Journal of Physical Science and Innovation
ISSN: 2277-0119
Volume 9, No. 2, 2017

EFFECT OF DIFFERENT ACTIVITY DURATION DISTRIBUTION ON PROJECT DURATION PREDICTION PROBLEM

Adepeju A. Opaleye[1] & Oliver E. Charles-Owaba[2]
[1 &2] Department of Industrial and Production Engineering,
University of Ibadan, Nigeria
Email: opaleye_adepeju@yahoo.com

ABSTRACT

Project scheduling researchers mostly rely on beta or triangular distribution for modelling activity duration. This reliance is associated with the underlying assumption of the traditional PERT model. Contemporary researches in this area have however partly refuted this concept and many different distributions proposed without empirical justification. In this study, availability of historical project activity duration data in construction industry was investigated and associated statistical distribution empirically determined. Effects of these distributions on the project completion time were also experimentally investigated. It is shown that about 49.23% of activities durations observed exhibited lognormal distribution and a significant difference in predicted project duration exist due to varying activity duration distribution.

INTRODUCTION

The dilemma posed by project completion time and cost over-runs is daunting to, not only project managers, but also to the scientific community. The literature establishes three prongs of the problem situation here; the first is the difficulty of developing and applying more accurate project activity duration estimating models. The second, which is closely related to the first, is the hitch associated with finding realistic set of input data for estimating activity duration. The third is lack of reliable contingency action plan to handle eventualities arising from the inevitable working environmental and technical uncertainties during project execution (Opaleye, et al., 2017). A good estimating model is usually a statistical distribution which adequately replicates the behavior of well executed work system of identical/similar projects in the environment. Most published literatures in project scheduling uses probabilistic distribution models to estimate expected activity duration. Such distribution must be continuous, limited between two positive time intercepts, have a unique mode in its defined range and capable of describing both skewed and symmetric activity time distributions (Trietsch et al., 2012) Although, it is argued that the second assumption may not necessarily hold because of the difficulty in determining the value of the range of the estimate, the authors of Program Evaluation and Review Technique (PERT) made these four assumptions regarding activity duration distributions which led to a particular form of beta distribution known as PERT-Beta. Beta distribution is one of the most commonly used in modeling uncertainty in activity duration. When a standardized beta distribution is bounded by zero and one [0,1] and it is referred to as a "two-parameter beta distribution". It is useful for representing uncertainty in a fraction that cannot exceed one while a beta distribution bounded by optimistic and pessimistic values is considered to be a "four-parameter beta" or generalized.

The beta distribution is also known to be asymmetrical. This property is desirable for modeling activity duration which is often skewed to the right by unlikely but severe overruns. It is flexible and can take on many different shapes, including flat, narrow, U and inverted-U shapes exhibited by other distribution models. Program Evaluation and Review Technique (PERT), is the first project planning technique to considered uncertainty in activity duration estimation. In the model formulation, the authors (Malcolm et al., 1959)require three points elicited values of activity duration; optimistic, most of the times and pessimistic values and implicitly assumed beta distribution function is suitable for schedule risk analysis. Using the pre-assumed probability $(\frac{1}{6},\frac{4}{6},\frac{1}{6})$ of observing each point value, expected activity duration was determined as a weighted average in PERT. Since the publication of the model over six decades ago, it remains a subject in many scheduling literatures; criticized based on the simplicity of its underlying assumptions. Critics reveal that fundamentals of PERT's activity duration estimation postulation: that actually observed duration may agree with well thought out plan (most of the time estimate); fall below plan (optimistic) or exceed plan (pessimistic) is not theoretically untrue but the choice of statistical distributions, manner of estimating parameter values and derivation of duration from parameter values are unjustified (Khamooshi and Cioffi, 2013). As such researchers propose new PERT approximations by modifying its underlying assumptions (Herrerías-velasco,et al.,2011,Sireesha and Shankar, 2010 Premachandra, 2001) or suggest alternative distributions such as; triangular, lognormal, normal distribution etc. (Cottrell, 1999; Jannat, 2012; Mccombs et al., 2009; Mohan et al., 2007; Trietsch et al., 2012)in an attempt to develop more accurate activity duration prediction models. The first group of authors to modify PERT model relaxed the restriction on the beta distribution shape parameters in order to arrive at their prediction models considered more accurate while retaining the traditional three PERT times elicited as minimum, maximum and modal values.

Approximations in this category include; Herrerías-velasco,et al.,(2011)and Premachandra(2001). A similar and more recent weighted average model approximation is by Sireesha and Shankar (2010) who suggests pre-assumed probability $(\frac{5}{27}, \frac{17}{27}, \frac{5}{27})$of observing each elicited value; optimistic, most of the times and pessimistic values. Since there is no empirical justification for the use of beta distribution, alternative distributions are considered. One of such employed in project management literature is the triangular distribution. It is considered easy to estimate and as good as other distributions proposed in project management literature. Back *et al* (2000) studied the use of beta and triangular distributions in estimating project cost data concluding that there were not significant differences but suggested triangular distribution as a preferred model because it's easy to estimate its parameters. Likewise,Kotz and

Van Dorp (2004) explored the advantages of using triangular distribution over beta distribution. Jaskowski*et al* (2011) develop a model for estimating activity duration distribution parameters based on the assumption that they are triangularly distributed. Holm and Barra (2011)on the other hand presented a model of the Emergency Department (ED) of a Norwegian hospital. They concluded that a model with beta distributions based on the SME estimates outperforms a model with the more frequently used triangular distributions. These results present a mixed feeling on the use of beta and triangular distribution. Another alternative to PERT-beta was proposed by Cottrell (1999) based on the use of normal distribution function. The approximation was to reduce the number of estimates required for activity durations from three, as in conventional PERT, to two. Cottrel model is however close to that of PERT only if data is not highly skewed. Mohan et al. (2007) and Trietsch et al., (2012) proposed the lognormal distribution as a simplified version of PERT . McConmbs et al ., (2009) suggest that weibull distribution is also a good alternative that does not require approximations for the mean and variance.

However, there is no evidence in the literature that any of these distributions has been empirically validated. Authors proposal are based on intuition and ease of distribution parameter estimation. Furthermore, for any of the prediction models to be accurate, underlying activity duration distribution must be as assumed in the model development and expert elicited values must be realistic. This study therefore investigates availability of historical project activity duration data in construction industry and determines associated statistical distributions. Effect of the choice of statistical distributions in project duration prediction problem was also experimentally investigated. It is noteworthy that the use of empirical data to validate the distribution of each project activity may be cumbersome and time-consuming however, the rigor may be necessary in making accurate prediction.

METHODOLOGY

In this section, procedures for sourcing of project activity duration historical data, empirical investigation of statistical distributions of observed activity duration and experimental investigation of the possible effects of the choice of statistical distributions on project duration prediction are presented.

Sourcing of Project Activity Historical Data

The study started with sourcing for project data. Information required is the observed activity duration data. Associated are observed activity cost and causes of variation from plan. Hence, it is necessary to determine the extent construction firms keep such data. First, letters were sent to some registered construction firms requesting their participation, particularly in volunteering the needed data. These companies were selected through a convenience method of non-probability sampling technique from the National Database of Federal Contractors and Service Providers (NDCCSPs) by the Bureau of Public Procurement (BPP) in Nigeria. Afterwards, a survey to determine data availability and accessibility in construction industry was developed. The form was administered through postal mail, e-mails, and personal contacts in companies earlier contacted. Records of companies who provided positive response on their willingness to assist with data were vetted and operations of ongoing activities observed. Sets of activity duration data collected was organized for analysis.

Determination of Statistical Distribution of Historical Data

The objective of this section is to investigate the pattern of statistical distribution of observed activity duration data. This is achieved by calculating the sample statistics and relating them to the statistics of well-known distributions using Kolmogorov-Smirnov (KS) test.

HYPOTHESIS TESTING

Kolmogorov-Smirnov (KS) test measures the probability that a chosen univariate dataset is drawn from a hypothesized distribution or of the same parent population as a second dataset (19) is used.

Given that $d_1 \leq \; d_2 \ldots d_i \ldots \leq . d_{k_i}$ are observed point durations of activity i for period k periods with observed/empirical cumulative distribution function (EDF), $F_O(\, d_i)$expressed as;

$$F_o(\, d_k) = \frac{1}{z_i} * [number\ of\ observations \;\; \leq \; d_k] \qquad (1.0)$$

KS test statistics which compares $F_o(\, d_k)$ and $F_o(\, d_{k-1})$ with a hypothesized/theoretical cumulative distribution function (CDF) $F_E(\, d_k)$ is given as; $D_i = max_{1\leq k\leq z_i} max[(F_o(d_k) - F_E(d_k) \;\; , \; F_o(d_{k-1}) - F_E(d_k))]$ (1.1)

Where D: Measures the maximum difference between the empirical cumulative distribution function (EDF) of the observed sample data and CDF of the hypothesized distribution. Therefore, equality of the empirical and hypothesized distribution can be tested by comparing the statistic D_i to 0 (if D_i is significantly larger than 0 and close to 1, then we might conclude that the distributions are not equal)(Arnold & Emerson, 2011). Feigelson and Babu (2012) pointed out that the theory underlying the KS test requires independence between the EDF and CDF curves under consideration. Thus, model parameters must be derived from another dataset, or the significance level of the difference between the curves can be estimated by bootstrap resamples of the original dataset. This study therefore uses bootstrapped KS test because model parameters are derived from the dataset. The Kolmogorov-Smirnov (KS) test is considered for the following hypothesis;

Hypothesis 1:

H_{01i}: The distribution of the observed data belongs to a class of beta distribution functions.

H_{11i}: The distribution of the observed data does not belong to a class of beta distribution functions.

In a similar manner, the statements for the null and alternative hypothesis are made for the following;

Hypothesis 2: Triangular distribution function

Hypothesis 3: Lognormal distribution function

Hypothesis 4: Normal distribution functions

Hypothesis 5: Uniform distribution functions

Hypothesis 6: Weibull distribution functions

Hypothesis 7: Exponential distribution functions

Model Fitting

Beta Distribution

The general characterization of the four-parameter beta distribution having shape parameters $\alpha, \beta > 0$ is

$$f_x(\alpha, \beta, a, b) = \begin{cases} \frac{(x-a)^{(\alpha-1)}(b-x)^{\beta-1}}{(b-a)^{\alpha+\beta-1}\,\hat{A}(\alpha,\beta)} & \text{if } a \le x \le b \\ 0 & \text{otherwise} \end{cases} \qquad (1.2)$$

where 'a' is the lower bound or optimistic value, 'b' is the upper bound or pessimistic value of the distribution. The Cumulative Distribution Function (CDF) over the range a , b and shape parameters α *and* β is given by

$$F_x(\alpha, \beta, a, b) = \begin{cases} 0, & if\ x < a \\ \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)} \int_b^x \frac{(x-a)^{(\alpha-1)}(a-x)^{\beta-1}}{(b-a)^{\alpha+\beta-1}}\, dt, & if\ a \le x \le b \\ 1, & if\ x > b \end{cases} \qquad (1.3)$$

With the stochastic characteristic parameters of the distribution of x given as:

Mode: $$m_x = \frac{\alpha-1}{\alpha+\beta-2}(b) + \frac{\beta-1}{\alpha+\beta-2}(a) \qquad (1.3)$$

Mean: $$\mu_x = E(x) = a + (b-a)\frac{\alpha}{\alpha+\beta} \qquad (1.5)$$

Variance: $$\sigma^2{}_x = \frac{\alpha\beta(b-a)^2}{(\alpha+\beta-1)(\alpha+\beta)^2} \qquad (1.6)$$

α *and* β, may be obtained as follows (Davis, 2008);

$$\alpha = \left(\frac{\mu-a}{b-a}\right)\left[\left(\frac{(\mu-a)(b-a)}{\sigma^2}\right)-1\right] \tag{1.7}$$

$$\beta = \left(\frac{b-\mu}{b-a}\right)\left[\left(\frac{(\mu-a)(b-\mu)}{\sigma^2}\right)-1\right] \tag{1.8}$$

Triangular Distribution

The probability density function of a triangular distribution is defined as

$$f(x|\mathrm{a,b,m}) = \begin{cases} \dfrac{2(x-\mathrm{a})}{(\mathrm{b-a})(\mathrm{m-a})}, & for\ \mathrm{a} \le x < \mathrm{m} \\ \dfrac{2}{\mathrm{b-a}}, & for\ \ x = \mathrm{m} \\ \dfrac{2(\mathrm{b}-x)}{(\mathrm{b-a})(\mathrm{b-m})}, & for\ \ \mathrm{m} < x \le \end{cases}$$

where$a$ $< b$ and $a \le m$ $\le b$. (1.9)

The cumulative function is given as

$$F(x|\mathrm{a,m,b}) = \begin{cases} \frac{(x-\mathrm{a})^2}{(\mathrm{b-a})(\mathrm{m-a})}, & for\ \mathrm{a} \le x < \mathrm{m} \\ \frac{\mathrm{m-a}}{\mathrm{b-a}}, & for\ \ x = \mathrm{m} \\ 1-\frac{(\mathrm{b}-x)^2}{(\mathrm{b-a})(\mathrm{b-m})} & for\ \mathrm{m} < x \le \mathrm{b} \end{cases} \tag{1.10}$$

(Garg et al.,2009).

'a' is defined as the minimum possible value of dataset x; 'm' is defined as the mode or most likely value of x and 'b' is defined as the maximum value of x

Normal Distribution

For normally distributed activity duration, the probability density function is defined as

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad ; \tag{1.11}$$

For all real numbers , 'μ' is the location parameter equal to the mean and 'σ' is the standard deviation and $-\infty<x<\infty$, $-\infty<\mu<\infty$, $\sigma>0$. The cumulative function is given as

$$F(x) = \int_{-\infty}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \mathrm{dx} \tag{1.12}$$

(Polyanin, and Manzhirov, 2008).

$$\mu = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad (1.13)$$

$$\sigma^2 = S^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \qquad (1.14)$$

Lognormal Distribution

The central limit theorem is also a basis for selecting the lognormal distribution. Unlike the normal distribution, it is a good representation of non-negative, positively skewed quantities. If the distribution of activity duration is assumed to be lognormal distributed, the probability density function is given as

$$y = f(x|\mu,\sigma) = \frac{1}{x\sigma_{\ln(x)}\sqrt{2\pi}} e^{\frac{-(lnx-\mu_{\ln(x)})^2}{2\sigma^2}} \qquad (1.15)$$

The cumulative function is $F(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\int_0^x \frac{e^{\frac{-(\ln(t)-\mu)^2}{(2\sigma)^2}}}{t}\, dt$ (1.16) (Frey and Rhodes, 1999).

Uniform Distribution

The uniform distribution has two parameters a and b with its probability density function expressed as

$$f(x|\mathbf{a},\mathbf{b}) = \begin{cases} \frac{1}{\mathbf{b}-\mathbf{a}} & , \mathbf{a} \le x \le \mathbf{b} \\ 0 \ , \ elsewhere \end{cases} \qquad (1.17)$$

and the cumulative distribution function (CDF) as,

$$F(x) = \begin{cases} 0, & x \le \mathbf{a} \\ \frac{x-\mathbf{a}}{\mathbf{b}-\mathbf{a}} & , \mathbf{a} \le x \le \mathbf{b} \\ 0, \mathbf{b} \le x \end{cases} \qquad (1.18)$$

'a' is defined as the minimum possible value of dataset x and 'b' is defined as the maximum value of x.

Weibull Distribution:

Weibull distribution has its two parameters both positive constants (α, β) that determine its location and shape. It is considered to accommodate a longer tail probability than is allowed by the beta distribution.

The probability density function (pdf) of weibull is

$$f(x) = \begin{cases} \alpha\beta^{\alpha}x^{\alpha-1}e^{-(\beta x)^{\alpha}} & x > 0 \\ 0, \quad x \leq 0 \end{cases} \tag{1.19}$$

The weibull cumulative distribution function (CDF) may be computed by integrating its pdf

$$F(x) = P(X \leq x) = \begin{cases} 1 - e^{-(\beta x)^{\alpha}} & x > 0 \\ 0 \quad x \leq 0 \end{cases} \tag{1.20}$$

(Nwobi,2014) .

Exponential Distribution

The probability density function (pdf) of exponential distribution is

$$f(x) = \frac{1}{\mu}e^{\frac{-x}{\mu}} \tag{1.21}$$

While the CDF is expressed as

$$F(x|\mu) = 1 - e^{\frac{-x}{\mu}} \tag{1.22}$$

Procedure for Bootstrap Analysis of the Kolmogorov-Smirnov (KS) test

The procedure for bootstrap KS test is as follows;

Step 1: Specify theoretical distribution, F_E.

Step 2: Compute nominal parameters values from the observed duration data i as appropriate for F_E

Step 3: Obtain $D_i$from the observed sample data, i.

Step 4: Independently draw with replacement Q samples of size z_i from the observed data$i$ (the bootstrap samples). Denote these $q$samples; $q \in \{1, \ldots \ldots . Q\}$

Step 5: From each of the bootstrap sample, evaluateD_q to obtain Q bootstrap replicates of D_i.

Step 6: Order KS statistics D_q

Step 7: Obtain an approximate $pvalue$

$$pvalue \approx \frac{Number \ \#(D_q \geq D_i)}{Q} \tag{1.23}$$

Step 8: If a test with a level of significance $\alpha = 0.05$ is desired, $F_E$is an 'Acceptable Fit' if $if\ pvalue > 0.05$ i.e.

reject$H_o, if\ pvalue < 0.05$,

Otherwise do not reject H_o(Stute et al., 1993)

Step 9: Repeat steps 2-8 for all specified distributions,F_E, or hypothesis for i.

Step 10:F_E **is a '**Best Fit' distribution for i if it has the maximum $pvalue$ of all specified distributions

Step 11: Tabulate results and show remark as the best fit distribution.

Experimentation and Statistical Investigation

In order to verify effect of underlying activity duration distribution on project duration prediction problem, we simulated critical paths with different numbers of activities and distributions. Statistical measures of the average inaccuracy associated with each choice of distribution-produced estimates was computed using three (3) different error terms. To summarize our approach, the necessary steps are given as follows.

Step 1: Specify an assumed distribution (F_E) for all project activities with parameters p_i . For example, beta distribution.

Step 2: Generate 'n' samples of critical activity duration from the assumed distribution (F_E).

Step 3: Compute nominal parameters values from the generated duration data$i$ as appropriate for alternative distributionF_E'.

Step 4: Compute expected project completion time (PCT) and predicted project completion time $(\widehat{\text{PCT}})$ for each set of sample given F_E and F_E' respectively.

Step 5: Evaluate consistency of alternative distribution,F_E' using the bias, mean absolute error (MAE) and mean square error (MSE) with the following expression;

$$Bias = \frac{\sum_{i=1}^{N} PCT - \widehat{\mathrm{PCT}}}{N} \qquad (1.24)$$

$$MAE = \frac{\sum_{i=1}^{N} |PCT - \widehat{\mathrm{PCT}}|}{N} \qquad (1.25)$$

$$MSE = \frac{\sum_{i=1}^{N} (PCT - \widehat{\mathrm{PCT}})^2}{N} \qquad (1.26)$$

Step 6: Repeat step 1 to 5 given n=10,20,30,40,50,60,70,80,90,100 samples of critical activity duration.

Step 7: Repeat steps 1 – 6 for $F_E$and F_E' given as Beta distribution, Triangular distribution, Lognormal distribution, Normal distribution, Uniform distribution, Weibull distribution and Exponential distribution.

Step 8: Tabulate the Bias, MAE and MSE across statistical distribution and sample size.

Step 9: Using F test statistics of the one way analysis of variance (ANOVA), investigate significance of Bias, MAE and MSE across distribution.

Data Collection, Presentation, Analysis and Discussion of Results

In this section, data collected were analyzed and the results presented and discussed.

Data Collection

The survey consists of questions on availability of observed activity duration data, observed activity cost data, causes of variation and the willingness to provide such for research purpose. A total of sixty – two contracting construction firms were contacted. Summary on the category of response and observed information is provided on Table 1.0.

Table 1.0: Summary of Survey Information

Category Description	Number	Percentage of respondents
Total number of construction firms contacted	62	-
Total number of firms that responded	42	
Number of firms willing to provide activity duration data	15	32%
Number of firms willing to provide activity cost data	22	89%
Number of firms willing to provide data on causes of variation	56	97%
Number of firms that provided observed activity duration data	3	7.1%
Number of firms that provided observed activity cost data	4	9.5%
Number of firms that provided observed causes of variation data	4	9.5%
Number of firms that provided observed activity duration & cost data	2	4.7%
Number of firms that provided observed activity duration & causes of variation data	-	-
Number of firms that provided observed activity duration, cost & causes of variation data	1	0.02

A five- period drilling activity duration data, eight-period building activity duration data, and twenty-five period building activity duration data were obtained from three different contracting firms. From the summary result of the survey on Table 1.0, a significant difference in the number of positive responses to the survey and those that provided requested (observed) data during visitation was observed. For instance, fifteen firms indicated willingness to provide historical (observed) activity duration data; however, only three (3) willingly provided some useful data on the observed duration data upon visitation to the company. Of the three firms, (Firm I, Firm II and Firm III) only two were willing to provide associated cost data.

Data Presentation, Analysis and Discussion

In this section the three sets of observed activity duration for the three projects types were analyzed and associated statistical distributions of each activity determined.

Parametric Analysis of Project Data

The analysis is achieved following the procedure in sub-section 2.2;

Step 1: Seven types of probability distributions, F_E; Beta, Triangular, Normal, Lognormal, Uniform, Weibull and Exponential distribution are specified.

Step 2: Nominal parameters values obtained using the Maximum Likelihood Estimation Method (MLE) and Methods of Moment (MoM) as appropriate.

Step 3-10: The bootstrap analysis of the dataset with Q=1,000 samples of each of project activity duration data was generated with a sample size Z_i, to develop 1,000 replications of the Kolmogorov Smirnov statistics (D_q). The p-values from 1000 replicates of D_q for activities of the three Firms are as shown on Table 1.1 to Table 1.4. These are used to test the hypothesis stated in section 2.2:3.

Table 1.1: Bootstrap KS Test Statistics, $p\ value$ (Firm I)

Activity	Beta	Triangular	Normal	Lognormal	Uniform	Weibull	Exponential	Remark
A	0.4952	0.0418	0.5181	0.8197	0.0418	0.6577	0.1264	Lognormal
B	0.8963	0.5143	0.8811	0.6752	0.5143	0.9542	0.2733	Weibull
C	0.1865	0.0048	0.2217	0.291	0.0048	0.2609	0.1521	Lognormal
D	0.9128	0.8023	0.9241	0.8455	0.8023	0.9519	0.0816	Weibull
E	0.9416	0.806	0.9636	0.9606	0.806	0.9351	0.1728	Normal
F	0.7519	0.156	0.8109	0.8612	0.156	0.8062	0.0633	Lognormal
G	0.895	0.7401	0.816	0.9087	0.5814	0.8696	0.5388	Lognormal
H	0.3681	0.0267	0.3907	0.8304	0.0267	0.6523	0.6413	Lognormal
I	0.6473	0.4158	0.6973	0.7883	0.4158	0.6196	0.1121	Lognormal

Table 1.2: Bootstrap KS Test Statistics, $p\ value$ (Firm II)

Activity	Beta	Triangular	Normal	Lognormal	Uniform	Weibull	Exponential	Remark
A1	0.2508	0.0865	0.2356	0.3619	0.0057	0.2843	0.0195	Lognormal
B2	0.7773	0.2183	0.9482	0.9889	0.2183	0.9962	0.6082	Weibull
C3	0.166	0.0202	0.2171	0.6373	0.0019	0.3924	0.1488	Lognormal
D4	0.1352	0.1321	0.1521	0.1718	0.0084	0.1854	0.0064	Weibull
E5	0.5065	3.62E-05	0.4567	0.3917	0.1616	0.4854	0.0426	Beta
F6	0.2855	0.3418	0.2847	0.3532	0.0226	0.2838	0.006	Lognormal
G7	0.0487	0.001	0.0543	0.0982	3.62E-05	0.0781	0.0111	Lognormal
H8	0.7696	0.1616	0.7434	0.8097	0.1616	0.692	0.0096	Lognormal
I9	0.2755	0.0015	0.2897	0.2626	0.1616	0.2137	0.0261	Normal
J10	0.6448	0.0901	0.6134	0.7621	0.0901	0.5707	0.0133	Lognormal
K11	0.1352	0.0015	0.1706	0.195	0.0015	0.1061	0.0531	Lognormal
L12	0.0379	3.62E-05	0.0491	0.0491	3.62E-05	0.0565	0.0016	Weibull
M13	0.6609	0.1513	0.6875	0.7467	0.0536	0.6544	0.0089	Lognormal
N14	0.6056	0.6863	0.5975	0.6159	0.3901	0.5979	0.0052	Triangular
O15	0.3845	0.0226	0.4754	0.4298	0.0226	0.554	0.0056	Weibull
P16	0.1878	0.073	0.1932	0.2394	0.0901	0.1774	0.0083	Lognormal
Q17	0.2463	0.0226	0.2968	0.29681	0.0226	0.2307	0.0083	Lognormal
R18	0.1039	0.0015	0.1309	0.131	0.0015	0.1194	0.0016	Lognormal
S19	0.9728	0.9178	0.9718	0.9714	0.6134	0.9145	0.0064	Beta
T20	0.469	0.2703	0.4856	0.5785	0.0901	0.4235	0.0052	Lognormal
U21	0.089	0.0655	0.0925	0.2014	0.0015	0.1208	0.0367	Lognormal
V22	0.3198	0.0226	0.3851	0.3519	0.0226	0.3658	0.0165	Lognormal
W23	0.6134	0.0226	0.6134	0.4129	0.6134	0.5133	0.0288	Normal
X24	0.6134	0.0226	0.6135	0.515	0.6134	0.4323	0.0074	Normal
Y25	0.3397	0.2703	0.3618	0.5276	0.01	0.3021	0.0064	Lognormal
Z26	0.1624	2.19E-05	0.1616	0.1345	0.01	0.3373	0.0061	Weibull
A27	0.1039	0.0015	0.1309	0.131	0.0015	0.1194	0.0014	Lognormal
B28	0.1059	0.0124	0.123	0.1489	0.0015	0.1049	0.0026	Lognormal
C29	0.8328	0.073	0.8257	0.8111	0.813	0.7268	0.0022	Beta
D30	0.2641	0.0015	0.255	0.3028	0.0015	0.1807	0.002	Lognormal
E31	0.089	0.0655	0.0925	0.1316	0.0015	0.09	0.0056	Lognormal
F32	0.1532	0.3418	0.1639	0.1927	0.0226	0.1828	0.0073	Triangular
G33	0.4057	0.0015	0.4312	0.3969	0.1616	0.5615	0.0028	Weibull
H34	0.2468	2E-08	0.2297	0.128	0.0084	0.1863	0.0895	Beta
I35	0.1722	1.19E-06	0.1639	0.1292	0.0041	0.1945	0.0431	Weibull
J36	0.4431	0.014	0.4209	0.4131	0.014	0.3591	0.0404	Beta
K37	0.4589	0.014	0.4381	0.395	0.014	0.3979	0.1405	Beta
L38	0.522	0.0226	0.5515	0.4318	0.1616	0.5637	0.0083	Weibull

Table 1.3: Bootstrap KS Test Statistics, $p\ value$ (Firm III)

Firm III								
Activity	Beta	Triangular	Normal	Lognormal	Uniform	Weibull	Exponential	BestFit
A	0.9005	0.0598	0.6815	0.6751	0.5532	0.7289	4.52E-05	Beta
B	0.8374	0.0021	0.884	0.7771	0.2363	0.9347	0.0011	Weibull
C	0.4616	3.56E-15	0.327	0.274	0.0944	0.3609	4.58E-05	Beta
D	0.0325	6.34E-05	0.0444	0.0629	6.34E-05	0.0301	0.000131	Lognormal
E	0.3405	0.0238	0.4547	0.7778	0.2179	0.3334	7.21E-06	Lognormal
F	0.765	0.000191	0.8938	0.8876	0.0172	0.5052	2.11E-08	Normal
G	0.7063	7.47E-06	0.5042	0.5895	0.4662	0.4958	0.000144	Beta
H	0.9688	0.7332	0.9723	0.9782	0.7332	0.9077	4.81E-08	Beta
I	0.1857	0.000402	0.3422	0.2173	0.5224	3.36E-05	0.3819	Uniform
J	0.5173	4.47E-06	0.3891	0.5578	0.2239	0.4173	0.0043	Lognormal
K	0.4219	1.61E-08	0.3565	0.2426	0.185	0.6188	2.17E-07	Weibull
L	0.9027	5.98E-10	0.9242	0.8142	0.4945	0.79	1.82E-05	Normal
M	0.1766	0.000282	0.3577	0.5793	8.07E-06	0.1827	6.93E-07	Lognormal
N	0.8834	0.3501	0.9092	0.5533	0.3501	0.9153	0.000334	Weibull
O	0.5246	6.2E-08	0.7625	0.2832	0.0354	0.8863	6.37E-05	Weibull
P	0.6009	0.0086	0.5864	0.7234	0.0607	0.6319	0.001	Lognormal
Q	0.3131	0.000224	0.3915	0.685	0.0082	0.2483	1.68E-06	Lognormal
R	0.9967	0.5521	0.9975	0.9853	0.552	0.9565	0.00001	Normal

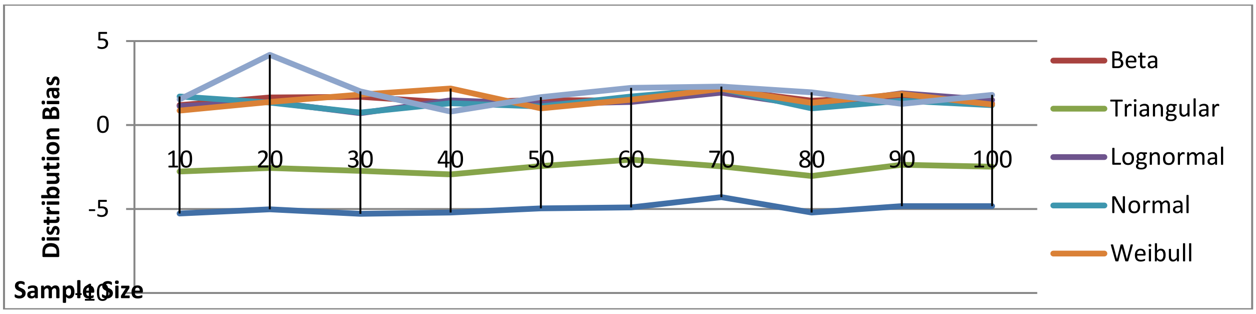
Table 1.4: Distribution Fit for Observed Dataset

Acceptable Fit					
Distribution	Firm I	Firm II	Firm III	Total No.	Percentage of fit
Beta	9	36	17	62	95.385
Triangular	6	16	5	26	40.000
Normal	9	37	17	63	96.923
Lognormal	9	37	18	64	98.462
Uniform	6	15	14	34	52.308
Weibull	9	38	16	63	96.923
Exponential	9	5	1	14	21.538
Best Fit					
Distribution	Firm I	Firm II	Firm III	Total No.	Percentage of fit
Beta	0	6	3	9	13.846
Triangular	0	2	0	2	3.077
Normal	1	3	4	7	12.308
Lognormal	6	19	7	32	49.231
Uniform	0	0	1	1	1.538
Weibull	2	8	3	14	20.000
Exponential	0	0	0	0	0.000
Total No. of Activities	9	38	18	65	100.000

In order to investigate associated statistical distribution of the observed activity duration data (beta, triangular, normal, lognormal, uniform, weibull and exponential distribution), the bootstrap Kolmogorov-Smirnov (KS) test statistic was used. The non-parametric bootstrap creates a large number of datasets that we might have observed and computes the KS statistic on each of these datasets. Thus, accurate asymptotic approximations of the p-value can be obtained following procedure in section 2.2:3. Following results on Table 1.1 to Table 1.4, each of the sixty-five activities from the three different projects may be best modeled with unique distribution and parameters. The summary result presented on Table 1.4 showed that from Firm I, a total of nine activities data were analyzed; Beta distribution is an acceptable fit for the nine activities but not the best fit for any. Similarly, the triangular, uniform and exponential distributions are acceptable for some of the activities but not the best fit for any. Normal, lognormal and weibull distribution on the other hand best fit one (1), six (6) and two (2) of the activities respectively. For Firm II, a total of thirty-eight activities data were analyzed; beta distribution best fit six (6) of the activities, triangular, normal, lognormal and weibull distribution best fit two (2), three (3),nineteen (19) and eight (8) of the activities respectively while uniform and exponential distributions fit none. In general, it can be observed that exponential distributions is not a best fit for any of the sixty-five activities considered while lognormal distribution best fit about 49.231% of the activities. Though, the beta distribution is one of the commonly used distributions in practice due to its versatility, the lognormal distribution seems more applicable for these sets of observed data. Similarly, the overwhelming dependence on the triangular distribution in practice, notwithstanding the fact that most data do not meet the criteria needed for the distribution to fit may be associated with its ease of parameter estimation.

Performance Evaluation

In this section error analysis of beta, triangular, lognormal, normal, Weibull, exponential and uniform distributions under ten (10) case situations (n=10, 20, 30, 40, 50, 60, 70, 80, 90 and 100) were considered. Following steps 1 to 9 in sub-section 2.3, error term across distribution and sample size. Is presented in the appendix while figure 1.0 shows the behavior of Bias, MAE and MSE across distribution and sample size.



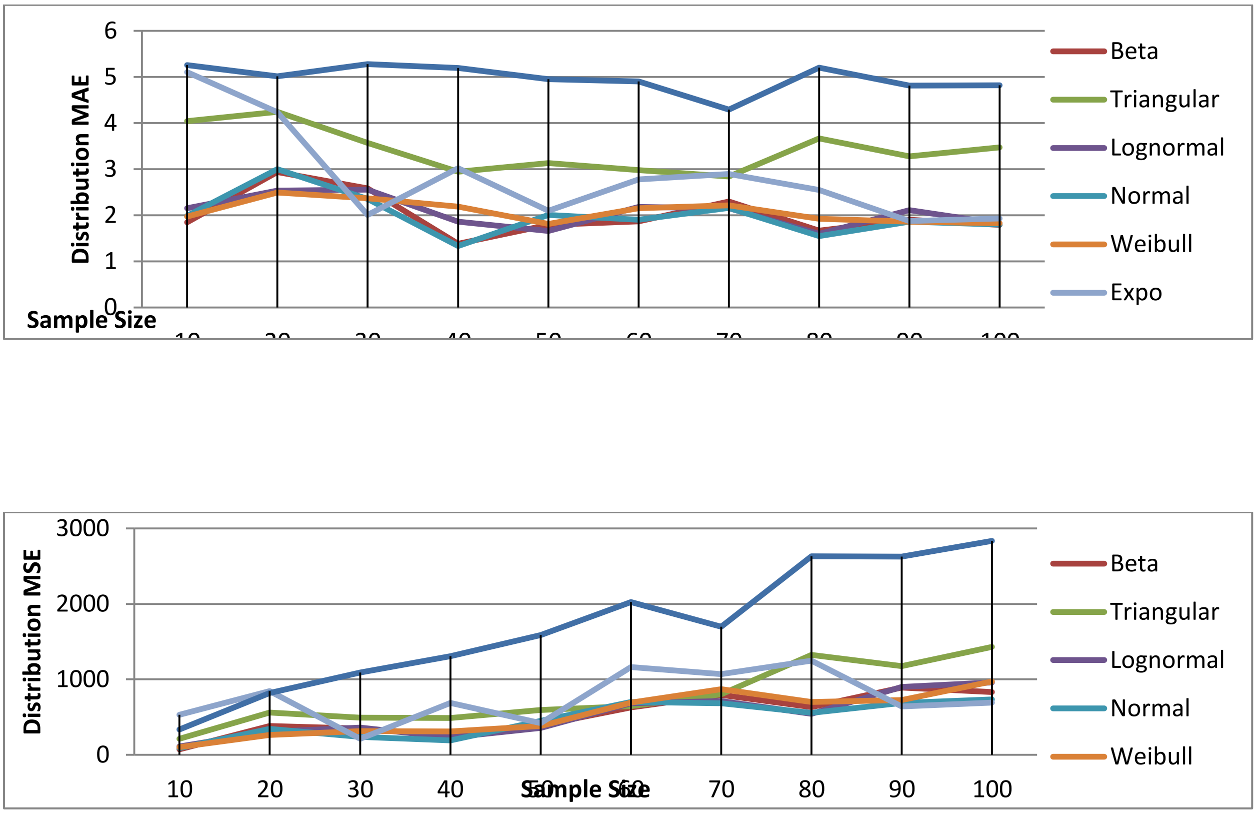


Figure 1.0: Behavior of Error term across distribution & sample size.

Figure 1.0 shows a similar behavior of error across distribution and sample size. Hypotheses were tested to understand effect of activity duration and number of critical activities on project completion time prediction problem.

Table 1.6: Statistical Analysis of Error Term Across Distribution and Sample Size

Error Term	Across	F	P-value	F critical
Bias	Distribution	328.4432	1.51E-45	2.246408
MAE	Distribution	42.65642	2.31E-20	2.246408
MSE	Distribution	10.33423	6.03E-08	2.246408

The summary result of the ANOVA on Table 1.6 shows that the F-ratio for the test at 0.05 significant levels is 328.44, 42.65 and 10.33 for the Bias, MAE and MSE with corresponding p-value of 1.51E-45, 2.31E-20 and 6.03E-08. This shows a significant difference in predicted project completion time exist with varying activity duration distribution.

CONCLUSIONS

Information on availability of historical data of observed activity duration was sought from sixty-two construction firms at home and abroad. Historical durations of sixty-five activities across three different projects and firm were obtained. The degrees of fit of activity duration to commonly applied statistical distributions such as Beta, Triangular, Normal, Lognormal, Uniform, Weibull and Exponential were determined using thebootstrapped kolmogorov-Smirnov test at p=0.05. Impact of each distribution on expected project completion time was experimentally investigated.

Based on these, the following conclusions were drawn:

1. Records of observed project activity duration exist but construction firms appear reluctant to share with others as evidence in Table 1.0.
2. Observed project activity duration exhibit different statistical distributions for different activities as shown in Table 1.4. With the sixty-five activities of the three different projects, about 49.231%, 20%, 13.846%, 12.308% , 3.077%and 1.538% of the activities maybe best modeled with lognormal, Weibull, beta, normal, triangular and uniform distribution respectively.
3. There is a significant difference in predicted project duration due to varying activity duration distribution.

REFERENCES

Arnold, T. B., and Emerson, J. W. 2011. Nonparametric Goodness-of-Fit Tests for Discrete Null Distributions, The R Journal, 3(2), 34–39.

Back, W. E., Boles, W. W., and Fry, G. T. (2000). "Defining triangular probability distributions from historical cost data."*Journal of Construction Engineering and Management-ASCE,*, 126(1), 29-37.

Cottrell, W. D. (1999). Simplified Program Evaluation and Review Technique (PERT), Journal of Construction Engineering and Management, 125(1), 16–22.

Davis, R. (2008). Teaching Note —Teaching Project Simulation in Excel Using PERT- Beta Distributions. INFORMS Transactions on Education, 8(3), 139–148.

Feigelson, E. D., Babu, G. J. (2012). Statistical inference. In *Modern Statistical Methods for Astronomy with R Applications* (1st ed., p. 490). UK: Cambridge University Press.

Frey, C., & Rhodes, D. S. (1999). Quantitative Analysis of Variability and Uncertainty in Environmental Data and Models. Volume 1-Theory and Methodology Based Upon Bootstrap Simulation., 1(April), 1–178. Retrieved from http://www4.ncsu.edu/~frey/reports/Frey_Rhodes_99.pdf

Garg,M., Choudhary,S. and Kalla, S.L. (2009). On the Sum of two Triangular Random Variable, International Journal of Optimisation: Theory, Methods and Application, 1(3), 279-290

Herrerías-velasco, J. M., Herrerías-pleguezuelo, R., & Dorp, J. R. Van. (2011). revisiting the PERT mean and variance.pdf. European Journal of Operational Research, 210, 448–451.

Holm, L.B. and Barra, M. (2011). The Consequences of How Subject Matter Expert Estimates Are Interpreted and Modelled, Demostrated By An Emergency Department DES Model Comparing Triangular and Beta Distributions. In Simulation Conference (WSC), Proceedings of the 2011 Winter (pp. 3654–3661).

Jannat, S. and A. G. Greenwood. (2012). Estimating Parameters Of The Triangular Distribution Using Nonstandard. Proceedings of the 2012 Winter Simulation Conference. Edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A.M. Uhrmacher.

Jaskowski, P, Biruk, S and A. Czarnigowska (2011), Estimating Distribution Parameters Of Schedule Activity Duration On The Basis Of Risk Related To Expected Project Conditions, International Journal Of Business And Management Studies, Vol 3, No 1,pp 299-307

Khamooshi, H. and, & Cioffi, D. F. 2013. Uncertainty in Task-Duration and

Cost Estimates: A Fusion of Probabilistic Forecasts and Deterministic Scheduling. Journal of Construction Engineering and Management, 139(May), 488–497.

Kotz, S., & Van Dorp, J. R. (2004). Beyound beta: Other continous families of distributions with bounded support and application World Scientific Singapore.

Malcolm, D., & Roseboom, J. (1959). Application of a technique for research and development program evaluation. Operations Research, 7, 646–669. Retrieved from http://pubsonline.informs.org/doi/abs/10.1287/opre.7.5.646

Mccombs, E. L., Elam, M. E., & Pratt, D. B. (2009). Estimating Task Duration in PERT using the Weibull Probability Distribution. Journal of Modern Applied Statistical Methods, 8(1), 282–288.

Mohan, S., Gopalakrishnan, M., Balasubramanian, H., & Chandrashekar, A. (2007). A lognormal approximation of activity duration in PERT using two time estimates. Journal of the Operational Research Society, 58(6), 827–831. http://doi.org/10.1057/palgrave.jors.2602204

Nwobi, F.N. and Ugomma, C.A (2014) A Comparison of Methods for the Estimation of Weibull Distribution Parameters. Metodološkizvezki, Vol. 11, No. 1, 2014, 65-78

Opaleye, A.A, Oliver E. Charles-Owaba and Bill Bender (2017), Relevance Of Historical-Data Based Activity Scheduling And Risk Mitigation Model, International Journal of Science and Technology Volume 6 No. 3, pp 728-732.

Parkinson distribution with a lognormal core: Theory and validation. European Journal of Operational Research, 216(2), 386–396. http://doi.org/10.1016/j.ejor.2011.07.054

Polyanin, A., and Manzhirov, A. V. (2008). Handbook of Integral Equations. 6000 Broken Sound Parkway NW, Suite 300 Boca Raton, FL: Chapman and Hall/CRC: Taylor & Francis Group.

Premachandra, I. M. (2001). An approximation of the activity duration distribution in PERT. Computers and Operations Research, 28(5), 443–452.

Sireesha, V., & Ravi Shankar, N. (2011). Activity Times in PERT. Journal of Statistics and Mathematics, 2(1), 15–22.

Trietsch, D., Mazmanyan, L., Gevorgyan, L., & Baker, K. R. (2012). Modeling activity times by the Parkinson distribution with a lognormal core: Theory and validation. European Journal of Operational Research, 216(2), 386–396. http://doi.org/10.1016/j.ejor.2011.07.054

Appendix I

Error term across distribution and sample size.

BIAS							
Sample Size	Beta	Triangular	Lognormal	Normal	Weibull	Expo	Uniform
10	1.185203	-2.76137	1.135812	1.693209	0.868035	1.533674	-5.25598
20	1.641129	-2.55883	1.375864	1.336665	1.378212	4.183878	-5.01622
30	1.684522	-2.71684	0.709136	0.744777	1.813098	2.004612	-5.27677
40	1.344575	-2.94301	1.465828	1.295378	2.186339	0.797806	-5.19908
50	1.500535	-2.44266	1.268642	1.119024	1.003167	1.670449	-4.95111
60	1.46221	-2.06293	1.372515	1.707198	1.512463	2.222823	-4.9019
70	2.131059	-2.45666	1.938928	2.159902	2.158735	2.287286	-4.29481
80	1.446137	-3.02262	1.143618	0.996439	1.325384	1.954119	-5.2054
90	1.683731	-2.35935	1.895356	1.459886	1.860119	1.258404	-4.81598
100	1.281682	-2.47236	1.491661	1.186782	1.219947	1.80524	-4.81903
MAE							
Sample Size	Beta	Triangular	Lognormal	Normal	Weibull	Expo	Uniform
10	1.845178	4.043968	2.152383	1.985575	1.969423	5.105881	5.255983
20	2.938174	4.242231	2.534274	3.004244	2.492135	4.234507	5.01622
30	2.58583	3.572953	2.556411	2.351793	2.369194	2.004612	5.276773
40	1.378328	2.943009	1.863996	1.331482	2.186339	3.021183	5.199076
50	1.789329	3.13615	1.656854	2.007518	1.815782	2.095438	4.951109
60	1.869596	2.979518	2.17848	1.903345	2.151548	2.778602	4.901898
70	2.30077	2.839497	2.150667	2.159902	2.212492	2.898129	4.294809
80	1.669766	3.666848	1.59812	1.546824	1.927	2.550345	5.205397
90	1.905262	3.282044	2.113865	1.862834	1.861256	1.878296	4.815981
100	1.789382	3.470248	1.822685	1.801465	1.826157	1.933793	4.819027

MSE							
Sample Size	Beta	Triangular	Lognormal	Normal	Weibull	Expo	Uniform
10	71.32445	213.5686	108.7802	86.87747	99.46184	528.473	336.2412
20	383.3967	560.7126	306.4858	343.5154	260.5551	841.9301	819.1676
30	347.8705	492.0414	358.8884	235.998	313.4739	210.9596	1089.295
40	256.8147	486.3312	229.8757	189.8983	310.6316	686.0947	1305.89
50	389.4909	593.9786	356.3486	450.5724	381.5567	420.9599	1587.05
60	628.0628	645.9099	682.7028	698.3165	691.503	1160.386	2025.506
70	792.3736	802.1603	711.5417	684.2305	870.4368	1069.105	1695.477
80	629.7799	1323.592	542.239	556.1872	701.4386	1247.094	2630.131
90	890.3892	1176.307	897.2969	688.5693	724.7221	638.6381	2626.381
100	829.303	1429.92	956.5594	732.2289	971.4797	692.2379	2834.059

Reference to this paper should be made as follows: Adepeju A. Opaleye & Oliver E. Charles-Owaba (2017) Effect of Different Activity Duration Distribution on Project Duration Prediction Problem. J. of Physical Science and Innovation, Vol. 9, No. 2, Pp. 85-108