Statistics and Probability Letters 105 (2015) 168–175

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

Likelihood-based inference for singly and multiply imputed synthetic data under a normal model

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

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

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

[c] Department of Mathematics and Statistics, University of Maryland, Baltimore County, Baltimore, MD 21250, USA

ARTICLE INFO

Article history:
Received 16 July 2014
Received in revised form 2 June 2015
Accepted 3 June 2015
Available online 11 June 2015

Keywords:
Maximum likelihood estimator
Pivot
Plug-in sampling
Posterior predictive sampling
Statistical disclosure control

ABSTRACT

Likelihood-based inference for both singly and multiply imputed synthetic data is developed in this paper under a univariate normal model and two distinct data generation scenarios, namely, posterior predictive sampling and plug-in sampling. We show that valid and exact inference can be drawn in both scenarios. Some theoretical issues of multiply imputed datasets under posterior predictive sampling are also pointed out.

Published by Elsevier B.V.

1. Introduction

To protect privacy and confidentiality of survey respondents, it is a standard practice of statistical agencies to use statistical disclosure control methodology. For magnitude microdata one popular method is to create multiple *synthetic* datasets based on a parametric model and apply suitable combination rules to draw inference about population parameters (Rubin, 1987, 1993; Raghunathan et al., 2003; Reiter, 2003; Reiter and Kinney, 2012; Drechsler, 2011).

There are two standard methods of generation of synthetic data. Assuming that the original data $\boldsymbol{x}$ are generated under a probability model f_{θ}, $\boldsymbol{\theta}$ being the unknown parameter, under *Case* 1: *Posterior Predictive Sampling*, a prior for $\boldsymbol{\theta}$ is assumed and converted to the posterior distribution, which is then used to draw independent replications $\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_m^*$ (known as posterior draws). Then for each such posterior draw $\boldsymbol{\theta}_i^*$ of $\boldsymbol{\theta}$, a corresponding replicate of $\boldsymbol{x}$ is generated from $f_{\boldsymbol{\theta}_i^*}$, resulting in multiply imputed synthetic data which are released to the public. Inference about a parameter of interest based on the multiply imputed synthetic data can be drawn following the (asymptotically valid) combination rules suggested by Reiter (2003). Under *Case* 2: *Plug-in Sampling*, a point estimator $\hat{\boldsymbol{\theta}}(\boldsymbol{x})$ of $\boldsymbol{\theta}$ is plugged into the joint probability density function (pdf) of $\boldsymbol{x}$, resulting in $f_{\hat{\boldsymbol{\theta}}(\boldsymbol{x})}$, which is used to generate synthetic data of any size. Here again the combination rules of Reiter (2003) can be used (see Reiter and Kinney, 2012).

The motivations for this current research are twofold. First, although synthetic data methodology calls for releasing multiple synthetic versions of the original data, there are situations where this is not feasible, perhaps due to severe privacy concerns (see Kinney et al., 2011 for an example). Second, since synthetic data generation is indeed *model-based*, one wonders if rigorous model-based *finite sample* inference can be developed. We demonstrate in this paper that, under a normal model, it is indeed possible to draw valid inference based on just one synthetic dataset!

* Corresponding author.
E-mail addresses: martin.klein@census.gov (M. Klein), sinha@umbc.edu (B. Sinha).

http://dx.doi.org/10.1016/j.spl.2015.06.003
0167-7152/Published by Elsevier B.V.

In Section 2 we consider Case 1 with $m = 1$, allowing a general form of the prior $\pi(\boldsymbol{\theta})$, involving a hyperparameter α, and making some recommendations about its choice. In Section 3 we deal with Case 2. Our comparison of the two approaches of synthetic data generation reveals some very interesting features. The entire treatment is non-asymptotic in nature. We also discuss Case 1 with $m > 1$ in Section 4 and point out some subtle theoretical issues. Section 5 presents numerical results to evaluate the performance of the proposed methods, and to compare the performance of Cases 1 and 2. We complete the paper with some concluding remarks in Section 6. Throughout we assume that the original data $\boldsymbol{x} = (x_1, \ldots, x_n)$ are such that $x_1, \ldots, x_n \sim$ i.i.d. $\sim N(\theta, \sigma^2)$ with $-\infty < \theta < \infty$ and $0 < \sigma^2 < \infty$. Let $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$, $\mathbb{S}_x^2 = \sum_{i=1}^n (x_i - \bar{x})^2$, and $s_x^2 = \mathbb{S}_x^2/(n-1)$.

2. Inference under posterior predictive sampling

Assume a joint prior $\pi(\theta, \sigma^2) \propto \frac{1}{\sigma^\alpha}$ and $n + \alpha > 7$ for valid inference about θ and σ^2.

Step 1. Draw $(\sigma^*)^2$ such that $\frac{\mathbb{S}_x^2}{(\sigma^*)^2} \sim \chi^2_{n+\alpha-3}$, then draw $\theta^* \sim N\left[\bar{x}, \frac{(\sigma^*)^2}{n}\right]$.

Step 2. Draw $\boldsymbol{z} = (z_1, \ldots, z_n)$ as i.i.d. from $N\left[\theta^*, (\sigma^*)^2\right]$.

Write $\bar{z} = \sum_{i=1}^n z_i/n$, $\mathbb{S}_z^2 = \sum_{i=1}^n (z_i - \bar{z})^2$, and $s_z^2 = \mathbb{S}_z^2/(n-1)$. It is easy to verify that $\bar{z}$ and $\mathbb{S}_z^2$ are jointly sufficient for (θ, σ^2). Here are the main *inferential* results.

1. The maximum likelihood estimator (MLE) of θ is $\bar{z}$, which is unbiased for θ with $\mathrm{Var}(\bar{z}) = \frac{2\sigma^2}{n} + \frac{(n-\alpha+3)\sigma^2}{n(n+\alpha-5)}$.
2. An unbiased estimator of σ^2 is $\hat{\sigma}^2_{\mathrm{U}}(\boldsymbol{z}) = \frac{(n+\alpha-5)}{(n-1)^2}\mathbb{S}_z^2$, since $E(\mathbb{S}_z^2) = (n-1)E[(\sigma^*)^2] = \frac{(n-1)^2\sigma^2}{n+\alpha-5}$. Furthermore, $\mathrm{Var}(\hat{\sigma}^2_{\mathrm{U}}(\boldsymbol{z})) = \left[\frac{(n+1)^2(n+\alpha-5)}{(n+\alpha-7)(n-1)^2} - 1\right]\sigma^4$. Observe that only the choice $\alpha = 4$ yields the usual unbiased estimator $\mathbb{S}_z^2/(n-1)$, pointing out the fact that one cannot ignore the importance of the choice of α.
3. The MLE of σ^2 is $\hat{\sigma}^2_{\mathrm{MLE}}(\boldsymbol{z}) = \frac{\mathbb{S}_z^2}{\psi_n}$, with its mean squared error (MSE) as $\mathrm{MSE}\left(\hat{\sigma}^2_{\mathrm{MLE}}(\boldsymbol{z})\right) = E\left[\frac{\mathbb{S}_z^2}{\psi_n} - \sigma^2\right]^2 = E\left[(n^2-1)\frac{(\sigma^*)^4}{\psi_n^2} - 2\sigma^2\frac{\mathbb{S}_z^2}{\psi_n} + \sigma^4\right] = \sigma^4[\frac{(n^2-1)^2}{(n+\alpha-5)(n+\alpha-7)\psi_n^2} - 2\frac{(n-1)^2}{(n+\alpha-5)\psi_n} + 1]$ where ψ_n is the value of ψ that maximizes the expression $Q_1(\psi) = \psi^{\frac{n}{2}}\int_0^\infty e^{-\frac{u}{2}}u^{n+\frac{\alpha-3}{2}-1}\frac{1}{\sqrt{\psi+\frac{u}{2}}}[\psi+u]^{-\frac{2n+\alpha}{2}+2}du$. In the MSE computation above we have used the facts that $(\sigma^*)^2|\mathbb{S}_x^2 \sim \frac{\mathbb{S}_x^2}{\chi^2_{n+\alpha-3}}$ and $\frac{\mathbb{S}_x^2}{\sigma^2} \sim \chi^2_{n-1}$.
4. A two-sided shortest length $(1-\gamma)$ level confidence interval for σ^2 based on $\boldsymbol{z}$ is
$$\left[\frac{\mathbb{S}_z^2}{d_{n,\alpha}}, \frac{\mathbb{S}_z^2}{c_{n,\alpha}}\right], \tag{1}$$
where the constants $c_{n,\alpha}$ and $d_{n,\alpha}$ satisfy $\int_{c_{n,\alpha}}^{d_{n,\alpha}} f_{n,\alpha}(v)dv = 1-\gamma$, $c^2_{n,\alpha}f_{n,\alpha}(c_{n,\alpha}) = d^2_{n,\alpha}f_{n,\alpha}(d_{n,\alpha})$, and $f_{n,\alpha}(v)$ is the pdf of $V = \mathbb{S}_z^2/\sigma^2$ given below in Theorem 2.2. The length of the confidence interval is $L_{\sigma^2}(\boldsymbol{z}) = \mathbb{S}_z^2[\frac{1}{c_{n,\alpha}} - \frac{1}{d_{n,\alpha}}]$, and the expected length is $E[L_{\sigma^2}(\boldsymbol{z})] = \sigma^2\left[\frac{1}{c_{n,\alpha}} - \frac{1}{d_{n,\alpha}}\right] \times \left[\frac{(n-1)^2}{n+\alpha-5}\right]$.
5. A two-sided shortest length $(1-\gamma)$ level confidence interval for θ based on $\boldsymbol{z}$ is
$$\left[\bar{z} - \frac{b_{n,\alpha}}{[n(n-1)]^{1/2}}\mathbb{S}_z,\ \bar{z} + \frac{b_{n,\alpha}}{[n(n-1)]^{1/2}}\mathbb{S}_z\right], \tag{2}$$
where the constant $b_{n,\alpha}$ satisfies $1-\gamma = 2\Pr[0 < t < b_{n,\alpha}]$, and the pdf of $t = \sqrt{n}(\bar{z}-\theta)/s_z$ is given below in Theorem 2.3. The length of the confidence interval is $L_\theta(\boldsymbol{z}) = \frac{2b_{n,\alpha}}{[n(n-1)]^{1/2}}\mathbb{S}_z$, and the expected length is $E[L_\theta(\boldsymbol{z})] = \frac{2\sigma b_{n,\alpha}}{[n(n-1)]^{1/2}} \times E[\{(\chi^2_{n-1})^{1/2}\}]^2 \times E[\{(\chi^2_{n+\alpha-3})^{-1/2}\}]$. To compute the expected confidence interval length, we have used the facts that $E(\mathbb{S}_z) = E(\sigma^*) \times [E\{(\chi^2_{n-1})^{1/2}\}]$, $E(\sigma^*) = E[(\mathbb{S}_x^2)^{1/2}] \times E[(\chi^2_{n+\alpha-3})^{-1/2}]$, and finally $E[(\mathbb{S}_x^2)^{1/2}] = \sigma E[(\chi^2_{n-1})^{1/2}]$.

The following theorems dealing with (1) the joint distribution of $\bar{z}$ and $\mathbb{S}_z^2$, (2) the distribution of $V = \mathbb{S}_z^2/\sigma^2$ (for confidence interval for σ^2), and (3) the distribution of $t = \sqrt{n}(\bar{z}-\theta)/s_z$ (for confidence interval for θ), are used to derive the above inferential results. We refer the reader to the *Supplementary materials* (see Appendix A) for the proofs of the theorems.

Theorem 2.1. *The joint pdf of $\bar{z}$ and $\mathbb{S}_z^2$ is*
$$f_{\theta,\sigma^2}(\bar{z}, \mathbb{S}_z^2) \propto \int_0^\infty \frac{\exp\left\{-\frac{1}{2}\left[\frac{n(\bar{z}-\theta)^2}{\sigma^2+2(\sigma^*)^2} + \frac{\mathbb{S}_z^2}{(\sigma^*)^2}\right]\right\}\frac{(\mathbb{S}_z^2)^{\frac{n-1}{2}-1}}{\sigma^n}\left[\frac{1}{(\sigma^*)^2}\right]^{n+\frac{\alpha-3}{2}-1}}{\left[\frac{1}{\sigma^2} + \frac{1}{2(\sigma^*)^2}\right]^{\frac{1}{2}}\left[\frac{1}{\sigma^2} + \frac{1}{(\sigma^*)^2}\right]^{\frac{2n+\alpha}{2}-2}} d\left(\frac{1}{(\sigma^*)^2}\right).$$