

# Mean Value Estimation Using Low Size Samples Extracted from Skewed Populations

JOÃO PAULO MARTINS<sup>1,3</sup>, MIGUEL FELGUEIRAS<sup>2,3</sup>, RUI SANTOS<sup>1,3</sup>

<sup>1</sup>School of Health, P. Porto, Porto, PORTUGAL

<sup>2</sup>School of Technology and Management, Polytechnic Institute of Leiria, Leiria, PORTUGAL

<sup>3</sup>CEAUL – Center of Statistics and Applications, Faculdade de Ciências, Universidade de Lisboa, Lisboa PORTUGAL

*Abstract:* - The use of the  $T$ -statistic in statistical inference procedures is usually restricted to normal populations or to large samples. However, these conditions may not be fulfilled in some situations: the population can be moderate/highly skewed, or the sample size can be small. In this work, we use the Pearson's system of distributions, namely, type IV distributions to model  $T$ . By some simulation work, it is shown that this approximation leads to confidence intervals whose coverage is close to the nominal coverage even for low sample sizes.

*Key-Words:* -  $T$ -statistic, Type IV distributions, Pearson's system, skewness, kurtosis, estimation, confidence interval, coverage, mean value, simulation

Received: May 20, 2021. Revised: January 23, 2022. Accepted: February 18, 2022. Published: March 17, 2022.

## 1 Introduction

Let  $X_1, \dots, X_n$  be a random sample with mean  $\bar{X}$  and standard deviation  $S$  drawn from a population  $X$  with finite mean  $\mu$  and standard deviation  $\sigma$ . The study of the distribution of the ratio

$$T = \sqrt{n} \frac{\bar{X} - \mu}{S} \quad (1)$$

under an underlying Normal distribution, presented by Gosset [1] (under the pseudonym Student) was one of the seeds of the development of Statistical Inference. However, the potentialities of using  $T$  have not (surprisingly?) been exploited outside the comfort of the Normal distribution or the scope of the Central Limit Theorem (CLT).

At the beginning of the last decade, [2] derived the first four moments of the Student's  $T$ -Statistic for any underlying population with finite first four moments. The derived approximations for all these four moments depend only on two measures: skewness and kurtosis. Skewness is defined as

$$\gamma_1(X) = \frac{E(X-\mu)^3}{\sigma^3} \quad (2)$$

and kurtosis is given by

$$\gamma_2(X) = \frac{E(X-\mu)^4}{\sigma^4} - 3. \quad (3)$$

For a Normal distribution both measures are equal to zero.

Using Delta method (cf. [3]), [2] derived expressions for the first four moments of  $T$  that are describe in expressions (4) to (7).

$$E(T) = -\frac{\gamma_1(X)}{\sqrt{n}} + O(n^{-3/2}) \quad (4)$$

$$E(T^2) = 1 + 2 \frac{\gamma_1^2(X)+1}{n} + O(n^{-2}) \quad (5)$$

$$E(T^3) = -\frac{7\gamma_1(X)}{2\sqrt{n}} + O(n^{-3/2}) \quad (6)$$

$$E(T^4) = 3 + 2 \frac{14\gamma_1^2(X)-\gamma_2(X)+9}{n-1} + O(n^{-2}) \quad (7)$$

Surprisingly, the first three moments estimates only depend on  $\gamma_1(X)$ .

From equations (4) to (7), clearly, as  $n$  increases the importance of skewness decreases. When  $n$  is large, Slutsky's theorem [4] allows the application of the Student's  $T$ -statistic to non-normal populations. However, statistical inference for small sample sizes may not be possible if the underlying distribution is not symmetric. This is also clear when we compute the first-order Edgeworth expansion of  $T$  (Edgeworth expansions are a particular case of the well-known Gram-Charlier series which allows to write a distribution function of some variable from a well-known one, usually

the Normal standard distribution). Let  $F$  be the cumulative distribution function (cdf) of  $T$ . Then,

$$F_T(x) = \Phi(x) - \frac{\gamma_1(X)\Phi^{(3)}(x)}{6\sqrt{n}} + O(n^{-1}) \quad (8)$$

where  $\Phi^{(3)}(x)$  is the third-order derivative of the cdf of a standard normal distribution, cf [5].

Pearson's system of distributions is a partition of the set of all distributions with finite first four moments cf. [6,7] whose probability/probability density function  $f$  satisfies the following differential equation:

$$[\ln f(x)]' = -\frac{x+a}{b_0+b_1x+b_2x^2} \quad (9)$$

where  $a, b_0, b_1$  and  $b_2$  are distribution parameters.

The solutions of equation (9) are divided into seven groups known as Pearson's type of distributions that range from I to VII.

Multiplying equation (9) by  $x^r$  (with  $r \in \{0,1,2,3\}$ ) and integrating it, it is possible to derive the relation between the four parameters and the first four raw moments [8]. [4] provide a more comprehensive overview on this subject. However, it should be noted that the expression presented by [4] for the coefficient  $b_1$  contains an inaccurate. Where it should be  $\gamma_1(X)$ , it appears  $\sqrt{\beta_1}$  where  $\beta_1 = \gamma_1^2(X)$ , which of course is not equivalent.

[4] performs the partition of the distributions considering the combination of parameters

$$k = \frac{b_1^2}{4b_0b_2} \quad (10)$$

The case  $0 < k < 1$  corresponds to the type IV distributions.

Let  $\Gamma(\cdot)$  stand for gamma function and consider the beta function

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)} \quad (11)$$

where  $a > 0$  and  $b > 0$ .

The probability density function is given by

$$f(x) = \frac{\left| \frac{\Gamma(m+\frac{\nu}{2}i)}{\Gamma(m)} \right|^2}{\alpha B\left(m - \frac{1}{2}, \frac{1}{2}\right)} \times \quad (12)$$

$$\times \left(1 + \left(\frac{x-\lambda}{\alpha}\right)^2\right)^{-m} \exp\left(-\nu \tan^{-1} \frac{x-\lambda}{\alpha}\right)$$

where  $m > 1/2$  and  $\nu > 0$ .

None of the most used distributions in Statistics verifies the density function defined in (12), i.e., none is a type IV distribution. Nevertheless, it is possible to identify in literature examples where these distributions are used to model real life problems (cf. [9,10]).

Several researchers have dedicated some of their attention to this type of distributions. [11] determined several values of type IV distribution functions. Later, [12] have constructed an algorithm for determination of some quantiles that can be applied to any of the types of distributions of the Pearson system. [13] analyzed the moments of type IV distributions. More recently, [14-16] determined approximate expressions for a type IV distribution function. Details of packages/macros developed to allow the use of type IV distributions can be found in [17] for R software and in [18] for SAS software.

Under broader conditions, [1] showed that  $F_T$  is a type IV distribution if  $X$  is non-symmetric, i.e.,  $\gamma_1(X) \neq 0$ . It is clear from equation (8) that as  $n$  increases  $F_T$  gets close to a normal distribution. Until recently, there was no closed form expression for the cdf of a type IV distribution. Moreover, this type of distribution depends on four parameters. This is important because if the sample size is high, it is possible to apply the CLT and if the size is low it may not be reasonable to estimate all four parameters. These two issues may help to explain the little of use of this kind of distributions in statistical inference. However, it is now possible to use software to easily compute probabilities or quantiles. Hence, in the recent years some applications involving the use of type IV distributions can be found. For instance, its use in econometric modelling [19] or in operating room management [20].

In practice, the problem of fitting a type IV distribution to  $T$  is that it requires finite first four moments of the underlying distribution and estimates of its skewness and kurtosis. In the context of a samples with low sizes this may be a challenge as previously discussed. Moreover, it is not clear if there is any advantage of using a type IV distribution instead of just using a Normal distribution even for small sample sizes.

In this work, we address the problem of using the  $T$ -statistic in small samples and skewed populations to perform statistical inference. A Bayesian approach to this matter can be found in [21].

The outline of this work is as follows. In Section 2, two confidence intervals are presented for the mean value of a population: one based in the Normal approximation of  $F_T$  and the other based on

the approximation to a type IV distribution. In Section 3, simulation work is presented to compare the two confidence intervals. In two of the studied cases studied, it also addressed the situation where the skewness can be written using the mean value. In the last Section, the obtained results are discussed.

## 2 Confidence Intervals for the Mean

Given a random sample  $X_1, \dots, X_n$  drawn from a population  $X$  with finite variance, the application of the CLT allows us to obtain the following confidence interval for the mean value  $\mu$  of  $X$ :

$$\left[ \bar{X} - z_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{X} + z_{\alpha/2} \frac{s}{\sqrt{n}} \right] \quad (13)$$


where  $z_\alpha$  is the quantile  $1 - \alpha$  of a standard normal distribution. This confidence interval is widely used. When the underlying population is normal the  $t$ -Student with  $n - 1$  degrees of freedom should be used if  $n$  is not large [4].

In a similar way, it is possible for small sample sizes to replace the quantiles used in (13) when dealing with skewed populations. Approximating  $F_T$  by a type IV distribution that verifies the estimates defined by equations (4) to (7), we get an alternative confidence interval for  $\mu$ :

$$\left[ \bar{X} - q_{1-\alpha/2} \frac{s}{\sqrt{n}}, \bar{X} + q_{\alpha/2} \frac{s}{\sqrt{n}} \right] \quad (14)$$

where  $q_\alpha$  is the quantile  $1 - \alpha$  of a type IV distribution. Clearly, this interval is no longer symmetric. In practice, the application of that interval requires some knowledge about the population kurtosis  $\gamma_2(X)$  and especially the population skewness  $\gamma_1(X)$ . Thus, its usefulness in practice needs to be assessed by some simulation work. It is not clear if there is any advantage in estimating both skewness and kurtosis in order to compute estimates for the bounds defined in (14).

The confidence interval defined in (14) only makes sense when  $\gamma_1(X) \neq 0$ . Otherwise, the approximation to a type IV distribution is no longer valid. The  $F_T$  distribution would be a type VII distribution which (with no surprise) corresponds to the  $t$ -Student distribution.

To compute the type IV distribution quantiles involved in (14),  package PearsonDS [17] was used.

## 3 Simulation Results

To analyse the performance of the confidence intervals (13) and (14) several underlying distributions were considered. The first choice was the exponential distribution  $\text{Exp}(\lambda)$ , where  $\mu = 1/\lambda$ , because both skewness and kurtosis do not depend on the parameter  $\lambda$ . The results concerning the estimated coverage probability, i.e., the proportion of intervals that contain  $\mu$  for a 95% confidence level when equation (13) (*Normal*) and equation (14) (*Type IV*) are used are presented in Table 1 ( $\lambda = 1$ ) and 2 ( $\lambda = 1000$ ). Samples sizes ranging from 5 to 200 ( $10^5$  replicas) were considered.

Table 1. Estimated coverage for  $X \sim \text{Exp}(1)$   
 $(\gamma_1(X) = 2, \gamma_2(X) = 6)$

$n$	<i>Normal</i>	<i>TypeIV</i>
5	81.11	89.44
10	86.95	91.09
15	89.21	91.00
20	90.20	91.09
30	91.79	91.09
50	92.92	91.16
200	94.39	94.01

Table 2. Estimated coverage for  $X \sim \text{Exp}(1000)$   
 $(\gamma_1(X) = 2, \gamma_2(X) = 6)$

$n$	<i>Normal</i>	<i>TypeIV</i>
5	81.21	89.63
10	86.81	90.93
15	89.13	91.48
20	90.56	91.90
30	91.79	92.17
50	93.00	92.76
200	94.74	94.15

Clearly, the observed coverages are very similar when the mean value of the underlying distribution changes. Comparing the performance of the confidence intervals (13) and (14), the coverage improves when a type IV distribution is used for low sample sizes. For large samples, the coverages of both intervals are similar with a little advantage to the Normal distribution.

Both confidence intervals tend to be liberal in the sense that its coverage is lower than the nominal probability.

As we are working in a skewed population setting two more skewed distributions were considered in the more likely situation of having to estimate both skewness and kurtosis. To assess the performance of the confidence intervals with

discrete and continuous underlying distributions it was chosen the Poisson  $P(\lambda)$  and Chi-square with  $m$  degrees of freedom  $\chi_m^2$  distributions. The usual confidence levels were used: 0.90, 0.95 and 0.99.

In those two distributions, skewness and kurtosis can be estimated different ways. It is possible to use:

- their observed values as estimates – strategy *TypeIV,o*;
- the sample mean to estimate both measures using the relation between each one and  $\mu$  – strategy *TypeIV,m*.

Recall that for a Poisson distribution  $X \sim P(\lambda)$ :

$$\gamma_1(X) = \mu^{-1/2} \text{ and } \gamma_2(X) = \mu^{-1} \quad (15)$$

For a Chi-square distribution  $X \sim \chi_m^2$ :

$$\gamma_1(X) = 2\sqrt{2}\mu^{-1/2} \text{ and } \gamma_2(X) = 12\mu^{-1} \quad (16)$$

Table 3 to Table 5 present the results of the estimated coverage probability for both strategies (compared to the use of the Normal approximation) at a 95% confidence level for two different Poisson distributions.

Table 3. Estimated coverage for  $X \sim P(1)$  ( $\gamma_1(X) = 1, \gamma_2(X)=1$ )

<i>n</i>	<i>Normal</i>	<i>TypeIV,o</i>	<i>Type IV,m</i>
10	91.04	92.76	93.23
15	91.45	92.51	92.85
20	92.67	93.15	93.05
30	93.26	93.84	93.73
50	93.96	93.96	93.79
200	94.62	94.63	94.75

Table 4. Estimated coverage for  $X \sim P(3)$  ( $\gamma_1(X) = 0.5774, \gamma_2(X)=0.3333$ )

<i>n</i>	<i>Normal</i>	<i>TypeIV,o</i>	<i>Type IV,m</i>
10	91.41	93.89	93.58
15	92.56	94.17	93.87
20	93.12	94.20	94.22
30	93.93	94.53	94.39
50	94.21	94.61	94.48
200	94.95	95.03	94.82

Clearly, using type IV distributions shortens the gap between the nominal and real confidence. For small sample sizes the gap is roughly half of what we would get using the Normal approximation. Comparing *TypeIV,o* to *TypeIV,m* there is no clear winner between the two strategies of estimating skewness.

Table 5. Estimated coverage for  $X \sim P(5)$  ( $\gamma_1(X) = 0.4472, \gamma_2(X)=0.2$ )

<i>n</i>	<i>Normal</i>	<i>TypeIV,o</i>	<i>Type IV,m</i>
10	91.76	92.96	94.78
15	92.89	94.16	94.29
20	93.25	94.48	94.09
30	93.98	94.44	94.52
50	94.47	94.69	94.70
200	94.89	94.88	94.87

The script of the simulations performed with an underlying Poisson distribution, due to discrete nature of the distribution, must take into account two issues than turn impossible the estimation of the skewness directly from the sample (strategy *TypeIV,o*). A finite value for the sample skewness cannot be computed if all values are equal because there is no possibility of estimating the sample variance. Another issue arises, if the sample is symmetric as the distribution from Pearson's system that is going to fit data is no longer a type IV distribution (it would be a type VII distribution). Every replica/sample that met one of those two criteria were excluded from the simulation and replaced by other (simulated) sample. Thus, in Tables 3 to 5 the sample size  $n = 5$  was not considered due to the high number of samples with all values equal or symmetric.

Table 6 to Table 8 are like the previous ones but consider underlying Chi-squared distributions

Table 6. Estimated coverage for  $X \sim \chi_6^2$  ( $\gamma_1(X) = 1.1547, \gamma_2(X)=2$ )

<i>n</i>	<i>Normal</i>	<i>TypeIV,o</i>	<i>Type IV,m</i>
5	85.49	89.85	91.75
10	90.11	92.3	92.53
15	91.59	92.95	92.74
20	92.41	93.31	93.25
30	93.38	93.78	93.51
50	93.90	94.06	93.76
200	94.81	94.72	94.56

All confidence intervals are once again liberal. For a given  $n$ , the real coverage tends to be closer to the nominal coverage when  $\gamma_1(X)$  and/or  $\gamma_2(X)$  are close to zero.

As expected, the strategies that use a type IV distribution overcome the approximation to the Normal distribution. However, for a large  $n$ , the performances are similar. When using a type IV distribution, it is not clear what is the best strategy to follow: *TypeIV,o* or *TypeIV,m*.

Table 7. Estimated coverage for  $X \sim \chi_{12}^2$  ( $\gamma_1(X) = 0.8165, \gamma_2(X)=1$ )

$n$	Normal	TypeIV,o	Type IV,m
5	86.51	91.14	91.51
10	91.13	93.48	93.06
15	92.30	93.74	93.40
20	92.98	94.00	93.73
30	93.74	94.36	94.00
50	94.26	94.55	94.28
200	94.57	94.73	94.64

Table 8. Estimated coverage for  $X \sim \chi_{18}^2$  ( $\gamma_1(X) = 0.6667, \gamma_2(X)=0.6667$ )

$n$	Normal	TypeIV,o	Type IV,m
5	89.92	91.63	91.38
10	91.18	93.70	93.46
15	92,54	94.10	93.82
20	92.97	94.11	94.13
30	93.78	94.45	94.29
50	94,27	94.68	94.47
200	94.85	94.90	94.92

All conclusions are similar for other levels of coverage: 0.90 and 0.99 (results not shown).

## 4 Discussion

More than one hundred years after Gosset's work, under the pseudonym of Student, the  $T$ -ratio potentiality has not been totally exploited yet. The  $t$ -Student statistic is in the genesis of what we now call Statistical Inference.

Trying to use very well-known methods to situations where assumptions are violated is common and, above all, a need in the way data is often messier than desired [22]. In the literature, several works regarding sample size calculation for skewed populations can be found (cf. [23,24]).

In [25], using simulation, it is described that the normative value of 50 for the sample size is not enough when the population is skewed.

However, when it is not possible to get a sample whose size is not equal or higher to the desired one inferential statistics may also be performed even if with some constraints.

This work showed how the  $t$ -Student statistics can be used outside the CLT assumptions. Population skewness should be considered and even in samples with only 5 individuals it is possible to improve the coverage of the confidence intervals (comparing to the straightforward application of the CLT). However, there was no clear winner between the two analyzed strategies that used type IV

distributions. This is, at some extent, surprising since the sample mean has some optimal proprieties as an estimator of the population mean value. Therefore, we would expect a better performance of strategy *TypeIV,m*.

When both skewness  $\gamma_1(X)$  and kurtosis  $\gamma_2(X)$  do not exceed 1, a coverage of about 94% is observed for samples sizes as low as 15 (Tables 4, 5 and 8). As those measures increase and are closer to 1, the required sample sizes for that coverage is about 20 (Table 7) or 50 (Table 3 and 6). Clearly, when skewness or kurtosis is very high, a large sample is required to achieve that level of coverage as seen when the underlying distribution was Exponential (Tables 1 and 2).

## 5 Conclusion

Clearly, the approximations to a type IV distribution is only a plus when the sample size is low and/or skewness is at least moderate. For instance, for  $n = 50$ , the performances were, in general, quite similar. In the future, it can be studied, in more detail, conditions where it makes sense to use the type IV distribution instead of the Normal distribution.

### References:

- [1] Student, On the probable error of the mean. *Biometrika*, Vol. 6, 1908, pp. 1-25.
- [2] Martins JP, Student t-statistic distribution for non-Gaussian populations', *Proceedings of the ITI 2010, 32nd International Conference on Information Technology Interfaces*, 2010, pp. 563-568.
- [3] Chandra T, *A First Course in Asymptotic Theory of Statistics*, Narosa, 1999.
- [4] Johnson N, Kotz S, Balakrishnan N, *Continuous Univariate Distributions*, Vol. I, 2<sup>nd</sup> edition, Wiley Interscience Publication, 1994.
- [5] Hall P, Edgeworth Expansion for Student's  $t$  Statistic Under Minimal Moment Conditions, *Ann Probab*, 1987, pp. 920-931.
- [6] Pearson K, Memoir on skew variation in homogeneous material, *Philosophical Transactions of the Royal Society of London*, 1985, pp. 343-414.
- [7] Singh V, Zhang L, Pearson System of Frequency Distributions', In *Systems of Frequency Distributions for Water and Environmental Engineering*, Cambridge University Press, 2020.
- [8] Andreev A, Kanto A, Malo P, Simple Approach for Distribution Selection in the

- Pearson System, Helsinki School of Economics-Electronic Working Papers, 2005, 388.
- [9] Chifurira R, Chinghamu K, Using the Generalized Pareto and Pearson type IV Distributions to Measure Value-At-Risk for the Daily South African Mining Index, *J Stud Econ Econom*, Vol. 41, 2020, pp. 33-54.
- [10] Saswat P, Revisiting value-at-risk and expected shortfall in oil markets under structural breaks: The role of fat-tailed distributions, *Energy Econ* 10, 2021, 105452.
- [11] Bouver H, Table of the cumulative standardized Pearson type IV distribution function, *Themis Report* 28, Technical Report 100, Department of Statistics and Computer Science, The University of Georgia, Athens, 1973.
- [12] Davis C, Stephens M, Approximate percentage points using Pearson curves. *J Appl Stat*, Vol. 32, 1983, pp. 322-327.
- [13] Stuart A, Ord J, *Kendall's Advanced Theory of Statistics*, Vol. 1, 6<sup>th</sup> edition, Oxford University Press, 1994
- [14] Woodward W, Approximation of Pearson type IV tail probabilities, *JASA*, Vol. 71, 1976, pp. 513-514.
- [15] Skates S, On secant approximations to cumulative distribution functions. *Biometrika*, Vol. 80, 1993, pp. 223-235.
- [16] Willink R, A closed-form expression for the Pearson type IV distribution function, *Aust N Z J Stat*, Vol. 50, 2008, pp. 199-205
- [17] Becker M, Klößner S, Heinrich J. *Package PearsonDS*, Available at <https://cran.r-project.org/web/packages/PearsonDS/PearsonDS.pdf>. Accessed February 11, 2022.
- [18] Yang Q, An X, Pan W, Computing and graphing probability values of pearson distributions: a SAS/IML macro, *Source Code Biol Med*, 2019, Vol. 14, pp. 1-6.
- [19] Stavroyiannis S, Makris I, Nikolaidis V, Zarangas L, Econometric modeling and value at-risk using the Pearson type-IV distribution, *Int Rev Financial Anal*, Vol. 22, 2012, pp. 10-17
- [20] Wang J, Yang K, Using type IV Pearson distribution to calculate the probabilities of underrun and overrun of lists of multiple cases, *Eur J Anaesthesiol*, Vol. 31, 2014, pp. 363-370.
- [21] Meeden G, Interval estimators for the population mean for skewed distributions with a small sample size, *Journal of Applied Statistics*, Vol. 26, 1999, pp. 81-96.
- [22] Hoaglin DC, Mosteller F, Tukey, JW, *Understanding Robust and Exploratory Data Analysis*, Wiley, 2000.
- [23] Gerlovina I, van der Laan MJ, Hubbard A, Big Data, Small Sample. *Int J Biostat*, Vol. 13, 2017, pp. 20170012.
- [24] Cundill B, Alexander ND, Sample size calculations for skewed distributions, *BMC Med Res Methodol*, Vol. 15, pp. 28
- [25] Piovesana A, Senior G, How Small Is Big: Sample Size and Skewness. Assessment, Vol. 25, 2018, pp. 793-800.

#### **Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)**

All authors equally contributed to this paper.

#### **Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself**

Funded by FCT – Fundação para a Ciência e a Tecnologia, Portugal, through the projects UIDB/00006/2020 and UIDP/00006/2020.

#### **Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)**

This article is published under the terms of the Creative Commons Attribution License 4.0

[https://creativecommons.org/licenses/by/4.0/de.ed.en\\_US](https://creativecommons.org/licenses/by/4.0/de.ed.en_US)