SELECTION OF A PROBABILTY DISTRIBUTION FUNCTION FOR CONSTRUCTION COST ESTIMATION

Rifat Sonmez[1]

[1]*Department of Civil Engineering, Middle East Technical University, Ankara, 06531, Turkey*

Probabilistic cost estimating techniques could be used during budgeting of construction projects. One of the difficulties in using probabilistic cost estimation techniques is to determine the probability distribution function for the project cost. Although assumptions are made about the probability distribution function of the project cost, there are only few studies in which the fit of data to the assumed distribution functions are actually studied. In this study cost data compiled from building projects are fitted against major probability distribution functions and the goodness of fit of the fitted functions are compared. A model distribution function is selected for the data and a probabilistic cost estimation is performed with the selected distribution.

Keywords: Cost estimation, probabilistic modeling, probability distribution function

INTRODUCTION

Probabilistic cost estimating techniques for construction projects could be used for several purposes including determination of the probability of a budget overrun, determination of the contingency amount, or determination of project ceiling price. The main focus of the probabilistic cost estimating techniques is to establish a probability distribution function (pdf) for the project cost. Once a pdf is selected cost estimation could be performed by calculating the cost estimates for different probability levels or by using simulation techniques. Although selection of a pdf for project cost is crucial for probabilistic cost estimation there are only limited number of studies in which the fit of major distribution functions for the actual projects costs are compared.

Data obtained from building projects were used by Touran and Wiser (1992) to conclude that lognormal distribution fits project cost data better than other commonly used pdfs including normal and beta distributions. Wang (2002) on the other hand used beta distribution function for probabilistic modeling of project ceiling price. Isidore, Back, and Fry (2000) used an empirical probability distribution function technique for integrated probabilistic cost and schedule estimation and argued that fitting a theoretical distribution function to a data set might not be an easy task.

Software developed in recent years is capable of assessing the goodness-of-fit of a data set to theoretical probability distribution functions. The software could also optimize the parameters for each theoretical distribution function. These developments allow fit and comparison of several pdfs to the data sample, which may aid to better selection of the model pdf. The main focus of this study is to study fit of

[1] rsonmez@metu.edu.tr

Sonmez, R (2004) Selection of a probability distribution function for construction cost estimation. *In:* Khosrowshahi, F (Ed.), *20th Annual ARCOM Conference*, 1-3 September 2004, Heriot Watt University. Association of Researchers in Construction Management, Vol. 1, 125-8.

major probability functions for the cost data of construction projects to provide input for the selection of a pdf function for probabilistic cost estimation.

PROJECT DATA

The data used in this study were acquired from thirty retirement community building projects of a US contractor. The projects were constructed in fourteen different states over a twenty-year period. The $/m^2 unit cost for each project was calculated by diving the project cost to the total building area. Historical cost indices were used to adjust the unit cost of projects to a reference year and, location indices were used to adjust the unit costs to a reference location (Means 1996). The inflation and location adjusted unit cost data for the thirty projects were than used for probabilistic modeling.

FIT OF PDFS

The cost data of thirty projects were compared with the ten major theoretical pdfs including normal, triangular, lognormal, uniform, exponential, weibull, beta, rayleight, logistic, and extreme value distributions with the use of a probability distribution fitting software. The software used optimizes the parameters for the probability distribution functions and provides goodness-of-fit test statistics. Goodness-of-fit tests are used to asses how well the theoretical pdf fits to the data sample. Three tests are generally used to evaluate goodness-of-fit: chi-square, Kolmogorov-Smirnov (K-S) and Anderson-Darling (A-D) tests.

In chi-square test, data is grouped and intervals need to be determined to evaluate the goodness-of-fit. This is an important limitation of chi-square test since there are no clear guidelines for selection of the intervals and test results may change depending on the selection of intervals. The K-S and A-D tests on the other hand do not require grouping of the data or determination of intervals. One of the major limitation of K-S test is that it does not detect the discrepancies at tails very well however the A-D test is mainly designed to detect the discrepancies in tails (Law and Kelton, 1991). When all of the three test results are used simultaneously a good picture for the goodness-of-fit could be observed.

The goodness-of-fit of the ten pdfs for the project unit cost data are given in Table 1. Beta distribution has the lowest chi-square (2.345) and A-D statistics (0.180) calculated among the ten theoretical pdfs fitted. Using the chi-square *P* value 0.8 it could be concluded that there is not sufficient evidence to reject the hypothesis that cost data has a beta distribution at *P* = 0.8 significance level. Beta distribution fit also gives a reasonably low K-S value (0.086) although this value is slightly higher than the value of triangular (0.081), weibull (0.082), normal (0.084) and logistic (0.085) distribution K-S values. As a conclusion beta distribution is selected for probabilistic cost estimation for the project cost data, however it should be pointed out that weibull, triangular, and normal distributions also provided reasonably good fits to the cost data.

PROBABILISTIC COST ESTIMATION

The plot of the data and fitted beta distribution is given in Figure 1. Probabilistic cost estimation could be performed using the fitted beta distribution. As an example a cost estimate with a probability of 90% would be $940.2 per m^2. If this estimate is used there is 90% chance that the actual cost would be lower than estimated value of $940.2 per m^2. A range estimate that would quantify possible variations in the project cost could also be developed using the fitted beta pdf. A cost estimate with a 10% chance of being lower than the actual cost is calculated as $687.5 per m^2. A 10%-90% range estimate for the project cost would be $687.5 - $940.2 and there is 80% chance that the actual cost would be within the range.

Table 1: Goodness-of-fit of pdfs

Distribution	Chi-Square	*P*-Chi-Square	A-D	K-S
Beta	2.345	0.800	0.180	0.086
Weibull	2.586	0.764	0.184	0.082
Triangular	2.680	0.750	0.186	0.081
Normal	2.832	0.726	0.187	0.084
Lognormal	3.596	0.610	0.274	0.107
Logistic	3.726	0.590	0.249	0.085
Rayleigh	5.160	0.397	0.555	0.152
Extreme Value	5.640	0.343	0.490	0.140
Uniform	7.051	0.217	1.032	0.170
Exponential	23.620	0.000	3.734	0.277

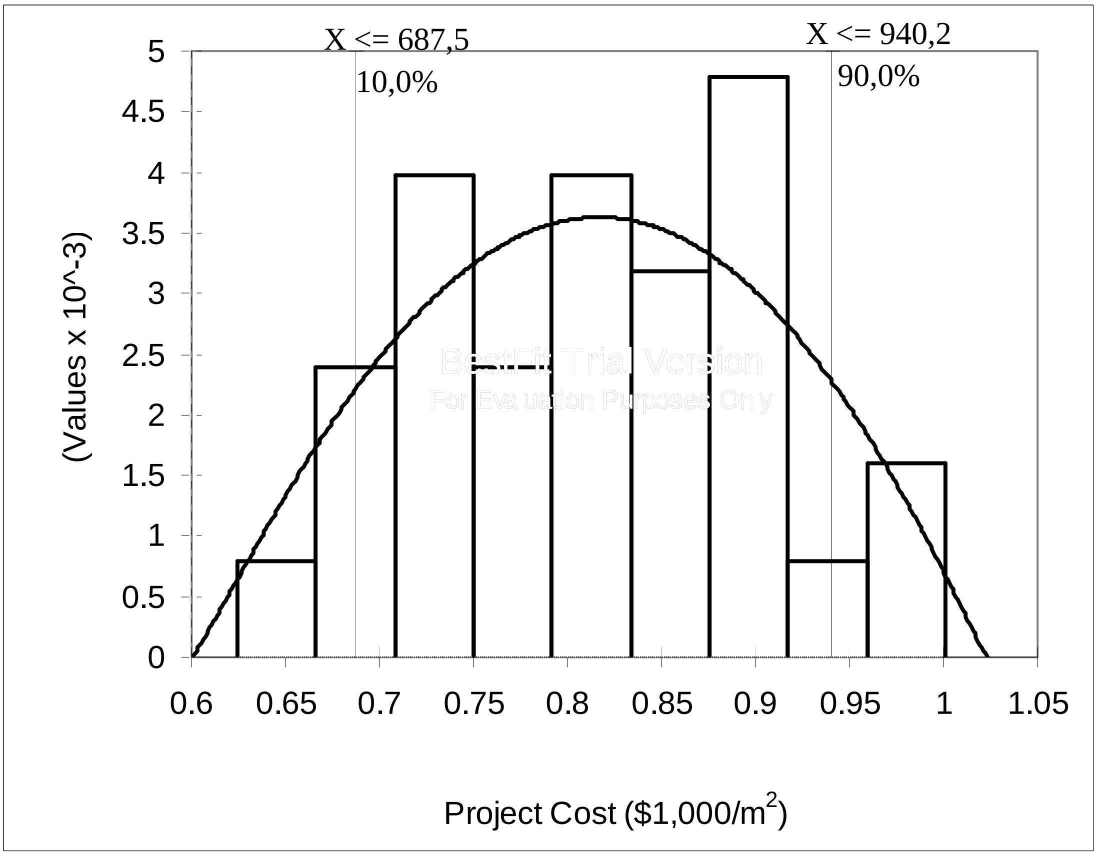

Figure 1. Plot of project cost data and fitted beta distribution

The range estimate developed could be useful especially during early stages of a project when initial budget considerations are made. The range estimate provides cost estimates at different probability levels which would more helpful than a point estimate for budgeting purposes. As an example an owner organization who would like to be 90% sure that their initial budget would not be exceeded could use $940.2 per m2 as the budgeted amount for the project.

CONCLUSIONS

Beta distribution provided the best fit to the project cost data, however weibull, triangular, and normal distributions also provided reasonably good fits. The methodology used in this study could be used for selection of probability distribution functions for project costs. The goodness-of-fit of the distributions depends on the characteristics of data set used. The total project cost for building projects could be divided in to systems such as civil, electrical and mechanical costs. At the system level different pdfs may be selected for different systems depending on the fit results. In this study a data set compiled from thirty building projects were used to determine the pdf of project cost. More research is needed to conclude on the specific probability distribution functions for probabilistic cost estimation of construction projects.

REFERENCES

Isidore, L. J., Back, W. E., Fry, G. T. (2001). Integrated probabilistic schedules and estimates from project simulated data. *Construction Management and Economics*, **19**, 417-426.

Law, A.M., and Kelton, W.D. (1991). Simulation modeling and analysis, McGraw-Hill, New York.

Means Building Construction Cost Data. 1996. R.S. Means Company, Kingston, MA.

Touran, A., and Wiser E. (1992). Monte Carlo technique with correlated random variables. *Journal of Construction Engineering and Management*, **118**(2), 258-272.

Wang, W. C. (2002). SIM-UTILITY: Model for project ceiling price determination. *Journal of Construction Engineering and Management*, **128**(1), 76-84.