PROBABILISTIC MEASUREMENT OF RISK ASSOCIATED WITH INITIAL COST ESTIMATES

Seokyon Hwang[1]

[1] Assistant Professor, Lamar University
Correspond to seokyon.hwang@lamar.edu

ABSTRACT: Accurate initial cost estimates are essential to effective management of construction projects where many decisions are made in the course of project management by referencing the estimates. In practice, the initial estimates are frequently derived from historical actual cost data, for which standard distribution-based techniques are widely applied in the construction industry to account for risk associated with the estimates. This approach assumes the same probability distribution of estimate errors for any selected estimates. This assumption, however, is not always satisfied. In order to account for the probabilistic nature of estimate errors, an alternative method for measuring the risk associated with a selected initial estimate is developed by applying the Bayesian probability approach. An application example include demonstrates how the method is implemented. A hypothesis test is conducted to reveal the robustness of the Bayesian probability model. The method is envisioned to effectively complement cost estimating methods that are currently in use by providing benefits as follows: (1) it effectively accounts for the probabilistic nature of errors in estimates; (2) it is easy to implement by using historical estimates and actual costs that are readily available in most construction companies; and (3) it minimizes subjective judgment by using quantitative data only.

Key words: Cost estimating; Risk management; Bayesian probability model.

1. INTRODUCTION

Predicting initial cost estimates is one of primary elements of construction project management. Many decisions are made based on the estimates in the course of management; to that end, the accuracy of initial cost estimates can affect overall performance of a project. An analysis of 290 sets of historical labor cost data that is conducted in the present study has revealed a large degree of deviation between the actual costs and the initial estimates—approximately a half of the samples missed the mark by plus and minus 50 percent of deviation.

Various methods have been proposed to date for cost estimating, to name a few notable previous efforts: Flood (1997) combined various distributions of cost per item into a single probability density function so as to model uncertainty in construction cost at project level; Haas and Einstein (2002) presented a method for predicting range estimates of construction costs for tunneling projects; Touran (2003) modeled random occurrence of change orders to estimate contingency costs. Other notable previous works include cost prediction models that are based on long-term trend of cost change (Dawood and Molson 1997; Wilmot and Cheng(2003). Despite the importance of accurate initial cost estimates, the problem of accuracy, however, remains unsolved (Trost and Oberlender 2003).

While there are various methods for cost estimating, standard statistical techniques have been accepted and applied the most widely in the construction industry. The standard techniques normally use basic statistics (mean and standard deviation) and probability distributions of historical actual costs. For instance, estimators incorporate probabilities and confidence intervals in the estimates based on the distributions of sample data. Meanwhile, it is not unusual to find stochastic nature from such historical data, which

was also identified by the afore-mentioned examination.

2. RESEARCH NEEDS AND OBJECTIVES

The above strongly suggests that estimators should carefully examine and measure how much potential error is associated with their estimates. In order to facilitate such a practice in the industry, there is a need for methods that can robustly analyze scattered, rather than correlated, data. Attempting to meet the need, this paper presents a quantitative method. The present study notes that the associated risk is conditional on the accuracy of initial estimates where the risk arises from erroneous estimates. The method adopts the Bayesian approach to measure the risk associated with initial cost estimates. Through the analysis of historical cost data collected from real projects, the method takes into consideration the stochastic and conditional nature of estimates without dealing with various factors.

3. A PROPOSED METHOD

As noted earlier, the purpose of the proposed method is to measure risk associated with initial cost estimates. In the context of the present study, the risk is measured through the calculation of probabilities for estimate errors, given a selected initial estimate.

3.1 Data Preparation

The first step is to retrieve historical cost data—initial cost estimate (***e***) and actual cost at completion (***a***) of each activity from project database, which is followed by the quantification of accuracy. When it comes to data format, unit cost was chosen to account for various scale of the size of work. The accuracy is quantified based on the error rate (***er***) that is deviation between the two costs—a numeric representation of the calculation is represented by Eq (1).

$$er_i = \frac{a_i - e_i}{e_i},\ i = 1,2,\ldots,n \qquad (1)$$

where of the past project (i), the initial estimate is e_i and the actual cost at completion is a_i. As its definition implies, the smaller the absolute value of er_i is, the more accurate the initial estimate is. In addition, the negative value of er_i indicates overestimate and the positive underestimate.

3.2 Examination of Data

The second step consists of three sequential examinations of data as follows.

Outlier Test: Because outliers may greatly influence the result of analysis, it is necessary to find and remove them. Box-plot is a general statistical tool that is useful to identify outliers in data, using boundary values like Lower Quartile (Q_1, value of 25th percentile) and Upper Quartile (Q_3, value of the 75th percentile). Outliers are defined as data points falling beyond the range between Upper Limit (Q_3+*1.5*IQR*) and Lower Limit (Q_1-*1.5*IQR*), where inter-quartile range (*IQR*) is Q_3 minus Q_1 (Nolan and Speed 2000).

Correlation Check: The degree of association between two random variables ***e*** and ***a*** needs to be checked, which can be done by measuring correlation coefficient. High correlation indicates evidence of strong association, for which variables can be explained by a relational model. Similarly, if high correlation exists between ***e*** and ***a***, then both ***er*** and ***a***, given ***e***, can be predicted by a relational model that is built on the relationship between the two random variables ***e*** and ***a***.

Distribution Check: Let us assume that random variable ***a*** consists of subsets of *a*'s. If the probability distributions of *er* for the subsets are the same or similar, then, the probability of *er*, given a subset, is equal to the probability of *er* for the entire sample space of *a*. This means that if the probability distribution of error for each subset of *a* is consistent with and follows that for the entire samples, the probability for a certain error level for any subset will be more-likely similar to that of the whole samples. However, if such a condition of consistent distributions is not satisfied across subsets, the conditional probability of error rate should be taken into consideration.

3.3 Bayesian Probability Calculation

The Bayesian approach is an application of the calculation of conditional probability, for which

Bayes' theorem is used as the basis. Calculated probability is often called the Bayesian probability. Eq. (2) represents the theorem that is evolved from the definition of conditional probability and the law of total probability where the sample space is partitioned into n mutually exclusive and exhaustive events $C_1, C_2,\ldots,C_n$. Using the theorem, the conditional probability of a particular event C_j, given event C, is calculated from the probability of each event $C_1, C_2,\ldots,C_n$ and the conditional probabilities of C, given each event C_i, $i= 1,2,\ldots,n$ (Hogg and Craig 1995).

$$P(C_j \mid C) = \frac{P(C \cap C_j)}{P(C)} = \frac{P(C_j)P(C \mid C_j)}{\sum_{i=1}^{n} p(C_i)P(C \mid C_i)} \qquad (2)$$

The sample space that comprises all error rate values needs to be partitioned into n mutually exclusive and exhaustive subsets of er_i, $i= 1,2,\ldots,n$. Each subset er_i and a particular estimate (event) e correspond to C_i and C in the above Eq. (2), respectively. Of interest in this model is the conditional probability of er_j, given an estimate e. Using Bayes' theorem, one can calculate the conditional probability $P(er_j|e)$ from $P(er_i)$, probabilities of event er_i and $P(e|er_i)$, conditional probabilities of e, given each event er_i, $i= 1,2,\ldots,n$. Theoretically, er_i and e can be the value of point or of range, which is dependent on the interest of decision makers and the number of data sets or the size of sample.

3.4 Interpretation of Results

The implementation of the proposed method results in the conditional probabilities of potential errors, given a selected initial estimate. The resulting Bayesian probability is interpreted as the probability of the expected error rate er_i, given the selected initial estimate e, by which represents risk associated with the selected estimate. In other words, the initial estimate e has a chance to yield an error as much as er_i, and the chance is quantified by the Bayesian probability $P(er_j|e)$. From this we can find out a cumulative probability, $P(er \leq er_j|e)$.

4. AN APPLICATION EXAMPLE

The proposed method was developed working with historical cost data collected from real projects executed by a general contractor. The following example explains the step-by-step procedure of the method. Hypothesis tests were conducted to compare the method with the existing standard distribution-based technique that is most widely used by practitioners.

An activity, forming concrete wall, is selected for the application example. Historical labor cost per square foot was collected from 30 projects. Error rates (er) of initial estimates are calculated for the project by using Eq. (1). Table 1 presents a complete set of samples.

Table 1. Samples of unit cost data

i	Estimate (e_i)	Actual (a_i)	Error Rate (er_i)
1	4.66	4.17	(0.10)
2	2.62	3.36	0.28
3	2.92	4.70	0.61
4	2.77	3.77	0.36
5	2.78	3.27	0.18
6	3.47	4.06	0.17
7	3.47	4.25	0.23
8	3.47	4.21	0.21
9	*3.47*	*7.39*	*1.13*
10	2.96	4.35	0.47
11	3.08	2.66	(0.14)
12	6.38	6.04	(0.05)
13	*5.51*	*13.45*	*1.44*
14	4.43	4.58	0.03
15	3.86	3.99	0.03
16	3.86	4.05	0.05
17	2.48	3.23	0.30
18	*2.48*	*7.31*	*1.95*
19	2.78	3.64	0.31
20	*2.78*	*7.04*	*1.53*
21	5.39	6.49	0.20
22	10.24	5.68	(0.45)
23	4.02	3.61	(0.10)
24	4.02	3.42	(0.15)
25	4.02	4.51	0.12
26	7.48	9.00	0.20
27	*6.60*	*23.29*	*2.53*
28	3.30	3.49	0.06
29	*3.30*	*8.46*	*1.56*
30	4.87	2.05	(0.58)

Note: i=sample number; number in parenthesis is negative; the underlined row is an outlier.

The results of a Box-Plot test identified six outliers out of 30 samples—sample number 9, 13, 18, 20, 27, and 29 (refer to Table 1). The correlation coefficient between ***e*** and ***a*** was measured to be 0.66. Additional checks were conducted regarding ***er*** versus ***e*** and ***er*** versus ***a***, and the results were -0.56 and 0.17 respectively. All of these do not show significant relationships, which suggests relational models may not be appropriate for this problem.

In this example, it was assumed that ***er*** has six ranges, considering its sample size and sample variance. The two chosen subsets of ***a*** are subset 1($3.0 < a \leq 4.0$) and subset 2 ($4.0 < a \leq 5.0$.) Fig. 1 illustrates the cumulative probabilities of *er's* for the two subsets and that for the whole ***a***.

Paired t-tests were conducted regarding cumulative probabilities of each error rate for the pairs: subset 1 and subset 1, subset 1 and the whole, and subset 2 and the whole. Given that alpha of 0.1 and 0.2 and degree of freedom of 5, *critical values* of two-sided test are $\pm\ t_{0.1,5} = \pm$ 2.015 and $\pm\ t_{0.2,5} = \pm$ 1.476 whereas resulting *t-values* for each test are 5.000, 1.083, and 2.007. Based on test results, consistent distribution of ***er*** against ***a*** is not obvious. The results of correlation and distribution checks support the choice of Bayesian approach to solving the research problem. Given the constructed sample space and subsets, probability of various error rates (er_i's) for a given estimate (e) can be calculated. For instance, assuming that the interested event of er_3 ($-0.25 < er_3 \leq 0.00$) and a given range estimate $4.0 < e_3 \leq 5.0$, the conditional probability of er_3, given the e_3 is mathematically expressed as $P(-0.25 < er_3 \leq 0.00 | 4.0 < e_3 \leq 5.0)$ using the Bayesian probability model. By repeating the same procedure for other error rates, all the conditional probabilities of six error rates and cumulative probabilities can be computed.

The results are interpreted as follows. For the given initial estimate e_3 ($4.0 < e_3 \leq 5.0$), there is a 67 percent chance that the given estimate can have an error rate less than 0.00; in other words, the actual cost will be within the budget estimate with 67 percent chance (Fig. 2).

5. HYPOTHESIS TESTS

Comparing the two distributions in Fig. 2, we can find that the probability distributions resulting from the two approaches are different. This test result concludes that relational models might not be appropriate methods for assessing risks associated with the initial cost estimates considering lack of correlations between the results from the two methods.

Another statistical test was conducted to examine the consistency of the probability distributions of ***er*** for ***a***, by which the belief of consistency was rejected again. Unlike the standard distribution-based techniques based on only ***a***, the new method considers the fact that $P(er_i)$ could be different for the various given estimates. This implies that resulting probabilities by two different methods should be statistically different. It can be examined by testing the null hypothesis: the probabilities of error rate calculated by two methods are same.

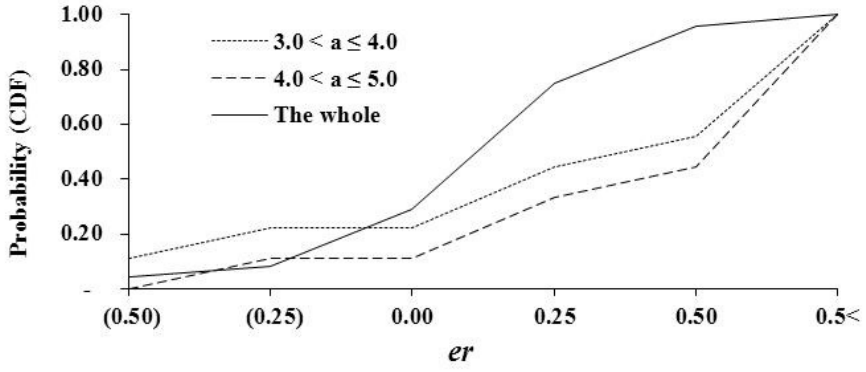

Fig. 1 Probability distributions of ***er*** for two subsets and the population

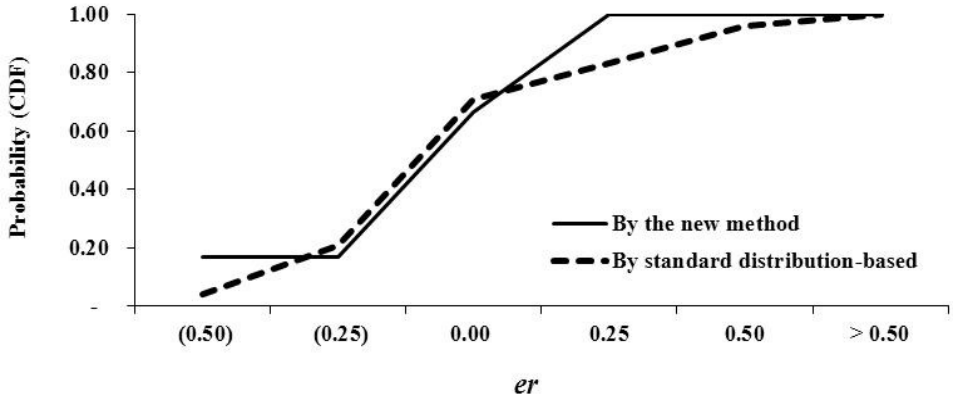

Fig. 2 Comparison of probability distribution

The second test used two-sided paired t-test to examine the cumulative probabilities that are calculated by two methods for the same estimates. Six estimates falling in the range $4.0 < e_3 \leq 5.0$ are used, and six error rates are applied to each estimate. Thus, there is a total of thirty six paired samples of cumulative probabilities. The results of two-sided paired t-test (*degree of freedom = 35; two given alphas = 0.1, 0.2; calculated t-value of test = ± 3.953; critical values of two-sided tests,* $\pm t_{0.1,35} = \pm 1.306$, $\pm t_{0.2,35} = \pm 1.689$) provide evidence to reject the null hypothesis.

6. CONCLUSIONS

In managing a construction project, the initial cost estimates are important early decisions that are made prior to construction using limited information at the time. The accuracy of the estimates is critical where the quality of the following decision-making processes is highly dependent on it. Therefore, lack of accuracy of the estimates adds risk to a project. Despite many methods developed to improve the accuracy, the situation to date has not been satisfactory. The analysis of historical data in the present study revealed that errors in estimates are significantly problematic. Under the circumstances, the industry needs a method that can evaluate potential risk associated with initial estimates due to the inherent errors from estimating. Knowing such risks in advance, project managers can take them into consideration, so they can prepare realistic targets while managing their projects.

Few methods are available for this purpose, and most of them do not effectively account for the probabilistic nature of errors in cost estimates. As a result of this study, a method was developed based on the Bayesian approach to effectively evaluate and measure risk associated with the selected initial cost estimate. The proposed method takes into consideration the probability of estimate errors. The method considers the distribution of actual costs as a prior belief and the conditional probability of the estimate errors as a posterior belief. Thereby, the Bayesian probability model enables project managers to calculate the conditional probabilities of various error levels, given the selected initial estimate. The hypothesis tests proved the appropriateness of the Bayesian probability model.

The new method is a practical tool that can be easily applied to analyze risk in the initial cost estimates. The method is envisioned to provide a few significant advantages while complementing methods that are currently in use. The advantages include: (1) it effectively accounts for the probabilistic nature of errors in estimates through its easy step-by-step procedure; (2) it minimizes the involvement of subjective judgment being objectively built upon actual historical data; (3) it is easy to acquire data for the implementation of the method because it uses only historical

estimates and actual costs at completion which are readily available in most construction companies.

REFERENCES

Dawood, N. and Molson, A. (1997). "An Integrated approach to cost forecasting and construction planning for the construction industry." Proc., 4th Congress on Computing in Civil Engineering, ASCE, New York, New York, 535-542.

Flood, I. (1997). "Modeling uncertainty in cost estimates: a universal extension of the central limit theorem." Proc., of 4th congress on Computing in Civil Engineering, ASCE, New York, New York, 551-558.

Haas, C. and Einstein, H. H. (2002). "Updating the decision aids for tunneling." Journal of Construction Engineering and Management, ASCE, 128 (1) 40-48.

Hogg, R. V. and Craig, A. T. (1995). Introduction to mathematical statistics, 5th Ed., Prentice-Hall, Inc., Upper Saddle River, New Jersy.

Nolan, D. and Speed, T. (2000). Stat labs - Mathematical statistics through applications. Springer, New York, New York.

Touran, A. (2003). "Probabilistic model for cost contingency." Journal of Construction Engineering and Management, ASCE, 129(3), 280-284.

Trost, S. M. and Oberlender, G. D. (2003). "Predicting accuracy of early cost estimates using factor analysis and multivariate regression." Journal of Construction Engineering and Management, ASCE, 129 (2), 198-204.

Wilmot, C. G. and Cheng, G. (2003). "Estimating future highway construction costs." Journal of Construction Engineering and Management , ASCE, 129(3), 272-279.