RANDOM-ROUNDING TO PREVENT STATISTICAL DISCLOSURES

M.S. Nargundkar, Statistics Canada
Walt Saveland, Indian and Northern Affairs

This is an extract from our "Random-rounding: A means of preventing disclosure of information about individual respondents in aggregate data". Anyone interested in using or critically examining the random-rounding procedure should request a copy of the full report from either author. The procedure is being used in the production of nearly all population and housing tabulations of the 1971 Census of Canada. Nevertheless, the authors are personally responsible for the assertions and arguments presented herein. Mr. Saveland was with Statistics Canada during most of the developmental work on random-rounding.

In Canada, our recently enacted Statistics Act requires (Queen's Printer for Canada, 1970, Section 16.1.b) that no sworn employee of Statistics Canada "shall disclose or knowingly cause to be disclosed, by any means, any information obtained under this Act in such a manner that it is possible from any such disclosure to relate the particulars obtained, from any individual return to any identifiable individual person, business or organization".

This requirement is very similar to its counterpart in the former Act. Its introduction, however, coincides with the increasing practical difficulty of preventing disclosures in statistical data. While the new Act was being drafted and legislated, planning and "tooling up" for the 1971 Census of Canada were also well under way. The Census tabulation programme is far more complex than that of any former Census. In addition to pre-planned tabulations, it includes efficient means of access to the data base in order to produce special-area and special-population tabulations as they are needed and requested. With outputs on paper, microfilm and magnetic tape and with the indeterminate accumulation of special requests in future years, the tabulation programme is much too large and interrelated to admit to manual checking for even the most rudimentary violations of individuals' confidences.

The rigorous wording of the new Act apparently requires prevention of possible disclosures in aggregate (tabulated) data, yet no concrete guides exist in legal precedent as to what would be and what would not be illegal disclosure. Therefore, it was the advantage of Statistics Canada to develop sufficient safeguards to avoid even the risk of violating the law or public trust. Because of the imminence and complexity of 1971 Census tabulations, an acceptable means of preventing disclosure had to be amenable to easy introduction into computer programmes which had already been developed. Also, an acceptable means of prevention must not distort the statistical meaning of data; it must be mathematically straightforward enough so that the statistical agency can give a general guarantee of its harmlessness.

1. Possibilities of disclosure

In order to have criteria with which to examine the effectiveness of random-rounding, we begin with a catalogue of the logical possibilities of disclosure in aggregate data. Adhering strictly to the Statistics Act, two assumptions underly the catalogue. First, a cell count or a measure derived from cell counts may be an illegal disclosure regardless of what additional information, from other counts and measures or from independent sources, must be mustered so that information about an individual person or organization can be deduced. Second, all cell counts and derived measures are treated as if exactly correct, as if completely free of error. Though the second assumption might be improbable in any given instance, the possibility of disclosure,not the likelihood, depends on the possibility of such exactitude. Hopefully, the caution which these assumptions prescribe will result in more than foiling the possibilities of disclosure. It should discourage most attempts to find disclosures and, perhaps, spare someone from being victimized by means of a seeming disclosure which is no less threatening just because it is inaccurate.

The simplest possibility is direct disclosure from a cell which either is empty or counts one individual (person, business or other organization). If a particular individual is known to be the only one who could be counted in that cell, the only one who could have the particular combination of attributes counted by that cell, then a count of 0 discloses that he does not have the particular combination of attributes, a count of 1 discloses that he does indeed have the particular combination of attributes.

Given the possibility of disclosure from a cell counts of 0 or 1, it is easy to imagine similar possibilities from counts of 1 or 2, 2 or 3, and so forth. If one individual is known to have a particular combination of attributes and a second individual is known to be the only other one who could have the same combination of attributes, then a cell count of 1 for the combination of attributes discloses that the second individual does not have them, a count of 2 discloses that he does have them. The same reasoning works if two, three, or any number of individuals have a particular combination of attributes and only one other individual could have the same combination. Because the individuals known to have the particular combination of attributes can be thought of as being eliminated from the cell count in order to see whether a 0 or 1 remains, we call this disclosure by elimination within a cell.

Given the possibility of elimination within a cell, it is again easy to imagine the next possibility - disclosure by elimination among cells. In the simplest instance, this would be to sub-

tract one cell count from another in order to find a difference of 0 or 1. Of course, the two cells must be related such that one is the subset of the other - so that differencing their counts yields the count of a second subset, the subset which complements the first with respect to the entire set. If a particular individual is known to be the only one who could have the particular combination of attributes counted by the second subset, then a difference of 0 discloses that he does not have the particular combination, a difference of 1 discloses that he does with this simple instance in mind, it is obvious that the count of the set and/or of the first subset could be sums of cell counts. Also, the difference between counts of the set and the first subset might be greater than 1 and still subject to disclosure by elimination within a cell. The full report gives a detailed discussion of various eventualities.

The above three possibilities seem to us fundamental, and we only mention two derivative varieties. First, given the possibilities of disclosure about individual persons or organizations through cell counts which reveal whether or not they have particular combinations of attributes, it follows that disclosures about organizations might be derived from disclosures about the attributes of members and that disclosures about persons or sub-organizations might be derived from disclosures about attributes of organizations of which they are members. Because of the necessity to extend information from members to organizations or vice-versa, we call this disclosure by extended correspondence. The second variety is disclosure from derived measures. It is simply the derivation of disclosing cell counts from measures which themselves had originally been derived from cell counts - weighted sample counts, percentiles, means, rates, and percentages. The full report discusses these two varieties of disclosure in detail.

2. Random-rounding

Recognition of the danger of disclosure has often been focussed on the very small cell counts; replacing these small counts by an asterisk or other cipher has been a common remedy. Our catalogue of possible disclosures shows small cell counts to be only part of the danger: the general danger lies in the possibility of exactly correct, error-free cell counts - of any size. The strategy of random-rounding is to render improbable and indeterminable the correctness of each issued cell count by introducing an acceptably small and unbiased error into its original value - a value which itself might or might not be free of error. Each individual random-rounding error must be indeterminable, but the resultant loss of data reliability must be estimable.

Random-rounding is a variation on systematic (or conventional) rounding. Partly because the error introduced by systematic rounding need not be free of bias, partly because the rigid rules of systematic rounding can sometimes be used to deduce pre-rounded values, random-rounding must have two properties. First, any value to be rounded may be rounded either "up" or "down" so that, in the long run, positive and negative errors balance to yield zero bias. Second, every single decision to round "up" or "down" must be indeterminable so that there exist no circumstances in which a pre-rounded value can be deduced from its position in a distribution or cross-classification and from knowledge or rounding rules.

2.1 Probabilities of rounding up and down -

Imagine that all the counts in a table of, for example, m x n cells are being rounded to multiples of some integer value, called the rounding base. For each cell count, the remainder to be rounded up or down is calculated:

(2-1) $r_{ij} = c_{ij} \bmod b$

where r_{ij} is the remainder for the ij^{th} cell in the mxn table, c_{ij} is the underlying cell count, and b is the rounding base. The calculation of r_{ij} is clarified by an identity:

(2-2) $r_{ij} = c_{ij} - bk_{ij}$

where k_{ij} is the largest integer such that $bk_{ij} \leq c_{ij}$. Clearly,

(2-3) $0 \leq r_{ij} < b$ and r_{ij} must also be an integer value because all cell counts c_{ij}, are themselves integer values. Given r_{ij}, rounding "up" is defined as

(2-4) $\hat{c}_{ij} = c_{ij} + (b - r_{ij})$

where $\hat{c}_{ij}$ is the rounded cell count. The error, e_{ij}, introduced by rounding up is

(2-5) $e_{ij} = b - r_{ij}$ Substituting the value of r_{ij} from Eq. 2-2 in Eq. 2-4 shows that

(2-6) $\hat{c}_{ij} = b\,(k_{ij} + 1)$, that the underlying c_{ij} has indeed been rounded to a multiple of b.

Similarly, rounding "down" is defined as

(2-7) $\hat{c}_{ij} = c_{ij} - r_{ij}$ The corresponding error is

(2-8) $e_{ij} = -r_{ij}$, and the rounded value is again a multiple of b:

(2-9) $\hat{c}_{ij} = bk_{ij}$

Given the two alternatives of rounding up or down, imagine a probability being attached to each alternative, such that the sum of the two probabilities is unity. Expected rounding error for each r_{ij} can then be calculated as the sum of two products, each probability times the error which it introduces:

(2-10) $E(e_{ij}|r_{ij}) = P_{ij}\,(b - r_{ij}) + (1 - P_{ij})\,(-r_{ij})$ where $E(e_{ij}|r_{ij})$ is the expected rounding error for the r_{ij}, P_{ij} is the probability of rounding c_{ij} up and the error terms come from Eqs. 2-5 and 2-8.

One means of excluding bias from the rounding error is to require the expected error for each r_{ij} to be zero:

(2-11) $0 = P_{ij}\,(b - r_{ij}) + (1 - P_{ij})\,(-r_{ij})$.

Then, solving for P_{ij} reveals the probability to be attached to rounding up:

(2-12) $P_{ij} = r_{ij}/b,$

and the probability for rounding down is obvious:

(2-13) $(1 - P_{ij}) = 1 - (r_{ij}/b).$

With the expected error for each $\hat{c}_{ij}$ equal to zero, the expected error for any sum of $\hat{c}_{ij}$'s will also equal zero. An appendix to the full report documents the superiority of random-rounding over systematic rounding in this respect.

2.2 Random decisions to round up and down -

Given probabilities to round up and down, such that positive and negative errors balance to yield zero bias, it remains to render indeterminable the direction of any particular instance of rounding. Then, no underlying count will be deducible from its position in a distribution or cross-classification and a knowledge of systematic-rounding rules.

An obvious means of making indeterminable each rounding decision (up? or down?) is to line the cell counts in a sequence and to match the sequence of cell counts to a sequence of random numbers: each random number in sequence will be unpredictable and will decide the rounding direction for the cell count with which it is matched. In order for the rounding decisions to conform with the rounding probabilities in Eqs. 2-12 and 2-13, the random numbers must be uniformly distributed, lying in real interval between zero and the rounding base:

(2-14) $0 \leq v_{ij} < b$

where v_{ij} is the random number matched with the ij^{th} cell count.

Given uniform distribution of the random numbers, the probability of v_{ij} being less than r_{ij} is as follows:

(2-15) $$P(v_{ij} < r_{ij}) = \frac{r_{ij} - 0}{b - 0} = r_{ij}/b$$

with r_{ij} as defined in Eqs. 1-1 and 1-2. By substitution from Eq. 2-12, this is clearly the probability of rounding "up".

(2-16) $P(v_{ij} < r_{ij}) = P_{ij}.$

Similarly, the probability of v_{ij} being greater than or equal to r_{ij} is as follows:

(2-17) $$P(v_{ij} \geq r_{ij}) = \frac{b - r_{ij}}{b-0} = 1 - (r_{ij}/b).$$

By substitution from Eq. 2-13, this is the probability of rounding down:

(2-18) $P(v_{ij} \geq r_{ij}) = 1 - P_{ij}.$

Given Eqs. 2-16 and 2-18, it is a simple task to tie the rounding decisions for each cell count to the random number with which the count is matched: if v_{ij} is less than r_{ij}, then round up; otherwise round down. Rounding up and down being defined, respectively, in Eqs. 2-4 and 2-7, this is the random-rounding procedure for each cell count:

(2-19) IF $v_{ij} < r_{ij}$

THEN $\hat{c}_{ij} = c_{ij} + b - r_{ij}$;

ELSE $\hat{c}_{ij} = c_{ij} - r_{ij}$

where IF, THEN, and ELSE have their common meanings. Substituting from Eq. 2-1:

(2-20) IF $v_{ij} < c_{ij} \bmod b$

THEN $\hat{c}_{ij} = c_{ij} + b - c_{ij} \bmod b$;

ELSE $\hat{c}_{ij} = c_{ij} - c_{ij} \bmod b.$

2.3 Generating random numbers -

The random-rounding procedure stated in Eq. 2-19 and again in Eq. 2-20 yields the two necessary qualities--zero bias and indeterminacy. As shown in Eq. 2-20, three values - c_{ij}, b and v_{ij} must be at hand in order to random-round each ij^{th} cell. The cell count, c_{ij}, is given in each instance, and the rounding base, b, is predetermined. (The choice of b is discussed in the full report).

Given the practical need to generate sequences of random numbers, v_{ij}'s, a brief survey of literature led us to RANDU--the "power-residue" routine published by IBM (1968, p.77). Its only distinctly non-random feature is that idential "starting values" yield identical sequences of (random) numbers. In order to circumvent this weakness, each starting value is determined by the respective computer run's exact START time--a truly unpredictable value.

RANDU generates values, y_{ij}, in the open interval between zero and one:

(2-21) $0 < y_{ij} < 1.$

Multiplication of these values by b almost transforms them into the values required to satisfy Eq. 2-14:

(2-22) $0 < by_{ij} < b.$

The discrepancy between Eqs. 2-14 and 2-22 is that the former includes the exact value of zero, {0,b), while the latter excludes it, (0,b). So trivial is this discrepancy that any effort to compensate for it would introduce additional error. For practical purposes:

(2-23) $v_{ij} = by_{ij}.$

It should be noted that computer execution of random-rounding, often as an addition to existing tabulation-producing programmes, is greatly eased by the inconsequence of exactly how sequences of random numbers are matched to cell counts. Any "route" may be taken through a table: our imaginary mxn table and ij^{th} cell have no other purpose than to indicate the independent and identical application of Eq. 2-20 to

each cell count.

3. Final remarks

In addition to the topics covered in this extract, the full report discusses the acceptability of random-rounding error compared to other forms of error in census data, the effectiveness of random-rounding in preventing disclosures, and *a fortiori* estimation of random-rounding error in tabulations of cell counts.

Many people at Statistics Canada helped us to progress from a vague grasp of the confidentiality problem to a rigorous development and justification of random-rounding. We mention only a few - and them only for their most important contributions. E.M. Murphy brought the problem to our attention and guided initial work with bright ideas and practical sense. Ivan Fellegi, through earlier drafts of the paper referenced below, outlined both the problem and the general practical requirements of any solution. L.O. Stone and F.G. Boardman gave us notions which pointed our minds at random-rounding. R.J. Davy provided us with crucial managerial support and criticism. M.A. Mocken relentlessly chained our speculations to the practical qualities of the data-processing systems. Others, whom we do not even know, have taken much time and trouble to make random-rounding part of the computer programmes which produce 1971 Census tabulations.

REFERENCES

[1] Fellegi, I.P.: "On the Question of Statistical Confidentiality", Journal of the American Statistical Association, 67,337 (March 1972).

[2] IBM: "System/360 Scientific Subroutine Package, (360A-CM-03X) Version III, Programmer's Manual", IBM, White Plains (New York), 1968.

[3] Parliament of Canada: "Statistics Act". The Queen's Printer for Canada, Ottawa, 1970.