WP. 31
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Geneva, Switzerland, 9-11 November 2005)

Topic (v): Confidentiality aspects of tabular data, frequency tables, etc.

EFFECTS OF ROUNDING ON DATA QUALITY

Invited Paper

Submitted by the U.S. National Center for Health Statistics, Center for Disease Control and Prevention, United States of America [1]

[1] Prepared by Jay J. Kim, Lawrence H. Cox, Myron Katzoff and Joe Fred Gonzalez, Jr.

Effects of Rounding on Data Quality

Jay J. Kim*, Lawrence H. Cox*, Myron Katzoff* and Joe Fred Gonzalez, Jr*

* U.S. National Center for Health Statistics, Center for Disease Control and Prevention, Hyattsville, MD 20782, USA. Contact: pzk3@cdc.gov

Abstract: Integer data such as frequency counts may be rounded to integer values for purposes including disclosure limitation. It may be necessary to round noninteger data to integer data (*base 1 rounding*) for various statistical purposes, e.g., rounding expected sample counts (noninteger) to actual sample counts (integer). We evaluate the effects of four methods of rounding data on data quality and utility in two ways: (1) bias and variance (increase in total mean squared error) and (2) effects on the underlying distribution of the data (as measured, e.g., by the distance measure which can be considered a proxy chi-square statistic). The four rounding methods are conventional rounding, modified conventional rounding, zero-restricted 50/50 rounding, and unbiased rounding.

1 Introduction

Data are often rounded. Sometimes it is necessary to round noninteger values to integer values for statistical purposes. For example, at the end of sample weighting, the fractions are rounded to integers, since the number of persons or establishments cannot be fractions. Also data is rounded to enhance readability of the data, to protect confidentiality of records in the file, or to keep the important digits only.

Integers can be expressed as $x = q_x B + r_x$, where q_x is the quotient, integer B is the *rounding base*, and r_x is the remainder. B is a constant, but q_x and r_x are random variables. When the subscript x is not needed, it will be ignored. Four rounding rules are considered for rounding the remainder *r*. Note we will use R(x) to denote the rounded number of x and subscript i can be added whenever needed. This implies that R(x) = qB + R(r). For concreteness, we illustrate the rules for B = 10, so that r = 0, 1, 2, . . ,9. Two important properties for evaluating and comparing rounding methods are as follows. *Unbiased rounding* satisfies: E[R(r) | r] = r. A weaker but still useful property is *sum-unbiasedness*: E[R(r)] = E[r].

The first rounding rule is *conventional rounding*: any r greater than or equal to B/2 = 5 is rounded up to B = 10; otherwise it is rounded down to zero. Conventional rounding is not unbiased but is sum-unbiased if and only if B is odd. The second rule is *modified conventional rounding*. This rule is the same as conventional rounding, except when r = B/2 (e.g., 5), r is rounded up to B = 10 or down to zero each with probability ½. Modified conventional rounding is sum-unbiased. The third is *zero-restricted 50/50 rounding*: r = 0 is rounded down and all nonzero r are rounded up or down with probabilities ½. It, too, is sum-unbiased. The last rule is *unbiased rounding* proposed by Nargundkar and Saveland. According to this rule, r

is rounded up with probability r/10 and down with probability 1-r/10. Consequently unbiased rounding is unbiased and therefore also sum-unbiased. These rules are easily restated for any positive integer rounding base B.

We evaluate the effects of these rounding methods on data quality and utility in two ways: (1) bias and variance (increase in total mean squared error) and (2) effects on the underlying distribution of the data as evaluated by a distance measure.

2 Bias and Variance of the Rounded and Unrounded Numbers

We consider various distributions for the data, but assume that the remainders r follow a discrete uniform distribution.

2.1 Mean and Variance of Unrounded Data x and Remainders r

Since r takes values 0, 1, 2, . . . , B-1 with uniform probability:

$$E(r) = \frac{B\text{-}1}{2} \tag{1}$$

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

Therefore,

$$E(x) = E\left[E(x \mid q)\right] = BE(q) + \frac{B-1}{2}. \tag{3}$$

In general,

$$V(x) = V\left[E(x \mid q)\right] + E\left[V(x \mid q)\right] \tag{4}$$

Formula (4) will be used for deriving variance formulas for all cases. We obtain:

$$V(x) = B^2 V(q) + \frac{B^2 - 1}{12}. \tag{5}$$

2.2 Mean and Variance of R(x) for Conventional Rounding

If B is even, under conventional rounding, r is rounded up to B if r is greater than or equal to B/2; otherwise, it is rounded down to 0. If B is odd, we require: r rounds up to B if $r \geq \frac{B+1}{2}$, and rounds down to 0 otherwise.

Case 1. B is an Even Integer

$$E[R(r)] = B/2 \tag{6}$$

Comparing the expression in equation (6) with that in equation (1), we can see that they differ by 1/2. The rounded data overestimate the mean by ½, viz., the *absolute bias* is ½. The variance of the rounded data is:

$$V[R(r)] = B^2 / 4 . \tag{7}$$

Note that the expression in equation (7) is approximately three times the variance of r in equation (2) where r is not rounded.

Using equation (4), we have

$$V\left[R(x)\right] = B^2 V(q) + \frac{B^2}{4} . \tag{8}$$

Hence,

$$MSE\left[R(x)\right] = B^2 V(q) + \frac{B^2 + 1}{4} \tag{9}$$

Case 2. B is an Odd Integer

$$E\left[R(r)\right] = \frac{B\text{-}1}{2} \tag{10}$$

This expected value is exactly the same as in equation (1). Thus the rounded data provide a sum-unbiased estimator of the original data.

$$V[R(r)] = \frac{B^2 - 1}{4} \tag{11}$$

The variance of the rounded remainders in equation (11) is exactly three times that of the unrounded r in equation (2).

$$V\left[R(x)\right] = B^2 V(q) + \frac{B^2 - 1}{4} \tag{12}$$

Since R(x) provides an unbiased estimator, the MSE of R(x) is the same as the variance of R(x) above.

2.3 Mean and Variance of R(x) for Modified Conventional Rounding

This rule is the same as 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:

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

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

Thus,

$$E\left[R(r)\right]=\frac{B\text{-}1}{2} \tag{13}$$

which is the same as for the unrounded r in equation (1), and

$$V[R(r)]=\frac{B^2-1}{4}. \tag{14}$$

This variance is exactly three times that for the unrounded remainders in (2).

2.4 Mean and Variance of R(x) for Zero-Restricted 50/50 Rounding

Except for zero, all remainders r are rounded up or down with probability ½. Of course, zero remains zero after rounding. The probability the rounded remainder is B or 0 is the same as that observed with the modified conventional rounding. Hence this rounding rule has the same mean, variance, and mean square error as those of conventional rounding when B is odd and modified conventional rounding.

2.5 Mean and Variance of R(x) for Unbiased Rounding

According to Nargundkar and Saveland's unbiased rounding rule, r is rounded up with probability r/B and rounded down with probability (B-r)/B. Thus,

$$P(r)=\frac{1}{B},\ P\left[R(r)=B\,|\,r\right]=\frac{r}{B}\ \text{and}\ P\left[R(r)=0\,|\,r\right]=\frac{B-r}{B},\ \text{for r}\geq 1,$$

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

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

Since the above probabilities are the same as those observed with the modified conventional rounding, this rounding rule again has the same mean, variance, and mean square error as those of the conventional rounding rule when B is odd.

3 Distance Measure

The quality of the rounded data can be measured by the values of the distance measure of the rounding rules mentioned above. In comparing the rounded number with the original number, we can use the following measure for every number or cell subject to rounding:

$$U=\frac{[R(x)-x]^2}{x} \tag{15}$$

The numerator can be re-expressed as $[R(r_x)-r_x]^2$. Since $R(r_x)$ can be either B or 0, the above can be further re-expressed as $(d_xB-r_x)^2$, where d_x is an indicator variable: d_x = 0 means round down and d_x = 1 means round up. We assume U = 0, when x = 0. The conditional expected value of U over d_x is:

$$E_d(U|\ x)=\sum_{d_x=0}^{1}\left[\frac{(d_xB-r_x)^2}{x}|\,x\right]P(d_x) \tag{16}$$

d_x is the only variable in the above. When B = 10, with conventional rounding, $P(d_x=1)=1$ with x = 5, 6, . . ., 9. Otherwise, $P(d_x=0)=1$ with x = 0, 1, 2, 3, 4.

3.1 Conventional Rounding

The expected value of U can be expressed as

$$E_q\left[E_r\left\{E_d(U\mid x)\right\}\right]=\sum_{q_x}\sum_{r_x}\sum_{d_x}[\frac{(d_xB-r_x)^2}{x}|\,x]P(d_x)P(r_x)P(q_x) \tag{17}$$

For conventional rounding, assuming B even,

$$U_1=E_r\left\{E_d(U\mid x)\,|\,q\right\}=\sum_{r_x}\sum_{d_x=0}^{1}[\frac{(d_xB-r_x)^2}{q_{xB}+r_x}|\,q_x]P(d_x)P(r_x), \tag{18}$$

which is

$$[\sum_{r_x=1}^{B/2-1}\frac{r_x^2}{q_xB+r_x}+\sum_{r_x=B/2}^{B-1}\frac{(B-r_x)^2}{q_xB+r_x}]\frac{1}{B} \tag{19}$$

By separating the term with q_x = 0 from those with $q_x \geq 1$, we have

$$\begin{aligned}U_1=&[\sum_{r_x=1}^{B/2-1}r_x+\sum_{r_x=B/2}^{B-1}\frac{(B-r_x)^2}{r_x}]\frac{1}{B}\\&+[\sum_{r_x=1}^{B/2-1}\frac{r_x^2}{q_xB+r_x}+\sum_{r_x=B/2}^{B-1}\frac{(B-r_x)^2}{q_xB+r_x}]\frac{1}{B}\end{aligned} \tag{20}$$

U_1 is bounded as follows.

$$U_1\leq[\sum_{r_x=1}^{B/2-1}r_x+\sum_{r_x=B/2}^{B-1}\frac{(B-r_x)^2}{r_x}]\frac{1}{B}$$

$$+ [\sum_{r_x=1}^{B/2-1} \frac{r_x^2}{q_x B} + \sum_{r_x=B/2}^{B-1} \frac{(B-r_x)^2}{q_x B}] \frac{1}{B}. \quad (21)$$

The expected value of the second term of the above equation over q_x reduces to

$$E_{q_x}[\frac{1}{q_x}][\sum_{r_x=1}^{B/2-1} r_x^2 + \sum_{r_x=B/2}^{B-1} (B-r_x)^2] \frac{1}{B^2} \quad (22)$$

The product of the second and third factors above, i.e., excluding the expected q-reciprocal in equation (22), is denoted by V as seen below.

$$V = [\sum_{r_x=1}^{B/2-1} r_x^2 + \sum_{r_x=B/2}^{B-1} (B-r_x)^2] \frac{1}{B^2} \quad (23)$$

After some algebra, the above equation (23) reduces to

$$V = \frac{B^2+2}{12B} \quad (24)$$

The first term of U_1 in equation (21) does not involve q_x, hence there is no need taking expectation over q_x. An upper and lower bounds for the sum of a harmonic series whose last integer is n are: $\ln(n+1) < H_n \le 1+\ln(n)$. Using the above upper bound, we obtain the upper bound for the first term of U_1 as:

$$B \ln[\frac{2(B-1)}{B-2}] - \frac{B+1}{2}$$

Of course, $E_q(\frac{1}{q_x})$ varies depending on the distribution of q_x. We examine the expected q-reciprocal for various distributions in a separate section.

3.2 Modified Conventional Rounding Rule

For modified conventional rounding, V in equation (23) is:

$$V = [\sum_{r_x=1}^{B/2-1} r_x^2 + \frac{B^2}{4} + \sum_{r_x=B/2+1}^{B-1} (B-r_x)^2] \frac{1}{B^2} \quad (25)$$

This reduces to the same expression as the one in equation (24).

The first term of U_1 in equation (21) for this rounding rule is the same as that for the conventional rounding rule.

3.3 Zero-restricted 50/50 Rounding Rule

For the zero-restricted 50/50 rounding rule, we have:

$$V = \sum_{r_x=1}^{B-1} [r_x^2 + (B - r_x)^2] \frac{1}{2B^2} \tag{26}$$

$$V = \frac{2B^2 - 3B + 1}{6B} \tag{27}$$

The first term of U_1 in equation (21) is

$$\frac{B}{2} \ln (B-1) + \frac{1}{2}.$$

3.4 Unbiased Rounding

For the unbiased rounding, the expectation of the numerator of U over d_x is:

$$\sum_{d_x} (d_x B - r_x)^2 P(d_x)$$

$$= (B - r_x)^2 \frac{r_x}{B} + r_x^2 (\frac{B - r_x}{B}).$$

Thus

$$V = \sum_{r_x=1}^{B-1} [(B - r_x)^2 \frac{r_x}{B} + r_x^2 (\frac{B - r_x}{B})] \frac{1}{B^2} \tag{28}$$

This reduces to

$$V = \frac{B^2 - 1}{6B} \tag{29}$$

The first term of U_1 in equation (21) is $\frac{B-1}{2}$.

Among the four rounding rules, the difference comes from V and the first term of U_1. The expected q-reciprocal remains the same over all the rules. Comparing (24), (27) and (29), it can be observed that conventional rounding has the minimal V between rounded and unrounded data. This is just over half of that for unbiased rounding. The zero-restricted 50/50 rounding rule has the highest expected distance. It is also observed that modified conventional rounding has the same expected distance as that for conventional rounding when B is even. Conventional rounding has the minimal value of the first term of U_1. Unbiased rounding has around 2.58 times higher value.

4 Expected Value of 1/q or An Upper Bound

For three distributions, we derived the expected value formula for 1/q, or the upper

bound for the expected value of 1/q when it is not possible to derive its expected value formula. They are shown below.

4.1 E(1/q) for the Lognormal Distribution

Let the normal density function g(y) be

$$g(y \mid m, s^2) = \left(s\sqrt{2p}\right)^{-1} e^{-\frac{1}{2}\left(\frac{y-m}{s}\right)^2}, \qquad -\infty < y < \infty.$$

The lognormal distribution f(x) has the following form:

$$f(x \mid m', s'^2) = (s'\sqrt{2p}\,x)^{-1} e^{-\frac{1}{2}\left(\frac{\ln x - m'}{s'}\right)^2}, \qquad 0 < x < \infty,$$

where m' and s' are the mean and variance of ln x.

Let

$$c = \int_1^{\infty} f(x \mid m', s'^2)\,dx.$$

Then

$$E\left(\frac{1}{q} \mid q \ge 1, m', s'^2\right) = \frac{1}{c}\int_1^{\infty} \frac{1}{q} f(q)\,dq$$

$$= \frac{1}{c}\frac{1}{s'\sqrt{2p}\,x}\int_1^{\infty} \frac{1}{x} e^{-\frac{1}{2}\left(\frac{\ln x - m'}{s'}\right)^2} dx \qquad (30)$$

To integrate by parts, set $n = \dfrac{\ln x - m'}{s'}$, so that $dn = \dfrac{1}{x s'}dx$ and $x = e^{s'n + m'}$.

Equation (30) becomes

$$E\left(\frac{1}{q} \mid q \ge 1, m', s'^2\right) = \frac{1}{c} e^{\frac{1}{2}s'^2 - m'}\left[1 - \Phi\left(s' - \frac{m'}{s'}\right)\right] \qquad (31)$$

For example, when m'=.25 and s'=.25, the probability of the truncated lognormal distribution is .52283.

4.2 E(1/q) for the Pareto Distribution

The Pareto distribution is sometimes used to fit the size of firms, personal incomes and stock price fluctuations, etc. The Pareto distribution of the second kind has the following form:

$$f(q)=\frac{ak^a}{q^{a+1}}, \quad a>0,\ q\geq k>0.$$

For the above distribution, we derive E(1/q).

$$E(\frac{1}{q})=\int_k^\infty \frac{1}{q}\frac{ak^a}{q^{a+1}}dq$$

$$= \frac{a\,k^a\,q^{-a-1}}{-a-1}\Big|_k^\infty = \frac{a\,k^{-1}}{a+1} \qquad (32)$$

where k above is the minimum value of q.

The method of moment's estimator of a, a^* is

$$a^*=\frac{n\bar{x}-x_1'}{n(\bar{x}-x_1')},$$

where x_1' is the smallest sample value.

The maximum likelihood estimator of a is

$$\hat{a}=n[\sum_{j=1}^{n}\log(x_j/\hat{k})]^{-1},$$

where $\hat{k}=\min_i x_i$, or the smallest sample value of x.

4.3 Upper Limit for E(1/q) for the Multinomial Distribution

Let $q_1, q_2, q_3, \ldots$ follow multinomial distribution. That is,

$$f(q_1,q_2,\ldots q_k \mid p_1,p_2,\ldots p_k)=\frac{n!}{\prod_{i=1}^{k} q_i!}\prod_{i=1}^{k} p_i^{q_i}, \quad q_i = 0, 1, 2, \ldots$$

Note

$$E(\frac{1}{q_i}) \leq E(\frac{1}{q_i+1}) + E[\frac{3}{(q_i+1)(q_i+2)}] \quad \text{for all i.} \qquad (33)$$

The expected value of interest is

$$E(\frac{1}{q_1 q_2 \ldots q_k})=\sum\sum \ldots \sum \frac{1}{q_1 q_2 \ldots q_k}\frac{n!}{\prod_{i=1}^{k} q_i!}\prod_{i=1}^{k} p_i^{q_i}.$$

The upper limit of the above expected value can be derived using equation (33). Let the size of the category i be n_i, $\sum_{i=1}^{k} n_i = n$ and n' = $n-\sum_{i=1}^{j-1} n_i$, j = 2, 3, . Using the multinomial distribution truncated at 0 for all q's, we have

$$E\left(\frac{1}{q_1 q_2 \cdot \cdot q_k}\right) \le \prod_{i=1}^{k}\left[\frac{5(n+1)(n+2)p_i^2 r_i^{\sum_{j=1}^{i-1} n_j} + 2(n+2)p_i + 6}{2(n+1)(n+2)\,p_i^2 (1 - r_i^{\sum_{j=1}^{i-1} n_j})}\right] \quad (34)$$

5 Concluding Comments

It is often necessary to transform statistical data, both count and continuous data, to integer values. Various methods of rounding and in some applications various choices for the rounding base B typically are available. The question becomes: which method and/or base is expected to perform best in terms of data quality and preserving distributional properties of original data and, quantitatively, what is the expected distortion due to the rounding? This paper provides a preliminary analysis towards answering these questions.

References

Grab, E.L & Savage, I.R. (1954), Tables of the Expected Value of 1/X for Positive Bernoulli and Poisson Variables, *Journal of the American Statistical Association* **49**, 169-177.

N.L. Johnson & S. Kotz (1969). **Distributions in Statistics, Discrete Distributions**, Boston: Houghton Mifflin Company.

N.L. Johnson & S. Kotz (1970). **Distributions in Statistics, Continuous Univariate Distributions-1**, New York: John Wiley and Sons, Inc.

Kim, Jay J., Cox, L.H., Gonzalez, J.F. & Katzoff, M.J. (2004), Effects of Rounding Continuous Data Using Rounding Rules, *Proceedings of the American Statistical Association, Survey Research Methods Section*, Alexandria, VA, 3803-3807 (available on CD).

Vasek Chvatal. Harmonic Numbers, Natural Logarithm and the Euler-Mascheroni Constant. See www.cs.rutgers.edu/~chvatal/notes/harmonic.html.