Available at
http://pvamu.edu/aam
Appl. Appl. Math.
ISSN: 1932-9466

Applications and Applied Mathematics:
An International Journal
(AAM)

Vol. 8, Issue 2 (December 2013), pp. 366 – 390

Application of Fractional Moments for Comparing Random Variables with Varying Probability Distributions

Munther R. Al Shami and A. R. Mugdadi
Department of Pharmaceutical Technology
Jordan University of Science and Technology
Irbid, Jordan
mralshami@just.edu.jo; aamugdadi@just.edu.jo

R. R. Nigmatullin and S. I. Osokin
Theoretical Physics Department
Kazan State University
Kazan, Tatarstan, Russian Federation
nigmat@knet.ru; osokin@hitv.ru

Received: January 14, 2013; Accepted: May 30, 2013

Abstract

New methods are being presented for statistical treatment of different random variables with unknown probability distributions. These include analysis based on the probability circles, probability ellipses, generalized mean values, generalized Pearson correlation coefficient and the beta-function analysis. Unlike other conventional statistical procedures, the main distinctive feature of these new methods is that no assumptions are made about the nature of the probability distribution of the random series being evaluated. Furthermore, the suggested procedures do not introduce uncontrollable errors during their application. The effectiveness of these methods is demonstrated on simulated data with extended and reduced sample sizes having different probability distributions.

Keywords: Normality assumptions, logarithmic transformation, fractional moments, generalized Pearson correlation coefficient, probability ellipses, probability distribution mapping

AMS-MSC 2010 No.: 62H20, 62P35

1. Introduction

It is common place knowledge that the probability distribution (PD) of random variables may acquire any of a dozen of well-defined patterns**.** These include, among others, uniform, normal, log-normal, beta, gamma, logistic, Weibull and chi-square. Notwithstanding, the identification of these distributions has not been systematically attempted before. Statistical methods dealing with the PD of real data are focused only upon determining the deviation of its PD from normality. The work of some scholars like Shapiro and Wilk (1965) was basically geared towards the estimation of such deviations without any concern, about the particular distribution causing them. Kolmogorov-Smirnov (KS) as discussed in Massey (1951) and Sheskin (2007) introduced what was claimed to be distribution-free procedures. Such claims were criticized by others, for example D'Agostino (1986) with a very strong statement: "The Kolmogorov-Smirnov test is only a historical curiosity. It should never be used". Moreover, McCulloch (1987) suggested a distribution free procedure based on Spearman's rank correlation coefficient to test the equality of the variance for data sets under consideration. It has been suggested that the procedure be applied when the normality assumptions of the random variable are serious violated. This work attempts to depict the actual PD of random variables for data sets generated with predefined distribution. The main rationale for such endeavor is to demonstrate that the suggested procedures in this work are insensitive, within boundaries, to the PD of data being evaluated.

Many conventional statistical procedures are based on the assumption that the distribution of random variables is normal. Assumptions on the probability distribution also apply to bioequivalence (BE) data which is believed to be log-normally distributed and are invariably skewed to the right. Although, such assumptions could hardly lend themselves for scientifically sound justification, they have indeed provided the basis for logarithmic transformation of such data prior to its statistical evaluation. The impact of logarithmic transformation on the restoration of normality for different sets of data is subject to critical examination in the present work. It is unlikely that normality assumptions are satisfied in the presence of extreme varying responses or outlying observations. The impact of such observations in BE data as well as in data pertaining to many complex systems would be very difficult to predict. Also, the conventional statistical procedures can't feel the changes in the noise distributions (parameters and even forms of distribution). Their effectiveness would be doubtful in cases where the sample size (SZ) of data considered is small. This is especially applicable to biopharmaceutics data, where the maximum SZ hardly exceeds 100 data points. Taking into consideration the above mentioned arguments, it becomes necessary to develop alternative methods for the statistical treatment and recognition of such data without any prior assumptions.

2. Method (Data Generation)

For the purpose of the present work, data sets having five different PDs were generated on a MathCad-14 platform. These are defined as extended SZ with 10^5 data points and reduced data sets with SZ of 50 data points. Typically, hypothetical reference data sets with normal distribution will be generated and contrasted with the test data sets having normal, lognormal,

uniform, logistic, and gamma distributions. All data sets will have the same mean and standard deviation values which will be set at 10 and 1 respectively. A double normalization step has been undertaken to ensure that all initial samplings have a mean value of 10 and Standard deviation of 1 and also to minimize the influence of the differences in mean and standard deviation values in the consequent analysis. This step has been done in accordance with the following algorithm:

$$New\,sampling = \frac{Old\ sampling - MeanValue}{StDev} + 10\,. \tag{19}$$

Typical extended data sets generated in accordance with this simulation procedure are presented in Figures 1 & 2. All data sets were generated in duplicates with the first designated as Reference data and the second as the Test data.

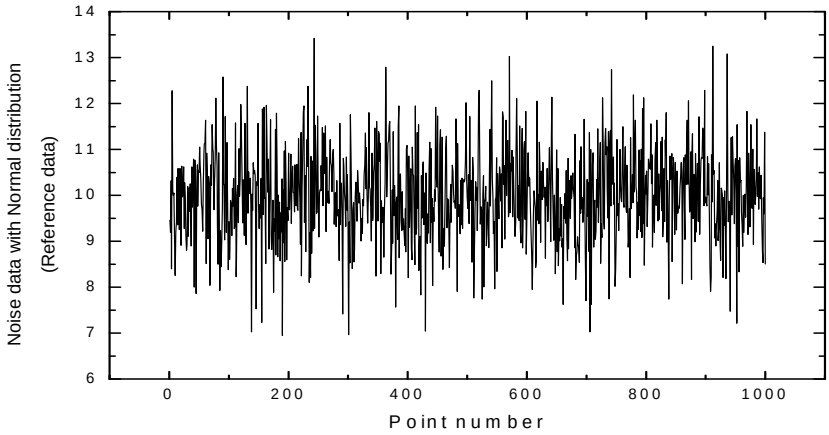

Figure 1. The first set of extended sample size random series with normal distribution and is designated as the reference data

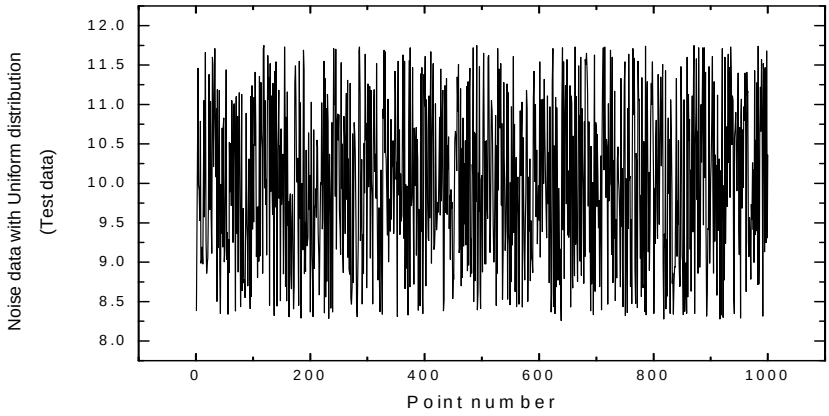

Figure 2. The second set of extended sample size random series with uniform distribution and is designated as test data

3. Theoretical Considerations

3.1. Probability Circles (PC) Analysis

PCs analysis for two sets of random series (y_{1i} and y_{2i}) could be established in accordance with the following expressions

$$\begin{aligned} X_1(\varphi) &= \Delta_1 + R_1 \cos(\varphi), \\ Y_1(\varphi) &= R_1 \sin(\varphi), \end{aligned} \tag{1}$$

where

$$\begin{aligned} &\Delta_j = \frac{1}{N}\sum_{i=1}^{N} y_{ji} - mean\,value, \\ &(M_2)_j = \frac{1}{N}\sum_{i=1}^{N} (y_{ji} - \Delta_j)^2 - standard\ deviation, \\ &R_j = \sqrt{(M_2)_j}\ and\ \varphi \subset [0, 2\pi] \end{aligned} \tag{2}$$

with j = 1. This is a parametric form of a circle with the center at $r_1(\Delta_1, 0)$ located on the horizontal axis with radius $R_1 = \sqrt{(M_2)_1}$, which is defined by the value of the standard deviation. The second stage is the construction of the comparable PCs for test data. The comparable PCs for the second (test) data set y_{2i} with the given sampling volume N (for this set $N = 10^5$) is presented by expressions (3):

$$\begin{aligned} X_2(\varphi) &= \Delta_2 \sin(\psi) + R_2 \cos(\varphi), \\ Y_2(\varphi) &= \Delta_2 \cos(\psi) + R_2 \sin(\varphi), \end{aligned} \tag{3}$$

where $\Delta_2, (M_2)_2\ and\ R_2$ are determined by equation (2) with j = 2. This is the parametric form of a circle with the center rotated relatively to the basic circle by an angle ψ. The magnitude of the angle ψ is defined by expression (5):

$$\cos(\psi) = \frac{mean(y_1 y_2) - \Delta_1 \Delta_2}{R_1 R_2}, \tag{5}$$

where y_1 and y_2 determine the corresponding vectors of the dimension N. PCs obtained in accordance with the above description for data representing test and reference treatments are shown on Figure 3a. Here one can see that these two circles located quite close to each other (circles overlap). If these circles are located inside of each other or crossed with each other than two comparable data sets cannot be differentiated. It means that they have approximately the same mean values and standard deviation values and in addition these two sets of data correlate to each other. When two circles do not overlap, these two data sets can be differentiated from each other in the frame of the PCs analysis. This means that they have different mean values and standard deviation values and do not correlate with one another.

The main evident shortcoming for the PCs is that they can be best constructed for a pair of data sets with different mean values. This implies that it is not possible to simultaneously compare more than two sets of data whose mean values are more or less the same. In the latter case, the circles will overlap. The second shortcoming is the low sensitivity of this approach since it cannot distinguish two data sets with different PDs. This is clearly demonstrated in Figure 3b where all PCs for data sets with the same mean and standard deviation were completely overlapped. Notwithstanding, the PCs can quantitatively indicate the magnitude of overlaps of data sets, with different mean values. As shown in Figures 3a and 3b, the overlaps between the circles of such sets could serve as a rough indicator of their proximity.

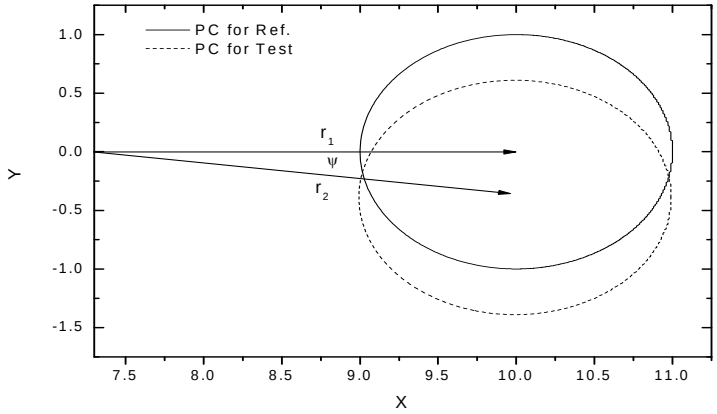

Figure 3a. PCs for reference and test data (see Figure 1 and Figure 2)
Module of the radius ($r_{1,2}$) of the circle center shows the mean value
Radius of the circle shows the standard deviation value
And angle between radii to the circle centers (Ψ) shows the correlation between two noise data (random series)

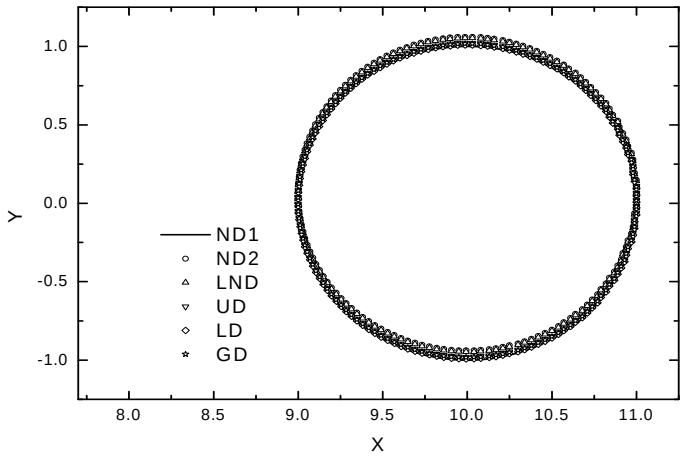

Figure 3b. PCs for different sets of data with different distributions (Normal, Log-Normal, Uniform, Logistics and Gamma)

3.2. Probability Ellipses (PEs) Analysis

To overcome the above mentioned shortcomings of the PCs analysis, an additional parameter for each circle must be added so that it is transformed into an ellipse. This will increase the sensitivity of the approach and uncouple one set of data from another and will allow the simultaneous presentation of more than two sets of data. This has been accomplished by the incorporation of the slopes of the ellipse (angle between 0X axis and major semi-axis). To obtain necessary parameters for ellipse one has to separate the present random series set on two data subsets. The results obtained by such procedure are shown on Figure 4. The first subset is the difference between the point values located above the general mean value and the general mean value itself. The second subset is the difference between the general mean value and the points values located below the general mean value. For both subsets one can calculate the mean values and standard deviation values to obtain four necessary parameters. The fifth parameter is related to the Pearson's correlation coefficient between two subsets.

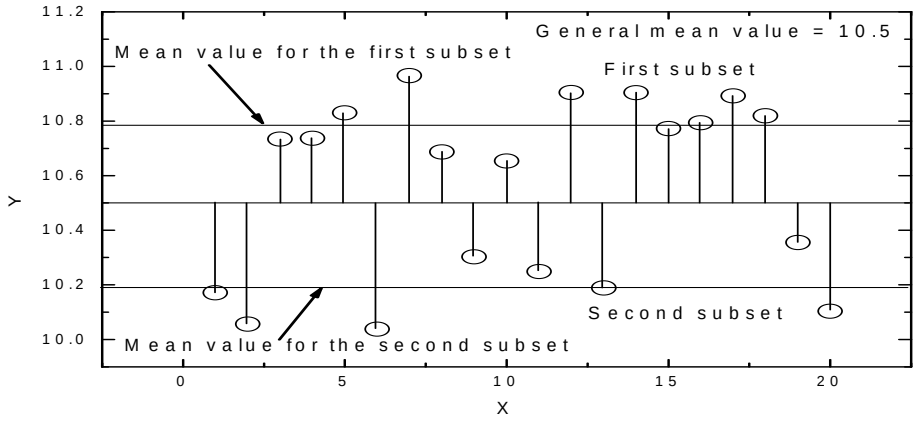

Figure 4. Presentation of the separation procedure.
One initial set of data is separated on two subsets.
First subset represents the difference between points values located above the general mean value and the general mean value itself.
Second subset represents the difference between the general mean value and the points values located below the general mean value

The PEs for the present data set y_i with the given sampling, having an N sample size, is given by expressions (6):

$$X(\varphi) = \Delta_1 + R_1 \cos(\varphi + \psi),$$
$$Y(\varphi) = \Delta_2 + R_2 \sin(\varphi), \tag{6}$$

where

$$\Delta_{1,2} = \frac{1}{N}\sum_{i=1}^{N} y_{1,2i} - mean\, values\ for\ first\ and\ second\ subsets$$
$$(M_2)_{1,2} = \frac{1}{N}\sum_{i=1}^{N} (y_{1,2i} - \Delta_{1,2})^2 - standard\ deviation\, values\ for\ first\ and\ second\ subsets \quad (7)$$
$$R_{1,2} = \sqrt{(M_2)_{1,2}}, \varphi \subset [0, 2\pi].$$

The magnitude of the angle ψ is defined by the next expression (8):

$$\cos(\psi) = \frac{mean(y_1 y_2) - \Delta_1 \Delta_2}{R_1 R_2}, \quad (8)$$

where y_1 and y_2 determine the corresponding vectors (first and second subsets) belonging to the given dimension *N*.

PEs are shown in Figure 5 with different centers and semi-axes for data representing the reference and test treatment which were represented in Figure 1 and Figure 2. In contrast to the PCs approach, this approach has the ability to differentiate the set of random series. In addition, it appears to be more sensitive in comparison to the previously developed rough PCs approach. The main common and evident shortcoming of the PCs and PEs approaches is that they can only give visual depiction of the statistical closeness or proximity of the random series being evaluated. But in most applications, a quantitative estimation of such proximity is deemed necessary.

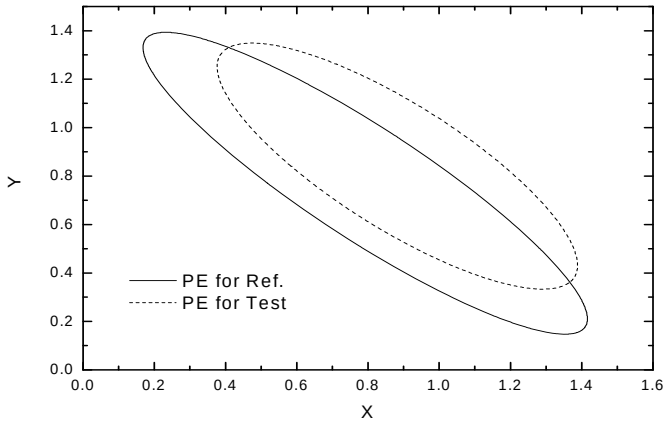

Figure 5. PEs for reference and test data (see Figure 1 and Figure 2)

3.3. General Mean Value (GMV)

In order to resolve the shortcomings associated with the PCs and PEs, a new approach based on the generalized mean value (GMV) function is introduced. This approach was initially suggested

by Nigmatulin (2006) and is based on the generalization of the arithmetic mean value and takes into consideration the total set of moments including the fractional and even complex values. The integer (or fractional) moment of the p^{th} order is determined as:

$$\Delta_p \equiv \Delta(m_p) = \frac{1}{N}\sum_{j=1}^{N}\left(\tilde{y}_j\right)^{m_p},\ 0 \le m_p \le Mx\ . \tag{9}$$

The generalized mean value (GMV)-function is related to the moment of the p^{th} order by the relationship.

$$GMV_N(m_p) = \left[\Delta\left(m_p\right)\right]^{1/m_p} \tag{10}$$

This function includes the harmonic mean (m_p = -1), geometric mean (at $m_p \to 0$), arithmetic mean (m_p = 1) as partial cases. The GMV-function is an increasing (monotonous) function that as $m_p \to \infty$ recovers the right limit of the sequence ($\tilde{y}_{\max}$ = *ymx*) and as $m_p \to -\infty$ tends to another limiting value $\tilde{y}_{\min}$ = *ymn*. These properties can be mathematically expressed by the following expressions

$$\begin{aligned} &GMV_N(m_1) > GMV_N(m_2),\ \ if\ \ m_1 > m_2, \\ &\lim_{m\to\pm\infty} GMV_N(m) = \begin{pmatrix} ymx \\ ymn \end{pmatrix}. \end{aligned} \tag{11}$$

This function has the remarkable ability, (being presented in the plot $G2_N(m) = G1_N(m)$) to show the *statistical proximity* of two random sequences that satisfy a linear relationship of the type

$$GMV2_N(m) = \lambda GMV1_N(m) + b\,. \tag{12}$$

The linear relationship (12) may be chosen as *a quantitative criterion* for the verification of the statistical proximity for two arbitrary random sequences. However, for detection of the statistical proximity it is necessary to calculate the expressions (10) for the two sequences and plot them with respect to each other for approximate verification of the expression (12). Another informative picture can be obtained from the plot of function $G2_N(m)/G1_N(m)$ against *m*. The GMV for a random series with reduced SZ (exemplifying test and reference data) are shown in Figure 6.

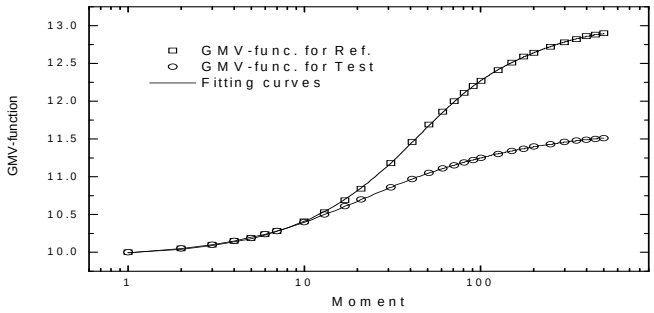

Figure 6. GMV functions for reference and test data with reduced sample size
Solid lines represent the fitting curves obtained by the use of the equation (13), where S=3
The values of the parameters are shown in Table 2

It could be seen that these GMV functions are very sensitive to the type of the PD considered and could be useful to distinguish one distribution from the other. The GMV curve itself may be fitted by the use of the expression of the type:

$$y(p) = a_0 + \sum_{i=1}^{S} a_i \exp(\lambda_i p)\,, \tag{13}$$

where $S \subset [1,\infty]$, see Nigmatulin (2006). Application of this function shows that for the most cases it is sufficient to take S=3. So, using the GMV approach and function (13) one can "read" quantitatively any distribution in terms of parameters a_i and λ_i from expression (13). The fitting curves obtained by the use of the equation (13), where S=3, for the reduced SZ data, are presented in Figure 6. The values of the parameters are shown in the caption to Figure 6. Another informative plot for the ratio of the GMV functions for the ratio of test to reference data is presented in Figure 7, together with the respective upper (120%) and the lower (80%) limits. This implies that if the curve is not contained within these limits, the data sets being contrasted will differ from each other more than on ±20% in terms of the GMV. The same picture has been reproduced for a data set with the same mean and standard deviation (10 ± 1) and is represented by the solid line in Figure 8. The dotted line in the same figure was obtained by setting the standard deviation of the same mean (10) to 3. It could be seen that this lines went out of the upper limit at moments higher than 6.

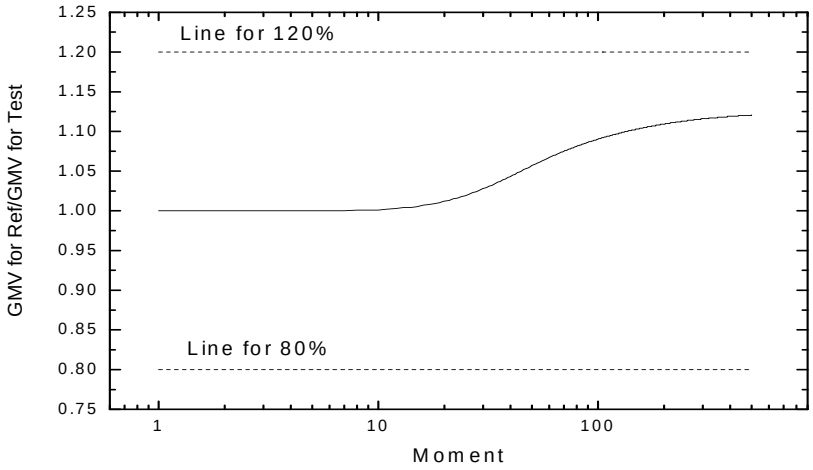

Figure 7. GMV functions for ratios of tests data with UD to reference with ND the same mean values of and the same standard deviations (see Figure 1 and 2) Dashed lines for 120% and for 80% represent the ±20% criterion

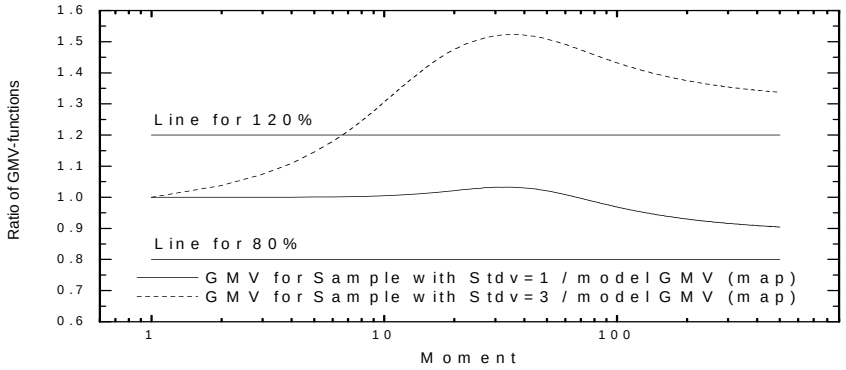

Figure 8. GMV functions for ratio of test data set with UD to reference data set with ND all having the same mean values of 10 and different standard deviations (1 for the reference and 3 for the test)
Dashed lines for 120% and for 80% represent the ±20% criterion

3.4. Generalized Pearson Correlation Coefficient (GPCC)

By analogy to the generalization of the conventional mean value it is possible to generalize the Pearson correlation coefficient (PCC). This generalization was introduced by Nigmatullin, Arbuzov, Nelson (2006). In common statistical analysis the Pearson correlation coefficient is calculated for the first moment by the use of the expression

$$PCC = \frac{\sum_{i=1}^{N} \left| F_i Q_i \right|}{\left(\sum_{i=1}^{N} \left| F_i \right|^2 \right)^{\frac{1}{2}} \left(\sum_{i=1}^{N} \left| Q_i \right|^2 \right)^{\frac{1}{2}}}, \tag{14}$$

where F_i and Q_i represent respectively a couple of the random sequences compared. When PCC = 1 we have the ideal correlation for any two sequences. The deviations from the ideal case could serve as a *quantitative* measure of the statistical proximity of the sequences.

The generalized expression of the PCC (GPCC) may be expressed as

$$GPCC(F,Q)=\frac{\left(\sum_{i=1}^{N}\left|F_iQ_i\right|^m\right)^{\frac{1}{m}}}{\left(\sum_{i=1}^{N}\left|F_i\right|^{2m}\right)^{\frac{1}{2m}}\left(\sum_{i=1}^{N}\left|Q_i\right|^{2m}\right)^{\frac{1}{2m}}}. \tag{15}$$

One can write down the following identities

$$GPCC_m(Q,Q)=1, \qquad \text{for all } m\geq 0, \tag{16}$$

$$GPCC_0(F,Q)=1, \qquad \text{for } m=0 \text{ and different sequences } F \text{ and } Q. \tag{17}$$

So, the function $GPCC_m(F,Q)$ at certain values of $m > 0$ can have a *minimal* value and would tend to a limit value of *true* correlations at $m >> 1$. It is self-evident that expression (16) is *correct* only in cases where the random sequences analyzed are strictly positive and do *not* have any zero values (F_i, Q_i) > 0. The advantage of this analysis in comparison to conventional analysis, based only on the single value of the Pearson coefficient (14), is obvious. The Pearson correlation coefficient that follows from (15) at $m = 1$ and coinciding with expression (14) gives only one correlation point in the space of the given moments. We think that this point is chosen for *traditional* reasons. There are no other arguments showing the advantage of this choice ($m = 1$) in comparison with other values of m (as far as we know) in the literature devoted to statistical analysis, see Nigmatullin (2010).

Instead of considering this "historical" point and constructing the correlation analysis based on selection of one point ($m = 1$) only, the GPCC-function (15), is applicable to *all* available values of the real moments located in the interval ($0 < m < \infty$). This correlation function will effectively define the correlation band as a region, thus permitting a more definite statement on the correlation that may exist between arbitrary random sequences. It is interesting to note that at $F_i = \lambda Q_i$ expression (15) is reduced to the value one. At $m = 1$ it reproduces the conventional definition of the PCC for *positive* sequences. So, based on new statistics of the fractional moments there is a possibility to develop the '*noninvasive*' (when any calculated error can be evaluated and controlled) and accurate *quantitative* methods, which can be used for more sensitive recognition of random sequences of *any* nature. The GPCC for test and reference data presented in Figure 1 & 2 are provided in Fig 9. It is evident that the GPCC is more sensitive and informative than the conventional PCC in the recognition of possible differences of PD of different sets of data, especially at higher moments. The conventional PCC can only account for a single value of the moment equaling unity.

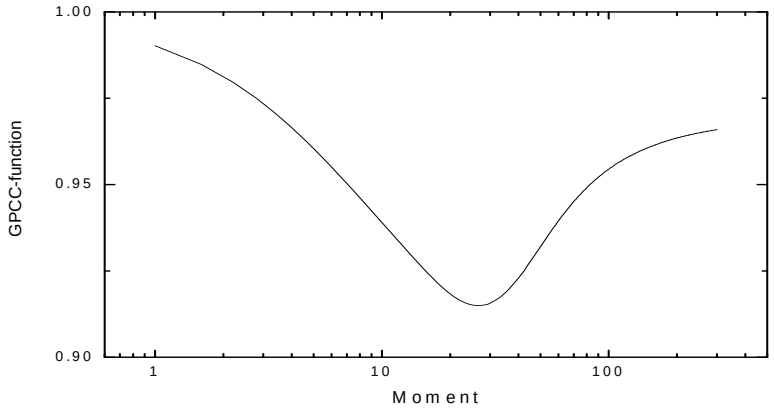

Figure 9. GPCC function for a pair of the random sequences represented on Figure 1 & Figure 2 (reference and test data sets).

3.5. Beta Function (BF) Analysis

This represents a completely new and purely mathematical approach for random series (noise data) treatment. It is an approach that was proposed by one of the authors of this paper (RRN). The effectiveness of this approach was demonstrated clearly in Nigmatulin (2008) and Massey (1951). It is based on transforming a set of random series to the curve that resembles the beta-function and can be fitted by this function. The parameters of the beta function can serve as quantitative parameters of the random series considered. This is achieved by forming a sequence of the ranged amplitudes (SRA) for the random series by the sorting procedure. Such a procedure helps to locate the positive and negative values of the random series in a decreasing order. If the series is integrated by the standard trapezoid method, then we obtain a bell-like curve located in the interval $[x_1, x_N]$. Based on the results of previous works Nigmatullin (2008) and Massey (1951), it can be proved that this integrated curve approximately satisfies the function

$$Jf(x) = b\left(x - x_1\right)^{\alpha}\left(x_N - x\right)^{\beta}. \tag{18}$$

The values of the fitting parameters b, α, β can be used as a specific quantitative label for identification of the random series considered. These are easily estimated by the eigen-coordinates (ECs) method. The details of this transformation is fully described elsewhere, see Nigmatullin (1998 and 2000), Abdul-Gader and Nigmatullin, (2001), and Al-Hasan and Nigmatullin (2003). The results could be seen in Figure 10 (bell-like curves) for reference and test data sets earlier presented in Figure 1 and 2. These curves were fitted by the use of the equation (18). The values of the parameters are shown in the caption to the Figure 10. Here one can see that this approach is very sensitive in the quantitative reading of distributions that correspond to the random series analyzed.

4. Data Treatment

The treatment will be applied to all data set in accordance with the above mentioned approached. The results of this treatment are presented as PCs and PEs in Figures 10 and 11 respectively.

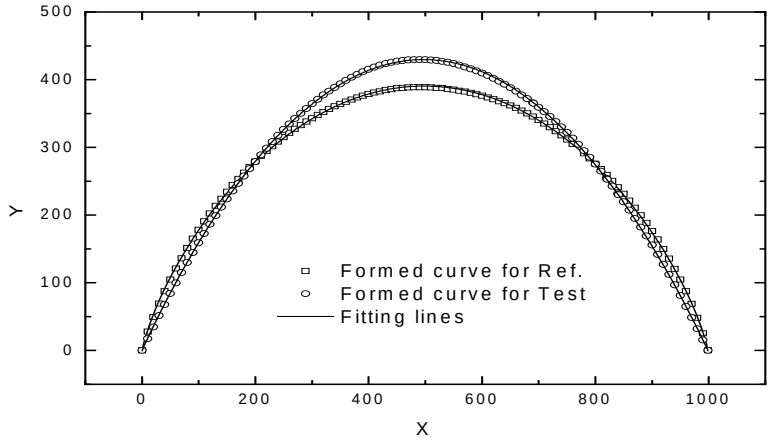

Figure 10. The bell-like curves for reference and test data sets (see Figure 1 and Figure 2)
Solid lines represent fitting curves obtained by the use of the equation (18)
The values of the parameters are shown in Table 2

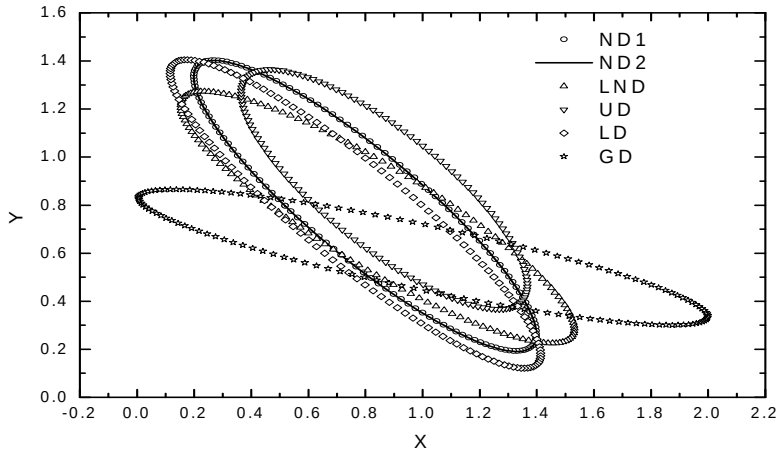

Figure 11. PEs for different sets of data with different distributions (Normal, Log-Normal, Uniform, Logistics and Gamma).

Contrary to the PEs, the PCs are lacking in its ability to discriminate between all the data sets being treated and have approximately the same positions of the center and radii. Such lack of sensitivity is not demonstrated with the PEs treatment. An interesting feature is shown in Figure 11 where the ellipses for two normally distributed data sets are completely overlapped. The PEs

are obtained upon contrasting the normally distributed data set with other sets with different distributions have different positions of the center, slopes and dimensions of the axes. This can serve as a pictorial tool for the evaluation of the distribution for any set of random series because these ellipses relating to different distributions possess striking stability in terms of center positions, slopes and dimensions.

The GMV functions for different sets of model data with different distributions are presented in Figure 12. It is evident that GMV functions for two sets of data (open circles and squares) with the same normal distribution are completely overlapped. The GMV functions for other distributions have distinctly different forms. These curves have been fitted by the use of the equation (13) with the value of the parameter *S set at* 3. The fitting values of the different parameters of this equation are provided in Table 2. These GMV functions can serve as quantitative as well as a pictorial tool (GMV-mapping) for evaluation of the distribution for any set of random series because these curves possess a striking stability in terms of parameters of equation (13). An important feature is noticed in Figure 12 where the first six moments can distinguish between any two distributions. Conventional statistical methods are usually concerned with the first four moments. So this figure gives us some evidence that conventional methods that are based only on the first four or five moments cannot distinguish different distributions.

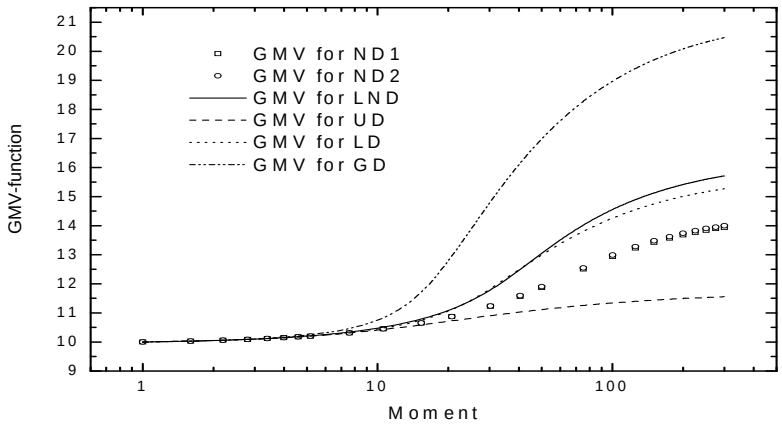

Figure 12. GMV functions for different sets of data with different distributions (Normal, Log-Normal, Uniform, Logistics and Gamma)
These curves were fitted by the use of the equation (13) with *S*=3
The values of the parameters are shown in Table 2

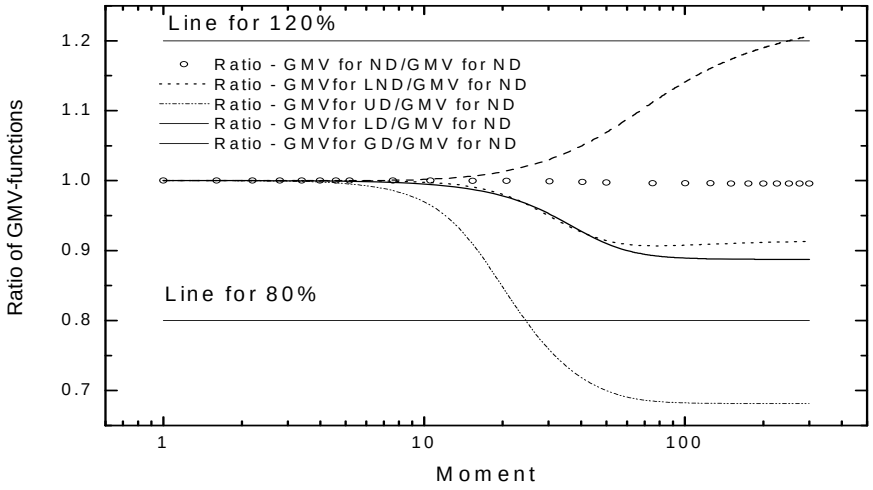

Figure 13. GMV functions for ratios of test data sets with different distributions (Normal, Log-Normal, Uniform, Logistics and Gamma) to reference sets of data with normal distribution
Solid lines for 120% and for 80% represent the ±20% criterion

The GMV for the ratios of pairs of data sets simulating test/reference ratios are presented in Figure 13. These ratios have been constructed by contrasting all five distributions with normally distributed set of data. Upper and lower solid lines represent the 120% and the 80% limits which reflect the ± 20% conventional criterion. This implies that if a curve goes outside of this ± 20% region then the two data sets being contrasted will differ from one another by ± 20% in terms of the GMV. This picture can also serve as a tool to evaluate the statistical closeness of any data set to the normal distribution.

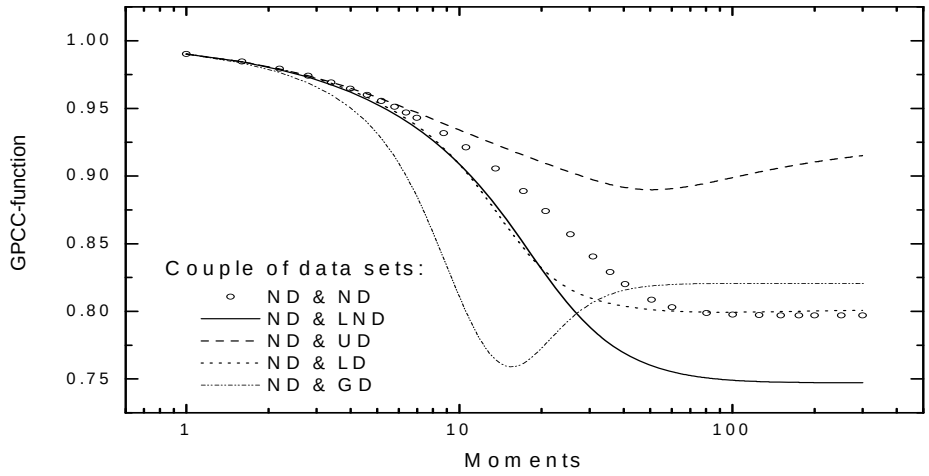

Figure 14. GPCC functions for different sets of data with different distributions (Normal, Log-Normal, Uniform, Logistics and Gamma)

The GPCC functions obtained for five sets of random series with different distributions in relation to one reference set of random series with normal distribution are shown in Figure 14. This picture can serve only as qualitative pictorial tool to compare the distribution of any data set with normal distribution of the model set. The curves associated with the different distributions demonstrate a good stability in terms of their forms. In this case we have to say that this tool works only on long (extended) samplings because here we have to use some real sampling with modeled sampling having normal distribution. Also, a plot for the GPCC obtained by contrasting the normal distribution with all other PDs after log-transformation is provided in Figure 15. The bell-like curves for the different sets of data with different distributions are presented in Figure 16. These curves were fitted by the use of the equation (18). The values of the fitting parameters are shown in Table 3. It must be emphasized that this method works only for data with extended data SZ.

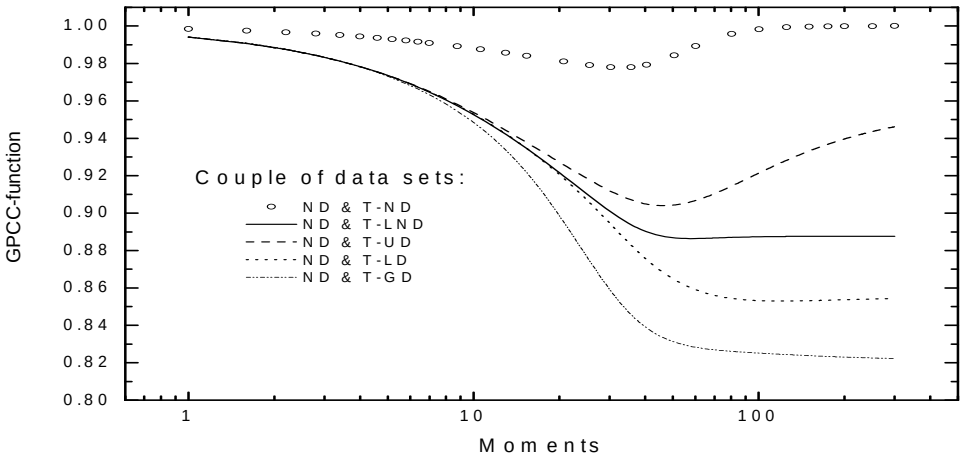

Figure 15. GPCC functions for different contrast for sets of reduced sample size data with different distributions consequent to log-transformation.

5. Method Robustness

The robustness or reproducibility of the above mentioned methods could be interpreted from two different perspectives. Namely, the ability of these methods to reproduce the same outcome in every occasion where the specific data sets, with distinct distribution, has been generated. Likewise, it is imperative to demonstrate that such methods will perform in a consistent manner for replicate data sets irrespective of its size. The validity of data generation is ascertained by the data present in Table 1. On the one hand, it may be seen that the point estimator or the percentage ratio of test to reference stands at exactly 100, as seen for the linear scale. On the other hand, the variance term (S^2) for the linear scale in SW results, is identical for all data sets. This represents a clear indication of the validity of the data generation procedure adopted for this work. It should also be noted in this context that logarithmic transformation has resulted in the distortion of these expectedly consistent patterns.

5.1. Extended sample size ($N = 10^5$)

Further evidence on the robustness of these new methods is demonstrated by using 10 different random series generated with normal and log-normal distributions. The plots for the PEs, GMV-functions and bell-like integrated beta curves were constructed and shown in Figs. 16, 17 & 18 respectively. It is evident that the pictorial patterns for 10 sets of data are effectively identical. Only slight difference, in the GMV are seen at moments higher than 20. It is also noted that the GMV functions for all samplings with normal distribution are distinctly distinguishable from the samplings having the Log-normal distribution. These results clearly suggest that the three methods are very stable and reproducible which qualifies them for the recognition of the different PDs.

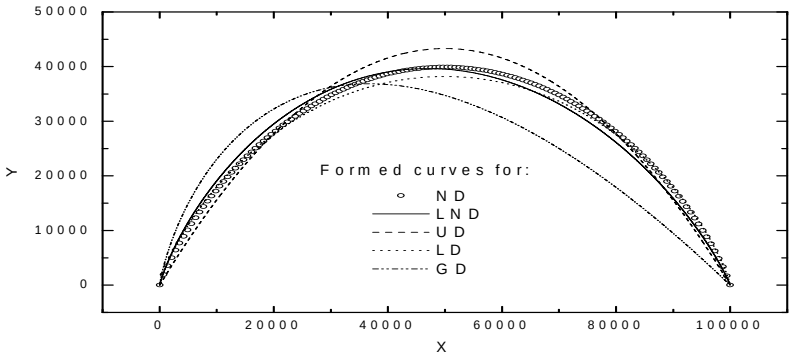

Figure 16. Formed bell-like curves for different sets of data with different distributions (Normal, Log-Normal, Uniform, Logistics and Gamma) These curves were fitted by the use of the equation (18) The values of the parameters are shown in Table 2

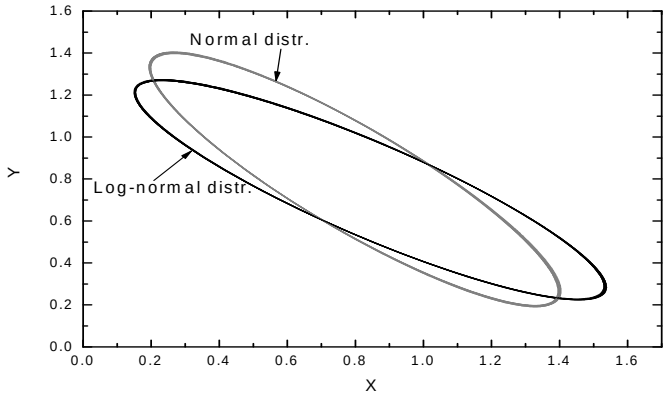

Figure 17. PEs for two sets of data with Normal and Log-normal distributions Each set has ten samplings of random series

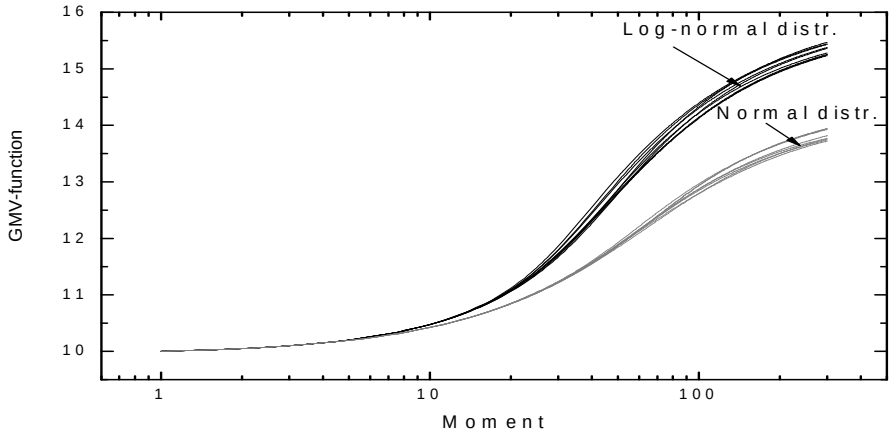

Figure 18. GMV functions for two sets of data with Normal and Log-normal distributions
Each set has ten samplings of random series

5.2. Reduced sample size (N = 50)

Five reduced SZ series (50 points each) with normal, log-normal, uniform, gamma and logistic distributions have been generated as explained above. Their respective means and standard deviation values are set at 10 and 1. The PD maps were constructed for the data set in a similar manner to those with extended SZ. The standard deviation for the reduced SZ data that submitted to the conventional analysis was set at 3 since the variability in real BE data, expressed in %CV, is invariably much higher that 10%. The PEs maps shown in Figure 19 have been stable for normal, log-normal, uniform and gamma distribution. This has also been the case for the random series with reduced SZ for the same distributions. Here one can see that this map could be useful to qualitatively distinguish these four distributions and evaluate the closeness to other random series of different distributions. The ellipse for logistic distribution is not provided in Figure 19 because its closeness to the ellipse of the log-normal distribution. The Ellipses mapping cannot help to distinguish the log-normal and logistic distributions. The Plots for the GMV for five different distributions (Normal, Log-normal, Uniform, Logistic & Gamma) with extended and reduced SZs are presented in Figure 20. It is obvious that the GMV lines obtained for reduced SZ samplings are quite close to the corresponding GMV lines for the corresponding extended SZ data sets in intermediate region from moments 7 to 50. This represents a clear indication that in such a regoin, the GMV map has the capability to distinctly recognize the different distributions for the reduced SZ series. Also, this picture shows that it is impossible to distinguish the Log-normal and Logistic distributions in the same region. This is because their GMV lines separated at higher moments (m>80). It may be postulated that this mapping approach could be useful to qualitatively distinguish four distributions and to evaluate their closeness or proximity to other random series without prior knowledge about their PD.

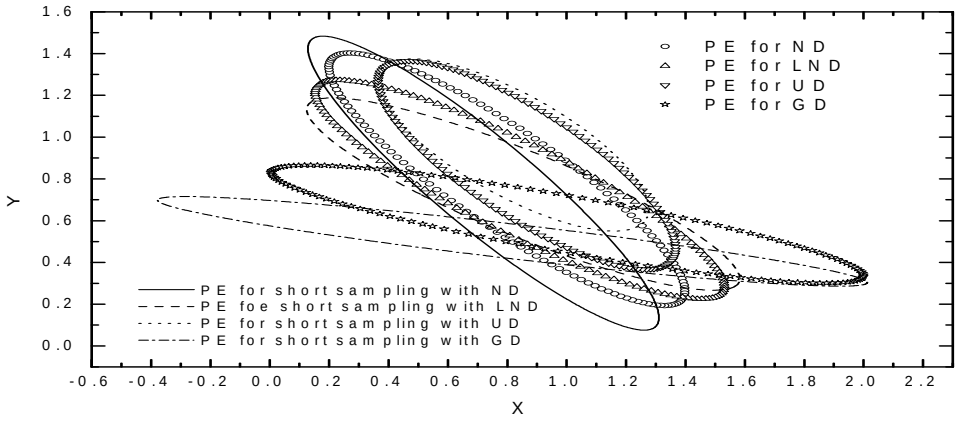

Figure 19. Ellipses map for four random series with ND, LND, UD and GD distribution
Points represent four stable ellipses for the extended sample size series and lines represent PEs calculated for the reduced sample size.

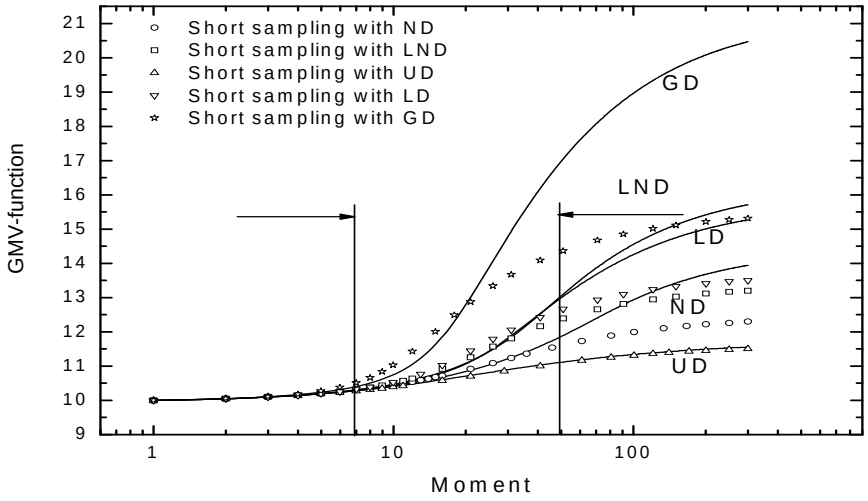

Figure 20. GMV map (curves) for five reduced sample size series with ND, LND, UD, LD and GD distributions
Solid lines represent five stable GMV map for extended and reduced sample size data sets
Arrows show the region for determination of the distribution closeness

6. Conventional Statistical Evaluation

The theory of statistical moments (Yamoaka and Nakagawa (1978), and Mayer and Brazzell (1988)) has been gaining significant grounds in the PK evaluation of bioequivalence data over the past three decades. Based on this theory, non-compartmental (NC) PK procedures have

become mandatory by many regulatory authorities like the American Food and Drug Administration (FDA), the International Conference of Harmonization (ICH). These procedures represented a drastic shift from the model-dependent PK approach for the assessment of plasma concentration-time data to a purely model-independent statistical approach considering that such raw data could be seen as the first (zero-moment) of other different moments. The other moments include the variance, skewness and kurtosis. Consequent to the PK evaluation of data, statistical inferences have to be made of the geometric ratio of the ratios of, or the differences between, the main PK metrics pertaining to the test and reference treatment.

The construction of the 90% confidence interval (CI) about the ratios of the geometric means of test to reference products (T/R) represents a worldwide procedure used to assess the statistical proximity or the so called bioequivalence of the two products. This assessment is usually conducted on the logarithmic ally transformed PK metrics of interest. Since it is widely believed that such data is generally skewed to the left, logarithmic transformation is undertaken so that the distribution of the metrics will be, invariably restored to normal. The authors of the present work are set to demonstrate that normality assumptions of the transformed data have no scientific ground. In order to achieve this objective, the 90% CI for data sets with different PDs, exemplifying the PK random variables, was estimated on the original (linear) and the logarithmic Scales. The data sets with different PDs and reduced SZ were used to conduct this statistical assessment. The results for the upper and lower limits of the CI as well as the point estimator for T/R are resented in Table 1. Also, the impact of logarithmic transformation on the SW normality indicator is presented in the same table. In addition, attempts were made to pictorially represent different distribution for the reduced sampling to establish the deviation of the typical Gaussian picture. Normalized probability plots are presented Figure 22 for both the linear and logarithmic scale.

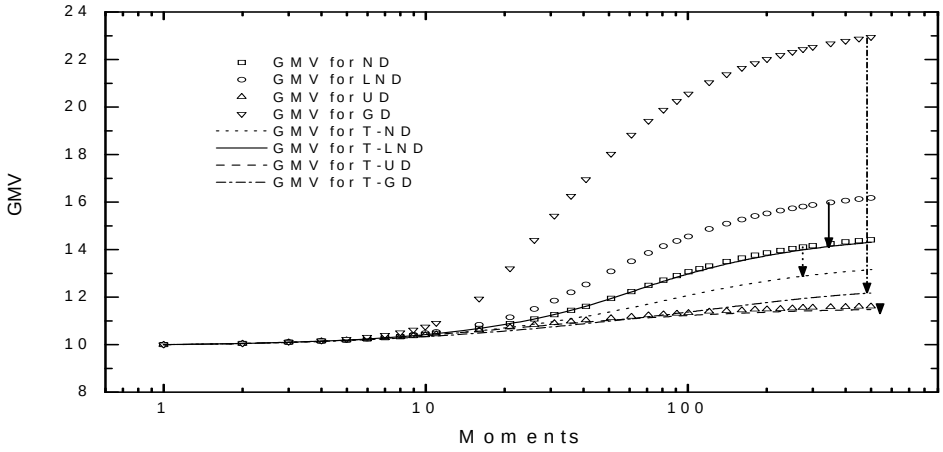

Figure 21. The impact of logarithmic transformation on GMV map for different distributions presented in Figure 12 with arrows showing the shift in the plot consequent to log-transformation

7. Discussion and Conclusions

As stated earlier, one of the prime concerns in this work is to demonstrate the shortcomings of some of the existing statistical methods used to contrast two random series with the purpose of establishing the equivalence or statistical proximity of such series. Officially accepted procedures like Schuirmann (1987) two one-sided test (TOST) and the classical shortest 90% CI are within the most commonly recommended procedures for the evaluation of BE of two drug products. However, their validity is strictly determined by the underlying assumptions inherent in the cross-over models adopted for such analysis.

Some of these assumptions are related to the homogeneity of the variance and the normality of PD of the error term associated with the intra-subject residual stemming from the analysis of the variance. So far, such assumptions have not become subject to verification since sponsors of BE studies or researchers are not encouraged by official bodies to test their validity. It is widely accepted that the skewness in such data may be rectified by logarithmic transformation. The recommendation to log-transform data implies that normal distribution is doubtful.

In addition, there is no scientific evidence to suggest that normality could be restored after such a crude and invasive procedure. This could be substantiated by considering the impact of the transformation step on SW normality indicator or statistic presented in Table 1. It could be seen that the gamma distribution is the only one that showed improvement according to SW normality statistic. Interestingly, this statistic became lower upon the transformation of data with normal distribution. Evidence on the uncertain outcome of logarithmic transformation is presented in Figures 21 and 22 where different distributions are pictorially presented as probability maps and normalized probability plots respectively. Logarithmic transformation has only caused the log-normally distributed data set to approximate normal distribution.

By contrast, logarithmic transformation has distorted some other distributions without bringing it close to the normal distribution. The distribution that has not been affected by logarithmic transformation is the uniform distribution. Also, it could be noticed in Figure 22 where the normalized probability plots are presented, that apart from the gamma distribution, such plots could hardly distinguish between different data sets. Intriguingly, the transformation of the normal data set has resulted in distinct skewness which coincides with the impact of such transformation on SW normality indicator mentioned above. This signifies the inadequacy of such graphic presentation to the probability plots.

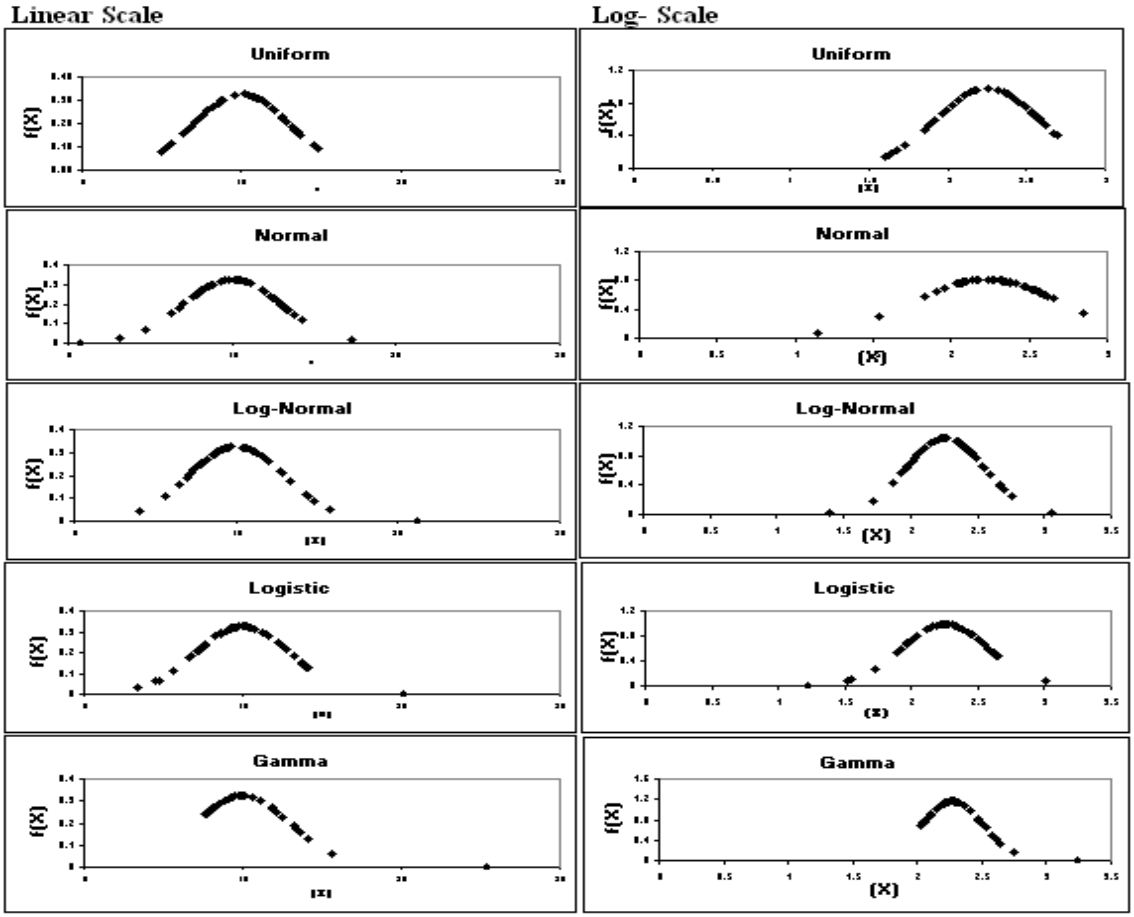

Figure 22. Normalized probability plots for different distribution constructed for both the linear and logarithmic scale

On the other hand, the impact of logarithmic transformation on the width of 90% confidence interval (WCI) is a very significant matter which must be carefully considered. Typically, the BE will be concluded if such interval is contained within the predefined limits set at 80 – 125 for the logarithmic scale. The effect of logarithmic transformation as shown in Table 1 clearly demonstrates the lack of a consistent pattern or trend with the distribution at hand. Estimates of the WCI for the linear scale have a narrow range between 20.22 and 23.81; whereas the same estimates consequent to logarithmic transformation are given at 16.02 and 34.88. The lack of a consistent pattern is evident since the WCI was increased in some data set and decreased in others. The most significant change was noticed for the data set with normal distribution as it is increased from 23.81 to 34.88. The serious implications of this effect are quite obvious since it may lead to concluding bio-in-equivalence of the drug products being evaluated. Another interesting observation could be seen in this table which relates the effect of logarithmic transformation on the point estimator of T/R which ranged from 99.71 to 101.48 consequent upon log-transformation. The impact on the decision-making related to bioequivalence is very obvious. Notably, the point estimator value was increased in the case of normal distribution and decreased for the log-normal set of data after log-transformation. This represents a clear indication of the invasiveness of this procedure and its ability to introduce uncontrollable error in the statistical evaluation of such data.

Another interesting feature of the 90% CI is the fact that it represents a direct reflection of the standard deviation used for data generation. In a previous exercise where the reduced SZ data were used, with a standard deviation of 1 for a mean of 10, the WCI approximated 7.5. In theory, the confidence interval should reflect the actual differences between random series and serves as a measure of their statistical proximity. This argument bears extreme significance within the context of BE data evaluation. The hypothesis may be postulated as follows; when both test and reference products have the same mean value for the pharmacokinetic metric and the same common variance or variability, common sense dictates that they be regarded equivalent. Notwithstanding, it is extremely likely that the test product would be concluded bio-in-equivalent in situations where the variability, expressed as the %CV, exceeded a value of 40.

One of the most obvious advantages of the methods suggested in the present work is its ability to characterize, and to discriminate between, different PDs of the data being evaluated. This would pave the way for more efficient transformation of such data if researches deem it imperative to restore its distribution to normality prior to any statistical evaluation. This represent an issue that is worthy of future consideration. Notwithstanding, the PEs and the GMV maps presented in this work offer a quantitative measure for the closeness of random series irrespective of the SZ of data at hand. Preliminary investigations show that these two methods are very sensitive to the changes in the PD and to the presence of small noise signals on the background of the large random series. Such signals are usually encountered in BE data as outlying observations. It is interesting to note that such responses have not affected the GMV analysis at moments lower than 10 for all PDs. It is beyond the scope of the present work to examine the impact of such observations on the outcome of analysis at higher moments. In addition, the authors are cognizant that the homogeneity of the variance for random series of data remains a very serious matter that must be duly dealt with in future endeavors.

Acknowledgment

The authors would like to express their gratitude to the Russian Ministry of Education & Science for part financial support of this work in the form of the grant "Development of Scientific Potential of Leading Higher Schools of Russian Federation" (grant number: 2.1.1/2474). Appreciation is also extended to the Jordanian Pharmaceutical Manufacturing Co. Ltd. for the continued interest and the scientific collaboration in different aspects of this work. Also, the authors would like to thank the three referees for their comments.

Table 1. The impact of logarithmic transformation on the 90% confidence interval and the Shapero-Wilk normality indicator

90% Confidence Interval

Linear Scale

Distribution	LL	Point Est	UL	WCI
Uniform	88.94	100.00	111.06	22.12
Normal	88.09	100.00	111.91	23.81
Log-Normal	89.89	100.00	110.11	20.22
Logistic	88.70	100.00	111.30	22.60
Gamma	89.47	100.00	110.53	21.06

Log-Scale

Distribution	LL	Point Est	UL	WCI
Uniform	92.44	100.13	108.46	16.02
Normal	85.52	101.48	120.41	34.88
Log-Normal	90.62	99.71	109.72	19.09
Logistic	89.39	100.20	112.32	22.93
Gamma	92.44	100.13	108.46	16.02

Shapiro-Wilk Normality Test

Linear Scale

Distribution	$(\sum \alpha_i Y_i)^2$	S^2	S-W Statistic
Uniform	382.19	423.00	0.90
Normal	409.47	423.00	0.97
Log-Normal	408.16	423.00	0.96
Logistic	406.66	423.00	0.96
Gamma	279.70	423.00	0.66

Log-Scale

Distribution	$(\sum \alpha_i Y_i)^2$	S^2	S-W Statistic
Uniform	4.28	4.80	0.89
Normal	5.97	7.18	0.83
Log-Normal	4.34	4.43	0.98
Logistic	4.51	4.72	0.96
Gamma	2.02	2.45	0.83

Table 2. Values of the parameters entering the equation (13) for fitting curves obtained for six sets of data with different distributions and for the same sets of data after log-transformation

Distribution	a_0	a_1	a_2	a_3	λ_1	λ_2	λ_3
Normal 1	14.53	-1.94	4.89	-7.37	$-5.5*10^{-3}$	-0.037	-0.027
Normal 2	14.3	-1.79	0.75	-3.23	$-5.1*10^{-3}$	-0.056	-0.022
Log Normal	16.26	-2.65	2.51	-6.25	$-6.42*10^{-3}$	-0.068	-0.03
Uniform	11.65	-0.36	-0.79	-0.59	$-4.6*10^{-3}$	-0.065	-0.02
Logistics	15.4	-2.74	3.74	-5.73	$-9.4*10^{-3}$	-0.1	-0.041
Gamma	22.74	0.035	-12.1	-4.72	$4*10^{-3}$	-0.038	$-8.8*10^{-3}$
Normal 1 (Log)	12.87	0.067	-2.16	-0.7	$3.66*10^{-3}$	-0.014	$-6.1*10^{-3}$
Normal 2 (Log)	13.4	-1.14	-0.18	-2.11	$-2.8*10^{-3}$	-0.108	-0.014
LogNormal (Log)	14.42	-1.97	1.34	-3.67	$-5.53*10^{-3}$	-0.056	-0.024
Uniform (Log)	11.52	-0.33	-0.69	-0.54	$-4.29*10^{-3}$	-0.066	-0.02
Gamma (Log)	12.36	-0.7	-0.49	-1.12	$-2.8*10^{-3}$	-0.047	$-.9*10^{-3}$

Table 3. Values of the two parameters (α and β) entering the equation (18) for fitting curves obtained for six sets of data with different distributions.

Distribution	α	β
Normal 1	0.802	0.8
Normal 2	0.805	0.802
Log Normal	0.757	0.845
Uniform	0.999	1.001
Logistics	0.737	0.74
Gamma	0.634	1.057

REFERENCES

Abdul-Gader, Jafar and Nigmatullin, R. R. (2001). Identification of a New Function Model for the AC-impedance of Thermally Evaporated (Undoped) Selenium Films Using the Eigen-Coordinates Method, Thin Solid Films, 396, 280-294.

Al-Hasan, M., and Nigmatulin, R. R. (2003). Identification of the Generalized Weibull Distribution in Wind Speed Data by the Eigen-Coordinates Mathod, Renewable Energy, 28 (1), 93-110

D'Agostino R. B. (1986). Tests for normal distribution in goodness-of-fit techniques, Marcel Decker.

Massey F. (1951). The Kolmogorov-Smirnov test for goodness of fit." Journal of the American Statistical Association, Vol. 46, No.253, 68–78.

Mayer, P. R. and Brazzell, R. K. (1988). Application of statistical moments theory to pharmacokinetics. J. Clin. Pharmacol., 28, 481-483.

McCulloch, C. E. (1987). Test for equality of the variance with paired data. Commun. Stat. Theory Methods, 16, 1377-1391.

Nigmatullin R. R. (1998). Eigen-Coordinates: New method of identification of analytical functions in experimental measurements, Applied Magnetic Resonance, 14, 601-633.

Nigmatullin R. R. (2000). Recognition of nonextensive statistic distribution by the eigen-coordinates method, Physica A, 285, 547-565.

Nigmatullin, R. R. (2006). The statistics of the fractional moments: Is there any chance to read "quantitatively" any randomness?, Journal of Signal Processing, 86, 2529-2547.

Nigmatullin R. R. (2008). Strongly correlated variables and existence of the universal disctribution function for relative fluctuations. Physics of Wave Phenomena, vol.16(2). 119-145.

Nigmatullin R. R. (2010). Universal distribution function for the strongly-correlated fluctuations: general way for description of random sequences, Communications in Nonlinear Science and Numerical Simulation. 15, 637-647.

Nigmatullin, R. R., Arbuzov A. A. and Nelson S. O. (2006). Dielectric Relaxation of Complex Systems: Quality sensing and dielectric properties of honeydew melons from 10 MHZ to 1.8 GHZ", Journal of Instruments JINST, N1-P10002.

Shapiro, S. And Wilk M. (1965). Analysis of the variance test for normality (complete samples), Biometrika, 52, (3-4), 591–611.

Sheskin, D. (2007). Handbook of parametric and nonparametric statistical procedures, 4th ed., Boca Raton, FL, Chapman & Hall. CRC, 241-255.

Schuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and power approach for assessing the equivalence of average bioavailability, J. Pharmacokinet. Biopharma., 15, 657-680.

Yamoaka, K. and Nakagawa, T. (1978). Statistical Moments theory in pharmacokinetics, J Pharmacokinetic Biopharm. 6(6), 548-57.