

REPLACEMENT OF BIASED ESTIMATORS WITH UNBIASED ONES IN THE CASE OF STUDENT'S t -DISTRIBUTION AND GEARY'S KURTOSIS

GERGELY TÓTH* AND PÁL SZEPESVÁRY

Institute of Chemistry, Faculty of Science, Eötvös Loránd University, 1/A Pázmány Péter sétány, Budapest, H-1117, HUNGARY

The use of biased estimators can be found in some historically and up to now important tools in statistical data analysis. In this paper their replacement with unbiased estimators at least in the case of the estimator of the population standard deviation for normal distributions is proposed. By removing the incoherence from the Student's t -distribution caused by the biased estimator, a corrected t -distribution may be defined. Although the quantitative results in most data analysis applications are identical for both the original and corrected t -distributions, the use of this last t -distribution is suggested because of its theoretical consistency. Moreover, the frequent qualitative discussion of the t -distribution has come under much criticism, because it concerns artefacts of the biased estimator. In the case of Geary's kurtosis the same correction results $(2/\pi)^{1/2}$ unbiased estimation of kurtosis for normally distributed data that is independent of the size of the sample. It is believed that by removing the sample-size-dependent biased feature, the applicability domain can be expanded to include small sample sizes for some normality tests.

Keywords: unbiased estimator, normal distribution, Anscombe-Glynn test, Jarque-Bera test, Bonett-Seier test

1. Introduction

The Student's t -distribution [1] is one of the most widely used statistical functions in experimental practice. It is well documented that experts active in analytical, physical and clinical chemistry, biology, agriculture, ecology, economy as well as forensic science or even legal representatives apply this tool to formulate solid statements, e.g. population means, to differentiate between two sample means, etc., which are in most cases necessarily based on a limited number of observations. Fortunately, scientists are in this respect soundly supported by a lot of textbooks, standards, software, software systems, and last but not least trained in elements of statistics. However, it should be a moral obligation to be aware of the evolution of this routinely used method, its principles and often tacitly supposed assumptions. Although not crucial in daily use, it is worthwhile to know that besides Student's t -distribution there are different functions which may be suitable for the determination of percentiles in the same way as the t -distribution. They may differ mainly in terms of alternate estimators for the population mean and population standard deviation. An attractive variant is presented in this work emphasizing its theoretically fully consistent feature on the contrary to the Student's t -distribution.

The Student's t -distribution corresponds to a ratio of normally distributed random variables to chi-distributed random variables (see the references in the historical review of Zabell [2]). Chi-distributed variables postulate normally distributed data as well. Gosset [1] used in his definition an estimate of the standard deviation which is biased relating the population standard deviation and even the variance in the case of normally distributed random variables as was shown earlier by Helmert [3-4]. Amazingly, when Fischer proposed a transformation of Gosset's original z variable to $t = z\sqrt{N-1}$ [5-6], he chose an estimate for the standard deviation which is also biased relating σ . A corrected t_c -distribution is proposed that fulfils all theoretical requirements and yields a more normal distribution-like shape. It is consequently based on normal sample data and uses an unbiased estimator.

A similar correction can be applied to Geary's kurtosis that is the ratio of the mean deviation to the standard deviation [7]. In this work the use of the correction is proposed in order to eliminate the sample-size dependency of the mean Geary's kurtosis on normally distributed data.

Finally, some remarks are made on statistical tests based on sample-size dependent values in order to extend their applicability to small sample sizes.

*Correspondence: toth@chem.elte.hu

2. Theoretical Background

The square root of the mean of the square of the centred observations of the random variable, Y ,

$$s_N = \sqrt{\frac{\sum_{i=1}^N (y_i - \bar{y})^2}{N}} \quad (1)$$

may be an estimator of the population standard deviation, σ , of a sample. N is the sample size and \bar{y} denotes the sample mean. Bessel pointed long before to the bias of s_N and proposed:

$$s = \sqrt{\frac{\sum_{i=1}^N (y_i - \bar{y})^2}{N-1}} \quad (2)$$

Although s^2 is already an unbiased estimator of the variance, σ^2 , irrespective of the distribution of Y , the statistics s_N and s are both biased in terms of the standard deviation, σ . However, there is a correction [5-6, 8-9]:

$$c_4(N) = \sqrt{\frac{2}{N-1}} \frac{\Gamma\left(\frac{N}{2}\right)}{\Gamma\left(\frac{N-1}{2}\right)}, \quad (3)$$

which applies to s by transforming it, in the case of normally distributed variables, into an unbiased statistic:

$$s_c = \frac{s}{c_4(N)}. \quad (4)$$

$c_4(N)$ follows from Helmert's papers or Cochran's theorem [10]. For normally distributed Y , $\sqrt{N-1}(s/\sigma)$ exhibits a chi distribution with $N-1$ degrees of freedom. $c_4(N)$ is the expected value of s/σ . There is a $c_2(N)$ value as well, that is equal to the expected value of s_N/σ . The correcting effect of $c_4(N)$ may be considerable for low values of N (Table 1).

Table 1. Selected $c_4(N)$ corrections

N	$c_4(N)$
2	0.7979
3	0.8862
4	0.9213
5	0.9400
10	0.9727
20	0.9869
30	0.9914

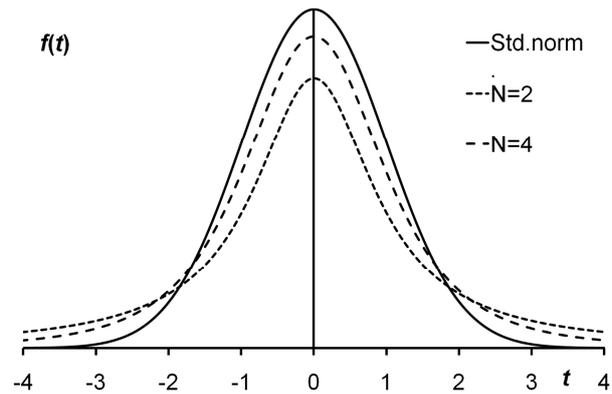


Figure 1. Student's t -distribution compared with the standard normal distribution

The correction term is frequently used in Statistical Process Control (SPC) to define $\pm 3\sigma$ intervals and process standard deviations that are determined for samples of different sizes.

3. Results and Discussion

3.1. Discussion of the density functions of the Student's t -distribution

Let Y be an independent $N(\mu, \sigma^2)$ variable. The

$$t = \frac{\bar{y} - \mu}{s/\sqrt{N}} \quad (5)$$

statistic exhibits the usual Student t -distribution with $N-1$ degrees of freedom. The density function (Eq.(6)) can be derived as the quotient of normal and chi-distributed random variables. The independency of the nominator and denominator can be shown and it is also proved by a series of theorems that t defined in Eq.(5) follows a Student's t -distribution with $N-1$ degrees of freedom (see the references in [2]):

$$f(t) = \frac{1}{\sqrt{\pi(N-1)}} \frac{\Gamma\left(\frac{N}{2}\right)}{\Gamma\left(\frac{N-1}{2}\right)} \left(1 + \frac{t^2}{N-1}\right)^{-\frac{N}{2}} \quad (6)$$

The distribution sketched in Fig.1 is widely known and needs no comments except for its flaw: it is based on the biased statistic (Eq.(2)) for σ .

By replacing the standard deviation of the sample, s , in Eq.(5) with the unbiased equivalent one, s_c , corrected by $c_4(N)$, results in a new value:

$$t_c = \frac{\bar{y} - \mu}{s_c/\sqrt{N}}, \quad (7)$$

and in the density function:

Table 2. Sample-size dependence of Geary's and Pearson's kurtosis. The $c_4(N)$ -corrected Geary's and the size-corrected Pearson's kurtosis [11] provided the ∞ limit values for all sample sizes within the statistical uncertainty of the 10^5 random samples.

N	Geary's kurtosis	Pearson's kurtosis
3	0.9004	1.5000
4	0.8659	1.8005
5	0.8489	1.9996
6	0.8385	2.1437
7	0.8319	2.2501
8	0.8267	2.3340
9	0.8233	2.4006
10	0.8200	2.4557
20	0.8085	2.7141
50	0.8021	2.8820
100	0.7999	2.9408
∞	0.7979	3.0000

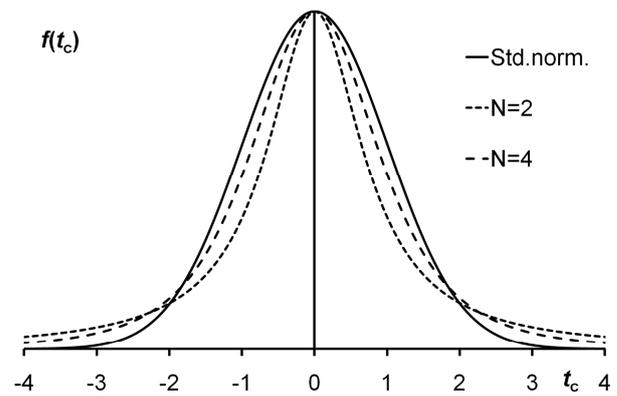


Figure 2. Corrected Student's *t*-distribution compared with the standard normal distribution

$$f(t_c) = \frac{1}{c_4(N)} \frac{\Gamma\left(\frac{N}{2}\right)}{\sqrt{(N-1)\pi} \Gamma\left(\frac{N-1}{2}\right)} \left(1 + \frac{t_c^2}{c_4^2(N-1)}\right)^{-\frac{N}{2}}$$

$$= \frac{1}{\sqrt{2\pi}} \left(1 + \frac{t_c^2}{c_4^2(N-1)}\right)^{-\frac{N}{2}} \quad (8)$$

which is shown in Fig.2. As expected, the corrected Student's *t*-distribution (Student's t_c -distribution) consists of more distinct peaks, it exhibits fatter tails than the standard normal distribution, but rather oddly at its maximum, i.e. when $t_c = 0$, the value $f(t_c)$ does not depend on N and is represented by the value $1/\sqrt{2\pi}$, the same as for those of normal distribution. The function $f(t_c)$ and the Student's *t*-distribution without a doubt differ in this respect.

The obvious differences between the Student and modified Student distributions do not complicate their daily usage. The confidence intervals calculated using s and t or s_c and t_c estimators and distributions, respectively,

$$\bar{y} \pm \frac{s_c t_c^{\alpha/2, (N-1)}}{\sqrt{N}} \quad \text{and} \quad \bar{y} \pm \frac{s t^{\alpha/2, (N-1)}}{\sqrt{N}} \quad (9)$$

do not differ, because

$$s_c t_c^{\text{crit}(\alpha/2, N-1)} = \frac{s}{c_4(N)} t^{\text{crit}(\alpha/2, N-1)} c_4(N)$$

$$= s t^{\text{crit}(\alpha/2, N-1)} \quad (10)$$

Evidently, corresponding estimators and functions should be used.

3.2. The effect of the correction to Geary's kurtosis

Geary's kurtosis [7], w_N , is the ratio of the mean absolute deviation (MAD) to the standard deviation (Eq.(11)). It is an alternative to Pearson's kurtosis based on the fourth moment. The expected value of Geary's kurtosis depends on the sample size even for normally distributed data [7]. The mean values of 10^5 random samples from standard normal distributions are shown in Table 2.

$$w_N = \frac{\frac{1}{N} \sum_{i=1}^N |y_i - \bar{y}|}{s_N} \quad (11)$$

If the $c_4(N)$ corrections (Table 1) are applied during the calculation of the nominator of the ratio as $w_{N, \text{corr}} = w_N c_4(N)$, the expected value of the kurtosis is $(2/\pi)^{1/2}$ for all sample sizes. This means that the platykurtic and the leptokurtic features of a sample can be found without searching for the size-dependent dividing value in tables.

3.3. Sample-size bias in statistical tests

Geary's kurtosis and its transformed values are used in normality tests due to their enhanced sensitivity to some leptokurtic deviations from normality [12]. Contrary to the case of the Student's *t*-distribution, where the correction has no effect on the *t*-test, here the effect of the correction is not cancelled. Generally, the size dependence decreases the performance of the tests for small sample sizes. This feature is interpreted by users as a recommendation that the tests are unsuitable for small sample sizes. In the same way, neglect of sample-size dependence is applicable in tests where Pearson's kurtosis is used. The calculated mean value of Pearson's kurtosis is shown in Table 2 but its convergence is rather weak to the theoretical value of 3. It should be noted here that the sample-size unbiased estimator of kurtosis can be easily calculated [11].

Table 3. Sample-size dependence of five normality tests based on unbiased or sample-size-dependent biased estimators. The numbers show the ratio of the rejected null hypotheses to all trials at a significance level of 0.05 from 10^5 random samples.

<i>N</i>	<i>Shapiro-Wilk</i>	<i>D'Agostino</i>	<i>Anscombe-Glynn</i>	<i>Bonett-Seier</i>	<i>Jarque-Bera</i>
4	0.0502				
5	0.0521			0.0373	
6	0.0477		0.0189	0.0087	
7	0.0492		0.0343	0.0282	
8	0.0505		0.0365	0.0348	
9	0.0507	0.0538	0.0369	0.0390	0.0024
10	0.0498	0.0528	0.0394	0.0417	0.0058
20	0.0497	0.0525	0.0406	0.0421	0.0089
50	0.0498	0.0500	0.0466	0.0470	0.0241
100	0.0499	0.0493	0.0533	0.0490	0.0368

In *Table 3* the type-I error of some normality tests calculated on 10^5 standard normal samples is shown. In the calculation the ‘moments’ package in R was used [13]. *Table 3* contains the ratio of the samples to all samples where the H_0 hypothesis of normality was rejected at the significance level of 0.05. The Shapiro-Wilk method [14] uses the ratio of two unbiased estimators of the variance, and is suitable for all data sizes. The skewness test of D’Agostino [15] slightly overestimates the number of rejected cases. It is based on the normalized third-order central moment definition of skewness, where the expected value is estimated without bias. The Anscombe-Glynn [16] test applies Pearson’s kurtosis without size correction using a biased estimation of normally distributed data. The Bonett-Seier [12] test shown here uses Geary’s kurtosis and the Jarque-Bera test [17] combines skewness and Pearson’s kurtosis. The performance of the last three tests was rather weak for small sample sizes, whereof one cause might be the lack of correction for small sample sizes even for normally distributed data. These tests are usually only recommended for medium and large sample sizes. The correction should extend the applicability domain to small sample sizes. Of course, the type-I error of normally distributed data is only one narrow aspect of a test, detailed analysis should be performed to investigate the effect of correction on many distributions, like, e.g. in ref. [12].

4. Conclusion

Nowadays, data are evaluated by computers and biased estimators can be replaced by unbiased ones, even if their calculation schemes are complex.

It has been shown that, in terms of Student’s t -distributions, to decide upon the confidences of statistics one has two functions which are completely equivalent as far as practical applicability is concerned. They can, however, be distinguished theoretically. The assertion that only the unbiased estimator should be recognized as the correct one implies the use of the corrected Student’s t -distribution, $f(t_c)$. In that case the known

shape of the Student’s t -distribution may be labelled as an artefact and the usual application of the Student’s t -distribution as a production of “correct numbers by an incoherent theory”.

In the case of Geary’s kurtosis, the correction removes the sample-size dependence from the expected value. This change of distinguishing platykurtic or leptokurtic features of the sample is simpler than using the original version of Geary’s kurtosis. Furthermore, subtracting $(2/\pi)^{1/2}$ results in a number to be interpreted in a similar way to the excess kurtosis obtained by subtracting 3 from the Pearson’s kurtosis.

As a further study, the use of unbiased/sample-size-dependent corrections to extend the applicability domain to small sample sizes in the case of normality tests is recommended. It is believed that the use of biased estimators was acceptable before the age of computers and a systematic change to unbiased ones might be necessary in terms of statistics and standards with regard to industrial processes.

Acknowledgement

The authors are sincerely grateful for the valuable comments of Prof. S. Kemény and other participants during the scientific discussion after the presentation concerning Student’s t -distribution at the SCAC 2010 conference in Budapest in September 2010.

REFERENCES

- [1] Student (Gosset, W.S.): The Probable Error of a Mean, *Biometrika* 1908 **6**(1), 1–25 DOI: 10.2307/2331554
- [2] Zabell, S.L.: On Student’s 1908 Article “The Probable Error of a Mean”, *J. Am. Stat. Assoc.* 2008 **103**(481), 1–7 DOI: 10.1198/016214508000000049
- [3] Helmert, F.R.: Über die Wahrscheinlichkeit der Potenzsummen der Beobachtungsfehler und über einige damit in Zusammenhang stehende Fragen, *Z. Math. Phys.* 1876 **21**, 192–218
- [4] Helmert, F.R.: Die Genauigkeit der Formel von Peters zur Berechnung des wahrscheinlichen Beobachtungsfehles directer Beobachtungen gleicher Genauigkeit, *Astronomische Nachrichten* 1876 **88**(8-9), 113–131 DOI: 10.1002/asna.18760880802
- [5] Fischer, R.A.: Applications of “Student’s” distribution, *Metron* 1925 **5**, 90–104
- [6] Fischer, R.A.: Statistical methods for research workers, (Oliver & Boyd, Edinburgh and London) 1925.
- [7] Geary, R.C.: Moments of the ratio of the mean deviation to the standard deviation for normal samples, *Biometrika* 1936 **28**(3/4), 295–307 DOI: 10.2307/2333953
- [8] Grubbs, F.E.; Weaver, C.L.: The best unbiased estimate of population standard deviation based on group ranges, *J. Am. Stat. Assoc.* 1947 **42**(238), 224–241 DOI: 10.1080/01621459.1947.10501922

- [9] Vincze, I.: Matematikai statisztika ipari alkalmazásokkal, (Műszaki Kiadó, Budapest), 1975 pp. 58, 89, 165–168 ISBN 963-10-0472-4
- [10] Cochran, W.G.: The distribution of quadratic forms in a normal system, with applications to the analysis of covariance, *Math. Proc. Cambridge* 1934 **30**(2), 178–191 DOI: 10.1017/S0305004100016595
- [11] Joanes, D.N.; Gill, C.A.: Comparing measures of sample skewness and kurtosis, *J. Roy. Stat. Soc. D-Stat.* 1998 **47**(1), 183–189 DOI:10.1111/1467-9884.00122
- [12] Bonett, D.G.; Seier, E.: A test of normality with high uniform power, *Comput. Stat. Data An.* 2002 **40**(3), 435–445 DOI: 10.1016/S0167-9473(02)00074-9
- [13] Komsta, L.; Novomestky, F.: Moments R package, version 0.14, 2015 at <http://cran.r-project.org>, accessed in October 2017
- [14] Shapiro, S.S.; Wilk, M.B.: An Analysis of Variance Test for Normality (Complete Samples), *Biometrika* 1965 **52**(3/4), 591–611 DOI: 10.1093/biomet/52.3-4.591
- [15] D'Agostino, R.B.: Transformation to Normality of the Null Distribution of G_1 , *Biometrika* 1970 **57**(3), 679–681 DOI: 10.1093/biomet/57.3.679
- [16] Anscombe, F.J.; Glynn, W.J.: Distribution of the kurtosis statistic b_2 for normal samples, *Biometrika* 1983 **70**(1), 227–234 DOI: 10.1093/biomet/70.1.227
- [17] Jarque, C.M.; Bera, A.K.: Efficient tests for normality, homoscedasticity and serial independence of regression residuals, *Economic Letters* 1980 **6**(3), 255–259 DOI: 10.1016/0165-1765(80)90024-5