

Inference for the Lorenz Curve and Gini Index under the Geometric Distribution

Abdul Sathar E I

sathare@gmail.com

University of Kerala

Jolly Kumari R

Mar Ivanios College

Sreekumar N V

Government Women's College

Research Article

Keywords: Lorenz curve, Gini index, Geometric distribution, Maximum likelihood estimation, Asymptotic normality, Exact distribution, Discrete models

Posted Date: July 10th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10136585/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Inference for the Lorenz Curve and Gini Index under the Geometric Distribution

Abdul Sathar E I^{*} Jolly Kumari R[†] Sreekumar N V[‡]

Abstract

This paper develops the exact and asymptotic distributional properties of the maximum likelihood estimators of the Lorenz curve and the Gini index under the geometric distribution. Explicit forms of the estimators are obtained via the invariance property of maximum likelihood estimation. Owing to the discrete nature of the model, the resulting sampling distributions are lattice-based and exhibit non-smooth characteristics; their exact forms are derived in closed form. The estimators are shown to be strongly consistent and to converge in L^k for all $k \geq 1$. Asymptotic distributions are established using the delta method, yielding asymptotic normality with explicit variance expressions. The finite-sample performance of the estimators is examined through a Monte Carlo simulation study, and the methodology is illustrated through an application to a real dataset. The proposed results provide a rigorous inferential framework for inequality measures in discrete settings and extend classical Lorenz-based analysis to a fundamental discrete count model.

Keywords: Lorenz curve, Gini index, Geometric distribution, Maximum likelihood estimation, Asymptotic normality, Exact distribution, Discrete models.

MSC (2020): 62F10, 62F12, 62E15, 62E20, 62P05.

JEL Classification: C13, C46, D63

1 Introduction

Measures of inequality, most notably the Lorenz curve and the Gini index, play a fundamental role in a wide range of disciplines, including economics, demography, reliability theory, actuarial science, and survival analysis. The Lorenz curve provides a graphical representation of inequality by describing the cumulative proportion of a variable attributable

^{*}Department of Statistics, University of Kerala, Thiruvananthapuram, India. Corresponding author. Email: sathare@email.com

[†]Department of Statistics, Mar Ivanios College, Thiruvananthapuram, India. Email: jollykumari.r@mic.ac.in.

[‡]Department of Statistics, Government Women's College, Thiruvananthapuram, India. Email: nvsreekumar@gmail.com

to the lowest fraction of a population, whereas the Gini index offers a scalar summary based on the area between the Lorenz curve and the line of perfect equality. Owing to their interpretability and broad applicability, these measures have been extensively studied in both theoretical and applied contexts (Lorenz 1905, Gini 1912).

For a non-negative continuous random variable with distribution function F and finite mean μ , the Lorenz curve is defined as

$$L(u) = \frac{1}{\mu} \int_0^u F^{-1}(t) dt, \quad 0 \leq u \leq 1, \quad (1.1)$$

where F^{-1} denotes the quantile function. The corresponding Gini index is given by

$$G = 1 - 2 \int_0^1 L(u) du. \quad (1.2)$$

While these formulations are well established in the continuous setting, their direct application to discrete distributions requires modification. For discrete models, F^{-1} is a step function, and the integral in (1.1) does not correctly account for the probability mass at individual points. Consequently, the Lorenz curve must be defined via summation over the support rather than integration, as detailed in Section 2. The resulting curve is inherently stepwise and depends on the cumulative structure of the underlying probability mass function.

A substantial body of literature has been devoted to the estimation and inferential properties of Lorenz curves and Gini indices for continuous distributions. Exact and asymptotic properties of estimators have been derived for several parametric families, including the exponential, lognormal, and Pareto distributions (Kleiber & Kotz 2003, Gastwirth 1971). Early contributions include the work of Moothathu Moothathu (1985), who obtained the exact sampling distributions of maximum likelihood estimators of the Lorenz curve and Gini index for the exponential model. Subsequent developments encompass uniformly minimum variance unbiased estimation for the lognormal distribution Moothathu (1989) and uniform strong consistency of parametric Lorenz curve estimators Moothathu & Prasad (1991). A comprehensive treatment of the mathematical and statistical properties of Lorenz curves and Gini indices is presented in Moothathu (1991).

In parallel, related distributional functionals and inequality measures have received considerable attention. Giorgi and Crescenzi Giorgi & Crescenzi (2001) examined the Bonferroni concentration index and its relationship to the Lorenz curve in reliability contexts. Arnold Arnold (2015) provided a comprehensive treatment of Pareto-type size distributions commonly used in inequality modeling. Extensions to the Bonferroni curve and Bonferroni index under specific parametric models have been explored in (Sobhanan & Abdul Sathar 2024, Abdul Sathar & Sobhanan 2025). Despite these advances, the overwhelming majority of existing results are confined to continuous distributions.

In many modern applications, however, observations arise naturally in discrete form. Examples include counts of failures in reliability systems, insurance claim frequencies,

epidemiological incidence data, queue lengths in service systems, and discrete-time survival outcomes. In such contexts, continuous formulations are not directly applicable, and discrete analogues must be developed. Although the Lorenz curve and Gini index remain conceptually meaningful in discrete settings, their statistical properties under discrete parametric models have not been studied with comparable depth. Notable exceptions include the work of Gastwirth (1971), who provided a general definition of the Lorenz curve applicable to mixed distributions, and Kleiber and Kotz (2003), who discussed size distributions in economics including discrete cases. However, a systematic inferential theory — covering exact sampling distributions, consistency, and asymptotic normality of estimators — has not been developed for discrete parametric models.

The geometric distribution constitutes one of the most fundamental discrete models owing to its simplicity, analytical tractability, and memoryless property. Let $X \sim \text{Geom}(p)$ denote a geometric random variable with parameter $p \in (0, 1)$ and support $\{0, 1, 2, \dots\}$, representing the number of failures preceding the first success in a sequence of independent Bernoulli trials. It arises naturally in reliability theory, queuing models, insurance claim frequency analysis, and text analysis, among other fields. Despite its central role in discrete modeling, a comprehensive inferential framework for inequality measures under the geometric distribution has not been developed. To the best of our knowledge, exact and asymptotic sampling distributions of estimators of the Lorenz curve and Gini index for the geometric model have not been established in the existing literature.

The discrete nature of the geometric distribution introduces several analytical challenges that do not arise in the continuous setting. The Lorenz curve is non-smooth and exhibits a stepwise structure that reflects the lattice-based (discrete-grid) nature of the distribution. Estimators derived via the invariance property of maximum likelihood estimation consequently possess sampling distributions with non-standard, lattice-based characteristics that require a careful treatment distinct from classical continuous approaches. In particular, the piecewise nature of the Lorenz curve as a function of the parameter introduces regularity conditions for asymptotic normality that have no counterpart in continuous models, and the practical implications of these conditions are non-trivial.

The primary objective of this paper is to develop the exact and asymptotic distribution theory of the maximum likelihood estimators of the Lorenz curve and the Gini index for the geometric distribution. Explicit expressions for the estimators are obtained using the invariance property of maximum likelihood estimation, and their exact sampling distributions are derived in closed form via the negative binomial distribution of the sufficient statistic. Strong consistency and convergence in L^k for all $k \geq 1$ are established. Asymptotic normality is obtained via the delta method, with explicit asymptotic variance expressions derived in closed form. A Monte Carlo simulation study examines the finite-sample performance of the estimators and investigates the practical relevance of the regularity conditions. An application to a real dataset illustrates the methodology.

The main contributions of this work are fourfold. First, we provide a systematic inferential framework for Lorenz-based inequality measures in a discrete parametric setting,

addressing a gap in the existing literature. Second, we derive exact distributional results that reveal the distinctive effects of discreteness on estimator behavior, including the connection to the negative binomial distribution. Third, we establish asymptotic properties — including a closed-form asymptotic variance for the Gini index estimator — that support practical inference and confidence interval construction. Fourth, through simulation, we demonstrate that while the Gini index estimator achieves reliable normal approximation at moderate sample sizes, the Lorenz curve estimator requires substantially larger samples due to discreteness effects, a finding with direct practical implications for applied work.

The remainder of the paper is organized as follows. Section 2 introduces the discrete Lorenz curve and develops the MLE together with its exact and asymptotic distributional properties. Section 3 presents corresponding results for the Gini index, including the closed-form Gini expression $G(p) = 1/(2 - p)$ and its inferential consequences. Section 4 reports the Monte Carlo simulation study. Section 5 illustrates the methodology through an application to the Federalist Papers word count dataset. Section 6 provides concluding remarks.

2 MLE of the Lorenz Curve

In this section, we introduce a discrete formulation of the Lorenz curve appropriate for count data and develop the maximum likelihood estimator under the geometric distribution.

Definition 2.1. *Let X be a non-negative integer-valued random variable with probability mass function $p_x = P(X = x)$ and finite mean $\mu = E(X)$. Let $F(x) = \sum_{k=0}^x p_k$ denote the distribution function. Define the generalized inverse of F by*

$$F^{-1}(u) = \inf\{x \in \mathbb{Z}_{\geq 0} : F(x) \geq u\}, \quad 0 \leq u \leq 1. \quad (2.1)$$

The Lorenz curve of X is defined as

$$L(u) = \frac{1}{\mu} \sum_{x=0}^{F^{-1}(u)} x p_x, \quad 0 \leq u \leq 1. \quad (2.2)$$

The above definition provides a discrete analogue of the continuous Lorenz curve by replacing integration with summation and using the generalized inverse of the distribution function. Owing to the discrete nature of X , the resulting Lorenz curve is a non-decreasing step function with discontinuities at points of the form $u = F(x)$.

We now consider the geometric distribution with parameter $p \in (0, 1)$, defined on $\{0, 1, 2, \dots\}$ by

$$P(X = x) = p(1 - p)^x, \quad x = 0, 1, 2, \dots, \quad (2.3)$$

with distribution function

$$F(x) = 1 - (1 - p)^{x+1}, \quad (2.4)$$

and mean

$$\mu = \frac{1-p}{p}. \quad (2.5)$$

It can be shown that, for any integer $m \geq 0$,

$$\sum_{x=0}^m xp(1-p)^x = \frac{(1-p) [1 - (m+1)(1-p)^m + m(1-p)^{m+1}]}{p}. \quad (2.6)$$

Substituting into the definition of the Lorenz curve, we obtain

$$L(u; p) = \frac{1}{\mu} \cdot \frac{(1-p) [1 - (m+1)(1-p)^m + m(1-p)^{m+1}]}{p}, \quad (2.7)$$

where $m = F^{-1}(u)$.

Let $X_1, X_2, \dots, X_n$ be a random sample from the geometric distribution. The likelihood function is given by

$$L(p) = \prod_{i=1}^n p(1-p)^{X_i} = p^n (1-p)^{\sum_{i=1}^n X_i}. \quad (2.8)$$

The corresponding log-likelihood function is

$$\ell(p) = n \log p + \left(\sum_{i=1}^n X_i \right) \log(1-p). \quad (2.9)$$

Differentiating with respect to p and solving the likelihood equation yields the maximum likelihood estimator

$$\hat{p} = \frac{n}{n + \sum_{i=1}^n X_i} = \frac{1}{1 + \bar{X}}. \quad (2.10)$$

By the invariance property of maximum likelihood estimation Casella & Berger (2024), the estimator of the Lorenz curve is obtained as

$$\hat{L}(u) = L(u; \hat{p}).$$

We now establish the strong consistency and L^k -convergence of the proposed estimator.

Theorem 2.1. *Let $0 < u < 1$ be fixed, and suppose that $u \neq F(m; p)$ for any non-negative integer m , so that $L(u; \cdot)$ is continuous at the true parameter value p . Then, as $n \rightarrow \infty$,*

$$\hat{L}(u) \xrightarrow{a.s.} L(u; p). \quad (2.11)$$

Moreover, for every integer $k \geq 1$,

$$E \left[\left| \hat{L}(u) - L(u; p) \right|^k \right] \rightarrow 0. \quad (2.12)$$

Proof. Since $E(X) = (1-p)/p < \infty$ for all $p \in (0, 1)$, the strong law of large numbers gives

$$\bar{X} \xrightarrow{a.s.} \frac{1-p}{p}.$$

Since $g(x) = 1/(1+x)$ is continuous at $x = (1-p)/p$, the continuous mapping theorem yields

$$\hat{p} = \frac{1}{1+\bar{X}} \xrightarrow{a.s.} p.$$

Under the stated condition on u , the function $L(u; \cdot)$ is continuous at p . A further application of the continuous mapping theorem gives

$$\hat{L}(u) = L(u; \hat{p}) \xrightarrow{a.s.} L(u; p).$$

For the L^k -convergence, since $0 \leq \hat{L}(u) \leq 1$ for all n , we have

$$\left| \hat{L}(u) - L(u; p) \right|^k \leq 1,$$

providing an integrable dominating function. Since $\hat{L}(u) \xrightarrow{a.s.} L(u; p)$, it follows that

$$\left| \hat{L}(u) - L(u; p) \right|^k \xrightarrow{a.s.} 0.$$

An application of the dominated convergence theorem yields

$$E \left[\left| \hat{L}(u) - L(u; p) \right|^k \right] \rightarrow 0.$$

□

We now establish the asymptotic distribution of the proposed estimator.

Theorem 2.2. *Let $0 < u < 1$ be fixed, and suppose that the mapping $p \mapsto F^{-1}(u; p)$ is constant in a neighborhood of the true parameter value p , so that $L(u; \cdot)$ is differentiable at p . Let $m = F^{-1}(u; p)$. Then*

$$\sqrt{n} \left(\hat{L}(u) - L(u; p) \right) \xrightarrow{d} N(0, \sigma^2(u)),$$

where

$$\sigma^2(u) = m^2(m+1)^2 p^4 (1-p)^{2m-1}.$$

Proof. From standard likelihood theory for the geometric distribution, the MLE satisfies

$$\sqrt{n}(\hat{p} - p) \xrightarrow{d} N(0, p^2(1-p)).$$

Under the stated condition, $m = F^{-1}(u; p)$ is locally constant, so $L(u; p)$ is differentiable in p with

$$\frac{\partial L(u; p)}{\partial p} = m(m+1)p(1-p)^{m-1}.$$

A first-order Taylor expansion gives

$$\sqrt{n}(\hat{L}(u) - L(u; p)) = \frac{\partial L(u; p)}{\partial p} \cdot \sqrt{n}(\hat{p} - p) + o_p(1).$$

An application of the delta method yields

$$\sqrt{n}(\hat{L}(u) - L(u; p)) \xrightarrow{d} N\left(0, \left(\frac{\partial L}{\partial p}\right)^2 p^2(1-p)\right) = N(0, m^2(m+1)^2 p^4(1-p)^{2m-1}).$$

□

We next characterize the distributional structure of the estimator $\hat{L}(u)$.

Proposition 2.1. *Let $X_1, \dots, X_n$ be a random sample from the geometric distribution with parameter $p \in (0, 1)$, and let $T = \sum_{i=1}^n X_i$. For each fixed $0 < u < 1$, the estimator $\hat{L}(u)$ is a discrete random variable whose distribution is induced by the transformation*

$$\hat{L}(u) = L\left(u; \frac{n}{n+T}\right),$$

where T follows a negative binomial distribution with parameters (n, p) , having PMF

$$P(T = t) = \binom{n+t-1}{t} p^n (1-p)^t, \quad t = 0, 1, 2, \dots$$

Proof. Since $X_1, \dots, X_n$ are i.i.d. geometric(p) random variables, the sum $T = \sum_{i=1}^n X_i$ follows a negative binomial distribution with parameters (n, p) as above, taking values in $\{0, 1, 2, \dots\}$.

The MLE of p is $\hat{p} = n/(n+T)$, and by the invariance property of MLE,

$$\hat{L}(u) = L(u; \hat{p}) = L\left(u; \frac{n}{n+T}\right).$$

Thus $\hat{L}(u)$ is a measurable function of the discrete random variable T , and is therefore itself a discrete random variable with support

$$\left\{ L\left(u; \frac{n}{n+t}\right) : t = 1, 2, \dots \right\},$$

where $t = 0$ is excluded since $\hat{p} = 1$ renders the mean $\mu = (1 - \hat{p})/\hat{p} = 0$, making $L(u; \cdot)$ degenerate. The exact distribution function of $\hat{L}(u)$ is derived in Theorem 2.3 below. □

We now derive the exact distribution function of the estimator $\hat{L}(u)$ based on the distribution of the sufficient statistic.

Theorem 2.3. *Let $X_1, \dots, X_n$ be a random sample from the geometric distribution with parameter $p \in (0, 1)$, and let $T = \sum_{i=1}^n X_i$, which is sufficient for p . For any fixed $0 < u < 1$ and $y \in [0, 1]$, the distribution function of $\hat{L}(u)$ is given by*

$$P(\hat{L}(u) \leq y) = \sum_{t=t^*(y)}^{\infty} \binom{n+t-1}{t} p^n (1-p)^t,$$

where

$$t^*(y) = \inf \left\{ t \in \mathbb{Z}_{\geq 0} : L\left(u; \frac{n}{n+t}\right) \leq y \right\}.$$

For $y < L(u; 1^-)$ the probability is 0, and for $y \geq L(u; 0^+)$ it equals 1.

Proof. Since $X_1, \dots, X_n$ are i.i.d. geometric(p), the likelihood factors as $p^n(1-p)^T$, so T is sufficient for p by the factorization theorem. Moreover T follows a negative binomial(n, p) distribution with PMF

$$P(T = t) = \binom{n+t-1}{t} p^n (1-p)^t, \quad t = 0, 1, 2, \dots$$

Since $\hat{p} = n/(n+T)$ is decreasing in T , and $L(u; p)$ is increasing in p (as $\partial L/\partial p = m(m+1)p(1-p)^{m-1} \geq 0$ for $m = F^{-1}(u)$ fixed), the mapping $t \mapsto L(u; n/(n+t))$ is decreasing in t . Therefore

$$\left\{ L\left(u; \frac{n}{n+T}\right) \leq y \right\} = \{T \geq t^*(y)\},$$

and hence

$$P(\hat{L}(u) \leq y) = P(T \geq t^*(y)) = \sum_{t=t^*(y)}^{\infty} \binom{n+t-1}{t} p^n (1-p)^t.$$

□

We now obtain an explicit expression for the moments of the estimator $\hat{L}(u)$.

Proposition 2.2. *Let $X_1, \dots, X_n$ be a random sample from the geometric distribution with parameter $p \in (0, 1)$, and let $T = \sum_{i=1}^n X_i$. For any fixed $0 < u < 1$ and any integer $k \geq 1$, the k -th moment of $\hat{L}(u)$ is given by*

$$E[\hat{L}(u)^k] = \sum_{t=0}^{\infty} \left[L\left(u; \frac{n}{n+t}\right) \right]^k \binom{n+t-1}{t} p^n (1-p)^t,$$

where $L(u; 1) = 0$ by convention for the $t = 0$ term. In particular, for $k = 1$, this yields the exact expected value of $\hat{L}(u)$, from which the finite-sample bias $E[\hat{L}(u)] - L(u; p)$ can be evaluated numerically.

Proof. As established in Proposition 2.1, T follows a negative binomial(n, p) distribution with PMF

$$P(T = t) = \binom{n+t-1}{t} p^n (1-p)^t, \quad t = 0, 1, 2, \dots,$$

and $\hat{L}(u) = L(u; n/(n+T))$. By the law of total expectation,

$$E \left[\hat{L}(u)^k \right] = \sum_{t=0}^{\infty} \left[L\left(u; \frac{n}{n+t}\right) \right]^k P(T = t).$$

Substituting the PMF of T gives the stated expression.

Absolute convergence follows since $0 \leq L(u; n/(n+t))^k \leq 1$ for all t , so each term is bounded by the corresponding NB probability, and

$$\sum_{t=0}^{\infty} \left[L\left(u; \frac{n}{n+t}\right) \right]^k \binom{n+t-1}{t} p^n (1-p)^t \leq \sum_{t=0}^{\infty} \binom{n+t-1}{t} p^n (1-p)^t = 1.$$

□

We now establish the asymptotic behavior of the moments of the estimator.

Corollary 2.1. *Let $0 < u < 1$ be fixed, and suppose that the mapping $p \mapsto F^{-1}(u; p)$ is constant in a neighborhood of the true parameter value p , so that $L(u; \cdot)$ is continuous at p . Then, as $n \rightarrow \infty$, for every integer $k \geq 1$,*

$$E \left[\hat{L}(u)^k \right] \rightarrow [L(u; p)]^k.$$

Proof. By Theorem 2.1, $\hat{L}(u) \xrightarrow{a.s.} L(u; p)$ under the stated condition. By the continuous mapping theorem, since $x \mapsto x^k$ is continuous,

$$\hat{L}(u)^k \xrightarrow{a.s.} [L(u; p)]^k.$$

Since $0 \leq \hat{L}(u)^k \leq 1$ for all n , the sequence is dominated by the constant function 1, which satisfies $E[1] = 1 < \infty$. The dominated convergence theorem then yields

$$E \left[\hat{L}(u)^k \right] \rightarrow [L(u; p)]^k$$

as $n \rightarrow \infty$.

□

3 MLE of the Gini Index

In this section, we define the Gini index for discrete random variables and develop the corresponding maximum likelihood estimator under the geometric distribution.

Definition 3.1. *Let X be a non-negative integer-valued random variable with probability mass function $p_x = P(X = x)$ and finite mean $\mu = E(X) < \infty$. The Gini index of X is defined as*

$$G = \frac{1}{2\mu} \sum_{x=0}^{\infty} \sum_{y=0}^{\infty} |x - y| p_x p_y. \quad (3.1)$$

This definition represents the normalized mean absolute difference between two independent copies of X , and serves as a natural discrete analogue of the corresponding continuous formulation.

Lemma 3.1. *For $X \sim \text{Geom}(p)$ on $\{0, 1, 2, \dots\}$, the Gini index is*

$$G(p) = \frac{1}{2 - p}.$$

Proof. Using the double sum definition and $p_x = p(1 - p)^x$:

$$\begin{aligned} \sum_x \sum_y |x - y| p_x p_y &= 2 \sum_{x=1}^{\infty} \sum_{y=0}^{x-1} (x - y) p^2 (1 - p)^{x+y} \\ &= 2p^2 \sum_{x=1}^{\infty} (1 - p)^x \sum_{y=0}^{x-1} (x - y) (1 - p)^y \\ &= \frac{2(1 - p)}{p(2 - p)}. \end{aligned}$$

Dividing by $2\mu = 2(1 - p)/p$ gives $G(p) = 1/(2 - p)$. □

Let $X_1, X_2, \dots, X_n$ be a random sample from the geometric distribution, and let $T = \sum_{i=1}^n X_i$. The maximum likelihood estimator of p is given by

$$\hat{p} = \frac{n}{n + T}. \quad (3.2)$$

By the invariance property of maximum likelihood estimation Casella & Berger (2024), the estimator of the Gini index is obtained as

$$\hat{G} = G(\hat{p}) = \frac{1}{2 - \hat{p}}.$$

We now establish the strong consistency and L^k -convergence of the proposed estimator of the Gini index.

Theorem 3.1. *Let $X_1, \dots, X_n$ be a random sample from the geometric distribution with parameter $p \in (0, 1)$. Then, as $n \rightarrow \infty$,*

$$\hat{G} \xrightarrow{a.s.} G(p).$$

Moreover, for every integer $k \geq 1$,

$$E \left[|\hat{G} - G(p)|^k \right] \rightarrow 0.$$

Proof. By Theorem 2.1, $\hat{p} \xrightarrow{a.s.} p$. Since $G(p) = 1/(2-p)$ is globally continuous on $(0, 1)$ — and unlike the Lorenz curve estimator, no regularity condition on u is required — the continuous mapping theorem gives

$$\hat{G} = G(\hat{p}) \xrightarrow{a.s.} G(p).$$

By the continuous mapping theorem applied to $x \mapsto |x - G(p)|^k$,

$$|\hat{G} - G(p)|^k \xrightarrow{a.s.} 0.$$

Since $\hat{G}, G(p) \in (1/2, 1)$, we have $|\hat{G} - G(p)| < 1/2$, so

$$0 \leq |\hat{G} - G(p)|^k \leq \left(\frac{1}{2}\right)^k,$$

providing an integrable dominating function. The dominated convergence theorem yields

$$E \left[|\hat{G} - G(p)|^k \right] \rightarrow 0$$

as $n \rightarrow \infty$. □

We now derive the exact sampling distribution of the estimator $\hat{G}$.

Theorem 3.2. *Let $X_1, \dots, X_n$ be a random sample from the geometric distribution with parameter $p \in (0, 1)$, and let $T = \sum_{i=1}^n X_i$. Then*

$$\hat{G} = \frac{n + T}{2n + T}$$

is a discrete random variable taking values in $[1/2, 1)$, and for $y \in [1/2, 1)$ its distribution function is

$$P(\hat{G} \leq y) = \sum_{t=0}^{\lfloor n(2y-1)/(1-y) \rfloor} \binom{n+t-1}{t} p^n (1-p)^t, \quad (3.3)$$

with $P(\hat{G} \leq y) = 0$ for $y < 1/2$ and $P(\hat{G} \leq y) = 1$ for $y \geq 1$.

Proof. Since $X_1, \dots, X_n$ are i.i.d. geometric(p), $T = \sum_{i=1}^n X_i$ follows a negative binomial(n, p) distribution with PMF

$$P(T = t) = \binom{n+t-1}{t} p^n (1-p)^t, \quad t = 0, 1, 2, \dots$$

The estimator $\hat{G} = (n+T)/(2n+T)$ is a measurable function of T . When $T = 0$, $\hat{G} = n/(2n) = 1/2$, and as $T \rightarrow \infty$, $\hat{G} \rightarrow 1$, so $\hat{G}$ takes values in $[1/2, 1)$.

The mapping $t \mapsto (n+t)/(2n+t)$ is strictly increasing in t since its derivative with respect to t equals $n/(2n+t)^2 > 0$. Therefore, for $y \in [1/2, 1)$:

$$\left\{ \frac{n+T}{2n+T} \leq y \right\} = \left\{ T \leq \frac{n(2y-1)}{1-y} \right\} = \left\{ T \leq \left\lfloor \frac{n(2y-1)}{1-y} \right\rfloor \right\},$$

where the last equality uses the fact that T is integer-valued. Hence

$$P(\hat{G} \leq y) = P\left(T \leq \left\lfloor \frac{n(2y-1)}{1-y} \right\rfloor\right) = \sum_{t=0}^{\lfloor n(2y-1)/(1-y) \rfloor} \binom{n+t-1}{t} p^n (1-p)^t,$$

which is the incomplete negative binomial CDF evaluated at $\lfloor n(2y-1)/(1-y) \rfloor$. $\square$

We now derive the exact distribution function of the estimator $\hat{G}$ based on its representation as a function of the sufficient statistic.

Remark 1. *Since the mapping $t \mapsto (n+t)/(2n+t)$ is injective, the estimator $\hat{G}$ has a one-to-one correspondence with T , and its PMF is given explicitly by*

$$P\left(\hat{G} = \frac{n+t}{2n+t}\right) = \binom{n+t-1}{t} p^n (1-p)^t, \quad t = 0, 1, 2, \dots$$

This makes $\hat{G}$ one of the rare cases where the exact finite-sample distribution of a Gini index estimator is fully tractable.

We now derive an explicit expression for the moments of the estimator $\hat{G}$.

Proposition 3.1. *Let $X_1, \dots, X_n$ be a random sample from the geometric distribution with parameter $p \in (0, 1)$, and let $T = \sum_{i=1}^n X_i$. Then, for any integer $k \geq 1$, the k -th moment of $\hat{G}$ is given by*

$$E[\hat{G}^k] = \sum_{t=0}^{\infty} \left(\frac{n+t}{2n+t} \right)^k \binom{n+t-1}{t} p^n (1-p)^t.$$

In particular, for $k = 1$, this yields the exact expected value of $\hat{G}$, from which the finite-sample bias $E[\hat{G}] - G(p)$ can be evaluated numerically.

Proof. As established in Theorem 3.2, T follows a negative binomial(n, p) distribution with PMF

$$P(T = t) = \binom{n+t-1}{t} p^n (1-p)^t, \quad t = 0, 1, 2, \dots$$

Writing $\hat{G} = g(T)$ where $g(t) = (n+t)/(2n+t)$, the law of total expectation gives

$$E[\hat{G}^k] = \sum_{t=0}^{\infty} g(t)^k P(T = t) = \sum_{t=0}^{\infty} \left(\frac{n+t}{2n+t} \right)^k \binom{n+t-1}{t} p^n (1-p)^t.$$

Absolute convergence follows since $0 \leq g(t)^k \leq 1$ for all $t \geq 0$, so each term is bounded by the corresponding NB probability, and

$$\sum_{t=0}^{\infty} \left(\frac{n+t}{2n+t} \right)^k \binom{n+t-1}{t} p^n (1-p)^t \leq \sum_{t=0}^{\infty} \binom{n+t-1}{t} p^n (1-p)^t = 1,$$

confirming finiteness of the moment. $\square$

We now establish the asymptotic behavior of the moments of the estimator.

Corollary 3.1. *Let $X_1, \dots, X_n$ be a random sample from the geometric distribution with parameter $p \in (0, 1)$. Then, as $n \rightarrow \infty$, for every integer $k \geq 1$,*

$$E[\hat{G}^k] \rightarrow (G(p))^k.$$

No regularity condition is required, since $G(\cdot)$ is globally continuous on $(0, 1)$.

Proof. By Theorem 3.1, $\hat{G} \xrightarrow{a.s.} G(p)$. Since $x \mapsto x^k$ is continuous, the continuous mapping theorem gives

$$\hat{G}^k \xrightarrow{a.s.} (G(p))^k.$$

Since $\frac{1}{2} \leq \hat{G} < 1$ for all n , the sequence $\{\hat{G}^k\}$ is dominated by the integrable constant 1. The dominated convergence theorem yields

$$E[\hat{G}^k] \rightarrow (G(p))^k$$

as $n \rightarrow \infty$. $\square$

We now establish the asymptotic distribution of the proposed estimator using the delta method.

Theorem 3.3. *Let $X_1, \dots, X_n$ be a random sample from the geometric distribution with parameter $p \in (0, 1)$. Then, as $n \rightarrow \infty$,*

$$\sqrt{n}(\hat{G} - G(p)) \xrightarrow{d} N(0, \tau^2), \tag{3.4}$$

where

$$\tau^2 = \frac{p^2(1-p)}{(2-p)^4}. \tag{3.5}$$

Proof. The maximum likelihood estimator $\hat{p}$ satisfies the asymptotic normality

$$\sqrt{n}(\hat{p} - p) \xrightarrow{d} N(0, p^2(1 - p)),$$

which follows from standard likelihood theory for the geometric distribution.

Define $G(p) = (2 - p)^{-1}$. Then $G(\cdot)$ is continuously differentiable on $(0, 1)$, with derivative

$$G'(p) = (2 - p)^{-2}.$$

Since $\hat{G} = G(\hat{p})$, an application of the delta method yields

$$\sqrt{n}(\hat{G} - G(p)) \xrightarrow{d} N\left(0, (G'(p))^2 p^2(1 - p)\right).$$

Substituting $G'(p)$, we obtain

$$\tau^2 = (2 - p)^{-4} p^2(1 - p) = \frac{p^2(1 - p)}{(2 - p)^4}.$$

□

Remark 2. *Theorem 3.3 yields an asymptotic 95% confidence interval for $G(p)$:*

$$\hat{G} \pm \frac{1.96 \hat{\tau}}{\sqrt{n}}, \quad \text{where } \hat{\tau}^2 = \frac{\hat{p}^2(1 - \hat{p})}{(2 - \hat{p})^4},$$

with $\hat{p}$ substituted for p by the continuous mapping theorem. This provides a practical inferential tool for the Gini index in discrete settings.

4 Simulation Study

In this section, we investigate the finite-sample performance of the proposed maximum likelihood estimators of the Lorenz curve and the Gini index for the geometric distribution. Performance is evaluated in terms of bias, variance, and mean squared error (MSE). Random samples are generated from the geometric distribution with parameter values $p = 0.2, 0.4$, and 0.6 , and sample sizes $n = 20, 50, 100$, and 200 . All experiments are based on 5000 Monte Carlo replications carried out in R. The estimator $\hat{G} = 1/(2 - \hat{p})$ is evaluated against the true value $G(p) = 1/(2 - p)$. Since $G(\cdot)$ is globally smooth in p , with no regularity condition required (see Theorem 3.5), the delta method approximation is expected to perform reliably at moderate sample sizes. The asymptotic variance is τ^2/n , where

$$\tau^2 = \frac{p^2(1 - p)}{(2 - p)^4}.$$

Table 1: Finite-sample performance of the Gini index estimator $\hat{G}$ for the geometric distribution, based on 5000 Monte Carlo replications. The column τ^2/n gives the asymptotic variance from Theorem 3.5.

p	n	$G(p)$	$\text{Bias}(\hat{G})$	$\text{Var}(\hat{G})$	$\text{MSE}(\hat{G})$	τ^2/n
0.2	20	0.5556	0.0025	0.000186	0.000192	0.000152
0.2	50	0.5556	0.0010	0.000068	0.000069	0.000061
0.2	100	0.5556	0.0004	0.000033	0.000033	0.000030
0.2	200	0.5556	0.0002	0.000016	0.000016	0.000015
0.4	20	0.6250	0.0067	0.000883	0.000929	0.000732
0.4	50	0.6250	0.0024	0.000309	0.000315	0.000293
0.4	100	0.6250	0.0014	0.000156	0.000158	0.000146
0.4	200	0.6250	0.0006	0.000075	0.000075	0.000073
0.6	20	0.7143	0.0091	0.002114	0.002197	0.001874
0.6	50	0.7143	0.0034	0.000774	0.000785	0.000750
0.6	100	0.7143	0.0021	0.000391	0.000396	0.000375
0.6	200	0.7143	0.0010	0.000193	0.000194	0.000187

Table 1 shows that $\hat{G}$ performs well across all configurations. The bias is small even at $n = 20$ and decreases steadily with n , consistent with the strong consistency established in Theorem 3.1. The simulated variance closely tracks the asymptotic approximation τ^2/n , with the ratio of simulated to asymptotic variance decreasing toward 1 as n increases — from approximately 1.22 at $n = 20$ to 1.03 at $n = 200$ for $p = 0.2$, with similar patterns for other parameter values. This confirms that the asymptotic normality result of Theorem 3.5 provides a reliable basis for inference on the Gini index from moderate sample sizes onward.

The estimator $\hat{L}(u)$ is evaluated at points u satisfying the regularity condition of Theorem 2.2, namely that $p \mapsto F^{-1}(u; p)$ is locally constant in a neighborhood of the true p . Since the geometric distribution is discrete, this condition can only be satisfied over a narrow range of u for each p ; points near a jump of $F^{-1}(u; \cdot)$ lead to substantially degraded finite-sample performance. Following a systematic search, we select $u = 0.25$ for $p = 0.2$, $u = 0.45$ for $p = 0.4$, and $u = 0.65$ for $p = 0.6$. In each case the generalized inverse satisfies $m = F^{-1}(u; p) = 1$, giving the true Lorenz curve value $L(u; p) = p^2$. The asymptotic variance is $\sigma^2(u)/n$, where

$$\sigma^2(u) = m^2(m+1)^2 p^4 (1-p)^{2m-1} = 4p^4(1-p) \quad (m = 1).$$

Table 2 reveals a markedly different picture from Table 1. While $\hat{L}(u)$ is consistent — the bias and MSE both decrease with n across all configurations, confirming Theorem 2.1 — the simulated variance substantially exceeds the asymptotic approximation $\sigma^2(u)/n$, and the gap widens for larger values of p . For $p = 0.2$, the ratio of simulated to asymptotic variance decreases from 1.17 at $n = 20$ to 1.04 at $n = 200$, indicating reasonable agreement at larger samples. For $p = 0.6$, however, this ratio is 2.42 at $n = 20$ and remains as large as

Table 2: Finite-sample performance of the Lorenz curve estimator $\hat{L}(u)$ for the geometric distribution, based on 5000 Monte Carlo replications. For each p , the point u is chosen so that $F^{-1}(u; p) = 1$. The column $\sigma^2(u)/n$ gives the asymptotic variance from Theorem 2.2.

p	u	n	$L(u; p)$	$\text{Bias}(\hat{L})$	$\text{Var}(\hat{L})$	$\text{MSE}(\hat{L})$	$\sigma^2(u)/n$
0.2	0.25	20	0.0400	-0.0076	0.000299	0.000357	0.000256
0.2	0.25	50	0.0400	-0.0017	0.000153	0.000156	0.000102
0.2	0.25	100	0.0400	0.0005	0.000063	0.000063	0.000051
0.2	0.25	200	0.0400	0.0004	0.000027	0.000027	0.000026
0.4	0.45	20	0.1600	-0.0567	0.005064	0.008274	0.003072
0.4	0.45	50	0.1600	-0.0303	0.003707	0.004628	0.001229
0.4	0.45	100	0.1600	-0.0132	0.002192	0.002365	0.000614
0.4	0.45	200	0.1600	-0.0023	0.000737	0.000742	0.000307
0.6	0.65	20	0.3600	-0.1319	0.025089	0.042474	0.010368
0.6	0.65	50	0.3600	-0.0730	0.018920	0.024250	0.004147
0.6	0.65	100	0.3600	-0.0425	0.013158	0.014963	0.002074
0.6	0.65	200	0.3600	-0.0156	0.005919	0.006163	0.001037

5.71 at $n = 200$, indicating that the asymptotic regime is far from being reached at these sample sizes.

This behavior is a direct consequence of the discrete structure of the geometric distribution. Even when u is chosen to satisfy the regularity condition of Theorem 2.2, the estimator $\hat{p}$ has sufficient variability in finite samples that it frequently crosses jump points of $F^{-1}(u; \cdot)$, causing $\hat{L}(u)$ to take discrete jumps that inflate the variance beyond what the local linearization underlying the delta method can capture. The effect is more pronounced for larger p because the distribution is more concentrated near zero, making $\hat{p}$ relatively more variable with respect to the spacing of jump points.

This finding carries an important practical implication: inference on $L(u; p)$ should be based on the exact distribution derived in Theorem 2.3 rather than the asymptotic normal approximation, particularly for smaller samples or larger values of p . The exact distribution results are therefore not merely of theoretical interest but constitute the appropriate inferential tool in the discrete setting.

The simulation results provide strong support for the theoretical properties of $\hat{G}$, confirming that the asymptotic normality of Theorem 3.5 holds reliably from moderate sample sizes. For $\hat{L}(u)$, the results demonstrate consistency as established in Theorem 2.1, but reveal that the delta method approximation of Theorem 2.2 requires substantially larger samples before it becomes practically reliable, particularly for larger values of p . This contrast between the Gini index and Lorenz curve estimators highlights a fundamental distinction between continuous and discrete inequality estimation, and underscores the practical value of the exact distributional results derived in this paper.

5 Application to Real Data

In this section, the proposed estimators of the Lorenz curve and Gini index are applied to a real dataset to illustrate the practical utility of the geometric model for characterizing inequality in discrete count data.

We consider the **Federalist** dataset available in the R package **vcd** Meyer et al. (2006). The dataset records the frequency distribution of the occurrence of the word *may* across 262 blocks of text drawn from the Federalist Papers, a collection of political essays authored by Hamilton, Madison, and Jay in 1787–1788. The variable of interest X denotes the number of occurrences of the word *may* in a text block, taking values in $\{0, 1, 2, \dots\}$. The observed frequency distribution is given in Table 3.

Table 3: Observed and expected frequencies under the fitted geometric distribution for the Federalist Papers word count data ($n = 262$).

Count x	0	1	2	3	4	5	6
Observed	156	63	29	8	4	1	1
Expected	158.2	62.7	24.8	9.8	3.9	1.5	0.6

The geometric distribution is a natural model for word occurrence counts of this type, as it arises from a memoryless discrete process in which each text block independently generates occurrences with constant probability. The parameter p is estimated by maximum likelihood:

$$\hat{p} = \frac{n}{n + \sum_{i=1}^n X_i} = \frac{262}{262 + 106} = 0.6037,$$

yielding a fitted mean of $(1 - \hat{p})/\hat{p} = 0.657$ occurrences per block.

Model adequacy is assessed using the Pearson chi-square goodness-of-fit statistic, the Akaike Information Criterion (AIC), and the Bayesian Information Criterion (BIC). The results are summarized in Table 4.

Table 4: Maximum likelihood estimate, model selection criteria, goodness-of-fit statistic, and estimated Gini index for the Federalist Papers dataset under the fitted geometric distribution.

n	$\hat{p}$	AIC	BIC	χ^2	df	p -value	$\hat{G}$	95% CI
262	0.6037	584.85	588.42	1.07	3	0.7834	0.7162	(0.6926, 0.7398)

The chi-square statistic of 1.07 on 3 degrees of freedom yields a p -value of 0.783, providing no evidence against the geometric model. The close agreement between observed and expected frequencies in Table 3 further confirms adequate fit.

The estimated Gini index under the geometric model is

$$\hat{G} = \frac{1}{2 - \hat{p}} = \frac{1}{2 - 0.6037} = 0.7162,$$

with asymptotic 95% confidence interval (0.6926, 0.7398) obtained from Theorem 3.5. This indicates a high degree of concentration in word usage: a relatively small proportion of text blocks accounts for a disproportionately large share of all occurrences of the word *may*. Specifically, since $\hat{p} = 0.6037$, approximately 60% of blocks contain no occurrence of the word, while the remaining 40% account for all occurrences.

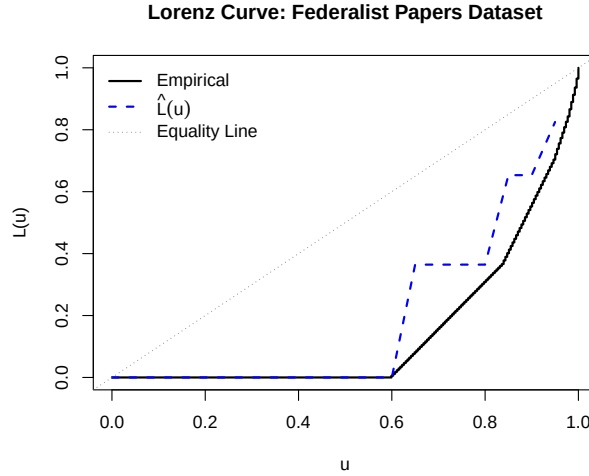


Figure 1: Empirical and fitted Lorenz curves for the Federalist Papers word count data under the geometric distribution. The dotted line represents perfect equality. Both curves lie substantially below the diagonal, reflecting the estimated Gini index of $\hat{G} = 0.7162$

The empirical and fitted Lorenz curves are displayed in Figure 1. The empirical Lorenz curve is constructed from the order statistics $X_{(1)} \leq \dots \leq X_{(n)}$ as

$$L_n\left(\frac{j}{n}\right) = \frac{\sum_{i=1}^j X_{(i)}}{\sum_{i=1}^n X_{(i)}}, \quad j = 1, \dots, n,$$

and the fitted curve $\hat{L}(u) = L(u; \hat{p})$ is obtained by substituting $\hat{p}$ into the closed-form expression derived in Section 2. Both curves lie substantially below the line of perfect equality, reflecting the high concentration indicated by $\hat{G} = 0.7162$.

A notable feature of the Lorenz curve in this setting is that $\hat{L}(u) = 0$ for $u \leq 1 - \hat{p} \approx 0.40$. This occurs because the lowest 40% of the population — the text blocks with $X = 0$ —

contribute nothing to the total count, a direct consequence of the geometric distribution placing substantial mass at zero. This behavior is a genuine feature of the discrete model and is correctly captured by the estimator $\hat{L}(u)$.

The analysis demonstrates that the geometric distribution provides an excellent fit to the Federalist Papers word count data, and that the proposed estimators yield interpretable and precise measures of concentration. The estimated Gini index of 0.7162 with a narrow 95% confidence interval reflects both the high inequality in word usage and the reliability of the asymptotic inferential framework developed in Section 3. The exact distributional results of Theorem 3.2 provide the theoretical foundation for this inference, while the asymptotic normality of Theorem 3.3 supports the construction of confidence intervals at this sample size.

6 Conclusion

This paper has developed a rigorous inferential framework for the Lorenz curve and Gini index under the geometric distribution, addressing a gap in the statistical literature on inequality measurement for discrete parametric models.

On the theoretical side, explicit maximum likelihood estimators of the Lorenz curve and Gini index were obtained via the invariance property of MLE. The exact sampling distributions of both estimators were derived in closed form through their representation as measurable functions of the sufficient statistic $T = \sum_{i=1}^n X_i$, which follows a negative binomial distribution. For the Gini index estimator, this yields a particularly tractable result: the exact distribution function admits a closed-form expression involving the incomplete negative binomial CDF, evaluated at an explicitly computable threshold. Strong consistency and L^k -convergence for all $k \geq 1$ were established via the strong law of large numbers and the dominated convergence theorem. Asymptotic normality was obtained via the delta method, with the asymptotic variance of the Gini index estimator given explicitly as

$$\tau^2 = \frac{p^2(1-p)}{(2-p)^4},$$

enabling direct construction of asymptotic confidence intervals for $G(p)$.

The simulation study revealed an important distinction between the two estimators. The Gini index estimator $\hat{G}$ performs well across all parameter configurations, with simulated variances closely tracking the asymptotic approximation τ^2/n from moderate sample sizes onward. The Lorenz curve estimator $\hat{L}(u)$, by contrast, exhibits substantially slower convergence to the asymptotic regime due to the discrete structure of the geometric distribution. Even when the evaluation point u satisfies the regularity condition of Theorem 2.2, the estimator $\hat{p}$ frequently crosses jump points of $F^{-1}(u; \cdot)$ in finite samples, inflating the variance beyond what the delta method captures. This finding carries a direct practical implication: inference on $L(u; p)$ should rely on the exact distributional results of Theo-

rem 2.3 rather than the asymptotic approximation, particularly at moderate sample sizes or larger values of p .

The application to the Federalist Papers word count dataset demonstrated the practical utility of the proposed methodology. The geometric distribution provided an excellent fit, with a chi-square goodness-of-fit p -value of 0.783. The estimated Gini index $\hat{G} = 0.7162$ with 95% confidence interval (0.6926, 0.7398) quantifies a high degree of concentration in word usage, illustrating how the geometric Lorenz curve and Gini index provide interpretable summaries of concentration in discrete count data beyond their traditional economic applications.

The results of this paper suggest that the inferential framework developed here can serve as a foundation for studying inequality measures under related discrete parametric models, and several extensions of the present work are currently under investigation.

Declarations

Competing interests. The authors have no relevant financial or non-financial interests to disclose.

Funding. No funding was received for conducting this study.

Author contributions. Abdul Sathar EI: Conceptualization, Methodology, Formal analysis, Writing – original draft, Supervision. Jolly Kumari R: Methodology, Software, Validation, Writing – review and editing. Sreekumar NV: Investigation, Data curation, Writing – review and editing.

Data availability. The dataset used in Section 5 is the `Federalist` dataset available in the R package `vcd` Meyer et al. (2006), freely accessible by running `data(Federalist)` in R. The R code used for simulation and data analysis is available from the corresponding author upon reasonable request.

Use of AI tools. The authors declare that no AI tools were used in the preparation, writing, analysis, of this manuscript. **Clinical Trial Number.** Not applicable.

Consent to Publish. Not applicable.

Consent to Participate. Not applicable.

Ethics declaration. Not applicable.

References

- Abdul Sathar, E. I. & Sobhanan, P. (2025), ‘Estimation of the bonferroni curve and bonferroni index for the power function distribution’, *International Journal of System Assurance Engineering and Management* **16**(12), 3956–3967.
- Arnold, B. C. (2015), *Pareto Distributions*, 2 edn, Chapman and Hall/CRC, Boca Raton.

- Casella, G. & Berger, R. L. (2024), *Statistical Inference*, 2 edn, Chapman and Hall/CRC., <https://doi.org/10.1201/9781003456285>.
- Gastwirth, J. L. (1971), ‘A general definition of the lorenz curve’, *Econometrica* **39**(6), 1037–1039.
- Gini, C. (1912), *Variabilità e Mutabilità*, Tipografia di Paolo Cuppini, Bologna. Reprinted in Studi Economico-Giuridici della R. Università di Cagliari, Vol. 3, Part 2.
- Giorgi, G. M. & Crescenzi, M. (2001), ‘A look at the bonferroni inequality measure in a reliability framework’, *Statistica* **61**(4), 571–583.
- Kleiber, C. & Kotz, S. (2003), *Statistical Size Distributions in Economics and Actuarial Sciences*, Wiley, New York.
- Lorenz, M. O. (1905), ‘Methods of measuring the concentration of wealth’, *Publications of the American Statistical Association* **9**(70), 209–219.
- Meyer, D., Zeileis, A. & Hornik, K. (2006), ‘The strucplot framework: Visualizing multi-way contingency tables with vcd’, *Journal of Statistical Software* **17**(3), 1–48.
- Moothathu, T. S. K. (1985), ‘Distributions of maximum likelihood estimators of lorenz curve and gini index of exponential distribution’, *Annals of the Institute of Statistical Mathematics* **37**, 483–494.
- Moothathu, T. S. K. (1989), ‘On unbiased estimation of gini index and yntema–pietra index of lognormal distribution and their variances’, *Communications in Statistics—Theory and Methods* **18**(2), 661–672.
- Moothathu, T. S. K. (1991), ‘Lorenz curve and gini index’, *Calcutta Statistical Association Bulletin* **41**, 1–24.
- Moothathu, T. S. K. & Prasad, M. S. (1991), ‘On uniformly consistent estimation of parametric lorenz curves’, *Communications in Statistics—Theory and Methods* **20**(3), 851–871.
- Sobhanan, P. & Abdul Sathar, E. I. (2024), ‘Estimation of bonferroni curve and bonferroni index of the pareto distribution’, *Stochastics and Quality Control* **39**(2), 119–129.