BUREAU OF THE CENSUS

STATISTICAL RESEARCH DIVISION REPORT SERIES

SRD Research Report Number: CENSUS/SRD/RR-85/03

A SELECTIVE REVIEW OF CONFIDENTIALITY RESEARCH PUBLISHED SINCE 1975

by

Philip J. Smith

U.S. Bureau of the Census
Washington, D.C. 20233

Recommended by: Nash J. Monsour

Report completed: April 12, 1985

Report issued: April 17, 1985

Abstract

This technical report is a selective review of confidentiality research which has been published since 1975. Rather than being a highly polished and final product of my work in confidentiality, it is a "milestone marker" along the course of my own research. We offer it in the SRD technical report series for documentation of this continuing research as well as to provide a synopsis of pertinant research for the interested reader.

This review is arranged in chronological order from the most recent publications to the oldest. Also, within each year reviews are arranged alphabetically by authors' last names.

For macrodata, among the methodologies to limit disclosure risk in this reviewed literature are cell suppression and complimentary cell suppression, data perturbation. For microdata, among the methodologies for limiting disclosure risk in this review are: data perturbation, collapsing, data suppression, swapping, and randomized response. Also, some of the literature reviewed here considers how these methodologies effect the integrity of commonly used statistics computed from microdata and also to what extent these methodologies limit disclosure.

Cox, L.H. (1984). Disclosure Control Methods for Frequency Count Data, U.S. Bureau of the Census, preliminary report.

I. Purpose of Paper

A. To present alternative methods for controlling statistical disclosure in frequency count data.

B. Disclosure occurs in frequency count data when small counts are released or may be derived from other related counts.

II. Types of Data are Distinguished According to

A. Their sensitivity, e.g., "age" vs. "has a criminal record".

B. Whether they are attribute or identifier data.

III. Operatonal Definition of Disclosure in Tables

Usually defined in terms of a threshold rule, e.g., if $\underline{X}$ denotes a set of attributes defining a cell, if the number of respondents with $X = \underline{X} < a$ threshold, n , then we may say that disclosure occurs.

IV. Different Methods for Controlling Disclosure

A. Suppressing cell values - used by U.S. Bureau of the Census.

B. Perturbing cell values and replacing cell values by intervals

1. random perturbation - used by Office of Population Censuses and Surveys of the United Kingdom and Statistics Sweden.

2. data rounding

a. random rounding--developed and used by Statistics Canada

b. controlled rounding--developed by U.S. Bureau of the Census

V. Cell Suppression

A. Primary disclosure cells have counts less than the threshold value and are suppressed. These suppressed counts are replaced by a "D" in the table. Let X denote the suppressed value which we would like to remain unknown.

B. Additive relations along rows and columns of the table generate a system of linear constraints $\underline{L}$ on the X variables. $\underline{L}$ contains certain additive constants, namely the values of the non-disclosure cells.

C. Once the primary disclosures have been suppressed, mathematical programming is employed as a disclosure audit on the table:

1. for each variable suppressed variable X above, the best linear range [min(X) , max(X)] of values for this suppressed value is derived from the constraints, L.

2. If these intervals is sufficiently broad, i.e., if

$$(0,n) \subseteq [\min(X) , \max(X)] \qquad (*)$$

then the disclosure analysis is complete.

3. If (*) fails for any X-variable then additional non-disclosure cells must be suppressed until (*) is satisfied for all primary disclosure cells X. This step is called complimentary suppression.

4. This procedure of suppressing primary and complimentary cells is aimed at making disclosure cells sufficiently ambiguous, i.e., (*) must be satisfied.

5. The "best" suppression procedure minimizes both

   a. the number of suppressed cells and
   b. the total value of suppressed cells.

   Typically it is not possible to satisfy both (a) and (b) simultaneously. Therefore, "best" complimentary cell suppressions often are defined to satisfy (a) first and, subject to (a), (b) secondly.

VI. Data Perturbation

A. Random perturbation

A fixed set of small perturbation values $\{v_i\}_{i=1}^{m}$ and an associated set of probabilities $\{p_i\}_{i=1}^{m}$ are chosen such that

$$E(v) = \sum_{i=1}^{m} p_i v_i = 0$$

Typically m is odd, m=2t+1 and $\{v_i\}$ is the set of integers from -t to t.

Instead of publishing the count in each cell, X, a choice v is made from $\{v_i\}$ according to the probabilities $\{p_i\}$. The quantity max $\{X+v, 0\}$ is published.

Consequently a user cannot infer what the actual count is except that it is in the interval

$$[\max\{X-t , 0\} , X+t]$$

(assuming the user knows t.)

Expected values are maintained for some but not all aggregates due to the upward bias introduced when small counts are perturbed.

The Office of Population Censuses and Surveys of Great Britain uses random perturbation with t=1 and $\{p_i\} = \{1/4, 1/2, 1/4\}$ (Newman, 1975). Apparently, their disclosure definiton is limited to preventing the inference that #(X)=1.

Problems

1. may change totals

2. if the published number is 0, or 1 then clearly #(X) $\leq$ n.

VII. Random Rounding

Beginning with a rounding base B (e.g., B=5) the remainder r of #(X) module B is computed (i.e., #(X) = qB + r for integers q,r with $0 \leq r <$ B).

In place of #(X), either qB or (q+1)B is published i.e., #(X) is rounded down or up to an adjacent integer multiple of R(X) of B.

Round down with probability (1-r/B)
Round up with probability r/B.

By choosing B=n+1 all published values are multiples of n and (*) is preserved.

Properties of random rounding

1. Random rounding is unbiased so that expected values are maintained.

2. In 1 way tables random rounding can be controlled so that the sum of the rounded entries equals the rounded value of the sum of the unrounded entries (Fellegi, 1975).

3. In 2 way tables random rounded cannot, in general, be controlled so as to preserve additivity. Application of controlled random rounding along rows improves additivity, but not entirely.

4. Beyond the expected value property, random rounding offers no control over the deviation between the rounded and unrounded arrays.

VIII. Controlled Rounding

- Replaces the entries in the original 2 way array by an array whose entries are rounded values adjacent to the corresponding original values.

However--

- the rounded array is guaranteed to be additive.

- rounding can be chosen so as to minimize any of a class of standard measures of deviation between the original and rounded tables. (Cox and Ernst, 1982).

- is based on a specialized linear programming model, the transportation model.

Cox, L.H. (1984). Methods for Controlling Statistical Disclosure in Microdata. U.S. Bureau of the Census draft version.

I. Methods for Controlling Disclosure of Microdata

A. Population size cutoffs 250,000 and 100,000.

B. Use of small sampling fractions.

C. If very accurate info is on the file further steps must be taken.

1. rounding
2. perturbation

D. Principal techniques

1. collapsing

a. use of population cutoffs.
b. top coding - determined by distributional properties.
c. collapsing $ figures to nearest $5,000.

2. perturbing

a. random $\varepsilon \sim (0_1\ ?)$
b. non-random
- controlled rounding - rounding to preserve additivity

rec #1	X_{11}	X_{12}	X_{13}	S_1
#2	X_{21}	X_{22}	X_{23}	S_2
	⋮			
#n	X_{n1}	X_{n2}	X_{n3}	S_n
#n	T_1	T_2	T_3	

round X_{ij} so that $\sum_j$ round $(X_{ij}) = S$ and

$\sum_i$ round $(X_{ij}) = T_j$

- microaggregation - leads to disturbed variances & destroyed correlation structure.
- data swapping
- microdata modelling

Stem, D.E. Jr. and Steinhorst, R.K. (1984). Telephone interview and mail questionnaire applications of the randomized response model, Journal of the American Statistical Association, Vol. 79, No. 387, pp. 555-564.

Summary

Randomization devises and methodologies for self administered mail questionnaire and telephone interview applications of the randomized response method are described. The results of field tests of both techniques indicate that if the serious problems of nonresponse and failure to follow instructions can be overcome, the techniques may serve as less expensive alternatives to personal interview randomized response studies. A multivariate version of Warner's linear model is also described.

Mugge, R.H. (1983). Issues in Protecting Confidentiality in National Health Statistics. Proceedings of the Section on Survey Research Methods, Washington, DC 20005.

I. Introduction. Administrators must set policies on:

A. Providing informed consent to study subjects;
B. Physical protection of records;
C. Training and supervising staff on confidentiality matters;
D. Avoiding statistical disclosures through published tables or data tapes;
E. Establishing appropriate limits to the sharing of data with other researchers;
F. Determining what else is necessary to meet the letter and spirit of existing laws and regulations;
G. Changes to be requested in laws and regulations.

II. Data on Public Use Microdata Tapes

A. To minimize the chance of disclosure they (NCHS) take the following precautons:

1. Cases with rare characteristics are either

a. deleted or
b. concealed in a broad coding category.

2. Cases from small geographic areas with populaton less than 100,000 are either

a. removed from the file or
b. the geographic code is encrypted.

3. Require the purchaser of a microdata tape to sign a statement in which he or she agrees to abide by the law which states that the data may be used only for the purpose for which they were obtained, i.e., for statistical purposes.

Spruill, N. (1983). The Confidentiality and Analytical Usefulness of Masked Business Microdata. Proceedings of the Section on Survey Research Methods.

I. Introduction

Microdata must

A. Provide confidentiality to individual firms

B. Give reliable microdata analysis.

II. Issues

A. The Small Business Administration wanted to expand their data base to include microdata on taxes paid, depreciation, etc.--tax data was to come from IRS data.

masking had to be applied to this data at IRS before it could be released.

B. Masking techniques used

1. adding normal random error
2. multiplying by normal random error
3. grouping
4. random rounding
5. data swapping.

C. Zeros were preserved.

D. Economic analyses included examination of

1. summary statistics

a. means
b. standard deviations
c. percent zero

2. hypotheses tests about the correlation coefficient

3. multiple regression analyses

III. Measures of Confidentiality

A. Select a firm and mask its data

B. Match masked to original data using the sum of absolute deviations or squared deviations for all common variables.

C. If the firm that minimizes the sum is the same firm on which the release data is based, a link has been made.

D. The confidentiality criteria is the percent of released firms for which a link cannot be made.

IV. Results - Test Data

A. For small number of common variables all releasing strategies provide confidentiality.

B. For large number of common variables grouping provides confidentiality.

C. All strategies provide

1. good estimates of means
2. % zero

D. Estimate of standard deviation

1. adding and multiplying by normal noise increase it
2. grouping decreases it

E. Estimates of the correlation coefficient

1. adding and multiplying by normal noise decrease it
2. grouping increases it
3. random rounding have little effect on it
4. data swapping destroys many correlations

F. Regression coefficient estimation

1. adding random error, grouping, random rounding, and data swapping give close estimates
2. multiplying by random error destroys the coefficients

V. Results - Tax Data (27 variables analyzed)

A. Actual data tended to be bimodal or trimodal with little data between the modes. Adding noise resulted in putting more data between the modes.

B. Results are similar to test data for summary statistics.

C. Correlation and regression coefficients

The effect of masking on these statistics depend on p, the proportion of zero changed to nonzeroes.

D. Regression analysis

1. adding and multiplying by random error had various effects:

a. sometimes relations were maintained
b. sometimes destroying it
c. sometimes introducing it spuriously, and
d. sometimes not effecting it

2. grouping either

    a. destroyed relations or
    b. left them unchanged

3. random rounding had <u>least</u> effect on regression results.

Spruill, N.L. (1983). Protecting Confidentiality of Business Microdata by Masking. Working Paper, The Public Research Institute, Center for Naval Analyses, 2000 North Beauregard Street, Alexandria, VA 22311.

Summary

This paper's objectives are to look at masking strategies (adding random error, multiplying by random error, grouping, random rounding, and data swapping) and to propose measures of confidentiality to determine whether the released data violate confidentiality restrictions. These measures of confidentiality are based on the sum of squared deviations and the sum of absolute deviations between the variable values for the true and the released data.

Thornburn, D. (1983). On Methods for Disclosure Control in Longitudinal Studies. Statistisk Tidskrift (Statistical Review) 21(2), pp. 93-101.

Summary

In longitudinal studies the same subjects are studied at different time points. For privacy reasons, it may be unwise to retain individual data over the whole period of time studies. Different methods are discussed to protect the privacy of the subjects without losing all merits of longitudinal studies. This situation is compared to that leading to randomized response and some ideas are used from that technique. The issue of how tables may be constructed so that longitudinal studies may be facilitated is discussed. Methods are illustrated by examples in which reasonable protection can be achieved by a loss in efficiency of between 50% and 80%.

Dalenius, T. and Denning, D.E. (1982). A Hybrid Scheme for Release of Statistics. Statistisk Tidskrift (Statistical Review)

Summary

This article introduces the idea of publishing statistical moments of continuous data rather than the microdata itself. If $\{x_{ij}\}$ $j=1,\ldots,M$ denotes the microdata for case i and there are $i=1,\ldots,N$ records of microdata, the authors propose releasing finite moments of the form

$$A(p_1,\ldots,p_M) = \sum_{i=1}^{N} x_{11}^{p_1} x_{12}^{p_2} \ldots x_{1M}^{p_M}$$

where the exponents $p_1,\ldots,p_M$ are nonnegative integers. The computational feasibility for computing all A-moments subject to the constraint $p_1+\ldots+p_M = k$ for $k=0,1,\ldots,K$ and $K>0$.

Poole, W.K. and Clayton, C.A. (1982). Generalizations of a Contamination Model for Continuous Type Random Variables. Communications in Statistics - Theory and Methods, 11(15), 1733-1742.

Summary

This paper focuses on the application of randomized response type models for preservation of confidentiality in existing data files.

Through a randomization process cases in a file are randomly selected and variables in these cases are "scrambled" (i.e., disturbed).

Poole and Clayton show how to recover the original multivariate distribution from the scrambled multivariate data.

Spruill, N.L. and Gastwirth, J.L. (1982). On the estimation of the correlation coefficient from grouped data. Journal of the American Statistical Association, Vol. 77, No. 379, pp. 614-620.

Summary

This article proposes two estimators for the correlation coefficient, ρ, when statisticians will not construct a master file on individuals because of confidentiality issues. The approach depends on grouping the data according to values of one of the variables. Group means and total variance for both variables are required to calculate the estimators. The complete bivariate sample never needs to be compiled. The estimators are shown to be asymptotically normal. Asymptotic and Monte Carlo results are examined. For 1000 observations in 10 groups, these results indicate that the ratio of the MSE for one of the proposed estimators relative to that of the usual estimator of ρ ranges from .82 to .95 as ρ ranges from .9 to .25. The proposed estimators are more efficient than previous estimators for grouped data.

Cox, L.H. (1980). Suppression Methodology and Statistical Disclosure Control. Journal of the American Statistical Association, Vol. 25, No. 370, pp. 377-385.

Summary

This article discusses theory and method of complementary cell suppression and related topics in statistical disclosure control. Emphasis is placed on the development of methods that are theoretically broad but also practical to implement. The approach draws from areas of discrete mathematics and linear optimization theory.

Note

A synopsis of this paper is given in the review of Cox's (1984) preliminary report on "Disclosure Control Methods for Frequency Count Data."

Cox, L.H. (1980). Suppression Methodology and Statistical Disclosure Control. Journal of the American Statistical Association, 75(370), 377-385.

I. The Disclosure Problem and Cell Suppression

A. Effective disclosure control techniques reduce to an acceptable level the likelihood that either

1. a respondent may be identified through its responses or that

2. data collected from an identifiable respondent may be determined or narrowly estimated from the published data.

B. In this paper the disclosure problem is studied within the context of published data in aggregate form, only.

Disclosure of

1. categorical and
2. magnitude data are considered.

C. Structure of Cox's paper (by section)

1. Section 2: Describes how in a census or major survey the typical number of large tabulation cells and linear relations between them necessitate partitioning a single disclosure problem into a well defined sequence of interrelated subproblems.

2. Section 3: Discusses mathematical techniques for measuring cell sensitivity.

3. Section 4: Discusses the problem of verifying, in part through linear estimation, whether adequate disclosure protection is in fact provided. This is referred to as the validation phase of the disclosure control process.

4. Section 5: Complementary cell suppression is discussed. Under cell suppression methodology, all cells identified as disclosure cells (sensitive cells) are suppressed from publication, together with a sufficient number of additional nonsensitive cells, called complementary suppressions, to ensure that the values of only nonsensitive cells and cell combinations may be derived from the published cell data.

For example, if a cell representing data for a particular county is sensitive, then a sufficient number of additional cells at the county level within the given state will be suppressed complimentarily until no union of county cells containing this sensitive cell is sensitive. Analogously, subcells of this sensitive cell as well perhaps as other cells must be suppressed complimentarily to ensure that a narrow lower estimate of the value of this sensitive cell cannot be derived.

II. Sensitivity Measures and Cell Value Estimation

A. (n,k) dominance rule: defines a cell to be sensitive for a particular statistical item if n or fewer respondents constitute greater than k% of the total cell value for this item. The value of n is chosen to be larger than the size of coalitions of respondents persumed to exist. Such coalitions pool their individual confidential data for the purpose of determining the confidential data of a competitor.

B. Cox (1979) gives other measures of sensitivity for aggregated data.

Bulmer, M. (ed.) (1979). Censuses, Surveys and Privacy. London, England: Holmes and Meier Publishers.

Summary

This book is divided into four parts, each preceded by an introduction written by the editor. Within each of the parts there are sections or chapters written by a total of 12 persons or groups. The idea is to give the reader a look at privacy from several different points of view and to afford a historical perspective of issues involved.

The articles discuss the following issues:

1. the impact of privacy upon social research,
2. privacy in modern industrial society,
3. does research threaten privacy or does privacy threaten research?,
4. survey research and privacy,
5. the issue of privacy on public opinion research,
6. resistance to community surveys,
7. is the public acceptability of social survey research declining?,
8. census confidentiality in Britain,
9. invasion of privacy: the case of the U.S. Census,
10. public confidence in social survey research,
11. protecting statistical and research data from improper use, and
12. method for assuring personal integrity in social research.

Dalenius, T. (1979). Data-Swapping: A Technique for Disclosure Control when Releasing Micro-Statistics. Statistisk Tidskrift (Statistical Review) 17(4) pp. 253-258.

Summary

This paper introduces the method of data-swapping: attribute values from one record are "swapped" for those of another record.

Cox, L.H. (1979). Confidentiality Problems in Microdata Release. Proceedings of the 3rd Annual Symposium on Computer Applications in Medical Care, IEEE Comp. Soc., pp. 397-402.

I. Confidentiality Problems in Microdata

A. The two competing issues in microdata release are

1. the confidentiality issue and
2. the data utility issue

B. The confidentiality issue

1. reduce to an acceptable level the likelihood that an exact or likely match may be made between an alternative data source and an arbitrary microrecord from the microdata file. Such matching would result in a reasonably certain identification of the corresponding respondent, thereby disclosing all data in this respondent's microrecord.

2. absolute protection of respondent data cannot be guaranteed under microdata release.

3. the aim is to balance the risk of disclosure with the agency's responsibility to provide information in as broadly useful and inexpensive a manner as is practicable.

II. Typology and Sensitivity of Data Items

Ideally, the disseminator of microdata would guarantee that respondent identification through matching with an alternative data file had been prevented. In a practical sense, this level of anonymization is virtually impossible to achieve, primarily because the disseminating agency cannot control or predict the sophistication of alternative data files and matching techniques. However, it may be possible for the agency to ensure that no respondent from the microdata file may be identified through the analysis of the microdata file alone or in combination with a reasonable amount of additional information and resources.

III. Likely Privacy - Invasion Motives

A. To obtain confidential information about individuals, i.e., personal identification.

- a reasonable counter to this motive is to render this approach more expensive and least likely of success than other methods of personal identification and privacy invasion.

B. To show that "it can be done", i.e., a "fishing expedition"

- the purpose is to discredit the agency.
- one counter is to purge the microdata file of all salient microrecords which may be easily identifiable in a "fishing expedition".

IV. Protective Techniques

A. Inherent protective techniques

1. response error
2. time lag between collection and dissemination
3. population size
4. sample design

B. Active protective techniques

1. suppression of unique identifiers such as

   a. name
   b. social security number

2. reduction of respondent recognizability through

   a. use of population cutoffs : 250,000
   b. suppression of unique identifiers at lower level of geographic detail
   c. collapsing of definitional categories--usually occur at upper end of scale : topcode
   d. addition of statistical noise
   e. rounding
   f. replacement by means or midpoints of class intervals
   g. data swapping

3. removal of salient microrecords from the microdata file

V. Recommendations

Cox concludes by recommending data analysis on microdata files to try to assess the severity or possibility of privacy invasion.

Cox, L.H. and Sande, G. (1979). Techniques for Preserving Statistical Confidentiality. Proceedings of the 42nd Session of the International Statistical Institute, Manilla.

I. Introduction

Statistical disclosure characterized by instances in which condifential data pertaining to individual respondents might be divulged through the uncensored release of statistical products.

II. Record Linkage and Respondent Recognizability

A. Record linkage (conflation) is the problem of conflating two bodies of data or files by associating corresponding records between them.

B. Techniques for preserving statistical confidentiality are aimed at introducing ambiguities into the released data at or near the individual respondent level so that linkages between identifying and confidential data are either ambiguous or spurious, and to do so at minimal disruption to the accuracy and utility of derived tallys and statistical moments.

C. Foremost in any discussion of statistical disclosure is the concept of recognizability of individual respondents, i.e., the degree and accuracy with which an individual respondent can be identified from the released data. There are

1. direct identifiers (e.g., name, address, social security number) and

2. surrogate identifiers (e.g., unusual occupation, high income, geographic area).

If sufficiently accurate data are present for correct identification of a respondent and a good approximation of confidential data, and if it is also possible to correctly associate that data with the respondent, then statistical disclosure has occurred.

III. Statistical Disclosure in Microdata Release

A. Protective techniques for preserving respondent confidentiality are aimed at reducing to an acceptable level of probability that a microrecord may be unambigouosly linked to data from an alternative source in a manner which reveals both the identity and confidential data of an individual respondent, or ensuring such linkages would yield unreliable information

1. removal of all direct identifiers.

2. collapsing of definitional categories.

3. perturbing data to reduce reliability of derived linkages

a. random.

b. replacement of values with midpoints or means.

4. randomized response.

5. randomized release.

IV. Other Issues Reviewed

A. Statistical Disclosure in Frequency Count Tabulations.

B. Statistical Disclosure in Tabulations of Magnitude Data.

Ove, Frank (1979). Infering Individual Information from Released Statistics. Proceedings of the 42nd Session of the International Statistics Institute, Manilla. pp. 487-498.

I. Introduction

The complexity of the problem indicates that the probability models must necessarily be very crude and can be used merely as tools as deriving various privacy and disclosure measures which might be worth investigating further in practical disclosure control.

II. Disclosure of Individual Data

A. Let $U = \{1,\ldots,N\}$ be a population of N individuals labelled by integers $1,\ldots,N$ and let $C = \{1,\ldots,r\}$ be a set of r categories labelled by integers $1,\ldots,r$. We may think of the categories as the cells in a multidimensional contingency table. The category to which individual u belongs is denoted by x_u and the vector of individual data is denoted by $\underset{\sim}{x} = (x_1,\ldots,x_N)$. The parameter θ denotes background information.

B. The distribution of x_u conditional on θ may be summarized by the conditional entropy:

$$H(x_u|\theta) = \sum_{i=1}^{r} O(P(x_u=i|\theta))$$

where

$$O(p) = -p \log p \quad \text{if } p > 0$$

$$= 0 \qquad \text{if } p = 0$$

The conditional entropy can be viewed as a measure of privacy, since it may be interpreted as the logarithm of the number of categories which are likely alternatives for the individual considered.

C. Quantities of interest

1. Minimum expected conditional entropy of an individual category

$$\min_{u \in U} E\, H\,(x_u|\theta)$$

2. The proportion of individuals in U for which the expected conditional entropy is below a certain threshold, t.

3. The disclosure set X is the set of those individuals in U for which, conditional on θ, the distribution of x_u degenerates to a one category distribution, i.e.,

$$X = \{u \in U : H(x_u|\theta) = 0\}$$

$$= \{\text{zero entropy set}\}$$

: specifies a degenerate distribution.

4. The probability distribution of the disclosure set X is conveniently summarized by the entropy

$$H(X) = \sum_{A \subseteq U} O(p(A))$$

Dalenius, T. and Reiss, S. (1978). Data Swapping--A Technique for Disclosure Control. Proceedings of the Survey Research Methods Section of the American Statistical Association, pp. 191-194.

Discussed the procedure of data swapping.

U.S. Department of Commerce (1978). Statistical Policy Working Paper 2: Report on Statistical Disclosure and Disclosure Avoidance Techniques. Superintendent of Documents, U.S. Government Printing Office, Washington, DC 20402.

Summary

This report is about techniques for avoiding disclosure of confidential information about individuals in connection with the release of statistical tabulations and microdata files. The major issues addressed in this report are about

- the nature of statistical disclosure,
- the pervasiveness of statistical disclosure,
- how agency requirements can be translated into specific disclosure-avoidance techniques,
- how agency requirements can be met without unduly restricting data releases, and
- how do agency disclosure-avoidance practices affect data subjects and data users.

Dalenius, T. (1977). Towards a Methodology for Statistical Disclosure Control. Statistik Tidskrift (Statistical Review), No. 5, pp. 429-444.

I. Introduction. Purpose of the Paper

A. To define statistical disclosure

B. To present a theory of statistical disclosure

C. To give some examples

D. To give possible methodologies for statistical disclosure control (SDC).

II. Definition of Statistical Disclosure

If the release of the statistics S makes it possible to determine the value D_k (an attribute value of a specific person) more accurately than is possible without access to S, a statistical disclosure has taken place.

III. Theory of Statistical Disclosure

A. Accessibility of disclosure

1. If the disclosure is given by explicitly released statistics, the disclosure will be called direct.

2. If the disclosure calls for carrying out certain operations on S in order to generate additional statistics, S', then the disclosure of confidential information is indirect.

3. Accuracy of disclosure: disclosed information may be known approximately. For example, a disclosed attribute may be known to belong to an interval.

4. External vs. internal disclosure: If knowledge of the confidential attribute value for one unit enables you to determine the confidential attribute of another object more accurately, then the disclosure is called an internal disclosure.

5. S vs. SxE-based disclosure

Here E denotes "extra objective" information about the sampled objects.

If the computation of the confidential attribute value makes use of the published statistics, S, only, the disclosure is designated as S-based.

If access to E enables one to compute the confidential attribute value more accurately than with the use of S only, the disclosure is designated as an SxE-based disclosure.

Flaherty, D.H. (1977). Privacy and Government Data Banks: An International Perspective. London: Mansell.

Summary

Results are given of an investigation of the problems of privacy and confidentiality involved in the collection and dissemination of microdata by national statistical agencies or sections of administrative departments in the United Kingdom, Sweden, the Federal Republic of Germany, Canada, and the United States.

On the basis of extensive interviews and reviews and agency files and publications in the five countries over the 4-year period of the investigation, during which each country was visited at least five times, the author argues that the custodians of microdata maintain a high level of concern for protection of confidentiality. Further, the author believes that more should be done, within the boundaries of maintaining confidentiality, to satisfy demands for access to personal data for various research and statistical purposes. The author also states that publication of aggregate statistics and the provision of special tabulations do not fully satisfy the needs of researchers in a number of social sciences.

IV. Examples--Microstatistics

"Disclosure in the context of microstatistics may be defined with reference to the notion of "equal vectors. ...If the microstatistics released from some area comprises one or more unique vectors, a disclosure has taken place."

V. Towards a Methodology for Statistical Disclosure Control

A. It is necessary to discard the notion of elimination of disclosure.

1. Such a goal is not operationally feasible

2. Elimination of disclosure is possible only by elimination of statistics

3. "it is not possible to protect against disclosure by collusion" - Hansen, 1971.

B. Measures required

1. The amount of disclosure associated with the release of some statistics and extra objective data

2. The benefit associated with the statistics released

3. Criterion: maximize the benefit subject to the constraint that disclosure must be no greater than a certain level.

C. Means of control

General purpose means for SDC

1. Training of statisticians--in disclosure analyses, for example.

2. Use of sampling

- leads to less detailed statistics

- reduces disclosure by collusion

Newman, Dennis (1975). Techniques for Ensuring the Confidentiality of Census Information in Great Britain. Meeting of the International Statistical Institute and the International Association of Survey Statisticians. Invited paper no. 71. Warsaw, Poland.

Summary

This paper discusses the method of random perturbation used by the Office of Population Censuses and Surveys of Great Britain.