J. Statist. Comput. Simul., Vol. 30, pp. 129–143
Reprints available directly from the publisher
Photocopying permitted by license only
© 1988 Gordon and Breach, Science Publishers, Inc.
Printed in Great Britain

Confidence Interval Scatterplots for Evaluating Confidence Interval Performance

MURALI SUBRAMANIAM
Lehigh University, Bethlehem, PA 18015, USA

and

LAWRENCE LEEMIS
School of Industrial Engineering, The University of Oklahoma, Norman, OK 73019, USA

(*Received March 1987*)

Confidence intervals are widely used to determine an interval, estimated from data, that contains a population parameter with a prescribed probability. Kang and Schmeiser (1986) introduce a scatterplot that is useful in comparing the effectiveness of different confidence interval procedures based on performance measures such as the coverage, the expected interval width and the variance of the interval width. This paper examines the usefulness of this confidence interval scatterplot (CIS) for evaluating approximate confidence intervals, as illustrated by the comparison of simulation output analysis techniques for correlated observations.

KEY WORDS: Censoring, halfwidth, Monte Carlo simulation, simulation output analysis.

1. INTRODUCTION

Confidence intervals are used to provide a range of values within which a population parameter of interest will lie with a prescribed probability. A confidence interval contains more information than a point estimate for a parameter as it provides information about the variability of the estimator, in addition to a point estimate. The

purpose of this paper is to illustrate the confidence interval scatterplot (CIS), developed by Kang and Schmeiser (1986), that is useful in evaluating approximate confidence intervals and in comparing alternative confidence interval procedures. Section 2 contains a description of the CIS and associated properties. Section 3 contains an assessment of several output analysis techniques used in discrete event simulation using the CIS, Section 4 applies the CIS to a queuing model and Section 5 contains conclusions.

2. CONFIDENCE INTERVAL SCATTERPLOTS

A confidence interval for a population parameter θ is given by

$$L \leqq \theta \leqq U$$

where L and U are functions of n, the sample size, $\mathbf{X}$, a random vector of n data values and $1-\alpha$, the nominal (or stated) coverage of the interval. The true value of the parameter θ is denoted by θ_0.

One simple diagram for communicating the concept of a confidence interval is shown in Figure 1, where seven confidence intervals are plotted on axes with the same units as θ. The confidence intervals numbered 4 and 6 do not contain the true parameter value θ_0. Confidence interval number 4 "misses low" and confidence interval number 6 "misses high". It is also apparent from Figure 1 that there is substantial variability in the widths of these confidence intervals.

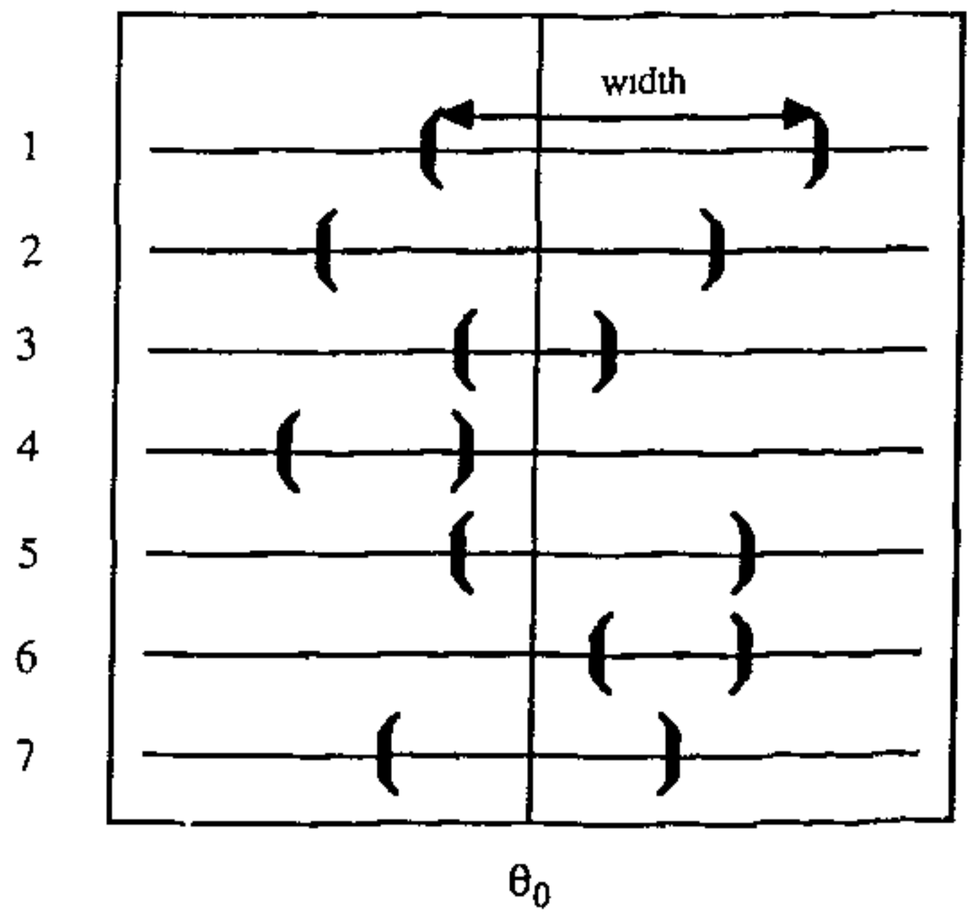


Figure 1 Confidence Intervals for the population parameter θ.

A simple example of an *exact* $(1-\alpha)*100\%$ confidence interval for the population parameter μ when $X_1, X_2, \ldots, X_n$ are independent and identically distributed normal random variables with mean μ and variance σ^2 is

$$L = \bar{X} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}} \quad \text{and} \quad U = \bar{X} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

where $t_{\alpha/2,n-1}$ is the value of the $1-\alpha/2$ percentile of a t random variable with $n-1$ degrees of freedom,

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \text{and} \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Many of the confidence intervals currently used are *approximate*, rather than exact, confidence intervals. A confidence interval is approximate if

$$\lim_{m\to\infty}\sum_{i=1}^{m}\frac{I[L_i, U_i]}{m} \neq 1-\alpha$$

for any sample size n, and I is an indicator function that assumes the value 1 if $L_i \leqq \theta_0 \leqq U_i$ and 0 otherwise for $i = 1, 2, \ldots, m$. Alternatively, a confidence interval is approximate if the actual coverage does not equal the stated coverage for all n. Relationships between the parameter θ, the stated coverage $1-\alpha$ and the actual coverage are given by Kang and Schmeiser (1986). Many confidence intervals are based on large sample methods and their coverage converges to the nominal coverage as $n\to\infty$, i.e. they are asymptotically exact. Since fewer exact confidence intervals are available due to lack of tractability, evaluating approximate confidence intervals becomes more important.

Another reason for a technique to evaluate the effectiveness of a confidence interval procedure is that a modeler is often faced with several different confidence interval procedures for one data set (e.g., simulation output analysis uses several techniques to evaluate correlated output). In this case, if the actual coverages of the procedures are close to their stated coverage, then other factors should be

considered in selecting a confidence interval. Two of these factors are the expected interval width and the variance of the interval width, where the interval width is $W = U - L$. Assuming two confidence interval procedures have the same coverage, the one with the smaller expected width or width variance should be chosen.

Level surfaces of the joint density of L and U are illustrated by the concentric curves in Figure 2. All confidence intervals must lie above the line $L = U$ since the lower limit is always less than the upper limit. The value of the true population parameter θ_0 is indicated on both axes, and any confidence interval that lies to the northwest of the point (θ_0, θ_0) contains the true parameter θ_0 since $L \leqq \theta_0 \leqq U$. Thus the probability of covering θ_0 is

$$P[L \leqq \theta \leqq U] = \int_{-\infty}^{\theta_0} \int_{\theta_0}^{\infty} f(l, u)\, du\, dl$$

and a confidence interval procedure is exact if this probability is equal to the stated coverage for all n. The mean interval width is

$$E[W] = \int_{-\infty}^{\infty} \int_{l}^{\infty} (u - l) f(l, u)\, du\, dl$$

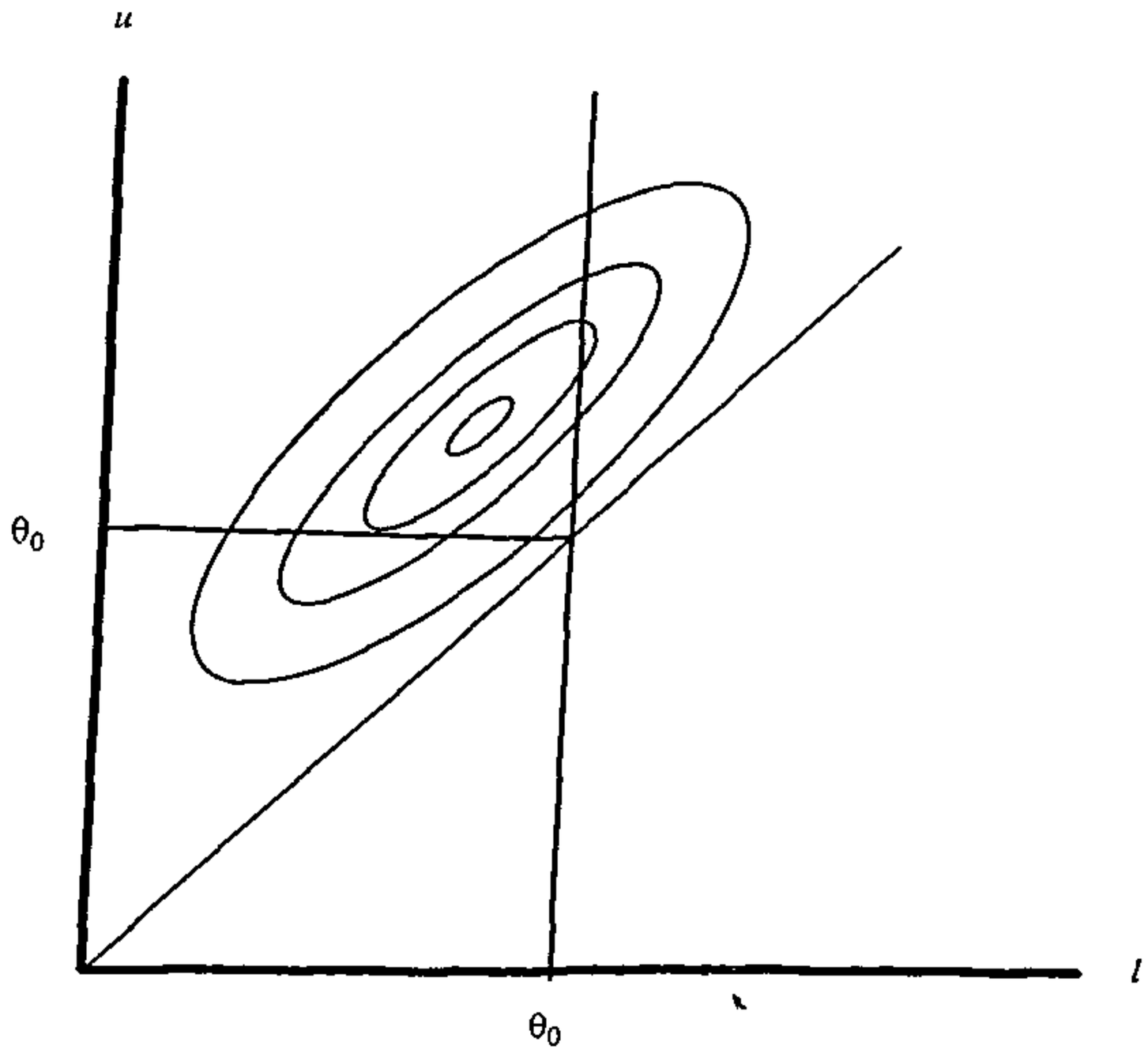


Figure 2 Joint density of l and u.

and the variance of the interval width is

$$V[W] = \int_{-\infty}^{\infty} \int_{l}^{\infty} [(u-l) - E[W]]^2 f(l,u)\, du\, dl.$$

A CIS is a plot of many confidence intervals, (l,u), on a plane. Points falling to the right of $l=\theta_0$ are confidence intervals "missing high", and points falling below $u=\theta_0$ are confidence intervals "missing low". Figure 3 is a CIS of two hundred 90% confidence intervals generated from data sets of size $n=100$ independent and identically distributed N(2,1) observations. Two variations of this CIS follow in Figures 4 and 5 using the same 200 confidence intervals.

Figure 4 is the previous figure rotated 45 degrees clockwise, yielding a graph with the midpoint of the interval (i.e., $(l+u)/2$) as the horizontal axis, and the halfwidth as the vertical axis (i.e., $(u-l)/2$). The "V" on the graph has its vertex at $(\theta_0, 0)$ and falling within the "V" indicates that a confidence interval contains the true parameter value, falling to the left of the "V" indicates that both l and u are less than the true parameter value (i.e., the interval

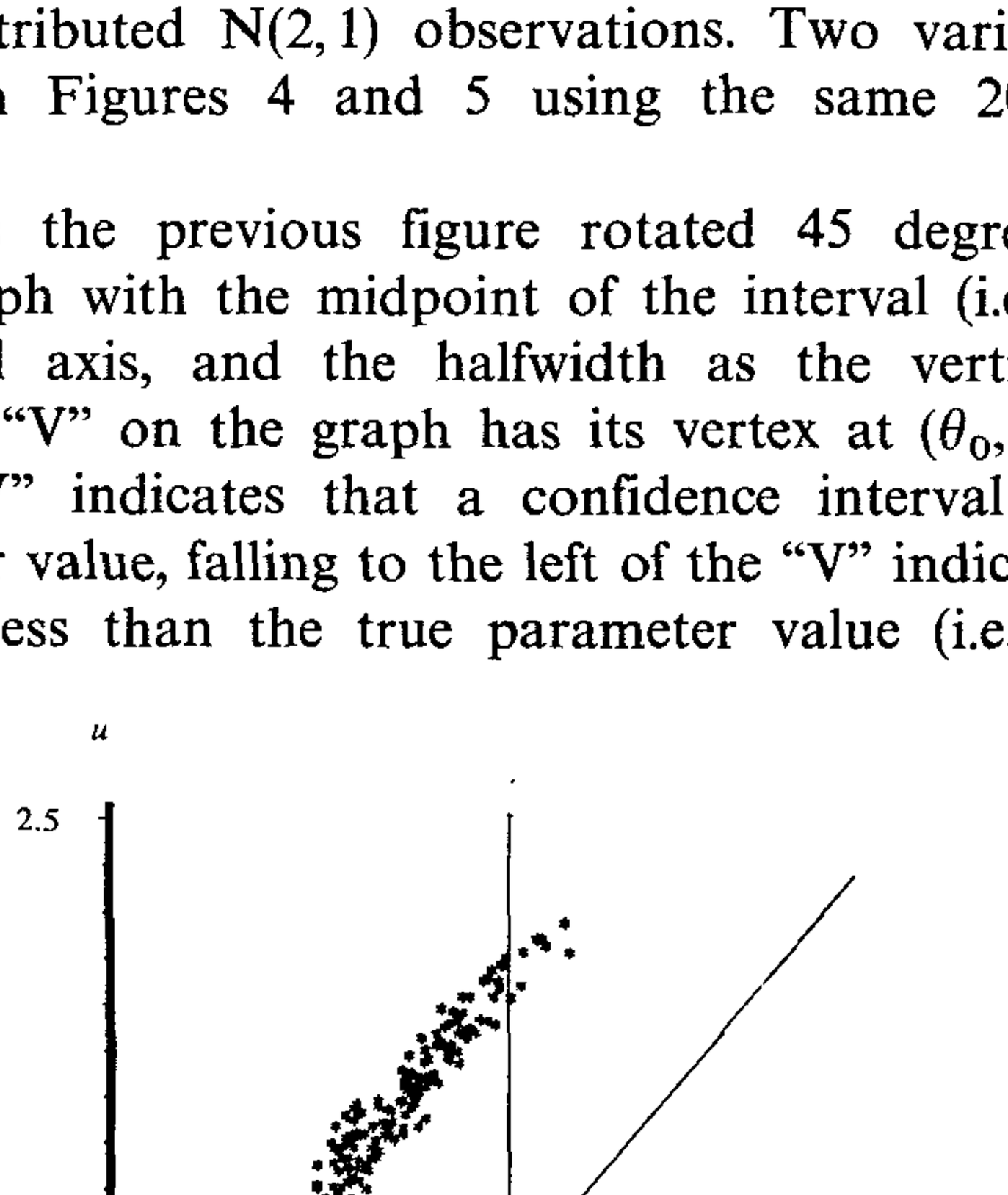


Figure 3 Confidence Interval Scatterplot of 200 (l,u) pairs.

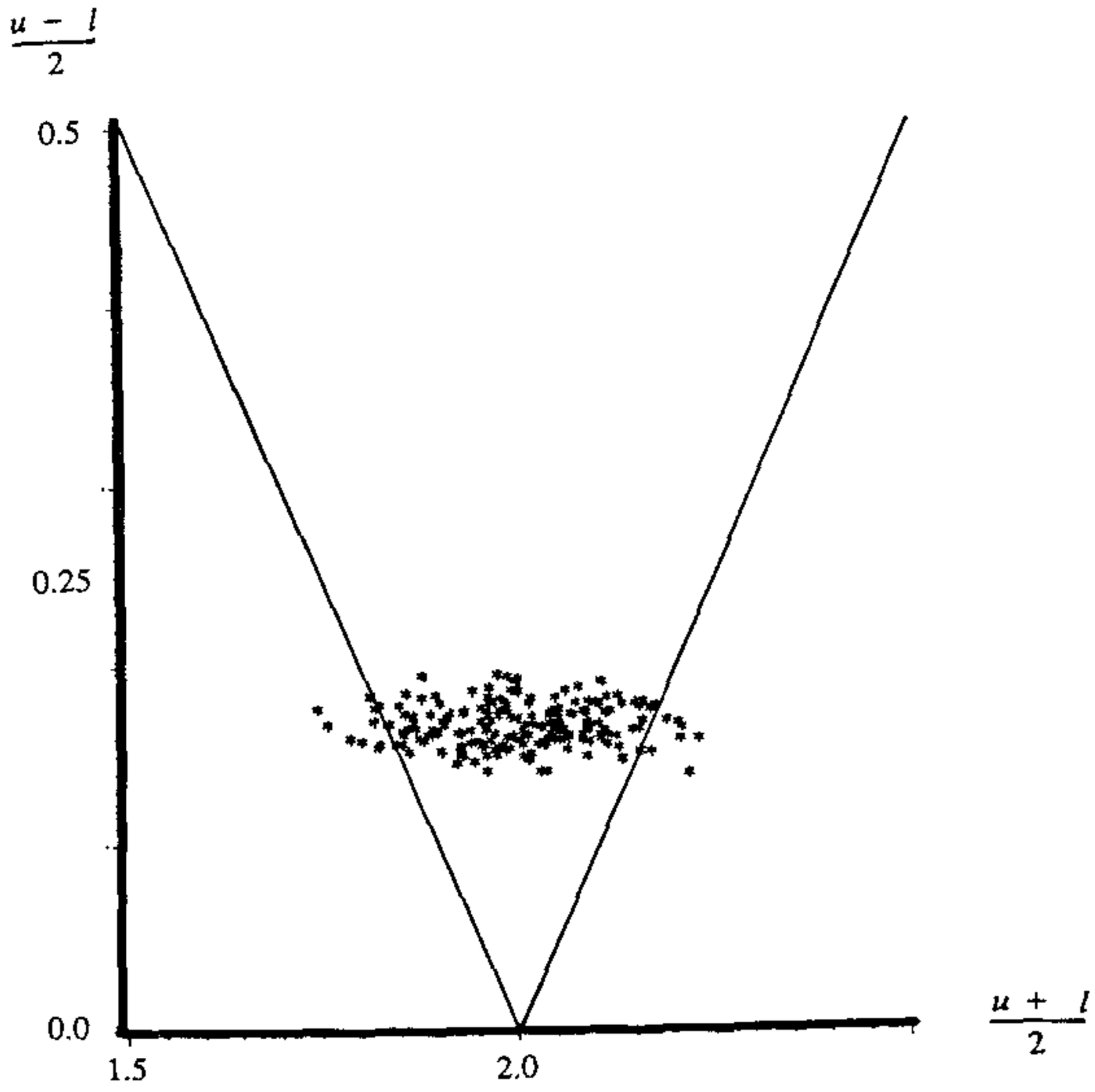


Figure 4 Transformed CIS (midpoint and halfway as axis).

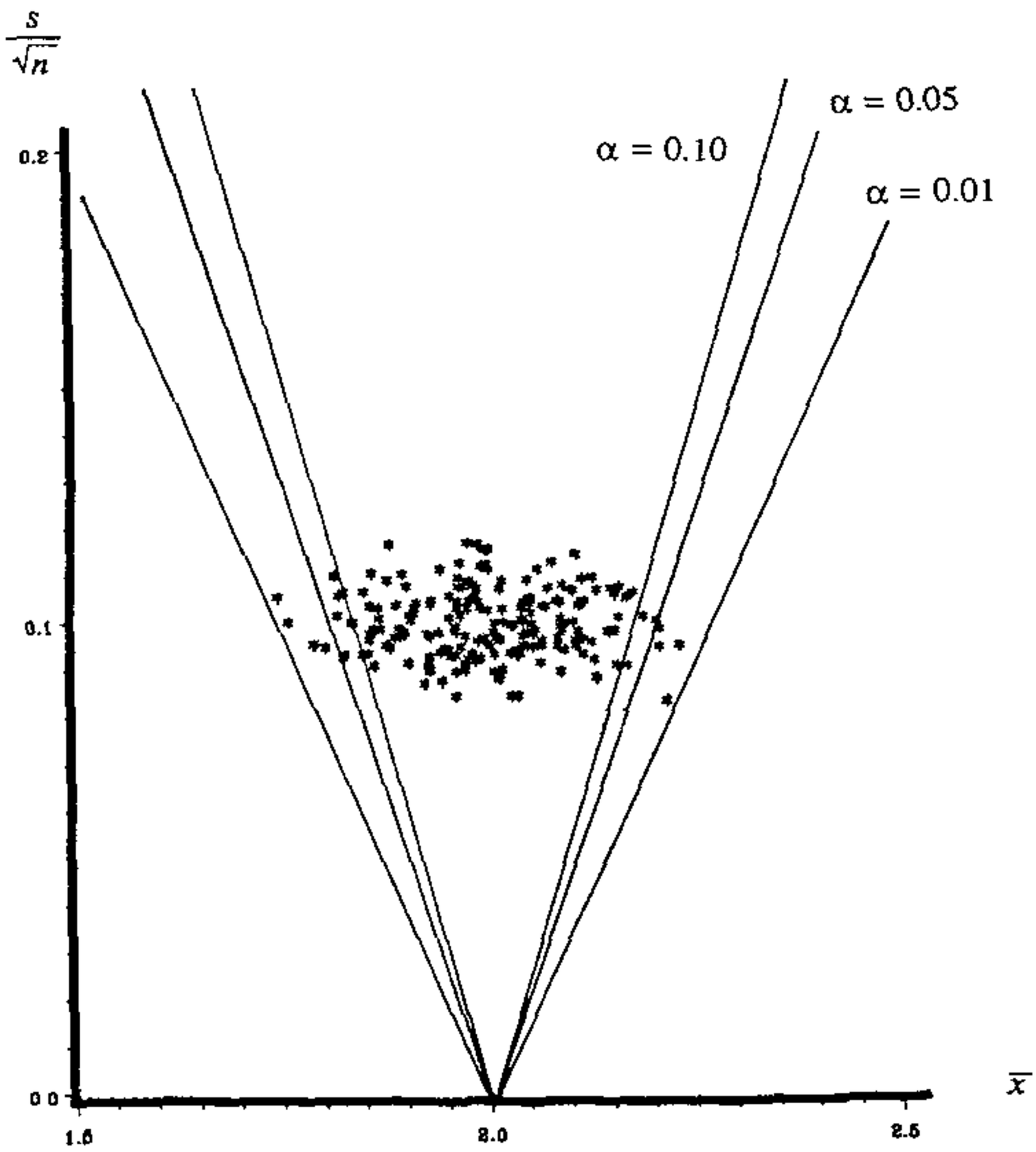


Figure 5 CIS Sensitivity Plots.

"missed low") and falling to the right of the "V" indicates that both l and u are greater than the true parameter value (i.e., the interval "missed high"). If the upper and lower limits of many confidence intervals are plotted on a CIS, the bias of the confidence interval procedure is apparent (since the midpoint may not be centered around the true parameter value), the actual coverage of the confidence interval procedure is apparent (by the fraction of points falling in the "V"), the expected halfwidth is apparent (by the average height of the points on the graph) and the variance of the halfwidth is apparent (by the dispersion of the heights of the plotted values). The effect of changing α for this type of plot is to shift the points up (for decreased α) or down (for increased α). If x is the horizontal axis and y is the vertical axis, then the equation of the lines on the CIS are

$$y = |x - \mu_0|$$

A slight modification to this plot allows the plotting of several α values on a single axis. The vertical axis of this plot can be changed to $s/\sqrt{n}$ instead of the halfwidth. As shown in Figure 5 this plot has the advantage of having several confidence levels on a single axis. Using this plot, the change in the coverage can be observed for a change in the "V" angle, giving the sensitivity of a confidence interval procedure with respect to the stated coverage. The lines in this plot are

$$y = t_{\alpha/2}^{-1} |x - \mu_0|$$

3. APPLICATION: SIMULATION OUTPUT ANALYSIS

Analyzing the output of a discrete event simulation is usually complicated by the presence of correlated observations (e.g., customer wait times in a queue). Most of the general purpose simulation languages use classical procedures to construct confidence intervals for an output measure of interest. Law and Kelton (1982) cite an example where ignoring autocorrelation in the customer delay times in a M/M/1 queue with traffic intensity 0.9 results in an estimate of

$V[\bar{X}]$ that is 0.34% of the true value, and confidence intervals that have an actual coverage far less than the stated coverage.

In this section, three confidence interval procedures are compared: classical, standardized time series (STS), and batch means. These three and other procedures for constructing a confidence interval for an output measure of interest are described in Law (1983). It is assumed that the interest is in the steady state, rather than the transient, behavior of the measure of interest and only one fixed length replication is performed.

Assuming that the queue delay times from the simulation experiment are $X_1, X_2, \ldots, X_n$, the classical procedure, described in Section 2, yields this $(1-\alpha)*100\%$ confidence interval for μ

$$L = \bar{X} - t_{\alpha/2,n-1}\frac{S}{\sqrt{n}} \quad \text{and} \quad U = \bar{X} + t_{\alpha/2,n-1}\frac{S}{\sqrt{n}}.$$

The STS and batch means procedures require the modeler to divide the observations into b batches consisting of k observations each. Both procedures assume independence between batches. The STS procedure (Schruben, 1983) yields the confidence interval

$$L = \bar{\bar{X}} - t_{b,\alpha/2}\sqrt{\frac{A}{nb}} \quad \text{and} \quad U = \bar{\bar{X}} + t_{b,\alpha/2}\sqrt{\frac{A}{nb}}$$

where $\bar{\bar{X}} = (1/bk)\sum_{i=1}^{n} X_i$ is the grand sample mean, $\bar{X}_j = (1/k)\sum_{i=1}^{k} X_{i+(j-1)k}$ is the jth batch mean for $j = 1, 2, \ldots, b$, $t_{b,\alpha/2}$ is the right hand critical value for the t distribution with b degrees of freedom, and

$$A = \frac{12}{k^3 - k}\sum_{j=1}^{b}\left\{\sum_{l=1}^{k}\sum_{i=1}^{l}\left[\bar{X}_j - X_{i+(j-1)k}\right]\right\}^2$$

The batch means procedure yields the confidence interval

$$L = \bar{\bar{X}} - t_{\alpha/2,b-1}\frac{S_{\bar{X}}}{\sqrt{b}} \quad \text{and} \quad U = \bar{\bar{X}} + t_{\alpha/2,b-1}\frac{S_{\bar{X}}}{\sqrt{b}}$$

where

$$s_{\bar{X}}^2 = \frac{1}{b-1} \sum_{j=1}^{b} (\bar{X}_j - \bar{\bar{X}})^2$$

is the standard deviation of the b batch means.

A simulation experiment was conducted with each of these three confidence interval procedures having input (i.e., $X_1, X_2, \ldots, X_n$) from one of three sources: independent and identically distributed normal observations, a MA(1) model and an AR(1) model. In all the experiments, $n = 100$ observations were used to compute each confidence interval, 1000 confidence intervals (at $\alpha = 0.10$) were generated, and the batch size was assumed to be $k = 10$ for the STS and batch means procedures. Standard Gaussian white noise, Z_t, drove the three time series models:

$$X_t = Z_t \qquad (i.i.d.)$$

$$X_t = Z_{t-1} + \tfrac{1}{2} Z_t \qquad (MA(1))$$

$$X_t = X_{t-1} + 0.856 Z_t \qquad (AR(1))$$

for $t = 1, 2, \ldots, n$, Z_0 is standard normal and the AR(1) process is warmed up. The parameter 0.856 in the AR(1) model was chosen so that the autocorrelation function for a realization is approximately that of a M/M/1 queue with traffic intensity of 0.9.

Figure 6 contains CISs of the three confidence interval procedures applied to the three types of inputs. The confidence interval midpoint ($\bar{x}$) is plotted on the horizontal axis and halfwidth is plotted on the vertical axis. The performance measures of interest that are apparent from these plots are the actual coverage, expected halfwidth and variance of the halfwidth. The axes on all of the plots have the same scaling, indicating the wide differences between the three procedures for correlated observations.

One common feature of all 9 CISs is that they are all symmetric about the true mean of the process. When the input consists of independent and identically distributed observations, all procedures have similar coverage. When the data are correlated (i.e., MA(1) and AR1)), the classical procedure's actual coverage is much lower than

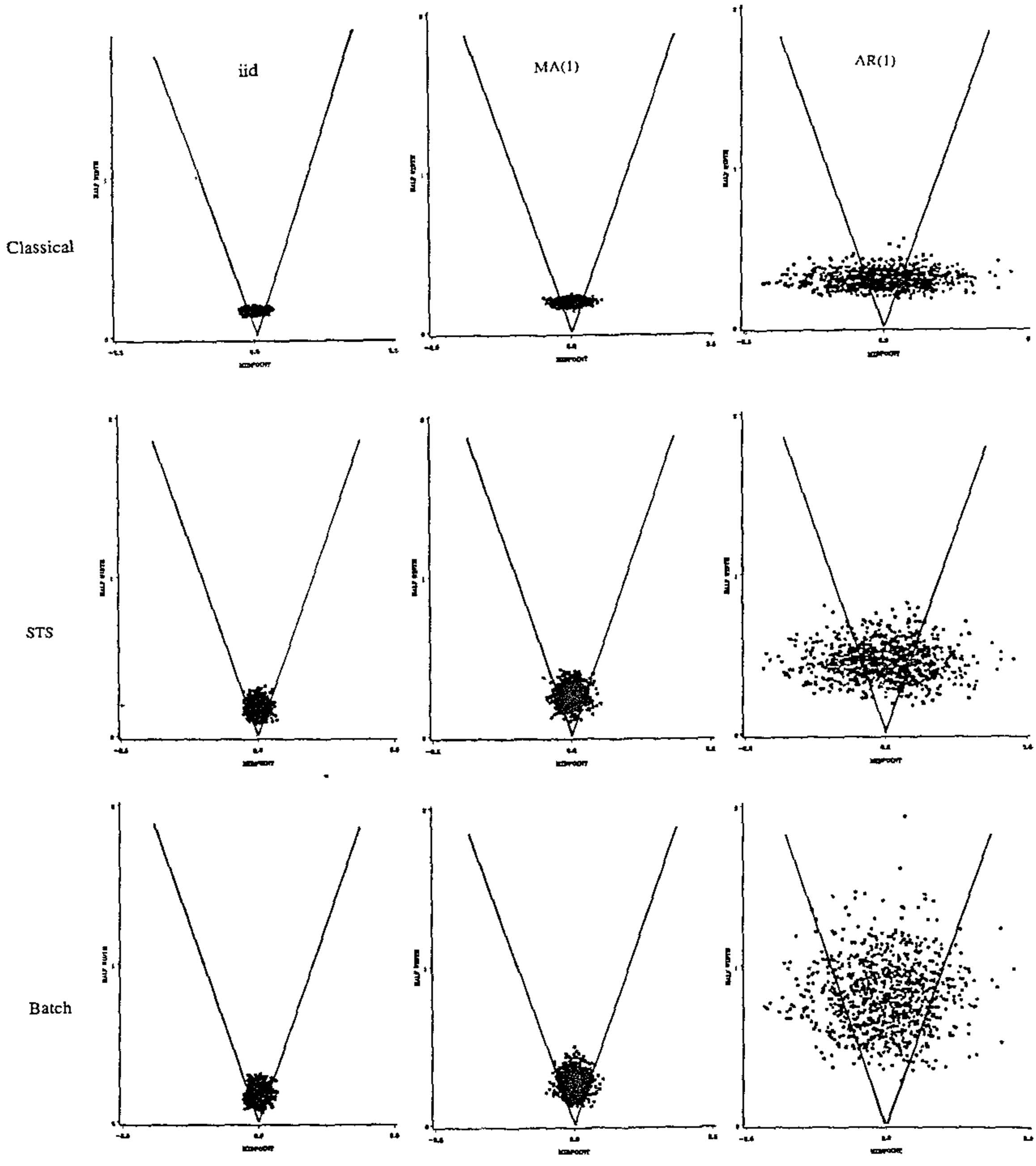


Figure 6.

90%. The difference between the STS and batch means procedures is apparent when the observations are taken from an AR(1) input. In this case, the expected halfwidth is lower for the STS procedure, and the actual coverage is closer to the stated coverage for batch means. The coverage, expected halfwidth and variance of the halfwidth are given in Tables 1, 2 and 3.

Table 1 Coverage for the simulation experiment

	Coverage		
	iid	MA(1)	AR(1)
Classical	89.8	78.8	33.0
STS	89.4	89.0	51.0
Batch means	89.2	91.2	74.2

Table 2 Expected halfwidth for the simulation experiment

	Expected halfwidth		
	iid	MA(1)	AR(1)
Classical	0.165	0.184	0.294
STS	0.177	0.247	0.450
Batch means	0.176	0.260	0.804

Table 3 Variance of the halfwidth for the simulation experiment

	Variance of the halfwidth		
	iid	MA(1)	AR(1)
Classical	0.0014	0.0002	0.0025
STS	0.0016	0.0032	0.0110
Batch means	0.0018	0.0038	0.0453

The relationship between input model and coverage is illustrated in Figure 7. The batch means procedure outperforms the STS procedure in terms of coverage. Figure 8 shows the expected halfwidth as a function of the input model. The STS procedure has a shorter expected halfwidth than the batch means procedure. It is not surprising that there is a positive correlation between coverage and expected halfwidth, and this correlation is dependent on the confidence interval procedure used and the type of input data. Wider confidence intervals are more likely to cover a parameter.

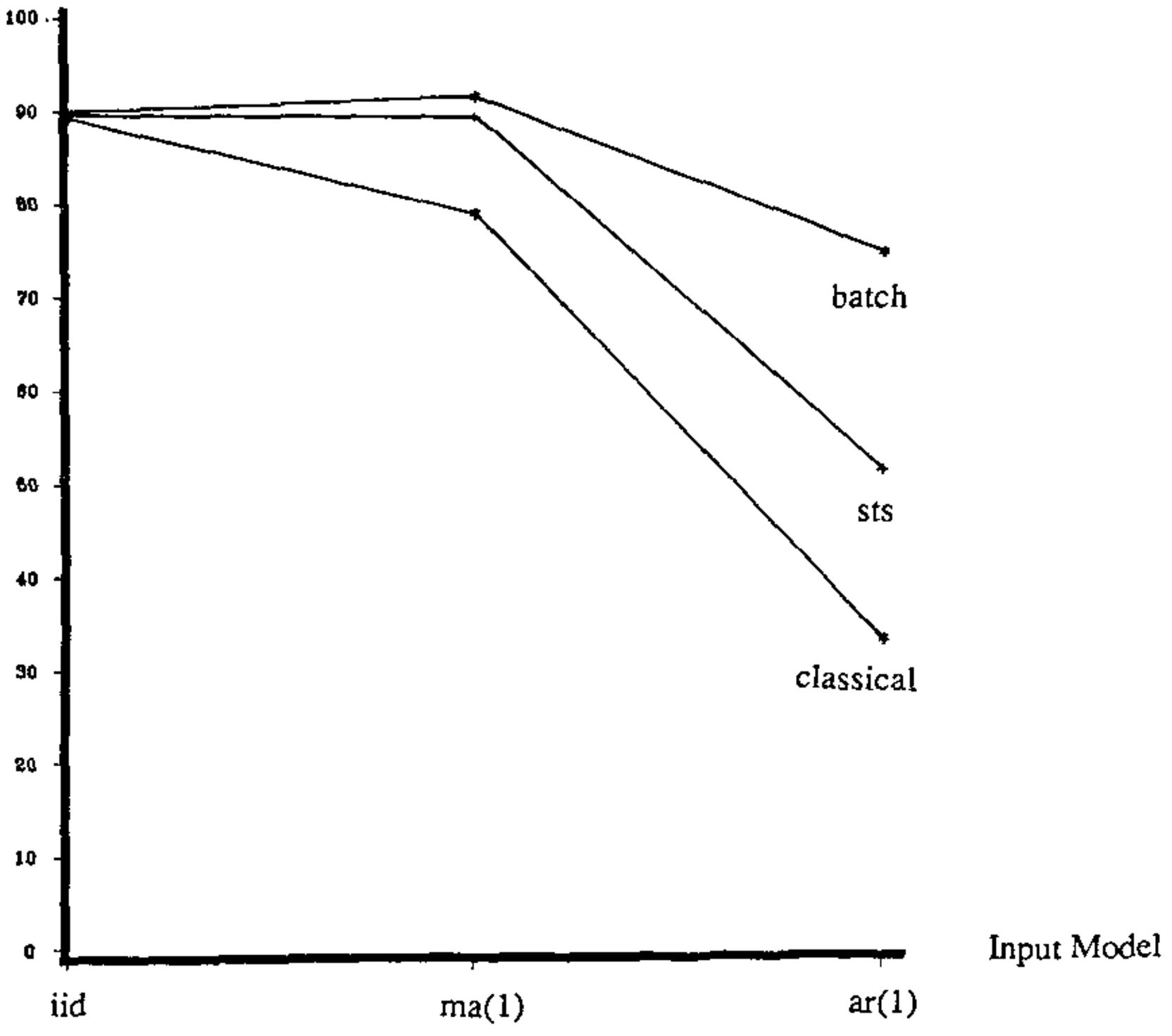


Figure 7 Input Model vs Coverage (stated coverage: 90%).

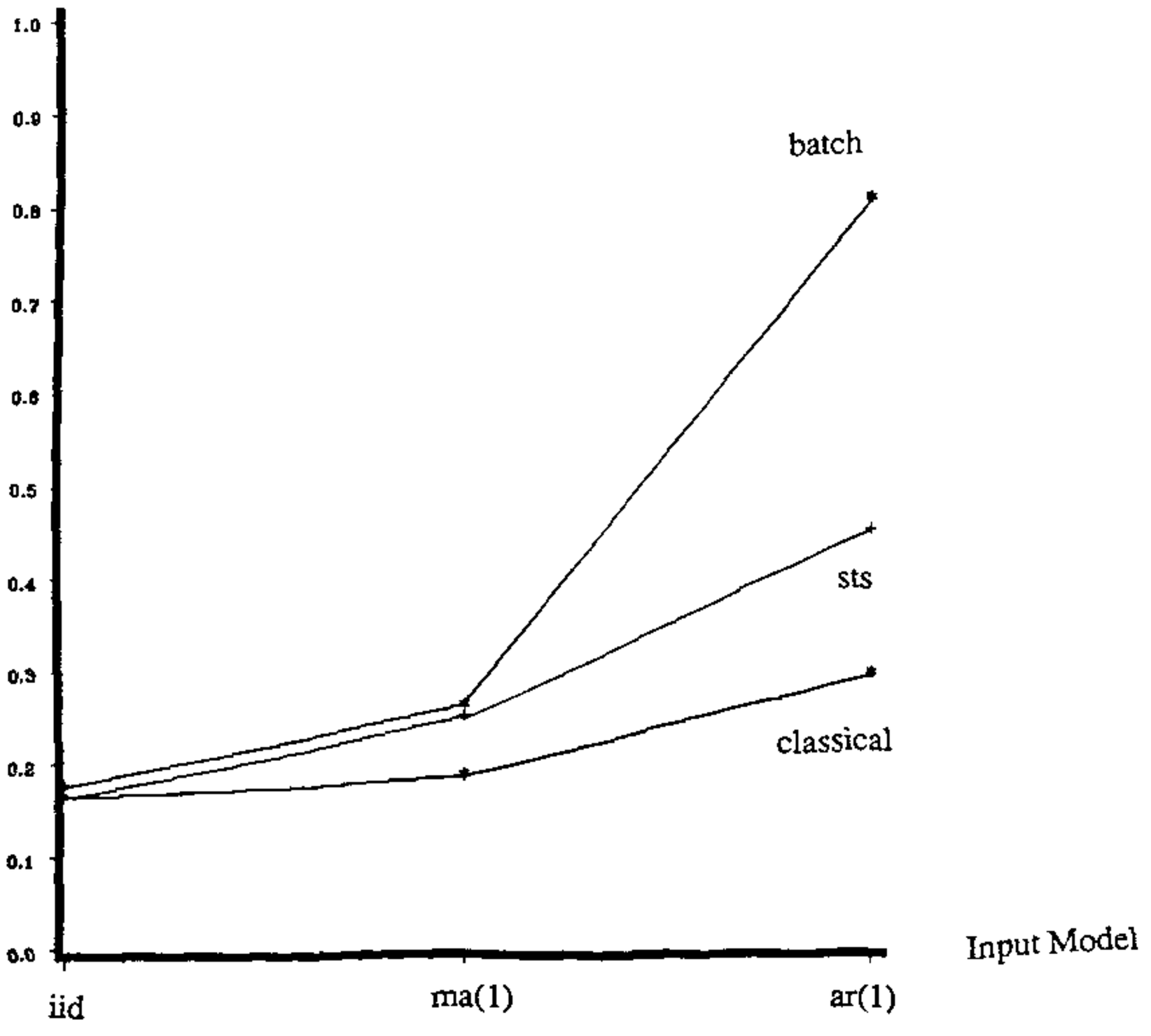


Figure 8 Input Model vs Expected Halfwidth (stated coverage: 90%).

4. APPLICATION: M/M/1 QUEUE

A CIS can be used to aid in the determination of an underlying stochastic process. The following example illustrates the use of a CIS in identifying the appropriate model for the wait times in a M/M/1 queue. The queue has a traffic intensity of 0.9. The parameter of interest is the expected delay in the queue for the first 100 customers. Figure 9, a CIS of 200 confidence intervals generated by classical methods, shows a significant positive correlation between confidence interval midpoint and halfwidth. Kang and Goldsman (1987) discuss the effects of the correlation between mean and variance estimators.

The autocorrelation function of the delay times reveals that the output process resembles an AR(1) time series model. However, a comparison of Figure 6 (where the delay times were approximated by an AR(1) process) and Figure 9 indicates that the autocorrelation function does not completely describe the stochastic process underlying the observations. It was hypothesized that the nonsymmetry in Figure 9 was due to the fact that the wait times were effectively truncated (or, censored) at time zero for those customers that did not wait.

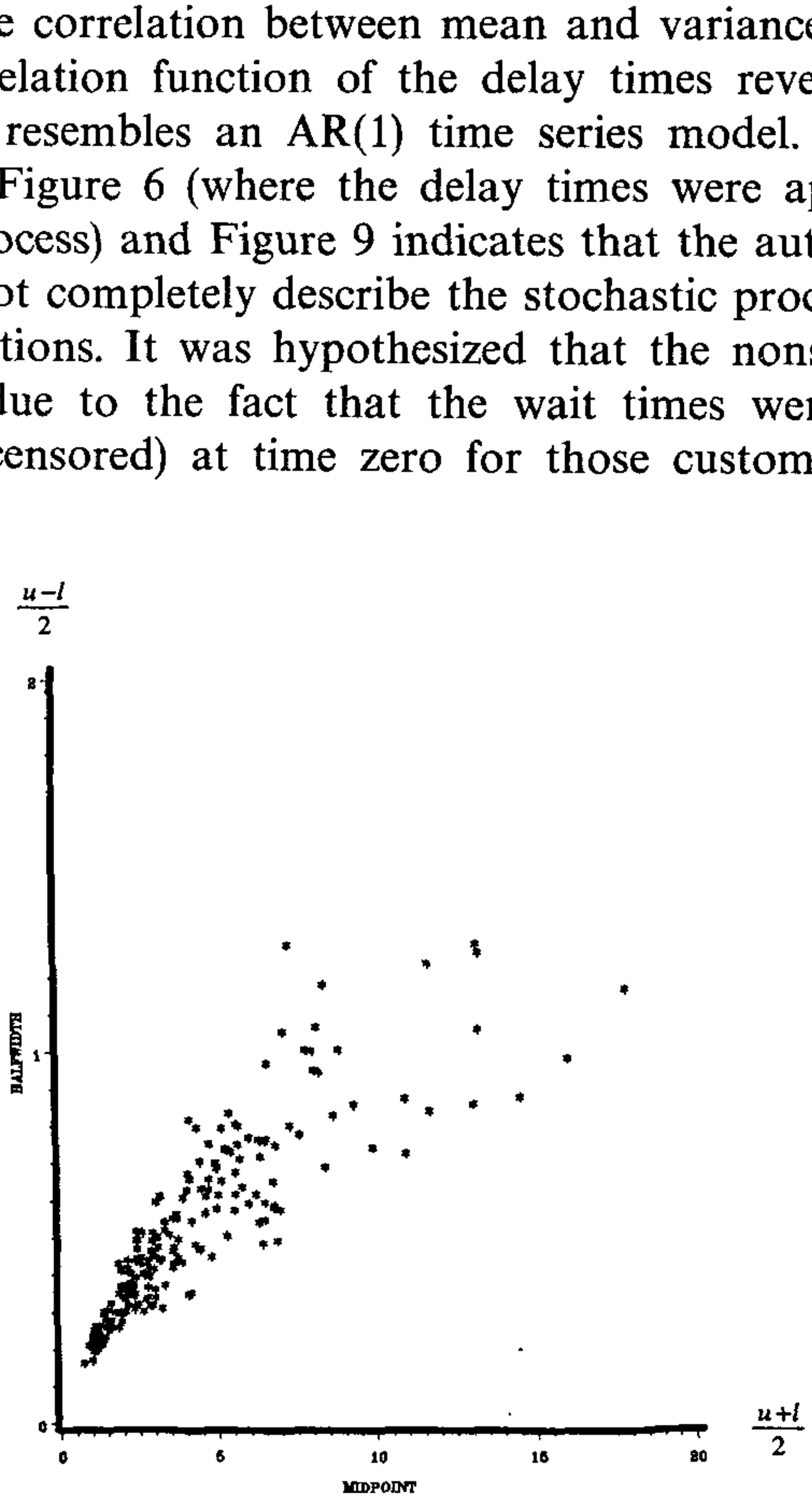


Figure 9 CIS for M/M/1 queue wait times.

The actual coverage for the M/M/1 queue wait times is lower than the AR(1) model. This shows that an AR(1) model does not approximate the process as well as the censored AR(1) model. Also, the AR(1) model censored at zero approximates the M/M/1 queue wait times, as illustrated in Figure 10, a CIS of a truncated AR(1) process. The reason for this behavior is that an AR(1) process may have negative values, whereas the smallest observation of a M/M/1 queue wait time is zero.

One conclusion that can be drawn from the M/M/1 queuing example is that a CIS can be used along with the autocorrelation function in determining the underlying stochastic model from a time series realization. There is clearly a nonlinear relationship between the midpoint and halfwidth of a confidence interval that is easily detected by a CIS.

5. CONCLUSIONS

A CIS is a useful tool for students to gain intuition concerning confidence intervals and for practitioners to compare several confi-

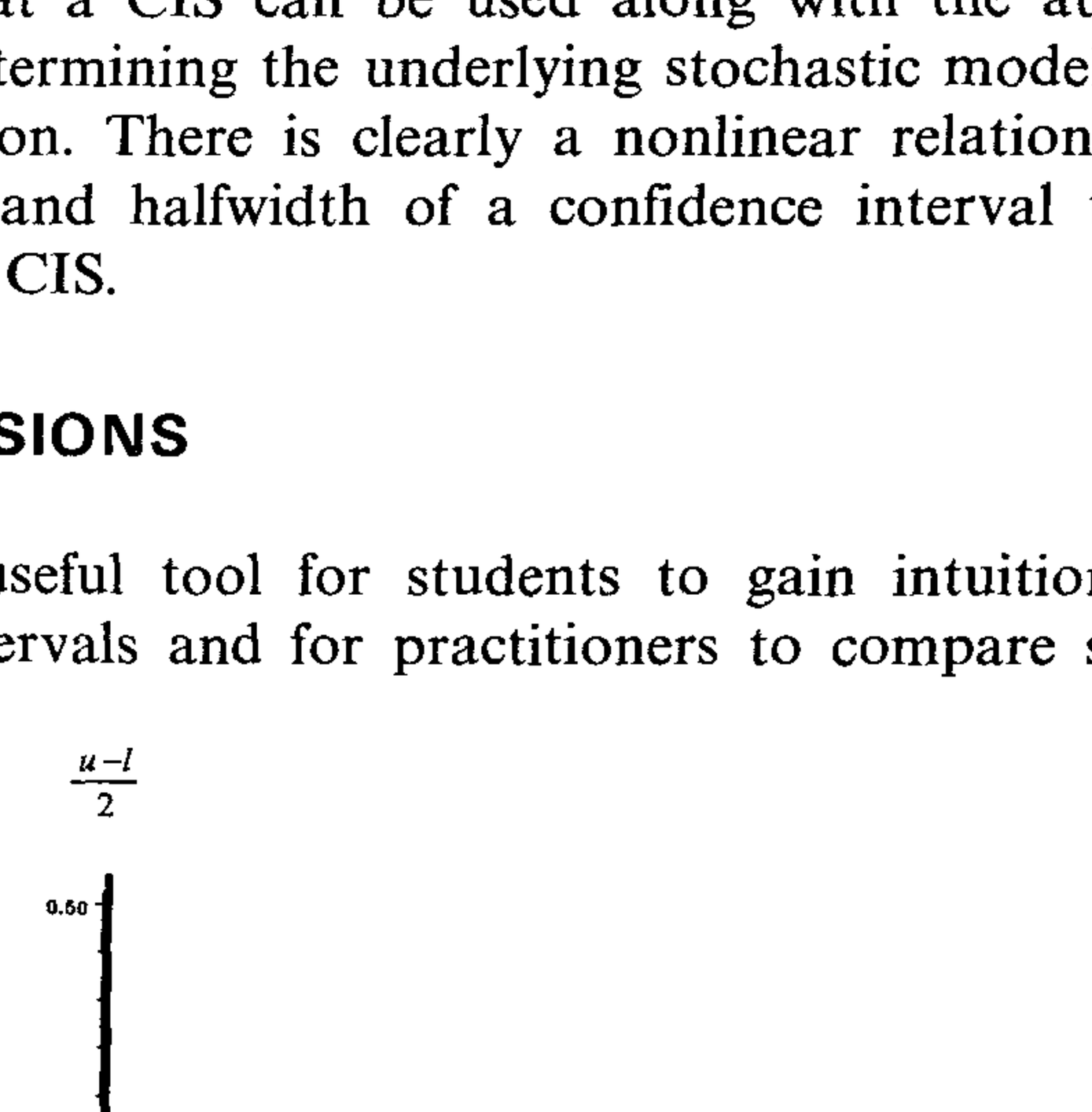


Figure 10 CIS for truncated AR(1) input model.

dence interval procedures. The scatterplot indicates four measures of interest: coverage, bias, expected halfwidth and variance of the halfwidth. Plotting CISs for three simulation output procedures revealed large differences in coverage and halfwidth when the observations were correlated. The CIS was used to show that a censored AR(1) process approximates the delay times in a M/M/1 queuing model.

Acknowledgement

The authors are grateful to Bruce Schmeiser and Keebom Kang for their helpful suggestions on this paper.

References

Kang, K. and Goldsman, D. (1987). The correlation between mean and variance estimators. Technical Report, Department of Industrial Engineering, University of Miami, Coral Gables, FL 33124.

Kang, K. and Schmeiser, B. W. (1986). Methods for evaluating and comparing confidence interval procedures. Technical Report, Department of Industrial Engineering, University of Miami, Coral Gables, FL 33124.

Law, A. M. and Kelton, W. D. (1982). *Simulation Modeling and Analysis*. McGraw-Hill.

Law, A. M. (1983). Statistical analysis of simulation output data. *Operations Research* (31) **6,** 983–1029.

Schruben, L. (1983). Confidence interval estimation using standardized time series. *Operations Research* (31) **6,** 1090–1108.