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 geometric mean for 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 $X_1, X_2, \ldots$ 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

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

by taking the logarithm

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

This is the mean of the logarithm of the random variable $X$, i.e,
\[
\log G = E(\log X) = E(\log x(F))
\]

Therefore,

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

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 \( G_Y = e^{E(\log \frac{b}{X})} = e^{E(\log b - \log X)} = \frac{b}{G_X}. \)
4. \( X_1, ..., X_r, \) and \( Y_1, ..., Y_k \) are jointly distributed random variables each with \( G_{X_i} \) and \( G_{Y_j} \) and \( Z = \frac{\prod_{i=1}^{r} X_i}{\prod_{i=1}^{k} Y_i}, \) then
   \[
   G_Z = e^{E(\log (\frac{\prod_{i=1}^{r} X_i}{\prod_{i=1}^{k} Y_i}))} = e^{E(\sum_{i=1}^{r} \log X_i - \sum_{i=1}^{k} \log Y_i)} = \frac{\prod_{i=1}^{r} G_{X_i}}{\prod_{i=1}^{k} G_{Y_i}}.
   \]
5. \( X_1, ..., X_r, \) are jointly distributed random variables with \( G_{X_i}, \) and \( Y = \prod_{i=1}^{r} X_i \) then \( G_Y = e^{E(\log \prod_{i=1}^{r} X_i)} = e^{E(\sum_{i=1}^{r} \log X_i)} = \prod_{i=1}^{r} G_{X_i}. \)
6. \( X_1, ..., X_r, \) are jointly distributed random variables each with \( E(X_i), \) \( c_i \) are constants, and \( Y = e^{a+b \sum_{i=1}^{r} X_i^c}, \) then \( G_Y = e^{E(a \sum_{i=1}^{r} X_i^c)} = e^{a+b \sum_{i=1}^{r} E(X_i^c)}. \)
7. \( X_1, ..., X_r, \) are independent random variables with \( E(X_i), \) and \( Y = e^{a+b \prod_{i=1}^{r} X_i}, \) then \( G_Y = e^{E(a \prod_{i=1}^{r} X_i)} = e^{a+b \prod_{i=1}^{r} E(X_i)}. \)

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

\[
N_{\text{odd}} \sqrt[\text{odd}]{x} = \frac{N_{\text{odd}}}{\sqrt{x}}
\]

2.1 Case 1: if all \( X < 0 \) and \( N \) is odd

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

\[
G = \prod_{i=1}^{N} \sqrt[\text{odd}]{(-X_i)} = \prod_{i=1}^{N} |X_i|
\]

This is minus the \( N^{th} \) root of the product of absolute values of \( X, \) then

\[
-G = \prod_{i=1}^{N} |X_i|
\]

Hence,

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

The geometric mean for negative values is

\[
-G = e^{E(\log|X|)} \quad \text{or} \quad G = -G|X|
\]
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

\[
\sqrt[N]{ab} = \sqrt[N]{a} \sqrt[N]{b}
\]

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

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

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

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

and

\[
-G_- = e^{E(\log|X^-|)} \quad \text{or} \quad G_- = -G_{|X^-|}
\]

The second for positive values is

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

and

\[
G_+ = e^{E(\log X^+)} = G_X^+
\]

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

\[
G = W_1 G_- + W_2 G_+ = \begin{cases} G_- & \text{with } p(-\infty < X < 0) \\ G_+ & \text{with } p(0 < X < \infty) \end{cases}
\]

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

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

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

\[
f(x) = \alpha \beta^{\alpha} x^{-\alpha - 1}, x > \beta, \quad \text{and} \quad x(F) = \beta (1 - F)^{-1/\alpha}
\]

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

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

The geometric mean is

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

and
The geometric mean is given by

\[ G = \beta e^\frac{1}{\alpha} \]  

(21)

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

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

(22)

\[ \text{Figure 1} \text{ ratio of geometric mean to mean and median from Pareto distribution} \]

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

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

\[ f(x) = \frac{1}{b-a}, \quad a < x < b \]  

(23)

then

\[ \log G_u = \int_a^b \log x \left( \frac{1}{b-a} \right) dx = \frac{b \log b - b}{b - a} \]  

(24)

and

\[ \log(-G_u) = \int_a^b \log|x| \left( \frac{1}{b-a} \right) dx = \frac{|a| \log|a| - |a|}{b - a} \]  

(25)

Two geometric means are

\[ G_u = -e \frac{|a| \log|a| - |a|}{b-a} = \left( \frac{|a|}{e} \right)^\frac{1}{b-a}, \text{ and } G_u = e \frac{b \log b - b}{b-a} = \left( \frac{b}{e} \right)^\frac{b}{b-a} \]  

(26)

The weights could be found as

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

(27)

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

\[ \text{(28)} \]
\[ G = W_2 \left( \frac{b}{e} \right)^{\frac{b-a}{b}} - W_1 \left( \frac{|a|}{e} \right)^{\frac{|a|}{b-a}} = \frac{b - a e^{b-a} - |a| e^{b-a}}{b - a} \]

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

<table>
<thead>
<tr>
<th>( (a, b) )</th>
<th>(3,10)</th>
<th>(0,1)</th>
<th>(-2,0)</th>
<th>(-1,0)</th>
<th>(-1,1)</th>
<th>(-11,2)</th>
<th>(-3,10)</th>
</tr>
</thead>
<tbody>
<tr>
<td>( G_+ )</td>
<td>6.429</td>
<td>0.368</td>
<td>0.736</td>
<td>0.368</td>
<td>0.736</td>
<td>1.061</td>
<td>3.263</td>
</tr>
<tr>
<td>( W_1 )</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0.5</td>
<td>0.6</td>
<td>0.846</td>
</tr>
<tr>
<td>( W_2 )</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0.5</td>
<td>0.4</td>
<td>0.154</td>
</tr>
<tr>
<td>( G )</td>
<td>6.429</td>
<td>0.368</td>
<td>-0.736</td>
<td>-0.368</td>
<td>0</td>
<td>-0.283</td>
<td>2.613</td>
</tr>
<tr>
<td>( \mu )</td>
<td>6.5</td>
<td>0.5</td>
<td>-1</td>
<td>-0.5</td>
<td>0</td>
<td>-3</td>
<td>3.5</td>
</tr>
</tbody>
</table>

For log-logistic distribution the density, cumulative and quantile functions with scale \( \beta \) and shape \( \alpha \) are

\[ f(x) = \frac{\alpha}{\beta} \left( \frac{x}{\beta} \right)^{\alpha-1} \frac{e^{-\left( \frac{x}{\beta} \right)^{\frac{1}{\alpha}}}}{1 + \left( \frac{x}{\beta} \right)^{\frac{1}{\alpha}}}, \quad x > 0, \]  
and \( x(F) = \beta \left( \frac{F}{1-F} \right)^{\frac{1}{\alpha}} \) \hspace{1cm} (29)

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

\[ \mu = \frac{\beta \pi / \alpha}{\sin(\pi / \alpha)}, \quad \text{and} \quad \nu = \beta \] \hspace{1cm} (30)

The geometric mean is

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

The ratios of geometric mean to mean and median are

\[ \frac{G}{\mu} = \frac{\beta}{\beta \pi / \alpha} = \frac{\sin(\pi / \alpha)}{\pi / \alpha}, \quad \text{and} \quad \frac{G}{\nu} = \frac{\beta}{\beta} = 1 \] \hspace{1cm} (32)

The Poisson distribution has a probability mass function

\[ p(x) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x = 0, 1, 2, ... \] \hspace{1cm} (33)

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

\[ G = \begin{cases} 
G_0 = 0, & \text{with prob.} = e^{-\lambda} \\
G_+ = \sum_{x=1}^\infty \log x \ e^{-\lambda x} \frac{\lambda^x}{x!}, & \text{with prob.} = 1 - e^{-\lambda}
\end{cases} \] \hspace{1cm} (34)

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

\[ p(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, 2, ... , n \] \hspace{1cm} (35)

The geometric mean is

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

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

$$\log g = \frac{1}{n} \sum_{i=1}^{n} \log x_i, \quad \text{and} \quad g = e^{\log x}$$  \hspace{1cm} (37)

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

3.2  Case 2: negative and positive values

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

$$\log(-g_-) = \frac{n_1}{n} \log |x^-| = \frac{1}{n} \sum_{i=1}^{n_1} \log |x^-|$$  \hspace{1cm} (38)

and

$$-g_- = e^{\frac{n_1}{n} \log |x^-|} \quad \text{or} \quad -g_- = -e^{\frac{n_1}{n} \log |x^-|}$$  \hspace{1cm} (39)

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

$$\log g_+ = \frac{n_2}{n} \log x_+ = \frac{1}{n} \sum_{i=1}^{n_2} \log x_+$$  \hspace{1cm} (40)

and

$$g_+ = e^{\frac{n_2}{n} \log x_+}$$  \hspace{1cm} (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

$$g = \frac{n_1 g_- + n_2 g_+}{n}$$  \hspace{1cm} (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

$$g = \frac{n_1 g_- + n_2 g_+ + n_3 (g_0 = 0)}{n} = \frac{n_1 g_- + n_2 g_+}{n}$$  \hspace{1cm} (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 \( \theta \). If \( G_{\hat{\theta}} = \theta \), \( \hat{\theta} \) is geometrically unbiased estimator to \( \theta \).

**Example**

Let \( \hat{\theta} = g = e^{\frac{1}{n} \sum \log x} \) and \( \theta = G \), then
\[
G_{\hat{\theta}} = e^{E[\log \theta]} = e^{E[\frac{1}{n} \sum \log x]} = \frac{1}{n} \sum E[\log x] = e^{\frac{1}{n} \sum \log G} = G \tag{45}
\]
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
\[
\sigma_g^2 \approx \frac{\sigma_{\log x}^2}{n} G^2, \quad \text{and} \quad \mu_g \approx G + \frac{\sigma_{\log x}^2}{2n} G \tag{46}
\]
This can be estimated from data as
\[
s_g^2 \approx \frac{s_{\log x}^2}{n} g^2, \quad \text{and} \quad \hat{\mu}_g \approx g + \frac{s_{\log x}^2}{2n} g \tag{47}
\]
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
\[
\sigma_g^2 = \frac{\sigma_{\log |x|}^2}{n} (G)^2, \quad \text{and} \quad \mu_g \approx G + \frac{\sigma_{\log |x|}^2}{2n} G \tag{48}
\]

**Proof**

By using the delta method on \( g = -e^{\log |x|} \) 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 \( Q(u; f(Y)) = f[q(Q(u; Y))] \). 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. \( \bar{y} = \log |x| \) has approximately normal distribution for large \( n \) with \( \mu_{\bar{y}} = \mu_{\log |x|} = \log G \) and variance \( \sigma_{\bar{y}}^2 = \frac{s_{\log x}^2}{n} \), 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 $e^y$ is an increasing function and $-e^y$ 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

\begin{equation}
\left( e^{\bar{Q}_N(x; \mu, \sigma)}, e^{\bar{Q}_N(1-a; \mu, \sigma)} \right)
\end{equation}

(51)

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

\begin{equation}
\left( -e^{\bar{Q}_N(1-a; \mu, \sigma)}, -e^{\bar{Q}_N(x; \mu, \sigma)} \right)
\end{equation}

(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

$E(g) = e^{\mu} + \frac{\sigma^2}{\pi} = Ge^{\frac{\sigma^2}{2}}$, $G(g) = e^\mu = G$, Mode($g$) = $e^{\mu - \sigma^2}$ and \(\sigma_g^2 = [E(g)]^2 \left[ \frac{\sigma^2}{\pi} - 1 \right]\)

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

$E(-g) = e^{\mu - \frac{\sigma^2}{2}}$, $G(-g) = G$, Mode($-g$) = $e^{\mu - \sigma^2}$ and $\sigma_{-g}^2 = [E(-g)]^2 \left[ \frac{\sigma^2}{\pi} - 1 \right]$.

and $\mu = E(\log|x|)$ and $\sigma^2 = \sigma^2_{\log |x|}$.

Proof
Since $g$ and $-g$ have lognormal distributions with mean $\mu$ and $\sigma^2$, 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 $\hat{\sigma}_g^2$ and distributional approximation $E(g)$ and $\hat{\sigma}_g^2$. The results using Pareto distribution with different choices of $\alpha$, $\beta$ 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. $\hat{G}$ 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.

<table>
<thead>
<tr>
<th>$n$</th>
<th>$G$</th>
<th>$\hat{\mu}_g$</th>
<th>$\hat{\sigma}_g^2$</th>
<th>$E(g)$</th>
<th>$\hat{\sigma}_g^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>14.918</td>
<td>17.478</td>
<td>4.872</td>
<td>15.175</td>
<td>15.042</td>
</tr>
<tr>
<td>25</td>
<td>14.918</td>
<td>15.765</td>
<td>1.596</td>
<td>15.017</td>
<td>14.967</td>
</tr>
<tr>
<td>50</td>
<td>14.918</td>
<td>15.310</td>
<td>0.7545</td>
<td>14.957</td>
<td>14.932</td>
</tr>
<tr>
<td>100</td>
<td>14.918</td>
<td>15.128</td>
<td>0.3695</td>
<td>14.955</td>
<td>14.943</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>11.535</td>
<td>11.707</td>
<td>0.3217</td>
<td>11.558</td>
<td>11.546</td>
</tr>
<tr>
<td>25</td>
<td>11.535</td>
<td>11.602</td>
<td>0.1168</td>
<td>11.549</td>
<td>11.544</td>
</tr>
<tr>
<td>50</td>
<td>11.535</td>
<td>11.564</td>
<td>0.0558</td>
<td>11.538</td>
<td>11.536</td>
</tr>
<tr>
<td>100</td>
<td>11.535</td>
<td>11.550</td>
<td>0.0276</td>
<td>11.537</td>
<td>11.536</td>
</tr>
</tbody>
</table>
Theorem 4
Under the assumptions of theorem 3 and $X \geq 0$, the geometric mean $g = \frac{n_2}{n} e^{\frac{n_2}{n} \log x^+}$ has approximately
1. a lognormal distribution with $\mu = \frac{n_2}{n} \mu_{\log x^+} + \log \frac{n_2}{n}$ and $\sigma^2 = \left(\frac{n_2}{n}\right)^2 \sigma_{\log x^+}^2$.
2. the lower and upper $(1 - \alpha)\%$ confidence intervals for $G$ are
\[
\left( e^{\frac{\mu}{2} - \frac{1}{2} \sigma^2}, e^{\frac{\mu}{2} + \frac{1}{2} \sigma^2} \right)
\]

Proof
If $X$ is lognormal ($\mu, \sigma^2$) then $aX$ has lognormal ($\mu + \log a, \sigma^2$); see Jonson et al. (1994). Since $e^{\frac{n_2}{n} \log x^+}$ has lognormal with $\mu_{\log x^+}, \sigma^2_{\log x^+}$, 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
\[
E(g) \approx \frac{n_1 \mu_{g_-} + n_2 \mu_{g_+}}{n} \quad (53)
\]
and
\[
\sigma^2_g \approx \left(\frac{n_1}{n}\right)^2 \sigma^2_{g_-} + \left(\frac{n_2}{n}\right)^2 \sigma^2_{g_+} - 2n_1 n_2 \frac{1}{n^2} \text{Cov}(|g_-|, g_+) \quad (54)
\]
where
\[
\sigma^2_{g_-} \approx \left(\frac{n_1}{n}\right)^2 \frac{\sigma^2_{\log |x^-|}}{n_1} \left[ \mu_{g_-} \right]^2 \quad \text{and} \quad \sigma^2_{g_+} \approx \left(\frac{n_2}{n}\right)^2 \frac{\sigma^2_{\log |x^+|}}{n_2} \left[ \mu_{g_+} \right]^2,
\]

\[
\text{Cov}(|g_-|, g_+) = E \left( e^{\frac{n_1}{n} \log |x^-| + \frac{n_2}{n} \log |x^+|} \right) - E \left( e^{\frac{n_1}{n} \log |x^-|} \right) E \left( e^{\frac{n_2}{n} \log |x^+|} \right),
\]
\[
E \left( e^{\frac{n_1}{n} \log |x^-|} \right) \approx |\mu_{g_-}| + \left(\frac{n_1}{n}\right)^2 \frac{\sigma^2_{\log |x^-|}}{2n_1} |\mu_{g_-}|,
\]
\[
E \left( e^{\frac{n_2}{n} \log |x^+|} \right) \approx |\mu_{g_+}| + \left(\frac{n_2}{n}\right)^2 \frac{\sigma^2_{\log |x^+|}}{2n_2} |\mu_{g_+}|,
\]

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

and
\[
c = e^{\frac{n_1}{n} \log |x^-| + \frac{n_2}{n} \log |x^+|}
\]

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 Poisson and Pareto distributions for different choices of parameters and sample sizes are reported in Table 4. For Poisson distribution, it is shown that if \( \lambda \) 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 \( \alpha \) (very skewed distribution), and negligible for large values of \( \alpha \) and \( n \).

| Table 4 bias and RMSE for \( g \) from Poisson and Pareto distributions |
|-------------------|-----------------|-----------------|------------------|-----------------|-----------------|
| \( \lambda = 0.1, G = 0.995 \) | \( \lambda = 1.0, G = 0.789 \) | \( \lambda = 3.0, G = 2.468 \) |
| \( n \) | 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 | 0.003 | 0.0614 | 0.009 | 0.1389 |

| Table 4 bias and RMSE for \( g \) from Poisson and Pareto distributions |
|-------------------|-----------------|-----------------|------------------|-----------------|-----------------|
| \( \beta = 1, \alpha = 1.5 \) | \( \beta = 1, \alpha = 3 \) | \( \beta = 1, \alpha = 7 \) |
| \( n \) | Bias | RMSE | Bias | RMSE | Bias | RMSE |
| 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 \( \alpha \), 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 |
| \( \beta = 1, \alpha = 0.5, G = 1 \) | \( \beta = 1, \alpha = 1, G = 1 \) | \( \beta = 1, \alpha = 1, G = 1 \) |
| \( n \) | Bias | RMSE | Bias | RMSE | Bias | RMSE |
| 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 |
| \( \beta = 1, \alpha = 5, G = 1 \) | \( \beta = 1, \alpha = 10, G = 1 \) |
| \( n \) | Bias | RMSE | Bias | RMSE | Bias | RMSE |
| 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.

<table>
<thead>
<tr>
<th>Mean</th>
<th>Median</th>
<th>Geometric mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>MSE</td>
<td>MSE eff.</td>
</tr>
<tr>
<td>25</td>
<td>1.082</td>
<td>0.633</td>
</tr>
<tr>
<td>51</td>
<td>0.498</td>
<td>0.665</td>
</tr>
<tr>
<td>101</td>
<td>0.251</td>
<td>0.652</td>
</tr>
<tr>
<td>201</td>
<td>0.133</td>
<td>0.676</td>
</tr>
</tbody>
</table>

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

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

Therefore,

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

Example

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

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

and

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

The ratios of exact to approximation are

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

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 distribution

<table>
<thead>
<tr>
<th>( \sigma )</th>
<th>0.01</th>
<th>0.05</th>
<th>0.10</th>
<th>0.20</th>
<th>0.50</th>
<th>0.70</th>
<th>0.90</th>
<th>1</th>
<th>1.10</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma / \mu )</td>
<td>0.01</td>
<td>0.05</td>
<td>0.10</td>
<td>0.20</td>
<td>0.50</td>
<td>0.70</td>
<td>0.90</td>
<td>1</td>
<td>1.10</td>
</tr>
<tr>
<td>( R_1 )</td>
<td>1</td>
<td>0.9987</td>
<td>0.9950</td>
<td>0.9802</td>
<td>0.8825</td>
<td>0.7827</td>
<td>0.6670</td>
<td>0.6065</td>
<td>0.546</td>
</tr>
<tr>
<td>( R_2 )</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1.017</td>
<td>1.074</td>
<td>1.245</td>
<td>1.432</td>
<td>1.771</td>
</tr>
</tbody>
</table>

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

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

and

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

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 \(n = 25\).

<table>
<thead>
<tr>
<th>(\sigma)</th>
<th>(\mu)</th>
<th>10</th>
<th>-10</th>
<th>-50</th>
<th>10</th>
<th>8</th>
<th>-12</th>
<th>-15</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>50</td>
<td>-10</td>
<td>-49.99</td>
<td>7.99</td>
<td>10.63</td>
<td>7.99</td>
<td>12.02</td>
<td>15.01</td>
</tr>
<tr>
<td>50</td>
<td>50</td>
<td>-100.04</td>
<td>-49.99</td>
<td>50.03</td>
<td>10.01</td>
<td>8.02</td>
<td>11.98</td>
<td>15.03</td>
</tr>
<tr>
<td>50</td>
<td>0.02</td>
<td>0.1</td>
<td>0.1</td>
<td>0.20</td>
<td>1</td>
<td>0.25</td>
<td>0.833</td>
<td>0.667</td>
</tr>
<tr>
<td>50</td>
<td>49.995</td>
<td>9.962</td>
<td>-9.955</td>
<td>-49.98</td>
<td>49.04</td>
<td>6.09</td>
<td>4.48</td>
<td>-7.43</td>
</tr>
<tr>
<td>50</td>
<td>0.004</td>
<td>10.002</td>
<td>-10.004</td>
<td>-49.99</td>
<td>50.03</td>
<td>10.62</td>
<td>8000</td>
<td>-10.56</td>
</tr>
</tbody>
</table>

7 APPLICATION

7.1 Estimation of the scale parameter of log-logistic distribution

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

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

The geometric mean and median for logistic distribution are

\[
G = \beta, \quad \text{and} \quad v = \beta
\]  \( (64) \)

For known values of the shape parameter \(\alpha\) the scale parameter \(\beta\) can be estimated as

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

Table 9 shows the bias, mean square error (MSE), efficiency \((MSE(g)_\text{MSE(med)})\) and geometric bias \((g - G)\) for \(\beta\) with known values of \(\alpha\) 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 \(\alpha\) and \(n\) and number of replication is 10000.

<table>
<thead>
<tr>
<th>(\alpha)</th>
<th>(n)</th>
<th>15</th>
<th>25</th>
<th>50</th>
<th>100</th>
<th>150</th>
<th>200</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>med</td>
<td>Bias</td>
<td>7.618</td>
<td>3.823</td>
<td>1.729</td>
<td>0.828</td>
<td>0.5704</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>703.93</td>
<td>190.92</td>
<td>52.15</td>
<td>20.624</td>
<td>12.689</td>
<td>9.3597</td>
</tr>
<tr>
<td>1</td>
<td>Bias</td>
<td>5.613</td>
<td>2.966</td>
<td>1.397</td>
<td>0.6570</td>
<td>0.4542</td>
<td>0.3524</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>396.08</td>
<td>148.13</td>
<td>40.41</td>
<td>16.548</td>
<td>10.214</td>
<td>7.3461</td>
</tr>
<tr>
<td>Efficiency ((MSE_g)/MSE_{med})</td>
<td>0.562</td>
<td>0.775</td>
<td>0.775</td>
<td>0.8023</td>
<td>0.8049</td>
<td>0.785</td>
<td></td>
</tr>
<tr>
<td>Geometric Bias</td>
<td>0.0198</td>
<td>-0.460</td>
<td>-0.006</td>
<td>-0.0308</td>
<td>0.0066</td>
<td>0.0020</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>med</td>
<td>Bias</td>
<td>1.37</td>
<td>0.877</td>
<td>0.425</td>
<td>0.239</td>
<td>0.107</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>42.233</td>
<td>21.997</td>
<td>9.090</td>
<td>4.416</td>
<td>2.779</td>
<td>2.055</td>
</tr>
<tr>
<td>1</td>
<td>Bias</td>
<td>1.09</td>
<td>0.699</td>
<td>0.331</td>
<td>0.189</td>
<td>0.096</td>
<td>0.085</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>31.561</td>
<td>17.157</td>
<td>7.507</td>
<td>3.531</td>
<td>2.262</td>
<td>1.690</td>
</tr>
<tr>
<td>Efficiency ((MSE_g)/MSE_{med})</td>
<td>0.75</td>
<td>0.78</td>
<td>0.82</td>
<td>0.80</td>
<td>0.81</td>
<td>0.822</td>
<td></td>
</tr>
<tr>
<td>Geometric Bias</td>
<td>-0.010</td>
<td>-0.054</td>
<td>-0.0127</td>
<td>-0.0112</td>
<td>0.0171</td>
<td>0.0040</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>med</td>
<td>Bias</td>
<td>0.027</td>
<td>0.016</td>
<td>0.0053</td>
<td>0.002</td>
<td>0.0035</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.4213</td>
<td>0.2505</td>
<td>0.1214</td>
<td>0.0618</td>
<td>0.0411</td>
<td>0.0312</td>
</tr>
<tr>
<td>8</td>
<td>Bias</td>
<td>0.0145</td>
<td>0.0098</td>
<td>0.0051</td>
<td>0.0007</td>
<td>0.0021</td>
<td>0.0025</td>
</tr>
<tr>
<td></td>
<td>MSE</td>
<td>0.3541</td>
<td>0.2051</td>
<td>0.1029</td>
<td>0.0522</td>
<td>0.0342</td>
<td>0.0255</td>
</tr>
<tr>
<td>Efficiency ((MSE_g)/MSE_{med})</td>
<td>0.84</td>
<td>0.82</td>
<td>0.84</td>
<td>0.84</td>
<td>0.83</td>
<td>0.82</td>
<td></td>
</tr>
<tr>
<td>Geometric Bias</td>
<td>-0.030</td>
<td>-0.003</td>
<td>0</td>
<td>-0.0002</td>
<td>0.0004</td>
<td>0.0001</td>
<td></td>
</tr>
</tbody>
</table>

431
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 \(\beta\).

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 \(N^\text{th}\) 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