Schneeweiß, Komlos, Ahmad:

Symmetric and Asymmetric Rounding

Sonderforschungsbereich 386, Paper 479 (2006)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

Symmetric and Asymmetric Rounding

H. Schneeweiss
Department of Statistics,
University of Munich LMU,
Munich, Germany,

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

A. S. Ahmad
RG Biostatistics,
Max Planck Institute of Psychiatry,
Munich, Germany

Abstract

If rounded data are used in estimating moments and regression coffiecients, the estimates are typically more or less biased. The purpose of the paper is to study the bias inducing effect of rounding, which is also seen when population moments intstead of their estimates are considered. Under appropriate conditions this effect can be approximately specified by versions of Sheppard's correction formula. We discuss the conditions under which these approximations are valid. We also investigate the efficiency loss that comes along with 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.

We generalize the concept of simple rounding to that of asymmetric rounding and study its effect on the mean and variance of a distribution under similar circumstances as with simple rounding.

1 Introduction

Data often contains rounding errors. Variables (like 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) or simply our preference of some numbers over other numbers. The reason may be the avoidance of costs associated with a fine measurement or the imprecise nature of the measuring instrument. The German military, for example, measures the height of recruits to the nearest 1 cm. 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.

In the following we analyze statistical characteristics of rounded data X^* and of the rounding error δ. We consider expectations, variances and 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 the so-called Sheppard's correction (1898), and we give conditions under which it can be applied. Finally, we discuss the problem of asymmetric rounding.

This report is to a large extent a review of the literature, but new results are presented concerning the rounding error and asymmetric rounding. A comprehensive review of the field can be found in Heitjan (1989). Earlier reviews are Eisenhart (1947, Section 4), Stuart and Ord (1987, Sections 3.18-3.30), Gjeddeback (1968), Haitvosky (1982).

A more general case than simple rounding not treated in this paper is the case of heaping. With heaping only part of the data is rounded and the rounding points (or points of attraction) are not evenly spaced on the line.

Section 2 introduces the concept of simple rounding. In Section 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 Section 5. Section 6 investigates the validity of approximations introduced earlier. Section 7 studies some special distributions where these approximatons are either exact or completely invalid. Estimating and testing with rounded data is studied in Section 8. Section 9 deals with ML estimation. Asymmetric rounding is the subject of Section 10. Some concluding remarks are found in Section 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 problem is as follows. Let there be given

a set (a grid) of equidistant points on the real line,

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

where h is the distance between two adjacent points of the grid and ah, $0 \leq a \leq 1$ is the origin of the grid. For simplicity and without loss of generalty we will assume, unless otherwise stated, that $a = 0$. For any value of X, the rounded value X^* is that pointed of $\mathbb{R}^*$ which is nearest to X. (If X happens to ly exactly in the middle between two adjacent points of the grid, then X^* is the larger of the two points). X^* is a function of X: If round(x) is the function that maps any real x onto its nearest integer, then (assumming $a = 0$)

$$X^* = h \text{ round}(\frac{X}{h}).$$

The rounding error δ is defined as

$$\delta = X^* - X. \tag{1}$$

The equation

$$X^* = X + \delta \tag{2}$$

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, clearly δ is not independent of X. Instead, δ is a function of X. But δ is also not independent of X^*. Instead, the conditional density of δ given X^* is (see also Figures 3 and 4)

$$h(\delta|X^*) = \begin{cases} \frac{\varphi(X^*-\delta)}{p(X^*)} & \text{for} \quad -\frac{h}{2} \leq \delta \leq \frac{h}{2} \\ 0 & \text{for} \quad \delta < -\frac{h}{2} \quad \text{or} \quad \delta > \frac{h}{2} \end{cases} \tag{3}$$

where

$$p(x^*) = \int_{x^*-\frac{h}{2}}^{x^*+\frac{h}{2}} \varphi(x)dx \tag{4}$$

is the probability that $X^* = x^*$. With $x^* = ih$, this probability can also be written as

$$P(X^* = ih) = \int_{ih-\frac{h}{2}}^{ih+\frac{h}{2}} \varphi(x)dx = \int_{-\frac{h}{2}}^{\frac{h}{2}} \varphi(ih+u)du, \qquad u = x - ih. \tag{5}$$

These probabilities are illustrated in Figure 1 as the shaded areas under the density function $\varphi(x)$.

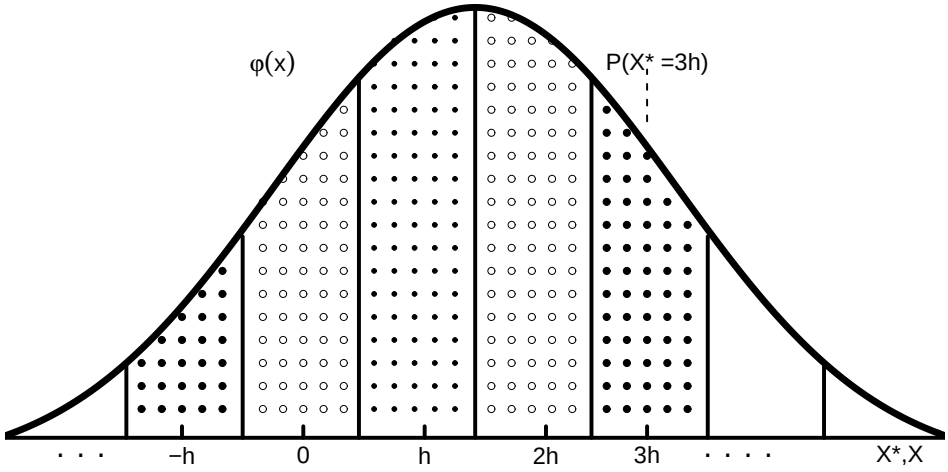

Figure 1: Density and probability function of the unrounded and rounded variables, X and X^*, respectively.

3 Moments of rounded values

3.1 Univariate moments

We next to relate the moments of X^* to those of X. The kth moment of the distribution of the rounded values X^* can be computed as:

$$\mathbb{E}X^{*k} = \sum_i (ih)^k P(X^* = ih) \tag{6}$$

$$= \sum_i (ih)^k \int_{-\frac{h}{2}}^{\frac{h}{2}} \varphi(ih+u)du. \tag{7}$$

We can approximate the sum in Equation (7) by an integral using the Euler-Maclaurin formula (see, e.g., Stuart and Ord, 1987, and Section 3.2 for more details). We rewrite each term of the sum in Formula (7) in the following way:

$$f(y) = y^k \int_{-\frac{h}{2}}^{\frac{h}{2}} \varphi(y+u)du \tag{8}$$

with $y = ih$. Then, according to the Euler-Maclaurin formula the sum in Equa-

tion (7) becomes:

$$\sum_i f(ih) = \frac{1}{h}\int_{-\infty}^{\infty} f(y)dy + R, \tag{9}$$

where R is a remainder term often quite small, which will be treated in more detail in Section 3.2. Ignoring the remainder term for the moment and coming back to Equation (7), we can write the kth moment of the rounded variable X^* as follows:

$$\mathbb{E}X^{*k} \approx \int_{-\infty}^{\infty} y^k \frac{1}{h}\int_{-\frac{h}{2}}^{\frac{h}{2}} \varphi(y+u)dudy. \tag{10}$$

Substituting $x = y + u$ and $v = u/h$ in the above formula we obtain:

$$\mathbb{E}X^{*k} \approx \int_{-\infty}^{\infty}\int_{-\frac{1}{2}}^{\frac{1}{2}} (x - vh)^k dv\ \varphi(x)dx. \tag{11}$$

Equation (11) can be used to compute approximately any k-th moment of the rounded data. E.g., for $k = 1$, which corresponds to the expected value of X^*, we obtain:

$$\begin{aligned}
\mathbb{E}X^* &\approx \int_{-\infty}^{\infty}\left[xv - \frac{v^2}{2}h\right]_{-\frac{1}{2}}^{\frac{1}{2}} \varphi(x)dx \\
&= \int_{-\infty}^{\infty} x\varphi(x)dx \\
&= \mathbb{E}X
\end{aligned} \tag{12}$$

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

$$\begin{aligned}
\mathbb{E}X^{*2} &\approx \int\left[x^2 v - xv^2 h + \frac{v^3}{3}h^2\right]_{-\frac{1}{2}}^{\frac{1}{2}} \varphi(x)dx \\
&= \int\left(x^2 + \frac{h^2}{12}\right)\varphi(x)dx \\
&= \mathbb{E}X^2 + \frac{h^2}{12}.
\end{aligned} \tag{13}$$

Because of (12) it follows that

$$\mathbb{V}X^* \approx \mathbb{V}X + \frac{h^2}{12}. \tag{14}$$

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

$$\mathbb{V}X \approx \mathbb{V}X^* - \frac{h^2}{12}. \tag{15}$$

This formula is known as Sheppard's correction, Sheppard (1898). Note that the term $\frac{h^2}{12}$ is just the variance of a variable uniformly distributed on the interval $[-\frac{h}{2}, \frac{h}{2}]$.

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

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

where $\mu_n = \mathbb{E}(X-\mu)^n$, $\mu_n^* = \mathbb{E}(X^*-\mu^*)^n$, $\mu = \mathbb{E}X$, $\mu^* = \mathbb{E}X^*$. For the correction of the 7th and 8th moments see Rietz (1924) p.94.

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

$$\begin{aligned}
\psi(t) &= \int e^{itx}\varphi(x)dx \\
\psi^*(t) &= \sum_j e^{itjh}p(jh)
\end{aligned}$$

be the characteristic functions of X and X^*, respectivley. Then

$$\psi^*(t) \approx \frac{2}{ht}\sin(\frac{ht}{2})\psi(t). \tag{16}$$

The r.h.s. of (16) is the characteristic function of $X+U$, where U is a random variable independent of X and uniformly distributed over $\left[-\frac{h}{2}, -\frac{h}{2}\right]$. Thus $X^* = X + \delta$ has approximately the same distribution as if δ were uniformly distributed and independent of X, see also Section 5. If follows that approximately

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

or

$$\kappa_m(X^*) \approx \kappa_m(X) + \kappa_m(U) = \kappa_m(X) + \frac{B_m}{m},$$

where κ_m is the m-th semi-invariant and B_m the m-th Bernoulli number. From this follow all the moment relations considered before.

An exact expression of the characteristic function of X^* in terms of trigonometric functions, for which the r.h.s. of (16) is the leading term, is given in Janson (2005), where also corresponding expressions for the moments of X^* are found.

3.2 The Remainder Term R

The sum in Equation (9) can be approximated by the integral on the r.h.s of (9) 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 grid. If the following two conditions are satisfied (see also Figure 2):

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

then the remainder term R equals

$$R = \frac{B_{2m+2}}{(2m+2)!}(b-a)h^{2m+1}f^{(2m+2)}(y_m), \tag{17}$$

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

The magnitude of the remainder term R (and thus the closeness of the approximation of the sum in (9) by the integral in (9)) depends on h and on the smoothness of f. The smaller h and the smaller $\max_{a \le y \le b} f^{(2m+2)}(y)$, the better the approximation.

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

Although the Euler-Maclaurin approximation has been stated only for integrals on a finite interval, one can expect that the approximation also holds good when the

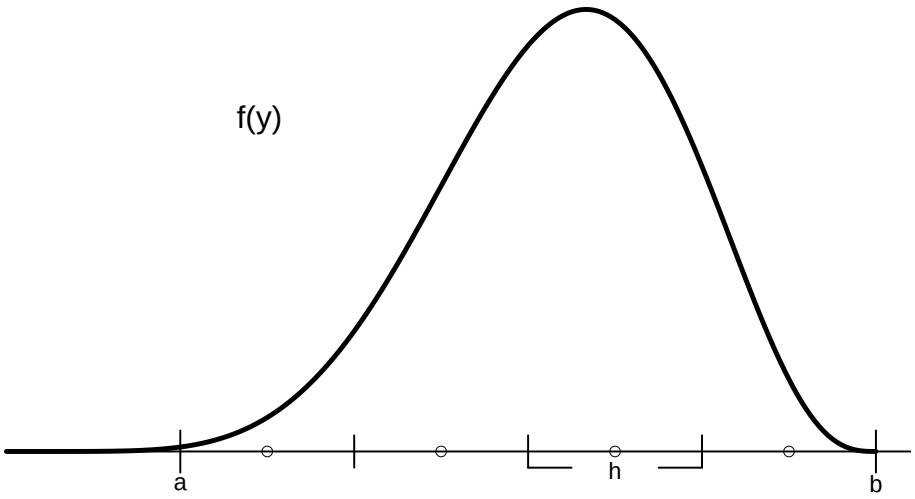

Figure 2: A function fulfilling the conditions for the Euler-Maclaurin formula

integral is taken over the whole real line. The conditions for that to hold is that the function f must be sufficiently smooth for the whole line and that the derivatives tend to zero when y tends to $+\infty$ or $-\infty$.

In applying this idea to the approximation (10) for moments of order k, one has to **make sure** that not only $\varphi(y)$ but $y^k\varphi(y)$ satisfies the Euler-Maclaurin conditions.

From (11) it is clear that the error of approximation (i.e., the difference of l.h.s. and r.h.s of (11)) changes by the factor λ^k if X and h are both multiplied by λ. In discussing the accuracy of the approximation (11) (and consequently of the mean equality (12) and of Sheppard's correction in (13)) it makes sense to restrict the discussion to the case of a standardized distribution with variance 1, see Section 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 grids with widths h and k, respectively, and origin (0,0) and let X^* and Y^* be the rounded variables. Their joint probability distribution is given by

$$p(ih, jk) = P(X^* = ih, Y^* = jk) = \int_{ih-\frac{h}{2}}^{ih+\frac{h}{2}} \int_{ik-\frac{k}{2}}^{ik+\frac{k}{2}} \varphi(x, y)dxdy.$$

Consider the bivariate moment of order (m, n) of (X, Y)

$$\mu_{mn} = \mathbb{E}(X^m Y^n) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^m y^n \varphi(x, y)dxdy$$

and the corresponding moment of (X^*, Y^*)

$$\begin{aligned} \mu^*_{mn} &= \mathbb{E}(X^{*m} Y^{*n}) = \sum_i \sum_j (ih)^m (jk)^n p(ih, jk) \\ &= \sum_i \sum_j (ih)^m (jk)^n \int_{-\frac{h}{2}}^{\frac{h}{2}} \int_{-\frac{k}{2}}^{\frac{k}{2}} \varphi(ih + u, jk + v)dudv. \end{aligned}$$

Aussuming the conditions for a good approximation by the Euler-Mclaurin formula to be satisfied the (double) sum can be approximated by a corresponding (double) integral:

$$\mu^*_{mn} \approx \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^m y^n \frac{1}{hk} \int_{-\frac{h}{2}}^{\frac{h}{2}} \int_{-\frac{k}{2}}^{\frac{k}{2}} \varphi(x + u, y + v)dudvdxdy$$

Substituting $t = x + u$, $s = y + v$, $\tilde{u} = \frac{u}{h}$, $\tilde{v} = \frac{v}{k}$, we get

$$\mu^*_{mn} \approx \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\frac{1}{2}}^{\frac{1}{2}} \int_{-\frac{1}{2}}^{\frac{1}{2}} (t - h\tilde{u})^m (s - k\tilde{v})^n d\tilde{u}d\tilde{v}\varphi(t, s)dtds.$$

Consider the special case $m = n = 1$. Then

$$\begin{aligned} \mu^*_{11} &= \mathbb{E}(X^*Y^*) \approx \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left[t\tilde{u} - h\frac{\tilde{u}^2}{2}\right]_{-\frac{1}{2}}^{\frac{1}{2}} \left[s\tilde{v} - k\frac{\tilde{v}^2}{2}\right]_{-\frac{1}{2}}^{\frac{1}{2}} \varphi(t, s)dtds \\ &= \int\int ts\varphi(t, s)dtds = \mathbb{E}(XY) = \mu_{11}. \end{aligned}$$

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

$$\mathbb{C}ov(X^*, Y^*) \approx \mathbb{C}ov(X, Y). \tag{18}$$

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

$$\mathbb{C}ov(X^*, Y) \approx \mathbb{C}ov(X, Y^*) \approx \mathbb{C}ov(X, Y). \tag{19}$$

4 The Influence of Rounding on Regression Estimates

Now we analyze the influence of rounding on the estimates of the regression coefficients. We always assume that the assumption for the application of the Euler-Maclaurin approximation are satisfied. Let Y be the unrounded response (or dependent) variable, and X^* (X) the rounded (unrounded) explanatory (or independent) variable and consider a simple linear regression model

$$Y = \alpha + \beta X + \varepsilon. \tag{20}$$

The corresponding regression for rounded data is

$$Y = \alpha^* + \beta^* X^* + \varepsilon^*. \tag{21}$$

The true regression coefficient is given by

$$\beta = \frac{\mathbb{C}ov(X, Y)}{\mathbb{V}X},$$

whereas the regression coefficient with the rounded variable X^* is

$$\beta^* = \frac{\mathbb{C}ov(Y, X^*)}{\mathbb{V}X^*}.$$

Then, using (19) and Sheppard's correction introduced in Equation (14) we get for β^*:

$$\beta^* \approx \frac{\mathbb{C}ov(X, Y)\mathbb{V}X}{\mathbb{V}X\mathbb{V}X^*} \approx \beta \, \frac{\mathbb{V}X^* - \frac{h^2}{12}}{\mathbb{V}X^*} = \beta \left(1 - \frac{1}{12}\left(\frac{h}{\sigma_{X^*}}\right)^2\right). \tag{22}$$

However, we are normally interested in β rather than β^*. The approximate bias correction for β due to rounding becomes:

$$\beta \approx \beta^* \left[1 - \frac{1}{12}\left(\frac{h}{\sigma_{X^*}}\right)^2\right]^{-1}. \tag{23}$$

This formula is very convenient since we normally know h and can estimate σ_{X^*} from the data. The naive estimate of β (i.e., $\hat{\beta}^* = \frac{s_{x^*y}}{s^2_{x^*}}$) is biased due to rounding. But it can be corrected according to (23), which leads to an approximately unbiased estimate of β:

$$\hat{\beta}^*_c := \hat{\beta}^* \left[1 - \frac{1}{12}\left(\frac{h}{s_{X^*}}\right)^2\right]^{-1} = \frac{s_{x^*y}}{s^2_{x^*} - \frac{h^2}{12}}. \tag{24}$$

Sometimes, the response variable Y is rounded and not the covariate X. In this case

$$\beta^* = \frac{\mathbb{C}ov(X, Y^*)}{\mathbb{V}X} \approx \frac{\mathbb{C}ov(X, Y)}{\mathbb{V}X} = \beta. \tag{25}$$

This means that rounding the response variable does not influence the regression estimates. The last possibility is that both the response and the explanatory variables are rounded. Then:

$$\beta^* = \frac{Cov(Y^*, X^*)}{\mathbb{V}X^*} \approx \frac{Cov(Y, X)}{\mathbb{V}X^*} = \beta\left(1 - \frac{1}{12}\left(\frac{h}{\sigma_{X^*}}\right)^2\right). \tag{26}$$

In a linear regression of Y on X, it is only rounding of X and not Y that has an effect on the value of the slope parameter.

5 Rounding Error

The rounding error δ was defined in (1) as the difference between the rounded and the true value. We have to keep in mind that δ is neither independent of X^* nor of X.

Globally, over the whole density of X^*, the expectation of δ equals zero approximately:

$$\mathbb{E}\delta \approx 0 \qquad \text{because} \qquad \mathbb{E}X^* \approx \mathbb{E}X.$$

However, locally (within an interval of length h; compare the shaded area in Figure 3 and the corresponding Figure 4) the expectation of δ may strongly differ from zero, e.g., there can be more positive values of δ than negative ones. The global expectation of zero results from the fact that the density $\varphi(x)$ fulfills the Euler-Maclaurin conditions.

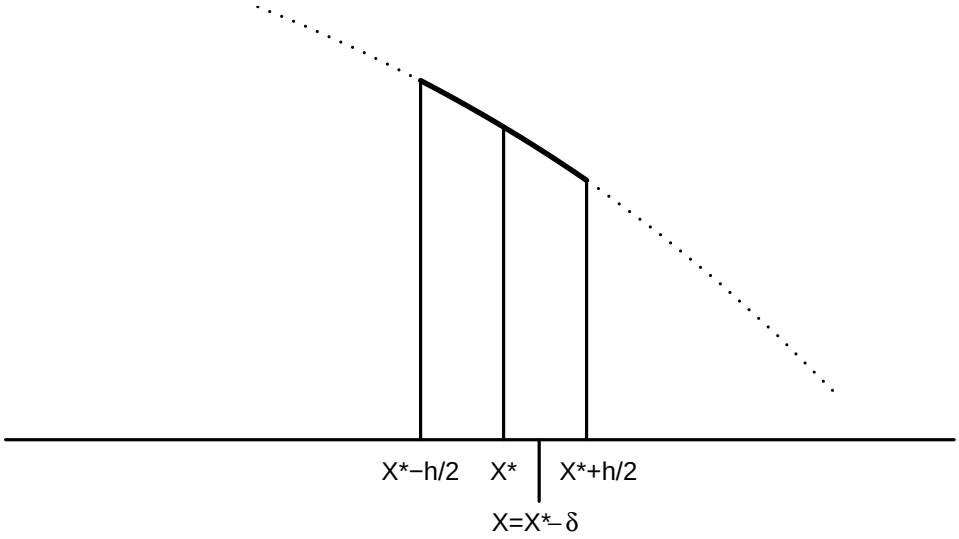

Figure 3: Distribution of X restricted to a rounding interval

Although X and δ are dependent, they are (approximately) uncorrelated:

$$\mathbb{C}ov(X, \delta) \approx 0$$

because, by (2),

$$\mathbb{C}ov(X, X^*) = \mathbb{V}X + \mathbb{C}ov(X, \delta),$$

but also, by (19) with $Y = X$,

$$\mathbb{C}ov(X, X^*) \approx \mathbb{V}X.$$

Moreover, the variance of δ is

$$\mathbb{V}\delta \approx \frac{h^2}{12}$$

because of (14) and because

$$\mathbb{V}X^* = \mathbb{V}X + 2\mathbb{C}ov(X, \delta) + \mathbb{V}\delta \approx \mathbb{V}X + \mathbb{V}\delta.$$

We can say even more: 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

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

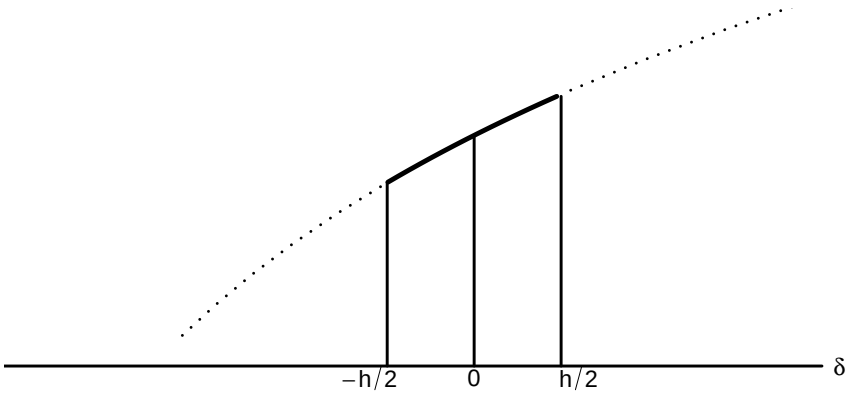

Figure 4: Conditional Distribution of δ given X^*

with $h(\delta|x^*)$ from (3), and concequently

$$g(\delta) = \sum_{x^*} \varphi(x^* - \delta) = \sum_i \varphi(ih - \delta). \tag{27}$$

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 $[-\frac{h}{2}, \frac{h}{2}]$.

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 want to compare the moments of X^* to those of X, depending on the width h of the rounding interval. We do this for the expected value and the variance of X. The difference of the moments of X and X^* depends not only on h but also on ah, the origin of the rounding grid, $0 < a < 1$. 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 distribution, see Tricker (1984).

From Equation (6) we can compute the exact expected value of the rounded data X^* when $X \sim N(0,1)$:

$$\mathbb{E}X^* = \sum_i (i+a)h[\Phi((i+a+\frac{1}{2})h) - \Phi((i+a-\frac{1}{2})h)], \tag{28}$$

where ah indicates by how much the rounding grid has been shifted away from the origin 0: $a = 0$ means that zero is a point of the rounding grid. With $a = 1$ the rounding grid is in the same position as with $a = 0$. Therefore $a \geq 1$ need not be considered.

Figure 5 shows the difference of the means of the unrounded and rounded data for various values of h as a function of the shift a. The rounding interval h varies from one to four standard deviations of the distribution. As we can see, rounding intervals of length up to 2.5 standard deviations have a rather small influence on the data mean. Furthermore, the bias disappears for $a = 0$, $a = 1$ as well as $a = 0.5$.

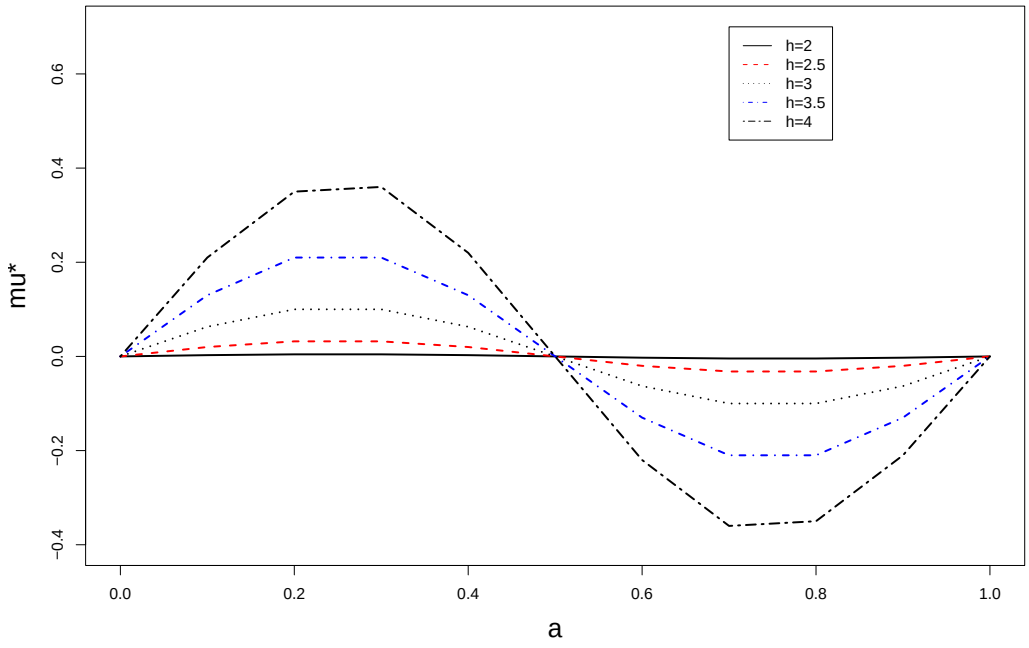

Figure 5: Differences of the means of the unrounded data, μ, and the rounded data, μ^*, as a function of a for $X \sim N(0,1)$.

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

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

Figure 6 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 highest 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 true variance particularly at $a = 0.5$ and the approximation is very poor.

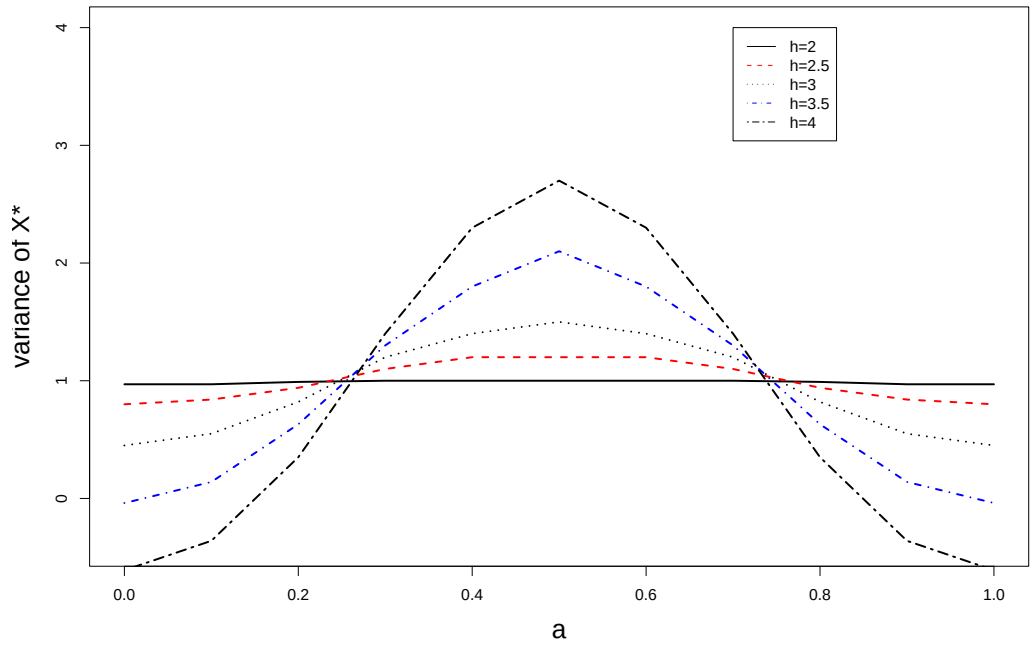

Figure 6: Sheppard-corrected variance of the rounded data, $\sigma^2_{X^*}$ as a function of a, for $X \sim N(0,1)$.

To sum up, we can say that the approximation for mean and variance work very well even for rather large values of h as long as the underlying distribution is the normal one. For other, in particular for skew distributions, Sheppard's correction only works well for considerably smaller values of h. Higher moments are also less well approximated by the corrsponding correction formulas.

7 Some special distribution

7.1 An "Exact Case" for Rounding

The approximation formulas (12) and (14) 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 width 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 grid consists of all midpoints of the subintervals, $x_i^* = (i + \frac{1}{2})h$, $i = 0, \cdots, n-1$ (compare Figure 7). Let us call such a density function a "piecewise linear grid density". In this situation, we can compute the exact values of the various moments using simple integration methods.

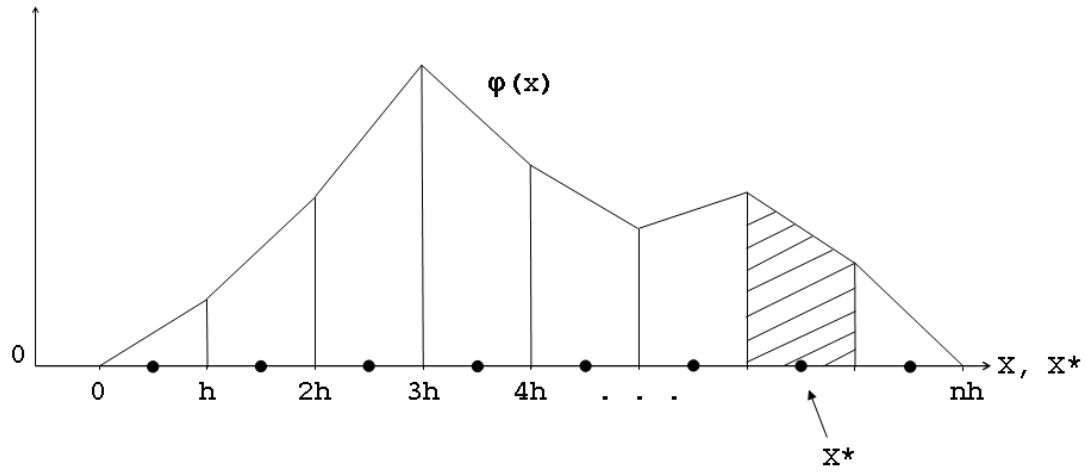

Figure 7: An "exact case" of rounding

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

$$\varphi(x) \;=\; \varphi(x^*) + \varphi'(x^*)(x - x^*),\; x^* - \frac{h}{2} \geq x \geq x^* + \frac{h}{2} \tag{29}$$

or

$$\varphi(x^* + u) \;=\; \varphi(x^*) + \varphi'(x^*)u, \qquad -\frac{h}{2} \geq u \geq \frac{h}{2} \tag{30}$$

The expected values of the unrounded and rounded data, respectively, are

$$\begin{aligned}
\mathbb{E}X &= \int_0^{nh} x\varphi(x)dx = \sum_{x^*} \int_{x^*-\frac{h}{2}}^{x^*+\frac{h}{2}} x\varphi(x)dx \\
&= \sum_{x^*} \int_{-\frac{h}{2}}^{\frac{h}{2}} (x^*+u)\varphi(x^*+u)du \qquad (31) \\
\mathbb{E}X^* &= \sum_{x^*} x^* \int_{x^*-\frac{h}{2}}^{x^*+\frac{h}{2}} \varphi(x)dx = \sum_{x^*} x^* \int_{-\frac{h}{2}}^{\frac{h}{2}} \varphi(x^*+u)du. \qquad (32)
\end{aligned}$$

To check whether the two expectations are equal, we look at the difference:

$$\begin{aligned}
\mathbb{E}X - \mathbb{E}X^* &= \sum_{x^*} \int_{-\frac{h}{2}}^{\frac{h}{2}} u\ \varphi(x^*+u)du \\
&= \sum_{x^*} \int_{-\frac{h}{2}}^{\frac{h}{2}} [u\ \varphi(x^*) + u^2\varphi'(x^*)]du \\
&= \sum_{x^*} \left[\frac{u^3}{3}\right]_{-\frac{h}{2}}^{\frac{h}{2}} \varphi'(x^*) = \frac{h^3}{12} \sum_{x^*} \varphi'(x^*) = 0. \qquad (33)
\end{aligned}$$

This is due to the fact that $\varphi(0) = \varphi(nh) = 0$, see Figure 7, and

$$\begin{aligned}
h\sum_{x^*} \varphi'(x^*) &= \varphi(h) - \varphi(0) + \varphi(2h) - \varphi(h) + \varphi(3h) - \varphi(2h) + \ldots \\
&+ \varphi(nh) - \varphi((n-1)h) = 0. \qquad (34)
\end{aligned}$$

Thus for the piecewise linear grid density

$$\mathbb{E}X^* = \mathbb{E}X. \qquad (35)$$

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

$$\begin{aligned}
\mathbb{E}X^2 &= \sum_{x^*} \int_{-\frac{h}{2}}^{\frac{h}{2}} (x^*+u)^2\varphi(x^*+u)du \qquad (36) \\
\mathbb{E}X^{*2} &= \sum_{x^*} x^{*2} \int_{-\frac{h}{2}}^{\frac{h}{2}} \varphi(x^*+u)du \qquad (37)
\end{aligned}$$

The difference is

$$\begin{aligned}
\mathbb{E}X^2 - \mathbb{E}X^{*2} &= \sum_{x^*} \int_{-\frac{h}{2}}^{\frac{h}{2}} (u^2 + 2ux^*)[\varphi(x^*) + \varphi'(x^*)u]du \\
&= \sum_{x^*} \int_{-\frac{h}{2}}^{\frac{h}{2}} u^2[\varphi(x^*) + 2x^*\varphi'(x^*)]du \\
&= \sum_{x^*} \frac{h^3}{12}[\varphi(x^*) + 2x^*\varphi'(x^*)] \qquad (38) \\
&= \frac{h^2}{12}\left[h\sum_{x^*}\varphi(x^*) + 2h\sum_{x^*} x^*\varphi'(x^*)\right]. \qquad (39)
\end{aligned}$$

To calculate the sums in Formula (39) we first recall that the area under the density function equals one, i.e.,

$$\sum_{x^*} h\varphi(x^*) = 1. \qquad (40)$$

Moreover,

$$\begin{aligned}
\sum_{x^*} x^*\varphi'(x^*) &= \frac{h}{2}\frac{\varphi(h) - \varphi(0)}{h} + \frac{3h}{2}\frac{\varphi(2h) - \varphi(h)}{h} + \frac{5h}{2}\frac{\varphi(3h) - \varphi(2h)}{h} + \dots \\
&= \frac{1}{2}[\varphi(h) + 3\varphi(2h) - 3\varphi(h) + 5\varphi(3h) - 5\varphi(2h) + \dots] \\
&= -[\varphi(h) + \varphi(2h) + \varphi(3h) + \dots] \\
&= -\sum_{x^*}\varphi(x^* + \frac{h}{2}) = -\sum_{x^*}\varphi(x^*) - \frac{h}{2}\sum_{x^*}\varphi'(x^*) \\
&= -\sum_{x^*}\varphi(x^*)
\end{aligned}$$

by (34). Substituting this result in Equation (39), the difference $EX^2 - EX^{*2}$ becomes:

$$\mathbb{E}X^2 - \mathbb{E}X^{*2} = \frac{h^2}{12}\left(\sum_{x^*} h\varphi(x^*) - 2\sum_{x^*} h\varphi(x^*)\right) = \frac{h^2}{12}(-1) = -\frac{h^2}{12}$$

and

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

Thus in the case of a piecewise linear grid density, Sheppard's correction holds exactly. The same is true for all the higher moments.

One can also show that in this case the marginal distribution of δ is not only approximately but exactly uniformly distributed on $[-\frac{h}{2}, \frac{h}{2}]$. Indeed, by (27),

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

by (34) and (40).

These results can be generalized to the case of a piecewise linear grid density, which is defined on the whole real line, as long as the slopes in the rounding intervals tend to zero with $x \to \pm\infty$ at a sufficiently large rate.

The piecewise linear grid density may be a rather artificial density function. But as far as other, more realistic, densities that tend to zero sufficiently fast as $x \to \pm\infty$ can be approximated by a piecewise linear grid density, the latter serves as a convenient model to explain the approximate relations between the moments of rounded and unrounded data.

7.2 Uniform distribution

In this section, we investigate a special case of the distribution of X, where Sheppard's correction does not work. We assume that X is uniformly distributed over a specified range. Note that this distribution does not satisfy the requirements for the Sheppard approximation. Therefore, different results must be expected.

With respect to the width of the rounding intervals at the end points of the distribution function, we can distinguish two cases: All intervals are of the same width or narrower intervals are placed at the end points of the distribution. Furthermore, within these two cases we can differentiate between two possible placing schemes of the rounding grid: a) the value zero is the midpoint of one rounding interval (i.e., a point of the grid) and b) zero lies on the border between two intervals (compare Figure 8, pictures a and b).

Case 1
In the first case all rounding intervals are of the same width, compare Figure 8. A simple example for rounding where the density function looks like the one in Figure 8a is rounding to integers with a rounding interval of width $h = 1$. The density shown in Figure 8b refers, e.g., to rounding to odd numbers with a rounding interval of width $h = 2$.

From Formula (2) we recall that:

$$X^* = X + \delta.$$

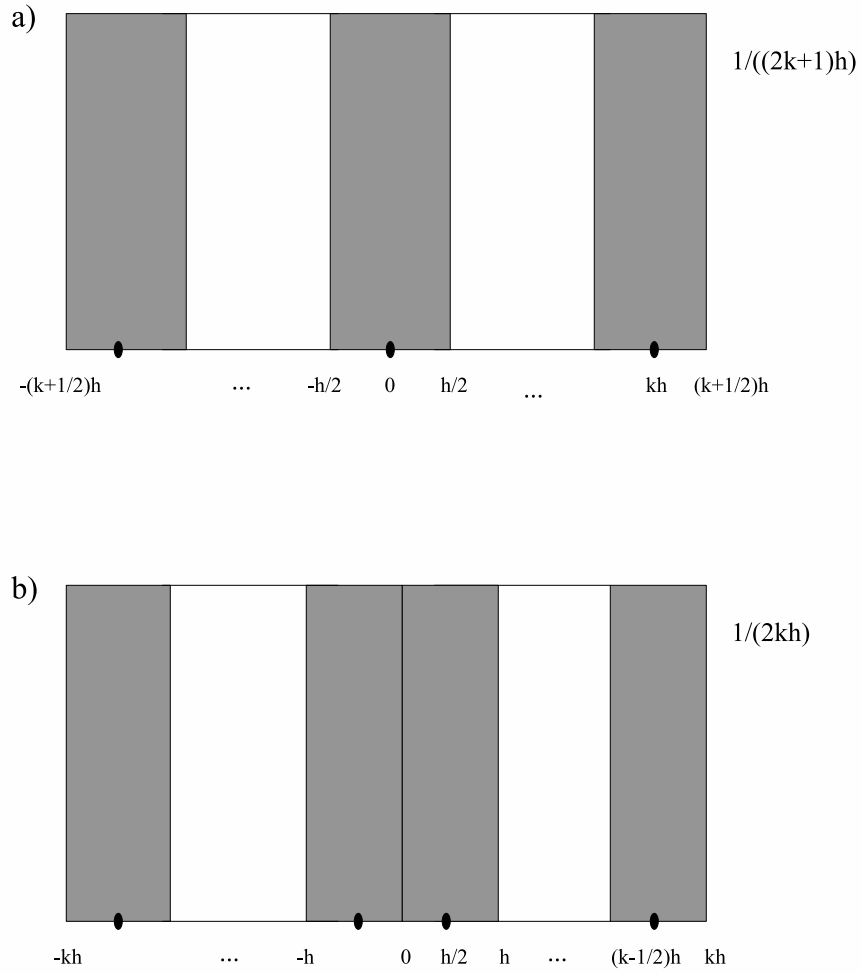

Figure 8: Uniform distribution, Case 1

Now, in this case, δ is not only globally (i.e., marginally) but also locally uniformly distributed. Indeed, by (3),

$$h(\delta|x^*) = \frac{1}{h}, \qquad -\frac{h}{2} \leq \delta \leq \frac{h}{2},$$

and therefore δ is independent of X^*. It also follows that

$$\mathbb{E}\delta = 0, \qquad \mathbb{V}\delta = \frac{h^2}{12}.$$

Then, for both of the two possibilities shown in Figures 8 a and b we obtain:

$$\begin{aligned} \mathbb{C}ov(X^*, \delta) &= 0 \\ \mathbb{V}X &= \mathbb{V}X^* + \mathbb{V}\delta \\ &= \mathbb{V}X^* + \frac{h^2}{12}. \end{aligned} \tag{41}$$

Thus Sheppard's correction is reversed in this special case. Instead of subtracting $\frac{h^2}{12}$, as in (15), this term has to be added in order to correct $\mathbb{V}X^*$ (see also Vardeman 2003).

We can also compute the covariance of X and δ by starting from

$$\begin{aligned} \mathbb{V}X^* &= \mathbb{V}X + \mathbb{V}\delta + 2\mathbb{C}ov(X, \delta) \\ &= \left(\mathbb{V}X^* + \frac{h^2}{12}\right) + \frac{h^2}{12} + 2\mathbb{C}ov(X, \delta) \end{aligned}$$

and thus

$$\mathbb{C}ov(X, \delta) = -\frac{h^2}{12},$$

in contrast to the general relation $\mathbb{C}ov(X, \delta) \approx 0$, see Section 5.

Now, consider a linear regression of Y on X:

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

with ε independent of X. Then ε is independent of X^*, too. For the slope parameter of the corresponding regression of Y and X^* we have

$$\beta^* = \frac{\mathbb{C}ov(X^*, Y)}{\mathbb{V}X^*} = \beta\frac{\mathbb{C}ov(X^*, X)}{\mathbb{V}X^*} + \frac{\mathbb{C}ov(X^*, \varepsilon)}{\mathbb{V}X^*} = \beta\frac{\mathbb{C}ov(X^*, X)}{\mathbb{V}X^*} = \beta,$$

where we used the identity

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

Thus in Case 1 there is no bias in the regression parameter resulting from rounding.

Now we move on to the situation where the end intervals are only half as wide as the internal intervals.

Case 2
This situation is plausible if there are natural limits of the variable values, e.g. $(-50, 50)$. Then, the last interval, say $(45, 50)$, is narrower than the last but one interval $(35, 45)$

since no unrounded values greater than 50 are allowed. Figure 9 shows the density functions of X^* for two interval placing schemes. In Case 2 we have again

$$\mathbb{E}\delta = 0, \qquad \mathbb{V}\delta = \frac{h^2}{12}.$$

However, since the rounding intervals are narrower at the end points, the conditional density of δ given X^* depends on X^* and thus δ and X^* are not independent. Indeed, in Case 2a (and similarly in Case 2b)

$$\mathbb{E}(\delta|X^* = kh) = \frac{h}{4}, \qquad \mathbb{E}(\delta|X^* = -kh) = -\frac{h}{4}.$$

and $\mathbb{E}(\delta|X^*) = 0$ for all other values of X^*. Consequently, we obtain for the previous covariances and the variance of X values different from those in Case 1. In particular for Case 2a we have

$$\begin{aligned}
\mathbb{C}ov(X^*, \delta) &= \mathbb{E}[X^*\mathbb{E}(\delta|X^*)] = \left[\frac{h}{4}kh - \frac{h}{4}(-kh)\right]\frac{h}{2}\cdot\frac{1}{2kh} = \frac{h^2}{8} \\
\mathbb{V}X &= \mathbb{V}X^* - 2\,\mathbb{C}ov(X^*, \delta) + \mathbb{V}\delta = \mathbb{V}X^* - \frac{h^2}{4} + \frac{h^2}{12} = \mathbb{V}X^* - \frac{h^2}{6} \\
\mathbb{C}ov(X^*, X) &= \mathbb{V}X^* - \mathbb{C}ov(X^*, \delta) = \mathbb{V}X^* - \frac{h^2}{8}
\end{aligned}$$

For a linear regression of Y on X with slope parameter β we get

$$\beta^* = \frac{\mathbb{C}ov(X^*, Y)}{\mathbb{V}X^*} = \beta\frac{\mathbb{C}ov(X^*, X)}{\mathbb{V}X^*} = \beta\frac{VX^* - \frac{h^2}{8}}{VX^*} = \beta\left(1 - \frac{1}{8}\left(\frac{h}{\sigma_{x^*}}\right)^2\right).$$

The last equation shows that the regression parameter estimated from the rounded data is asymptotically smaller than β, just as in the case where Sheppard's correction applies, but with quite a different correction (compare to (26)). Analogous formulas can be derived for the Case 2b, where the value zero is a border of two adjacent intervals instead of being the midpoint of an interval.

These examples show that completely different results with respect to Sheppard's correction can be obtaind when the assumptions for the Euler-Maclaurin formula are not satisfied. See also Example 2.9 in Janson (2005).

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 in a smooth way. 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).

A generalization of the uniform distribution is the so-called histogram density function, where the density is constant on each rounding interval, see Figure 10. In this

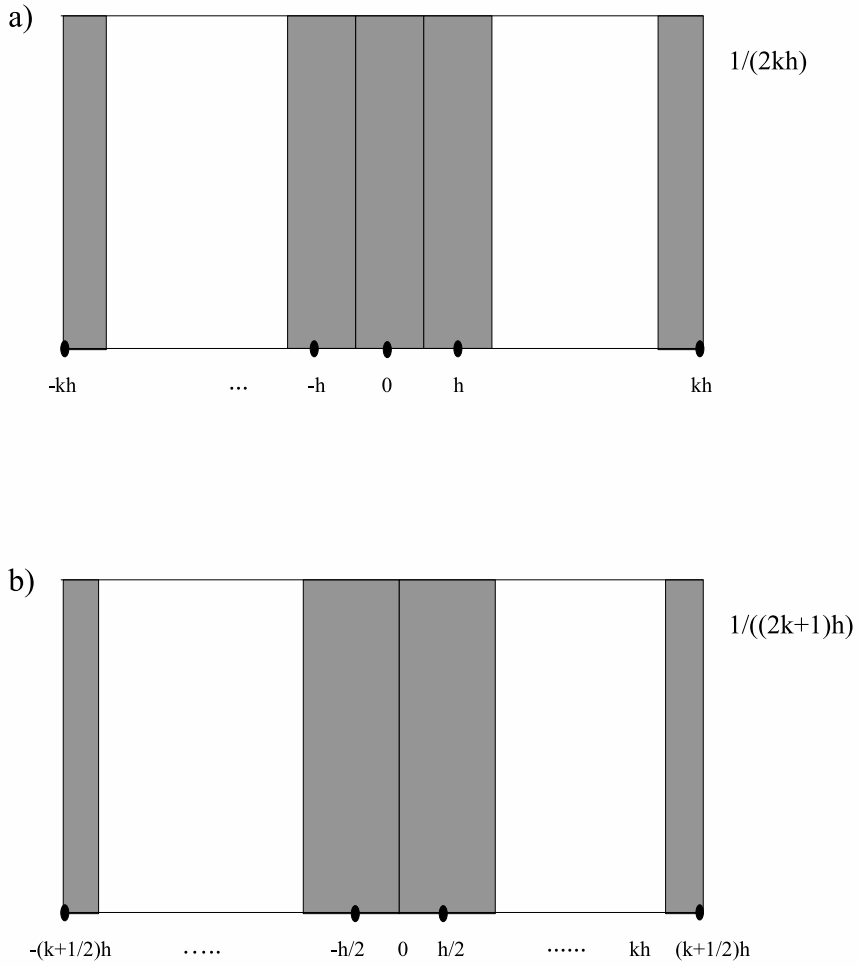

Figure 9: Uniform distribution, Case 2

case (just like in Case 1 of the uniform distribution)

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

and Sheppard's correction would give a completely wrong result. The histogram density is the counterpart to the piecewise linear grid density.

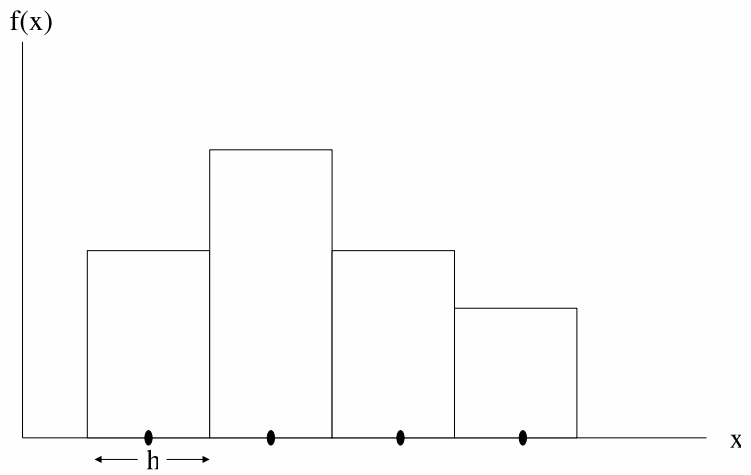

Figure 10: Histogram density

8 Estimation and Testing

8.1 Estimation

Up to now we only dealt with various population parameters (moments and regression coefficients) of rounded und unrounded random variables and their relations to each other. We did not consider estimation and testing problems.

Let us now consider 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^*, $i = 1, \cdots, n$, in order to estimate μ. Let us assume that the conditions for Sheppard's correction are satisfied.

Now if the original, unrounded, data x_i, $i = 1, \cdots, n$, is an iid sample, so is the rounded data x_i^*, $i = 1, \cdots, n$. The arithmetic mean $\bar{x}^*$ of the x_i^* is therefore an unbiased as well as a strongly consistent estimate of $\mu^* = \mathbb{E}X^*$. If the Euler-Mclaurin 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:

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

In a similar way we can use the rounded data to estimate the variance of X. However, here we must observe Sheppard's correction. Thus

$$\mathbb{E}s_{x^*}^2 = \operatorname{plim} s_{x^*}^2 = \sigma_{x^*}^2, \;\; \sigma_x^2 \approx \sigma_{x^*}^2 - \frac{h^2}{12}.$$

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

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, although unbiased (in the finite sample or asymptotic sense), 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 $\frac{1}{n}\sigma^2_{x^*}$, while the variance of $\hat{\mu} = \bar{x}$ is $\frac{1}{n}\sigma^2_x$ and $\sigma^2_{x^*} \approx \sigma^2_x + \frac{h^2}{12} > \sigma^2_x$. 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,

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

is always larger than the corresponding confidence interval from the unrounded data,

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

The efficiency loss is measured by the ratio, see (15) (see also Gjeddeback (1956))

$$\frac{\sigma^2_X}{\sigma^2_{X^*}} \approx \left(1 - \frac{h^2}{12\sigma^2_X}\right)^{-1} \approx 1 + \frac{h^2}{12\sigma^2_X}$$

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 $N(\mu, \sigma^2)$ distribution. In the two-sided case, the null hypothesis to be tested is

$$H_0 : \mu = \mu_0. \tag{42}$$

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

$$\tau = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}_{\hat{\mu}}}, \tag{43}$$

where $\hat{\mu} = \bar{x}$, $\hat{\sigma}_{\hat{\mu}} = \frac{\hat{\sigma}_X}{\sqrt{n}}$ and $\hat{\sigma}^2_X = \frac{1}{n-1}\sum_i^n (x_i - \bar{x})^2$ is the sample variance of the data.

Under H_0, the test statistic τ has the Student t-distribution, $\tau \sim t$. Now, for rounded data, the null hypothesis can be stated as

$$H_0 : \mu^* = \mu_0 \tag{44}$$

because $\mu^* \approx \mu$ (assuming that the Euler-Maclaurin conditions are satisfied).

The corresponding test statistic is:

$$\tau^* = \frac{\hat{\mu}^* - \mu_0}{\hat{\sigma}_{\hat{\mu}^*}}, \tag{45}$$

where $\hat{\mu}^* = \overline{x^*}$, $\hat{\sigma}_{\hat{\mu}^*} = \frac{\hat{\sigma}_{X^*}}{\sqrt{n}}$ and $\hat{\sigma}^2_{X^*} = \frac{1}{n-1}\sum_i^n (x_i^* - \overline{x^*})^2$. 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 of H_0. (Econometricians use to call such a test still a t-test, although it is actually an asymptotic Gauss-test). The test is constructed such that H_0 is rejected whenever $|\tau^*| > t_{1-\frac{\alpha}{2}}$, where $t_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$ 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. In order to compare the power functions of both t-tests, we consider a one sided test, which tests the hypothesis

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

against the alternative

$$H_1 : \mu > \mu_0.$$

We use the same test statistics τ and τ^* for this test depending on whether we base the test on unrounded or on rounded data.

The corrsponding power functions of these tests are given by:

$$\begin{aligned} \pi(\mu) &= P(\tau > t_{1-\alpha}|\mu) \\ \pi^*(\mu) &= P(\tau^* > t_{1-\alpha}|\mu). \end{aligned}$$

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

$$\begin{aligned} \pi(\mu) &= P(\frac{\hat{\mu} - \mu_0}{\hat{\sigma}_{\hat{\mu}}} > t|\mu) \\ &= P(\frac{\hat{\mu} - \mu}{\hat{\sigma}_{\hat{\mu}}} + \frac{\mu - \mu_0}{\hat{\sigma}_{\hat{\mu}}} > t|\mu). \end{aligned}$$

For sufficiently large n, $\hat{\sigma}_{\hat{\mu}}$ can be replaced with $\sigma_{\hat{\mu}} = \sigma_x/\sqrt{n}$ and

$$\frac{\hat{\mu} - \mu}{\hat{\sigma}_{\hat{\mu}}} \sim N(0,1).$$

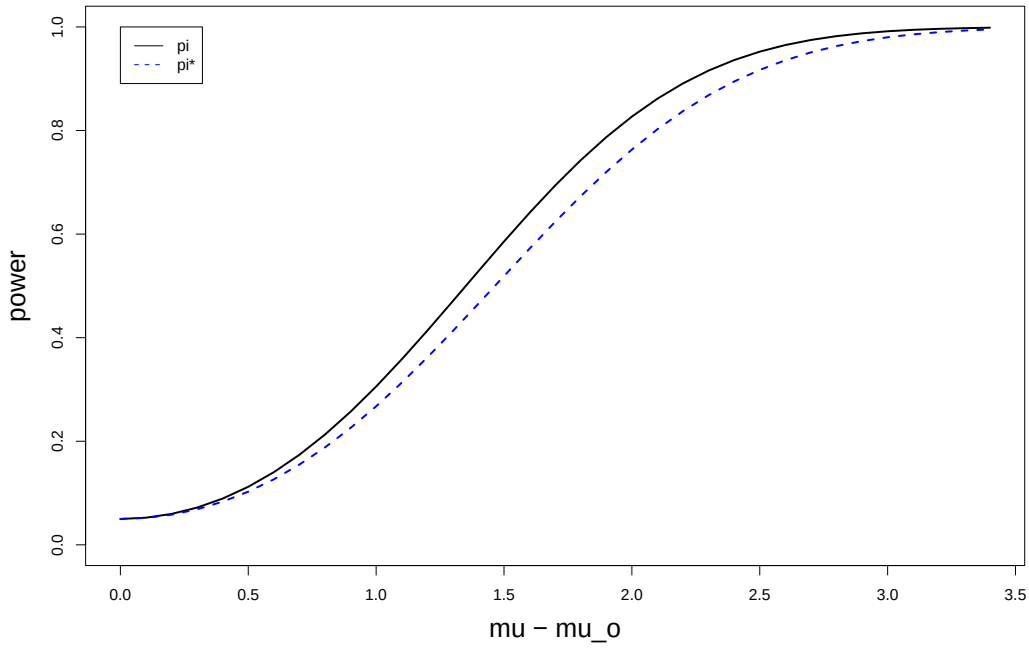

Figure 11: Two power functions

Thus for large n,

$$\pi(\mu) = 1 - \Phi(t - \frac{\mu - \mu_0}{\sigma_x}\sqrt{n}). \tag{46}$$

similarly

$$\pi^*(\mu) = 1 - \Phi(t - \frac{\mu - \mu_0}{\sigma_{x^*}}\sqrt{n}). \tag{47}$$

But since, by Sheppard's correction,

$$\sigma_{x^*} \approx \sigma_x\sqrt{1 + \frac{1}{12}(\frac{h}{\sigma_x})^2} > \sigma_x,$$

obviously

$$\pi^*(\mu) < \pi(\mu) \quad \text{under} \quad H_1 : \mu > \mu_0. \tag{48}$$

So the test with rounded data has smaller power than the test with unrounded data and is thus less efficient. Figure 11 shows the power functions π and π^* as functions of $\mu - \mu_0$. π and π^* have been computed according to (46) and (47) with $n = 100$, α =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 *et al* (1947).

9 ML Estimation of μ and σ when h is large

The approximation of the moments of the rounded data for large rounding intervals h is rather poor, since the remainder term R is proportional to h^{2m-1} (see (17)). For this reason, it is sometimes better to estimate the parameters of the data using the Maximum Likelihood method. By way of example we present the estimation of μ and σ^2 of a normal distribution.

Suppose the variable X is normally distributed $N(\mu, \sigma^2)$. The discrete distribution of X^* is given by:

$$p(x^*|\mu,\sigma^2) = \int_{x^*-\frac{h}{2}}^{x^*-\frac{h}{2}} \varphi_{\mu,\sigma}(x)dx = \Phi_{\mu,\sigma}\left(x^* + \frac{h}{2}\right) - \Phi_{\mu,\sigma}\left(x^* - \frac{h}{2}\right), \tag{49}$$

where

$$\varphi_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

is the density function of the normal distribution and $\Phi_{\mu,\sigma}$ the corresponding distribution function. Given a sample of data x_i^*, $i = 1, \cdots, n$, the likelihood function becomes:

$$L = L(\mu,\sigma^2) = \prod_i p(x_i^*|\mu,\sigma^2),$$

and the log-likelihood, $l = \log(L)$, is given by:

$$l = l(\mu,\sigma^2) = \sum_i \log p(x_i^*|\mu,\sigma^2) \to \max_{\mu,\sigma^2}. \tag{50}$$

Solving this optimization problem yields the ML estimates $\hat{\mu}_{\text{ML}}$ and $\hat{\sigma}^2_{\text{ML}}$. One can find the ML estimates by simply computing $l(\mu_i, \sigma_j^2)$ on a sufficiently fine grid of parameter points (μ_i, σ_j^2) and picking the point with the largest value of l.

Figur 12 depicts $l(\mu, \sigma^2)$ for a simulated set of height in inch as of 272 men with mean 67.5 and variance 7.29. The idea of this simulation has been taken from a historical example of slave men data in the USA, where the men have classified into three groups, namely small medium and tall, Komlos (1994).

We rounded the simulated data with $h = 6$ to the points 60, 66, 72 so that we also have three groups: 60 for small, 66 for medium, and 72 for tall. The sample consists of 14, 183, and 75 individuals for small, medium and tall, respectively. Although h is rather large ($h = 2.2\sigma$), we first try to estimate μ and σ^2 with the rounded data.

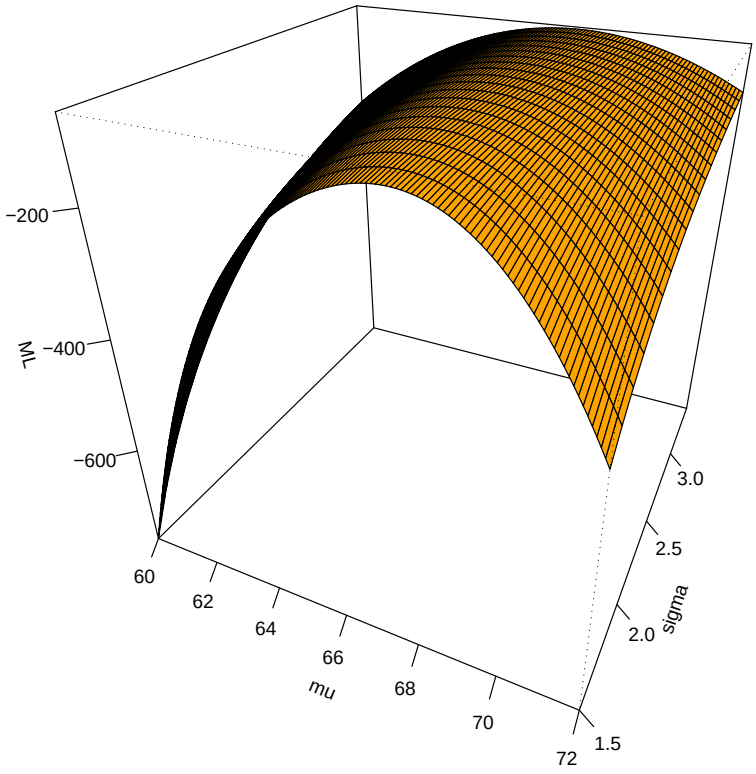

Figure 12: ML for μ and σ

We get $\bar{x^*} = 67.4$ and $s^2_{x^*} = 10.3$. correcting $s^2_{x^*}$ by substracting $\frac{h^2}{12}$ we receive the estimate $\hat{\sigma}^2 = 7.3$, which comes very close to the true value of σ^2. Nevertheless, we also apply ML to estimate μ and σ^2. Figure 12 shows the likelihood function. Its maximum is found by a grid search and we find $\hat{\mu}_{\mathrm{ML}} = 67.4$ and $\hat{\sigma}^2_{\mathrm{ML}} = 7.29$.

A more systematic way of finding the maximum of l is to use Newton's method (or some other iterative algorithm): Let $\theta = (\mu, \sigma^2)^\top$ and start with the simple estimate

$\theta_0 = (\bar{x}^*, s^2_{x^*})^\top$ as an initial parameter value. An improved estimate is given by

$$\theta_1 = \theta_0 - \left(\frac{\partial^2 l}{\partial \theta_0 \partial \theta_0^\top} \right)^{-1} \frac{\partial l}{\partial \theta_0}. \tag{51}$$

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 working of this method for the normal distribution can be studied in Gjeddeback (1949) and for other distributions, in particular the exponential distribution, in Kulldorff (1961) and Tallis and Young (1962), who both also consider unequal rounding intervals.

The derivatives in (51) are not always easy to compute, in particular when other distributions besides the normal are considered. But for small h, following Lindley (1949), approximations to these derivatives can be computed easily using a Taylor expansion of $\varphi(x) := \varphi_{\mu,\sigma}(x)$ at x^*:

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

Then, omitting terms of higher order in h, (49) yields

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

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

$$\begin{aligned} \log p(x^*) &\approx \log h + \log \varphi(x^*) + \log\left(1 + \frac{h^2}{24}\frac{\varphi''(x^*)}{\varphi(x^*)}\right) \\ &\approx \log h + \log \varphi(x^*) + \frac{h^2}{24}\frac{\varphi''(x^*)}{\varphi(x^*)}, \end{aligned} \tag{52}$$

and this expression can be substiuted in (50). Now by taking derivatives in (52), we obtain for the derivatives in (51)

$$\frac{\partial l}{\partial \theta_0} \approx \frac{h^2}{24} \sum_i \frac{\partial}{\partial \theta_0} \frac{\varphi''(x_i^*)}{\varphi(x_i^*)} \tag{53}$$

$$\frac{\partial^2 l}{\partial \theta_0 \partial \theta_0^\top} \approx \sum_i \frac{\partial^2}{\partial \theta_0 \partial \theta_0^\top} \log \varphi(x_i^*). \tag{54}$$

In (53) we made use of the fact that

$$\frac{\partial}{\partial \theta_0} \sum \log \varphi(x_i^*) = 0. \tag{55}$$

Indeed, the simple estimate θ_0 is found by solving the likelihood score equation (55) of the orignal model with the rounded data x_i^* in place of the orignal data x_i. In (54) terms of order h^2 were omitted.

Substituting (53) and (54) in (51) yields a first step approximation to the ML estimator of θ:

$$\theta_1 \approx \theta_0 - \frac{h^2}{24}\left(\sum_i \frac{\partial^2}{\partial\theta_0\partial\theta_0^\top}\log\varphi(x_i^*)\right)^{-1}\left(\sum_i \frac{\partial}{\partial\theta_0}\frac{\varphi''(x_i^*)}{\varphi(x_i^*)}\right). \tag{56}$$

The difference $\theta_1-\theta_0$ in (56) can be regarded as an analogue to Sheppard's correction stemming from ML estimation theory rather than from moment considerations. In the case of estimating $\theta = (\mu,\sigma^2)^\top$ from a normal distribution, we find with some algebra from (53) and (54)

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

(56) then yields

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

and this is just Sheppard's correction for μ and σ^2. For extensions to other distribution see also Tallis (1967) and for a generalization to the multivariate case see Don (1981). Fryer and Pethybridge (1972) extend Sheppard's correction to higher orders of h and do the same for the estimates of a linear regression.

In a similar manner one can derive Sheppard's correction from the first step of an EM algorithem so solve the ML equations for μ and σ^2 of a normal distribution, cf. Dempster and Rubin (1983).

10 Asymmetric Rounding

Sometimes the data have been rounded asymmetrically, which symmetricmeans that the rounding intervals about x^* and x^*+h are not equal. A simple example of asymmetric rounding is a situation where even numbers are preferred and values (0.75 to 1.25) are rounded to 1 while values (1.25 to 2.75) are rounded to 2, see Figure 13. 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 (Section 3). Let us assume that all values of X in the interval $[2ih - rh, 2ih + rh]$, $0 \le r \le 1$, 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 $x^* + h = (2i+1)h$, $i = 0, \pm 1, \pm 2, \cdots$, see Figure 13(Here we assume without loss of generality that $a = 0$, i.e., the origin 0 is a point of the grid).

First note that

$$\begin{aligned} p(2ih) &:= \mathbb{P}(X^* = 2ih) = \int_{-rh}^{rh} \varphi(2ih + u)du \\ p((2i+1)h) &:= \mathbb{P}(X^* = (2i+1)h) = \int_{-(1-r)h}^{(1-r)h} \varphi(2ih + u)du. \end{aligned}$$

Then the k-th moments of X^* is

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

Using the Euler-Maclaurin approximation, we obtain

$$\begin{aligned} \mathbb{E}X^{*k} &\approx \int_{-\infty}^{\infty} \frac{y^k}{2h} \int_{-rh}^{rh} \varphi(y + u)dudy + \int_{-\infty}^{\infty} \frac{y^k}{2h} \int_{-(1-r)h}^{(1-r)h} \varphi(y + u)dudy \\ &= \frac{1}{2} \int_{-\infty}^{\infty} \left[\int_{-r}^{r} (x - vh)^k dv + \int_{-(1-r)}^{1-r} (x - vh)^k dv \right] \varphi(x)dx. \end{aligned}$$

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

$$\begin{aligned} \mathbb{E}X^* &\approx \frac{1}{2} \int_{-\infty}^{\infty} \left(\left[xv - \frac{v^2}{2}h \right]_{-r}^{r} + \left[xv - \frac{v^2}{2}h \right]_{-(1-r)}^{1-r} \right) \varphi(x)dx \\ &= r \int_{-\infty}^{\infty} x\varphi(x)dx + (1-r) \int_{-\infty}^{\infty} x\varphi(x)dx = \mathbb{E}X. \end{aligned} \tag{57}$$

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

Similarly, for $k = 2$:

$$\begin{aligned}\mathbb{E}X^{*2} &\approx \frac{1}{2}\int_{-\infty}^{\infty}\left(\left[x^2v - 2x\frac{v^2}{2}h + \frac{v^3}{3}h^2\right]_{-r}^{r} + \left[x^2v - 2x\frac{v^2}{2}h + \frac{v^3}{3}h^2\right]_{-(1-r)}^{1-r}\right)\varphi(x)dx \\ &= r\mathbb{E}X^2 + \frac{r^3}{3}h^2 + (1-r)\mathbb{E}X^2 + \frac{(1-r)^3}{3}h^2 \\ &= \mathbb{E}X^2 + \frac{1}{3}(1 - 3r + 3r^2)h^2. \qquad (58)\end{aligned}$$

Together with (57) this implies

$$\mathbb{V}X^* \approx \mathbb{V}X + \frac{h^2}{3}(1 - 3r + 3r^2) =: \mathbb{V}X + f(r)h^2. \qquad (59)$$

The shape of the function $f(r) := r^2 - r + \frac{1}{3}$ is shown in Figure 14 for $0 \le r \le 1$. $f(r)$ 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^* or $X^* + h$, respectively, has a rounding interval of width zero.

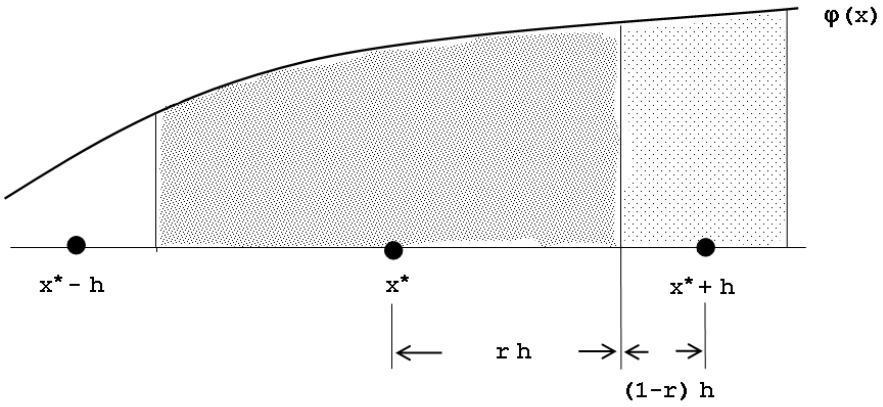

Figure 13: Asymmetric rounding

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

$$\begin{aligned}\mathbb{E}(X^*Y) &= \sum_i 2ih\int_{-\infty}^{\infty} y\int_{-rh}^{rh}\varphi(2ih + u, y)dudy \\ &+ \sum_i (2i+1)h\int_{-\infty}^{\infty} y\int_{-(1-r)h}^{(1-r)h}\varphi((2i+1)h + u, y)dudy.\end{aligned}$$

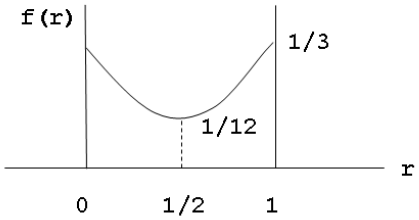

Figure 14: Shape of the function $f(r)$

With Euler-Maclaurin we get

$$\begin{aligned}\mathbb{E}(X^*Y) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-rh}^{rh}\frac{z}{2h}y\varphi(z+u,y)dudzdy \\ &+ \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-(1-r)h}^{(1-r)h}\frac{z}{2h}y\varphi(z+u,y)dudzdy.\end{aligned}$$

With $x = z + u$, $\tilde{u} = \frac{u}{h}$ the right hand side becomes

$$\frac{1}{2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\left[\int_{-r}^{r}(x-\tilde{u}h)d\tilde{u} + \int_{-(1-r)}^{1-r}(x-\tilde{u}h)d\tilde{u}\right]y\varphi(x,y)dxdy =$$

$$= \frac{1}{2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[2rx+2(1-r)x]y\varphi(x,y)dxdy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}xy\varphi(x,y)dxdy = \mathbb{E}(XY).$$

Thus

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

Similar relations hold if both X and Y or if only Y is (asymmetrically) rounded. As a consequence, for the covariances of rounded and unrounded variables, the previous esymmetricquations (18) and (19) hold also true in the case of asymmetric rounding.

Now we turn to linear regression in the case of asymmetric rounding. We consider the case where only X is rounded.

As in Section 4, see (22), we have

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

with $f(r) = r^2 - r + \frac{1}{3}$.

11 Conclusion

Rounding of data has the inevitable consequence that their statistical moments, in particular mean and variance, (and consequantly also regression parameters) computed from such data are more or less distorted in comparison to the moments of the unrounded data. This survey looks into 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.

Most of the paper is concerned with population moments. But 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 get consistent estimates.

Sheppard's correction is generalized to the case of asymmetric rounding. The correction turns out to be a function of the symmetry portion r.

12 References

1. Baten, W. D., (1931). Correction for the moments of a frequency distribution in two variables. *Annals of Mathematical Statistics* **2** 309-312.

2. Dempster, A. P., & Rubin, D. B., (1983). Rounding error in regression: The appropriateness of Sheppard's correction. *Journal of the Royal Statistical Society* **B 45** 51-59.

3. Don, F. J. H., (1981). A note on Sheppard's corrections for grouping and maximum likelihood estimation. *Journal of Multivariate Analysis* **11** 452-458.

4. Eisenhart, C., Hastay, M. W., & Wallis, W. A., (1947). *Techniques of statistical analysis*, first edition. McGRAW-HILL Book Company, Inc. New York and London.

5. Fryer, J. G., & Pethybridge, R. J., (1972). Maximum likelihood estimation of a linear regression function with grouped data. *Applied Statistics* **21** 142-154.

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

7. Gjeddebaek, N. F., (1956). Contribution to the study of grouped observations II. Loss of information caused by grouping of normally distribution observations. *Skandinavisk Aktuarietidskrift* **39** 154-159.

8. Gjeddebaek, N. F., (1968). Statistical analysis: Grouped observation. *In International Encyclopedia of Social Sciences* (D. R. Sills, ed.) **15**. Macmillan and Free Press, New York.

9. Haitovsky, Y., (1982). Grouped data. *Encyclopedia of Statistical Sciences* **3** 527-536. Wiley, New York.

10. Heitjan, D, F., (1989). Inference from grouped continuous data: A review. *Statistical science* **4** 164-179.

11. Janson, S., (2005). Rounding of continuous random variables and oscillatory asymptotic. *www.math.uu.se/ svante/papers/sj175.pdf.*

12. Kendall, M. G., (1938). The conditions under which Sheppard's corrections are valid. *Journal of the Royal Statistical Society* **101** 592-605.

13. Kenney, J. F. and Keeping, E. S, (1962). Sheppard's Correction for Grouping Errors. *§7.6 in Mathematics of Statistics, Pt. 1, 3rd ed. Princeton, NJ: Van Nostrand.*

14. Komlos, J. (1994). The Stature of Runaway Slaves in Colonial America, in J. Komlos (ed.) Stature, *Living Standards, and Economic Development: Essays in Anthropometric History, Chicago: University of Chicago Press*, pp. 93-116.

15. Kullback, S., (1935). A note on Sheppard's coorections. *Annals of Mathematical Statistics* **6** 158-159.

16. Kulldorff, G., (1961). *Contributions to the theory of estimation from grouped and partially grouped samples.* Almqvist and Wiksell, Stockholm.

17. Lindley, D. V., (1950). Grouping corrections and maximum likelihood equations. *Proceeding of the Cambridge Philosophical Society* **46** 106-110.

18. Pairman, E. & Pearson, K. (1919). 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.

19. Rietz, H. L., (1924). *Handbook of Mathematical Statistics.* Houghton Mifflin Company, Boston.

20. Sheppard, W. F., (1898). On the calculation of the most probable values of frequency constants for data arranged according to equidistant division of a scale. *Proceeding of the London Mathematical Society* **29** 353-380.

21. Stuart, A., & Ord, J. K., (1987). *Kendall's Advanced Theory of Statistics.* V0l. 1, Distribution Theory, 5th Edition, London: Charles Griffin & Company.

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

23. Tallis, G. M., (1967). Approximate maximum likelihood estimation from grouped data. *ometrics* **9** 599-606.

24. Tallis, G. M., & Young, S. S., (1962). Maximum likelihood estimation of parameters of the normal, log-normal, truncated normal and bivariate normal distributions from grouped data. *The Australian Journal of Statistics* **4** 49-54.

25. Tricker, A. R., (1984). Effects of rounding on the moments of a probability distribution. *The Statistician* **33** 381-390.

26. Varedeman, S. B., (2003). Sheppard's correction for variances and the "Quantization Noise Model". *www.public.iastate.edu/ vardeman/Sheppard.pdf.*

27. Wold, H., (1934). Sheppard's correction formulae in several variables. *Skandinavisk Aktuarietidskrift* **17** 248-255.