

A SIMPLE MONTE CARLO METHODOLOGY TO CALCULATE GENERALIZED APPROXIMATE CONFIDENCE INTERVALS

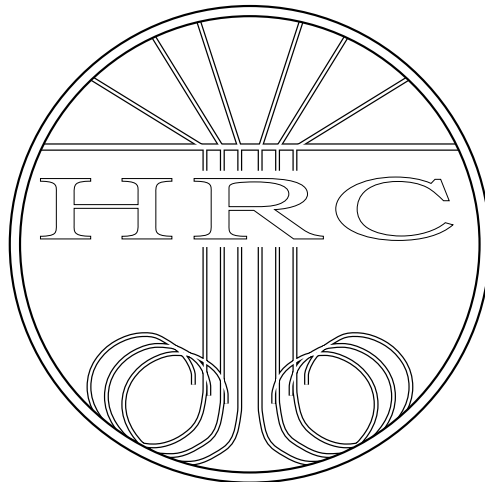
Demetris Koutsoyiannis

Visiting scholar, Hydrologic Research Center
and Department of Water Resources, National Technical University of Athens, Greece

and

Stefanos Kozanis

Department of Water Resources, National Technical University of Athens, Greece



HRC Technical Note No. 25

Hydrologic Research Center
12780 High Bluff Drive, Suite 250
San Diego, CA 92130

October 2005

HRC Technical Note No. 25

A SIMPLE MONTE CARLO METHODOLOGY TO CALCULATE GENERALIZED APPROXIMATE CONFIDENCE INTERVALS

Demetris Koutsoyiannis

Visiting scholar, Hydrologic Research Center

and Department of Water Resources, National Technical University of Athens, Greece

and

Stefanos Kozanis

Department of Water Resources, National Technical University of Athens, Greece

Contents

Acknowledgements

Executive Summary

1. Definitions and introductory comments
2. Confidence limits for one unknown distributional parameter
3. Confidence limits for many unknown distributional parameters
4. Method validation for one unknown distributional parameter
5. Method validation for many unknown distributional parameters

References

October 2005

ACKNOWLEDGEMENTS

This work by Professor Demetris Koutsoyiannis was authored while at the Hydrologic Research Center as a Visiting Scholar, along with Stefanos Kozanis from the Department of Water Resources, National Technical University of Athens, Greece

Executive Summary

Determination of confidence limits of distributional parameters (either marginal or dependence) and derivative quantities (e.g. distribution quantiles) is crucial for estimation of uncertainty and risk. Analytical determination is possible in few cases only. Monte Carlo simulation is a numerical method with the potential to determine confidence limits without restrictions. However, even Monte Carlo simulation is not as direct, general and easily applicable as it may seem. Existing direct solutions are exact only in limited cases whereas if applied in other cases may result in significant errors. Extending and generalizing existing solutions, a simple Monte Carlo simulation technique is studied that can determine good approximations of confidence limits in a general setting. The proposed method is partly heuristic and simultaneously so general that needs no assumptions about the statistical behavior of the statistics under study, i.e. it can perform for any distribution with any number of parameters, and for any distributional or derivative parameter. Only the theoretical probabilistic model is needed and all other calculations are done by a number of Monte Carlo simulations without additional assumptions. Some tests of the method in cases with analytically determined confidence limits indicate impressively good performance. Even though the method has been tested for independent sequences of random variables (random samples) its general formulation allows direct application in stochastic processes with any dependence structure, provided that a stochastic generator of the process of interest exists.

TABLE OF CONTENTS

Acknowledgements	i
Executive Summary	ii
1. Definitions and Introductory Comments	1
2. Confidence Limits for One Unknown Distributional Parameter	3
3. Confidence Limits for Many Unknown Distributional Parameters	5
4. Method Validation for One Unknown Distributional Parameter	7
5. Method Validation for Many Unknown Distributional Parameters	11
References	13

1. Definitions and Introductory Comments

Let β be an (unknown) true (population) parameter whose confidence limits are sought. Let B and b denote the estimator and the point estimate (PE), respectively, of β , based on a random sample of size n ; it is reminded that B is a random variable and b is a numerical value. If β were known then it would be generally easy to infer the probability distribution of the estimator B and, consequently, to find to numbers λ and v , so that that for a certain $\alpha < 1$,

$$P\{\lambda < B < v\} = \alpha \quad (1)$$

where P denotes probability. The determination of λ and v is called prediction and the number α is the confidence coefficient of the prediction (Papoulis, 1990, p. 241). The numbers λ and v are the lower and upper prediction limits, respectively. Typically, they are estimated such as

$$P\{B < \lambda\} = (1 - \alpha)/2, \quad P\{B < v\} = (1 + \alpha)/2 \quad (2)$$

If analytical determination of the prediction limits is difficult or intractable, Monte Carlo simulation can be readily compute them numerically; in the latter case these will be referred to as Monte Carlo prediction limits (MCPL). Although prediction limits provide some indication of uncertainty, useful in sensitivity studies, strictly they are not the appropriate measure of parameter uncertainty. Instead, confidence limits should be used, which are defined in terms of some random variables L and U that embrace the unknown parameter β with a given probability or confidence coefficient α , i.e.,

$$P\{L < \beta < U\} = \alpha \quad (3)$$

The numerical values l and u of the random variables L and U , respectively, calculated from the available sample, are the required confidence limits for confidence coefficient α .

The determination of L and U , and the calculation of l and u , are difficult tasks except for certain cases that can be treated analytically (e.g. for the mean and variance of a normal variable based on an independent sample) and unless very sophisticated numerical methods are used (e.g. based on Markov chain Monte Carlo methods), which however are out of the scope of this study. Nevertheless, Ripley (1987) provides a simple approximate way to transform MCPL into Monte Carlo confidence limits (MCCL), which is the reflection of the prediction limits about the estimate of β , i.e.,

$$l = 2b - v, u = 2b - \lambda \quad (4)$$

This method however may not give good approximations for models with more than one parameter (see Figure 5 and its discussion). Also, the method does not work in cases that the estimator B has asymmetric distribution (see Figure 6 and its discussion). For such cases, Ripley (1987) gives a second alternative, i.e. a logarithmic reflection,

$$l = b^2 / v, u = b^2 / \lambda \quad (5)$$

This is not free of the weakness of the first alternative; in addition, the selection of one of the two methods is not easy, unless there is concrete theoretical knowledge of the model behavior, which cannot be the case even in models with moderate complexity.

The methods described by equations (4) and (5) will be referred to as Methods 1 and 2, respectively and the resulting confidence limits as MCCL1 and MCCL2. The method proposed here, which will be referred to as Method 3, is a unification, generalization and extension of Methods 1 and 2 done two steps: In the first step both Methods 1 and 2 are unified and generalized in problems involving a single parameter and in the second one the method is extended to incorporate more than one parameter.

As it will be demonstrated in the following sections, Method 3 is partly heuristic and simultaneously so general that needs no assumptions about the statistical behavior of the statistics under study. Only the theoretical model of the process is needed and all other calculations are done by a number of Monte Carlo simulations free of additional assumptions. Some tests of the method in cases with analytically determined confidence limits indicate impressive performance.

2. Confidence limits for one unknown distributional parameter

We consider a random variable X whose distribution depends on a single parameter β , which is unknown. If we assume a certain value of β , then using Monte Carlo simulation we can readily calculate an estimate b of β for a sample of size n and by repeating the simulation for a large number of times we can estimate the average $\eta = E[B]$ (which may not be identical to β if the estimator B is biased) as well as the prediction limits λ and v for a certain confidence coefficient α . Theoretically, by doing the same procedure for a number of different values of β , we can construct the curves $\eta(\beta)$, $\lambda(\beta)$ and $v(\beta)$ shown in Figure 1. The prediction limits for $\beta = b$ are marked in this figure as points A and A' . The confidence limits of β for the given estimate b are given by points C and C' , which are the intersections of the horizontal line at abscissa b with the curves $v(\beta)$ and $\lambda(\beta)$, respectively.

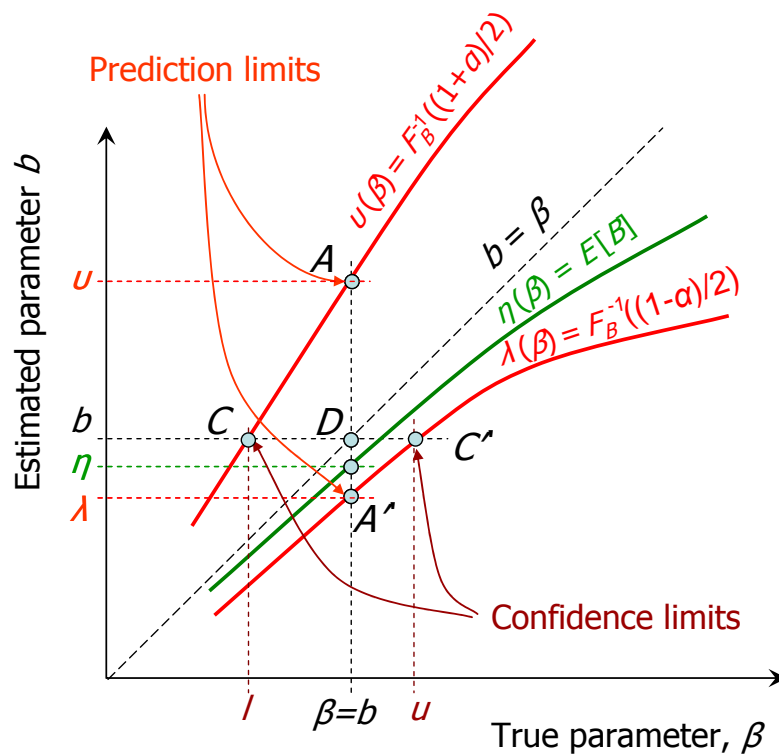


Figure 1 Explanation sketch for the determination of confidence limits for one unknown distributional parameter.

Now from the triangle ADC we obtain

$$\frac{v - b}{b - l} = \frac{AD}{DC} \approx \frac{dv}{d\beta} \quad (6)$$

where the derivative is meant at point A (i.e. for $\beta = b$). Solving for the lower confidence limit l we find

$$l = b + \frac{b - v}{dv/d\beta} \quad (7)$$

and in a similar manner, the upper confidence limit u will be

$$u = b + \frac{b - \lambda}{d\lambda/d\beta} \quad (8)$$

In this way, to estimate the confidence limits l and u we need to estimate the prediction limits λ and v at $\beta = b$ as well as their derivatives at the same value of β . To estimate the latter an additional ensemble of Monte Carlo simulations is required at a point near to $\beta = b$. Thus the method is computationally efficient as only two Monte Carlo simulation ensembles are required.

It can be readily verified that when $dv/d\beta = d\lambda/d\beta = 1$, this method yields Method 1 as a special case, whereas when $dv/d\beta = v/\beta$ and $d\lambda/d\beta = \lambda/\beta$ it yields Method 2.

3. Confidence Limits for Many Unknown Distributional Parameters

Let us now assume that the model under study involves k parameters (referring to the marginal distribution or the dependence structure) that form the vector $\boldsymbol{\theta} = [\theta_1, \theta_2, \dots, \theta_k]'$, whose estimator is the vector $\mathbf{T} = [T_1, T_2, \dots, T_k]'$ (where the prime denotes the transpose of a vector or matrix). We wish to determine confidence limits for a single parameter β dependent on θ , i.e. $\beta = h(\boldsymbol{\theta})$, whose estimator is $B = h(\mathbf{T})$ and its estimate is $b = h(t)$. For a given $\boldsymbol{\theta}$, it is easy to compute by Monte Carlo simulation the prediction limits λ and v of b as in the single parameter case. However, to extend Method 3, described by equations (7) and (8) in the multiple parameter case, the derivatives $d\lambda/d\beta$ and $dv/d\beta$ should be evaluated at appropriate directions \mathbf{d}_λ and \mathbf{d}_v .

Let $\boldsymbol{\gamma} = [\lambda, \beta, v]'$ the vector consisting of β and its prediction limits (λ, v) for confidence α and let $\text{Var}[\mathbf{T}] = \text{diag}(\text{Var}[T_1], \dots, \text{Var}[T_k])$. The latter can be easily computed during the same Monte Carlo simulation that is performed to compute $\boldsymbol{\gamma}$. It is reasonable to assume that \mathbf{d}_λ and \mathbf{d}_v will depend on $\text{Var}[\mathbf{T}]$ as well as of the matrix of derivatives of $\boldsymbol{\gamma}$,

$$\frac{d\boldsymbol{\gamma}}{d\boldsymbol{\theta}} = \begin{bmatrix} \frac{d\lambda}{d\boldsymbol{\theta}} \\ \frac{d\beta}{d\boldsymbol{\theta}} \\ \frac{dv}{d\boldsymbol{\theta}} \end{bmatrix} = \begin{bmatrix} \frac{\partial\lambda}{\partial\theta_1} & \frac{\partial\lambda}{\partial\theta_2} & \dots & \frac{\partial\lambda}{\partial\theta_k} \\ \frac{\partial\beta}{\partial\theta_1} & \frac{\partial\beta}{\partial\theta_2} & \dots & \frac{\partial\beta}{\partial\theta_k} \\ \frac{\partial v}{\partial\theta_1} & \frac{\partial v}{\partial\theta_2} & \dots & \frac{\partial v}{\partial\theta_k} \end{bmatrix} \quad (9)$$

Heuristically, we can assume a simple relation of the form

$$\mathbf{d}_\lambda = \text{Var}[\mathbf{T}] \left(\frac{d\boldsymbol{\gamma}}{d\boldsymbol{\theta}} \right)' \mathbf{e}_\lambda \quad (10)$$

where \mathbf{e}_λ is a size 3 vector of constants needed to transform the matrix product of the first two terms of the right hand side into a vector. The elements of this vector could be thought of as weights corresponding to each of the derivatives of the three elements of $\boldsymbol{\gamma} = [\lambda, \beta, v]'$. The simplest choice is to assume equal weights, i.e. $\mathbf{e}_\lambda = [1, 1, 1]'$. However, numerical investigations showed that the choice $\mathbf{e}_\lambda = [0, 1, 1]'$ yields better approximations. The derivative of λ and β with respect to $\boldsymbol{\theta}$ on direction \mathbf{d}_λ will then be

$$\left(\frac{d\lambda}{d\boldsymbol{\theta}} \right) \mathbf{d}_\lambda = \left(\frac{d\lambda}{d\boldsymbol{\theta}} \right) \text{Var}[\mathbf{T}] \left(\frac{d\boldsymbol{\gamma}}{d\boldsymbol{\theta}} \right)' \mathbf{e}_\lambda, \quad \left(\frac{d\beta}{d\boldsymbol{\theta}} \right) \mathbf{d}_\lambda = \left(\frac{d\beta}{d\boldsymbol{\theta}} \right) \text{Var}[\mathbf{T}] \left(\frac{d\boldsymbol{\gamma}}{d\boldsymbol{\theta}} \right)' \mathbf{e}_\lambda \quad (11)$$

and are both scalars, so by taking their ratio we can calculate $d\lambda/d\beta$. By symmetry, similar relationships can be written for v and \mathbf{d}_v with $\mathbf{e}_v = [1, 1, 0]'$. The two groups of relationships can be unified in terms of the 3×3 matrix \mathbf{q} defined as

$$\mathbf{q} := \frac{d\boldsymbol{\gamma}}{d\boldsymbol{\theta}} \text{Var}[\mathbf{T}] \left(\frac{d\boldsymbol{\gamma}}{d\boldsymbol{\theta}} \right)' \quad (12)$$

It can then be easily shown that on the directions \mathbf{d}_λ and \mathbf{d}_v ,

$$\frac{d\lambda}{d\beta} = \frac{q_{12} + q_{13}}{q_{22} + q_{23}}, \quad \frac{dv}{d\beta} = \frac{q_{31} + q_{32}}{q_{21} + q_{22}} \quad (13)$$

It can be observed that the evaluation of all terms used in the method requires $1 + k$ simulation ensembles, one for the calculation of γ and $\text{Var}[\mathbf{T}]$ at point $\boldsymbol{\theta} = \mathbf{t}$, and k for the calculation of the derivatives and eventually of the matrix \mathbf{q} .

4. Method Validation for One Unknown Distributional Parameter

As a first example, we consider a random variable x exponentially distributed with density

$$f(x) = e^{-x/\theta} / \theta, x > 0 \quad (14)$$

where θ is the single parameter of the distribution. It is known that the mean $\mu \equiv \beta$ as well as the standard deviation σ are both equal to θ . We seek confidence limits for the parameter $\mu \equiv \beta \equiv \theta$, assuming that its estimator is the mean of a sample x_1, \dots, x_n , i.e.,

$$\bar{x} = \frac{x_1 + \dots + x_n}{n} \quad (15)$$

The classic theoretical confidence limits given in the statistical literature are (Papoulis, 1990, p. 281):

$$l, u = \frac{\bar{x}}{1 \pm z_{(1+\alpha)/2} \sqrt{n}} \quad (16)$$

where z_p denotes the p -quantile of the standard normal distribution. These, however, are obtained assuming normal distribution of the mean, which is not the case for small n ; the convergence of the distribution of \bar{x} to the normal distribution is very slow due to the skewness of the exponential distribution. However, the exact distribution of \bar{x} is known to be gamma with shape parameter n and scale parameter θ/n (symbolically $F_G(\bar{x}; n, \theta/n)$). From this it can be deduced that the exact confidence limits l, u of the mean can be determined by solving numerically the equations

$$F_G(\bar{x}, n, l/n) = (1 + \alpha)/2, \quad F_G(\bar{x}, n, u/n) = (1 - \alpha)/2 \quad (17)$$

For demonstration, the theoretical confidence limits given by both (16) and (17) assuming $\bar{x} = 1$ and $n = 10$ are plotted against α ranging from 0.80 to 0.99 in Figure 2. Clearly, the classic statistical confidence limits are very poor for such a low sample size. At the same figure, the MCPL as well as MCCL for all three methods 1, 2 and 3 are also plotted. It can be observed that the curves of both MCCL2 and MCCL3 are indistinguishable from the exact curves and outperform significantly the classic theoretical limits. In contrast, MCCL1 are poor, worse than the classic theoretical limits.

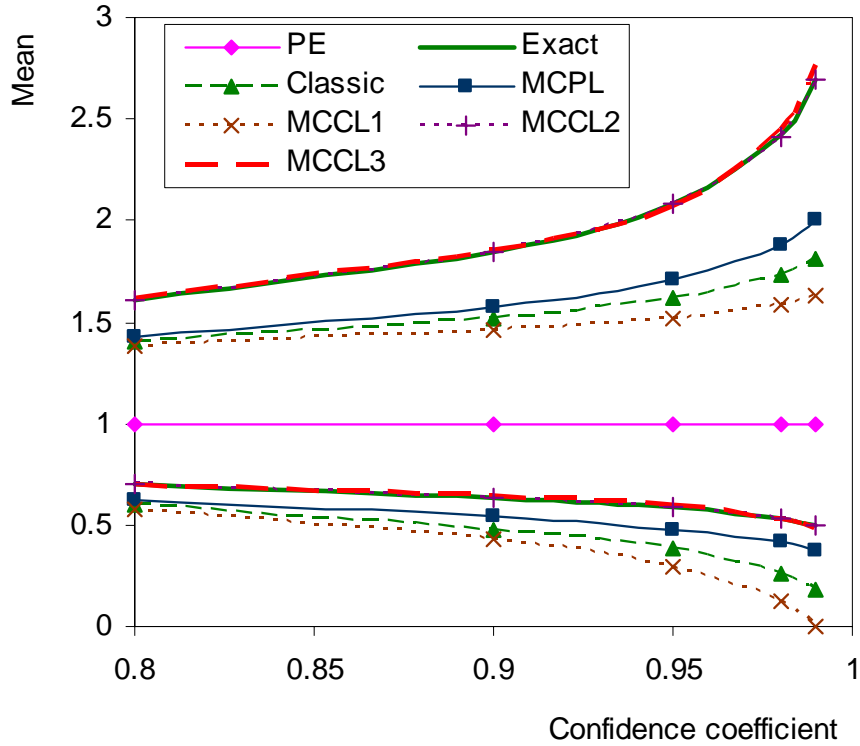


Figure 2 Point estimate, prediction limits and confidence limits for the mean of an exponentially distributed variable against the confidence coefficient α , as estimated by various methods assuming $\bar{x} = 1$ and $n = 10$. MCC2 and MCC3 are indistinguishable from the exact theoretical confidence limits.

The second example is related to the coefficient limits of a probability $p \equiv \beta$ of a certain event A , which is estimated by the mean \bar{x} of a binary random process X , assuming $X = 1$ for the occurrence of this event and $X = 0$ otherwise. Obviously $b \equiv \bar{x} = n_s/n$, where n_s is the number of the occurrences of A in n total trials. For relatively large n , the theoretical confidence limits (TCL) are given as the two roots of equation

$$(p - \bar{x})^2 = \frac{z_{(1+\alpha)/2}^2}{n} p(1-p) \quad (18)$$

solved for p (Papoulis, 1990, p. 284). Figure 3 provides a demonstration of the resulting MCCL for all three methods in comparison to TCL. Here the performance of Method 1 is rather satisfactory where that of Method 2 is poor. Method 3 outperforms the two yielding confidence limits very close to the exact ones (TCL).

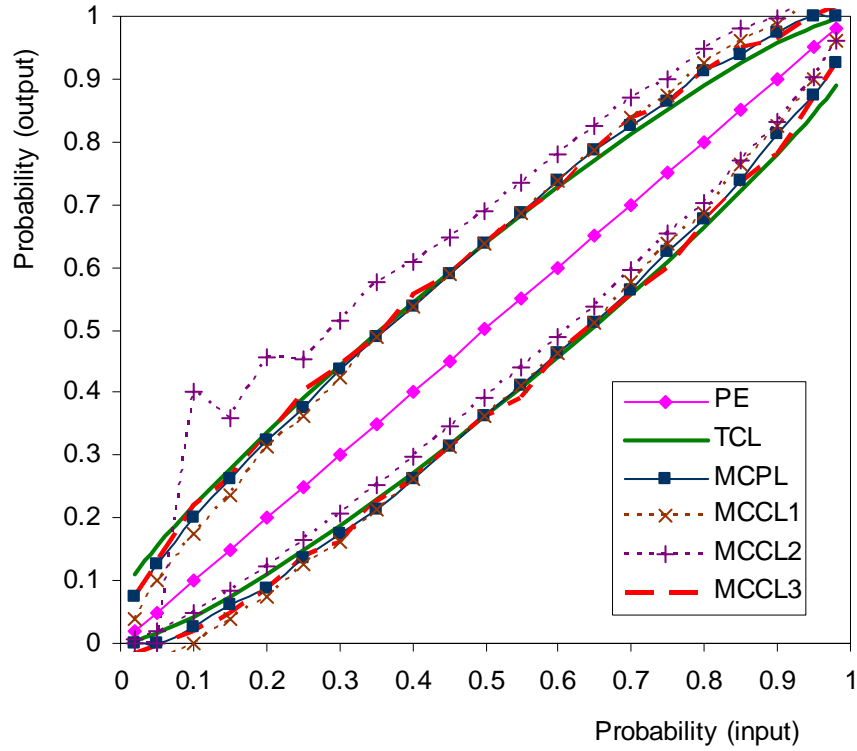


Figure 3 Point estimate, prediction limits and confidence limits for the probability of an event assuming confidence coefficient $\alpha = 0.99$ and $n = 80$. For MCPL input probability is the theoretical probability p and output probabilities are the prediction limits λ and v ; for all other cases input probability is the estimate \bar{x} and output probabilities are the estimates of p (PE) and the confidence limits l and u (TCL, MCCL).

The third example deals with the most common case of the statistical literature, which is the mean $\mu \equiv \beta$ of a normally distributed variable whose variance σ^2 is known, so that the model has again a unique unknown parameter. This case is so common because of its simplicity; however, it is almost never met in real world problems. (The case where both μ and σ^2 are unknown is examined in the next section).

TCL in this case are given by

$$l, u = \bar{x} \pm \frac{z_{(1+\alpha)/2} \sigma}{\sqrt{n}} \quad (19)$$

Figure 4 provides a demonstration of the resulting MCCL in comparison to TCL. Here the performance of Methods 1 and 3 are excellent whereas that of Method 2 is poor.

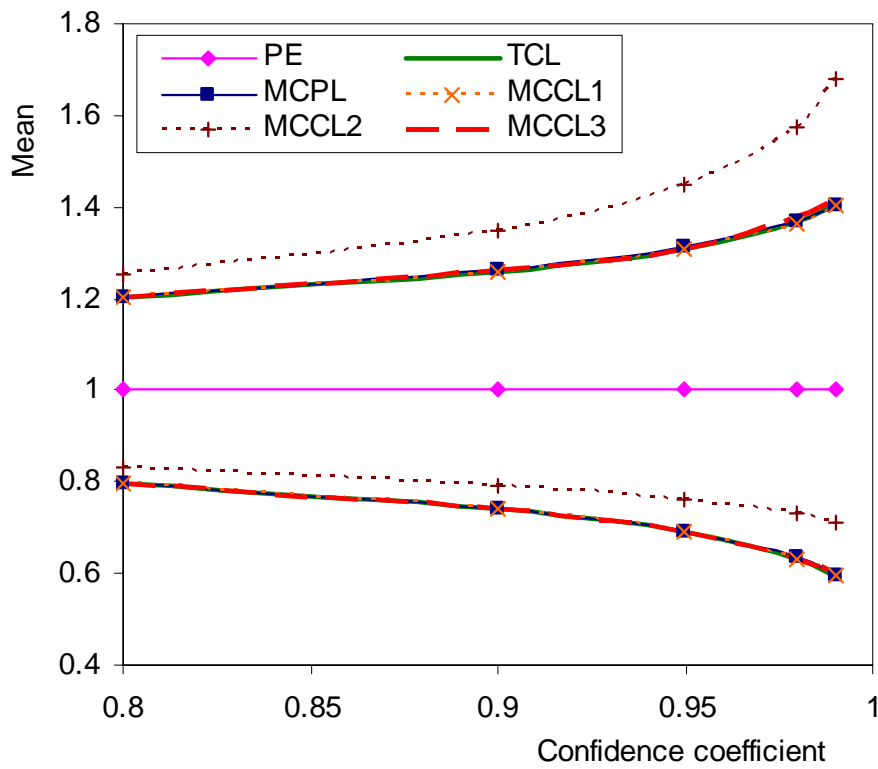


Figure 4 Point estimate, prediction limits and confidence limits for the mean of a normally distributed variable against the confidence coefficient α , as estimated by various methods assuming $\bar{x} = 1$, $\sigma = 0.5$ (assumed known) and $n = 10$. All curves except MCC2 are indistinguishable from the exact TCL.

5. Method Validation for Many Unknown Distributional Parameters

In the first example of the multiple parameter setting of the methods we consider again the mean of a normally distributed variable whose variance σ^2 is now unknown, being estimated by the sample variance s^2 , and we wish to find the confidence limits for its mean $\mu \equiv \beta$. TCL in this case are given by

$$l, u = \bar{x} \pm \frac{t_{(1+\alpha)/2}(n-1) s}{\sqrt{n}} \quad (20)$$

where $t_p(m)$ denotes the p -quantile of the Student t distribution with m degrees of freedom. Figure 5 provides a demonstration of the resulting MCCL for all three methods in comparison to TCL. Here the performance Method 3 is excellent whereas that of Methods 1 and 2 is poor.

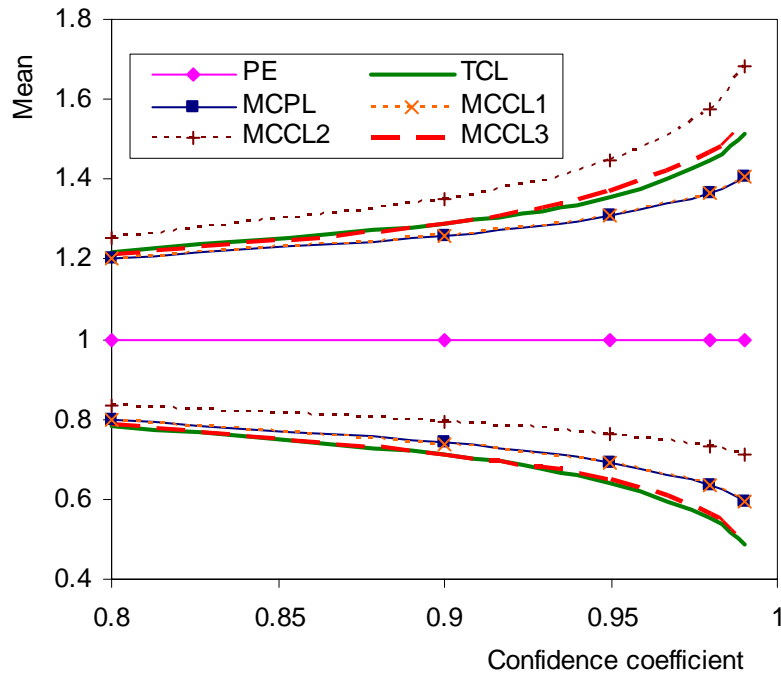


Figure 5 Point estimate, prediction limits and confidence limits for the mean of a normally distributed variable against the confidence coefficient α , as estimated by various methods assuming $\bar{x} = 1$, $s = 0.5$ (both assumed unknown) and $n = 10$. MCC3 are indistinguishable from the exact TCL.

In the second example we consider a normally distributed variable with unknown mean and variance, and we wish to find the confidence limits for its variance $\sigma^2 \equiv \beta$. TCL in this case are given by

$$l = \frac{n s^2}{\chi_{(1-\alpha)/2}^2(n)}, \quad u = \frac{n s^2}{\chi_{(1+\alpha)/2}^2(n)} \quad (21)$$

where $\chi_p^2(n)$ denotes the p -quantile of the χ^2 distribution with n degrees of freedom. Figure 6 provides a demonstration of the resulting MCCL for all three methods in comparison to TCL. Here the performance Methods 2 and 3 is excellent whereas that of Method 1 is poor.

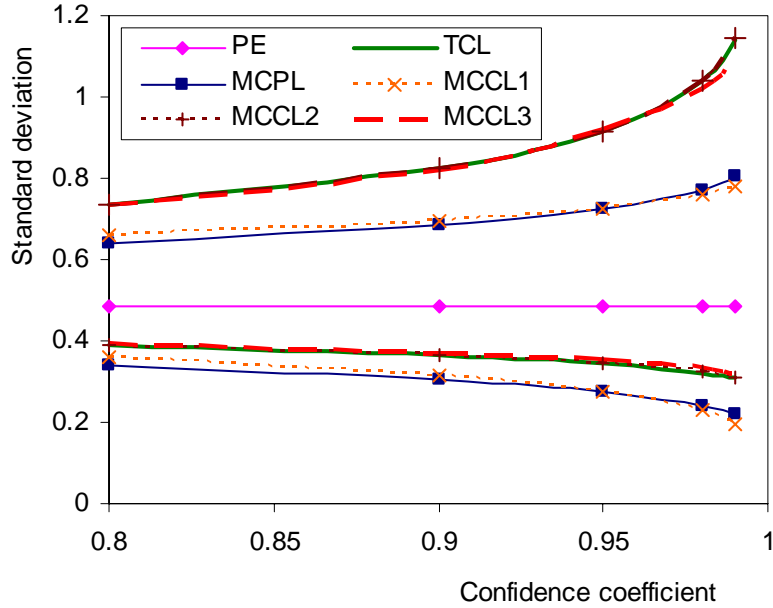


Figure 6 Point estimate, prediction limits and confidence limits for the variance of a normally distributed variable against the confidence coefficient α , as estimated by various methods assuming $\bar{x} = 1$, $s = 0.5$ (both assumed unknown) and $n = 10$. MCC2 and MCC3 are indistinguishable from the exact TCL.

In conclusion, all verification comparisons showed that the proposed Method 3 had in all cases excellent performance whereas each of the other methods in some cases failed to approximate the exact confidence limits that were obtained theoretically. Obviously, however, these demonstrations are not a mathematical proof of the exactness of the proposed method, which, in some degree, is heuristic.

References

Papoulis, A., *Probability and Statistics*, Prentice-Hall, 1990.

Ripley, B. D., *Stochastic Simulation*, Wiley, New York, 1987