MATHEMATICAL SCIENCES

THE CENTRAL LIMIT THEOREM AND THE MEASURES OF CENTRAL TENDENCY

Cvetkov V.
Chief Assist. Prof. PhD Eng.
University of Architecture, Civil Engineering and Geodesy.
1164, 1 Hristo Smirnenski Blvd.
Sofia, Bulgaria
ORCİD: 0000-0001-9628-6768

Abstract

The current study demonstrates that for a large number of independent Gamma (2, 1) distributed random variables $X_1, X_2, \ldots, X_n$ with finite variance, the measures of central tendency, e.g., mean, median, center, and C_2 have a normal distribution. In addition, when the size of a sample of the Gamma (2, 1) distributed random variables tends to infinity (n→∞) each of the above measures tends to a strictly expected value. For example, the mean tends to 2.000, the median tends approximately to 1.678, and both new statistics the center and the C_2 tend to 1.814. It is also illustrated that when n→∞ the ratio between the standard deviations based on the mean and the center statistic tend approximately to 1.47 regarding the Gamma(2,1) distribution. Using the above findings, the main goal of the article is to demonstrate that the central limit theorem is valid for both the center and the C_2 statistics, even in the matter of strongly skewed distributions.

Keywords: center statistic; expectation; gamma distribution; weighted average;

Introduction

One of the most useful theorems in statistics is the central limit theorem. According to it, if $X_1, X_2, \ldots, X_n$ is a random sample of size n taken from a population (either finite or infinite) with mean μ and finite variance σ^2 and if $\bar{X}$ is the sample mean, the limiting form of the distribution of

$$Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \tag{1}$$

is the standard normal distribution [4, 6] as n→∞.

Using (1), we can express the sample mean $\bar{X}$ by equation (2).

$$\bar{X} = \mu + Z.\sigma/\sqrt{n} \tag{2}$$

Since (2), it follows that $\bar{X}$ approximately has an $N(\mu, \sigma^2/n)$ [3]. It is usually true if n ≥ 30, even for samples taken from strongly skewed populations.

It would be interesting whether the central limit theorem is valid for both new statistics the center and C_2 [2], that is to say, whether the distribution of the differences between these statistics and their common expectation E [$\breve{X}$], when multiplied by the factor $\sqrt{n}$ ($\sqrt{n}(\breve{X} - E[\breve{X}])$), approximates the normal distribution with an expectation E [$\breve{X}$] and variance $\breve{\sigma}^2$. In other words, whether equation (3) is valid.

$$\breve{X} = E[\breve{X}] + Z.\breve{\sigma}/\sqrt{n} \tag{3}$$

According to [2], the center of the sample will tend to the C_2 statistic as the sample size n → ∞. In addition, the standard error of the center tends to the standard error of the C_2. So, let us denote both statistics by $\breve{X}$. The expectation E [$\breve{X}$], the standard error of the $\breve{X}$ ($\breve{\sigma}_{\breve{X}}$) and the variance $\breve{\sigma}^2$ of a population can be expressed by equations (4)-(6).

$$\breve{X} = \sum_{i=1}^{n} w_i.X_i \,/ \sum_{i=1}^{n} w_i \tag{4}$$

$$\breve{\sigma}^2 = \sum_{i=1}^{n} w_i.(X_i - \breve{X})^2/(n-1) \tag{5}$$

$$\breve{\sigma}_{\breve{X}} = \breve{\sigma}/\sqrt{n} \tag{6}$$

The weights w_i in equations (4)-(6) are the weights w_i', expressed by equation (25) in [2]. More detailed information on how the above weights concerning the center and C_2 statistics one can find in [2].

Simulations and Results

In order to check whether the mean, the median, the center and the C_2 statistic of a data sample drawn from the Gamma (2, 1) distribution have finite values, we generated two sets with 500 random variables. For each data set we yeilded the realizations of the first n values of the above mentioned statistics in the form of the order $X_1, X_2, \ldots, X_{500}$ and plotted them against n. Thus, we yielded the top charts in Figure 1. Simultaneously, we calculated the standard errors of each statistic of the order $X_1, X_2, \ldots, X_{500}$ and plotted them against n. The results are presented by the bottom charts in Figure 1.

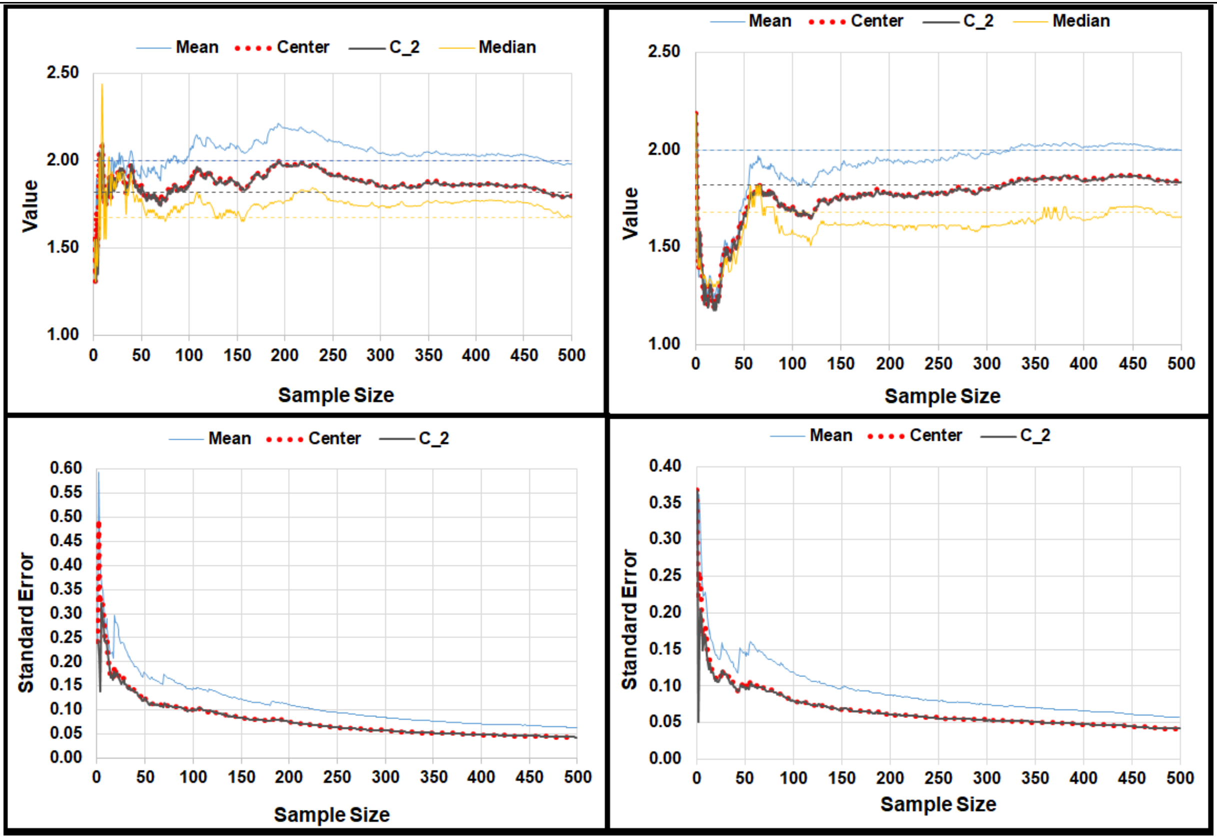

Figure 1: Statistics and their standard errors of two realizations of sequence of Gamma (2, 1) distributed random numbers. Top: Statistics. Bottom: Standard Errors. Left: The First Realization. Right: The Second Realization.

Looking at Figure 1, one can see that not only the means but also the meadians, the centers and the C_2 statistics of the samples converge to finite values. These values for the mean and the median of the Gamma (2, 1) distribution are 2.000 [3] and approximately 1.678 [1], respectively. The variance of the Gamma (κ, θ) distribution is $\sigma^2 = k.\theta^2$ [5]. Therefore, the variance of the Gamma (2, 1) distribution is $\sigma^2 = 2.1^2 = 2$, which means that the standard deviation is σ ≈ 1.414. Using the software available on [7] we generated 20000 Gamma (2, 1) random numbers and found that the expected value E [$\breve{X}$] of both the center and C_2 is approximately equal to 1.814 and the standard deviation is $\breve{\sigma} \approx 0.961$. Thus, the standard error of both the center and C_2 statistics is approximately 1.47 times less than the standard error of the mean concerning samples derived from the Gamma (2, 1) distribution. Besides, the standard errors of the center and C_2 statistics tend to zero as n → ∞. The curves of the standard errors of the analyzed statistics given by the charts in the bottom of Figure 1 clearly illustrate this fact. If one multiplies the factor $\sqrt{n}$ by the standard error of the center of a sample of size n, taken from Figure 1 or more precisely from the chart in the bottom right corner of Figure 2, they will obtain a value close to 0.961. This result is equal to the square root of the variance of the Gamma (2, 1) distribution calculated by equation (5). As a result, we can claim that the variance $\breve{\sigma}^2$ has a finite value. Consequently, equation (3) is valid as n → ∞.

Figure 2 below shows the convergence of the analyzed statistics obtained by 500 independent Gamma (2, 1) distributed samples whose sizes vary from 1 to 500. These charts reconfirm the above findings.

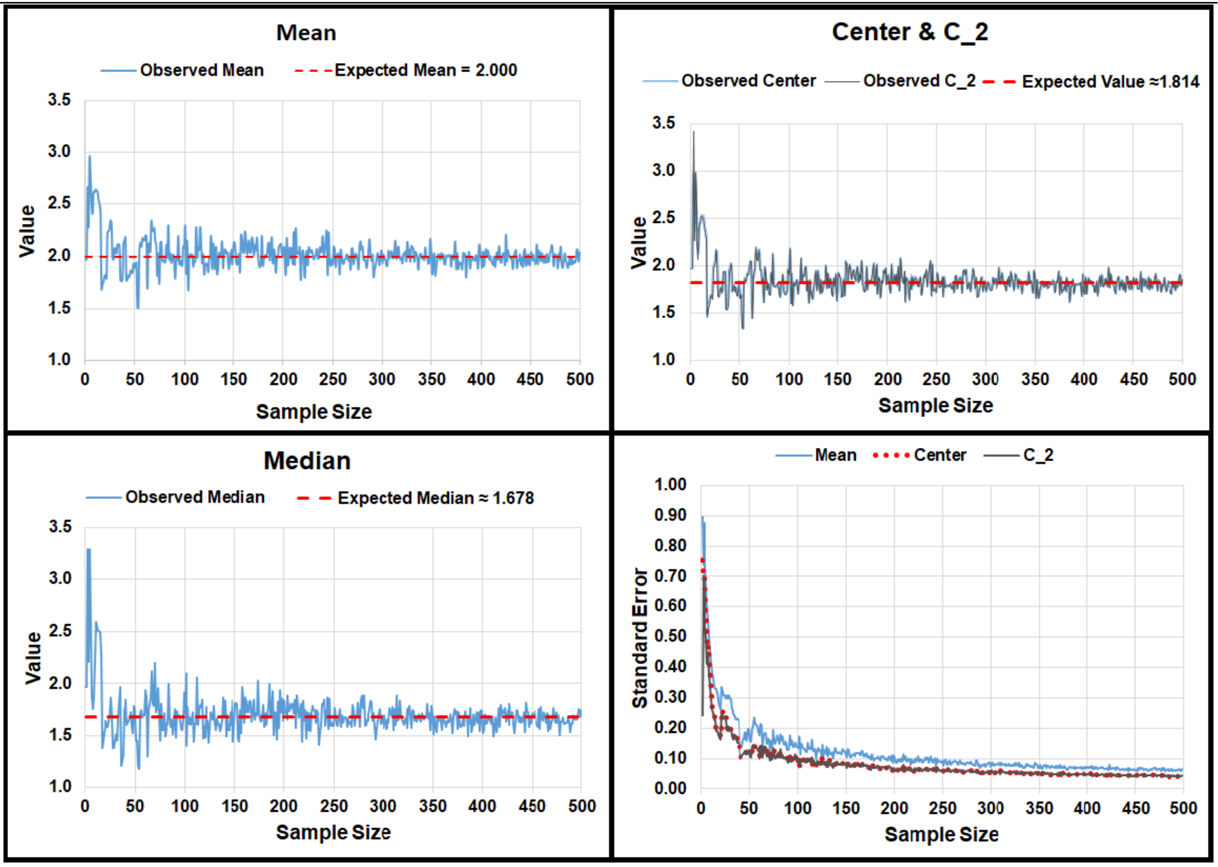

Figure 2: Statistics and their standard errors of independent samples with different sizes of Gamma (2, 1) distributed random numbers. Top Left: Mean. Top Right: Median. Bottom Left: Center and C_2 Statistics. Bottom Right: Standard Errors of Averages, Centers and C_2 Statistics.

In order to demonstrate graphically the validity of equation (3) in the matter of both the center and C_2 statistics Figures 3-6 below present the histograms of the analyzed statistics based on 30 independent samples of size n = { 10, 30, 50, 100, 150, 200, 250 } drawn from the Gamma (2, 1) distribution. The normality of the statistics was also checked by both a χ^2 and an one-sample K-S tests and the results are embedded into the corresponding charts. Deviations from the normal distribution were not detected.

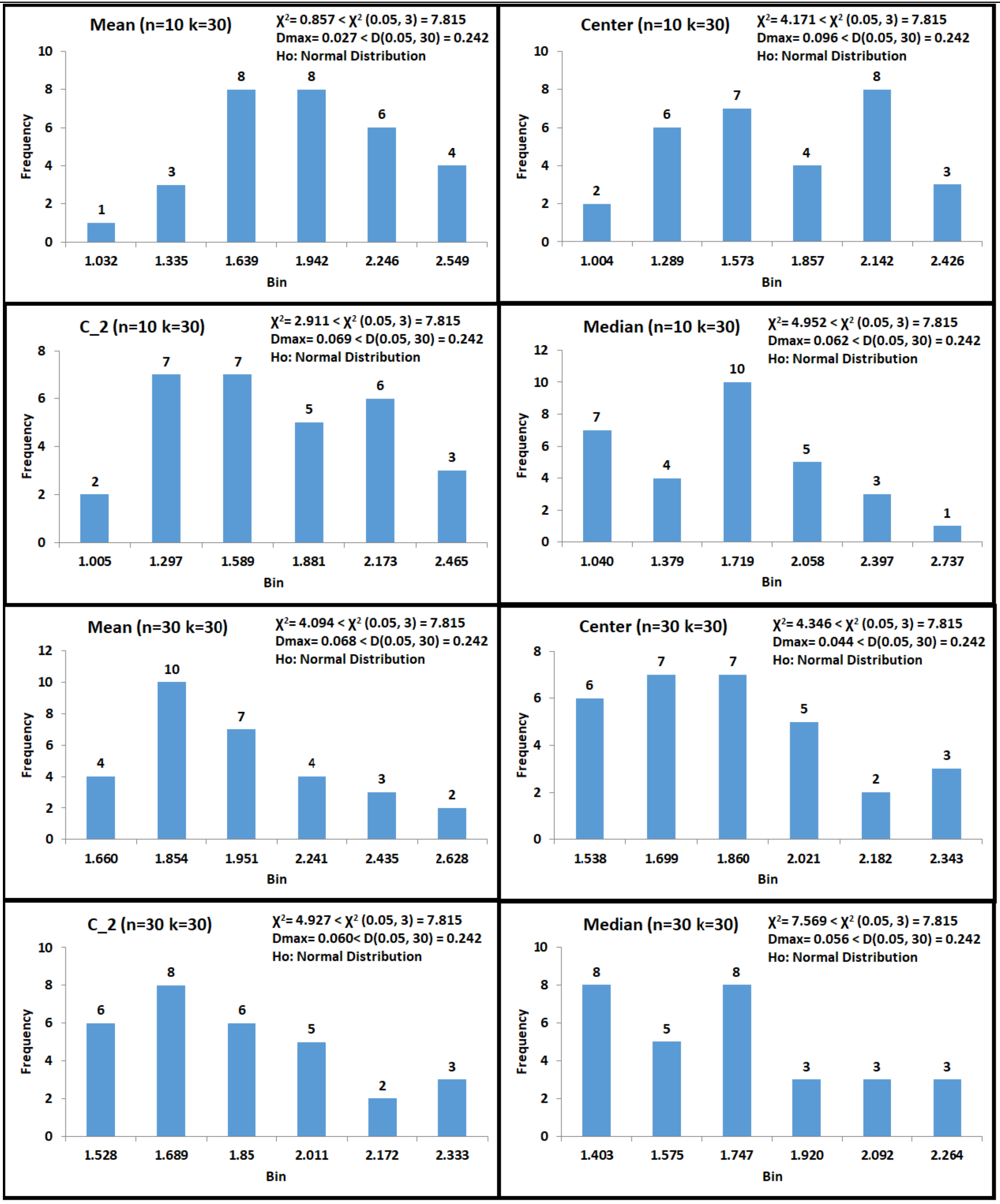

Figure 3: Histograms of the Means, Centers, C_2 and Medians based on k=30 samples of Gamma (2, 1) distributed random numbers of sizes n=10 and n=30.

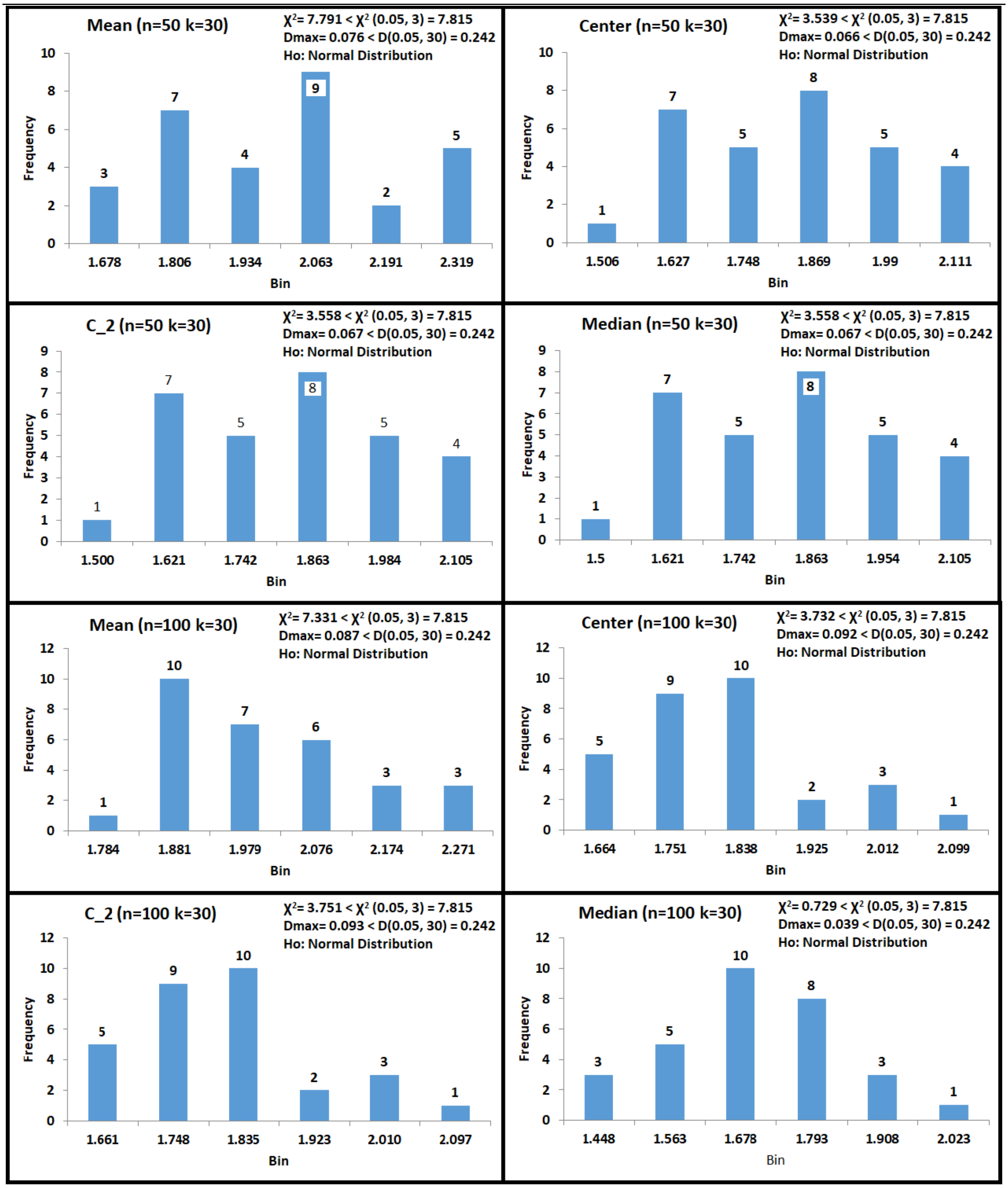

Figure 4: Histograms of the Means, Centers, C_2 and Medians based on k=30 samples of Gamma (2, 1) distributed random numbers of sizes n=50 and n=100.

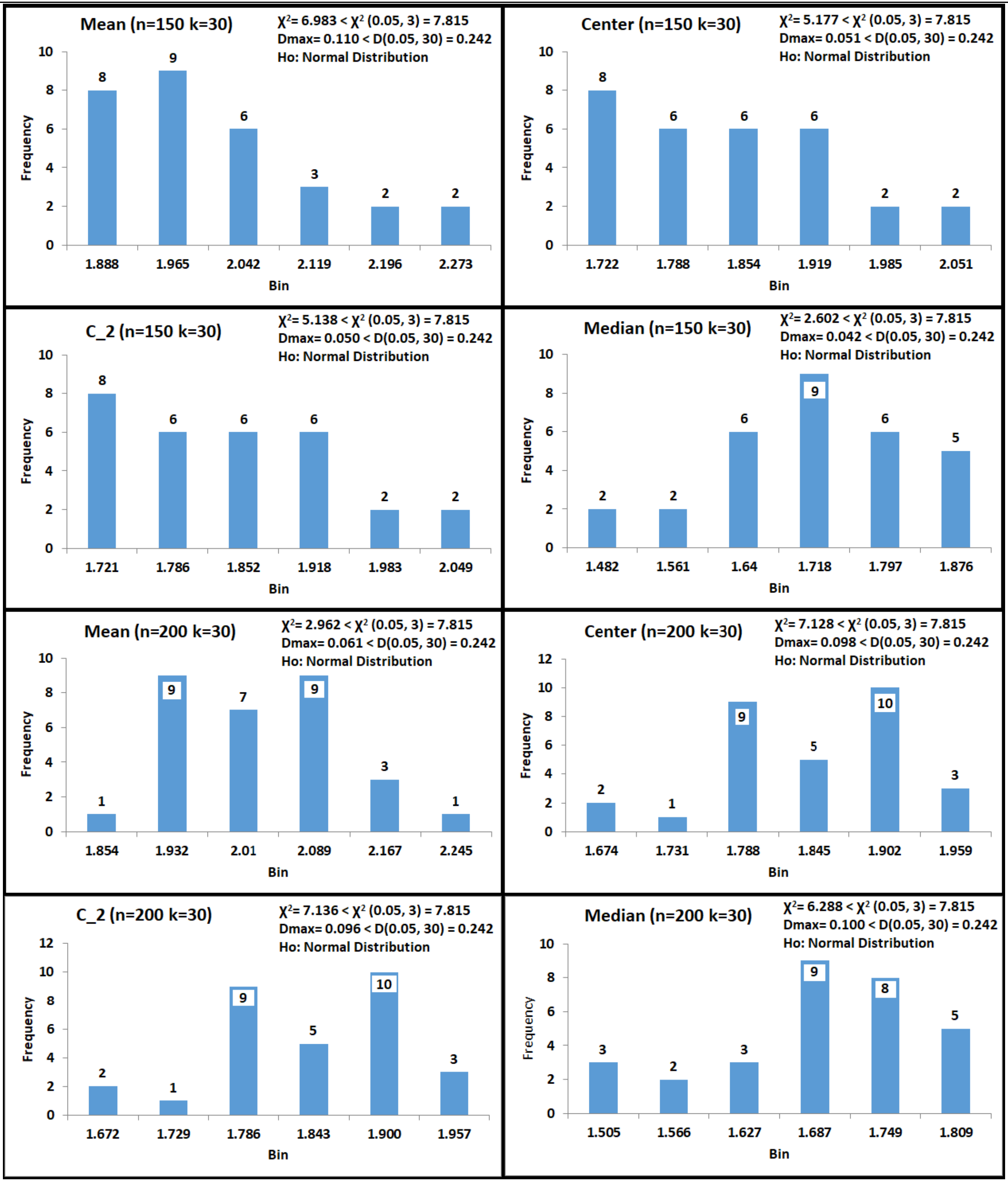

Figure 5: Histograms of the Means, Centers, C_2 and Medians based on k=30 samples of Gamma (2, 1) distributed random numbers of sizes n=150 and n=200.

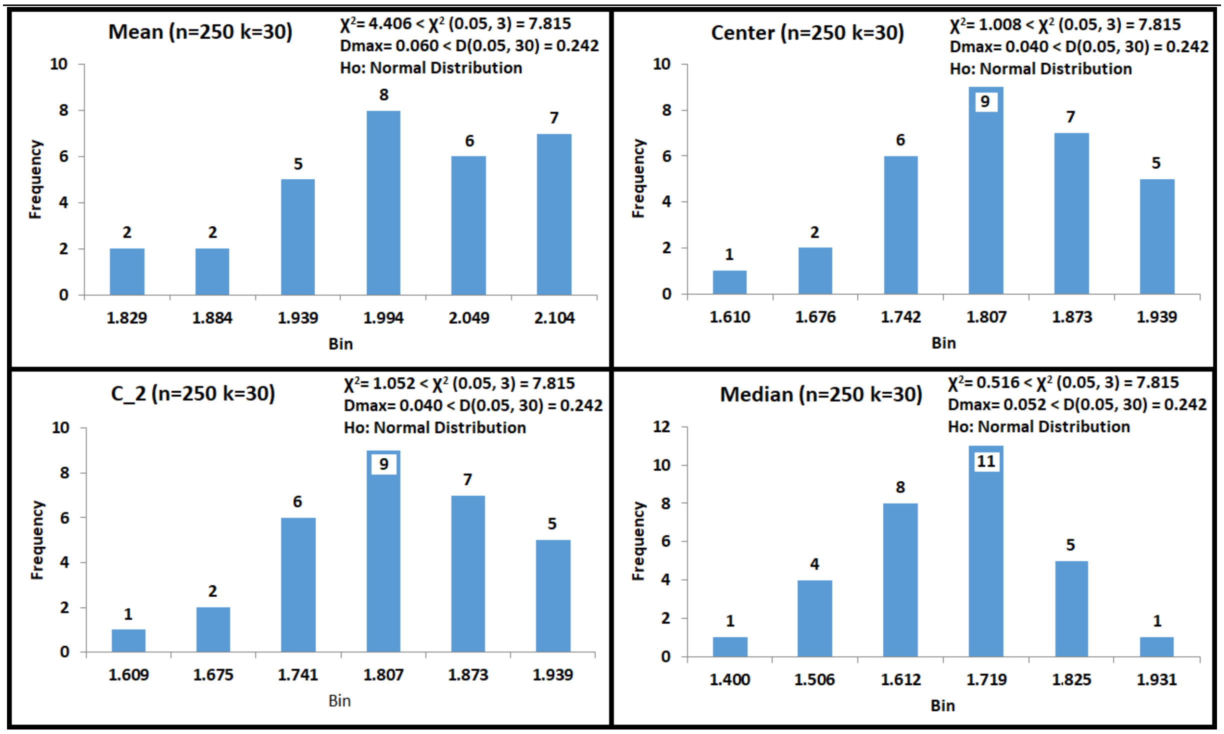

Figure 6: Histograms of the Means, Centers, C_2 and Medians based on k=30 samples of Gamma (2, 1) distributed random numbers of sizes n=250.

Additional data about standard errors of the means, the centers and the C_2 statistics of the independent samples, which distribution are illustrated by Figures 3-6, are given in Table 1 below.

Table 1

Observed and Expected Standard Errors of the Mean, the Center and the C_2 statistic of Gamma (2, 1) distributed random sample of size n

n	Observed St. Errors			Expected St. Errors		Observed-Expected	
	Mean	Center	C_2	Mean	Center	Mean	Center
10	0.420	0.284	0.251	0.447	0.304	-0.028	-0.020
30	0.266	0.180	0.174	0.258	0.175	0.008	0.005
50	0.193	0.133	0.131	0.200	0.136	-0.007	-0.003
100	0.144	0.097	0.096	0.141	0.096	0.003	0.001
150	0.116	0.080	0.079	0.115	0.078	0.001	0.002
200	0.102	0.069	0.069	0.100	0.068	0.002	0.001
250	0.090	0.061	0.061	0.089	0.061	0.000	0.000

The observed standard errors of the analyzed measures of central tendency of the independent samples show:

- The variance of the Gamma (2, 1) distribution based on equation (5) and weights regarding the center and C_2 statistics is finite and $\breve{\sigma}^2 \approx 0.923$.
- The variance of the Gamma (2, 1) distribution based on the mean, that is to say calculated by the use of equal weights in equation (5), is finite and $\sigma^2 = 2.000$. Therefore, this variance presents the spread of the Gamma (2, 1) distributed data more than twice widely. The reason is the equal weighting of all variables and as a result the overweighting of the right-tailed variables in this right-skewed distribution.

Conclusions

According to the results presented above, the following conclusions can be made:

- Regarding the Gamma (2, 1) distribution, which is a heavy right-skewed distribution, the center and the C_2 statistics have a finite expectation E $[\breve{X}] \approx 1.814$ and a finite variance $\breve{\sigma}^2 \approx 0.923$ as the size of the analyzed sample n ⟶ ∞.
- When the sample size n ≥ 30 equation (3) is valid, that is to say, the central limit theorem is valid in the matter of these statistics. According to the results given above, this statement is true even for n ≥ 10.
- The expectation of both the center and the C_2 statistics E $[\breve{X}]$ is more close to the expected value of the median of the Gamma (2, 1) distributed samples, which is approximately 1.678 than that of the mean E $[\bar{X}] = 2.00$. If one generates a Gamma (2, 1) distributed sample of size n=20000, they will find that the mean is approximately the 60th percentile. The center is approximately the 54th. This fact raises the question if the arithmetic mean is the most suitable measure of central tendency.
- The standard error of both the center and the C_2 statistics is 1.47 times less than the standard error

of the mean of a Gamma (2, 1) distributed sample, i.e. $\breve{\sigma}^2/\sigma^2 = 0.923/2.00 \approx 0.4615$. Thus, the relative efficiency of the center and the C_2 statistics over the mean is 46.15%. Therefore, both the center and the C_2 are more efficient estimators of central tendency than the arithmetic mean [6].

- If the size of a data set is getting bigger, the values and standard errors of the C_2 and the center will be getting closer. According to the results given above, when the size of a sample n ≥ 100 we can use its center statistic instead of its C_2, because of the lower computational cost of the first one.

References

1. Berg, Christian & Pedersen, Henrik L. (2006). The Chen–Rubin conjecture in a continuous setting. Methods and Applications of Analysis, 13 (1): 63–88. https://dx.doi.org/10.4310/MAA.2006.v13.n1.a4
2. Cvetkov V. (2022). ALTERNATIVE MEASURES OF CENTRAL TENDENCY. Deutsche Internationale Zeitschrift Für Zeitgeössische Wissenschaft, 38, 4-8. https://doi.org/10.5281/zenodo.7002877
3. Dekking, F. M., Kraaikamp, C., Lopuhaä, H. P., and Meester, L. E. (2005), A Modern Introduction to Probability and Statistics, London, Springer – Verlag.
4. https://en.wikipedia.org/wiki/Central_limit_theorem, (visited on 14.01.2023)
5. https://en.wikipedia.org/wiki/Gamma_distribution, (visited on 14.01.2023)
6. Montgomery, Douglas C.; Runger, George C. (2014). Applied Statistics and Probability for Engineers (6th ed.). Wiley. ISBN-13 9781118539712.
7. https://center-based-statistics.com/html/download.php, (visited on 14.01.2023)