symmetry

MDPI

Article

A Test Detecting the Outliers for Continuous Distributions Based on the Cumulative Distribution Function of the Data Being Tested

Lorentz Jäntschi [1,2]

[1] Department of Physics and Chemistry, Technical University of Cluj-Napoca, 400641 Cluj, Romania; lorentz.jantschi@chem.utcluj.ro or lorentz.jantschi@gmail.com

[2] Chemical Doctoral School, Babeș-Bolyai University, 400084 Cluj-Napoca, Romania

Received: 12 June 2019; Accepted: 24 June 2019; Published: 25 June 2019

Abstract: One of the pillars of experimental science is sampling. Based on the analysis of samples, estimations for populations are made. There is an entire science based on sampling. Distribution of the population, of the sample, and the connection among those two (including sampling distribution) provides rich information for any estimation to be made. Distributions are split into two main groups: continuous and discrete. The present study applies to continuous distributions. One of the challenges of sampling is its accuracy, or, in other words, how representative the sample is of the population from which it was drawn. To answer this question, a series of statistics have been developed to measure the agreement between the theoretical (the population) and observed (the sample) distributions. Another challenge, connected to this, is the presence of outliers - regarded here as observations wrongly collected, that is, not belonging to the population subjected to study. To detect outliers, a series of tests have been proposed, but mainly for normal (Gauss) distributions—the most frequently encountered distribution. The present study proposes a statistic (and a test) intended to be used for any continuous distribution to detect outliers by constructing the confidence interval for the extreme value in the sample, at a certain (preselected) risk of being in error, and depending on the sample size. The proposed statistic is operational for known distributions (with a known probability density function) and is also dependent on the statistical parameters of the population—here it is discussed in connection with estimating those parameters by the maximum likelihood estimation method operating on a uniform U(0,1) continuous symmetrical distribution.

Keywords: test for outliers; order statistics; extreme values; confidence intervals; Monte-Carlo simulation

1. Introduction

Many statistical techniques are sensitive to the presence of outliers and all calculations, including the mean and standard deviation can be distorted by a single grossly inaccurate data point. Therefore, checking for outliers should be a routine part of any data analysis.

To date, several tests have been developed for the purpose of identifying outliers of certain distributions. Most of the studies are connected with the Normal (or Gauss) distribution [1]. The first paper that attracted attention on this matter is [2] and this was followed by studies that identified the derivation of the distribution of the extreme values in samples taken from Normal distributions [3]. Then, a series of tests were developed by Thompson in 1935 [4], these were subjected to evaluation [5], and revised [6,7].

For other distributions such as the Gamma distribution, procedures for detecting outliers were proposed [8], revised [9], and unfortunately proved to be inefficient [10].

The first attempt to generalize the criterion for detecting outliers for any distribution can be found in [11], but further research on this subject is scarce apart from a notable recent attempt by Bardet and Dimby [12].

The Grubbs test is a frequently used test for detecting the outliers of a Normal distribution [7]. For a sample (x), the Grubbs' test statistic takes the largest absolute deviation from the sample mean ($\overline{x}$) in units of the sample standard deviation (s) in order to calculate the risk of being in error (α_G) when stating that the most departed values from the mean (min(x), max(x) or both) are not outliers (see Table 1). The associated probabilities of the observed (p_G) are obtained from the Student t distribution [13].

Table 1. The Grubbs statistic.

Sample statistic (G)	Associated probability (p_G = 1-α_G)	Equation
$G_{"min"} = \frac{\overline{x}-\min(x)}{s}$ $G_{"max"} = \frac{\max(x)-\overline{x}}{s}$	$\alpha_G = n \cdot \mathrm{CDF}_{"\text{Student t}"}\left(-\sqrt{\frac{n(n-2)}{(\frac{n-1}{G})^2-n}}, n-2\right)$	(1)
$G_{"all"} = \max(G_{"min"}, G_{"max"})$	$\alpha_G = 2n \cdot \mathrm{CDF}_{"\text{Student t}"}\left(-\sqrt{\frac{n(n-2)}{(\frac{n-1}{G})^2-n}}, n-2\right)$	(2)

One should note that the Grubbs test statistic produces a symmetrical confidence interval (see Equations (1) and (2)). The Grubbs statistic as given in Table 1, is intended to be used with the parameters of the population (μ and σ), which are determined using the central moments (CM) method ($\hat{\mu} = \overline{x} = \sum x/\mathrm{n}$; $\hat{\sigma} = \mathrm{s} = \left(\sum (x-\overline{x})^2\right)^{1/2}/\mathrm{n}$).

Here, a method is proposed for constructing the confidence intervals for the extreme values of any continuous distribution for which the cumulative distribution function is also obtainable. The method involves the direct application of a simple test for detecting the outliers. The proposed method is based on deriving the statistic for the extreme values for the uniform distribution. Also, the proposed method provides a symmetrical confidence interval in the probability space.

2. Materials and Methods

The Grubbs test (Table 1) is based on the fact that if outliers exist, then these are "localized" as the maximum value and/or the minimum value in the dataset. Thus, the Grubbs test is essentially a sort of order statistic [14].

Some introductory elements are required for describing the proposed procedure. When a sample of data is tested under the null hypothesis that it follows a certain distribution, it is intrinsically assumed that the distribution is known. The usual assumption is that we possess its probability density function (PDF, for a continuous distribution) or its probability distribution function (PDF for a discrete distribution). The discussion below relates to continuous distributions, although the treatment of discrete distributions are similar to certain degree. Nevertheless, a major distinction between continuous and discrete distributions in the treatment of data is made here; that is, a continuous distribution is "dense", e.g., between any two distinct observations it is possible to observe another while in the case of a discrete distribution, this is generally not true.

Even when the PDF is known (possibly intrinsically), its (statistical) parameters may not necessarily be known, and this raises the complex problem of estimating the parameters of the (population) distribution from the sample; however, this issue is outside the scope of this paper. In general, the estimation of the parameters of the distribution of the data is biased by the presence of the outliers in the data, and thus, identifying the outliers along with the estimation of the parameters of the distribution is a difficult task because two statistical hypotheses are operating. Assuming that the parameters ("parameters") of the distribution (of the PDF) are obtained using the maximum likelihood estimation method (MLE, Equation (3); see [15]), there is some suggestion that the uncertainty accompanying this estimation is transmitted to the process of detecting the outliers.

$$\prod \mathrm{PDF}(\mathrm{X};\ \text{“parameters”}) \rightarrow \max. \ \Rightarrow \ \sum \ln\left(\mathrm{PDF}(\mathrm{X};\ \text{“parameters”})\right) \rightarrow \min. \quad (3)$$

It should be noted that Equation (3) is a simplified version of the MLE method, since the real use of it requires and involves partial derivatives of the parameters; see Source code (MathCad language) for the MLE estimations in the Supplementary Materials available online.

Either way (whether the uncertainty accompanying this estimation is transmitted to the process of detecting the outliers or not), once an estimate for the parameters of the distribution is available, a test (most desirably, a test based on a statistic) for detecting the presence of an outlier must provide the probability of observing that (assumed) "outlier" as a randomly drawn value from the distribution. What to do next with the probability is another statistical "trick": to observe a value with a probability less than an imposed "level" (usually 5%) is defined as an unlikely event, and therefore, the suspicion regarding the presence of the outlier is justified. With regard to the statistical "trick" mentioned above, the opinion of the author of this manuscript is that one "observation" is not enough. Actually, there should be a series of observations, that come from a series of statistics, each providing a probability. Then, the unlikeliness of the event can be safely ascertained by using Fisher's "combining probability from independent tests" method (FCS, Equation (4); see [16–18]:

$$-\sum_{i=1}^{\tau} \ln(p_i) \sim \chi^2(\tau) \rightarrow \alpha_{FCS} = 1 - CDF_{\chi^2}(-\sum_{i=1}^{\tau} \ln(p_i); \tau) \quad (4)$$

where $p_1, \ldots, p_\tau$ are probabilities from τ independent tests, CDFχ^2 is the χ^2 cumulative distribution function (see also up until Equation (6) below), and p_{FCS} is the combined probability from independent tests.

Taking the general case, for $(x_1, \ldots, x_n)$ as n independent draws (or observations) from a (assumed known) continuous distribution defined by its probability density function, PDF $(x; (\pi_j)_{1\le j\le m})$ where $(\pi_j)_{1\le j\le m}$ are the (assumed unknown) m statistical parameters of the distribution, by way of integration for a (assumed known) domain (D) of the distribution, we may have access to the associated cumulative density function (CDF) CDF$(x; (\pi_j)_{1\le j\le m}$; PDF), simply expressed as (Equation (5)):

$$CDF(x; (\pi_{j)1\le j\le m}) = \int_{\inf(D)}^{x} PDF(x; (\pi_j)_{1\le j\le m}) \quad (5)$$

where inf(D) was used instead of min(D) to include unbounded domains (e.g., when inf(D) = -∞; "inf" stands for infimum, "min" stands for minimum). Please note that having the PDF and CDF does not necessarily imply that we have an explicit formula (or expression) for any of them. However, with access to numerical integration methods [19], it is enough to have the possibility of evaluating them at any point (x).

Unlike PDF$(x; (\pi_j)_{1\le j\le m})$, CDF$(x; (\pi_j)_{1\le j\le m})$ is a bijective function and therefore, it is always invertible (even if we do not have an explicit formula; let "InvCDF" be its inverse, Equation (6)):

$$\text{if } p = CDF(x; (\pi_j)_{1\le j\le m}), \text{ then } x = InvCDF(p; (\pi_j)_{1\le j\le m}), \text{ and vice-versa} \quad (6)$$

CDF$(x; (\pi_j)_{1\le j\le m}$; "PDF") is a strong tool that greatly simplifies the problem at hand: the problems of analyzing any distribution function (PDF) are translated such that only one needs to be analyzed (the continuous uniform distribution). That is, a series of observed data $(x_i)_{1\le i\le n}$ is expressed through their associated probabilities $p_i = CDF(x_i; (\pi_j)_{1\le j\le m})$ (for $1\le i\le n$) and the analysis can be conducted on the $(p_i)_{1\le i\le n}$ series instead.

Since the analysis of the $(p_i)_{1\le i\le n}$ series of probabilities is a native case of order statistics, the discussion now turns to order statistics. The first studies in this area were by the fathers of modern statistics, Karl Pearson [20] and Ronald A. Fisher [3] while the first order statistic applicable to any distribution (not only the normal distribution) was first studied by Cramér and Von Mises (see [21,22]).

An order statistic operating on probabilities $((p_i)_{1\le i\le n})$ will sort the values (let $(q_i)_{1\le i\le n}$ be the series of sorted $(p_i)_{1\le i\le n}$ values, Equation (7)) and will assess its departure from the continuous uniform distribution (where it is assumed that SORT is a procedure that sorts ascending the values).

$$(q_i)_{1\le i\le n} \leftarrow \mathrm{SORT}((p_i)_{1\le i\le n}) \tag{7}$$

Since the assessment of the departure from the continuous uniform distribution cannot be made directly, the use of a series of order statistics was proposed by several authors including: Cramér and Von Mises [21,22], Kolmogorov-Smirnov [23–25], Anderson-Darling [26,27], Kuiper V [28], Watson U^2 [29], and the H1 Statistic [18]; see Equation (8). They remain in use today.

For instance, the Kolmogorov-Smirnov (KS) method (see Equation (8); the Kolmogorov-Smirnov statistic) calculates the $KS_{Statistic}$ and later tests the value (from a sample) against the threshold of a chosen significance level (usually 5%).

In order to have certain thresholds for a series of significance levels, these statistics can be derived from Monte-Carlo ("MC") simulations [30], and deployed for a large number of samples in order to reflect, as best as possible, the state of the population.

$$\begin{aligned}
KS_{Statistic} &= \sqrt{n}\cdot\max_{1\le i\le n}\left(q_i - \tfrac{i-1}{n}, \tfrac{i}{n} - q_i\right)\\
KV_{Statistic} &= \sqrt{n}\cdot\left(\max_{1\le i\le n}\left(q_i - \tfrac{i-1}{n}\right) + \max_{1\le i\le n}\left(\tfrac{i}{n} - q_i\right)\right)\\
AD_{Statistic} &= -n - \tfrac{1}{n}\cdot\sum_{i=1}^{n}(2i-1)\cdot\ln(q_i\cdot(1-q_{n-i}))\\
CM_{Statistic} &= \tfrac{1}{12n} + \sum_{i=1}^{n}\left(\tfrac{2\cdot i-1}{2\cdot n} - q_i\right)^2\\
WU_{Statistic} &= \mathrm{CM}_{\mathrm{Statistic}} + \left(\tfrac{1}{2} - \tfrac{1}{n}\sum_{i=1}^{n} q_i\right)^2\\
H1_{Statistic} &= -\sum_{i=1}^{n} q_i\cdot\ln(q_i) - \sum_{i=1}^{n}(1-q_i)\cdot\ln(1-q_i)
\end{aligned} \tag{8}$$

3. Proposed Outlier Detection Statistic

A statistic was developed to be applicable to any distribution. For a series of probabilities $((p_i)_{1\le i\le n})$ or (sorted probabilities, $(q_i)_{1\le i\le n}$) associated with a series of (repeated drawing) observations $((x_i)_{1\le i\le n})$, the $(r_i)_{1\le i\le n}$ differences are calculated as Equation (9):

$$r_i = \left|p_i - 0.5\right|, \text{ for } 1 \le i \le n \tag{9}$$

The statistic called "g1" (see below) was generated based on the formula given in Equation (9) (given as Equation (10)).

$$g1 = \max_{1\le i\le n} r_i \tag{10}$$

It should be noted that Equations (9) and (10) provide the same result regardless of whether the calculation is made on a sorted series of probabilities $((q_i)_{1\le i\le n})$ or not (then it is made on $(p_i)_{1\le i\le n}$).

Regarding the name of this new proposed statistic ("g1"), when Equations (1) and (2) ($G_{"min"}$, $G_{"max"}$, $G_{"all"}$) and Equation (9) are compared, for a standard normal distribution N(x; μ=0,σ=1) the equation defining $G_{"all"}$ becomes much more like Equation (9), with the difference being that in Equation (2) the sample mean ($\bar{x}$) is used as an estimate for the mean of the population (μ) and the sample standard deviation (s) is used as an estimate for the standard deviation of the population (σ) while Equation (9) basically expresses the same in terms of associated probabilities ($p_i = P(X \le x_i) = \mathrm{CDF}_{"Normal"}(x_i; \mu,\sigma)$, $0.5 = P(X \le \mu) = \mathrm{CDF}_{"Normal"}(\mu; \mu,\sigma)$).

Therefore, the proposed statistic very much resembles the Grubbs test for normality (and hence its name). One difference is that in the Grubbs test sample statistics are used to calculate the sample $G_{"all"}$ value ($\bar{x}$ and s), thereby reducing the degrees of freedom associated with the value (from n to n-2) while

for the g1 value (and statistic) the degrees of freedom remain unchanged (n). The major difference is actually the one that makes the proposed statistic generalizable to any distribution—the mean used in the Grubbs test is replaced by the median—the beauty of this change is that for symmetrical distributions (including a Normal distribution) these two coincide.

A further connection with other statistics must also be noted. If any sample is resampled by extracting only the smallest and the largest of its values, then the Kolmogorov-Smirnov statistic for those subsamples almost perfectly resembles (by setting n = 2 in Equations (8)–(1)) the proposed "g1" statistic.

Since CDF is a bijective function (see Equation (6)), the proposed generalization of the Grubbs test for detecting the outliers for Normal distribution into the "g1" statistic for detecting the outliers for any distribution is a natural extension of it. The "g1" test associated with the "g1" statistic will be able to operate in the probability space $((p_i)_{1\le i\le n}$ or $(q_i)_{1\le i\le n})$ instead of the observed space $((x_i)_{1\le i\le n})$, the calculation formula (Equations (9) and (10)) is slightly different (to those given in Equations (1) and (2)), and the probability associated with the departure will no longer be extracted from the Student t distribution (as in Equations (1) and (2)). The change from mean (μ for $G_{"all"}$) to median (0.5 in Equation (9)) is a safe extension for any distribution type, since Equation (9) measures (or accounts for) the extreme departures from the equiprobable point—having an observation y ($y \leftarrow X$) with $y \le \text{InvCDF}_{"\text{Any distribution}"}(0.5;$ "parameters") and an observation z ($z \leftarrow X$) with $z \ge \text{InvCDF}_{"\text{Any distribution}"}(0.5;$ "parameters") is equiprobable.

One way to associate a probability with the "g1" statistic is to do a Monte-Carlo (MC) simulation.

4. Simulation Study

A MC study was conducted. Two different strategies were developed in order to deal efficiently with a very large amount of data, and specifically, to solve the order statistics problem (that is, first sampling from the uniform distribution, and later using Equations (7)–(10). One of those alternatives has been described in [14] and the other is described below. Table 2 shows the details of the conducted MC study.

Table 2. Details of the MC simulation on "g1" outlier detection statistic.

Parameter	Meaning	Setting
n	sample size of the observed	from 2 to 12
m	sample size of the MC simulation	10^8
p	control points for the probability	999
resa	internal resamples (repetitions)	10
repe	external repetitions	7

For each sample size of the observed *n* in each run *m* samples (see Table 2) were generated from the standard uniform continuous distribution (e.g., from the [0, 1] interval). The outlier detection statistic "g1" was calculated (Equations (9) and (10)). From a large pool of sampled and resampled data (m·resa·repe = $7{\cdot}10^9$ in Table 2, repetitions were joined *(n, p, g1)* as pairs from the *p·n* control points, that is, where the probability was from 0.001 to 0.999 with a step of 0.001 for each n (from 2 to 12). The external repetitions (resa = 7 in Table 2) were joined together by taking the median (since the median is a sufficiency statistic [31] for any order statistic such as in the extraction of (*n*, *p*, *g1*) pairs from the *p·n* control points). The MC simulation was conducted with the configuration set as defined in Table 2. The obtained data were recorded in separate files by sample size and analyzed as such.

The objective associated (with any statistic) is to obtain the cumulative distribution function (CDF, Equation (5)), and thus by evaluating the CDF for the value of the statistic obtained from the sample (Equations (9) and (10)) to obtain a probability for the sampling. Please note that only in the lucky cases are we able to do this; Generally only the critical values (values corresponding to certain risks

of being in error) or approximation formulas are available (see for instance [21,24,26,28,29]). Here, the analytical CDF formula was obtained for the "g1" outlier detection statistic.

5. The Analytical Formula of CDF for g1

The "g1" statistic have a very simple calculation formula (see Equation (9)) and, as expected, its CDF formula is also very simple (see Equation (11)). Thus, for a calculated sample statistic g1 (x ← g1 in Equation (11)), the significance level ($\alpha \leftarrow$ 1-p) is immediate (Equation (11), where P represents the probability that the random variable X takes on a value less than or equal to x).

$$p = \text{CDF}_{\text{"g1"}}(x; n) = \text{P}(\text{X} \leq \text{x}) = (2{\cdot}x)^n,\ \alpha = 1 - p = 1 - (2{\cdot}x)^n \tag{11}$$

6. Simulation Results for the Distribution of the "g1" Statistic

The results of the simulation for n varying from 2 to 10 were sufficient to provide a clear indication of the analytical formula for the CDF of "g1". Descriptive statistics including Standard Error (SE, the standard error formula is given as Equation (12)) between the expected probability (from MC simulation) and the calculated probability (from Equation (11), $\hat{p}_i \leftarrow (2{\cdot}x_i)^n$) and the highest positive and highest negative departures are given in Table 3.

$$SE = \sqrt{\frac{1}{999}\sum_{i=1}^{999}(p_i - \hat{p}_i)^2},\ p_i = \frac{i}{1000} \tag{12}$$

Table 3. Descriptive statistics for the agreement in the calculation of the "g1" statistic (Equation (10) vs. Equation (11)).

n	SE	$\min(p_i - \hat{p}_i)$		$\max(p_i - \hat{p}_i)$	
2	$\mathbf{2.9 \times 10^{-6}}$	-7.9×10^{-6}	at p = 0.694	5.7×10^{-6}	at p = 0.427
3	5.6×10^{-6}	-1.2×10^{-5}	at p = 0.787	1.6×10^{-6}	at p = 0.118
4	2.2×10^{-6}	-5.6×10^{-6}	at p = 0.234	3.7×10^{-6}	at p = 0.613
5	6.0×10^{-6}	-1.2×10^{-5}	at p = 0.546	2.3×10^{-6}	at p = 0.080
6	3.5×10^{-6}	-5.8×10^{-6}	at p = 0.797	9.2×10^{-6}	at p = 0.196
7	5.0×10^{-6}	-9.6×10^{-6}	at p = 0.777	3.8×10^{-6}	at p = 0.035
8	4.2×10^{-6}	-8.4×10^{-6}	at p = 0.675	3.9×10^{-6}	at p = 0.948
9	3.3×10^{-6}	-9.1×10^{-6}	at p = 0.269	7.9×10^{-6}	at p = 0.689
10	2.8×10^{-6}	-6.4×10^{-6}	at p = 0.443	6.6×10^{-6}	at p = 0.652

As can be observed in Table 3 the standard error (SE) slowly decreases beginning with n = 7, being two orders of magnitude smaller (actually it is about 200 times smaller) than the step from the MC experiment. Since the standard error alone is not proof that Equation (11) is the true CDF formula for providing the probability for the g1 statistic, the smallest and the highest difference between the observed and the expected probabilities are also given in Table 3. They substantiate that Equation (11) is indeed the right estimate for the CDF of g1. For convenience, Figure 1 shows the value of the error in each observation point (999 points corresponding to p = 0.001 up to p = 0.999 for each n from 2 to 12).

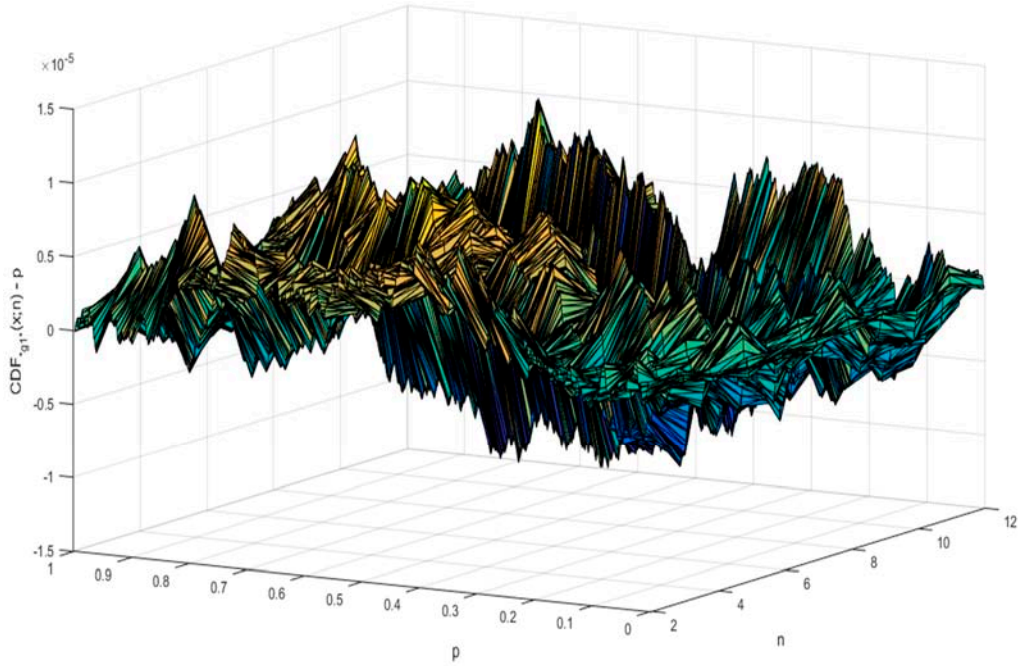

Figure 1. Departures between expected and observed probabilities for g1 statistic (Equation (10) vs. Equation (11)).

Regarding the estimation error (of the "g1" statistic) depicted in Figure 1, the "g1" statistic is rarely bigger than 10^{-5}, never bigger than $1.5 \cdot \times 10^{-5}$ and tends to become smaller with the increase in sample size (n). Using Equation (11), Figure 2 depicts the shape of the $CDF_{"g1"}(x;n)$.

With regard to the "g1" statistic (depicted in Figure 2), the domain for a variable distributed by the "g1" statistic (see Equation (11)) has values between 0 and 0.5 with the mode at p = 0 (a vertical asymptote at p = 0), a median of $n^{-1} \cdot 2^{-1/n}$ (and having a left asymmetry decreasing with the increasing of n and converging (for $n \to \infty$) to symmetry) and mean of 1/2(n+1).

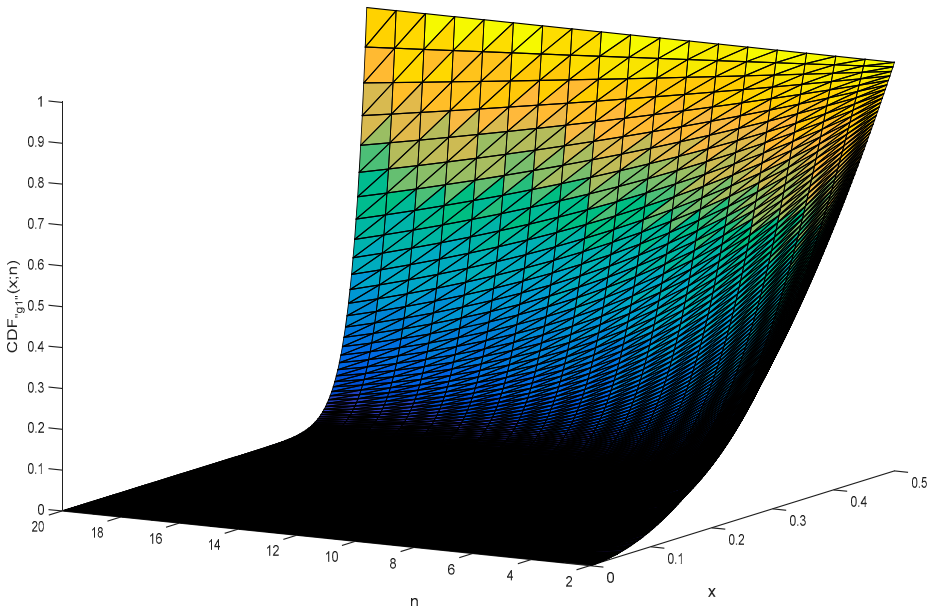

Figure 2. $CDF_{"g1"}(x;n)$ for n = 2 to n = 20.

The expression of $CDF_{"g1"}$ is easily inverted (see Equation (13)).

$$CDF_{"g1"}(x;\ n)\ =\ (2x)^n\ \rightarrow\ InvCDF_{"g1"}(p;\ n)\ =\ \sqrt[n]{p}/2 \tag{13}$$

7. from "g1" Statistic to "g1" Confidence Intervals for the Extreme Values

Equation (13) can be used to calculate the critical values of the "g1" statistic for any values of α ($\alpha \leftarrow$ 1-p) and n. The critical values of the "g1" statistic acts as the boundaries of the confidence intervals.

By setting the risk of being in error α (usually at 5%), then p = 1-α and Equation (13) can be used to calculate the statistic associated with it ($InvCDF_{"g1"}$(1-α;n) = $\sqrt[n]{1-\alpha}/2$). By placing this value into Equations (9) and (10), the (extreme) probabilities can be extracted (Equation (14)).

$$\max_{1\le i\le n}\left|p_i - 0.5\right| = \sqrt[n]{1-\alpha}/2.\ \rightarrow\ p_{extreme}(\alpha)\ =\ 0.5\ \pm\ \sqrt[n]{1-\alpha}/2 \tag{14}$$

One should note that the confidence interval defined by Equation (14) is symmetric.

In order to arrive at the confidence intervals for the extreme values in the sampled data (Equation (15)) it is necessary to use the inverse of the CDF (again), and for the distribution of the sampled data.

$$x_{extreme}(\alpha) = InvCDF_{"Distribution"}(0.5 \pm \sqrt[n]{1-\alpha}/2; "parameters") \tag{15}$$

To illustrate the calculation of the confidence intervals for the extreme values in the sampled data, a series of 206 data was chosen from [32]. The data were tested against the assumption that it follows a generalized Gauss-Laplace distribution (Equation (16), a symmetrical distribution), and later if there were some observations suspected to be outliers. The steps of this analysis and the obtained results are given in Table 4.

$$PDF_{"GL"}(x;\mu,\sigma,k) = c_1\sigma^{-1}e^{-|c_0 z|^k},\ c_0 = \left(\frac{\Gamma(3/k)}{\Gamma(1/k)}\right)^{1/2},\ c_1 = \frac{kc_0}{2\Gamma(1/k)},\ z = \frac{x-\mu}{\sigma} \tag{16}$$

The greatest departure from the median (0.5) for the 206 PCB dataset (Table 4) was 9.603 ($CDF_{"GL"}$(9.603; μ = 6.47938, σ = 0.82828, k = 1.79106) = 0.9998). Due to the force of this deviation from the median, 9.603 was suspected as being an outlier and was removed (it should be noted that in a broader context, an outlier can be also seen as an atypical observation, correctly collected from the population observation, as part of the data generation process and thus it may be maintained in the sample but probably with a less weight). The same procedure (as in Table 4) can be applied to the remaining data (205 observations). Then, $InvCDF_{"g1"}$(1-0.05; 205) = 0.499875, p_{min}(n=205) = 0.0001251; and p_{max}(n=205) = 0.9998749. The MLE estimates for the parameters of the Gauss-Laplace distribution remain unchanged (μ = 6.47938, σ = 0.82828, k = 1.79106) and the removed observation (9.603) is still not an outlier (x_{max} = $InvCDF_{"GL"}$(0.9998749; μ = 6.47938, σ = 0.82828, k = 1.79106) = 9.7166 > 9.603).

Table 4. Distribution analysis for a series of 206 measurements for the octanol water partition coefficient (K_{ow}) of polychlorinated biphenyls expressed in logarithmic scale ($\log_{10}(K_{ow})$)

Step	Results
Dataset (given for convenience)	4.151; 4.401; 4.421; 4.601; 4.941; 5.021; 5.023; 5.150; 5.180; 5.295; 5.301; 5.311; 5.311; 5.335; 5.343; 5.404; 5.421; 5.447; 5.452; 5.452; 5.481; 5.504; 5.517; 5.537; 5.537; 5.551; 5.561; 5.572; 5.577; 5.577; 5.627; 5.637; 5.637; 5.667; 5.667; 5.671; 5.677; 5.677; 5.691; 5.717; 5.743; 5.751; 5.757; 5.761; 5.767; 5.767; 5.787; 5.811; 5.817; 5.827; 5.867; 5.897; 5.897; 5.904; 5.943; 5.957; 5.957; 5.987; 6.041; 6.047; 6.047; 6.047; 6.057; 6.077; 6.091; 6.111; 6.117; 6.117; 6.137; 6.137; 6.137; 6.137; 6.137; 6.142; 6.167; 6.177; 6.177; 6.177; 6.204; 6.207; 6.221; 6.227; 6.227; 6.231; 6.237; 6.257; 6.267; 6.267; 6.267; 6.291; 6.304; 6.327; 6.357; 6.357; 6.367; 6.367; 6.371; 6.427; 6.457; 6.467; 6.487; 6.497; 6.511; 6.517; 6.517; 6.523; 6.532; 6.547; 6.583; 6.587; 6.587; 6.587; 6.607; 6.611; 6.647; 6.647; 6.647; 6.647; 6.647; 6.657; 6.657; 6.671; 6.671; 6.677; 6.677; 6.677; 6.697; 6.704; 6.717; 6.717; 6.737; 6.737; 6.737; 6.747; 6.767; 6.767; 6.767; 6.797; 6.827; 6.857; 6.867; 6.897; 6.897; 6.937; 6.937; 6.957; 6.961; 6.997; 7.027; 7.027; 7.027; 7.057; 7.071; 7.087; 7.087; 7.117; 7.117; 7.117; 7.121; 7.123; 7.147; 7.151; 7.177; 7.177; 7.187; 7.187; 7.207; 7.207; 7.207; 7.211; 7.247; 7.247; 7.277; 7.277; 7.277; 7.281; 7.304; 7.307; 7.307; 7.321; 7.337; 7.367; 7.391; 7.427; 7.441; 7.467; 7.516; 7.527; 7.527; 7.557; 7.567; 7.592; 7.627; 7.627; 7.657; 7.657; 7.717; 7.747; 7.751; 7.933; 8.007; 8.164; 8.423; 8.683; 9.143; 9.603
For n = 206 calculate the probability that the extreme values contain an outlier by using Equation (13)	At α = 5% risk being in error $InvCDF_{"g1"}$(1-0.05; 206) = 0.498755

Table 4. *Cont.*

Step	Results
Calculate the critical probabilities for the extreme values by using Equations (9) and (10)	g1 = 0.498755 → |0.5 - $p_{min/max}$| = 0.498755 → 1 - 2$p_{min/max}$ = ± 0.99751 → p_{min} = 0.0001245; p_{max} = 0.9998755
Estimate the parameters of the distribution fitting the dataset (distribution: Gauss-Laplace; μ - location parameter; σ - scale parameter; k - shape parameter)	Initial estimates (from a hybrid CM & MLE method): μ = 6.4806; σ = 0.83076; k = 1.4645; MLE estimates (by applying eq.3): μ = 6.47938; σ = 0.82828; k = 1.79106;
Calculate the lower and the upper bound for the extreme values by using InvCDF of the distribution fitting the data (Equation (15))	$InvCDF_{"GL"}$(0.0001245; μ = 6.47938, σ = 0.82828, k = 1.79106) = 3.2409 $InvCDF_{"GL"}$(0.9998755; μ = 6.47938, σ = 0.82828, k = 1.79106) = 9.7178
Make the conclusion regarding the outliers	Since the smallest value in the dataset is 4.151 (> 3.24) and the largest value is 9.603 (< 9.71), at 5% risk being in error there are no outliers in the dataset on the assumption that data follows the Gauss-Laplace distribution

8. Proposed Procedure for Detecting the Outliers

The procedure for detecting the outliers should start with measuring the agreement between the observed and estimated (Figure 3).

Figure 3 contains a statistical "trick", namely, when there are no outliers the statistics measuring the gap between the observation and the model (order statistics, Equation (6)) are in agreement (their associated probabilities are not too far from each other). When outliers exist, the order statistics are also sensitive to their presence. Since this is a separate subject, for further discussion please see the series of papers beginning with [32–34].

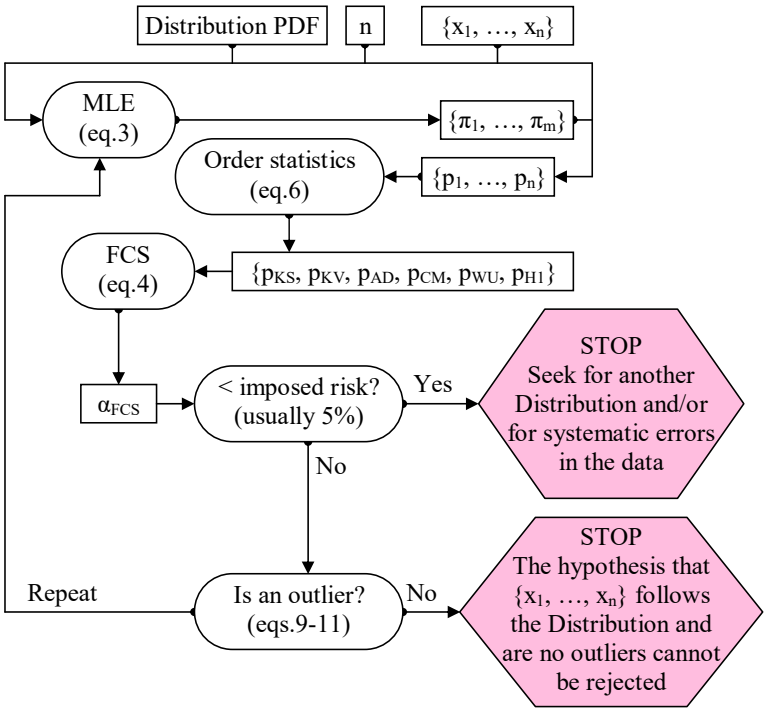

Figure 3. The procedure for detecting outliers.

9. Second Simulation Assessing "Grubbs" and "g1" Outlier Detection Alternatives

Another MC study was designed to test the claim that the proposed method provides consistent results. This second MC simulation is much simpler than the one used to derive the data for constructing the outlier statistics (Figure 4).

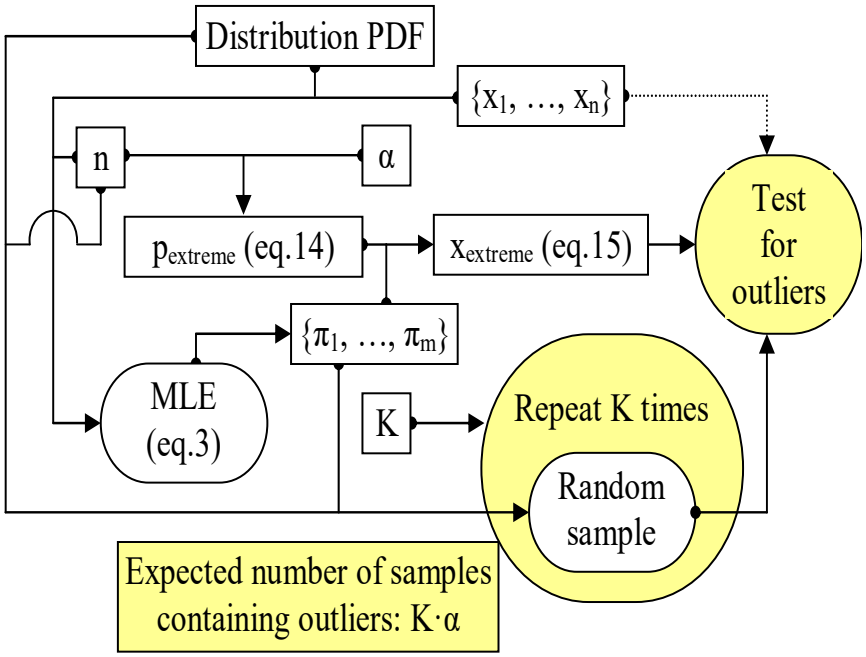

Figure 4. The procedure for testing the outlier statistics.

The data used here as a proof of the facts are from [7] and all cases involve a Normal distribution (Distribution = Normal in Equation (15); PDF and CDF for Normal distribution in Equation (18); a symmetrical distribution) with α = 5% risk being in error. The parameters of the Normal distribution (μ and σ) are determined for each case, as well as the sample size (Equation (17)).

$$x_{\text{extreme}}(\alpha) = \text{InvCDF}_{\text{“Normal”}}(0.5 \pm 0.5 \cdot \sqrt[n]{1-\alpha}; \mu, \sigma) \tag{17}$$

$$\text{PDF}_{\text{“Normal”}}(x; \mu, \sigma) = \frac{e^{-\frac{(x-\mu)^2}{\sigma^2}}}{\sigma\sqrt{2\pi}}, \ \text{CDF}_{\text{“Normal”}}(x; \mu, \sigma) = \int_{-\infty}^{x} \text{PDF}_{\text{“Normal”}}(t; \mu, \sigma)dt \tag{18}$$

For comparison, the same strategy for calculating the confidence intervals of the extreme values for the Normal distribution with the Grubbs test statistic (Equation (2)) was used to provide an alternate result (Equation (19)).

$$x_{\text{crit}}(\alpha) = \overline{x} \pm G_{\text{crit}}(\alpha) \cdot s, G_{\text{crit}}(\alpha) = \frac{n-1}{\sqrt{n}} \sqrt{\frac{t_G^2(\alpha)}{n-2+t_G^2(\alpha)}}, \ t_G = \text{InvCDF}_{\text{“Student t”}}(\frac{\alpha}{2n}, n-2) \tag{19}$$

The steps followed in this analysis are given in the Table 5.

Table 5. Comparison of the steps of the analysis and simulation for extreme values confidence intervals (proposed method vs. Grubbs test)

Step	Action (step 0 is setting the dataset; $\alpha \leftarrow 0.05$)
1	Estimate (with MLE, Equation (3)) parameters (μ, σ) of the Normal distribution; calculate the associated CDFs (Equation (18))
2	Calculate the order statistics, their associated risks being in error, FCS and p_{FCS} (Equations (6) and (4))
3	For n and α calculate the confidence intervals for the extreme values by using (a) Equation (6) and (17) and (b) Equation (19)
4	Run the MC experiment (Figure 4) for K = 10000 (and then the expected number of outliers is 500) samples and count the samples containing outliers for the existing method (Grubbs, Equation (19); with μ and σ from CM method) and for the proposed method (g1, Equations (13)–(15) and (17); with μ and σ from the MLE method)

Results of the analysis using the steps given in Table 5 for the first dataset are given in Table 6.

Table 6. Outlier analysis results for {568, 570, 570, 570, 572, 572, 572, 578, 584, 596} dataset.

Step	Results (for α = 5%)							
1	μ = 575.2; σ = 8.256 (MLE) → CPs = {0.1916, 0.2644, 0.2644, 0.2644, 0.3492, 0.3492, 0.3492, 0.6328, 0.8568, 0.9941}							
2	**Statistic**	**AD**	**KS**	**CM**	**KV**	**WU**	**H1**	**FCS**
	Value	1.137	1.110	0.206	1.715	0.182	5.266	12.293
	$\alpha_{Statistic}$	0.288	0.132	0.259	0.028	0.049	0.343	0.056
3	$x_{crit}(5\%) = 575.2 \pm 2.29 \cdot 8.7025$; $p_{extreme}(5\%) = 0.5 \pm InvCDF_{"g1"}(1\text{-}0.05; 10)$; $x_{extreme}(5\%) = \{552.086, 598.314\}$							
4	**Number of samples containing outliers**		**Existing method (Grubbs)**		**Proposed method (g1)**			
	First run		1977 (19.77%)		510 (5.1%)			
	Second run		2009 (20.09%)		526 (5.26%)			

In regard to the results given in Table 6:

At step 1, CPs are the cumulative probabilities ($\{p_1, \ldots, p_{10}\}$ in Figure 3) associated with the series of the observations from the sample ($\{x_1, \ldots, x_{10}\}$ in Figure 3).

At step 2, the data passes the normality test (α_{FCS} = 7% > 5% = α, see Figure 3).

Step 3 was made for n = 10 (see Figure 4). (a) The proposed method does not detect outliers in the sample (552.086 < 568, 596 < 598.314); (b) Grubbs test detect 596 as being an outlier (596 > 595.13).

At step 4 (see Figure 4), since {510, 526} are comparable with 500 and {1977, 2009} are much greater than 500, the results lead to the conclusion that the existing method produces type I errors by leading to false positive detection of outliers in the samples while the proposed method does not.

10. Going Further with the Outlier Analysis

What if "596" is removed from the sample? The following table provides mirror-like results for this scenario (Table 7).

Table 7. Outlier analysis results for {568, 570, 570, 570, 572, 572, 572, 578, 584} dataset.

Step	Results (for α = 5%)							
1	μ = 572.889; σ = 4.725 (MLE) → CPs = {0.1504, 0.2705, 0.2705, 0.2705, 0.4254, 0.4254, 0.4254, 0.8603, 0.9907}							
2	**Statistic**	**AD**	**KS**	**CM**	**KV**	**WU**	**H1**	**FCS**
	Value	0.935	1.057	0.174	1.535	0.155	4.678	9.715
	$\alpha_{Statistic}$	0.389	0.167	0.327	0.082	0.088	0.394	0.137
3	$x_{crit}(5\%) = 572.89 \pm 2.215 \cdot 5.011$; $p_{extreme}(5\%) = 0.5 \pm InvCDF_{"g1"}(1\text{-}0.05; 9)$; $x_{extreme}(5\%) = \{559.822, 585.956\}$							
4	**Number of samples containing outliers**		**Existing method (Grubbs)**		**Proposed method (g1)**			
	First run		2341 (23.41%)		563 (5.63%)			
	Second run		2333 (23.33%)		543 (5.43%)			

As can be observed in Table 7, the data is not in good agreement with normality (α_{FCS} in Table 6 is 7%, while in Table 7 it is 16%) and there is no change in the accuracy of the classification ({563, 543} comparable with 500, {2341, 2333} is much greater than 500; the existing method produces type I errors by leading to false positive detection of outliers in the samples, while the proposed method does not). When comparing the results given in Table 6 with the results given in Table 7 it should be noted that both tests (Grubbs and the newly proposed g1) produce somewhat confusing results (see Table 8 for side-by-side outcomes).

Table 8. Side-by-side comparison of the analysis of the samples.

Sample	{568, 570, 570, 570, 572, 572, 572, 578, 584, 596}	{568, 570, 570, 570, 572, 572, 572, 578, 584}
At 5% risk being in error can the hypothesis that the sample was drawn from a normal distribution be rejected?	No (α_{FCS} = 7%)	No (α_{FCS} = 15.8%)
Grubbs confidence interval for 'no outliers' at 5% risk being in error	(555.27, 595.13) 596 is detected as being outlier	(561.79, 583.99) 584 is detected as being outlier
g1 confidence interval for 'no outliers' at 5% risk being in error	(552.08, 598.32) no outliers	(559.82, 585.96) no outliers

Table 8 highlights the fact that based on the {568, 570, 570, 570, 572, 572, 572, 578, 584} sample, the g1 test may be interpreted as identifying 596 as being an outlier. This is not quite true because the g1 test was not intended to be used in this way. That is, 596 is outside of the dataset, so at the time of constructing the confidence intervals for the extreme values, the information regarding its observation was missing.

Another trial was done, this time with **601** replacing **596** in the initial dataset (Table 9).

Table 9. Outlier analysis results for the {568, 570, 570, 570, 572, 572, 572, 578, 584, **601**} dataset.

Step	Results (for α = 5%)							
1	From the CM method: μ = 575.7; σ = 10.067; from MLE method: μ = 575.7; σ = 9.550							
2	**Statistic**	**AD**	**KS**	**CM**	**KV**	**WU**	**H1**	**FCS**
	Value	1.267	1.109	0.225	1.774	0.198	5.411	13.652
	$\alpha_{Statistic}$	0.241	0.132	0.226	0.018	0.035	0.254	0.034
3	Grubbs confidence interval for 'no outliers' at 5% risk being in error: (552.647,598.753); **601** is an outlier g1 confidence interval for 'no outliers' at 5% risk being in error: (548.963, 602.437); no outliers							

In a further trial, **604** replaced **596** in the initial dataset (Table 10).

Table 10. Outlier analysis results for the {568, 570, 570, 570, 572, 572, 572, 578, 584, 604} dataset.

Step	Results (for α = 5%)							
1	From the CM method: μ = 576.0; σ = 10.914; from MLE method: μ = 576.0; σ = 10.354							
2	**Statistic**	**AD**	**KS**	**CM**	**KV**	**WU**	**H1**	**FCS**
	Value	1.348	1.108	0.238	1.803	0.209	5.481	14.468
	$\alpha_{Statistic}$	0.216	0.133	0.206	0.015	0.028	0.215	0.025
3	Grubbs confidence interval for 'no outliers' at 5% risk being in error: (551.00, 601.00); 604 is an outlier g1 confidence interval for 'no outliers' at 5% risk being in error: (547.01, 604.99); no outliers							

The conclusion is simple (see the results in the Tables 6, 7, 9 and 10): A test will hardly ever detect an outlier for a small sample; it is more likely to reject the hypothesis of the sample drawn from the distribution itself!

The same trick was used on a bigger sample and the results are shown in Table 11 (the dataset is from Table 4).

Table 11. Outlier analysis results for Table 4 dataset under the assumption of normal distribution.

Step	Results (for α = 5%)							
1	Table 5 Dataset; Normal distribution → CM: μ = 6.481; σ = 0.831; MLE: μ = 6.481; σ = 0.829							
2	**Statistic**	**AD**	**KS**	**CM**	**KV**	**WU**	**H1**	**FCS**
	Value	0.439	0.484	0.049	0.952	0.047	104.2	1.276
	$\alpha_{Statistic}$	0.812	0.965	0.886	0.852	0.743	0.641	0.973
3	Grubbs confidence interval for 'no outliers' at 5% risk being in error: (3.492, 9.470); 9.603 is an outlier g1 confidence interval for 'no outliers' at 5% risk being in error: (3.444, 9.517); 9.603 is an outlier							
4	**Number of samples containing outliers**			**Existing method (Grubbs)**		**Proposed method (g1)**		
	First run			637 (6.37%)		511 (5.11%)		
	Second run			630 (6.3%)		481 (4.81%)		

On one hand, as the results in Table 11 prove, the proposed method correctly identifies the confidence interval for the extreme values, while the existing method does not.

On the other hand, the results in Table 11 also show that the likelihood of identifying the outliers increases with the sample size, making it perfectly possible to identify outliers with the proposed method, although this is not the case in small samples. It is possible to detect the outliers in small samples as well, but not when the parameters of the distribution are derived from the sample data—only when the parameters of the distribution are known a priori or determined from other samples (the results given in Tables 6–10 are proof of this).

11. Further Discussion

The obtained expression for CDF of "g1" (Equation (11)) reveals the domain of a random variable distributed by the "g1" statistic ([0, 0.5]), which is consistent with the definition of "g1" (Equations (9) and (10)).

Independently of the shape of the theoretical distribution being tested (the generic case is defined by Equation (5)), as defined by Equations (9) and (10), the newly proposed statistic "g1" defines a symmetric confidence interval for the extreme values in samples in the probability space (Equation (14)). Later, this symmetric confidence interval may be changed back into an asymmetrical one when it is expressed in the domain of the theoretical distribution being tested (Equation (15)). It should be recognized that "g1" uses a symmetrization strategy to obtain the confidence interval for the extreme values in samples.

It might seem that the literature on robust statistics was ignored in this work, however, this is not entirely true. In fact, a whole pool of robust statistics was used extensively in the study (see Equation (8)), introduced as a tool in Table 5 and involved in the later calculations (Tables 6, 7 and 9, Tables 10 and 11). Also, it should be noted that the substitution of the mean by the median is not a new idea; it is well known in the field of robust statistics (for example, Watson U^2 [29], the $WU_{Statistic}$ in Equation (8), uses it).

A short literature survey provides several of examples of current real applications that require the proposed method. Thus, in signal processing, non-stationary, non-Gaussian, spiky signals are usually regarded as outliers and thus discarded (see [35–38] as typical cases). In this context, it should be noted that Mood's median test is preferred to the Kruskal-Wallis test when outliers are present [39]. The identification of outliers is also recognized as an issue in the validation of protein structures, and the current methods are revised in [40]. Other examples can be found in [41].

In the wider context, an alternate window-based strategy has been proposed in which outliers are detected in each window by the Tukey method and labeled so that they can be excluded from the realization of the process points to be used for model identification [42]. A contingency-based strategy proposes maximization of true positive (TP) values and minimization of false negative (FN) and false positive (FP) values [43]. Finally, another distribution testing procedure has been proposed in [44].

12. Conclusions

A new method for detecting outliers was proposed in this paper. The method is applicable to any continuous distribution at any risk being in error. It was proved that the method correctly detects the outliers. For a normal distribution at 5% risk being in error, it was also shown that the proposed method outperforms the classical Grubbs test for detecting the outliers.

Supplementary Materials: Details of the software used for deriving the results given in the figures and tables, algorithms and source codes are given as supplementary material available online at http://www.mdpi.com/2073-8994/11/6/835/s1.

Funding: This research received no external funding.

Acknowledgments: Thanks to my colleague S.D. Bolboacă and for our fruitful discussions during the development stage of the study, which helped and motivated the author to complete the study.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Gauss, C.F. *Theoria Motus Corporum Coelestium*; (Translated in 1857 as "Theory of Motion of the Heavenly Bodies Moving about the Sun in Conic Sections" by C. H. Davis. Little, Brown: Boston. Reprinted in 1963 by Dover: New York); Perthes et Besser: Hamburg, Germany, 1809; pp. 249–259.
2. Tippett, L.H.C. The extreme individuals and the range of samples taken from a normal population. *Biometrika* **1925**, *17*, 151–164. [CrossRef]
3. Fisher, R.A.; Tippett, L.H.C. Limiting forms of the frequency distribution of the largest and smallest member of a sample. *Proc. Camb. Philos. Soc.* **1928**, *24*, 180–190. [CrossRef]
4. Thompson, W.R. On a criterion for the rejection of observations and the distribution of the ratio of the deviation to the sample standard deviation. *Ann. Math. Stat.* **1935**, *6*, 214–219. [CrossRef]
5. Pearson, E.; Sekar, C.C. The efficiency of the statistical tools and a criterion for the rejection of outlying observations. *Biometrika* **1936**, *28*, 308–320. [CrossRef]
6. Grubbs, F.E. Sample criteria for testing outlying observations. *Ann. Math. Stat.* **1950**, *21*, 27–58. [CrossRef]
7. Grubbs, F.E. Procedures for detecting outlying observations in samples. *Technometrics* **1969**, *11*, 1–21. [CrossRef]
8. Nooghabi, M.; Nooghabi, H.; Nasiri, P. Detecting outliers in gamma distribution. *Commun. Stat. Theory Methods* **2010**, *39*, 698–706. [CrossRef]
9. Kumar, N.; Lalitha, S. Testing for upper outliers in gamma sample. *Commun. Stat. Theory Methods* **2012**, *41*, 820–828. [CrossRef]
10. Lucini, M.; Frery, A. Comments on "Detecting Outliers in Gamma Distribution" by M. Jabbari Nooghabi et al. (2010). *Commun. Stat. Theory Methods* **2017**, *46*, 5223–5227. [CrossRef]
11. Hartley, H. The range in random samples. *Biometrika* **1942**, *32*, 334–348. [CrossRef]
12. Bardet, J.-M.; Dimby, S.-F. A new non-parametric detector of univariate outliers for distributions with unbounded support. *Extremes* **2017**, *20*, 751–775. [CrossRef]
13. Gosset, W. The probable error of a mean. *Biometrika* **1908**, *6*, 1–25.
14. Jäntschi, L.; Bolboacă, S.-D. Computation of probability associated with Anderson-Darling statistic. *Mathematics* **2018**, *6*, 88. [CrossRef]
15. Fisher, R. On an Absolute Criterion for Fitting Frequency Curves. *Messenger Math.* **1912**, *41*, 155–160.
16. Fisher, R. Questions and answers #14. *Am. Stat.* **1948**, *2*, 30–31.
17. Bolboacă, S.D.; Jäntschi, L.; Sestraș, A.F.; Sestraș, R.E.; Pamfil, D.C. Supplementary material of 'Pearson-Fisher chi-square statistic revisited'. *Information* **2011**, *2*, 528–545. [CrossRef]
18. Jäntschi, L.; Bolboacă, S.D. Performances of Shannon's Entropy Statistic in Assessment of Distribution of Data. *Ovidius Univ. Ann. Chem.* **2017**, *28*, 30–42. [CrossRef]
19. Davis, P.; Rabinowitz, P. *Methods of Numerical Integration*; Academic Press: New York, NY, USA, 1975; pp. 51–198.
20. Pearson, K. Note on Francis Gallon's problem. *Biometrika* **1902**, *1*, 390–399.
21. Cramér, H. On the composition of elementary errors. *Scand. Actuar. J.* **1928**, *1*, 13–74. [CrossRef]
22. Von Mises, R.E. *Wahrscheinlichkeit, Statistik und Wahrheit*; Julius Springer: Berlin, Germany, 1928; pp. 100–138.

23. Kolmogorov, A. Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari* **1933**, *4*, 83–91.
24. Kolmogorov, A. Confidence Limits for an Unknown Distribution Function. *Ann. Math. Stat.* **1941**, *12*, 461–463. [CrossRef]
25. Smirnov, N. Table for estimating the goodness of fit of empirical distributions. *Ann. Math. Stat.* **1948**, *19*, 279–281. [CrossRef]
26. Anderson, T.W.; Darling, D.A. Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. *Ann. Math. Stat.* **1952**, *23*, 193–212. [CrossRef]
27. Anderson, T.W.; Darling, D.A. A Test of Goodness-of-Fit. *J. Am. Stat. Assoc.* **1954**, *49*, 765–769. [CrossRef]
28. Kuiper, N.H. Tests concerning random points on a circle. *Proc. Koninklijke Nederlandse Akademie van Wetenschappen Series A* **1960**, *63*, 38–47. [CrossRef]
29. Watson, G.S. Goodness-Of-Fit Tests on a Circle. *Biometrika* **1961**, *48*, 109–114. [CrossRef]
30. Metropolis, N.; Ulam, S. The Monte Carlo Method. *J. Am. Stat. Assoc.* **1949**, *44*, 335–341. [CrossRef] [PubMed]
31. Fisher, R.A. On the mathematical foundations of theoretical statistics. *Philos. Trans. R. Soc. A* **1922**, *222*, 309–368. [CrossRef]
32. Jäntschi, L. Distribution fitting 1. Parameters estimation under assumption of agreement between observation and model. *Bull. UASVM Hortic.* **2009**, *66*, 684–690.
33. Jäntschi, L.; Bolboacă, S.D. Distribution fitting 2. Pearson-Fisher, Kolmogorov-Smirnov, Anderson-Darling, Wilks-Shapiro, Kramer-von-Misses and Jarque-Bera statistics. *Bull. UASVM Hortic.* **2009**, *66*, 691–697.
34. Bolboacă, S.D.; Jäntschi, L. Distribution fitting 3. Analysis under normality assumption. *Bull. UASVM Hortic.* **2009**, *66*, 698–705.
35. Liu, K.; Chen, Y.Q.; Domański, P.D.; Zhang, X. A novel method for control performance assessment with fractional order signal processing and its application to semiconductor manufacturing. *Algorithms* **2018**, *11*, 90. [CrossRef]
36. Paiva, J.S.; Ribeiro, R.S.R.; Cunha, J.P.S.; Rosa, C.C.; Jorge, P.A.S. Single particle differentiation through 2D optical fiber trapping and back-scattered signal statistical analysis: An exploratory approach. *Sensors* **2018**, *18*, 710. [CrossRef] [PubMed]
37. Teunissen, P.J.G.; Imparato, D.; Tiberius, C.C.J.M. Does RAIM with correct exclusion produce unbiased positions? *Sensors* **2017**, *17*, 1508. [CrossRef] [PubMed]
38. Pan, Z.; Liu, L.; Qiu, X.; Lei, B. Fast vessel detection in Gaofen-3 SAR images with ultrafine strip-map mode. *Sensors* **2017**, *17*, 1578. [CrossRef] [PubMed]
39. Vergura, S.; Carpentieri, M. Statistics to detect low-intensity anomalies in PV systems. *Energies* **2018**, *11*, 30. [CrossRef]
40. Chen, L.; He, J.; Sazzed, S.; Walker, R. An investigation of atomic structures derived from X-ray crystallography and cryo-electron microscopy using distal blocks of side-chains. *Molecules* **2018**, *23*, 610. [CrossRef] [PubMed]
41. Bolboacă, S.D.; Jäntschi, L. The effect of leverage and influential on structure-activity relationships. *Comb. Chem. High Throughput Screen.* **2013**, *16*, 288–297. [CrossRef] [PubMed]
42. Faes, L.; Porta, A.; Nollo, G.; Javorka, M. Information decomposition in multivariate systems: Definitions, implementation and application to cardiovascular networks. *Entropy* **2017**, *19*, 5. [CrossRef]
43. Li, G.; Wang, J.; Liang, J.; Yue, C. Application of sliding nest window control chart in data stream anomaly detection. *Symmetry* **2018**, *10*, 113. [CrossRef]
44. Paolella, M.S. Stable-GARCH models for financial returns: Fast estimation and tests for stability. *Econometrics* **2016**, *4*, 25. [CrossRef]

symmetry

MDPI

Supplementary material

Supplementary material for "A test detecting the outliers for continuous distributions based on the cumulative distribution function of the data being tested"

Lorentz Jäntschi [1,2,*]

[1] Department of Physics and Chemistry, Technical University of Cluj-Napoca, 400641 Cluj, Romania; lorentz.jantschi@chem.utcluj.ro

[2] Chemical Doctoral School, Babeș-Bolyai University, 400084 Cluj-Napoca, Romania

* Correspondence: lorentz.jantschi@gmail.com; Tel.: +4-0264-401-775

Received: 12 June 2019; Accepted: 21 June 2019; Published: date

Summary: Software usage details for deriving the results given in figures and tables, algorithms and source codes are given in this supplementary material online available.

Introduction

FreePascal [1] were used to compile the executable for Monte-Carlo (MC) sampling (implements the peculiarities of MC simulation given in Table 2 in the paper). Data were generated by running the program (32bit executable). PHP v.7.3.1 cli [2] was used to assess numerically the agreement for "g1" statistic (Table 3), to calculate the statistics (eqs.6; used as given in Fig.3, Step 2 in Tables 5, 6, 7, 9, 10, and 11), and to implement (Fig.4) and run the second MC simulation study (Step 4 in Tables 5, 6, and 7). EasyFit [3] was used to obtain initial estimates (uses a hybrid CM-MLE method) for the population parameters. MS Excel [4] was used to do simple calculations, including FCS (eq.4). EasyFitXL [5] was used to calculate CDF of Gauss Laplace (Table 4), of χ^2 (eq.4; results of those calculations appears on Tables 5, 6, 7, 9, 10, and 11 in Step 2), and of Normal and Student (eq.17 and 19; Step 3 in Tables 5, 6, 7, 9, 10, and 11). Mathcad [6] was used to obtain MLE estimates (eq.3) for population parameters (Table 4). l.academicdirect.org/Statistics/tests [7] was used to calculate the associated probabilities with the statistics (eqs.6; used as given in Fig.3, Step 2 in Tables 5, 6, 7, 9, 10, and 11). Matlab [8] was used to obtain Figs. 1 and 2.

Algorithms and Source codes

*The algorithm for the first MC simulation study (referred in Table 2 in the paper; N ← 10; R ← 10; S ← 1000*100000)*

```
For n from 2 to N
  For resampling from 0 to R
    For sampling from 0 to S
      sample ← Generate n probabilities from uniform distribution
      g1[sampling] ← g1 statistic for the sample //eqs. 9-10 in the paper
    EndFor
    Sort the g1[0..S] array of statistics
    For grid_point from 1 to 999
      Extract g1g[grid_point][resampling] from g1[0..S]
    EndFor
```

```
  EndFor //collect 11 resamples of the g1 statistic in the given grid points
  For each grid_point from 1 to 999
    Sort the g1g[grid_point][0..R] array of statistics
    Extract g1g[grid_point][R/2] from g1g[grid_point][0..R]
    Output n, grid_point, g1g[grid_point][R/2] //output data of MC simulation
  EndFor //median (5 for 0..10) as statistic for order statistics
EndFor
```

The source code for the first MC simulation study software (referred in Table 2 in the paper; FreePascal language)

```
const
  s_resa=10; //'resa' in Table 2 in the paper (T2paper)
  s_leve=1000; //'p'+1 in T2paper
  s_bloc=100000; s_size=s_leve*s_bloc; s1size=s_size-1; //'m' in T2paper
  n_min=2; n_max=60; // 'n' in T2paper
type
  st_type=array[0..s1size]of extended;
  praguri=array[1..s_leve-1]of extended; //1..999
  probabi=array[0..n_max]of extended; //one sample of probabilities
  npragur=array[1..s_leve-1]of probabi; //grid for distribution of statistic
function Rnd2:Extended;var i:byte;r:Extended; begin r:=0.0; //binary random
  for i:=0 to 63 do r:=r/2.0+random(2);r:=r/2.0;if(r>1.0)then r:=1.0;Rnd2:=r;
end; //FreePascal random uses Mersenne twister
procedure QuickSortP(var A:probabi;Lo,Hi:LongInt);
  procedure SortP(l,r:LongInt);var i,j:LongInt;x,y:Extended; begin
    i:=l;j:=r;x:=a[(l+r)DIV 2];
    repeat while(a[i]<x)do i:=i+1; while(x<a[j])do j:=j-1;
      if(i<=j)then begin y:=a[i];a[i]:=a[j];a[j]:= y;i:=i+1;j:=j-1;end;
    until(i>j); if(l<j)then SortP(l,j);if(i<r)then SortP(i,r);
  end; begin SortP(Lo,Hi);
end; //sorts an array of probabilities
procedure QuickSortA(var A:st_type;Lo,Hi:LongInt);
  procedure SortP(l,r:LongInt);var i,j:LongInt;x,y:Extended; begin
    i:=l;j:=r;x:=a[(l+r)DIV 2];
    repeat while(a[i]<x)do i:=i+1; while(x<a[j])do j:=j-1;
      if(i<=j)then begin y:=a[i];a[i]:=a[j];a[j]:= y;i:=i+1;j:=j-1;end;
    until(i>j);if(l<j)then SortP(l,j);if(i<r)then SortP(i,r);
  end; begin SortP(Lo,Hi);
end; //sorts an array of statistics
function G1S(n:byte;var f:prob):Extended;var S,T:Extended;i:byte; begin
  S:=abs(f[n]-0.5);
  for i:=n-1 downto 1 do begin T:=abs(f[i]-0.5);if(T>S)then S:=T;end;
  G1s:=S;
```

```
end; //calculates the 'g1' statistic for a sample 'f' of size 'n'
procedure ST_Sample(n: LongInt; var A2:st_type; var A2_low, A2_upp: praguri);
  var i:LongInt;j,k:LongInt;p:probabi; begin
  for j:= s1size downto 0 do begin
    for i:= n-1 downto 0 do p[i]:=Rnd2; A2[j]:=G1S(n,p);
  end; QuickSortA(A2,0,s1size);
  for i:= s_leve-1 downto 1 do begin
    A2_low[i]:=A2[i*s_bloc-1]; A2_upp[i]:=A2[i*s_bloc];
  end;
end; //samples the statistic
procedure ST_Resample(n:LongInt; var A2:st_type; var A2lr,A2ur:praguri);
  var A2l_r,A2u_r:npragur;j,i:LongInt; begin
  for j:= 0 to s_resa do begin
    ST_Sample(n,A2,A2lr,A2ur);
    for i:= s_leve-1 downto 1 do begin
      A2l_r[i][j]:=A2lr[i];A2u_r[i][j]:=A2ur[i];
    end;
  end;
  for i:= s_leve-1 downto 1 do begin
    QuickSortP(A2l_r[i],0,s_resa);A2lr[i]:=A2l_r[i][s_resa div 2];
    QuickSortP(A2u_r[i],0,s_resa);A2ur[i]:=A2u_r[i][s_resa div 2];
  end;
end; //resamples the statistic
var
 b:st_type;
 b_low,b_upp:praguri;
 i,j:LongInt;
 f:text;s:string[2];
begin Randomize;
  for i:= n_min to n_max do begin
    str(i,s);
    ST_Resample(i,b,b_low,b_upp);
    assign(f,s+'_bin_rnd_'+'G1S.txt'); rewrite(f);
    for j:= 1 to s_leve-1 do writeln(f,j,chr(9),b_low[j],chr(9),b_upp[j]);
    close(f); writeln('resample for ',i,' completed.');
  end;
end. //b_low[j] and b_upp[j] have the same digits till a point (of accuracy)
```

Source code for calculation of the assessing of the agreement between observed and expected (implementing eq.12, as eq10 vs. eq.11 from the paper; PHP language)

```
for($n=10;$n>1;$n--){ $x=array();
  $a=explode("\r\n",file_get_contents($n."_bin_rnd_G1S.txt"));
  for($p=1;$p<1000;$p++){$b=explode("\t",$a[$p]);$x[$p]=$b[1];}
```

```
    $ss=0;
    for($p=999;$p>0;$p--) $ss+=pow($p/1000-pow(2*$x[$p],$n),2);
    $se=sqrt($s/999); echo($n."\t".$s."\t".$se."\r\n");
}
```

Source code for the second MC simulation study software (Fig.4; PHP language; it uses NormalDistribution.php [9])

```
<?php
  include("NormalDistribution.php");//implements normal distribution object
  define("K_runs",10000);
  define("S_size",10);//this is for Tables 6 and 7 calculations
  // define("S_size",206) for Table 11 calculations
  //below are data calculated outside and given here
  $po_gr=array(575.200, 8.702); //MC estimations for μ and σ
  $mm_gr=array(555.271, 595.129);//CI extreme val. grubbs
  $po_g0=array(575.200, 8.256); //MLE estimations for μ and σ
  $mm_g0=array(552.086, 598.314); //CI extreme val. g1
  //data from above is for Table 6 calculations given in the paper
  /*
  $po_gr=array(572.889, 5.011); //MC estimations for μ and σ
  $mm_gr=array(561.789, 583.989);//CI extreme val. grubbs
  $po_g0=array(572.889, 4.725); //MLE estimations for μ and σ
  $mm_g0=array(559.821, 585.957); //CI extreme val. g1
  //data from above is for Table 7 calculations given in the paper
  */
   /*
  $po_gr=array(6.481, 0.831); //MC estimations for μ and σ
  $mm_gr=array(3.492, 9.470); //CI extreme val. grubbs
  $po_g0=array(6.481, 0.829); //MLE estimations for μ and σ
  $mm_g0=array(3.444, 9.517); //CI extreme val. g1
  //data from above is for Table 11 calculations given in the paper
  */
  $st_gr=0;
  $st_g0=0;
  $my_norm = new NormalDistribution(0,1);
  for($i=0;$i<K_runs;$i++){//samples
     $s=array(); for($j=0;$j<S_size;$j++) $s[]=$my_norm->_getRNG();
     //_getRNG() uses mt_rand(); mt_rand() implements Mersenne twister
     for($j=0;$j<S_size;$j++){
        if($po_gr[0]+$s[$j]*$po_gr[1]<$mm_gr[0]){$st_gr++;break;}
        if($po_gr[0]+$s[$j]*$po_gr[1]>$mm_gr[1]){$st_gr++;break;}
     }//st_grubbs
     for($j=0;$j<S_size;$j++){
```

```
        if($po_g1[0]+$s[$j]*$po_g1[1]<$mm_g1[0]){$st_g1++;break;}
        if($po_g1[0]+$s[$j]*$po_g1[1]>$mm_g1[1]){$st_g1++;break;}
      }//st_g1
    }
    echo("outliers grubbs: ".$st_gr."\r\n");
    echo("outliers g1: ".$st_g1."\r\n");
    echo("expected number of outliers: ".(0.05*K_runs)."\r\n");
?>
```

It should be noted that the second MC simulation study software uses a mirror (or symmetrical) strategy for comparison of the "g1" and Grubbs methods - the same U(0,1) random drawings are used to feed the both methods (see $s[]=$my_norm->_getRNG(); in the code above).

Source code (MathCad language) for the MLE estimations for Gauss-Laplace distribution (eq.3; results given in Table 4 in the paper)

$X := \text{READPRN("d\_206.txt")}$

$c0(ka) := \sqrt{\frac{\Gamma\left(\frac{3}{ka}\right)}{\Gamma\left(\frac{1}{ka}\right)}} \quad c1(ka) := \frac{ka \cdot c0(ka)}{2 \cdot \Gamma\left(\frac{1}{ka}\right)}$

$MleGgl(x,md,si,ka) := -\ln(si) - \ln(c1(ka)) + \left(\left|c0(ka) \cdot \frac{x-md}{si}\right|\right)^{ka}$

$Mle(md,si,ka) := \sum_{i=0}^{rows(X)-1} MleGgl(X_i,md,si,ka)$

$md := 4806 \qquad si := 0.83017 \quad ka := 1.4645$

Given

$\frac{d}{dmd} Mle(md,si,ka) = 0 \quad \frac{d}{dsi} Mle(md,si,ka) = 0 \quad \frac{d}{dka} Mle(md,si,ka) = 0$

$Y := Find(md,si,ka)$

$Y = \begin{pmatrix} 6.47938 \\ 0.82828 \\ 1.79106 \end{pmatrix}$

Source code (MathCad language) for the MLE estimations for Normal distribution (eq.3; results given in Tables 6, 7, 9, 10 and 11 in the paper)

$X := \text{READPRN(Datafile)}$

$MleNor(x,md,se) := -\ln\left(\sqrt{2 \cdot \pi}\right) - \ln(se) - \frac{1}{2} \cdot \left(\frac{x-md}{se}\right)^2$

$Mle(md,se) := \sum_{i=0}^{rows(X)-1} MleNor(X_i,md,se)$

$md := mean(X) \qquad se := Stdev(X)$

Given

$\frac{d}{dmd} Mle(md,se) = 0 \quad \frac{d}{dse} Mle(md,se) = 0$

$Y := Find(md,se)$

$Y = \begin{pmatrix} \end{pmatrix}$

Datafiles and the tables containing the results

Datafile	Table in the paper
"d_10.txt"	Table 6
"d_10_9.txt"	Table 7
"d_10_601.txt"	Table 9
"d_10_604.txt"	Table 10
"d_206.txt"	Table 11

Source code for the plots (to do Fig.1 and Fig.2; Matlab language)

<table>
<tr><th>Source code for Fig.1</th><th>Source code for Fig.2</th></tr>
<tr><td>XYZ = load('g1t_err_plot.txt');</td><td>XYZ = load('g1t_plot.txt');</td></tr>
<tr><td colspan="2">x = XYZ(:,1);
y = XYZ(:,2);
z = XYZ(:,3);
plot3(x,y,z,'.-')
tri = delaunay(x,y);
plot(x,y,'.')
h = trisurf(tri, x, y, z);
set(0,'defaulttextInterpreter','latex');</td></tr>
<tr><td>xlabel('n');
ylabel('p');
zlabel('CDF_{"g1"}(x;n) - p');</td><td>xlabel('x');
ylabel('n');
zlabel('CDF_{"g1"}(x;n)');</td></tr>
<tr><td colspan="2">view(-60,15);</td></tr>
</table>

References

1. Carl E. CODÈRE, Daniël MANTIONE, Florian KLÄMPFL, Jonas MAEBE, Michael Van CANNEYT, Peter VREMAN, Pierre MULLER, Marco van de VOORT, Leon de BOER, Armin DIEHL, Casey DUNCAN, Berczi GABOR, Sebastian GUENTHER, Tomas HAJNY, John LEE, Mark MAY, Mazen NEIFER, Olle RAAB, Thomas SCHATZL, Balazs SCHEIDLER, Nils SJOHOLM, MH SPIEGEL, Gernot TENCHIO, Erik WACHTMEESTER, Frank ZAGO, Gertjan SCHOUTEN, Karoly BALOGH, 1998. FreePascal: open source compiler for Pascal and Object Pascal (v.0.99.10); 2000: FreePascal v.1.0; 2005: FreePascal v.2.0; 2009: FreePascal v.2.2.4; 2011: FreePascal v.2.4.2; 2012: FreePascal v.2.6.0. 2017: FreePascal v.3.0.4.
2. Rasmus LERDORF, 1994. PHP/FI - initially from Personal Home Page tools (open source online since 1995; PHP v.1.0); Andi GUTMANS, Zeev SURASKI, 1997. PHP3 (open source online since 1998; PHP v.3.0); Andi GUTMANS, Zeev SURASKI, 1998. PHP4 (with 'Zend' engine; open source online since 1999; PHP v.4.0); Andi GUTMANS, Zeev SURASKI, 2004. PHP5 (with 'Zend' engine 2.0; open source online since 2004; PHP v.5.0). Dmitry STOGOV, Xinchen HUI, Nikita POPOV, 2014. PHP7 (with 'Zend' engine 3.0; open source online since 2015; PHP v.7.3.1).
3. MathWave Technologies, 2009. EasyFit Professional v.5.2 (© 2004-2009).
4. Microsoft, 1987. MS Excel v.2.05. Microsoft, 2013. MS Excel v.15.0.5127.
5. MathWave Technologies, 2009. EasyFitXL (MS Excel add-on), v.5.2.
6. Allen RAZDOW, 1986. MathCad v.0.3 (beta on 5 1/4 floppy, DOS version). PTC (Parametric Technology Corporation), 2007. Mathcad v.14.0.0.
7. Lorentz JÄNTSCHI, 2008. Statistics tests. Online: http://l.academicdirect.org/Statistics/tests
8. Jack LITTLE, Steve BANGERT, James Hardy WILKINSON, 1984. MATLAB v.1.0 (MatLab - from Matrix Laboratory); MathWorks Inc, 2015. MATLAB v.8.5.0 (version R2015a, 32bit, win32).
9 . Jaco van KOOTEN, Paul MEAGHER, 2013. Version 1.3 of NormalDistribution class: an object for encapsulating normal distributions (PHP implementation).