A SIMPLE APPROXIMATION TO THE LOWER TRUNCATED CUMULATIVE NORMAL DISTRIBUTION BASED ON MILL’s RATIO

Mohammad M. Hamasha¹ and Mohammad Al-Rabayah²

¹ Department of Engineering Management
Prince Sultan University, Riyadh 11586, Saudi Arabia
Email: mhamash1@binghamton.edu
² Department of Communications and Networks Engineering
Prince Sultan University, Riyadh 11586, Saudi Arabia

ABSTRACT

In this paper, two very high accurate models to approximate the lower truncated normal cumulative distribution have been developed. Mill’s ratio with an order of 1 and 2 is used to develop two models to approximate the density of normal cumulative distribution, and then these two models are modified to approximate the density of the lower truncated normal cumulative distribution. The first model (i.e., order of 1) was very simple and very accurate with maximum absolute error of about 0.0015 over the domain \([Z_L : \infty]\). \(Z_L\) is the truncation point and it takes any value from negative infinity to zero (i.e., \(Z_L \in [-\infty:0]\)). The second model (i.e., order of 2) is more advanced and can provide a superior accuracy of a maximum absolute error of less than 0.00004. Particularly, the first model can be used whenever an industrial engineer/practitioner needs to estimate probabilities and statistics associated with the lower truncated normal distribution due to its simplicity and accuracy. Further, we strongly recommend using the first model in the case of manual solutions. Although the second model provides superior accurate results, it is complicated and hard to be used manually.

KEYWORDS

Normal Distribution, Normal Cumulative Distribution, Mill’s Ratio, Truncated Normal Distribution.

1) INTRODUCTION

The normal distribution is perhaps the most important probability distribution used in engineering and science [1]. This is because the normal distribution can accurately describe many phenomena in nature, business and industry [2]. Further, measurement of errors is well approximated by the normal distribution. For example, the normal distribution can be used to model the level of sugar in human’s blood and to estimate the expected lifetime or cost of many products. Also, there are many characteristics of the normal distribution which make it unique compared to other distributions. First, the normal distribution is bell-shaped and symmetric around its center (mean). Second, the equal values of the normal distribution mean, median, and mode. Last, the normal
distributions have a higher density at the center, and the density decreases when approaching toward the tails.

Assuming a normally distributed random variable $X$ with mean $\mu$ and variance $\sigma^2$, then we can express its density function as:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-1/2(x-\mu/\sigma)^2}, -\infty \leq x \leq \infty.$$  \hspace{1cm} (1)

The normal distribution with $\sigma$ equals to 1, and $\mu$ equals to 0, is called the standard normal distribution. The standard normal distribution function is addressed in Equation 2.

$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-t^2/2} dt,$$  \hspace{1cm} (2)

The standard normal distribution is usually expressed in the variable $Z$ (i.e., $Z$-score). To convert non-standard normal distribution expressed in $X$ to standard expressed in $Z$-score, we need to follow the transformation formula, $z = (x-\mu)/\sigma$.

In many situations, one may be interested in calculating the probability of which the required variable is equal to or less than a certain value. For this purpose, the cumulative distribution function (CDF) is defined as

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt,$$  \hspace{1cm} (3)

The cumulative distribution function for the standard normal distribution can be denoted by $\Phi(z)$ and as shown below:

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

Equation (4) cannot be solved manually due to the complexity of its integration. Therefore, we can only estimate the values of the cumulative distribution function by statistics tables, particularly the $Z$-table.

In many situations, scientists may be interested in finding probabilities associated with a distribution with a truncated part of the population. The truncated population can be higher than a certain value, lower than a certain value or truncated from both sides. The distribution after truncating a part of the population is called truncated distribution. The range of the new distribution can be either less than a certain value, say max, greater than a certain value, say min, or between the two values (min, max). The truncated distribution is discussed intensively in the literature. [3] and [4] are examples on truncated distribution from Pakistan Journal of Statistics. Moreover, truncated normal distribution can be used in many applications in various fields of economy, engineering, and statistics. For example, truncated normal distribution can be used to estimate the expected life duration of used electronic devices, as well as, estimate probabilities after removing products which largely deviate from the required mean (i.e., not within the specification limit).

If a distribution is truncated from left, right or both sides, with a truncated range of interest $A$, the truncated distribution can be given as:
Further, the cumulative distribution function (CDF) of the truncated distribution is given in Equation 6.

\[ F_T(z) = \int_{-\infty}^{x} f_T(t)dt, \quad Z_L \leq z \]  

Equation (6) is a complex integration function, and needs a numerical solution to solve it. Furthermore, there are no popular statistical tables to handle it. Therefore, engineers usually use sophisticated computer programs or specialized software packages to handle the truncated normal distribution. However, if the CDF of the original distribution before truncation is approximated with a simple function, the CDF of the truncated distribution may be approximated by a function, depending on the simplicity of the original approximation. If the differentiation result of CDF approximation function of the original function can be integrated on a certain domain, then the approximation of the truncated distribution CDF is possible.

Many papers have been written about the truncated normal distribution, some of them even deeply analyzed many aspects of the distribution. To mention some, Ke et al. [5] have studied the action reliability of ammunition swing of large caliber gun using a double-sided truncated normal distribution. The moments of the truncated normal distribution have been derived by [6]. Those derived expressions have provided a deep insight into skewness and kurtosis dynamics of impatient customers in single server queues. In [7], a stochastic model has been developed to distribute a group of facies over an area of interest based on spatial distribution and proportions of facies to produce various textures. The variance and covariance inequalities of truncated normal distribution have been studied in [8]. Independent truncated normal populations have been studied in [9], by developing probabilities, ranking, and selection rules to describe such populations. The truncated multivariate normal distributions have been studied in [10], where they provided a simplified derivation to serve the field of environmental sciences, and also showed how to such distributions can be related to this field.

2) MILL’S RATIO AND NORMAL APPROXIMATION MODEL

Approximations to standard normal distribution can be classified into two main groups [11]. The first group is called numerical algorithms, which consist of approximation with high precision that requires massive computations. The second group is more popular and is called ad-hoc approximations, which are approximations with acceptable accuracy that depends on short formulas and few carefully selected numeric constants. However, exceptions for the previously mentioned classification can be found. For example, some numerical algorithms are accurate and simple to be calculated, while some ad-hoc approximations can provide high accuracy and a large number of computations.

Most of the proposed approximations have been designed to work accurately within a certain range of values, but the accuracy will deteriorate outside the range, and the relative errors will become unbounded.
For example, in Taylor’s polynomial approximation [12] shown below, the size of the polynomial $N$ can be freely adjusted for each $z$. The formula looks very simple, but in practice, it is difficult to be calculated for large $z$ values, even when evaluated for integer $z$ values using computerized tools like Maple and Mathematica.

$$\Pr(Z > z) \approx \frac{1}{2} - \frac{1}{\sqrt{2\pi}} \sum_{k=0}^{N} (-1)^k \frac{z^{2k+1}}{k!(2k+1)2^k}$$ (7)

A class of sigmoid approximations [13] derived from the Taylor expansion is suitable for the use with programs that can compute Taylor expansions of certain orders, and they perform well with pocket calculator approximations.

$$\Pr(Z > z) = 1/\left(1 + \exp(\sum_{k=0}^{\infty} a_k z^{2k+1})\right)$$ (8)

Strecock’s approximation [14, 15] can be considered the most accurate of the ad-hoc methods, but this is true for only a small range of $z$.

$$\Pr(Z > z) \approx \frac{1}{2} - \frac{1}{\pi} \left(\frac{z}{3\sqrt{2}} + \sum_{n=1}^{12} \frac{1}{n} e^{n^2/9} \sin(nx\sqrt{2}/3)\right)$$ (9)

This approximation is accurate for its small $z$ range, but outside that range, the accuracy decreases rapidly.

Also, Hart [16] presented two approximation models that cover the range $0 \leq z < \infty$, the first one (i.e., Equation 10) is a very simple and provides a relative errors of about 2%, while the other one (i.e., Equation 11) is more complex but provides a higher accuracy 0.055%:

$$\Pr(Z > z) \approx \frac{e^{-z^2/2}}{\sqrt{2\pi} z + 0.8e^{-0.4z}}$$ (10)

$$\Pr(Z > z) \approx \frac{e^{-z^2/2}}{z\sqrt{2\pi}} \left(1 - \frac{\sqrt{1 + b z^2/(1 + a z^2)}}{P_0 z + \sqrt{P_0^2 + \exp(-z^2/2)\sqrt{1 + b z^2/(1 + a z^2)}}}\right)$$ (11)

where $P_0 = \sqrt{\pi}/2$, $a = \frac{1 + \sqrt{1 + 6\pi + 2\pi^2}}{2\pi}$, $b = 2\pi a^2$.

The normal cumulative distribution can be approximated by estimating Mill’s ratio of the rational function. The rational function is addressed in Equation 11.

$$\Pr(Z > z) \approx \frac{\sum_{n=1}^{\infty} a_n z^{n-1} + \cdots + a_0}{z^{n-1} + b_n z^n + \cdots + b_0}$$ (12)

To have Equation 12 in the form of useful approximation, two positive integers, $n_1$ and $n_2$ must be selected. But we must also ensure that $n_1$ and $n_2$ are the two highest order terms to ensure the correct asymptotes as $z$ approaching infinity. Remember that the asymptotic theory is a generic framework for the assessment of properties of estimators and statistical tests. Within this framework, it is typically assumed that the sample size $n$ grows indefinitely, and the properties of statistical procedures are evaluated in the limit as $n$ approaches infinity. The result of coefficient (i.e., $a_{n-1}$, $a_{n-2}$, ..., $a_0$, $b_n$, $b_{n-1}$, ..., $b_0$) are determined by matching the derivatives of order 1, 2, 3,
etc. at $z = 0$. No model can be derived if only one order is chosen, so we need to choose at least two orders. If the chosen orders are $n = 1$ and $n = 2$, two approximations for cumulative normal distribution can be developed, one approximation for each order. In this paper, we will refer to the approximation of the order $n = 1$ as Model A and the approximation of order $n = 2$ as Model B. Model A is addressed in Equation 13 and Model B is addressed in Equation 14.

$$\Pr(Z > z) \approx \frac{z + 3.333}{\sqrt{2\pi z^2 + 7.32z + 3.333}} e^{-\frac{z^2}{2}}, z > 0,$$

(13)

$$\Pr(Z > z) \approx \frac{z^2 + 5.575192695z + 12.77436324}{\sqrt{2\pi z^3 + 14.38718147z^2 + 31.5351977z + 2 \times 12.77436324}} e^{-\frac{z^2}{2}}, z > 0$$

(14)

Model A is very simple and very accurate with a maximum absolute error of 0.0007. Furthermore, Model B is not very simple as Model A, but it has a superior accuracy with a maximum absolute error of 0.000019. Both models are defined on the domain $z \in [0:\infty]$. However, the models can be redefined to cover the negative $z$-score using the following two facts.

1) $\Pr(Z > z) = 1 - \Pr(Z < z)$

2) $\Pr(Z < -z) = 1 - \Pr(Z < z)$

(15)

(16)

The first fact is shared with all distributions, while the second one is valid only for standard normal distribution. We utilize these two properties to extend the model’s definition to the negative side. Keeping in mind to reverse the sign of $z$ on the domain of $z < 0$. However, if the power of the $z$ term is even, no need to reverse the sign as both signs have the same effect. Model A and B will have new shapes as addressed in Equation 17 and Equation 18, respectively.

$$\Pr(Z < z) \approx \Phi_A(z) = \begin{cases} \frac{(-z) + 3.333}{\sqrt{2\pi z^2 + 7.32(-z) + 3.333}} e^{-\frac{z^2}{2}}, z < 0 \\ 1 - \frac{z + 3.333}{\sqrt{2\pi z^2 + 7.32z + 3.333}} e^{-\frac{z^2}{2}}, z > 0 \end{cases}$$

(17)

$$\Pr(Z < z) \approx \Phi_B(z) = \begin{cases} \frac{z^2 + 5.575192695(-z) + 12.77436324}{\sqrt{2\pi(-z)^3 + 14.38718147z^2 + 31.5351977(-z) + 2 \times 12.77436324}} e^{-\frac{z^2}{2}}, z < 0 \\ 1 - \frac{z^2 + 5.575192695z + 12.77436324}{\sqrt{2\pi z^3 + 14.38718147z^2 + 31.5351977z + 2 \times 12.77436324}} e^{-\frac{z^2}{2}}, z > 0 \end{cases}$$

(18)

To analyze the accuracy change with $Z$-score, two deviation functions for Model A and Model B are proposed, as addressed in Equation 19 and Equation 20, respectively.

$$D_1(z) = \Phi_A(z) - \Phi(z)$$

(19)

$$D_2(z) = \Phi_B(z) - \Phi(z)$$

(20)

where $\Phi(z)$ is the real value of normal cumulative distribution density. The plots of both functions $D_1(z)$ and $D_2(z)$ over the range $-4 \leq z \leq 4$ are shown in the Figures, 1 and 2, respectively.
The shape of the deviation curve of Model A is a mirror image of the curve of Model B, besides the significant difference in the scale between the two curves (Model A has a much larger scale). For further clarity, both curves are plotted in the same figure (i.e., Figure 3). The difference in the scale is very noticeable in this figure.
3) LOWER TRUNCATED NORMAL CUMULATIVE DISTRIBUTION

In this section, two models are developed for the truncated cumulative normal distribution, one is for the order $n = 1$ and the other is for the order $n = 2$. The name Model A2 is selected for the first model, and Model B2 is selected for the second model.

At each $n$ value, an approximation to normal cumulative distribution density is derived in the previous section. Figure 4 illustrates the concept of lower truncation normal distribution, compared to the untruncated distribution. The solid line refers to the un-truncated distribution, the large segments discrete line refers to the lower truncated normal distribution at $z = -1$, and the small segments discrete line refers to the lower truncated normal distribution at $z = 0$.
The lower truncated normal distribution density is given by Equation 21. Further, the lower truncation normal cumulative distribution density is given by Equation 22.

\[ f_T(x) = \frac{f(x)}{\int_{x_L}^{x} f(x) dx}, x_L < x \quad (21) \]

\[ F_T(x) = \int_{x_L}^{x} f_T(t) dt, x_L < x \quad (22) \]

We have an approximation to \( F(x) \) of the standard normal distribution (i.e., \( \Phi(z) \)), and we can derive it to estimate \( f(x) \) (i.e., \( \varphi(z) \)). Solving Equation 23 leads to Equation 24.

\[ \Phi_A(z) \frac{d[\Phi_A(z)]}{dz} + \left( \frac{3(z-3.333)}{2\sqrt{\pi}x-7.32} \right) e^{-\frac{z^2}{2}}, z < 0 \quad (23) \]

\[ \Phi_T(A)_L(z) = \begin{cases} \frac{(z(z-3.333)+1)e^{-\frac{z^2}{2}}}{\sqrt{2\pi}z^2-7.32z+3.333} + \frac{(z-3.333)\left(\frac{3}{2\sqrt{\pi}}\sqrt{\pi}x-7.32\right)}{(\sqrt{2\pi}z^2-7.32z+3.333)^2}, z < 0 \\ \frac{(z(z+3.333)-1)e^{-\frac{z^2}{2}}}{\sqrt{2\pi}z^2+7.32z+3.333} + \frac{(z+3.333)\left(\frac{3}{2\sqrt{\pi}}\sqrt{\pi}x+7.32\right)}{(\sqrt{2\pi}z^2+7.32z+3.333)^2}, z > 0 \end{cases} \quad (24) \]

To calculate the first approximation (i.e., Model A2), Equation 25 is solved leading to Equation 26.

\[ \Phi_T(z) = \int_{z_L}^{z} \Phi_T(z) dt, z_L < z \quad (25) \]

\[ \Phi_T(A)_L(z) = \begin{cases} \frac{(3.333-z)e^{-\frac{z^2}{2}}}{\sqrt{2\pi}z^2-7.32z+3.333} - \frac{(3.333-z_L)e^{-\frac{z_L^2}{2}}}{\sqrt{2\pi}z_L^2-7.32z_L+3.333}, z < 0 \\ 1 - \frac{1}{\sqrt{2\pi}z_L^2-7.32z_L+3.333} - \frac{1}{\sqrt{2\pi}z^2+7.32z+3.333}, z > 0 \end{cases} \quad (26) \]

The same way is conducted to approximate Model B2 (i.e., order of \( n = 2 \)). \( \Phi_T(B)_L(z) \) is derived first, then \( \Phi_T(B)_R(z) \) is derived.

\[ \varphi_T(B)_L(z) = \begin{cases} \left[ (2z-5.575192695) - z \left( z^2 - 5.575192695 + 12.77436324 \right) \right] e^{-\frac{z^2}{2}} \\ -\sqrt{2\pi}z^3 + 14.38718147z^3 - 31.5351977z + 25.54872648 \\ (z^2 - 5.575192695 + 12.77436324)(-3\sqrt{2\pi}z^2 + 28.7743628z - 31.5351977) \left( -\sqrt{2\pi}z^3 + 14.38718147z^3 - 31.5351977z + 25.54872648 \right)^2, z < 0 \\ \left[ z \left( z^2 + 5.575192695 + 12.77436324 \right) - (2z + 5.575192695) \right] e^{-\frac{z^2}{2}} \\ -\sqrt{2\pi}z^3 + 14.38718147z^3 + 31.5351977z + 25.54872648 \\ (z^2 + 5.575192695 + 12.77436324)(3\sqrt{2\pi}z^2 + 28.7743628z + 31.5351977) \left( -\sqrt{2\pi}z^3 + 14.38718147z^3 + 31.5351977z + 25.54872648 \right)^2, z > 0 \end{cases} \quad (27) \]
\[ \Phi_{TB}(z) = \int_{z_L}^{T} \phi_{TB}(t) \, dt \]

Solving and simplifying the equation leads to the following

\[
\Phi_{TB}(z) = \begin{cases} 
\frac{\left( z^2 - 5.575192695z + 12.77436324 \right) e^{-\frac{z^2}{2}}}{\sqrt{2\pi} z} + 14.38718147z^2 - 31.5351977z + 2 \times 12.77436324} & , z < 0 \\
1 - \frac{\left( z_L^2 - 5.575192695z_L + 12.77436324 \right) e^{-\frac{z_L^2}{2}}}{\sqrt{2\pi} z_L^3} + 14.38718147z_L^2 - 31.5351977z_L + 2 \times 12.77436324} & , z > 0 \\
\end{cases}
\]

(28)

In the following discussion, an example about the application of our model in industry is given. There are many advance situations where the model can be used, but we will take a simple example to just give an idea. Assume the life time of an air conditioner follows normal distribution with a mean of 25.3 years and standard deviation of 7. A company is looking to buy 100 units of this conditioner, but they will buy 5 year used of this brand to reduce the price since they have financial problems. An engineer in the company is interested to estimate the chance that each conditioner will survive another 7 years hoping their financial problems will be resolved at that time. As a summary of the question, what is the chance that the 5 years used air conditioner will survive for another 7 years if the life distribution of new devices is normal with a mean of 25.3 years and standard deviation of 7 years?

This problem falls in the area of Reliability Engineering. Reliability Engineering 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 [17]. Since the proposed model is based on standard normal distribution, we have to replace \( z \) with \( (x - \mu)/\sigma \) and
with \((x_L - \mu)/\sigma\) in Model A2. The used parameters in details are as follows: \(x = 5 + 7 = 12\), \(\mu = 25.3\), \(\sigma = 7\), and \(x_L = 5\). Solving Model A2 with these parameters gives the value of 0.02721 referring to the chance of failing before the end of next 7th year. The chance of survive is just the complement of value 0.02721 to 1.0, which is 0.97279. In other word, there is 97.279% chance to survive for more than another 7 years for any of these conditioners. The true result of this value is 0.0973099 which is very close to the model result. The error is only 0.000309. This level of error is very ignorable for most business and engineering applications. It is difficult to use Model B2 to solve manual solution case like the previous example.

4) MODEL’S ACCURACY

The accuracy of the model is evaluated through the maximum absolute deviation, which changes with \(z_L\). This maximum absolute deviation is shown in Figures 5 and 6 for models A2 and B2, respectively. For Model A2, Figure 5 shows the maximum absolute deviation curves at the following truncation point (\(Z_L\)): -4,-3,-2, -1, and 0. It can be seen that the maximum absolute deviation is about -0.0014 for \(z_L = 0\) curve. Also, a similar figure for Model B2 is shown in Figure 6. The maximum absolute deviation is about 3.7 *10^{-5} for \(z_L = 0\) curve. The accuracy of the second model is superior.
Figures 6 and 7 show a 3D visualization of the deviation for both models, A2 and B2, respectively with varying values of $z$ and $z_L$.

Figure 6: The Deviation for Model B2 Results from Real Results at Different $z_L$ Values

Figure 7: 3D Model of the Maximum Absolute Deviation for Model A2
Finally, the maximum absolute deviation is plotted versus $Z_L$ for both models A2 and B2 on the same graph. See Figure 9. The figure clearly shows the higher accuracy which can be achieved by using model B2 in comparison to model A2.
5) CONCLUSION

This paper has proposed very high accuracy mathematical models for lower truncated normal distribution. This proposed model has used Mill’s ratio with an order of 1 and 2 to develop two models for standard normal distribution, and then built on those models to develop lower truncated normal distribution models. The first model is very simple, while the second model is more advanced and accurate one. The first model (i.e., order of 1) was very accurate with a maximum absolute error of about 0.0015 over the domain $[Z_L : \infty]$; $Z_L [-\infty : 0]$. The second model (i.e., order of 2) is a superior accurate with a maximum absolute error of less than 0.00004 over the same domain. We recommend practitioners to use the first model in their hand calculation.

REFERENCES


