DOI NUMBER: 10.19255/JMPM01210

KEYWORDS Software cost estimation Software risk Project management Statistical computing

COST ESTIMATION OF SOFTWARE PROJECTS: A SUB-ADDITIVE APPROACH

Dr. Masood Uzzafer

Project Management Consultant
uzzafer@alumni.nottingham.ac.uk

ABSTRACT

Sub-additivity in estimations suggests that the sum of sub-estimates is a good approximation of the overall estimate; alternatively, an overall estimate decomposes into well represented sub-estimates. Traditionally from probability distributions, estimated costs are percentiles of probability distributions; however, such estimates may not be sub-additive. This paper presents a model which produces sub-additive cost estimates from probability distributions. The proposed model relies on expectations as oppose to percentiles of probability distributions. For bottom-up cost estimation scenarios, the proposed model ensures that sub-estimates are sub-additive such that the sum of sub-estimates is a good approximation of the overall cost. For top-down cost estimation scenarios, the model ensure that the overall estimate is sub-additive and decomposing the overall estimate into sub-estimates provides a good representation of sub-estimates. A case-study proves that the proposed model produces sub-additive estimates for bottom-up and top-down cost estimation scenarios while percentile based estimates are susceptible to sub-additivity. Violation of sub-additivity contributes towards under-estimation of sub-estimates for bottom-up scenarios and over-estimation of overall estimates for top-down scenarios. Therefore, sub-additivity consideration is critical in estimation which helps to avoid understated or overstated estimates.

1. Introduction

The cost is the man-months effort required to complete a software project (Navlakha, 1990). The cost of software projects is measured in terms of lines-of-code or function point count which are converted to man-month effort (Sommerville, 2007). Standish's report (Galorath, 2008; Haughey, 2009) stated that software projects are notorious for cost overruns; the report further revealed that a noticeable number of software projects showed underrun cost trends. Under-estimation causes actual cost to over-run the estimate; whereas, over-estimation causes actual cost to under-run the estimate. Under-estimated costs cause lack of project contingency reserves (Uzzafer, 2013b) and over-estimated costs lead to losses of potential business opportunities (Moataz, 2013).

Cost estimates are prone to uncertainty; uncertainty is well represented with probability distributions (Kitchenham, et al., 1997). Probability distributions capture a range of random cost estimates to counter the uncertainty. Each random estimate has an associated probability which adds a probabilistic confidence to estimates. Traditionally, estimated costs of software projects are percentiles of probability distributions (Moataz, et al., 2013). However, Acerbi et al. (2001, 2003) and Yamai, (2005) explained that percentiles of distributions may not always produce sub-additive results; especially, the percentiles of non-parametric distributions are susceptible to violate sub-additivity. Such estimates raise following questions about the integrity of estimates: is aggregating estimated sub-estimates produces a good approximation of the overall estimate or whether decomposing the estimated overall estimate into sub-estimates provides a good representation of the sub-estimates. In general are estimates sub-additive? Therefore, percentile based estimated cost from probability distributions may lead to misleading estimates since estimates may not be sub-additive. Sub-additivity is well established in the field of finance (Artzner et al., 1999).

Software practitioners experience that the cost of a portfolio of software projects is less or not more than the sum of the costs of all the software projects within the portfolio (Kitchenham et al. (2003), Bannerman (2008), Abdul-Rahmana et al. (2012), Costa et al. (2007); this is sub-additive behavior of cost of software projects.

There are two cost estimation scenarios: namely bottom-up and top-down (Pfleeger, et al., 2006). The bottom-up cost estimation technique focuses on estimating sub-estimates which are aggregated to get the overall cost. While, the top-down cost estimation technique requires estimating the overall cost which is then decomposed into sub-estimates.

1.1 Related Work

There are various cost estimation models available for software development projects (Pfleeger and Atlee, 2006; Robert et al. 2002). Common approaches to software cost estimation are: expert judgment (Jorgensen, 2005; Jorgensen et al. 2007), analogy based estimation, e.g. (Li, et al. 2007; Shepperd et al. 1997), algorithmic models like COCOMO and COCOMO-II (Boehm 2000) and SLIM (Putnam, 1978), machine learning techniques like Bayesian belief networks (Hamdan, et al. 2009; Leea, et al., 1998), fuzzy logic (Nisar, et al., 2008) and artificial intelligence (Park et al. 2008). These models produce single-value estimates of costs (Lum et al., 2003; Dillibabu and Krishnaiah, 2005; Karen et al., 2003; Evelyn et al., 2002); however, single-value estimates are uncertain. Costs are represented with probability distributions to handle uncertainty (Kitchenham et al., 1997; Stein et al., 2006, Uzzafer, 2013b). Researchers in the field of software engineering are proposing different probability distributions to represent the cost of software projects (Kitchenham, et al. 2003, 1997; Moataz, et al, 2003, Pendharkar et al. 2005). Kitchenham et al. (1997, 2003) suggested gamma distribution represention for the cost of software projects. Gamma distribution is continuous having an extended tail on the right side of the distribution. Gamma distribution is represented as $\Gamma(k,\theta)$ where k and θ are the shape and the scale parameters of the distribution, respectively. The expectation of gamma distribution is defined as follows: $E[\Gamma(k,\theta)] = k\theta$ [Appendix B]. Fairley (1995) and Connor (2005) adopted Monte Carlo simulations which generates discrete non-parametric probability distributions to represent costs of software projects. Moataz et al. (2013) and Braga (2007) relied on the Gaussian probability distribution to reduce the uncertainty in the estimate; whereas, Parag (2005) proposed a Bayesian probabilistic model and Kathleen (2012) adopted the Weibull probability distribution to represent the cost of software projects. Jørgensen et al. (2004a) investigated the causes of estimation errors which leads to overestimation and underestimation of the cost of software projects. Moløkken-Østvold et al. (2005) explained that flexible software development models, i.e., incremental models, are better in dealing with cost overruns. In another study, Jørgensen et al. (2004b) investigated that the overconfidence plays a role in underestimation of efforts of software projects and concluded that the tendency to learn about maximum efforts from historical projects' data is low.

Stewart et al. (1995) classified software cost estimation models into three categories where each category can adopt a top-down or a bottom-up scenario for estimation. They further presented a top-down cost estimation scenario which explains that a software project of 10,000 lines will take 100 person-months based on a productivity of 100 lines of code per man-month. While, using the bottom-up scenario, the same software is decomposed into five components, The man-month effort for each component is estimated as 40.0, 3.3, 20.0 50.0 and 20.0 man-months where each component has a different productivity level. Therefore, the overall estimate for this project is the sum of the sub-estimates which comes to 133 man-months.

While, research continues to find probability models for a better representation of the cost of software development projects, not much has been contributed to ascertain the integrity of estimated costs that are originating from probability distributions. Therefore, cost overruns and underruns is a challenge for the development of software projects. Such unexpected overruns and underruns of costs are generally associated with the accuracy of the estimation models (Alkoffash, et al., 2008) while issues related to sub-additivity of estimates have not been thoroughly explored.

1.2 Probabilistic cost representation

Estimated cost of a software project is represented with a random variable X where X is mapped to a parametric continuous probability distribution. The estimated cost at probability q is xq, such that:

$$xq = supremum\ \{x{:}P\,[X \le x] \le q\} \qquad q \in [0,1] \qquad (1)$$

Where x is a realization of X at any probability, supremum is the upper-limit among all the values of X for which the probability $P[X \le x] \le q$. The random variable X that is mapped to a gamma distribution is represented as $X \sim \Gamma(k,\theta)$. Uzzafer (2013a) explained that a single-point estimate can be represented with gamma distribution where the estimate $\mathfrak{E}$ is mapped to the expectation of the gamma distribution, i.e., $\mathfrak{E} \mapsto \mathbb{E}[\,\Gamma(k,\theta)]$, then letting $k = 2$, θ can be estimated from $\mathfrak{E} = k\theta$ (Guo, 2010; Kitchenham, 1997).

Furthermore, discrete estimated costs are represented with a discrete random variable $\mathbb{X}_i$ which is mapped to a discrete probability distribution, where i is sample's index and $\mathbb{x}q_i$ is the estimated cost at probability q [**Appendix A**]:

$$\mathbb{x}q_i = supremum\ \{\mathbb{x}_i{:}P[\mathbb{X}_i \le \mathbb{x}_i] = q\} \qquad q \in [0,1] \qquad (2)$$

1.3 Sub-additivity

In mathematics, sub-additivity states that the result of a function applied to a whole should be less or at-most-equal to the sum of the results of the function applied to parts (Royden et. al., 2010). For example, a function $\Delta(\cdot)$ applied to $(A+B)$ produces $\Delta(A+B)$. Therefore, $\Delta(A+B)$ is sub-additive when it is less or at-most-equal to the sum of the results $\Delta(A)$ and $\Delta(B)$ of the function; i.e., $\Delta(A+B) \le \Delta(A)+\Delta(B)$ (Acerbi et al., 2001, 2003; Uzzafer, 2010b). Similarly, the results $\Delta(A)$ and $\Delta(B)$ are considered sub-additive when their sum leads to $\Delta(A+B) \le \Delta(A)+\Delta(B)$. In general, sub-additivity is expressed as follows: $\Delta\Sigma(\cdot) \le \Sigma\Delta(\cdot)$. Therefore, estimates of an estimation function $\rho(\cdot)$ are sub-additive when the overall estimate $\rho\Sigma(\cdot)$ is less or at-most-equal to the sum of sub-estimates $\Sigma\rho(\cdot)$ or vice-versa:

$$\rho\Sigma(\cdot) \le \Sigma\rho(\cdot) \qquad (3)$$

In bottom-up cost estimation scenarios, sub-estimates are first estimated then the overall estimate is the sum of the sub-estimates, i.e., $\Sigma\rho(\cdot)$. For such scenarios, sub-additivity ensures that the sum of sub-estimates is a good representation of the overall estimate. In a real software project, the overall estimate is not estimated rather it is approximated through the sum of sub-estimates; therefore, with sub-additivity assured the sum is a good representation of the overall estimate. Sub-additivity is violated when sub-estimates are under-estimated and their sum is an overall estimate which is under-estimated. Therefore, when sub-additivity is not supported, it can be deduced that the sum of sub-estimates does not represent the overall estimate. Consider the bottom-up scenario presented in **Figure 1**. The sub-estimates $\rho(A1)$, $\rho(A2)$, $\rho(B1)$ and $\rho(B2)$ are aggregated to get the overall estimate, i.e., $\Sigma\rho(\cdot)=\rho(A1)+\rho(A2)+\rho(B1)+\rho(B2)$. With sub-additivity supported the sum $\Sigma\rho(\cdot)$ is a good representation of the overall estimate; whereas, violation of sub-additivity results in the sum that does not represent the overall estimate.

Top-down cost estimation scenario focus on estimating the overall estimate first which is then decomposed into sub-estimates. For such scenarios, sub-additivity ensures that sub-estimates are well represented after decomposition of the overall estimate. Consider the top-down scenario presented in **Figure 1**; the overall estimate of $\rho(A1+A2+B1+B2)$ is decomposed to $\rho(A1)$, $\rho(A2)$, $\rho(B1)$ and $\rho(B2)$ which are a good representation of sub-estimates with sub-additivity ensured. Sub-additivity is violated when the overall estimate is over-estimated and the decomposition of the overall estimate into sub-estimated is not a good representation of sub-estimates. In real top-down scenarios, the overall cost is estimated and sub-estimates are the decomposition of the overall estimates with sub-additivity assured decomposition produces a good representation of sub-estimates.

Consider a software project of three tasks; the costs are represented with random variables $\mathbb{X}1_i$, $\mathbb{X}2_i$ and $\mathbb{X}3_i$. The probability distributions $f(\cdot)$ and cumulative distribution $F(\cdot)$ of these random variables are identical which are tabulated in **Table 1**.

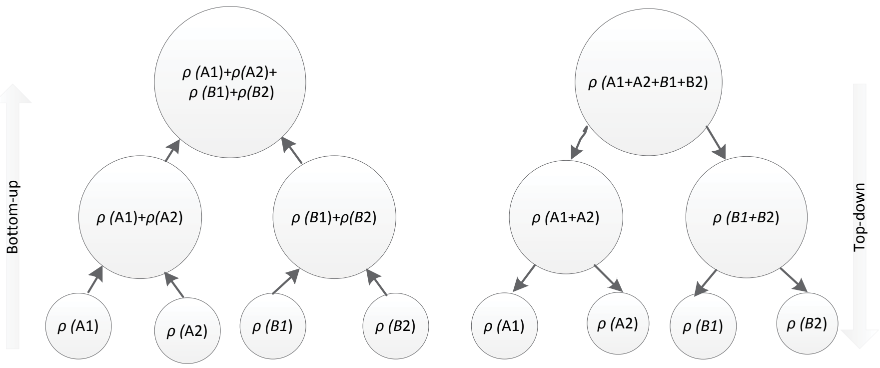

FIGURE 01. *Bottom-up and Top-down cost estimation scenarios*

The probability distributions explain that there is a 95% probability that a task can take 1 man-month of effort to complete while there is a 5% probability that the task can take up to 1.25 man-month. Therefore, using the percentile based approach for this bottom-up scenario, the sub-estimates at 90% probability from equation (2) are as follows:

$\mathbb{X}1_i, \mathbb{X}2_i, \mathbb{X}3_i$	$f(\mathbb{X}1_i), f(\mathbb{X}2_i), f(\mathbb{X}3_i)$	$F(\mathbb{X}1_i), F(\mathbb{X}2_i), F(\mathbb{X}3_i)$
1	0.95	0.95
1.25	0.05	1

TABLE 01. *Probability and Cumulative Distributions of* $\mathbb{X}1_i$, $\mathbb{X}2_i$, $\mathbb{X}3_i$ *(man-months)*

$\rho(\mathbb{X}1_i)$ = *supremum* $\{x_i : P[\mathbb{X}1_i \le x_i]=0.9\}$=1 man-months

$\rho(\mathbb{X}2_i)$ = *supremum* $\{x_i : P[\mathbb{X}2_i \le x_i]=0.9\}$=1 man-months

$\rho(\mathbb{X}3_i)$ = *supremum* $\{x_i : P[\mathbb{X}3_i \le x_i]=0.9\}$=1 man-months

Therefore, the estimated overall cost of the software project is the sum $\Sigma\rho(\cdot)$, i.e., $\rho(\mathbb{X}1_i) + \rho(\mathbb{X}2_i) + \rho(\mathbb{X}3_i)$ = 3 man-months. The sub-estimates $\rho(\mathbb{X}1_i)$, $\rho(\mathbb{X}2_i)$ and $\rho(\mathbb{X}3_i)$ may not be sub-additive and their sum $\Sigma\rho(\cdot)=\rho(\mathbb{X}1_i)+\rho(\mathbb{X}2_i)+\rho(\mathbb{X}3_i)$ may not represent the overall cost of $\rho\Sigma(\cdot)$. To test the sub-additivity, the random variables $\mathbb{X}1_i$, $\mathbb{X}2_i$ and $\mathbb{X}3_i$ are added which forms a random variable $(\mathbb{X}1_i+\mathbb{X}2_i+\mathbb{X}3_i)$. The probability distribution $f(\mathbb{X}1_i+\mathbb{X}2_i+\mathbb{X}3_i)$ of random variable $(\mathbb{X}1_i+\mathbb{X}2_i+\mathbb{X}3_i)$ is the convolution of the distributions $f(\mathbb{X}1_i)$, $f(\mathbb{X}2_i)$ and $f(\mathbb{X}3_i)$ (Papoulis, 1991). The probability distribution $f(\mathbb{X}1_i+\mathbb{X}2_i+\mathbb{X}3_i)$ and cumulative distribution $F(\mathbb{X}1_i+\mathbb{X}2_i+\mathbb{X}3_i)$ is tabulated in **Table 2**. Now, the estimated overall cost at 90% probability is from equation (2) is:

$\rho(\mathbb{X}1_i+\mathbb{X}2_i+\mathbb{X}3_i)$ = *supremum* $\{x\_i : P[(\mathbb{X}1_i+\mathbb{X}2_i+\mathbb{X}3_i) \le x_i]=0.9\}$= 3.25 man-months

These results show that the estimated overall cost from probability distribution $f(\mathbb{X}1_i+\mathbb{X}2_i+\mathbb{X}3_i)$ is more than the sum of the sub-estimates $\rho(\mathbb{X}1_i)$, $\rho(\mathbb{X}2_i)$ and $\rho(\mathbb{X}3_i)$, i.e., $\{\rho(\mathbb{X}1_i+\mathbb{X}2_i+\mathbb{X}3_i=3.25\}>\{\rho(\mathbb{X}1_i)+\rho(\mathbb{X}2_i)+\rho(\mathbb{X}3_i)=3\}$. It suggests that the sub-estimates $\rho(\mathbb{X}1_i)$, $\rho(\mathbb{X}2_i)$ and $\rho(\mathbb{X}3_i)$ are not sub-additive and their sum is not a good representation of the overall estimate.

This example can be reversed for top-down scenario where at 90% probability the estimated overall cost is $\rho\Sigma(\cdot)=\rho(\mathbb{X}1_i+\mathbb{X}2_i+\mathbb{X}3_i)$=3.25 man-months. This overall estimate is equally decomposed into estimates of 1.083 man-months for each task. However, when the sub-additivity is not supported these decomposed estimates may not represent the sub-estimates. To test the sub-additivity, the random variable $(\mathbb{X}1_i+\mathbb{X}2_i+\mathbb{X}3_i)$ is decomposed into random variables $\mathbb{X}1_i$, $\mathbb{X}2_i$ and $\mathbb{X}3_i$ and their probability distributions $f(\mathbb{X}1_i)$, $f(\mathbb{X}2_i)$ and $f(\mathbb{X}3_i)$ are de-convolved from the distribution $f(\mathbb{X}1_i+\mathbb{X}2_i+\mathbb{X}3_i)$ which are tabulated in **Table 1**. The sub-estimates are estimated at 90% probability from probability distributions $f(\mathbb{X}1_i)$, $f(\mathbb{X}2_i)$ and $f(\mathbb{X}3_i)$ which produces $\rho(\mathbb{X}1_i)=\rho(\mathbb{X}2_i)=\rho(\mathbb{X}3_i)$=1 man-months. These results explains that the decomposed estimates of 1.083 man-months are more than the estimated sub-estimates of 1 man-months which violates sub-additivity. Therefore, the overall estimate is not sub-additive and is over estimated; hence, the decomposed estimates are over estimated. Therefore, this scenario experiences $\{\rho(\mathbb{X}1_i+\mathbb{X}2_i+\mathbb{X}3_i)=3.25=1.083+1.083+1.083\}>\{\rho(\mathbb{X}1_i)+\rho(\mathbb{X}2_i)+\rho(\mathbb{X}3_i)=1+1+1\}$.

Sub-additivity of estimates is critical which protects against the under-estimation of sub-estimates in bottom-up and over-estimation of overall estimate in top-down scenario. This paper presents a model to estimate the cost of software projects from the probabilistic representation of cost and aims to produce sub-additive estimates. The proposed model exploits expectations of probability distributions.

The second section of this article presents the model and discusses examples. The third section presents a case study where the proposed model is deployed using the estimated cost data of real software projects. The case study proves that the model generates sub-additive estimates. Finally, the fourth section draws some conclusions.

2. The Model

The proposed model aims to ensure sub-additive estimates of costs. Furthermore, the proposed model aims to be generic and independent of the type and the shape of the adopted probability distribution.

While the percentiles of probability distributions may not be sub-additive, the expectations of probability distributions are sub-additive (Lange, 2003). A random variable $(X1+...+Xn)$ of expectation $\mathbb{E}[X1+...+Xn]$ can be decomposed into random variables $X1,...,Xn$ such that $\mathbb{E}[X1+...+Xn] = \mathbb{E}[X1]+...+\mathbb{E}[Xn]$, which explains that expectations are sub-additive (Lange, 2003). Applying this to the estimation suggests that the expectation $\mathbb{E}[X1+...+Xn]$ of a random variable $(X1+...+Xn)$ which represents cost should be less or at most equal to the sum of expectations $\mathbb{E}[X1]+...+\mathbb{E}[Xn]$ of random variables $X1,...,Xn$ which represents sub-estimates:

$$\mathbb{E}[\Sigma\cdot] \le \Sigma\,\mathbb{E}[\cdot] \qquad (4)$$

$(\mathbb{X}1_i + \mathbb{X}2_i + \mathbb{X}3_i)$	$f(\mathbb{X}1_i + \mathbb{X}2_i + \mathbb{X}3_i)$	$F(\mathbb{X}1_i + \mathbb{X}2_i + \mathbb{X}3_i)$
3	0.857375	0.857375
3.25	0.135375	0.9927
3.5	0.00713	0.9998
3.75	0.000125	1

TABLE 02. *Probability and Cumulative Distribution* $\mathbb{X}1_i$, $\mathbb{X}2_i$, $\mathbb{X}3_i$ *(man-months)*

The model defines a range of random variables. Consider a random variable X and let xw be the minimum and xa be the maximum range with probabilities of w and a, respectively, $w, a \in [0,1]$. This range forms a bounded random variable $X_{[xw,xa]}$. **Figure 2** illustrates a probability distribution X and highlights the range $X_{[xw,xa]}$.

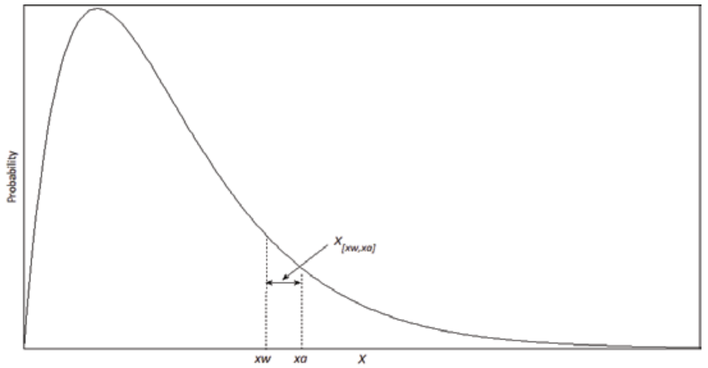

FIGURE 03. *A parametric probability distribution representing the estimated cost X (man-months)*

The proposed model defines the estimated cost as the expectation of bounded random variable $X_{[xw,xa]}$; therefore, the estimated cost $\rho(X)$ is the mapping of $X_{[xw,xa]}$ to its expectation, as expressed as follows:

$$\rho(X): X_{[xw,xa]} \; \mathbb{E}[X_{[xw,xa]}] \qquad (5)$$

From equation (5), the following computational model is constructed:

$$\rho(X)=\frac{\int_{w}^{a} x\,f(x)\,dx}{\int_{w}^{a} f(x)\,dx} \qquad (6)$$

Where x is a value of X and $f(x)$ is the probability distribution of X. Similarly, for a discrete random variable $\mathbb{X}_i$, the bounded random variable $X_{[xw,xa]}$ is between $\mathbb{x}w_i$ and $\mathbb{x}a_i$ with probabilities w and a, respectively; the random variable $X_{[xw_i,xa_i]}$ is shown in **Figure 3**.

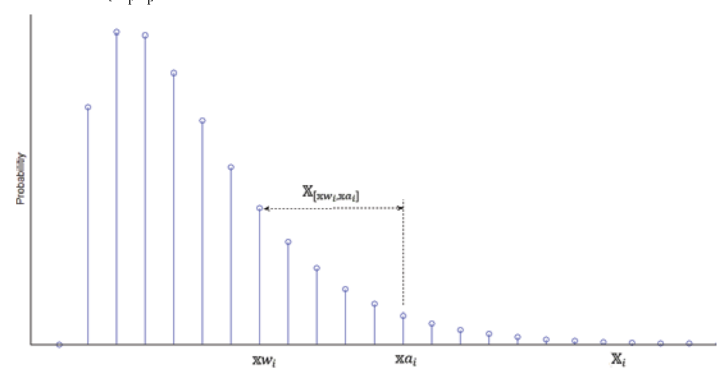

FIGURE 03. *Discrete Probability Distribution* X_i *(man-months)*

For discrete probability distribution X_i, the estimated cost $\rho(\mathbb{X}_i)$ is the following mapping:

$$\rho(\mathbb{X}_i): \mathbb{X}_{[xw_i,xa_i]} \mapsto \mathbb{E}[\mathbb{X}_{[xw_i,xa_i]}] \qquad (7)$$

Equation (7) results in the following computational model:

$$\rho(\mathbb{X}_i)=\frac{\sum_{i=iw}^{ia} \mathbb{x}_i f(\mathbb{x}_i)}{\sum_{i=iw}^{ia} f(\mathbb{x}_i)} \qquad (8)$$

Where i_w and i_a are the indexes of $\mathbb{x}w_i$ and $\mathbb{x}a_i$, respectively, and $f(\mathbb{x}_i)$ is the probability distribution of $\mathbb{X}_i$. Note that the estimated probability of $\mathbb{x}a_i$, i.e., $P\{\mathbb{X}_i \le \mathbb{x}a_i\}$ may be more than the probability a, i.e., $P\{\mathbb{X}_i \le \mathbb{x}a_i\}>a$ (Acerbi, 2003); similarly, the probability $P\mathbb{X}_i>\mathbb{x}w_i\}$ may be less, i.e., $P\{\mathbb{X}_i>\mathbb{x}w_i\}<1-w$ **[Appendix A]**. These values are adjusted in equation (8) resulting in the following:

$$\rho(\mathbb{X}_i)=\frac{\sum_{i=iw}^{ia} \mathbb{x}_i f(\mathbb{x}_i)}{\sum_{i=iw}^{ia} f(\mathbb{x}_i)} - \mathbb{x}a_i\,(P\{\mathbb{X}_i \le \mathbb{x}a_i\} - a) + \mathbb{x}w_i\,(1-P\{\mathbb{X}_i > \mathbb{x}w_i\} - (1-w))$$

$$\rho(\mathbb{X}_i)=\frac{\sum_{i=iw}^{ia} \mathbb{x}_i f(\mathbb{x}_i)}{\sum_{i=iw}^{ia} f(\mathbb{x}_i)} - \mathbb{x}a_i\,(P\{\mathbb{X}_i \le \mathbb{x}a_i\} - a) + \mathbb{x}w_i\,(1-P\{\mathbb{X}_i \le \mathbb{x}w_i\} - 1+w)$$

$$\rho(\mathbb{X}_i)=\frac{\sum_{i=iw}^{ia} \mathbb{x}_i f(\mathbb{x}_i)}{\sum_{i=iw}^{ia} f(\mathbb{x}_i)} - \mathbb{x}a_i\,(P\{\mathbb{X}_i \le \mathbb{x}a_i\} - a) + \mathbb{x}w_i\,(w - P\{\mathbb{X}_i \le \mathbb{x}w_i\}) \qquad (9)$$

The model $\rho(\cdot)$ presented in equations (6) and (9) are independent of the shape of the underlying probability distributions which can be continuous, discrete, parametric and non-parametric.

Consider the bottom-up scenario example discussed in section 1 and estimate the cost at 90% probability using the proposed model; therefore, a=0.9 also assume w =0.9. The estimated values of $\mathbb{x}a_i$ and $\mathbb{x}w_i$ are $\mathbb{x}a_i=\mathbb{x}w_i$=1 man-months and their estimated probabilities are $P\{\mathbb{X}_i \le \mathbb{x}a_i\}= P\{\mathbb{X}_i \le \mathbb{x}w_i\}$=0.95, see **Table 1**. Note that the probability $(P\{\mathbb{X}_i \le \mathbb{x}a_i\}=0.95) > (a=0.9)$ and the probability $(P\{\mathbb{X}_i > \mathbb{x}a_i\}=1- P\{\mathbb{X}_i \le \mathbb{x}\,w_i\}=1-0.95=0.05) < (1-w=1-0.9=0.1)$. The sub-estimates at 90% probability using the model presented in equation (9) are calculated as follows:

$$\rho(\mathbb{X}1_i)=\rho(\mathbb{X}2_i)=\rho(\mathbb{X}3_i) = \frac{\sum_{i=iw}^{ia} \mathbb{x}_i f(\mathbb{x}_i)}{\sum_{i=iw}^{ia} f(\mathbb{x}_i)} - \mathbb{x}a_i\,(P\{\mathbb{X}_i \le \mathbb{x}a_i\}-a)+\mathbb{x}w_i\,(\varepsilon-P\{\mathbb{X}_i \le \mathbb{x}w_i\})$$

$$\rho(\mathbb{X}1_i)=\rho(\mathbb{X}2_i)=\rho(\mathbb{X}3_i) = \left(\frac{1\text{x}0.95}{0.95}\right) - (1\times(0.95-0.9))+(1\times(0.9-0.95))$$

$$\rho(\mathbb{X}1_i)=\rho(\mathbb{X}2_i)=\rho(\mathbb{X}3_i) = 0.945 \text{ man-months}$$

The overall cost is the sum of the sub-estimates which is $\rho(\mathbb{X}1_i)+\rho(\mathbb{X}2_i)+\rho(\mathbb{X}3_i)$ = 2.835 man-months.

To test the sub-additivity, the overall cost is estimated from the random variable $(\mathbb{X}1_i+\mathbb{X}2_i+\mathbb{X}3_i)$ assuming the same parameters, i.e., a=0.9, w= 0.9. The estimated values are $\mathbb{x}a_i= \mathbb{x}w_i$=3.25 man-months and the estimated probabilities are $P\{\mathbb{X}_i \le \mathbb{x}a_i\} = P\{\mathbb{X}_i \le \mathbb{x}w_i\}$=0.9927. Furthermore, it is observed that $(P\{\mathbb{X}_i \le \mathbb{x}a_i\}= 0.9927)>(a=0.9)$ and $(P\{\mathbb{X}_i > \mathbb{x}wa_i\}=1-P\{\mathbb{X}_i \le \mathbb{x}w_i\}=1-0.9927=0.0073)<(1-\varepsilon=1-0.9=0.1)$. The estimated overall cost at 90% probability from the probability distribution $f(\mathbb{X}1_i+\mathbb{X}2_i+\mathbb{X}3_i)$ using the model presented in equation (9) is:

$$\rho(\mathbb{X}1_i)=\rho(\mathbb{X}2_i)=\rho(\mathbb{X}3_i) = \frac{\sum_{i=iw}^{ia} \mathbb{x}_i f(\mathbb{x}_i)}{\sum_{i=iw}^{ia} f(\mathbb{x}_i)} - \mathbb{x}a_i\,(P\{\mathbb{X}_i \le \mathbb{x}a_i\}-a)+\mathbb{x}w_i\,(\varepsilon-P\{\mathbb{X}_i \le \mathbb{x}w_i\})$$

$$= \frac{3\times0.857375+3.25\times0.135375}{(0.857375+0.135375)} - (3.25\times(0.9927-0.9))+(3.25\times(0.9-0.9927))$$

$$\rho(\mathbb{X}1_i+\mathbb{X}2_i+\mathbb{X}3_i)= 2.43 \text{ man-months.}$$

The estimated overall cost of 2.43 man-months is less than the aggregated sum of the sub-estimates of 2.835 man-months, i.e., $\{\rho(\mathbb{X}1_i+\mathbb{X}2_i+\mathbb{X}3_i)=2.43\} \le \{\rho(\mathbb{X}1_i)+\rho(\mathbb{X}2_i)+\rho(\mathbb{X}3_i)=2.835\}$; these results are in compliance with equation (3) and confirms that the sub-estimates $\rho(\mathbb{X}1_i)$, $\rho(\mathbb{X}2_i)$ and $\rho(\mathbb{X}3_i)$ are sub-additive.

Using the proposed model, this bottom-up example show sub-estimates of 0.945 man-months for each task and their aggregation leads to the overall cost of 0.945+0.945+0.945=2.835 man-months; whereas an assessment of the overall cost suggests the overall estimate of 2.43 man-months conforming to sub-additivity, i.e., 2.43≤(0.945+0.945+0.945=2.835). These results may lead to a question: which estimate is the overall estimate 2.43 or 2.835? To answer this question lets understand real bottom-up cost estimation scenario where project managers assesses sub-estimates and the overall cost is the sum of the sub-estimates. Therefore, with sub-additivity assured, the sum of the sub-estimate of 2.835 man-months is considered a good approximation of the overall estimate.

Similarly, for the top-down example, the estimated overall cost is $\rho(\mathbb{X}1_i+\mathbb{X}2_i+\mathbb{X}3_i)$= 2.43 man-months at 90% probability, i.e., a=0.9 and let w=0.9. This estimate is decomposed to equal sub-estimates of 0.81 man-months for each task. For sub-additivity check, the random variable $(\mathbb{X}1_i+\mathbb{X}2_i+\mathbb{X}3_i)$ is decomposed to random variables $(\mathbb{X}1_i)$, $(\mathbb{X}2_i)$ and $(\mathbb{X}3_i)$ and estimated sub-estimates are $\rho(\mathbb{X}1_i)=\rho(\mathbb{X}2_i)=\rho(\mathbb{X}3_i)$=0.945 man-months. These results are sub-additive since each decomposed sub-estimate of 0.81 man-months is less than the estimated sub-estimate of 0.945 man-months and overall estimate is not over-estimated, i.e., $\{\rho(\mathbb{X}1_i+\mathbb{X}2_i+\mathbb{X}3_i)$=2.43=0.81+0.81+0.81$\}>\{\rho(\mathbb{X}1_i)+\rho(\mathbb{X}2_i)+\rho(\mathbb{X}3_i)$=0.945+0.945+0.945$\}$. Therefore, decomposition of the overall estimate is a good representation of sub-estimated.

3. Case Study

A case study is conducted to investigate the sub-additivity of estimates originating from probability distributions in real software development projects. Case study investigates bottom-up and top-down scenarios. The case-study estimate costs using the proposed model and using the percentile based approach. Case-study aims to show that the proposed model generates sub-additive estimates whereas percentile based estimates are susceptible to sub-additivity. The case-study uses the dataset from Kitchenham et al. (2001, 2002) study which presents estimated function points effort data of 144 software projects after outliers are removed. Beside other fields, the dataset contains the following fields of interest: estimated overall cost (hours), total adjusted function point (*FP*), total unadjusted function point (*UFP*) and unadjusted function point elements, i.e., Internal Logical Files (*ILF*), External Interface Files (*EIF*), External Inputs (*EI*), External Outputs (*EO*) and external Enquiries (*EQ*). The data of total unadjusted function point counts together with the decomposed function point elements is ideal to study the bottom-up and top-down cost estimation scenarios of real software development projects.

Function point Description

Function point describes the size of the software using five elements (Pfleeger, et al., 2006): *ILF, EIF, EI, EO* and *EQ*. Function point elements are counted and assigned a complexity level (*Low, Average, High*) based on their associated file number such as Data Element Type (DET), File Type Referenced (FTR) and Record Element Types (RET). The complexity metrics for five elements is shown in **Table 3**. Each function point element is then assigned a weight according to its complexity shown in **Table 4**. The unadjusted function point (*UFP*) is the sum of function point elements which is computed from equation 10:

$$UFP = \sum_{i=1}^{5}\sum_{j=1}^{3} w_{ij}\, v_{ij} \qquad (10)$$

Where w_ij is the complexity weight and v_ij is the count for each function element. UFP is then multiplied by the Value Adjustment Factor (VAF) to get the function point (FP) count.

$$FP = UFP \times VAF \qquad (11)$$

The VAF is calculated from 14 General System Characteristics (GSC) using equation (12).

$$VAF = 0.65 + 0.01 \sum_{i=1}^{14} c_i \qquad (12)$$

Where c_i are values of GSC characteristic on a scale of 0 to 5 which are described as: 1) Data Communication 2) Distributed Functions 3) Performance 4) heavily used configuration 5) transaction rate 6) on-line data entry 7) end user efficiency 8) on-line update 9) complex processing 10) reusability 11) installation ease 12) operational ease 13) multiple sites and 14) facilities change. These values are summed and modified to calculate the VAF.

ILF/EIF	DET			EI	DET			EO/EQ	DET		
RET	1-19	20-50	51+	FTR	1-4	5-15	16+	FTR	1-5	6-19	20
1	Low	Low	Avg	0-1	Low	Low	Avg	0-1	Low	Low	Avg
2-5	Low	Avg	High	2	Low	Avg	High	2-3	Low	Avg	High
6+	Avg	High	High	3+	Avg	High	High	4+	Avg	High	High

TABLE 03. *Function Point element complexity metrics*

Component	Low	Average	High
External Inputs	3	4	6
External Outputs	4	5	7
External Inquiries	3	4	6
Internal Logical Files	7	10	15
External Interface Files	5	7	10

TABLE 04. *Function Point element complexity weights*

	Actual (man days)	Estimated (man days)	*UFP*	*FP*	*ILF*	*EIF*	*EI*	*EO*	*EQ*
Mean	292.94	290.67	405.39	394.69	132.2	101.9	59.3	13.8	87.4
Median	193.06	215.75	259.59	267.5	75.5	59.5	28	0	49
Min	27.37	25	15.36	15	0	0	0	0	0
Max	1959.12	1778.25	2075.8	1940	850	627	555	614	618

TABLE 05. *Function point descriptive statistics*

Case-Study Design

The case study classifies each project of the dataset to either bottom-up or top-down scenario. The projects where the actual overall costs overruns the estimated overall cost are classified as bottom-up cost estimation scenarios because the overall cost is under-estimated. This could be due to under-estimated sub-estimates which violates sub-additivity. Whereas, the projects where the actual overall costs underruns the estimated overall costs are classified as top-down cost estimation scenario because the estimated overall cost is over-estimated which could be due to sub-additivity violation. Therefore, out of 144 projects, 55 projects were classified as the bottom-up scenarios while the rest of 89 projects were classified as the top-down scenarios. **Table 5** shows the summary statistics of the dataset.

For each bottom-up scenario, function point elements *ILF, EIF, EI, EO*, and *EQ* are converted to efforts which are then represented with probability distributions. From the probability distributions sub-estimates are estimated and the overall cost is the sum of the sub-estimates. For sub-additivity, the probability distributions of the function point elements are convolved which generates the probability distribution of *FP*. The overall cost is estimated from the distribution of *FP* which is compared with the sum of the sub-costs. These steps are repeated for the proposed model and for the percentile based cost estimation approach.

Similarly, for top-down scenario, the value of *FP* is first converted to effort which is then mapped to a probability distribution and the overall cost is estimated. The distribution of *FP* is then de-convolved to get the distributions of each function point elements and sub-estimates are estimated from each respective probability distribution. These sub-estimates are summed and compared with the estimated overall cost for sub-additivity. These steps are repeated using the proposed model and using the percentile based approach.

Following steps were taken to implement the procedure described above for bottom-up and top-down scenarios. The unadjusted function point elements (*UILF, UEIF, UEI, UEO, UEQ*) are converted to adjusted function point counts as follows, where, *UFP= UILF+ UEIF+ UEI+ UEO+ UEQ* and *VAF=FP/UFP*:

$$[ILF, EIF, EI, EO, EQ] = VAF \times [UILF, UEIF, UEI, UEO, UEQ] \qquad (13)$$

The estimated effort (man-days) required to develop an adjusted function point is:

$$per\ FP\ effort = \frac{overall\ estimated\ cost\ (man\ days)}{FP} \qquad (14)$$

Then the overall effort (man-days) required for total function point element FP and sub-efforts required for each adjusted function point element ILF,EIF,EI,EO and EQ is:

$$eFP = per\ FP\ effort \times [FP] \qquad (15)$$

$$[eILF, eEIF, eEI, eEO, eEQ] = per\ FP\ effort \times [ILF, EIF, EI, EO, EQ] \qquad (16)$$

These estimated efforts of *eILF, eEIF, eEI, eEO, eEQ* and *eFP* are represented with appropriate probability distributions. Parametric probability distributions often possess a sub-additive behavior; for example, a gamma distribution has the following property: $\Gamma(k,\theta_1+...+\theta_n)=\Gamma(k,\theta_1)+...+\Gamma(k,\theta_n)$ (Veerarajan, 2008) which is sub-additive. Whereas, the aim of the case study is to observe situations of sub-additivity violations to study the model's ability to handle such violations. Therefore, a non-parametric distribution $\tilde{X}$ is defined which is a mixture of a gamma distribution X and a discrete probability distribution $\mathbb{X}_i$ such that (Daníelsson et al., 2005),

$$\tilde{X}=X\sim\Gamma(k,\theta)+\mathbb{X}_i \begin{cases} 0 & 0.991 \\ \mathbb{E}[X]\text{x}10 & 0.991 \end{cases} \qquad (17)$$

The probability distribution $\tilde{X}$ explains: that there is a 99.1% probability that the estimated cost from X is the right estimate and $\mathbb{X}_i$ takes a value 0; furthermore, there is a 0.9% probability that the estimate is 10 times the expectation of X, i.e., $\mathbb{E}[X]$. At 99% probability a value of $\tilde{X}$ could experience a large value leading to sub-additivity violation; therefore, the case study estimates the costs at 99% probability, i.e., a=0.99.

Bottom-up scenarios

For bottom-up cost scenarios, the estimated sub-efforts of each function point element (*eILF, eEIF, eEI, eEO, eEQ*) are mapped to gamma distributions and mixed with respective discrete distributions of $\mathbb{X}_i$. For example, the estimated effort *eILF* of function point element *ILF* is mapped to the gamma distribution, i.e., $eILF \mapsto \mathbb{E}[XeILF \sim \Gamma(k,\theta)]$. Therefore, *eILF* is modeled with the random variable XeILF which is mixed with the respective discrete random variable, $\mathbb{X}_i$ and the random variable $\tilde{X}eILF$ is generated. The estimated efforts of *eILF, eEIF, eEI, eEO* and *eEQ* are represented with random variables $\tilde{X}eILF$, $\tilde{X}eEIF$, $\tilde{X}eEI$, $\tilde{X}eEO$ and $\tilde{X}eEQ$, respectivley. The sub-estimates of $\rho(\tilde{X}eILF)$, $\rho(\tilde{X}eEIF)$, $\rho(\tilde{X}eEI)$, $\rho(\tilde{X}eEO)$ and $\rho(\tilde{X}eEQ)$ are then estimated at 99% probability keeping *w*=89%, i.e., *a*=0.9,*w*=0.89, respectivley, using equation (9). The overall cost is the sum of the sub-estimates, i.e., $\rho(\tilde{X}eILF)+\rho(\tilde{X}eEIF)+\rho(\tilde{X}eEI)+\rho(\tilde{X}eEO)+\rho(\tilde{X}eEQ)$.

For sub-additivity, the random variable $\tilde{X}eFP$ to represent the overall cost is generated which is the sum of the random variables of sub-estimates, i.e., $\tilde{X}eFP=\tilde{X}eILF+\tilde{X}eEIF+\tilde{X}eEI+\tilde{X}eEO+\tilde{X}eEQ$. The probability distribution of $\tilde{X}$eFP is the convolution of the probability distributions of $\tilde{X}eILF$, $\tilde{X}eEIF$, $\tilde{X}eEI$, $\tilde{X}eEO$ and $\tilde{X}eEQ$. Then the overall cost $\rho(\tilde{X}eFP)$ is estimated at 99% probability keeping *w*=0.89 using equation (9). The sub-additivity is tested as follows:

$$\rho(\tilde{X}eFP) \le \rho(\tilde{X}eILF)+\rho(\tilde{X}eEIF)+\rho(\tilde{X}eEI)+\rho(\tilde{X}eEO)+\rho(\tilde{X}EQ) \qquad (18)$$

The same procedure is repeated for percentile based cost estimation approach where the sub-costs and the overall costs are estimated at 99% probability from the respective probability distributions and sub-additivity is tested using equation (18).

Top-Down scenarios

The estimated overall effort *eFP* is represented with the random variable $\tilde{X}eFP$ and the overall cost $\rho(\tilde{X}eFP)$ is estimated at 99% keeping *w*=0.89 using equation (9). For sub-additivity test, the random variable $\tilde{X}eFP$ is decomposed into random variables $\tilde{X}eILF$, $\tilde{X}eEIF$, $\tilde{X}eEI$, $\tilde{X}eEO$ and $\tilde{X}eEQ$ based on their respective shares of *ILF, EIF, EI, EO* and *EQ*. The probability distributions of random variables $\tilde{X}eILF$, $\tilde{X}eEIF$, $\tilde{X}eEI$, $\tilde{X}eEO$ and $\tilde{X}eEQ$ are de-convolved from the probability distribution of $\tilde{X}eFP$. Then the sub-estimates of $\rho(\tilde{X}eILF)$, $\rho(\tilde{X}eEIF)$, $\rho(\tilde{X}eEI)$, $\rho(\tilde{X}eEO)$ and $\rho(\tilde{X}eEQ)$ are estimated at 99% at *w*=0.89 using the proposed model defined by equation (9). Sub-additivity is tested using equation (18). The same procedure is repeated for the percentile based approach where overall cost and sub-estimates are estimated at 99% probability. Then the sub-additivity is tested using equation (18).

Case-Study results

Table 6 presents the results from the proposed model which includes the sub-estimates $\rho(\tilde{X}eILF)$, $\rho(\tilde{X}eEIF)$, $\rho(\tilde{X}eEI)$, $\rho(\tilde{X}eEO)$ and $\rho(\tilde{X}eEQ)$ and the estimated overall cost of $\rho(\tilde{X}eFP)$ of bottom-up scenarios at 99% probability. To elaborate on the calculations, consider project 10, refer to Kitchenham (2001, 2002) for function point data:

$UFP = UILF + UEIF + UEI + UEO + UEQ$ =4+26 +37+0+0=67

FP=84.42

$VAF=FP/UFP$=1.26

The adjusted function point elements *ILF, EIF, EI, EO* and *EQ* are calculated as follows:

$[ILF, EIF, EI, EO, EQ] = VAF \times [UILF, UEIF, UEI, UO, UEQ]$ = [5.04,32.76,46.62,0,0]

The estimated effort required to complete one adjusted function point element is:

$$FP_{effort} = \frac{overall\ estimated\ cost\ (man\ days)}{FP} = \frac{885/8}{84{,}42} = 1.31\ man\ days/\ function\ point$$

The effort required for each function point element:

$[eILF, eEIF, eEI, eEO, eEQ] = FP_{effort}\times[ILF, EIF, EI, EO,EQ]$ = [6.60,42.93,61.1,0,0] *man days*

These efforts are mapped to respective gamma distributions $\Gamma(k,\theta)$:

$eILF$=6.60$\mapsto \mathbb{E}\,[XeILF\sim\Gamma(k,\theta)]=XeILF\sim\Gamma(k=2,\theta=3.3)$

$eEIF$= 42.93$\mapsto \mathbb{E}\,[XeEIF\sim\Gamma(k,\theta)]=XeEIF\sim\Gamma(k=2,\theta=21.46)$

eEI= 61.1$\mapsto \mathbb{E}\,[XeEI\sim\Gamma(k,\theta)]=XeEI\sim\Gamma(k=2,\theta=30.55)$

Where *eEO* and *eEQ* are zero. These gamma distributions are mixed with the respective discrete distributions to get $\tilde{X}$:

$$\tilde{X}eILF = XeILF + \mathbb{X}_i \begin{cases} 0 \\ 6.60\text{ x}10 \end{cases}$$

$$\tilde{X}eEIF = XeEIF + \mathbb{X}_i \begin{cases} 0 \\ 42.93\text{ x}10 \end{cases}$$

$$\tilde{X}eEI = XeEI + \mathbb{X}_i \begin{cases} 0 \\ 61.1\text{ x}10 \end{cases}$$

Then the estimated sub-estimates using the proposed model are:

$\rho(\tilde{X}eILF)$=16.14 *man days*

$\rho(\tilde{X}eEIF)$=107.06 *man days*

$\rho(\tilde{X}eEI)$=152.54 *man days*

$\rho(\tilde{X}eEO)$=0 *man days*

$\rho(\tilde{X}eEQ)$=0 *man days*

Therefore, for the bottom-up scenario of project 10, the overall cost is the sum of the sub-estimates:

$\rho(\tilde{X}eILF)+\rho(\tilde{X}eEIF)+\rho(\tilde{X}eEI)+\rho(\tilde{X}eEO)+\rho(\tilde{X}eEQ)$ =276 *man days*

For sub-additivity test, the random variables $\tilde{X}eILF$, $\tilde{X}eEIF$, $\tilde{X}eEI$, $\tilde{X}eEO$ and $\tilde{X}eEQ$ are summed and the random variable $\tilde{X}eFP$ is generated, i.e., $\tilde{X}eILF+\tilde{X}eEIF+\tilde{X}eEI+\tilde{X}eEO+\tilde{X}eEQ=\tilde{X}eFP$; the probability distribution of $\tilde{X}eFP$ is the convolution of the distributions of $\tilde{X}eILF$, $\tilde{X}eEIF$, $\tilde{X}eEI$, $\tilde{X}eEO$ and $\tilde{X}eEQ$. Then the estimated overall cost from $\tilde{X}eFP$ is using the proposed model:

$\rho(\tilde{X}eFP)$=248 *man days*

Now the sub-additivity can be tested as follows:

$\{\rho(\tilde{X}eFP)$ =248$\} \le \{\rho(\tilde{X}eILF)+\rho(\tilde{X}eEIF)+\rho(\tilde{X}eEI)+\rho(\tilde{X}eEO)+\rho(\tilde{X}eEQ)$ =276$\}$

The sub-additivity equality holds and fulfills the sub-additivity property. For illustration, the probability distribution of the random variable $\tilde{X}eEI$ is shown in **Figure 4** which has the estimated effort of *eEI*=61.1 man-months.

The hump on the tail of the probability distribution of $\tilde{\mathbb{X}}eEI$ is due to convolving the distribution $XeEI \sim \Gamma(k=2,\theta=30.55)$ with the discrete distribution

$$\mathbb{X}eEI_i \begin{cases} 0 & 0.991 \\ 61.1\text{ x}10 & 0.009 \end{cases}$$

which converts the distribution $XeEI \sim \Gamma(k=2,\theta=30.55)$ into a non-parametric distribution.

Furthermore, **Table 7** present the results of percentile based cost estimation approach for the bottom-up scenarios. The costs are estimated at 99% probability from the respective distributions of $\tilde{\mathbb{X}}eILF$, $\tilde{\mathbb{X}}eEIF$, $\tilde{\mathbb{X}}eEI$, $\tilde{\mathbb{X}}eEO$ and $\tilde{\mathbb{X}}eEQ$. For example, the project 10 has the following percentile based sub-estimates from the distributions of $\tilde{\mathbb{X}}eILF$, $\tilde{\mathbb{X}}eEIF$, $\tilde{\mathbb{X}}eEI$, $\tilde{\mathbb{X}}eEO$ and $\tilde{\mathbb{X}}eEQ$: 30, 196, 279, 0 and 0 man-days, respectively. The overall cost is the sum of the sub-estimates, i.e., 30+196+279+0+0=505 man-days. However, the percentile based estimated overall cost from the distribution of $\tilde{\mathbb{X}}eFP$ at 99% probability is 607 man-months. These results show a violation of sub-additivity since: 607>(30+196+279+0+0=505).

The results of top-down scenarios using the proposed model at 99% probability are tabulated in **Table 8**; for example, consider project 8 which has an estimated overall effort in terms of adjusted function point of FP=225.54

The estimated effort required to complete one adjusted function point element is:

$$FP_{effort} = \frac{\text{overall estimated cost (man days)}}{FP} = \frac{1800/8}{225{,}54} = 0.9976 \text{ man days/function point}$$

$$eFP=FP_{effort}\times FP=225.54\times 0.9976= 225 \textit{ man days}$$

Which is mapped to the expectation of gamma distribution $\Gamma(k,\theta)$ such that:

$eFP=225 \mapsto \mathbb{E}[XeFP\sim\Gamma(k,\theta)]=XeFP\sim\Gamma(k=2,\theta=112.5)$

The distribution of $XeFP$ is mixed with $\mathbb{X}_i$ as follows:

$$\tilde{\mathbb{X}}eFP=XeFP+ \mathbb{X}_i \begin{cases} 0 \\ 225\text{ x}10 \end{cases}$$

Then the estimated overall cost at 99% using the proposed model is $\rho(\tilde{\mathbb{X}}eFP)$=492 man-months. For sub-additivity, the random variable $\tilde{\mathbb{X}}eFP$ is decomposed into random variables $\tilde{\mathbb{X}}eILF$, $\tilde{\mathbb{X}}eEIF$, $\tilde{\mathbb{X}}eEI$, $\tilde{\mathbb{X}}eEO$ and $\tilde{\mathbb{X}}eEQ$ based on the respective shares of ILF=63, EIF=5, EI=72, EO=0 and EQ=39, respectively (Kitchenham, 2001, 2002) such that $\tilde{\mathbb{X}}eFP=\tilde{\mathbb{X}}eILF+\tilde{\mathbb{X}}eEIF+\tilde{\mathbb{X}}eEI+\tilde{\mathbb{X}}eEO+\tilde{\mathbb{X}}eEQ$. The probability distributions of $\tilde{\mathbb{X}}eILF$, $\tilde{\mathbb{X}}eEIF$, $\tilde{\mathbb{X}}eEI$, $\tilde{\mathbb{X}}eEO$ and $\tilde{\mathbb{X}}eEQ$ were de-convolved from the distribution of $\tilde{\mathbb{X}}eFP$.

Then the estimated sub-estimates using the proposed model are as follows:

$\rho(\tilde{\mathbb{X}}eILF)$=197.94 *man days*

$\rho(\tilde{\mathbb{X}}eEIF)$=15.10 *man days*

$\rho(\tilde{\mathbb{X}}eEI)$=226.40 *man days*

$\rho(\tilde{\mathbb{X}}eEO)$=0 *man days*

$\rho(\tilde{\mathbb{X}}eEQ)$=122.48 *man days*

Therefore, the estimated overall cost is the sum of the sub-estimates:

$\rho(\tilde{\mathbb{X}}eILF)+\rho(\tilde{\mathbb{X}}eEIF)+\rho(\tilde{\mathbb{X}}eEI)+\rho(\tilde{\mathbb{X}}eEO)+\rho(\tilde{\mathbb{X}}eEQ)$=561 *man days*

These results confirms the sub-additivity, i.e., $\{\rho(\tilde{\mathbb{X}}eFP)=492\} \le \{\rho(\tilde{\mathbb{X}}eILF) +\rho(\tilde{\mathbb{X}}eEIF) +\rho(\tilde{\mathbb{X}}eEI) +\rho(\tilde{\mathbb{X}}eEO) +\rho(\tilde{\mathbb{X}}eEQ) = 561\}$.

The percentile based cost estimates for top-down scenarios at 99% probability are given is **Table 9**. For example for project 8, the estimated overall cost from the distribution of $\tilde{\mathbb{X}}eFP$ at 99% probability is 1053 man-days. The sub-estimates are estimated from the respective distributions of $\tilde{\mathbb{X}}eILF$, $\tilde{\mathbb{X}}eEIF$, $\tilde{\mathbb{X}}eEI$, $\tilde{\mathbb{X}}eEO$ and $\tilde{\mathbb{X}}eEQ$, at 99% which are 362, 29, 414, 0 and 224 man-days, respectively. The overall cost is the sum of the sub-estimates, i.e., 362+29+414+0+224=1029 man-days. This is a violation of the sub-additivity since the estimated overall cost of 1053 man-days is more than the sum of the sub-estimates of 1029 man-months, i.e., 1053>(362+29+414+0+224=1029).

The results of the case-study show that for the bottom-up scenarios, the proposed model holds the sub-additivity property; whereas the percentile based estimates violates sub-additivity for the projects 10, 18, 20, 24, 38, 42, 52, 81, 90, 103 and 122; therefore, out of 55 projects 11 projects violated the sub-additivity. Similarly, for the top-down scenarios, while the proposed model fulfills the sub-additivity for all the projects the percentile based estimates fails the sub-additivity for the projects 8, 34, 36, 50, 51, 59, 69, 76, 83, 93, 104, 117, 134, 138, and 145; altogether 15 out of 89 projects failed the percentile based sub-additivity. Therefore, altogether 26 out of 144, i.e., 18%, projects experienced sub-additive adjustments in estimated costs.

Note the large differences between the estimated costs from the proposed model and the estimated costs from the percentile based approach. For example, the estimated overall cost

(Sub-additivity: pass ✓, fail ×)								
	$\rho(\tilde{X}eFP)$	$\rho(\tilde{X}eILF)$	$\rho(\tilde{X}eEIF)$	$\rho(\tilde{X}eEI)$	$\rho(\tilde{X}eEO)$	$\rho(\tilde{X}eEQ)$	$\rho(\tilde{X}eILF)+\rho(\tilde{X}eEIF)+\rho(\tilde{X}eEI)+\rho(\tilde{X}eEO)+\rho(\tilde{X}eEQ)$	
3	2186	662	141	359	46	1374	2582	✓
6	269	263	0	0	0	11	274	✓
9	319	251	10	27	0	72	360	✓
10	248	16	107	153	0	0	276	✓
11	568	128	242	87	0	206	663	✓
17	140	111	28	0	0	19	158	✓
18	255	30	102	120	0	44	296	✓
19	101	38	15	21	42	0	116	✓
20	62	44	0	24	0	0	68	✓
24	566	345	211	33	57	19	665	✓
28	511	209	138	49	17	188	601	✓
29	1482	203	360	927	11	263	1764	✓
30	1036	508	373	219	85	42	1227	✓
31	522	257	128	120	10	107	622	✓
33	880	490	161	77	0	305	1033	✓
35	1043	547	293	96	0	291	1227	✓
37	1286	290	161	709	56	334	1550	✓
38	362	200	36	0	0	168	404	✓
41	1136	40	1046	116	0	28	1230	✓
42	681	269	361	69	11	89	799	✓
45	1482	387	531	93	187	569	1767	✓
46	480	0	480	0	0	0	480	✓
52	482	72	216	12	0	246	546	✓
54	353	111	45	23	0	232	411	✓
55	966	219	699	40	0	159	1117	✓
56	779	288	268	130	8	216	910	✓
60	1352	76	68	102	1117	202	1565	✓
70	461	139	146	6	0	235	526	✓
73	2095	530	891	356	0	681	2458	✓
75	996	505	345	69	0	242	1161	✓
77	115	24	13	19	0	79	135	✓
81	219	122	52	0	0	75	249	✓
82	595	186	233	16	0	238	673	✓
85	695	281	242	89	0	197	809	✓
87	1081	593	217	97	0	364	1271	✓
89	556	365	161	57	0	72	655	✓
90	893	150	398	54	0	426	1028	✓
92	3721	1167	2070	509	100	609	4455	✓
94	1049	385	240	250	133	257	1265	✓
95	171	53	11	48	45	43	200	✓
103	425	229	42	151	0	77	499	✓
105	156	18	42	100	24	0	184	✓
108	489	122	231	80	18	134	585	✓
112	1116	511	380	124	0	292	1307	✓
113	593	138	265	132	0	164	699	✓
114	261	172	59	20	0	54	305	✓
119	53	7	30	19	5	0	61	✓
121	436	68	43	329	62	12	514	✓
122	219	110	5	14	0	114	243	✓
124	985	647	27	273	132	93	1172	✓
125	187	85	75	33	11	16	220	✓
128	1065	484	287	229	56	226	1282	✓
132	701	486	10	162	0	146	804	✓
133	352	122	36	185	0	71	414	✓
144	272	68	68	68	68	50	322	✓

TABLE 06. *Bottom-up estimates at 99% using the proposed model*

for the bottom-up scenario of project 10 using the proposed model is 248 man-days while the percentile based approach produces the overall cost of 607 man-months. Similarly, for the top-down scenario, the overall estimated cost using the proposed model is 492 man-days while the percentile based overall cost is 1053 man-days.

(Sub-additivity: pass ✓, fail ×)								
	FP	*ILF*	*EIF*	*EI*	*EO*	*EQ*	*ILF+EIF+EI+EO+EQ*	
3	4269	1207	257	654	86	2505	4709	✓
6	486	481	0	0	0	21	502	✓
9	561	460	20	49	0	134	663	✓
10	607	30	196	279	0	0	505	×
11	1142	235	442	160	0	376	1213	✓
17	241	204	51	0	0	35	290	✓
18	555	56	187	220	0	81	544	×
19	208	71	28	39	79	0	217	✓
20	143	81	0	45	0	0	126	×
24	1258	630	384	61	104	35	1214	×
28	1017	382	252	91	32	343	1100	✓
29	2611	372	658	1690	21	481	3222	✓
30	2218	928	681	400	156	78	2243	✓
31	939	470	234	219	20	197	1140	✓
33	1860	895	295	142	0	556	1888	✓
35	2024	998	535	176	0	531	2240	✓
37	2353	529	294	1294	103	610	2830	✓
38	882	366	67	0	0	307	740	×
41	2000	73	1907	213	0	52	2245	✓
42	1558	492	660	127	20	164	1463	×
45	2852	707	969	171	342	1037	3226	✓
46	875	0	875	0	0	0	875	✓
52	1121	132	394	23	0	450	999	×
54	709	204	83	43	0	423	753	✓
55	1678	400	1276	74	0	291	2041	✓
56	1456	526	489	238	16	395	1664	✓
60	2292	139	126	186	2037	368	2856	✓
70	943	255	267	11	0	429	962	✓
73	4074	967	1626	650	0	1241	4484	✓
75	2097	922	630	127	0	443	2122	✓
77	197	45	25	35	0	145	250	✓
81	459	223	96	0	0	138	457	×
82	1194	341	426	31	0	436	1234	✓
85	1351	513	442	164	0	360	1479	✓
87	2243	1082	397	178	0	664	2321	✓
89	1065	666	295	104	0	132	1197	✓
90	2013	273	726	99	0	778	1876	×
92	7413	2128	3774	929	183	1112	8126	✓
94	1822	702	438	457	245	469	2311	✓
95	289	97	22	88	84	79	370	✓
103	915	417	79	276	0	141	913	×
105	289	33	78	183	44	0	338	✓
108	921	223	422	147	33	246	1071	✓
112	2273	932	693	227	0	533	2385	✓
113	1111	252	484	242	0	301	1279	✓
114	458	314	110	37	0	99	560	✓
119	114	14	56	34	10	0	114	✓
121	725	126	80	600	114	23	943	✓
122	534	201	9	26	0	209	445	×
124	1842	1180	50	500	241	170	2141	✓
125	403	155	138	60	22	30	405	✓
128	1963	883	524	418	103	413	2341	✓
132	1234	887	18	296	0	268	1469	✓
133	746	223	66	338	0	130	757	✓
144	428	125	125	125	125	92	592	✓

TABLE 07. *Bottom-up percentile estimates for at 99% probability*

This phenomenon is due to long extended tail of the gamma distribution which reaches to large value with small changes in probabilities. The proposed model has a maximum and minimum range which keeps the estimates well within the defined region. Whereas, the percentile based estimates reaches to large amounts of estimated costs.

Furthermore, the case-study focused on the estimated costs at 99% probability to stress the sub-additivity, while project managers are keen to estimate the cost at 70% probability (Fairley, 1995). Therefore, let's consider the bottom-up scenario of project 10 and estimate the cost at 75% probability, i.e., a=0.75, and letting w=0.65, so that the probability of the estimate is approximately around 70%. The estimated sub-estimates at 75% probability are 7.64, 52.55, 75.07, 0 and 0 man-days, and their sum comes to 135.26 man-days. While, the estimated overall cost from the aggregated distribution is 133 man-days. Therefore, for project 10 at 75% probability the sub-estimates using the proposed model are sub-additive, i.e., 133≤(7.64+52.55+75.07+0+0=135.26).

Similarly, for the top-down scenario of project 8, the estimated overall cost at 75% probability using the proposed model is 268 man-days. Furthermore, the de-convolved distributions produce the following sub-estimates: 97.45, 7.24, 111.43, 0 and 60.14 man-days. These results show that the estimated overall estimate is sub-additive, i.e., 268≤(97.45 +7.24+111.43+0+60.14=276.26).

Case Study Limitations

The case-study focuses on the gamma probability distribution; nonetheless, other parametric distributions, i.e., Gaussian and Weibull, are good candidates to represent the cost of software projects. Generally, parametric distributions are sub-additive; however, there are different shape and scale parameters involved in the construction of different probability distributions. Therefore, careful consideration should be given to the sub-additive behavior of estimates originating from different probability distributions.

Furthermore, different confidence levels produce different estimates; the case study tested the sub-additivity at 99% probability because the probability distributions are constructed to violate sub-additivity at 99% probability and how the proposed model handles sub-additivity violations. Testing the proposed model with different probabilistic levels will further enhance our understanding of the behavior of the modes.

4.Conclusions

Software development projects are known for runaway costs; therefore, cost estimation is a critical activity for the management of software development projects. As software engineering moves into the probabilistic domain; the cost of software projects is represented with probabilistic models. Traditionally, the estimated cost is a percentile of a probability distribution. However, such percentiles based estimated costs may not be sub-additive. The sub-additivity becomes more evident with the bottom-up and top-down cost estimation processes due to aggregations and decompositions of costs. A model to estimate the cost of software projects has been proposed which produces sub-additive cost estimates. For bottom-up cost estimation scenarios, the model ensures that the sum of the sub-estimates provides a good representation of the overall estimate. Similarly, for top-down cost estimation scenarios the model ensures that the decomposition of an estimated overall estimate is a good representation of sub-estimates.

Sub-additivity plays a critical part in the cost estimation as it may contribute to under-estimations and over-estimations of costs which leads to cost overruns and underruns. While research continues to improve representation of cost and to find better cost estimation models, care should be taken to validate the sub-additivity of estimates originating from cost estimation models.

Sub-additivity only ensures the natural aggregation and decomposition of estimates (Evans, et al, 1984). It is possible for cost estimation models to produce perfectly sub-additive estimates that are either over-estimated or under-estimated. Sub-additivity is one factor which could contributes to under-estimation and over-estimation besides other factors. Lederer, et al., (1995) presented a list of 24 factors which are responsible for inaccurate estimates and Jørgensen (2004a) investigated factors related to effort estimation errors of software projects. The proposed model establishes sub-additivity as a factor for the construction of cost estimation models. Furthermore, sub-additivity of estimates is not only related to estimates originating from probability distributions; rather, estimates from any cost estimation technique can violate sub-additivity.

(Sub-additivity: pass ✓, fail ×)								
	$\rho(\tilde{X}eFP)$	$\rho(\tilde{X}eILF)$	$\rho(\tilde{X}eEIF)$	$\rho(\tilde{X}eEI)$	$\rho(\tilde{X}eEO)$	$\rho(\tilde{X}eEQ)$	$\rho(\tilde{X}eILF) + \rho(\tilde{X}eEIF) + \rho(\tilde{X}eEI) + \rho(\tilde{X}eEO) + \rho(\tilde{X}eEQ)$	
1	133	43	11	90	0	8	152	✓
2	365	91	106	206	0	23	426	✓
4	438	72	322	103	0	0	497	✓
5	1022	302	426	169	0	300	1197	✓
7	777	434	272	45	0	155	906	✓
8	492	198	15	226	0	122	561	✓
12	361	92	151	125	22	39	429	✓
13	332	286	0	0	0	71	357	✓
14	525	400	0	140	0	53	593	✓
15	357	122	89	159	0	47	417	✓
16	123	0	94	19	0	25	138	✓
21	777	32	418	204	0	246	900	✓
22	399	52	116	155	0	139	462	✓
23	440	205	103	103	0	110	521	✓
25	316	196	92	0	13	65	366	✓
26	589	154	281	63	0	194	692	✓
27	607	419	45	49	0	189	702	✓
32	594	341	149	77	17	128	712	✓
34	222	131	0	19	0	98	248	✓
36	1861	930	683	43	0	476	2132	✓
39	1513	725	196	345	0	515	1781	✓
40	1388	373	696	378	0	194	1641	✓
43	403	89	268	0	74	45	476	✓
44	273	188	71	14	0	45	318	✓
47	400	116	63	84	98	116	477	✓
48	89	79	6	6	5	0	96	✓
49	128	0	127	0	0	1	128	✓
50	199	85	89	12	0	42	228	✓
51	221	0	104	0	0	134	238	✓
53	189	79	62	0	0	72	213	✓
57	196	72	72	0	0	72	216	✓
58	199	59	65	0	0	101	225	✓
59	486	183	251	0	68	68	570	✓
61	694	238	176	154	42	216	826	✓
62	667	395	142	124	0	131	792	✓
63	604	182	172	18	0	325	697	✓
64	267	104	67	85	0	55	311	✓
65	436	137	108	52	0	216	513	✓
66	287	40	213	19	0	59	331	✓
67	608	203	169	101	0	237	710	✓
68	183	69	69	0	0	66	204	✓
69	200	68	97	0	0	61	226	✓
71	624	132	324	63	0	215	734	✓
72	598	326	88	112	24	169	719	✓
74	526	351	17	124	88	46	626	✓
76	344	181	75	12	0	128	396	✓
78	215	62	121	48	0	20	251	✓
79	233	93	37	73	20	55	278	✓
80	771	0	7	0	0	767	774	✓
83	480	119	167	14	0	253	553	✓
84	217	11	31	0	0	194	236	✓
86	308	225	29	29	0	75	358	✓
88	248	175	35	25	0	55	290	✓
91	1241	240	168	62	0	961	1431	✓
93	1610	819	663	168	59	193	1902	✓
96	218	90	28	50	0	85	253	✓
97	1659	667	221	670	109	316	1983	✓
98	84	7	69	0	16	0	92	✓
99	86	24	24	48	0	0	96	✓
100	216	0	62	0	172	0	234	✓
101	237	50	64	49	0	116	279	✓
104	459	62	175	267	0	27	531	✓
106	573	227	159	129	8	149	672	✓
107	271	168	71	0	0	71	310	✓
109	661	389	68	216	24	93	790	✓
110	710	55	577	61	122	0	815	✓
111	3291	2021	931	302	218	504	3976	✓
115	1000	749	214	132	0	68	1163	✓
116	793	627	59	66	0	160	912	✓
117	99	69	8	34	0	0	111	✓
118	991	333	404	154	0	266	1157	✓
120	155	46	57	54	0	23	180	✓
123	2300	620	770	561	106	665	2722	✓
126	228	160	56	23	0	26	265	✓
127	635	287	145	187	28	114	761	✓
129	1373	943	338	94	71	192	1638	✓
130	766	480	45	263	0	104	892	✓
131	327	52	240	37	11	48	388	✓
134	919	367	554	25	18	93	1057	✓
135	399	198	109	107	8	51	473	✓
136	598	325	123	82	0	177	707	✓
137	750	194	322	72	0	285	873	✓
138	364	164	178	0	3	64	409	✓
139	600	89	498	50	27	27	691	✓
140	90	36	7	20	0	40	103	✓
141	249	198	72	0	0	0	270	✓
142	120	77	34	0	0	26	137	✓
143	727	187	224	324	0	120	855	✓

TABLE 08. *Top-down estimate using the proposed model at 99% probability*

Under-estimation of sub-estimates in bottom-up scenario and over-estimation of overall estimate in top-down scenario violates sub-additivity this leads to miss-representations of costs in both the scenarios. Therefore, software practitioner should adopt tools and techniques that ensure sub-additive estimates. There can be scenarios where estimates fulfill the sub-additivity but they may still be under-estimated or over-estimated. For example, for a bottom-up scenario, the sub-estimates fulfilling the sub-additivity property means they are not under-estimated just to satisfy the sub-additivity. They can still be under-estimated or over-estimated due to other factors despite sub-additivity is fulfilled. Similarly, for a top-down scenario where the overall estimate fulfills sub-additivity ensuring that it is not over-estimated, it may still be over-estimated or under-estimated. Sub-additivity ensures the natural aggregation and decomposition of estimates. Sub-additivity is not the accuracy of the estimate; accuracy of the estimates is related to estimation models, tools and techniques of cost estimation processes.

Furthermore, it is observed that for few samples the range of random variable, i.e., difference between w and a, should be small, since due to small number of samples the probability changes significantly from sample to sample. Therefore, for the example discussed in section 2, the lower probability bound of w is set as a. While for large number of samples or for continuous probability distributions this restriction can be relaxed.

It is recommended that future research work should adopt different parametric and non-parametric probabilistic representations of costs to test the sub-additivity of estimates. For example, researchers have proposed Gaussian and Weibull distributions to represent the cost of software development projects. Furthermore, different probabilistic levels should be tested against the sub-additive behavior of estimates.

Acknowledgement

Authors would like to thank Dr. Barbara Kitchenham for sharing the TRSE0102 dataset.

(Sub-additivity: pass ✓, fail ×)								
	FP	ILF	EIF	EI	EO	EQ	ILF+EIF+EI+EO+EQ	
1	272	79	21	166	0	16	282	✓
2	706	167	195	376	0	42	780	✓
4	772	133	588	190	0	0	911	✓
5	1950	551	778	310	0	548	2187	✓
7	1649	791	496	83	0	283	1653	✓
8	1053	362	29	414	0	224	1029	×
12	717	168	277	230	42	73	790	✓
13	588	522	0	0	0	130	652	✓
14	965	730	0	255	0	97	1082	✓
15	722	223	163	292	0	87	765	✓
16	213	0	173	35	0	47	255	✓
21	1561	59	763	373	0	450	1645	✓
22	766	96	213	284	0	255	848	✓
23	803	374	189	189	0	200	952	✓
25	605	358	169	0	25	119	671	✓
26	1204	281	512	116	0	355	1264	✓
27	1208	765	83	90	0	346	1284	✓
32	1070	622	272	143	32	233	1302	✓
34	550	240	0	35	0	180	455	×
36	4003	1696	1245	80	0	868	3889	×
39	3129	1321	358	630	0	939	3248	✓
40	2646	681	1270	690	0	355	2996	✓
43	695	163	490	0	136	82	871	✓
44	496	343	130	26	0	82	581	✓
47	657	212	117	154	180	212	875	✓
48	154	145	13	13	9	0	180	✓
49	234	0	232	0	0	3	235	✓
50	433	157	163	23	0	78	421	×
51	563	0	191	0	0	245	436	×
53	383	145	113	0	0	131	389	✓
57	387	133	133	0	0	133	399	✓
58	412	109	120	0	0	185	414	✓
59	1080	335	459	0	124	124	1042	×
61	1226	435	322	281	77	394	1509	✓
62	1147	720	260	227	0	240	1447	✓
63	1202	333	314	34	0	594	1275	✓
64	499	190	123	157	0	101	571	✓
65	872	251	198	95	0	396	940	✓
66	490	74	390	35	0	110	609	✓
67	1149	370	309	185	0	432	1296	✓
68	363	127	127	0	0	120	374	✓
69	419	125	178	0	0	112	415	×
71	1312	241	592	117	0	392	1342	✓
72	1128	596	160	206	46	310	1318	✓
74	917	640	32	226	162	84	1144	✓
76	760	331	138	23	0	235	727	×
78	414	115	223	89	0	38	465	✓
79	442	170	70	135	39	100	514	✓
80	1403	0	15	0	0	1399	1414	✓
83	1012	218	304	26	0	462	1010	×
84	380	22	58	0	0	355	435	✓
86	546	412	53	53	0	137	655	✓
88	427	320	65	48	0	101	534	✓
91	2109	438	308	114	0	1753	2613	✓
93	3667	1495	1209	308	110	352	3474	×
96	446	166	52	91	0	157	466	✓
97	3393	1217	403	1223	200	578	3621	✓
98	146	13	127	0	31	0	171	✓
99	170	46	46	89	0	0	181	✓
100	411	0	113	0	315	0	428	✓
101	440	91	119	91	0	212	513	✓
104	1035	114	320	488	0	50	972	×
106	1048	415	292	237	16	272	1232	✓
107	510	308	130	0	0	130	568	✓
109	1357	709	125	395	44	169	1442	✓
110	1206	101	1052	112	224	0	1489	✓
111	6210	3684	1699	550	399	919	7251	✓
115	1717	1367	390	241	0	126	2124	✓
116	1359	1144	108	121	0	292	1665	✓
117	211	127	16	63	0	0	206	×
118	1911	607	738	282	0	486	2113	✓
120	293	84	105	98	0	42	329	✓
123	3980	1130	1405	1023	193	1212	4963	✓
126	406	293	104	43	0	49	489	✓
127	1221	524	264	342	52	208	1390	✓
129	2431	1719	617	172	130	351	2989	✓
130	1615	875	83	480	0	191	1629	✓
131	542	96	438	68	20	88	710	✓
134	2133	670	1011	47	33	170	1931	×
135	764	362	200	197	16	95	870	✓
136	1152	594	226	150	0	323	1293	✓
137	1532	355	588	131	0	519	1593	✓
138	832	300	326	0	6	117	749	×
139	1019	163	909	92	49	49	1262	✓
140	191	67	15	37	0	73	192	✓
141	478	361	133	0	0	0	494	✓
142	229	141	63	0	0	48	252	✓
143	1420	342	408	590	0	219	1559	✓
145	1012	346	16	0	0	434	796	×

TABLE 09. *Top-down percentile based estimates at 99% probability*

author

Masood Uzzafer has more than 20 years of software development experience in designing and development of different software applications including multimedia, DSP, embedded, mobile and web applications. Mr. Masood has worked for Philips Semiconductors, California, and AlliedSignal Aero-Space, Florida. His research interests are risk management and strategic planning. Dr. Masood earned his Ph.D. in computer science from the University of Nottingham, U.K. in 2015. He also has a M.S. in Electrical Engineering from Wayne State University, Michigan, USA and B.E. in Electrical Engineering from NED University, Karachi, Pakistan. Dr. Masood is a certified Project Management Professional (PMP) since 2014. uzzafer@alumni.nottingham.ac.uk

APPENDIX A

X is a random variable where xq is the value of X at probability $q \in [0,1]$, such that $P[X \leq xq]=q$. The expectation of X is defined as $\mathbb{E}[X] = \int xf(x)dx$, where $f(x)$ is the probability density function of X (Papoulis, 1991).

$\mathbb{X}_i$ is a discrete sequence of samples where the index i is defined as i=supremum $\{ k: P\{\mathbb{X}_i \leq \mathbb{x}_i \} \leq q \}$. The sample $\mathbb{x}q_i$ is the value $\mathbb{X}_i$ at probability q such that $P\{\mathbb{X}_i \leq \mathbb{x}_i \}=q$ (Ross, 2007; Delbaen, 2002; Montgomery and Runger, 2007; Papoulis, 1991). The expectation of $\mathbb{X}_i$ is defined as $E[\mathbb{X}_i] = \sum_k x_k P\{\mathbb{X}_i \leq \mathbb{x}_k \}$.

For X, the probability q is equal to the estimated probability $P[X \leq xq]=q$. However, for $\mathbb{X}_i$, the probability $P\{\mathbb{X}_i \leq \mathbb{x}_i\}$ may exceed q, i.e. $P\{\mathbb{X}_i \leq \mathbb{x}_i\}>q$. Furthermore, the probability $P\{\mathbb{X}_i > \mathbb{x}_i\} = 1-q$ may be under-estimated (Acerbi et al., 2001).

Therefore, a sample of $\mathbb{X}_i$ at a probability q may not have the exact probability q it may have probability $P\{\mathbb{X}_i \leq \mathbb{x}_i\}>q$, which causes over-estimated value of $\mathbb{x}_i$. As a consequence the probability $P\{\mathbb{X}_i > \mathbb{x}_i\}<1-q$ is less than $1-q$, i.e., $P\{\mathbb{X}_i > \mathbb{x}_i\}<1-q$ and the value of the sample at probability $1-q$ is less.

APPENDIX B

A continuous random variable X can be modeled with a gamma distribution, i.e., $X \sim \Gamma(k,\theta)$, where k and θ are the shape and spread parameters of the gamma distribution, respectively. The expectation of the gamma distribution is defined as $\mathbb{E}[X \sim \Gamma(k,\theta)] = k\,\theta$ (Ross, 2007).

The gamma distribution has the following probability density function (Papoulis, 1991):

$$f_X(x) = \frac{1}{\theta^k \, \Gamma(k)} x^{k-1} e^{-x/\theta}, \qquad x \in \mathbb{R}^+, k \in \mathbb{N}, \theta \in \mathbb{R}^+ \qquad \text{(B.1)}$$

Where $\Gamma(k)$ is the gamma function that is defined as follows:

$$\Gamma(k) = \int_0^\infty t^{k-1} e^{-t} dt, \qquad k \in \mathbb{R}^+, t \in \mathbb{R}^+$$

The expectation, $\mathbb{E}[X \sim \Gamma(k,\theta)]$, of the gamma distribution can be estimated using $\int x\, f_X(x)dx$; therefore, from equation (B.1):

$$\mathbb{E}[X \sim \Gamma(k,\theta)] = \int_0^\infty x \frac{1}{k^\theta \, \Gamma(k)} x^{k-1} e^{-x/\theta} dx$$

$$\mathbb{E}[X \sim \Gamma(k,\theta)] = \int_0^\infty \frac{1}{\theta^k \, \Gamma(k)} x e^{-x/\theta} dx = k\,\theta$$

references

Abdul-Rahmana, H., Mohd-Rahima, F.A., Chen, W., (2012). Reducing failures in software development projects: effectiveness of risk mitigation strategies. Journal of Risk Research, 15(4), pages 417–433.

Acerbi, C., Nordio, C., Sirtori, C., (2001). Expected Shortfall as a tool for financial risk management. IDEAS, February 2001. [Accessed on: 10 October 2010].

Acerbi, C., Tasche, D., (2003). Expected Shortfall: A natural coherent alternative to value at risk. Economic Notes, 31(2), pages 379–388.

Alkoffash, M.S., Bawaneh, M.J., Al Rabea, A.I., (2008). Which Software Cost Model to Choose in a Particular Project. Journal of Computer Science, 4(7), pages 606-612.

Artzner, P.; Delbaen, F.; Eber, J. M.; Heath, D., (1999). Coherent Measures of Risk, Mathematical Finance, 9 (3), pages 203-228.

Aven, T., Renn, O., (2009). On risk defined as an event where the outcome is uncertain. Journal of Risk Research, 12(1), pages 1–11.

Bakker de, K., Boonstra, A., Wortmann, H., (2010). Does risk management contribute to IT project success? A meta-analysis of empirical evidence, International Journal of Project Management, 28(5), pages 493–503.

Daníelsson, Jón, Jorgensen, Bjørn N., (2005). Subadditivity Re–Examined: the Case for Value–at–Risk, Cornell University Open access.

Bannerman, P., (2008). Risk and risk management in software projects: a reassessment. Journal of Systems and Software, 81(12), pages 2118-2133.

Barry, Evelyn J., Mukhopadhyay, Tridas, Slaughter, Sandra A., (2002). Software Project Duration and Effort: An Empirical Study. Information Technology and Management, 3(1-2), pages 113-136.

Baucus, D.A., Golec, J.H., Cooper, J.R., (1993). Estimating risk–return relationships: an analysis of measures. Strategic Management Journal, 14(5), pages 387–396.

Boehm, B.W., (1991). Software risk management: principles and practices. IEEE Software, 8(1), pages 32–41.

Boehm, B., Abts, C., Brown, A., Chulani, S., Clark, B., Horowitz, E., Madachy, R., Reifer, D., (2000). Software Cost Estimation with COCOMO II, Prentice Hal.

Boehm, B., Sullivan, K., (1999). Software economics: status and prospects. Information and Software Technology, 41(14), pages 937–946.

Braga, Petrônio L., Oliveira, Adriano L. I., (2007). Software Effort Estimation using Machine Learning Techniques with Robust Confidence Intervals, 7th International Conference on Hybrid Intelligent Systems, pages 352 – 357.

Chakravarthy, B.S., (1986). Measuring strategic performance. Strategic Management Journal, 7(5), pages 437–458.

Connor, AM, MacDonell, SG, (2005). Stochastic cost estimation and risk analysis in managing software projects, Proceedings of the ISCA 14th International Conference on Intelligent and Adaptive Systems and Software Engineering (IASSE), Toronto, Canada, ISCA, pages 140-144.

Costa, H.R., Barros, M.O., Travassos G.H., (2007). Evaluating software project portfolio risks. Journal Systems and Software, 80(1), pages 16 – 31.

Dangle, Kathleen, (2012). Developing a Cost Estimation Probability Model of a Large Multi-Year System – An Experience report. [Accessed on: March 2015]

Delbaen, F., (2002). Coherent risk measures on general probability space, Advances in Finance and Stochastics, pages 1-37.

Dey, P., Tabucanon, M., Ogunlana, S., (1994). Planning for project control through risk analysis: a petroleum pipeline-laying project. International Journal of Project Management, 12(1), pages 23–33.

Dillibabu, R., Krishnaiah, K., (2005). Cost estimation of a software product using COCOMO II 2000 model - A case study. International Journal of Project Management, 23(4), pages 297–307.

Evans, D. S. and J. J. Heckman., (1984). A Test for Subadditivity of the Cost Function with an Application to the Bell System. American Economic Review, 74, pages 615-623.

Fairley, R., (1995). Risk management for software projects. IEEE Software, 11(3), pages 57–67.

Fichman, Robert G., Kemerer, Chris F., (2002). Activity Based Costing for Component-Based Software Development. Information Technology and Management, 3(1-2), pages 137-160.

Foss, T., Stensrud, E., Kitchenham, B., Myrtveit, I., (2003). A simulation study of the model evaluation criterion MMRE, IEEE Transactions on Software Engineering, 29(11), pages 985 - 995.

Fujita, H., (2010). New Trends in Software Methodologies, Tools and Techniques: Proceedings of the 9th SoMeT_10, vol. 217, Frontiers in Artificial Intelligence and Applications, IOS Press, 1st edition.

Galorath, D., (2008). Software project failure costs billions, better estimation & planning can help. [Accessed on: 10 October 2011].

Georgakopoulos, Nicholas L., (2005). Principles and Methods of Law and Economics. Cambridge University Press.

Grimstad, Stein, Jørgensen, Magne, Moløkken-Østvold, Kjetil, (2006). Software effort estimation terminology: The tower of Babel, Information and Software Technology, 48(4), pages 302-310.

Guo, Y., (2010). Deeper understanding, faster calculation - Exam P/1 Insights & shortcuts, study manual. Actuarial Outpost, May 25. [Accessed on: 5 October 2011]

Hamdan, K., Bibi, S., Angelis, L., Stamelos, I., (2009). A Bayesian belief network cost estimation model that incorporates cultural and project leadership factors, IEEE Symposium on Industrial Electronics & Applications, pages 985–989.

Haughey, D., (2009). Why software projects fail and how to make them succeed. ProjectSmart, December. [10 December 2010].

Huang, S.J., Han, W.M., (2008). Exploring the relationship between software project duration and risk exposure: a cluster analysis. Information and Management, 45(3), pages 175–182.

Jørgensen, M., (2005). Practical Guidelines for Expert-Judgment-Based Software Effort

Estimation, IEEE Software 22 (3), pages 57–63.

Jørgensen, M., Shepperd, M., (2007). A systematic review of software development cost estimation studies, IEEE Transactions on Software Engineering, 33 (1), pages 33–53.

Jørgensen, Magne, Moløkken-Østvold, Kjetil, (2004a). Reasons for Software Effort Estimation Error: Impact of Respondent Role, Information Collection Approach, and Data Analysis Method, IEEE Transactions on Software Engineering, 30(12).

Jørgensen, Magne, Moløkken-Østvold, Kjetil, (2004b). Eliminating Over-Confidence in Software Development Effort Estimates, vol. 3009, Lecture Notes in Computer Science, pages 174-184.

Karen, L., Bramble, M., Hihn, J., Hackney, J., Khorrami, M., Monson, E., (2003). Handbook for Software Cost Estimation, Jet Propulsion Laboratory.

Keil, Mark, Wallace, Linda, Turk, Dan, Dixon-Randall, Gayle, Nulden, Urban, (2000). An investigation of risk perception and risk propensity on the decision to continue a software development project, Journal of Systems and Software, 53(2), pages 145-157.

Khamooshi, H., Cioffi, D.F., (2009). Program risk contingency budget planning. IEEE Transaction on Engineering Management, 56(1), pages 171–179

Kitchenham, B., Linkman, S., (1997). Estimates, uncertainty and risk. IEEE Software, 3, pages 69¬–74.

Kitchenham, B., Pfleeger, Shari L., McColl, B., Eagan, S., (2001). An Empirical Study of Maintenance and Development Estimation Accuracy, Technical report-TR/SE-0102.

Kitchenham, B., Pfleeger, Shari L., McColl, B., Eagan, S., (2002). An empirical study of maintenance and development estimation accuracy, The Journal of Systems and Software, 64, pages 57–77

Kitchenham, B., Pickard, M., Linkman, S., Jones, W.P., (2003). Modeling software bidding risks. IEEE Transactions on Software Engineering, 29(6), pages 542 – 554.

Lange, Kenneth, (2003). Applied Probability, Springer, 1st edition.

Lederer, Albert L., Prasad, Jayesh, (1995). Causes of inaccurate software development cost estimates, 31(2), November 1995, pages 125–134.

Leea, Anita, Cheng, Chun H., Balakrishnanc, Jaydeep, (1998). Software development cost estimation: Integrating neural network with cluster analysis, Information and Management, 34(1), pages 1–9.

Li, J., Ruhe, G., Al-Emran, A., Richter, M.M., (2007). A flexible method for software effort estimation by analogy, Journal of Empirical Software Engineering, 12 (1).

Lindland, O.I., Sindre, G., Solvberg, A., (1994). Understanding quality in conceptual modeling. IEEE Software, 2(11), pages 42–49.

Lum, K., Bramble, M., Hihn, J., Hackney, J., Khorrami, M., Monson, E., (2003). Handbook for Software Cost Estimation, Pasadena, California, Jet Propulsion Laboratory.

Lum, K., Powell, J., Hihn, J., (2002). Validation of spacecraft software cost estimation models for flight and ground systems. [Accessed on: 5 December 2010].

Miller, K.D., Reuer, J.J., (1996). Measuring organizational downside risk. Strategic Management Journal, 17(9): pages 671–691.

Moataz, Ahmed, Ahmed, Irfan, AlGhamdi, Jarallah, (2013). Probabilistic size proxy for software effort prediction: A framework, Information and Software Technology, 55, pages 241-251.

Moløkken-Østvold, Kjetil, Jørgensen, Magne, (2005). A comparison of software project overruns - flexible versus sequential development models, IEEE Transactions on Software Engineering 31(9), pages 754-766.

Montgomery, D.C., Runger, G.C., (2007). Applied Statistics and Probability for Engineers, Hoboken, NJ: John Wiley & Sons Inc.

Navlakha, Jainendra K., (1990). Choosing a software cost estimation model for your organization: A case study, Information and Management, 18(5), pages 255–261.

Nisar, M.W., Yong-Ji, W., Elahi, M., (2008). Software development effort estimation using fuzzy logic – a survey, Fifth International Conference on Fuzzy Systems and Knowledge Discovery, FSKD-2008, pages 421–427.

Papoulis, A., (1991). Probability, Random Variables, and Stochastic Processes, 3rd edition, McGraw Hill Companies.

Park, H., Baek, S., (2008). An empirical validation of a neural network model for software effort estimation, Journal of Expert Systems with Applications 35 (3), pages 929–937.

Pendharkar, Parag C., Subramanian, Girish H., Rodger, James A., (2005). A probabilistic model for predicting software development effort, IEEE Transactions on Software Engineering, 31(7).

Pfleeger, S.L., Atlee, J.M., (2006). Software Engineering, Theory and Practice, 3rd edition, Pearson International Edition.

Putnam, L.H., (1978). A general empirical solution to the macro software sizing and estimating problem, IEEE Transactions on Software Engineering, (4) 345–361.

Ross, S.M., (2007). Introduction to Probability Models, 8th edition, Elsevier.

Royden, Halsey, Fitzpatrick, Patrick, (2010). Real Analysis, 4th Edition, Pearson.

Ruefli, T.W., Collins, J.M., Lacugna, J.R., (1999). Risk measures in strategic management research: auld lang syne? Strategic Management Journal, 20(2), pages 167–194.

Touran, A., (2003). Calculation of contingency in construction projects. IEEE Transactions on Engineering Management, 50(2), pages 135–140.

Shepperd, M, Schofiel, C., (1997). Estimating software project effort using analogies,

IEEE Transactions on Software Engineering. 23 (11), pages 736–743.

Sherer, Susan A., (1994). Measuring software failure risk: Methodology and an example Original Research Article, Journal of Systems and Software. 25(3), pages 257-269.

Sommerville, Ian (2007). Software Engineering, 7th edition, Pearson Education.

Stein, Grimstad, Jørgensen, Magne, Moløkken-Østvold, Kjetil, (2006). Software effort estimation terminology: the tower of Babel, Information and Software Technology. 48, pages 302-310.

Stewart, Rodney D., Wyskida, Richard M., Johannes, James D., (1995). Cost Estimator's Reference Manual, John Wiley & Sons.

Uzzafer, Masood, (2010a). A Financial tool for Software Risk Measurement, International Conference on Information Science and Applications, ICISA-2010. Seoul, Korea. ISBN: 978-1-4244-5941-4: pages 1–6.

Uzzafer, Masood, (2010b). A pitfall of software estimated cost, Proceedings of IEEE International Conference on Information Management and Engineering , ICIME-2010, Chengdu, China, pages 578-582.

Uzzafer, Masood, (2013a). A simulation model for strategic management of software projects, Journal of System and Software. 86(1), pages 21–37.

Uzzafer, Masood, (2013b). A Contingency Estimation Model for Software Projects, International Journal of Project Management, 31(7), October 2013, pages 981–993.

Veerarajan, T., Probability, Statistics, and Random Processes, 3rd edition, (2008). Tata McGraw Hill Education.

Wit, A. de, (1988). Measurement of project success, International Journal of Project Management, 6(3), pages 164–170.

Yamai, Y., Yoshiba, T., (2005). Value-at risk versus expected shortfall: a practical perspective. Journal of Banking and Finance, 29(4), pages 997–1015.