Likelihood-Based Finite Sample Inference for Synthetic Data Based on Exponential Model

Martin Klein [a]* **and Bimal Sinha** [b,c]

[a] Center for Statistical Research and Methodology, U.S. Census Bureau, Washington, DC 20233, U.S.A.

[b] Center for Disclosure Avoidance Research, U.S. Census Bureau, Washington, DC 20233, U.S.A.

[c] Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD 20250, U.S.A.

*Author for correspondence; e-mail: martin.klein@census.gov

Abstract

Likelihood-based finite sample inference based on synthetic data under the exponential model is developed in this paper. Two distinct synthetic data generation scenarios are considered, one based on posterior predictive sampling, and the other based on plug-in sampling. It is demonstrated that valid inference can be drawn in both scenarios, even for a singly imputed synthetic dataset. The usual combination rules for drawing inference under multiple synthetic datasets are discussed in the context of likelihood-based data analysis.

Disclaimer: This article is released to inform interested parties of ongoing research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

Keywords: Exponential distribution, Maximum likelihood estimator, Plug-in sampling, Posterior predictive sampling, Statistical disclosure control, Synthetic data, Uniformly minimum variance unbiased estimator.

1 Introduction

Statistical agencies are often faced with two conflicting objectives: (1) collect and publish useful datasets for designing public policies and building scientific theories, and (2) protect confidentiality of survey respondents which is essential to uphold public trust, leading to better response rates and data accuracy. Although cell suppression and swapping are two popular methods for statistical disclosure control, use of noise-perturbed and synthetic datasets has gained considerable popularity and importance in recent times.

In regard to noise perturbation of original microdata to protect confidentiality (Kim [1]; Kim and Winkler [2], [3]; Little [4]), recently Nayak et al. [5] and Sinha et al. [6] discussed some salient features and properties of noise-multiplied data in general terms; Lin and Wise [7] considered estimation of regression parameters based on noise-multiplied data; Klein et al. [8] developed likelihood-based data analysis methods under noise-multiplication based on samples from exponential, normal and lognormal populations; and Klein and Sinha [9] proposed an approach to disseminate and analyze noise-multiplied data using multiple imputation.

The focus of this paper is to address some inferential aspects of statistical analysis based on synthetic data when real datasets are not released and, as a substitute, synthetic datasets based on the real data are created for publication and analysis. Rubin [10] first advocated use of synthetic data for statistical disclosure control, using the framework of multiple imputation (Rubin [11]), and argued that synthetic data so created do not correspond to any actual sampling unit, thus preserving the confidentiality of the respondents. Inferential methods for fully synthetic data were developed by Raghunathan et al. [12], and inferential methods for partially synthetic data were developed by Reiter [13]. Reiter [14] presented an illustration and empirical study of fully synthetic data. An overview of multiple imputation techniques, including its use in statistical disclosure control, is provided by Reiter and Raghunathan [15]. There has been much research to further develop synthetic data methodology, and a systematic account of the developments is provided by Drechsler [16]. The methodology of partially synthetic data has been successfully applied to a number of data products in the United States as described by Reiter and Kinney [17] and the references therein.

The two methods considered in this paper for generation and analysis of synthetic data are denoted by Case 1 and Case 2. To describe these two methods, suppose that $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ are

the original microdata which are jointly distributed according to the probability density function (pdf) $f_{\boldsymbol{\theta}}(\boldsymbol{x})$, where $\boldsymbol{\theta}$ is the unknown (scalar or vector) parameter.

Case 1: Posterior Predictive Sampling. Assume a prior $\pi(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$, then the posterior distribution of $\boldsymbol{\theta}$ given $\boldsymbol{x}$ is derived and used to draw m independent replications $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_m^*$ (known as posterior draws). Next, for each such posterior draw of $\boldsymbol{\theta}$, a corresponding replicate of $\boldsymbol{x}$ is generated, namely, $\boldsymbol{z}_i = (\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{in})$ is drawn from the pdf $f_{\boldsymbol{\theta}_i^*}(\boldsymbol{x})$, where $f_{\boldsymbol{\theta}_i^*}(\boldsymbol{x})$ denotes the joint pdf of the original data $\boldsymbol{x}$, with the unknown $\boldsymbol{\theta}$ replaced by the posterior draw $\boldsymbol{\theta}_i^*$. The synthetic data $\mathfrak{Z} = \{\boldsymbol{z}_i = (\boldsymbol{z}_{i1}, \ldots, \boldsymbol{z}_{in}) : i = 1, \ldots, m\}$ are then released to the public. For the scenario described here, the usual practice for drawing inference on $\boldsymbol{\theta}$ from the synthetic data is based on the methods of Reiter [13] for partially synthetic data. To summarize, suppose $Q = Q(\boldsymbol{\theta})$ is a scalar parameter of interest. Let $\eta = \eta(\boldsymbol{x})$ denote a point estimator of Q and let $V = V(\boldsymbol{x})$ denote an estimator of the variance of η, both computed on the original data set $\boldsymbol{x}$. To draw inference on $\boldsymbol{\theta}$ based on the synthetic data $\mathfrak{Z}$, one would compute $\eta_i = \eta(\boldsymbol{z}_i)$ and $V_i = V(\boldsymbol{z}_i)$, the analogs of η and V, respectively, computed on the ith synthetic data set $\boldsymbol{z}_i$. Then the estimator of Q based on the entire synthetic data $\mathfrak{Z}$ is

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

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

$$T_m = \frac{B_m}{m} + \bar{V}_m, \tag{2}$$

where $B_m = \frac{1}{m-1} \sum_{i=1}^{m} (\eta_i - \bar{\eta}_m)^2$ and $\bar{V}_m = \frac{1}{m} \sum_{i=1}^{m} V_i$. An approximate level $(1-\gamma)$ confidence interval for Q can be computed as $\bar{\eta}_m \pm t_{\gamma/2,v} T_m^{1/2}$ where $t_{\gamma/2,v}$ is the upper $\gamma/2$ quantile of the t distribution with degrees of freedom $v = (m-1)(1 + R_m^{-1})^2$ with $R_m = B_m (m\bar{V}_m)^{-1}$.

Case 2: Plug-in Sampling. An alternative way to generate synthetic data is to take the observed value of a point estimator $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ of $\boldsymbol{\theta}$, and plug it into the joint pdf of $\boldsymbol{x}$. The resulting pdf, with the unknown $\boldsymbol{\theta}$ replaced by the observed value of the point estimator $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\boldsymbol{x})$, is denoted by $f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})$. The synthetic data, namely, $\mathcal{Y} = \{\boldsymbol{y}_i = (\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{in}) : i = 1, \ldots, m\}$ are generated by drawing each $\boldsymbol{y}_i$ independently from the joint pdf $f_{\hat{\boldsymbol{\theta}}}(\boldsymbol{x})$. As discussed by Reiter and Kinney [17], in this scenario the combination rules of Reiter [13] appear to remain valid. Thus to draw inference for the scalar parameter $Q = Q(\boldsymbol{\theta})$, one can use the combination formulas of equations (1) and (2) along with

the t confidence interval discussed above (obviously, with $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m$ replaced by $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$, that is, $\eta_i = \eta(\boldsymbol{y}_i)$ and $V_i = V(\boldsymbol{y}_i)$).

The motivations for this current research are twofold. First, although synthetic data methodology calls for releasing $m > 1$ synthetic versions of the original data, there are situations where this is not feasible, perhaps due to severe privacy concerns. For example, the Synthetic Longitudinal Business Database, accessible through the VirtualRDC at Cornell University, is a synthetic version of the U.S. Census Bureau's Longitudinal Business Database (LBD). As discussed in Kinney et al. [18], the decision was made to release only a single version of the LBD in the synthetic file, instead of multiple copies, to avoid the perception of high disclosure risk. The usual combining rules are not applicable when only a single synthetic data set is released (i.e. when $m = 1$), so one wonders if it is possible to get a valid inference in this case. The results of this paper show that it is indeed possible in some cases, if one fully utilizes the model structure. Secondly, irrespective of how the synthetic data are generated, admittedly it is *model-based* and hence one wonders if rigorous model-based *finite sample* inference about $Q(\boldsymbol{\theta})$ can be developed based on $\mathfrak{Z}$ (Case 1) or $\mathfrak{Y}$ (Case 2). The results developed in this paper are used to obtain such a finite sample inference.

The organization of the paper is as follows. We develop likelihood-based inference for exponential mean in Section 2, and provide some concluding remarks in Section 3. Throughout, we derive the *exact* likelihood of synthetic data for both Cases 1 and 2, and carry out inference for the exponential mean. In the sequel we also allow a general form of the prior $\pi(\boldsymbol{\theta})$ under Case 1, involving a hyperparameter α, and make some recommendations about its choice. Our comparison of the two approaches of synthetic data generation reveals some very interesting features. The entire treatment is non-asymptotic in nature. We assume that the data user has knowledge of the form of the parametric model $f_{\boldsymbol{\theta}}(\boldsymbol{x})$ of the original data, and that this model is used to create the synthetic data as described above. Furthermore, in Case 1, we assume that knowledge of the underlying prior is available to the data user for conducting the analysis.

2 Methodology for Drawing Likelihood Based Inference

Throughout this section, we work under the following notation and model. Suppose that the original data $x_1, \ldots, x_n$ are independent and identically distributed (*iid*) according to the exponential probability density function (pdf) $f_\theta(x) = \frac{1}{\theta} e^{-x/\theta}$, $x > 0$, where $\theta > 0$ is the unknown param-

eter. Writing $\boldsymbol{x} = (x_1, \ldots, x_n)$ and $u = \sum_{i=1}^{n} x_i$, note that u is a sufficient statistic for θ, with u distributed as Gamma(n, θ), and the maximum likelihood estimator (MLE) of θ based on $\boldsymbol{x}$ is $\hat{\theta}_{\text{MLE}}(\boldsymbol{x}) = \bar{x} = u/n$.

2.1 Case 1: Posterior Predictive Sampling

Inference Based on a Singly Imputed Synthetic Data Set. Under a Bayesian setting, we would say that the synthetic data $\boldsymbol{z}$ are drawn from the posterior predictive distribution of $\boldsymbol{x}$. We take the prior distribution on θ as $\pi(\theta) \propto \theta^{-\alpha}$, $\theta > 0$. Using this prior and noting that u is sufficient for θ, suppose that, conditional on u, a synthetic dataset $\boldsymbol{z} = (z_1, \ldots, z_n)$ is generated as follows.

Step 1. Draw θ^* from the posterior distribution of θ given u. The posterior takes the form of the inverse gamma distribution having parameters $(n + \alpha - 1)$ and u; that is, draw θ^* from the pdf $\pi(\theta^*|u) = \frac{u^{n+\alpha-1}}{\Gamma(n+\alpha-1)}(\theta^*)^{-(n+\alpha-1)-1}e^{-\frac{u}{\theta^*}}$, $\theta^* > 0$.

Step 2. Given the value of θ^* drawn in step 1, draw $z_1, \ldots, z_n$ as *iid* from the exponential density $f_\theta(z) = \frac{1}{\theta}e^{-z/\theta}, z > 0$ with the unknown θ replaced by θ^*.

Central to our analysis based on $\boldsymbol{z}$ is its joint pdf or the likelihood function of θ based on $\boldsymbol{z}$, given by the following, where we write $Z = \sum_{i=1}^{n} z_i$.

Theorem 2.1 *The joint pdf of $\boldsymbol{z}$ is given by*

$$h_\theta(\boldsymbol{z}) = \int_0^\infty \left[\frac{e^{-\frac{Z\xi}{\theta}}\xi^n}{\theta^n}\right]\left[\frac{\xi^{n+\alpha-2}}{B(n, n+\alpha-1)(1+\xi)^{2n+\alpha-1}}\right] d\xi$$

which, interestingly enough, is a scale mixture of gamma with an F-type mixing distribution.

Proof. The proof of this result follows from the fact that the marginal pdf of $\boldsymbol{z}$ is given by

$$h_\theta(\boldsymbol{z}) = \int_0^\infty \int_0^\infty \left[\frac{e^{-\frac{1}{\theta^*}Z}}{(\theta^*)^n}\right]\left[\frac{u^{n+\alpha-1}}{\Gamma(n+\alpha-1)}(\theta^*)^{-(n+\alpha-1)-1}e^{-\frac{u}{\theta^*}}\right]\left[\frac{u^{n-1}e^{-\frac{u}{\theta}}}{\Gamma(n)\theta^n}\right] d\theta^* du,$$

for $z_1 > 0, \ldots, z_n > 0$, and thus, upon integrating out θ^*, we get

$$h_\theta(\boldsymbol{z}) = \int_0^\infty \left[\frac{\Gamma(2n+\alpha-1)}{(Z+u)^{2n+\alpha-1}}\right]\left[\frac{u^{n+\alpha-1}}{\Gamma(n+\alpha-1)}\right]\left[\frac{1}{\Gamma(n)\theta^n}u^{n-1}e^{-\frac{u}{\theta}}\right] du.$$

Finally, we make a transformation from u to $\xi = u/Z$ which yields the desired result. □

The MLE of θ which is obtained by maximizing this pdf is readily given by $\hat{\theta}_{\text{MLE}}(\boldsymbol{z}) = \frac{Z}{\eta_{n,\alpha}}$ where $\eta_{n,\alpha}$ is the value of η that maximizes

$$Q_1(\eta) = \eta^n \int_0^\infty \frac{e^{-\eta\xi}\xi^{2n+\alpha-2}}{(1+\xi)^{2n+\alpha-1}} d\xi.$$

The mean squared error (MSE) of the MLE is computed as (using expressions for $E(Z)$ and $E(Z^2)$ derived below)

$$\text{MSE}(\hat{\theta}_{\text{MLE}}(\boldsymbol{z})) = \theta^2 \left[\frac{n^2(n+1)^2}{\eta_{n,\alpha}^2(n+\alpha-2)(n+\alpha-3)} - \frac{2n^2}{\eta_{n,\alpha}(n+\alpha-2)} + 1 \right].$$

Remark 2.1. One can verify that Z, which is obviously sufficient for θ, is also complete. Here is an outline of the proof. Assume $E_\theta[g(Z)] = 0$ for all $\theta > 0$. Writing $\frac{\xi}{\theta} = \eta$ and changing the order of integration, this is equivalent to

$$0 = \int_0^\infty \left[\int_0^\infty e^{-Z\eta} g(Z) dZ \right] \frac{\eta^{2n+\alpha-2}}{(1+\theta\eta)^{2n+\alpha-1}} d\eta.$$

Upon defining $\lambda(\eta) = \eta^{2n+\alpha-2}[\int_0^\infty e^{-Z\eta} g(Z) dZ]$, and noting that $\int_0^\infty e^{-u(1+\theta\eta)} u^{2n+\alpha-2} du = \Gamma(2n+\alpha-1)(1+\theta\eta)^{-(2n+\alpha-1)}$, we get

$$0 = \int_0^\infty \left[\int_0^\infty e^{-u(1+\theta\eta)} u^{2n+\alpha-2} du \right] \lambda(\eta) d\eta.$$

Finally, writing $v = u\theta$, and changing the order of integration, we get

$$0 = \int_0^\infty e^{-\frac{v}{\theta}} v^{2n+\alpha-2} \left[\int_0^\infty e^{-v\eta} \lambda(\eta) d\eta \right] dv.$$

Now completeness of v implies that the inner integral above is 0 for all v (almost everywhere (a.e.)), which in turn implies $\lambda(\eta) = 0$ for all η (a.e.), and hence $g(Z) = 0$ for all Z (a.e.).

Thus, while the computation of the MLE of θ based on the joint pdf of $\boldsymbol{z}$ is not explicit, the uniformly minimum variance unbiased estimator (UMVUE) of θ based on Z is obtained as

$$\hat{\theta}_{\text{UMVUE}}(\boldsymbol{z}) = \frac{n+\alpha-2}{n} \bar{Z}, \tag{3}$$

where $\bar{Z} = \frac{Z}{n}$. The variance of the UMVUE is computed (in several steps) as follows. Note that

$$E[\mathrm{Var}(\bar{Z}|\theta^*)] = E\left[\frac{(\theta^*)^2}{n}\right] = \frac{\theta^2(n+1)}{(n+\alpha-2)(n+\alpha-3)},$$
$$\mathrm{Var}[E(\bar{Z}|\theta^*)] = \mathrm{Var}(\theta^*) = E[\mathrm{Var}(\theta^*|\boldsymbol{x})] + \mathrm{Var}[E(\theta^*|\boldsymbol{x})],$$

$$E[\mathrm{Var}(\theta^*|\boldsymbol{x})] = E\left[\frac{u^2}{(n+\alpha-2)(n+\alpha-3)} - \frac{u^2}{(n+\alpha-2)^2}\right] = \frac{n(n+1)\theta^2}{(n+\alpha-2)^2(n+\alpha-3)},$$
$$\mathrm{Var}[E(\theta^*|\boldsymbol{x})] = \mathrm{Var}\left[\frac{u}{n+\alpha-2}\right] = \frac{n\theta^2}{(n+\alpha-2)^2}.$$

Combining the above terms, we get

$$\begin{aligned}\mathrm{Var}(\bar{Z}) &= \theta^2\left[\frac{(n+1)}{(n+\alpha-2)(n+\alpha-3)} + \frac{n(n+1)}{(n+\alpha-2)^2(n+\alpha-3)} + \frac{n}{(n+\alpha-2)^2}\right] \\ &= \theta^2\frac{(n+\alpha)(2n+1)+n^2-4n-2}{(n+\alpha-2)^2(n+\alpha-3)}.\end{aligned}$$

Hence, we get

$$\mathrm{Var}(\hat{\theta}_{\mathrm{UMVUE}}(\boldsymbol{z})) = \frac{\theta^2}{n^2}\left[\frac{(n+\alpha)(2n+1)+n^2-4n-2}{(n+\alpha-3)}\right] = \frac{\theta^2}{n^2}\left[2n+1+\frac{(n+1)^2}{n+\alpha-3}\right]. \quad (4)$$

To construct a confidence interval for θ, one can verify that $Z^* = Z/\theta$ is a pivot with its distribution given as

$$h(z^*) = \int_0^\infty \left[\frac{e^{-z^*\xi}(z^*)^{n-1}\xi^n}{\Gamma(n)}\right]\left[\frac{\xi^{n+\alpha-2}}{B(n,n+\alpha-1)(1+\xi)^{2n+\alpha-1}}\right]d\xi.$$

If $c_{n,\alpha}$ and $d_{n,\alpha}$ satisfy:

$$\int_{c_{n,\alpha}}^{d_{n,\alpha}} h(z^*)dz^* = 1-\gamma, \;\; c_{n,\alpha}^2 h(c_{n,\alpha}) = d_{n,\alpha}^2 h(d_{n,\alpha}),$$

then the shortest $1-\gamma$ level confidence interval for θ based on Z and the expected length of the confidence interval are obtained, respectively, as

$$\left[\frac{Z}{d_{n,\alpha}}, \frac{Z}{c_{n,\alpha}}\right] \;\text{ and }\; E[L_1(\boldsymbol{z})] = \frac{n^2\theta}{n+\alpha-2}\left[\frac{1}{c_{n,\alpha}} - \frac{1}{d_{n,\alpha}}\right],$$

where $L_1(\boldsymbol{z}) = Z(1/c_{n,\alpha} - 1/d_{n,\alpha})$.

Remark 2.2. It follows from (3) that only the choice $\alpha = 2$ makes the standard estimator of θ, namely, $\bar{Z}$, unbiased for θ. This shows that the usual combination rule (suggesting $\bar{Z}$) will not be unbiased in this case unless α is appropriately chosen. However, we note that for fixed α, the bias of $\bar{Z}$ converges to zero as $n \to \infty$.

Inference Based on a Multiply Imputed Synthetic Dataset. Now we suppose that conditional on u, the synthetic dataset consists of $m > 1$ replications of the original dataset generated by repeating Steps 1 and 2 (from the beginning of Section 2.1) a total of m times to get the synthetic data: $(z_{11}, \ldots, z_{1n})$, ..., $(z_{m1}, \ldots, z_{mn})$. Thus, for multiple replications of z-values, which is the usual synthetic data scenario, we denote by z_{ij} the jth synthetic value from the ith replication, $j = 1, \ldots, n$, $i = 1, \ldots, m$. Let $Z_i = \sum_{j=1}^{n} z_{ij}$ and $\boldsymbol{Z} = (Z_1, \ldots, Z_m)$. One can check that the vector $\boldsymbol{Z}$ is jointly sufficient for θ.

Theorem 2.2 *The joint pdf of* $\boldsymbol{Z} = (Z_1, \ldots, Z_m)$ *is*

$$h_\theta(Z_1, \ldots, Z_m) = \int_0^\infty \left[\prod_{i=1}^{m} \frac{u^{n+\alpha-1} Z_i^{n-1}}{B(n, n+\alpha-1)(u+Z_i)^{2n+\alpha-1}}\right] \left[\frac{e^{-\frac{u}{\theta}} u^{n-1}}{\Gamma(n)\theta^n}\right] du.$$

Proof. The proof follows upon noting that the conditional joint pdf of $(Z_1, \ldots, Z_m)$, given u, is the product of individual densities of the form $h(Z_i|u) = \frac{u^{n+\alpha-1} Z_i^{n-1}}{B(n,n+\alpha-1)(u+Z_i)^{2n+\alpha-1}}$. □

The MLE of θ, which is not explicit, can be obtained by maximizing this joint pdf with respect to θ. Unlike in the case of $m = 1$, here $Z_1, \ldots, Z_m$ are jointly sufficient for θ, and obviously the joint distribution is *not* complete, implying there is no obvious estimator of θ based on the Z_i's. Letting $\bar{\bar{Z}} = \frac{1}{mn} \sum_{i=1}^{m} Z_i$, it can be shown that

$$\tilde{\theta}_1 = \frac{n+\alpha-2}{n} \bar{\bar{Z}} \tag{5}$$

is an unbiased estimator of θ with

$$\mathrm{Var}(\tilde{\theta}_1) = \frac{\theta^2}{mn^2} \left[mn + n + 1 + \frac{(n+1)^2}{n+\alpha-3}\right]. \tag{6}$$

Likewise, using the fact that $E[Z_1^\gamma|u] = \left[\frac{B(n+\gamma, n+\alpha-1-\gamma)}{B(n,n+\alpha-1)}\right] u^\gamma$, it follows that

$$E\left[\prod_{i=1}^{m} Z_i^{\gamma}\right] = \left[\frac{B(n+\gamma, n+\alpha-1-\gamma)}{B(n, n+\alpha-1)}\right]^m E[u^{m\gamma}].$$

Taking $\gamma = \frac{1}{m}$ and noting that $E(u) = n\theta$, a second unbiased estimator of θ based on the geometric mean of $(Z_1, \ldots, Z_m)$ is given by

$$\tilde{\theta}_2 = \frac{1}{n}\left[\prod_{i=1}^{m} Z_i^{\frac{1}{m}}\right]\left[\frac{B(n+\frac{1}{m}, n+\alpha-1-\frac{1}{m})}{B(n, n+\alpha-1)}\right]^{-m}, \tag{7}$$

and its variance is

$$\mathrm{Var}(\tilde{\theta}_2) = \theta^2\left[\frac{n+1}{n}\left\{\frac{B(n, n+\alpha-1)B(n+\frac{2}{m}, n+\alpha-1-\frac{2}{m})}{B^2(n+\frac{1}{m}, n+\alpha-1-\frac{1}{m})}\right\}^m - 1\right]. \tag{8}$$

It is also possible to suggest other unbiased estimators of θ based on $Z_{(1)} = \min\{Z_1, \ldots, Z_m\}$ and $Z_{(m)} = \max\{Z_1, \ldots, Z_m\}$.

Since $\boldsymbol{V} = (V_1, \ldots, V_n) = (\frac{Z_1}{\theta}, \ldots, \frac{Z_m}{\theta})$ is a pivot with the joint pdf

$$h(\boldsymbol{v}) = \int_0^{\infty}\left[\prod_{i=1}^{m}\frac{t^{n+\alpha-1}v_i^{n-1}}{B(n, n+\alpha-1)(t+v_i)^{2n+\alpha-1}}\right]\left[\frac{e^{-t}t^{n-1}}{\Gamma(n)}\right]dt,$$

confidence intervals for θ based on suitable combinations of them (arithmetic mean, geometric mean, minimum, maximum) can be derived, and these can be compared with the one based on *Plug-in Sampling* method discussed in the next section.

2.2 Case 2: Plug-in Sampling

Following Reiter and Kinney [17], here a synthetic data set $\boldsymbol{y} = (y_1, \ldots, y_N)$ of size N is generated by drawing $y_1, \ldots, y_N$ as *iid* from the exponential density $f_\theta(y) = \frac{1}{\theta}e^{-y/\theta}$, $y > 0$, with the unknown parameter θ set equal to $\hat{\theta}_{\mathrm{MLE}}(\boldsymbol{x}) = \bar{x} = u/n$. Notice that N, the size of the synthetic sample, is not necessarily taken to be equal to n, the size of the original sample. In the case of m multiply imputed synthetic data sets, one would take $N = nm$, while for a singly imputed synthetic data set, one would simply take $N = n$. Regardless of the choice of N, it is assumed that the value of the original sample size n is known to the data analyst, as this value will be needed to apply the methodology developed in this section. The goal now is to draw inference on θ based on the

synthetic data $\boldsymbol{y}$. Central to this goal is the joint pdf of $\boldsymbol{y}$, or the likelihood function of θ based on $\boldsymbol{y}$, which is stated below.

Theorem 2.3 *The joint pdf of* $\boldsymbol{y}$ *is given by*

$$g_\theta(\boldsymbol{y}) = \int_0^\infty \left[\frac{n^N}{u^N} e^{-\frac{n}{u}\sum_1^N y_i}\right] \left[\frac{1}{\Gamma(n)\theta^n} u^{n-1} e^{-u/\theta}\right] du. \tag{9}$$

Proof. The proof depends on the simple fact that the conditional pdf of $\boldsymbol{y}$, given u, is

$$g(\boldsymbol{y}|u) = \prod_{i=1}^N \left[\frac{n}{u} e^{-ny_i/u}\right] = \frac{n^N}{u^N} e^{-\frac{n}{u}\sum_1^N y_i},\ y_1 > 0, \ldots, y_N > 0.$$

□

Trivially, the statistic $t = \sum_{i=1}^N y_i$ is sufficient for θ in model (9), and its pdf (by the same conditional argument) is given by

$$g_\theta(t) = \int_0^\infty \left[\frac{n^N}{\Gamma(N)u^N} t^{N-1} e^{-\frac{nt}{u}}\right] \left[\frac{1}{\Gamma(n)\theta^n} u^{n-1} e^{-u/\theta}\right] du,\ t > 0, \tag{10}$$

which is a scale mixture of gamma. The MLE of θ can be obtained by maximizing $g_\theta(t)$ with respect to θ. Writing $\psi = \frac{u}{\theta}$, $g_\theta(t)$, apart from a constant, can be expressed as

$$g_\theta(t) \propto \frac{t^{N-1}}{\theta^N} \int_0^\infty e^{-\psi - \frac{nt}{\theta\psi}} \psi^{n-N-1} d\psi.$$

Putting $\eta = t/\theta$, we choose η by maximizing

$$Q_2(\eta) = \eta^N \int_0^\infty e^{-\psi - \frac{n\eta}{\psi}} \psi^{n-N-1} d\psi$$

over $0 < \eta < \infty$. If $\eta_{n,N}$ is the maximizer, the MLE of θ is given by $\hat{\theta}_{\text{MLE}}(\boldsymbol{y}) = \frac{t}{\eta_{n,N}}$. The mean squared error (MSE) of the MLE is computed as

$$\text{MSE}(\hat{\theta}_{\text{MLE}}(\boldsymbol{y})) = E\left[\frac{t^2}{\eta_{n,N}^2} - 2\frac{t\theta}{\eta_{n,N}} + \theta^2\right] = \theta^2\left[\frac{N(n+1)(N+1)}{n\eta_{n,N}^2} - 2\frac{N}{\eta_{n,N}} + 1\right].$$

On the other hand, it is easy to verify that the pdf of t given by (10) is complete. This is because if $\omega(t)$ satisfies $E_\theta[\omega(t)] = 0$, for all θ, by changing the order of integration, it follows that

$\int_0^\infty \omega(t) e^{-\frac{nt}{u}} t^{N-1} dt = 0$ for all u, which implies $\omega(t) = 0$ (a.e.). Hence the UMVUE of θ based on t and its variance are readily obtained as

$$\hat{\theta}_{\text{UMVUE}}(\boldsymbol{y}) = \bar{t} = \frac{t}{N}, \tag{11}$$

$$\text{Var}(\hat{\theta}_{\text{UMVUE}}(\boldsymbol{y})) = \text{Var}(\bar{t}) = \theta^2 \left[\frac{1}{n} + \frac{1}{N} + \frac{1}{nN}\right]. \tag{12}$$

In the above, $\text{Var}(\bar{t})$ is obtained by using the facts that $E(t) = N\theta$ and $E(t^2) = \frac{N(N+1)(n+1)}{n}\theta^2$, which follow by the usual conditional argument based on u.

To construct a confidence interval for θ based on t, we note that $t^* = t/\theta$ is a *pivot*. This is because the marginal pdf of t^* is

$$\begin{aligned} g_\theta(t^*) &= \int_0^\infty \left[\frac{n^N}{\Gamma(N)u^N}(\theta t^*)^{N-1} e^{-\frac{n}{u}\theta t^*}\theta\right]\left[\frac{1}{\Gamma(n)\theta^n} u^{n-1} e^{-u/\theta}\right] du \\ &= \int_0^\infty \left[\frac{n^N}{\Gamma(N)u^N}\theta^N (t^*)^{N-1} e^{-\frac{n}{u}\theta t^*}\right]\left[\frac{1}{\Gamma(n)\theta^n} u^{n-1} e^{-u/\theta}\right] du,\ t^* > 0. \end{aligned}$$

Writing $\xi = \frac{u}{\theta}$, $d\xi = \frac{du}{\theta}$, we can express $g_\theta(t^*)$ as

$$\begin{aligned} g(t^*) &= \int_0^\infty \left[\frac{n^N}{\Gamma(N)\xi^N}(t^*)^{N-1} e^{-\frac{n}{\xi}t^*}\right]\left[\frac{1}{\Gamma(n)}\xi^{n-1} e^{-\xi}\right] d\xi \\ &= \frac{n^N (t^*)^{N-1}}{\Gamma(N)\Gamma(n)} \int_0^\infty e^{-\xi - \frac{nt^*}{\xi}} \xi^{n-N-1} d\xi \end{aligned}$$

which is clearly free of θ, and can be used to construct a confidence interval for θ. Thus, if $a_{n,N}$ and $b_{n,N}$ satisfy

$$\int_{a_{n,N}}^{b_{n,N}} g(t^*) dt^* = 1 - \gamma, \ \ a_{n,N}^2 g(a_{n,N}) = b_{n,N}^2 g(b_{n,N}),$$

then the shortest $1-\gamma$ level confidence interval for θ based on t and its expected length are obtained, respectively, as

$$\left[\frac{t}{b_{n,N}},\ \frac{t}{a_{n,N}}\right] \ \text{ and } \ E[L_2(\boldsymbol{y})] = N\theta\left[\frac{1}{a_{n,N}} - \frac{1}{b_{n,N}}\right],$$

where $L_2(\boldsymbol{y}) = t(1/a_{n,N} - 1/b_{n,N})$.

Remark 2.3. Taking $N = n$ and comparing (4) and (12), it follows that $\hat{\theta}_{\text{UMVUE}}(\boldsymbol{y})$ has a smaller variance than $\hat{\theta}_{\text{UMVUE}}(\boldsymbol{z})$, whatever be α.

Remark 2.4. Table 1 presents a comparison of Cases 1 and 2 based on their expected length of the 95% confidence intervals for θ, scaled by θ, when $N = n$. The results in the table clearly indicate that Case 2 yields shorter expected length of confidence intervals than Case 1.

Remark 2.5. It is interesting to compare the unbiased estimator of θ, namely $\bar{t}$ (defined in (11)), with the unbiased estimators $\tilde{\theta}_1$ and $\tilde{\theta}_2$ (defined in (5) and (7)), based on the arithmetic mean and geometric mean, respectively. Some numerical values of the variances of these three estimators appear in Table 2 for $N = nm$. Surprisingly enough, $\tilde{\theta}_2$ turns out to be marginally better than $\tilde{\theta}_1$ in the scenarios considered, implying that the usual arithmetic mean combination approach need not always be preferable. Both $\tilde{\theta}_1$ and $\tilde{\theta}_2$ are found to be inferior to $\bar{t}$; that is, the unbiased estimator under Case 2 is more efficient than those obtained under Case 1.

Table 1: Cut-off points and scaled expected length (scaled by θ) of the 95% confidence interval for the exponential mean θ

	Plug-in sampling			Posterior predictive sampling					
				$\alpha = 1$			$\alpha = 2$		
n	a_n	b_n	$\frac{E[L_2(\boldsymbol{y})]}{\theta}$	$c_{n,\alpha}$	$d_{n,\alpha}$	$\frac{E[L_1(\boldsymbol{z})]}{\theta}$	$c_{n,\alpha}$	$d_{n,\alpha}$	$\frac{E[L_1(\boldsymbol{z})]}{\theta}$
10	4.1	28.2	2.1	3.7	45.0	2.5	3.4	39.1	2.7
15	7.4	33.8	1.6	6.7	46.9	1.9	6.3	43.4	2.0
20	10.8	39.9	1.3	9.9	51.7	1.6	9.6	48.3	1.7
25	14.4	46.3	1.2	13.3	57.2	1.4	12.9	54.1	1.5
30	18.2	52.2	1.1	16.9	62.6	1.3	16.4	60.2	1.3
50	34.1	75.8	0.8	31.9	86.5	1.0	31.5	84.2	1.0

Table 2: Numerical values of $\mathrm{Var}(\bar{t})$, $\mathrm{Var}(\tilde{\theta}_1)$ and $\mathrm{Var}(\tilde{\theta}_2)$

		Plug-in sampling	Posterior predictive sampling			
			$\alpha = 1$		$\alpha = 2$	
m	n	$\mathrm{Var}(\bar{t})$	$\mathrm{Var}(\tilde{\theta}_1)$	$\mathrm{Var}(\tilde{\theta}_2)$	$\mathrm{Var}(\tilde{\theta}_1)$	$\mathrm{Var}(\tilde{\theta}_2)$
5	10	0.122	0.152	0.147	0.149	0.145
	15	0.081	0.098	0.096	0.097	0.095
	20	0.061	0.073	0.072	0.072	0.071
	25	0.048	0.058	0.057	0.057	0.057
	30	0.040	0.048	0.047	0.048	0.047
	50	0.024	0.028	0.028	0.028	0.028

3 Concluding Remarks

In this paper, we have derived finite sample likelihood based methods of inference for synthetic data when the original data follow the exponential model and the synthetic data are generated either by posterior predictive sampling (Case 1) or by plug-in sampling (Case 2). We provided some comparisons between Case 1 and Case 2, and found that in general plug-in sampling yields more efficient inference than posterior predictive sampling. We have found that in Case 1, for finite n, the usual suggested estimator of θ based on single or multiple imputation exhibits bias unless α is suitably chosen. Also, in Case 2, if the original data are *iid* and m synthetic data sets are generated, then there is an arbitrariness in the usual combining rule for estimating variance, since there is no unique way to partition the data into m synthetic data sets. The methods developed in this paper however do not have this arbitrary nature.

The inferential methods developed in this paper are naturally somewhat more complicated to apply than the standard inferences based on the simple multiple imputation combining formulas. However, the methods in this paper have the desirable property that they are exact, and based on sufficient statistics. Furthermore, these methods allow a data user to draw valid inference when only a single synthetic data set is released which is useful in cases where (perhaps due to privacy concerns or limitations in resources) a statistical agency releases a single synthetic data set instead of multiple synthetic copies.

Acknowledgments

The authors thank the editor and two anonymous referees for providing helpful comments on a previous draft. The authors thank Paul Massell, Laura McKenna, Joseph Schafer, Eric Slud, Yves Thibaudeau, and Tommy Wright for encouragement.

References

[1] Kim, J. 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, 1986; American Statistical Association: 370-374.

[2] Kim, J.J., and Winkler, W.E. Masking Microdata Files, *Proceedings of the American Statistical Association, Section on Survey Research Methods*, Alexandria, VA, 1995; American Statistical Association: 114-119.

[3] Kim, J.J., and Winkler, W.E. Multiplicative Noise for Masking Continuous Data, *Statistical Research Division, Research Report Series (Statistics #2003-01), U.S. Census Bureau*: 2003. Available from URL: `http://www.census.gov/srd/papers/pdf/rrs2003-01.pdf` [accessed October 28, 2013].

[4] Little, R.J.A. Statistical Analysis of Masked Data, *Journal of Official Statistics*, 1993; 9: 407-426,

[5] Nayak, T., Sinha, B.K., and Zayatz, L. Statistical Properties of Multiplicative Noise Masking for Confidentiality Protection, *Journal of Official Statistics*, 2011; 27: 527-544.

[6] Sinha, B.K., Nayak, T., and Zayatz, L. Privacy Protection and Quantile Estimation From Noise Multiplied Data, *Sankhya, Series B*, 2011; 73: 297-315.

[7] Lin, Y.-X., and Wise, P. Estimation of Regression Parameters from Noise Multiplied Data, *Journal of Privacy and Confidentiality*, 2012; 4: 61-94.

[8] Klein, M., Mathew, T., and Sinha, B. Likelihood Based Inference Under Noise Multiplication, *Thailand Statistician: Journal of the Thai Statistical Association*, 2014; 12: 1-23.

[9] Klein, M., and Sinha, B. Statistical Analysis of Noise Multiplied Data Using Multiple Imputation, *Journal of Official Statistics*, 2013; 29: 425-465.

[10] Rubin, D.B. Discussion: Statistical Disclosure Limitation, *Journal of Official Statistics*, 1993; 9: 461-468.

[11] Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*, Hoboken, NJ: John Wiley & Sons, 1987.

[12] Raghunathan, T.E., Reiter, J.P., and Rubin, D.B. Multiple Imputation for Statistical Disclosure Limitation, *Journal of Official Statistics*, 2003; 19: 1-16.

[13] Reiter, J.P. Inference for Partially Synthetic, Public Use Microdata Sets, *Survey Methodology*, 2003; 29: 181-188.

[14] Reiter, J.P. Releasing Multiply Imputed, Synthetic Public Use Microdata: An Illustration and Empirical Study, *Journal of Royal Statistical Society, Series A*, 2005; 168: 185-205.

[15] Reiter, J.P., and Raghunathan, T.E. The Multiple Adaptations of Multiple Imputation, *Journal of American Statistical Association*, 2007; 102: 1462-1471.

[16] Drechsler, J. *Synthetic Datasets for Statistical Disclosure Control*, New York, NY: Springer, 2011.

[17] Reiter, J.P., and Kinney, S.K. Inferentially Valid, Partially Synthetic Data: Generating From Posterior Predictive Distributions Not Necessary, *Journal of Official Statistics*, 2012; 28: 583-590.

[18] Kinney, S.K., Reiter, J.P., Reznek, A.P., Miranda, J., Jarmin, R.S., and Abowd, J.M. Towards Unrestricted Public Use Business Microdata: The Synthetic Longitudinal Business Database, *International Statistical Review*, 2011; 79: 362-384.