REVSTAT – Statistical Journal
Volume 0, Number 0, Month 0000, 000-000

LIKELIHOOD-BASED FINITE SAMPLE INFERENCE FOR SYNTHETIC DATA FROM PARETO MODEL

Authors: Nutan Mishra
– Department of Mathematics and Statistics
University of South Alabama
Mobile, AL 36688, USA (email: nmishra@southalabama.edu)

Sandip Barui
– Quantitative Methods and Operations Management Area
Indian Institute of Management Kozhikode
Kozhikode, Kerala 673570, India (email: sandipbarui@iimk.ac.in)

Received: November 2019 Revised: February 2021 Accepted: Month 0000

Abstract:

- Statistical agencies often publish microdata or synthetic data to protect confidentiality of survey respondents. This is more prevalent in case of income data. In this paper, we develop likelihood-based finite sample inferential methods for a singly imputed synthetic data using plug-in sampling and posterior predictive sampling techniques under Pareto distribution, a well known income distribution. The estimators are constructed based on sufficient statistics and the estimation methods possess desirable properties. For example, the estimators are unbiased and confidence intervals developed are exact. An extensive simulation study is carried out to analyze the performance of the proposed methods.

Key-Words:

- *Maximum likelihood estimation; plug-in sampling; posterior predictive sampling; synthetic data; exact confidence interval; pivotal quantity.*

AMS Subject Classification:

- 62H10, 62H15, 62H12, 62J05, 62F10, 62E15, 62E10, 62E17, 62D99.

1. INTRODUCTION

Collecting and publishing public data (e.g., tax data) relevant to national interests, or to human race in a broader sense, thereby establishing transparency in government policies and aiding socio-economic development have been primary objectives of many statistical organizations. However, they are also responsible for protecting survey respondents' confidentiality since that leads to greater rates and accuracy in responses. The aforementioned issues have collectively led to the origin of synthetic data where sensitive values are treated like missing values and are imputed based on the underlying data distribution. Mere elimination of the key identifiers, e.g., name, address, unique government identification number, age, etc. may not be sufficient to provide full protection to respondent's identity, and hence, additional steps should be taken to this end. Synthetic data are generated in a way that the privacy and confidentiality of general public are not compromised, however, keeping underlying structure of the stochastic model that generated the data, intact. The inferences drawn from synthetic data are expected to reveal similar characteristics as the ones obtained actual data, hence, any decisions or actions based on synthetic data remain valid. Some well known techniques in this front involve cell suppression (method of withholding values of the primary risky cells and secondary nearly-risky cells by some random mechanism; [1], [2]), data swapping (a perturbation method of creating pairs with similar attributes and interchanging sensitive values between them; [3]), top coding/bottom coding (replacing confidential values of an attribute with the maximum or minimum or some other threshold values), random noise perturbation (method of contaminating data with random noises following some known distribution and applying statistical methods to estimate the true values ignoring the noises; [4], [5], [6]) and multiple imputation (replacing sensitive values with some aggregated measure obtained from multiple imputed values by utilizing the underlying stochastic nature of the data; [7],[8]) have been implemented widely for statistical disclosure control.

In this context, application of noise perturbed data and synthetic data have gained recognition only in the recent years. Under these techniques, random errors or noises are generated from a well known probability distribution and applied on quantitative data that need to be masked, either additively or multiplicatively. Many inherent characteristics and principal features of noise-perturbed data obtained from the actual microdata in order to protect privacy were studied by [4], [5], [6], [9], [10], [11], to name a few. Recently, [12] in their paper, developed a likelihood based inferential method under the assumption of multiplicative noise where data is obtained from a parametric model.

One of the early works to implement synthetic data for statistical disclosure control was accomplished by [13] where synthetic data is generated with a concept similar to multiple imputation [7]. Multiple imputation provides a framework in which each datum is replaced by a vector of m values sampled from a known probability distribution. In [13], the author suggested that multiple-imputation technique results in synthetic data that do not resemble any actual sampling unit while preserving inherent properties of the underlying distribution and confidentiality of the respondents. Detailed parametric and non-parametric inferential methods of analyses based on synthetic data were examined by [14].

An illustration on multiply imputed fully synthetic public use microdata with respect to inferences on various descriptive and analytic estimands, and degree of protection of confidentiality, was carried out by [15]. Modified adaptations on multiple imputation based framework in context of missing data, data confidentiality and measurement error was discussed in [16]. A likelihood-based finite sample inference was studied by [17] for a synthetic data obtained from an exponential distribution. Similar studies were carried out by [18] and [19] where synthetic data are generated from a normal distribution using posterior predictive sampling and plug-in sampling methods. Further discussions and developments in synthetic data methodology could be found in [20], [21] and [22].

Following the line of work similar to [17], in this paper we develop a likelihood-based inferential procedure for synthetic data using plug-in sampling and posterior predictive sampling where the true population is a two-parameter Pareto distribution. Define $x = (x_1, \dots, x_n)^{\mathrm{T}}$ as the original microdata with a probability density function (pdf) given by $f_\theta(x)$ where θ is the parameter characterizing the underlying population. To illustrate the mechanism of plug-in sampling, let $\hat{\theta} = \hat{\theta}(x)$ be a point estimate of θ. Then, for a positive integer m, a synthetic data is given by $Y = (y_1, \dots, y_m)$ where $y_i = (y_{i1}, \dots, y_{in})^{\mathrm{T}}; i = 1, \dots, m$ is a random sample generated from $f_{\hat{\theta}}(.)$. On the other hand, posterior predictive sampling method assumes an appropriate prior distribution $\pi(\theta)$ of θ. θ^* is chosen randomly from the posterior distribution $\pi(\theta|x)$ of θ given x. A synthetic data is given by $Y = (y_1, \dots, y_m)$ where $y_i = (y_{i1}, \dots, y_{in})^{\mathrm{T}}; i = 1, \dots, m$ is a random sample generated from $f_{\theta_i^*}(.)$ where θ_i^* is the value of θ obtained by sampling from $\pi(\theta|x)$ at ith draw.

As discussed by [20], [21] and [22], for multiple imputed data sets, one may develop inference based on a scalar parameter $Q = Q(\theta)$. Let $\eta = \eta(x)$ and $\nu = \nu(x)$ be point estimator of $Q(\theta)$ and estimator of variance of η, respectively. An estimator of Q obtained from the synthetic data Y is given by

$$\bar{\eta}_m = \frac{1}{m}\sum_{i=1}^{m} \eta_i \tag{1.1}$$

and an estimator of variance of $\bar{\eta}_m$ is given by

$$V_m = \frac{1}{m(m-1)}\sum_{i=1}^{m}(\eta_i - \bar{\eta}_m)^2 + \frac{1}{m}\sum_{i=1}^{m}\nu_i \tag{1.2}$$

where $\eta_i = \eta(y_i)$ and $\nu_i = \nu(y_i)$ for $i = 1, \ldots, m$. For the upper $\gamma/2$th quantile $t_{\gamma/2;\nu}$ for a t-distribution with degrees of freedom

$$\nu = (m-1)\left[1 + \frac{(m-1)\sum_{i=1}^{m} \nu_i}{\sum_{i=1}^{m} (\eta_i - \bar{\eta}_m)^2}\right]^2,$$

an approximate interval estimate of $Q(\theta)$ can be evaluated using $\left(\bar{\eta}_m \pm t_{\gamma/2;\nu}\sqrt{V_m}\right)$.

Income data are often published by the statistical agencies as aggregates to ensure confidentiality at the cost of huge information loss. In order to circumvent this problem, these agencies use microdata in form of individual income data published synthetically. Again, Internal Revenue Service (IRS) releases tax return records of chosen individuals by masking their key identifiers because these are important source of information for policy makers, academicians or non-profit research organizations to analyze the influences of variation of tax policies on revenues or burden of tax on different social strata [23]. It is widely known that individual income can be well-modeled by Pareto distribution ([24]; [25]; [26]; [27]). The pdf of a random variable X following a Pareto distribution is given by

$$f_\theta(x) = \frac{\psi C^\psi}{x^{\psi+1}} \tag{1.3}$$

where $x > C$, C is a scale parameter that denotes minimum threshold value for x, $\psi > 0$ is a shape parameter, and $\theta = (C, \psi)^{\mathrm{T}}$. In economics, ψ is known as the Pareto index [28] which is a measure related to breadth of the income distribution.

Though synthetic or imputed data are widely used to mask income related information of individuals [29], inferential procedures for a synthetic data generated from a Pareto model have not been studied yet, to the best of our knowledge. Therefore, in this paper, we study and develop inferential methods based on likelihood function for a model-based singly imputed synthetic data using plug-in and posterior predictive sampling methods when the original data is obtained from a Pareto distribution. The formulation and derivations of the inferential methodologies are mathematically more intensive, complex and challenging in comparison to the exponential [17] or normal [19] distributions, owing to the dependency between the scale parameter C and the Pareto random variable. In particular, for posterior predictive sampling, expressions for the estimators are either implicit or their derivations are intractable. However, the estimators that could be derived are sufficient for the concerned parameter and mostly exact in nature, except few which are build based on asymptotic normality of the ML estimators. Moreover, as argued by [18], developing inferential methods based on synthetic data requires generation of m random samples of size n with $m > 1$. However, situations arise when m may not be greater than one due to stricter privacy policies or to avoid high disclosure risks [17], and only a single synthetic version of the original data is available for study. Thus, a major motivation of this work is to establish valid inferential results based on a single synthetic data by properly utilizing the underlying model structure.

The rest of the paper is arranged as follows. In section 2, discussion on methodology to estimate the parameters is provided. Section 3 deals with a simulation study which is carried out to validate the performance of our proposed method of estimation. Interpretation of the results of the simulation study are also discussed. Finally, concluding remarks are made in section 4.

2. METHODOLOGY FOR DRAWING LIKELIHOOD BASED INFERENCE

Let $X = (X_1, \ldots, X_n)^{\mathrm{T}}$ represent the original data of size n where $X_1, \ldots, X_n$ are independent and identically distributed (iid) according to Pareto distribution with a pdf given in 1.3. The maximum likelihood (ML) estimators of C and ψ are, respectively, given by $\hat{C} = X_{(1)} = \min\{X_1, \ldots, X_n\}$ and $\hat{\psi} = n\left[\sum_{i=1}^{n} \log\left(\frac{X_i}{X_{(1)}}\right)\right]^{-1}$. Note that the sampling distribution of $\hat{C}$ is Pareto with scale parameter C and shape parameter $n\psi$. On the other hand, $\hat{\psi}$ follows Inverse-Gamma (IG) distribution with parameters n and $n\psi$ when C is known, and IG distribution with parameters $n-1$ and $n\psi$ when C is unknown [30]. Moreover, $\hat{C}$ and $\hat{\psi}$ are stochastically independent ([26]; [30]). Furthermore, $\hat{C} = X_{(1)}$ is sufficient for C when ψ is known, $\left(\prod_{i=1}^{n} X_i\right)^{1/n} = Ce^{1/\hat{\psi}}$ is sufficient for ψ when C is known, and $\hat{\theta} = \left(X_{(1)}, \sum_{i=1}^{n} \log\left(\frac{X_i}{X_{(1)}}\right)\right)^{\mathrm{T}} = (\hat{C}, n/\hat{\psi})^{\mathrm{T}}$ is jointly sufficient for $\theta = (C, \psi)^{\mathrm{T}}$ when both C and ψ are unknown ([26]). Finally, $\hat{C}$ and $\hat{\psi}$ are both individually complete whereas $(\hat{C}, \hat{\psi})^{\mathrm{T}}$ is jointly complete [30]. With this background, the following results are developed for synthetic data based on plug-in sampling.

2.1. Plug-in sampling

Let $Y = (Y_1, Y_2, ..., Y_N)^{\mathrm{T}}$ be a synthetic data of size N obtained by generating a random sample from a Pareto distribution with parameters $\hat{C}$ and $\hat{\psi}$. For m multiply imputed synthetic data sets, N is generally taken as nm. However, our interest lies in the case where $m = 1$ to incorporate stricter confidentiality as mentioned earlier. Hence, assuming the value of n known, N is considered to be equal to n. Once the synthetic data $Y = (Y_1, Y_2, ..., Y_n)^{\mathrm{T}}$ is obtained, our objective is to provide inference on $\theta = (C, \psi)^{\mathrm{T}}$ based on Y. In the following subsections, we describe methodologies to draw inference on θ under three scenarios, *viz.*, inference on C when ψ is known, inference on ψ when C is known, and inference on θ when both C and ψ are unknown.

2.1.1. Inference on ψ when C known

Under this scenario, Y is generated from Pareto distribution with the value of C known. Let us define $A = C^{-n} \prod_{i=1}^{n} Y_i$.

Theorem 2.1. *For $i = 1, \ldots, n$, $y_i > C > 0$, $\psi > 0$ and C known, the pdf of Y is given by*

$$g_\psi(y) = \frac{2(\psi n)^n}{AC^n\Gamma(n)} BesselK\left(0, 2\sqrt{n\psi \log A}\right) \tag{2.1}$$

where BesselK (., .) is the modified-Bessel function of second kind defined as

$$BesselK(n, z) = \sqrt{\frac{\pi}{2z}} \frac{e^{-z}}{(n-\frac{1}{2})!} \int_0^\infty e^{-t} t^{n-1/2} \left(1 - \frac{t}{2z}\right)^{n-1/2} dt \tag{2.2}$$

for $n \in \mathbb{R}$ and $z \in \mathbb{C}$.

Proof: For $y_i > C > 0$, $i = 1, \ldots, n$, $\psi > 0$ and known C, the conditional pdf of Y given $\hat{\psi}$ is given by

$$g_1(y|\hat{\psi}) = \hat{\psi}^n C^{n\hat{\psi}} \left(\prod y_i\right)^{-\hat{\psi}-1}$$

and the conditional pdf of $\hat{\psi}$ given ψ is given by

$$g_2(\hat{\psi}|\psi) = \frac{\psi^n n^n}{\Gamma(n)} \hat{\psi}^{-n-1} \exp(-\psi n/\hat{\psi}).$$

Thus,

$$\begin{aligned} g_\psi(y) &= g_1(y|\hat{\psi}) \times g_2(\hat{\psi}|\psi) \\ &= \frac{\psi^n n^n}{\Gamma(n)} \int_0^\infty C^{n\hat{\psi}} (\prod y_i)^{-\hat{\psi}-1} \exp(-n\psi/\hat{\psi}) \hat{\psi}^{-1} d\hat{\psi} \\ &= \frac{\psi^n n^n}{\Gamma(n) C^n} \int_0^\infty \hat{\psi}^{-1} A^{-\hat{\psi}-1} \exp(-n\psi/\hat{\psi}) d\hat{\psi}. \end{aligned}$$

□

Many well known distributions can be expressed in the form of Bessel function. This special function, namely, modified Bessel function of second kind expressed in (2.2) can be computed for specified values of its argument using Mathematica version 12.2 [31].

Uniformly minimum variance unbiased estimator and exact confidence interval for ψ

As discussed in ([26]), A is sufficient for ψ and complete. Let us define

$$\tilde{\psi} = \frac{n}{\sum_{i=1}^{n} \log{(Y_i/C)}} = n[\log(A)]^{-1}. \tag{2.3}$$

$\tilde{\psi}$ is also sufficient for ψ and complete. Hence,

$$E\{\tilde{\psi}\} = E\{E\{\tilde{\psi}|\hat{\psi}\}\} = E\left\{\frac{n\hat{\psi}}{n-1}\right\} = \frac{n^2}{(n-1)^2}\psi. \tag{2.4}$$

An unbiased estimator of ψ is $\psi_u = \frac{(n-1)^2}{n^2}\tilde{\psi}$. ψ_u is also a sufficient and complete statistic. Further, *Lehmann Scheffé theorem* ([32, Chapter 6]), implies that ψ_u is the uniformly minimum variance unbiased estimator (UMVUE) of ψ ([32, Chapter 7]). The variance of ψ_u is given by

$$\begin{aligned}
V(\psi_u) &= V(E\{\psi_u|\hat{\psi}\}) + E\{V(\psi_u|\hat{\psi})\} \\
&= \left(\frac{n-1}{n}\right)^4 \left[V(E\{\tilde{\psi}|\hat{\psi}\}) + E\{V(\tilde{\psi}|\hat{\psi})\}\right] \\
&= \left(\frac{n-1}{n}\right)^4 \left[V\left(\frac{n\hat{\psi}}{n-1}\right) + E\left\{\frac{n^2\hat{\psi}^2}{(n-1)^2(n-2)}\right\}\right] \\
&= \left\{\frac{2n-3}{(n-2)^2}\right\}\psi^2. \qquad (2.5)
\end{aligned}$$

An estimate $\widehat{V(\psi_u)}$ of $V(\psi_u)$ is obtained using (2.5) by replacing ψ with $\tilde{\psi}$.

To find an exact CI for ψ, we construct a pivotal quantity based on the sufficient statistic $\tilde{\psi}$. Recall that $\tilde{\psi}$ follows IG distribution with parameters $(n, n\hat{\psi})$ when C is known. Then the conditional pdf of $\tilde{\psi}$ is

$$g_2(\tilde{\psi}|\hat{\psi}) = \frac{\hat{\psi}^n n^n}{\Gamma(n)}\tilde{\psi}^{-n-1}\exp\left(-\frac{\hat{\psi}n}{\tilde{\psi}}\right). \tag{2.6}$$

Again, the conditional pdf of $\hat{\psi}$ given ψ is given by

$$g_2(\hat{\psi}|\psi) = \frac{\psi^n n^n}{\Gamma(n)}\hat{\psi}^{-n-1}\exp\left(-\frac{\psi n}{\hat{\psi}}\right). \tag{2.7}$$

Combining (2.6) and (2.7), we obtain

$$h_\psi(\tilde{\psi}) = \frac{\psi^n n^{2n}}{[\Gamma(n)]^2}\int_0^\infty \hat{\psi}^{-1}\tilde{\psi}^{-n-1}\exp\left(-n\left[\frac{\psi}{\hat{\psi}} + \frac{\hat{\psi}}{\tilde{\psi}}\right]\right)d\hat{\psi}. \tag{2.8}$$

Taking substitution $\psi_a = \frac{\hat{\psi}}{\psi}$, (2.8) can be written as

$$h_\psi(\tilde{\psi}) = \frac{\psi^n n^{2n}}{[\Gamma(n)]^2}\tilde{\psi}^{-n-1}\int_0^\infty \psi_a^{-1}\exp\left(-n\left[\frac{1}{\psi_a}+\frac{\psi\psi_a}{\tilde{\psi}}\right]\right)d\psi_a. \tag{2.9}$$

Further, considering a transformation of variable $\tilde{\psi} \to W$ where $W = \frac{\tilde{\psi}}{\psi}$ we obtain

$$h(w) = \frac{n^{2n}}{[\Gamma(n)]^2}w^{-n-1}\int_0^\infty \psi_a^{-1}\exp\left(-n\left[\frac{1}{\psi_a}+\frac{\psi_a}{w}\right]\right)d\psi_a \tag{2.10}$$

which is independent of ψ. Hence, $W = \frac{\tilde{\psi}}{\psi} = \frac{n(\log A)^{-1}}{\psi}$ is a pivot for ψ. For a given level of significance $\gamma \in (0,1)$, we may obtain $\kappa_2 > \kappa_1 > 0$ such that

$$\int_{\kappa_1}^{\kappa_2} h(w)dw = 1-\gamma. \tag{2.11}$$

Therefore, an exact $(1-\gamma)100\%$ CI for ψ is given by

$$\left(\frac{n(\log A)^{-1}}{\kappa_2}, \frac{n(\log A)^{-1}}{\kappa_1}\right). \tag{2.12}$$

κ_1 and κ_2 are chosen such that the CI in (2.12) has the shortest length. For achieving that, we define the length of the CI in (2.12) as

$$L_\psi = n(\log A)^{-1}\left[\frac{1}{\kappa_1} - \frac{1}{\kappa_2}\right] = \tilde{\psi}\left[\frac{1}{\kappa_1} - \frac{1}{\kappa_2}\right].$$

The objective is to find κ_1 and κ_2 such that the expected value of L_ψ is minimum subject to (2.11). Applying Lagrangian multiplier technique, the Lagrangian function $L_\psi(\kappa_1, \kappa_2, \lambda)$ is obtained as

$$L_\psi(\kappa_1, \kappa_2, \lambda) = \left[\frac{1}{\kappa_1} - \frac{1}{\kappa_2}\right] + \lambda\left(H_W(\kappa_2) - H_W(\kappa_1) - (1-\gamma)\right) \tag{2.13}$$

where λ is the Lagrangian multiplier and $H_W(w) = \int_0^w h(u)du$. On taking partial derivatives of $L_\psi(\kappa_1, \kappa_2, \lambda)$ in (2.13) with respect to κ_1, κ_2 and λ we solve (2.14) for κ_1 and κ_2 where

$$\begin{aligned} \kappa_1^2 h(\kappa_1) - \kappa_2^2 h(\kappa_2) &= 0 \\ H_W(\kappa_2) - H_W(\kappa_1) - (1-\gamma) &= 0. \end{aligned} \tag{2.14}$$

Maximum likelihood estimator and asymptotic confidence interval for ψ

The ML estimator of ψ is obtained as usual by taking the partial derivative of the (2.1) and equating to zero. That is, solving

$$\psi - \frac{n}{\log A}\left(\frac{\text{BesselK}[0, 2\sqrt{n\psi\log A}]}{\text{BesselK}[1, 2\sqrt{n\psi\log A}]}\right)^2 = 0 \tag{2.15}$$

for ψ, the ML estimator $\tilde{\psi}_{syn}$ of ψ can be obtained. It is well known that under certain regularity conditions $\tilde{\psi}_{syn}$ follows an asymptotic normal distribution ([33, Chapter 6.3]) with mean ψ and variance $\sigma^2(\tilde{\psi}_{syn}) = I(\psi)^{-1}$ where $I(\psi) = -E\left[\left\{\frac{\partial^2 \log g_\psi(y)}{\partial \psi^2}\right\}\right]$ is the information at the true value of ψ. Since $\sigma^2(\tilde{\psi}_{syn})$ depends on unknown ψ, an estimate of $\sigma^2(\tilde{\psi}_{syn})$ is given by $\hat{\sigma}^2(\tilde{\psi}_{syn}) = -\left\{\frac{\partial^2 \log g_\psi(y)}{\partial \psi^2}\right\}\Big|_{\psi=\tilde{\psi}_{syn}}$ ([34, Chapter 35]). Therefore, an asymptotic $100(1-\gamma)\%$ CI for ψ is given by $\left(\tilde{\psi}_{syn} \pm z_{\gamma/2}\hat{\sigma}^2(\tilde{\psi}_{syn})\right)$.

2.1.2. Inference on C when ψ is known

Under this scenario, a synthetic data y is generated from Pareto distribution with the scale parameter $\hat{C} = X_{(1)}$ and the shape parameter as ψ. The goal is to derive inference on C based on y. Central to this goal is the joint pdf $g_C(y)$ which can be used to obtain the likelihood function $L(C|y)$. Let us define $\tilde{C} = Y_{(1)} = \min\{Y_1, \ldots, Y_n\}$ and $B = \prod_{i=1}^n y_i$.

Theorem 2.2. *The joint pdf of Y is given by*

$$g_C(y) = \frac{n\psi^{n+1}C^{n\psi}}{B^{\psi+1}} \times \log\left(\frac{\tilde{C}}{C}\right) \tag{2.16}$$

where $y_i > C > 0$ for $i = 1, \ldots, n$, $\tilde{C} > C$ and $\psi > 0$.

Proof: Note that $\tilde{C} > \hat{C} > C$. Let $g_3(y|\hat{C})$ and $g_4(\hat{C}|C)$ be the conditional pdfs of y given $\hat{C}$ and $\hat{C}$ given C respectively. Also, $g_4(\hat{C}|C)$ is Pareto with parameters C and $n\psi$. For $\tilde{C} > C$, the joint pdf of Y is expressed as

$$\begin{aligned} g_C(y) &= \int_C^{\tilde{C}} g_3(y|\hat{C}) g_4(\hat{C}|C) d\hat{C} \\ &= \int_C^{\tilde{C}} \frac{\psi^n \hat{C}^{n\psi}}{(\prod y_i)^{\psi+1}} \times \frac{n\psi C^{n\psi}}{\hat{C}^{n\psi+1}} d\hat{C} = \frac{n\psi^{n+1}C^{n\psi}}{(\prod y_i)^{\psi+1}} \int_C^{\tilde{C}} \frac{d\hat{C}}{\hat{C}} \\ &= \frac{n\psi^{n+1}C^{n\psi}}{B^{\psi+1}} \times \log\left(\frac{\tilde{C}}{C}\right). \end{aligned} \tag{2.17}$$

□

Uniformly minimum variance unbiased estimator and exact confidence interval for C

Since $\tilde{C} = Y_{(1)}$ is a complete sufficient statistic for C when ψ is known, $C_u = \frac{(n\psi-1)^2}{(n\psi)^2}\tilde{C}$ is an unbiased estimator of C as shown below.

$$E\left\{\tilde{C}_u\right\} = E\left\{E\left\{\tilde{C}_u|\hat{C}\right\}\right\} = \frac{(n\psi-1)^2}{(n\psi)^2}E\left\{E\left\{\tilde{C}|\hat{C}\right\}\right\} = \frac{(n\psi-1)^2}{(n\psi)^2}E\left\{\frac{n\psi\hat{C}}{n\psi-1}\right\} = C. \tag{2.18}$$

By *Lehmann Scheffé theorem* ([32, Chapter 6]), $\tilde{C}_u$ is the UMVUE of C when ψ is known. The variance of $\tilde{C}_u$ is given by

$$\begin{aligned}
V\left\{\tilde{C}_u\right\} &= V\left\{E\left\{\tilde{C}_u|\hat{C}\right\}\right\} + E\left\{V\left\{\tilde{C}_u|\hat{C}\right\}\right\} \\
&= \frac{(n\psi-1)^4}{(n\psi)^4} \times \left[V\left\{E\left\{\tilde{C}|\hat{C}\right\}\right\} + E\left\{V\left\{\tilde{C}|\hat{C}\right\}\right\}\right] \\
&= \frac{(n\psi-1)^4}{(n\psi)^4} \times \left[V\left\{\frac{n\psi\hat{C}}{n\psi-1}\right\} + E\left\{\frac{n\psi\hat{C}^2}{(n\psi-1)^2(n\psi-2)}\right\}\right] \\
&= \left\{2 - \frac{1}{(n\psi-1)^2}\right\}\frac{C^2}{(n\psi)^2}. \qquad (2.19)
\end{aligned}$$

The development of an exact confidence interval for C involves construction of a pivot for C from its sufficient statistic $\tilde{C} = Y_{(1)}$. For $\tilde{C} > C$, the pdf of $\tilde{C}$ is given by

$$\begin{aligned}
h_C(\tilde{C}) &= \int_C^{\tilde{C}} g_4(\tilde{C}|\hat{C})g_4(\hat{C}|C)d\hat{C} = \int_C^{\tilde{C}} \frac{n\psi\hat{C}^{n\psi}}{\tilde{C}^{n\psi+1}} \times \frac{n\psi C^{n\psi}}{\hat{C}^{n\psi+1}}d\hat{C} \\
&= \frac{n^2\psi^2C^{n\psi}}{\tilde{C}^{n\psi+1}} \times \log\left(\frac{\tilde{C}}{C}\right). \qquad (2.20)
\end{aligned}$$

Let $T = \log\left(\frac{\tilde{C}}{C}\right)$, then the pdf of T as

$$\tilde{h}(t) = n^2\psi^2te^{-n\psi t}, \text{ for } t > 0$$

and $\tilde{h}(t)$ is independent of C. For some $\kappa_2 > \kappa_1 \geq 1$ and $\gamma \in (0,1)$, we obtain

$$\int_{\kappa_1}^{\kappa_2} \tilde{h}(t)dt = 1 - \gamma.$$

Therefore, an exact $100(1-\gamma)\%$ CI for C is given by $\left(\tilde{C}e^{-\kappa_2}, \tilde{C}e^{-\kappa_1}\right)$. Define $\tilde{H}_C(c) = \int_0^c \tilde{h}(u)du$. Following the steps as discussed in section 2.1.1, the shortest length $100(1-\gamma)\%$ for C is obtained by solving

$$\begin{aligned}
e^{\kappa_1}\tilde{h}(\kappa_1) - e^{\kappa_2}\tilde{h}(\kappa_2) &= 0 \\
\tilde{H}_C(\kappa_2) - \tilde{H}_C(\kappa_1) - (1-\gamma) &= 0 \qquad (2.21)
\end{aligned}$$

for κ_1 and κ_2.

Maximum likelihood estimation of C

The usual method of derivative based ML estimation cannot be applied here to obtain the ML estimate of C. However, noting

$$\frac{\partial L(C|y)}{\partial C} = \frac{\partial g_C(y)}{\partial C} = -\frac{n\psi^{n+1}C^{-(n+1)}}{A^{\psi+1}}\left\{1+\log\left(\frac{\tilde{C}}{C}\right)\right\} < 0,$$

i.e., $L(C|y)$ is decreasing in C with $0 < C < \tilde{C}$, the ML estimator of C is obtained as $\tilde{C} = Y_{(1)}$. The exact distribution of $\tilde{C}$ is given by equation (2.20). An estimate of the variance of $\tilde{C}$ can be derived from (2.19) as

$$\widehat{V(\tilde{C})} = \left[\frac{(n\psi)^2}{(n\psi-1)^4}\left\{2-\frac{1}{(n\psi-1)^2}\right\}\right]\tilde{C}^2.$$

2.1.3. Inference on $\theta = (C, \psi)^{\mathrm{T}}$ when both C and ψ are unknown

To develop inference on θ, the joint pdf of y given θ in Theorem 2.3, where $y = (y_1, \ldots, y_n)^{\mathrm{T}}$ is a synthetic data obtained from Pareto distribution with parameters $\hat{C}$ and $\hat{\psi}$. Define $\psi^* = \frac{n}{\sum_{i=1}^n \log(Y_i/Y_{(1)})}$. ψ^* follows IG with parameters $n-1$ and $n\hat{\psi}$.

Theorem 2.3. *The joint pdf of Y is given by*

$$g_\theta(y) = \frac{(n\psi)^n C^{n\psi}}{\Gamma(n-1)}\int_0^\infty \left(\frac{\tilde{C}^{n(\hat{\psi}-\psi)} - C^{n(\hat{\psi}-\psi)}}{n(\hat{\psi}-\psi)}\right)\frac{\exp\left\{-\frac{n\psi}{\hat{\psi}}\right\}}{\left(\prod_{i=1}^n y_i\right)^{\hat{\psi}+1}}d\hat{\psi} \tag{2.22}$$

where $\tilde{C} = \min\{y_1, \ldots, y_n\} > C > 0$ and $\psi > 0$.

Proof: The conditional pdf of y given $\hat{\theta} = (\hat{C}, \hat{\psi})^{\mathrm{T}}$ is expressed as

$$g_5(y|\hat{\theta}) = \frac{\hat{\psi}^n \hat{C}^{n\hat{\psi}}}{\left(\prod_{i=1}^n y_i\right)^{\hat{\psi}+1}} \tag{2.23}$$

where $y_i > \hat{C} > C > 0$ for $i = 1, \ldots, n$ and $\hat{\psi} > 0$. Again, the conditional pdf of $\hat{\theta}$ given θ is

$$g_6(\hat{\theta}|\theta) = \frac{(n\psi)^n C^{n\psi}\exp\left\{-\frac{n\psi}{\hat{\psi}}\right\}}{\hat{C}^{n\psi+1}\hat{\psi}^n\Gamma(n-1)} \tag{2.24}$$

for $0 < C < \hat{C} < \tilde{C}$, $\hat{\psi} > 0$ and $\psi > 0$. Equation (2.24) is obtained using the fact that $\hat{C}$ and $\hat{\psi}$ are stochastically independent where $\hat{C}$ follows Pareto distribution with scale C and shape $n\psi$, and $\hat{\psi}$ follows IG distribution with parameters $n-1$ and $n\psi$. Finally, the pdf of y is given by

$$\begin{aligned} g_\theta(y) &= \int_0^\infty \int_C^{\tilde{C}} g_5(y|\hat{\theta}) \times g_6(\hat{\theta}|\theta) d\hat{C} d\hat{\psi} \\ &= \frac{(n\psi)^n C^{n\psi}}{\Gamma(n-1)} \int_C^{\tilde{C}} \left(\hat{C}^{n(\hat{\psi}-\psi)-1}\right) \int_0^\infty \frac{\exp\left\{-\frac{n\psi}{\hat{\psi}}\right\}}{\left(\prod_{i=1}^n y_i\right)^{\hat{\psi}+1}} d\hat{\psi} d\hat{C} \end{aligned} \tag{2.25}$$

which can be further simplified to

$$g_\theta(y) = \frac{(n\psi)^n C^{n\psi}}{\Gamma(n-1)} \int_0^\infty \left(\frac{\tilde{C}^{n(\hat{\psi}-\psi)} - C^{n(\hat{\psi}-\psi)}}{n(\hat{\psi}-\psi)} \right) \frac{\exp\left\{-\frac{n\psi}{\hat{\psi}}\right\}}{\left(\prod_{i=1}^n y_i\right)^{\hat{\psi}+1}} d\hat{\psi}. \tag{2.26}$$

□

Construction of a pivot for θ

Let us define $\tilde{\theta} = (\tilde{C}, \psi^*)^{\mathrm{T}}$. The pdf of $\tilde{\theta}$ is given by

$$\begin{aligned} h_\theta(\tilde{\theta}) &= \int_0^\infty \int_C^{\tilde{C}} g_6(\tilde{\theta}|\hat{\theta}) \times g_6(\hat{\theta}|\theta) d\hat{C} d\hat{\psi} \\ &= \int_0^\infty \int_C^{\tilde{C}} \frac{(n\hat{\psi})^n \hat{C}^{n\hat{\psi}} \exp\left\{-\frac{n\hat{\psi}}{\psi^*}\right\}}{\tilde{C}^{n\hat{\psi}+1} \psi^{*n} \Gamma(n-1)} \times \frac{(n\psi)^n C^{n\psi} \exp\left\{-\frac{n\psi}{\hat{\psi}}\right\}}{\hat{C}^{n\psi+1} \hat{\psi}^n \Gamma(n-1)} d\hat{C} d\hat{\psi} \\ &= \frac{n^{2n} C^{n\psi} \psi^n}{\{\Gamma(n-1)\}^2 \psi^{*n}} \int_0^\infty \int_C^{\tilde{C}} \frac{\hat{C}^{n(\hat{\psi}-\psi)-1}}{\tilde{C}^{n\hat{\psi}+1}} \exp\left[-n\left\{\frac{\hat{\psi}}{\psi^*} + \frac{\psi}{\hat{\psi}}\right\}\right] d\hat{C} d\hat{\psi}. \end{aligned} \tag{2.27}$$

Substituting $t = \frac{\hat{\psi}}{\psi}$, we obtain

$$\begin{aligned} h_\theta(\tilde{\theta}) &= \frac{n^{2n} C^{n\psi} \psi^n}{\{\Gamma(n-1)\}^2 \psi^{*n}} \int_0^\infty \int_C^{\tilde{C}} \frac{\hat{C}^{n\psi(t-1)-1}}{\tilde{C}^{n\psi t+1}} \exp\left[-n\left\{\frac{\psi t}{\psi^*} + \frac{1}{t}\right\}\right] d\hat{C} \times \psi dt \\ &= \frac{n^{2n} C^{n\psi} \psi^{n+1}}{\{\Gamma(n-1)\}^2 \psi^{*n}} \int_0^\infty \frac{1}{\tilde{C}^{n\psi t+1}} \exp\left[-n\left\{\frac{\psi t}{\psi^*} + \frac{1}{t}\right\}\right] \times \int_C^{\tilde{C}} \hat{C}^{n\psi(t-1)-1} d\hat{C} dt \\ &= \frac{n^{2n} C^{n\psi} \psi^{n+1}}{\{\Gamma(n-1)\}^2 \psi^{*n}} \int_0^\infty \frac{1}{\tilde{C}^{n\psi t+1}} \exp\left[-n\left\{\frac{\psi t}{\psi^*} + \frac{1}{t}\right\}\right] \times \left[\frac{\tilde{C}^{n\psi(t-1)} - C^{n\psi(t-1)}}{n\psi(t-1)}\right] dt. \end{aligned} \tag{2.28}$$

Considering a bivariate transformation $(\tilde{C}, \psi^*) \to (U, V)$ where

$$U = \left(\frac{\tilde{C}}{C}\right)^\psi \text{ and } V = \frac{\psi^*}{\psi} \tag{2.29}$$

we obtain pdf of (U, V) which is independent of θ. The Jacobian of the transformation is $Cu^{\frac{1}{\psi}-1}$. From (2.28), the joint pdf of (U, V) is

(2.30)

$$h_{U,V}(u,v) = \frac{n^{2n-1}}{\{\Gamma(n-1)\}^2 v^n} \int_0^\infty \exp\left[-n\left\{\frac{t}{v}+\frac{1}{t}\right\}\right] \times \left[\frac{u^{n(t-1)}-1}{u^{nt+1}(t-1)}\right] dt, u > 1 \text{ and } v > 0$$

which is independent of θ. The marginal pdfs of U and V are obtained from (2.30) as follows:

$$h_U(u) = \frac{n^n}{\Gamma(n-1)} \int_0^\infty \frac{\{u^{n(t-1)}-1\}\exp(-n/t)}{u^{nt+1}t^{n-1}(t-1)} dt, u > 1 \tag{2.31}$$

and

$$h_V(v) = \frac{n^{2n-2}}{\{\Gamma(n-1)\}^2 v^n} \int_0^\infty t^{-1} \exp\left[-n\left\{\frac{t}{v}+\frac{1}{t}\right\}\right] dt, v > 0. \tag{2.32}$$

The marginal cdfs U and V are, respectively, $H_U(u) = \int_0^u h_U(a)da$ and $H_V(v) = \int_0^v h_V(a)da$. Now, we proceed as follows.

Estimation of ψ when C is unknown

The expected value of ψ^* is derived as

$$E\{\psi^*\} = E\{E\{\psi^*|\hat{\theta}\}\} = E\left\{\frac{n\hat{\psi}}{n-2}\right\} = \frac{n^2}{(n-2)^2}\psi. \tag{2.33}$$

Hence, an unbiased estimator ψ_u^* of ψ is $\frac{(n-2)^2}{n^2}\psi^*$. The variance of ψ_u^* is

$$V(\psi_u^*) = V(E\{\psi_u^*|\hat{\psi}\}) + E\{V(\psi_u^*|\hat{\psi})\} = \frac{(2n-5)}{(n-3)^2}\psi^2. \tag{2.34}$$

An estimate $\widehat{V(\psi_u^*)}$ of $V(\psi_u)$ is obtained by replacing ψ with ψ^* in (2.34). Mimicking steps in section 2.1.1, a $100(1-\gamma)\%$ CI for ψ has the following form:

$$\left(\frac{\psi^*}{\kappa_2}, \frac{\psi^*}{\kappa_1}\right) \tag{2.35}$$

where κ_1 and κ_2 are the roots of

$$\begin{aligned} \kappa_1^2 h_V(\kappa_1) - \kappa_2^2 h_V(\kappa_2) &= 0 \\ H_V(\kappa_2) - H_V(\kappa_1) - (1-\gamma) &= 0. \end{aligned} \tag{2.36}$$

Estimation of C when ψ is unknown

For $C < \tilde{C}$, we derive the marginal pdf of $\tilde{C}$ from (2.28) as

$$\begin{aligned} q_\theta(\tilde{C}) &= \int_0^\infty h_\theta(\tilde{\theta}) d\psi^* \\ &= \frac{n^n C^{n\psi}\psi}{\Gamma(n-1)} \int_0^\infty \frac{\exp\{-n/t\}}{(t-1)t^{n-1}} \times \left[\frac{\tilde{C}^{n\psi(t-1)} - C^{n\psi(t-1)}}{\tilde{C}^{n\psi t+1}} \right] dt. \end{aligned} \tag{2.37}$$

Note that U in (2.29) is not independent of ψ. Hence, in an effort to construct CI for C, we further take the transformation:

$$W^* = V \log U = \psi^* \log \frac{\tilde{C}}{C}$$

where the pdf of W^* is

$$h_{W^*}(w^*) = \frac{n^{(n-1)}(n-1)}{\Gamma(n-1)} \int_0^\infty \frac{\exp(-n/t)}{(t-1)} [(t+w^*)^{-n} - (t(w^*+1))^{-n}] dt \tag{2.38}$$

for $w^* > 0$. Therefore, a $100(1-\gamma)\%$ CI for C is calculated using the following:

$$\left(\tilde{C} \exp\{-\kappa_2/\psi^*\}, \tilde{C} \exp\{-\kappa_1/\psi^*\} \right). \tag{2.39}$$

κ_1 and κ_2 are calculated from $\int_0^{\kappa_1} h_{W^*}(w^*) dw^* = \gamma/2$ and $\int_{\kappa_2}^\infty h_{W^*}(w^*) dw^* = \gamma/2$.

2.2. Posterior Predictive Sampling

This is the second method of sampling to draw synthetic data based on original data. Under a Bayesian setting, the synthetic data $z = (z_1, \ldots, z_n)^{\mathrm{T}}$ comes from the posterior predictive distribution of θ given x. Here, we discuss the method of drawing inference on ψ when C is known.

2.2.1. Inference on ψ when C is known

We utilize the fact that the posterior distribution of ψ given U is Gamma with parameters $(n + c_0, u + d)$ as given by [35]. Here, $c_0 > 0$ and $d > 0$ are the hyper parameters obtained using a Gamma prior with parameters c_0 and d, and $U = \sum_{i=1}^n \log(X_i/C)$. Below we discuss the procedure for posterior predictive sampling.

Step 1: Draw ψ^* from the posterior distribution of ψ given u.

Step 2: Given value of ψ^* in Step 1, draw $z = z_1, ..., z_n$ as iid from the Pareto density $f_\theta(z_i) = \frac{\psi^* C^{\psi^*}}{z_i^{\psi^*+1}}$.

For the purpose of analysis based on z, we develop the joint pdf of z in Theorem 2.4. In order to prove the theorem, the following three facts are used.

- $z_i|\psi^*, i = 1, \ldots, n$ are iid with each following Pareto distribution with parameters C and ψ^*.
- $\psi^*|u$ follows Gamma distribution with parameters $(n + c_0, u + d)$.
- $U|\psi$ is Gamma distribution with parameters (n, ψ).

Theorem 2.4. *The joint pdf of z is given by*

$$f_\psi(z) = \frac{\psi^n}{\Gamma n \Gamma(n + c_0) \left(\prod_{i=1}^n z_i\right)}$$

(2.40)

$$\times \int_0^\infty \left[\int_0^\infty \psi^{*(2n+c_0-1)} \left\{ \frac{c^n}{(\prod_{i=1}^n z_i)} e^{-(u+d)} \right\}^{\psi *} d\psi^* \right] u^{n-1} (u + d)^{n+c_0} e^{-u\psi} du$$

where $\psi > 0$ and $z_i > C, i = 1, \ldots, n$.

Proof: The above theorem can be proved by considering

$$f_\psi(z) = \int_0^\infty \int_0^\infty f(z|\psi^*) \times f(\psi^*|u) \times f(u|\psi) d\psi^* du$$

where f denotes the corresponding pdfs as usual. □

Define $\tilde{\psi} = \dfrac{n}{\sum \log(Z_i/C)}$ as an estimator of ψ and $\tilde{\psi}|\psi^*$ follows IG distribution with parameters n and $n\psi^*$. The expected value of $\tilde{\psi}$ is obtained as

$$\begin{aligned} E\{\tilde{\psi}\} &= E\{E\{\tilde{\psi}|\psi^*\}\} = E\left\{ \frac{n}{(n-1)} \psi^* \right\} = \frac{n}{(n-1)} E\left\{ \frac{n + c_0}{u + d} \right\} \\ &= \frac{n(n + c_0)\psi^n}{(n-1)\Gamma n} \int_0^\infty \frac{u^{n-1} e^{-u\psi}}{(u + d)} du = \frac{n(n + c_0)\psi^n}{(n-1)\Gamma n} M_1(\psi, n, d) \end{aligned}$$

where the term

$$M_1(\psi, n, d) = \int_0^\infty \frac{u^{n-1} e^{-u\psi}}{(u + d)} du.$$

Further, the variance of $\tilde{\psi}$ is computed follows:

$$V(\tilde{\psi}) = V(E\{\tilde{\psi}|\psi^*\}) + E\{V(\tilde{\psi}|\psi^*)\}$$

where

$$\begin{aligned} V(E\{\tilde{\psi}|\psi^*\}) &= \frac{n^2(n+c_0)(n+c_0+1)\psi^n}{(n-1)^2(n-2)\Gamma n}\int_0^\infty \frac{u^{n-1}\exp(-u\psi)}{(u+d)^2}du \\ &= \frac{n^2(n+c_0)(n+c_0+1)\psi^n}{(n-1)^2(n-2)\Gamma n}M_2(\psi,n,d) \end{aligned}$$

with

$$M_2(\psi,n,d) = \int_0^\infty \frac{u^{n-1}e^{-u\psi}}{(u+d)^2}du,$$

and

$$E\{V(\tilde{\psi}|\psi^*)\} = \frac{n^2(n+c_0)\psi^n}{(n-1)^2\Gamma n}\left[(n+c_0+1)M_2(\psi,n,d) - \frac{(n+c_0)}{\Gamma n}\psi^n M_1^2(\psi,n,d)\right].$$

Hence, we can express the variance of $\tilde{\psi}$ as

(2.41)

$$V(\tilde{\psi}) = \frac{n^2(n+c_0)\psi^n}{(n-1)^2\Gamma n}\left[\frac{(n-1)(n+c_0+1)}{(n-2)}M_2(\psi,n,d) - \frac{(n+c_0)}{\Gamma n}\psi^n M_1^2(\psi,n,d)\right].$$

Shortest confidence interval for ψ

Applying the same concept used in Theorem 2.4 and considering $\tilde{\psi}|\psi^*$ follows IG distribution with parameters n and $n\psi^*$, the pdf of $\tilde{\psi}$ is given by

$$f_\psi(\tilde{\psi}) = \frac{n^n\psi^n\Gamma(2n+c_0)}{(\Gamma n)^2\Gamma(n+c_0)\tilde{\psi}^{n+1}}\int_0^\infty \frac{u^{n-1}e^{-u\psi}(u+d)^{n+c_0}}{(n/\tilde{\psi}+u+d)^{2n+c_0}}du. \tag{2.42}$$

For computational convenience, we consider $d = 0$ which leads to the prior density of the parameter ψ to be a Jeffreys prior. However, the posterior density of ψ still follows a Gamma distribution. For a detailed discussion, refer to section 2.1 in [35]. Henceforth, we assign $d = 0$. For $\omega > 0$, considering the transformations $t = u\left[\frac{n}{\tilde{\psi}} + u\right]^{-1}$ and $\omega = \dfrac{\tilde{\psi}}{\psi}$ sequentially in (2.42), we get the pdf

$$f_W(\omega) = \frac{n^n\Gamma(2n+c_0)}{(\Gamma n)^2\Gamma(n+c_0)\omega^{n+1}}\int_0^1 \frac{t^{2n+c_0-1}\exp\{-\frac{1}{\omega}[\frac{n-t}{1-t}])\}}{t-1}dt \tag{2.43}$$

independent of ψ. Hence, ω is a pivotal quantity and the shortest distance $(1-\gamma)100\%$ CI for ψ is

$$\left(\tilde{\psi}\omega_2^{-1}, \tilde{\psi}\omega_1^{-1}\right) \tag{2.44}$$

where ω_1 and ω_2 are obtained by solving

$$\begin{aligned} &\omega_1^2 f_W(\omega_1) - \omega_2^2 f_W(\omega_2) = 0 \\ &F_W(\omega_2) - F_W(\omega_1) - (1-\gamma) = 0 \end{aligned} \tag{2.45}$$

and $F_W(\omega) = \int_0^{\omega} f_W(u)du$. The discussion on constructing the shortest CI can be found in section 2.1.1.

Remark: *In practice, it is unrealistic to assume that the shape parameter ψ is known and C is not known. Once we have data then the minimum value in the data is sufficient for C. As per [35] the posterior distribution of C given the original data is a power function distribution with two hyper parameters namely $\delta \geq 0$ and $\sigma_0 > 0$. One of the parameters of the posterior distribution of C depends on $min\{\sigma_0, x_{(1)}\}$. While computing the unconditional pdf of $\tilde{C}$, an explicit expression could not be obtained since the integrals involved in the derivation often have limits depending on the original data x. Hence we do not discuss this case here. On the other hand, the case of joint posterior distribution when both parameters are unknown, becomes extremely complex due to the same issue, and hence, it is not discussed either in this paper.*

3. SIMULATION STUDY AND RESULTS

To study the performance of the proposed estimation methods, we carry out an extensive simulation study. For all scenarios, *viz.* only ψ unknown (Scenario 1), only C unknown (Scenario 2), both C and ψ unknown (Scenario 3) in case of plug-in sampling, and only ψ unknown in case of posterior predictive sampling (Scenario 4), few candidate true values of C and ψ are chosen. True values of C are taken as 1 and 100, while true values of ψ are selected to be 1.5 and 3. To study the effect of smaller and larger sample sizes on estimation, $n = 50$ and $n = 100$ are considered. Under these parameter settings, we examine the performance and robustness of our estimation methods with respect to singly imputed synthetic data based on one thousand Monte-Carlo simulation runs. Mathematica 12.2 and R-4.0.1 [36] software packages are employed for coding.

For these settings, parameter estimate (EST), empirical standard error (ESE), average model based standard error (ASE), average bias of the estimator (BIAS), average root mean squared error (RMSE), and coverage rate (CR) of 90 % and 95% nominal level are provided in Tables 1 - 6. Based on simulation results, estimates are found to be accurate with low bias and low standard errors in all cases. As one would expect, increasing sample size results in more precise estimates with improved coverage probabilities, and with noticeable reduction in BIAS, ASE and RMSE. Estimates are less precise for estimating C when ψ is unknown, and for estimating ψ when C is unknown than their corresponding known counterparts. This can be attributed to the fact that estimating associated parameter instead of using their known values introduces more variability to the data, resulting in less accuracy in estimation of the primary parameter. ASE and RMSE obtained for estimating C are high when the true value of $C = 100$ than when the true value of $C = 1$. A similar trend is observed for estimating ψ

as well; ASE and RMSE are high when true $\psi = 3$ as compared to the case when true $\psi = 1.5$.

c	ψ	n	$1-\gamma$	EST	ESE	ASE	BIAS	RMSE	ECR
1	1.5	50	0.90	1.001	0.020	0.019	-0.001	0.020	0.893
1	1.5	50	0.95	1.000	0.019	0.019	0.000	0.019	0.948
1	1.5	100	0.90	0.999	0.008	0.009	0.001	0.009	0.911
1	1.5	100	0.95	1.001	0.009	0.009	-0.001	0.010	0.948
1	3.0	50	0.90	1.000	0.009	0.009	0.000	0.009	0.892
1	3.0	50	0.95	1.000	0.009	0.009	0.000	0.009	0.946
1	3.0	100	0.90	1.000	0.004	0.004	0.000	0.005	0.899
1	3.0	100	0.95	1.000	0.004	0.004	0.000	0.004	0.948
100	1.5	50	0.90	99.950	1.818	1.962	0.050	2.673	0.913
100	1.5	50	0.95	100.011	1.948	1.963	-0.011	2.764	0.948
100	1.5	100	0.90	99.950	0.917	0.961	0.050	1.326	0.904
100	1.5	100	0.95	100.011	0.982	0.962	-0.011	1.371	0.952
100	3.0	50	0.90	100.044	0.968	0.962	-0.044	1.363	0.892
100	3.0	50	0.95	99.992	0.941	0.961	0.008	1.345	0.949
100	3.0	100	0.90	100.020	0.501	0.476	-0.020	0.685	0.892
100	3.0	100	0.95	100.009	0.475	0.476	-0.009	0.670	0.951

Table 1: EST, ASE, ESE, BIAS, RMSE, ACR and ECR for C when ψ is known.

c	ψ	n	$1-\gamma$	EST	ESE	ASE	BIAS	RMSE	ECR
1	1.5	50	0.90	1.001	0.018	0.019	0.001	0.026	0.909
1	1.5	50	0.95	1.002	0.019	0.019	0.002	0.027	0.954
1	1.5	100	0.90	1.000	0.010	0.009	0.000	0.014	0.897
1	1.5	100	0.95	1.000	0.009	0.009	0.000	0.013	0.951
1	3.0	50	0.90	1.000	0.009	0.009	0.000	0.013	0.902
1	3.0	50	0.95	1.000	0.009	0.009	0.000	0.013	0.951
1	3.0	100	0.90	1.000	0.005	0.005	0.000	0.007	0.904
1	3.0	100	0.95	1.000	0.005	0.005	0.000	0.007	0.952
100	1.5	50	0.90	100.096	1.963	1.888	0.096	2.753	0.904
100	1.5	50	0.95	100.026	1.886	1.883	0.026	2.692	0.951
100	1.5	100	0.90	100.034	0.958	0.943	0.034	1.350	0.910
100	1.5	100	0.95	100.008	0.960	0.938	0.008	1.349	0.950
100	3.0	50	0.90	100.010	0.900	0.922	0.010	1.301	0.910
100	3.0	50	0.95	100.078	0.956	0.919	0.079	1.341	0.949
100	3.0	100	0.90	100.011	0.486	0.468	0.011	0.678	0.901
100	3.0	100	0.95	100.015	0.480	0.467	0.015	0.673	0.947

Table 2: EST, ASE, ESE, BIAS, RMSE, ACR and ECR for C when ψ is unknown.

c	ψ	n	$1-\gamma$	EST	ESE	ASE	BIAS	RMSE	ECR
1	1.5	50	0.90	1.486	0.291	0.305	0.014	0.426	0.857
1	1.5	50	0.95	1.494	0.309	0.307	0.006	0.440	0.953
1	1.5	100	0.90	1.497	0.212	0.214	0.003	0.303	0.917
1	1.5	100	0.95	1.494	0.217	0.214	0.006	0.306	0.966
1	3.0	50	0.90	3.020	0.604	0.620	-0.020	0.874	0.860
1	3.0	50	0.95	3.009	0.610	0.617	-0.009	0.877	0.952
1	3.0	100	0.90	2.999	0.427	0.429	0.001	0.609	0.910
1	3.0	100	0.95	3.018	0.439	0.432	-0.018	0.619	0.965
100	1.5	50	0.90	1.488	0.285	0.305	0.012	0.422	0.868
100	1.5	50	0.95	1.482	0.298	0.304	0.018	0.430	0.953
100	1.5	100	0.90	1.498	0.215	0.215	0.002	0.306	0.910
100	1.5	100	0.95	1.497	0.227	0.214	0.003	0.314	0.963
100	3.0	50	0.90	2.988	0.573	0.613	0.012	0.847	0.864
100	3.0	50	0.95	3.011	0.596	0.618	-0.011	0.867	0.958
100	3.0	100	0.90	2.988	0.427	0.428	0.012	0.608	0.903
100	3.0	100	0.95	3.019	0.448	0.432	-0.019	0.626	0.968

Table 3: EST, ASE, ESE, BIAS, RMSE, ACR and ECR for ψ when C is known.

c	ψ	n	$1-\gamma$	EST	ESE	ASE	BIAS	RMSE	ECR
1	1.5	50	0.90	1.494	0.307	0.310	0.006	0.440	0.852
1	1.5	50	0.95	1.506	0.313	0.312	-0.006	0.447	0.955
1	1.5	100	0.90	1.490	0.202	0.214	0.010	0.296	0.913
1	1.5	100	0.95	1.505	0.231	0.217	-0.005	0.318	0.952
1	3.0	50	0.90	2.997	0.587	0.621	0.003	0.863	0.859
1	3.0	50	0.95	3.027	0.656	0.628	-0.027	0.918	0.949
1	3.0	100	0.90	2.994	0.430	0.431	0.006	0.612	0.907
1	3.0	100	0.95	3.013	0.450	0.434	-0.013	0.628	0.959
100	1.5	50	0.90	1.500	0.289	0.311	0.000	0.429	0.860
100	1.5	50	0.95	1.498	0.306	0.311	0.002	0.441	0.953
100	1.5	100	0.90	1.508	0.216	0.217	-0.008	0.308	0.899
100	1.5	100	0.95	1.493	0.217	0.215	0.007	0.307	0.958
100	3.0	50	0.90	2.996	0.617	0.621	0.004	0.884	0.855
100	3.0	50	0.95	3.004	0.615	0.623	-0.004	0.884	0.951
100	3.0	100	0.90	2.983	0.421	0.429	0.017	0.605	0.900
100	3.0	100	0.95	2.986	0.446	0.430	0.014	0.623	0.956

Table 4: EST, ASE, ESE, BIAS, RMSE, ACR and ECR for ψ when C is unknown.

C	ψ	n	$1-\gamma$	UEST	ESE	ASE	BIAS	RMSE	ECR
1	1.5	50	90	1.540	0.379	0.409	0.040	0.345	0.905
1	1.5	50	95	1.544	0.394	0.410	0.044	0.359	0.947
1	1.5	100	90	1.513	0.266	0.273	0.013	0.152	0.895
1	1.5	100	95	1.521	0.265	0.274	0.021	0.153	0.952
1	3	50	90	3.086	0.794	0.820	0.086	1.445	0.908
1	3	50	95	3.048	0.755	0.810	0.048	1.356	0.959
1	3	100	90	3.000	0.515	0.541	0.000	0.585	0.907
1	3	100	95	3.035	0.522	0.547	0.035	0.601	0.963
100	1.5	50	90	1.554	0.385	0.413	0.054	0.355	0.926
100	1.5	50	95	1.549	0.401	0.411	0.049	0.367	0.946
100	1.5	100	90	1.513	0.257	0.273	0.013	0.147	0.908
100	1.5	100	95	1.494	0.261	0.269	-0.006	0.148	0.951
100	3	50	90	3.061	0.807	0.813	0.061	1.368	0.920
100	3	50	95	3.116	0.809	0.827	0.116	1.441	0.955
100	3	100	90	3.053	0.546	0.550	0.053	0.646	0.891
100	3	100	95	2.997	0.537	0.540	-0.003	0.593	0.951

Table 5: Inference for ψ when C is known, under Bayesian predictive sampling with hyper parametric values $d = 0$ and $c_0 = 0$.

C	ψ	n	$1-\gamma$	UEST	ESE	ASE	BIAS	RMSE	ECR
1	1.5	50	90	1.535	0.367	0.408	0.035	0.334	0.918
1	1.5	50	95	1.543	0.387	0.410	0.043	0.353	0.953
1	1.5	100	90	1.514	0.267	0.273	0.014	0.153	0.909
1	1.5	100	95	1.524	0.264	0.275	0.024	0.153	0.955
1	3	50	90	3.062	0.755	0.813	0.062	1.365	0.904
1	3	50	95	3.060	0.766	0.813	0.060	1.382	0.950
1	3	100	90	3.032	0.522	0.547	0.032	0.601	0.904
1	3	100	95	3.021	0.531	0.545	0.021	0.607	0.949
100	1.5	50	90	1.521	0.369	0.404	0.021	0.331	0.921
100	1.5	50	95	1.537	0.379	0.408	0.037	0.344	0.960
100	1.5	100	90	1.521	0.268	0.274	0.021	0.154	0.904
100	1.5	100	95	1.518	0.278	0.274	0.018	0.160	0.953
100	3	50	90	3.065	0.775	0.814	0.065	1.368	0.909
100	3	50	95	3.055	0.774	0.811	0.055	1.449	0.947
100	3	100	90	3.046	0.539	0.549	0.046	0.647	0.898
100	3	100	95	3.013	0.538	0.543	0.013	0.601	0.948

Table 6: Inference for ψ when C is known, under Bayesian predictive sampling with hyper parametric values $d = 0$ and $c_0 = 1$.

The coverage rates are mostly close to the nominal level throughout all scenarios, further suggesting the estimation method is robust and the estimates are accurate. More specifically, CRs corresponding to C behave quite well for both cases when ψ is known or unknown. However, though rare, there are some instances of slight under-coverage for ψ when employing our estimation method, specifically when C is unknown (see Table 4). A probable reason can be the mathematical dependence of the estimator of ψ on C (known or unknown). But, we would like emphasize that this under-coverage reduces as the sample size increases, validating that for large enough sample size confidence intervals provided by our estimation method are quite precise and reliable.

In Tables 5 and 6, we list the estimation results on ψ when C is known under posterior predictive sampling. Throughout, we assign $d = 0$ that results in unbiased estimates of ψ. Simulation results corresponding to $c_0 = 0$ and $c_0 = 1$ are presented in Tables 5 and 6, respectively. The bias in the estimates are of the order of 10^{-2} and coverage rates are close to the specified values of confidence level. Impact of increase in sample size can be seen in the reduction of BIAS and RMSE.

4. CONCLUDING REMARKS

In this paper, we have derived likelihood based methods of inference for synthetic data when the original data comes from a two parameter Pareto model. To this end, synthetic data were generated by two different methods, *viz.* plug-in sampling and posterior predictive sampling. For the plug-in sampling method, we have developed unbiased estimators for the parameters, and obtained the expressions of the corresponding variances and shortest distance CIs under three possible scenarios (inference on ψ when C is known, inference on C when ψ is known and inference on θ when both parameters are unknown). On the other hand, under posterior predictive sampling, inference has been drawn only for the shape parameter ψ when C is known. The methods have been discussed based on a single synthetic data set.

Results from the simulation study have shown that the plug-in sampling exhibits less bias, ASE and RMSE than posterior predictive sampling. A similar observation has been reported by [17] for a synthetic data from exponential distribution.

The developed estimators are unbiased in nature, and have been developed based on sufficient statistics. Exact shortest distance confidence intervals for parameters have been constructed for all methods of sampling, except for C when ψ is unknown in plug-in sampling. The primary strength of these methods is that they are based on a single synthetic data set, which is advantageous when release of multiple data sets is not allowed due to privacy concerns.

Despite observing actual microlevel data, the methodologies developed in this paper would allow researchers and policy makers to gain insights into the extent of financial burden tax payers face by filing income tax, or distribution of income or wealth across various strata of the society. The mathematical expressions provided in the paper would enable them to estimate the key parameters of the distribution relatively accurately, thereby, necessitating appropriate economic policy changes, or identifying gaps in a financial program or strategy. Computations of confidence intervals may require evaluating implicit integrals or solving non-linear simultaneous equations. However, these can be carried out easily by any established statistical software. It is recommended that users should carry out proper hypothesis test to verify whether a Pareto model fits data for a particular location or period. For researchers in government agencies, who have the access to actual data, could verify the precision of our estimates, and assess merits in our techniques. Future work may include developing estimation procedures in case of posterior predictive sampling with different informative priors.

ACKNOWLEDGMENTS

This work has not been supported by any grant or external funding source. We acknowledge the valuable suggestions provided by the referees.

REFERENCES

[1] Kelly, J. P., Golden, B. L. and Assad, A. A. (1992). Cell suppression: Disclosure protection for sensitive tabular data, *Journal of Political Economy*, **22**, 4, 397-417.

[2] Evans, T., Zayatz, L. and Slanta, J. (1996). Using noise for disclosure limitation of establishment tabular data, *Proceedings of the Annual Research Conference, US Bureau of the Census, Washington, DC*, **20233**, 4, 65-86.

[3] Dalenius, T. and Reiss, S. P. (1982). Data-swapping: A technique for disclosure control, *Journal of Statistical Planning and Inference*, **6**, 1, 73-85.

[4] Kim, J. J. (1986). A method for limiting disclosure in microdata based on random noise and transformation, *Proceedings of the American Statistical Association, Section on Survey Research Methods*, Alexandria, VA, American Statistical Association, 370–374.

[5] Kim, J. J and Winkler, W. (1995). Masking microdata files, *Proceedings of the American Statistical Association, Section on Survey Research Methods.* Alexandria, VA, American Statistical Association, 114–119.

[6] Kim, J. and Winkler, W. (2003). Multiplicative noise for masking continuous data, *Statistical Research Division, Research Report Series, U.S. Census Bureau.*

[7] Rubin, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*, John Wiley & Sons, Hoboken, NJ, USA.

[8] Willenborg, L. and De Waal, T. (2012). *Elements of Statistical Disclosure Control*, Springer Science & Business Media, NY, USA

[9] Little, R.J.A. (1993). Statistical analysis of masked data, *Journal of Official Statistics*, **9**, 2, 407–426.

[10] Nayak, T. K., Sinha, B. and Zayatz, L. (2011). Statistical properties of multiplicative noise masking for confidentiality protection, *Journal of Official Statistics*, **27**, 3, 527–544.

[11] Sinha, B., Nayak, T. K. and Zayatz, L. (2011). Privacy protection and quantile estimation from noise multiplied data, *Sankhya B*, **73**, 2, 297–315.

[12] Klein, M., Mathew, T. and Sinha, B. (2014). Likelihood based inference under noise multiplication, *Thailand Statistician: Journal of the Thai Statistical Association*, **12**, 1–23.

[13] Rubin, D. B. (1993). Statistical disclosure limitation, *Journal of Official Statistics*, **9**, 2, 461–468.

[14] Raghunathan, T. E., Reiter, J. P. and Rubin, D. B. (2003). Multiple imputation for statistical disclosure limitation, *Journal of Official Statistics*, **19**, 1–16.

[15] Reiter, J. P. (2004). Releasing multiply imputed, synthetic public-use microdata: an illustration and empirical, *Journal of the Royal Statistical Society, Series A* **168** 185–205.

[16] Reiter, J.P. and Raghunathan, T. E. (2007). The multiple adaptations of multiple imputation, *Journal of the American Statistical Association*, **102**, 1462–1471.

[17] Klein, M. and Sinha, B. (2015). Likelihood-based finite sample inference for synthetic data based on exponential model, *Thailand Statistician*, **13**, 1, 33–47.

[18] Klein, M. and Sinha, B. (2015). Inference for singly imputed synthetic data based on posterior predictive sampling under multivariate normal and multiple linear regression models, *Sankhya B*, **77**, 2, 293–311.

[19] Klein, M. and Sinha, B. (2015). Likelihood-based inference for singly and multiply imputed synthetic data under a normal model, *Statistics & Probability Letters*, **105**, 168–175.

[20] Reiter, J. P. and Kinney, S. K. (2012). Inferentially valid, partially synthetic data: Generating from posterior predictive distributions not necessary, *Journal of Official Statistics*, **28**, 4, 583–590.

[21] Reiter, J. P. (2003). Inference for partially synthetic, public use microdata sets, *Survey Methodology*, **29**, 2, 181–188.

[22] Kinney, S. K., Reiter, J.P., Reznek, A. P., Miranda, J., Jarmin, R. S. and Abowd, J. M. (2011). Towards unrestricted public use business microdata: The synthetic longitudinal business database, *International Statistical Review*, **79**, 3, 362–384.

[23] Bowen, C. M., Bryant, V., Burman, L., Khitatrakun, S., McClelland, R., Stallworth, P., Ueyama, Kyle and Williams, A. R. (2020, September). A synthetic supplemental public use file of low-income information return data: methodology, utility, and privacy implications. *In International Conference on Privacy in Statistical Databases* Springer, Cham, 257–270.

[24] Hagstroem, K.G. (1960). Remarks on Pareto distributions, *Scandinavian Actuarial Journal*, **1**, 59–71.

[25] Mandelbrot, B. (1963). New methods in statistical economics, *Journal of Political Economy*, **71**, 5, 421–440.

[26] Malik, H. J. (1970). Estimation of the parameters of the Pareto distribution, *Metrika*, **15**, 1, 126–132.

[27] Arnold, Barry C (2015). *Pareto Distributions*, Chapman and Hall/CRC, Boca Raton, FL, USA.

[28] Souma, W. (1970). Universal structure of the personal income distribution, *Fractals*, **9**, 4, 463–470.

[29] Drechsler, J. (2011). *Synthetic Datasets for Statistical Disclosure Control: Theory and Implementation* (Vol. 201). Springer Science & Business Media.

[30] Saksena, S.K. and Johnson, A.M. (1984). Best unbiased estimators for the parameters of a two-parameter Pareto distribution, *Metrika*, **31**, 1, 77–83.

[31] Wolfram Research, Inc., Mathematica, Version 12.2, Champaign, IL (2020).

[32] Casella, G. and Berger, R. L. (2002). **2**, *Statistical Inference*, Duxbury Pacific Grove, CA, USA

[33] Lehmann, L. and Casella, G. (2006). *Theory of Point Estimation*, Springer Science & Business Media.

[34] Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing, *Handbook of Econometrics*, **4**, 2111–2245.

[35] Arnold, B. C. and Press, S. J. (1983). Bayesian inference for Pareto populations, *Journal of Econometrics*, **21**, 3, 287–306.

[36] R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.