GEOMETRIC MEAN FOR NEGATIVE AND ZERO VALUES

Elsayed A. E. Habib
Department of Mathematics and Statistics, Faculty of Commerce, Benha University,
Egypt & Management & Marketing Department, College of Business, University of Bahrain, P.O. Box 32038,
Kingdom of Bahrain

ABSTRACT
A geometric mean tends to dampen the effect of very high values where it is a log-transformation of data. In this paper, the geometric mean for data that includes negative and zero values are derived. It turns up that the data could have one geometric mean, two geometric means or three geometric means. Consequently, the geometric mean for discrete distributions is obtained. The concept of geometric unbiased estimator is introduced and the interval estimation for the geometric mean is studied in terms of coverage probability. It is shown that the geometric mean is more efficient than the median in the estimation of the scale parameter of the log-logistic distribution.

Keywords: Coverage probability; geometric mean; lognormal distribution; robustness.

INTRODUCTION

Geometric mean is used in many fields, most notably financial reporting. This is because when evaluating investment returns and fluctuating interest rates, it is the geometric mean that gives the average financial rate of return; see, Blume (1974), Cheng and Karson (1985), Poterba (1988) and Cooper (1996). Many wastewater dischargers, as well as regulators who monitor swimming beaches and shellfish areas, must test for fecal coliform bacteria concentrations. Often, the data must be summarized as a geometric mean of all the test results obtained during a reporting period. Also, the public health regulations identify a precise geometric mean concentration at which shellfish beds or swimming beaches must be closed; see, for example, Draper and Yang (1997), Elton (1974), Jean (1980), Moore (1996), Michoud (1981), Limpert et al. (2001) and Martin (2007).
In this paper, the geometric mean for zero and negative values is derived. The data could have one geometric mean, two geometric means (bi-geometrical) or three geometric means (tri-geometrical). The overall geometric mean might be obtained as a weighted average of all the geometric means. Therefore, the geometric mean for discrete distributions is obtained. The concept of geometric unbiased estimator is introduced and the point and interval estimation of the geometric mean are studied based on lognormal distribution in terms of coverage probability.
The population geometric mean and its properties are defined in Section 2. The geometric mean for negative and zero values are derived in Section 3. Estimation of the geometric mean is defined in Section 4. The sampling distribution of geometric mean is obtained in Section 5. Simulation study is presented in Section 6. Approximation methods are presented in Section 7. An application to estimation of the scale parameter of log-logistic distribution is studied in Section 8. Section 9 is devoted for conclusions.

1 GEOMETRIC MEAN

Let $X_1, X_2, ...$be a sequence of independent random variables from a distribution with, probability function $p(x)$, density function $f(x)$, quantile function $x(F) = F^{-1}(x) = Q(u)$ where $0 < u < 1$, cumulative distribution function $F(x) = F_X = F$, the population mean $\mu = \mu_X$ and the population median $\nu = \nu_X$.

1.1 Population geometric mean

The geometric mean for population is usually defined for positive random variable as

$$G = G_X = \sqrt[N]{\prod_{i=1}^{N} X_i} = \left(\prod_{i=1}^{N} X_i\right)^{1/N} \tag{1}$$

by taking the logarithm

$$\log G = \frac{1}{N}\sum_{i=1}^{N} \log X_i \tag{2}$$

This is the mean of the logarithm of the random variable X, i.e,

$$\log G = E(\log X) = E(\log x(F)) \tag{3}$$

Therefore,

$$G = e^{E(\log X)} = e^{[E(\log x(F))]} \tag{4}$$

See; for example, Cheng and Karson (1985).

1.2 Properties of geometric mean

The geometric mean has the following properties: if

1. $x = a$, a constant, then $G_a = e^{E \log a} = e^{\log a} = a$.
2. $Y = bX$, $b > 0$ constant, then $G_Y = e^{E(\log bX)} = e^{E(\log b + \log X)} = bG_X$
3. $Y = \frac{b}{X}$, $b > 0$, then $G_Y = e^{E\left(\log \frac{b}{X}\right)} = e^{E(\log b - \log X)} = \frac{b}{G_X}$.
4. $X_1, \dots, X_r$ and $Y_1, \dots, Y_k$ are jointly distributed random variables each with G_{X_i} and G_{Y_j} and $Z = \frac{\prod_{i=1}^r X_i}{\prod_{j=1}^k Y_i}$, then $G_Z = e^{E\left(\log \frac{\prod_{i=1}^r X_i}{\prod_{j=1}^k Y_i}\right)} = e^{E\left(\sum_{i=1}^k \log X_i - \sum_{j=1}^k \log Y_j\right)} = \frac{\prod_{i=1}^r G_{X_i}}{\prod_{i=1}^k G_{Y_i}}$.
5. $X_1, \dots, X_r$ are jointly distributed random variables with G_{X_i}, and $Y = \prod_{i=1}^r X_i$ then $G_Y = e^{E\left(\log \prod_{i=1}^r X_i\right)} = e^{E\left(\sum_{i=1}^r \log X_i\right)} = \prod_{i=1}^r G_{X_i}$
6. $X_1, \dots, X_r$ are jointly distributed random variables each with $E(X_i)$, c_i are constants, and $Y = e^{a + b \sum_{i=1}^r X_i^{c_i}}$, then $G_Y = e^{E\left(\log e^{a + b \sum_{i=1}^r X_i^{c_i}}\right)} = e^{a + b \sum_{i=1}^r E\left(X_i^{c_i}\right)}$.
7. $X_1, \dots, X_r$ are independent random variables with $E(X_i)$, and $Y = e^{a + b \prod_{i=1}^r X_i}$, then $G_Y = e^{E\left(\log e^{a + b \prod_{i=1}^r X_i}\right)} = e^{E\left(a + b \prod_{i=1}^r X_i\right)} = e^{a + b \prod_{i=1}^r E(X_i)}$.

2 GEOMETRIC MEAN FOR NEGATIVE AND ZERO VALUES

The geometric mean for negative values depends on the following rule. For odd values of N, every negative number x has a real negative N^{th} root, then

$$\sqrt[N_{odd}]{-X} = -\sqrt[N_{odd}]{X} \tag{5}$$

2.1 Case 1: if all $X < 0$ and N is odd

The geometric mean in terms of N^{th} root is

$$G = \sqrt[N]{\prod_{i=1}^{N} (-X_i)} = -\sqrt[N]{\prod_{i=1}^{N} |X_i|}, \tag{6}$$

This is minus the Nth root of the product of absolute values of X, then

$$-G = \sqrt[N]{\prod_{i=1}^{N} |X_i|} \tag{7}$$

Hence,

$$\log(-G) = \frac{1}{N} \sum_{i=1}^{N} \log|X_i| = E(\log|X|) = E(\log|x(F)|) \tag{8}$$

The geometric mean for negative values is

$$-G = e^{E(\log|X|)} \quad \text{or} \quad G = -G_{|X|} \tag{9}$$

This is minus the geometric mean for the absolute values of X.

2.2 Case 2: negative and positive values (bi-geometrical).

In this case it could use the following

$$\sqrt[N]{ab} = \sqrt[N]{a}\sqrt[N]{b} \tag{10}$$

Consequently, under the conditions that N and N_1 are odd the geometric mean is

$$G = \sqrt[N]{\prod_{i=1}^{N} X_i} = \sqrt[N]{\prod_{i=1}^{N_1} X_i^- \prod_{i=N_1+1}^{N} X_i^+} = \sqrt[N]{\prod_{i=1}^{N_1} X_i^-}\sqrt[N]{\prod_{i=N_1+1}^{N} X_i^+} \tag{11}$$

There are two geometric means (bi-geometrical). The geometric mean for negative values is

$$G_- = \sqrt[N]{\prod_{i=1}^{N_1} X_i^-}, \qquad \log(-G_-) = E(\log|X^-|). \tag{12}$$

and

$$-G_- = e^{E(\log|X^-|)} \quad \text{or} \quad G_- = -G_{|X^-|} \tag{13}$$

The second for positive values is

$$G_+ = \sqrt[N]{\prod_{i=N_1+1}^{N} X_i^+}, \qquad \log(G_+) = E(\log X^+) \tag{14}$$

and

$$G_+ = e^{E(\log X^+)} = G_{X^+} \tag{15}$$

If one value is needed it might obtain an overall geometric mean as a weighted average

$$G = W_1 G_- + W_2 G_+ = \begin{cases} G_-, & \text{with } p(-\infty < X < 0) \\ G_+, & \text{with } p(0 < X < \infty) \end{cases} \tag{16}$$

where $W_1 = \frac{N_1}{N} = p(-\infty < X < 0)$ and $W_2 = \frac{N_2}{N} = p(0 < X < \infty)$.

2.3 Case 3: zero included in the data (tri-geometrical)

With the same logic there are three geometric means (tri-geometrical). G_- for negative values with numbers N_1, G_+ for positive values with numbers N_2, and $G_0 = 0$ for zero values with numbers N_3. It may write the overall geometric mean as

$$G = W_1 G_- + W_2 G_+ + W_3 G_0 = \begin{cases} G_-, & \text{with } p(-\infty < X < 0) \\ G_+, & \text{with } p(0 < X < \infty) \\ G_0 = 0, & \text{with } p(X = 0) \end{cases} \tag{17}$$

where $W_3 = \frac{N_3}{N} = p(X = 0)$ and $N = N_1 + N_2 + N_3$ are the total numbers of negative, positive and zero values.

2.4 Examples

The density, cumulative and quantile functions for Pareto distribution are

$$f(x) = \alpha\beta^{\alpha}x^{-\alpha-1}, x > \beta, \text{ and } x(F) = \beta(1-F)^{-1/\alpha} \tag{18}$$

with scale β and shape α; see, Elamir (2010) and Forbes et al. (2011). The mean and the median are

$$\mu = \frac{\alpha\beta}{\alpha - 1}, \text{ and } \nu = \beta\sqrt[\alpha]{2} \tag{19}$$

The geometric mean is

$$\log G = \int_0^1 \log\left[\beta(1-F)^{-1/\alpha}\right]dF = \log\beta + \frac{1}{\alpha} = \log\left(\beta e^{\frac{1}{\alpha}}\right), \tag{20}$$

and

$$G = \beta e^{\frac{1}{\alpha}} \tag{21}$$

The ratios of geometric mean to the mean and the median are

$$C_{\frac{G}{\mu}} = \frac{G}{\mu} = \frac{(\alpha-1)e^{\frac{1}{\alpha}}}{\alpha}, \text{ and } C_{\frac{G}{\nu}} = \frac{G}{\nu} = \frac{e^{\frac{1}{\alpha}}}{\sqrt[\alpha]{2}} \tag{22}$$

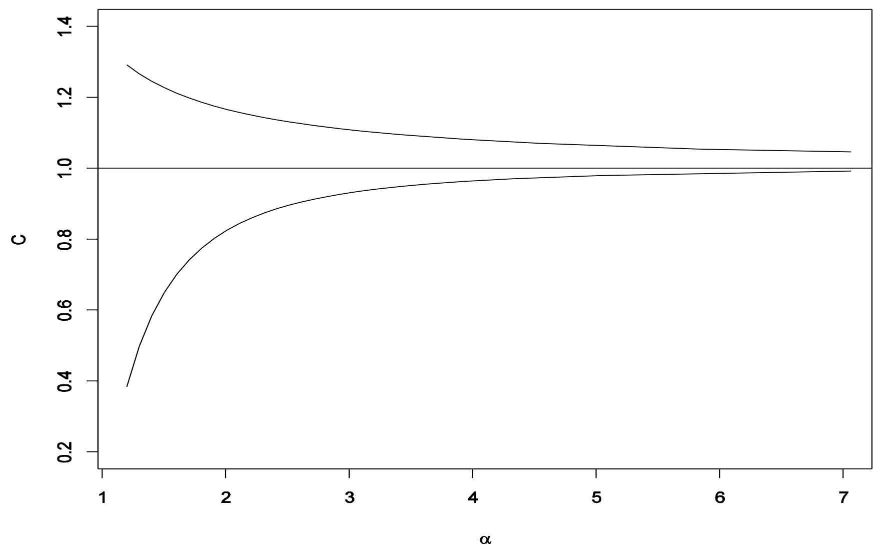

Figure 1 ratio of geometric mean to mean and median from Pareto distribution

Figure 1 shows that the geometric mean is quite less than the mean for small α and approaches quickly with increasing α. On the other hand, the geometric mean is more than the median for small α and approaches slowly for large α.

For uniform distribution with negative and positive values, the density is

$$f(x) = \frac{1}{b-a}, \quad a < x < b \tag{23}$$

then

$$\log G_{+} = \int_0^b \log x \left(\frac{1}{b-a}\right) dx = \frac{\mathrm{b} \log b - b}{b-a} \tag{24}$$

and

$$\log(-G_{-}) = \int_a^0 \log|x| \left(\frac{1}{b-a}\right) dx = \frac{|\mathrm{a}| \log|a| - |a|}{b-a} \tag{25}$$

Two geometric means are

$$G_{-} = -e^{\frac{|\mathrm{a}| \log|a| - |a|}{b-a}} = -\left(\frac{|a|}{e}\right)^{\frac{|a|}{b-a}}, \text{and } G_{+} = e^{\frac{\mathrm{b} \log b - b}{b-a}} = \left(\frac{b}{e}\right)^{\frac{b}{b-a}} \tag{26}$$

The weights could be found as

$$W_1 = p(a < X < 0) = \frac{|a|}{b-a}, \text{ and } W_2 = p(0 < X < b) = \frac{b}{b-a} \tag{27}$$

The overall geometric mean in terms of weighted average may be found as

(28)

$$G = W_2 \left(\frac{b}{e}\right)^{\frac{b}{b-a}} - W_1 \left(\frac{|a|}{e}\right)^{\frac{|a|}{b-a}} = \frac{b\left(\frac{b}{e}\right)^{\frac{b}{b-a}} - |a|\left(\frac{|a|}{e}\right)^{\frac{|a|}{b-a}}}{b-a}$$

Table 1 gives the values of geometric mean and mean from uniform distribution for different choices of a and b.

Table 1 geometric mean and mean from uniform distribution

(a,b)	(3,10)	(0,1)	(-2,0)	(-1,0)	(-1,1)	(-3,2)	(-11,2)	(-3,10)
G_-	-	-	-0.736	-0.368	-0.736	-1.061	-3.263	-1.023
G_+	6.429	0.368	-	-	0.736	0.884	0.954	2.724
W_1	0	0	1	1	0.5	0.6	0.846	0.23
W_2	1	1	0	0	0.5	0.4	0.154	0.77
G	6.429	0.368	-0.736	-0.368	0	-0.283	-2.613	1.862
μ	6.5	0.5	-1	-0.5	0	-0.5	-3	3.5

For log-logistic distribution the density, cumulative and quantile functions with scale β and shape α are

$$f(x) = \frac{\frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1}}{\left[1+\left(\frac{x}{\alpha}\right)^{\beta}\right]^2}, \ x > 0, \text{ and } \quad x(F) = \beta\left[\frac{F}{1-F}\right]^{\frac{1}{\alpha}} \tag{29}$$

See, Johnson et al. (1994). The mean and median are

$$\mu = \frac{\beta\pi/\alpha}{\sin(\pi/\alpha)}, \quad \text{and} \quad \nu = \beta \tag{30}$$

The geometric mean is

$$\log G = \int_0^1 \log\left[\beta\left[\frac{1}{F}-1\right]^{-\frac{1}{\alpha}}\right] dF = \log\beta, \text{ and } \quad G = \beta \tag{31}$$

The ratios of geometric mean to mean and median are

$$\frac{G}{\mu} = \frac{\beta}{\frac{\beta\pi/\alpha}{\sin(\pi/\alpha)}} = \frac{\sin(\pi/\alpha)}{\pi/\alpha}, \text{ and } \frac{G}{\nu} = \frac{\beta}{\beta} = 1 \tag{32}$$

The Poisson distribution has a probability mass function

$$p(x) = e^{-\lambda}\frac{\lambda^x}{x!}, \ x = 0,1,2,\ldots \tag{33}$$

See, Forbes et al. (2011). The geometric mean is

$$G = \begin{cases} G_0 = 0, & \text{with prob.} = e^{-\lambda} \\ G_+ = \sum_{x=1}^{\infty} \log x \ e^{-\lambda}\frac{\lambda^x}{x!}, & \text{with prob.} = 1 - e^{-\lambda} \end{cases} \tag{34}$$

Table 2 gives the geometric mean and mean from Poisson distribution for different choices of λ and the number of terms used in the sum is 100. The binomial distribution has a probability mass function

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \ x = 0,1,2,\ldots,n \tag{35}$$

The geometric mean is

$$G = \begin{cases} G_0 = 0, & \text{with prob.} = (1-p)^n \\ G_+ = \sum_{x=1}^{n} \log x \ \binom{n}{x} p^x (1-p)^{n-x}, & \text{with prob.} = 1-(1-p)^n \end{cases} \tag{36}$$

Table 2 gives the values of geometric mean and mean from binomial distribution for $n = 10$ and different choices of p.

Table 2 geometric mean and mean from Binomial and Poisson distributions

				Poisson				
λ	0.1	0.5	0.75	1	1.5	3	7	10
G^0	0	0	0	0	0	0	0	0
$W_3 = p_0$	0.905	0.606	0.472	0.368	0.223	0.050	0.0009	0.00004
G^+	1.003	1.071	1.149	1.249	1.504	2.597	6.447	9.463
$W_2 = 1 - p_0$	0.095	0.394	0.528	0.632	0.777	0.950	0.9991	0.99995
G	0.095	0.421	0.606	0.789	1.168	2.468	6.442	9.462
μ	0.1	0.5	0.75	1	1.5	3	7	10
				Binomial, $n = 10$				
p	0.05	0.20	0.35	0.50	0.60	0.70	0.75	0.95
G^0	0	0	0	0	0	0	0	0
W_3	0.60	0.107	0.013	0.0009	0.0001	0	0	0
G^+	0.065	0.607	1.150	1.550	1.753	1.922	1.996	2.248
W_2	0.40	0.893	0.987	0.9991	0.9999	1	1	1
G	0.026	0.542	1.135	1.548	1.753	1.922	1.996	2.248
μ	0.5	2	3.5	5	6	7	7.5	9.5

3 ESTIMATION OF GEOMETRIC MEAN

Consider a random sample $X_1, X_2, \ldots, X_n$ of size n from the population.

3.1 Case 1: positive values

If all $X > 0$, the nonparametric estimator of the geometric mean is

$$\log g = \overline{\log x} = \frac{1}{n}\sum_{i=1}^{n} \log x_i\,, \quad \text{and} \qquad g = e^{\overline{\log x}} \tag{37}$$

and $\overline{\log x}$ is the sample mean of logarithm of x, hence, $E\left(\overline{\log x}\right) = \mathrm{E}(\log \mathrm{g}) = \log G$.

3.2 Case 2: negative and positive values

If there are negative and positive values and under the conditions n and $n_1$are odds, the estimator of geometric mean for negative values is

$$\log(-g_-) = \frac{n_1}{n}\overline{\log|x^-|} = \frac{1}{n}\sum_{i=1}^{n_1} \log|x_i^-| = \frac{1}{n}\sum_{i=1}^{n} I_{x<0} \log|x_i|\,, \tag{38}$$

and

$$-g_- = e^{\frac{n_1}{n}\overline{\log|x_-|}} \;\text{ or }\; g_- = -e^{\frac{n_1}{n}\overline{\log|x_-|}} \tag{39}$$

$I_{x<0}$ is the indicator function for x values less than 0. For positive values

$$\log g_+ = \frac{n_2}{n}\overline{\log x_+} = \frac{1}{n}\sum_{i=1}^{n_2} \log x_i = \frac{1}{n}\sum_{i=1}^{n} I_{x>0} \log x_i\,, \tag{40}$$

and

$$g_+ = e^{\frac{n_2}{n}\overline{\log x_+}} \tag{41}$$

where x_-, n_1, x_+, n_2 are the negative values and their numbers and the positive values and their numbers, respectively. The estimated overall geometric mean might be obtained as weighted average

$$g = \frac{n_1 g_- + n_2 g_+}{n} \tag{42}$$

3.3 Case 3: negative, positive and zero values

When there are negative, positive and zero values in the data and under the conditions that n and $n_1$are odd, the estimated weighted average geometric mean is

$$g = \frac{n_1 g_- + n_2 g_+ + n_3(g_0 = 0)}{n} = \frac{n_1 g_- + n_2 g_+}{n} \tag{43}$$

Note that if there are negative values and n and n_1 are even numbers it might delete one at random. If there are zero and positive values only in the data, the weighted average geometric mean is

$$g = \frac{n_2 g_+ + n_3 g_0}{n} = \frac{n_2}{n} g_+ \tag{44}$$

Definition 1

Let $\hat{\theta}$ is an estimator of θ. If $G_{\hat{\theta}} = \theta$, $\hat{\theta}$ is geometrically unbiased estimator to θ.

Example

let $\hat{\theta} = g = e^{\frac{1}{n}\sum \log x}$ and $\theta = G$, then

$$G_g = e^{E[\log g]} = e^{E\left[\log e^{\frac{1}{n}\sum \log x}\right]} = e^{\frac{1}{n}\sum E[\log x]} = e^{\frac{1}{n}\sum \log G} = G \tag{45}$$

Then g is geometrically an unbiased estimator for G.

4 SAMPLING DISTRIBUTIONS

In this section the sampling distribution of g is studied for negative, zero and positive values.

Theorem 1 (Norris 1940)

Let all $X > 0$ be a real-valued random variable and $E(\log X)$ and $E(\log X)^2$ exist, the variance and expected values of g are

$$\sigma_g^2 \approx \frac{\sigma_{\log x}^2}{n} G^2, \text{ and } \mu_g \approx G + \frac{\sigma_{\log x}^2}{2n} G \tag{46}$$

This can be estimated from data as

$$s_g^2 \approx \frac{s_{\log x}^2}{n} g^2, \text{ and } \hat{\mu}_g \approx g + \frac{s_{\log x}^2}{2n} g \tag{47}$$

where $\sigma_{\log x}^2$ and $s_{\log x}^2$ are the population and sample variances for $\log x$, respectively.

Corollary 1

If all $X < 0$, the variance and the expected values of g are

$$\sigma_g^2 = \frac{\sigma_{\log|x|}^2}{n} (G)^2, \text{ and } \mu_g \approx G + \frac{\sigma_{\log|x|}^2}{2n} G \tag{48}$$

Proof

By using the delta method on $g = -e^{\overline{\log|x|}}$ the result follows.

Theorem 2 (Parzen 2008)

If f is the quantile like (non-decreasing and continuous from the left) then $f(Y)$ has quantile $Q\big(u; f(Y)\big) = f[Q(u;Y)]$. If f is decreasing and continuous, then $Q[u; f(Y)] = f[Q(1-u;Y)]$.

Theorem 3

Under the assumptions of

1. $E(\log|X|)$ and $E(\log|X|)^2$ exist and
2. $\bar{y} = \overline{\log|x|}$ has approximately normal distribution for large n with $\mu_{\bar{y}} = \mu_{\overline{\log|x|}} = \log G$ and variance $\sigma_{\bar{y}}^2 = \frac{\sigma_{\log x}^2}{n}$, then

For all $X > 0$, the geometric mean $g = e^{\bar{y}}$ has approximately the quantile function

$$Q(u; g) = e^{Q_N(u; \mu_{\bar{y}}, \sigma_{\bar{y}})} \tag{49}$$

For all $X < 0$, the geometric mean $g = -e^{\bar{y}}$ has approximately the quantile function

$$Q(u; g) = -e^{Q_N(1-u; \mu_{\bar{y}}, \sigma_{\bar{y}})} \tag{50}$$

Q_N is the quantile function for normal distribution.

Proof
Where the function $e^{\bar{y}}$ is an increasing function and $-e^{\bar{y}}$ is a decreasing function and by using theorem 2 the result follows; see, Gilchrist (2000) and Asquith (2011).

Corollary 2
If all $X > 0$, the lower and upper $(1-\alpha)\%$ confidence intervals for G are

$$\left(e^{Q_N(\alpha;\,\mu_{\bar{y}},\sigma_{\bar{y}})},\ e^{Q_N(1-\alpha;\,\mu_{\bar{y}},\ \sigma_{\bar{y}})}\right) \tag{51}$$

If all $X < 0$, the lower and upper $(1-\alpha)\%$ confidence intervals for G are

$$\left(-e^{Q_N(1-\alpha;\,\mu_{\bar{y}},\sigma_{\bar{y}})}, -e^{Q_N(\alpha;\,\mu_{\bar{y}},\sigma_{\bar{y}})}\right) \tag{52}$$

Proof
The result follows directly from the quantile function that obtained in theorem 3.

Corollary 3
Under the assumptions of theorem 3
1. if all $X > 0,$ the distributional moments of g are
$$E(g) = e^{\mu+\frac{\sigma^2}{2n}} = Ge^{\frac{\sigma^2}{2n}}, G(g) = e^{\mu} = G,\ \text{Mode}(g) = e^{\mu-\sigma^2}\ \text{ and }\ \sigma_g^2 = [E(g)]^2\left[e^{\frac{\sigma^2}{n}} - 1\right]$$
and $\mu = \log G = E(\log x)$ and $\sigma^2 = \sigma^2_{\log\, x}$.
2. if all $X < 0,$ the distributional moments of $-g$ are
$$E(-g) = e^{\mu+\frac{\sigma^2}{2n}},\quad G(-g) = G,\ \text{Mode}(-g) = e^{\mu-\sigma^2}\ \text{ and }\ \sigma_{-g}^2 = [E(-g)]^2\left[e^{\frac{\sigma^2}{n}} - 1\right]$$
and $\mu = \mathrm{E}(\log|x|)$ and $\sigma^2 = \sigma^2_{\log|x|}$.

Proof
Since g and – g have lognormal distributions with mean μ and σ^2, the results follow using the moments of lognormal distribution; see, Forbes et al. (2011) for moments of lognormal.

It is interesting to compare the estimation using Norris's approximation $\hat{\mu}_g$ and s_g^2 and distributional approximation $\hat{E}(g)$ and $\hat{\sigma}_g^2$. The results using Pareto distribution with different choices of α, β and n are given in Table 3:
1. The main advantage of distributional approximation over Norris's approximation is that the distributional approximation has much less biased until in small sample sizes and very skewed distributions ($\alpha = 1.5$ and 2.5).
2. The distributional and Norris's approximations have almost the same variance.
3. $g = \hat{G}_g$ is geometrically unbiased to G.

Table 3 comparison between Norris and distributional approximations mean and variance for g using Pareto distribution and the number of replications is 10000.

			$\beta = 10$			
	G	$\hat{\mu}_g$	s_g^2	$\hat{E}(g)$	$\hat{G}_g$	$\hat{\sigma}_g^2$
n			$\alpha = 1.5$			
10	19.477	33.596	27.278	20.477	19.967	28.977
25	19.477	23.658	8.014	19.851	19.664	8.163
50	19.477	21.381	3.6707	19.635	19.545	3.656
100	19.477	20.389	1.761	19.566	19.522	1.761
			$\alpha = 2.5$			
10	14.918	17.478	4.872	15.175	15.042	4.619
25	14.918	15.765	1.596	15.017	14.967	1.553
50	14.918	15.310	0.7545	14.957	14.932	0.7437
100	14.918	15.128	0.3695	14.955	14.943	0.3667
			$\alpha = 7$			
10	11.535	11.707	0.3217	11.558	11.546	0.2913
25	11.535	11.602	0.1168	11.549	11.544	0.1123
50	11.535	11.564	0.0558	11.538	11.536	0.0547
100	11.535	11.550	0.0276	11.537	11.536	0.0273

Theorem 4

Under the assumptions of theorem 3 and $X \geq 0$, the geometric mean $g = \frac{n_2}{n} e^{\frac{n_2}{n}\overline{\log x^+}}$ has approximately

1. a lognormal distribution with $\mu = \frac{n_2}{n}\mu_{\overline{\log x^+}} + \log\frac{n_2}{n}$ and $\sigma^2 = \left(\frac{n_2}{n}\right)^2 \sigma^2_{\overline{\log x^+}}$,
2. the lower and upper $(1-\alpha)\%$ confidence intervals for G are

$$\left(e^{Q_N(\alpha;\,\mu,\sigma)},\ e^{Q_N(1-\alpha;\,\mu,\sigma)}\right)$$

Proof

If X is lognormal (μ, σ^2) then aX has lognormal $(\mu + \log a\,, \sigma^2)$; see Jonson et al. (1994). Since $e^{\frac{n_2}{n}\overline{\log x^+}}$has lognormal with $\mu_{\frac{n_2}{n}\overline{\log x^+}}$, $\sigma^2_{\frac{n_2}{n}\overline{\log x^+}}$, and $a = n_2/n$ then the result follows.

Theorem 5

For all values of X, the expected and variance values of $g = \frac{n_1 g_- + n_2 g_+}{n}$ are

$$E(g) \approx \frac{n_1\mu_{g_-} + n_2\mu_{g_+}}{n} \tag{53}$$

and

$$\sigma_g^2 \approx \left(\frac{n_1}{n}\right)^2 \sigma^2_{g_-} + \left(\frac{n_2}{n}\right)^2 \sigma^2_{g_+} - \frac{2n_1n_2}{n^2} Cov(|g_-|, g_+) \tag{54}$$

where

$$\sigma^2_{g_-} \approx \left(\frac{n_1}{n}\right)^2 \frac{\sigma^2_{\overline{\log|x^-|}}}{n_1}\left[\mu_{g_-}\right]^2 \text{ and } \sigma^2_{g_+} \approx \left(\frac{n_2}{n}\right)^2 \frac{\sigma^2_{\overline{\log|x^+|}}}{n_2}\left[\mu_{g_+}\right]^2,$$

$$Cov(|g_-|, g_+) = E\left(e^{\frac{n_1}{n}\overline{\log|x^-|} + \frac{n_2}{n}\overline{\log|x^+|}}\right) - E\left(e^{\frac{n_1}{n}\overline{\log|x^-|}}\right) E\left(e^{\frac{n_2}{n}\overline{\log|x^+|}}\right),$$

$$E\left(e^{\frac{n_1}{n}\overline{\log|x^-|}}\right) \approx |\mu_{g_-}| + \left(\frac{n_1}{n}\right)^2 \frac{\sigma^2_{\overline{\log|x^-|}}}{2n_1}|\mu_{g_-}|,$$

$$E\left(e^{\frac{n_2}{n}\overline{\log|x^+|}}\right) \approx \mu_{g_+} + \left(\frac{n_2}{n}\right)^2 \frac{\sigma^2_{\overline{\log|x^+|}}}{2n_2}\mu_{g_+},$$

$$E\left(e^{\frac{n_1}{n}\overline{\log|x^-|} + \frac{n_2}{n}\overline{\log|x^+|}}\right) \approx \mu_c + \frac{\mu_c}{2}\left[\left(\frac{n_1}{n}\right)^2 \frac{\sigma^2_{\overline{\log|x^-|}}}{n_1} + \left(\frac{n_2}{n}\right)^2 \frac{\sigma^2_{\overline{\log|x^+|}}}{n_2}\right]$$

and

$$c = e^{\frac{n_1}{n}\overline{\log|x^-|} + \frac{n_2}{n}\overline{\log|x^+|}}$$

Proof

Using the delta method for variance and expected values the results follow; see Johnson et al. (1994).

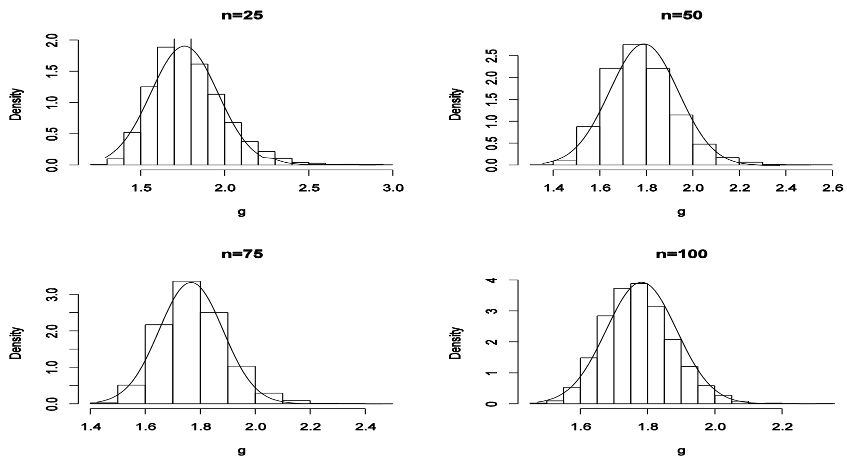

Figure 2 distribution of g with lognormal distribution superimposed using 5000 simulated data from Pareto distribution with $\beta = 1$ and $\alpha = 1.75$ and different n.

The lognormal distribution gives a good approximation to the distribution of g. Figure 2 shows the distribution of g using simulated data from Pareto distribution with $\beta = 1$, $\alpha = 1.75$ and different choices of n and Figure 3 shows the distributions of $-g$ and g using simulated data from normal distribution with $(-100, 10)$ and $(0,1)$ and $n = 25$ and 50, respectively.

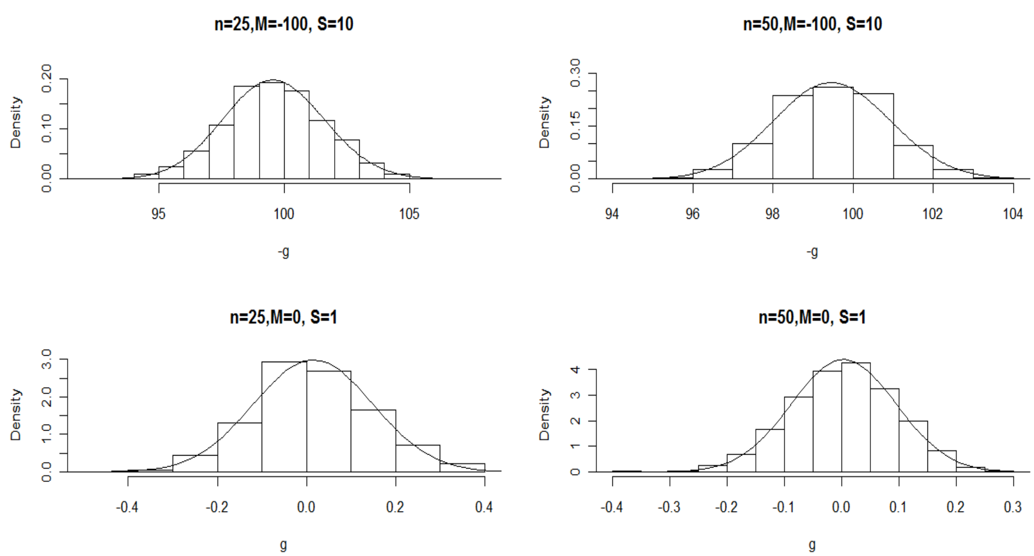

Figure 3 distribution of g with lognormal distribution superimposed using 5000 simulated data from normal distribution with mean=M, and standard deviation=S and $n = 25$ and 50.

Figure 3 distribution of g with lognormal distribution superimposed using 5000 simulated data from normal distribution with mean=M, and standard deviation=S and $n = 25$ and 50.

5 SIMULATION PROCEDURES

In order to assess the bias and root mean square error (RMSE) of g and the coverage probability of the confidence interval of G, a simulation study is built. Several scenarios are considered and in each scenario the simulated bias, RMSE are calculated. Further, the coverage probability of the confidence interval is evaluated for estimating the actual coverage of the confidence interval by the proportion of simulated confidence interval containing the true value of G. The design of the simulation study is

• sample sizes: 25, 50, 75, 100;
• number of replications: 10000;
• nominal coverage probability of confidence interval: 0.90, 0.95 and 0.99.

The bias and RMSE for Poison and Pareto distributions for different choices of parameters and sample sizes are reported in Table 4. For Poisson distribution, it is shown that if λ is near zero the bias is negligible and when as it becomes far away from zero the bias start to increase for small sample sizes and negligible for large sample sizes. For Pareto distribution, the bias is slightly noticeable for small α (very skewed distribution), and negligible for large values of α and n.

Table 4 bias and RMSE for g from Poisson and Pareto distributions

	Poisson					
	$\lambda = 0.1, G = 0.095$		$\lambda = 1, G = 0.789$		$\lambda = 3, G = 2.468$	
n	Bias	RMSE	Bias	RMSE	Bias	RMSE
25	0.0005	0.0481	0.007	0.1272	0.025	0.2812
50	0.0004	0.0338	0.003	0.0885	0.015	0.1985
75	0.0004	0.0269	0.002	0.0725	0.010	0.1633
100	-.0006	0.0234	0.0003	0.0614	0.009	0.1389
	Pareto					
	$\beta = 1, \alpha = 1.5$		$\beta = 1, \alpha = 3$		$\beta = 1, \alpha = 7$	
25	0.0224	0.2713	0.0010	0.0938	0.0010	0.0335
50	0.0094	0.1853	0.0011	0.0657	0.0010	0.0237
75	0.0023	0.1509	0.0005	0.0534	0.0008	0.0186
100	0.0031	0.1291	0.0003	0.0459	0.0007	0.0165

The simulation result for coverage probability using log-logistic distribution for different choices of n is given in Table 5. As suggested by the results obtained from the log-logistic distribution for different values of α, the small n the worst is the coverage probability. While the sample size increases, the coverage probability is improved. Then, a relatively small sample size of 50 is sufficient in order to assure a good coverage probability of the confidence interval.

Table 5 coverage probability of the confidence interval for G from log-logistic distribution and the number of replication is 10000

	Coverage probability					
	0.90	0.95	0.99	0.90	0.95	0.99
n	$\beta = 1, \alpha = 0.5, G = 1$			$\beta = 1, \alpha = 1, G = 1$		
25	0.884	0.939	0.984	0.888	0.938	0.987
50	0.897	0.948	0.988	0.898	0.948	0.988
75	0.898	0.947	0.988	0.897	0.947	0.988
100	0.899	0.949	0.989	0.902	0.951	0.989
	$\beta = 1, \alpha = 5, G = 1$			$\beta = 1, \alpha = 10, G = 1$		
25	0.890	0.940	0.985	0.889	0.940	0.983
50	0.892	0.945	0.989	0.894	0.946	0.987
75	0.901	0.950	0.989	0.899	0.949	0.987
100	0.900	0.951	0.990	0.900	0.950	0.989

Moreover, Table 6 shows comparison between the geometric mean, sample mean and median for negative values from normal distribution with $\mu = -100$ and $\sigma = 5$ in terms of mean square error. The table shows the geometric mean is more efficient that the median and very comparable to the sample mean where efficiency is 0.988 at $n = 25$ and 0.95 at $n = 101$.

Table 6 Mean square error (MSE) and efficiency (eff.) with respect to mean using simulated data from normal distribution and the number of replications is 5000.

	Normal $(-100, 5)$				
	Mean	median		geometric mean	
n	MSE	MSE	eff.	MSE	eff.
25	1.082	1.708	0.633	1.095	0.988
51	0.498	0.749	0.665	0.513	0.971
101	0.251	0.404	0.652	0.263	0.953
201	0.133	0.196	0.676	0.148	0.907

6 APPROXIMATION METHODS

When it is difficult to obtain geometric mean exactly, the approximation may be useful in some cases; see, Cheng and Lee (1991), Zhang and Xu (2011) and Jean and Helms (1983). Let X be a real-valued random variable. If $E(X)$ and $E(X^2)$ exist, the first and the second order approximation of geometric mean are

$$\log G_X \approx \log E(X)\,, \text{ and } \quad \log G_X \approx \log\left[E(X)e^{-\frac{\sigma_X^2}{2[E(x)]^2}}\right] \tag{55}$$

Therefore,

$$G_X \approx E(X), \text{ and } G_X \approx E(X)e^{-\frac{\sigma_X^2}{2[E(X)]^2}} \tag{56}$$

Example

For lognormal distribution $E(X) = e^{\mu+\frac{\sigma^2}{2}}$, $\sigma_X^2 = \mu^2\left[e^{\sigma^2} - 1\right]$ and the exact geometric mean is $G_X = e^{\mu}$. The first and the second order approximations are

$$\log G_X \approx \log E(X) = \mu + \frac{\sigma^2}{2}, \text{ and } \quad G_X \approx e^{\mu+\frac{\sigma^2}{2}} \tag{57}$$

and

$$\log G_X \approx \mu + \frac{\sigma^2}{2} - \left[\frac{e^{\sigma^2} - 1}{2}\right], \text{ and } \quad G_X \approx e^{\mu+\frac{\sigma^2}{2}-\left[\frac{e^{\sigma^2}-1}{2}\right]} \tag{58}$$

The ratios of exact to approximation are

$$R_1 = \frac{G(\text{exact})}{G(\text{approx.})} = \frac{e^{\mu}}{e^{\mu+\frac{\sigma^2}{2}}} = e^{-\frac{\sigma^2}{2}} \text{ and } R_2 = e^{-\frac{\sigma^2}{2}+\left[\frac{e^{\sigma^2}-1}{2}\right]} \tag{59}$$

Table 7 shows the first and the second order approximations for g from lognormal distribution. The first order approximation is good as long as $\frac{\sigma}{|E(X)|} < 0.10$ and the second order approximation is very good as long as $\frac{\sigma}{|E(X)|} < 0.50$.

Table 7 ratios of exact to approximated geometric mean from lognormal dsitribution

σ	0.01	0.05	0.10	0.20	0.50	0.70	0.90	1	1.10
σ/μ	0.01	0.05	0.10	0.20	0.53	0.80	1.12	1.31	1.53
R_1	1	0.9987	0.9950	0.9802	0.8825	0.7827	0.6670	0.6065	0.546
R_2	1	1	1	1	1.017	1.074	1.245	1.432	1.771

For normal distribution $E(X) = \mu$ and $\sigma_X^2 = \sigma^2$. The first and the second order approximations are

$$\log G_X \approx \log E(X) = \log \mu\,, \text{ and } \quad G_X \approx \mu \tag{60}$$

and

$$\log G_X \approx \log\left[\mu e^{-\frac{\sigma^2}{2\mu^2}}\right], \text{ and } \quad G_X \approx \mu e^{-\frac{\sigma^2}{2\mu^2}} \tag{61}$$

then,

$$\mu \approx G e^{\frac{\sigma^2}{2\mu^2}} \tag{62}$$

Table 8 shows the simulation results of mean, median, the first and second order approximations for g from normal distribution.

Table 8 sample mean, median, first (g_1), second order (g_2) approximations of g and coefficient of variation for simulated data from normal distribution and $n = 25$.

	$\sigma = 1$				$\sigma = 10$				
μ	50	10	-10	-50	50	10	8	-12	-15
$\bar{x}$	50.005	10.002	-10.004	-49.99	50.05	10.63	7.99	-12.02	-15.01
med	50.003	10.001	-10.001	-49.99	50.03	10.01	8.02	-11.98	-15.03
$\sigma/\lvert\mu\rvert$	0.02	0.1	0.1	0.02	0.20	1	**1.25**	0.833	0.667
g_1	49.995	9.962	-9.955	-49.98	49.04	6.09	**4.48**	-7.43	-10.48
g_2	50.004	10.002	-10.004	-49.99	50.03	10.62	**8000**	-10.56	-14.60

7 APPLICATION

7.1 Estimation of the scale parameter of log-logistic distribution

From Forbes et al. (2011) the log-logistic distribution with scale β and shape α is defined as

$$f(x) = \frac{\frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1}}{\left[1+\left(\frac{x}{\alpha}\right)^{\beta}\right]^2}, \quad x > 0, \text{ and } \quad x(F) = \beta\left[\frac{F}{1-F}\right]^{\frac{1}{\alpha}} \tag{63}$$

The geometric mean and median for logistic distribution are

$$G = \beta, \quad \text{and} \quad \nu = \beta \tag{64}$$

For known values of the shape parameter α the scale parameter β can be estimated as

$$\hat{\beta} = g, \quad \text{and } \hat{\beta} = med(x) \tag{65}$$

Table 9 shows the bias, mean square error (MSE), efficiency$\left(\frac{MSE(g)}{MSE(med)}\right)$ and geometric bias $(g - G)$ for β with known values of α using simulated data from log-logistic distribution and different choices of n and the number of replications is 10000.

Table 9 bias, MSE, efficiency and geometric bias for $\beta = 10$ from log-logistic distribution and different choice of α and n and number of replication is 10000

		n					
		15	25	50	100	150	200
		$\alpha = 0.5$					
med	Bias	7.618	3.823	1.729	0.828	0.5704	0.4267
	MSE	703.93	190.92	52.15	20.624	12.689	9.3597
g	Bias	5.613	2.966	1.397	0.6570	0.4542	0.3524
	MSE	396.08	148.13	40.41	16.548	10.214	7.3461
Efficiency $\left(\frac{MSE_g}{MSE_{med}}\right)$		**0.562**	**0.775**	**0.775**	**0.8023**	**0.8049**	**0.785**
Geometric Bias		0.0198	-.0460	-.0006	-.0308	0.0066	0.0020
		$\alpha = 1$					
Med	Bias	1.37	0.877	0.425	0.239	0.107	0.116
	MSE	42.233	21.997	9.090	4.416	2.779	2.055
g	Bias	1.09	0.699	0.331	0.189	0.096	0.085
	MSE	31.561	17.157	7.507	3.531	2.262	1.690
Efficiency $\left(\frac{MSE_g}{MSE_{med}}\right)$		**0.75**	**0.78**	**0.82**	**0.80**	**0.81**	**0.822**
Geometric Bias		-.0010	-.0054	-.0127	-0.0112	0.0171	0.0040
		$\alpha = 8$					
Med	Bias	0.027	0.016	0.0053	0.002	0.0035	0.0024
	MSE	0.4213	0.2505	0.1214	0.0618	0.0411	0.0312
g	Bias	0.0145	0.0098	0.0051	0.0007	0.0021	0.0025
	MSE	0.3541	0.2051	0.1029	0.0522	0.0342	0.0255
Efficiency $\left(\frac{MSE_g}{MSE_{med}}\right)$		**0.84**	**0.82**	**0.84**	**0.84**	**0.83**	**0.82**
Geometric Bias		-.0030	-.0003	0	-0.0002	0.0004	0.0001

Table 9 shows that

1. In general the geometric mean (g) is less biased than median (med).
2. g has less mean square error (MSE) and, therefore, is more efficient than med.
3. g is geometrically unbiased estimator for the parameter β.

8 CONCLUSION

There are many areas in economics, chemical, finance and physical sciences in which the data could include zero and negative values. In those cases, the computation of geometric mean presents a much greater challenge. By using the rule: for odd numbers every negative number x has a real negative N^{th} root, it is derived to the geometric mean as a minus of geometric mean for absolute values. Therefore, the data could have one geometric mean, two geometric means and three geometric means. The overall geometric mean is obtained as a weighted average of all geometric means. Of course different rules could be used. The sample geometric mean is proved to be geometrically unbiased estimator to population geometric mean. Moreover, it is shown that the geometric mean is outperformed the median in estimation the scale parameter from log-logistic distribution data in terms of the bias and the mean square error where geometric mean tends to dampen the effect of very high values by taking the logarithm of the data. Its interval estimation is obtained using lognormal distribution and it is shown that the geometric mean had a good performance for large and small sample sizes in terms of coverage probability.

ACKNOWLEDGMENT

The author is greatly grateful to associate editor and reviewers for careful reading, valuable comments and highly constructive suggestions that have led to clarity, better presentation and improvements in the manuscript.

REFERENCES

[1]. Norris, N. (1940) The standard errors of the geometric and harmonic means and their application to index numbers. *The Annals of Mathematical Statistics*, 11, 445-448.
[2]. Blume, M.E. (1974) Unbiased estimators of long-run expected rates of return. *Journal of the American Statistical Association*, 69, 634-638.
[3]. Elton, E.J., and Gruber, M.I. (1974) On the maximization of the geometric mean with a lognormal return distribution. *Management Science*, 21, 483-488.
[4]. Jean, W.H. (1980) The geometric mean and stochastic domaince. *Journal of Finance*, 35, 151-158.
[5]. Michaud, R.O. (1981) Risk policy and long-term investment. *Journal of Financial and Quantitative Analysis*, XVI, 147-167.
[6]. Jean, W.H., and Helms, B.P. (1983) Geometric mean approximations. *Journal of Financial and Quantitative Analysis*, 18, 287-293.
[7]. Cheng, D.C. and Karson, M.J. (1985) Concepts, theory, and techniques on the use of the geometric mean in long-term investment. *Decision Sciences*, 16, 1-13.
[8]. Poterba, J.M. and Summers, L.H. (1988) Mean reversion in stock prices. *Journal of Financial Eonomics*, 22, 27-59.
[9]. Cheng, D.C. and Lee, C.F. (1991) Geometric mean approximations and risk policy in long-term investment. *Advances in Quantitative Analysis of Finance and Accounting*, 26, 1-14.
[10]. Johnson N.L., Kotz S., and Balakrishnan N. (1994) *Continuous univariate distributions.* Vol. 1 and 2, Wiley Jhon & Sons.
[11]. Cooper, I. (1996) Arithmetic versus geometric mean estimators: setting discount rates for capital budgeting. *European Financial Management*, 2, 157-167.
[12]. Moore, R.E (1996) Ranking income distributions using the geometric mean and a related general measure. *Southern Economic Journal*, 63, 69-75.
[13]. Draper, N.R. and Yang, Y. (1997) Generalization of the geometric mean functional. *Computational Statistics & Data Analysis*, 23, 355-372.
[14]. Gilchrist W.G. (2000) *Statistical modeling with quantile functions*. Chapman & Hall.
[15]. Limpert, E., Stahel, W.A. and Abbt, M. (2001) log-normal distributions across the sciences: keys and clues. *BioScience*, 51, 341-352.
[16]. Martin M.D. (2007) The geometric mean versus the arithmetic mean. *Economic Damage*, 2401-2404.
[17]. Parzen,E. (2008) United statistics, confidence quantiles, Bayesian statistics. *Journal of Statistical Planning and Inference*, 138, 2777-2785.
[18]. Elamir, E.A.H. (2010) Optimal choices for trimming in trimmed L-moment method. *Applied Mathematical Sciences*, 4,2881-2890.
[19]. Asquith, W.H. (2011) *Univariate distributional analysis with L-moment statistics using R*. Ph.D. Thesis, Department of Civil Engineering, Texas Tech University.
[20]. Forbes G., Evans M., Hasting N. and Peacock B.(2011) *Statistical distributions*. Fourth Edition, Wiley Jhon & Sons.
[21]. Zhang, Q. and Xu, Bing (2011) An invariance geometric mean with respect to generalized quasi-arithmetic-mean. *Journal of Mathematical and Applications*, 379, 65-74.