Statistical Methodology 6 (2009) 577–593

Contents lists available at ScienceDirect

Statistical Methodology

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

Probabilistic rounding and Sheppard's correction

H. Schneeweiss *, J. Komlos

Department of Statistics, University of Munich Akademiestr. 1 I, D-80799, Munich, Germany

ARTICLE INFO

Article history:
Received 4 February 2009
Received in revised form
27 May 2009
Accepted 19 June 2009

Keywords:
Rounding
Asymmetric rounding
Mixture rounding
Heaping
Profile function
Sheppard's correction
Moments
Linear regression

ABSTRACT

When rounded data are used in place of the true values to compute the variance of a variable or a regression line, the results will be distorted. Under suitable smoothness conditions on the distribution of the variable(s) involved, this bias, however, can be corrected with very high precision by using the well-known Sheppard's correction. In this paper, Sheppard's correction is generalized to cover more general forms of rounding procedures than just simple rounding, viz., probabilistic rounding, which includes asymmetric rounding and mixture rounding.

1. Introduction

Data often contain rounding errors. Variables (such as heights or weights) that by their very nature are continuous are, nevertheless, typically measured in a discrete manner. They are rounded to a certain level of accuracy, often to some preassigned decimal point of a measuring scale (e.g., to multiples of 10 cm, 1 cm, or 0.1 cm). The reason may be the avoidance of costs associated with a fine measurement or the imprecise nature of the measuring instrument. Even if precise measurements are available, they are sometimes recorded in a coarsened way in order to preserve confidentiality or to compress the data into an easy to grasp frequency table.

Two recent reviews of the field are [3,8]. Most of the literature is concerned with simple rounding as described above. But there are other types of rounding procedures, where certain numbers are preferred over others. In asymmetric rounding, for example, more than half of the rounding interval is rounded to one of the round values and less than half to the neighboring round value [6]. In mixture rounding, different portions of the population round in different ways, e.g. some preferring even

* Corresponding author. Tel.: +49 0 89 2180x3224, +49 0 89 2180x5386; fax: +49 0 89 2180x5044.
E-mail address: schneew@stat.uni-muenchen.de (H. Schneeweiss).

1572-3127/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.stamet.2009.06.005

values, some odd values, again some preferring zeros or fives as the last digit [13]. [4] and similarly [11] use a variant of mixture rounding, where the mixing parameters (the proportions mentioned above) depend on the value of the true variable.

We generalize these approaches by introducing the concept of probabilistic rounding. Numbers are rounded to a round value with certain probabilities which depend on the distance from the original value to the round value. The probability as a function of the distance is given by a so-called rounding profile function. Again there may be several profile functions depending on whether some rounded values are preferred over other ones. We only consider two profile functions below, one for even and one for odd numbers. Profile functions, though not with this name, were employed by [10] to describe heaping in unemployment duration data.

Whereas probabilistic rounding generalizes simple, asymmetric, and mixture rounding, it can again be generalized. Indeed, it can be viewed as a special form of heaping; e.g. see [12]. It is special in that (1) the heaping points (here the round values) form a regular lattice of equidistant points with regularly alternating profile functions and that (2) the two types of profile functions are the same for each of the two types of heaping points. By contrast, e.g., in [11] the profile functions are different for different values of the true variable.

The mean and the variance, and also higher moments of a variable X calculated using rounded data X^* instead of the original data X will be biased. However, under certain smoothness conditions (see, e.g., [5,9,8]), the means of X and X^* do not differ very much and can be considered as almost equal. Yet, the variances differ markedly. However, the difference is captured, to a high degree of accuracy, by a very simple term, $h^2/12$, where h is the distance between neighboring values of X^*. This is the famous Sheppard's correction (1898). The purpose of the present paper is to extend Sheppard's correction to the case of probabilistic rounding. We derive a similar, though more complicated, correction term for the variance (and in principle also for higher moments) which depends on the profile function of the rounding procedure.

This result is then used to show how the estimation of the slope parameter of a linear regression based on rounded data can be corrected in order to obtain an essentially unbiased estimate. In such cases, the variance of a rounded independent variable has to be considered and also its covariance with the dependent variable, which may or may not be rounded. However, the covariance is essentially not affected by rounding and so only the effect of rounding on the variance of the independent variable has to be taken into account.

Section 2 introduces probabilistic rounding together with the special cases of simple, asymmetric, and mixture rounding. Section 3 derives a Sheppard-like correction term, which is used in Section 4 to work out a correction formula for linear regression analysis based on rounded data. Section 5 deals with the problem of finding the correct correction formula when the rounding procedure is only partially known. Section 6 has an example, and Section 7 concludes. Some technical details are presented in the Appendix.

2. Probabilistic rounding

2.1. Simple rounding

Let X be a continuous random variable. The values of X are not reported in their original form but only as rounded values. Rounding is a procedure that shifts the value of X to values on a rounding lattice of equidistant points in a prescribed manner. The rounding lattice is defined as the following set:

$$G = \{ih | i \in \mathbb{Z}\},$$

where h is the distance between two adjacent lattice points and is also called the width of the rounding intervals. We distinguish between even and odd lattice points, $2ih$ and $(2i + 1)h$, respectively. (Note that 0 is a point of the lattice. More generally we could define a rounding lattice by shifting the above lattice away from the origin by some amount a. However, we can restrict our discussion to the special case $a = 0$ without loss of generality.) Let X^* be the rounded variable. The various rounding