Journal of Statistical Research *ISSN 0256 - 422 X*
2007, Vol. 41, No. 1, pp. 59–67
Bangladesh

APPROXIMATING THE CUMULATIVE DISTRIBUTION FUNCTION OF THE NORMAL DISTRIBUTION

Amit Choudhury
Department of Statistics
Gauhati University, Guwahati 781014, India
Email: achoudhury@rediffmail.com

Subhasis Ray
ICFAI Business School, Kolkata 700091, India.
Email: subhasisr@ibsindia.org

Pradipta Sarkar
The Procter and Gamble Company, Cincinnati, OH 45217, USA.
Email: sarkar.p@pg.com

SUMMARY

Over the years, a number of approximations to the cdf of the Normal distribution have been proposed. How does one make a choice among them? This paper compares their performance with a view to identifying the best among them. Our analysis reveals that a uniformly best approximation formula does not exist. Locally best are identified. Finally we combine the locally best approximations to obtain a combined formula with a very high accuracy. A subroutine is presented.

Keywords and phrases: Normal distribution, approximating cdf.

1 Introduction and Motivation

The normal distribution is perhaps the most widely used of all statistical distributions. Area under the normal probability curve, also known as the normal cdf, is a measure that almost all of us have dealt with at some point or other. The normal cdf does not have a closed form solution and requires numerical techniques to evaluate the associated integral. Therefore, unlike many other probability distributions, one cannot quickly compute the normal cdf without consulting the normal probability table or using software. This inconvenience in readily evaluating the normal cdf together with its widespread use has led to a few attempts at constructing approximations. These approximations vary in accuracy and complexity.

There appears to be three broad approaches to such an exercise, the most popular being construction of approximation formulas. The second approach involves the use of

© Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka 1000, Bangladesh.

distributions that closely resemble the normal distribution under specific conditions. The third approach involves construction of bounds for the normal cdf. In all these approaches, the approximations or bounds have been for the standard normal variety.

Simple bounds such as the ones found in probability textbooks such as Feller (1968) are very popular. This approach is perhaps a century old and belongs to the pre–scientific calculator age. Later on different authors have provided sharper bounds and other approximations. Examples include approximation formula by Tocher (1963), Zelen and Severo (1964), Page (1977), Hammakar (1978), Abernathy (1988), Lin (1989, 1990), Bagby (1995), and Byrc (2001). Szarek (1999) proposed bounds to this function. Johnson and Kotz (1994) have listed a few distributions that are close to the standard normal probability distribution function under certain specific conditions. Of these probability distributions we have selected the standard Logistic for our review as it is the closest.

With such a long menu, how does one make a choice? Simply put, which approximation formula works best? To the best of our knowledge, such an exercise has not been carried out. The usefulness however is obvious.

The purpose of this article therefore is to review available literature on normal cdf approximations, identify regions where one of them outperforms others and finally combine them, if possible, to obtain an approximation that is uniformly better than all of the above. An additional aim is significant. Academician and practitioners are often required to write programs in different languages (FORTRAN, C etc) which require the cdf of normal distribution. Presently, libraries of such programming languages do not offer any in built subroutine or function to compute the normal cdf. Consequently, an algorithm for a highly accurate approximation formula will also be presented.

2 Overview of Approximations

Various approximation formula to the standard normal cdf culled from literature are enumerated below.

1. Tocher (1963):
$$\Phi(x) \approx e^{2kx}/(1+e^{2kx}), \text{where } k = \sqrt{2/\pi}.$$

2. Zelen and Severo (1964):
$$\Phi(x) \approx 1 - (0.4361836t - 0.1201676t^2 + 0.9372980t^3)\left(\sqrt{1/2\pi}\right)e^{-x^2/2},$$
where $t = (1 + 0.33267x)^{-1}$.

3. A popular bound found in many probability texts (for example, Feller 1968).
For $x > 0$
$$(x^{-1} - x^{-3})\left(\sqrt{2/\pi}\right)^{-1} e^{-x^2/2} < 1 - \Phi(x) < x^{-1}\left(\sqrt{2/\pi}\right)^{-1} e^{-x^2/2}.$$

4. Page (1977):
$$\Phi(x) \approx 0.5\left\{1+\tanh(y)\right\}$$
where $y = \left(\sqrt{2/\pi}\right)x(1+0.044715x^2)$.

5. Hammakar (1978):
$$1-\Phi(x) \approx 0.5\left\{1-\left(1-e^{-y^2}\right)^{0.5}\right\}, y = 0.806x(1-0.018x).$$

6. Abernathy (1988):
$$\Phi(x) \approx 0.5 + \frac{1}{\sqrt{2\pi}}\sum_{n=0}^{\infty}\frac{(-1)^n x^{2n+1}}{2^n n!(2n+1)}, x > 0.$$

7. Lin (1989):
$$1-\Phi(x) \approx 0.5\left(e^{-0.717x-0.416x^2}\right), x > 0.$$

8. Lin (1990):
$$1-\Phi(x) \approx \frac{1}{1+e^y} \text{where } y = 4.2\pi\left(\frac{x}{9-x}\right), x > 0.$$

9. Bagby (1995):
$$\Phi(x) \approx 0.5 + 0.5\left(1-\frac{1}{30}\left\{7e^{-x^2/2}+16e^{-x^2(2-\sqrt{2})}+(7+\frac{\pi x^2}{4})e^{-x^2}\right\}\right)^{0.5}, x > 0.$$

10. Szarek bounds (1999): For $x > -1$,
$$\frac{2}{x+(x^2+4)^{0.5}} \leq e^{x^2/2}\int_x^{\infty} e^{-t^2/2}dt \leq \frac{4}{3x+(x^2+8)^{0.5}}$$

11. Byrc (2001A):
$$\Phi(x) \approx 1 - \frac{(4-\pi)x+\sqrt{2\pi}(\pi-2)}{(4-\pi)x^2\sqrt{2\pi}+2\pi x+2\sqrt{2\pi}(\pi-2)}e^{-x^2/2}.$$

12. Byrc (2001B):
$$\Phi(x) \approx 1 - \frac{x^2+5.575192695x+12.77436324}{x^3\sqrt{2\pi}+14.38718147x^2+31.53531977x+25.548726}e^{-x^2/2}.$$

13. Standard Logistic cdf:
$$\Phi(x) \approx F(x) = \left(1+e^{-\pi x/\sqrt{3}}\right)^{-1}$$

Table 1: Maximum Absolute Errors for Different Approximation Methods

	Range of the Standard Normal Variable		
Approximation Method	0 - 1.0	1.0 - 3.0	3.0 - 4.0
Tocher (1963)	9.919×10^{-3}	1.767×10^{-2}	6.912×10^{-3}
Zelen and Severo (1964)	1.120×10^{-5}	1.095×10^{-5}	4.990×10^{-6}
Page (1977)	1.530×10^{-4}	1.791×10^{-4}	1.373×10^{-4}
Hammakar (1978)	6.229×10^{-4}	3.852×10^{-4}	2.800×10^{-6}
Lin (1989)	6.585×10^{-3}	2.374×10^{-3}	2.690×10^{-5}
Lin (1990)	6.688×10^{-3}	2.538×10^{-3}	1.220×10^{-5}
Bagby (1995)	3.040×10^{-5}	2.960×10^{-5}	2.710×10^{-6}
Byrc (2001B)	1.185×10^{-5}	1.873×10^{-5}	2.051×10^{-6}
Std Logistic	2.266×10^{-2}	1.846×10^{-2}	2.963×10^{-3}

We need to compare the accuracy of these approximations. We have chosen the NORMDIST function of SAS software as the gold standard for computing the cdf of the standard normal distribution and consequently for comparing the accuracy of different approximation formulas. The choice of NORMDIST function has largely been dictated by the fact that SAS uses a highly accurate Monte Carlo technique for computing area under the N(0,1) curve thereby determining the area with very high precision. For determining the accuracy of various approximation formulas, we have determined absolute errors of each of these approximations for $x = 0(0.0005)4$. Above 4, one can take the cdf as 1. Consistent with the symmetric property of the distribution, we have restricted our comparison to $x > 0$. The absolute errors have been computed for each of the approximations with reference to the NORMDIST function. Some summary statistics are placed in Table 1.

Summary statistics have not been computed for Feller (1968) bounds and Szarek (1999) as they perform poorly – the bounds are quite wide. Besides, practitioners will perhaps be more comfortable with a point approximation of the area under normal curve rather than a bound, notwithstanding the effort. Summary statistics have not been calculated for Abernathy (1988) too as the absolute errors are not small as the proposer has shown (refer to Table 1 of his paper). Further as shown by Byrc (2001), his second formula is more accurate than his first and hence summary statistics for Byrc (2001A) has not been shown.

In Table 1 the choice of upper boundary of the first interval (0,1] has been dictated by the fact that 1 is the inflexion point of N(0,1) distribution. Upper boundary of the second interval (1,3] has been so chosen as it is well known that 99.73% of the area under standard normal curve lies between ± 3 and for most practical applications is considered as the effective domain of the variable of interest.

Table 2: Mean Absolute Error for Different Approximation Methods in the range (0,4]

Approximation Method	Mean Absolute Error
Tocher (1963)	8.592×10^{-3}
Zelen and Severo (1964)	5.980×10^{-6}
Page (1977)	9.470×10^{-5}
Hammakar (1978)	1.682×10^{-4}
Lin (1989)	1.342×10^{-3}
Lin (1990)	1.365×10^{-3}
Bagby (1995)	1.160×10^{-5}
Byrc (2001B)	6.921×10^{-6}
Std Logistic	7.311×10^{-3}

3 A Uniformly Best Approximation Formula

It is clear from Table 1, Table 2 and Fig.1 that a uniformly best approximation formula does not exist. It can however be observed that different approximations work best in different segments of the range (0,4]. We have been able to identify the locally best approximating formulas for different segments. These are:

1. Byrc (2001B) works best for (0, 0.7315] in the sense that its absolute error is uniformly lower than that of all other formulas in this range.

2. In the interval (0.7315, 1.726], Zelen and Severo (1964) works best. However, there are three small sub–intervals viz [0.791, 0.822], [1.437, 1.442] and [1.548, 1.5525] where Bagby (1995), Lin (1990) and Lin (1989) have lower absolute errors respectively. Nevertheless from the practicality point of view, we suggest that Zelen and Severo (1964) be used even in these small sub–intervals. Even if one were to use Zelen and Severo (1964), the approximations would be worse off by at most $1.44719 \times 10^{-6}, 7.08 \times 10^{-6}$ and 3.48×10^{-6} respectively which are small enough to be ignored.

3. Bagby (1995) works best in the range (1.726, 1.8135].

4. In (1.8135, 2.2075], Zelen and Severo (1964) again works best except for one small sub–interval [1.8715, 1.9] where Page (1977) has a lower absolute error. Again from the practicality point of view, we suggest that Zelen and Severo (1964) be used in this small sub–interval. The approximation would be worse off at most by 6.22823×10^{-6}

5. In (2.2075, 2.7245], Byrc (2001B) is best. Here too, there are three small sub intervals viz [2.3605,2.448], [2.448,2.518] and [2.5965,2.661] where Hammakar (1978), Lin (1990)

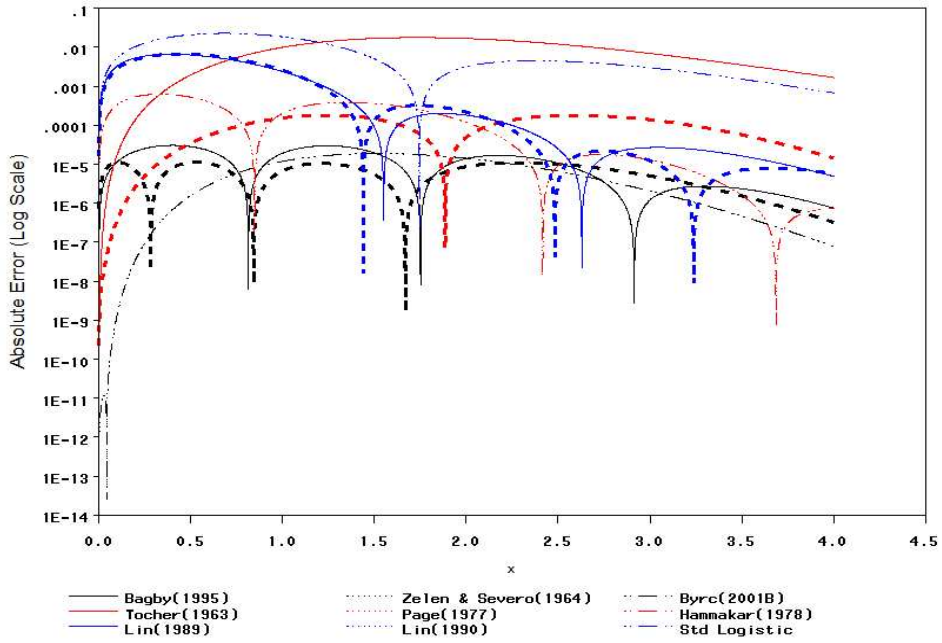

Figure 1: Plot of Absolute Errors of Different Approximations

and Lin (1989) have lower absolute errors respectively. Again from practical necessity, we suggest the use of Byrc (2001B) all through as the errors would be worse off by at most $7.63225 \times 10^{-5}, 6.7193 \times 10^{-6}$ and 4.9911×10^{-6} respectively; all too small to be of any consequence.

6. Bagby (1995) is again best in the range (2.7245, 3.056].

7. Lastly, Byrc (2001B) is best in the range (3.056, 4] except for one sub–interval [3.64, 3.72] where Hammakar (1978) has lower absolute error. We however, suggest use of Byrc (2001B) all along (3.056, 4] as the error would at most worsen by 2.19914×10^{-7}.

We recommend that different approximation formulas be used in different ranges as detailed above. An improved approximation formula for the standard normal cdf can now be constructed using this recommendation. This combined approximation is as follows:

Table 3: Summary Statistics on Combined Approximation Formula (3.1)

Range of the Standard Normal Variable	0.0 - 1.0	1.0 - 3.0	3.0 - 4.0
Maximum Absolute Error	6.77732×10^{-6}	1.07936×10^{-5}	1.76549×10^{-6}
Mean Absolute Error in the range (0,4]	3.74037×10^{-6}		

$$h(x) = \begin{cases} 1 - \frac{x^2+5.575192695x+12.77436324}{x^3\sqrt{2\pi}+14.38718147x^2+31.53531977x+25.548726} e^{-x^2/2}. & \text{if } x \in (0, 0.7315] \\ & \text{or } (2.2075, 2.7245] \\ & \text{or } (3.056, 4] \\ 1 - (0.4361836t - 0.1201676t^2 + 0.9372980t^3)\left(\sqrt{1/2\pi}\right) e^{-x^2/2}, & \text{if } x \in (0.7315, 1.726] \\ \text{where } t = (1 + 0.33267x)^{-1} & \text{or } (1.8135, 2.2075] \\ 0.5 + 0.5\left(1 - \frac{1}{30}\left\{7e^{-x^2/2} + 16e^{-x^2(2-\sqrt{2})} + (7 + \frac{\pi x^2}{4})e^{-x^2}\right\}\right)^{0.5} & \text{if } x \in (1.726, 1.8135] \\ & \text{or } (2.7245, 3.056] \end{cases} \quad (3.1)$$

Compared to existing approximations, this combined approximation formula provides better accuracy as Table 3 shows. An algorithm for a subroutine determining the combined approximation formula is placed in the appendix.

4 Conclusion

In effect, a fresh formula for approximating the area under the standard normal curve has been proposed. As is apparent from Table 3 and Figure 2, this combined formula performs better than other others currently available in the literature. While we do not claim to have rendered the normal probability table redundant, our combined approximation is a very close competitor. The algorithm should be of use too.

5 Acknowledgments

We would like to thank Hongjun Wang from University of Cincinnati for his help with graphs. Also we thank the referees and the members of the editorial board for their help.

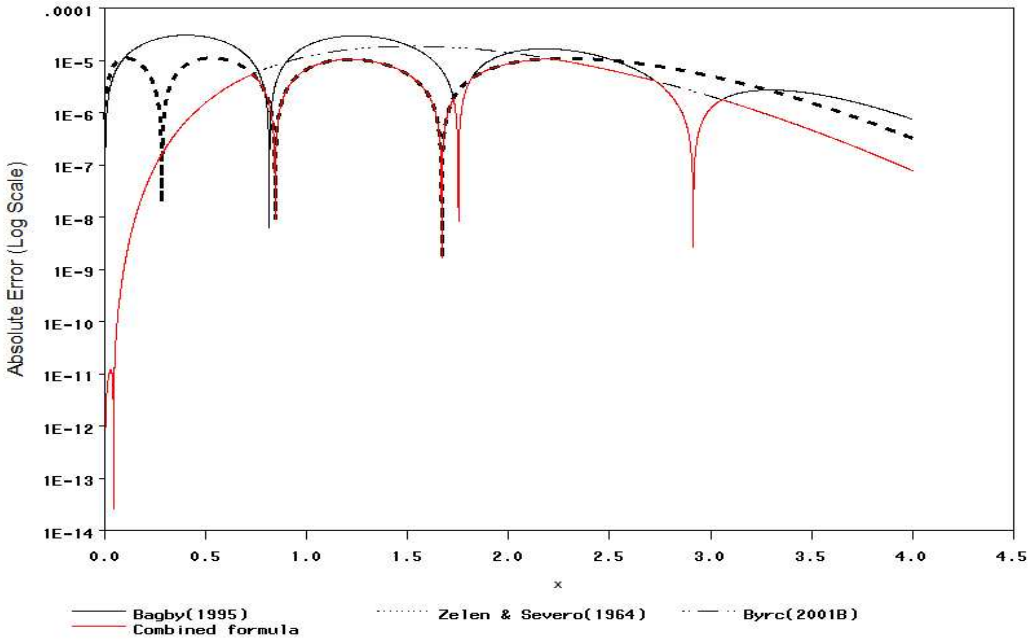

Figure 2: Plot of Absolute Errors of Three Locally Best Approximations and the Combined Approximation Formula

References

[1] Abernathy, R.W. (1988). Finding Normal Probabilities with Hand held Calculators. *Mathematics Teacher,* **81**, *651–652.*

[2] Bagby, R.J. (1995). Calculating Normal Probabilities. *American Math Monthly,* **102**, *46–49.*

[3] Byrc, W. (2001). A uniform approximation to the right normal tail integral. *Applied Mathematics and Computation,* **127**, *365–374.*

[4] Feller. W. (1968). An Introduction to Probability Theory and its Applications, Vol. 1, Wiley Eastern Limited.

[5] Hammakar, H. C. (1978). Approximating the Cumulative Normal Distribution and its Inverse. *Applied Statistics,* **27**, *76–77.*

[6] Johnson, N.L. and Kotz, S. (1994). Continuous Univariate Distribution Vol. 1. John Wiley and Sons.

[7] Page, E. (1977). Approximations to the cumulative normal function and its inverse for use on a pocket calculator. *Applied Statistics,* **26***, 75–76.*

[8] Lin, J.T. (1989). Approximating the Normal Tail Probability and its Inverse for use on a Pocket Calculator. *Applied Statistics,* **38***, 69–70.*

[9] Lin, J.T. (1990). A Simpler Logistic Approximation to the Normal Tail Probability and its Inverse. *Applied Statistics,* **39***, 255–257.*

[10] Szarek, S.J. (1999). A Nonsymmetric Correlation Inequality for Gaussian Measure. *Journal of Multivariate Analysis,* **68***, 193–211.*

[11] Tocher, K.D. (1963). The Art of Simulation. English University Press, London.

[12] Zelen, M. and Severo, N.C. (1964). Probability Function. *Handbook of Mathematical Functions, Edited by M. Abramowitz and I. A. Stegun, Applied Mathematics Series,* **55***, 925–995.*

Appendix

```
Subroutine NormalCDF(x : Float as input, z : Float as output)
Variables: y, t : Float
z = y = t = 0
Begin Case
  Case: 0 < mod(x) < 0.7315 or 2.2075 < mod(x) < 2.7245 or 3.056 < mod(x)
    y = (exp(-sqr(x)/2))*(sqr(x) + 5.575192695*x + 12.77436324)/
          (power(x,3)*sqrt(2*3.141592654) + 14.38718147*sqr(x)
            + 31.53531977*x + 25.548726)
  Case: 0.7315 < nod(x) <= 1.726 or 1.8135 < mod(x) <= 2.2075
    t = 1/(1 + 0.33267*x)
    y = 1 - sqrt(3.141592654/2)*(exp(-sqr(x)/2))*(0.4361836*t
          - 0.1201676*sqr(t) + 0.9372980*power(t,3))
            Case: 1.726 < mod(x) <= 1.8135 or 2.7245 < mod(x) <= 3.056
    y = 0.5 + 0.5*sqrt((1 - (1/30)*(7*(exp(-sqr(x)/2))
         + 16*(exp((sqrt(2)-2)*sqr(x))+(7 + 3.141592654/4*sqr(x))*exp(-sqr(x)))))
End case
If x < 0
      Then z = 1 - y
Else if x = 0
      Then z = 0.5
Else z = y
Return (z)
End of Subroutine NormalCDF
```