This PDF is a selection from an out-of-print volume from the National Bureau of Economic Research

Volume Title: Annals of Economic and Social Measurement, Volume 3, number 2

Volume Author/Editor: Sanford V. Berg, editor

Volume Publisher: NBER

Volume URL: http://www.nber.org/books/aesm74-2

Publication Date: April 1974

Chapter Title: Statistical Confidentiality: Some Theory and Applications to Data Dissemination

Chapter Author: I. P. Fellegi, J. L. Phillips

Chapter URL: http://www.nber.org/chapters/c10117

Chapter pages in book: (p. 399 - 409)

Annals of Economic and Social Measurement, 3/2, 1974

STATISTICAL CONFIDENTIALITY: SOME THEORY AND APPLICATIONS TO DATA DISSEMINATION

I. P. Fellegi and J. L. Phillips

Potential disclosure has been a problem with cross-tabulations. With computers, both the problems and potential uses are multiplied. This paper indicates how to eliminate the possibility of direct and residual disclosure without limiting a statistical agency's dissemination capability. This price paid for the mass produced flexibility is a loss of some reliability for very small frequencies.

1. Introduction

Statistical offices traditionally carefully scrutinize their publications to insure that there is no disclosure, i.e., disclosure of information about individual respondents. This task has never been easy or straightforward. Yet, in the past the technical limitations of available tabulating equipment put a rather low ceiling over the number of tabulation cells that could economically be compiled from any single survey; moreover, these were preplanned tabulations, which typically repeated themselves from survey to survey. Under these circumstances the problem of scrutinizing the tabulations, prior to their release, for potential disclosure was more manageable.

Computers have now become such powerful tools in the hands of both users and producers of data that the dimensions (though not the substance) of the confidentiality problem have been transformed. Computers, by enabling users to apply analytical and decision-assisting techniques to a variety of statistical information, typically in highly disaggregated form, have stimulated an increasing demand for detailed information which is often of an *ad hoc* nature rather than pre-planned. Similarly, the increasing role of governments at all levels puts additional demands on statistical agencies for data to support their administrative, regulatory, policy-making, judical and planning activities. Computers have also provided statistical agencies with a tool for processing, storing and retrieving information from a variety of separate or linked files, possibly collected over long periods of time. Thus computers have altered the statistical "market" on both sides: on the side of the "purchaser" as well as on the side of the "producer."

The statistical "market" situation, however, has a third dimension as well: the needs and concerns of the general public. The public benefits indirectly from the legitimate uses of statistics by governments, businesses, nonprofit organizations, academic users, etc.; yet, the public is also concerned about the increasing burden of providing the required statistics and about the real or imagined possibility of the misuse of the data provided by them. The explosive increase in the demands for more statistics can only be met, without impossible response burdens being put on the public, through a more effective exploitation of the data. This increases geometrically the magnitude of the problem of checking tabulations for disclosure. It will not be possible to continue very long with the manual methods of checking; the development of mass production techniques cannot be put off. This is necessary not only because of the legal requirement but also because statistical offices must make the concern of the public their own; the real foundation of reliable statistics is public cooperation, not the threat of prosecution under a Statistics Act.

This paper, the first half of which is based on an article by one of the authors (Fellegi, 1972), discusses the nature of the disclosure problem, provides a necessary and sufficient condition for residual (or indirect) disclosure to occur and concludes by the discussion of an approach which can be applied on a mass production scale and which eliminates the possibility of both direct and residual disclosure without limiting the statistical agency's dissemination capability. The price to be paid for the mass produced flexibility is a loss of some reliability for very small frequencies.

2. Direct and Residual Disclosure: Definitions and Tests

2.1 *Considerations Concerning a Possible Definition of Inadvertent Direct Disclosure*

Inadvertent direct disclosure, or i.d.d. for short, involves making available information concerning a unique and identifiable individual. The statistical office, of course, never discloses information about an individual who is identified by his name. But the concept of disclosure also implies restrictions on disclosure of information on an individual who can be identified through his characteristics. In this latter case, therefore, disclosure occurs when a user can identify a respondent by recognizing him through his characteristics and learning something about him. From this point of view violation of confidentiality might be defined as the disclosure of information that goes beyond that required for identification alone.

In the case of a tabulation of counts (frequencies) from a census one may argue, therefore, that a table in which some of the cells contain entries of one, but in which none of the marginal totals are ones, does not represent a violation of confidentiality. In fact, in this case a particular entry of one in a table can only be recognized as referring to a unique identifiable person if the reader knows *a priori* that the particular person has all the characteristics indicated by the table. However, should another dimension of breakdown be superimposed on the table, then disclosure would clearly occur: at that point, information is disclosed about a person which goes beyond that required for his unique identification. In a sense at that point the reader may learn something new about the particular person.

To illustrate by an example, suppose that a census table published for a given municipality is as follows.

TABLE 1

Number of Persons by Industry and Occupation
Municipality *X*

Industry/Occupation	...	i	...	Total
⋮	⋮	⋮	⋮	⋮
j	·	1	·	139
⋮	⋮	⋮	⋮	⋮
Total	...	5	...	4329

Suppose that the entry of 1 in row *j*, column *i* refers to the synthetic textile industry, occupation statistician. The reader may recognize the person to whom the entry of one refers: he may say, "Joe Smith is a statistician working in a synthetic textile mill; the table shows that there is one such person; that entry must therefore refer to Joe Smith." For the reader to recognize this entry as referring to Joe Smith, he must know in advance both that Joe Smith is a statistician and that he works in a synthetic textile mill.

Since in this example none of the marginal totals are equal to one, neither the industry nor the occupation identifies Joe Smith by itself: both are needed simply for identification. If this table, however, is extended to a cross-classification of industry by occupation by age, at that point the reader may learn Joe Smith's age.

TABLE 2

NUMBER OF PERSONS BY INDUSTRY, OCCUPATION AND AGE
MUNICIPALITY *X*

Industry	Occupation	.....	*i*							...
			Age							
			<20	20–29	30–39	40–49	50–59	60+	Total	...
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
	j	....	0	0	1	0	0	0	1	...
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Total			0	2	3	0	0	0	5	⋮

Clearly, Table 2 discloses Joe Smith's age group: he is between 30 and 39 years of age.

I.d.d. in the case of count (frequency) tables based on a census could therefore be defined as an entry of one in a table, provided that at least one of the corresponding possible marginal totals is also one. Given such a precise definition of i.d.d. (or one similar to it) checking for it can relatively easily be automated.

This argument, of course, does not stand or fall on whether or not the definition is in terms of an entry of one. It may be argued that if there are two statisticians in the synthetic textile industry, the age distribution would disclose to the other statistician the age of Joe Smith. The definition may be extended and slightly generalized to read: i.d.d. occurs when in an *n*-dimensional cross-tabulation of counts one of the possible marginal totals (in the dimension $n - 1$) is equal to a specified number (presumably one, two, or at most three).

In the case of tabulations based on sample data, the condition stated above may be relaxed. No i.d.d. occurs so long as two conditions are satisfied: the identity of respondents in the sample is held confidential and all tabulation cells are such that the corresponding population values are greater than one (or some other specified number). Of these two conditions the first one seems to be very important; the second is probably automatically satisfied since only very detailed tables would violate it and these would not be disseminated anyway due to the sampling errors involved.

Finally, in the case of tabulated quantities (in contrast with counts), one may arrive at appropriate translations of the guidelines indicated above if one assumes that at least the order of magnitude of the individual quantities involved may be known to the reader *a priori*. For example, if in a tabulation of industry by occupation the total income earned is tabulated instead of the number of persons, it is probable that a well-informed reader may know the order of magnitude of the average income in each cell so that he may be able to deduce the number of people involved from the total income that is reported (at any rate, he might even be able to obtain from the statistical office a separate tabulation showing the corresponding number of people). In this case i.d.d. occurs as soon as identification is possible. Hence in our previous example of the industry by occupation table, if total income earned is tabulated rather than the number of persons, then the entry in the cell corresponding to the synthetic textile industry and statistician occupation would disclose Joe Smith's income: the definition of the cell identifies Joe Smith, the entry in the cell discloses his income.

The operationally important point is to stop further cross-classifications just prior to any individual response becoming identifiable in a tabulation cell. Identifiability in the case of tabulated aggregates, of course, may depend on considerations other than the number of observations in the cell. Even where a cell involves several observations, a uniquely large reported quantity (e.g., income) may be identifiable. In such cases definitions of identifiability other than those based on the number of observations alone must be adopted, e.g., that no single response in a tabulated aggregate may account for more than a specified percent of the total.

A special type of indirect disclosure is worth at least mentioning; the so-called *negative disclosure*. Negative disclosure occurs when a tabulated zero for a well-defined population in effect discloses that no one in the population has the particular characteristic. It only takes a moment of reflection to realize that, for example, if a tabulation showing the number of persons by income group indicates that there are no persons in a given area earning over $40,000 per annum, then this tabulation in effect discloses about every individual in the given area that his income is below $40,000. Although from a purely legal sense such a disclosure would probably not be prosecutable, from a strictly substantive point of view it is disclosure nonetheless: information is provided, although indirectly, about identifiable individual respondents.

2.2. *Checking for Residual Disclosure* (*Counts or Aggregates*)

Residual disclosure occurs when two or more sets of published data, taken together, enable the user to identify information pertaining to an individual respondent even though none of the published data, by itself, is a direct disclosure.

Given an unambiguous definition of inadvertent direct disclosure exists, a precise treatment of the problem of residual disclosure is feasible (equally valid whether counts or aggregates are tabulated). In fact, in a previous paper (Fellegi, 1972) a precise mathematical treatment was provided dealing with the problem of detecting residual disclosure. In that paper a theorem is stated and proved which provides a necessary and sufficient condition for residual disclosure to occur. The condition is expressed in terms of the rank of a matrix whose *smaller* dimension is equal to the number of published cells derived from the given survey data base.

Clearly the calculation of the rank of a matrix as large as the one implied by the previous paragraph represents a potentially large amount of computation, particularly if the number of retrieved tabulation cells is large. This is, however, a consequence of the nature of the problem rather than of the complexity of the theorem. In fact, since the theorem provides a necessary and sufficient condition for residual disclosure to occur, the procedures implied by the theorem are logically equivalent to any other set of test procedures. With a modest and predetermined publication plan, particularly if the publication cells correspond to generally non-overlapping sets of respondents testing for residual disclosure is feasible.

Even where testing for residual disclosure is feasible, we are left with the problem: when we discover that a tabulation, taken together with all previously published tabulations, is a disclosure, what should we do? The options which are available are explored in the next section.

3. A Flexible Data Dissemination Program which Overcomes the Problem of Disclosure

3.1 *STATPAK: a Flexible Retrieval and Tabulation Program*

As part of the general strategy of dissemination of the 1971 Census data, a very flexible retrieval system was designed. This system, called the Geographically Referenced Data Storage and Retrieval System (GRDSR) is described in more detail elsewhere (Fellegi and Goldberg, 1969; and Statistics Canada, 1972). The system is a complex one which assigns geographic coordinates to all households in the Census and then enables us to tabulate data for any special area that a user wants to identify by outlining its boundaries on a map or by other means. One of the modules of the GRDSR system is called STATPAK, which is a general tabulation program that can, in fact, be used whether or not the data base is geographically coded.

Thus STATPAK is a data retrieval package which produces cross classified tables for arbitrary areas. It produces frequency *counts*, and *sums* of quantitative data (such as income) for any specified breakdown (up to ten dimensions). Options exist for the computation of subtotals, totals, averages, percentages and in general for the tabulation of functions, counts or sums at the row, column, hyper-plane, etc. level.

In the face of the technical flexibility of STATPAK and the substantive flexibility adopted as the official policy of Statistics Canada in relation to the 1971 data dissemination program, the question of disclosure became an issue of paramount importance. Clearly, testing for residual disclosure along the lines of Section 2 is computationally impossible. But dealing even with inadvertent direct disclosure on a mass production scale would be quite a task. Some of the alternative methods which were considered are outlined below.

3.2. *Alternative Methods of Dealing with Disclosure*

(a) *Suppression*

Superficially, preventing disclosure would appear to be quite simple: suppressing those numbers which would represent direct disclosure. Unfortunately, subtotals and totals can often be used to "fill in the blanks." Suppressing entire tables containing disclosure is not a pleasant possibility. Neither of these prevents residual disclosure.

(b) *Grouping*

The entries in rows, columns, hyper-planes, etc., could be aggregated (grouped) together with adjacent rows, columns, etc., until the numbers become large enough to print. The main problems here are the possible loss of meaningful data (i.e., the loss of separate break-outs of data which could be shown but which have to be grouped together with other data to prevent the disclosure of the other data), difficult computer implementation, possible additional burden on the user who might have to supply constraints to avoid undesirable groupings, and most importantly no protection against residual disclosure.

(c) *Rounding*

All numbers could be rounded up or down to some multiple of a base number in the usual way. In tests with census data, this method produced biased estimates in the sense that, since a disproportionate number of tabulation cells involved small last digits, the rounding was more often a rounding down than a rounding up. It is also easy to show that rounding does not necessarily prevent residual disclosure. If Table 3 below is known to have been obtained by rounding each entry (including totals) to multiples of five or zero, it is easy to deduce that the underlying unrounded numbers are those shown in Table 4. The reason why Table 4 can immediately be deduced from Table 3 is the completely predictable nature of rounding.

TABLE 3

A Hypothetical Table in Which Every Entry is Rounded to a Multiple of 5 or Zero

			Total
	0	0	5
	0	0	5
Total	5	0	5

TABLE 4

Unrounded Hypothetical Table Underlying Table 3

			Total
	2	1	3
	2	1	3
Total	4	2	6

(d) *Random rounding*

Given the fact that ordinary rounding has a tendency to result in biased estimates, and given that due to its predictability it does not always prevent disclosure, it seemed like a natural idea to introduce randomization into the rounding process. Given a *rounding base b*, such that every number is rounded to its multiples, and if r is the remainder of a number when divided by b, an unbiased random rounding procedure could be defined as follows:

(i) round up with probability equal to r/b $(0 < r < b)$

(ii) round down with probability equal to $1 - r/b$ $(0 < r < b)$

(iii) do not round if $r = 0$.

It is easy to show that this procedure is unbiased in the sense that the expected value of any number so rounded is equal to its original unrounded value. It follows that this is also true for the sum of randomly rounded numbers.

The choice of a base for random rounding is obviously critical. The larger the base, the larger the variance will be due to random rounding. On the other hand a small base, such as 2 or 3, does not introduce enough uncertainty to effectively prevent disclosure. The detailed considerations relating to the choice of an appropriate rounding base go beyond the scope of the present paper: they are extensively discussed by Nargundkar and Saveland in an unpublished paper whose summary is listed as a reference (Nargundkar, Saveland; 1972). For purposes of the present paper we just mention that the rounding base for the 1971 Census data dissemination program was chosen to be 5. This is a large enough base to effectively prevent disclosure and at the same time its effect on data reliability is acceptably small (except for very small numbers). It also has the advantage that in a publication its effect is immediately visible since every number ends in a digit of 0 or 5.

Some of the advantages of the random rounding techniques are:

(i) It is easy to understand and is intuitively appealing.

(ii) The expected value of every rounded count is equal to the original count; that is, the rounded count is an unbiased estimator of the original count. This property is particularly important if the rounded numbers are used to produce other statistics.

(iii) Direct, residual and negative disclosure are all prevented. For example, with base 5, a tabulated zero may now represent any number between 0 and 4, a 5 may represent any number between 1 and 9, etc.

(iv) The error introduced by random rounding using a small rounding base has minimal effect on the data.

(v) Computer implementation is simple.

As a matter of official policy, all publications, summary tapes and user-requested special tabulations from the 1971 Population Census of Canada are subject to random rounding.

3.3. *The Implementation of Random Rounding in STATPAK*

While the random rounding technique is straightforward and easy to apply, a few additional requirements make it impossible to simply random round every printed number individually. Some of the considerations in the design are outlined here:

1. Averages should be maintained.

For example, if the original tables showed 3 individuals earning a total of $33,003.00, thus an average of $11,001.00 each, we do not want to random round these figures separately, giving, for example, 5 persons earning $33,000.00, an average of $6,600 each. Rather we would like to show either 5 persons earning $55,005 or 0 persons earning $0.00.

2. The rounding error of totals and subtotals should be minimized.

Suppose, totalling and subtotalling would be done following random rounding. We might, for example, have had the following unrounded table:

					Total
7	11	13	4	22	57

Now if we random rounded each of the frequency counts and then totalled, we might get:

					Total
5	10	10	5	20	50

But as can be seen, because by chance we rounded down more often than up, the TOTAL value contains the accumulated error and the rounded total is outside the range we want. In order to minimize the rounding error of totals and subtotals they are first accumulated, then all tabulation cells in the resulting tables are independently rounded, *including the totals*. In the example above, 57 would be rounded to either 55 or 60. Thus, in order to minimize the rounding error of totals and subtotals we have to sacrifice the reassuring feature that the "totals add up." This will be true now *only by chance*. As will be seen later, the effect of this on tables involving small counts (frequencies) can be startling, at least until one gets used to it.

3. Percentages should not "give the game away." This applies also to functions of rows, columns, hyper-planes, etc.

For example, suppose we had the following two corresponding tables:

TABLE 5
Hypothetical Table of Frequencies (Count)

			Total
4	1	0	5

TABLE 6
Hypothetical Table of Income Totals

			Total
44,000	15,000	0	59,000

As percentages the table of frequencies would be:

TABLE 7
Hypothetical Table of Percentages

			Total
80%	20%	0%	100%

Now suppose we round while maintaining averages. We might get the following:

TABLE 8
Rounded Hypothetical Table of Counts (Frequencies)

			Total
5	0	0	5

TABLE 9
Rounded Hypothetical Table of Income Totals

			Total
55,000	0	0	59,000

If we published the unrounded Table 7 of percentages, together with Tables 8 and 9, on the basis of a minimal understanding of the mechanics of random rounding it would be obvious that the unrounded counts must have been 4 and 1. Now using Table 9, the unrounded Table 6 of incomes can be obtained.

To avoid such problems, all percentages, and other functions (except averages) of rows, columns, etc. are computed after random rounding of the count and sum tables.

4. The result of repeated random rounding of the same table should be the same.

This avoids the undesirable nuisance of getting different answers on reruns. It also provides some deterrent to attempts to obtain an arbitrarily close approximation of the unrounded value by repeatedly retrieving the same table and computing the sample mean.

To satisfy these objectives the following basic procedure is followed:

1. Retrieve data and accumulate the required table of frequencies, called the *count matrix* (in the case of weighted files, the count matrix is the sum of the weights) and any *sum matrices* required (including those needed to compute averages, percentages, etc.).
2. Compute any totals or subtotals required.
3. Divide the elements of each sum matrix by the corresponding elements in the count matrix to obtain averages.
4. Random round each element in the count matrix. The first non-zero number in the count matrix is used as a starting number for a random number generator.
5. Multiply the temporary average matrices (from step 3) by the rounded count matrix to give rounded sum matrices.
6. Compute ratios, percentages, averages, etc. using the rounded count and sum matrices.
7. Do traditional rounding to produce figures rounded to the number of decimal places the user has specified; print the matrices.

3.4. *The Impact of Random Rounding*

Understanding the random rounding procedure is no challenge. Accepting some of the tables as valid products of this procedure is more difficult. Tables whose related count matrix contains very small numbers can be severely distorted. Looking at some tables (even with a complete theoretical understanding of the random rounding implementation) can produce a "what happened?" reaction.

The following are some examples of random rounding. Consider Table 10 which is unrounded. A hypothetical random rounding of Table 10 is shown in Table 11. Note that in Table 11 the subtotal happened to be rounded up while the total happened to be rounded down, with the result that the subtotal is 200 percent of the total. As can be seen, very small counts rounded can produce useless results. With even moderately large counts the distortion is minimal. For example, consider Tables 12 and 13. The second table is a rounded version of the first where, as it happened, the worst possible random rounding occurred in that both the subtotal

TABLE 10

Hypothetical Unrounded Tabulation of Income by Age Group for a Small Subpopulation

	Count	Percentages on Count	Sum (Income)	Average (Income)	Percentages on (Income)
Age 11–20	—	—	—	—	—
Age 21–30	6	100	13,800	2,300	100
Age 31–40	—	—	—	—	—
Subtotal	6	100	13,800	2,300	100
All Other Ages	—	—	—	—	—
Total	6	100	13,800	2,300	100

TABLE 11

Hypothetical Random Rounded Tabulation of Income by Age Groups for a Small Subpopulation

	Count	Percentages on Count	Sum (Income)	Average (Income)	Percentages on (Income)
Age 11–20	—	—	—	—	—
Age 21–30	5	100	11,500	2,300	100
Age 31–40	—	—	—	—	—
Subtotal	10	200	23,000	2,300	200
All Other Ages	—	—	—	—	—
Total	5	100	11,500	2,300	100

and total have been rounded down while the rest of the counts were rounded up. Clearly for most uses, Tables 12 and 13 are identical.

In pondering the impact of random rounding, one has to keep in mind that it is the mean squared error of the final numbers that matters, not the error associated with a particular operation. Estimates based on small numbers (even from a census) usually have relatively large errors associated with them due to response, sampling, processing and other errors. On the basis of more detailed studies (Nargundkar and Saveland, 1972), the increase in the mean squared error of census estimates due to random rounding is negligible for estimates based on moderately large frequencies (10–15 persons or more).

TABLE 12

Hypothetical Unrounded Tabulation of Income by Age Groups

	Count	Percentages on Count	Sum (Income)	Average (Income)	Percentages on (Income)
Age 11–20	17	1	17,000	1,000	0
Age 21–30	321	36	1,284,000	4,000	30
Age 31–40	163	18	978,000	6,000	23
Subtotal	501	56	2,279,000	4,549	53
All Other Ages	398	44	1,990,000	5,000	57
Total	899	100	4,269,000	4,749	100

TABLE 13

Hypothetical Random Rounded Tabulation of Income by Age Groups

	Count	Percentages on Count	Sum (Income)	Average (Income)	Percentages on (Income)
Age 11–20	20	2	20,000	1,000	0
Age 21–30	325	36	1,300,000	4,000	31
Age 31–40	165	18	990,000	6,000	23
Subtotal	500	56	2,274,500	4,549	54
All Other Ages	400	44	2,000,000	5,000	57
Total	895	100	4,250,355	4,749	100

Conclusion

The implementation of random rounding in STATPAK provides adequate safeguarding of confidential data. Very small numbers are relatively severely distorted. This provides good protection against both direct and indirect (residual as well as negative) disclosure. Thus, when designing a STATPAK tabulation, care should be taken to specify a breakdown in keeping with the number of data units being retrieved. At any rate, the increased mean squared error (very slight for counts exceeding, say, 10) is a price that has to bc paid for the almost unlimited retrieval and tabulation flexibility which a retrieval program like STATPAK can provide.

Statistics Canada

References

Bachi, R. and Baron, R., "Confidentiality Problems Related to Data-Banks," *Bulletin of the International Statistical Institute*, Vol. 43, Book 1 (1969), 225–41.

Fellegi, I. P. and Goldberg, S. A., "Some Aspects of the Impact of the Computer on Official Statistics," *Bulletin of the International Statistical Institute*, Vol. 43, Book 1 (1969), 157–75.

Fellegi, I. P., "On the Question of Statistical Confidentiality," *Journal of the American Statistical Association*, 67 (March 1972), 7–18.

Nargundkar, M. S. and Saveland, W., "Random Rounding: A Means of Preventing Disclosure of Information About Individual Respondents in Aggregate Data," *American Statistical Association Proceedings*, Social Statistics Section, 1972.

Statistics Canada, "GRDSR: Facts by Small Areas", June 1972.