study on variance components, the two experimental factors are the sample size patterns $P = (n_1, \ldots, n_a)$, and the variance component ratios σ_a^2/σ_e^2. Setting the levels of experimental factors is always a problem in experimental design. How did Swallow & Monahan arrive at their levels? Well, they following some previous analysis with some modifications. Following precedents allows for easy comparison as well as verification of results. Textbook examples and examples from research papers form an easy starting point. At some point, nonetheless, the usefulness of the experiment depends on choosing levels that match situations arising in practice. Since inference on which method is better often depends on design space, researchers should set the factor levels to permit comparison with the situations that practitioners – potential users of their conclusions – will face. Someone working on problems with only mild imbalance may criticize the S & M study for overweighting extreme imbalance and disregard its conclusions as inappropriate. Someone looking at small group/treatment effects may say that the large values of σ_a^2 were overemphasized.

For practical purposes, the number of levels is limited by the total computation resources available. The range should be large enough to see an effect of the factor for a reasonable replication size. And if the number of factors becomes large, keep in mind that that there is no law saying that the design must be a full factorial. The full toolbox of experimental designs is available, including screening designs to reduce the number of factors as well as fractional factorial designs.

3.2 Approximate normality by batching

Full analysis of an experiment with a response Y_{mijk} as in [4] often rely on assumptions of additive effects, homoskedasticity, and the normal distribution. While variable transformations, such as log, can help simplify the modeling, a simple tool for improving the distribution is batching. In this problem, there is no basis for assuming that $\left[t_m(\mathbf{X}^{(i,j,k)}) - \theta^*\right]^2$ has a normal distribution, especially when $t_m(\mathbf{X})$ is an estimate of a scale parameter, such as a variance component. However, if the sample size N were decomposed $N = N_1 \times N_2$, then the distribution of

$$Y_{m,i,j,k_1} = \frac{1}{N_2} \sum_{k_2=1}^{N_2} \left[t_m(\mathbf{X}^{(i,j,k_1,k(2))}) - \theta^*\right]^2$$

would be closer to normality, with the loss of sample size. In the common case of $N = 100 = 5 \times 20$ still leaves 5 replicates of a random variable closer to normality.

3.3 Accuracy

Good statistical practice requires that any estimate be accompanied with some measure of its accuracy. The standard error $se_m(\mathbf{X})$ of estimator $t_m(\mathbf{X})$ attempts use the current sample $\mathbf{X}$ to estimate $\sqrt{Var(t_m(\mathbf{X}))}$, the standard deviation of the estimator. This variance would naturally be estimated by the sample Monte Carlo variance,

$$\frac{1}{N-1} \sum_{k=1}^{N} \left[t_m(\mathbf{X}^{(k)}) - \bar{t}_m\right]^2.$$

The fidelity of the standard error to this sample standard deviation is a natural comparison of interest. Often this is best measured by a side-by-side table of mean standard error $\frac{1}{N} \sum_k se_m(\mathbf{X}^{(k)})$ or the ratio of these two.

The choice of the replication size N depends upon the computational resources that are available and the goals of the study. But its determination follows the same logic as in other designed experiments. If we want to be able to determine whether a test was liberal, say, then we are looking for the sample size that would give high power of rejecting the hypothesis that the size were the same as the level. If we say we want high power, meaning .80, at the alternative size .06 compared to level .05, then the calculation follows the usual path:

$$\frac{.06 - .05}{\sqrt{.05 \times .95 \times N}} = z_{.80}.$$

Taking the sample size too small means that the experiment may be unable to determine liberal or conservative tests, or whether one method is better than another. In other words, the experiment may be worthless. On the other hand, taking the sample size too large may mean wasting costly, shared resources unnecessarily. Judicious use of resources may mean balancing the replication size with the number and levels of factors. Rarely can this be done without knowledge of the situation; in other words, pilot studies must be part of the experimental plan.

4 Computational issues

4.1 Random numbers and seeds

Monte Carlo methods rely on pseudorandom methods for the generation of random variables from the uniform distribution on the interval $(0, 1)$, from which all other distributions arise. All methods that are in common use are, in fact, algebraic methods on the finite ring of integers which are periodic and have a starting value or *seed*. Changing the seed generates a new set of random variables that are for most purposes effectively independent. Setting the seed fixes the entire sequence of random variables, so that the simulation experiment can be reproduced in its entirety. Documenting and managing the seeds used in a Monte Carlo experiment are essential for replication of the experiment, either by other scientists, or by the researcher himself/herself.

4.2 Algorithm performance: convergence and failure

These days, most sophistical statistical methods are the result of some iterative method or search procedure that may not work adequately for all possible datum $\mathbf{X}$. Many black box codes for searching or optimization return codes that describe the performance as a success (or possible success), or varieties of failures (overflow, failure to converge, etc.) Potential uses want to know how the success rate may be related to experimental factors. Moreover, algorithm failures present grave difficulties for the researcher. For example, in logistic regression a case of complete separation means that the log-likelihood increase without bound as a parameter value grows infinitely large or small. In other words, with certain samples, the MLE of a parameter is infinite. In some circumstances, the researcher may choose to define a modified MLE and analyze the performance of this modified method. An alternative is to describe the performance of an estimator conditional on the successful results of an algorithm, accompanied by analysis of the algorithm failure rate.

4.3 Starting values

As many methods require search or optimization, codes for these algorithms often require starting values. In fact, choosing good starting values is often essential for a successful search. In a simulation experiment, however, the researcher may be tempted to use information that would not be available to a practitioner. In particular, the researcher may be tempted to use the true parameter value as the start of a search. But the starting value, since it affects the performance of the statistical procedure, is also part of the statistical procedure. Since a practitioner would not have that information, this is not a valid statistical method. The starting values must arise from only the information that a practitioner would have available.

5 Presentation

The perfect simulation experiment would be able to concisely describe the performance of statistical procedure across several factors. In practice, however, a good simulation experiment would produce a wealth of information about the procedures. The only way to summarize a great deal of information in a small space is the use of graphs. A goal for graphics is that a reader should not need the caption to understand the results. Nonetheless, titles, captions, labels, and symbols are essential to clearly describe the results.

More often, however, an important part of the results will be presented in tables. To present results from a study of the performance of an estimator, such a table may take the following form:

Typical Table

true	mean	mean	standard deviation
β^*	$\hat\beta$	se($\hat\beta$)	(MC) of $\hat\beta$
0.35	0.37	0.13	0.17
0.50	0.54	0.15	0.23
1.00	1.11	0.19	0.31

Note that except for the first column, all of the other entries are Monte Carlo estimates whose variability must be expressed in some way. The second column is composed of sample means whose standard errors are, respectively, $0.17/\sqrt{N}$, $0.23/\sqrt{N}$, and $0.31/\sqrt{N}$. The entries in the fourth column are sample standard deviations whose approximate standard errors are $0.17/\sqrt{2N}$, $0.23/\sqrt{2N}$, and $0.31/\sqrt{2N}$.

For another example, Swallow & Monahan present tables of mean squared errors across factors. A simplified form is

Mean squared errors			
	ratio σ_a^2/σ_e^2		
n-pattern	0	.1	.2
(3,5,7)	0.98	1.03	1.00
(2,4,6)	1.05	1.04	1.03
(2,10,18)	1.04	1.01	1.00
(3,15,27)	1.01	1.00	1.02

The accuracy of the entries to this table must be expressed in some fashion. Primarily, the number of digits given should reflect the accuracy. That is, entries in the tables should be two or three digits, and only four if the standard error is in the fourth digit. Do not write 1.26 when the standard error is .5; instead round to 1.3. Often in this situation including standard errors for each entry just clutters the table to make it unreadable. Include in the caption a typical standard error for an entry, or the range.

In most cases, entries in the same row or column are either independent or positively correlated. Including the standard error will permit the reader to do some comparisons through an approximate (and/or conservative) t-test. However, in the table above, the differences across rows or columns may be significant because of the strong positive correlation induced by blocking by common random numbers. For comparisons whose test results are not apparent, the statistical analysis and test results must be clearly elaborated in the text:

> as a randomized complete block experiment with xx as blocks and factors yy and zz. Treatments $yy1$ and $yy2$ are significantly different at level $\alpha = 0.05$ using a t-test with k degrees of freedom and $v1m2$ as the error variance.

with the same details as if it had appeared in *Journal of Plant Science.*

6 Advice

The simulation experiments described in the literature are rarely the result of a single carefully planned experiment. In reality, the researcher tests various aspects of the experiment before planning the experiment that appears in print. First on the agenda is rigorous debugging, documentation, and testing of the computer code. Pilot studies or screening experiments are performed to determine information critical to the planning of a good experiment: screening factors, estimating variances, testing sensitivity of the code. Only with this information can the factors and their levels be chosen and the replication size set to judiciously use the available computer resources.

After the 'final' design is elaborated, lots of numbers crunched, plots and tables carefully designed and documented, the effort will not be complete. Whether the experiment is part of thesis research or for publication, the work will be reviewed and the researcher should expect to modify the experiment some time in the future: add another factor, include a competitor for comparison, etc. Before anything is submitted for further review, the computer code and accompanying documentation must include enough information that another researcher could replicate the study. In our experience, often either an advisor or collaborator may attempt to correct, modify, or expand the simulation experiment in response to a review or to continue that avenue of research. The original researcher may not have the time, inclination, software, or hardware available to do that work. Additionally, in our experience, often the interval of time for review is sufficient for the researcher to completely forget how and what was done in the original experiment. As a result of these experiences, some recommended practices are:

- Document your computer code, add more comments to the code, add still more comments, and when you think you've overdone it and begun to obscure the code with too many comments, then go back and cross-reference both your code and research document/paper/thesis. Equation labels in the document refer to functions or blocks of code, and vice versa.

- Create a list of the files used in the simulation experiment, including debugging trials. Describe what each file does and what data/tables/figures/files are produced.

- Backup your files. You've put a lot of work into it, and the cost of losing that work is high. Keep a second backup of the essentials at a second location – mail a jumpdrive to your mother.

- Ensure that each task stands by itself. This is sometimes best done by packaging the code to run as batch jobs. If advisor's/reviewer's/co-author's comments lead to changes, the entire task could be repeated and everything updated with minimal effort.