LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

INSTITUT FÜR STATISTIK

Schneeweiss, H. & Komlos, J.

# Probabilistic Rounding and Sheppard's Correction

Technical Report Number 045, 2008
Department of Statistics
University of Munich

http://www.stat.uni-muenchen.de

# Probabilistic rounding and Sheppard's correction

Schneeweiss, H. and Komlos, J.

December 16, 2008

## Abstract

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

*Keywords*: Rounding, asymmetric rounding, mixture rounding, heaping, profile function, Sheppard's correction, moments, linear regression.

## 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). The reason may be the avoidance of costs associated with a fine measurement or the imprecise nature of the measuring instrument. Even if precise measurements are available, they are sometimes recorded in a coarsened way in order to preserve confidentiality or to compress the data into an easy to grasp frequency table.

Two recent reviews of the field are Heitjan (1989) and Schneeweiss et al. (2006).

Most of the literature is concerned with simple rounding as described above. But there are other types of rounding procedures, where certain numbers are preferred over others. In asymmetric rounding, for example, more than half of the rounding interval is rounded to one of the the round values and less than half to the neighboring round value, Komlos (1999). In mixture rounding, different portions of the population round in different ways, e.g. some preferring even values, some odd values, again some preferring zeros or fives as the last digit, Wright and Bray (2003).

We generalize these approaches by introducing the concept of probabilistic rounding. Numbers are rounded to a round value with certain probabilities which depend on the distance of the original value to the round value. The probability as a function of the distance is given by a so-called rounding profile function. Again there may be several profile functions depending on whether some rounded values are preferred over other ones. We only consider two profile functions below, one for even and one for odd numbers. Profile functions, though not with this name, were employed by Torelli and Trivellato (1993) to describe heaping in unemployment duration data. A somewhat different model of probabilistic heaping was used by Heitjan and Rubin (1991).

The mean and the variance and also higher moments of a variable $X$ calculated using rounded data $X^*$ instead of the original data $X$ will be biased. However, under certain smoothness conditions, see, e.g., Kendall(1938), Schneeweiss *et al.*(2006), the means of $X$ and $X^*$ do not differ very much and can be considered as almost equal. Yet, the variances differ markedly. However, the difference is captured, to a high degree of accuracy, by a very simple term, $h^2/12$, where $h$ is the distance between neighboring values of $X^*$. This is the famous Sheppard's correction (1898). The purpose of the present paper is to extend Sheppard's correction to the case of probabilistic rounding. We derive a similar, though more complicated, correction term for the variance (and in principle also for higher moments) which depends on the profile function of the rounding procedure.

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

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

# 2 Probabilistic rounding

## 2.1 Simple rounding

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

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

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

$$X^* = ih \text{ if } X \in [ih \pm h/2]. \quad (1)$$

Here we use a simplifying notation: $[m \pm d]$ denotes the interval $[m - d, m + d]$. As $X$ is a continuous variable, the case that $X$ is equally distant from two lattice points has probability 0 and will therefore be disregarded. (In practice, of course, this case can come up and then it is common practice to shift $X$ to the higher lattice point.) It is well-known that the mean and variance computed from the rounded variable $X^*$ differ from the mean and variance of the original variable $X$. For the mean, the difference is typically very small if the distribution of $X$ is smooth enough and $h$ is not too large. E.g., if $X$ is Gaussian and $h$ is not larger than twice the standard deviation of $X$, then the mean of $X^*$ hardly differs from the mean of $X$:

$$\mathbb{E}X^* \approx \mathbb{E}X.$$

However, the variances differ more pronouncedly. In fact, for a smooth distribution and not too large rounding width $h$, the variances differ by an amount $h^2/12$, the so-called Sheppard's correction:

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

That is to say, if the variance has been computed from a set of rounded data, the true variance (i.e. the variance of the original variable) is found by "correcting" the rounded variance by subtracting the amount $h^2/12$. This has consequences for the estimation of linear regressions. The slope $\beta$ of a linear regression line $y = \alpha + \beta x$, if computed from rounded data, has to be corrected by multiplying it with the factor

$$\left[1 - \frac{1}{12}\left(\frac{h}{s_{x^*}}\right)^2\right]^{-1}, \tag{3}$$

where $s_{x^*}$ is the standard deviation of $X^*$.

## 2.2 Asymmetric rounding

Simple rounding treats all lattice points in the same way. Sometimes, however, there is a preference for one type of lattice points over the other, Komlos (1999). Let us suppose that even and odd lattice points have different preferences such that a smaller or larger portion $r$ of the rounding interval is shifted to the even lattice point and the remaining part to the odd lattice point. In this case,

$$X^* = \begin{cases} 2ih & \text{if } X \in [2ih \pm rh] \\ (2i+1)h & \text{if } X \in [(2i+1)h \pm (1-r)h], \end{cases} \tag{4}$$

where $0 \le r \le 1$. The case $r = 1/2$ corresponds to simple rounding, while $r = 1$ (or 0) means that all values are rounded to the nearest even (or odd) lattice point. For asymmetric rounding, again $\mathbb{E}X^* \approx \mathbb{E}X$ under similar conditions as for simple rounding. There is also a Sheppard-like correction for the variance, however with a different correction term, cf. Schneeweiss *et al.*(2006):

$$\mathbb{V}X^* \approx \mathbb{V}X + \frac{1}{3}(1 - 3r + 3r^2)h^2. \tag{5}$$

For $r = 1/2$ this reduces to Sheppard's correction.

## 2.3 Mixture rounding

There is another deviation from simple rounding that may occur in practice. Suppose $X$ is a characteristic feature of the members of some population, and suppose that some portion $m$, $0 \le m \le 1$, of the population always rounds $X$ to the nearest even lattice point, while the other part $1-m$ rounds $X$ to the nearest odd lattice point. The distribution of $X$ is the same in both subpopulations. We then have a mixture of rounding procedures with mixing parameters $m$ and $1-m$, see also Wright and Bray (2003). We can give this situation a probability interpretation. If a member of the population is randomly chosen, then $X$ is a random variable, which is rounded to an even or odd lattice point with probability $m$ or $1-m$, respectively, i.e.,

$$\mathbb{P}(X^* = 2ih|X) = \begin{cases} m & \text{if } X \in [2ih \pm h] \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

$$\mathbb{P}(X^* = (2i+1)h|X) = \begin{cases} 1-m & \text{if } X \in [(2i+1)h \pm h] \\ 0 & \text{otherwise} \end{cases}$$

We shall see that for this kind of rounding procedure, again $\mathbb{E}X^* \approx \mathbb{E}X$ under suitable conditions, and again there is a Sheppard-like correction for the computation of the variance. Indeed (see Section 3),

$$\mathbb{V}X^* \approx \mathbb{V}X + \frac{h^2}{3}, \tag{7}$$

irrespective of the parameter $m$. The first two, deterministic, rounding procedures can also be given a probabilistic interpretation. Indeed, for e.g. asymmetric rounding,

$$\mathbb{P}(X^* = 2ih|X) = \begin{cases} 1 & \text{if } X \in [2ih \pm rh] \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

$$\mathbb{P}(X^* = (2i+1)h|X) = \begin{cases} 1 & \text{if } X \in [(2i+1)h \pm (1-r)h] \\ 0 & \text{otherwise} \end{cases}$$

It therefore seems natural to look for a general probabilistic rounding procedure that comprises the procedures studied up to now.

### 2.4 General probabilistic rounding

We introduce a so-called rounding profile function $q(u)$ with the properties that

$$0 \leq q(u) \leq 1$$

$$q(u) = 0 \text{ for } |u| > 1$$

$$q(-u) = q(u).$$

Typically $q(u)$ will be a decreasing function for $0 \leq u \leq 1$, but this property will most often not be needed in the following. The function $q(u)$ can be thought of as the probability that a random variable $U$ is shifted to 0, given that $U$ takes the value $u$, when the lattice width is $h = 1$:

$$\mathbb{P}(U^* = 0|U = u) = q(u).$$

More generally, let us suppose that the rounding to even lattice points is performed probabilistically according to the conditional probability

$$\mathbb{P}(X^* = 2ih|X) = q\left(\frac{X - 2ih}{h}\right). \tag{9}$$

Rounding to odd lattice points is done in a similar way with the help of the following "complementary" profile function

$$\overline{q}(u) = \begin{cases} 1 - q(1 - |u|) & \text{if } |u| \leq 1 \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

The function $\overline{q}$ has the same properties as $q$. In addition,

$$\overline{q}(u) + q(1-u) = 1 \text{ for } 0 \le u \le 1. \tag{11}$$

We then have

$$\mathbb{P}(X^* = (2i+1)h|X) = \overline{q}\left(\frac{X-(2i+1)h}{h}\right). \tag{12}$$

For any $X$ between the two lattice points $2ih$ and $(2i+1)h$, due to (10),

$$\mathbb{P}(X^* = 2ih|X) + \mathbb{P}(X^* = (2i+1)h|X) = 1, \tag{13}$$

as it should be, and similarly if $(2i-1)h < X < 2ih$. The rounding procedures described in Section 2.1 to 2.3 are all special cases of probabilistic rounding. We obtain

- simple rounding for $q(u) = \begin{cases} 1 & \text{if } |u| < \frac{1}{2} \\ 0 & \text{otherwise,} \end{cases}$
- asymmetric rounding for $q(u) = \begin{cases} 1 & \text{if } |u| < r \\ 0 & \text{otherwise,} \end{cases}$
- mixture rounding for $q(u) = \begin{cases} m & \text{if } |u| < 1 \\ 0 & \text{otherwise.} \end{cases}$

Probabilistic rounding can be symmetric, in which case

$$q(u) = \overline{q}(u) \text{ for all } u.$$

Symmetric probabilistic rounding is characterized by a profile function $q$ with the property

$$q(u) = 1 - q(1-u) \text{ for } 0 \leq u \leq 1. \tag{14}$$

Simple rounding and mixture rounding with $m = \frac{1}{2}$ are both symmetric probabilistic rounding procedures.

An interesting application of probabilistic rounding is found in M'Raïhi *et al.* (2001), where two profile functions are considered: $q(u) = \frac{1}{2}$ and $q(u) = 1 - u$, $0 \leq u \leq 1$.

# 3 Sheppard's correction

We compare the mean and variance and, more generally, any moment of the rounded variable $X^*$ to the corresponding moment of the original variable $X$. We first compute the probability distribution of $X^*$ from the distribution of $X$, which is given by its density function $\varphi(x)$.

$$\begin{aligned} \mathbb{P}(X^* = 2ih) &= \int_{-\infty}^{\infty} \mathbb{P}(X^* = 2ih | X = x)\varphi(x)dx \\ &= \int_{-\infty}^{\infty} q(\frac{x - 2ih}{h})\varphi(x)dx \\ &= h \int_{-1}^{1} q(u)\varphi(2ih + hu)du \qquad (15) \end{aligned}$$

and similarly

$$\mathbb{P}(X^* = (2i+1)h) = h \int_{-1}^{1} \overline{q}(u)\varphi((2i+1)h + hu)du. \tag{16}$$

We can now compute the $k$-th moment of $X^*$:

$$\begin{aligned}\mathbb{E}X^{*k} &= \sum_{i=-\infty}^{\infty}(2ih)^k\mathbb{P}(X^*=2ih)+\sum_{i=-\infty}^{\infty}[(2i+1)h]^k\mathbb{P}(X^*=(2i+1)h)\\ &= \sum_{i=-\infty}^{\infty}(2ih)^k h\int_{-1}^{1}q(u)\varphi(2ih+hu)du+\sum_{i=-\infty}^{\infty}[(2i+1)h]^k h\int_{-1}^{1}\overline{q}(u)\varphi((2i+1)h+hu)du.\end{aligned}$$

Under the condition that the functions under the two sums are sufficiently smooth and the lattice width $h$ is not too large, the two sums can be approximated by corresponding integrals. For the first sum, replace $2ih$ with the continuous variable $t$ and $2h$ with the differential $dt$. Similarly for the second sum, replace $(2i+1)h$ with $t$ and again $2h$ with $dt$. Finally replace $\sum_{i=-\infty}^{\infty}$ with $\int_{-\infty}^{\infty}$. We then obtain approximately

$$\mathbb{E}X^{*k} \approx \frac{1}{2}\int_{-\infty}^{\infty}t^k\int_{-1}^{1}q(u)\varphi(t+hu)dudt+\frac{1}{2}\int_{-\infty}^{\infty}t^k\int_{-1}^{1}\overline{q}(u)\varphi(t+hu)du\,dt.$$

This approximation can be justified by invoking the Euler-Maclaurin formula, according to which a sum of the form $h\sum_{i=-\infty}^{\infty}f(ih)$ can be approximated by the integral $\int_{-\infty}^{\infty}f(t)dt$ if $f$ is sufficiently smooth and $h$ is not too large. For details, in particular for the conditions involved, see Schneeweiss *et al.* (2006). Note that in our application $h$ has to be replaced with $2h$. Now, by a change of variables $(z=t+hu, u=u)$, the two double integrals can be transformed into

$$\mathbb{E}X^{*k} \approx \frac{1}{2}\int_{-\infty}^{\infty}\int_{-1}^{1}(z-hu)^k[q(u)+\overline{q}(u)]du\varphi(z)dz. \tag{17}$$

Considering to the first two moments ($k=1$ and $k=2$), we obtain the main result of the paper.

**Theorem 1.** *For the mean and the variance of $X^*$ the following approximate relations hold:*

$$\mathbb{E}X^* \approx \mathbb{E}X \tag{18}$$

$$\mathbb{V}X^* \approx \mathbb{V}X + \left(\frac{1}{3} - Q_0 + 2Q_1\right) h^2 =: \mathbb{V}X + Qh^2, \tag{19}$$

*where* $Q_0 = \int_0^1 q(u)du$, $Q_1 = \int_0^1 uq(u)du$, *and* $Q = \frac{1}{3} - Q_0 + 2Q_1$.

Thus the means of the rounded and the unrounded variables are approximately the same and the variances differ by the amount $Qh^2$, which is Sheppard's correction for probabilistic rounding.

*Proof.* Let us start with some preliminary results, which will be useful also in Section 5. In evaluating the right side of (17), one has to compute integrals of the form $\int\limits_{-1}^{1} u^j[q(u) + \overline{q}(u)]du$. This integral is zero for odd $j$. So we need only consider even $j$. If we define

$$Q_j := \int\limits_0^1 u^j q(u)du, j = 0, 1, 2, ..., \tag{20}$$

as the $j$-th "half moment" of $q(u)$, then for $j$ even,

$$
\begin{aligned}
&\frac{1}{2}\int_{-1}^{1} u^j[q(u)+\overline{q}(u)]du \\
&= \int_{0}^{1} u^j[q(u)+\overline{q}(u)]du \\
&= \int_{0}^{1} u^j[q(u)+1-q(1-u)]du \qquad \text{due to (11)} \\
&= \int_{0}^{1} u^j du + \int_{0}^{1} u^j q(u)du - \int_{0}^{1} (1-u)^j q(u)du \qquad \text{by change of variable} \\
&= \frac{1}{j+1} - \sum_{i=0}^{j-1} (-1)^i \binom{j}{i} Q_i.
\end{aligned}
$$

In particular, we have for

$$
j=0: \quad \frac{1}{2}\int_{-1}^{1} [q(u)+\overline{q}(u)]du = 1, \tag{21}
$$

$$
j=2: \quad \frac{1}{2}\int_{-1}^{1} u^2[q(u)+\overline{q}(u)]du = \frac{1}{3} - Q_0 + 2Q_1 =: Q, \tag{22}
$$

$$
j=4: \quad \frac{1}{2}\int_{-1}^{1} u^4[q(u)+\overline{q}(u)]du = \frac{1}{5} - Q_0 + 4Q_1 - 6Q_2 + 4Q_3 =: Q'. \tag{23}
$$

Now with $k=1$, we obtain the mean of $X^*$ from (17):

$$
\mathbb{E}X^* \approx \frac{1}{2}\int_{-\infty}^{\infty}\int_{-1}^{1} (z-hu)[q(u)+\overline{q}(u)]du\varphi(z)dz = \int_{-\infty}^{\infty} z\varphi(z)dz = \mathbb{E}X,
$$

where we used (21).

With $k = 2$, we obtain the second moment of $X^*$:

$$\begin{aligned}\mathbb{E}X^{*2} &\approx \frac{1}{2}\int\limits_{-\infty}^{\infty}\int\limits_{-1}^{1}(z^2 - 2zhu + h^2u^2)[q(u) + \overline{q}(u)]du\phi(z)dz \\ &= \int\limits_{-\infty}^{\infty} z^2\varphi(z)dz + \frac{h^2}{2}\int\limits_{-1}^{1} u^2[q(u) + \overline{q}(u)]du \int\limits_{-\infty}^{\infty} \varphi(z)dz \\ &= \mathbb{E}X^2 + \left(\frac{1}{3} - Q_0 + 2Q_1\right)h^2,\end{aligned}$$

where we used (21) and (22). Subtracting on the left side $(\mathbb{E}X^*)^2$ and on the right side $(\mathbb{E}X)^2$, which are approximately equal, we obtain (19) □

If we specialize to the rounding procedures of Section 2, we see (see Appendix A.2) that for

- simple rounding, $Q_0 = \frac{1}{2}$, $Q_1 = \frac{1}{8}$, and $Q = \frac{1}{12}$,
- asymmetric rounding, $Q_0 = r$, $Q_1 = \frac{r^2}{2}$, and $Q = \frac{1}{3} - r + r^2$,
- mixture rounding, for $Q_0 = m$, $Q_1 = \frac{m}{2}$, and $Q = \frac{1}{3}$,

thereby verifying (2), (5), and (7).

For symmetric probabilistic rounding, $Q_0 = \frac{1}{2}$.

We need to know the two key parameters $Q_0$ and $Q_1$ in order to apply Sheppard's correction. The question is whether $Q_0$ and $Q_1$ can be derived from the distribution of the rounded data. This can be done for $Q_0$. If we consider only the probabilities of the even lattice points, then their sum is approximately equal to $Q_0$:

(24)
$$\sum_{i=-\infty}^{\infty} \mathbb{P}(X^* = 2ih) \approx Q_0.$$

Indeed, by (15),

$$\sum_{i=-\infty}^{\infty} \mathbb{P}(X^* = 2ih) = h \sum_{i=-\infty}^{\infty} \int_{-1}^{1} q(u)\varphi(2ih + hu)du \approx \frac{1}{2} \int_{-\infty}^{\infty} \int_{-1}^{1} q(u)\varphi(t + hu)dudt$$
$$= \frac{1}{2} \int_{-\infty}^{\infty} \int_{-1}^{1} q(u)\varphi(z)dudz = \frac{1}{2} \int_{-1}^{1} q(u)du = \int_{0}^{1} q(u)du = Q_0.$$

(Doing the same for the odd lattice points leads to a similar approximate equation: $\sum \mathbb{P}(X^* = (2i+1)h) \approx \overline{Q}_0 =: \int_0^1 \overline{q}(u)du = 1 - Q_0$ because of (11)). Unfortunately there does not seem to be a similarly simple relation for $Q_1$. Only if we know the general form of the rounding profile, can we hope to derive $Q_1$ from the rounded data. An example is given in Section 5. However, we can find upper and lower bounds for $Q_1$ given $Q_0$ (which imply corresponding bounds for $Q$):

**Theorem 2.** *Suppose $q(u)$ is a monotonely decreasing function for $0 \le u \le 1$. Then*

$$\frac{1}{2}Q_0^2 \le Q_1 \le \frac{1}{2}Q_0.$$

*The lower bound, $Q_{1min} = \frac{1}{2}Q_0^2$, is obtained for asymmetric rounding with $r = Q_0$ and the upper bound, $Q_{1max} = \frac{1}{2}Q_0$, for mixture rounding with $m = Q_0$.*

Thus asymmetric rounding and mixture rounding are the two extreme cases of probabilistic rounding.

*Proof.* The proof comes in two parts.

1) Let $Q_0 = r$. Then

$$\begin{aligned}
Q_1 - \frac{1}{2}Q_0^2 &= Q_1 - \frac{r^2}{2} \\
&= \int_0^1 uq(u)du - \int_0^r udu \\
&= -\int_0^r u(1-q(u))du + \int_r^1 uq(u)du \\
&\geq -r\int_0^r (1-q(u))du + r\int_r^1 q(u)du \\
&= r\left(-\int_0^r du + \int_0^1 q(u)du\right) \\
&= r(-r + Q_0) = 0.
\end{aligned}$$

Thus

$$Q_1 \geq \frac{1}{2}Q_0^2.$$

2) Let $Q_0 = m$. Since $q(u)$ was supposed to be monotonely decreasing, there exists $u_0$, $0 \leq u_0 \leq 1$, such that

$$q(u) \geq m \text{ for } u \leq u_0$$

$$q(u) \leq m \text{ for } u \geq u_0.$$

Then

$$
\begin{aligned}
\frac{1}{2}Q_0 - Q_1 &= \frac{m}{2} - \int_0^1 uq(u)du \\
&= \int_0^1 u(m - q(u))du \\
&= -\int_0^{u_0} u(q(u) - m)du + \int_{u_0}^1 u(m - q(u))du \\
&\geq -u_0 \int_0^{u_0} (q(u) - m)du + u_0 \int_{u_0}^1 (m - q(u))du \\
&= u_0 \int_0^1 (m - q(u))du \\
&= u_0(m - Q_0) = 0.
\end{aligned}
$$

Thus

$$
Q_1 \leq \frac{1}{2}Q_0.
$$

□

# 4 Regression

Doing regressions with rounded data may lead to biased estimates. For a linear regression of $Y$ on $X$, we need to compute not only means and variances but also the covariance of $X$ and $Y$. Let us assume that $X$ and $Y$ are two continuous random variables with joint density $\varphi(x, y)$. We consider two cases, one where both variables have been rounded and another where only one variable, $X$ say, has been rounded. In the first case, assume that $X$ has been rounded to $X^*$ just as in Section 2 with rounding width $h_x$ and rounding profiles

$q_x$ and $\overline{q}_x$ and that $Y$ has been rounded to $Y^*$ with width $h_y$ and profiles $q_y$ and $\overline{q}_y$. We assume that the two rounding procedures are stochastically independent. Then

$$
\begin{aligned}
(25) \qquad \mathbb{E}X^*Y^* = &\sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} 2ih_x 2jh_y \mathbb{P}(X^* = 2ih_x, Y^* = 2jh_y) \\
&+ \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} (2i+1)h_x 2jh_y \mathbb{P}(X^* = (2i+1)h_x, Y^* = 2jh_y) \\
&+ \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} 2ih_x (2j+1)h_y \mathbb{P}(X^* = 2ih_x, Y^* = (2j+1)h_y) \\
&+ \sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} (2i+1)h_x (2j+1)h_y \mathbb{P}(X^* = (2i+1)h_x, Y^* = (2j+1)h_y).
\end{aligned}
$$

Let us consider only the first double sum. With the help of the profile functions $q_x$ and $q_y$ it can be written as

$$
\begin{aligned}
&\sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty} 2ih_x 2jh_y \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathbb{P}(X^* = 2ih_x, Y^* = 2jh_y)|X = x, Y = y)\varphi(x,y)dxdy \\
&= \sum_i \sum_j 2ih_x 2jh_y \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} q_x(\frac{x - 2ih_x}{h_x}) q_y(\frac{y - 2jh_y}{h_y})\varphi(x,y)dxdy \\
&= \sum_i \sum_j 2ih_x 2jh_y h_x h_y \int_{-1}^{1}\int_{-1}^{1} q_x(u)q_y(v)\varphi(2ih_x + h_x u, 2jh_y + h_y v)dudv.
\end{aligned}
$$

Replacing $2ih_x$ and $2jh_y$ with continuous variables $t$ and $s$ and $2h_x$ and $2h_y$ with the differentials $dt$ and $ds$, respectively, and the double sum with a double integral, the double sum becomes approximately

$$
\frac{1}{4}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} ts \int_{-1}^{1}\int_{-1}^{1} q_x(u)q_y(v)\varphi(t + h_x u, s + h_y v)dudvdtds.
$$

The other three double sums of (25) can be transformed in the same way. We get similar results, except that the product term $q_x(u)q_y(v)$ is replaced with $\overline{q}_x(u)q_y(v)$, $q_x(u)\overline{q_y}(v)$, and $\overline{q_x}(u)\overline{q_y}(v)$, respectively. Collecting terms, we have the approximation

$$\mathbb{E}X^*Y^* \approx \frac{1}{4}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} ts \int_{-1}^{1}\int_{-1}^{1} [q_x(u)+\overline{q_x}(u)][q_y(v)+\overline{q_y}(v)]\varphi(t+h_x u, s+h_y v)du\, dv\, dt\, ds.$$

With the change of variables $z = t + h_x u$, $w = s + h_y v$, $u = u$, $v = v$,

$$\begin{aligned}
\mathbb{E}X^*Y^* &\approx \frac{1}{4}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\int_{-1}^{1}\int_{-1}^{1}(z-h_x u)(w-h_y v)[q_x(u)+\overline{q_x}(u)][q_y(v)+\overline{q_y}(v)]du\, dv\, \varphi(z,w)dz\, dw \\
&= \frac{1}{4}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} zw\varphi(z,w)dz\, dw \int_{-1}^{1}\int_{-1}^{1}[q_x(u)+\overline{q_x}(u)][q_y(v)+\overline{q_y}(v)]du\, dv \\
&= \frac{1}{4}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} zw\varphi(z,w)dz\, dw \cdot 4 = \mathbb{E}XY
\end{aligned}$$

because of (21). It follows that a similar approximation also holds for the covariances:

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

In the second case, where only $X$ has been rounded with $h$ as rounding width and $q$ as rounding profile, we have

$$
\begin{aligned}
\mathbb{E}X^*Y =& \sum_{i=-\infty}^{\infty} \int_{-\infty}^{\infty} 2ihy\mathbb{P}(X^* = 2ih|X = x)\varphi(x,y)dx\,dy \\
&+ \sum_{i=-\infty}^{\infty} \int_{-\infty}^{\infty} (2i+1)hy\mathbb{P}(X^* = (2i+1)h|X = x)\varphi(x,y)dx\,dy \\
=& \sum_{i=-\infty}^{\infty} \int_{-\infty}^{\infty} 2ihyh \int_{-1}^{1} q(u)\varphi(2ih + hu, y)du\,dy \\
&+ \sum_{i=-\infty}^{\infty} \int_{-\infty}^{\infty} (2i+1)hyh \int_{-1}^{1} \overline{q}(u)\varphi((2i+1)h + hu, y)du\,dy \\
\approx& \frac{1}{2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} ty \int_{-1}^{1} [q(u) + \overline{q}(u)]\varphi(t + hu, y)du\,dt\,dy \\
=& \frac{1}{2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-1}^{1} (z - hu)y[q(u) + \overline{q}(u)]\varphi(z,y)du\,dz\,dy \\
=& \frac{1}{2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} zy \int_{-1}^{1} [q(u) + \overline{q}(u)]du\varphi(z,y)dz\,dy \\
=& \mathbb{E}XY,
\end{aligned}
$$

because $\int_{-1}^{1} uq(u)du = \int_{-1}^{1} u\overline{q}(u)du = 0$ and because of (21). Again it follows that

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

So the covariance between two continuous variables never changes much by rounding, regardless of whether only one or both variables are rounded, so long as the regularity conditions for good approximations in rounding are satisfied. The consequences for linear regression analysis are clear. Suppose we want to evaluate the parameters of a linear regression

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

by least squares and we use rounded data, the results may be biased. But they can easily be corrected with Sheppard's correction if the distributions of $X$ and $Y$ are sufficiently smooth and $h_x$ and $h_y$ are not too large. If $X$ has been rounded (but not $Y$) and if

$$\mathbb{V}_c X^* = \mathbb{V}X^* - Qh^2 \tag{26}$$

is the "corrected" variance of the rounded variable $X$ using a Sheppard-like correction, so that $\mathbb{V}_c X^* \approx \mathbb{V}X$, then a corrected slope parameter is given by

$$\beta_c = \beta^* \frac{\mathbb{V}X^*}{\mathbb{V}_c X^*}, \tag{27}$$

where $\beta^* = \frac{\mathbb{C}ov(X^*,Y)}{\mathbb{V}X^*}$ is the slope parameter of the regression with the rounded variable: $Y = \alpha^* + \beta^* X^* + \epsilon^*$. Then $\beta_c$ will be approximately equal to $\beta = \frac{\mathbb{C}ov(X,Y)}{\mathbb{V}X}$. The same conclusion follows if also $Y$ has been rounded, in which case $\beta^* = \frac{\mathbb{C}ov(X^*,Y^*)}{\mathbb{V}X^*}$ and $\beta_c$ is again given by (27). However, if $X$ has not been rounded but only $Y$, $\beta^*$ will be approximately equal to $\beta$ and no correction is needed. In summary, a correction is only necessary if $X$ has been rounded and then (27) together with (26) gives the correction formula.

If the residual variance

$$\sigma_\epsilon^2 = \mathbb{V}Y - \frac{\mathbb{C}ov(X,Y)^2}{\mathbb{V}X}$$

is computed from rounded variables $X^*$ and $Y^*$ in place of $X$ and $Y$, then $\sigma_{\epsilon^*}^2$ will be biased.

But a corrected version

$$\sigma^2_{\epsilon,c} = \mathbb{V}_c Y^* - \frac{\mathbb{C}ov(X^*, Y^*)^2}{\mathbb{V}_c X^*}$$

will be approximately unbiased. If $X$ or $Y$ have not been rounded the corresponding variance corrections are, of course, not needed.

# 5 Determining $Q_1$ and $Q$

In Section 3 we noted that we need to know the general form of $q(u)$ in order to be able to compute $Q_1$ and $Q$ from a given distribution of the rounded variable $X^*$. If $q(u)$ depends on only one unknown parameter, which is related one-to-one to $Q_0$, then $Q_1$ can be determined from $Q_0$, which in turn can be computed from the distribution of $X^*$, see (24). Examples are asymmetric rounding and mixture rounding.

However, if $q(u)$ depends on two (or more) unknown parameters, $\alpha$ and $r$, say, things are more difficult. One might still be able to determine $Q_1$ if in addition to knowing the function $q(u; \alpha, r)$ one also knows the distribution of the unrounded variable $X$ apart from a few unknown parameters.

To be more specific, suppose we know that $X \sim N(\mu, \sigma^2)$ with unknown $\mu$ and $\sigma^2$ and suppose we can plausibly specify a function $q(u; \alpha, r)$ of the following form

$$(28) \qquad q(u; \alpha, r) = \begin{cases} 1 - \frac{1}{2}(\frac{u}{r})^\alpha & \text{for } 0 \le u \le r \\ \frac{1}{2}(\frac{1-u}{1-r})^\alpha & \text{for } r \le u \le 1 \end{cases}$$

with $0 \le r \le 1, \alpha \ge 0$. Then using higher moments of $X^*$, one can derive equations for the unknown parameters $r$ and $\alpha$. If these were solved for $r$ and $\alpha$, the profile function $q(u)$ would be completely specified and $Q_1$ could be determined.

This procedure, however, is rather complicated. But a simplified problem can be solved. Suppose we know that the rounding procedure (28) is a symmetric one, then $r = \frac{1}{2}$ and the function $q(u; \alpha, r)$ reduces to a one-parameter family:

$$q(u;\alpha) = \begin{cases} 1 - 2^{\alpha-1}u^{\alpha} & \text{for } 0 \le u \le \frac{1}{2} \\ 2^{\alpha-1}(1-u)^{\alpha} & \text{for } \frac{1}{2} \le u \le 1. \end{cases} \tag{29}$$

The profile function specializes to the case of simple rounding if $\alpha = \infty$ and to the case of mixture rounding with $m = \frac{1}{2}$ if $\alpha = 0$. For $\alpha = 1$, the profile function becomes $q(u) = 1 - |u|$, $-1 \le u \le 1$.

The parameter $\alpha$ is not in a one-to-one relation with $Q_0$. Indeed, $Q_0 = \frac{1}{2}$, whatever the value of $\alpha$. But $\alpha$ determines $Q_1$. So, once $\alpha$ is known, $Q_1$ is also known and the generalized Sheppard correction can be used. Indeed with some algebra one obtains (see appendix)

$$Q_1 = \frac{1}{4}\left(\frac{1}{2} + \frac{1}{(\alpha+1)(\alpha+2)}\right) \tag{30}$$

and consequently

$$Q = \frac{1}{12} + \frac{1}{2(\alpha+1)(\alpha+2)}. \tag{31}$$

In order to set up an equation for $\alpha$, we derive an approximate expression for the fourth moment of the rounded variable $X^*$ by setting $k = 4$ in (17) and using (22) and (23):

$$EX^{*4} \approx EX^4 + 6EX^2Qh^2 + Q'h^4. \tag{32}$$

As $EX^* \approx EX$ we may as well replace $X$ with $X - EX$ in the moment expressions. Denoting the second and fourth central moments of $X$ and $X^*$ by $\mu_2$, $\mu_4$, $\mu_2^*$, and $\mu_4^*$, respectively, we can now rewrite (19) and (32) as

$$\mu_2^* \approx \mu_2 + Qh^2$$

$$\mu_4^* \approx 3\mu_2^2 + 6\mu_2 Q h^2 + Q' h^4,$$

where we used the fact that $\mu_4 = 3\mu_2^2$ for a normal variable $X$. Substituting the first equation in the second one, we obtain

$$\mu_4^* - 3\mu_2^{*2} \approx (Q' - 3Q^2)h^4. \tag{33}$$

The expression $b(\alpha) := Q' - 3Q^2$ is a function of $\alpha$, which can be determined by computing $Q_j$, $j = 0, 1, 2, 3$, according to (20) with the profile function $q(u; \alpha)$. $Q_0 = \frac{1}{2}$ and $Q_1$ have already been given, see (30). In addition (see Appendix A.1):

$$Q_2 = \frac{1}{4}\left(\frac{1}{6} + \frac{1}{\alpha+1} - \frac{1}{\alpha+2}\right)$$
$$Q_3 = \frac{1}{32}\left(\frac{1}{4} + \frac{4}{\alpha+1} - \frac{6}{\alpha+2} + \frac{3}{\alpha+3} - \frac{1}{\alpha+4}\right).$$

From this and (31) we derive $Q$ and $Q'$ (Appendix A.1) and finally get

$$b(\alpha) = Q' - 3Q^2 = \frac{1}{4}\left(-\frac{1}{30} + \frac{1}{\alpha+1} - \frac{3}{\alpha+2} + \frac{3}{\alpha+3} - \frac{1}{\alpha+4} - \frac{3}{(\alpha+1)^2(\alpha+2)^2}\right).$$

Denoting $\delta := (\mu_4^* - 3\mu_2^{*2})/h^4$, we have the approximate equation

$$\delta \approx b(\alpha),$$

which we can solve for $\alpha$. The approximation is very accurate for $h$ not too large ($h < 1.5$). So one might expect a rather accurate solution. Unfortunately, the function $b(\alpha)$ is not monotone for $0 \leq \alpha \leq \infty$. It increases from $b(0) = -\frac{2}{15} = -0.133$ up to a maximum value of $b_{max} = -0.00826$ at $\alpha = \alpha_{max} = 4.71$ and then decreases to the asymptotic value of $-\frac{1}{120} = -0.00833$. So for all values $\delta > -0.00833$, the equation has two solutions $\alpha$. Before it reaches its maximum, the curve $b(\alpha)$ intersects the asymptotic line (-0.00833) at $\alpha = 3.38$. All solutions $\alpha > 3.38$ have a second solution in the same range $\alpha > 3.38$. If the first solution is near the critical value 3.38, the other one is extremely large. Fortunately, we are not so much interested in a precise value of $\alpha$ but rather in the constant $Q$ of Sheppard's correction. But the function $Q = Q(\alpha)$ becomes very flat for $\alpha > 3.88$ and its values come close to the asymptotic value $Q(\infty) = \frac{1}{12}$, which is Sheppard's correction for simple rounding. Thus it is suggested to use Sheppard's correction for simple rounding whenever one gets a solution for $\alpha$ (or rather two solutions) greater than 3.88.

Let us consider another case, where asymmetry is involved. Suppose we only want to decide between the two extreme rounding models: asymmetric rounding with asymmetry parameter $r$ or mixture rounding with mixture parameter $m$. In both cases, $r$ or $m$ can be found, at least approximately, from the rounded data via $Q_0$, see (24), where $Q_0 = r$ or $Q_0 = m$, respectively. Again a decision between these two cases can be made if $X \sim N(\mu, \sigma^2)$ by looking at the quantity $\mu_4^* - 3\mu_2^{*2}$. Computing this quantity for the two cases we get (see Appendix A.2)

$$(34) \quad \mu_4^* - 3\mu_2^{*2} \approx \begin{cases} -\frac{2}{15}h^4 & \text{for mixture rounding,} \\ -\left[\frac{1}{3}\{r^3 + (1-r)^3\}^2 - \frac{1}{5}\{r^5 + (1-r)^5\}\right] h^4 & \text{for asymm. rounding.} \end{cases}$$

For $r = 0$ or $r = 1$, these two expressions become identical. In practice none of these two expressions may come close to $\mu_4^* - 3\mu_2^{*2}$. But we may choose that rounding model for which

(34) is best satisfied.

# 6 Example

In Figure 1, a distribution of 1576 height measurements (in inches) is shown with integer and half-integer values, see Coclanis and Komlos (1995). The integer values show a clear preponderance over the half-integer values. So here is a case for (asymmetric) probabilistic rounding. The rounding lattice consists of all integer and half integer values and the lattice width is $h = 0.5$. The mean of the rounded data is 70.325, which we can take as an (essentially) unbiased estimate of the true mean. The variance of the rounded data is $\sigma^{*2} = \mu_2^* = 5.647$. Using Sheppard's correction $h^2/12 = 0.0208$ for *simple* rounding we obtain a corrected variance $\sigma_{cs}^2 = 5.626$. However, because of the preference for integer valued measurements over measurements ending with 0.5 inches, Sheppard's simple correction would give a wrong result. Instead some kind of probabilistic rounding must be assumed, where the integer-valued measurements correspond to the even lattice points and the measurements ending with 0.5 to the odd lattice points of our general rounding model. We here consider only the possibility of either asymmetric rounding or mixture rounding. In order to decide which one to use, we rely on the criterion (34). First we compute the proportion of integer measurements, which is 0.8230, and this is the value of $r$ in asymmetric rounding or of $m$ in mixture rounding due to (24). The fourth central moment of the rounded data is computed as $\mu_4^* = 91.5782$ and we thus have

$$\mu_4^* - 3\mu_2^{*2} = -4.0891.$$

The right side of (34) is $-0.00833$ for mixture rounding and $-0.00188$ for asymmetric rounding. Both values are quite far away from the value of $\mu_4^* - 3\mu_2^{*2}$. This may have several reasons: sampling errors, departure from normality, a more complex rounding scheme. But since the value for mixture rounding comes closer to the value of $\mu_4^* - 3\mu_2^{*2}$ we decide to use mixture rounding. The corrected variance, corrected with $h^2/3 = 0.0833$, then is:

$$\sigma_{cm}^2 = 5.564.$$

(For asymmetric rounding with $r = 0.8230$ the corrected variance would be $\sigma_{ca}^2 = 5.600$.)

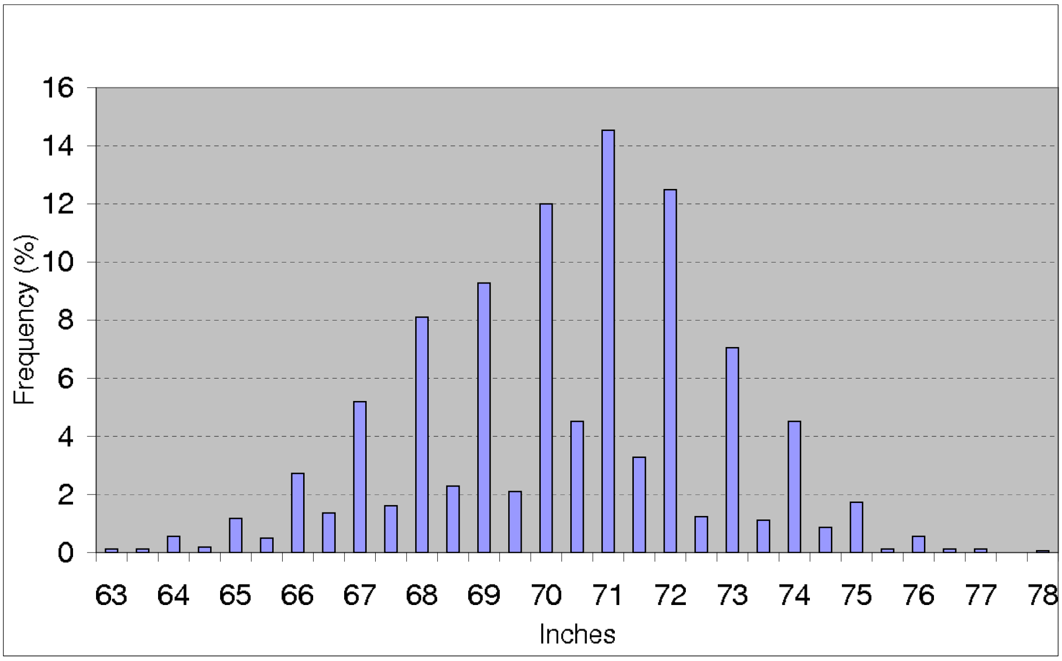


Figure 1: Height of 18-year-old students in the Citadel Military Academy, measured 1920-1940

# 7 Conclusion

Sheppard's correction for determining the variance of a random variable $X$ which is only available as rounded values has been developed in the literature for simple rounding only. But there are many other forms of more complicated rounding procedures including asymmetric rounding, by which some numbers (such as even numbers) are preferred to others (e.g., odd numbers). Probabilistic rounding is a convenient way to summarize all these various rounding procedures.

A generalized Sheppard correction formula has been derived for probabilistic rounding, which works under the same conditions as the formula for simple rounding: the distribution of $X$ should be sufficiently smooth and the interval between neighboring rounding points should not be too large.

Generalized Sheppard's correction also serves to correct regression estimates when these are computed from rounded data.

Although no explicit correction formulas for higher moments are given, these can be derived from general formulas of higher moments presented in the paper.

The generalized Sheppard's correction depends on the form of the rounding profile function used to perform probabilistic rounding. An attempt is made to determine this form in one particular case from the rounded data, which are the only data available to the statistician.

In this paper, we focused on a particular form of asymmetric rounding, where only even and odd rounding points were involved. More generally, one could have different preferences for rounding to even or odd points and in addition to points ending with the digit 5 and still other preferences for rounding to points ending with the digit 0. All the different preferences could be modeled probabilistically with different rounding profile functions. Such a scenario of multiple asymmetric rounding could be on the agenda for future research.

# Appendix

## A.1 Half moments of a symmetric rounding profile

The half moments of $q(u)$ of (29) are given by

$$\begin{aligned}
Q_k &= \int_0^1 u^k q(u) du \\
&= \int_0^{1/2} u^k (1 - 2^{\alpha-1} u^\alpha) du + \int_{1/2}^1 2^{\alpha-1} u^k (1-u)^\alpha du \\
&= \int_0^{1/2} u^k du - 2^{\alpha-1} \int_0^{1/2} u^{\alpha+k} du + 2^{\alpha-1} \int_0^{1/2} (1-u)^k u^\alpha du \\
&= \frac{1}{2^{k+1}(k+1)} - \frac{1}{2^{k+2}(\alpha+k+1)} + 2^{\alpha-1} \int_0^{1/2} (1-u)^k u^\alpha du.
\end{aligned}$$

$$\begin{aligned}
Q_1 &= \frac{1}{2^2 \cdot 2} - \frac{1}{2^3(\alpha+2)} + 2^{\alpha-1} \int_0^{1/2} (u^\alpha - u^{\alpha+1}) du \\
&= \frac{1}{8} - \frac{1}{4(\alpha+2)} + \frac{1}{4(\alpha+1)} \\
&= \frac{1}{4} \left( \frac{1}{2} + \frac{1}{(\alpha+1)(\alpha+2)} \right).
\end{aligned}$$

$$\begin{aligned}
Q_2 &= \frac{1}{2^3 \cdot 3} - \frac{1}{2^4(\alpha+3)} + 2^{\alpha-1} \int_0^{1/2} (1 - 2u + u^2) u^\alpha du \\
&= \frac{1}{24} - \frac{1}{16(\alpha+3)} + \frac{1}{4(\alpha+1)} - \frac{1}{4(\alpha+2)} + \frac{1}{16(\alpha+3)} \\
&= \frac{1}{24} + \frac{1}{4(\alpha+1)} - \frac{1}{4(\alpha+2)}.
\end{aligned}$$

$$Q_3 = \frac{1}{2^4 \cdot 4} - \frac{1}{2^5(\alpha+4)} + 2^{\alpha-1} \int_0^{1/2} (1 - 3u + 3u^2 - u^3) u^\alpha du$$
$$= \frac{1}{2^4} \left( \frac{1}{4} - \frac{1}{2(\alpha+4)} + 2^{\alpha+3} \left[ \frac{1}{2^{\alpha+1}(\alpha+1)} - \frac{3}{2^{\alpha+2}(\alpha+2)} + \frac{3}{2^{\alpha+3}(\alpha+3)} - \frac{1}{2^{\alpha+4}(\alpha+4)} \right] \right).$$
$$= \frac{1}{2^4} \left( \frac{1}{4} + \frac{4}{\alpha+1} - \frac{6}{\alpha+2} + \frac{3}{\alpha+3} - \frac{1}{\alpha+4} \right).$$

From these half moments we derive

$$Q = \frac{1}{3} - Q_0 + 2Q_1 = \frac{1}{12} + \frac{1}{2(\alpha+1)} - \frac{1}{2(\alpha+2)}$$
$$Q' = \frac{1}{5} - Q_0 + 4Q_1 - 6Q_2 + 4Q_3$$
$$= \frac{1}{80} + \frac{1}{2(\alpha+1)} - \frac{1}{\alpha+2} + \frac{3}{4(\alpha+3)} - \frac{1}{4(\alpha+4)}$$
$$= \frac{1}{4} \left( \frac{1}{20} + \frac{2}{\alpha+1} - \frac{4}{\alpha+2} + \frac{3}{\alpha+3} - \frac{1}{\alpha+4} \right).$$

## A.2 Moments with asymmetric and mixture rounding

The profile functions $q$ for asymmetric and for mixture rounding are given in the paragraph following (13). As before, assume $X \sim N(\mu, \sigma^2)$. Then again (33) is valid (it is valid for any profile function $q$).

Now, for asymmetric rounding,

$$Q_k = \int_0^r u^k du = \frac{r^{k+1}}{k+1}$$

and, by (22) and (23),

$$Q = \frac{1}{3} - r + r^2 = \frac{1}{3}[r^3 + (1-r)^3]$$
$$Q' = \frac{1}{5} - r + 2r^2 - 2r^3 + r^4 = \frac{1}{5}[r^5 + (1-r)^5],$$

and thus by (33),

$$\mu_4^* - 3\mu_2^{*2} \approx \left(\frac{1}{5}[r^5 + (1-r)^5] - \frac{1}{3}[r^3 + (1-r)^3]^2\right) h^4.$$

For mixture rounding,

$$Q_k = \int_0^1 mu^k du = \frac{m}{k+1}$$

and

$$Q = \frac{1}{3}$$

$$Q' = \frac{1}{5}.$$

It follows from (33) that

$$\mu_4^* - 3\mu_2^* \approx -\frac{2}{15}h^4.$$

# 8 References


1. Coclanis, P. and Komlos, J., (1995). Nutrition and economic development in post-reconstruction South Carolina: an anthropometric approach. *Social Science History* **19** 91-116.

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

3. Heitjan, D, F. and Rubin, D. B., (1991). Ignorability and coarse data. *Annals of Statistics* **19** 2244 - 2253.

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

5. Komlos, J. (1999). On the nature of the Malthusian threat in the eighteenth century. *Economic History Review* **52** 730-48.

6. M'Raïhi, D., Naccache, D., and Tunstall, M. (2001). Asymmetric Currency Rounding. In: Y. Frankel (ed.): *Financial Cryptography.* Lecture Notes in Computer Science **1962**. Springer, Berlin, 192-201.

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

8. Torelli, N. and Trivellato, H. (1993). Modelling inaccuracies in job-search duration data. *Journal of Econometrics* **59** 187-211.

9. Wang, H. and Heitjan, D. F., (2008). Modeling heaping in self-reported cigarette counts. *Statistics in Medicine* **27** 3789-3804.

10. Wright, D.E. and Bray, Isabelle (2003). A mixture model for rounded data. *The Statistician* **52** 3-13.