Proceedings of the 2015 Winter Simulation Conference
L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, eds.

BOOTSTRAP CONFIDENCE BANDS AND GOODNESS-OF-FIT TESTS IN SIMULATION INPUT/OUTPUT MODELLING

Russell Cheng

School of Mathematics
University of Southampton
Highfield
Southampton, SO17 1BJ, UNITED KINGDOM

ABSTRACT

In the analysis of input and output models used in computer simulation, parametric bootstrapping provides an attractive alternative to asymptotic theory for constructing confidence intervals for unknown parameter values and functions involving such parameter values, and also for calculating critical values of EDF statistics used in goodness-of-fit tests, such as the Anderson-Darling A^2 statistic. This latter is known to give a GoF test that clearly out-performs better known tests such as the chi-squared test, but is hampered by having a null distribution that varies with different null hypotheses including whether parameters are estimated or not. Parametric bootstrapping offers an easy way round the difficulty, so that the A^2 test can routinely be applied. Moreover we show that bootstrapping is probabilistically *exact* for location-scale models, and so in general will be reasonably accurate using a mean and standard deviation parametrization. A numerical example is given.

1 PARAMETRIC SAMPLING

The purpose of this article is to point out the ease and effectiveness with which parametric sampling can be used to calculate confidence intervals (CI) and confidence bands (CB), and to make goodness-of-fit (GoF) tests in the analysis of input and output models used in simulation experimentation.

The fundamental statistical problem we face is as follows. We view a statistic T as a function of a sample $\mathbf{Y} = (Y_1, Y_2, ..., Y_n)$, so that $T = T(\mathbf{Y})$. The Y_i values are assumed to be independent observations from the same distribution with cumulative distribution function (CDF) $F(\cdot, \theta)$. To make any statistical inference with T we will need to know the distribution of T itself. Let $G_n(\cdot, \theta)$ be its CDF, the notation making it clear that this will depend on n as well as θ. The objective of parametric sampling is to estimate $G_n(\cdot, \theta)$. We distinguish two forms of parametric sampling (i) *Monte-Carlo estimation* and (ii) *parametric bootstrapping* depending on whether θ is known or not.

1.1 Monte Carlo Estimation

We call the version of parametric sampling, when θ is known, *Monte Carlo estimation.* In this version we generate B independent values of T : $T_1, T_2, ..., T_B$ then estimate $G_n(t, \theta)$ by the empirical distribution function (EDF) formed from the $T_1, T_2, ..., T_B$, namely

$$\tilde{G}_n(t, \theta) = \frac{\#\text{ of } T_j \le t}{B}, \quad -\infty < t < \infty.$$

The Glivenko-Cantelli lemma states that $\tilde{G}_n(t,\theta) \to G_n(t,\theta)$ *uniformly* over all t, with probability 1 as $B \to \infty$. Thus in principle $G_n(t,\theta)$, for all practical purposes, can be found numerically to any given accuracy by choosing B sufficiently large. This holds for any given n no matter how large or small.

In what follows we shall speak of this use of the EDF to estimate the CDF as being *probabilistically exact*, or just simply as being *exact*, in the sense that any desired accuracy is achievable in principle, whatever the n, by taking B sufficiently large. For the purposes of statistical inference, a value of $B = 1000$ is usually sufficient. For exploratory work or for illustrative purposes, where results only need to be indicative rather than sufficiently precise for formal decision taking, a value of B of just a few hundred is often sufficient.

1.2 Parametric Bootstrapping

The second case is when θ is unknown, so that Monte Carlo estimation cannot be applied directly, but we suppose that we have an actual sample $\mathbf{y} = (y_1, y_2, ..., y_n)$ from which to estimate θ. In what follows, we will use the *maximum likelihood estimator* (MLE), denoted by $\hat{\theta}$.

Parametric bootstrapping is simply Monte Carlo estimation of $G_n(t,\hat{\theta})$ *with the assumption that this will then provide an estimate of* $G_n(t,\theta)$. We generate by computer, B independent random samples $\mathbf{Y}_j^* = (Y_{ij}^*,$ $i = 1,2,...,n)$, $j = 1,2,...,B$, with the Y_{ij}^* all generated from the fitted distribution $F(\cdot,\hat{\theta})$. We then calculate the statistic of interest T from each sample $\mathbf{Y}_j^*$: $T_j^* = T(\mathbf{Y}_j^*)$ $j = 1,2,...,B$. Here the superscript * indicates a computer generated 'bootstrap' quantity obtained in a statistically identical way to the corresponding quantity obtained in the original process, except that $\hat{\theta}$ replaces the unknown true value θ. The EDF $\tilde{G}_n^*(t,\hat{\theta})$ of the $T_1^*, T_2^*, ..., T_B^*$ estimates $G_n(t,\hat{\theta})$ to whatever accuracy desired, by making B sufficiently large. We have written $\tilde{G}_n^*(t,\hat{\theta})$ with the added superscript *, instead of simply $\tilde{G}_n(t,\hat{\theta})$, to emphasize that this EDF has been obtained by bootstrapping.The EDF $\tilde{G}^*(t,\hat{\theta})$ only estimates $G(t,\hat{\theta})$, but as $\hat{\theta}$ estimates the unknown true θ it is reasonable to consider $\tilde{G}^*(t,\hat{\theta})$ as an estimator of $G(t,\theta)$ and not just of $G(t,\hat{\theta})$.

Much of bootstrap theory is given to establishing conditions under which $\tilde{G}_n^*(t,\hat{\theta})$ is a satisfactory estimator of $G_n(t,\theta)$. The ideal case is when $\lim_{B\to\infty} \tilde{G}_n^*(t,\hat{\theta}) = G_n(t,\theta)$ for all t, θ and n. Like Monte-Carlo estimation the parametric bootstrap is then (*probabilistically*) *exact*. We emphasize that only $B \to \infty$ is needed to achieve actual exactness, which will then hold for all n. Thus $n \to \infty$ is *not* required as well.

Under suitable regularity conditions the MLEs of θ and smooth functions $g(\cdot,\theta)$ will be asymptotically normally distributed, so that $|\hat{\theta} - \theta| = O_p(n^{-1/2})$ as $n \to \infty$, and, assuming $G_n(t,\theta)$ is twice differentiable in θ , also that $|\tilde{G}_n(t,\hat{\theta}) - G_n(t,\theta)| = O_p(n^{-1/2})$, as $n \to \infty$. In this case $\tilde{G}_n^*(t,\hat{\theta})$ can be expected to match, at least asymptotically, the behaviour of any theoretically derived estimator of $G_n(t,\theta)$. We shall not discuss bootstrap theory in much detail. Hjorth (1994) provides a clear and succinct account from a viewpoint well suited to that of this article. A clear but brief summary is provided by Young and Smith (2005). There is now a vast literature on the subject. Chernick (2008) provides over 140 *pages* of selected references covering research and practice up to 2007.

We discuss the use of parametric sampling for calculating confidence intervals (CIs) and for GoF tests, including an important case where it is exact rather than only approximate.

For simplicity we will omit the subscript n from now on, writing $\tilde{G}_n(t,\hat{\theta})$, $G_n(t,\theta)$ and so on, simply as $\tilde{G}(t,\hat{\theta})$, $G(t,\theta)$.

1.3 Bootstrap Confidence Intervals

In standard asymptotic theory the simplest two-sided $(1-\alpha)100\%$ confidence interval for φ_i, the ith component of θ (where we have used φ_i rather than θ_i for symbolic clarity), takes the well-known form

$$\left(\hat{\varphi}_i + z_{\alpha/2}\sqrt{\hat{V}_{ii}},\quad \hat{\varphi}_i + z_{(1-\alpha/2)}\sqrt{\hat{V}_{ii}}\right), \tag{1}$$

where $\hat{V}_{ii}$ is the ith main diagonal entry of $\hat{V} = [\mathrm{J}(\hat{\theta})]^{-1}$, the estimated covariance matrix of $\hat{\theta}$ obtained by inverting the observed information matrix $\mathrm{J}(\hat{\theta})$, and z_α is the α quantile of the standard normal distribution.

Note that here and in the rest of the paper we have used z_α to denote the α quantile. Often z_α is used to denote the *upper* α quantile, that is $z_{1-\alpha}$ in our notation, to take advantage of the fact that, for a distribution symmetric about zero, the lower quantile can then be written as $-z_\alpha$. We have not used this notation as we will be dealing with quantiles of some distributions not symmetric about zero.

This can be extended to a function $\lambda(x,\theta)$, $a \le x \le b$, that is dependent on the parameter values. A first order correct two-sided $(1-\alpha)100$ % confidence interval for $\lambda(x,\theta)$ is

$$(\lambda(x,\hat{\theta}) + z_{\alpha/2}\delta, \quad \lambda(x,\hat{\theta}) + z_{(1-\alpha/2)}\delta), \tag{2}$$

where

$$\delta = \sqrt{\frac{\partial\lambda(x,\theta)}{\partial\theta}\bigg|_{\hat{\theta}}^{T} V(\hat{\theta}) \, \frac{\partial\lambda(x,\theta)}{\partial\theta}\bigg|_{\hat{\theta}}}.$$

Under the same standard conditions we can construct a CI using bootstrapping. Several forms of CIs have been proposed, see Hjorth (1994) for a clear description. We describe just the simplest form.

We assume that the situation is exactly as just described with the observations in the data sample $\mathbf{y} = (y_1, y_2, ..., y_n)$ assumed to have the distribution with CDF $F(y,\theta)$ and that estimation of θ by ML yields the estimator $\hat{\theta}$. We then obtain B samples, called *bootstrap* (BS) samples, $\mathbf{y}_j^*$, $j = 1,2,...,B$, each of size n, by sampling from the fitted distribution $F(y,\hat{\theta})$. Now take T equal to the MLE $\hat{\theta}$ itself, that is $T(\mathbf{Y}) = \hat{\theta}(\mathbf{Y})$ so that from each BS sample $\mathbf{y}_j^*$, $j = 1,2,...,B$ we calculate $T_j^* = \hat{\theta}_j^*$, the MLE of θ based on that BS sample. A CI with confidence level $(1-\alpha)$ for an individual component, denoted by φ for symbolic clarity, can be constructed by arranging the bootstrap ML estimates of φ : $\hat{\varphi}_j^*$, $j = 1,2,...,B$ in rank order: $\hat{\varphi}_{(1)}^* \le \hat{\varphi}_{(2)}^* \le ... \le \hat{\varphi}_{(B)}^*$, then taking the $(1-\alpha)100\%$ confidence inequality $\hat{\varphi}_{(l)} \le \hat{\varphi}^* \le \hat{\varphi}_{(m)}^*$, where $l = (\alpha/2)B$ and $m = (1-\alpha/2)B$, and replacing $\hat{\varphi}^*$ by the unknown true φ. The resulting confidence interval for φ takes the form

$$\hat{\varphi}_{(l)}^* \le \varphi \le \hat{\varphi}_{(m)}^*.$$

This is called the *percentile CI.*

The bias of $\hat{\varphi}^*$ can be estimated by

$$\hat{\beta} = B^{-1}\sum_{j=1}^{B}\hat{\varphi}_j^* - \hat{\varphi},$$

though it has to be said that this bias correction give rather mixed results in practice.

The following flowchart summarizes the steps of the percentile CI calculation including the bias correction.

Flowchart for calculating a $(1-\alpha)100\%$ BS CI for a component φ of θ (3)

Given $\mathbf{y} = (y_1, y_2, ..., y_n)$, a random sample drawn from distribution with CDF$F(y,\theta)$

$$\hat{\theta} = \hat{\theta}(\mathbf{y}) \qquad \text{MLE}$$

For $j = 1$ to B

$$\mathbf{y}_j^* = (y_{ij}^* \sim F(y,\hat{\theta}),\ i = 1,2,...,n) \quad j\text{th BS sample}$$

$$\downarrow$$

$$\hat{\theta}_j^* = \hat{\theta}(\mathbf{y}_j^*) \qquad\qquad j\text{th BS MLE}$$

$$\downarrow$$

$$\hat{\varphi}_j^* \qquad j\text{th BS MLE of } \varphi\text{, the component of interest}$$
$$\downarrow$$
$$\hat{\beta} = n^{-1}\sum_{j-1}^{n}\hat{\varphi}_j^* - \hat{\varphi} \qquad \text{Estimate of bias due to bootstrapping}$$

End j

$$\hat{\varphi}^*_{(1)} \le \hat{\varphi}^*_{(2)} \le \ldots \le \hat{\varphi}^*_{(B)} \qquad \text{Ordered BS MLEs of component of interest}$$
$$\downarrow$$
$$l = \max[1,(\alpha/2)B],\ m = (1-\alpha/2)B \qquad \text{Rounded integer subscripts}$$
$$\downarrow$$
$$\hat{\varphi}^*_{(l)} - \hat{\beta} \le \varphi \le \hat{\varphi}^*_{(m)} - \hat{\beta} \qquad \text{Bias corrected } (1-\alpha)100 \text{ percentile CI for } \varphi$$

1.4 Toll Booth Example

We illustrate the discussion so far with an example arising in a study made of the operation of toll booths of the (old) Severn River bridge in the United Kingdom, see Griffiths and Williams (1984). Each toll booth is modelled as a single server queue and data was collected of the service time of vehicles, that is the time taken for a vehicle to pay at the toll booth before crossing the bridge. As the example is for illustration only we use a small sample of size $n = 47$. The observations are in seconds.

4.3	4.7	4.7	3.1	5.2	6.7	4.5	3.6	7.2	10.9	6.6	5.8
6.3	4.7	8.2	6.2	4.2	4.1	3.3	4.6	6.3	4.0	3.1	3.5
7.8	5.0	5.7	5.8	6.4	5.2	8.0	4.9	6.1	8.0	7.7	4.3
12.5	7.9	3.9	4.0	4.4	6.7	3.8	6.4	7.2	4.8	10.5	

The service times are treated as independent random variates with a gamma distribution $G(a,b)$ with probability density function (PDF)

$$f(y) = \frac{1}{\Gamma(a)b^a}y^{a-1}\exp(-y/b), \quad a,b>0,\ 0<y<\infty. \tag{4}$$

with parameters a and b that are unknown. The MLEs are $\hat{a} = 9.20$ and $\hat{b} = 0.63$. The left-hand plot in Figure 1 compares the fitted CDF with the EDF, and the right-hand plot compares the fitted PDF with a frequency histogram of the sample; 90% CIs calculated from (1) gave $(6.14, 12.27)$ for a and $(0.41, 0.85)$ for b. We also calculated the BS percentile CIs for a and b, with $\alpha = 0.1$, $B = 1000$.The BS CI for a was $(7.06, 13.93)$ showing this to be biased to the right compared with the CI using the asymptotic formula. The CI for b was $(0.42, 0.85)$ which is almost identical to the asymptotic CI.

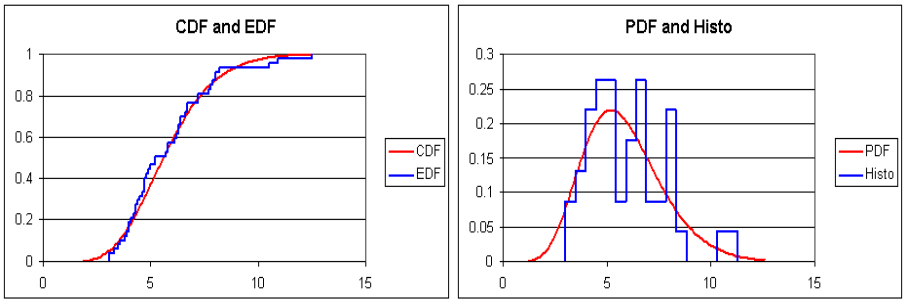

Figure 1: Toll Booth Data. Fitted CDF, EDF, fitted PDF and frequency histogram

However the main interest is not the parameter estimates themselves but a quantity such as the steady-state expected overall waiting time, $w(x,\alpha,b)$, of customers in the queue, regarded as a function of the

average arrival rate x. For instance, if the arrival pattern is Poisson with rate x, then it is known that

$$w(x,a,b) = \frac{(1+a)ab^2x}{2(1-abx)}. \tag{5}$$

We would be interested in examining the value of $w(x,a,b)$ over a range, $x_0 \le x \le x_1$, of arrival rates. The MLE of w is simply $\hat{w} = w(x,\hat{a},\hat{b})$, $x_0 \le x \le x_1$, with a CI for any given x that can be calculated using asymptotic theory by (2). We postpone presenting the results for $\hat{w}$ until after we have considered calculating CIs using of bootstrapping, so as to directly compare the results using the two approaches.

1.5 Coverage Error and Scatterplots

The quality of CIs is usually discussed in terms of their *coverage error*, that is the difference between the actual confidence level achieved and the target value of $1-\alpha$. If its coverage error tends to zero as $n \to \infty$, a CI said to be *consistent.* The coverage error is usually given in terms of a probabilistic order of magnitude, which can be determined by asymptotic theory that also extends to cover bootstrap CIs especially for parametric bootstrapping where the MLE $\hat{\theta}$ is used. In general the coverage error is typically $O(n^{-1/2})$, see Young and Smith (2005, Section 11.1), a disappointing result given that $\left|\hat{\theta} - \theta^0\right| = O_p(n^{-1/2})$ as previously discussed.

However there is the proviso that for balanced two-sided CIs where the target confidence level is the same (i.e. $\alpha/2$) at both ends of the interval both, the coverage error is then usually reduced to $O(n^{-1})$. We do not go into details, but this improved performance of a balanced CI arises because coverage error comes mainly from bias in the estimators. However the effect of this bias is in opposite directions at the two ends of the CI, so that when the CI is balanced and the distribution of $\hat{\theta}$ is asymptotically normal and so symmetric, then they cancel sufficiently precisely to reduce the coverage error to $O(n^{-1})$.

Coverage error can be usefully viewed as the discrepancy between the coverage actually achieved and the target value when the latter is derived under asymptotic normality assumptions. As a general guide one would therefore expect the coverage to be lower the closer that the distribution of the estimated parameter of interest, $\hat{\theta}$ say, is to being normal.

Scatterplots are a useful feature of BS analysis for indicating if the distribution of $\hat{\theta}$ is normal or not.

As an illustration consider the toll booth example again but where we consider two alternative parametrizations. Figure 2 gives the scatterplots of the $B = 1000$ BS MLEs $(\hat{a}_j^*, \hat{b}_j^*)$, $j = 1,2,...,B$ and of $(\hat{\mu}_j^*, \hat{\sigma}_j^*)$, $j = 1,2,...,B$ where $\mu = ab$ and $\sigma = \sqrt{ab}$ are the mean and standard deviation (SD) of the gamma distribution. The points in each plot are divided into two types. The black points corresponding to $(1-\alpha)100\%$ (where $\alpha = 0.1$, in the plots) of the total number of points. These points estimate what is called a $(1-\alpha)100\%$ *loglikelihood-based confidence region* as discussed by Hall (1987) the construction of which we will describe in Section 2. The green points lie outside the confidence region. For now we simply look at all the points in each plot as a whole and note that the scatter plot of the original $(\hat{a}_j^*, \hat{b}_j^*)$ has a very noticeable asymmetrically curved 'banana' form. This is typical of such plots, see Hjorth (1994, Fig 6.1) or Davison and Hinkley (1997, Fig 7.11, left lower plot). In contrast, the scatter of $(\hat{\mu}_j^*, \hat{\sigma}_j^*)$ points, though indicating strong correlation between $\hat{\mu}_j^*$ and $\hat{\sigma}_j^*$, has a much more elliptically symmetrical form, indicative of normality.

In Section 2 we will consider a practical use of such scatterplots not often discussed.

In general the coverage error of the percentile CI and the simple CI can be noticeable especially when n is relatively small, but perhaps not so large as to preclude their use. As an example we consider fitting the gamma distribution $G(a,b)$ with PDF (4) to a sample of size $n = 50$ drawn from the standard exponential distribution, so that the true parameter values are $a = 1$ and $b = 1$. The value of b is unimportant being a scale parameter, so that the form of the distribution of the MLEs depends only on the value of a. The value $a = 1$ is a fairly extreme choice making the joint distribution of $(\hat{a},\hat{b})$ distinctly non-normal when $n = 50$, so provides a difficult coverage error test. Table(1) gives the results.

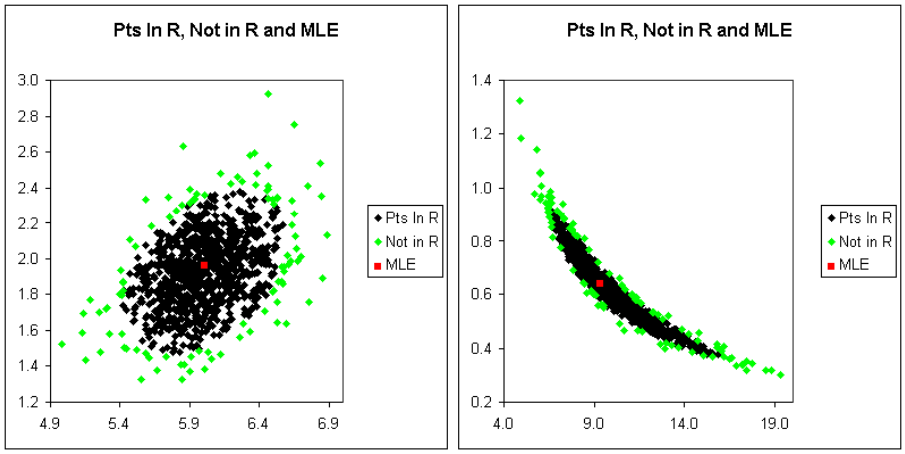

Figure 2: Toll Booth Example. Left: Scatterplots of 1000 BS MLEs $\hat{\mu}$ and $\hat{\sigma}$ and Right: 1000 BS MLEs $\hat{a}$ and $\hat{b}$. The MLE from the original sample is the red square. Each plot divides the points into those estimated to be in a likelihood-based region R_α (black) and those not in R_α (green). Here $\alpha = 0.1$

Table 1: Coverage Error Results: Actual coverage of 1000 CIs for parameters a, b, μ, σ of G(1,1) distribution with nominal coverage level $(1-\alpha) = 0.9$

		a	b	μ	σ
BS not bias corrected	In CI	862	858	887	863
	Not in CI	138	142	113	137
BS bias corrected	In CI	916	871	884	866
	Not in CI	84	129	116	134
Asymptotic	In CI	899	860		
	Not in CI	101	140		

It will be seen that the bias-corrected BS CI gives is not too unsatisfactory taking the coverage of both a and b into account.

2 CONFIDENCE BANDS FOR FUNCTIONS

We can extend comparison of the asymptotic theory and bootstrapping approaches to estimation of a function $\lambda(x,\theta)$, $x_0 \le x \le x_1$, taking as our example, the waiting time function distribution in the toll booth example, $w(x,a,b)$, $x_0 \le x \le x_1$, as given in (5).

Consider first a two-sided CI for a given x. Using asymptotic theory the MLE is $w(x,\hat{a},\hat{b})$, $x_0 \le x \le x_1$ and we can apply (2) directly to construct an asymptotic $(1-\alpha)$ level CI for $w(x,a,b)$ at any *fixed* x. The left-hand plot in Figure 3 shows the MLE of the function $w(x,\hat{a},\hat{b})$ (red line) calculated at 10 equally spaced values of x in the range $0 \le x \le 0.1$. The solid (green) upper line and solid (blue) lower line, give the corresponding limits of the CI at level $(1-\alpha)$, which allows the limits to be read off for any *one* given x. It should be stressed that the two lines do *not* enable CI's to be calculated at *several different* x with confidence level $(1-\alpha)$ holding *simultaneously* across all values. A simple but very conservative (simultaneous) Bonferroni confidence level (see for example Miller, 1981) can be constructed for M different x's simultaneously by setting the level to be $(1-M^{-1}\alpha)$ for each x with overall level $\ge (1-\alpha)$. The right-hand plot in Figure 3 shows the CI limits calculated by bootstrapping. As in the asymptotic formula case, the BS CI limits were calculated over ten equally spaced $x-$ intervals. At each x_i, the $B = 1000$ BS waiting time values $w^*_j(x_i) = w(x_i,\hat{a}^*_j,\hat{b}^*_j)$, $j = 1,2,...,B$ were calculated then ordered

so that $w^*_{(1)}(x_i) \le w^*_{(2)}(x_i) \le ... \le w*_{(B)}(x_i)$ and the confidence interval at that x_i taken as

$$w^*_{(l)}(x_i) \le w(x_i, a, b) \le w^*_{(m)}(x_i)$$

where l and m are as given in the flowchart (3).

It will be seen that in the bootstrap case the CI is slightly asymmetric at each x value.

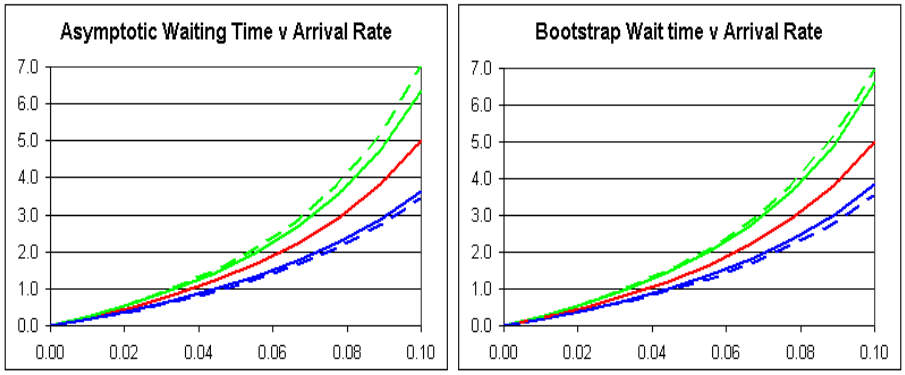

Figure 3: Toll Booth Example. Plots of expected waiting time versus arrival rate. From calculations using asymptotic normal theory - left graph; using bootstrapping - right graph. Upper CB limit - dashed green, Upper CI Limit - green, MLE - red, Lower CI Limit - blue, Lower CB Limit - dashed blue

For simultaneous CI's, rather than use a Bonferroni CI, it is more efficient to calculate a *confidence band*, which with given confidence level, will *entirely contain* the unknown curve $w(x,a,b)$, $x_0 \le x \le x_1$. Simultaneous confidence intervals are discussed by Miller (1981). The construction of confidence bands based on asymptotic theory has been described by Cheng and Iles (1983) for continuous CDFs and by Cheng (1987) more generally. Setting $\theta = (a,b)$, we have asymptotically as $n \to \infty$, that

$$(\hat{\theta} - \theta)^{\mathrm{T}} \mathrm{J}(\hat{\theta})(\hat{\theta} - \theta) \sim \chi^2_2, \tag{6}$$

where χ^2_2 is the chi-squared distribution with two degrees of freedom, that is the exponential distribution. Writing $\chi^2_2(\alpha)$ for the α quantile, inversion of the probability statement $\Pr((\hat{\theta} - \theta)^{\mathrm{T}} \mathrm{J}(\hat{\theta})(\hat{\theta} - \theta) \le \chi^2_2(1 - \alpha)) = 1 - \alpha$, in the usual way gives the $(1-\alpha)$ confidence region

$$R_{1-\alpha} = \{\theta : (\hat{\theta} - \theta)^{\mathrm{T}} \mathrm{J}(\hat{\theta})(\hat{\theta} - \theta) \le \chi^2_2(1-\alpha)\}. \tag{7}$$

Thus if

$$w_{\min}(x) = \min_{\theta \in R_{1-\alpha}} w(x,\theta),\ w_{\max}(x) = \max_{\theta \in R_{1-\alpha}} w(x,\theta),\ x_0 \le x \le x_1, \tag{8}$$

then the condition

$$w_{\min}(x) \le w(x,\theta) \le w_{\max}(x),\ x_0 \le x \le x_1 \tag{9}$$

holds *simultaneously* for all $\theta \in R_{1-\alpha}$ with asymptotic confidence level no less than $(1-\alpha)$, so that (9) is an *asymptotic confidence band* containing $w(x,\theta)$, for all $x_0 \le x \le x_1$, with confidence level no less than $(1-\alpha)$ as $n \to \infty$.

We can obtain a bootstrap version of the confidence band by using B BS MLEs $\hat{\theta}^*_j$, $j = 1,2,...,B$ to construct a bootstrap equivalent of $R_{1-\alpha}$. When $\hat{\theta}$ is exactly normally distributed, the region $R_{1-\alpha}$ has the property that all points in $R_{1-\alpha}$ have higher likelihood values than all those outside it. Cox and Hinkley (1974, p.218) call such a region a *likelihood-based region*. Methods for calculating such a region are described by Hall (1987) who recommends the percentile-t bootstrap method of generating a set of θ^*_j

points and then use of a non-parametric kernel smoothing method to identify the boundary of $R_{1-\alpha}$. This method will usually have the attractive property of making the coverage error in $R_{1-\alpha}$ of order $O(n^{-1})$.

In our example we use a significantly simpler BS approach to obtain an estimate of $R_{1-\alpha}$ which nevertheless has the same motivation as the approach proposed by Hall. First note that we are free to choose the parametrization θ in constructing the confidence band. We so therefore choose θ to make the distribution of its MLE $\hat{\theta}$ as close to being normal as possible when (6) will be a good approximation. In the toll booth example if we compare the scatterplot of $(\hat{a}^*, \hat{b}^*)$ with that of $(\hat{\mu}^*, \hat{\sigma}^*)$ in Figure 2 the latter is clearly much more normally distributed. To construct an estimate of $R_{1-\alpha}$ we note that the quadratic in (7) is the leading term in the Taylor expansion of the usual loglikelihood ratio $2(L(\hat{\theta}) - L(\theta))$, corresponding to distributions that are asymptotically normal. The equation

$$q(\theta) = (\hat{\theta} - \theta)^{\mathrm{T}} \mathrm{J}(\hat{\theta})(\hat{\theta} - \theta) = c,$$

with c a constant thus gives loci where $L(\theta)$ is constant with smaller $q(\theta) = c$ corresponding to larger $L(\theta)$. Therefore if we calculate $q_j^* = (\hat{\theta} - \hat{\theta}_j^*)^{\mathrm{T}} \mathrm{J}(\hat{\theta})(\hat{\theta} - \hat{\theta}_j^*),\ j = 1, 2, ..., B$, corresponding to the BS $\hat{\theta}_j^*$ and order the points so that

$$q_{(1)}^* \leq q_{(2)}^* \leq \ldots \leq q_{(B)}^*,$$

then the points $\hat{\theta}_{(j)}^*$ corresponding to the first $m = \lfloor (1-\alpha)B \rfloor$ $q_{(j)}^*$ values, that is $q_{(1)}^* \leq q_{(2)}^* \leq \ldots \leq q_{(m)}^*$, will estimate a likelihood-based region of level $(1-\alpha)$. These points can be used directly in (8) to represent $R_{1-\alpha}$ to calculate $w_{\min}(x)$ and $w_{\max}(x)$.

The black points in the right-hand plot of Figure 2 are for the case $B = 1000$, $\alpha = 0.1$ in the toll booth example. The left-hand plot are the *same* selection of bootstrap points in their original parametrization $\theta = (a, b)$. These are the points used in calculating the CB. Figure 3 shows the upper and lower CI and CB limit curves (solid for CI, dashed for CB) using both the asymptotic and BS approaches. It will be seen that the BS curves compared with the asymptotic curves are skewed slightly higher, and the CB limits are wider than the CI limits for either approach.

3 CONFIDENCE INTERVALS USING PIVOTS

Coverage error can be entirely eliminated if the CI can be constructed using a *pivotal quantity.* This has interesting repercussions not only for CI construction but for parametric bootstrap GoF tests so we will discuss this rather more fully here.

A pivotal statistic is usually defined somewhat cryptically and rather unspecifically as a function of a random sample, $\mathbf{Y}$, that also depends on θ the parameter of interest but whose distribution does not depend on the value of θ.

In practice the form of pivotal statistics tends to follow the two used in constructing CIs for the parameters μ and σ^2 from a sample $\mathbf{Y}$ drawn from the normal distribution $N(\mu, \sigma^2)$, namely the celebrated studentized statistics T and W defined by

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{\nu n}}, \quad W = \frac{S^2}{\sigma^2}, \quad \text{where } S^2 = \sum_{i=1}^{n} (Y_i - \bar{Y})^2, \text{ with } \nu = n - 1. \tag{10}$$

Note that we have used S^2 to denote a sum of squares. This is not to be confused with the lower case notation s^2 conventionally used to denote the sample variance.

If μ and σ^2 are the true values then T has the Student's t-distribution and W has the chi-squared distribution both with ν degrees of freedom. We shall write $t_\nu(\alpha)$ and $\chi^2_\nu(\alpha)$ for the α quantiles of these two distributions. As with z_α, many authors use $t_\nu(\alpha)$ and $-t_\nu(\alpha)$ to denote the upper and lower t_ν quantiles when $\alpha < 0.5$ as the t-distribution is symmetric about zero. We have not used this notation in view of our discussion of non-symmetric pivot distributions shortly to follow. An exact balanced two

sided $(1-\alpha)100\%$ CI for μ and for σ^2 is obtained by replacing the random variables in the probability statements (assuming $\alpha < 0.5$)

$$\Pr\{t_\nu(\alpha/2) \le \frac{\bar{Y}-\mu}{S/\sqrt{\nu n}} \le t_\nu(1-\alpha/2)\} = 1-\alpha \text{ and } \Pr\{\chi^2_\nu(\alpha/2) \le \frac{S^2}{\sigma^2} \le \chi^2_\nu(1-\alpha/2)\} = 1-\alpha$$

by their sample values and inverting the inequalities to give the well-known *studentized* confidence intervals for μ and σ^2:

$$\left(\bar{y} - t_\nu(1-\alpha/2)S/\sqrt{\nu n},\ \bar{y} - t_\nu(\alpha/2)S/\sqrt{\nu n}\right) \text{ and } \left(\frac{S^2}{\chi^2_\nu(1-\alpha/2)},\ \frac{S^2}{\chi^2_\nu(\alpha/2)}\right). \tag{11}$$

In this particular case, where the Y's are normal, the distributions of T and W are well-known with Student's t-quantiles and χ^2_ν quantiles well tabulated and available in standard computer routines. There is therefore no need to use Monte-Carlo estimation to estimate them.

However there is a more general situation where Y is not normal, but the same statistics T and W can still be used to construct CIs, if their CDFs are known. If not known, they can easily be estimated by Monte-Carlo estimation.

Consider a sample $Y_1, Y_2, ..., Y_n$ drawn from the location-scale model $Y = \mu + \sigma X$, where μ and σ are fixed but unknown parameters, and X is a random variable with CDF $F(x)$, whose form $F(\cdot)$ is completely known and not dependent on unknown parameters. The normal model $N(\mu, \sigma^2)$ just considered is an example with X simply the standard normal distribution $N(0,1)$. Other examples include the exponential, logistic, extreme value and Weibull distributions, (the last under an appropriate transformation so that it becomes the extreme value distribution). Consider

$$T' = \frac{\bar{Y}-\mu}{S/n},\ \ W = \frac{S^2}{\sigma^2} \text{ where } S^2 = \sum_{i=1}^{n}(Y_i - \bar{Y})^2,$$

these being essentially the T and W statistics defined in (10) with a slight simplification in T where the factor ν is replaced by n. Now write $Y_i = \mu + \sigma X_i$ and we get

$$T' = \frac{\bar{X}}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}},\ \ W = \sum_{i=1}^{n}(X_i - \bar{X})^2, \tag{12}$$

showing that T' and W are pivotal quantities. The confidence sets for μ and σ^2 take the same form as (11) for the normal model, except that appropriate quantiles of $G_{T'}(\cdot)$, the CDF of T', and of $G_W(\cdot)$, the CDF of W, need to be substituted for the Student's t and χ^2-quantiles appearing in (11).

If these CDFs are not known, Monte-Carlo estimation, as described in subsection 1.1 provides a simple way to produce samples T'_j, W_j , $j = 1, 2, ..., B$ from (12) with the EDF of the $\{T'_j\}$ estimating $G_{T'}(\cdot)$ and of the $\{W_j\}$ estimating $G_W(\cdot)$, from which the required quantiles can be read off.

The EDFs are probabilistically exact. Therefore, for location-scale models, CIs calculated in this way have no coverage error, whatever the sample size n.

4 BOOTSTRAP GOODNESS OF FIT

Once we have fitted a model, the natural question is: Does the model that we have fitted actually match the data very well? For instance when we fitted a gamma distribution to toll booth service time data, does the fitted gamma distribution capture the characteristics of the data properly?

The classical way to answer this question is to use a GOF test. Let $Y = (Y_1, Y_2, ..., Y_n)$ where our *null hypothesis*, denoted by H_0, is that $Y_i \sim F(\cdot, \theta)$, where the form of F is known. We consider GoF tests based on an EDF test statistic of the form $T = T(Y)$ that measures the difference between the EDF $\tilde{F}(y)$ of the

sample values $(Y_1, Y_2, ..., Y_n)$ and the fitted $F(y, \theta)$. We will consider both the case where θ is known, and where it is not known. In the latter case we use the MLE $\hat{\theta}$ to estimate it. The distribution of T under the null hypothesis has to be known to apply the GoF test. We simply calculate $T = T(\mathbf{Y})$ for the given sample, and compare the value with a high quantile $T_{1-\alpha}$ of the null distribution, for which $\Pr(T \leq T_{1-\alpha}) = 1 - \alpha$. A small value of α is used, $\alpha = 0.1$ or 0.05 or 0.01 is typical, so that if the null hypothesis is true then it is unlikely that the T value calculated from the actual observations $\mathbf{Y}$ will be greater than $T_{1-\alpha}$. If $T > T_{1-\alpha}$ the GoF test has failed at the $(1-\alpha)$ level, and H_0 is rejected with the inference that the Y_i have *not* been drawn from $F(\cdot, \theta)$.

Two well-known EDF test statistics are the Cramér-von Mises W^2 statistic and the Anderson-Darling A^2 statistic both of which have the form

$$T = \int_{-\infty}^{\infty} \psi(y, \theta)[\tilde{F}(y) - F(y, \theta)]^2 dF(y, \theta),$$

where $\psi(y, \theta) = 1$ for the W^2 statistic and $\psi(y, \theta) = [F(y, \theta)(1 - F(y, \theta))]^{-1}$ for the A^2 statistic. Equivalent versions that are computationally more convenient are, for the W^2 case

$$W^2 = \frac{1}{12n} + \sum_{i=1}^{n} [Z_i - \frac{2i-1}{2n}]^2$$

and for the A^2 case

$$A^2 = -n - \sum_{i=1}^{n} (2i-1)[\ln Z_i + \ln(1 - Z_i)]/n \tag{13}$$

where, in either case, $Z_i = F(y_{(i)}, \theta)$ if θ is known and θ is replaced by $\hat{\theta}$ if not, with $y_{(i)}$ the ith order statistic.

Another well-known EDF statistic is the Kolmogorov-Smirnov D statistic. Perhaps the best known test is the chi-squared GoF test. The main reason for its popularity is that it is relatively easy to implement. The test statistic is easy to calculate and moreover it has a known chi-squared distribution under the null, which makes critical values easy to obtain. However the chi-squared test has two obvious weaknesses. It is significantly less powerful than EDF tests, and it has a certain subjective element because the user has to divide the data into groups of her/his own choosing.

It is quite clear from work done notably by Stephens, see Stephens (1970, 1974) and D'Agostino and Stephens (1986), that A^2 yields the most powerful test and so should be the test statistic of choice.

The problem with the W^2 and A^2 statistics is that their distribution does not remain the same but changes depending on the null. Thus different critical values are needed. D'Agostino and Stephens (1986) have tabulated a range of critical values for different GoF statistics and different nulls including those for the exponential, gamma, Weibull, extreme value, logistic and Cauchy distributions. Though the situation has improved, statistical packages that offer GoF tests do not always make it clear what critical values are implemented.

The parametric bootstrap offers a very practical way out of the difficulty, whether parameters have to be estimated or not. Stute, Mantega and Quindimil (1993) show that under regularity conditions, so that the MLE, $\hat{\theta}$, is a consistent estimator, and provided $F(y, \theta)$ is a sufficiently smooth function of y and θ, then the distribution of an EDF GoF test statistic is consistently estimated by its bootstrap version. Stute, Mantega and Quindimil (1993) demonstrate this specifically for the D and W^2 statistics. Babu and Rao (2004) give a rigorous proof of the weak consistency of the parametric bootstrap, examining a number of examples, including the normal and Cauchy distributions.

In what follows we focus on the Anderson-Darling statistic A^2. In view of its similarity, our discussion applies also to W^2 but for reasons of space we do not discuss W^2 further.

We will write $A^2(\mathbf{y}, \theta)$ for the value of A^2 as calculated in (13) from the EDF $\tilde{F}$ of a sample $\mathbf{y}$ and the CDF $F(\cdot, \theta)$. For the case when the MLE $\hat{\theta}$ is used instead of θ we write $A^2(\mathbf{y}, \hat{\theta})$.

In the case where θ is known, we use Monte-Carlo estimation to generate samples $\mathbf{y}_j = (y_{1j}, y_{2j}, ..., y_{nj})$, $j = 1, 2, ..., B$, with all $y_{ij} \sim F(\cdot, \theta)$ drawn from the known distribution, from which we calculate test statistic values $T_j = A_j^2 = A^2(\mathbf{y}_j, \theta)$, $j = 1, 2, ..., B$. The EDF of the A_j^2 is (probabilistically) exact for estimating the null distribution of the test statistic in this case.

If parameters have to be estimated we still generate B samples $\mathbf{y}_j^*$, $j = 1, 2, ..., B$, each of size n, but now with sampled values $y_{ij}^* \sim F(\cdot, \hat{\theta})$ where the MLE $\hat{\theta}$, calculated from the original sample, replaces the unknown θ. For each j we have to again estimate θ, calculating the MLE $\hat{\theta}_j^* = \hat{\theta}(\mathbf{y}_j^*)$ based on the jth sample. We can then calculate $T_j^* = A_j^{*2} = A^2(\mathbf{y}_j^*, \hat{\theta}_j^*)$, $j = 1, 2, ..., B$, using the formula for A^2 with $z_{ij} = F(y_{(i),j}^*, \hat{\theta}_j^*)$, where $y_{(1),j}^*, y_{(2),j}^*, ..., y_{(n),j}^*$ is the ordered jth BS sample. The EDF of the sample $A_j^{*2} = A^2(\mathbf{y}_j^*, \hat{\theta}_j^*)$, $j = 1, 2, ..., B$, estimates the null distribution of A^2 for the case where the parameters have been estimated by MLE. Writing the GoF critical value at level $(1-\alpha)$ as $A_{(1-\alpha)}^2$, this is estimated by the $(1-\alpha)$ quantile of the EDF, so if the A_j^{*2} are placed in order $A_{(1)}^{*2} \le A_{(2)}^{*2} \le ..., \le A_{(B)}^{*2}$, the estimate is then $A_{(1-\alpha)}^2 = A_{(m')}^{*2}$, where $m' = \lfloor (1-\alpha)B \rfloor$. These calculations are set out in the GoF flowchart (14).

Flowchart for the A^2 GoF test when θ is estimated by the MLE $\hat{\theta}$ (14)

Given $\mathbf{y} = (y_1, y_2, ..., y_n)$, a random sample drawn from distribution with CDF $F(y, \theta)$, θ unknown

$\downarrow$

$\hat{\theta} = \hat{\theta}(\mathbf{y})$ MLE

$\downarrow$

$A^2(\mathbf{y}, \hat{\theta})$ A^2 GoF test statistic

For $j = 1$ to B

$\mathbf{y}_j^* = (y_{ij}^* \sim F(y, \hat{\theta}),\ i = 1, 2, ..., n)$ jth BS sample

$\downarrow$

$\hat{\theta}_j^* = \hat{\theta}(\mathbf{y}_j^*)$ jth BS MLE

$\downarrow$

$A_j^{*2} = A^2(\mathbf{y}_j^*, \hat{\theta}_j^*)$ jth BS A^2 GoF test statistic

End j

$A_{(1)}^{*2} \le A_{(2)}^{*2} \le ... \le A_{(B)}^{*2}$ Ordered null-distributed A_j^{*2} values

$\downarrow$

$m = (1-\alpha)B$ Rounded integer subscript

$\downarrow$

$A_{(1-\alpha)}^{*2} = A_{(m)}^{*2}$ BS estimate of GoF critical value at level $(1-\alpha)$

$\downarrow$

GoFTest: If $A_{(1-\alpha)}^{*2} < A^2(\mathbf{y}, \hat{\theta})$ Reject null hypothesis H_0 at level $(1-\alpha)$.

In general, the bootstrap at each j introduces an approximation error as $\hat{\theta}$ has to be used in place of the unknown true θ^0 in forming the sample $\mathbf{y}_j^*$. This introduces an error of order $|\hat{\theta} - \theta^0| = O_p(n^{-1/2})$. It is relatively easy to incorporate a sensitivity assessment into the procedure to gauge the importance of this error. We will discuss how this is done, but before we do so we discuss an important family of models where the bootstrapping remains exact even when parameters are estimated.

For location-scale models the use of MLE's in EDF GoF tests is equivalent to studentization so that GoF test is exact. This property has been recognized and commented on in Stephens (1974) and D'Agostino and Stephens (1986), however their focus was on the calculation of tables of critical values and so attention was not drawn to the usefulness of the property in carrying out BS GoF tests. The property is also pointed out

by Babu and Rao (2004), but their proof is quite technical. We give a simpler, we hope more transparent, proof that covers the case where the location and scale parameters are both unknown.

Consider the location-scale model where the observations have the form $Y_i = \mu + \sigma X_i$. We write X_i as $X_i = F^{-1}(U_i)$, with F the completely known continuous CDF of X. We can think of the U_i, $i = 1,2,...,n$ as being the fundamental random variables from which the X_i and then the Y_i are formed. Thus there is a one-to-one correspondence between each sample of U_i's and the observed sample $\{Y_i\}$ that it gives rise to. Our key result is that whatever the underlying sample of uniforms $\{U_i\}$, using MLE to estimate μ and σ, yields an estimated standardized set of Y_i's, $(Y_i - \hat{\mu})/\hat{\sigma}$, $i = 1,2,...,n$, each value of which, whatever the i, is the same whether calculated for the original sample or under bootstrapping. We show this as follows.

Let $\{u_i^0\}$ be the uniforms that give rise to $\{y_i^0\}$ our observed sample, so that $y_i^0 = \mu^0 + \sigma^0 x_i^0$ where $x_i^0 = F^{-1}(u_i^0)$ and μ^0, σ^0 are the true but unobserved μ and σ values. Let f be the PDF. The loglikelihood in terms of the y_i^0 is $L = -n\ln\sigma + \sum_{i=1}^n \ln f[(y_i^0 - \mu)/\sigma]$ for which the likelihood equations $\partial L/\partial\mu = 0$ and $\partial L/\partial\sigma = 0$ can be written as

$$-\sum_{i=1}^{n} \frac{1}{f[(y_i^0-\mu)/\sigma]} \left.\frac{\partial f}{\partial y}\right|_{\frac{y_i^0-\mu}{\sigma}} = 0 \text{ and } -n + \sum_{i=1}^{n} \frac{(y_i^0-\mu)/\sigma}{f[(y_i^0-\mu)/\sigma]} \left.\frac{\partial f}{\partial y}\right|_{\frac{y_i^0-\mu}{\sigma}} = 0, \tag{15}$$

showing they depend on the y_i^0, μ and σ only through the standardized values $(y_i^0 - \mu)/\sigma$ for $i = 1,2,...,n$.

Let the MLE's obtained from the original sample be $\hat{\mu}^0$ and $\hat{\sigma}^0$ so that they satisfy (15) .

Now suppose that, using bootstrap sampling with $\hat{\mu}^0$ and $\hat{\sigma}^0$ in place of the unknown μ^0, σ^0, we had generated the *same* set of uniforms $\{u_i^0\}$ as in the original sample. The BS y_i^* would take the form $y_i^* = \hat{\mu}^0 + \hat{\sigma}^0 x_i^0$. The MLE's $\hat{\mu}^*$ and $\hat{\sigma}^*$ corresponding to these y_i^* must satisfy (15) only with y_i^0 replaced by y_i^*. Thus the only difference is that, for the original observations, the standardized values are $(y_i^0 - \mu)/\sigma$ whilst they are $(y_i^* - \mu)/\sigma$ in the bootstrap version. However these are identical simultaneously for all i if

$$\hat{\mu}^* = \hat{\mu}^0 + \frac{\hat{\sigma}^0}{\sigma^0}(\hat{\mu}^0 - \mu^0) \text{ and } \hat{\sigma}^* = \frac{\hat{\sigma}_0^2}{\sigma_0},$$

so that this must be the form of the MLE's in the BS version. We therefore have

$$(y_i^0 - \hat{\mu}^0)/\hat{\sigma}^0 = (y_i^* - \hat{\mu}^*)/\hat{\sigma}^*, \ \ i = 1,2,...,n. \tag{16}$$

This equality of the standardized versions holds for any given uniform sample $\{u_i^0\}$. Allowing $\{u_i^0\}$ to vary over its sample space $[0,1]^n$ clearly generates the entire joint distribution of both $(Y_i^0 - \hat{\mu}^0)/\hat{\sigma}^0$, $i = 1,2,..,n$ and $(Y_i^* - \hat{\mu}^*)/\hat{\sigma}^*$, $i = 1,2,...,n$. The equality (16) therefore shows that the joint distribution of the BS standardized quantities is identical to that of the original.

We now apply this result to the A^2 statistic. This depends only on the quantities $Z_i = F(y_{(i)}, \hat{\theta})$. For the location-scale model we have from (16) that $Z_i^* = F((y_{(i)}^* - \hat{\mu}^*)/\hat{\sigma}^*) = F(y_i^0 - \hat{\mu}^0)/\hat{\sigma}^0) = Z_i$ for all i. Thus, for any uniform sample $\{u_i^0\}$, the bootstrap version A^{*2} is identical in value to the A^2 value obtained for the original sample. The bootstrap EDF of a BS sample of A^{*2} values therefore remains an exact estimator of the null distribution of A^2 for any n (> 2) when unknown parameter location and scale parameters are replaced by their MLEs.

The same result holds for the W^2 as this also depends only on the Z_i quantities.

For the general case where F may not be a location-scale model, and when θ^0 is unknown, we need to identify the error that arises from using the BS estimated critical value $A^{*2}_{(1-\alpha)}$ in the GoF test instead of the true value $A^2_{(1-\alpha)}$. As $A^{*2}_{(1-\alpha)}$ is a random quantity based on sampling from $F(\cdot, \hat{\theta})$ we therefore need to estimate its distribution.

The flowchart (14) shows how the critical value $A^{*2}_{(1-\alpha)}$ actually used in the GoF test can be calculated as the $(1-\alpha)$ *quantile* of the EDF formed from a bootstrap random sample, $A^2(\mathbf{y}_j^*, \hat{\theta}_j^*)$ $j = 1,2,...,B$ of

A^2 values; this EDF being a consistent estimator of the desired null distribution. In this calculation each BS sample $\mathbf{y}_j^*$ is drawn from $F(\cdot,\hat{\theta})$ so that $\hat{\theta}$ is the underlying parameter value determining just this *one* value of $A^{*2}_{(1-\alpha)}$. We can obtain an estimate the distribution of $A^{*2}_{(1-\alpha)}$ by adding an *additional level of bootstrapping* in the flowchart to form a *random sample* of critical values, $A^{*2}_{(1-\alpha),j}$ $j=1,2,...,B'$, where each $A^{*2}_{(1-\alpha),j}$ is calculated in the same way as the original $A^{*2}_{(1-\alpha)}$, but using a BS value $\hat{\theta}_j^*$ that varies with j as the underlying parameter value instead of $\hat{\theta}$ to reflect the additional variation in calculating $A^{*2}_{(1-\alpha)}$ from bootstrap samples rather than from the true samples. This calculation can be regarded as an *inner bootstrap* executed at each step j of the now *outer bootstrap* of (14). The flowchart for the calculation is as follows.

Flowchart for BS estimation of the distribution of $A^{*2}_{(1-\alpha)}$, the BS estimate of the critical value when θ is estimated by the MLE $\hat{\theta}$ (17)

Given $\mathbf{y}=(y_1,y_2,...,y_n)$, a random sample drawn from distribution with CDF $F(y,\theta)$, θ unknown

$\downarrow$

$\hat{\theta}=\hat{\theta}(\mathbf{y})$ MLE

$\downarrow$

$m=(1-\alpha)B$ Rounded integer subscript

For $j=1$ to B' Outer bootstrap [typically $B'\le B$]

$\mathbf{y}_j^*=(y_{ij}^*\sim F(y,\hat{\theta}),\ i=1,2,...,n)$ jth BS sample

$\downarrow$

$\hat{\theta}_j^*=\hat{\theta}(\mathbf{y}_j^*)$ jth BS MLE

For $k=1$ to B Inner bootstrap

$F(\cdot,\hat{\theta}_j^*)$

$\downarrow$

$\mathbf{y}_{jk}^{**}=(y_{1jk},y_{2jk},...,y_{njk})$ (j,k)th BS sample with all $y_{ijk}\sim F(\cdot,\hat{\theta}_j^*)$

$\downarrow$

$\hat{\theta}_{jk}^{**}$ MLE of θ, calculated from $\mathbf{y}_{jk}^{**}$

$\downarrow$

$A_{jk}^{**2}=A^2(\mathbf{y}_{jk}^{**},\hat{\theta}_{jk}^{**})$ (j,k)th BS A^2 GoF test statistic

Next k

$A^{**2}_{j,(1)}\le A^{**2}_{j,(2)}\le ...\le A^{**2}_{j,(B)}$ B ordered null-distributed BS A_j^{**2} values

$\downarrow$

$A^{**2}_{(1-\alpha),j}=A^{**2}_{j,(m)}$ jth GoF critical value at level $(1-\alpha)$

Next j

EDF of $A^{**2}_{(1-\alpha),1}$, $A^{**2}_{(1-\alpha),2}$, ..., $A^{**2}_{(1-\alpha),B'}$ is the BS estimate of the distribution of $A^2_{(1-\alpha)}$

A double superscript '**' is used to denote quantities calculated in the inner BS loop.

At the end of the outer BS loop there are B' BS estimated critical values $A^{**2}_{(1-\alpha),j}$ $j=1,2,...,B$, the EDF of which estimates the distribution of $A^2_{(1-\alpha)}$. If a formal assessment of the accuracy of $A^2_{(1-\alpha)}$ is needed this can be provided by calculating a percentile CI from this EDF. We have not included this calculation in the flowchart.

Taken together, the outer and inner bootstraps, are commonly called a *double bootstrap.*

The double bootstrap is a misnomer as the effort required is not $2B$ but $B'B$, so is computationally expensive if B' is large. In practice B' only needs to be large for a formal estimate of the BS error. The

tabulated critical values given in D'Agostino and Stephens (1986) for different distributions suggests that this BS error will not be large in general. For instance, in the case of the gamma distribution, when parameters are estimated, the GoF critical values are affected by the shape parameter a, but the effect is not great with, for instance, the critical values when $\alpha = 0.1$ only varying from 0.657 at $a = 1$ to 0.631 at $a = \infty$. Therefore a less formal more flexible approach to the sensitivity analysis is simply to increase B' by adding replications one at a time, stopping once the magnitude of the error is clear.

We conducted such a double bootstrap experiment for the case $\theta^0 = (a,b) = (1,1)$, with $B = 1000$ and in fact with $B' = 10$, small but fixed. As we know the value of θ^0, and the example is for illustration only, we generated each initial $\mathbf{y}$ directly from $F(\cdot,\theta^0)$ rather than $F(\cdot,\hat{\theta})$. The $B' = 10$ BS EDFs of A^{*2} are depicted in Figure 4.

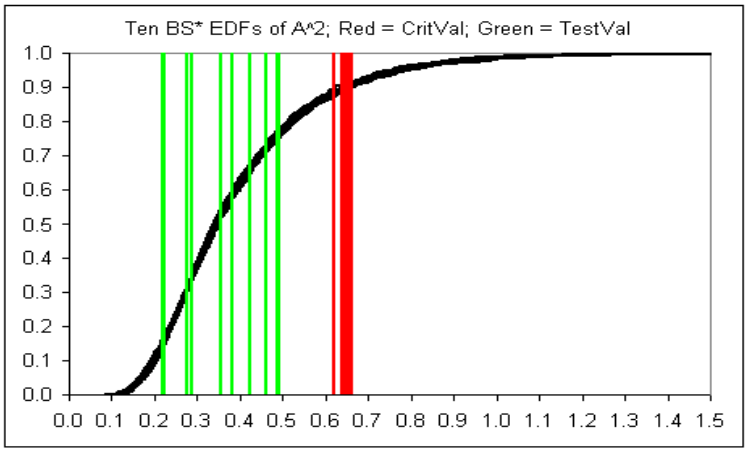

Figure 4: Bootstrap EDFs from ten runs of the BS GoF test set out in the flowchart (3) Also shown are the $\alpha = 0.1$ level critical values obtained from the EDFs and the A^{*2} values corresponding to each $\mathbf{Y}$ and MLE $\hat{\theta}$ calculated from $\mathbf{Y}$ as indicated in the square brackets in the flowchart.

The ten critical $A^{*2}_{(1-\alpha)=0.9}$ values corresponding to the $(1-\alpha) = 0.9$ critical value calculated from the EDFs are shown in red. These values of $A^2_{(1-\alpha)=0.9}$ fell in the range $0.629 - 0.660$, straddling the value of 0.657 tabulated in D'Agostino and Stephens (1986, Table 4.21). Also, as an added indication that there is nothing untoward about the replications, the Figure also shows, in green, the ten 'test' $A^2(\mathbf{Y}, \hat{\theta}(\mathbf{Y}))$ values calculated from the initial sample $\mathbf{Y}$ and MLE $\hat{\theta}(\mathbf{Y})$, this $\hat{\theta}$ being used to generate the BS samples in the replication.

For the tollbooth data $\hat{a} = 9.20$ and $\hat{b} = 0.63$ for which the BS estimated $(1-\alpha) = 0.9$ critical value using $B = 1000$, was $A^2 = 0.631$. This is greater than the test value of $A^2 = 0.497$, so that the null hypothesis that the data are gamma distributed is not rejected. In comparision the tabulated critical values in D'Agostino and Stephens (1986, Table 4.21) are 0.634 at $\hat{a} = 8$ and 0.633 at $\hat{a} = 10$.

In addition we tested fitting the normal distribution. The MLEs were $\hat{\mu} = 5.80$ and $\hat{\sigma} = 2.04$, with a test value of $A^2 = 1.109$. Bootstrapping using $B = 1000$, gave a critical test value at the $(1-\alpha) = 0.9$ level of $A^2 = 0.620$. The test value is greater, so we would reject the hypothesis that the data is normally distributed. The D'Agostino and Stephens (1986, Table 4.7) tabulated critical value at $\alpha = 0.1$ is $A^2 = 0.631$, but to apply the test with this critical value, the test value has to be inflated by a factor of $(1+0.75/n+2.25/n^2) = 1.017$ at $n = 47$.This is equivalent to modifying the critical value to $0.631/1.017 = 0.620$ if the test value is retained at $A^2 = 1.109$, making it the same as the BS version.

5 SUMMARY

We have shown that parametric bootstrapping provides a very simple way of assessing the quality of parametric models fitted to input and output data in simulation experiments. In particular we have (i) given details of how to construct bootstrap confidence intervals (CIs) for parameters of interest and confidence bands (CBs) for output functions of interest, and (ii) described how to carry out an Anderson-Darling GoF test, using bootstrapping to calculate GoF critical values.

We have shown that bootstrap CIs and the Anderson-Darling GoF test are probabilistically exact for location-scale models.

REFERENCES

Anderson, T.W. and Darling, D.A. 1952. "Asymptotic Theory of Certain "Goodness of Fit" Criteria Based on Stochastic Processes". *The Annals of Mathematical Statistics*, 23:193-212.

Babu, G.J. and Rao, C.R. 2004. "Goodness-of-fit Tests when Parameters are Estimated". *Sankhyā* 66:63-74.

Cheng, R.C.H. and Iles, T.C. 1983. "Confidence Bands for Cumulative Distribution Functions of Continuous Random Variables". *Technometrics* 25:77-86.

Cheng, R.C.H. 1987. "Confidence Bands for Two-stage Design Problems". *Technometrics* 29:301-309.

Chernick, M.R. 1999. *Bootstrap Methods, A Practitioner's Guide.* New York: Wiley.

Cox, D.R. and Hinkley, D.V. 1974. *Theoretical Statistics.* London: Chapman and Hall.

D'Agostino, R.B. and Stephens, M.A. 1986. *Goodness-of-Fit Techniques.* New York: Marcel Dekker, Inc.

Davison, A.C. and Hinkley, D.V. 1997. *Bootstrap Methods and their Application.* Cambridge: Cambridge University Press.

Griffiths, J.D. and Williams J.E. 1984. "Traffic Studies on the Severn Bridge". *Traffic Engineering and Control* 25:268-71, 274.

Hall, P. 1987. "On the Bootstrap and Likelihood-based Confidence Regions". *Biometrika* 74:481-93.

Hjorth, U. 1994. *Computer Intensive Statistical Methods.* London: Chapman & Hall.

Miller, R.G. Jr 1981. *Simultaneous Statistical Inference, 2nd Ed.* New York: Springer-Varlag.

Stephens, M.A. 1970. "Use of the Kolmogorov-Smirnov, Cramér-von Mises and Related Statistics Without Extensive Tables". *Journal of the Royal Statistical Society, Series B* 32:115-122.

Stephens, M.A. 1974. "EDF Statistics for Goodness-of-fit and Some Comparisons". *Journal of the American Statistical Association*, 69:730-737.

Stute, W., Mantega, W.G. and Quindimil, M.P. 1993. "Bootstrap Based Goodness-of-fit-Tests". *Metrika* 40:243-256.

Young, G.A. and Smith, R.L. 2005. *Essentials of Statistical Inference.* Cambridge: Cambridge University Press.

AUTHOR BIOGRAPHY

RUSSELL C. H. CHENG is Emeritus Professor of Operational Research at the University of Southampton. He has an M.A. and the Diploma in Mathematical Statistics from Cambridge University, England. He obtained his Ph.D. from Bath University. He is a former Chairman of the U.K. Simulation Society, a Fellow of the Royal Statistical Society and Fellow of the Institute of Mathematics and Its Applications. His research interests include: design and analysis of simulation experiments and parametric estimation methods. He was a Joint Editor of the IMA Journal of Management Mathematics. His email address is cheng@btinternet.com.