Computational Statistics and Data Analysis 65 (2013) 4–12

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Density estimation for data with rounding errors[☆]

B. Wang[a,*], W. Wertelecki[b]

[a] Department of Mathematics & Statistics, University of South Alabama, Mobile, AL 36688, USA

[b] Department of Medical Genetics, University of South Alabama, Mobile, AL 36688, USA

ARTICLE INFO

Article history:
Received 26 April 2011
Received in revised form 15 February 2012
Accepted 16 February 2012
Available online 3 March 2012

Keywords:
Bootstrapping
Kernel density estimation
Measurement error
Fast Fourier transformation
Deconvolution density estimation

ABSTRACT

Rounding of data is common in practice. The problem of estimating the underlying density function based on data with rounding errors is addressed. A parametric maximum likelihood estimator and a nonparametric bootstrap kernel density estimator are proposed. Simulations indicate that the maximum likelihood approach performs well when prior information on the functional form of the underlying distribution is available, while the kernel-type estimator attains stable and good performance in various cases. The proposed methods are further applied to detect the distributional difference of head circumferences from two Chernobyl impacted regions of Ukraine.

1. Introduction

Measurement errors exist in many applications due to improper instrument calibration or operator errors, among many other reasons. In practice, it is not uncommon that data are rounded. In the US, birth weights show ounces, for example 5 lb 3 oz, while in Europe, for example in Ukraine, intervals could be in 50 g or nearly 1.8 oz. Compared with some extreme cases, rounding might be the mildest form of binning. In heath and social research, due to confidentiality reasons, individual level data are often not accessible, and data must be aggregated to a sufficient level. The problem of density estimation based on binned data has been investigated in the literature. Scott and Sheather (1985) used standard kernel density estimation, and Härdle and Scott (1992) proposed a kernel-type estimator called "weighted average of rounded points". A spline density estimation was also proposed for binned data in Minnotte (1998), and a log-spline density estimator by Koo and Kooperberg (2000). Blower and Kelsall (2002) proposed a nonlinear binned kernel estimator.

Let $X = \{X_1, X_2, \dots, X_n\}$ be a random sample from an underlying (latent) continuous distribution with density f. X is pre-binned (we have no control over the rounding process), and the rounding mechanism is known. Let $W = \{W_1, W_2, \dots, W_n\}$ be the corresponding rounded values of X. This paper proposes two methods to estimate f based on rounded data W. The remainder of this paper is organized as follows. The impact of rounding errors on some statistical inferences are illustrated via a simulation study in Section 2. A parametric density estimator is developed in Section 3, and a nonparametric kernel density estimator is proposed in Section 4. The performances of the new estimators are evaluated via a simulation study in Section 5, and the new approaches are applied to a real birth data analysis in Section 6. Finally, a discussion is provided in Section 7.

[☆] All algorithms have been implemented and collected in a CRAN R software package bda, the R codes for the figures and simulation study is available at http://gauss.usouthal.edu/publ/bootKDE/.

[*] Correspondence to: Mathematics and Statistics Department, 411 University BLVD N, Room 325, Mobile, AL 36688, USA. Tel.: +1 251 460 6264x2626; fax: +1 251 460 7969.

E-mail address: bwang@jaguar1.usouthal.edu (B. Wang).

0167-9473/$ – see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2012.02.016

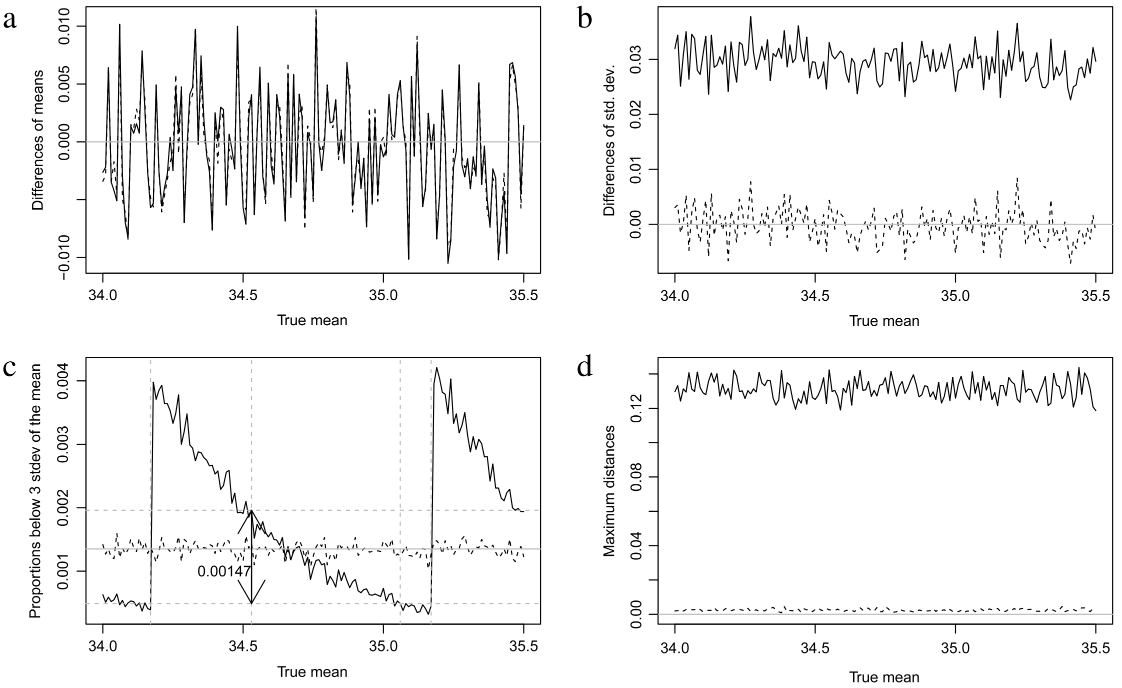

Fig. 1. The effects of rounding errors in OFC data analysis. The solid curves show the results based on the data rounded to the nearest centimeter, and the dashed curves show the results based on the data without rounding. Plot (a) compares the differences of between the sample means and the true population means. Plot (b) compares the differences between the sample standard deviations and the true standard deviations. Plot (c) compares the proportions of observations that fall below three standard deviations of the means. Plot (d) compares maximum distances between the empirical distribution functions and the true distribution functions.

2. The impact of rounding errors on some statistical inferences

Head circumference or occipito-frontal circumference (OFC hereafter) measurements are important markers of cerebral development. Wertelecki (2010) reported elevated rates of certain congenital malformations among children born in the Rivne Province (or Oblast) of Ukraine. Nearly half of Rivne Province is heavily impacted by ionizing radiation from the 1986 Chernobyl disaster (Chornobyl in Russian). Reductions of OFC are well-documented effects of severe prenatal exposures to ionizing radiation. However, to what extent OFC reductions reflect degrees of lesser exposures to ionizing radiation is less well known (Murphy et al., 1942; Burrow et al., 1964; Wood et al., 1967). In Rivne Province, preliminary studies suggest that microcephaly (OFC three standard deviations below the norm) may be more prevalent in severely radiation impacted areas. Investigations to confirm rates of microcephaly in Rivne oblast are ongoing.

The actual OFC differences among the populations in different regions might be small and are affected by various factors. The OFC of live births in Chernobyl impacted regions can be affected by the increase in the body burden of ^{137}Caesium (^{137}Cs) of the mothers as an aftermath of the Chernobyl disaster, or an outcome of forest fires. It can also be caused by the inhalation of the smoke from the burning of potato stems and other vegetative matter (Dancause et al., 2010). Other causes such as alcohol exposure and malnutrition may affect OFC as well. During standard neonatal medical examinations, OFC, birth weight and gestational age are routinely recorded. The OFC measures are customarily recorded to the nearest centimeter (cm). Of concern are whether rounding itself will mask such relatively mild contrasts, and how we can distinguish such contrasts by taking the rounding errors into consideration.

For illustration purposes, we assume that OFC measures follow a normal distribution, $N(\mu, \sigma^2)$, with mean μ and standard deviation σ. A random sample X of size $n = 100{,}000$ is drawn from $N(\mu, \sigma^2)$. We take different values of μ from 34 cm to 35.5 cm with an increment 0.01 cm and fix σ to be 1.39 cm. Let W be the sample by rounding all X values to the nearest centimeter. We then compare various statistics based on X and W, respectively. The following can be found from the results shown in Fig. 1.

1. Rounding does not affect the sample mean too much when the underlying distribution is normal, which is expected for other symmetric distributions as well. Plot (a) shows the differences between the sample means and the true mean μ. We see that for the same sample, the difference between the sample means based on W and X are not very different—the solid curve and dashed curve stay close to each other. Although the absolute difference between the sample mean $\bar{X}$ and μ can be as large as 0.0489 cm, and 0.0494 cm between $\bar{W}$ and μ, the largest absolute difference between $\bar{X}$ and $\bar{W}$ is only 0.0072 cm.