AStA Adv Stat Anal (2010) 94: 247–271
DOI 10.1007/s10182-010-0125-2

ORIGINAL PAPER

Symmetric and asymmetric rounding: a review and some new results

H. Schneeweiss · J. Komlos · A.S. Ahmad

Received: 26 February 2010 / Accepted: 2 March 2010 / Published online: 17 March 2010

Abstract Using rounded data to estimate moments and regression coefficients typically biases the estimates. We explore the bias-inducing effects of rounding, thereby reviewing widely dispersed and often half forgotten results in the literature. Under appropriate conditions, these effects can be approximately rectified by versions of Sheppard's correction formula. We discuss the conditions under which these approximations are valid and also investigate the efficiency loss caused by rounding. The rounding error, which corresponds to the measurement error of a measurement error model, has a marginal distribution, which can be approximated by the uniform distribution, but is not independent of the true value. In order to take account of rounding preferences (heaping), we generalize the concept of simple rounding to that of asymmetric rounding and consider its effect on the mean and variance of a distribution.

Keywords Simple rounding · Asymmetric rounding · Euler–Maclaurin · Moments · Sheppard's correction · Maximum likelihood

1 Introduction

Data often contains 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).

H. Schneeweiss (✉)
Department of Statistics, University of Munich LMU, Munich, Germany
e-mail: schneew@stat.uni-muenchen.de

J. Komlos
Department of Economic History, University of Munich LMU, Munich, Germany

A.S. Ahmad
Institute of Cancer Research, 237 Fulham Road, London SW3 6JB, UK

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. Sometimes the rounding process shows some asymmetry due to the preferences of the recorder for even numbers or for multiples of five; see, e.g., Myers (1954).

There are quite a few research areas where very crudely rounded data are common and where the effect of rounding is not negligible, e.g., in social sciences, in economic history or in demography, e.g., Crockett and Crockett (2006). Rounding and more general quantization procedures play a fundamental role in the discretization of continuous data in information processing, see Gray and Neuhoff (1998) and Widrow et al. (1996).

In the following we analyze statistical characteristics of rounded data X^* of a variable X and of the rounding error δ. We consider means, variances, and higher moments, and also regression parameters obtained from rounded variables and show how they are related to the parameters of unrounded data. We study in particular the approximations that arise when the rounding interval is small. These approximations are governed by Sheppard's (1898) correction. We give conditions under which Sheppard's correction can be applied.

This report is to a large extent a review of the literature. It highlights some results which are not so often discussed in the literature, such as the distribution of the rounding error and the efficiency loss due to rounding. In addition, new results concerning asymmetric rounding are presented. Earlier reviews are Eisenhart (1947), Stuart and Ord (1987, Sects. 3.18–3.30), Gjeddebaek (1968), Haitovsky (1982), and Heitjan (1989). The latter also deals with the more general concept of grouped data. For more details with respect to the present paper, see Schneeweiss et al. (2006).

Although rounding is widely applied in real data—in fact any histogram is a rounded data set—, most of the literature is of a theoretical nature. But there are some exceptions. Sheppard (1898) gives typical examples of rounded data, e.g., weight frequencies in intervals of 10 lb.s and frequencies of examination marks in intervals of 100 points; Gjeddebaek (1968) analyzes a set of weight data coarsened in several different ways; Heitjan (1989) analyzes Fisher's famous *Iris* data; Johnson et al. (2004) discuss parturition data in wild sheep that have been sampled with a coarse scale and cite many other studies on live-history events; Braun et al. (2005) apply their method to interval-censored HIV infection time data and to a histogram for body mass index; Lambert and Eilers (2009) study the distribution of lead concentration in blood samples and certain mortality statistics.

Heaping, in which some numbers are preferred, is a more general case of rounding, and is treated only partly in this paper. Asymmetric rounding is a special kind of heaping, where the preferred points are evenly spaced on the line. In general, heaping points need not be evenly spaced; see e.g., Heitjan and Rubin (1991), Wolff and Augustin (2003), Augustin and Wolff (2004), Wang and Heitjan (2008).

Section 2 introduces the concept of simple rounding. In Sect. 3, approximate expressions of the moments of rounded data are derived. Section 4 pertains to the effect of rounding on regression results. The rounding error δ itself is analyzed in Sect. 5.

Section 6 investigates the validity of the approximations introduced earlier. Section 7 examines some special distributions where these approximations are either exact or completely invalid. Estimating and testing with rounded data is considered in Sect. 8. Section 9 is devoted to ML estimation and related methods. Asymmetric rounding is the subject of Sect. 10. Some concluding remarks are found in Sect. 11.

2 Simple rounding

Let X be a continuous random variable with density $\varphi(x)$ and let X^* be the corresponding rounded variable. The rounding model is as follows. Let there be given a set of equidistant points on the real line, the rounding lattice,

$$\mathbb{R}^* := \mathbb{R}^*_{a,h} := \big\{(a+i)h, i \in \mathbb{Z}\big\},$$

where h is the rounding width, i.e., the distance between two adjacent points of the rounding lattice, and ah, $0 \le a \le 1$, is the origin of the lattice. We call a the shift parameter of the lattice. For simplicity, we will assume, unless otherwise stated, that $a = 0$. Most of the following approximate results do not depend on the value of a. For any value of X, the rounded value X^* is that point of $\mathbb{R}^*$ which is nearest to X. Using the floor function, this can be expressed as

$$X^* = h\left\{a + \text{floor}\left(\frac{X}{h} - a + \frac{1}{2}\right)\right\},$$

where floor(x) is the largest integer $\le x$.

The rounding error δ is defined as $\delta = X^* - X$. The equation

$$X^* = X + \delta \tag{1}$$

looks like the measurement equation of a classical measurement error model with X being the unobservable variable, X^* the observable variable, and δ the measurement error. However, unlike the measurement error in such a model, δ is not independent of X. It is a function of X because X^* is a function of X, see also Vardeman (2005). But δ is not independent of X^* either. The conditional density of δ given $X^* = x^*$, where $x^* \in \mathbb{R}^*$, is given by

$$h\big(\delta|x^*\big) = \begin{cases} \frac{\varphi(x^*-\delta)}{p(x^*)} & \text{for } -\frac{h}{2} \le \delta \le \frac{h}{2} \\ 0 & \text{for } \delta < -\frac{h}{2} \text{ or } \delta > \frac{h}{2}, \end{cases} \tag{2}$$

where

$$p\big(x^*\big) := P\big(X^* = x^*\big) = \int_{x^*-\frac{h}{2}}^{x^*+\frac{h}{2}} \varphi(x)\,dx = \int_{-\frac{h}{2}}^{\frac{h}{2}} \varphi\big(x^* + u\big)\,du. \tag{3}$$

3 Moments of rounded values

3.1 Univariate moments

We relate the moments of X^* to those of X. The kth moment of the distribution of the rounded values X^* (assuming $a = 0$) can be computed as

$$\mathbb{E}X^{*k} = \sum_i (ih)^k P\big(X^* = ih\big) = \sum_i (ih)^k \int_{-\frac{h}{2}}^{\frac{h}{2}} \varphi(ih + u)\, du. \tag{4}$$

We can approximate the sum in (4) by an integral using the Euler–Maclaurin formula (see, e.g., Stuart and Ord 1987, and Sect. 3.2 for more details). We rewrite each term of the sum in (4) as follows:

$$f(y) = y^k \int_{-\frac{h}{2}}^{\frac{h}{2}} \varphi(y + u)\, du$$

with $y = ih$. Then, according to the Euler–Maclaurin formula, the sum in (4) becomes

$$\sum_i f(ih) = \frac{1}{h} \int_{-\infty}^{\infty} f(y)\, dy + R, \tag{5}$$

where R is a remainder term (often quite small) to be treated in more detail in Sect. 3.2. Ignoring it, we can write the kth moment of the rounded variable X^* in (4) as follows:

$$\mathbb{E}X^{*k} \approx \int_{-\infty}^{\infty} y^k \frac{1}{h} \int_{-\frac{h}{2}}^{\frac{h}{2}} \varphi(y + u)\, du\, dy. \tag{6}$$

Substituting $x = y + u$ and $v = u/h$ into (6) we obtain

$$\mathbb{E}X^{*k} \approx \int_{-\infty}^{\infty} \int_{-\frac{1}{2}}^{\frac{1}{2}} (x - vh)^k\, dv\, \varphi(x)\, dx. \tag{7}$$

Equation (7) can be used to compute approximately any kth moment of the rounded data. For example, for $k = 1$, we obtain

$$\mathbb{E}X^* \approx \int_{-\infty}^{\infty} \left[xv - \frac{v^2}{2} h \right]_{-\frac{1}{2}}^{\frac{1}{2}} \varphi(x)\, dx = \int_{-\infty}^{\infty} x\varphi(x)\, dx = \mathbb{E}X. \tag{8}$$

Thus, the expectations of the rounded and unrounded data are approximately equal. For the second moment we obtain

$$\mathbb{E}X^{*2} \approx \int \left[x^2 v - xv^2 h + \frac{v^3}{3} h^2 \right]_{-\frac{1}{2}}^{\frac{1}{2}} \varphi(x)\, dx = \int \left(x^2 + \frac{h^2}{12} \right) \varphi(x)\, dx = \mathbb{E}X^2 + \frac{h^2}{12}.$$

Because of (8) it follows that

$$\mathbb{V}X^* \approx \mathbb{V}X + \frac{h^2}{12}. \tag{9}$$

Thus the variance of the rounded data has to be "corrected" by the term $-\frac{h^2}{12}$, known as Sheppard's (1898) correction, in order to derive an approximate value for the (unobservable) variance of the unrounded data:

$$\mathbb{V}X \approx \mathbb{V}X^* - \frac{h^2}{12}. \tag{10}$$

Note that the term $\frac{h^2}{12}$ is just the variance of a variable uniformly distributed on the interval $[-\frac{h}{2}, \frac{h}{2}]$; see also Sect. 5.

An interesting application of Sheppard's correction is to decide how many significant figures in experimental data should be recorded, when these data are imprecise anyway (Wilrich 2005; see also Wimmer et al. 2000 for a similar idea).

For the third to sixth common moments the corresponding formulas are (cf. Kendall 1938, and Stuart and Ord 1987):

$$\begin{aligned}
\mu_3 &\approx \mu_3^* - \frac{1}{4}\mu_1^* h^2, \\
\mu_4 &\approx \mu_4^* - \frac{1}{2}\mu_2^* h^2 + \frac{7}{240}h^4, \\
\mu_5 &\approx \mu_5^* - \frac{5}{6}\mu_3^* h^2 + \frac{7}{48}\mu_1^* h^4, \\
\mu_6 &\approx \mu_6^* - \frac{5}{4}\mu_4^* h^2 + \frac{7}{16}\mu_2^* h^4 - \frac{31}{1344}h^6,
\end{aligned} \tag{11}$$

where $\mu_k = \mathbb{E}X^k$, $\mu_k^* = \mathbb{E}X^{*k}$. In general, see Stuart and Ord (1987),

$$\mu_k = \sum_{j=0}^{k} \binom{k}{j} \left(2^{1-j} - 1\right) B_j \mu_{k-j}^* h^j, \tag{12}$$

where B_j is the jth Bernoulli number. These are found from the recurrence relation $\sum_{i=0}^{j-1} \binom{j}{i} B_i = 0$ for $j > 1$, starting with $B_0 = 1$. By omitting all terms of higher powers of h^2, which is justified if h is small, we obtain the approximate equation

$$\mu_k \approx \mu_k^* - \binom{k}{2} \frac{h^2}{12}\, \mu_{k-2}^*. \tag{13}$$

Similar relations for the central moments are found by setting $\mu_1^* = 0$ in the above formulas and redefining $\mu_k := \mathbb{E}(X - \mathbb{E}X)^k$ and $\mu_k^* := \mathbb{E}(X^* - \mathbb{E}X^*)^k$.

By the same principles, one can also derive a relation between the characteristic functions of the unrounded and rounded variables, respectively, cf. Kullback (1935).

Let $\psi(t)=\int e^{itx}\varphi(x)\,dx$ and $\psi^*(t)=\sum_j e^{itjh}p(jh)$ be the characteristic functions of X and X^*, respectively. Then

$$\psi^*(t)\approx\frac{2}{ht}\sin\left(\frac{ht}{2}\right)\psi(t). \tag{14}$$

The r.h.s. of (14) is the characteristic function of $X+U$, where U is a random variable independent of X and uniformly distributed over $[-\frac{h}{2},\frac{h}{2}]$. Thus $X^*=X+\delta$ has approximately the same distribution as if δ were uniformly distributed and independent of X, see also Sect. 5. If follows that approximately

$$\mathbb{E}X^{*m}\approx\mathbb{E}(X+U)^m$$

or

$$\kappa_m\left(X^*\right)\approx\kappa_m(X)+\kappa_m(U)=\kappa_m(X)+\frac{B_m}{m}h^m,\quad m>1,$$

where κ_m is the mth semi-invariant. From this follow all the moment relations considered before.

An exact expression of the characteristic function of X^* (including the shift parameter a) is

$$\psi^*(t)=\sum_{j=-\infty}^{\infty}\exp(-i2\pi ja)\psi\left(t+\frac{2\pi j}{h}\right)\frac{\sin[(ht+2\pi j)/2]}{(ht+2\pi j)/2}, \tag{15}$$

from which exact expressions for the moments can be derived (Tricker 1984b; Widrow et al. 1996; and Janson 2006). Note that the term for $j=0$ in the sum of (15) corresponds to (14). If $\psi(t)$ happens to be "limited" in the sense that $\psi(t)=0$ for $|t|\geq\frac{2\pi}{h}$, then the above approximate relations for the moments become exact (Tricker 1984b). In practice, $\psi(t)$ will hardly ever be limited, but it will often go to zero quite rapidly for $|t|\to\infty$, which again justifies the above approximations.

3.2 The remainder term R

The sum in (5) can be approximated by the integral on the r.h.s. of (5) only if R is small. Suppose for the moment that the function f is restricted to a finite interval $[a,b]$ with $f(a)=f(b)=0$ and that $a+\frac{h}{2}$ and $b-\frac{h}{2}$ are points of the rounding lattice. If the following two conditions are satisfied:

- $f(y)$ is differentiable on the interval $[a,b]$ to the order $2m+2$,
- all derivatives of odd order to the order $2m-1$ vanish at the points a and b,

then the remainder term R equals

$$R=\frac{B_{2m+2}}{(2m+2)!}(b-a)h^{2m+1}f^{(2m+2)}(y_m),$$

where $y_m\in[a,b]$ and B_{2m+2} is the $(2m+2)$-th Bernoulli number (Stoer and Bulirsch 1980).

The magnitude of the remainder term R, and thus the closeness of the approximation of the sum in (5) by the integral in (5), depends on h and on the smoothness of f. The smaller are h and $\max_{a \le y \le b} f^{(2m+2)}(y)$, the better is the approximation.

Clearly, a sum can always be approximated by a corresponding integral if h is sufficiently small, no matter if the conditions for the Euler–Maclaurin formula are satisfied or not. However, if the conditions *are* satisfied, then the Euler–Maclaurin approximation is typically extremely good even if h is (moderately) large. See also Sect. 6.

3.3 Multivariate moments

The analysis of moments of rounded and unrounded data can be extended to the multivariate case, cf. Baten (1931), Wold (1934). Here we restrict our account to the bivariate case.

Let $\varphi(x, y)$ be the joint density of the random variables X and Y. Let these be rounded according to two rounding lattices with widths h and k, respectively, and origin $(0, 0)$ and let X^* and Y^* be the rounded variables. By arguments similar to those in Sect. 3.1 we derive an expression for mixed moments of X^* and Y^* analogous to (7):

$$\mathbb{E}\big(X^{*m} Y^{*n}\big) \approx \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{-\frac{1}{2}}^{\frac{1}{2}} (t - uh)^m (s - vk)^n \, du \, dv \, \varphi(t, s) \, dt \, ds.$$

Specializing to the case $m = n = 1$, we find

$$\mathbb{E}\big(X^* Y^*\big) \approx \mathbb{E}(XY).$$

As $\mathbb{E}X^* \approx \mathbb{E}X$ and $\mathbb{E}Y^* \approx \mathbb{E}Y$, it follows that

$$\mathbb{C}ov\big(X^*, Y^*\big) \approx \mathbb{C}ov(X, Y). \tag{16}$$

By the same arguments, a similar relation holds if only one variable is rounded. Thus

$$\mathbb{C}ov\big(X^*, Y\big) \approx \mathbb{C}ov\big(X, Y^*\big) \approx \mathbb{C}ov(X, Y). \tag{17}$$

4 The influence of rounding on regression estimates

We are now ready to analyze the influence of rounding on the estimates of the coefficients of a linear regression. We always assume that the assumption for the application of the Euler–Maclaurin approximation is satisfied. Let Y be the unrounded response variable, and X^* (X) the rounded (unrounded) explanatory variable, and consider a simple linear regression model

$$Y = \alpha + \beta X + \varepsilon.$$

The corresponding regression for rounded data is $Y = \alpha^* + \beta^* X^* + \varepsilon^*$. The true regression coefficient is given by $\beta = \mathbb{C}ov(X, Y)/\mathbb{V}X$, whereas the regression coefficient computed from the rounded variable X^* is $\beta^* = \mathbb{C}ov(X^*, Y)/\mathbb{V}X^*$ which differs from β. However, one can retrieve β using Shepard's correction:

$$\beta = \frac{\mathbb{C}ov(X,Y)}{\mathbb{V}X} \approx \frac{\mathbb{C}ov(X^*,Y)}{\mathbb{V}X^* - \frac{h^2}{12}} = \beta^* \left(1 - \frac{1}{12}\frac{h^2}{\mathbb{V}X^*}\right)^{-1}. \tag{18}$$

This formula is convenient since we normally know h and can estimate $\mathbb{V}X^*$ from the data. The naive estimate of β (i.e., $\hat{\beta}^* = s_{x^*y}/s^2_{x^*}$), which is biased due to rounding, can be corrected according to (18), which leads to an approximately unbiased estimate of β:

$$\hat{\beta}^c := \hat{\beta}^* \left(1 - \frac{1}{12}\frac{h^2}{s^2_{x^*}}\right)^{-1} = \frac{s_{x^*y}}{s^2_{x^*} - \frac{h^2}{12}}. \tag{19}$$

If the response variable Y is rounded but not the covariate X, then because of (17), $\beta \approx \beta^*$. If both the response and the explanatory variables are rounded, the same result as (18) is obtained. Thus, in a linear regression of Y on X, only rounding of X and not of Y affects the value of the slope parameter.

These results can be easily generalized to a multiple linear regression $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$. If the variables X_i have been rounded to X_i^* with rounding width h_i, corrected values for the regression coefficients are found in the usual way as solutions to the normal equations, where the variances $\mathbb{V}X_i^*$ have been corrected by subtracting $h_i^2/12$ while leaving all the covariances unchanged. Using (11), one can find similar correction formulas for a polynomial regression model; see Müller (2008).

5 Rounding error

As mentioned in Sect. 2, the rounding error $\delta = X^* - X$ is not independent of X. Nevertheless, X and δ are approximately uncorrelated:

$$\mathbb{C}ov(X, \delta) = \mathbb{C}ov\big(X, X^* - X\big) = \mathbb{C}ov\big(X, X^*\big) - \mathbb{V}X \approx 0,$$

because, by (17) with $Y = X$, $\mathbb{C}ov(X, X^*) \approx \mathbb{V}X$. Moreover, the marginal distribution of δ is approximately the uniform distribution on the interval $[-\frac{h}{2}, \frac{h}{2}]$. Indeed, the marginal distribution of δ is given by

$$g(\delta) = \sum_{x^*} h\big(\delta | x^*\big) p\big(x^*\big), \quad -\frac{h}{2} < \delta < \frac{h}{2},$$

with $h(\delta|x^*)$ from (2), and consequently

$$g(\delta) = \sum_{x^*} \varphi\big(x^* - \delta\big) = \sum_{i} \varphi(ih - \delta). \tag{20}$$

Using the Euler–Maclaurin formula, the sum can be approximated by a corresponding integral

$$g(\delta) \approx \frac{1}{h}\int_{-\infty}^{\infty} \varphi(y-\delta)\,dy = \frac{1}{h}, \qquad -\frac{h}{2} < \delta < \frac{h}{2}.$$

Thus δ is approximately uniformly distributed on $[-\frac{h}{2}, \frac{h}{2}]$. It follows that $\mathbb{E}\delta \approx 0$ and $\mathbb{V}\delta \approx \frac{h^2}{12}$, which is Sheppard's correction. The representation (14) of the characteristic function of X^* confirms this result.

6 Goodness of the approximation

For practical purposes it is important to know by how much the moments computed from the rounded data differ from those of the original data. We compare the mean and variance of X^* with the mean and variance of X. The difference depends not only on h but also on the shift parameter a. It also depends on the underlying distribution φ of the unrounded data. Here we only study the standard normal distribution $X \sim N(0,1)$. For other distributions, see Tricker (1984b). From (4) we can compute the exact expected value of the rounded data X^* when $X \sim N(0,1)$:

$$\mathbb{E}X^* = \sum_i (i+a)h\left[\Phi\left(\left(i+a+\frac{1}{2}\right)h\right) - \Phi\left(\left(i+a-\frac{1}{2}\right)h\right)\right].$$

Figure 1 shows the mean of the rounded data for various values of h as a function of the shift parameter a. We need to consider only the interval $0 \le a < 1$, as with $a = 1$ the rounding lattice is in the same position as with $a = 0$. The length of the rounding interval h varies from two to four standard deviations of the distribution. We see that

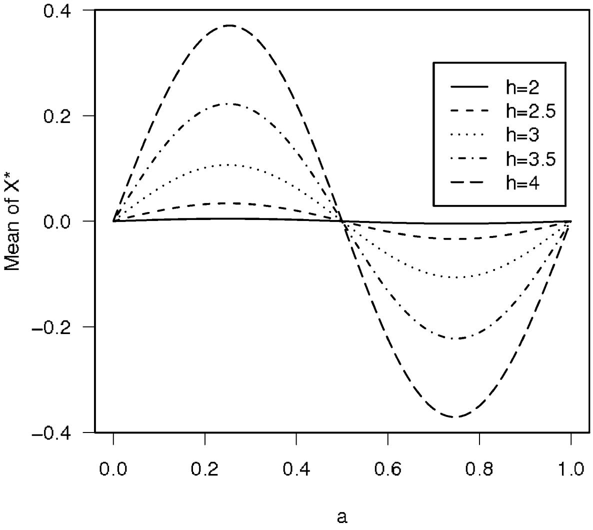


Fig. 1 Mean of the rounded data, μ^*, as a function of the shift parameter a for $X \sim N(0,1)$ and for various interval lengths h

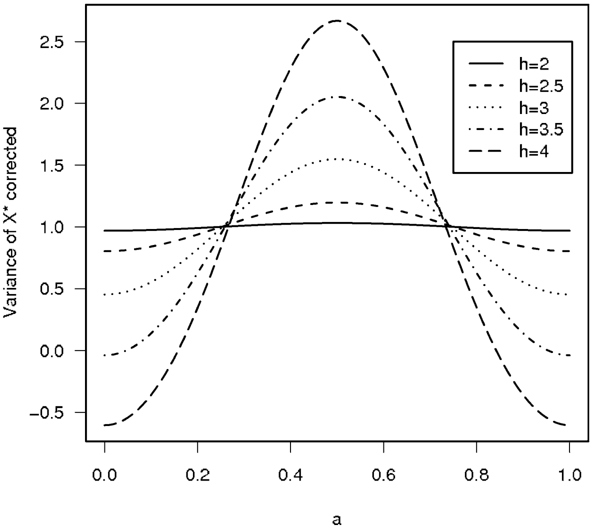


Fig. 2 Sheppard-corrected variance of the rounded data, $\sigma^2_{X^*} - \frac{h^2}{12}$, as a function of a, for $X \sim N(0,1)$

a rounding interval of length up to two standard deviations has a negligible influence on the data mean. The curve for $h = 2$ is almost indistinguishable from the zero line. Furthermore, the bias disappears for $a = 0$, $a = 1$ as well as $a = 0.5$.

The variance of the rounded data can be computed using the following equations:

$$\mathbb{E}X^{*2} = \sum_i \big[(i+a)h\big]^2 \left[\Phi\left(\left(i+a+\frac{1}{2}\right)h\right) - \Phi\left(\left(i+a-\frac{1}{2}\right)h\right)\right],$$

$$\mathbb{V}X^* = \mathbb{E}X^{*2} - \big(\mathbb{E}X^*\big)^2.$$

Figure 2 shows the Sheppard-corrected variance of the rounded data as a function of a for $X \sim N(0,1)$. In this case, the deviation from the variance of the unrounded data is largest for $a = 0.5$. Again, the correction performs quite well for rounding intervals h less than or equal to about two standard deviations. For larger rounding intervals, the Sheppard-corrected variance functions deviate from the true variance, particularly at $a = 0$, 0.5, and 1, where the approximation becomes very poor.

We obtain practically the same results if we use sampled data instead of the underlying distribution, even if the sample size n is small. In a simulation with 1000 replications we generated, for each replication, a sample of size $n = 30$ $N(0,1)$-variates and rounded these data with various values of h and a, where the origin of the rounding lattice was set at 0 for $a = 0$. For each replication, we computed the empirical means and variances of the rounded and unrounded data, denoted by m^*, m, v^*, v, respectively, and also the corrected variance v^c from the rounded data. We then computed the differences $d_{m^*} := m^* - m$, $d_{v^*} := v^* - v$, $d_{v^c} := v^c - v$ and averaged them over the replications. The results were in complete agreement with Figs. 1 and 2. For example, for $h = 1, 2, 3$ and $a = 0, 0.25, 0.5$, we found that the average difference $\bar{d}_{m^*}$ was always 0.00 except for $h = 3$, $a = 0.25$, where $\bar{d}_{m^*} = 0.11$, just as in Fig. 1. As to the differences in variances, we have the following table for $\bar{d}_{v^*}$ (in parentheses) and $\bar{d}_{v^c}$. Clearly, Sheppard's correction works extremely well for

Table 1 Differences of variances: $\bar{d}_{v^c}$ and $\bar{d}_{v^*}$ (in parentheses)

a	h		
	1	2	3
0	(0.09)	(0.30)	(0.20)
	0.00	−0.03	−0.55
0.25	(0.09)	(0.34)	(0.75)
	0.00	0.01	0.00
0.5	(0.08)	(0.37)	(1.29)
	0.00	0.04	0.54

Table 2 Regression coefficients

h_x	2	2	1
ρ	0.50	0.90	0.90
b	0.50	0.90	0.90
b^*	0.37	0.73	0.81
b^c	0.53	1.03	0.89

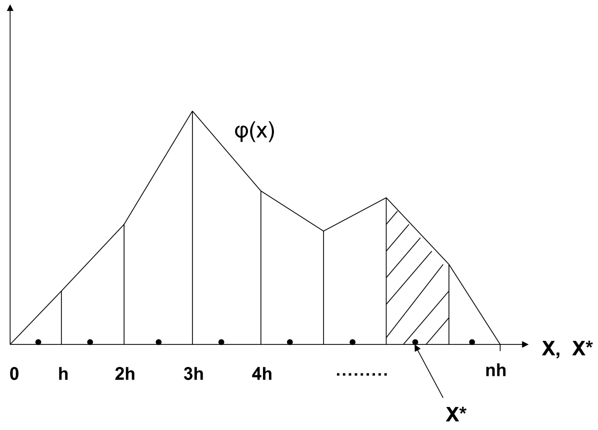
$h = 1$ and nicely even for $h = 2$, while not so well for $h = 3$. Note the rather large differences $\bar{d}_{v^*}$ of v^* and v for $h = 2$ in Table 1 and how these differences almost vanish after Sheppard's correction has been applied.

Summarizing, the approximation of the mean and variance is quite accurate even for values of h greater than 1 if the underlying distribution is Gaussian. For other, in particular for skew, distributions, Sheppard's correction only works well for considerably smaller values of h (Tricker 1984b). Higher moments are also less well-approximated by the corresponding correction formulas.

As to the accuracy of Sheppard's correction in regression analysis, we here report just a few selected results from a simulation study, see also Dempster and Rubin (1983) and Liu et al. (2007). We generated a sample of $n = 30$ jointly normally distributed pairs (X_i, Y_i) with means 0, variances 1 and covariance ρ. Both variables were rounded with rounding widths h_x and h_y, respectively, and origins of the rounding lattices at $(0, 0)$. We fixed $h_y = 2$. We computed estimates of the regression coefficient of a linear regression of Y on X with unrounded data, b, rounded data, b^*, and corrected, b^c. These were averaged over $N = 1000$ replications.

In all three cases considered (see Table 2) the uncorrected regression estimates b^* are far from the true value b and the corrected estimates b^c come much closer. However, for $\rho = 0.9$ and $h_x = 2$ the correction overshoots. This is due to the fact that with high correlation the joint distribution of (X, Y) does not comply very well with the Euler–Maclaurin condition and Sheppard's correction does not yield good results in this case, see also Daniels (1947). But reducing the rounding width h_x to $h_x = 1$ produces excellent results again.

Fig. 3 Piecewise linear lattice density



7 Some special distributions

7.1 Piecewise linear lattice density

The approximation formulas (8) and (9) for mean and variance become exact equalities when the density of the unrounded variable is a continuous, piecewise linear function on a finite interval $[c,d]$ with the following properties. The interval $[c,d]$ is subdivided into n subintervals of equal length h. Within each interval the density is a linear function. For simplicity let $c=0$, then $d=nh$. The density is zero at the endpoints of the interval $[c,d]$. The rounding lattice consists of all midpoints of the subintervals, $x_i^*=(i+\frac{1}{2})h$, $i=0,\dots,n-1$ (Fig. 3). Let us call such a density function together with its rounding lattice a "piecewise linear lattice density" (*plld*). In this case, the approximate relations for mean and variance become exact equations as long as $a=0$.

The value of the density function is given, for each rounding interval, by

$$\varphi(x^*+u)=\varphi(x^*)+\varphi'(x^*)u, \quad -\frac{h}{2}\le u\le\frac{h}{2}. \tag{21}$$

We use the following identities:

$$h\sum\varphi(x^*)=1, \qquad \sum\varphi'(x^*)=0, \qquad \sum x^*\varphi'(x^*)=-\sum\varphi(x^*), \tag{22}$$

which follow from the definition of the *plld*. The expected values of the unrounded and rounded data, respectively, are

$$\mathbb{E}X=\sum_{x^*}\int_{x^*-\frac{h}{2}}^{x^*+\frac{h}{2}} x\varphi(x)\,dx=\sum_{x^*}\int_{-\frac{h}{2}}^{\frac{h}{2}}(x^*+u)\varphi(x^*+u)\,du$$

$$\mathbb{E}X^*=\sum_{x^*}x^*\int_{x^*-\frac{h}{2}}^{x^*+\frac{h}{2}}\varphi(x)\,dx=\sum_{x^*}x^*\int_{-\frac{h}{2}}^{\frac{h}{2}}\varphi(x^*+u)\,du,$$

and the difference is

$$\begin{aligned}\mathbb{E}X - \mathbb{E}X^* &= \sum_{x^*} \int_{-\frac{h}{2}}^{\frac{h}{2}} u\varphi\big(x^* + u\big)\,du \\ &= \sum_{x^*} \int_{-\frac{h}{2}}^{\frac{h}{2}} \big[u\varphi\big(x^*\big) + u^2\varphi'\big(x^*\big)\big]\,du = 0,\end{aligned}$$

where we used (21) and (22). Thus for the *plld*, $\mathbb{E}X^* = \mathbb{E}X$.

To compute the variance of the unrounded and rounded data, respectively, we first analyze the second moments of X and X^*:

$$\begin{aligned}\mathbb{E}X^2 &= \sum_{x^*} \int_{-\frac{h}{2}}^{\frac{h}{2}} \big(x^* + u\big)^2 \varphi\big(x^* + u\big)\,du, \\ \mathbb{E}X^{*2} &= \sum_{x^*} x^{*2} \int_{-\frac{h}{2}}^{\frac{h}{2}} \varphi\big(x^* + u\big)\,du.\end{aligned}$$

The difference is

$$\begin{aligned}\mathbb{E}X^2 - \mathbb{E}X^{*2} &= \sum_{x^*} \int_{-\frac{h}{2}}^{\frac{h}{2}} (u^2 + 2ux^*)\big[\varphi\big(x^*\big) + \varphi'\big(x^*\big)u\big]\,du \\ &= \frac{h^2}{12}\Big[h\sum_{x^*}\varphi\big(x^*\big) + 2h\sum_{x^*} x^*\varphi'\big(x^*\big)\Big] = -\frac{h^2}{12},\end{aligned}$$

where we used (21) and (22). It follows that

$$\mathbb{V}X^* = \mathbb{V}X + \frac{h^2}{12}.$$

Thus in the case of a *plld*, Sheppard's correction for the variance holds exactly.

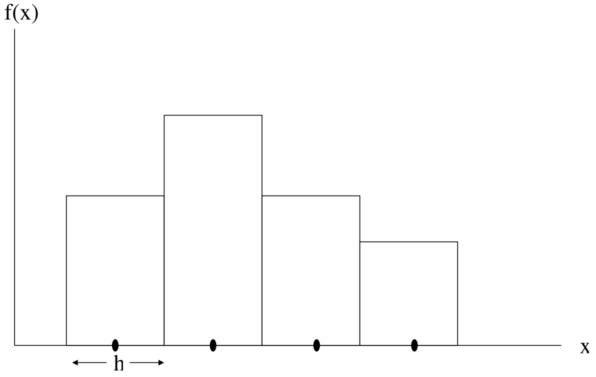
This cannot be said of the moments higher than the third. For these moments, the correction formula for the *plld* differs from the approximate general formula, but they agree up to the h^2 term, see (13), and only differ in terms of powers higher than h^2.

One can also show that for the *plld* the marginal distribution of δ is not only approximately but also exactly uniformly distributed on $[-\frac{h}{2}, \frac{h}{2}]$. Indeed, by (20), (21), and (22)

$$g(\delta) = \sum_{x^*}\big[\varphi\big(x^*\big) - \delta\varphi'\big(x^*\big)\big] = \frac{1}{h}, \quad -\frac{h}{2} \le \delta \le \frac{h}{2}.$$

The piecewise linear lattice density may be a rather artificial density function. But as far as other, more realistic, densities can be approximated by a *plld*, the latter serves as a convenient model to explain the approximate relations between the moments of rounded and unrounded data.

Fig. 4 Histogram density



7.2 Histogram density

A completely different picture arises if we consider a distribution of X which has a histogram density instead of a *plld*, where the rounding lattice is made up by the midpoints of the histogram intervals (the bins), Fig. 4. The uniform distribution over an interval which is subdivided into rounding intervals of equal length is a special case. Note that this distribution does not satisfy the requirements for the Euler–Maclaurin approximation. For more details, see Schneeweiss et al. (2006).

For a variable X following a histogram density, the measurement equation (1) can be written as $X = X^* - \delta$. But now δ is uniformly distributed on the interval $[-\frac{h}{2}, \frac{h}{2}]$ and is independent of X^*. (In the theory of measurement error models this is Berkson's case.) It follows that $\mathbb{E}X = \mathbb{E}X^*$ and

$$\mathbb{V}X = \mathbb{V}X^* + \mathbb{V}\delta = \mathbb{V}X^* + \frac{h^2}{12}. \tag{23}$$

This is just the reverse of Sheppard's correction formula (10). Instead of subtracting $\frac{h^2}{12}$, we have to add this term. Other correction terms arise if the rounding lattice is shifted ($a \neq 0$), see Liu et al. (2007).

This example shows that completely different results with respect to Sheppard's correction can be obtained when the assumptions for the Euler–Maclaurin formula are not satisfied; see also Example 2.9 in Janson (2006).

The same can be said of any distribution of X which, like the uniform, is restricted to a finite interval and does not tend to 0 at the endpoints smoothly. A normal distribution truncated at one or both sides is a case in point. The breakdown of Sheppard's correction for this case has been studied in Pairman and Pearson (1919).

8 Estimation and testing

8.1 Estimation

Up to now we only dealt with various population parameters (moments and regression coefficients) of rounded and unrounded random variables and their relations to each other. Let us now consider estimating and testing problems, in particular, estimating the mean $\mu = \mathbb{E}X$ of the underlying random variable X.

As we do not observe X but rather the rounded variable X^*, we have to use the rounded data x_i^*, $i = 1, \dots, n$, in order to estimate μ. Let us assume that the conditions for Sheppard's correction are satisfied. If the original, unrounded, data x_i, $i = 1, \dots, n$, is an iid sample, so is the rounded data x_i^*, $i = 1, \dots, n$. The arithmetic mean $\bar{x}^*$ of the x_i^* is therefore an unbiased as well as strongly consistent estimate of $\mu^* = \mathbb{E}X^*$. If the Euler–Maclaurin conditions are satisfied, μ^* and μ are approximately equal, and therefore $\bar{x}^* =: \hat{\mu}^*$ is also an approximately unbiased and consistent estimate of μ. So we can estimate μ (at least approximately) without bias even if only rounded data are available:

$$\mathbb{E}\hat{\mu}^* = \operatorname*{plim}_{n\to\infty} \hat{\mu}^* = \mu^* \approx \mu.$$

It should be noted that the bias due to rounding, though small, becomes noticeable if n is large, while for small n it is swamped with sample randomness.

In a similar way, we can use the rounded data to estimate the variance of X, $\sigma_x^2 = \mathbb{V}X$. However, here we must observe Sheppard's correction. Thus

$$\mathbb{E}s_{x^*}^2 = \operatorname*{plim}_{n\to\infty} s_{x^*}^2 = \sigma_{x^*}^2, \quad \sigma_x^2 \approx \sigma_{x^*}^2 - \frac{h^2}{12},$$

where $s_{x^*}^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i^* - \bar{x}^*)^2$. Hence $s_{x^*}^2 - \frac{h^2}{12}$ is an approximately unbiased estimate of σ_x^2.

Similarly, the slope parameter β of a linear regression can be consistently estimated with rounded data as long as Sheppard's correction is taken into account.

It should, however, be kept in mind that all these estimates are less efficient than those computed from the unrounded data. Rounding leads to an efficiency loss. This can be clearly seen in the case of estimating the mean. The variance of $\hat{\mu}^* = \bar{x}^*$ is $\frac{1}{n}\sigma_{x^*}^2$, while the variance of $\hat{\mu} = \bar{x}$ is $\frac{1}{n}\sigma_x^2$, and $\sigma_{x^*}^2 \approx \sigma_x^2 + \frac{h^2}{12} > \sigma_x^2$. Thus the estimate from the rounded data has a larger variance than the estimate from the unrounded data. A confidence interval constructed from the rounded data,

$$\bar{x}^* \pm t_{1-\frac{\alpha}{2}} \frac{s_{x^*}}{\sqrt{n}},$$

tends to be systematically larger than the corresponding confidence interval from the unrounded data,

$$\bar{x} \pm t_{1-\frac{\alpha}{2}} \frac{s_x}{\sqrt{n}}.$$

The efficiency loss is measured by the ratio, see also Gjeddebaek (1956),

$$\frac{\sigma_{x^*2}}{\sigma_x^2} \approx 1 + \frac{h^2}{12\sigma_x^2}.$$

8.2 t-Test

The efficiency loss due to rounding can also be seen in parameter tests. As an example, consider testing the mean of a $N(\mu, \sigma_x^2)$ distribution. In the one-sided case, the null hypothesis to be tested is

$$H_0 : \mu \leq \mu_0.$$

Then, the t-test statistic, τ, computed from the unrounded data is given by

$$\tau = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}_{\hat{\mu}}},$$

where $\hat{\mu} = \bar{x}$, $\hat{\sigma}_{\hat{\mu}} = \frac{\hat{\sigma}_x}{\sqrt{n}}$, and $\hat{\sigma}_x^2 = \frac{1}{n-1}\sum_i^n (x_i - \bar{x})^2$. Under H_0, the test statistic τ has the Student t-distribution, $\tau \sim t$. For rounded data, the null hypothesis can be stated as

$$H_0 : \mu^* \leq \mu_0$$

because $\mu^* \approx \mu$.

The corresponding test statistic is

$$\tau^* = \frac{\hat{\mu}^* - \mu_0}{\hat{\sigma}_{\hat{\mu}^*}},$$

where $\hat{\mu}^* = \bar{x}^*$, $\hat{\sigma}_{\hat{\mu}^*} = \frac{\hat{\sigma}_{x^*}}{\sqrt{n}}$ and $\hat{\sigma}_{x^*}^2 = \frac{1}{n-1}\sum_i^n (x_i^* - \bar{x}^*)^2$. This test statistic is no longer t-distributed because the rounded data are no longer normally distributed—they follow a discrete distribution. However, for large n, the distribution of τ^* converges to the standard normal distribution $N(0, 1)$, and this distribution can be used to construct an asymptotic Gauss-test[1] of H_0. The test is constructed such that H_0 is rejected whenever $\tau^* > t_{1-\alpha}$, where $t_{1-\alpha}$ is the $(1 - \alpha)$ quantile of the $N(0, 1)$ distribution.

This test is unbiased (at least approximately so), but it has smaller power than the corresponding test with the unrounded data. The power functions of these two tests are given by

$$\pi(\mu) = P\big(\tau > t_{1-\alpha} \mid \mu\big),$$
$$\pi^*(\mu) = P\big(\tau^* > t_{1-\alpha} \mid \mu\big).$$

We compute $\pi(\mu)$ for $\mu > \mu_0$. For simplicity of notation we denote $t_{1-\alpha}$ by t

$$\pi(\mu) = P\left(\frac{\hat{\mu} - \mu_0}{\hat{\sigma}_{\hat{\mu}}} > t \,\middle|\, \mu\right) = P\left(\frac{\hat{\mu} - \mu}{\hat{\sigma}_{\hat{\mu}}} + \frac{\mu - \mu_0}{\hat{\sigma}_{\hat{\mu}}} > t \,\middle|\, \mu\right).$$

[1] Econometricians often call such a test still a t-test, although it is actually an asymptotic Gauss-test.

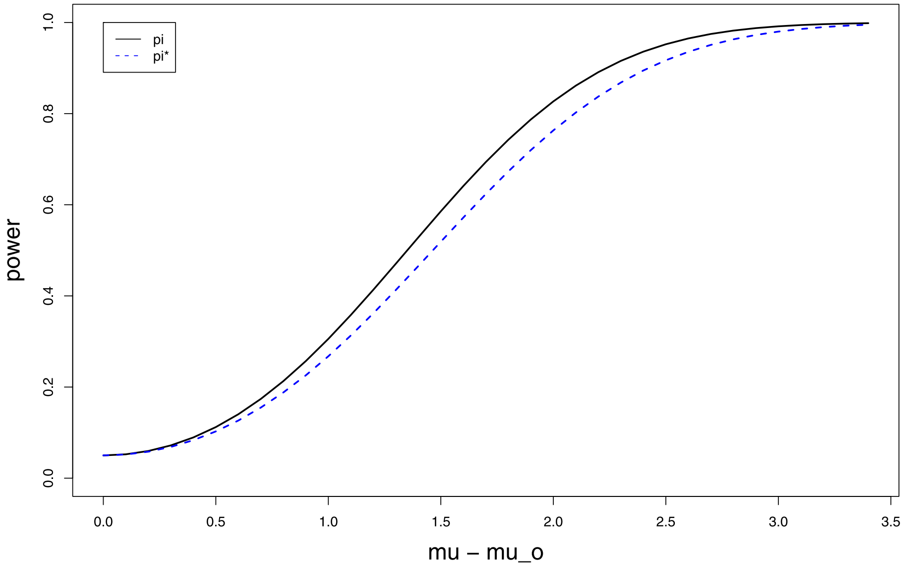


Fig. 5 Two power functions for the same test with rounded and unrounded data

For sufficiently large n, $\hat{\sigma}_{\hat{\mu}}$ can be replaced with $\sigma_{\hat{\mu}} = \sigma_x/\sqrt{n}$ and $(\hat{\mu} - \mu)/\hat{\sigma}_{\hat{\mu}} \sim N(0, 1)$. Thus for large n,

$$\pi(\mu) \approx 1 - \Phi\left(t - \frac{\mu - \mu_0}{\sigma_x}\sqrt{n}\right). \tag{24}$$

Similarly,

$$\pi^*(\mu) \approx 1 - \Phi\left(t - \frac{\mu - \mu_0}{\sigma_{x^*}}\sqrt{n}\right). \tag{25}$$

But since, by Sheppard's correction, $\sigma_{x^*} > \sigma_x$, obviously

$$\pi^*(\mu) < \pi(\mu) \quad \text{under } H_1 : \mu > \mu_0.$$

So the test with rounded data has smaller power than the test with unrounded data and is thus less efficient. Figure 5 shows the power functions π and π^* as functions of $\mu - \mu_0$. π and π^* have been computed according to (24) and (25) with $n = 100$, $\alpha = 5\%$, $h = 10$, $\sigma_x^2 = 47.5$.

A discussion of the t-test with rounded data for small sample size is given in Eisenhart (1947). Tricker (1990a, 1990b) studied the performance of the t-, double t-, χ^2-, and F-tests for rounded Gaussian variables, however, without adjusting the variances in the last two tests. A comparison of the confidence intervals for μ and σ^2 based on the t- and χ^2-pivotals, respectively, and on the likelihood ratio test can be found in Lee and Vardeman (2001, 2002); see also Vardeman and Lee (2005).

9 ML estimation when h is large

The approximation of the moments of the rounded data for large rounding intervals h is often rather poor. For this reason, it is sometimes better to estimate the parameters of the data using the Maximum Likelihood (ML) method. The disadvantage of ML is that it needs a parametric distribution for X in order to work.

Thus let $\varphi(x,\theta)$ be the density of X with unknown parameter vector θ. The discrete distribution $p(x^*|\theta)$ of X^* is given by (3). Given a sample of data x_i^*, $i = 1, \dots, n$, the log-likelihood is

$$l = l(\theta) = \sum_i \log p\left(x_i^* | \theta\right), \tag{26}$$

which has to be maximized in order to obtain the ML estimate $\hat{\theta}_{\mathrm{ML}}$.

There is a link from the ML approach to Sheppard's correction. Suppose one wants to find the maximum of l by Newton's method. Start with the naive estimate θ_0 as an initial parameter estimate, where θ_0 is computed as the ML estimate of θ from the distribution of X, but with X replaced by X^*. An improved estimate is given by

$$\theta_1 = \theta_0 - \left(\frac{\partial^2 l}{\partial\theta_0 \partial\theta_0^\top}\right)^{-1} \frac{\partial l}{\partial\theta_0}. \tag{27}$$

This procedure may be repeated with θ_1 in place of θ_0, and so on. But θ_1 is often good enough, in particular if h is small.

The derivatives in (27) are not always easy to compute. But for small h, following Lindley (1950), approximations to these derivatives can be found easily using a Taylor series expansion of $\varphi(x)$ at x^*:

$$\varphi(x) = \varphi\left(x^*\right) + \varphi'\left(x^*\right)\left(x - x^*\right) + \frac{1}{2}\varphi''\left(x^*\right)\left(x - x^*\right)^2 + \cdots .$$

Then, omitting terms of higher order in h, (3) yields

$$p\left(x^*\right) \approx h\varphi\left(x^*\right) + \frac{h^3}{24}\varphi''\left(x^*\right).$$

Taking logarithms, we obtain (again omitting terms of higher order in h)

$$\begin{aligned} \log p\left(x^*\right) &\approx \log h + \log\varphi\left(x^*\right) + \log\left(1 + \frac{h^2}{24}\frac{\varphi''(x^*)}{\varphi(x^*)}\right) \\ &\approx \log h + \log\varphi\left(x^*\right) + \frac{h^2}{24}\frac{\varphi''(x^*)}{\varphi(x^*)}, \end{aligned}$$

and this expression can be substituted into (26). We then obtain approximate expressions for the derivatives in (27):

$$\frac{\partial l}{\partial\theta_0} \approx \frac{h^2}{24}\sum_i \frac{\partial}{\partial\theta_0}\frac{\varphi''(x_i^*)}{\varphi(x_i^*)}, \tag{28}$$

$$\frac{\partial^2 l}{\partial\theta_0 \partial\theta_0^\top} \approx \sum_i \frac{\partial^2}{\partial\theta_0 \partial\theta_0^\top} \log\varphi\left(x_i^*\right). \tag{29}$$

In obtaining (28), we made use of the fact that

$$\frac{\partial}{\partial\theta_0} \sum \log\varphi\left(x_i^*\right) = 0. \tag{30}$$

Indeed, the initial estimate θ_0 is found by solving the likelihood score equation (30) of the original model with the rounded data x_i^* in place of the original data x_i. In (29) terms of order h^2 were omitted.

Substituting (28) and (29) into (27) yields a first step approximation to the ML estimator of θ:

$$\theta_1 \approx \theta_0 - \frac{h^2}{24}\left(\sum_i \frac{\partial^2}{\partial\theta_0 \partial\theta_0^\top} \log\varphi\left(x_i^*\right)\right)^{-1}\left(\sum_i \frac{\partial}{\partial\theta_0} \frac{\varphi''(x_i^*)}{\varphi(x_i^*)}\right). \tag{31}$$

The difference $\theta_1 - \theta_0$ in (31) can be regarded as an analogue to Sheppard's correction stemming from ML estimation theory rather than from moment considerations. Note, however, that unlike Sheppard's correction, this correction depends on the distribution involved.

In the case of estimating $\theta = (\mu, \sigma^2)^\top$ from a normal distribution $N(\mu, \sigma^2)$, we find with some algebra from (28) and (29) that

$$\frac{\partial l}{\partial\theta_0} \approx -\frac{h^2}{12 s_{x^*}^4} \sum_i \begin{pmatrix} x_i^* - \bar{x}^* \\ \frac{(x_i^* - \bar{x}^*)^2}{s_{x^*}^2} - \frac{1}{2} \end{pmatrix} = -\frac{h^2 n}{12 s_{x^*}^4} \begin{pmatrix} 0 \\ \frac{1}{2} \end{pmatrix},$$

$$\frac{\partial^2 l}{\partial\theta_0 \partial\theta_0^\top} \approx -\frac{1}{s_{x^*}^4} \sum_i \begin{pmatrix} s_{x^*}^2 & x_i^* - \bar{x}^* \\ x_i^* - \bar{x}^* & \frac{(x_i^* - \bar{x}^*)^2}{s_{x^*}^2} - \frac{1}{2} \end{pmatrix} = -\frac{n}{s_{x^*}^4} \begin{pmatrix} s_{x^*}^2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}.$$

Then (31) yields

$$\begin{pmatrix} \mu_1 \\ \sigma_1^2 \end{pmatrix} \approx \begin{pmatrix} \bar{x}^* \\ s_{x^*}^2 \end{pmatrix} - \frac{h^2}{12} \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

and this is just Sheppard's correction for μ and σ^2.

Fryer and Pethybridge (1972) extend this result to higher orders of h and to the multivariate case and apply it to the estimation of a normal linear regression, see also Don (1981). Johnson et al. (2004) derive Sheppard's correction in a similar way for unequal rounding intervals and apply their method to data from a Beta distribution. For extensions to other distributions, see Tallis (1967). Alternatively, one can derive Sheppard's correction from the first step of an EM algorithm to solve the ML equations for μ and σ^2 of a normal distribution, see Dempster and Rubin (1983). McNeil (1966) finds correction terms by using a method of moments for the rounded data instead of ML.

Apart from providing a new justification for Sheppard's correction in the case of a Gaussian distribution and for Sheppard-like corrections in other distributions, ML is a general procedure to estimate unknown parameters with rounded data without the need to take refuge in Sheppard's correction. One may even have unequal rounding intervals. Gjeddebaek (1949) estimates the normal distribution by ML. Kulldorff (1961) and Tricker (1984a) consider the ML estimation of the exponential, Tallis and Young (1962) of the log-normal and the truncated normal, and Tricker (1992) of the Gamma distribution. As the two parameters (shape α and scale θ) of the Gamma distribution can be estimated from first and second moments, Sheppard's correction can also be applied when rounded data are used. But for $h > 1$, the bias (and also the variance) of the Sheppard corrected moment estimators of α and θ turn out to be much larger than for the ML estimators when α is small, i.e., when the Gamma distribution is very skew. Similarly, the exponential distribution with rounded data is better estimated by ML rather than just by using the mean.

There are several numerical methods for carrying out ML like Newton–Raphson, Fisher Scoring, EM, etc. For a comparison see Schader and Schmid (1984).

In the context of a linear regression model, Liu et al. (2007) apply an approximate (so-called two-stage) ML approach and compare it with Sheppard's correction. ML estimation of stochastic processes, in particular ARMA processes, has been studied by Zhang et al. (2010).

It is worth repeating: as long as there is a parametric model of X, ML can be employed to estimate the distribution consistently with the rounded data, and that is possible even if the rounding intervals are not of equal size, thus bypassing Sheppard's correction.

If, however, the model is of a non-parametric nature, corresponding methods, like kernel density or spline methods, suitably modified can be applied to the rounded data, see Braun et al. (2005), Lambert and Eilers (2009). For example, in the latter case, the idea is, roughly speaking, to estimate the probabilities on a very fine grid, which is supposed to mimic the domain of the continuous variable, using the observed frequencies on the coarse grid by applying a P-spline approach in a Bayesian framework. But unlike non-parametric models with *unrounded* data, these methods will not lead to consistent estimates in general, see also Hall (1982).

10 Asymmetric rounding

Sometimes rounded data show a marked preference for particular numbers, e.g., a preference for even over odd lattice points. A typical example is illustrated in Fig. 6, where a frequency distribution of historical height measurements is shown (Wu 1994). The measurements ending with one half inch, of which there were only very few, were deleted. The remaining measurements show a slight but marked preponderance of even values.

There are several ways to address such a problem. One possibility is to introduce a mixture of rounding procedures, see Wright and Bray (2003), Wang and Heitjan (2008). One might also think of describing the rounding procedure by a rounding probability function, see Rietveld (2002), Schneeweiss and Komlos (2009). Here we

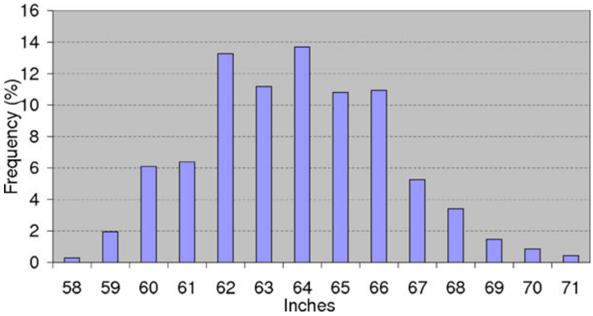


Fig. 6 Height distribution of black female voters in Pittsburgh, born 1887–1955

follow another suggestion: asymmetric rounding, see Komlos (1999). This means that the rounding intervals about x^* and $x^* + h$ are not equal. An example would be if even numbers were preferred and values (0.75 to 1.25) were rounded to 1 while values (1.25 to 2.75) were rounded to 2. Asymmetric rounding has no implications for the expected value, but it does have an influence on the variance.

We can determine the moments of X^* in the same way as for the symmetric case (Sect. 3). Let us assume that all values of X in the interval $[2ih - rh, 2ih + rh]$, $0 \leq r \leq 1$, are rounded to $x^* = 2ih$ while all values of X in the interval $[(2i+1)h - (1-r)h, (2i+1)h + (1-r)h]$ are rounded to $x^* + h = (2i+1)h$, $i \in \mathbb{Z}$. Again we assume without loss of generality that $a = 0$. First note that

$$p(2ih) := \mathbb{P}\big(X^* = 2ih\big) = \int_{-rh}^{rh} \varphi(2ih + u)\, du,$$

$$p\big((2i+1)h\big) := \mathbb{P}\big(X^* = (2i+1)h\big) = \int_{-(1-r)h}^{(1-r)h} \varphi\big((2i+1)h + u\big)\, du.$$

Then the kth moment of X^* is

$$\mathbb{E}X^{*k} = \sum_i (2ih)^k \int_{-rh}^{rh} \varphi(2ih + u)\, du + \sum_i \big[(2i+1)h\big]^k \int_{-(1-r)h}^{(1-r)h} \varphi\big((2i+1)h + u\big)\, du.$$

Using the Euler–Maclaurin approximation, we obtain

$$\mathbb{E}X^{*k} \approx \frac{1}{2} \int_{-\infty}^{\infty} \left[\int_{-r}^{r} (x - vh)^k\, dv + \int_{-(1-r)}^{1-r} (x - vh)^k\, dv \right] \varphi(x)\, dx.$$

Setting $k = 1$, we obtain for the mean of X^*:

$$\mathbb{E}X^* \approx r \int_{-\infty}^{\infty} x\varphi(x)\, dx + (1-r) \int_{-\infty}^{\infty} x\varphi(x)\, dx = \mathbb{E}X. \tag{32}$$

Thus the means of X^* and X are approximately equal.

For $k=2$, we obtain:

$$\begin{aligned}\mathbb{E}X^{*2} &\approx r\mathbb{E}X^2 + \frac{r^3}{3}h^2 + (1-r)\mathbb{E}X^2 + \frac{(1-r)^3}{3}h^2\\ &= \mathbb{E}X^2 + \frac{1}{3}\bigl(1-3r+3r^2\bigr)h^2.\end{aligned}$$

Together with (32) this implies

$$\mathbb{V}X^* \approx \mathbb{V}X + \frac{h^2}{3}\bigl(1-3r+3r^2\bigr) =: \mathbb{V}X + f(r)h^2. \tag{33}$$

The function $f(r)$ has a minimum at $r=\frac{1}{2}$, which corresponds to symmetric rounding, and at this point $f(\frac{1}{2})=\frac{1}{12}$, which is Sheppard's correction. Thus the term $f(r)h^2$ is a generalization of Sheppard's correction to the case of asymmetric rounding. The function $f(r)$ reaches its maximum for $r=0$ and $r=1$, which means that $X^*=2ih$ or $X^*=(2i+1)h$, respectively, has a rounding interval of width zero.

In the example of Fig. 6, we find $\mathbb{V}X^*=6.050$, which has to be corrected according to (33). The value of r can be found approximately through the equation

$$r \approx \sum_i \mathbb{P}\bigl(X^*=2ih\bigr),$$

which can be derived using the Euler–Maclaurin formula again. So we estimate r by the proportion of the frequencies at even values. We find $r\approx 0.564$ and thus $f(r)h^2 = 0.088$ with $h=1$, so that $\mathbb{V}X \approx 5.962$. The difference to the uncorrected variance is not very large in this example, but this is due to the relatively small value of h.

Multivariate moments are treated in the same way, see also Sect. 3.3. We find, e.g., for the second mixed moment of X^* and Y:

$$\mathbb{E}\bigl(X^*Y\bigr) \approx \mathbb{E}(XY).$$

Similar relations hold if both X and Y are or if only Y is (asymmetrically) rounded. As a consequence, the previous equations (16) and (17) for the covariances of rounded and unrounded variables, hold also true in the case of asymmetric rounding.

As in Sect. 4, we can use these results to study the distortion of the slope parameter in a linear regression due to asymmetric rounding. When only X is rounded, we have, see also (18),

$$\beta \approx \frac{\mathbb{C}ov(X,Y)}{\mathbb{V}X} \approx \frac{\mathbb{C}ov(X^*,Y)}{\mathbb{V}X^* - f(r)h^2} = \beta^*\left(1 - f(r)\frac{h^2}{\mathbb{V}X^*}\right)^{-1}.$$

11 Conclusion

Rounding of data has the inevitable consequence that their statistical moments, in particular the variance (and consequently also regression parameters), computed from such data are somewhat distorted in comparison to the moments of the unrounded data. This survey explores the magnitude of this distortion and when and how it can be approximated by simple expressions depending on the length of the rounding interval. Sheppard's correction for the variance is the best known approximation in this context. We study cases where it is appropriate and, indeed, where it is exact and other cases where it is completely misleading.

We also consider estimating and testing moments (and regression parameters) on the basis of a random sample of rounded data. Clearly, rounding implies a loss of efficiency, even though the bias may often be negligible (after appropriate correction). When rounding is so coarse that the approximation formulas fail, maximum likelihood must be employed to obtain consistent estimates.

Sheppard's correction is generalized to the case of asymmetric rounding. The correction turns out to be a function of the asymmetry portion r. Asymmetric rounding can be extended to more general rounding procedures, which, however, is a subject for further research.

Acknowledgement We thank Thomas Augustin for helpful discussions. We also appreciate some well-founded suggestions by two anonymous referees to improve an earlier version of this paper.

References

Augustin, T., Wolff, J.: A bias analysis of Weibull models under heaped data. Stat. Pap. **45**, 211–229 (2004)

Baten, W.D.: Correction for the moments of a frequency distribution in two variables. Ann. Math. Stat. **2**, 309–312 (1931)

Braun, J., Duchesne, T., Stafford, J.E.: Local likelihood density estimation for interval censored data. Can. J. Stat. **33**, 39–59 (2005)

Crockett, A., Crockett, R.: Consequences of data heaping in the British religious census of 1851. Hist. Methods **39**, 24–47 (2006)

Daniels, H.E.: Grouping correction for high autocorrelations. J. R. Stat. Soc. B **9**, 245–249 (1947)

Dempster, A.P., Rubin, D.B.: Rounding error in regression: the appropriateness of Sheppard's correction. J. R. Stat. Soc. B **45**, 51–59 (1983)

Don, F.J.H.: A note on Sheppard's corrections for grouping and maximum likelihood estimation. J. Multivariate Anal. **11**, 452–458 (1981)

Eisenhart, C.: Effects of rounding or grouping data. In: Eisenhart, C., Hastay, M.W., Wallis, W.A. (eds.) Selected Techniques of Statistical Analysis, pp. 185–223. McGraw-Hill, New York/London (1947). Chapter 4

Fryer, J.G., Pethybridge, R.J.: Maximum likelihood estimation of a linear regression function with grouped data. Appl. Stat. **21**, 142–154 (1972)

Gjeddebaek, N.F.: Contribution to the study of grouped observations: I. Application of the method of maximum likelihood in case of normally distributed observations. Skand. Aktuarietidskrift **32**, 135–159 (1949)

Gjeddebaek, N.F.: Contribution to the study of grouped observations: II. Loss of information caused by groupings of normally distributed observations. Skand. Aktuarietidskrift **39**, 154–159 (1956)

Gjeddebaek, N.F.: Statistical analysis: III. Grouped observations. In: Sills, D.R. (ed.): International Encyclopedia of Social Sciences, vol. 15, pp. 193–196. Macmillan/Free Press, New York (1968)

Gray, R.M., Neuhoff, D.L.: Quantization. IEEE Trans. Inf. Theory **44**, 1–63 (1998)

Haitovsky, Y.: In: Grouped Data. Encyclopedia of Statistical Sciences, vol. 3, pp. 527–536. Wiley, New York (1982)

Hall, P.: The influence of rounding errors on some nonparametric estimators of a density and its derivatives. SIAM J. Appl. Math. **42**, 390–399 (1982)

Heitjan, D.F.: Inference from grouped continuous data: a review. Stat. Sci. **4**, 164–179 (1989)

Heitjan, D.F., Rubin, D.B.: Ignorability and coarse data. Ann. Stat. **19**, 2244–2253 (1991)

Janson, S.: Rounding of continuous random variables and oscillatory asymptotics. Ann. Probab. **34**, 1807–1826 (2006)

Johnson, D.S., Barry, R.P., Bowyer, R.T.: Estimating timing of life-history events with coarse data. J. Mammal. **85**, 932–939 (2004)

Kendall, M.G.: The conditions under which Sheppard's corrections are valid. J. R. Stat. Soc. **101**, 592–605 (1938)

Komlos, J.: On the nature of the Malthusian threat in the eighteenth century. Econ. Hist. Rev. **52**, 730–748 (1999)

Kullback, S.: A note on Sheppard's corrections. Ann. Math. Stat. **6**, 158–159 (1935)

Kulldorff, G.: Contributions to the Theory of Estimation from Grouped and Partially Grouped Samples. Almqvist and Wiksell, Stockholm (1961)

Lambert, P., Eilers, P.H.C.: Bayesian density estimation from grouped continuous data. Comput. Stat. Data Anal. **53**, 1388–1399 (2009)

Lee, C.-S., Vardeman, S.B.: Interval estimation of a normal process mean from rounded data. J. Qual. Technol. **33**, 335–348 (2001)

Lee, C.-S., Vardeman, S.B.: Interval estimation of a normal process standard deviation from rounded data. Commun. Stat., Part B: Simul. Comput. **31**, 13–34 (2002)

Lindley, D.V.: Grouping corrections and maximum likelihood equations. Proc. Camb. Philos. Soc. **46**, 106–110 (1950)

Liu, T.Q., Zhang, B.X., Hu, G.R., Bai, Z.D.: Revisit of Sheppard corrections in linear regression. RMI Working Paper 07/06, Berkeley-NSU (2007)

McNeil, D.R.: Consistent statistics for estimating and testing hypotheses from grouped samples. Biometrika **53**, 545–557 (1966)

Müller, S.: Zuverlässige statistische Modellierung bei gerundeten Daten. Diplomarbeit. Department of Statistics, Ludwig-Maximilian University Munich (2008)

Myers, R.J.: Accuracy of age reporting in the 1950 United States census. J. Am. Stat. Assoc. **49**, 826–831 (1954)

Pairman, E., Pearson, K.: On correcting for the moment-coefficients of limited range frequency-distributions when there are finite or infinite ordinates and any slopes at the terminals of range. Biometrika **12**, 231–258 (1919)

Rietveld, P.: Rounding of arrival and departure times in travel surveys: an interpretation in terms of scheduled activities. J. Transp. Stat. **5**, 71–82 (2002)

Schader, M., Schmid, F.: Computation of maximum likelihood estimates for μ and σ from a grouped sample of a normal population. A comparison of algorithms. Stat. Pap. **25**, 245–258 (1984)

Schneeweiss, H., Komlos, J.: Probabilistic rounding and Sheppard's correction. Stat. Methodol. **6**, 577–593 (2009)

Schneeweiss, H., Komlos, J., Ahmad, A.S.: Symmetric and asymmetric rounding. Discussion Paper 479, Sonderforschungsbereich 386, University of Munich (2006)

Sheppard, W.F.: On the calculation of the most probable values of frequency constants for data arranged according to equidistant divisions of a scale. Proc. Lond. Math. Soc. **29**, 353–380 (1898)

Stuart, A., Ord, J.K.: Kendall's Advanced Theory of Statistics. Distribution Theory, vol. 1, 5th edn. Charles Griffin, London (1987)

Stoer, J., Bulirsch, R.: Introduction to Numerical Analysis. Springer, New York (1980)

Tallis, G.M.: Approximate maximum likelihood estimation from grouped data. Technometrics **9**, 599–606 (1967)

Tallis, G.M., Young, S.S.: Maximum likelihood estimation of parameters of the normal, log-normal, truncated normal and bivariate normal distributions from grouped data. Aust. J. Stat. **4**, 49–54 (1962)

Tricker, A.R.: Estimation of rounding data sampled from the exponential distribution. J. Appl. Stat. **11**, 51–87 (1984a)

Tricker, A.R.: Effects of rounding on the moments of a probability distribution. Statistician **33**, 381–390 (1984b)

Tricker, A.R.: The effect of rounding on the significance level of certain normal test statistics. J. Appl. Stat. **17**, 31–38 (1990a)

Tricker, A.R.: The effect of rounding on the power level of certain normal test statistics. J. Appl. Stat. **17**, 219–227 (1990b)

Tricker, A.R.: Estimation of parameters for rounded data from non-normal distributions. J. Appl. Stat. **19**, 465–471 (1992)

Vardeman, S.B.: Sheppard's correction for variances and the "Quantization Noise Model". IEEE Trans. Instrum. Meas. **54**, 2117–2119 (2005)

Vardeman, S.B., Lee, C.-S.: Likelihood-based statistical estimation from quantization data. IEEE Trans. Instrum. Meas. **54**, 409–414 (2005)

Wang, H., Heitjan, D.F.: Modeling heaping in self-reported cigarette counts. Stat. Med. **27**, 3789–3804 (2008)

Widrow, B., Kollar, I., Liu, M.-C.: Statistical theory of quantization. IEEE Trans. Instrum. Meas. **45**, 353–361 (1996)

Wilrich, P.Th.: Rounding of measurement values or derived values. Measurement **37**, 21–30 (2005)

Wimmer, G., Witowsky, V., Duby, T.: Proper rounding for the measurement results under the assumption of uniform distribution. Meas. Sci. Technol. **11**, 1659–1665 (2000)

Wold, H.: Sheppard's correction formulae in several variables. Skand. Aktuarietidskrift **17**, 248–255 (1934)

Wolff, J., Augustin, T.: Heaping and its consequences for duration analysis: a simulation study. Allgemeines Stat. Arch. **87**, 59–86 (2003)

Wright, D.E., Bray, I.: A mixture model for rounded data. Statistician **52**, 3–13 (2003)

Wu, J.: How severe was the Great Depression? Evidence from the Pittsburgh region. In: Komlos, J. (ed.) Stature, Living Standards, and Economic Development: Essays in Anthropometric History, pp. 129–152. University of Chicago Press, Chicago (1994)

Zhang, B.X., Liu, T.Q., Bai, Z.D.: Analysis of rounded data from dependent sequences. Ann. Inst. Stat. Math. (2010, to appear)