Thailand Statistician
2014; 12(1): 1-23
http://statassoc.or.th
Contributed paper

Likelihood Based Inference Under Noise Multiplication

Martin Klein $^*[a]$**, Thomas Mathew** $[a, b]$**, Bimal Sinha** $[b, c]$

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

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

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

* corresponding author; e-mail: martin.klein@census.gov

Received: 19 June 2013
Accepted: 30 September 2013

Abstract

When statistical agencies release microdata to the public, a major concern is the control of disclosure risk, while ensuring utility in the released data. Often some statistical disclosure control methods such as data swapping, multiple imputation, top coding, and perturbation with random noise, are applied before releasing the data. This article develops methodology for data analysis when each original observation is multiplied by random noise for the purpose of statistical disclosure control. A parametric model is assumed, and specific details are provided for the exponential, normal and lognormal models. Our analysis shows that noise multiplied data can yield accurate inferences, and detailed simulation results provide guidance as to how the dispersion of the noise generating distribution affects accuracy of the inference.

Keywords: Confidentiality, EM algorithm, microdata, statistical disclosure limitation.

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.

1. Introduction

When survey organizations and statistical agencies release microdata to the public, a major concern is the control of disclosure risk, while simultaneously ensuring quality and utility of the released data. Very often some popular statistical disclosure control methods such as data swapping, multiple imputation (MI), top coding/bottom coding (especially for income data), and perturbation by random noise, are applied before releasing the data. Multiple imputation has been in existence for some time as a viable methodology to handle missing data (see Rubin [1]). Rubin [2] proposed to use multiple imputation as a method for sensitive data protection. The rigorous foundations for multiple imputation as a method for data protection were further developed (e.g., Drechsler [3]; Drechsler and Reiter [4]; Raghunathan, Reiter, and Rubin [5]; Reiter [6], [7], [8], [9]), and this still continues to be an active area of research. Noise perturbation by addition or multiplication has also been advocated by some statisticians as a possible data confidentiality protection mechanism (Hwang [10]; Little [11]; Kim and Winkler [12]); and this also continues to be an active area of research (e.g., Lin and Wise [13], Nayak, Sinha and Zayatz [14]; Sinha, Nayak and Zayatz [15]).

This article provides a comprehensive account of likelihood based data analysis methods under noise multiplication for drawing inference about unknown parameters and specific details are provided for the exponential, normal and lognormal models. We assume that the entire data set is noise multiplied, and our analysis shows that noise multiplication can provide accurate results by appropriately adjusting the variance of the noise generating distribution while protecting privacy of respondents. Detailed simulation results provide guidance as to how the noise variance affects accuracy of inference in several parametric settings.

Regarding application of noise multiplied data, the first public use microdata sample (PUMS) produced from the Survey of Business Owners (SBO) was released in August 2012 (`http://www.census.gov/econ/sbo/`), and noise multiplication was employed for confidentiality protection of some variables. Here

each record corresponds to a business surveyed in the 2007 SBO, and a number of variables are provided relating to firm size, business characteristics, and business owner characteristics. In this data product, a number of steps are taken to protect confidentiality of businesses, and the variables relating to receipts, payroll, and employment are rounded and multiplied by random noise prior to release [16].

Instead of applying noise multiplication (NM) to the *entire* data, resulting in *fully* noise-perturbed data, often times there are situations when a part of the data is sensitive and *must* not be released while the rest of the data can be used/released without any compromise. This is the set up of top coding, where values above a certain threshold C are suppressed and only the number of values in the data set above C are reported along with the actual values below C. This is precisely the scenario considered by An and Little [17], and they have developed data analysis methods based on multiple imputation of the data above C, in combination with the original values below C. In a separate communication we have developed in detail the likelihood based data analysis methods under noise multiplication of the values above C, along with the actual observations below C, and provided a comparison with An and Little's [17] procedure. Note that top coding is akin to the method of type I censoring which is widely used in reliability studies.

Here is the organization of the paper. In Section 2 we provide details of the statistical analysis for *fully* noise-perturbed data. After giving the general framework for the proposed methodology in Subsection 2.1, specific results for exponential, normal and lognormal appear in Subsections 2.2, 2.3, and 2.4, respectively. Simulation results for estimation of mean and variance under a lognormal distribution are presented in Section 3. Other simulation results, including an application using data from the 2000 U.S. Current Population Survey to illustrate the scope of the methods developed here and also a comparison with results obtained under multiple imputation, appear in Klein, Mathew, and Sinha [18]. We conclude the paper with some discussion in Section 4.

We end this section with two general observations. First, while standard and often *optimum* parametric inference can be drawn for the three chosen standard probability models based on unperturbed data, such an analysis is far from being close to optimum or even simple when noise multiplication (NM) is used. We have essentially relied on the asymptotic theory, providing enough

computational details of the maximum likelihood estimators (MLEs) and observed Fisher information matrices in each case. Second, we should point out that our approach to modify the microdata to protect the confidentiality of all records and carry out the analysis based on noise-modified microdata data is different from modifying the microdata when the goal is to release tables with frequency counts (Evans, Zayatz, and Slanta [19]). Moreover, the focus of this paper is on data analysis methods based on noise-modified microdata rather than on a study of the effectiveness of the procedures in protecting the data.

2. Data Analysis Under Full Noise Multiplication

2.1 General Framework

Perturbation of data by random noise for the purpose of statistical disclosure limitation has been discussed by Hwang [10]; Little [11]; and Kim and Winkler [12]. More recently, some results have emerged in a *nonparametric* setup for estimation of the moments, and for inference about the quantiles of a variable Y based on noise *multiplied* data (Nayak, Sinha and Zayatz [14]; Sinha, Nayak and Zayatz [15]). Briefly, Nayak, Sinha, and Zayatz [14] discussed at length various issues related to the statistical properties of random noise perturbation methods for data masking. Under the noise multiplication scenario, issues such as confidentiality protection, moment estimation, properties of balanced noise distribution, and effects on data quality and privacy protection in the context of tabular data were addressed at length. In a subsequent paper, Sinha, Nayak, and Zayatz [15] proposed some inferential procedures for quantile estimation based on noise multiplied micro data. It turns out that this is indeed a difficult inferential problem, and an empirical Bayes solution based on a nonparametric model was developed by the authors. Furthermore, Lin and Wise [13] developed methods for estimating regression parameters using noise multiplied data.

Consider a random variable Y having the density $f_\theta(y)$ where $f_\theta(y)$ is a parametric model with unknown parameter vector θ. Let R be a noise random variable having the completely known density $h(r)$. We assume that both $f_\theta(y)$ and $h(r)$ are densities of continuous distributions and we assume the support of $h(r)$ is nonnegative. The noise multiplied random variable is $Z = Y \times R$ and its

probability density function (pdf) is $g_\theta(z) = \int f_\theta(\frac{z}{r})h(r)r^{-1}dr$. Let $y_1, \ldots, y_n \sim iid \sim Y$ denote a random sample of size n from the distribution $f_\theta(y)$, and let $r_1, \ldots, r_n \sim iid \sim R$ be a set of noise random variables generated from $h(r)$. The noise multiplied version of y_i is $z_i = y_i \times r_i$, and thus, $z_1, \ldots, z_n$ can be thought of as a random sample from the distribution $g_\theta(z)$. For convenience, let $\mathbf{y} = (y_1, \ldots, y_n)$, $\mathbf{r} = (r_1, \ldots, r_n)$, and $\mathbf{z} = (z_1, \ldots, z_n)$.

When the components of the parameter vector θ have moment-type interpretations based on Y, they admit simple unbiased (not necessarily *optimum*) estimates based on $\mathbf{z}$ (Hwang [10]; Nayak, Sinha, and Zayatz [14]); however, efficient inference for θ based on $\mathbf{z}$ is far from being simple due mainly to the possible complexity of $g_\theta(z)$. In the context of masking by noise multiplication, unlike our setup where R has a specified noise distribution, independent of the data Y, Kim [20], Sullivan and Fuller [21], [22] and Little [11] dealt with the case when R is made data-dependent. This procedure, while it keeps intact certain basic moments of the original data, obviously renders considerable difficulty in the inference process. In this context, it is rather interesting to quote Little [11], which indeed provides a compelling motivation for our research: *Although a full likelihood-based analysis may not be feasible in many settings, I think the modeling perspective provides a useful basis for assessing simpler approximate methods. Future work might provide more detailed applications of the modeling approach to specific masking procedures*. Our goal here is to provide exact and approximate efficient inference procedures when noise multiplication is used as a data masking mechanism.

Here is an outline of our approach. We assume that the information available to the data user consists of the noise multiplied data $\mathbf{z}$, knowledge of the form of the parametric model $f_\theta(y)$, and knowledge of the noise generating distribution $h(r)$. Then, to circumvent the complexity of the *marginal* likelihood based on $\mathbf{z}$, we apply the EM algorithm to compute the MLE of θ (Dempster, Laird, and Rubin [23]; Little and Rubin [24]). We derive the observed Fisher information matrix and use it to estimate the standard deviation of the MLE. We construct approximate $1 - \alpha$ level confidence intervals for scalar parameters as (MLE $\pm\, z_{\alpha/2} \times$ estimated standard deviation of the MLE) where $z_{\alpha/2} = \Phi^{-1}(1 - \alpha/2)$ and $\Phi(\cdot)$ is the standard normal cumulative distribution function. We note that it is also possible to use a bootstrap procedure to estimate the variance of the MLE and to compute confidence intervals for unknown parame-

ters (Efron [25]; Efron and Tibshirani [26]). To apply the EM algorithm, we frame the analysis of noise multiplied data as a missing data problem. In order to do so, we define $\mathbf{u}_{\rm obs} = (z_1, ..., z_n)$, $\mathbf{u}_{\rm mis} = (r_1, ..., r_n)$, $\mathbf{u}_{\rm c} = (\mathbf{u}_{\rm obs}, \mathbf{u}_{\rm mis})$, to denote the observed data, missing data, and complete data, respectively. Using the notations in the previous paragraph, the complete data likelihood can obviously be expressed as $L(\theta|\mathbf{u}_{\rm c}) = \prod_{i=1}^{n} \left[f_\theta \left(\frac{z_i}{r_i} \right) \frac{h(r_i)}{r_i} \right]$ and the observed data likelihood as $L(\theta|\mathbf{u}_{\rm obs}) = \prod_{i=1}^{n} \left[\int f_\theta \left(\frac{z_i}{r_i} \right) \frac{h(r_i)}{r_i} dr_i \right]$.

Taking logarithm, if $\ell(\theta|\mathbf{u}_{\rm c}) = \ln L(\theta|\mathbf{u}_{\rm c}) = \sum_{i=1}^{n} \ln f_\theta(\frac{z_i}{r_i})$ (ignoring constants), the E-step is then carried out by starting with the estimate $\theta^{(t)}$ (at the $t^{\rm th}$ step) and computing

$$Q\left(\theta|\theta^{(t)}\right) = E_{\theta^{(t)}}\left[\ell(\theta|\mathbf{u}_{\rm c})|\mathbf{u}_{\rm obs}\right] = \sum_{i=1}^{n} E_{\theta^{(t)}}\left[\ln f_\theta\left(\frac{z_i}{r_i}\right)\Big| z_i\right] \tag{1}$$

and the M-step is carried out by maximizing $Q(\theta|\theta^{(t)})$ with respect to θ, resulting in $\theta^{(t+1)}$. It would be rather easy to evaluate the one dimensional integral (with respect to r) in $Q(\theta|\theta^{(t)})$ above, either explicitly or numerically. The iteration from $\theta^{(t)}$ to $\theta^{(t+1)}$ defined through the E and M-steps can then be run until a stopping criterion is met. For many choices of $f_\theta(y)$, the M-step will have a closed form. The E-step is also quite feasible since the one-dimensional integrals appearing in (1) are straightforward to evaluate using numerical or Monte Carlo methods. Details of computation of the MLEs, and the derivation of the observed Fisher information for the exponential, normal and lognormal models appear in the following subsections.

Remark. We have assumed above that the information available to the data user consists of (i) noise multiplied data $\mathbf{z}$, (ii) knowledge of the form of the underlying parametric model which generates the original data, and (iii) knowledge of the noise generating distribution. One of the reviewers has correctly pointed out that the assumption about the availability of this complete knowledge to the data users may not always hold! If data users are provided only with the noise-perturbed data (to protect confidentiality) and nothing else, perhaps not much can be done in terms of drawing valid statistical inference. However, with the added information about the noise distribution, some features (such as moments) of the original data can be estimated based on the noise multiplied data (Nayak, Sinha and Zayatz [14]).

2.2 Details for Exponential Data

We assume $Y \sim f_\theta(y) = \frac{1}{\theta} e^{-\frac{y}{\theta}}$, $0 < y < \infty$, where θ is the unknown positive scalar parameter. Then the joint pdf $g_\theta(z, r)$ of (Z, R) is $g_\theta(z, r) = \frac{1}{\theta} e^{-\frac{z}{r\theta}} h(r) r^{-1}$, and hence we have the following expressions for the marginal pdf $g_\theta(z)$ of Z, and for the conditional pdf $g_\theta(r|z)$ of R, given $Z = z$:

$$g_\theta(z) = \int_0^\infty \frac{1}{\theta} e^{-\frac{z}{r\theta}} h(r) r^{-1} dr, \;\; g_\theta(r|z) = \frac{e^{-\frac{z}{r\theta}} h(r) r^{-1}}{\int_0^\infty e^{-\frac{z}{w\theta}} h(w) w^{-1} dw}, \qquad (2)$$

for $0 < r < \infty$ and $0 < z < \infty$. Hence the complete data likelihood $L(\theta|\mathbf{u}_{\rm c})$ and loglikelihood $\ell(\theta|\mathbf{u}_{\rm c})$ can be expressed, respectively, as

$$L(\theta|\mathbf{u}_{\rm c}) = \frac{1}{\theta^n} e^{-\frac{1}{\theta}\sum_{i=1}^n \frac{z_i}{r_i}}, \;\; \ell(\theta|\mathbf{u}_{\rm c}) = -n \ln \theta - \frac{1}{\theta} \sum_{i=1}^n \frac{z_i}{r_i}.$$

The E and M-steps for computing the MLE of θ based on $\mathbf{z}$ are as follows.
E-step. We compute

$$Q(\theta|\theta^{(t)}) = E_{\theta^{(t)}}\left[\ell(\theta|\mathbf{u}_{\rm c})|\mathbf{u}_{\rm obs}\right] = -n\ln\theta - \frac{1}{\theta}\sum_{i=1}^n z_i E_{\theta^{(t)}}\left[\frac{1}{r_i}\Big| z_i\right] = -n\ln\theta - \frac{1}{\theta}\sum_{i=1}^n z_i \psi\left(\theta^{(t)}, z_i\right)$$

where $\psi(\theta^{(t)}, z_i) = E_{\theta^{(t)}}[\frac{1}{r_i}|z_i]$.
M-step. By maximizing $Q(\theta|\theta^{(t)})$ with respect to θ, we obtain the following equation which defines the sequence of EM iterations:

$$\theta^{(t+1)} = \frac{1}{n}\sum_{i=1}^n z_i \psi\left(\theta^{(t)}, z_i\right).$$

The expression for $\psi(\theta^{(t)}, z)$ follows directly from the conditional pdf of R, given $Z = z$, which is given by (2). Note that in the special case when $h(r)$ is the uniform$(1-\epsilon, 1+\epsilon)$ density, upon direct integration, $\psi(\theta^{(t)}, z)$ simplifies to

$$\psi_{\rm uniform}(\theta^{(t)}, z) = \frac{\frac{\theta^{(t)}}{z}\left[e^{-\frac{z}{\theta^{(t)}(1+\epsilon)}} - e^{-\frac{z}{\theta^{(t)}(1-\epsilon)}}\right]}{\int_{1-\epsilon}^{1+\epsilon} e^{-\frac{z}{w\theta^{(t)}}} \frac{dw}{w}}.$$

The observed Fisher information is $-\left[\sum_{i=1}^n \frac{\partial^2 \ln g_\theta(z_i)}{\partial\theta^2}\right]_{\theta=\hat\theta(\mathbf{z})}$ where $\hat\theta(\mathbf{z})$ is the MLE of θ computed based on $\mathbf{z}$. The expressions for the derivatives are:

$$\frac{\partial \ln g_\theta(z)}{\partial\theta} = -\frac{1}{\theta} + \frac{\int_0^\infty e^{-\frac{z}{r\theta}} \frac{z h(r)}{r^2\theta^2} dr}{\int_0^\infty e^{-\frac{z}{r\theta}} \frac{h(r)}{r} dr},$$

$$\frac{\partial^2 \ln g_\theta(z)}{\partial\theta^2} = \frac{1}{\theta^2} - \frac{2z}{\theta^3}\frac{\int_0^\infty e^{-\frac{z}{r\theta}} \frac{h(r)}{r^2} dr}{\int_0^\infty e^{-\frac{z}{r\theta}} \frac{h(r)}{r} dr} + \frac{z^2}{\theta^4}\frac{\int_0^\infty e^{-\frac{z}{r\theta}} \frac{h(r)}{r^3} dr}{\int_0^\infty e^{-\frac{z}{r\theta}} \frac{h(r)}{r} dr} - \frac{z^2}{\theta^4}\left[\frac{\int_0^\infty e^{-\frac{z}{r\theta}} \frac{h(r)}{r^2} dr}{\int_0^\infty e^{-\frac{z}{r\theta}} \frac{h(r)}{r} dr}\right]^2.$$

A remark is now in order because a special choice of the noise generating distribution $h(r)$ will lead to simplified inferential methods. We refer to this choice of $h(r)$ as a *customized* noise distribution, which is defined by

$$h_\delta(r) = \frac{\delta^{\delta+1}}{\Gamma(\delta+1)} r^{-\delta-2} e^{-\frac{\delta}{r}},\ 0 < r < \infty,\ \delta > 1.$$

It is easy to verify that $E(R) = 1$, as desired, and var$(R) = \sigma_r^2 = \frac{1}{\delta-1}$. We choose $\delta > 1$ for a desirable level of noise variation. A direct computation shows that the pdf of $g_\theta(z)$ takes the form

$$g_\theta(z) = \frac{\delta+1}{\theta} \times \frac{\delta^{\delta+1}}{(\frac{z}{\theta}+\delta)^{\delta+2}},\ 0 < z < \infty.$$

This readily leads to the following likelihood and log-likelihood based on $\mathbf{z}$:

$$L(\theta|\mathbf{u}_{\rm obs}) = \theta^{-n} \times \prod_{i=1}^{n} \left(\frac{z_i}{\theta}+\delta\right)^{-\delta-2},\ \ \ell(\theta|\mathbf{u}_{\rm obs}) = -n\ln\theta - (\delta+2)\sum_{i=1}^{n} \ln\left(\frac{z_i}{\theta}+\delta\right).$$

Then $\hat{\theta}(\mathbf{z})$, the maximum likelihood estimate of θ, can be directly computed in this case by solving the equation:

$$\frac{\partial \ell(\theta|\mathbf{u}_{\rm obs})}{\partial\theta} = -\frac{n}{\theta} + \frac{\delta+2}{\theta^2}\sum_{i=1}^{n} \frac{z_i}{(\frac{z_i}{\theta}+\delta)} = 0,$$

which can be simplified as

$$\sum_{i=1}^{n} \frac{z_i}{z_i+\theta\delta} = \frac{n}{\delta+2}.$$

The observed Fisher information about θ contained in the data $\mathbf{u}_{\rm obs}$, is now obtained as

$$-\left[\frac{\partial^2 \ell(\theta|\mathbf{u}_{\rm obs})}{\partial\theta^2}\right]_{\theta=\hat{\theta}(\mathbf{z})} = \left[\frac{n(\delta+1)}{\theta^2} - \delta^2(\delta+2)\sum_{i=1}^{n}\frac{1}{(z_i+\theta\delta)^2}\right]_{\theta=\hat{\theta}(\mathbf{z})}.$$

2.3 Details for Normal Data

We assume $Y \sim f_\theta(y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$, where $\theta = (\mu, \sigma^2)$ is the unknown parameter vector. Then the joint pdf $g_\theta(z,r)$ of (Z,R) is $g_\theta(z,r) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(\frac{z}{r}-\mu)^2}{2\sigma^2}} h(r) r^{-1}$, and hence we have the following expressions for the marginal pdf $g_\theta(z)$ of Z,

and for the conditional pdf $g_\theta(r|z)$ of R, given $Z = z$:

$$g_\theta(z) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^\infty e^{-\frac{(\frac{z}{r}-\mu)^2}{2\sigma^2}} h(r) r^{-1} dr, \;\; g_\theta(r|z) = \frac{e^{-\frac{(\frac{z}{r}-\mu)^2}{2\sigma^2}} h(r) r^{-1}}{\int_0^\infty e^{-\frac{(\frac{z}{w}-\mu)^2}{2\sigma^2}} h(w) w^{-1} dw}, \tag{3}$$

for $0 < r < \infty$ and $-\infty < z < \infty$. Hence the complete data likelihood and loglikelihood can now be expressed, respectively, as

$$L(\theta|\mathbf{u}_{\rm c}) = \frac{1}{\sigma^n} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n \left(\frac{z_i}{r_i}-\mu\right)^2}, \;\; \ell(\theta|\mathbf{u}_{\rm c}) = -n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^n \left(\frac{z_i}{r_i}-\mu\right)^2.$$

The E and M-steps for computing the MLE of θ based on $\mathbf{z}$ are as follows.

E-step. We compute

$$Q(\theta|\theta^{(t)}) = E_{\theta^{(t)}}\left[\ell\left(\theta|\mathbf{u}_{\rm c}\right)|\mathbf{u}_{\rm obs}\right] = -n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^n \left[z_i^2\psi_2\left(\theta^{(t)}, z_i\right) - 2\mu z_i \psi_1(\theta^{(t)}, z_i) + \mu^2\right],$$

where $\psi_1\left(\theta^{(t)}, z_i\right) = E_{\theta^{(t)}}\left[\frac{1}{r_i}|z_i\right]$ and $\psi_2\left(\theta^{(t)}, z_i\right) = E_{\theta^{(t)}}\left[\frac{1}{r_i^2}|z_i\right]$, and expressions for these two quantities can be obtained based on $g_\theta(r|z)$ given in (3).

M-step. By maximizing $Q(\theta|\theta^{(t)})$ with respect to θ we obtain the following equations which define the sequence of EM iterations:

$$\mu^{(t+1)} = \frac{1}{n}\sum_{i=1}^n z_i \psi_1\left(\theta^{(t)}, z_i\right), \;\; [\sigma^{(t+1)}]^2 = \frac{1}{n}\sum_{i=1}^n z_i^2 \psi_2\left(\theta^{(t)}, z_i\right) - [\mu^{(t+1)}]^2.$$

The observed Fisher information is

$$-\sum_{i=1}^n \begin{pmatrix} \frac{\partial^2 \ln g_\theta(z_i)}{\partial\mu^2} & \frac{\partial^2 \ln g_\theta(z_i)}{\partial\mu\partial(\sigma^2)} \\ \frac{\partial^2 \ln g_\theta(z_i)}{\partial\mu\partial(\sigma^2)} & \frac{\partial^2 \ln g_\theta(z_i)}{\partial(\sigma^2)^2} \end{pmatrix}_{\theta=\hat{\theta(\mathbf{z})}}$$

where $\hat{\theta}(\mathbf{z})$ is the MLE of θ based on $\mathbf{z}$. The expressions for the derivatives are:

$$\frac{\partial \ln g_\theta(z)}{\partial\mu} = \frac{1}{\sigma^2}\left[\frac{\int_0^\infty g_\theta(z,r)\frac{z}{r}dr}{g_\theta(z)} - \mu g(z)\right],$$

$$\frac{\partial^2 \ln g_\theta(z)}{\partial\mu^2} = \frac{z^2}{\sigma^4 g_\theta(z)}\left[\int_0^\infty g_\theta(z,r)\frac{dr}{r^2} - \frac{(\int_0^\infty g_\theta(z,r)\frac{dr}{r})^2}{g_\theta(z)}\right] - \frac{1}{\sigma^2},$$

$$\frac{\partial \ln g_\theta(z)}{\partial\sigma^2} = \frac{1}{2\sigma^4}\frac{\int_0^\infty g_\theta(z,r)(\frac{z}{r}-\mu)^2 dr}{g_\theta(z)} - \frac{1}{2\sigma^2},$$

$$\frac{\partial^2 \ln g_\theta(z)}{(\partial\sigma^2)^2} = -\frac{1}{\sigma^6}\frac{\int_0^\infty g_\theta(z,r)(\frac{z}{r}-\mu)^2 dr}{g_\theta(z)} + \frac{1}{4\sigma^8}\frac{\int_0^\infty g_\theta(z,r)(\frac{z}{r}-\mu)^4 dr}{g_\theta(z)} - \frac{1}{4\sigma^8}\left[\frac{\int_0^\infty g_\theta(z,r)(\frac{z}{r}-\mu)^2 dr}{g_\theta(z)}\right]^2 + \frac{1}{2\sigma^4},$$

$$\begin{aligned}\frac{\partial^2 \ln g_\theta(z)}{\partial\mu\partial\sigma^2} &= \frac{\partial}{\partial\mu}\left[\frac{1}{2\sigma^4}\frac{\int_0^\infty g_\theta(z,r)(\frac{z}{r}-\mu)^2 dr}{g_\theta(z)}\right] \\ &= \frac{1}{2\sigma^4 g_\theta(z)}\left[-2\int_0^\infty g_\theta(z,r)(\frac{z}{r}-\mu)dr + \frac{1}{\sigma^2}\int_0^\infty g_\theta(z,r)(\frac{z}{r}-\mu)^3 dr\right] \\ &\quad - \frac{1}{2\sigma^6}\frac{[\int_0^\infty g_\theta(z,r)(\frac{z}{r}-\mu)dr][\int_0^\infty g_\theta(z,r)(\frac{z}{r}-\mu)^2 dr]}{g_\theta^2(z)} \\ &= -\frac{1}{\sigma^4}\frac{\int_0^\infty g_\theta(z,r)(\frac{z}{r}-\mu)dr}{g_\theta(z)} + \frac{1}{2\sigma^6}\frac{\int_0^\infty g_\theta(z,r)(\frac{z}{r}-\mu)^3 dr}{g_\theta(z)} \\ &\quad - \frac{1}{2\sigma^6}\frac{[\int_0^\infty g_\theta(z,r)(\frac{z}{r}-\mu)dr][\int_0^\infty g_\theta(z,r)(\frac{z}{r}-\mu)^2 dr]}{g_\theta^2(z)}.\end{aligned}$$

2.4 Details for Lognormal Data

We assume $Y \sim f_\theta(y) = \frac{1}{y\sigma\sqrt{2\pi}}e^{-\frac{(\ln y-\mu)^2}{2\sigma^2}}$, $0 < y < \infty$, where $\theta = (\mu, \sigma^2)$ is the unknown parameter vector. Then the joint pdf $g_\theta(z,r)$ of (Z,R) is $g_\theta(z,r) = \frac{1}{z\sigma\sqrt{2\pi}}e^{-\frac{(\ln z-\ln r-\mu)^2}{2\sigma^2}}h(r)$, and hence we have the following expressions for the marginal pdf $g_\theta(z)$ of Z, and for the conditional pdf $g_\theta(r|z)$ of R, given $Z = z$:

$$g_\theta(z) = \frac{1}{z\sigma\sqrt{2\pi}}\int_0^\infty e^{-\frac{(\ln z-\ln r-\mu)^2}{2\sigma^2}}h(r)dr, \quad g_\theta(r|z) = \frac{e^{-\frac{(\ln z-\ln r-\mu)^2}{2\sigma^2}}h(r)}{\int_0^\infty e^{-\frac{(\ln z-\ln r-\mu)^2}{2\sigma^2}}h(r)dr}, \tag{4}$$

for $0 < r < \infty$ and $0 < z < \infty$. The complete data likelihood and log-likelihood can be expressed as

$$L(\theta|\mathbf{u}_c) = \frac{1}{\sigma^n}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n(\ln\frac{z_i}{r_i}-\mu)^2}, \quad \ell(\theta|\mathbf{u}_c) = -n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^n(\ln\frac{z_i}{r_i})^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^n\ln\frac{z_i}{r_i} - \frac{n\mu^2}{2\sigma^2}.$$

The E and M-steps for computing the MLE of θ based on $\mathbf{z}$ are as follows.
E-step. We compute

$$Q(\theta|\theta^{(t)}) = E_{\theta^{(t)}}\left[\ell(\theta|\mathbf{u}_c)|\mathbf{u}_{obs}\right] = -n\ln\sigma - \frac{1}{2\sigma^2}\sum_{i=1}^n\left[\psi_2\left(\theta^{(t)}, z_i\right) + \frac{\mu}{\sigma^2}\psi_1\left(\theta^{(t)}, z_i\right)\right] - \frac{n\mu^2}{2\sigma^2},$$

where $\psi_1(\theta^{(t)}, z_i) = E_{\theta^{(t)}}[\ln \frac{z_i}{r_i}|z_i]$ and $\psi_2(\theta^{(t)}, z_i) = E_{\theta^{(t)}}[(\ln \frac{z_i}{r_i})^2|z_i]$, and these two quantities can be readily computed based on the conditional pdf of R, given z, mentioned in (4).

M-step. By maximizing $Q(\theta|\theta^{(t)})$ with respect to θ, we obtain the following equations which define the sequence of EM iterations:

$$\mu^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n}\psi_1(\theta^{(t)}, z_i), \quad (\sigma^{(t+1)})^2 = \frac{1}{n}\sum_{i=1}^{n}\psi_2(\theta^{(t)}, z_i) - (\mu^{(t+1)})^2.$$

The observed Fisher information is

$$-\sum_{i=1}^{n}\begin{pmatrix} \frac{\partial^2 \ln g_\theta(z_i)}{\partial \mu^2} & \frac{\partial^2 \ln g_\theta(z_i)}{\partial \mu \partial(\sigma^2)} \\ \frac{\partial^2 \ln g_\theta(z_i)}{\partial \mu \partial(\sigma^2)} & \frac{\partial^2 \ln g_\theta(z_i)}{\partial(\sigma^2)^2} \end{pmatrix}_{\theta=\hat{\theta(\mathbf{z})}}$$

where $\hat{\theta}(\mathbf{z})$ is the MLE of θ based on $\mathbf{z}$. The expressions for the derivatives are:

$$\frac{\partial \ln g_\theta(z)}{\partial \mu} = \frac{\int_0^\infty e^{-\frac{(\ln z - \ln r - \mu)^2}{2\sigma^2}} \frac{\ln z - \ln r - \mu}{\sigma^2} h(r)dr}{\int_0^\infty e^{-\frac{(\ln z - \ln r - \mu)^2}{2\sigma^2}} h(r)dr},$$

$$\frac{\partial^2 \ln g_\theta(z)}{\partial \mu^2} = -\frac{1}{\sigma^2} + \frac{\int_0^\infty e^{-\frac{(\ln z - \ln r - \mu)^2}{2\sigma^2}} (\frac{\ln z - \ln r - \mu}{\sigma^2})^2 h(r)dr}{\int_0^\infty e^{-\frac{(\ln z - \ln r - \mu)^2}{2\sigma^2}} h(r)dr} - \left[\frac{\int_0^\infty e^{-\frac{(\ln z - \ln r - \mu)^2}{2\sigma^2}} \frac{\ln z - \ln r - \mu}{\sigma^2} h(r)dr}{\int_0^\infty e^{-\frac{(\ln z - \ln r - \mu)^2}{2\sigma^2}} h(r)dr}\right]^2,$$

$$\frac{\partial^2 \ln g_\theta(z)}{\partial \mu \partial \sigma^2} = -\frac{1}{\sigma^4} \times \frac{\int_0^\infty e^{-\frac{(\ln z - \ln r - \mu)^2}{2\sigma^2}} (\ln z - \ln r - \mu) h(r)dr}{\int_0^\infty e^{-\frac{(\ln z - \ln r - \mu)^2}{2\sigma^2}} h(r)dr}$$
$$+ \frac{1}{2\sigma^6} \times \frac{\int_0^\infty e^{-\frac{(\ln z - \ln r - \mu)^2}{2\sigma^2}} (\ln z - \ln r - \mu)^3 h(r)dr}{\int_0^\infty e^{-\frac{(\ln z - \ln r - \mu)^2}{2\sigma^2}} h(r)dr}$$
$$- \frac{1}{2\sigma^6} \frac{\left[\int_0^\infty e^{-\frac{(\ln z - \ln r - \mu)^2}{2\sigma^2}} (\ln z - \ln r - \mu) h(r)dr \times \int_0^\infty e^{-\frac{(\ln z - \ln r - \mu)^2}{2\sigma^2}} (\ln(z/r) - \mu)^2 h(r)dr\right]}{\left[\int_0^\infty e^{-\frac{(\ln(z/r) - \mu)^2}{2\sigma^2}} h(r)dr\right]^2},$$

$$\frac{\partial \ln g_\theta(z)}{\partial \sigma^2} = -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4} \frac{\int_0^\infty e^{-\frac{(\ln z - \ln r - \mu)^2}{2\sigma^2}} (\ln z - \ln r - \mu)^2 h(r)dr}{\int_0^\infty e^{-\frac{(\ln z - \ln r - \mu)^2}{2\sigma^2}} h(r)dr},$$

$$\frac{\partial^2 \ln g_\theta(z)}{(\partial\sigma^2)^2} = \frac{1}{2\sigma^4} + \frac{1}{\sigma^6}\frac{\int_0^\infty e^{-\frac{(\ln z - \ln r - \mu)^2}{2\sigma^2}}(\ln z - \ln r - \mu)^2 h(r)dr}{\int_0^\infty e^{-\frac{(\ln z - \ln r - \mu)^2}{2\sigma^2}} h(r)dr}$$
$$+ \frac{1}{4\sigma^8}\frac{\int_0^\infty e^{-\frac{(\ln z - \ln r - \mu)^2}{2\sigma^2}}(\ln z - \ln r - \mu)^4 h(r)dr}{\int_0^\infty e^{-\frac{(\ln z - \ln r - \mu)^2}{2\sigma^2}} h(r)dr}$$
$$- \frac{1}{4\sigma^8}\left[\frac{\int_0^\infty e^{-\frac{(\ln z - \ln r - \mu)^2}{2\sigma^2}}(\ln z - \ln r - \mu)^2 h(r)dr}{\int_0^\infty e^{-\frac{(\ln z - \ln r - \mu)^2}{2\sigma^2}} h(r)dr}\right]^2.$$

A customized noise distribution, permitting closed form evaluation of $g_\theta(z)$, is obtained as follows. Since Y follows a lognormal distribution, implying $\ln Y \sim N(\mu, \sigma^2)$ and since $Z = Y \times R$, we choose $R \sim \text{lognormal}$ with $\ln R \sim N(-\frac{\psi^2}{2}, \psi^2)$. Then $E(R) = 1$ and var$(R) = \sigma_r^2 = e^{\psi^2} - 1$. This readily yields

$$\ln Z \sim N(\mu_z = \mu - \frac{\psi^2}{2}, \sigma_z^2 = \sigma^2 + \psi^2). \tag{5}$$

Since the MLEs of μ_z and σ_z^2 are $\hat{\mu}_z = \frac{1}{n}\sum_{i=1}^n \ln z_i$ and $\hat{\sigma}_z^2 = \frac{1}{n}\sum_{i=1}^n (\ln z_i - \hat{\mu}_z)^2$, we obtain the MLEs of μ and σ^2 as $\hat{\mu} = \hat{\mu}_z + \frac{\psi^2}{2}$ and $\hat{\sigma}^2 = \hat{\sigma}_z^2 - \psi^2$. The estimated variance-covariance matrix of $(\hat{\mu}_z, \hat{\sigma}_z^2)$ is obtained from the observed Fisher information matrix for (μ_z, σ_z^2), given by

$$I_{\text{obs}}(\hat{\mu}_z, \hat{\sigma}_z^2) = \begin{pmatrix} \frac{n}{\hat{\sigma}_z^2} & 0 \\ 0 & \frac{n}{2\hat{\sigma}_z^4} \end{pmatrix}.$$

The estimated variances of meaningful and useful functions $g(.,.)$ of μ_z and σ_z^2 can be easily obtained by the delta method. We provide below examples of such functions.

1. $g(\mu_z, \sigma_z^2) = e^{\mu + \frac{\sigma^2}{2}} = e^{\mu_z + \frac{\psi^2}{2} + \frac{\sigma^2}{2}}$
2. $g(\mu_z, \sigma_z^2) = e^{2\mu + 2\sigma^2} - e^{2\mu + \sigma^2} = e^{2\mu_z + 2\sigma_z^2 - \psi^2} - e^{2\mu_z + \sigma_z^2}$
3. $g(\mu_z, \sigma_z^2) = e^{\mu + c\sigma} = e^{\mu_z + \frac{\psi^2}{2} + c(\sigma_z^2 - \psi^2)^{\frac{1}{2}}}$

Note that the first and second functions are, respectively, the mean and variance of the original lognormal random variable Y, expressed in terms of the mean and variance of the random variable $\ln Z$, specified in (5). Similarly, the third function is a percentile of Y. Recall once again that data are not available on Y, but on the perturbed quantity Z. In other words, inference on the above

functions related to the distribution of Y can now be carried out using the data on Z.

3. Simulation Study

In order to evaluate the finite sample performance of our proposed likelihood based analysis of noise multiplied data, we performed a simulation study in the context of the lognormal distribution with $f_\theta(y) = \frac{1}{y\sigma\sqrt{2\pi}} \exp[-(\log y - \mu)^2/(2\sigma^2)]$, $y > 0$, $\theta = (\mu, \sigma^2)$, and we fix $\mu = 0$, $\sigma^2 = 1$. Details of a simulation study that includes the exponential and normal populations appear in the technical report Klein, Mathew, and Sinha [18]. In the case of the exponential and normal populations, the major findings of the simulation study were generally similar to those for the lognormal. Our observations on the simulation results are based on the results appearing in Tables 1 - 4. The rows in Tables 1 - 4 show results for the following methods:

UD: Analysis based on unperturbed data with no masking.

MI: Analysis based on synthetic data created via multiple imputation using the methodology of Reiter [6]. Here, multiple imputation is used to create $m = 50$ synthetic datasets. The value $m = 50$ may be larger than what is typically used in practice, but we select a large value in order to get a clear picture of the accuracy of inference under the MI method. Specific details of our implementation of the MI method in this scenario appear in the technical report Klein, Mathew, and Sinha [18]. As discussed in Section 1, the method of creating synthetic data via multiple imputation is a popular technique for statistical disclosure limitation. Therefore we include this method in the simulation study so that we may compare it with our proposed noise multiplication method.

NM10U, NM20U, NM30U, NM40U, NM50U, NM60U, NM70U, NM80U, NM90U: These nine rows indicate analysis based on the proposed noise multiplication method presented in Section 2 (specifically, Subsection 2.4 for the lognormal population), where $h(r)$ is the uniform$(1-\epsilon, 1+\epsilon)$ density. For NM10U we take $\epsilon = 0.10$, for NM20U we take $\epsilon = 0.20$, and so on.

NM10C, NM20C, NM30C, NM40C, NM50C, NM60C, NM70C, NM80C, NM90C: These nine rows indicate analysis based on the proposed noise multiplication method presented in Section 2 where $h(r)$ is the customized

noise distribution for the lognormal as defined in Subsection 2.4 (that is, $\ln R \sim N(-\frac{\psi^2}{2}, \psi^2)$). For NM10C we choose ψ so that the variance of R equals the variance of the uniform$(1-0.1, 1+0.1)$ distribution, for NM20C we choose ψ so that the variance of R equals the variance of the uniform$(1-0.2, 1+0.2)$ distribution, and so on.

Tables 1 and 2 show results when the parameter of interest is μ, and Tables 3 and 4 show results when the parameter of interest is σ^2. Tables 1 and 3 give results for the sample sizes $n = 30$ and 50, respectively; while Tables 2 and 4 show results for $n = 100$ and 200, respectively. For the point estimators of the parameters, the following quantities were estimated by Monte Carlo simulation based on 5000 iterations: the root mean squared error (RMSE), bias, standard deviation (SD), mean of estimated standard deviation ($\widehat{\text{SD}}$), coverage probability of the nominal 0.95 level confidence interval (Cvg.), and expected length of the confidence interval relative to the expected length of the confidence interval computed on the unperturbed data (Rel. Len.). To facilitate a comparison of results, the results for unperturbed data are based on MLEs, observed Fisher information, and confidence intervals of the form (MLE $\pm 1.96 \times$ estimated standard deviation of the MLE). For the EM algorithm, the stopping criterion used was $\max\left\{\left|\mu^{(t)} - \mu^{(t+1)}\right|, \left|(\sigma^{(t)})^2 - (\sigma^{(t+1)})^2\right|\right\} \leq 10^{-5}$. The statistical computing software R [27] was used for all computations, and the *integrate* function in R was used to evaluate the required univariate integrals that could not be obtained in a closed form. The following is a summary of the findings of the simulation study.

1. For estimating μ, Tables 1 and 2 indicate that the noise multiplication method gives valid results in the chosen simulation settings. In each of the scenarios of Tables 1 and 2, the bias is close to zero, SD for the noise multiplication method is similar to SD for the unperturbed data, and Cvg. is close to the nominal coverage probability of 0.95. Comparing $\widehat{\text{SD}}$ and SD in Tables 1 and 2, we see that under the noise multiplication method, the bias in the estimated standard deviation is generally small. As one would expect for noise multiplication, RMSE, SD, and Rel. Len. increase as the dispersion in the noise generating distribution increases.

2. For estimating σ^2, most of the observations noted above (for the case of estimating μ) continue to hold, except that in Table 3 when $n = 30$ and

50, we notice that Cvg. under the noise multiplication method is below the nominal level of 0.95. However, Cvg. under the unperturbed data (which is also based on a Wald-type confidence interval) is also below 0.95 because the sample sizes $n = 30$ and 50 are too small for the sampling distribution of the MLE of σ^2 to be well approximated by normality. Thus, in Table 3, even though the noise multiplication method yields confidence intervals with low coverage, it still provides inference comparable to that from unperturbed data.

3. Under noise multiplication we find that for the methods NM10U and NM10C where the dispersion in the noise generating distribution is small, the results are nearly identical to the results for the unperturbed data.

4. Generally we observe that when the variance of the noise generating distribution is small, the customized noise distribution of Subsection 2.4, and the uniform$(1-\epsilon, 1+\epsilon)$ noise distribution both tend to give similar results. As this variance increases though, we find that the customized noise distribution yields *more* accurate inference than the uniform$(1-\epsilon, 1+\epsilon)$ noise.

5. In all the simulation settings considered, the MI method tends to give valid results in terms of the quantities considered in Tables 1 and 4. In each case of noise multiplication, one can select a value for the variance of the noise distribution, at which noise multiplication gives similar results as MI. For instance, in Table 1 when $n = 50$, the methods NM50U and NM50C give similar inference as the MI method. We also notice that in Table 3, while the unperturbed data and noise multiplication method give Cvg. below the nominal value, the MI method maintains the nominal coverage probability more closely.

6. Finally, we remark that the EM algorithm used to calculate the MLE tended to be quite stable, and also converged rapidly.

4. Discussion

How to alter data before releasing it to the public continues to be a vital concern of statistical agencies in order to minimize the risk of disclosure. However, this goal has to be balanced with the interest of preserving the utility of

the released data. Our main goal in this paper has been to develop rigorous data analysis methodology for noise multiplied data under noise multiplication of each observation in the dataset. We have developed all the necessary theoretical results for analyzing noise multiplied data, when the original (unmasked) sample arises from three standard parametric distributions: exponential, normal and lognormal. Furthermore, our basic framework is quite general, and can be applied to other parametric models as well. Our simulation results indicate that when a large noise variance providing a high degree of protection against disclosure is used, inference is typically less accurate but still valid (i.e., bias is small, confidence interval coverage probability is maintained at the nominal level, etc.). The results of this article provide guidance regarding the effect of noise variance on data quality, however the impact of the noise variance on disclosure risk requires further investigation. It should be noted that the methodology presented in Section 2 applies to a general noise generating distribution $h(r)$. Therefore, the statistical agency can choose this distribution in order to achieve a desired balance between accuracy of inference and protection against disclosure. For instance, a choice of $h(r)$ that has no mass around 1 would provide a lower bound on the relative distance between each original and its noise multiplied value. An appealing feature of noise multiplication is its flexibility; the noise distribution $h(r)$ acts as a tuning mechanism.

Acknowledgement: The authors thank Paul Massell and three anonymous referees for providing helpful comments on a previous draft, and they thank Joseph Schafer, Eric Slud, Yves Thibaudeau, Tommy Wright and Laura Zayatz for encouragement.

References

[1] Rubin, D.B. *Multiple Imputation for Nonresponse in Surveys*, New York, NY: Wiley, 1987.

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

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

[4] Drechsler, J., and Reiter, J.P. Sampling With Synthesis: A New Approach for Releasing Public Use Census Microdata, *Journal of the American Statistical Association*, 2010; 105: 1347-1357.

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

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

[7] Reiter, J.P. Simultaneous Use of Multiple Imputation for Missing Data and Disclosure Limitation, *Survey Methodology*, 2004; 30: 235-242.

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

[9] Reiter, J.P. Significance Tests for Multi-Component Estimands From Multiply-Imputed, Synthetic Microdata, *Journal of Statistical Planning and Inference*, 2005; 131: 365-377.

[10] Hwang, J.T. Multiplicative Errors-in-Variables Models With Applications to Recent Data Released by the U.S. Department of Energy, *Journal of American Statistical Association*, 1986; 81: 680-688.

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

[12] Kim, 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 `www.census.gov/srd/papers/pdf/rrs2003-01.pdf` [accessed September 3, 2013].

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

[14] 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.

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

[16] *2007 Survey of Business Owners (SBO) Public Use Microdata Sample (PUMS) Data Users Guide (2012)*, Available from URL: `http://www2.census.gov/econ/sbo/07/pums/2007_sbo_pums_users_guide.pdf` [accessed September 3, 2013].

[17] An, D., and Little, R.J.A. Multiple Imputation: An Alternative to Top Coding for Statistical Disclosure Control, *Journal of the Royal Statistical Society*, Series A, 2007; 170: 923-940.

[18] Klein, M., Mathew, T., and Sinha, B. A Comparison of Statistical Disclosure Control Methods: Multiple Imputation Versus Noise Multiplication, *Center for Statistical Research & Methodology, Research and Methodology Directorate Research Report Series (Statistics #2013-02), U.S. Census Bureau*: 2013. Available from URL: `http://www.census.gov/srd/papers/pdf/rrs2013-02.pdf` [accessed September 3, 2013].

[19] Evans, T., Zayatz, L., and Slanta, J. Using Noise for Disclosure Limitation of Establishment Tabular Data, *Journal of Official Statistics*, 1998; 14: 537-551.

[20] Kim, J. A Method for Limiting Disclosure of Microdata Based on Random Noise and Transformation, *Proceedings of the Survey Research Methods Section*, Alexandria, VA, 1986; American Statistical Association: 370-374.

[21] Sullivan, G., and Fuller, W.A. The Use of Measurement Error to Avoid Disclosure, *Proceedings of the Survey Research Methods Section*, Alexandria, V.A, 1989; American Statistical Association: 802-807.

[22] Sullivan, G., and Fuller, W.A. Construction of Masking Error for Categorical Variables, *Proceedings of the Survey Research Methods Section*, Alexandria, V.A, 1990; American Statistical Association: 435-439.

[23] Dempster, A.P., Laird, N.M., and Rubin, D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm, *Journal of the Royal Statistical Society. Series B*, 1977; 39: 1-38.

[24] Little, R.J.A., and Rubin, D.B. *Statistical Analysis With Missing Data*, second edition, New York, NY: Wiley, 2002.

[25] Efron, B. Bootstrap Methods: Another Look at the Jackknife, *Annals of Statistics*, 1979; 7: 1-26.

[26] Efron, B., and Tibshirani, R.J. *An Introduction to the Bootstrap*, New York, NY: Chapman & Hall/CRC, 1994.

[27] R. Development Core Team (2011). R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. URL http://www.R-project.org/.

Table 1: Inference for the log scale mean μ based on fully masked $LN(\mu = 0, \sigma^2 = 1)$ data

	Results for $n = 30$						Results for $n = 50$					
	RMSE	Bias	SD	$\widehat{\text{SD}}$	Cvg.	Rel.	RMSE	Bias	SD	$\widehat{\text{SD}}$	Cvg.	Rel.
	$\times 10^3$	$\times 10^3$	$\times 10^3$	$\times 10^3$	%	Len.	$\times 10^3$	$\times 10^3$	$\times 10^3$	$\times 10^3$	%	Len.
UD	180.08	-0.10	180.08	178.24	94.36	1.000	142.92	0.95	142.92	139.29	94.06	1.000
MI	184.65	-0.71	184.65	188.34	94.92	1.057	145.79	1.19	145.78	145.04	94.64	1.041
NM10U	180.29	-0.32	180.29	178.55	94.18	1.002	143.21	0.90	143.21	139.51	94.06	1.002
NM10C	180.46	-0.20	180.46	178.58	94.32	1.002	143.27	1.02	143.26	139.52	94.14	1.002
NM20U	180.99	-0.37	180.99	179.34	93.98	1.006	143.30	1.07	143.29	140.24	94.10	1.007
NM20C	181.34	-0.13	181.34	179.35	94.10	1.006	143.90	0.86	143.90	140.26	93.90	1.007
NM30U	183.34	0.36	183.34	180.93	94.20	1.015	144.94	1.68	144.93	141.52	94.22	1.016
NM30C	182.12	0.77	182.11	180.89	94.42	1.015	145.26	1.37	145.25	141.35	93.88	1.015
NM40U	184.28	0.46	184.28	183.26	94.32	1.028	145.78	1.85	145.77	143.30	94.42	1.029
NM40C	184.37	-0.38	184.37	182.93	94.16	1.026	146.27	1.28	146.26	142.78	94.24	1.025
NM50U	189.38	-0.24	189.38	186.52	94.10	1.046	149.82	0.85	149.81	145.60	93.96	1.045
NM50C	188.88	-0.51	188.88	185.02	94.40	1.038	149.54	1.31	149.54	144.87	93.80	1.040
NM60U	193.14	-0.34	193.14	190.76	94.24	1.070	152.49	1.37	152.48	149.14	93.82	1.071
NM60C	189.61	0.43	189.61	188.21	94.10	1.056	151.02	0.31	151.02	147.18	93.96	1.057
NM70U	199.34	1.13	199.34	196.17	93.90	1.101	159.06	-0.00	159.06	153.78	93.44	1.104
NM70C	195.07	-1.62	195.06	191.19	93.88	1.073	153.84	1.00	153.83	149.67	93.90	1.075
NM80U	207.71	2.08	207.70	204.14	93.90	1.145	165.26	0.68	165.26	159.88	94.06	1.148
NM80C	199.22	0.09	199.22	194.36	93.42	1.090	157.14	0.27	157.14	152.23	93.92	1.093
NM90U	223.68	3.98	223.64	217.41	93.98	1.220	172.53	2.11	172.52	170.03	94.06	1.221
NM90C	201.57	1.48	201.56	198.56	94.04	1.114	159.67	1.63	159.66	154.99	93.56	1.113

Table 2: Inference for the log scale mean μ based on fully masked $LN(\mu = 0, \sigma^2 = 1)$ data

	Results for $n = 100$						Results for $n = 200$					
	RMSE	Bias	SD	$\widehat{\text{SD}}$	Cvg.	Rel.	RMSE	Bias	SD	$\widehat{\text{SD}}$	Cvg.	Rel.
	$\times 10^3$	$\times 10^3$	$\times 10^3$	$\times 10^3$	%	Len.	$\times 10^3$	$\times 10^3$	$\times 10^3$	$\times 10^3$	%	Len.
UD	101.60	0.84	101.59	99.27	94.00	1.000	70.05	0.68	70.05	70.51	95.10	1.000
MI	104.05	0.75	104.04	102.27	94.26	1.030	71.57	0.79	71.57	72.30	95.24	1.025
NM10U	101.71	0.92	101.70	99.45	94.22	1.002	70.15	0.61	70.15	70.63	95.22	1.002
NM10C	101.80	1.05	101.80	99.43	94.22	1.002	70.21	0.66	70.20	70.64	95.08	1.002
NM20U	102.54	0.98	102.54	99.96	94.32	1.007	70.43	0.62	70.43	70.98	95.14	1.007
NM20C	102.23	1.04	102.22	99.95	94.20	1.007	70.54	0.62	70.53	70.99	95.18	1.007
NM30U	102.90	0.55	102.90	100.77	94.24	1.015	71.00	0.42	71.00	71.61	95.24	1.016
NM30C	102.83	0.90	102.82	100.71	94.34	1.014	71.07	0.82	71.06	71.55	95.30	1.015
NM40U	104.33	1.00	104.32	102.07	94.32	1.028	72.49	0.57	72.49	72.53	95.24	1.029
NM40C	103.45	0.78	103.45	101.80	94.12	1.025	71.89	0.24	71.89	72.36	95.42	1.026
NM50U	106.55	1.28	106.55	103.87	94.12	1.046	73.22	0.62	73.21	73.76	95.16	1.046
NM50C	105.12	0.59	105.12	103.13	94.08	1.039	73.05	0.65	73.05	73.31	95.12	1.040
NM60U	108.20	2.03	108.18	106.27	94.24	1.070	75.41	0.55	75.40	75.44	95.16	1.070
NM60C	107.52	0.50	107.52	104.81	94.08	1.056	74.51	0.08	74.51	74.38	94.84	1.055
NM70U	111.48	-0.52	111.48	109.50	94.06	1.103	78.32	1.17	78.31	77.79	94.66	1.103
NM70C	109.21	0.79	109.21	106.49	94.14	1.073	75.00	0.53	75.00	75.63	95.28	1.073
NM80U	116.57	0.63	116.57	114.05	93.90	1.149	81.71	0.26	81.71	81.00	94.88	1.149
NM80C	110.15	0.46	110.15	108.44	94.50	1.092	76.57	1.14	76.56	76.95	94.72	1.091
NM90U	123.52	1.24	123.51	121.14	94.56	1.220	85.86	-0.16	85.86	86.05	95.00	1.220
NM90C	112.41	0.31	112.40	110.63	94.26	1.114	77.98	0.74	77.97	78.44	95.00	1.112

Table 3: Inference for the log scale variance σ^2 based on fully masked $LN(\mu = 0, \sigma^2 = 1)$ data

	Results for $n = 30$						Results for $n = 50$					
	RMSE	Bias	SD	$\widehat{\text{SD}}$	Cvg.	Rel.	RMSE	Bias	SD	$\widehat{\text{SD}}$	Cvg.	Rel.
	$\times 10^3$	$\times 10^3$	$\times 10^3$	$\times 10^3$	%	Len.	$\times 10^3$	$\times 10^3$	$\times 10^3$	$\times 10^3$	%	Len.
UD	257.50	-30.14	255.73	250.42	89.60	1.000	199.30	-19.91	198.30	196.02	91.48	1.000
MI	283.05	40.68	280.11	293.92	94.14	1.174	211.95	21.45	210.86	216.95	94.24	1.107
NM10U	258.02	-30.13	256.25	251.28	89.54	1.003	200.32	-20.09	199.31	196.65	91.34	1.003
NM10C	257.82	-29.83	256.09	251.36	89.48	1.004	200.29	-19.97	199.29	196.67	91.32	1.003
NM20U	260.03	-31.82	258.07	253.48	89.20	1.012	202.43	-20.00	201.44	198.71	91.46	1.014
NM20C	261.97	-31.11	260.12	253.59	89.72	1.013	201.96	-19.44	201.02	198.76	91.58	1.014
NM30U	264.21	-32.18	262.25	257.93	89.64	1.030	206.21	-19.50	205.28	202.32	91.58	1.032
NM30C	266.40	-30.53	264.65	257.95	89.56	1.030	204.69	-20.33	203.68	201.85	91.56	1.030
NM40U	272.10	-32.54	270.14	264.54	89.38	1.056	210.89	-20.43	209.90	207.30	91.52	1.058
NM40C	272.07	-30.22	270.38	263.81	89.48	1.053	209.70	-22.07	208.53	205.98	91.48	1.051
NM50U	279.67	-32.89	277.73	273.75	90.14	1.093	215.93	-24.10	214.58	213.73	91.60	1.090
NM50C	279.85	-34.70	277.69	269.91	89.10	1.078	213.91	-20.04	212.97	212.00	91.88	1.082
NM60U	299.21	-34.27	297.24	285.98	88.96	1.142	227.60	-22.26	226.51	223.82	91.62	1.142
NM60C	286.55	-32.00	284.76	279.20	89.24	1.115	221.25	-19.18	220.42	218.83	91.76	1.116
NM70U	309.44	-42.62	306.50	300.97	89.40	1.202	238.43	-22.68	237.34	237.02	91.52	1.209
NM70C	298.14	-35.21	296.06	288.17	89.60	1.151	230.41	-19.67	229.57	226.32	91.50	1.155
NM80U	334.66	-49.35	331.00	322.94	88.10	1.290	260.19	-30.19	258.44	253.81	90.72	1.295
NM80C	305.85	-40.30	303.19	297.72	89.56	1.189	238.91	-22.74	237.82	234.13	91.32	1.194
NM90U	366.82	-52.33	363.06	356.45	88.38	1.423	286.05	-30.25	284.45	280.63	91.30	1.432
NM90C	317.95	-35.60	315.95	310.72	89.22	1.241	247.29	-25.57	245.96	242.69	91.48	1.238

Table 4: Inference for the log scale variance σ^2 based on fully masked $LN(\mu = 0, \sigma^2 = 1)$ data

	Results for $n = 100$						Results for $n = 200$					
	RMSE	Bias	SD	$\widehat{\text{SD}}$	Cvg.	Rel.	RMSE	Bias	SD	$\widehat{\text{SD}}$	Cvg.	Rel.
	$\times 10^3$	$\times 10^3$	$\times 10^3$	$\times 10^3$	%	Len.	$\times 10^3$	$\times 10^3$	$\times 10^3$	$\times 10^3$	%	Len.
UD	138.01	-9.72	137.67	140.05	94.04	1.000	102.64	-2.90	102.60	99.71	93.36	1.000
MI	144.13	10.64	143.73	148.71	95.28	1.062	106.04	7.62	105.77	103.76	94.08	1.041
NM10U	138.46	-9.59	138.13	140.54	93.74	1.004	102.85	-2.87	102.81	100.05	93.42	1.003
NM10C	138.56	-9.92	138.20	140.49	94.06	1.003	102.97	-2.79	102.93	100.05	93.36	1.003
NM20U	140.11	-9.53	139.79	141.99	93.84	1.014	104.11	-3.25	104.06	101.03	93.18	1.013
NM20C	139.62	-9.47	139.29	141.95	93.90	1.014	103.74	-2.78	103.70	101.05	93.58	1.013
NM30U	143.65	-10.86	143.24	144.28	93.50	1.030	106.28	-2.87	106.25	102.82	93.36	1.031
NM30C	142.46	-10.38	142.08	144.13	93.44	1.029	106.02	-2.82	105.98	102.67	93.20	1.030
NM40U	146.98	-10.83	146.58	147.94	93.68	1.056	108.47	-2.93	108.43	105.40	93.62	1.057
NM40C	145.02	-10.60	144.64	147.27	93.76	1.052	107.08	-1.90	107.06	105.01	93.66	1.053
NM50U	151.89	-10.51	151.52	153.05	93.60	1.093	111.76	-3.92	111.70	108.87	93.58	1.092
NM50C	149.72	-11.34	149.29	151.14	93.62	1.079	111.15	-2.41	111.13	107.76	94.04	1.081
NM60U	159.34	-10.91	158.96	159.85	93.38	1.141	117.71	-4.52	117.62	113.65	93.50	1.140
NM60C	155.28	-9.43	154.99	156.11	93.80	1.115	113.63	-3.82	113.56	110.95	93.58	1.113
NM70U	169.59	-11.88	169.17	169.05	93.00	1.207	123.41	-3.63	123.36	120.33	93.84	1.207
NM70C	163.02	-11.60	162.60	161.18	93.28	1.151	117.71	-4.22	117.63	114.71	93.28	1.150
NM80U	182.80	-13.72	182.29	181.81	93.10	1.298	131.21	-5.25	131.10	129.34	93.40	1.297
NM80C	166.68	-11.54	166.28	167.14	93.24	1.193	121.30	-6.16	121.15	118.72	93.28	1.191
NM90U	201.43	-15.06	200.87	201.05	92.90	1.436	148.31	-3.72	148.26	143.20	92.94	1.436
NM90C	170.98	-9.24	170.73	173.92	93.86	1.242	126.36	-5.13	126.26	123.39	94.04	1.237