Pak. J. Statist.
2019 Vol. 35(4), 315-320

FITTING A PIECE WISE CUBIC REGRESSION MODEL TO A CUMULATIVE NORMAL DISTRIBUTION

I.H. Tajuddin
College of Computer Science & Information System
Institute of Business Management (IoBM), Karachi, Pakistan
Email: tajuddini@yahoo.com

ABSTRACT

The standard cumulative normal distribution has been approximated by over 70 formulas. Very recently, Eidous and Abu-Shareefa (2019) have reviewed 45 competitive formulas given in literature from 1945 to 2019, to approximate the standard cumulative normal distribution. They introduced nine more accurate but quite complicated formulas in computation. However, they have recommended an approximation “$\Phi_{\mathrm{SE}}(\mathrm{z})$" given by Soranzo and Epure (2014) on the basis of simplicity and the low maximum absolute error. This has motivated us to re-examine some competitive approximations and propose a new approximation based on fitting a cubic regression model, to fit the normal cumulative probabilities over some ranges of the standardized values.

1. INTRODUCTION

In the last 50 years, computing facilities have been enhanced exponentially. Modern calculators can carry on calculations at much more accuracy and speed than the 50 year old big computers could do. This has led researchers to solve difficult problems in all scientific areas, in particular approximating a cumulative normal distribution function "$\Phi(z)$". The normal table provides probabilities correct to four decimal places whereas some authors have claimed that their approximations are more accurate. Most authors have not discussed the rationale behind the formula given. Some have used Logistic distribution or its modified forms [see, Lin (1990), Bowling et al. (2009)] to approximate $\Phi(z)$. Some have used some terms from Taylor’s expansion of the normal distribution’s exponent term. Shore (2005) has used RMM based approximation. Eidous and Abu-Shareefa (2019) have given a review of competitive formulas given in literature from 1945 to 2019, to approximate the standard cumulative normal distribution. They introduced nine more accurate formulas but quite complicated in computation. The authors have mentioned for formula 12 described in their paper that Max. AE < MAE which is not possible. Nonetheless, they have recommended the use of “$\Phi_{\mathrm{SE}}(\mathrm{z})$" given by Soranzo and Epure (2014) on the basis of simplicity and the low maximum absolute error. This has led us to re-examine some simple competitive approximations and propose a new approximation. In this paper, we use a regression model to approximate $\Phi(z)$ over different ranges of z. In section 2, we mention the approximation formulas considered in our study, and in section 3, we discuss the proposed formula to approximate $\Phi(z)$ for different range of z. In section 4, we compare the performance of the proposed approximation with some approximations considered in section 3. In section 5, we conclude the paper with our recommendations.

2. SOME APPROXIMATION FORMULAS CONSIDERED

Eidous and Abu-Shareefa (2019) have computed the maximum absolute error, (Max. AE) and the mean absolute error (MAE) for different approximations for z over [0,5]. In addition, they have given 9 new formulas and the ninth one gives an excellent approximation. However as their proposed formulas are complicated, they recommended the formula by Soranzo and Epure (2014), given below.

$$\Phi_{\mathrm{SE}}(\mathrm{z}) = 2^{-\mathrm{g(z)}}, \text{where } \mathrm{g(z)} = 22^{1-41^{0.1\mathrm{z}}}, \forall\, \mathrm{z} \geq 0. \tag{1}$$

We have included the following approximations by different researchers in our study. Choudhury (2014) has given a simple and competitive formula:

$$\Phi_{\mathrm{C}}(z) = 1 - \frac{\exp\{-0.5\, z^2\}}{\sqrt{2\pi}\,(0.226 + 0.64\, z + 0.33\, \sqrt{z^2 + 3}\,)}, -\infty < z < \infty \tag{2}$$

Yerukala and Boiroju (2015) have modified the above formula (2), which results in more accuracy, is as given below.

$$\Phi_{\mathrm{YB}}(z) = 1 - \frac{\exp\{-0.5\, z^2\}}{\frac{44}{79} + 1.6\, z + \frac{5}{6}\sqrt{z^2+3}}, \quad -\infty < z < \infty \tag{3}$$

Bowling et al. (2009) have given an approximation based on a modified form of the logistic distribution.

$$\Phi_{\mathrm{Bet.al}}(z) = \frac{1}{1+\exp\{-(1.5976\, z + 0.07056\, z^3)\}}, -\infty < z < \infty \tag{4}$$

We propose the following formula based on piece wise fitting of regression cubic curve. This will be discussed in the next section.

$$\Phi_{\mathrm{proposed}}(z) = \begin{cases} 0.5 + 0.40054\, z - 0.0103\, z^2 - 0.0503\, z^3, & 0 \leq z \leq 0.8 \\ 0.44351 + 0.57918\, z - 0.20597\, z^2 + 0.024929\, z^3, & 0.8 < z \leq 2.8 \\ 0.80987 + 0.15999z - 0.045228\, z^2 + 0.004291z^3 & 2.8 < z \leq 3.5 \\ 0.965807 + 0.0240906\, z - 0.005674z^2 + 0.000446z^3 & 3.5 < z \leq 4.4 \\ 1 & z > 4.4 \end{cases} \tag{5}$$

The above formula (5) can be presented as follows.

5a: $\Phi_{\mathrm{proposed}}(z) = 0.5 + 0.40054\, z - 0.0103\, z^2 - 0.0503\, z^3, \quad 0 \leq z \leq 0.8$

5b: $\Phi_{\mathrm{proposed}}(z) = 0.44351 + 0.57918\, z$
$-0.20597\, z^2 + 0.024929\, z^3, 0.8 < z \leq 2.8$

5c: $\Phi_{\mathrm{proposed}}(z) = 0.80987 + 0.15999z$
$-0.045228\, z^2 + 0.004291z^3, 2.8 < z \leq 3.5$

5d: $\Phi_{\mathrm{proposed}}(z) = 0.965807 + 0.0240906\, z$
$-0.005674z^2 + 0.000446z^3, 3.5 < z \leq 4.4$

For comparison purpose, we have studied the following formulas given by Zogheib and Hlynka (2009) claimed to perform good in the specified domain of z.

$$\Phi_{\mathrm{ZH1}}(z) = 0.5 + 0.398942\, z - 0.06649\, z^3 + 0.09974\, z^5, 0 \leq \mathrm{z} < 1 \tag{6}$$

$$\Phi_{\mathrm{ZH2}}(z) = 0.5 + 0.368929\, z - 0.037758\, z^3 + 0.0645\, z^5,\ 0 \leq z < 1.2 \tag{7}$$

We mention here that the coefficients of z given in (6) and (7) above, are positive; not as are in Zogheib and Hlynka (2009).

Dombi and Jonas (2018) have given a simple approximation to $\Phi(z)$ for z ∈ $(-\pi, \pi)$. Their formula given below, is simple and satisfies some desirable properties not satisfied by other approximations.

$$\Phi_{\mathrm{DJ}}(z) = \frac{1}{1+\left(\frac{\pi-z}{\pi+z}\right)^{\sqrt{(2\pi)}}},\ -\pi < z < \pi. \tag{8}$$

Shore (2005) has used the response modeling methodology (RMM) to approximate the standard normal cumulative probabilities using the following formula:

$$\Phi_{\mathrm{S}}(z) = [1 - \mathrm{g(z)} + \mathrm{g(-z)}]/2\,, \tag{9}$$

where $\mathrm{g(z)} = \exp\{-\log(2)\exp\left\{\left(\frac{\alpha S_1}{\lambda}\right)\left[(1+\ S_1 z)^{\frac{\lambda}{S_1}} - 1\right] + S_2 z\right\}$,
where $\lambda = -0.6122883;\ S_1 = -0.11105481;\ S_2 = 0.4434159$; and $\alpha = 6.37309208$.

Formula (9) is very much complicated, so we did not include this as well as other complicated formulas in our study. Formulas which are very simple are not accurate and thus are not included in our study.

Following the approach of Eidous and Abu-Shareefa (2019), we obtained maximum absolute error, (Max. AE) and the mean absolute error (MAE), and in addition the mean square error (MSE) for formulas (1) to (8) and present the results in section 3.

3. METHOD USED

We used Minitab to get the values of $\Phi(z)$ for different values of z, with steps of 0.001.

We fitted a cubic regression for the response variable $\Phi(z)$ for the predictor z in [-1, 1] and upon removing outliers, we found a cubic linear regression that fits well on [-0.8, 0.8]. In a similar manner, we ended up with the cubic regression over different ranges as described in formula (5). Firstly, we obtained the following fitted models over different ranges of z values.

I: For $0 \leq z \leq 0.8$, we find the following fitted model.

$\Phi(z) = 0.499916 - 0.401034\, z - 0.0107521 z^2 - 0.0504206\, z^3$,
with R^2= 100.0%

Models for other ranges given by Minitab were as follows.

II: $\Phi(\mathrm{z}) = 0.441541 + 0.582678\ \mathrm{z} - 0.207941 z^2$
$-0.0252824 \mathrm{z}^3, 0.8 < z \leq 2.8$

III: $\Phi(\mathrm{z}) = 0.499916 - 0.401034\mathrm{z} - 0.0107521 z^2$
$-0.0504206 \mathrm{z}^3, 2.8 < z \leq 3.5$

IV: $\Phi(\mathrm{z}) = 0.966351 + 0.0236887\mathrm{z} - 0.0055726 z^2$
$+0.00043789\ \mathrm{z}^3, 3.5 < z \leq 4.4$

All models gave R^2 = 100.0%.However, the constant term in model I should be 0.5, moreover other coefficients are rounded figures given by Minitab. The precise use of given coefficients did not give R^2 = 100.0%. We used the symmetry and fitted a cubic model for $-0.8 \le \mathrm{z} \le 0.8$ and upon some other considerations, we obtained fixed coefficients to approximate $\Phi_{\text{proposed}}(z)$ for $0 \le \mathrm{z} \le 0.8$, given below.

$$\Phi_{\text{proposed}}(z) = 0.5 + 0.40054\,\mathrm{z} - 0.0103\,\mathrm{z}^2 - 0.0503\,\mathrm{z}^3 \text{ as in formula (5a).}$$

In a similar manner we ended up with the proposed piece-wise approximation, for other ranges, given by formula 5b, 5c and 5d, in section 3.

Remark:

Coefficients of terms in models I to IV are different from those given in formula (5) or given by formula 5a, 5b, 5c and 5d.

4. COMPARISONS

We have included the Mean Square Error (MSE) in our computation. We believe that taking z in [0,5] will under estimate the MSE and the MAE. In table 1, we present the results of MSE, MAE and the max. error for z in [0,4.4] for formulas (1) to (5) and for the specified ranges for formulas (6) and (7).

Table 1
Some Statistics for Different Formulas

Approximations given by different Formulas for Specified Ranges	**MSE**	**MAE**	**Max. Error**
1. Soranzo and Epure (2014) for $0 \le z \le 4.4$	0.000000004	0.000066	0.000127
2. Choudhary (2014) for $0 \le z \le 4.4$	0.000000003	0.000042	0.000193
3. Yerukala and Boiroju (2015) for $0 \le z \le 4.4$	0.000000003	0.000040	0.000107
4. Bowling et al. (2009) for $0 \le z \le 4.4$	0.000000002	0.000085	0.000141
5. Proposed formula for $0 \le \mathrm{z} \le 4.4$	0.000000006	0.000079	0.000348
6. Zogheib & Hlynka (2009)-1 for $0 \le z \le 1.0$	0.000158556	0.015094	0.033358
7. Zogheib & Hlynka (2009)-2 for $0 \le z \le 1.2$	0.000039073	0.008959	0.033471
8. Dombi and Jonas (2018) for $0 \le z \le 3.141$	0.000002116	0.001262	0.002357

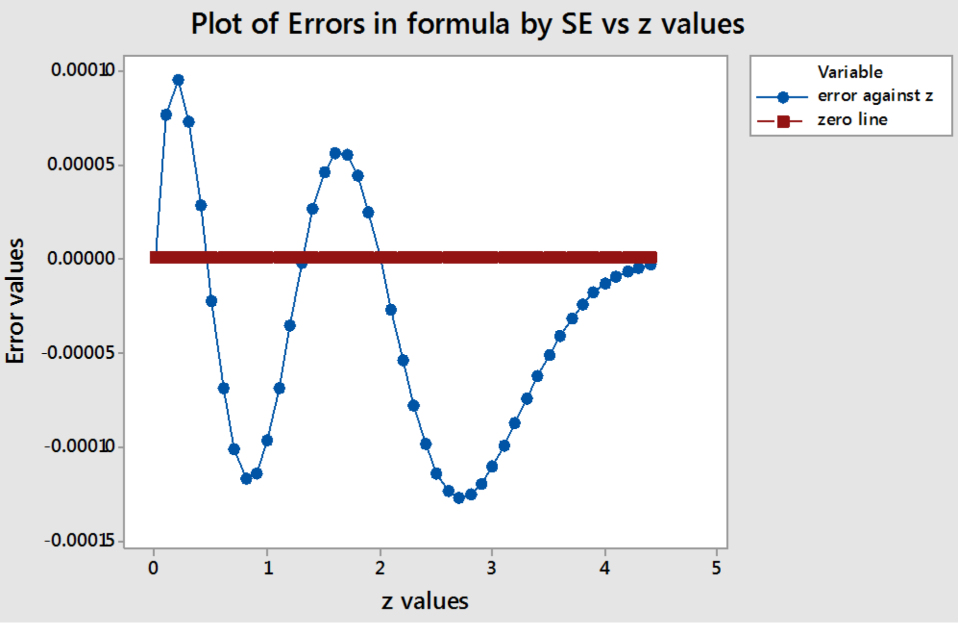
Some of the results given in table 1 can be compared with those given by Eidous and Abu-Shareefa (2019).

The graph below shows errors for the formula (1) for different values of z in [0, 4.4].

From Table 1, we observe the following.

1. The formula (3)-Yerukala and Boiroju (2015) is better than the recommended formula (1) given by Soranzo and Epure (2014), and it is easy to use.
2. Formula (4) given by Bowling et al. (2009) has the least MSE but the highest MAE among the five formulas.
3. Both the formulas (6) and (7) due to Zogheib and Hlynka (2009) do not stand any merit compared to other approximations.
4. Formula 8 has desirable properties but not as accurate as is the proposed formula
5. The proposed formula performs better than other formulas for the tail areas as well

as for area closer to the center as can be seen for results of 5a, 5c and 5d.



The graph clearly shows that for many values of z, error is more than 0.0001.

5. DISCUSSIONS AND CONCLUSION

Discussion:

1. The proposed formula gives an approximation to $\Phi(z)$ for z ≥ 0. One needs to use symmetry to obtain $\Phi(z)$ for $z < 0$, using $\Phi(z) = 1 - \Phi(-z)$.
2. The proposed formula does not satisfy the desirable property of symmetry $\Phi(z) = 1 - \Phi(-z)$ and $\emptyset(0) = \frac{1}{\sqrt{2\pi}}$ satisfied by formula (8). However, it satisfies the desirable property $\Phi(0) = 0.5$ not satisfied by most of approximations given in the literature.
3. The proposed formula in fact gives an approximation to half normal distribution.
4. One of the application of the proposed approximation is to compute probabilities for gamma distribution with shape parameter $\alpha = 0.5$. It is easy to see U = 0.5 Z^2 follows gamma distribution with $\alpha = 0.5$.
5. The formulas included in this study and the proposed one are correct to three places of decimals; the usual normal table stands merit over these formulas.

Conclusion:

We recommend the use of the proposed formula for computing p values in testing procedures. However, we believe an approximation must be as simple as given by Hoyt (1968), Lew (1981), Lin (1988) and given by others. Unfortunately, the very simple formulas are hardly correct to 2 places of decimals on specified ranges.

In general, if one needs more accurate values beyond table values, without using any package, one should use the ninth formula proposed in Eidous and Abu-Shareefa (2019).

Future Work:

It will remain a desire to obtain a simple formula for the whole real line to replace the normal table. The method described in section 3, can be used to obtain an approximate difficult integrals such as incomplete gamma distribution.

ACKNOWLEDGEMENT

Author is very much thankful for the referees for their comments.

REFERENCES

1. Bowling, S.R., Khasawneh, M.T., Kaewkuekool, S. and Cho, B.R. (2009). A logistic approximation to the cumulative normal distribution. *Journal of Industrial Engineering and Management*, 2, 114-127.
2. Choudhury, A. (2014). A simple approximation to the area under the normal curve. *Mathematics and Statistics*, 2(3), 147-149.
3. Dombi, J. and Jonas, T. (2018). Approximations to the normal probability distribution function using operators of continuous-valued logic. *Acta Cybernetica*, 23, 829-852.
4. Eidous, O.M. and Abu-Shareefa, R. (2019). New Approximations for standard normal distribution function. *Communications in Statistics–Theory and Methods*, https://doi.org/10.1080/03610926.2018.1563166.
5. Hoyt, J.P. (1968). Simple approximations to the normal distribution function. *American Statistician*, 22(3), 25-26.
6. Lew, R.A. (1981). An approximation to the cumulative normal distribution with simple coefficients. *Applied Statistics*, 30(3), 299-301.
7. Lin, J.T. (1988). Alternative to Hamaker's approximation to the cumulative normal distribution and its inverse. *Applied Statistics*, 37, 413-414.
8. Lin, J.T. (1990). A simpler logistic approximation to the normal tail probability and its inverse. *Applied Statistics*, 39(2), 255-257.
9. Minitab 17 Statistical Software (2010). [Computer software]. State College, PA: *Minitab, Inc*. (www.minitab.com).
10. Shore, H. (2005). Accurate RMM-based approximations for the CDF of the normal distribution. *Communications in Statistics–Theory and Methods*, 34(3), 507-13.
11. Soranzo, A. and Epure, E. (2014). Very simply explicitly invertible approximations of normal cumulative and normal quantile function. *Applied Mathematical Sciences* 87, 4323-41.
12. Yerukala, R. and Boiroju, N.K. (2015). Approximations to standard normal distribution function. *International Journal of Scientific & Engineering Research*, 6, 515-518.
13. Zogheib, B. and Hlynka, M. (2009). *Approximations of the standard normal distribution*. University of Windsor, Dept. of Mathematics and Statistics, Online available at http://www1.uwindsor.ca/math/sites/uwindsor.ca.math/files/09-09.pdf