

A Mathematical Approximation of the Left-sided Truncated Normal Distribution Using the Cadwell Approximation Model

Mohammad M. Hamasha

Department of Engineering Management
Prince Sultan University
Riyadh, Saudi Arabia
mhamash1@binghamton.edu

Abstract—In the case that life distribution of new devices follows the normal distribution, the life distribution of the same brand used devices follows left-sided truncated normal distribution. In spite of many mathematical models being available to approximate the normal distribution density functions, there is a few work available on modeling/approximating the density functions of left-sided truncated normal distribution. This article introduces a high accuracy mathematical model to approximate the cumulative density function of left-sided truncated standard normal distribution defined on the range of [truncation point (Z_L): ∞]. The introduced model is derived from the Cadwell approximation of the normal cumulative density. The accuracy level change with Z score is discussed in details. The maximum deviation of the model results, from the real results for the whole region of [$-\infty < Z_L < -2$: ∞], is 0.006877.

Keywords—truncated normal distribution; normal distribution; mathematical approximation

I. INTRODUCTION

If is a continuous random variable X having a normal distribution with mean of μ and standard deviation of σ , the normal curve can be graphed using (1)

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

The normal distribution can take any mean and any positive standard deviation. These two parameters determine the shape of the normal curve. The mean provides the location of middle of distribution, and the standard deviation defines how much the data are spread from the mean.

Normal distribution is used to model many sets of measurements in industry, business and nature. For instance, the human blood sugar level, the lifetime of radio sets, and even automobile costs are all normally distributed. A normal distribution has many unique properties including that 1) the value for mean, median, and mode are equal, 2) the distribution curve is bell shaped and symmetric about the mean, 3) the total area under the normal curve equals to 1 (all probability

distributions have this property), 4) the normal curve approaches to zero with X going to $-\infty$ or ∞ , and finally, 5) around the center of the curve, the graph curves downward (concave). Further, the graph curves upward at the right and the left sides. The inflection points are the points at which the curve changes from curving downward to upward. The normal distribution with σ equals to 1 and μ equals to 0 is called the standard normal distribution. The standard normal distribution function is shown in (2).

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \quad (2)$$

The standard normal distribution horizontal scale is well known as Z score. Any not standard normal distribution can be transferred into Z score with using of the following transformation formula: $z = (x - \mu) / \sigma$.

There are many situation where the interest is to compute the probability that the observed value X is less than or equal to some real number x. Therefore, the cumulative distribution function is defined as $F(x) = P(X \leq x)$. For every real number x, $F(x)$ is defined, as shown in (3).

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t) dt \quad (3)$$

The cumulative distribution function for the normal distribution is denoted by $\Phi(z)$ as addressed in the following equation:

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{t^2}{2}} dt \quad (4)$$

The solution of the above mentioned equation is complex and cannot be solved manually. Therefore, the statisticians prepared table to estimate the cumulative distribution function values for the standard distribution. This table is called a Z-table.

The normal distribution can be truncated in various ways as needed. For example, if the distribution of the GRE scores is

normally distributed and a graduate school has a minimum acceptable score, then the distribution of the admitted student scores is basically left-sided truncated normal distribution. This is because the left tail of the normal distribution (e.g., scores lower than minimum acceptable GRE scores) is truncated. Another example, if the life in years, of a television set is normally distributed, then the distribution of a two years used television set from the same brand is left-sided truncated normal distribution. The tail of distribution at $P(X < 2)$ is truncated. The following figure (Figure 1) is an example on the left-sided truncated normal distribution.

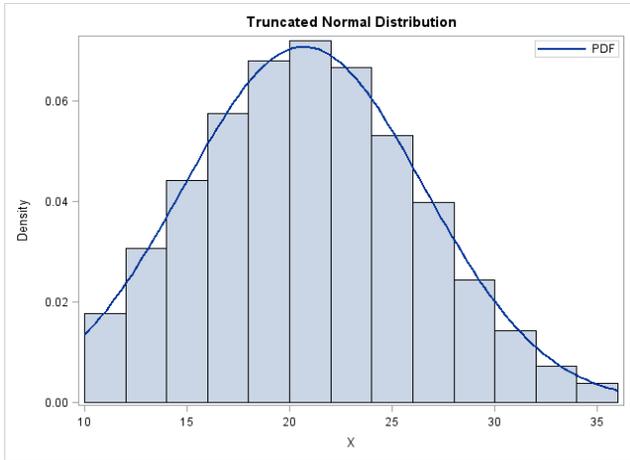


Fig. 1. Left-sided truncated normal distribution (e.g., $\mu=21.1$, $\sigma=5.3$, truncation point (X_L)=10).

The left-sided truncated standard normal distribution is defined on the region $(Z_L: \infty)$. Since the final area under the curve of truncated distribution must equal to 1, the new curve is stretched up to compensate the lost truncated area over the region $(-\infty: Z_L)$. The probability density function of the left-sided truncated normal distribution is addressed in (5)

$$\phi_T(z) = \frac{\phi(z)}{\int_{z_L}^{\infty} \phi(t) dt}, \quad z_L \leq z < \infty \quad (5)$$

The cumulative distribution function for the left-sided truncated normal distribution can be defined as the cumulative density function divided by the area under the original distribution function over the region $(z_L: \infty)$. Equation (6) represents the cumulative distribution function of left-sided truncated standard normal distribution.

$$\Phi_T(z) = \frac{\int_{z_L}^z \phi(t) dt}{\int_{z_L}^{\infty} \phi(t) dt}, \quad z_L \leq z < \infty \quad (6)$$

Equation (6) is a complex integration function, and need a numerical solution to be solved. Therefore, engineers usually use sophisticated computer programs or specialized software package to handle this kind of equations. However, in the case that $\phi(z)$ is approximated with function that is able to be

differentiated and integrated, then the truncated normal distribution can be estimated in a function.

II. REVIEW OF PRIOR RESEARCH

Much research on mathematical approximation of normal distribution functions is available in the literature. At least 15 highly accurate models were introduced. Cadwell (1949) [1-3] introduced a very nice approximating model, as shown in (7)

$$\Phi(z) = \begin{cases} \frac{1}{2} \left[1 + \left\{ 1 - \exp\left(\frac{-2z^2}{\pi} - \frac{2(\pi-3)z^4}{3\pi^2}\right) \right\}^{\frac{1}{2}} \right], & z \geq 0 \\ 1 - \frac{1}{2} \left[1 + \left\{ 1 - \exp\left(\frac{-2z^2}{\pi} - \frac{2(\pi-3)z^4}{3\pi^2}\right) \right\}^{\frac{1}{2}} \right], & z < 0 \end{cases} \quad (7)$$

The following 6 models are mentioned in [1, 2]: Polya (1945), 2) Hart (1963), 3) Zelen and Severo (1966), 4) Schucany and Gray (1968), 5) Hoyt (1968), 6) Hamaker (1978), 7) Lin (1988), 8) Lin (1989), and 9) Lin (1990). Further, there are few recent models such as Vazquez-Leal (2012) [4], Choudhury (2014) [5] and Luis (2015) [6].

Many authors discussed in details many aspect about the truncated normal distribution. For example, Ke et al. [7] used double-sided truncated normal distribution on action reliability for ammunition swing device of large caliber gun. Pender [8] provided an exact expression for the moments of the truncated normal distribution using Stein's lemma. His moment expressions provide insight into the steady state skewness and kurtosis dynamics of single server queues with impatient customers. Sun et al. [9] worked on finite fault modelling for the Wenchuan earthquake using hybrid slip model with truncated normal distributed source parameters. Beucher and Renard [10] described a stochastic model called truncated Gaussian simulations (TGS), which distributes a collection of facies or lithotypes over an area of interest. This method is based on facies proportions, spatial distribution and relationships, which can be easily tuned to produce numerous different textures. Mukerjee and Ong [11] worked on variance and covariance inequalities for truncated joint normal distribution via monotone likelihood ratio and log-concavity. Horrace [12] developed probability statements and ranking and selection rules for independent truncated normal populations. Sharples and Pezzey [13] presented results pertaining to truncated multivariate normal distributions, some of which already appear in the mathematical literature. They focused to make these types of results more accessible to the environmental science community and to this end, they included a conceptually simple alternative derivation of an important result. Furthermore, they illustrated how the theory of truncated multivariate normal distributions is employed in the environmental sciences.

We can conclude from the above discussion that there is a lot on work on mathematical aspects and applications of the normal cumulative distribution function. Even so, mathematical approximations on the truncated normal distribution are not that frequent. In this paper, a mathematical approximation model of left-sided truncated normal cumulative

distribution function is proposed. The model is derived from Cadwell's model of approximating the normal cumulative distribution function.

III. PROPOSED MODEL

Since (7) is separated into two parts according to z value, the model will be divided on two parts. We are going to formulate a model for the left part (Z<0) first. The left part is recalled, as extracted and simplified in (8).

$$\Phi(z) = \frac{1}{2} - \frac{1}{2} \left\{ 1 - \exp\left(\frac{-2z^2}{\pi} - \frac{2(\pi-3)z^4}{3\pi^2}\right) \right\}^{\frac{1}{2}}, z < 0 \quad (8)$$

The left-sided truncated standard normal distribution can be obtained through integration as explained in (6). Remember that $\phi(z) = d\Phi(z)/dz$. The following equation can be used to obtain the first part of our model, as addressed in (9).

$$\Phi_T(z) = \int_{z_L}^z \frac{d \left(1 - \frac{1}{2} \left[1 + \left\{ 1 - \exp\left(\frac{-2t^2}{\pi} - \frac{2(\pi-3)t^4}{3\pi^2}\right) \right\}^{\frac{1}{2}} \right] \right)}{0.5 + \int_{z_L}^0 \frac{d \left(1 - \frac{1}{2} \left[1 + \left\{ 1 - \exp\left(\frac{-2t^2}{\pi} - \frac{2(\pi-3)t^4}{3\pi^2}\right) \right\}^{\frac{1}{2}} \right)}{dt}} dt, z < 0 \quad (9)$$

The solution of the equation will give the following formula (written in the simplest form).

$$\Phi_T(z) = 1 - \frac{\left[1 + \left\{ 1 - \exp\left(\frac{-2z^2}{\pi} - \frac{2(\pi-3)z^4}{3\pi^2}\right) \right\}^{\frac{1}{2}} \right]}{\left[1 + \left\{ 1 - \exp\left(\frac{-2z_L^2}{\pi} - \frac{2(\pi-3)z_L^4}{3\pi^2}\right) \right\}^{\frac{1}{2}} \right]} z < 0 \quad (10)$$

If Z>0, then approximating $\Phi_T(z)$ needs to extract the positive z part of (7) and substitute it in (6), exactly as we did in (9). The result is addressed in (11).

$$\Phi_T(z) = 1 + \frac{\left[\left\{ 1 - \exp\left(\frac{-2z^2}{\pi} - \frac{2(\pi-3)z^4}{3\pi^2}\right) \right\}^{\frac{1}{2}} - 1 \right]}{\left[1 + \left\{ 1 - \exp\left(\frac{-2z_L^2}{\pi} - \frac{2(\pi-3)z_L^4}{3\pi^2}\right) \right\}^{\frac{1}{2}} \right]}, z \geq 0 \quad (11)$$

IV. NUMERICAL EXAMPLES

Example 1: If the life of a television set is normally distributed with a mean of 22 years and standard deviation of 7. The interest of this question is to find the 6 year reliability of 6 years used television from the same brand. In other words, what is the chance that the 6 years old television will be survived for another 6 years?

Solution: Reliability is defined as the probability that a device, part or system will perform their function for a given

period of time when operated under stated conditions [14]. The requirement can be written in statistics notation as $P(X > 6+6)$. Since this distribution is not standard, the transformation formula, $z=(x-\mu)/\sigma$ will be used to find z_L and z that are corresponding to $x=6$ and $x=12$, respectively.

$$z_L = \frac{6-22}{7} = -2.286$$

$$z = \frac{12-22}{7} = -1.249$$

The probability $P(X>12)$ is corresponding to $P(Z>-1.249)$ or $1-\Phi_T(-1.249)$. By using (11), the result is 0.905713. The model result is very close to the actual result which is 0.904229. The deviation of the model result from the actual result (i.e., error) is only 0.00148

Example 2: Let's assume that the distribution of scores of GRE analytical part test out of 800 is normally distributed with a mean of 540 points and standard deviation of 40 points, and let's assume that the acceptable score for some graduate school is 460. If an admitted student is selected randomly, what is the probability that his score is below 620?

Solution: Since the value of probability density at any X greater than 800 (more than is 6.5 sigma) is very close to zero, we will not assume that the distribution is truncated from the right side at 800. The distribution of the admitted students is left-sided truncated normal distribution at the acceptable level (i.e., 460). $z_L = -2$ is corresponding to $x_L = 460$ by using the transformation formula (i.e., $z=(x-\mu)/\sigma$). Further, $z=2$ is corresponding to $x=620$. In statistic notation, the requirement is $P(X<620)$ which is corresponding to $P(Z<2)$ or $\Phi_T(-1.249)$. By using (11), the result is 0.9826. The model result is very close to the actual result for this example. The actual result is 0.97672, and deviation of the model result from the actual result is only 0.00588.

It is noticeable from both examples, that the error level is very low. Indeed, this level of error is very ignorable for most applications of reliability engineering as well as other probability fields.

V. ACCURACY ANALYSIS

In this section, the accuracy of the model is presented in term of approximating the left-sided truncated normal distribution. Mainly, the change of deviation of the model results from the real results (i.e., error) with Z and Z_L is focused. To make sure that the introduced model is good at all value Z and Z_L , we are concerning about the maximum error over the whole defining range, $(Z_L:\infty)$ at any $Z_L \in (-\infty:-1.5)$.

Figure 2 presents the error versus Z at $Z_L = -4, -3.5, -3, -2.5, -2,$ and -1.5 . We can clearly note that there are two maximum peaks (i.e., maximum deviation or error) for every curve. The first peak is noticed at somewhere between $Z=-1.6$ and $Z=-1.7$, and the second peak is noticed at somewhere between $Z=1.6$ and $Z=1.7$ for all curves. The maximum error for each curve is as follows: for $Z_L = -4$, the maximum absolute error is about 0.006452, for $Z_L = -3.5$, the maximum absolute error is about 0.006456, for $Z_L = -3$, the maximum absolute error is about 0.0065, for $Z_L = -2.5$, the maximum absolute error is about

0.00663, for $Z_L = -2$, the maximum absolute error is about 0.006887, and for $Z_L = -1.5$, the maximum absolute error is about 0.0071251. Usually, the literature does not focus on any truncation point higher than $Z_L = -2$ as it is rarely needed in most applications. In this study, we took $Z_L = -1.5$ as the final truncation point.

In Figure 3, the error versus Z is constructed for the original approximation of standard normal distribution that our model is based on (i.e., Cadwell approximation). We can see that the maximum error is 0.006466 at $Z = -1.655$ and at $Z = 1.655$. The accuracy of the Cadwell model is close to the accuracy of the truncated Cadwell model at any considered Z_L . The current model is more accurate than the logistic function-based truncated normal distribution which introduced in [1]. The current model leads to a maximum absolute error of about 0.0071, while the maximum absolute error in the logistic function-based model is close to 0.02 at the range of $-\infty < Z_L < -1.5$. Furthermore, the current model is easier to implement

using simple calculator. Figure 4 refers to a comparison graph of the model and real results of $Z_L = -2$ left-sided truncated normal cumulative distribution function. We can notice that they are very close to each other. We can't distinguish the difference between the two curves with the normal paper scale.

VI. CONCLUSIONS

In this article, an approximation to the left-sided truncated cumulative distribution function is introduced. The model can be used by reliability and quality engineers in order to avoid the sophisticated computer programs or software when having such distributions in their work. The model was build based on Cadwell approximation to the normal distribution. This study is unique because it provided model with a higher accuracy than any similar study in the literature. It provided a maximum error of 0.006877 for $Z \in (-2; \infty)$ while logistic function-based approximation [1], for example, provides a maximum error 0.017 on the same range of Z .

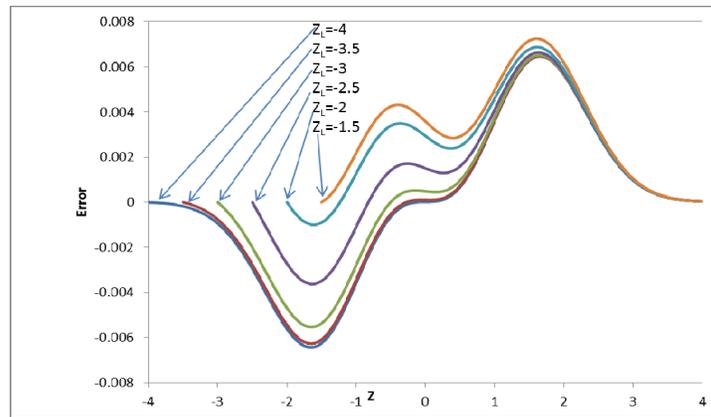


Fig. 2. Error (i.e., different between model and real results) versus Z at $Z_L = -4, -3.5, -3, -2.5, -2,$ and -1.5 .

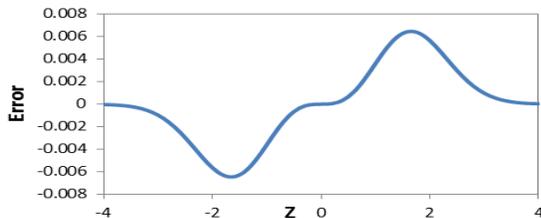


Fig. 3. Error versus Z for Cadwell approximation (i.e., untruncated standard normal distribution).

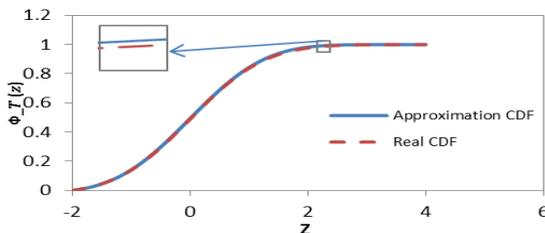


Fig. 4. Truncated normal distribution approximation versus the real truncated normal distribution at $Z_L = -2$.

REFERENCES

- [1] M. M. Hamasha, "Practitioner advice: approximation of the cumulative density of left-sided truncated normal distribution using logistic function and its implementation in Microsoft Excel", Qual. Eng., 2016
- [2] S. R. Bowling, M. T. Khasawneh, S. Kaewkuekool, B. R. Cho, "A logistic approximation to the cumulative normal distribution", J. Ind. Eng. Manage., Vol. 2, No. 1, pp. 114-127, 2009
- [3] J. H. Cadwell, "The bivariate normal integral", Biometrika, Vol. 38, pp. 475-479, 1951
- [4] H. Vazquez-Leal, R. Castaneda-Sheissa, U. Filobello-Nino, A. Sarmiento-Reyes, J. S. Orea, "High accurate simple approximation of normal distribution integral", Math. Prob. Eng., Vol. 2012, Article ID 124029, pp. 1-22, 2016
- [5] A. Choudhury, "A simple approximation to the area under standard normal curve", Math. Stat., Vol. 2, No. 3, pp. 147-149, 2014
- [6] A. J. Luis, "An Approximation to the Probability Normal Distribution and Its Inverse", Ingeniería, Investigación y Tecnología, Vol. 16, No. 4, pp. 605-611, 2015
- [7] X. Ke, J. Hou, T. Chen, "Notice of retraction research on action reliability for ammunition swing device of large caliber gun based on doubly truncated normal distribution", International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering (QR2MSE), July 15-18, 2013 [retracted]

- [8] J. Pender, "The truncated normal distribution: Applications to queues with impatient customers", *Oper. Res. Lett.*, Vol. 43 No. 1, pp. 40-45, 2015
- [9] X. D. Sun, X. X. Tao, C. Q. Liu, "Finite fault modelling for the Wenchuan earthquake using hybrid slip model with truncated normal distributed source parameters", *Appl. Mech. Mater.*, Vol. 256-259, pp. 2161-2167, 2013
- [10] H. Beucher, D. Renard, "Truncated Gaussian and derived methods", In Press, *C. R. Geosci.*, Vol. 348, No. 7, pp. 510-519, 2016
- [11] R. Mukerjee, S. H. Ong, "Variance and covariance inequalities for truncated joint normal distribution via monotone likelihood ratio and log-concavity", *J. Multi. Var. Anal.*, Vol. 139, pp. 1-6, 2015
- [12] W. C. Horrace, "On ranking and selection from independent truncated normal distributions", *J. Econometrics*, Vol. 126, pp. 335-354, 2005
- [13] J. J. Sharples, J. C. V. Pezzey, "Expectations of linear functions with respect to truncated multinormal distributions—With applications for uncertainty analysis in environmental modelling", *Environ. Modell. Software*, Vol. 22, No. 7, pp. 915-923, 2007
- [14] C. E. Ebeling, *An introduction to reliability and maintainability*, 2nd Ed. Waveland Pr Inc, USA, 2009

AUTHORS PROFILE

Dr. Mohammad M. Hamasha received his PhD in Industrial and Systems Engineering from the State University of New York at Binghamton, New York, USA, in December 2011. He also received his MS in Industrial Engineering and BS in Biosystems Engineering from Jordan University of Science and Technology (JUST) in 2008 and 2005, respectively. He worked for two years from 2011 to 2013 as a postdoctoral associate and team leader with the Center for Autonomous Solar Power (CASP) in Binghamton NY doing a research related to reliability of solar materials and cell, and optimization of smart grid. Up to August 2016, Dr. Hamasha has published about 45 publications. Over the years, he has been involved in a variety of projects that range from experimental-based research to operations research, management, simulation, material science, thin films and photovoltaic research. His current research interests are in the areas of Simulation, operational research, reliability engineering, design and analysis of production systems, Stochastic and Simulation Optimization and Lean Six Sigma.