

COMPARISON OF SIMULATION OUTPUT SERIES USING BOOTSTRAPPING

Christine S.M. Currie

School of Mathematics
 University of Southampton
 Southampton, SO17 1BJ, U.K.

Lanting Lu

School of Mathematics
 University of Southampton
 Southampton, SO17 1BJ, U.K.

ABSTRACT

We describe a method for comparing stochastic outputs of simulation models. The method is distribution-free and allows the comparison of sets of data with different numbers of data points. This makes it ideal for performing comparisons between simulation output and the real output of the system being modelled, when often there are many more data points available from the output of the simulation model than present in the real data. We calculate the two-sample Cramér-von Mises goodness-of-fit statistic between the two sets of data, using bootstrapping to find the distribution of the statistic, and so the probability that the two sets of data were drawn from the same distribution.

1 INTRODUCTION

We describe a method for comparing stochastic outputs of simulation models. The Cramér-von-Mises goodness-of-fit statistic is used to assess the similarity of the two data sets, with bootstrapping being used to estimate the significance level of the calculated statistic. Being distribution-free and allowing for different numbers of data points in the two data sets, this method is ideally suited to checking that the output of a simulation model matches any available real system output data.

Bootstrapping or resampling is well-known in the statistical literature but its use is less common in simulation, despite the ease with which it can be implemented. A thorough introduction is given by (Efron and Tibshirani 1994) and a description of its use in both input and output modeling for simulation is given in (Cheng 2006), with some interesting practical examples. We concentrate here on comparing series of output data, where the distribution of the data is unknown.

We describe the method in Section 2, before discussing its implementation on a real example from manufacturing in Section 3, where we wish to compare three different input models. Finally, we conclude in Section 4.

2 METHODS

2.1 Calculating the Goodness-of-Fit

We wish to measure the similarity of two samples of output data (x_1, x_2, \dots, x_n) , and (y_1, y_2, \dots, y_m) , and so obtain a measure of their similarity. In order to avoid making assumptions about the underlying distribution of the data we use the Cramér-von-Mises goodness-of-fit statistic as our measure. We did consider using the Anderson-Darling statistic (Stephens 1974) but this requires some information about the hypothesized distribution in order to calculate the goodness-of-fit. The Cramér-von Mises T criterion for testing that the two samples come from the same unspecified continuous distribution is

$$T = (nm/(n+m)) \int_{-\infty}^{\infty} (F_n(x) - G_m(x))^2 dH_{n+m}(x), \quad (1)$$

where $F_n(x)$ is the empirical distribution function (EDF) of the first sample; that is, $F_n(x) = (\text{no. of } x_i \leq x)/n$; $G_m(x)$ is the EDF of the second sample and $H_{n+m}(x)$ is the EDF of the two samples together; that is, $(n+m)H_{n+m}(x) = nF_n(x) + mG_m(x)$.

Let r_i and s_j be the ranks in the pooled sample of the ordered observations of the two samples X and Y , respectively, where $i = 1, 2, \dots, n$ and $j = 1, 2, \dots, m$. Then

$$F_n(x) - G_m(x) = i/n - (r_i - i)/m \quad (2)$$

at the i th x -observation and

$$F_n(x) - G_m(x) = (s_j - j)/n - j/m \quad (3)$$

at the j th y -observation. Thus we can write the criterion T as

$$T = \frac{U}{nm(n+m)} - \frac{4nm-1}{6(n+m)}, \quad (4)$$

where

$$U = n \sum_{i=1}^n (r_i - i)^2 + m \sum_{j=1}^m (s_j - j)^2. \quad (5)$$

To test the null hypothesis that the two samples are drawn from the same distribution, all of the observations are ordered, the ranks $r_1 < r_2 < \dots < r_n$ of the n observations from the first sample and the ranks $s_1 < s_2 < \dots < s_m$ of the m observations from the second sample are then determined and U is computed. If U is too large, we reject the null hypothesis, that the samples are drawn from the same distribution.

Tabulated criterion values are not very extensive and do not cover the samples that we are dealing with and so we use bootstrapping to determine the distribution of T , $\Phi(T)$, and hence the significance level. This is a well-known application of bootstrapping, as described very clearly by Cheng.

2.2 Bootstrapping

To carry out the bootstrapping, let $Z = (z_1, z_2, \dots, z_{n+m})$ be the pooled sample of breakdown data from machines X and Y . The EDF of Z is denoted by $H_{n+m}(z)$. We generate two samples of size n and m from the original pooled set of observations, Z , with replacement and call this our *bootstrap sample*, written as $X^* = (x_1^*, x_2^*, \dots, x_n^*)$ and $Y^* = (y_1^*, y_2^*, \dots, y_m^*)$. By comparing X^* and Y^* , we can calculate the Cramér-von Mises statistic T^* for the bootstrap sample. In order to estimate $\Phi(T)$, we generate a number, B , pairs of bootstrap samples from $Z : (X^{*1}, Y^{*1}), (X^{*2}, Y^{*2}), \dots, (X^{*B}, Y^{*B})$ and calculate the statistic T^{*j} for each pair of samples. The EDF of the sample $T^* = (T^{*1}, T^{*2}, \dots, T^{*B})$ is then written as

$$\Phi_B(T) = \frac{\text{(no. of } T^{*j} \leq T)}{B} \quad (6)$$

The bootstrap distribution $\Phi_B(T)$ will converge to $\Phi(T)$ with probability one as B tends to infinity (Cheng 2006) and we use $\Phi_B(T)$ as our estimate of $\Phi(T)$. (We consider $B = 5000$ as a large enough number for our bootstrap analysis.)

The Bootstrapping Process is

For $j = 1$ to B

For $i = 1$ to n

Draw x_i^{*j} from Z (with replacement)

Next i

For $i = 1$ to m

Draw y_i^{*j} from Z (with replacement)

Next i

Calculate T^{*j} by comparing X^{*j} with Y^{*j}

Next j

Form the EDF of T^* , $\Phi_B(T)$.

The p-value obtained at the end of the bootstrap analysis can be interpreted as the probability that the two sets of

output data are drawn from the same distribution, and so gives a good indication of the similarity of the two data sets. The whole process is illustrated in Figure 1.

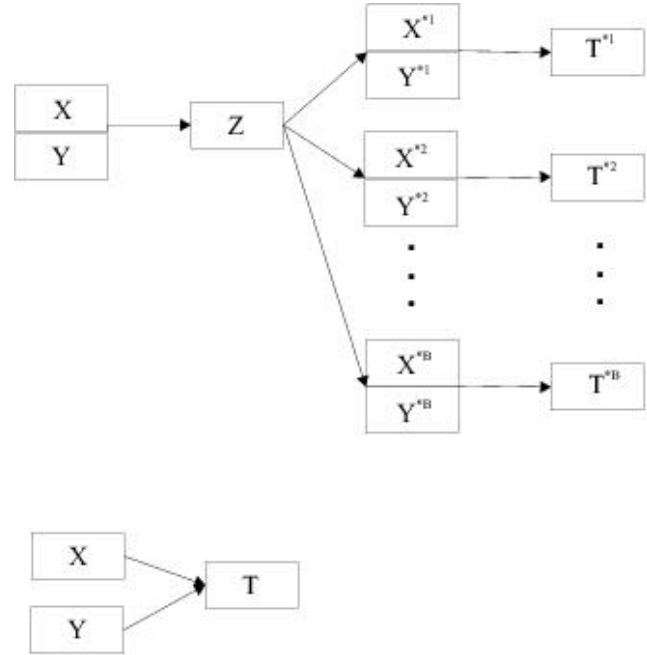


Figure 1: (a) The bootstrapping process used to determine the null distribution of T , $\Phi(T)$, and (b) the evaluation of the Cramér-von-Mises statistic for the original samples, which is compared with $\Phi(T)$ to determine the p-value for the similarity of the two sets of data

3 EXAMPLE

We consider a simulation model of an engine assembly line, where the output of interest is the throughput of the line or the number of jobs completed per hour (JPH). The aim of the exercise is to compare three different methods for generating machine breakdown durations within the simulation model:

1. Sampling from historical data on machine breakdown durations;
2. Sampling from finite mixture distributions fitted to each individual machine's breakdown duration data;
3. Sampling from finite mixture distributions fitted to breakdown duration data from groups of machines.

More information on the modeling of machine breakdown durations can be found in (Lu, Currie, Cheng, and Ladbrook 2007). Currently, the company use method 1; method 2 provides a very accurate description of the data but fitting the finite mixture models is computationally intensive; method 3

is preferred as it reduces the time spent fitting finite mixture distributions. The Arrows classification method is used to group the machines in method 3, based on the similarity of their breakdown duration data (Lu, Currie, Cheng, and Ladbrook 2007). See (Cheng and Currie 2003) for more information on finite mixture distributions and the method we use for fitting them.

3.1 Results

After a warm up period, we make a single run of 36 weeks for each of the three different methods. We make 36 observations in each run, each observation being the averaged number of jobs shipped per hour (JPH) in each of the 36 weeks. Thus, we obtain 36 averaged JPH observations for each model. Figure 2 is a boxplot of the simulation output for the three methods and suggests there is a high degree of similarity between outputs.

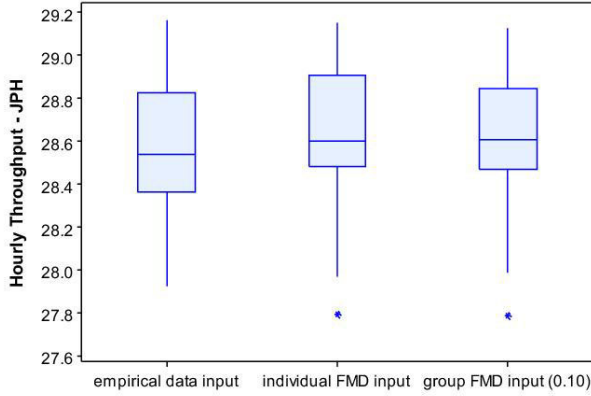


Figure 2: Boxplot of simulation output JPH using the three methods for sampling breakdown durations. The central line shows the median and the box spans the inter-quartile range.

The bootstrapping comparison we describe in this paper allows us to examine the similarities between the underlying distributions of the JPH outputs of the models using the three different breakdown duration inputs, where the similarities are measured by the possibilities that any two sets of the JPH observations have been drawn from the same distribution. The larger the possibility, the more similar the two sets of JPH outputs and thus the more similar the two breakdown duration inputs. We perform a pairwise comparison for the three methods and the resultant p-values are given in Table 1. As shown in this table, the p-values are all quite high, which indicates that the distributions of the JPH outputs of the three simulation models using different breakdown inputs are all very similar to each other and thus suggests the three representations of the breakdown durations as simulation

input have a similar effect on the whole system’s production performance.

Table 1: The p-values obtained from the bootstrapping process of comparison between the outputs of models using the three breakdown duration inputs.

Comparison	p-Value
Empirical data input vs. individual FMD input	0.693
Empirical data input vs. group FMD input	0.459
individual FMD input vs. group FMD input	0.746

As a further check of our method we use a paired t-test for testing the mean difference between paired observations of the JPH outputs using the three different input methods, as recommended by Law and Kelton (Law and Kelton 2000). The null hypothesis is

$$H_0 : \mu_d = 0,$$

where μ_d is the population mean of the differences between pairs of observations.

The results of the paired t-tests are given in Table 2. The confidence intervals for the mean difference between any two output processes of the model using any two breakdown duration inputs all include zero, which suggests there is no obvious difference between any two of the simulation outputs. The fairly high p-values further suggest that the data are consistent with $H_0 : \mu_d = 0$. It can be seen that the difference between the outputs of the model using historical data and the model using individual FMD is larger than the difference between the outputs of the model using historical data and the model using group FMD.

Table 2: The results of the paired t-tests between the outputs of models using the three breakdown duration inputs.

Paired T-Test	95% CI for	
	Mean Differences	p-Value
Empirical data input vs. individual FMD input	(-0.0922, 0.0061)	0.084
Empirical data input vs. group FMD input(0.10)	(-0.0811, 0.0214)	0.245
individual FMD input vs. group FMD input	(-0.0054, 0.0318)	0.159

4 DISCUSSION

We have described a method for estimating the similarity of different sets of output data. By using the Cramér-von-Mises statistic for estimating the goodness of fit of the two sets of

data, we need make no assumptions about the underlying distribution of the data. This is not true of other comparison statistics; for example, the paired t-test assumes the data are normally distributed. In addition, the data sets being compared can be of different sizes. This makes it versatile and perhaps particularly useful for comparing differences between the output of the real system and outputs of a simulation model, where typically more data points are available from the simulation.

Using bootstrapping allows us to calculate the distribution of the Cramér-von-Mises test statistic numerically, hence obtaining the critical value for the comparison. As the tabulated critical values for the test are not extensive, bootstrapping is particularly important for this approach.

The results of the example suggest that the outputs produced by each of the three methods for modeling machine breakdown durations in the simulation model are statistically similar.

REFERENCES

- Cheng, R. 2006. Validating and comparing simulation models using resampling. *Journal of Simulation* 1:53–63.
- Cheng, R. C. H., and C. S. M. Currie. 2003. Prior and candidate models in the Bayesian analysis of finite mixtures. In *Proceedings of the 2003 Winter Simulation Conference*, ed. S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, 392–398. IEEE.
- Efron, B., and R. J. Tibshirani. 1994. *An introduction to the bootstrap*. Boca Raton: FL: CRC Press.
- Law, A. M., and W. D. Kelton. 2000. *Simulation Modeling and Analysis*. 3rd ed. McGraw-Hill series in industrial engineering and management science. New York: McGraw-Hill.
- Lu, L., C. S. M. Currie, R. C. H. Cheng, and J. Ladbrook. 2007. Classification analysis for simulation of machine breakdowns. In *Proceedings of the 2007 Winter Simulation Conference*, 480–487. Piscataway, NJ, USA: IEEE Press.
- Stephens, M. A. 1974. EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association* 69:730–737.

AUTHOR BIOGRAPHIES

CHRISTINE CURRIE is a lecturer in Operational Research at the University of Southampton. Her e-mail address is christine.currie@soton.ac.uk.

LANTING LU is a PhD student in Operational Research at the University of Southampton. Her e-mail address is l.lu@soton.ac.uk.