CONTROLLED RANDOM ROUNDING

I.P. Fellegi
Assistant Chief Statistician, Statistical Services Field

Random rounding is a technique to ensure confidentiality of aggregate statistics. By randomly rounding all the components of a total, independently, together with the random rounding of the total itself, substantial discrepancies may arise when aggregating the published data. This paper presents a procedure which avoids substantial discrepancies while still protecting the concept of confidentiality.

1. INTRODUCTION

Random rounding is a technique to prevent statistical disclosure, both direct and residual. It consists of rounding published (or otherwise released) statistical aggregates to a multiple of some chosen base number -- but carrying out the rounding through a random mechanism which ensures that each randomly rounded published aggregate has as its expected value the corresponding unrounded (and, of course, unpublished) aggregate. This ensures that the rounding process is unbiased. For a more detailed description of the technique, the reader is referred to [2] and [3].

The particular problem addressed in this note can be summarized as follows. Given that each of a number of statistical aggregates has to be random rounded, can this be done in such a way that the sum of the individually random rounded numbers is equivalent to the random rounding of the sum of the unrounded numbers, i.e. if e_i (i=1, 2, ..., n) are unrounded numbers, and e_i^* are the corresponding random rounded numbers, can we carry out the random rounding in such a way that

$$\sum_{i=1}^{n} e_i^*$$

is equivalent to (i.e. has the same distribution as)

$$\left(\sum_{i=1}^{n} e_i\right)^*.$$

The question so stated grew out of a very concrete problem. Several countries have adopted the practice of releasing so-called summary tapes after their decennial or quinquennial population censuses. These tapes contain tabulations (aggregates) at the level of very small geographic areas, usually corresponding to the work assignment of one census enumerator. These small area data are used by research personnel as "building blocks" to aggregate data for their respective areas of interest. At least two countries, the United Kingdom [4] and Canada, have adopted the practice of introducing a small random disturbance into these small area level aggregates in order to safeguard against statistical disclosure, and the Bureau of the Census is at least contemplating a similar procedure for the 1980 Census [1].

Even though the level of such random errors is small, when the random rounded numbers are aggregated, their variances aggregate also. When several small area tabulations are aggregated in order to obtain a tabulation for a large area, say a municipality, the variance may become quite large (although, of course, the relative variance declines). So when users compare their own tabulations prepared from the summary tapes for, say, a municipality, with the corresponding tables actually published at the level of a municipality, substantial discrepancies may be observed. The reason is that the published municipality-level tabulations were random rounded directly, while the tabulations prepared by users from the summary tapes were random rounded at the level of the component small area level.

The following procedure ensures that when the small-area tabulations are random rounded, the cumulative impact of such errors is controlled at the level of some predefined higher level geographical areas. Of course, for other than the predefined larger areas the variance due to random rounding is probably unaffected.

An attempt to contain the cumulative impact of random errors is given in [4], but only for a situation where the amount of random error is +1, -1 or 0.

2. THE CONSTRAINTS

If the base of random rounding is an integer b (be was equal to 5 in the 1971 Census in Canada), suppose that a table entry is e. We compute the residual r of e after division by b:

$$e = k \times b + r \qquad 0 \leq r < b$$

It is this residual which is "rounded" at random to either 0 or b. Let the probability of rounding up to b be p, the probability of rounding down to zero being (1-p). The randomly rounded e, e* can be written as

$$e^* = k \times b + r^*$$

where r* = b with probability p and it is equal to 0 with probability (1-p). The expected value of e* can be written as

$$E(e^*) = k \times b + [p \times b + (1-p) \times 0]$$

If we want e* to be unbiased, we must set

$$E(e^*) = e$$

i.e.

$$p \times b = r \quad \text{or} \quad p = \frac{r}{b}$$

This is the first constraint we impose on a desirable random rounding procedure. The argument above also shows that if e* is to be an unbiased estimate of e, the only way e can be altered to become a multiple of b while changing it at random by an amount which is less than b in absolute value, is by a random rounding process with probabilities as shown above.

If we want to preserve the unbiasedness of random rounding, this constraint must, therefore, not be violated.

Next, suppose there are a series of n tabulation cells (each corresponding to one small area aggregate in a municipality) which are to be rounded. Denote these by e_i $(i = 1, 2, \ldots, n)$.

Let

$$e_i = k_i x b + r_i \qquad 0 \leq r_i < b$$

and the randomly rounded corresponding value as

$$e_i^* = k_i x b + r_i^*$$

where $r_i^* = b$ with probability $p_i = r_i/b$ and is equal to 0 with probability $1 - p_i$.

Their sum, e is

$$e = \sum_{i=1}^{n} e_i$$

which can be written as

$$e = k \times b + r \qquad 0 \leq r < b \tag{2.1}$$

and its rounded value is

$$e^* = k \times b + r^*$$

where r* is equal to b with probability of r/b and is zero otherwise. Ideally, one would like to have

$$e^* = \sum_{i=1}^{n} e_i^*$$

in the sense that $\Sigma\, e_i^*$ and e* assume the same values with the same probabilities. This is the second constraint we impose on a desirable random rounding procedure.

The procedure below satisfies both of these.

3. THE PROCEDURE

Consider the numbers r_i and cumulate them:

$$s_i = \sum_{j=1}^{i} r_j$$

$$S_n = \sum_{j=1}^{n} r_j$$

$$S_o = 0$$

Select a random integer between 1 and b, say R_1:

$$1 \leq R_1 \leq b$$

Consider S_1, S_2, S_3, ..., in order until

$$S_{i_1-1} < R_1 \leq S_{i_1}$$

Next let

$$R_2 = R_1 + b$$

and select i_2 so that

$$S_{i_2-1} < R_2 \leq S_{i_2}$$

Next let

$$R_3 = R_2 + b = R_1 + 2b$$

and select i_3 so that

$$S_{i_3-1} < R_3 \leq S_{i_3}$$

etc. Continue until the L-th step so defined that

$$R_L \leq S_n$$

but

$$R_{L+1} > S_n$$

Now round up the units so selected, down the others. In other words,

$$e_i^* = k_i x b + r_i^*$$

where

$$r_i^* = b \qquad \text{if} \qquad i = i_1, i_2, i_3, \ldots$$

$$= 0 \qquad \text{otherwise.}$$

The procedure is illustrated in Table 1.

4. PROOF THAT THE PROCEDURE SO DEFINED SATISFIES THE CONSTRAINTS

It is easy to verify, using arguments which are standard in selecting with probabilities proportional to a measure of size, that the probability

$$P(r_i^* = b) = r_i/b$$

so that the first constraint is satisfied.

As far as the second constraint is concerned, the following simple argument shows that it, too, is satisfied.

Since from (2.1)

$$e = \sum_{i=1}^{n} e_i = kb + r$$

and also

$$e = \sum_{i=1}^{n} e_i = \sum_{i=1}^{n} (k_i\, b + r_i) = b \sum_{i=1}^{n} k_i + \sum_{i=1}^{n} r_i$$

$$= b \sum_{i=1}^{n} k_i + S_n \qquad (4.1)$$

it follows that the integer remainder of S_n, when divided by b, must also be r. So we must have, for some integer m,

$$S_n = mb + r \qquad 0 \leq r < b \qquad (4.2)$$

So from (4.1) we obtain

$$e = \sum_{i=1}^{n} e_i = b \left(\sum_{i=1}^{n} k_i + m \right) + r$$

i.e.

$$k = \sum_{i=1}^{n} k_i + m$$

It immediately follows from (4.2) that the number of steps, L, required to complete the procedure is related to m, r and R_1 as follows:

$$\text{Prob } (L = m + 1) = \text{Prob } (1 \leq R_1 \leq r) = \frac{r}{b}$$

$$\text{Prob } (L = m) = \text{Prob } (r < R_1 \leq b) = 1 - \frac{r}{b}$$

Since

$$\sum_{i=1}^{n} e_i^* = b \sum_{i=1}^{n} k_i + \sum_{i=1}^{n} r_i^* = b \sum_{i=1}^{n} k_i + Lb$$

we have

$$\sum_{i=1}^{n} e_i^* = \begin{cases} b \sum_{i=1}^{n} k_i + (m+1)\, b & \text{with probability } \frac{r}{b} \\ b \sum_{i=1}^{n} k_i + mb & \text{with probability } 1 - \frac{r}{b} \end{cases}$$

$$= \begin{cases} kb + b & \text{with probability } \frac{r}{b} \\ kb & \text{with probability } 1 - \frac{r}{b} \end{cases} \qquad (4.3)$$

Also,

$$\left(\sum_{i=1}^{n} e_i\right)^* = kb + r^*$$

where

$$P(r^* = b) = \frac{r}{b}$$

$$P(r^* = 0) = 1 - \frac{r}{b}$$

so that

$$\left(\sum_{i=1}^{n} e_i\right)^* = \begin{cases} kb + b & \text{with probability } \frac{r}{b} \\ kb & \text{with probability } 1 - \frac{r}{b} \end{cases} \qquad (4.4)$$

Comparing (4.3) and (4.4) we obtain immediately that the random variables $\Sigma\, e_i^*$ and $(\Sigma\, e_i)^*$ have the same distribution.

Thus the net effect of the procedure on a predefined aggregate of randomly rounded individual numbers is equivalent to the random rounding of the aggregate itself.

It can also be shown quite readily that the same argument holds for the sum of any consecutive numbers e_t, e_{t+1}, ..., e_{t+5}. Thus controlled random rounding results in a desirable reduction of rounding variance not only for a predefined aggregate, but also for any user-defined area consisting of the union of consecutive "building block" areas.

Table 1: Example of Controlled Random Rounding

											Total
Unrounded E.A. total (e_i)	12	23	34	3	49	23	50	17	8	13	232
Unroundable "base" ($k_i b$)	10	20	30	0	45	20	50	15	5	10	205
Residual (r_i)	2	3	4	3	4	3	0	2	3	3	27
Cumulative Residual (S_i)	2	5	9	12	16	19	19	21	24	27	
$R_1 = 1$	*		*	*	*			*		*	
Rounded Residual (r_i^*)	5	0	5	5	5	0	0	5	0	5	30
Rounded E.A. total (e_i^*)	15	20	35	5	50	20	50	20	5	15	235
$R_1 = 2$	*		*	*		*			*	*	
Rounded Residual (r_i^*)	5	0	5	5	0	5	0	0	5	5	30
Rounded E.A. total (e_i^*)	15	20	35	5	45	25	50	15	10	15	235
$R_1 = 3$		*	*		*	*			*		
Rounded Residual (r_i^*)	0	5	5	0	5	5	0	0	5	0	25
Rounded E.A. total (e_i^*)	10	25	35	0	50	25	50	15	10	10	230
$R_1 = 4$		*	*		*	*			*		
Rounded Residual (r_i^*)	0	5	5	0	5	5	0	0	5	0	25
Rounded E.A. total (e_i^*)	10	25	35	0	50	25	50	15	10	10	230
$R_1 = 5$		*		*	*			*		*	
Rounded Residual (r_i^*)	0	5	0	5	5	0	0	5	0	5	25
Rounded E.A. total (e_i^*)	10	25	30	5	50	20	50	20	5	15	230
No. of times Rounded up	2	3	4	3	4	3	0	2	3	3	2
No. of times Rounded down	3	2	1	2	1	2	5	3	2	2	3

RESUME

L'arrondissement aléatoire est une technique qui vise à assurer la confidentialité des agrégats ou groupes de statistiques. En appliquant cette technique à tous les éléments d'un total, d'une part, et au total lui-même, d'autre part, des divergences importantes peuvent se produire au moment de regrouper les données publiées. La méthode décrite dans ce document permet d'éviter ces divergences tout en assurant la confidentialité des données.

REFERENCES

[1] Barabba, V.P. and Kaplan, D.C., "U.S. Census Bureau statistical techniques to prevent disclosure -- the right to privacy vs. the need to know". Paper presented at the 2nd meeting of the international Association of Survey Statisticians, Warsaw (1975).

[2] Fellegi, I.P. and Phillips, J.L., "Statisticians Confidentiality: Some Theory and Applications to Data Dissemination". Annals of Economic and Social Measurement, pp. 399-409, 1974.

[3] Nargundkar, M.S. and Saveland, W., "Random Rounding: A Means of Preventing Disclosure of Information about Individual Respondents in Aggregate Data". Proceedings of the Social Statistics Section of ASA, 1972.

[4] Newman, D., "Rounding and Error Injection of Preserving Confidentiality of Census Data". Paper presented at the 2nd meeting of the International Association of Survey Statisticians, Warsaw (1975).