Proceedings of the 2002 Winter Simulation Conference
E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, eds.

ON CHOOSING A SINGLE CRITERION FOR CONFIDENCE-INTERVAL PROCEDURES

Bruce Schmeiser

School of Industrial Engineering
Purdue University
West Lafayette, IN 47907–1287, U.S.A.

Yingchieh Yeh

Department of Business Administration
Yuan Ze University
Taoyuan, TAIWAN

ABSTRACT

Stating a confidence interval is a traditional method of indicating the sampling error of a point estimator of a model's performance measure. We propose a single dimensionless criterion, inspired by Schruben's coverage function, for evaluating and comparing the statistical quality of confidence-interval procedures. Procedure quality is usually thought to be multidimensional, composed of the mean (and maybe the variance) of the interval-width distribution and the probability of covering the performance measure (and maybe other values). Our criterion, which we argue lies at the heart of what makes a confidence-interval procedure good or bad, compares a given procedure's intervals to those of an "ideal" procedure. For a given point estimator (such as the sample mean) and given experimental data process (such as a first-order autoregressive process with specified parameters), our single criterion is a function of only the sample size (or other rule that ends sampling).

1 INTRODUCTION

Wilson and Pritsker (1978) propose a single dimensionless criterion for comparing methods for dealing with the initial transient in steady-state simulation experiments. In a similar spirit, but with a different approach, we propose a single dimensionless criterion for evaluating and comparing confidence-interval procedures (CIPs).

In this section we introduce notation and terminology associated with estimation in statistical inference. In Section 2 we discuss issues associated with evaluating and comparing CIPs, in Section 3 we propose a graphical approach and an associated single criterion for evaluating and comparing CIPs, and in Section 4 we list some additional thoughts.

1.1 Estimation

We consider a statistical experiment that estimates the value of a *performance measure* θ by creating a set of *output data* Y and computing from it a *point estimator* $\widehat{\theta}$. The *sampling distribution* of $\widehat{\theta}$ is often normal, with the statistical quality of $\widehat{\theta}$ summarized by its *bias*

$$\text{bias}(\widehat{\theta}, \theta) = \text{E}(\widehat{\theta}) - \theta$$

and its *variance*

$$\text{var}(\widehat{\theta}) = \text{E}((\widehat{\theta} - \text{E}(\widehat{\theta}))^2),$$

which can be combined into its *mean squared error*

$$\text{mse}(\widehat{\theta}, \theta) = \text{E}((\widehat{\theta} - \theta)^2) = \text{bias}^2(\widehat{\theta}, \theta) + \text{var}(\widehat{\theta}).$$

Often both bias and variance are $O(n^{-1})$, where n is the *sample size* (or other measure of size of the output-data set, such as computation time for Monte Carlo simulation experiments). The bias contribution to the mse is then negligible, allowing the quality of $\widehat{\theta}$ to be measured by its *standard error*,

$$\text{ste}(\widehat{\theta}) = \sqrt{\text{var}(\widehat{\theta})}.$$

Alternatively, the quality of $\widehat{\theta}$ can be indicated by a *confidence interval* (L_η, U_η), where the random variables L_η and U_η are functions of the output data Y with the goal of achieving

$$\Pr(L_\eta \le \theta \le U_\eta) = \eta,$$

where η is the *nominal coverage probability*.

Extension to higher dimensions has a vector performance measure θ estimated by a random vector point estimator $\widehat{\theta}$, with sampling distribution summarized by a covariance matrix and a confidence region rather than a confidence interval. In this paper, we assume a scalar performance measure, although the main ideas extend directly to higher dimensions.

1.2 Prototypical Example

In the prototypical example, the performance measure is $\theta = \mu$, the mean of a stationary time series with marginal variance σ^2 and lag-h autocorrelations

$$\rho_h = \frac{\mathrm{E}(Y_i\, Y_{i+h}) - \mu^2}{\sigma^2},$$

for $h = \ldots, -2, -1, 0, , 2, \ldots$. For data $Y_1, Y_2, \ldots, Y_n$, the point estimator is the sample average

$$\overline{Y} = \sum_{i=1}^{n} Y_i / n,$$

which is unbiased with variance

$$\mathrm{var}(\overline{Y}) = \frac{\sigma^2}{n}[1 + 2 \sum_{h=1}^{n}(1 - \frac{h}{n})\, \rho_h].$$

The standard error, as always, is

$$\mathrm{ste}(\overline{Y}) = \sqrt{\mathrm{var}(\overline{Y})}.$$

If $\overline{Y}$ is normally distributed with known variance $\mathrm{var}(\overline{Y})$, the confidence interval (L_η, U_η) achieves the nominal coverage probability by choosing

$$L_\eta = \overline{Y} - z_{(1+\eta)/2}\, \mathrm{ste}(\overline{Y})$$

and

$$U_\eta = \overline{Y} + z_{(1+\eta)/2}\, \mathrm{ste}(\overline{Y}),$$

where z_q denotes the qth quantile of the standard normal distribution; i.e.,

$$\Phi(z_q) = \int_{-\infty}^{z_q} \phi(z) dz = q,$$

where $\phi(z) = \mathrm{e}^{-z^2/2}/\sqrt{2\pi}$ is the standard normal density function.

The interval width is

$$W_\eta = U_\eta - L_\eta = 2\, z_{(1+\eta)/2}\, \mathrm{ste}(\overline{Y}),$$

a constant. The coverage indicator C_η is random, with $C_\eta = 1$ if and only if the nominal coverage probability η is less than Δ, where

$$\Delta = 2\, \Phi\left(\frac{\overline{Y} - \mu}{\mathrm{ste}(\overline{Y})}\right) - 1,$$

the *coverage value* that yields the shortest interval that covers the performance measure μ.

1.3 Point Estimators

In general, the performance measure θ can be any property of the (joint) distribution yielding the output data Y and the point estimator can be any function of the observations in the output data set Y. Given a performance measure (such as the standard deviation, the coefficient of variation, a quantile, or a correlation) creating an appropriate point estimator is often as simple as using the sample analog; the choice has little to do with the output-data process. For example, if the performance measure is the marginal variance, $\theta = \sigma^2$, the usual point estimator is the sample variance

$$S^2 = \frac{\sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2}{n-1}.$$

Creating an appropriate CIP, however, involves careful thought about the output-data process.

1.4 Confidence-Interval Procedures

The prototypical example is misleading in that the CIP to compute (L_η, U_η) is so simple that the term *interval estimator* is often used; the word *procedure* is seldom used when L_η and U_η are such simple functions of Y.

Generalizing the prototypical example quickly requires more-complicated functions, for which the word *procedure* seems appropriate. Suppose the simplest case, where the output data are independent and identically distributed (iid) and normal. If the variance $\mathrm{var}(\overline{Y})$ is unknown, z_q is replaced by a Student-t quantile. If the performance measure is the marginal variance, the usual confidence interval is

$$L_\eta = \frac{(n-1)S^2}{\chi^2_{(1-\eta)/2,n-1}}$$

and

$$U_\eta = \frac{(n-1)S^2}{\chi^2_{(1+\eta)/2,n-1}},$$

which is not symmetric around the point estimator.

Even when the performance measure is a mean, creating a CIP becomes yet more difficult for data sets Y that are not iid or normally distributed. The sample size might be a random variable, such as when an experiment simulates n factory shifts and the ith output observation Y_i is the time that the ith part spends waiting for processing. The data might be quite non-normal, as illustrated by the same example. The data might be autocorrelated, with unknown autocorrelations. The data might not be identi-

cally distributed, possibly due to the initial transient in a steady-state simulation experiment.

One approach to confidence intervals is to avoid them. Schmeiser (2001) discusses several well-known disadvantages of confidence intervals. Song and Schmeiser (1994) discuss various alternatives for reporting point-estimator precision. One alternative is to report the estimated standard error, without the additional computation of the confidence interval, which requires specifying a value for η and additional assumptions about the output-data process. Methods for estimating the standard error of the sample mean are fairly well developed (Example 1 of Calvin et al. 1999, Goldsman and Meketon 1986, Song and Schmeiser 1995, Pedrosa 1994, Yeh 2002). These ideas and methods focus on obtaining a small value for $\text{mse}(\widehat{\text{var}}(\overline{Y}))$. Having the single criterion simplifies procedure design, evaluation, and comparison.

Despite their disadvantages, the attraction of confidence intervals is strong, maybe because they are covered in most introductory statistics courses. In the systems-simulation community, for example, a large CIP literature, spanning decades, has been created for the mean μ with time-series data that are assumed to be identically distributed, but non-normal and autocorrelated (Fishman and Yarbarry 1997, Law and Carson 1979, Steiger et al. 2002).

Given that such CIPs exist and continue to be created, and despite our lack of enthusiasm for confidence intervals in practice, we proceed to Section 2 where we argue for a new view about what makes a CIP good.

2 COMPARING CONFIDENCE-INTERVAL PROCEDURES

We argue here that the usual CIP criteria are deficient and argue that a good CIP should be valid (in a certain sense) and should provide intervals appropriate to the observed data rather than to the underlying process.

2.1 The Usual CIP Criteria

Evaluating and comparing CIPs is complicated by multiple criteria. The quality of a CIP is usually viewed as being a function of the joint distribution of the *interval width* $W_\eta = U_\eta - L_\eta$ and of the *coverage indicator* C_η, which is one if $L_\eta \le \theta \le U_\eta$ and otherwise zero. The two most-commonly used CIP criteria are the marginal means: the expected width, $\text{E}(W_\eta)$, and the *actual coverage probability*, $\eta' = \text{E}(C_\eta) = \Pr(L_\eta \le \theta \le U_\eta)$. A third CIP criterion is $\text{var}(W_\eta)$, with a small value indicating interval stability; a high value makes the CIP useless as an indication of point-estimator quality. Schmeiser (1982) also discusses the probability of covering points other than θ, analogous to operating characteristic curves in hypothesis testing, but even the first three CIP criteria lead to inconclusive comparisons.

It might seem that a good CIP, for a given confidence η, would have the smallest possible expected width $\text{E}(W_\eta)$, the smallest possible width variance $\text{var}(W_\eta)$, and the largest possible actual coverage probability η'. The choice $(L_\eta, U_\eta) = (\widehat{\theta}, \widehat{\theta})$ yields $W_\eta = 0$ for the smallest possible mean and variance but with actual coverage probability $\eta' = 0$. The choice $(L_\eta, U_\eta) = (-\infty, \infty)$ yields $W_\eta = \infty$ for the largest possible actual coverage probability $\eta' = 1$ but with infinite interval width.

The fundamental conflict between the criteria complicates CIP design because only truly bad CIPs are dominated in all criteria, especially considering that comparisons must be made for various output-data processes, amount of sampling n, and nominal confidence values η.

Kang and Schmeiser (1990) argue for a graphical approach to comparing CIPs, in the hope that the visualization of the multiple criteria would yield clearer conclusions about CIP quality. Nevertheless, most empirical CIP comparisons are reported with tables containing estimates of expected width, $\text{E}(W_\eta)$, and actual coverage probability, η', for many cases defined by data process Y, amount of sampling n, and nominal coverage probability η.

2.2 Shortcomings of the Usual CIP Criteria

The usual striving for short and stable interval widths and for high coverage probabilities leads to CIP-design conflicts. Worse, the desire for short intervals and for high coverage probabilities is in error. The best value for actual coverage probability, η', is not one, but rather the nominal value, η. The best interval width is not zero, but whatever width correctly indicates the sampling error in the point estimator $\widehat{\theta}$. A short interval is bad if it misleads the practitioner to think that the sampling error is small when it is not. To argue for "shorter is better" is to forget the purpose of the confidence interval. Therefore, a CIP that is "conservative" in the sense of providing actual coverage probability greater than the nominal coverage probability is not as good as a similar CIP that matches the actual to the nominal coverage probability

One-sided confidence intervals provide another argument against using interval width W_η for CIP evaluation, because the intervals $(-\infty, U_\eta)$ and (L_η, ∞) both yield $W_\eta = \infty$.

2.3 What Should a Good CIP Do?

If short interval widths and high actual coverage probabilities are not good, then what is good? First, a CIP needs to be good not only for a particular value of η, but for all values of η between zero and one. Schruben (1980) addresses this point with his definition of coverage function. As in Section 1.2 for Δ in the prototypical CIP, but now for any specified CIP, define Ψ to be the value of η that, for this CIP and this

particular realization, yields the shortest confidence interval that covers the performance measure θ. If the assumptions of the CIP are true, then Ψ is uniformly distributed on $(0, 1)$, and we say that the CIP is *valid*. (The coverage value Ψ is analogous to the p value of hypothesis testing, where if H_0 holds then the p value is $U(0, 1)$.) To graphically illustrate and statistically test the assumptions underlying the CIP, Schruben defines the *coverage function* to be the empirical cumulative distribution function (cdf) of Ψ (for which he uses η^*). Whenever all assumptions underlying the CIP are true, the cdf is a straight line from $(0, 0)$ to $(1, 1)$.

We argue now for a second property of a good CIP. This property is that, for each realization, a good CIP should provide an interval that is appropriate in light of the information that is provided to it; the interval should not necessarily be good in terms of the output-data process, which is unknowable to the CIP. In addition to assumptions, the information is the output-data set Y. If a particular realization of Y is misleading, then a good CIP should provide a misleading interval. To hope for a good interval based on bad sample data is to hope for magic.

If we accept the second property as being desirable, then we need an implementable definition of *appropriate*. We argue that an interval is appropriate if, for this particular realization, its interval matches that of an expert. In this case, the expert is an *ideal* CIP, one for which all statistical assumptions are true and that is allowed to use information unavailable to a real-world CIP. The ideal CIP should perform better than any real-world CIP in the sense that its coverage-function random variable Δ is $U(0, 1)$ and in the sense that its intervals are the length appropriate for indicating the sampling error of $\widehat{\theta}$.

For example, the prototypical example of Section 1.2 could be used as the ideal CIP for evaluating and comparing real-world CIPs that mut estimate $\text{var}(\overline{Y})$. If someone proposed a CIP to compete with the usual Student-t distribution, then that CIP would be better if, and only if, it returned intervals closer to the those of the ideal known-variance CIP than does the Student-t intervals.

An important distinction is that we are not arguing for comparing the distributions of (L_η, U_η) or of (W_η, C_η) from many realizations. Rather, we are arguing for a paired comparison of intervals from two CIPs based on the same output-data set Y. We now combine this non-traditional view with Schruben's coverage-function idea to propose a single CIP criterion.

3 COMPARING TO AN IDEAL CIP

We assume that, for a particular situation, an ideal CIP has been chosen. We wish to measure the difference between its intervals and the intervals from a specified real-world CIP. The comparison is pairwise, in that for each realization of Y and nominal confidence η, both the ideal CIP and the specified CIP return an interval.

Like Schruben (1980), we do not specify a value of η. Rather than comparing the pairs of lower bounds L_η, or pairs of upper bounds U_η, or pairs of widths W_η, or pairs of coverage indicators C_η, all of which depend upon the nominal η, we compare the coverage values. Let Δ denote the coverage value from the ideal CIP; let Ψ denote the coverage value from the specified CIP. Each realization of Y yields one pair (Δ, Ψ), which is not a function of a nominal η value.

In practice, and as illustrated in Subsections 3.2 and 3.3, we advocate comparing a specified CIP to an ideal CIP with a meta experiment of r Monte Carlo practitioners, each of whom uses both CIPs. For $j = 1, 2, \ldots, r$, the jth practitioner observes (δ_j, ψ_j). A scatter plot of the r pairs illustrates the empirical performance of the specified CIP. Points close to the diagonal represent confidence intervals that are good in the sense of mimicking the ideal CIP.

Similarly, two specified CIPs can be compared to each other by comparing each to the ideal CIP. Plotting the $2r$ points (δ_j, ψ_j), using two symbols, visually compares the empirical performance of the two CIPs. We do not plot the r pairs of ψ_js, since we are arguing that a CIP is good (or bad) only with respect to an ideal CIP. If there is no agreed upon common ideal CIP, then we do not provide a method for comparison.

3.1 The Single Criterion

Although we like visual illustration of performance, a single numerical criterion is also desirable. The single criterion should measure how well the specified CIP intervals match the ideal CIP intervals. We propose using $\text{E}((\Delta - \Psi)^2)$, the expected squared error between the paired coverage values. Small values are good. The ideal CIP scores $\text{E}((\Delta - \Delta)^2) = 0$. Other CIPs have positive scores, possibly because Ψ is not $U(0, 1)$ due to violated assumptions. Even if a CIP has no violated assumptions (i.e., is valid), however, it will have a positive score unless $\Psi = \Delta$ for every realization.

Using squared error, rather than an alternative such as absolute deviation, yields the decomposition

$$\text{E}((\Delta - \Psi)^2) = \text{mse}(\Psi, 1/2) + \frac{1}{12} - 2\,\text{cov}(\Delta, \Psi),$$

obtained by subtracting $1/2$ from both Ψ and Δ, expanding the square, using $\text{E}(\Delta) = 1/2$ and $\text{var}(\Delta) = 1/12$, and simplifying. This decomposition yields two terms corresponding to the two properties that we have argued are important. First, a valid method has $\text{mse}(\Psi, 1/2) = \text{var}(\Psi) = 1/12$ because Ψ is $U(0, 1)$; this term does not depend upon the

choice of ideal CIP. Second, a good CIP provides intervals similar to the ideal CIP, as measured by $2\,\mathrm{cov}(\Delta, \Psi)$.

We have tried to obtain a decomposition with two positive terms, much like the mse decomposes into squared bias and variance. We would prefer to have a decomposition with a non-negative term indicating departure from uniformity and a non-negative term indicating coverage-value distances, but we have not found one.

For more insight into the single criterion, we briefly consider eight special cases, all of which are simple and none of which are realistic. Throughout, Δ from the ideal CIP is assumed to be $U(0, 1)$.

- As discussed above, but included here for completeness, if $\Psi = \Delta$, then $\mathrm{E}((\Delta - \Psi)^2) = 0$.
- If a CIP is valid, then Ψ is $U(0, 1)$, which implies that $\mathrm{E}((\Delta - \Psi)^2) = (1 - \mathrm{corr}(\Delta, \Psi))/6$. (If all CIPs were valid, then either $\mathrm{cov}(\Delta, \Psi)$ or $\mathrm{corr}(\Delta, \Psi)$ would be an appropriate single criterion.)
- If a CIP returns intervals that are independent of those of the ideal CIP (despite having common data Y), then $\mathrm{E}((\Delta - \Psi)^2) = \mathrm{mse}(\Psi, 1/2) + 1/12$, in which case the single criterion cannot be less than $1/12$.
- Combining the previous two cases (a valid CIP that is independent of the ideal) yields $\mathrm{E}((\Delta - \Psi)^2) = 1/6$, quite a large value. Despite returning intervals independent of those of the ideal CIP, Schruben's coverage function is a straight line because it considers only the effects of incorrect assumptions.
- A perversely bad case is $\Psi = 1 - \Delta$, which also corresponds to a valid CIP. Now the CIP returns a large interval when the ideal CIP returns a short interval, and vice versa. Here $\mathrm{E}((\Delta - \Psi)^2) = 1/6$.
- A trivially easy CIP to implement is to assume that $\mathrm{ste}(\widehat{\theta}) = 0$. Then the confidence interval has zero width at $\widehat{\theta}$, yielding $\Psi = 1$ except in the (unlikely) realizations in which $\widehat{\theta} = \theta$. Here $\mathrm{E}((\Delta - \Psi)^2) = 1/3$.
- Also easy to implement is the more-general CIP that flips a coin to return a zero-width interval at $\widehat{\theta}$ with probability $1 - \eta$ and $(L_\eta, U_\eta) = (-\infty, \infty)$ with probability η. For any value of η, this CIP provides the nominal coverage probability η. Nevertheless, the distribution of Ψ is not $U(0, 1)$. Rather, $\Pr(\Psi = 0) = \eta$ and $\Pr(\Psi = 1) = 1 - \eta$. For every value of $\eta \in [0, 1]$, the single criterion is $\mathrm{E}((\Delta - \Psi)^2) = 1/3$.
- The worst case in terms of maximizing the single criterion is $\Psi = 1$ whenever $\Delta < 1/2$ and $\Psi = 0$ whenever $\Delta \geq 1/2$. Here $\mathrm{E}((\Delta - \Psi)^2) = 7/12$.

In summary, $0 \leq \mathrm{E}((\Delta - \Psi)^2) \leq 7/12$, with every reasonable CIP satisfying $0 \leq \mathrm{E}((\Delta - \Psi)^2) \leq 1/3$. Every valid CIP satisfies $0 \leq \mathrm{E}((\Delta - \Psi)^2) \leq 1/6$. Every competitive CIP will have a single-criterion value substantially less than $1/6$.

3.2 Example: Student-t CIPs

To gain more insight into the single criterion, we consider the performance of Student-t confidence intervals for the mean μ of iid normal data. The ideal CIP is the prototypical example in Section 1, with all zero autocorrelations. That is, the ideal CIP produces the confidence interval centered at $\overline{Y}$ with half width $z_{(1+\eta)/2}\,\sigma/\sqrt{n}$. The Student-t CIP produces the confidence interval centered at $\overline{Y}$ with half width $t_{(1+\eta)/2,n-1}\,S/\sqrt{n}$, where $t_{q,\nu}$ is the qth quantile of the Student-t distribution with ν degrees of freedom.

The coverage-function values for the Student-t CIP are

$$\Psi = 2\,F_{n-1}\left(\frac{|\overline{Y} - \mu|}{S/\sqrt{n}}\right) - 1,$$

where F_ν is the Student-t cumulative distribution function (cdf) with ν degrees of freedom. The analogous formula for the ideal CIP's Δ is in Subsection 1.2.

Consider samples of size $n = 10$. Figure 1 is a scatter plot of 100 Monte Carlo observations (δ_j, ψ_j), representing the simulated experience of $r = 100$ practitioners. Because the ideal CIP and the Student-t CIP are valid, the marginal distributions of both Δ and Ψ are $U(0, 1)$, which is consistent with the $r = 100$ points. The points cluster around the diagonal, as is required of a CIP that mimics the ideal CIP.

Because Student-t CIPs are valid, $\mathrm{E}((\Delta - \Psi)^2) = (1/6)(1 - \mathrm{corr}(\Delta, \Psi))$. In this example, the empirical covariance is $\widehat{\mathrm{cov}}(\Delta, \Psi) \approx 0.080$ and the corresponding correlation is $\widehat{\mathrm{corr}}(\Delta, \Psi) = 12\,\widehat{\mathrm{cov}}(\Delta, \Psi) \approx 0.960$. The corresponding empirical value of the single criterion is $\widehat{\mathrm{E}}((\Delta - \Psi)^2) \approx 0.0069$. (These empirical results are based on $r = 20000$ simulated practitioners; all digits shown are correct to within one unit.)

How does this value of the single criterion relate to more-traditional measures of CIP performance? Although $\widehat{\mathrm{E}}((\Delta - \Psi)^2) \approx 0.0069$ seems close to zero, this CIP is far from the ideal. From Table 1 of Schmeiser (1982), for any coverage probability η, the confidence-interval widths W_η have coefficient of variation 0.24, whereas the ideal has zero coefficient of variation. For $\eta = 0.95$, the expected interval width is about 4.4 standard errors, whereas the ideal CIP has expected interval width of about 3.9 standard errors.

Now consider the effect of taking batches of size $m = 2$. The $n = 10$ observations form five batch means, with four degrees of freedom. Figure 2 compares the Student-t CIPs for ten batches of size $m = 1$ and five batches of size

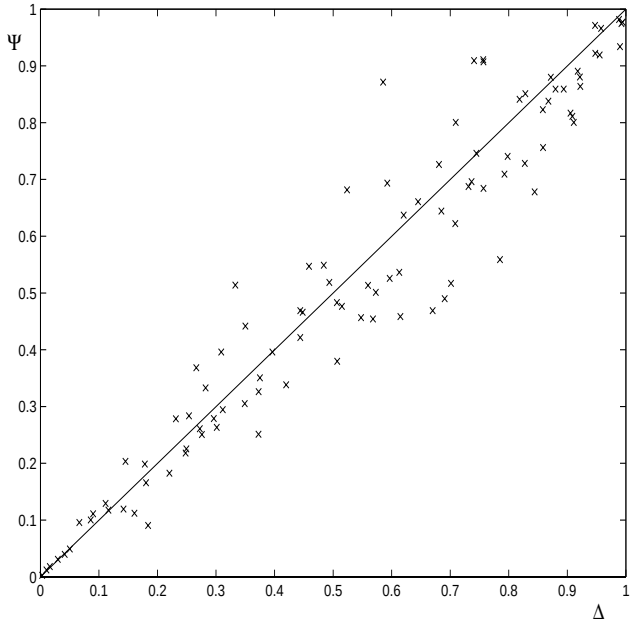

Figure 1: For $n = 10$ and IID Normal Data, Student-t Coverage Values Ψ Plotted against Normal Coverage Values Δ

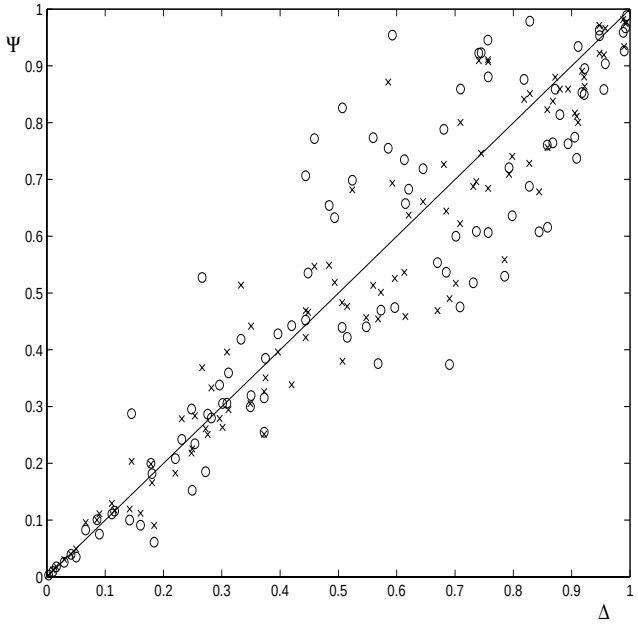

Figure 2: For $n = 10$ and IID Normal Data, Student-t Coverage Values Plotted against Normal Coverage Values, for Batch Sizes $m = 1$ ($\times$) and $m = 2$ ($\circ$)

$m = 2$. The $r = 100$ points (shown as $\times$) from Figure 1 for batches of size one are shown with the corresponding $r = 100$ points (shown as $\circ$) for batches of size $m = 2$, based on the same Monte Carlo realizations. Because of the common Δ values, the points appear as vertical pairs. The points corresponding to batches of size $m = 2$ cluster less around the diagonal, as is to be expected, with empirical covariance $\widehat{\text{cov}}(\Delta, \Psi) \approx 0.076$, empirical correlation is $\widehat{\text{corr}}(\Delta, \Psi) \approx 0.908$, and empirical value of the single criterion $\widehat{\text{E}}((\Delta - \Psi)^2) \approx 0.0154$.

3.3 Example: Invalid-normal CIPs

Now consider the (invalid) CIP for the mean μ of iid normal data that uses an estimated variance S^2, but assumes that $\sigma^2 = S^2$ by using half width $z_{(1+\eta)/2} S/\sqrt{n}$ rather than the valid half width from Subsection 3.2. Despite this CIP being invalid, using the traditional criteria it is better than the Student-t CIP in that its widths W_η are shorter and have smaller variance. It is worse than the Student-t CIP in that its actual coverage probabilities are less than nominal.

This invalid-normal CIP is compared to the Student-t CIP for $n = 10$ observations with batches of size $m = 2$ in Figure 3, using the same $r = 100$ realizations as Figures 1 and 2. The $r = 100$ Student-t points from Figure 2 (still shown as $\circ$) are shown with the corresponding $r = 100$ points (shown as $+$) for the invalid-normal CIP. Even with only four degrees of freedom, the difference between the two CIPs is not immediately visually obvious. Closer examination shows that for each of the $r = 100$ realizations, the invalid-normal $+$ lies above the Student-t $\circ$. The shorter widths yield larger Ψ coverage values, which is good when the data yield a variance estimate S^2 that is substantially larger than the true variance σ^2. On the whole, however, the inappropriately short intervals perform worse.

Based on $r = 20000$ Monte Carlo practitioners, with negligible standard errors, the $(n = 10, m = 2)$ invalid-normal empirical covariance is $\widehat{\text{cov}}(\Delta, \Psi) \approx 0.081$, the empirical correlation is $\widehat{\text{corr}}(\Delta, \Psi) \approx 0.909$, and the empirical single-criterion value is $\widehat{\text{E}}((\Delta - \Psi)^2) \approx 0.0182$.

The invalid-normal single criterion is about 18% larger than for the Student-t single criterion. It is larger because the underlying assumptions are violated, resulting in the distribution of Ψ being non-uniform. In particular, although the empirical mean $\widehat{\text{E}}(\Psi)$ is one half, the empirical value of the invalid-normal variance is $\widehat{\text{var}}(\Psi) \approx 0.096$, larger than $1/12$.

Because Student-t confidence intervals have substantial foundation in both theory and practice, it would be surprising if any alternative CIP were better. Therefore, if an alternative CIP, such as the invalid-normal CIP considered in this subsection, produced a lower single criterion, that would be a strong argument that our single criterion is not adequate for evaluating CIPs.

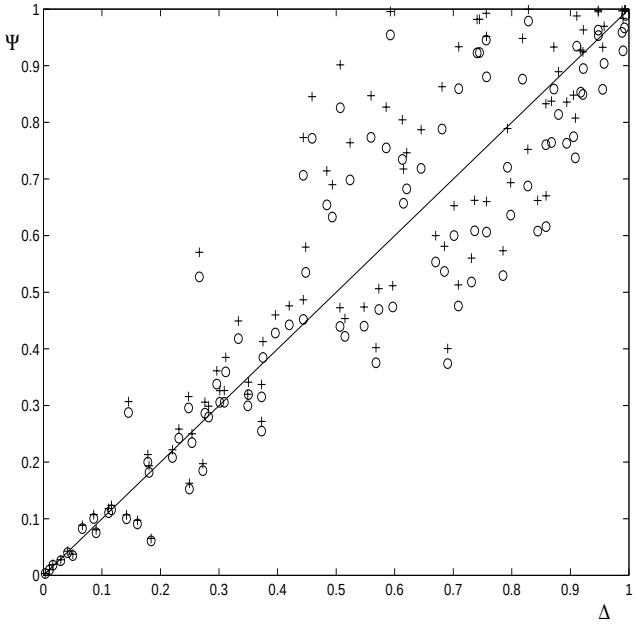

Figure 3: For $n = 10$, $m = 2$ and IID Normal Data, Student-t Coverage Pairs ($\circ$) Compared to Invalid-Normal Coverage Pairs ($+$)

4 ADDITIONAL THOUGHTS

Throughout this paper we have assumed that the purpose of a CIP is to indicate the precision of a point estimator $\widehat{\theta}$. Competing CIPs therefore are assumed to be based on a common data set Y of size n, from which a common point estimator $\widehat{\theta}$ is computed. The community of researchers with competing CIPs is expected to agree on the ideal CIP to be used for the single criterion.

Another purpose of a CIP is to provide a confidence interval as the primary product, rather than as an auxiliary to the point estimator. Here competing CIPs use a common data set Y of size n, but the point estimator might differ (or, in rare cases, such as quantile estimation, the CIP might not be a function of a point estimator). Here, again, the community of researchers with competing CIPS is expected to agree on the ideal CIP to be used for the single criterion.

Some CIPs determine the sample size n as part of the procedure. Because competing CIPs then have unequal sample sizes, they cannot be compared using exactly the same data set Y. How to choose an ideal CIP is less obvious in this setting.

Other than suggesting the prototypical example of Subsection 1.2 for the mean, we have said nothing about the nature of the ideal CIP except the assumption that it is usually valid and is always the standard of comparison. We have not been able to define a procedure for automatically determining an ideal CIP. Our impression is that it usually will arise from assuming knowledge of unknowable values about either the data process or the point-estimator's sampling distribution.

Another use of a CIP is as a means to evaluate procedures for other statistical problems. For example, Wilson and Pritsker (1978) base their start-up criterion on confidence-interval properties and Wilson and Pritsker (1984) compare variance-reduction techniques using CIP properties (as well as on the standard-error of the resulting point estimator). In these cases there is only one CIP, to which different data sets Y are input. Our single criterion has, we think, nothing to say about this use of CIPs.

Finally, we mention that maybe the single criterion can be used to evaluate and compare individual CIPs across sample sizes and families of CIPs that are parameterized by degrees of freedom. For example, for Student-t CIPs applied to iid normal data, the single criterion decreases to zero as sample size (and degrees of freedom) go to infinity. In Subsection 3.2, for both $\nu = 9$ and $\nu = 4$, the empirical value of the product of ν and the single criterion $\widehat{\mathrm{E}}((\Delta - \Psi)^2)$ is 0.062, which suggests $O(\nu^{-1})$. Whenever this rate holds, the product of sample size and the single criterion (or of degrees of freedom and the single criterion) would be a single criterion for the entire family of CIPs.

APPENDIX: WHY NOT USE COVARIANCE?

We choose $\mathrm{E}((\Delta - \Psi)^2)$ as our single criterion after also considering $\mathrm{cov}(\Delta, \Psi)$, for which large values would indicate good CIP performance. If, as is ideal, $\Psi = \Delta$, then $\mathrm{cov}(\Delta, \Psi) = \mathrm{var}(\Delta) = 1/12$ because Δ is $U(0, 1)$. For a CIP with Ψ independent of Δ, the covariance is zero. All reasonable CIPs have positive values.

The covariance is not a good single criterion, however, because it is not maximized by $\mathrm{cov}(\Delta, \Delta) = 1/12$. A counter example is to define $\Psi = \Delta^b$, which corresponds to the ideal CIP for $b = 1$. Because Δ is $U(0, 1)$,

$$\begin{aligned} \mathrm{cov}(\Delta, \Psi) &= \mathrm{E}(\Psi\Delta) - \frac{\mathrm{E}(\Psi)}{2} \\ &= \mathrm{E}(\Delta^{b+1}) - \frac{\mathrm{E}(\Delta^b)}{2} \\ &= \frac{1}{b+2} - \frac{1}{2(b+1)}. \end{aligned}$$

Rather than being maximized at $b = 1$, the maximum covariance value occurs at $b = \sqrt{2}$ and yields $\mathrm{cov}(\Delta, \Psi) \approx 0.086$, which is larger than $\mathrm{cov}(\Delta, \Delta) = 1/12$.

The deficiency in using covariance is that it does not include our desire for Ψ to be $U(0, 1)$, as is required of every valid CIP. At the optimal power $b = \sqrt{2}$, the mean is $\mathrm{E}(\Psi) \approx 0.414$, not $\mathrm{E}(\Psi) = 0.5$.

As another example of the deficiency of using covariance, consider the invalid-normal CIP from Subsection 3.3.

There, the invalid-normal covariance is 0.081, which is larger than the Student-t covariance of 0.076 in Subsection 3.2.

Another brief thought was to use $\text{corr}(\Delta, \Psi)$. A quick counter example is to consider $\Psi = b\Delta$, which yields $\text{corr}(\Delta, \Psi) = 1$ for every $b \in (0, 1]$. Because $b = 1$ is not the unique optimal coefficient, correlation is not an adequate criterion for evaluating CIPs. Also, both the invalid-normal CIP and the Student-t CIP yielded a correlation of about 0.91.

REFERENCES

Calvin, J. M., P. W. Glynn and M. K. Nakayama. 1999. On the small-sample optimality of multiple-regeneration estimators. In *Proceedings of the 1999 Winter Simulation Conference*, ed. P. A. Farrington, H. B. Nembhard, D. T. Sturrock, and G. W. Evans, 655–661, Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Fishman, G. S. and L. S. Yarberry. 1997. An implementation of the batch means method. *INFORMS Journal on Computing* 9 (3): 296–310.

Goldsman, D. M. and M. S. Meketon. 1986. A comparison of several variance estimators. Technical Report J–85–12, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia.

Kang, K. and B. Schmeiser. 1990. Graphical methods for evaluating and comparing confidence-interval procedures. *Operations Research Letters* 38: 546–553.

Law, A. M. and J. S. Carson. 1979. A sequential procedure for determining the length of a steady-state simulation. *Operations Research* 27 (5): 1011–1025.

Pedrosa, A. 1994. *Automatic Batching in Simulation Output Analysis*, Doctoral Dissertation, School of Industrial Engineering, Purdue University, West Lafayette, Indiana.

Schmeiser, B. W. 1982. Batch size effects in the analysis of simulation output. *Operations Research* 30: 556–568.

Schmeiser, B. W. 2001. Some myths and common errors in simulation experiments. In *Proceedings of the 2001 Winter Simulation Conference*, ed. B.A. Peters, J.S. Smith, D.J. Medeiros and M.W. Rohrer, 39–46. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Schruben, L. W. 1980. A coverage function for interval estimators of simulation response. *Management Science* 26: 18–27.

Song, W. T. and B. W. Schmeiser. 1994. Reporting the precision of simulation experiments. In *New Directions in Simulation for Manufacturing and Communications*, ed. S. Morito, H. Sakasegawa, K. Yoneda, M. Fushimi and K. Nakano, 402–407, Operations Research Society of Japan.

Song, W. T. and B. W. Schmeiser. 1995. Optimal mean-squared-error batch sizes. *Management Science* 41: 110–123.

Steiger, N. M., E. K. Lada, J. R. Wilson, C. Alexopoulos, D. Goldsman and F. Zouaoui. 2002. In *Proceedings of the 2002 Winter Simulation Conference*, ed. E. Yücesan, C.-H. Chen, J. L. Snowdon, and J. M. Charnes, this volume, Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Wilson, J. R. and A. A. B. Pritsker. 1978. Evaluation of startup policies in simulation experiments. *Simulation* 31 (3): 79–89.

Wilson, J. R. and A. A. B. Pritsker. 1984. Experimental evaluation of variance reduction techniques for queueing simulation using generalized concomitant variables. *Management Science* 30 (12): 1459–1472.

Yeh, Y. 2002. *Steady-State Simulation Output Analysis: MSE-Optimal Dynamic Batch Means with Parsimonious Storage*, Doctoral Dissertation, School of Industrial Engineering, Purdue University, West Lafayette, Indiana.

AUTHOR BIOGRAPHIES

BRUCE SCHMEISER is a professor in the School of Industrial Engineering at Purdue University. His interests lie in applied operations research, with emphasis in stochastic models, especially the probabilistic and statistical aspects of stochastic simulation. Since 1976, he has been an active participant in the Winter Simulation Conference, including being Program Chair in 1983 and chairing the Board of Directors during 1988–1990.

YINGCHIEH YEH is an assistant professor in the Department of Business Administration at Yuan Ze University. In 2002, he received a Ph.D. degree from the School of Industrial Engineering at Purdue University. His primary research interests are the probabilistic and statistical aspects of stochastic simulation, especially simulation output analysis.