ISSN 2229-5518

A Combined Approximation to t-distribution

Naveen Kumar Boiroju, R. Ramakrishna

Abstract— In this paper, a simple function developed for computing probability values of t-statistics. This function corrects the function proposed by Gleason (2000) with good accuracy and also it provides comprehensive t-statistic probability values without further check of the statistical tables. Probability values for any t-test statistic could be readily obtained from the suggested function and the proposed approximation guarantees atleast three decimal point accuracy, which is more than sufficient to compare the probability value with the level of significance in statistical hypothesis testing.

Index Terms— t-distribution, CDF, Maximum absolute error.

1 Introduction

IT is common knowledge that the t-statistic plays a key role in statistics and is the mostly used statistic in the statistical inference of a population mean or comparison of two population means. Therefore an accurate approximation to its cumulative distribution function (CDF) is very much needed in the statistical hypothesis testing (Jing et al., 2004, Johnson et al., 1995). Two independent variables X and Y such that $X \sim N(0,1)$ and $Y \sim \chi^2_{(n)}$ respectively, the statistic $t = X/\sqrt{(Y/n)}$ is said to have a t-distribution with n degrees of freedom. The probability density function of t-distribution with ν degrees of freedom is given by

$$f(t) = \frac{1}{\sqrt{\nu}\, B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \frac{1}{\left(1+\frac{t^2}{\nu}\right)^{\frac{\nu+1}{2}}}; -\infty < t < \infty \qquad (1)$$

There is no closed form to the CDF of t-distribution which show the way to refer the cumbersome and insufficient statistical tables. Hence, an approximation of CDF could provide the probability values for a t-statistic and often plays a key role in statistical inference. Recently, the approximations of t-distribution function discussed by Yerukala et al. (2013) and their paper motivated us to develop a new approximation function to the CDF of t-distribution. In this paper, an improved function suggested by correcting the Gleason (2000) function, then a new combined approximation discussed for $3 \le \nu \le 30$ and for all $t \ge 0$.

2 Methods

It is well known that the t-distribution is symmetric distribution and tends to follow normal distribution for large degrees of freedom (say n>30). The case t<0 can be handled by symmetry property of the distribution. Gleason (2000) proposed two approximations with two decimal point accuracy.

$$F_1 = F_\nu(t) = \Phi(Z_\nu(t)) \qquad (2)$$

where $\Phi(.)$ is the CDF of standard normal distribution,

$$Z_\nu(t) = \sqrt{\frac{\ln(1+t^2/\nu)}{g(\nu)}} \text{ and } g(\nu) = \frac{\nu-1.5}{(\nu-1)^2}. \qquad (3)$$

The second function defined by Gleason (2000) is given by substituting $g*(\nu) = \frac{\nu-1.5-(0.1/\nu)+0.5825/\nu^2}{(\nu-1)^2}$ in place of $g(\nu)$ in equation (3).

$$F_2 = F_\nu(t) = \Phi(Z_\nu(t)) \text{ with } Z_\nu(t) = \sqrt{\frac{\ln(1+t^2/\nu)}{g*(\nu)}} \qquad (4)$$

We propose a better approximation function by subtracting a non-linear component to the function F_2 and the resulting function is given as

- *Naveen Kumar Boiroju, Department of Statistics, Osmania University, Hyderabad, India. E-mail: nanibyrozu@gmail.com*
- *R. Ramakrishna, Vidya Jyothi Institute of Technology, Post, Aziznagar, Hyderabad, India. E-mail: ramakrishnaraavi9292@gmail.com*

ISSN 2229-5518

$$F_3 = F_\nu(t) = F_2 - [\left(\begin{matrix} 7.9 + \\ 7.9\tanh(3 - 0.63x - 0.52\nu) \end{matrix}\right) / 10000]$$

where $x = \begin{cases} 9 & if\ t = 0 \\ t & otherwise \end{cases}$ (5)

Li and Moor (1999) suggested a natural modification of the ordinary normal approximation to t-distribution.

$$F_4 = F_\nu(t) = \Phi(Z_\nu(t)),$$

where $Z_\nu(t) = t(4\nu + t^2 - 1)/(4\nu + 2t^2)$ (6)

A combined function defined based on the errors of these functions as

$$F_5 = \begin{cases} F_4 \ ; & 0 \le t < 1.3 + 0.04\nu \\ F_3 \ ; & 1.3 + 0.04\nu \le t < 5.94 - 0.04\nu \\ F_1 \ ; & t \ge 5.94 - 0.04\nu \end{cases} \quad (7)$$

The efficiency of these functions measured using the minimum of maximum absolute error and the error is computed as the difference between the probability of the given function and with that of the TDIST() function available in Microsoft office Excel 2007 software.

3 Results and Discussion

The maximum absolute error of these functions observed at 3 degrees of freedom and the Figure 1 presents the absolute errors of the functions at 3 degrees of freedom. It is evident that the corrected function and combined function has lowest absolute errors as compared with other approximations. At 3 degrees of freedom, Function F_4 has the maximum absolute error 0.0069818 observed at t=3.8, function F_1 has the maximum absolute error 0.0049514 observed at t=1 and the function F_2 has the maximum absolute error 0.0025012 observed at t=0.9. The corrected function F_3 has the maximum absolute error 0.0011699 observed at t=1 and the combined function (F_5) has the maximum absolute error 0.0008117 observed at t=1.5. The proposed combined function also accurate to the three decimal points as like of the functions defined in Yerukala et al. (2013). It is also observed that the proposed functions performing well at the tail probabilities.

Atleast two decimal point accuracy is obtained at 3 and 4 degrees of freedom for the functions F_1, F_2 and F_3 where as the function F_4 has the same accuracy for the degrees of freedom between 3 and 5. The function F_1 provides the three decimal point accuracy when the degrees of freedom lie in between 5 and 12 whereas the functions F_2 and F_3 provide atleast three decimal point accuracy for the degrees of freedom lie in between 5 and 14. The function F_4 provides three decimal value accuracy for degrees freedom from 6 to 11 whereas the function F_5 gives the same accuracy for degrees of freedom from 3 to 9. The four decimal point accuracy for the function F_1 is obtained for degrees of freedom from 13 to 30, for the functions F_2 and F_3, it is obtained for the degrees of freedom from 15 to 30. The function F_4 gives four decimal point accuracy when the degrees of freedom from 12 to 21 whereas the same is observed for the function F_5 in between 10 to 21 degrees of freedom. Only two functions F_4 and F_5 provide the accuracy up to five decimal points when the degrees of freedom are greater than or equal to 22. From the Table 1, it is observed that the proposed combined function F_5, guaranty the three decimal point accuracy and it may be treated as a competitor for the functions proposed by Yerukala et al. (2013).

4 Conclusion

The proposed combined function (F_5) guaranties the accuracy up to three decimal points to the CDF of t-distribution where as the corrected function F_3 is the efficient function as compared with the other two functions at lower degrees of freedom (Table 1). The function F_5 is better than the functions F_1, F_2, F_3 and F_4 for all $\nu \le 30$. The accuracy of F_4 and F_5 is almost equivalent for all $\nu > 16$. The functions F_1 and F_2 are better than the function F_4 for all $\nu < 8$ and F_1 is better than the functions F_2 and F_3 for all $\nu > 5$. The accuracy of the functions F_2 and F_3 are same for all $\nu > 11$. The proposed two functions are guarantying the accuracy up to three decimal points at the tails of the distribution and it is more than sufficient in the testing of hypothesis using t-statistics.

ISSN 2229-5518

TABLE 1
Maximum absolute errors of the approximations

df	Gleason (2000)-F_1	Gleason (2000)-F_2	Corrected Model-F_3	Li & Moor (1999)-F_4	Combined Model-F_5
3	0.004951	0.002501	0.001170	0.006982	0.000812
4	0.001901	0.001370	0.001080	0.003216	0.000413
5	0.000984	0.000874	0.000880	0.001659	0.000326
6	0.000595	0.000608	0.000531	0.000931	0.000214
7	0.000396	0.000448	0.000323	0.000557	0.000203
8	0.000283	0.000344	0.000296	0.000351	0.000148
9	0.000212	0.000273	0.000255	0.000230	0.000122
10	0.000165	0.000221	0.000215	0.000156	0.000083
11	0.000132	0.000183	0.000181	0.000109	0.000068
12	0.000108	0.000154	0.000154	0.000077	0.000056
13	0.000089	0.000132	0.000132	0.000056	0.000040
14	0.000076	0.000114	0.000114	0.000041	0.000032
15	0.000065	0.000099	0.000099	0.000030	0.000027
16	0.000056	0.000087	0.000087	0.000022	0.000022
17	0.000049	0.000078	0.000078	0.000018	0.000018
18	0.000043	0.000069	0.000069	0.000015	0.000015
19	0.000038	0.000062	0.000062	0.000013	0.000013
20	0.000034	0.000056	0.000056	0.000011	0.000011
21	0.000031	0.000051	0.000051	0.000010	0.000010
22	0.000028	0.000047	0.000047	0.000008	0.000008
23	0.000025	0.000043	0.000043	0.000008	0.000008
24	0.000023	0.000039	0.000039	0.000007	0.000007
25	0.000021	0.000036	0.000036	0.000007	0.000007
26	0.000019	0.000033	0.000033	0.000006	0.000006
27	0.000018	0.000031	0.000031	0.000006	0.000006
28	0.000017	0.000029	0.000029	0.000005	0.000005
29	0.000015	0.000027	0.000027	0.000005	0.000005
30	0.000014	0.000025	0.000025	0.000004	0.000004

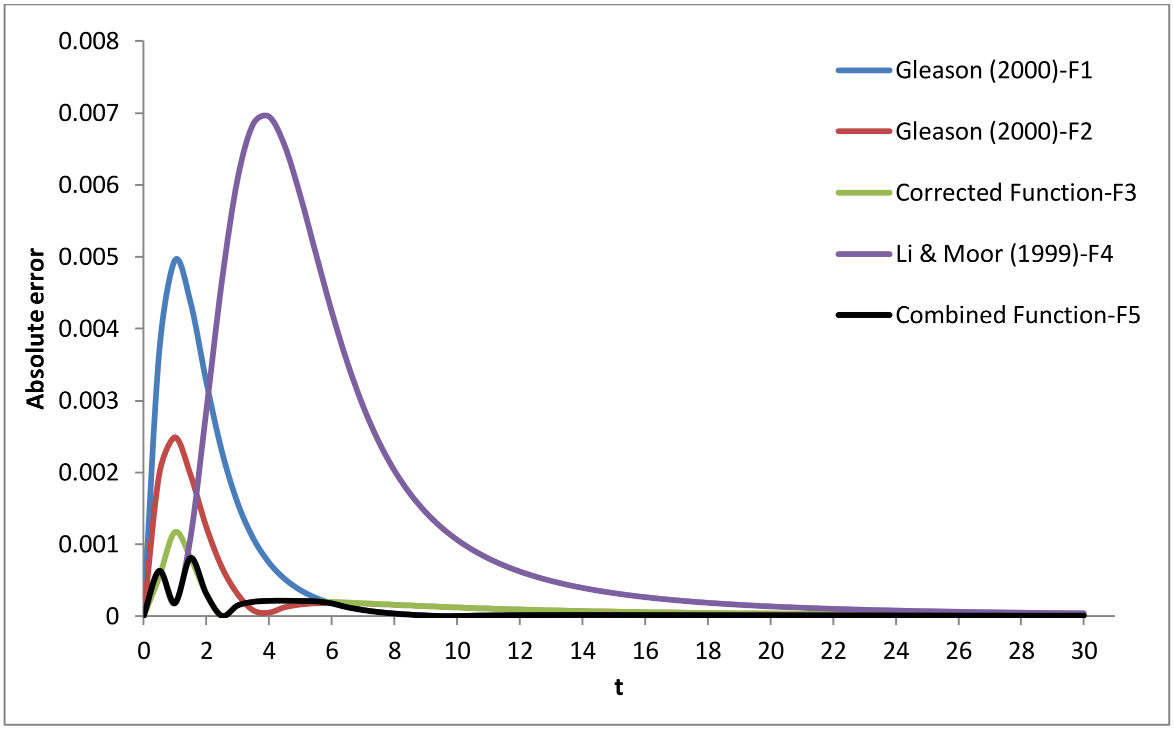

Fig. 1. Maximum absolute error of the approximations for 3 degrees of freedom

REFERENCES

[1] B.Y. Jing, Shao, Q.M. and Zhou, W. Saddle-point approximation for student's t-statistic with no moment conditions, *The Annals of Statistics*, 32 (6), pp2679-2711, 2004.

[2] N.L. Johnson, Kotz, S. and Balakrishnan, N., *Distributions in Statistics: Continuous Univariate Distributions*, Vol. 2, Second edition, New York. Wiley, 1995.

[3] R. Yerukala, Boiroju, N.K. and Reddy, M.K., Approximations to the t-distribution, *International Journal of Statistika and Mathematika*, Vol. 8 (1), pp19-21, 2013.

[4] J.R. Gleason, A note on a proposed student t approximation, *Computational Statistics & Data Analysis*, 34, pp63-66, 2000.

[5] B. Li and Moor, B.D., A corrected normal approximation for the Student's t distribution, *Computational Statistics & Data Analysis*, 29, pp213-216, 1999.