Automation and Robotics in Construction XVI © 1999 by UC3M

SIMULATION MODEL PROBABILITY DISTRIBUTION FUNCTIONS -- RELATIONSHIPS BETWEEN INPUT DATA AND GOODNESS-OF-FIT TESTS

Cristiano Maio[1] and Cliff Schexnayder[2]

[1]Fulbright Graduate Student and [2]Eminent Scholar, Del E. Webb School of Construction, Box 870204, Ariz. State Univ., Tempe, AZ 85287-0204, USA. cliff.s@asu.edu

Abstract: Advances in computer technologies now allow for the accumulation of construction process data in large quantities. The disadvantage of having a large number of observations is that there are problems with goodness-of-fit tests. This paper addresses the issue of the effect goodness-of-fit tests have on selecting a probability distribution function for use in construction process control, automation or in the development of simulation models of processes. In the case of a large number of observations the sum of small variations can yield goodness-of-fit results which could cause one to believe that the Probability Distribution Function does not accurately represent the underlying population.

Keywords: Probability, Distributions, Functions, Construction, Simulation, Chi-square, Kolmogorov-Smirnov, Anderson-Darling.

1 INTRODUCTION

New technologies are allowing researchers and managers to accumulate large quantities of automated real-time project process data. A trace-driven process control or simulation using these large data sets can be developed, however, there are major drawbacks to such a course of action as the process control or simulation will reproduce solely what has already happened. There is also the issue that using large amounts of data may require the use of time-consuming computational procedures. Therefore, standard probability distribution functions (PDF) are often used to represent this empirical data, in fact, standard distributions level data irregularities which can derive from field observations.

This paper addresses the issue of the effect goodness-of-fit tests have on selecting a probability distribution function for construction process control and simulation modeling.

1.1 Project Data

The data used in this research was acquired from the Atkinson-Washington-Zachry (AWZ) joint venture on the Eastside Reservoir Project in California, U.S.A. AWZ operates a fleet of Caterpillar Inc. (CAT) 785 trucks on this project. These trucks are equipped with Caterpillar's Vital Information Management System (VIMS) and the CAT Truck Payload Monitoring System (TPMS). These two systems automatically record a large amount of truck performance data. Eight months of TPMS data from this project, representing 54,000 truck cycles, is the original data for all of the subsequent statistical research presented in this paper.

1.2 Probability Distribution Functions

The problem of collecting and analyzing data confronts all researchers trying to model real world activities. The required process inputs for a simulation model are usually approached by fitting a statistical distribution to a collection of sample observations. Since many classical distribution functions could fit the sample, goodness-of-fit tests are performed on the constructed probability distribution function [1].

2 GOODNESS-OF-FIT TESTS

The methods available for investigating the quality of the fitted distributions can be divided into heuristic procedures and goodness-of-fit hypothesis tests. Heuristic procedures include frequency comparison and probability plots. Frequency comparisons are graphical comparisons between the histograms of the fitted and the original distribution. Probability plots are graphical comparisons of the data distribution with the fitted distribution, these can be either probability-probability (P-P), or quantile-quantile (Q-Q) plots. P-P and Q-Q plots reduce the problem of comparing distribution functions to comparing a straight line (the fitted distribution) to a curve (the data distribution). While admitting the existence of these alternate methods those techniques are not the issue under consideration in this research.

2.1 *Goodness-of-fit Tests*

Goodness-of-fit tests represent a statistical hypothesis test used to assess if the input data is an independent sample from a particular distribution function. Three tests have been developed for this purpose: Chi-square, Kolmogorov-Smirnov, and Anderson-Darling.

The Chi-square test, developed by K. Pearson in 1900, is a formal comparison of an input data histogram with the fitted distribution. It should be noted that the choice of the number and size of class intervals greatly influences the results of the Chi-square test. To ensure validity of the test, Law and Kelton suggest the use of a number of intervals k, so that $np \geq 5$, where n = number of data points and $p = 1/k$ [7]. The Chi-square test was derived for the case of estimating parameters by maximum likelihood and is best used with a large number of samples, >50 [1]. But there appears to be problems when the sample size becomes "very" large.

The Kolmogorov-Smirnov (K-S) test compares an empirical distribution function with the distribution function of the hypothesized distribution. The K-S test does not require grouping of the data and eliminates the problem of interval specification. The K-S test checks if the empirical data could have originated from a theoretical distribution with the estimated parameters. Law and Kelton [7] state that the K-S test has several advantages over the Chi-square test: 1) the test does not require grouping of the data in any way, 2) it is valid for the exact sample size and 3) it is more powerful against alternative distributions.

The Anderson-Darling (A-D) test differs from the K-S test on the weight given to the tails of the distributions. In fact, it was designed to detect discrepancies in the tails [7].

2.2 *Number of Class Intervals*

There is no definitive guide for choosing the number of intervals. Several methods have been suggested. Montgomery and Runger [8] suggest that, to avoid uninformative histograms with too many or too few class intervals, between 5 and 20 intervals should be used. They further state that the researcher should calculate the number of intervals by taking the square root of the number of observations. Abourizk and Halpin suggest the use of Sturges' Rule for selecting the number of class intervals. Sturges' Rule states that for n observations X_i to be summarized in a frequency distribution, the number and width of class intervals for the distribution should be calculated with the following equations:

$$\text{Number of class intervals} = 1 + 3.322 \log_{10} n$$

$$\text{Width class interval} = X_{max} - X_{min} / \text{No. class intervals}$$

Where n = sample size

X_{max} and X_{min} = highest and lowest observations

Figure 1 compares the difference in the number of class intervals resulting from the square root and the Sturges' rule approaches. It is obvious that for data containing fewer than 100 observations the rules will furnish similar results, but beyond that point they diverge rapidly. With 100 observations the square root value would be 10 and the Sturges' 7 but with 500 observations the values are 22 and 10. The square root guidance suggests a number twice that of the Sturges' rule.

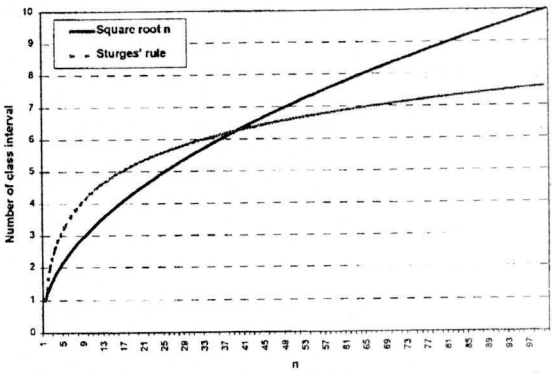

Figure 1. Number of class intervals according to square root and Sturges' rules.

The documentation for the statistical computer program BestFit [4] suggests that goodness-of-fit tests are very sensitive to the magnitude of n (the number of data points).

> *If* n *is small, the goodness-of-fit will only measure large differencse between the input data and the distribution function. As* n *increases, the modified test statistics increase and the null hypothesis will be rejected more often.* The results produced by these tests should be considered "guidelines" in selecting a fit. Always evaluate the results by comparing statistics and graphs before accepting or rejecting a fit* [4].
>
> * Our emphasis.

When the parameters of the distribution must be estimated from the sample data, it can be expected that the goodness-of-fit tests will give more meaningful information if small samples are used. Phillips [9] confirms this assumption saying "slight modifications of the calculated statistics are given to enable the points to be used with small samples."

2.3 *Goodness-of-fit Computer Programs*

Computer software programs have been developed that automatically assess the goodness-of-fit of sample data to distribution functions. The program Visual Interactive Beta Estimation System (VIBES) was used by Abourizk, Halpin and Wilson [2] to fit and shape beta distributions to their sample

data. The program BestFit compares the sample data with up to 26 different distributions [4]. The parameters for each distribution are compared to the sample data using maximum likelihood estimators; then they are optimized and the Chi-square test, the Kolmogorov-Smirnov test and the Anderson-Darling test statistics are calculated.

The goodness-of-fit statistic tells how probable it is that a given distribution function reproduces the data set. Which translates into: whether input data was created by the distribution function reported by calculated distributions [4, 6]. The critical value involved is a goodness-of-fit measurement that is compared to the goodness-of-fit of the selected distribution. This measurement is calculated differently for the Chi-square, Kolmogorov-Smirnov and Anderson-Darling tests.

2.4 Distribution fitting

The raw project data was formatted and entered into the BestFit program to obtain a fitted distribution. Three different methods for testing the match of the suggested distributions to the input data were utilized to rank suggested distributions; Chi-square, Kolmogorov-Smirnov and Anderson-Darling. The on-line self help for the BestFit program offers suggestions about each of the tests:

> *There is no specific goodness-of-fit test that will give you the "best" result. Each test has its strengths and weaknesses. You must decide which information is most important to you when considering which test to use.*
>
> *The Chi-square test is the most common goodness-of-fit test. It can be used with any type of input data (sample, density or cumulative) and any type of distribution function (discrete or continuous). A weakness of the Chi-square test is that there are no clear guidelines for selecting intervals. In some situations, you can reach different conclusions from the same data depending on how you specified the intervals (number of classes).**
>
> *The Kolmogorov-Smirnov test does not depend on the number of intervals, which makes it more powerful than the Chi-square test. This test can be used with any type of input data but cannot be used with discrete distribution functions. A weakness of the Kolmogorov-Smirnov test is that it does not detect tail discrepancies very well.**
>
> *The Anderson-Darling test is very similar to the Kolmogorov-Smirnov test, but it places more emphasis on tail values. It does not depend on the number of intervals. A weakness of the Anderson-Darling test is that it can only be used with sample input data* [4].
>
> *Our emphasis

The choice of which test to use for selecting probability distribution functions to use in controlling automated processes or for simulation modeling is based on different factors:

- **Goodness-of-fit test used in previous researches.** The Chi-square and Kolmogorov-Smirnov tests have been used in previous construction simulation modeling research. The Chi-square test was used by Clemmens and Willenbrock [5]. More recently, probably thanks to the new computer possibilities, Kolmogorov-Smirnov was used by Abourizk [3]. The Anderson-Darling test has not been used extensively in construction based research.
- **Outcome of consistent results.** The Chi-square test may be biased by the choice of the number of class intervals [7]. The number of class intervals in which the data is divided does not influence Kolmogorov-Smirnov and Anderson-Darling test.
- **Behaviour of test regarding goodness-of-fit.** The Anderson-Darling test was specifically designed to detect discrepancies in the tails of the distribution [8]. Chi-square and Kolmogorov-Smirnov, on the other hand, give a lower weight to the differences of the tails of the distributions.

Evaluating the advantages and disadvantages of the different goodness-of-fit test methods, it appears that the Kolmogorov-Smirnov test is the most appropriate for construction models because: 1) previously documented applications, 2) the consistency of results, and 3) the weight given to the differences of the distribution tails. However, the results of using all three of the goodness-of-fit tests are presented in this research to emphasis the effect the decision to use a specific test has on probability distribution function selection.

3 INPUT DATA FITTING RESULTS

The research used five categories of data: load time, haul time, dump time, return time and cycle time. The traveled distance does not influence load time and dump time, therefore the analysis of such data does not consider the distance parameter. Haul time, return time and cycle time are analyzed considering the traveled distance.

An attempt to fit a distribution to the input data was made only if:

1) There was a minimum of 100 data points in the case of haul, return, and cycle times (where the number of data points is influenced by the distance parameter).
2) There was a minimum of 500 data points in the case of load and dump times (where the number of data points is not influenced by the distance parameter).

3.1 *Class Interval Effect*

The determination of differences between fittings based on a specific class interval decision rule, Square root or Sturges' yielded results confirming the information in the literature:

- The decision rule influences the results primarily when the Chi-square fitting method is utilized (almost 25% of the time).

Fitted probability distribution functions obtained by entering the raw haul, return, and cycle time data into the BestFit program were developed using both of the class interval decision rules. Then the resulting first ranked distribution types for both the Square root and Sturges rules were compared. The ranking were acquired using each of the three goodness-of-fit tests, so the comparisons of the class interval rule effects are based on the same goodness-of-fit test. Figure 2 presents the number of times the first ranked probability distribution function resulting from both the class interval decision rules matched (matching/non matching fits).

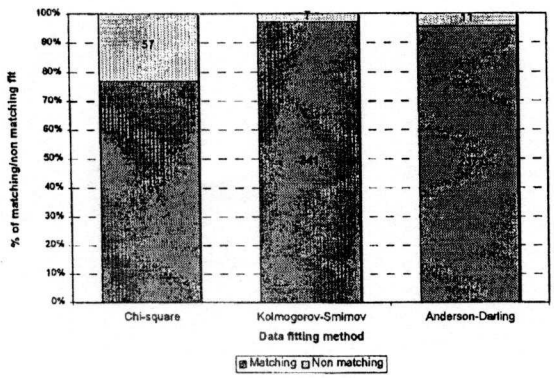

Figure 2. Allotment based on goodness-of-fit rule of matching/non matching haul, return, and cycle time distribution fits.

The analysis reveals that almost 25% of the resulting theoretical distributions (57 out of 248) may change if the Chi-square test is used to select the best matching probability distribution function. When using either the Kolmogorov-Smirnov or Anderson-Darling methods the change would occur in less than 5% of the cases. This data seems to support the warnings found in the literature about the weakness of the Chi-square test. *"In some situations, you can reach different conclusions from the same data depending on how you specified the intervals. [4]"*

3.2 *Load time data*

The results of the load-time data distribution fitting based on the three goodness-of-fit tests are shown in Table 1

In only one case (the 130-140 ton range) out of the six investigated was the Chi-square test not influenced by the choice of class interval. In the case of both the Kolmogorov-Smirnov or Anderson-Darling tests the class rules did not change the outcome of the best-ranked probability distribution. Again the data seems to support the warnings found in the literature about the Chi-square test.

Table 1. Distribution load time data fitting results.

LOAD TIME Weight (tons)	Class Rule	Data points	Chi-square	K-S	A-D
110-120	Square	2420	Triangular	Logistic	Normal
	Sturges	*2420*	*Rayleigh*	*Logistic*	*Normal*
130-140	Square	7000	Extreme val.	Pearson VI	Pearson V
	Sturges	*7000*	*Extreme val.*	*Pearson VI*	*Pearson V*
140-150	Square	8854	Extreme val.	Pearson V	Pearson V
	Sturges	*8854*	*Lognorm*	*Pearson V*	*Pearson V*
150-160	Square	7197	Chi-square	Pearson V	Pearson V
	Sturges	*7197*	*Pearson V*	*Pearson V*	*Pearson V*
160-170	Square	3308	Extreme val.	Pearson V	Pearson V
	Sturges	*3308*	*Loglogistic*	*Pearson V*	*Pearson V*
170-180	Square	626	Extreme val.	Loglogistic	Loglogistic
	Sturges	*626*	*Loglogistic*	*Loglogistic*	*Loglogistic*

3.3 *Haul time data*

The haul time depends on the distance traveled by the trucks. Therefore, the data was divided and grouped in ranges of 0.10 miles and a distribution fitting was attempted only for ranges with at least 100 data points. Table 2 shows the results of the haul-time data distribution fitting based on the three goodness-of-fit tests. These results fail to support in such a dramatic manner the sensitivity of the Chi-square test as only one data set displays a difference in selected distribution. However, this is one in four or 25%.

Figure 3 analyses how the types of distributions are apportioned based on the distribution fitting test, Chi-square, Kolmogorov-Smirnov, or Anderson-Darling. Extreme value and Beta distributions represent more than 60% of the fits whatever fitting test is implemented. The Beta distribution exhibits the greatest variation based on the fitting test employed, ranging from less than 20% of the fits in the case of Chi-square test, to more than 40% when the Anderson-Darling test is used.

The analysis presented in Figure 2 demonstrates that probability distribution fits obtained with the Kolmogorov-Smirnov method are not influenced by

the class interval selection method. In fact the K-S test seems almost immune to class interval variation.

The question of haul distance variation was also investigate to determine if it is possible to identify a trend in the probability distribution function fits based on the distance traveled by the haul trucks.

Figure 4 analyses the haul data based on four different distance ranges: 0.5-1 miles, 1-1.5 miles, 1.5-2 miles, and 2-2.5 miles. For clarity, only distributions with at least 10% of the fits are shown. It is clear that Beta distribution is fairly constant and not influenced by the traveled distance. Extreme Value distribution percentages vary with a maximum at the 1-1.5 mile range, but it does not seem to show a trend.

Table 2. Distribution fitting results for haul time data, distance range 0.5-0.6 miles.

HAUL Weight Class (tons)	Rule	Data points	Distance Chi-square	0.5 to 0.6 K-S	miles A-D
130-140	Square	688	Extreme Val.	Beta	Extreme Val.
	Sturges	*688*	*Extreme Val.*	*Beta*	*Extreme Val.*
140-150	Square	1162	Extreme Val.	Extreme Val.	Pearson V
	Sturges	*1162*	*Extreme Val.*	*Extreme Val.*	*Pearson V*
150-160	Square	940	Extreme Val.	Beta	Beta
	Sturges	*940*	*Extreme Val.*	*Beta*	*Beta*
160-170	Square	405	Extreme Val.	Beta	Beta
	Sturges	*405*	*Triangular*	*Beta*	*Beta*

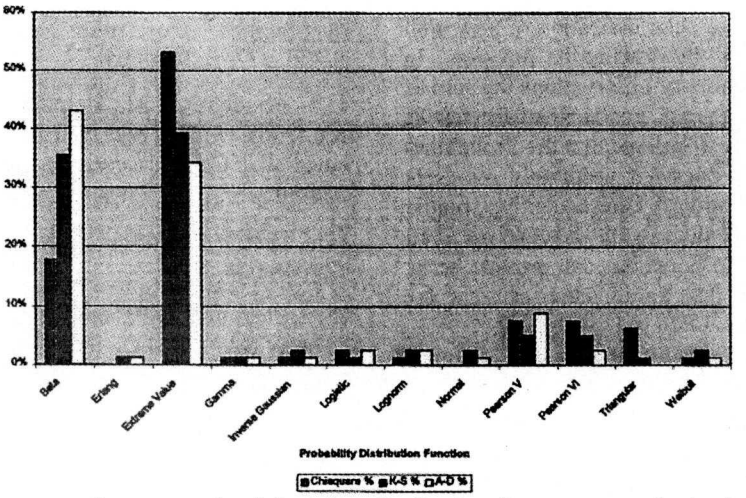

Figure 3. Percentages of aggregate haul time data distribution fits per test method using Sturges' rule.

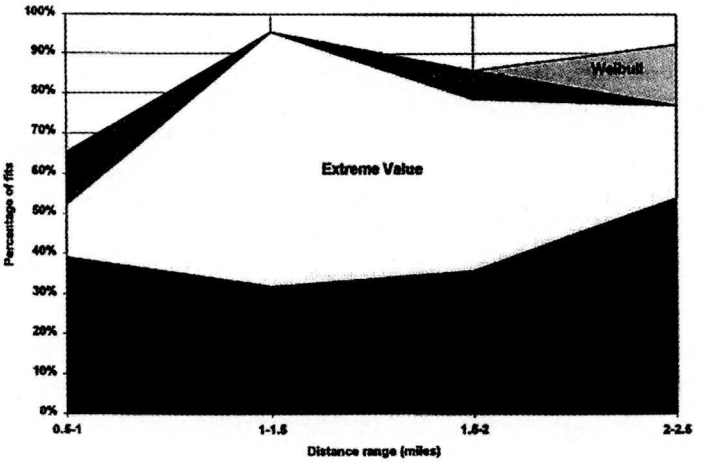

Figure 4. Effect of Haul distance on PDS selection, based on K-S distribution fit test and using Sturges' rule for class interval selection.

4 CONCLUSION

The need for reliable process control and simulation models of construction operations has been discussed and proposed by many researchers. Many of these same researchers have stated that the quality of the process control or simulation output is strictly related to the quality of the filtered data. Furthermore, it is necessary that the distribution functions used represent activity durations be appropriate. It is also a fundamental requirement that engineers understand the effect that a probability distribution function choice will have upon the controlled process.

Advances in computer technologies now permit the accumulation of construction process data in large quantities and the reduction of that data into appropriate distributions for use in real-time process control or for simulation modeling. The disadvantage of having such a large number of observations is that there are problems with matching large data sets to statistical PDFs using goodness-of-fit tests. Goodness-of-fit tests compare the proposed Probability Distribution Function to the input observations. When the number of observations is small the tests provide a clear indication of how well the observations match the distribution function. In the case of a large number of observations the sum of small variations can yield goodness-of-fit results which could cause one to believe that the Probability Distribution Function does not accurately represent the underlying population. Consequently, smaller data samples that are less well matched to their Probability Distribution Functions can provide better levels of confidence than large data sets that are better matched.

The research confirmed that the use of Chi-square fitting procedure is subjective regarding the choice of the number of class intervals. This produces results that, depending on the class interval used, differ almost 25% of the time a fitting procedure is attempted. The use of the Kolmogorov-Smirnov fitting test seems appropriate and gives results that fluctuate much less. The results of the fitting procedure seemed to be influenced by the number of observations that is included in the data set. Larger data sets, such as the load and dump cycle times, never returned the Beta distribution function as the first ranking distribution in the fitting procedure. On the other hand, limited data sets, such as haul and return cycle times, had the Beta distribution ranked first using BestFit on some occasions. A more in depth study is needed to evaluate the influence of the quantity of data on the results of the fitting procedure.

When fitted haul time distributions were plotted against haul distance no discernable trends were evident. It appears, therefore, that haul distance does not influence the selected the distribution function. The Beta distribution function appeared without peaks or lows, but in a small proportion, about 20%; the Extreme Value distribution appeared with higher fluctuation and higher percentage values.

The results of the fitting procedure appeared to be influenced by the dimension of the data sets. Large data sets returned theoretical distributions that were rejected by BestFit for their low confidence level. BestFit rarely rejected the results obtained from smaller sets. An investigation of the influence data set size has on fitting confidence may reveal better methods for evaluating the most appropriate method for fitting a theoretical distribution function to large data sets.

5 REFERENCES

(1) AbouRizk, S.M., and Halpin, D.W. (1990). "Probabilistic simulation studies for repetitive construction processes." *J. of Constr. Engrg. and Mgmt., ASCE, 116(4),* 575-594

(2) AbouRizk, S.M., Halpin, D. W., and Wilson J. R. (1991). "Visual interactive fitting of beta distributions." *J. of Constr. Engrg. and Mgmt., ASCE, 117(4),* 589-605

(3) AbouRizk, S.M., Halpin, D. W., and Wilson J. R. (1994). "Fitting beta distributions based on sample data." *J. of Constr. Engrg. and Mgmt., ASCE, 120(2),* 288-305

(4) *BestFit Software on Line Help.* (1993). Palisade Corporation, NY

(5) Clemmens, J.P., and Willenbrock, J.H. (1978). "The SCRAPESIM Computer Simulation." *J. of the Constr. Division, ASCE, 104(CO4),* 419-435

(6) D'Agostino, R.B., and Stephens, M.A. (1986). *Goodness-of-fit techniques.* Marcel Dekker Inc., NY

(7) Law, A.M. and Kelton, W.D. (1991). *Simulation modeling and analysis.* McGraw-Hill, NY

(8) Montgomery, D.C., Runger, G.C. (1994). *Applied Statistics and Probability for Engineers.* John Wiley, NY

(9) Phillips, D.T. (1972). "Applied goodness-of-fit testing." *Publication No.1, Operation Research Division,* American Institute of Industrial Engineers, Atlanta, Georgia