Proceedings of the 1989 Winter Simulation Conference
E.A. MacNair, K.J. Musselman, P. Heidelberger (eds.)

ANALYSIS OF SIMULATION DATA USING SANDIE

Arne Thesen
Department of Industrial Engineering
University of Wisconsin-Madison, WI 53706, USA.

ABSTRACT

Sandie is a PC based system designed to perform many of the data analysis chores of simulators. We first show how Sandie is used to analyze input data and to fit theoretical distributions to empirical data. We then show how Sandie can be used to perform simple simulations. Finally, we show how to use Sandie to analyze output data. Special care was made in the development of Sandie to design an easy-to-use system that would serve as an aid in the development of an intuitive "sense of numbers". Pull down menus and a dynamic help-line constantly inform the user of all currently operative options.

1. INTRODUCTION

Sandie is an interactive IBM/PC based system designed to perform many of the data analysis chores of simulators. Among the analysis capabilities of Sandie are;

- Auto-correlation
- Batch means analysis
- Compare two data sets
- Confidence intervals
- Data input and editing
- Fitting of data to distributions
- Histograms
- One-way analysis of variance
- Regression
- Runs tests
- Simulation of simple queueing systems
- Two-way analysis of variance

Special care was made in the development of Sandie to design an easy-to-use system that would serve as an aid in the development of an intuitive "sense of numbers". This goal was achieved by providing a system with minimal input requirements, easy to use graphics, and commands that encourage the casual exploration of properties of data sets. Sandie uses pull-down menus and single keystroke commands to invoke all its features.

A Typical Display. A typical Sandie screen is shown in Figure 1. This screen has four main areas:

1. The top line shows the main menu.
2. The next 17 lines is Sandie's main display screen.
3. The next five lines gives statistical summaries for data in four different columns.
4. The bottom line is a help line showing the currently active menu and the currently available commands.

The Main Menu. We see that the main menu has twelve options:

?	Show summary of express keys
Anova	One and two way analysis of variance, regression
Disk	File I/O
Edit	Enter/Edit Data, delete columns
Fit	See if data fits one of five distributions
Generate	Generate data from 9 distributions
Hypothesis	Investigate properties of data sets
List	Display current data
Simulate	Simulate queuing systems
Transform	Make new data from old columns
Zap	Delete data columns
Quit	Exit to DOS

The current main menu selection is indicated by a large *red background cursor* on the main menu line. The selection is changed by using the left or right arrow key to move the background cursor. For example hitting → twice will change the selection from **Disk** to **Fit**.

Pull-Down Menus. Each main-menu selection has an affiliated *pull-down menu*. This menu is displayed by hitting the **<Enter>** or ↓ key. For example, the **Generate** pull-down menu is given in Figure 2. A (red) block cursor behind the first entry (**Beta**) indicates that this is the current selection from this pull-down menu. The ↓ and ↑ keys are used to change the current pull-down menu selection. For example, the Exponential distribution is selected if ↓ is hit once.

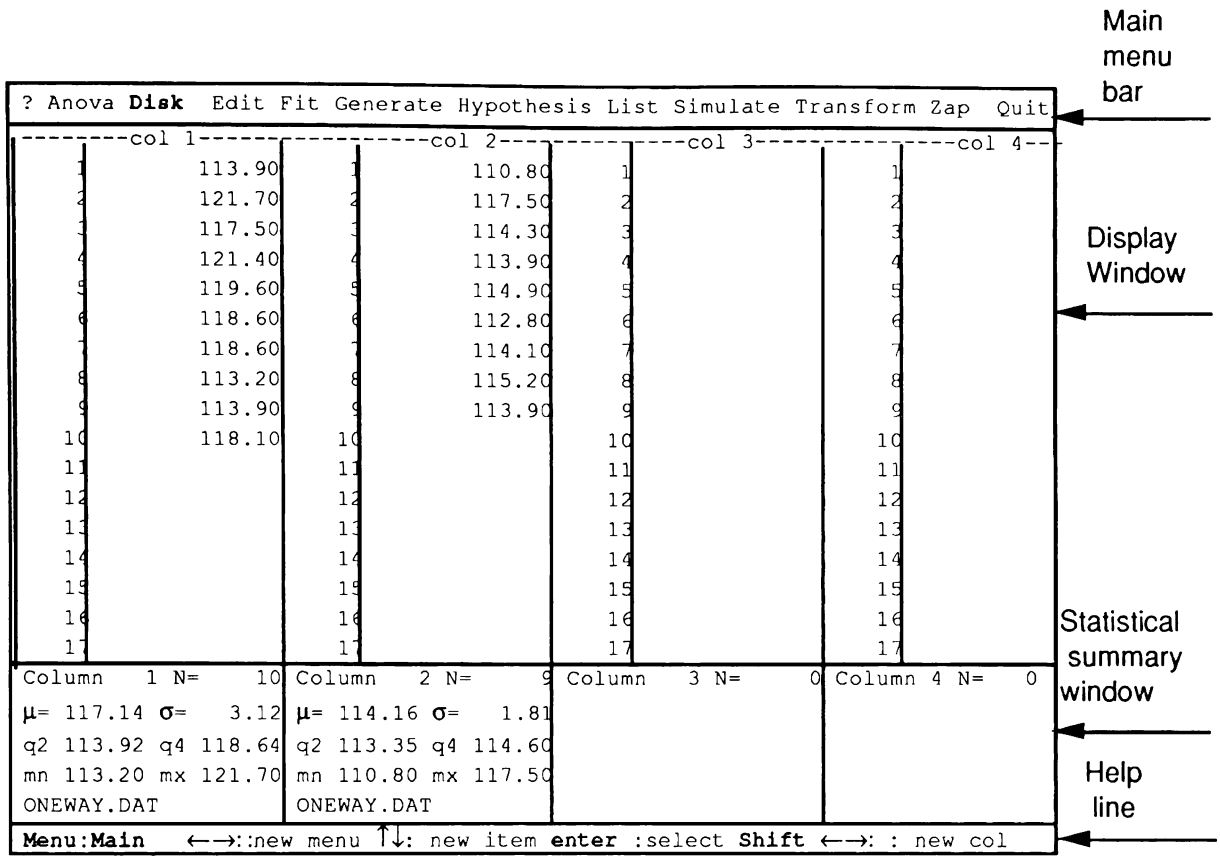


Figure 1: Sandie's display of two data sets, one containing ten observations and one containing nine observations.

The current menu selection is invoked by hitting the **<Enter>** key. Thus exponentially distributed random variables are generated if **<Enter>** is hit at this time.

The **?** pull-down menu differs from the other menus in that it does not provide invocable "action" items. Rather, it provides a summary of time-saving shortcuts programmed into the keyboard function keys

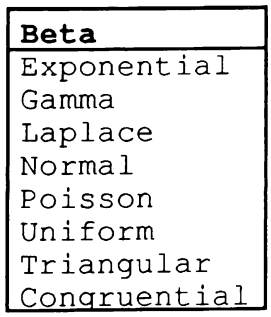


Figure 2: The Generate Pull-down menu.

Column Orientation. Sandie stores data in *columns*. Each column contains related information. For example, the data in column one in Figure 1 contains the results of replications of a simulation of a three-station job shop using a scheduling policy of First In/First Out.

Labels. Sandie assigns default labels as data is entered or generated. Sandie's text editor has options for user defined labels.

The Current Column. Most of Sandie's commands perform operations on all the data in a column at one time. To eliminate the repeated reference to this column, Sandie introduces the concept of the *current column*. This is the column that an operation will be applied to unless an other column is specified. A red statistical summary window is used to identify the current column. The current column is changed when the un-shifted **<4>** and **<6>** keys (corresponding to left and right arrows on the key pad) are hit. Sandie automatically shows the histogram of the the new current column, and the statistical summary windows are updated to show the new current column with a red background.

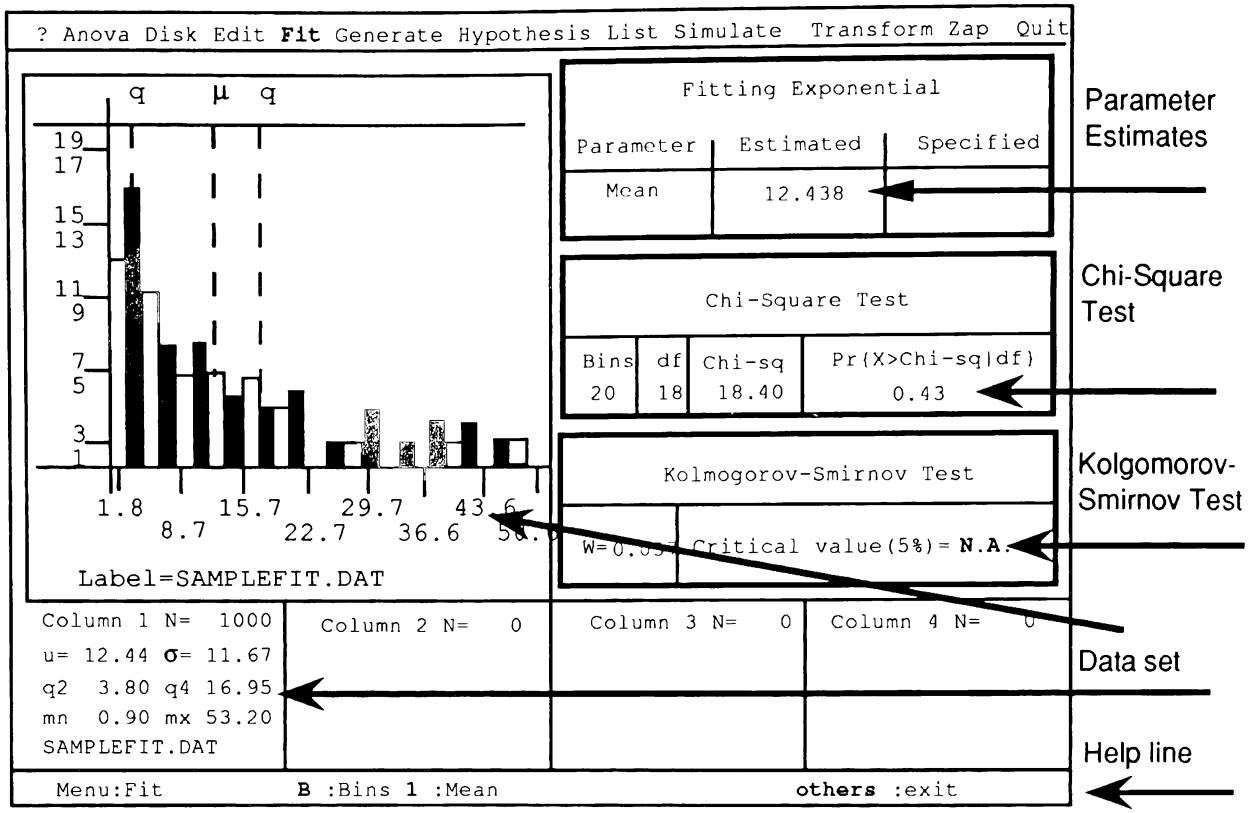


Figure 3: Fitting data to an exponential distribution.

The histogram. Sandie usually displays a histogram of the data in the current column. One such histogram is shown in Figure 3. Ten, fifteen, or thirty equally spaced bins, or intervals, including an overflow and an underflow bin can be used.

2. FITTING DATA TO DISTRIBUTIONS

Sandie is able to fit data to the exponential, gamma, normal, triangular and uniform distributions. Both a Kolmogorov-Smirnov test and a chi-square test is performed. Bin-widths resulting in bins with equal expectations are used for the chi-square test. Twenty bins are initially used. The user may change this as desired.

To illustrate Sandie's distribution fitting capabilities, we will fit an exponential distribution to a data set containing 1000 randomly generated values. A histogram of this data and the result of the analysis is given in Figure 3.

Parameter Estimation. The top window on the right hand side of the display window tells what distribution is being fit and the value that **Sandie** estimated for the parameter(s). Maximum likelihood estimators are usually used.

It is also possible to specify the value(s) of the parameter(s) to be used. This is important for the Kolmogorov-Smirnov test.

The Chi-Square Test. Sandie computes the value of chi-square using bins with equal expectations. The number of bins that were used, the degrees of freedom, the computed chi-square value, and the percentage of random variables from the appropriate chi-square are then displayed.

The Kolmogorov-Smirnov Test. The value of the Kolmogorov-Smirnov test statistic (W) is also computed and displayed. However, since the Kolmogorov-Smirnov test is usually meaningful only when parameter values are known with certainty, critical values are only given when parameters are specified. A critical value based on estimated parameter values is given when a normal distribution is being fitted.

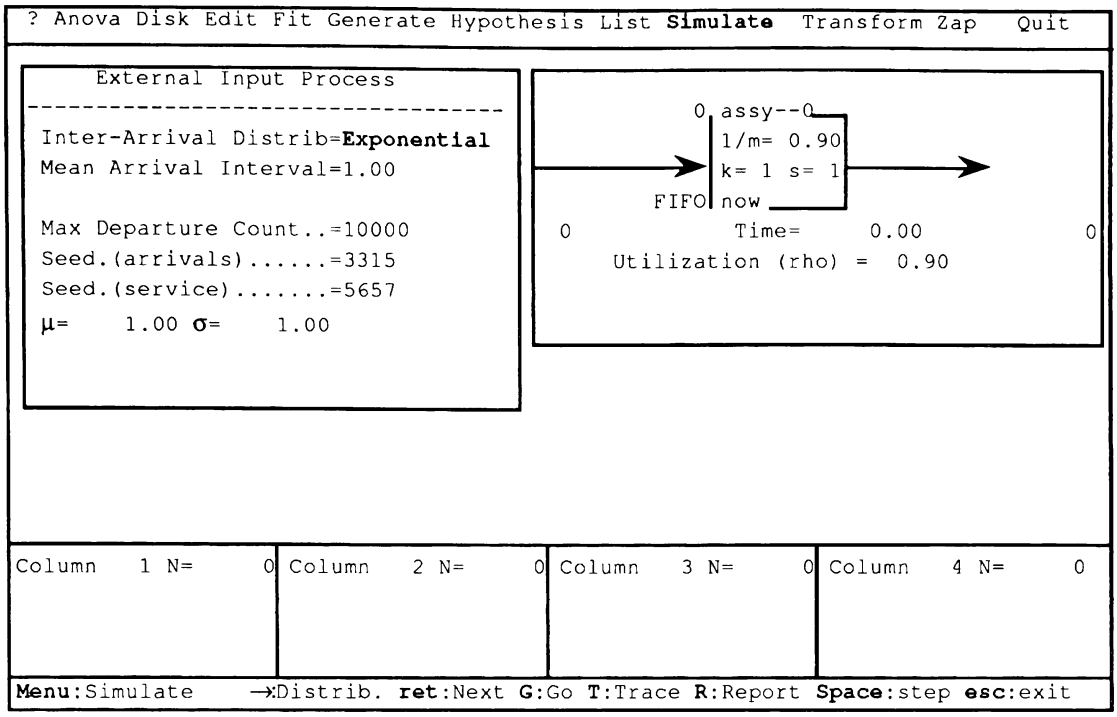


Figure 4: Sandie's initial simulation model.

3. SIMULATION

In addition to being able to generate random variables from nine different distributions (the Generate menu was shown in Figure 2), Sandie is able to perform simulation of two different classes of simple queueing systems, and to capture the systems times for individual customers. The resulting data sets can then be analyzed using Sandie's other features.

3.1 Single Station Queuing Systems

A screen such as the one shown in Figure 4 is displayed when the single station queuing simulation option is selected. The left portion of the screen defines the arrival process. Any of the parameters listed here can be changed simply by placing the cursor on the item and entering the new value. Four different arrival distributions are available.

The queuing model is shown on the right side on the screen. Again, any of the given parameters can be changed simply by positioning the cursor and entering the new value. Only Erlang distributed service times are allowed. Six different queuing disciplines are supported.

Figure 4 indicates that we are about to simulate 10,000 departures of a single server queuing system with exponentially distributes service and inter-arrival times and a specified utilization of 90%. From the help-line we see that <G> will start the simulation. After 15 seconds (PS2/80) the display shown in Figure 6 is seen.

We see that the time in the system for 10,000 customers has been captured in column one. This data will be analyzed in Section 4.

3.2 Inspect/Repair Model

Sandie's other simulation model is an inspect/repair loop (Figure 5). The main work station is an inspection station. Parts failing inspection are repaired at a separate work station and returned for a another inspection. As for the single station queuing model, all parameters can be changed by positioning the cursor and entering new values. The input process is defined in the same way as for the single station queueing system.

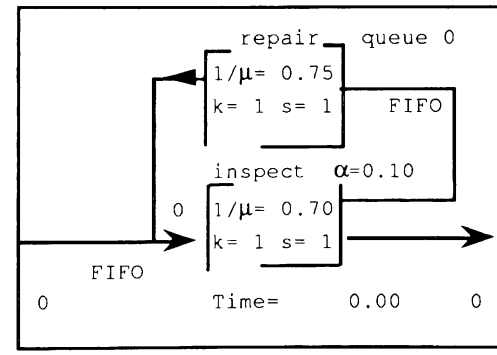


Figure 5: The Inspect/Repair model.

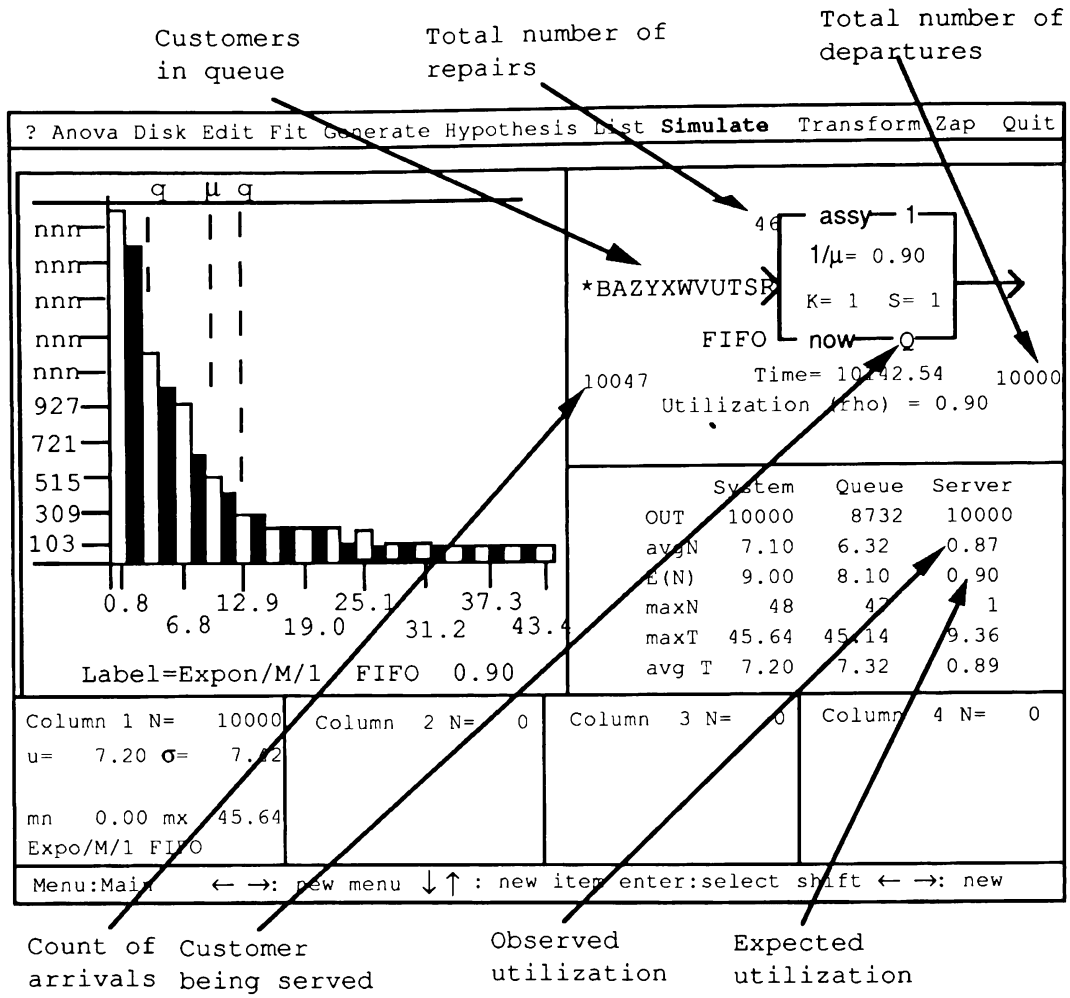


Figure 6: A completed simulation. A histogram of the collected data is given, as is a conventional statistical summary of the observed performance data.

4. OUTPUT ANALYSIS

Simulation data is often difficult to analyze. One reason for this is the presence of auto-correlation. Sandie is able to identify this problem, and it provides a technique, *batch means analysis*, that has been shown to be useful in many cases to deal with the problems introduced by positive auto correlation. These capabilities are discussed in this section.

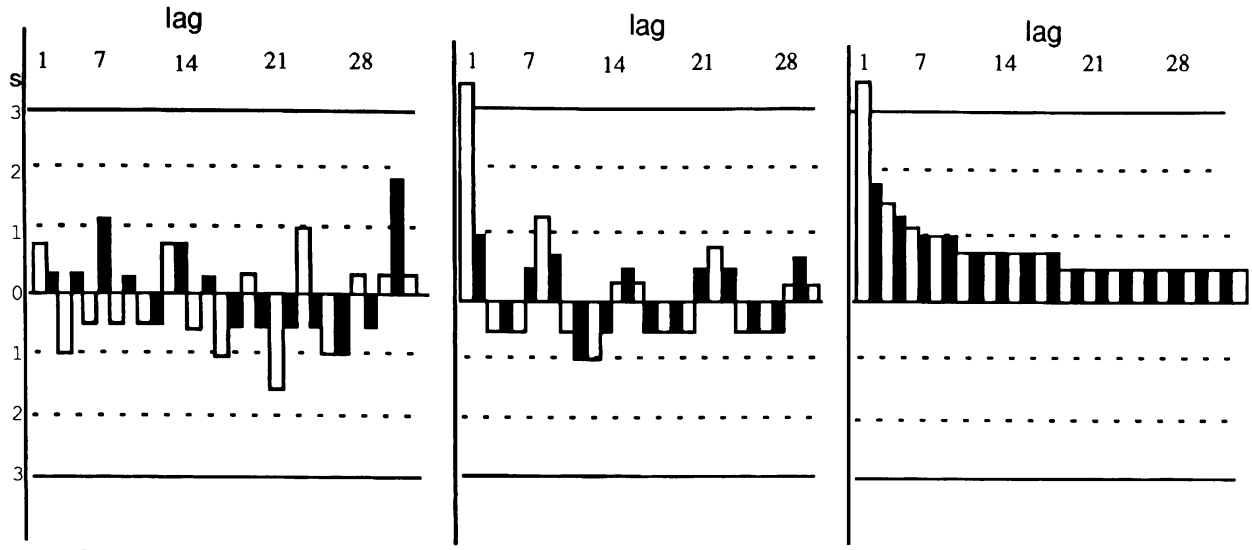
4.1 Auto-correlation

The unified auto-correlation coefficient. The unified auto-correlation coefficient for lag i (uacf(i)) shows the degree to which there is a relationship between an arbitrary variable in a data set and the variable i positions higher up in the data set. If there is no relationship then uacf(i) should be equal to zero, if there is a positive correlation, then uacf(i) should have a positive value, conversely if there is a negative relationship then uacf(i) should have a negative value.

The values of the uacf(i)'s for different lags can be estimated from the data. If there is no auto-correlation present in the data, then the resulting values will not be equal to zero, rather, they will contain arbitrary noise, showing no particular pattern, and few if any extreme values. The result of an analysis of data without any auto-correlation is shown in Figure 7a.

A plot of the uacf(i)'s may be used to detect the presence of cyclical patterns in data. For example, we show in Figure 7b, an analysis of hospital census data. We see from the pattern that there is a strong relationship between the census on consecutive days (lag 1) and between the census for consecutive identical days of the week (lag 7).

An analysis of time-in-system data for simulation of 10,000 departures in a simple queueing system is shown in Figure 7c. The presence of strong positive auto-correlation is indicated by the strong lag 1 correlation and by the gradually declining value of uacf(i) with increasing lags.



a). Independent observation b) Hospital census c)System times in queuing simulation

Figure 7: Auto-correlation analysis of three data sets. The first has no auto-correlation present, the second has a cyclic pattern, the third has strong positive auto-correlation.

4.2 Batch Means Analysis

Batch means analysis is used to develop confidence intervals about the true mean of auto-correlated data. The analysis has three steps. First, the original data set is partitioned into n consecutive batches. For example the first 1000 observations in the data set can go to batch 1 the second 1000 observations goes to batch 2 etc. Second, the average value of the items in each batch is computed. Finally, the new data set is used to estimate the underlying variance of the data and to compute a confidence interval about its true mean.

Sandie's batch means analysis of data generated for 10,000 departures for an MM1 queueing simulation is shown in Figure 8. The computed confidence interval is <4.34,10.05>. Figure 8 also shows individual (conventional) confidence intervals for the 10 batch means. Note that these intervals are quite narrow, and that few of the intervals overlap. This is a reflection of the auto-correlated nature of the raw data. Similar confidence intervals computed on data without auto-correlation would overlap.

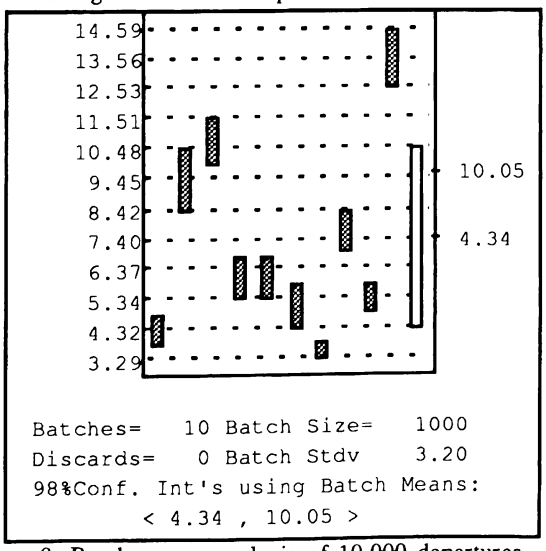


Figure 8: Batch means analysis of 10,000 departures from an MM1 queuing simulation with 90% utilization.

5. CONCLUDING REMARKS

We have seen that Sandie is a useful tool for analysis of simulation input and output data. To support these activities, Sandie also provides a built in data editor and easy to use data file capabilities. A number of other capabilities are also provided. Among these are transformations, one and two-way analysis of variance, regression and Knuth's runs test.

AUTHORS BIOGRAPHY

ARNE THESEN, a Professor of Industrial Engineering and Computer Sciences at the University of Wisconsin-Madison, has been active in the simulation field since 1964. His current research interests are in the area of expert scheduling systems. He is the co-author with Professor Laurel E. Travis of a forthcoming text *Simulation for Decision Making* (Academic Press 1990).

Arne Thesen
Department of Industrial Engineering
1513 University Ave,
Madison, WI 53706
(608) 262-3960.