Iraqi Journal of Statistical Sciences

www.stats.mosuljournals.com

New Approach to Approximating the Cumulative Function for the t-Distribution

A.N. SH. AL-SHALLAWI

Department of Statistics & Informative techniques, Northern Technical University, Mosul, Iraq.

Article information

Article history:
Received June 8 ,2023
Accepted August 13,2023
Available online December 1,2023

Keywords:
Cumulative distribution function,
Polya's formula,
t-distribution

Correspondence:
A N SH. AL-SHALLAWI
mohammedrahawy@yahoo.com

Abstract

The focus of this paper is to approximate the cumulative distribution function (CDF) of the t distribution, which represents a combined distribution of the normal distribution and gamma distribution. The study utilizes the approximate formula proposed by Polya for the normal distribution, originally introduced in 1945. By applying this final formula to various points and comparing the results with the tabulated values of the t distribution, the researchers found that the absolute error between the two sets of values is negligible. It should be noted that this error slightly increases with higher degrees of freedom. Furthermore, the study observed that the absolute errors remain consistent when multiple points are selected at the same degrees of freedom. These findings have practical implications for statistical analysis, as they offer a time and effort-saving approach for obtaining CDF values associated with the t-distribution.

DOI: https://doi.org/10.33899/iqjoss.2023.0181184

Introduction

T-distribution is one of the important distributions that have a role in statistical analysis. It is very similar to the normal distribution and is used instead if the population standard deviation is unknown or when the sample size is less than 30. [1]

One of the characteristics of this distribution is that it has a symmetrical frequency curve around the mean, but it is heavier at the edges than the normal curve. There is one parameter that determines the shape of the distribution, which is called degrees of freedom. [2]

The t distribution is a special case of the generalized hyperbolic distribution [3]. It is important to use it in many fields, such as t-test is used to find significant differences between sample means, and to find the confidence intervals to the difference between two population means, also it is used in linear regression and Bayesian analysis. [4].

The main aim of this research is to compute the cumulative function to t-distribution as an approximation in a simple way for obtaining an easy formula more than the old one which depended to a hypergeometric (as will be mentioned) and more difficult in practical application.

Material and methods

Theoretical Part

Materials and Methods

The probability density function for t- distribution is :

$$f(t) = \frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\,\Gamma\left(\frac{v}{2}\right)}\left(1+\frac{x^2}{v}\right)^{-\frac{v+1}{2}} \qquad -\infty < t < \qquad (1)$$

Whereas:

v: degree of freedom, it is a positive integer.

Γ : Complete gamma function

Now, we can define a cumulative function as a function that determines what is the probability that the value of any random variable (T) is less than or equal to a given value. It is written as [1] :

$$F(t) = \int_{-\infty}^{t} f(u)du = \frac{1}{2} + t\ \frac{\Gamma(\frac{1}{2}(v+1))}{\sqrt{\pi v}\,\Gamma(\frac{v}{2})}\ 2F_1\left(\frac{1}{2},\frac{1}{2}(v+1);\frac{3}{2};\frac{t^2}{v}\right) \qquad (2)$$

Where 2F1 is a special case of the hyper geometric function [5].

Review of approximations to the CDF.

Many formula to the approximation for CDF t-distribution proposed, [6] gave a list of various approximation to cumulative function for t-distribution and proposed a simple approximation of F(x;v) as :

$$F_1(x,v) \approx \Phi(x\lambda) for\ v \geq 3$$

$$with\ \lambda = \lambda(x,v) = \frac{(4v+x^2-1)}{(4v+2x^2)} \qquad 0 < \lambda < 1 \qquad (3)$$

for n=1 and 2, they have suggested the following exact formulae:

$$F(x,1) = \frac{1}{2} + \frac{1}{\pi}\tan^{-1}x \text{ and } F(x,2) = \frac{1}{2} + \frac{x}{2}\frac{1}{\left(2+x^2\right)^{\frac{1}{2}}}$$

And he gave an absolute error for many values to v and x.

[6] proposed a new formula to compute CDF for t-distribution by using neural networks when $x \geq 0\ and\ v \geq 3$, depending to the absolute error and compute approximation accuracy for $F_1(x,v)$, $F_2(x,v), F_3(x,v), F_4(x,v), F_5(x,v)$, he compared among them to choose a minimum absolute error .

Mixed Model:

Mixed model or compound distributions are one of the important distributions in modelling many phenomena because these phenomena are more flexible than standard distributions, and many researchers have been interested in studying this type of distributions, whether they are continuous or discontinuous.

It is used to represent some data that cannot be represented by standard statistical distributions as required because the nature of these data or phenomena necessitates the use of mixed distributions that are more flexible than standard distributions [7]. Therefore, the mixed model is a model that consists of two or more probability distributions [8], and it should be noted that it is not necessary that these mixed distributions belong to the same family [9]. If we have Z and Y as two independent variables, Z is normally distributed with mean = 0 and variance σ2 = 1, y has a chi-square distribution with n degrees of freedom, then:

$$t = \frac{Z}{\sqrt{y/n}} \qquad (4)$$

According to references [10], the t-distribution is characterized by its degrees of freedom "n". It can be represented as a mixed distribution, providing more versatility than the standard form. This mixed

(4 a)

distribution of the t-distribution combines elements of the normal distribution and the inverse gamma distribution, allowing for a broader range of applications. The random variable Z is normally distributed with certain arithmetic mean and variance, as follows:
X ~ t (μ, σ2, v)

X can be expressed as:

$x = \mu + \sqrt{\tau}\, z$

Whereas:

τ ~ IG $(\frac{v}{2}, \frac{v}{2})$

Z~ N (0, σ2)

Since z is independent of τ, and the variable τ has positive values ($\tau > 0$) and has a probability density function as follows[11]:

$$f(\tau|\alpha,\beta) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}\tau^{-(\alpha+1)}e^{\frac{-\beta}{\tau}} \quad (5)$$

This formula is used in statistical modelling of the t-distribution in conventional statistics and Bayesian statistics [12]

Approximation of cumulative function to t-distribution

In this part, we will propose a new approximation for a cumulative function to t-distribution by using a mixed model between normal distribution and gamma distribution as (4a) equation

The cumulative function for t distribution is:

$$F(x) = \int_a^b t(x|\mu, \sigma^2, v)dx \quad (6)$$

Above, we say that we can represent t-distribution as a mixed distribution

as:

$$F(x) = \int_a^b N(x|\tau,\mu,\sigma^2)\left[\int_0^{\infty} g(\tau)d\tau\right]dx \quad (7)$$

Whereas:

$N(x|\tau,\mu,\sigma^2)$: is the conditional normal distribution.

$g(\tau)$: is the gamma function.

We will use Polya's formula to find the solution for the Normal distribution function (Hermuz,1990):

$$F(z) = \frac{1}{2}\left[1 + \left(1 - e^{-\frac{2z^2}{\pi}}\right)^{1/2}\right] \qquad \infty \le z \le \infty \quad (8)$$

As the approximation formula, which will be updated, relies on the mixed distribution, we will utilize the poly(a) formula in equation (8) as a mixed distribution along with the gamma distribution. In this context, the variable Z will be conditioned on the variable τ, which follows a t-distribution according to equation (4a), as follows:

$$p(a < z|\tau < b) = p\left(\frac{a-\mu}{\sigma\,\tau^{1/2}} < z|\,\tau < \frac{b-\mu}{\sigma\,\tau^{1/2}}\right)$$

$$= F(b^* \mid \tau) - F(a^* \mid \tau) \quad (9)$$

Where:

$$b^* = \frac{b - \mu}{\sigma\, \tau^{1/2}} \quad (10)$$

$$a^* = \frac{a - \mu}{\sigma\, \tau^{1/2}} \quad (11)$$

$$\therefore F(z|\tau) = \frac{1}{2} + \frac{1}{2}\left(1 - e^{-\frac{2z^2}{\pi}}\right)^{1/2}$$

$$= \frac{1}{2} + \frac{1}{2}\, g(z \mid \tau) \quad (12)$$

Where: $g(z \mid \tau) = \left(1 - e^{-\frac{2z^2}{\pi}}\right)^{1/2}$

By using Tylor 's series as:

$$g(z) = g(z_0) + g'(z_0)(z - z_0) + g''(z_0)\frac{(z - z_0)^2}{2!} \ldots \quad (13)$$

$$g(z_0) = \left(1 - e^{-\frac{2z_0^2}{\pi}}\right)^{1/2} \quad (14)$$

$$g'(z_0^2) = \frac{2z}{\pi}\, e^{-\frac{2z_0^2}{\pi}} \left(1 - e^{-\frac{2z_0^2}{\pi}}\right)^{-1/2} \quad (15)$$

$$g''(z) = \frac{2}{\pi} e^{-\frac{2z_0^2}{\pi}} \left(1 - e^{-\frac{2z^2}{\pi}}\right)^{-1/2} \left(1 - \frac{4z^2}{\pi}\right) - \frac{4z^2}{\pi}\, e^{-\frac{4z^2}{\pi}} \left(1 - e^{-\frac{2z^2}{\pi}}\right)^{-3/2} \quad (16)$$

We will put $g''(z) = c$

We substitute (14), (15) and (16) in (13), and we get:

$$\therefore g(z) = \left(1 - e^{-\frac{2z^2}{\pi}}\right)^{1/2} + \frac{2}{\pi} z e^{-\frac{2z^2}{\pi}} \left(1 - e^{-\frac{2z^2}{\pi}}\right)^{-1/2} (z - z_0) + c\, \frac{(z - z_0)^2}{2} + \cdots \quad (17)$$

But:

$$F(z|\tau) = \frac{1}{2} + \frac{1}{2}\, g(z \mid \tau)$$

Then:

$$F(a^*|\tau) = \frac{1}{2} + \frac{1}{2}\, g(a^* \mid \tau)$$

$$F(b^*|\tau) = \frac{1}{2} + \frac{1}{2}\, g(b^* \mid \tau)$$

$$\therefore g(a^* \mid \tau) = \frac{1}{2} + \frac{1}{2}\left[\left(1 - e^{-\frac{2z^2}{\pi}}\right)^{1/2} + \frac{2}{\pi} z e^{-\frac{2z^2}{\pi}} \left(1 - e^{-\frac{2z^2}{\pi}}\right)^{-1/2} (a^* - z_0) + c\, \frac{(a^* - z_0)^2}{2} + \cdots\right] \quad (18)$$

and:

$$\therefore \boldsymbol{g}(\boldsymbol{b}^{*}|\boldsymbol{\tau}) = \frac{\mathbf{1}}{\mathbf{2}} + \frac{\mathbf{1}}{\mathbf{2}}\left[\left(\mathbf{1} - \boldsymbol{e}^{-\frac{2z^2}{\pi}}\right)^{1/2} + \frac{\mathbf{2}}{\boldsymbol{\pi}}\boldsymbol{z}\boldsymbol{e}^{-\frac{2z^2}{\pi}}\left(\mathbf{1} - \boldsymbol{e}^{-\frac{2z^2}{\pi}}\right)^{-1/2}(\boldsymbol{b}^{*} - \boldsymbol{z_0}) + \boldsymbol{c}\,\frac{(b^{*}-z_0)^2}{\mathbf{2}} + \cdots\right] \quad (19)$$

We put (18) and (19) in (9) we get:

$$F(b^{*}|\tau) - F(a^{*}|\tau) = \frac{-2}{\pi}\,e^{-\frac{2}{\pi}}\left(1 - e^{-\frac{\pi}{2}}\right)^{-1/2}(b^{*} - a^{*}) + \frac{c}{2}[(b^{*} - a^{*})(b^{*} + a^{*} - 1)] \quad (20)$$

We substitute (10) and (11) in (20) and we integrate the expression with respect to τ , we get the marginal cumulative function :

$$F(b^{*}) - F(a^{*}) = \int_0^{\infty}\left(F(b^{*}|\tau) - F(a^{*}|\tau)\right)g(\tau)d\tau$$
$$= \frac{-\mathbf{2}}{\boldsymbol{\pi}}\,\boldsymbol{z}\boldsymbol{e}^{-\frac{2z^2}{\pi}}\left(\mathbf{1} - \boldsymbol{e}^{-\frac{z^2}{\pi}}\right)^{-1/2}\left(\frac{\boldsymbol{b}-\boldsymbol{a}}{\boldsymbol{\sigma}}\right)\frac{\left(\frac{\nu}{2}\right)^{\left(\frac{1}{2}\right)}\Gamma\left(\frac{\nu-1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} + \frac{\boldsymbol{c}}{\mathbf{2}\boldsymbol{\sigma}^{2}}(\boldsymbol{b} + \boldsymbol{a} - \mathbf{2}\boldsymbol{\mu} - \boldsymbol{\sigma}(\boldsymbol{z_0}))\frac{\left(\frac{\nu}{2}\right)}{\left(\frac{\nu}{2}-\mathbf{1}\right)} \quad (21)$$

Where:

$$\int_0^{\infty}\tau^{\left(-\frac{1}{2}\right)}\tau^{\left(\frac{\nu}{2}-1\right)}e^{\left(-\frac{\nu}{2}\tau\right)}d\tau = \int_0^{\infty}\tau^{\left(\frac{\nu-1}{2}-1\right)}e^{\left(-\frac{\nu}{2}\tau\right)}d\tau = G\left(\alpha = \frac{\nu-1}{2}, \beta = \frac{\nu}{2}\right) = \frac{\Gamma(\alpha)}{\beta^{\alpha}} = \frac{\Gamma\left(\frac{\nu-1}{2}\right)}{\left(\frac{\nu}{2}\right)^{\frac{\nu-1}{2}}}$$

and:

$$\int_0^{\infty}\tau^{(-1)}\tau^{\left(\frac{\nu}{2}-1\right)}e^{\left(-\frac{\nu}{2}\tau\right)}d\tau = \int_0^{\infty}\tau^{\left(\frac{\nu-2}{2}-1\right)}e^{\left(-\frac{\nu}{2}\tau\right)}d\tau = \frac{\Gamma\left(\frac{\nu-2}{2}\right)}{\left(\frac{\nu}{2}\right)^{\frac{\nu-2}{2}}} = \frac{\Gamma\left(\frac{\nu}{2}-1\right)}{\left(\frac{\nu}{2}\right)^{\left(\frac{\nu}{2}-1\right)}}$$

$$\Gamma\left(\frac{\nu}{2}\right) = \left(\frac{\nu}{2}-1\right)\Gamma\left(\frac{\nu}{2}-1\right)$$

Then (21) is the final formula to approximation of the cumulative function for t- distribution.

Practical Side

As an application of equation (21), we follow the following algorithm:

1- Determine the degree of freedom.

2- We choose two values for a and b, we wanted to choose standard values to apply the equation (21), such as (1, -1).

2- After selecting these two values, we choose a value for z. (negative and positive values)

3- We get the value of the cumulative function between (a,b).

4- We compare the resulting value in (3) with the tabular value which obtained from the statistical tables, and find the difference between them.

5- If the difference is large, we choose another value for z and recalculate the algorithm until we get a value close to the tabular value.

Thus, whenever we choose values for a and b, or different values for the degrees of freedom, we recalculate the value of the cumulative function.

The researcher concluded that the value of z which $\left(\frac{4}{\sqrt[2]{\pi}}\right)$ is the one that achieved the best result of equation (21) if we compare it with the tabular value,at (-1 , 1) , (-1 , 0) , (0 , 1) with many degrees of freedom.

The algorithm was applied using (Matlab,2020a).

The Algorithm:

As an application of equation (21), we follow the steps below:

1. Determine the degrees of freedom.
2. Choose two values for "a" and "b," and for our implementation, we have opted for standard values such as (1, -1).
3. After selecting these values, choose a value for "z" (positive and negative values).
4. Calculate the cumulative function's value between (a, b).
5. Compare the result obtained in step (4) with the tabulated value obtained from statistical tables and find the difference between them.
6. If the difference is significant, choose another value for "z" and recalculate the equation until obtaining a value close to the tabulated value.
7. Repeat the process for different values of "a," "b," and degrees of freedom to recalculate the cumulative function's value.

Through this procedure, we found that the value (0.3334) yielded the best results with the least possible difference when compared to the tabulated value. Therefore, the value (0.3334) achieved the best outcome for equation (21) when compared to the tabulated value.

The results are as Table (1) :

Table (1) The comparative between (CDF original) and (CDF approximation)

DF	CDF approximation	CDF original	AE
(-1 , 1)			
10	0.6561	0.6591	0.003
15	0.6639	0.6668	0.0029
20	0.6675	0.6707	0.0032
25	0.6695	0.6731	0.0036
(-1 , 0)			
10	0.3280	0.3296	0.0016
15	0.3320	0.3334	0.0014
20	0.3337	0.3354	0.0017
25	0.3348	0.3366	0.0018
(0 , 1)			
10	0.3280	0.3296	0.0016
15	0.3320	0.3334	0.0014
20	0.3337	0.3354	0.0017
25	0.3348	0.3366	0.0018

The first column in Table (1) represents different values of the degrees of freedom that were chosen, the second column represents the results of the proposed equation, Equation (21), and the third column represents the tabular cumulative value taken from the statistical tables. As for the fourth column, it represents the absolute difference (AE) between the value according to the proposed equation and the tabular value.

It is noted from Table (1) that the tabular value is close to the value of t according to equation (21), and it is noted that the error value is equal in the tested truncation points (-1,0) and (0,1) , This means that the value of the cumulative function on the left side is equal to the value of the cumulative function on the right side.

4- Conclusion:

In this research, a new approximation was found for the CDF for t distribution, and this value was compared with the original value of the (CDF) by using (Matlab, 22), we concluded that:

1- It was noted that the results are close and the absolute error is very small.
2- The absolute error value is equal for the different truncation points that were used in the test.
3- Also, the value of cdf at points (-1,0) is the same as its value at truncation points (0, 1).

References

[1] Kotz S, Nadarajah S. (2004):" Multivariate t-distributions and their applications", Cambridge University Press; 2004.

[2] Hurst S. (1995):" The characteristic function of the student t distribution. Centre for Mathematics and its Applications", School of Mathematical Sciences; 1995.

[3] Frhr.Ernst August v. Hammerstein (2010): " Generalized hyperbolic distributions: Theory and applications to CDO pricing", Department of Mathematical Stochastics, Faculty of Mathematics and Physics. Albert-Ludwigs-University Freiburg. German.

[4] Nadarajah S, Kotz S. (2008):" Estimation methods for the multivariate t distribution", Acta Applicandae Mathematicae. 2008;102(1):99-118.

[5] Bagdasaryan A. (2009):" A note on the 2F1 hypergeometric function"m arXiv preprint arXiv:09120917. 2009.

[6] Yerukala R, Boiroju NK, Reddy MK. (2013):" Approximations to the t-distribution", International Journal of Statistika and Mathematika. 2013;8(1).

[7] Nascimento A, Rêgo LC, Silva JW. (2022):" Compound truncated Poisson gamma distribution for understanding multimodal SAR intensities", Journal of Applied Statistics. 2022:1-20.

[8] Booth JG, Casella G, Friedl H, Hobert JP.(2003):" Negative binomial loglinear mixed models", Statistical Modelling. 2003;3(3):179-91.

[9] Garcia V, Nielsen F.(2010):" Simplification and hierarchical representations of mixtures of exponential families", Signal Processing. 2010;90(12):3197-212.

[10] Weisstein EW.(2001):" Student's t-Distribution". https://mathworld wolfram com/. 2001.

[11] Arellano-Valle RB, Bolfarine H.(1995):" On some characterizations of the t-distribution", Statistics & Probability Letters. 1995;25(1):79-85.

[12] Arellano-Valle RB, Castro LM, González-Farías G, Muñoz-Gajardo KA.(2012):"Student-t censored regression model: properties and inference", Statistical Methods & Applications. 2012;21(4):453-73.

[13] Hermuz, Amir Hanna (1990). "Mathematical Statistics", Directorate of Printing and Publishing House, University of Mosul.

اسلوب جديد لتقريب الدالة التراكمية لتوزيع T

احمد نجم شيت الشلاوي

الكلية التقنية الادارية– الموصل، الجامعة التقنية الشمالية

الخلاصة : فكرة البحث هي ايجاد تقريب دالة التوزيع التراكمي لتوزيع t ، والتي تمثل توزيعًا مشتركًا للتوزيع الطبيعي وتوزيع جاما. وباستخدام الصيغة التقريبية التي اقترحها Polya للتوزيع الطبيعي ، والتي تم تقديمها في الأصل عام 1945. من خلال تطبيق هذه الصيغة النهائية على نقاط مختلفة ومقارنة النتائج مع القيم المجدولة لتوزيع t ، تبين أن الخطأ المطلق بين قيم المجموعتين لا يكاد يذكر. وتجدر الإشارة إلى أن هذا الخطأ يزداد زيادة قليلة جداً مع زيادة درجات الحرية. علاوة على ذلك ، لاحظت الدراسة أن الأخطاء المطلقة تظل ثابتة عند اختيار نقاط متعددة بنفس درجات الحرية. هذه النتائج لها آثار عملية على التحليل الإحصائي ، لأنها توفر نهجًا لتوفير الوقت والجهد للحصول على قيم دالة التوزيع التراكمي المرتبطة بتوزيع t.

الكلمات الدالة: دالة الكتلة الاحتمالية، صيغة بوليا، توزيع t