The Effect of Input Distribution Skewness on the Output Distribution for Total Project Schedule Simulation

Krige Visser
Department of Engineering and Technology Management, University of Pretoria, South Africa
krige.visser@up.ac.za

Abstract

A Monte Carlo simulation is useful to determine the probability of completing a project within budget and on time at various stages of the project. This paper discusses a research study to determine and compare the distribution for total project duration using ten different probability distributions for fourteen activities of a project. Triangular distributions were used as reference and parameter values for each activity duration were assumed. Parameter values of nine other distributions were calculated from the mean and standard deviations for the triangular distribution. The study indicated that P90 and P95 values of the output distributions, using different input distributions, differed by up to 4,8%. It was also found that there is a positive correlation between the mean skewness of the fourteen activities and the P90 and P95 values. The skewness of the output distribution showed a strong positive correlation with the mean skewness of the input distributions.

Keywords: risk, simulation, duration, projects

1 Introduction

1.1 Background

Monte Carlo simulation is mentioned in the Project Management Body of Knowledge (PMBoK) guide (PMI, 2014) as one of the quantitative risk analysis tools. A detailed discussion of cost and schedule risk analysis and simulation is provided by Cooper et al. (2014). A schedule risk simulation is typically done before project implementation and can be updated with actual activity duration values to replace the input distributions as the project is executed. The probability of finishing the project by the due date can thus be captured at various stages or phases of the project and plotted on a timeline to determine the trend.

Two notable projects in which schedule simulation was used extensively are the Øresund bridge project in Denmark and Sweden (Christensen and Rydberg, 2001), and the Gotthard base tunnel in Switzerland (Ehrbar et al., 2016). In the former the bridge and tunnel was completed 5 months ahead of schedule and in the latter the 57km tunnel underneath the Alps was completed a year earlier than initially planned.

It is unusual for a person to perform the same activity in exactly the same time when repeated, therefore, probability distributions are used to describe this uncertainty in duration. Various distributions have been proposed to model the uncertainty in the duration of project activities, e.g. the betapert, triangular, normal and lognormal distributions (Raydugin, 2013; Munier, 2014). The normal distribution is symmetric, the lognormal distribution is right skewed and the betapert and triangular distributions can be left skewed, symmetric, or right skewed.

If all the activity durations of a project are modelled with the normal distribution, one would expect the total project duration to be symmetric as well. This is the case if all activities are performed in series. If some activities are performed in parallel, the output distribution could be skewed even if all input distributions are symmetric. If all the activities are modelled with a right skewed distribution one would expect the distribution for the total project duration to also be right skewed. The skewness of the input distributions would therefore determine the skewness of the output distribution. Schedule simulation is mostly used to determine the 90% (P90) or 95% (P95) certainty duration of the project and the skewness of the output distribution affects the value of the P90 and P95 values. The choice of input distribution is therefore important to obtain a good approximation of the P90 and P95 values.

1.2 Objectives of Study

The main objective of this study was to determine the effect of the skewness of input distributions of a project network on the skewness of the output distribution provided by a Monte Carlo simulation in Excel. A further objective was to investigate the differences in the skewness of the output distributions for ten different input distributions that have different skewness values for fourteen project activities. A project network with 14 activities in series and parallel was selected as a case study for this research project.

2 Literature

2.1 Schedule Risk

Several high-profile projects have suffered delays and eventual slippage. A notable example of severe slippage is the construction of the Sydney Opera House that took 10 years longer to finish (Steyn et al., 2016). Another classical example is the Central Artery/Tunnel project in Boston, USA. The project duration was planned for about 10 years but eventually took 17 years to complete (National Research Council, 2003).

Managing the schedule for a project is of great importance, especially for mega projects and very long projects. Slippage in a large or long transport project will inevitably lead to cost overruns and loss of revenue since toll fees cannot be collected (Love et al. 2014).

Vanhoucke (2015) mentioned the importance of schedule risk analysis in projects and said a sound baseline schedule is critical for successful execution of the subsequent phases of the project. Nicholas and Steyn (2017) said "project scheduling is an integral part of project planning".

Uncertainty in the total duration of a project is influenced by two factors. The first is uncertainty in the duration of individual activities and tasks to be performed. This type of uncertainty is treated by means of suitable probability distributions. The second uncertainty is due to the occurrence of random events that impact the duration of an activity. This type of uncertainty can be incorporated in a simulation by means of a probability of occurrence and estimates for the consequence should the event occur. Discrete distributions like the binomial distribution is useful for incorporating random events in a schedule simulation (Damnjanovic and Reinschmidt, 2020).

2.2 Probability Distributions

Various probability distributions are available to model activity duration in projects as well as operational and maintenance tasks in an enterprise. Only distributions that have two parameters were selected for this study, except the triangular distribution which has three parameters. Some information on four popular probability distributions are given below.

The triangular distribution is a versatile continuous distribution that is appropriate for cost and schedule simulations (Scherer et al., 2003). The parameters are easy to estimate but one drawback is the fact that the maximum duration of an activity is capped. Even if very skewed values are selected for the triangular distribution, the inverse of the cumulative distribution function cannot produce a higher value than the maximum or upper bound value.

The lognormal distribution is quite popular for cost and schedule risk simulation and was used extensively by Bowden et al. (2001). It is a right-skewed distribution that is often used to model the duration of activities that are performed by novice artisans or technicians that perform complex tasks.

The normal distribution is probably the most popular continuous distribution used in risk simulation. It is symmetric and the parameters are quite easy to estimate, especially if similar activities have been performed in the project organization in the past.

The Fréchet distribution, also known as the inverse Weibull distribution, is an interesting distribution that is regarded as an extreme value distribution that is often applied to modeling extreme events. It is a "fat-tailed" distribution that can be useful to describe the uncertainty in the duration of new activities that have not been done before in a project organization.

Data on most of the probability distributions used in this study, i.e. formulae for the mean value, standard deviation, skewness, excess kurtosis, and the inverse variant, are provided by Evans et al. (2000).

2.3 Activity Duration

Many project network modelers use symmetric distributions to express uncertainty in activity duration, e.g. the normal distribution or logistic distribution. However, the distribution of activity durations in projects are often not symmetric but rather skewed to the right (positive skewness). According to Martens (2017), most activity distributions in projects are right skewed. This view is shared by Shankar (2011).

With many different distributions available to describe the uncertainty of activity duration, the question could be posed whether certain distributions are more suited for certain project situations. Various factors could influence this decision, e.g. ease of use, availability of inverse variants within add-in software, ease of estimating the parameters of the distribution, symmetry or not, positive inverse variants, and availability of an explicit formula to calculate the inverse variant if not available in add-in software.

Ferson et al. (1998) said "the results of probabilistic risk analyses are known to be sensitive to the choice of distributions used as inputs, an effect which is undoubtedly even stronger for the tail probabilities". The authors argue that, in the absence of information regarding the uncertainty of an input value, the uniform distribution with minimum and maximum bounds is the best approach.

Hajdu and Bokoro (2014) compared the results of simulations using the uniform, triangular and beta distributions as input. They found that the difference in the project duration for the three distributions was less than the effect on the duration due to a 10% variation in the values of the three-point estimates for the triangular distribution.

Sherer et al. (2003) approximated the normal distribution with a symmetric triangular distribution and approximated the lognormal distribution with a non-symmetric triangular distribution. The authors found

that the symmetric triangular distribution provides a good approximation for the normal distribution in the range of the mean ± 2,44σ.

Wood (2002) compared the output cumulative distributions of a 12-activity project with two paths using the triangular, uniform, normal and lognormal distributions as input. He found that P90 values using different input distributions varied by as much as 10%.

Visser (2016) performed simulations for two project networks and compared the results for the triangular, normal, lognormal and betapert distributions. The P80 and P90 values of the output distribution with 20000 trials were compared. The study found no significant difference in the output for these four input distributions.

3 Methodology

A theoretical project network with 14 activities in series and parallel was chosen to study the effect of skewness of input distributions on the skewness of the total project duration. Project managers seem to agree that project activity durations mostly have right skewed distributions (Damnjanovic and Reinschmidt, 2020). The parameters of the triangular distributions for all activities were therefore chosen to provide positive skewness, i.e. right skewed distributions.

The mean and standard deviation values for the input distributions were used to determine the parameters of nine other distributions. Only distributions with two parameters were considered for this study. The betapert distribution that is popular in schedule simulation was excluded.

A model that considers the logic of the project network was developed in Excel. The schedule simulation was performed with 100000 trials using the SimVoi add-in for Excel (Treeplan, 2020). One output of the simulation was a chart of the cumulative distribution for the total project duration. The duration values at increments of 5% probability were also provided by the add-in. The project network that was used for this study is shown in Figure 1.

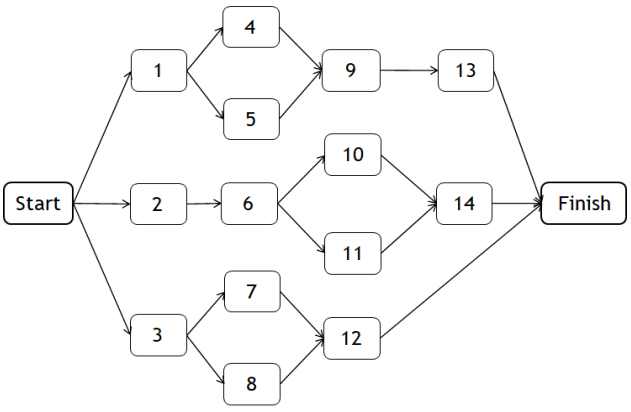

Figure 1. Project network used for simulation study

In this network there are six 'paths' that need to be executed in parallel to finish the total project. Values for the three parameters of the triangular distribution, i.e. *a* (lower bound), *m* (most likely) and *b* (upper bound), were chosen such that any of the six paths could be the critical path for the project. The mean value for each path of the network is shown in Table 1.

Table 1. Mean values of six paths in project network

Path	Lower bound	Most likely	Upper bound	Mean
1-4-9-13	27	35	62	41,33
1-5-9-13	25	32	57	38,00
2-6-10-14	26	33	58	39,00
2-6-11-14	28	35	59	40,67
3-7-12	28	35	54	39,00
3-8-12	29	35	53	39,00

The three parameter values for the triangular distribution, mean value (μ), standard deviation (σ) and the skewness (Skew) are shown in Table 2. The units for these parameters would typically be days or weeks if it is a long project, but it could represent any other time measurement.

Table 2. Input values for 14 activities

Act.	a	m	b	μ	σ	Skew
1	4	6	12	7,333	1,699	0,422
2	5	7	14	8,666	1,929	0,454
3	9	11	16	12,000	1,472	0,376
4	7	9	16	10,666	1,929	0,454
5	5	6	11	7,333	1,312	0,505
6	7	8	14	9,666	1,546	0,522
7	9	12	18	13,000	1,871	0,305
8	10	12	17	13,000	1,472	0,376
9	9	12	18	13,000	1,871	0,305
10	6	8	14	9,333	1,700	0,422
11	8	10	15	11,000	1,472	0,376
12	10	12	20	14,000	2,160	0,476
13	7	8	16	10,333	2,014	0,540
14	8	10	16	11,333	1,700	0,422
Mean Values				**10,762**	**1,725**	**0,426**

The mean and standard deviation values for each of the activities were used to calculate the parameters of the other probability distributions. Formulae for the mean and standard deviation of these nine distributions were mostly obtained from Evans et al. (2000) and the NIST e-Handbook of Statistical Methods (2012).

4 Results

4.1 Descriptive Statistics

Using the SimVoi add-in for Excel, the simulation provided the following values in Table 3 for the project duration distribution. The add-in provides inverse functions for most of the distributions and formulas were used for those not available. The truncated inverse function for the normal distribution was used to prevent negative duration values.

Table 3. Descriptive statistics for output distribution

Distribution	μ	σ	Skew	P90	P95
Fisk	43,719	3,03	+0,90	47,61	49,17
Frechet	43,771	3,49	+1,30	48,22	50,17
Gamma	43,723	2,76	+0,45	47,35	48,58
Gumbel	43,774	3,11	+0,74	47,87	49,43
Logistic	43,671	2,68	+0,49	47,15	48,38
Lognormal	43,742	2,84	+0,51	47,46	48,80
Normal	43,694	2,62	+0,33	47,13	48,22
Pareto	43,737	3,88	+1,77	48,59	50,89
Triangular	43,760	2,77	+0,32	47,41	48,56
Weibull	43,625	2,44	+0,20	46,80	47,80

The probability distributions in Table 3 are arranged alphabetically. The add-in provides the values for the mean value of the distribution (μ), the standard deviation (σ), skewness (Skew), P90, and P95. Other percentiles in increments of 5% are also provided. All the distributions produced a positive skewness for the output distribution. Even though all activities had a negative skewness for the Weibull distribution, the output distribution had a positive skewness.

4.2 Skewness of Total Project Duration

The skewness of the total project duration is illustrated graphically in Figure 2.

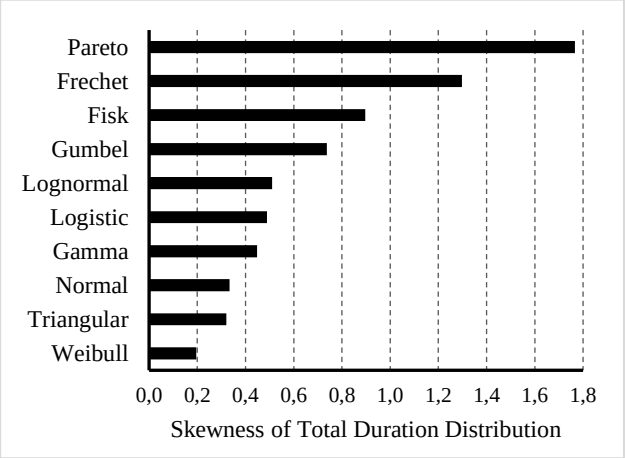

Figure 2. P90 and P95 values for ten input distributions

4.3 P90 and P95 Values

In some situations, a project manager might want to estimate the total duration with a 90% or 95% certainty (often when the project is nearing completion). The P90 and P95 values for the project duration using different input distributions are shown in Figure 3.

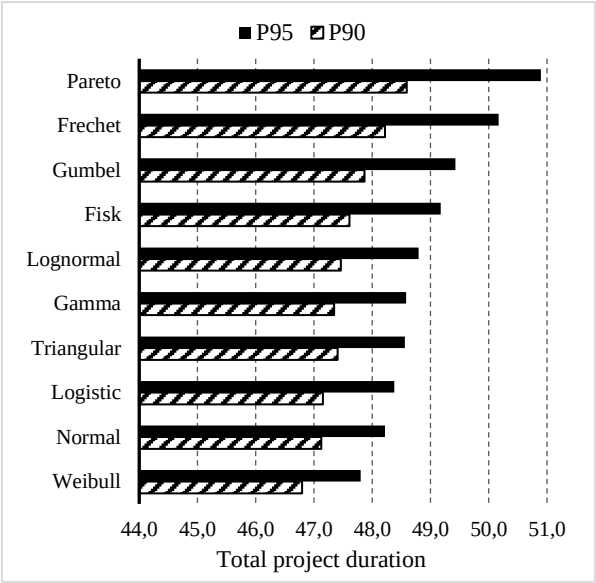

Figure 3. P90 and P95 values for 10 input distributions

The P95 values differed from 47,80 for the Weibull to 50,89 for the Pareto distribution. The value for the Weibull is 1,57% less than the value obtained for the triangular distribution while the value for the Pareto is 4,80% higher than that of the triangular. The Pareto is somewhat unpractical to describe the uncertainty in duration but the Fréchet distribution is well-known as a 'fat-tailed' distribution and is used in many engineering and scientific applications. The P95 value for the Fréchet distribution is 3,3% higher than the value of the triangular distribution. The P90 values for symmetric and slightly skewed distributions, i.e. the lognormal, gamma, logistic and normal distributions, did not differ significantly.

4.4 Effect of Skewness on P90 and P95

The effect of skewness of the input distributions of the activities was investigated. The mean skewness of all the activities was determined for all the distributions investigated and the relationship with the P90 and P95 values is shown in Figure 4.

It is evident from Figure 4 that there is a strong positive correlation between the skewness of the input distributions and the project duration as indicated by the P90 and P95 values. A correlation value of 0,94 was obtained for the relationship between the mean skewness of the input distributions and P90. A correlation value of 0,99 was obtained for the relationship between the mean skewness of the input distributions and P95. A least squares curve fit could provide an empirical formula for this relationship. This

will enable one to predict what the P90 and P95 values for the project duration will be for some other distribution if the mean skewness for that distribution can be calculated.

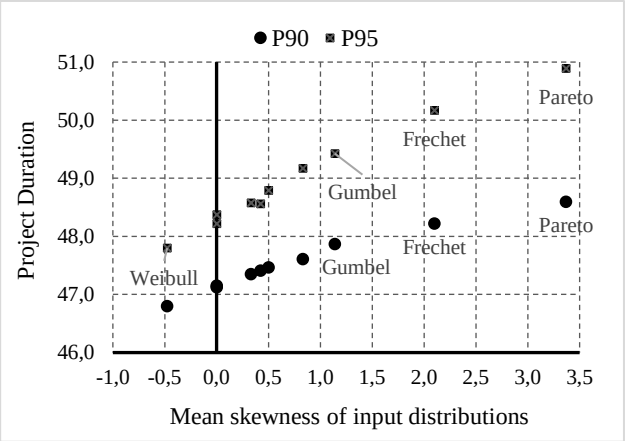

Figure 4. P90 and P95 values of project duration vs. the mean skewness of input distributions

4.5 Effect of Input Distribution Skewness

The relationship between the skewness of the output distribution for the project duration and the mean skewness of the 14 activities was also determined and the result is shown in Figure 5.

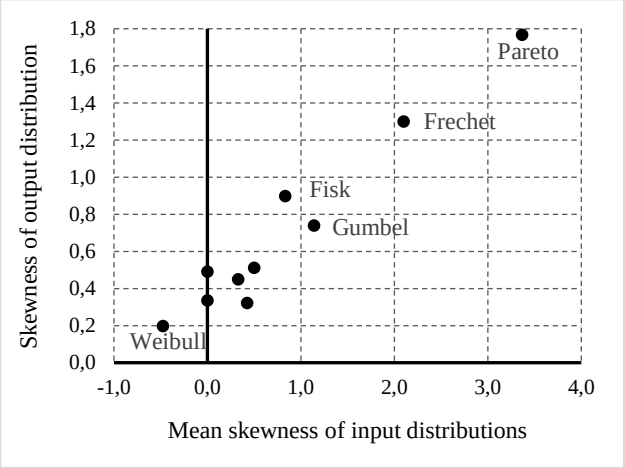

Figure 5. Relationship between mean skewness of input distributions and output distribution skewness

From Figure 5, it is seen that a strong positive correlation exists between the output distribution skewness and the mean skewness of the input distribution for the 14 activities in the network. A correlation value of 0,975 was found for this relationship. Within the range of -0,5 - +3,5 skewness this relationship is close to linear. A straight-line, least squares fit had a regression coefficient of 0,95.

5 Conclusion

The output of a Monte Carlo simulation of a network of activities provides a distribution of the total project duration. The skewness of the output distribution is positively correlated with the mean skewness of the input distributions. Even if the mean skewness of the input distributions is zero, e.g. for the normal and logistic distribution, the output distribution could be skew. In this study, the skewness of the output distribution is caused by the multiple paths of the project network.

The sum of the mean values of each path is very similar. The data in Table 3 indicates that any of the six paths are possible for one trial of the simulation. This means the critical path varies between the six paths and this causes the skewness of the output distribution, even if the input distributions are all symmetric. It is interesting to note that even for the Weibull distribution, with all activities having a negative skewness, the output distribution has a positive skewness.

The P90 and P95 values of the output distribution are positively correlated with the mean skewness of the input distributions. The correlation values were 0,94 and 0,99 respectively for P90 and P95.

The skewness of the output distribution is strongly correlated with the mean skewness of the activity distributions. The correlation value for this relationship was 0,975 in the range of -0,5 - +3,5. A straight line fit of the data had a regression coefficient of 95%.

The results of this study indicate that the choice of input distribution for cost or schedule simulation should be carefully considered by the risk panel. Total project duration could be underestimated if symmetric or near-symmetric input distributions are selected.

6 Recommendations

In practical applications of Monte Carlo simulation, e.g. in projects, operations or business simulation, a risk panel would decide on which distribution should be used to model the uncertainty in the duration of the activities. Different distributions could be selected for the activities. The next step is to allocate values for the parameters of the distribution that was chosen. The parameters of the triangular or normal distributions are fairly easy to estimate since the scale or location parameters point directly to actual duration. Shankar (2011) commented “the beta distribution can be estimated relatively easily from data on just the optimistic, pessimistic and most likely values”. The same applies to the triangular distribution although the mean and variance values differ from the betapert distribution. A stepwise procedure to elicit or estimate values for the triangular distribution is provided by Greenberg (2017).

If actual data for the duration of similar activities in previous projects performed by a company is available, the approach should be to perform a maximum likelihood fit for the data for several distributions and to use the one with the best fit. For novel projects, this data

is seldom available and expert opinion is the only option.

In projects where there is a high degree of certainty in activity duration, e.g. in outage projects for units of a power station, the normal and triangular distributions are recommended for schedule simulations. However, if there is much uncertainty in the duration of some activities in a project, it might be better to use the Fréchet, Fisk (log-logistic) or Gumbel distribution. However, it is not easy to estimate the parameters of these distributions. These distributions have two parameters, typically a shape and scale parameter. The scale parameter relates to the duration of an activity, but the shape parameter is difficult to estimate.

It is therefore recommended that the risk panel start by estimating the parameters of the triangular distribution for all activities, incorporating skewness in the choice of the lower and upper bound values. The mean and standard deviation values of the triangular distributions can then be used to calculate the parameters of the Fréchet, Fisk or Gumbel distribution. These distributions should then be used to run the schedule simulations for a project.

Acknowledgements

The Department of Engineering and Technology Management of the University of Pretoria provided support for this study.

References

A. R. Bowden, M. R. Lane, and J. H. Martin. *Triple Bottom Line Risk Management*. John Wiley & Sons Ltd., New York, USA, 2001.

P. J. Christensen and J. Rydberg. overcoming obstacles: strategic risk management in the Øresund bridge project, *PM Network*, 15 (11): 30-36, 2001.

D. Cooper, P. Bosnich, S. Grey, G. Purdy, G. Raymond, P. Walker, and M. Wood. *Project Risk Management Guidelines. Managing Risk with ISO 31000 and IEC 62198*. 2nd Edition, John Wiley & Sons Ltd., UK, 2014.

C. Croarkin and P. Tobias. *e-Handbook of Statistical Methods.* National Institute of Standards and Technology (NIST/SEMATECH), USA, 2012. doi: 10.18434/M32189.

I. Damnjanovic and K. Reinschmidt. *Data Analytics for Engineering and Construction Project Risk Management*. Springer, Switzerland, 2020. doi:10.1007/978-3-030-14251-3.

H. Ehrbar, L. R. Gruber, and A. Sala. *Tunnelling the Gotthard - Gotthard Base Tunnel*. Swiss Tunneling Society, Bauverlag Gütersloh, Switzerland, 2016.

M. Evans, N. Hastings, and B. Peacock. *Statistical Distributions.* 3rd Edition, John Wiley & Sons, New York, 2000.

S. Ferson, L. R. Ginzburg, and H. R. Akçakaya. Whereof one cannot speak: When input distributions are unknown. http://www.ramas.com/whereof.pdf. 1998. (Retrieved 06/07/2020).

M. W. Greenberg. A Step-wise Approach to Elicit Triangular Distribution. *NASA Technical Reports Server*, Bibliogov. 2013.

M. Hajdu and O. Bokor. The effects of different activity distributions on project duration in pert networks, *Procedia - Social and Behavioral Sciences*, 119:766-775, 2014. doi:10.1016/j.sbspro.2014.03.086.

P. E. D. Love, C-P. Sing, X. Wang, Z. Irani, and D. W. Thwala. Overruns in transportation infrastructure projects. *Structure and Infrastructure Engineering*, 10 (2):141-159, 2014. doi:10.1080/15732479.2012.715173.

L. Martens. Schedule Risk Analysis: Case Studies, *Masters Dissertation.* Universiteit Gent, 2017.

N. Munier. *Risk Management for Engineering Projects: Procedures, Methods and Tools*. Springer International Publishing, Switzerland, 2014. doi:10.1007/978-3-319-05251-9.

National Research Council (NRC). *Completing the "Big Dig": Managing the Final Stages of Boston's Central Artery/Tunnel Project*. The National Academies Press. Washington, 2003.

J. M. Nicholas and H. Steyn. *Project Management for Engineering, Business and Technology*. 5th Edition. Routledge, New York, 2017. doi:10.4324/9781315676319.

Project Management Institute (PMI). *A Guide to the Project Management Body of Knowledge (PMBOK Guide)*. 5th Edition, 2013. doi:10.1002/pmj.21345.

Y. Raydugin. *Project Risk Management: Essential Methods for Project Teams and Decision Makers*. John Wiley & Sons Inc., New Jersey, 2013.

W. T. Scherer, T. A. Pomroy, and D. N. Fuller. The triangular density to approximate the normal density. *Reliability Engineering and System Safety*, 82 (3):331–341, 2003. doi:10.1016/j.ress.2003.08.003.

N. R. Shankar, S. S. Babu, Y.I.P. Thorani, and D. Raghuram. Right skewed distribution of activity times in PERT. *International Journal of Engineering Science and Technology*, 3 (4): 2932-2938, 2011.

H. Steyn, M. Carruthers, A. Dekker, Y. Du Plessis, D. Kruger, B. Kuschke, A. Sparrius, S. van Eck, and K. Visser. *Project Management: A Multi-disciplinary Approach*. 4th Edition. FPM Publishing, Pretoria, South Africa, 2016.

Treeplan Software. SimVoi Monte Carlo Simulation Add-in for Excel. https://treeplan.com/simvoi/. (Retrieved 07/06/2020).

M. Vanhoucke. On the use of schedule risk analysis for project management, *The Journal of Modern Project Management*, 2(3):108-117, 2015.

J. K. Visser, Suitability of different probability distributions for performing schedule risk simulations in project management. In Proceedings - *Portland International Conference on Management of Engineering and Technology, PICMET 2016, 4-8 September 2016, Honolulu*, Hawaii, USA, pages 2031-2039, 2016. doi:10.1109/PICMET.2016.7806608.

D. A. Wood, Risk simulation techniques to aid project cost-time planning and management. *Risk Management*, 4(1):41-60, 2002. doi:10.1057/palgrave.rm.8240108.