applied sciences

MDPI

Article

A Simulation-Based Testing of Difference in the Means of Gamma-Distributed Positive Quantities

Filip Tošenovský

Department of Quality Management, Faculty of Materials Science and Technology, VSB—Technical University of Ostrava, 17. Listopadu 15/2172, 708 00 Ostrava, Czech Republic; filip.tosenovsky@vsb.cz

Featured Application: Applications can be found in industries, finance, economics, health sciences, and other sectors where the interest is centered on detecting a systematic difference in positive quantities, variables, or measures. This concerns many situations occurring in natural sciences, social sciences, and others. Some examples of real-life applications are presented in the introductory part of this text.

Abstract: This paper presents a simulation-based testing procedure that can be easily applied by practitioners who try to determine whether two gamma-distributed variables have the same expected values. From both theoretical and practical points of view, the gamma distribution and the testing in question have been of interest for some time given the many applications they can be used for, which include problems in the fields of economics, industrial statistics, life sciences, and others. The efforts to achieve the stated statistical objective have been focused throughout the years either on performing nontrivial, approximating mathematical steps or on simulations based on resampling techniques of various kinds. This text works with simulations that try to get closer to the true distributions of the quantities of interest so that a test can be designed rather than using samples generated out of samples, as the resampling techniques perform this by taking the initial samples for an approximation of the populations. The results presented in this text were validated, and they were also compared to other methods where possible. The resulting technique was looked upon as a complement to all the techniques that have been presented on this subject. The major advantage of the proposed procedure is seen in its simplicity. Since simulations are the basis for the presented conclusions, the results are unsurprisingly not as general as what could be achieved by exact mathematical deduction, but they do cover a reasonable range of situations that can serve as a basis on which to analogously build further research if desired.

Keywords: gamma distribution; two-sample test for means; simulation; type I and II errors

check for updates

Citation: Tošenovský, F. A Simulation-Based Testing of Difference in the Means of Gamma-Distributed Positive Quantities. *Appl. Sci.* **2023**, *13*, 9497. https://doi.org/10.3390/app13179497

Academic Editor: Andrea Prati

Received: 16 May 2023
Revised: 17 August 2023
Accepted: 17 August 2023
Published: 22 August 2023

1. Introduction

In many areas of human presence, situations arise when a subject is interested in comparing different scenarios they can choose from, the scenarios being related to their activity. The objective is to opt for the right scenario to ensure that a proper strategy is pursued in the future. The output of each scenario is often expressed by a measure that can take on only positive values. The measure allows one to reflect on the consequences of the scenarios so that a proper decision can ultimately be made in advance. However, making that right decision is often not a straightforward task despite the introduced measure because each scenario manifests itself in the end with more than one value of the measure due to its dependence on random events, and so a more rigorous, statistical approach should be adopted prior to making the selection. Let us look at some important real-life examples of such situations.

In industries, a quality characteristic of a product can be monitored, for instance, and two different production technologies might be capable of fabricating the product.

The company's management wants to know which technology leads to a more reliable output as far as the stability of the observed characteristic is concerned. To tackle this problem, a variability of the characteristic can be measured for each production batch to assess the reliability or stability in question. The variability will generally be different for each batch and can therefore be regarded as a random variable, variable X for production technology 1 and variable Y for production technology 2. The question then is whether the two production reliabilities are the same in the sense that their expected values are the same: $E(X) = E(Y)$. If this is true, the company's management can go for the cheaper of the two technologies since the more expensive one will not bring anything new regarding reliability. The ultimate result of this decision will then be financial savings.

In finance, an insurance company may strive to decide which of two specific and different insurance products it should offer to its customers. It will naturally prefer the one that exposes it to lesser risk, the risk being measured by how much money it may need to pay out within its insurance coverage liabilities. Here, the amount paid out due to an event covered by an insurance product represents a positive random variable, the value of which differs from one event to another, and so the company can ask itself the question, "Do we spend, on average, the same amount of money on the two insurances?"

An investor cares about the yield of a financial investment of a certain type; the yield is always observed after a specified period of time. The yield history has so far been positive values for the investments. Such an investor may look at the past yields to decide which investment to go for by answering the question "Do the various investments provide the same average yield for the observed time period?"

In seismology, the strength of tremor is certainly one of many positive variables of interest, and it can be accompanied by the question: "Do the two analyzed geographic regions experience earthquakes of the same magnitude, on average?"

The examples above work with two positive random variables X and Y, and it could also be said without assuming too much that one of their distinctive features is that they take on lower values probably more often than very high values. This is true for all the given examples. Thus, a right-skewed probability distribution, such as the gamma distribution [1], might be an appropriate probabilistic model for the description of the variables. If such a model is acceptable, the uttered questions come down to comparing the expected values of two gamma-distributed random variables. The standard procedure that usually follows in these cases takes the form of an assessment of whether the expression $\overline{X} - \overline{Y}$, or a transformation thereof, is too far from a set constant, usually zero. This is known in statistics as the two-sample test of the hypothesis $H_o: \ E(X) = E(Y)$. Such a technique, however, is no simple task when the two averages are calculated from samples drawn from gamma populations because for the procedure to be applicable, the probability distribution of $\overline{X} - \overline{Y}$ must be known under H_o so that the important judgment can be made on whether $\overline{X} - \overline{Y}$ has happened to take on a value that is very unlikely under H_o.

If $X = (X_1, X_2, \ldots, X_n)$ is a random sample from the $\Gamma(k = shape, \theta = scale)$ gamma distribution, the characteristic function of each X_j is

$$\psi_{X_j}(t) = (1 - i\theta t)^{-k}, \tag{1}$$

and so

$$\psi_{\sum X_j}(t) = (1 - i\theta t)^{-kn}, \tag{2}$$

$$\psi_{\overline{X}}(t) = (1 - it\theta/n)^{-kn}, \tag{3}$$

as is known from the theory of characteristic functions [2–4]. In other words, $\overline{X}$ follows the distribution $\Gamma(kn, \theta/n)$. Taking the analogous result for $\overline{Y}$, $\overline{X} - \overline{Y}$ is the difference of two independent gamma-distributed random variables if the two samples are independent. When the assumption of independence holds, it is straightforward to derive the characteristic function of the difference, but it is not clear anymore what distribution it belongs to. The

problem of how the difference is distributed under H_o could also be solved by deriving the density of $-\overline{Y}$, applying the transformation theorem [5], since the general form of the density of $\overline{Y}$ is known, and then one can use the convolution theorem [3,6] to derive the density of $\overline{X} + \left(-\overline{Y}\right)$ and hence its distribution. But the result for this density in the form of an integral containing two nontrivial densities and their multiplication is too complex to evaluate in practice, even in theory. This means it is difficult to derive an exact test for the stated purpose since the distribution of $\overline{X} - \overline{Y}$ under H_o derived from its complex density is not very clear. On the other hand, it is certainly possible to analyze the distribution of $\overline{X} - \overline{Y}$ under H_o through simulations and, based on the simulation results, to formulate recommendations on how $H_o : E(X) = E(Y)$ could be tested. This is especially beneficial when dealing with smaller data samples since for very large data sets, the problem of evaluating the significance of the difference $\overline{X} - \overline{Y}$ can be resolved with the central limit theorems and the associated higher-precision normality approximations.

The parameters of the gamma distribution have been under the spotlight for quite some time now given the increasing practical importance of this stochastic model. Yet the interest has not always been focused on two-sample testing, and when it was, as well as when it was not, the scientific output either bore greater complexity for practitioners or did not always deliver precise results simply because the form of the gamma distribution was too involved to do so. To provide some excursion into the past in this regard, Grice and Bain [7] proposed an approximate test for the mean of a gamma distribution with both its parameters unknown, which is a different problem than the one analyzed in this text. Shiue and Bain [8] then proposed an approximate test for testing the equality of the scale parameters of two gamma distributions with unknown and equal shape parameters. Shiue et al. [9] also proposed an approximate test for testing the equality of the means of two gamma distributions with unknown and unequal shape parameters. In the 1990s, Tripathi et al. [10] suggested an asymptotic test, that is, only an approximate procedure for many small-sample situations for the parameters of two gamma distributions. Bhaumik et al. [11] proposed testing procedures for each parameter of the gamma distribution individually, that is, for the scale parameter, shape parameter, and also the mean of the distribution (a different problem). Chang, Lin, and Pal [12] presented a bootstrap-based procedure to test the equality of more gamma distribution means. This is a paper of interest that will be used for comparative analysis in this text. Other mentions can also include the possibility of using nonparametric tests for the subject of interest, such as the Wilcoxon rank-sum test, approximate in nature for small samples, however, or Thiagarajah's [13] test of the homogeneity of gamma distributions, based on combining the likelihood ratio and score statistics: Their behavior is yet again known only for large samples.

It seems that some theoretical outcomes are available even though they result either from mathematical approximations or resampling techniques [12,14], such as the bootstrap. It is worth noting that resampling techniques also rely on large-sample properties. It can also be seen that the two-sample testing of means has not always been the subject matter under scrutiny.

This text adopts a different and more "hands-on" approach compared to the two just mentioned and focuses on smaller- and medium-sized data samples for which the testing problem is more pressing. The findings presented in this text are based on simulations that try to identify properties of the true distribution of $\overline{X} - \overline{Y}$ under H_o, striving to avoid various approximations that result either from complicated mathematical operations or potentially also from resampling techniques. Additionally, one of the major objectives of this paper is to provide as simple a testing technique as possible so that it can be used by practitioners without the need to code ad-hoc simulations or comprehend and evaluate more complex mathematical theories and the implied algorithms. As with any simulation-based research, the limitation of the presented technique lies in the fact that not all theoretically conceivable combinations of the gamma parameters can be handled while deriving the testing procedure, naturally, since there is an infinite number of them, but many of them are handled in this text. Regarding the number of parameter combinations

covered in this text, a certain aggregation of the results had to be adopted in the end when formulating the final recommendations due to a large number of simulation-based outcomes. The individual results are available in the attached datasets, however. The logic of the entire procedure is discussed below in the methodology section. Although the aggregation of results can be looked at as a way to blur back the detailed and more precise simulation outcomes, the resulting testing procedure based on the aggregation is still subject to verification so that its potential is indicated.

Last but not least, the aim of the paper is not to criticize the techniques presented so far in the literature—quite the opposite. This paper is regarded as their complement, which, after all, will be obvious once the method is validated, as it shows good potential in many situations, while in others, it may be less desirable.

In summary, a simpler method with desirable statistical properties for many smaller- and medium-sized samples and a reasonable range of the gamma-distribution parameters is the objective of this paper.

2. Materials and Methods

To design the required test for a number of situations, it is not necessary to know the true probability distribution of the test statistic $\overline{X} - \overline{Y}$ under $H_0 : \ E(X) = E(Y)$ in full. All that is necessary is a proper percentile, or percentiles, of that distribution for accepting or rejecting the null hypothesis. This is in line with the theory of hypothesis testing, the percentiles being also known as the "critical values". The methodology used in this text therefore aims at running R-coded simulations that try to identify the necessary percentiles of the distribution of $\overline{X} - \overline{Y}$ under H_0 for various sample sizes and gamma distributions from which many realizations of the sample averages $\overline{X}$, $\overline{Y}$ must be calculated. The distributions do not necessarily have the same parameters, but their expected values are the same for the null hypothesis to hold. Further, since it is inconvenient to have a lot of percentile estimates for different set-ups of the two gamma distributions, a simple formula is also pursued, which could be used in practice to calculate back the percentiles identified by the simulations so that the testing can be performed. However, since the formula at first relies on an unknown population quantity, the suitability of such a formula must be verified once the unknown quantity is replaced with an appropriate estimate. This verification is implemented in this text and takes the form of estimating the type I error probability of the formula-based procedure, as well as its test power. Hence, the entire analysis consists of several steps.

In the first step, the analysis focuses on detecting the 2.5% and 97.5% percentiles p_1 and p_2 of the distribution that $\overline{X} - \overline{Y}$ follows under H_0. This can be performed with large enough simulations that lead, for each considered set-up of the two gamma distributions and each considered sample size, to many realized values of the difference under H_0, giving a good estimate of the $\overline{X} - \overline{Y}$ distribution and its percentiles under the null hypothesis. Now, if it was known in practice that the found p_1, p_2 values are the two percentiles of interest under H_0, the null hypothesis could be tested: If $\overline{x} - \overline{y}$ calculated in practice from a single realization of the sample averages is below p_1 or above p_2, the hypothesis is rejected. In all other cases, it is accepted. The problem is that this is never known because the true gamma distributions are unknown, and so the question arises which pair p_1, p_2 of all the pairs found through the simulations should be actually used for the test.

To answer this pressing question, one could think and proceed as follows: The normal and other distributions are known for the two-sigma rule or a similar rule that states that once the population standard deviation is multiplied by two and minus two or by a different number and these multiples are added to the population mean, two percentiles are obtained with the property so that it is almost certain that the random variable will take on a value lying between the two percentiles. These rules can be described by the following formula:

$$p = d \cdot var^{1/2}, \tag{4}$$

where *var* is the population variance, and d is the proper multiple. Applying this formula to the distribution of $\overline{X} - \overline{Y}$ under H_0, *var* is known by design, while the percentiles p are obtained via the simulations, so the multiple(s) d can be calculated from (4) through the simulations, as well. Of course, using Formula (4) to calculate the percentiles in practice based on such research will at the moment still not help since not knowing the gamma distributions the samples came from means not knowing which d to choose, let alone the fact that *var* is also unknown. The problem here would just be shifted from the unknown p to the unknown d even if *var* was known. This situation can be helped, however, if it turns out that the multiple d happens to manifest a certain stability across various gamma-distribution set-ups and/or various sample sizes. If this property is present, the knowledge of what distributions the samples came from is not so relevant anymore since d is more or less the same regardless of the distributions. In this fortuitous instance, the only remaining problem is then the unknown population variance *var*, which must be estimated. This logic leads to the suggestion of using the formula $p = d \cdot \widehat{var}^{1/2}$ for testing purposes as long as d does have the property just described. Finding the ds for various situations and examining their stability ends the second part of the analysis.

Since $\widehat{var}$ is used instead of *var* in the percentile-generating formula, the potential of this method should be verified as well even if d is relatively stable. This constitutes the final stage of the analysis. The verification takes the form of analyzing the error rates of the method [15] and comparing it with other procedures.

Regarding the technical details, using the common symbols $\Gamma(k_1,\ \theta_1)$ and $\Gamma(k_2,\ \theta_2)$ for the two gamma distributions worked with, the distribution of the variable $\overline{X} - \overline{Y}$ is analyzed under $H_o:\ E(X) = E(Y)$ or equivalently [1]:

$$H_o : k_1\theta_1 = k_2\theta_2. \tag{5}$$

The following distribution set-ups are analyzed with the distribution parameter combinations grouped into several "scenarios":

$$\text{scenario } 1 - k_1, k_2 \in K = \{1, 2, \ldots, 10\},\ \theta_1 \in M = \{1, 2, \ldots, 30\}, \tag{6}$$

$$\text{scenario } 2 - k_1, k_2 \in L = \{11,\ 12, \ldots,\ 20\},\ \theta_1 \in M, \tag{7}$$

$$\text{scenario } 3 - k_1, k_2 \in N = \{21,\ 22, \ldots,\ 30\},\ \theta_1 \in M, \tag{8}$$

$$\text{scenario } 4 - k_1 \in K,\ k_2 \in L,\ \theta_1 \in M, \tag{9}$$

$$\text{scenario } 5 - k_1 \in K,\ k_2 \in N,\ \theta_1 \in M, \tag{10}$$

$$\text{scenario } 6 - k_1 \in L,\ k_2 \in N,\ \theta_1 \in M. \tag{11}$$

All parameter combinations work with all sample sizes from the set $n \in \{30,\ 40,\ ...,\ 80\}$. Both samples always have the same size. This is usually not a problem to achieve in practice as long as a controlled experiment is possible. Since the focus is on the percentiles of $\overline{X} - \overline{Y}$ under H_o, a specific sample size is chosen and scenario-based k_1, k_2, θ_1 parameters are set up, determining also $\theta_2 = k_1\theta_1/k_2$ so that the null hypothesis holds whereupon thousands of realizations of the statistic $\overline{X} - \overline{Y}$ are generated from two independent random samples drawn from the set-up gamma distributions. This experiment suggests to an acceptable level of detail the distribution of the difference under H_o. Once the distribution is detected,

its 2.5% and 97.5% p percentiles are calculated and hence also the multiples $d = p \cdot var^{-1/2}$ where

$$var = var(\overline{X}) + var(\overline{Y}) = n^{-1}\left(k_1\theta_1^2 + k_2\theta_2^2\right) \tag{12}$$

is calculated directly since its form is known within the given parameter setting [1].

At the end of these calculations, one should rather talk about entire distributions of the identified ds, one d having been calculated for the 2.5% percentile and the other for the 97.5% percentile, because each d is generally different for each parameter set-up and sample size. Therefore, the findings presented in this text are in the form of "the average lower d" and "the average upper d", each average calculated for a given scenario and sample size. The average results from the aggregation of the within-scenario results. At this stage, the averaging, which by its very nature always hides a large amount of information, is not an imprecision in the whole procedure, as this stage of analysis aims primarily to suggest which multiple d could be used in the testing. Its suitability is only scrutinized later.

Once the average ds are identified, the focus can be shifted to the formula

$$d \cdot \widehat{var}^{1/2}, \tag{13}$$

where

$$\widehat{var} = n^{-1}(\widehat{var}(X) + \widehat{var}(Y)), \tag{14}$$

and

$$\widehat{var}(X) = (n+1)^{-1}\sum_i \left(X_i - \overline{X}\right)^2,\ \widehat{var}(Y) = (n+1)^{-1}\sum_i \left(Y_i - \overline{Y}\right)^2. \tag{15}$$

The term $(n+1)^{-1}$ is used instead of the usual $(n-1)^{-1}$ since it gives the sample variance that optimizes (minimizes) the general MSE criterion [16] of estimate quality. As outlined, the property of interest will then be the frequency of cases when the testing procedure, based on the percentiles calculated from (13)–(15), incorrectly rejects H_o. This type I error frequency depends again on the specific set-up of k_1, k_2, θ_1, as well as the selected sample size. Hence again, within the given scenario and sample size, the average type I error probability is identified but this time also together with the 2.5% and 97.5% percentiles of the observed k_1, k_2, θ_1-dependent type I error probabilities so that the spread of the probabilities is seen, and the scenario-based aggregation does not mask the suitability of the test when it comes to its type I error. Each type I error is identified through thousands of employed tests.

At the final stage, the power of the test using the estimated percentiles is also the subject of interest. The six scenarios and the sample sizes are used again, but now, the null hypothesis does not hold, which is conveyed by introducing a constant to differentiate $E(X)$ from $E(Y)$: $k_2\theta_2 = k_1\theta_1 +$ const. Since the test power differs within a scenario due to different parameter combinations, given the sample size and constant, the average per-scenario and sample size power is considered. For each scenario–sample size combination, a specific and fixed constant is used. More interestingly, however, the method is also compared to another methodology regarding its test power. This is performed for specific parameter set-ups and sample sizes, and no averaging is involved.

If the method proves to be viable, the resulting algorithm can be described as follows:

1. Calculate the sample averages and variances for the two obtained random samples where the sample variances are evaluated according to (15).
2. Calculate $\hat{k}_1, \hat{k}_2, \hat{\theta}_1$ estimates: For instance, by estimating $E(X) = k\theta$ with $\overline{x}$ and $var(X) = k\theta^2$ with the sample variance s^2, the common estimates $\hat{\theta} = s^2/\overline{x}$ and $\hat{k} = \overline{x}^2/s^2$ are obtained for both samples. This helps detect which scenario is worked with.
3. Evaluate the expression $d\widehat{var}^{1/2}_{\overline{X}-\overline{Y}}$ where $\widehat{var}_{\overline{X}-\overline{Y}} = n^{-1}(\widehat{var}(X) + \widehat{var}(Y))$ using the two MSE-minimizing sample variances from step 1. n is the size of any of the two equally large samples; the $d\widehat{var}^{1/2}_{\overline{X}-\overline{Y}}$ is calculated twice with two generally different multiples d and the same $\widehat{var}^{1/2}_{\overline{X}-\overline{Y}}$ so that two percentile estimates $\hat{p}_{2.5}$ and $\hat{p}_{97.5}$ are

obtained. The multiples d are selected from the tables below according to the k_1, k_2, θ_1 scenario the practitioner is in, as suggested by the $\hat{k}_1, \hat{k}_2, \hat{\theta}_1$ estimates.

4. Perform the test: If $\hat{p}_{2.5} < \overline{x} - \overline{y} < \hat{p}_{97.5}$, the hypothesis of equal means is accepted; otherwise, it is rejected.

The next section shows the results. Where necessary, a comment is attached, although the contents are self-explanatory through the table headlines. The computer code for the simulations is in Appendix A of this paper, and the underlying data are stored in a referenced public depository.

3. Results

Given the findings contained in Tables 1–4, which show that the multiple d is stable across sample sizes and also across scenarios, as well as within each scenario, based on the underlying data that show the multiple moves within one-tenth to the left or right of its average, it can be selected to be used in the $d \cdot \widehat{var}^{1/2}$-based testing procedure. Once such a procedure is designed, its type I error should be analyzed. Before doing so, due to the stability of d, a further simplification of the testing procedure can be performed as follows: For the first 3 parameter scenarios, the lower $d = -1.96$ and the upper $d = 1.96$ are selected; for the fourth scenario, the lower $d = -1.92$ and the upper $d = 2.00$ (see Table 2) are used; for the fifth scenario, the lower $d = -1.91$ and the upper $d = 2.01$ (see Table 3) are selected; and for the sixth scenario, the lower $d = -1.95$ and the upper $d = 1.97$ (see Table 4) are chosen. The error probabilities of this procedure are contained in Tables 5–10.

Table 1. Averaged ds giving the 2.5%, 97.5% percentiles (k_1, k_2, θ_1 from scenario 1 here; for scenarios 2 and 3, the results are identical).

Sample Size n	Average Lower d	Average Upper d
30	−1.96	1.96
40	−1.96	1.96
50	−1.96	1.96
60	−1.96	1.96
70	−1.96	1.96
80	−1.96	1.96

Table 2. Average ds giving the 2.5%, 97.5% percentiles (k_1, k_2, θ_1 from scenario 4).

Sample Size n	Average Lower d	Average Upper d
30	−1.90	2.01
40	−1.91	2.01
50	−1.92	2.00
60	−1.92	2.00
70	−1.92	2.00
80	−1.93	1.99

Table 3. Average ds leading to the 2.5%, 97.5% percentiles (k_1, k_2, θ_1 from scenario 5).

Sample Size n	Average Lower d	Average Upper d
30	−1.89	2.03
40	−1.90	2.02
50	−1.91	2.01
60	−1.91	2.01
70	−1.92	2.00
80	−1.92	2.00

Table 4. Average ds leading to the 2.5%, 97.5% percentiles (k_1, k_2, θ_1 from scenario 6).

Sample Size n	Average Lower d	Average Upper d
30	−1.95	1.97
40	−1.95	1.97
50	−1.95	1.97
60	−1.95	1.97
70	−1.95	1.97
80	−1.95	1.97

Table 5. Type I error probability alpha for the test (lower $d = -1.96$, upper $d = 1.96$): average alpha (averaged across k_1, k_2, θ_1 from scenario 1) and the percentiles for alpha.

Sample Size n	Average Alpha	2.5%; 97.5% Percentiles
30	0.064	0.057; 0.070
40	0.059	0.053; 0.066
50	0.057	0.052; 0.064
60	0.057	0.050; 0.063
70	0.056	0.049; 0.062
80	0.055	0.049; 0.061

Table 6. Type I error probability alpha for the test (lower $d = -1.96$, upper $d = 1.96$): average alpha (averaged across k_1, k_2, θ_1 from scenario 2) and percentiles for alpha.

Sample Size n	Average Alpha	2.5%; 97.5% Percentiles
30	0.063	0.056; 0.069
40	0.059	0.053; 0.066
50	0.058	0.051; 0.064
60	0.056	0.050; 0.062
70	0.055	0.049; 0.062
80	0.050	0.049; 0.061

Table 7. Type I error probability alpha for the test (lower $d = -1.96$, upper $d = 1.96$): average alpha (averaged across k_1, k_2, θ_1 from scenario 3) and percentiles for alpha.

Sample Size n	Average Alpha	2.5%; 97.5% Percentiles
30	0.063	0.056; 0.069
40	0.060	0.053; 0.066
50	0.058	0.051; 0.064
60	0.056	0.050; 0.063
70	0.055	0.049; 0.062
80	0.055	0.048; 0.061

Table 8. Type I error probability alpha for the test (lower $d = -1.92$ upper $d = 2.00$): average alpha (averaged across k_1, k_2, θ_1 from scenario 4), and percentiles for alpha.

Sample Size n	Average Alpha	2.5%; 97.5% Percentiles
30	0.069	0.059; 0.090
40	0.064	0.055; 0.081
50	0.061	0.053; 0.076
60	0.060	0.052; 0.073
70	0.058	0.051; 0.070
80	0.058	0.050; 0.068

Table 9. Type I error probability alpha for the test (lower $d = -1.91$ upper $d = 2.01$): average alpha (averaged across k_1, k_2, θ_1 from scenario 5) and percentiles for alpha.

Sample Size n	Average Alpha	2.5%; 97.5% Percentiles
30	0.070	0.060; 0.092
40	0.066	0.056; 0.084
50	0.063	0.054; 0.078
60	0.061	0.053; 0.074
70	0.060	0.052 0.071
80	0.059	0.051; 0.069

Table 10. Type I error probability alpha for the test (lower $d = -1.95$ upper $d = 1.97$): average alpha (averaged across k_1, k_2, θ_1 from scenario 6) and percentiles for alpha.

Sample Size n	Average Alpha	2.5%; 97.5% Percentiles
30	0.064	0.057; 0.070
40	0.060	0.053; 0.067
50	0.058	0.051; 0.065
60	0.057	0.050; 0.063
70	0.056	0.049; 0.062
80	0.055	0.049; 0.061

It can be seen from Table 5 that for any considered sample size and almost (95%) any within-scenario 1 set-up of k_1, k_2, θ_1, the error probability is between the reasonable levels of 0.049 and 0.07. The average alpha is also not far from 0.05. Similar results follow.

The tables clearly suggest that in a majority of the analyzed situations (in 95% of them), the testing procedure returns a probability of type I error between approximately 0.05 and 0.09. In light of these probabilities, the method is quite usable. The average alphas also circle almost always around 0.06 regardless of the scenario and sample size and regardless of the specific within-scenario parameter combination.

Another set of results in Table 11 suggests which data sample sizes might be sought in order for the procedure to have a reasonable power of at least 0.7. Nevertheless, the table shows the average test power given the sample size, scenario, and constant c in $H_1 : k_2\theta_2 = k_1\theta_1 +$ c, the constant being fixed within a scenario. The power is averaged across different within-scenario k_1, k_2, θ_1 set-ups.

Table 11. Average test power per scenario "sc" and constant c: sc1~c = 20, sc2~c = 30, sc3~c = 38, sc4~c = 14, sc5~c = 16, sc6~c = 28. The average taken across within-scenario combinations of k_1, k_2, θ_1.

Sample Size n	Avg. Power (sc1; sc2; sc3)	Avg. Power (sc4; sc5; sc6)
30	0.59; 0.56; 0.56	0.60; 0.68; 0.59
40	0.64; 0.62; 0.61	0.65; 0.72; 0.65
50	0.68; 0.66; 0.66	0.69; 0.76; 0.69
60	0.72; 0.70; 0.70	0.72; 0.79; 0.73
70	0.74; 0.73; 0.73	0.75; 0.82; 0.76
80	0.77; 0.76; 0.76	0.77; 0.84; 0.79

It can be seen that for sample sizes of around 60 and the listed scenario-dependent values of the constant c, the average power equals or exceeds 0.7. Examples of situations when the mean difference equals the scenario-specific constant are for scenario 1 ... $[k_1, \theta_1] = [2, 3]$ and $[k_2, \theta_2] = [4, 6.5]$ or $[k_1, \theta_1] = [4, 2]$ and $[k_2, \theta_2] = [5, 5.6]$; for scenario 2 ... $[k_1, \theta_1] = [12, 14]$ and $[k_2, \theta_2] = [13, 15.3]$ or $[k_1, \theta_1] = [16, 11]$ and $[k_2, \theta_2] = [16, 12.9]$; for scenario 3 ... $[k_1, \theta_1] = [21, 25]$ and $[k_2, \theta_2] = [21.5, 26.2]$; and for scenario 6 ... $[k_1, \theta_1] = [23, 21]$, $[k_2, \theta_2] = [23, 22.3]$.

To conclude the results so far, a few remarks are due. First, when the shape parameter exceeds thirty, the normality approximation of the gamma distribution is generally expected

to perform. Therefore, the scenarios when both the shape parameters exceed thirty were not analyzed. Normality-based testing could be used instead. Secondly, in situations when the first scale parameter is above thirty, a change in the physical units in which the variables are observed to a higher order of magnitude squeezes this parameter for the transformed variable, so the results of the analysis expressed in different units may be applicable after the transformation as well. Thirdly, the sample sizes used range from thirty to eighty. Sample sizes of one hundred and more were not analyzed due to the possibility of a central limit-based normality approximation in the testing [17]. To check this possibility, an experiment was run: The Shapiro–Wilk test [18,19] was carried out for the sample average difference normality using all the considered combinations of k, theta, and the combinations set up so that H_0 held. For each combination, the p-value of the test was stored. Figure 1 shows the distribution of the p-values, a majority of them leading to accepting normality. Specifically, 95 percent of them are 0.011 or higher. In other words, for samples of n = 100, the test in a majority of k–theta cases accepted the normality of the sample average difference at the 1% sig. level when the samples came from gamma distributions with the same expected values. It seems that for samples of at least n = 100, the central limit theorem [20] can be used to test H_o instead.

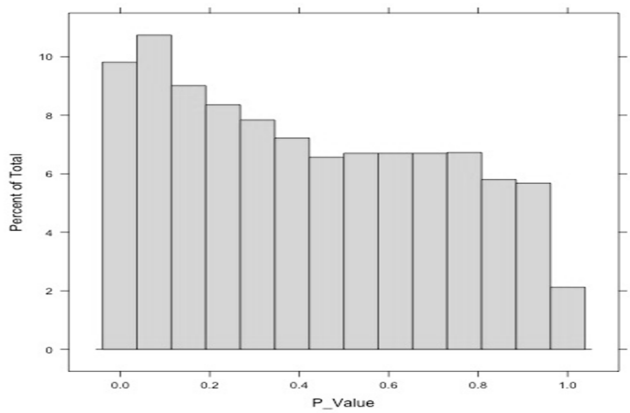

Figure 1. The Shapiro–Wilk test p-value distribution when testing normality of $\overline{X} - \overline{Y}$ under H_0 : $E(X) = E(Y)$; sample sizes of n = 100 used.

When it comes to comparing the method with others, a good place seems to be the methodology suggested by [12], which claims that it gives results "as good as, if not better than, the other methods discussed in (its) literature…". Even though a direct comparison is usually not possible in this case, since the cited paper uses mostly different parameter set-ups and smaller sample sizes, the sizes also being mostly unequal unlike the cases covered in this paper, a similar set-up can in some instances be detected. This allows us to shed some light on the functionality of the procedure suggested in this text. The authors of [12] use the delta symbol as the shape parameter (which equals k in this text) and the lambda letter as the scale parameter (here, the equivalent is the theta symbol). For the comparisons, the test power Table 1 published in [12], page 64, was selected, specifically the row where both sample sizes are equal to 25. This is the number closest to the sample sizes of 30 analyzed in this paper. The reproduction of their results is below in Table 12. Their methodology was also validated with simulations.

Table 12. A simulation-based estimate of the test power of the procedure proposed in [12]. Both samples have the same size of 25 datapoints, θ_1 = 1. The table is taken from [12].

$k_1 = k_2$	θ_2 = 1.25	θ_2 = 2.0	θ_2 = 5.0
2.5	0.243	0.969	1.0
5.0	0.417	0.999	1.0
7.5	0.566	1.000	1.0
10	0.688	1.000	1.0

Table 13 below is an equivalent of Table 12 constructed for the method proposed in this text. It contains a simulation-based estimate of its test power. The only difference in the design of the table is its second column where $\theta_2 = 1.5$ is used instead of $\theta_2 = 1.25$; the difference is explained below the table.

Table 13. A simulation-based estimate of the test power of the procedure proposed in this text. Both samples have the same size of 25 datapoints; θ_1 = 1. The simulations resulting in Table 13 can be viewed in their running mode at the address https://doi.org/10.6084/m9.figshare.23920197.v1 (accessed on 16 August 2023), the code applied is an extract of the code shown on the lower half of page 15. Major differences between Tables 12 and 13 are highlighted with bold lettering.

$k_1 = k_2$	θ_2 = 1.5	θ_2 = 2.0	θ_2 = 5.0
2.5	**0.05**	**0.41**	99
5.0	**0.25**	0.91	1.0
7.5	0.51	0.99	1.0
10	0.73	1.00	1.0

Given that $E(X) = k_1\theta_1 \neq E(Y) = k_2\theta_2$ in the test power simulations, the comparison of Tables 12 and 13 shows that the method of [12] is better when the differences in the means $E(X)$, $E(Y)$ of the two gamma distributions are rather small because the method of [12] is able to detect smaller differences in the means more often with its greater power as long as the simulation results in [12] are correct. In such instances, the method proposed here needs a somewhat larger difference in the means to reach similar test powers (for instance, the levels 0.51 and 0.73). For medium- and larger-sized differences, however, the methods seem to have the same test power—the best possible, nearing the level of 1. This is the case for population mean differences of the magnitude given by $\theta_2 \geq 2$ and $k_1 = k_2 \geq 5$, as the figures show. As an example, related to smaller differences, for $k_1 = k_2 = 7.5$ and $\theta_2 = 1.25$, the difference in the means is $|k_1\theta_1 - k_2\theta_2| = |7.5{\cdot}1 - 7.5{\cdot}1.25| = 1.875$, whereas with $\theta_2 = 1.5$, the difference is 3.75. These differences lead to similar test powers between 0.5 and 0.6 for both methods, as the tables show. This suggests that for not too small of differences in the means (a difference of at least five based on the tables), the occurrence of which can be assessed with parameter estimates, the method proposed here is quite usable, as its test power seems similar to other methods; it is very high, yet the proposed procedure is far simpler. Regarding small differences, they must be admittedly to some degree more discernible for the proposed method to be more competitive. Otherwise, the only limitation of the proposed method, as already outlined in the methodology section, is that its properties relate at the moment only to the parameter scenarios considered in the analysis of this paper. This is not the case for methods based on mathematical derivation, the validity of which is universal, unless mathematical restrictions are placed on them too. However, the presented methodology coupled with the attached computer code may be used for researching other scenarios as well, including even smaller sample sizes than thirty datapoints.

Another comparison using the same computer code may be performed for cases when the scale parameters k_1, k_2 differ, which is also considered in [12], in Table 5. The comparison, shown in Table 14 below, is again not direct, as the samples in [12] are equal to 10 for both data samples, whereas the smallest sample size considered in this

text is 30. Thus, the comparison puts the method of [12] at a disadvantage. Even so, the results in Table 14 show two facts: (a) the method of [12] unsurprisingly does not always shine despite its sophistication (test powers well below 0.8), and (b) the simple method proposed here can still perform well as long as the differences in the means are not diminutive. Further, the performance of any method measured by its power will depend on the parameter combination, which is why the resulting per-scenario average power of the method proposed here is not in the range of, say, 0.8–0.9 but is around 0.7 or slightly more for a distinct enough mean difference and not too small of sample sizes, as shown in Table 11. Clearly, there are parameter combinations within each scenario that bring trouble to the proposed method. The methodology in [12], unfortunately, does not employ the number of parameter combinations worked with in this paper to show how its method performs in more situations. Nevertheless, as shown in Table 14, there will be parameter combinations that do not particularly please the method proposed in [12] either—see, for instance, the last row of Table 14.

Table 14. Comparison of test powers for the method given in [12] and the method proposed in this text (the last column). For the method in [12], the sample sizes are both equal to 10, and the powers are taken from [12]. For the method proposed here, the sample sizes are both equal to 30.

k_1	k_2	θ_1	θ_2	Δ in Mean	Power [12]	Power
1	2	1	0.75	0.50	0.679	0.024
1	2	1	1.00	1.00	0.410	0.190
1	2	1	1.25	1.50	0.225	0.490
1	5	1	0.75	2.75	0.999	0.999
1	5	1	1.00	4.00	0.994	1.000
1	5	1	1.25	5.25	0.965	1.000
2.5	5	1	0.75	1.25	0.967	0.150
2.5	5	1	1.00	2.50	0.768	0.800
2.5	5	1	1.25	3.75	0.450	0.999
5	5	1	0.75	1.25	0.278	0.050
5	5	1	1.00	0.00	0.170	0.000
5	5	1	1.25	1.25	0.900	0.015
5	10	1	0.75	2.50	0.998	0.530
5	10	11	1.00	5.00	0.967	0.990
5	10	1	1.25	7.50	0.574	1.000

4. Examples of the Procedure

Three examples are now shown to demonstrate the simple technique. The first two examples used simulated data, and the third used real-life datasets.

First, let a random sample 1 of size 40 for a variable X from $\Gamma(k_1, \theta_1) = \Gamma(2, 4)$ be 7.4323, 8.7579, 3.8367, 24.3859, 4.1931, 1.4071, 2.1765, 13.8280, 2.1577, 8.9442, 3.6841, 3.6455, 5.8649, 2.2743, 3.5843, 4.0561, 8.9005, 10.6677, 12.9202, 21.4182, 4.9585, 7.7933, 1.4325, 2.4112, 11.1183, 11.2643, 7.2044, 6.7496, 7.0007, 2.2562, 2.7139, 9.4639, 4.3929, 9.1965, 16.9821, 4.4714, 12.9340, 7.5178, 13.2414, 12.3011.

Let a random sample 2 of size 40 for a variable Y from $\Gamma(k_2, \theta_2) = \Gamma(4, 2)$ be 3.5289, 7.5509, 15.9686, 2.9356, 4.2433, 22.0422, 12.0441, 3.2122, 6.0766, 3.7120, 17.4808, 14.7796, 17.8730, 3.8967, 7.1522, 10.7450, 7.7302, 9.9103, 6.2068, 3.4237, 12.2881, 3.8782, 5.0942, 13.1173, 4.6774, 7.0516, 11.4397, 4.5981, 6.1976, 12.9560, 15.7683, 9.1313, 6.0904, 3.1014, 4.2042, 13.4389, 4.5231, 4.5774, 10.8752, 5.0082.

The original values were rounded off to four decimal places. Their visualization is shown in Figure 2.

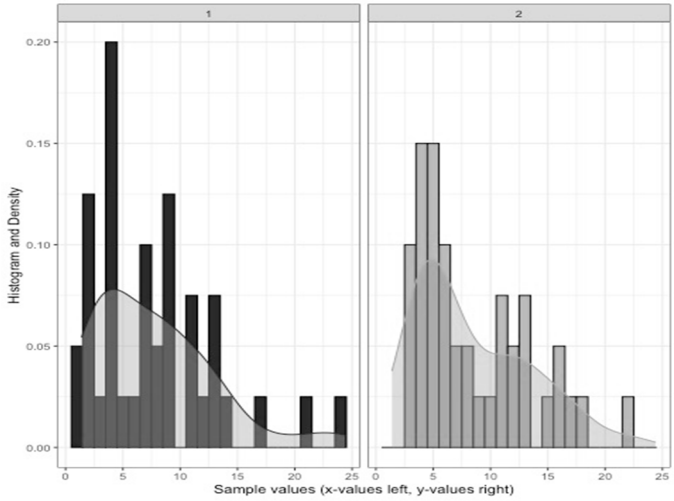

Figure 2. Visualization of the two data samples used for the first example.

The two distributions have the same means of 8; hence, the null hypothesis holds, and the variances are 32 and 16, respectively. The test criterion in abs. value is $|\overline{x} - \overline{y}| = |7.7385 - 8.4634| = 0.725$, the percentile-based critical value equals $1.96\widehat{var}^{1/2} = 1.96\sqrt{40^{-1}[\widehat{var}(X) + \widehat{var}(Y)]} = 1.96\sqrt{40^{-1}[27.206 + 23.587]} = 2.208$, and the hypothesis is correctly accepted. The variances $\widehat{var}(X)$, $\widehat{var}(Y)$ were calculated with the $(n+1)^{-1}$ divisor. The d multiples were taken as ± 1.96 since scenario 1 is observed here.

In the second example, let a sample 1 of 40 for a variable X from $\Gamma(k_1, \theta_1) = \Gamma(3, 4)$ be 20.2621, 5.0067, 6.8178, 17.9724, 5.7980, 47.9481, 24.0818, 16.9974, 15.6469, 2.2784, 7.0815, 16.5170, 12.8472, 20.7937, 11.4743, 10.6894, 10.7756, 15.0273, 6.6600, 21.2111, 10.5329, 18.7154, 12.5374, 8.01664, 5.4640, 26.8381, 16.0145, 20.7234, 5.3591, 11.9667, 14.6437, 8.9770, 9.0279, 10.5961, 17.2995, 9.2122, 27.5701, 9.9853, 15.5422, 2.3258.

Let a sample 2 of size 40 for a variable Y from $\Gamma(k_2, \theta_2) = \Gamma(4, 2.5)$ be 16.7113, 9.3561, 19.1426, 6.7874, 8.5002, 7.2641, 29.7096, 12.4152, 20.7611, 4.3977, 6.2623, 9.2883, 7.7831, 16.7462, 3.8046, 7.8470, 4.5086, 11.9305, 18.3177, 12.5083, 4.6306, 5.6554, 17.0151, 6.7073, 6.0264, 10.4029, 15.1457, 14.8109, 7.0398, 3.3200, 10.5470, 5.5386, 10.2651, 6.7952, 17.0514, 3.7160, 4.1363, 13.9024, 8.0307, 12.7817. The data visualization is in Figure 3.

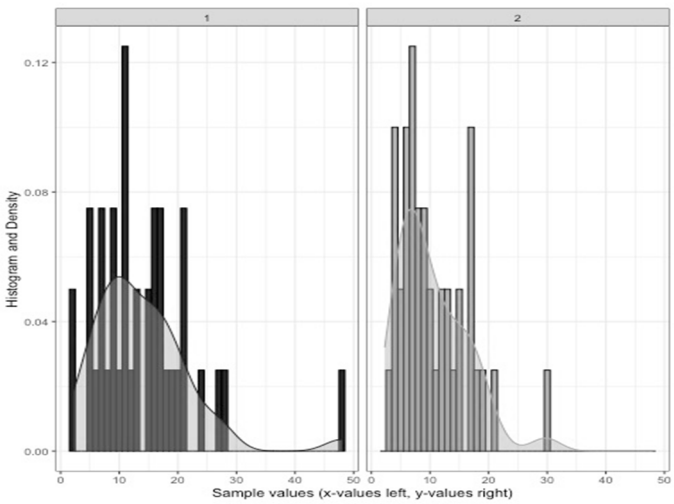

Figure 3. Visualization of the two data samples used for the second example.

The population means are 12 and 10 this time, respectively, so H_1 holds. The variances are 48 and 25, respectively. Now, $|\overline{x} - \overline{y}| = |13.931 - 10.439| = 3.492$, $1.96\widehat{var}^{1/2} =$

$1.96\sqrt{40^{-1}[\widehat{var}(X)+\widehat{var}(Y)]} = 1.96\sqrt{40^{-1}[67.482+31.899]} = 3.089$, and the null hypothesis is correctly rejected.

Taking real-life data, the interest is now focused on whether there was a significant across-time difference in the amount of July rainfall in all of the Czech Republic, the first data sample spanning the time period 1979–1999 and the second the years 2000–2020. Both data samples are of size n = 21. The officially published data are available at https://www.chmi.cz/historicka-data/pocasi/uzemni-srazky (accessed on 16 August 2023). The data published by the Czech Hydrometeorological Institute are (in millimeters and chronologically) 66, 154, 163, 72, 29, 72, 79, 74, 86, 88, 71, 34, 76, 69, 96, 56, 61, 89, 204, 93, 86, for 1979–1999; 121, 119, 87, 81, 64, 131, 38, 84, 86, 111, 118, 145, 113, 34, 102, 36, 115, 90, 42, 58, 61 for 2000–2020. Employing the gamma_test function in R, the procedure returned the p-values 0.2956 and 0.316, respectively, for the two samples, allowing us to model them with gamma distributions. Since $\overline{x}_1 = 86.57$, $s_1^2 = 1685.46$, the parameter estimates for the first-sample gamma distribution are $\hat{k}_1 = 86.57^2/1685.46 = 4.44$, $\hat{\theta}_1 = 1685.46/86.57 = 19.47$. For the second, the estimates are $\overline{x}_2 = 87.43$, $s_1^2 = 1131.76$, $\hat{k}_2 = 87.43^2/1131.76 = 6.75$, $\hat{\theta}_2 = 1131.76/87.43 = 12.94$. The sample variances were calculated with the term $(n-1)^{-1}$ for the moment. This brings the analyst to scenario 1, for which the multiples to calculate the critical percentiles are equal to -1.96 and 1.96. The test criterion in absolute value is $|\overline{x}_1 - \overline{x}_2| = 0.86$, and by (13)–(15), $1.96\sqrt{21^{-1}\left[(n-1)/(n+1)\left(s_1^2+s_2^2\right)\right]} = 21.65$ can be used as the critical value of the test. Since the test criterion in absolute value is much lower than the critical value, or equivalently, the test criterion lies in the interval $\left(-1.96\widehat{var}^{1/2},\ 1.96\widehat{var}^{1/2}\right)$, the null hypothesis of no shift in the average amount of July rainfall across time is accepted.

5. Conclusions

This paper presented a simple method that can be used to test the statistical significance of the difference between two independent gamma-distributed random variables in a small-to-medium-sized data setting. There are many real-life situations when such a technique might be useful provided the objective is to discern situations or scenarios from one another, the scenarios being evaluated or described by a positive or nonnegative measure. Some interesting applications were mentioned in the first part of this text. The method was designed based on simulations that were also used for its validation, a standard procedure in statistics, and its comparison with other methods. The simulation-based approach was adopted because the usual techniques of deriving the precise distribution of a test criterion under H_0: "No difference in the expected values" are too complex and incomparable to the presented technique in terms of the difficulty. The analysis was made possible through a code written in the R statistical software development environment, version 2022.07.2 Build 576, using suitable API functions for working with the gamma distribution, as well as API graphical tools (see Appendix A of this text). The presented conclusions concern small- and medium-sized samples of 30 to 80 datapoints coming independently from gamma distributions. For larger samples (at least a hundred datapoints), it seems that $\overline{X} - \overline{Y}$ is approximately normally distributed under H_0 in a majority of considered cases, as was confirmed with the Shapiro–Wilk test. This suggests that an approximate normality-based test could be applied for such sample sizes with the same objective in mind instead of the proposed procedure. The normality approach to testing can also be used once the gamma distribution shape parameters are, say, 30 or higher. The resulting method has a very good type I error probability within the considered situations, while its power averages above 0.7 for sample sizes of at least 60 for reasonably pronounced differences in the expected values (see Tables 8 and 9). The test power of the method was compared to the methodology proposed in [12] that claims to be as good as other methods, if not better. The method proposed here seems to possess comparable power unless the means differ little. This was suggested by the comparisons that could be made, at least approximately, for the parameter scenarios and sample sizes used in [12] were not always close to what was considered

in this paper. The analyses suggest that in many cases, it is possible to use the method designed in the paper, the major advantage of it being its simplicity (see page 6). Table 15 summarizes the recommendations. In practice, the true scenario behind the parameters can be identified with parameter estimates. Once the scenario estimate is given, the proper row in Table 15 is selected and the ds identified, leading to the percentiles and the test implementation.

Table 15. The multiples d to obtain the percentiles $\hat{p}_{2.5}$, $\hat{p}_{97.5}$. If $\hat{p}_{2.5} < \overline{x} - \overline{y} < \hat{p}_{97.5}$, the hypothesis H_0: $E(X) = E(Y)$ is accepted; in all other cases, it is rejected. For calculation of $\widehat{var}^{1/2}_{\overline{X}-\overline{Y}}$, see page 6.

k_1,k_2,θ_1 Scenarios (See Page 5)	d in $\hat{p}_{2.5} = d\hat{var}^{1/2}_{\overline{X}-\overline{Y}}$	d in $\hat{p}_{97.5} = d\hat{var}^{1/2}_{\overline{X}-\overline{Y}}$
1	−1.96	1.96
2	−1.96	1.96
3	−1.96	1.96
4	−1.92	2.00
5	−1.91	2.01
6	−1.95	1.97

Generally speaking, if the practitioner wants to keep the procedure even simpler, it is conceivable to use a constant d between 1.95 and 2, whatever the parameter scenario. The results obtained in this way should still possess reasonable statistical properties. In the current context, it is also possible to think of the Vysochanskij–Petunin inequality, which implies that under H_0, the difference in the sample averages will fall outside the interval $\left(-3var^{1/2},\ 3var^{1/2}\right)$ with a probability of at most 0.05. The result is valid for unimodal distributions with finite variance. This instead suggests using a multiple of three to get at the proper percentiles for the testing. But of course, the result concerns the unknown population variance var and not its estimate that must be used in practice. And even if var was known, the "at most" theoretical conclusion additionally suggests that the multiple of three is probably too high. It must be stressed that the obtained results concern specific situations, although their number runs into thousands. Nevertheless, for other situations not covered here, the Appendix A code can still be used to expand the results.

Funding: This research was funded by VSB—Technical University of Ostrava through its project "Development of Integrated Management of Industrial Processes through Linking the Principles of Smart Production and Quality Management", grant number SGS SP2023/043.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are openly available at https://doi.org/10.6084/m9.figshare.23899668.v1 (accessed on 16 August 2023), https://doi.org/10.6084/m9.figshare.23899524.v1 (accessed on 16 August 2023), https://doi.org/10.6084/m9.figshare.23899455.v1 (accessed on 16 August 2023). The first reference stores files "power_sc number1_number2.xlsx", each such file containing test powers for different parameter combinations within scenario "sc number1" for sample size "number2". The second reference stores files "d_sc number1_number2.xlsx", and each such file contains the two d multiples for the lower and upper percentiles for each scenario "sc number1" parameter combination and sample size "number2". The third reference stores files of the form "a_sc number1_number2.xlsx", which contains significance levels alpha for each scenario "sc number1" parameter combination and sample size "number2".

Conflicts of Interest: The author declares no conflict of interest. The funders had no role in the design of the study, in the collection, analysis, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

Appendix A

This supplement contains three R functions that can be used to both check the results presented in this paper and enhance at will those results for parameter scenarios other than the ones stated. The first function sets the k, theta parameters in a certain way, here, corresponding to the first parameter scenario (see page 4), then realizes tens of thousands of values of $\overline{X} - \overline{Y}$ under H_0 to estimate properly the true distribution of the difference and then calculates the lower and upper d multiple leading to the distribution percentiles relevant for the testing of $H_{0.}$ The second function takes the d-based proposed procedure described in the paper and uses it many times given a scenario-specific parameter set-up so that H_0 holds, counting the number of cases when the hypothesis is wrongly rejected with the proposed test; i.e., it estimates the type I error of the test. The third function runs the test many times for a given H_1 situation and estimates the test power.

```
# R function to identify the lower and upper d for the given scenario and sample
# size; here, the first parameter scenario is shown in the cycles below.

d <- function(n)                                    { #sample size n given

i <- 0; avgX_minus_avgY <- d_lower <- d_upper <- c(0)

for (k1 in 1:10)                    {
  for (k2 in 1:10)                    {
   for (theta1 in 1:30)                  {

theta2 <- k1 * theta1/k2                              #theta2 calculated, so that Ho holds
varXY <- (1/n) * (k1 * theta1^2 + k2 * theta2^2)     #known variance of X̄ − Ȳ under Ho

  for (draw in 1:10,000)                    #ten thousand realizations of X̄ − Ȳ under Ho

{

X <- rgamma(n,shape = k1, scale = theta1)
Y <- rgamma(n,shape = k2, scale = theta2)
avgX_minus_avgY[draw] <- mean(X) − mean(Y)

}
i <- i + 1
d_lower[i] <- quantile(avgX_minus_avgY, 0.025) * varXY ^ (−0.5)
d_upper[i] <- quantile(avgX_minus_avgY, 0.975) * varXY ^ (−0.5)

                                  }
                                }
                              }

return(list(d_lower = d_lower, d_upper = d_upper))    }

# R function to estimate the type I error alpha for the proposed test
# the code concerns the first parameter scenario (see page 4), shown here
# in the cycles below; for other scenarios, this part must be changed
# as well as the 1.96 mutliples used in the cycles
# n = sample size

alpha_check <- function(n, numberOfTests)                        {

bigger <- c(0); i <- 0
```

```
empir_quantupper <- empir_quantlower <- varXY <- avgX_minus_avgY <-
numeric(numberOfTests)

for (k1 in 1:10)                { #the first-scenario parameter combinations
  for (k2 in 1:10)          {
   for (theta1 in 1:30)      {
theta2 <- k1 * theta1/k2              #calculation of theta2, so that Ho holds

for (j in 1: numberOfTests)          { #many tests for the given k1,k2, theta 1,2

X <- rgamma(n, shape = k1, scale = theta1)
Y <- rgamma(n, shape = k2, scale = theta2)
avgX_minus_avgY[j] <- mean(X) - mean(Y)
varXY[j] <- (1/n) * ((n − 1)/(n + 1)) * (var(X) + var(Y)) #the min MSE estimate of var
(X̄ − Ȳ)
empir_quantlower[j] <- −1.96 * sqrt(varXY[j]) #the lower percentile for testing Ho
empir_quantupper[j] <- 1.96 * sqrt(varXY[j]) #the upper percentile for testing Ho

                                  }

i <- i + 1 #saving in"bigger" the relative frequency of cases when Ho is rejected
bigger[i] <- (sum( avgX_minus_avgY > empir_quantupper )/numberOfTests) +
(sum( avgX_minus_avgY < empir_quantlower )/numberOfTests)

                                 }
                               }
                            }
return( bigger )
                                                          }

#"bigger" is a vector whose i-th component gives an estimate of the type I error
#probability corresponding to the i-th parameter set-up within the scenario

# R function estimating the test power (here, for the 1st parameter scenario set-ups)
# n=sample size, const = the constant in H1 (see the top of page 5), const > 0
# the higher the "number" of tests, the more precise the test power estimation
power_check <- function(n, const, number)                    {

test_power <- c(0); i <- 0

for (k1 in 1:10)                        { #the first-scenario parameter combinations
  for (k2 in 1:10)                       {
   for (theta1 in 1:30)                        {

theta2 <- (k1*theta1 + const)/k2          # the set-up of a H1 case
numberOfSuccesses <- 0

for (test_run in 1:number) { #test power for given k1, k2, theta 1, 2, n, const
X <- rgamma(n,shape = k1, scale = theta1)
Y <- rgamma(n,shape = k2, scale = theta2)
percentile <- 1.96 * sqrt((1/n) * (n − 1)/(n + 1) * (var(X) + var(Y))) #test percentile,
                                                                  # here for the 1st scenario

if (abs(mean(X)-mean(Y)) >= percentile)
{ numberOfSuccesses <- numberOfSuccesses + 1}
```

```
# this condition is valid for the 1st scenario only; for others, it must be changed

                    }
i <- i + 1; test_power[i] <- numberOfSuccesses /number #saving the test power
                                                        #for the given set-up
                  }
                }
              }
return(test_power)                    }

# Some parts of the code must be changed for the appropriate scenario, e.g.:
# percentile1 <- −1.91 * sqrt((1/n) * (n − 1)/(n + 1) * (var(X) + var(Y))) for the 5th scenario
# percentile2 <- 2.01 * sqrt((1/n) * (n − 1)/(n + 1) * (var(X) + var(Y))) for the 5th scenario, and:
# if (( mean(X)-mean(Y) <= percentile1) | ( mean(X)-mean(Y) >= percentile2)) { . . . }
```

References

1. Krishnamoorthy, K. *Handbook of Statistical Distributions with Applications*; CRC Press: Boca Raton, FL, USA, 2019; pp. 186–187.
2. Handbook on Statistical Distributions for Experimentalists. p. 70. Available online: https://www.stat.rice.edu/~dobelman/textfiles/DistributionsHandbook.pdf (accessed on 10 May 2023).
3. Athreya, K.B.; Laihiri, S.N. *Measure Theory and Probability Theory*; Springer: New York, NY, USA, 2006; pp. 232–322.
4. Taboga, M. *Lectures on Probability Theory and Mathematical Statistics*; CreateSpace: Scotts Valley, CA, USA, 2017; p. 317.
5. Probability Density under Transformation. Available online: https://www.cs.cornell.edu/courses/cs6630/2015fa/notes/pdf-transform.pdf (accessed on 8 May 2023).
6. Ash, R.B.; Doleans-Dade, C.A. *Probability and Measure Theory*; Academic Press: San Diego, CA, USA, 2000; p. 329.
7. Grice, J.V.; Bain, L.J. Inferences Concerning the Mean of the Gamma Distribution. *J. Am. Stat. Assoc.* **1980**, *75*, 929–933. [CrossRef]
8. Shiue, W.K.; Bain, L.J. A Two-Sample Test of Equal Gamma Distribution Scale Parameters with Unknown Common Shape Parameter. *Technometrics* **1983**, *25*, 377–381. [CrossRef]
9. Shiue, W.K.; Bain, L.J.; Engelhardt, M. Test of Equal Gamma Distribution Means with Unknown and Unequal Shape Parameters. *Technometrics* **1988**, *30*, 169–174. [CrossRef]
10. Tripathi, R.C.; Gupta, R.C.; Pair, R.K. Statistical Tests Involving Several Independent Gamma Distributions. *Ann. Inst. Stat. Math.* **1993**, *45*, 773–786. [CrossRef]
11. Bhaumik, D.K.; Kapur, K.; Gubbons, R.D. Testing Parameters of a Gamma Distribution for Small Samples. *Technometrics* **2009**, *51*, 326–334. [CrossRef]
12. Chang, C.; Lin, J.J.; Pal, N. Testing the Equality of Several Gamma Means: A parametric bootstrap method with applications. *Comput. Stat.* **2011**, *26*, 55–76. [CrossRef]
13. Thiagarajah, K. Testing Homogeneity of Gamma Populations. *J. Stat. Manag. Syst.* **2013**, *16*, 433–444. [CrossRef]
14. Testing a Difference in Means. Available online: https://colab.research.google.com/github/AllenDowney/ElementsOfDataScience/blob/master/testing_means.ipynb (accessed on 10 May 2023).
15. Freund, J.E. *Mathematical Statistics with Applications*; Pearson: New York, NY, USA, 2014; pp. 337–342.
16. Variance Estimators That Minimize MSE. Available online: https://davegiles.blogspot.com/2013/05/variance-estimators-that-minimize-mse.html (accessed on 10 May 2023).
17. Freund, R.J.; Wilson, W.J.; Mohr, D.L. *Statistical Methods*; Academic Press: Cambridge, UK, 2010; p. 213.
18. Shapiro, S.S.; Wilk, M.B. An Analysis of Variance Test for Normality. *Biometrika* **1965**, *52*, 591–611. [CrossRef]
19. Razali, N.; Wah, Y.B. Power Comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling Tests. *J. Stat. Model. Anal.* **2011**, *2*, 21–33.
20. Wackerly, D.; Mendenhall, W.; Scheaffer, R.L. *Mathematical Statistics with Applications*; Brooks/Cole, Cengage Learning: Belmont, CA, USA, 2008; p. 370.