Effects of Rounding Continuous Data Using Specific Rules

Jay J. Kim, Lawrence H. Cox, Joe Fred Gonzalez, Jr., and Myron J. Katzoff, National Center for Health Statistics
Joe Fred Gonzalez, Jr., NCHS, 3311 Toledo Rd. Room 3121, Hyattsville, MD 20782

KEY WORDS: Rounding, Integer, Variance, Uniform Distribution

Summary

Data such as income are frequently rounded. Rounding may be done to protect the confidentiality of records in a file, to enhance the readability of the data, or to simplify the data values under the notion that the digits subject to rounding are inconsequential. The rounding may not have any effect on the bias of an estimator, but it may have a large impact on variance. Integers can be expressed as X = qB + r, where q is the quotient, B is the base, and r is the remainder. B is a constant, but q and r are random variables. We will investigate four rules for rounding "r" above and observe the effects of rounding on bias and variance. We will assume a uniform distribution for r, but no specific distributional assumption will be made for "q." When q =0, we will show that the variance of the data after rounding is three times the variance before rounding. As the variance of q gets larger, the effect of rounding on the variance decreases.

1. Introduction

Data are often rounded. The purpose for rounding can be to protect confidentiality of records in a file, to enhance readability of the data, or to simplify the presentation of data. Previous work on rounding includes Nargundkar and Saveland (1972), Fellegi (1975), Cox and Ernst (1982), and Cox (1987).

Integers can be expressed as X = q B + r, where q is the quotient, B is the base, and r is the remainder. B is a constant, but q and r are random variables. In this paper, four rounding rules will be considered for rounding "r." The four rounding rules are summarized as follows:

a.) *Conventional Rounding Rule*: For example, suppose B = 10, then r = 0, 1, 2, . . . , 9. In this situation, according to the conventional rounding, any r greater than or equal to 5 is rounded up to B. Otherwise, r is rounded down to zero (0).

b.) *Sum-Unbiased Conventional Rounding Rule*: This rule is the same as the conventional rounding rule, except when r is 5. According to this rule, when r = 5, it is rounded up to B or rounded down to zero (0) with probability ½.

c.) *Zero-Restricted 50/50 Rule*: Any number other than zero is rounded up or down with probability ½.

d.) *Unbiased Rounding Rule*: This rule is discussed in Cox (1987) and is also referred to as unbiased controlled rounding. According to this rule, a number r is rounded up with probability r/B and rounded down with probability (B-r)/B.

2. Conditional Mean and Variance of X

The variance of X = qB + r can be evaluated in two ways:

a.) $V(X) = E(X^2) - [E(X)]^2$

b.) $V(X) = V[E(X|q)] + E[V(X|q)]$.

NOTES: The second approach (b) involves less labor, and is adopted here.
1.) Since $E(X|q)=E(qB + r|q)=qB+E(r)$, $E(X|q)$ treats only r as a random variable and becomes a constant. Thus, $V[E(X|q)]=V[qB+E(r)]$ and $V[E(X|q)]$ captures the variance due to the variability of q.
2.) Since $V(X|q)= V(qB+ r|q)=V(r)$, $V(X|q)$ treats only r as a random variable and thus $E[V(X|q)]=E[V(r)]$ captures the variance component due to the variability of r. Therefore, to find E(X) and V(X), we must first find the mean and variance of r which will be derived in the following section.

2.1 Mean and Variance of r or Conditional Mean and Variance of X

We assume that r is uniformly distributed for the derivation of its mean and variance.

2.1.1 Mean and Variance of r Unrounded

Under the assumption that r can take on values 0, 1, 2, . . . , B-1 with uniform probability,

$$E(r) = \sum_{r=0}^{B-1} r\,P(r) = \frac{B-1}{2} \qquad (1)$$

$$E(r^2) = \sum_{r=0}^{B-1} r^2\,P(r) = \frac{(B-1)(2B-1)}{6}$$

and thus

$$V(r) = \frac{B^2-1}{12}. \qquad (2)$$

Let R(r) be the rounded number of r. The mean and variance of the rounded numbers R(r) will be compared with those of the unrounded numbers (r) shown above.

2.1.2 Mean and Variance of R(r) by Conventional Rounding Approach

The conventional rounding works as follows. When B is even, we round r up to B, if it is greater than or equal to $\frac{B}{2}$; otherwise, we round r down to 0. However, when B is odd, we will consider the rounding rule which rounds r up to B, when $r \geq \frac{B+1}{2}$, or rounds it down to 0.

Case 1. B is an Even Integer

Again, let R(r) be the rounded number of r.

$$E[R(r)] = \frac{B}{2} \qquad (3)$$

Comparing the expression in equation (3) with that in equation (1), we see that they differ by 1/2. The rounded data overestimate the mean.
It can be shown that

$$E[R^2(r)] = \frac{B^2}{2}$$

and thus,

$$V[R(r)] = \frac{B^2}{4}. \qquad (4)$$

Note that the expression in equation (4) is approximately three times the variance of r in equation (2) where r is not rounded.

Case 2. B is an Odd Integer

$$E[R(r)] = \frac{B\text{-}1}{2} \qquad (5)$$

The expected value of R(r) above is exactly the same as that in equation (1) which is for unrounded r. It can be shown that

$$E[R^2(r)] = \frac{B(B-1)}{2}$$

and thus

$$V[R(r)] = \frac{B^2-1}{4}. \qquad (6)$$

The variance of the rounded numbers in equation (6) is exactly three times the variance of the unrounded r in equation (2).

2.1.3 Mean and Variance of r Rounded by Sum-Unbiased Rounding Rule

This rule is the same as the conventional rounding rule, except that it allows for rounding r = B/2 up to B and down to 0, each with probability ½. It can be shown that

$$P[R(r) = B] = \frac{B\text{-}1}{2B}$$

and

$$P[R(r) = 0] = \frac{B+1}{2B}.$$

Thus,

$$E[R(r)] = \frac{B\text{-}1}{2} \qquad (7)$$

which is the same expectation as for the unrounded r in equation (1). It can be shown that

$$E[R^2(r)] = \frac{B(B-1)}{2}$$

and thus

$$V[R(r)] = \frac{B^2-1}{4}. \qquad (8)$$

The variance in equation (8) is exactly three times that for the unrounded data in equation (1).

2.1.4 Mean and Variance of r Rounded by Zero-Restricted 50/50 Rounding Rule

Except for zero (0), all numbers are rounded up or down with probability ½. Of course, zero (0) remains 0 throughout the rounding. Thus,

$$P[R(r) = B] = \frac{B-1}{2B}$$

and

$$P[R(r) = 0] = \frac{B+1}{2B}.$$

These probabilities are the same as those observed with the Sum-Unbiased Rounding Rule. As the probabilities of the data values becoming B or 0 are the same, the mean would be the same as those for the Sum-Unbiased Rounding Rule, or for the unrounded data, but the variance is three times that of the unrounded data.

2.1.5 Mean and Variance of r Rounded by Cox's Unbiased Rounding Rule

According to the Unbiased Rounding Rule, r is rounded up with probability r/B and rounded down with probability (B-r)/B. Thus,

$$P(r) = \frac{1}{B} \text{ and } P[R(r) = B \mid r] = \frac{r}{B}, \text{ for } r \geq 1,$$

$$P[R(r)=B] = \sum_{r=1}^{B-1} P(r)\, P[R(r) = B \mid r] = \sum_{r=1}^{B-1} \frac{1}{B}\frac{r}{B} = \frac{B-1}{2B},$$

and

$$P[R(r)=0] = \sum_{r=0}^{B-1} P(r)\, P[R(r) = 0 \mid r] = \sum_{r=0}^{B-1} \frac{1}{B}\frac{B-r}{B} = \frac{B+1}{2B}.$$

Since the above probabilities are the same as those observed with the Sum-Unbiased Rounding Rule, the mean and variance of the data rounded by the Unbiased Rounding Rule are the same as those for the Sum-Unbiased Rounding Rule.

Theorem 1. Suppose we have uniformly distributed data 0, 1, 2, . . . , B-1. Then, the variance of the rounded data is three times that of the unrounded data, if the numbers are rounded by the Conventional Rounding with B odd, Sum-Unbiased Rounding, 50/50 Rounding, and the Unbiased Rounding Rule.

Proof. Compare equation (2) with equations (6) and (8) and see the paragraphs right after P[R(r)=0] in sections 2.1.4 and 2.1.5.

If we use the conventional rounding rule with even B, the variance of the rounded number is greater than three times that of the unrounded numbers.

Theorem 2. Suppose we have the same data as used in Theorem 1. The mean of the rounded data rounded by the Conventional Rounding with B odd, Sum-Unbiased Rounding, 50/50 Rounding, and the Unbiased Rounding Rule is the same as that of the unrounded.

Proof. Compare equation (1) with equations (5) and (7) and the Conventional Rounding with odd B, Sum-Unbiased Rounding, 50/50 Rounding, and the Unbiased Rounding Rule.

The conventional rounding rule when B is even provides a mean which is greater than that of the unrounded data by ½ .

In short, the mean of the rounded r is the same as that of the unrounded data, regardless of which rounding rule we use, except for the conventional rounding rule when B is even. The variance of the rounded data is three times larger than that of the unrounded data irrespective of rounding rules, except the conventional rounding rule when B is even. For the latter, the variance of the rounded data is slightly greater than three times that of the unrounded.

3. (Unconditional) Mean and Variance of X

Using V(X) = V[E(X |q)] + E[V(X |q)], we have the following (unconditional) mean and variance of X.

3.1 Mean and Variance of X not Rounded

It can be shown that

$$E(X)=E[E(X \mid q)] = E(qB + \frac{B-1}{2})$$

$$= BE(q) + \frac{B-1}{2} \qquad (9)$$

and

$$V[E(X|q)] = V(qB + \frac{B-1}{2}) = B^2 V(q), \qquad (10)$$

and

$$E[V(X|q)] = E(\frac{B^2-1}{12}) = \frac{B^2-1}{12}.$$

Thus, $$V(X) = B^2 V(q) + \frac{B^2-1}{12}. \qquad (11)$$

3.2 Mean and Variance of X Rounded by Conventional Approach

Note that R(X) = qB + R(r).

Case 1. B is an Even Integer

$$E[R(X)] = E\{E[R(X)|q]\} = E(qB + \frac{B}{2})$$

$$= BE(q) + \frac{B}{2}, \qquad (12)$$

$$V\{E[R(X)|q]\} = V(qB + \frac{B}{2}) = B^2 V(q),$$

and

$$E\{V[R(X)|q]\} = E(\frac{B^2}{4}) = \frac{B^2}{4}.$$

Thus, $$V[R(X)] = B^2 V(q) + \frac{B^2}{4}. \qquad (13)$$

Case 2. B is an Odd Integer

$$E[R(X)] = E\{E[R(X)|q)]\}$$

$$= E(qB + \frac{B-1}{2})$$

$$= BE(q) + \frac{B-1}{2}, \qquad (14)$$

$$V\{E[R(X)|q]\} = V(qB + \frac{B\text{-}1}{2}) = B^2 V(q),$$

and

$$E\{V[R(X)|q]\} = E(\frac{B^2-1}{4}) = \frac{B^2-1}{4}.$$

Thus,

$$V[R(X)] = B^2 V(q) + \frac{B^2-1}{4}. \qquad (15)$$

3.3 Mean and Variance of R(X) by Sum-Unbiased Rounding Rule, Zero-Restricted 50/50 Rounding Rule, or the Unbiased Rounding Rule

All three rounding rules have the same conditional mean and variance as those of the conventional rounding rule when B is odd. Thus the mean and variance of R(X) are the same as those given in equations (14) and (15).

3.3.1 How Much Precision Can Be Lost Due to Rounding

Ehrenberg (1981) quoted a writer, who noted that "the approximate loss in accuracy when rounding a certain number to two digits was no more than 3.41 percent (p 70)." Can it be true? As shown before, the unconditional variance of x is the aggregate of the variance of qB and the variance of r. Recall that we assumed a uniform distribution for r. The variance of qB can vary depending on the distribution of q. The variance of r, however, remains the same, regardless of the distribution of q.

Suppose we have 100 observations which lie between 100 and 199 and we round the third digit by the conventional rounding rule. By changing the distributional form for q, in the following, we investigate the amount of increase in variance due to rounding.

Note that there is no difference between the rounded and unrounded numbers in the variance of qB. The variance of the unrounded r is 8.25, while the variance of the rounded r is 25.

If q is uniformly distributed, the variance of qB of the above mentioned data set is 2,185. The variance of the rounded X is 2.0 percent higher than that of the unrounded X.

When normally distributed random numbers with mean 0 and variance 1 are generated, which are multiplied by 3.1 and added by 10, the variance of qB is 395.017. The increase in variance due to rounding is 3.99 percent of the variance of the unrounded data.

We also generated log-normally distributed random numbers with mean 1 and variance 2.718. They were multiplied by .8, to which 10 was added. The resulting numbers have the variance of 235.71. The increase in variance due to rounding is 6.42 percent of the variance of the unrounded data. This and the findings concerning the normally distributed variable above prove that the claim Ehrenberg cited in his paper is untenable.

4. Conclusion

We have investigated the impact of rounding using four rounding rules on the mean and variance. Considering r only, or q = 0, the conventional rounding with even B behaves a little differently than the others. The rounding with B even is ½ greater than that of the unrounded and the variance of the rounded data rounded by the conventional rounding with B even is .25 higher than that of the rounded data by the rest of the rounding rules.

When q is non-zero, the expected value of x has a common term, BE(q), and the variance of x has also a common term, $B^2V(q)$, in all cases considered above. Thus, when we consider the difference in mean between the rounded and unrounded data, we will have the same observations as before when we considered the mean of x. The same can be said for the variance of x.

When q gets larger, the impact of rounding gets smaller. Ehrenberg cites an author who says that loss in accuracy when rounding a certain number to two digits is no more than 3.41 percent. However, we have found the cases where the loss could be more than that. In one case we observed 6.42 percent loss.

The q-stationary rounding rule seems to maintain the mean of the unrounded number. However, we have not investigated its impact on the variance. This will be investigated later.

5. References

Cox, L.H. (1987), "A Constructive Procedure for Unbiased Controlled Rounding," Journal of the American Statistical Association, 82, 520-524.

Cox, L.H. and Ernst, L. R. (1982), "Controlled Rounding," INFOR, 20, 423-432.

Ehrenberg, A.S.C. (1981), "The Problem of Numeracy," The American Statistician, 35, 67-71.

Fellegi, I.P., (1975), "Controlled Random Rounding," Survey Methodology, 1, 123-135.

Nargundkar, M.S., and Saveland, W. (1972), "Random Rounding to Prevent Statistical Disclosure," in Proceedings of the Social Statistics Section, American Statistical Association, 382-385.