29

Rapid #: -2430351

Ariel IP: 128.253.70.20

Status	Rapid Code	Branch Name	Start Date
Pending	GZM	Memorial Library	3/25/2009 7:35:13 AM

CALL #: **HA1523 A42**

LOCATION: **GZM :: Memorial Library :: memorial library stacks regular size shelving**

TYPE:	Article CC:CCL
JOURNAL TITLE:	Statistisk tidskrift
USER JOURNAL TITLE:	Statistisk tidskrift. Statistical review.
GZM CATALOG TITLE:	Statistisk tidskrift =
ARTICLE TITLE:	Towards a Methodology for Statistical Disclosure Control
ARTICLE AUTHOR:	Dalenius, T.
VOLUME:	15
ISSUE:	
MONTH:	
YEAR:	1977
PAGES:	429-444
ISSN:	0039-7261
OCLC #:	1639866
CROSS REFERENCE ID:	692462
VERIFIED:	

BORROWER: **COO :: Olin**

PATRON: **Abowd,John**

PATRON ID:	
PATRON ADDRESS:	
PATRON PHONE:	
PATRON FAX:	
PATRON E-MAIL:	john.abowd@cornell.edu
PATRON DEPT:	Labor Economics
PATRON STATUS:	F - Faculty
PATRON NOTES:	

RAPID ILL

This material may be protected by copyright law (Title 17 U.S. Code)
System Date/Time: 3/25/2009 7:54:38 AM MST

This material may be protected by copyright law (Title 17 U.S. Code).

3. References

[1] Andrew, A. M.: "A variant of modulus 11 checking", The Computer Bulletin (August 1970).

[2] Black, L. W.: "Error Detection in Decimal Numbers", Proceedings of the IEEE, Vol 60, No 3, March 1972, pp. 331—332 (The Institute of Electrical and Electronics Engineers).

[3] Blake, I. F. and Mullin, R. C.: "The Mathematical Theory of Coding", Academic Press, New York 1975.

[4] Block, H.: "En ny kontrollsiffermetod", Data 1-2 1977.

[5] Felme, S.: "Feltyper och felfrekvenser vid dataregistrering", Intern rapport, SCB 1976-04-28, 19 pages.

[6] Felme, S.: "Kontrollsiffror och personnummer", Intern rapport, SCB 1976-12-23, 39 pages.

[7] Freeman: "Detection of Transposition Errors in Decimal Numbers", Proceedings of the IEEE, August 1967, Vol 55, No 8, pp. 1500—1501.

[8] Larsen, L.: "Optimale Kontrollciffersystemer", Data 11-1967, pp. 14—17.

[9] Peterson, W. W. and Weldon, E. J.: "Error Correcting Codes", Second edition, MIT press, Cambridge, Massachusetts, 1972.

[10] Private Correspondense Riskalla and Taylor, 1976-08-12, "Foundation for Check Digit System's Evaluation".

[11] Tang, D. T. and Lum, V. Y.: "Error control for Terminals with human Operators", IBM Journal of research and development, Vol 14, No 4, July 1970, pp. 409—416.

[12] Taylor, A.: "Taylor Report", Computer World. Note in particular the reports of 1975-09-17, 1975-10-22, 1975-12-31, 1976-04-19, 1976-08-09, 1976-11-08 and a letter from John Beidler 1975-12-03.

Towards a methodology for statistical disclosure control

by Tore Dalenius[1]

A. INTRODUCTION

1. The problem of statistical disclosure—then and now

The term "statistical disclosure"—typically referred to simply as disclosure—will be used in this paper[2] in accord with its use in the context of releasing results (tabulations, microdata, etc.) of sample and census surveys.

The phenomenon of disclosure attracted the attention of survey statisticians long before the present era of public concern about "invasion of privacy". By the same token, survey statisticians early took special actions to *control*[3] disclosure, as evidenced by special statutes, regulations and policy statements. As an example in kind, Title 13, (U.S. Code), which deals with the work of the U.S. Bureau of the Census, dates back to 1929.

In recent years, some events have, however, occurred which have made it urgent to strengthen the efforts to control disclosure. Thus, one decisive event is represented by the lively public debate about various threats to the citizens' privacy; the proliferation of computerized information system has no doubt served to enhance the public concern about statistical information systems. One of the threats identified in this debate is indeed *disclosure*.[4]

While survey statisticians have shown their understanding of the public concern about disclosure, they have also emphasized the risk of an oversimplified debate of the disclosure problem. More specifically, they have pointed to two shortcomings of the debate:

i. Some cases of alleged disclosure have proved to have no or very little support in facts.[5]

ii. Many critics fail to discuss the problem of

[1] Brown University and University of Stockholm.

[2] This paper is virtually identical with report No. 19 of the research project Confidentiality in Surveys, financed by a generous grant from the Bank of Sweden Tercentenary Foundation.

[3] For reasons which will be touched upon in section 18, "control" is used in preference to "prevention" or "avoidance".

[4] Another decisive event is represented by the change that has taken place in the field statistics, with respect to the volume and detail of statistics produced, thus enhancing the risks for accidental disclosures; see Dalenius (1974).

[5] The following citation is a case in kind; it is taken from Miller (1971), p. 136:
"Some deficiencies inevitably crop up even in the Census Bureau. In 1963, for example, it reportedly provided the American Medical Association with a "statistical" list of one hundred and eighty-eight doctors residing in Illinois. The list was broken down into more than two dozen income categories, and each category was further subdivided by medical speciality and area of residence; as a result, identification of individual doctors was possible. In addition, there probably has been a fair amount of data disclosed at the information-gathering level by the large corps of enumerators employed to carry out the periodic canvassing. It is difficult to believe that all census takers are immune from gossiping or impervious to the entreaties by one neighbour for information concerning the replies of another. Of course, if direct-mail techniques prove successful, this type of abuse should be reduced."
Federal statisticians who have thoroughly investigated this specific case, have been unable to substantiate Miller's criticism!

disclosure in the context of a reasoned balance between the right to privacy and the need to know; they may also lump together *all* instances of disclosure, be they serious or harmless.

2. The purpose of this paper

It is the prime purpose of this paper to contribute to a better understanding of the phenomenon of disclosure. The achievement of this purpose should help the survey statisticians to cope more successfully with today's disclosure problem, and hopefully provide a basis for an *informed* public debate.

With this purpose in mind, the paper has been organized as follows:

In part B, we will suggest a *definition* of "statistical disclosure".

In part C, we will present a *theory* of statistical disclosure.

In parts D—F, we will give some examples.

In part G, finally, we consider the possibility of developing a *methodology for statistical disclosure control*, SDC for short.

B. THE CONCEPT OF "STATISTICAL DISCLOSURE"

3. The insufficiency of prevailing definitions

Statistical disclosure is used in the literature in a way which parallels its use in non-statistical contexts. Thus, in Webster's Third New International Dictionary, "disclosure" is defined as:

(1) the act or an instance of opening up to view, knowledge or comprehension

(2) something that is disclosed.

This definition is, indeed, general; it is by and large consistent with definitions of disclosure in the context of releases of statistical results. As an example, Title 13, (U.S. Code), Section 9-a-2, gives an implicit definition of disclosure; it states that there shall not be:

"... any publication whereby the data furnished by any particular establishment or individual under this title can be identified."

The definition just quoted is less general than the definition taken from Webster's dictionary, by making *identification* of the object(s) concerned an element of the definition. While this is indeed a crucial difference, it does not make the resulting definition sufficiently specific to serve as a basis for regulations and/or procedures aiming at disclosure control; it does not easily and unambiguously lend itself to implementation.

In section 4—6 an effort will be made to deal with the conceptual problem thus present.

4. A framework for defining "statistical disclosure"

As stated in section 1, "statistical disclosure" is used here in accord with the use of this term in the context of releasing statistics from a survey. In line with this notion of disclosure, the following four components are used to provide the conceptual framework called for:

i. A frame comprising certain objects
ii. Data associated with these objects
iii. Statistics released from a survey
iv. Extra-objective data

4.1. The frame

Consider a set of identifiable objects, to be referred to as the total population and denoted by $\{O\}_T$. In a typical case, $\{O\}_T$ may be "all Swedish citizens".

The survey concerns a subset of this total population, viz. that subset which is accessible by means of a certain frame F; for convenience, this subset will be denoted by $\{O\}_F$. In a specific case, $\{O\}_F$ may be "Swedish citizens living in Sweden".

The complimentary subset—i.e., the subset made up by objects in $\{O\}_T$ which are not in $\{O\}_F$—is denoted by $\{O\}_{\bar{F}}$. Thus $\{O\}_T$ is the "union" of $\{O\}_F$ and $\{O\}_{\bar{F}}$:

$$\{O\}_T = \{O\}_F \cup \{O\}_{\bar{F}} = \{O\}_F + \{O\}_{\bar{F}}$$

In the case of a *sample* survey, it may prove useful to make an additional distinction, viz. between objects selected for the sample, say $\{O\}_{F,s}$ and those not selected, $\{O\}_{F,\bar{s}}$.

4.2. Data associated with the objects in the frame

With each object in $\{O\}_F$, we associate data, which serves three different *functions*:

i. *Identifying function:*
We will denote the data serving this function by the identifier I. In a specific case, I may appear as a (registration) number, or as name and street address.

ii. *Classifying function:*
For purposes of presenting the "details" of the statistics to be released, the objects in $\{O\}_F$ will be associated with certain classes, defined by reference to some classifier C. In a specific case, C may appear as a "code" identifying a subset of $\{O\}_F$, for example a subset defined with reference to the sex and age of the objects in $\{O\}_F$.

iii. *Information function:*
The survey is carried out in order to provide information in terms of certain "survey characteristics" $X, Y, \ldots, Z$. For the object O_J in $\{O\}_F$, $J = 1, \ldots, N$, the values of these characteristics are denoted by $X_J, Y_J, \ldots, Z_J$. Typically (but not exclusively), these values may

statistical results. As an example, Title 13, (U.S. Code), Section 9-a-2, gives an implicit definition of disclosure; it states that there shall not be:

"... any publication whereby the data furnished by any particular establishment or individual under this title can be identified."

The definition just quoted is less general than the definition taken from Webster's dictionary, by making *identification* of the object(s) concerned an element of the definition. While this is indeed a crucial difference, it does not make the resulting definition sufficiently specific to serve as a basis for regulations and/or procedures aiming at disclosure control; it does not easily and unambiguously lend itself to implementation.

In section 4—6 an effort will be made to deal with the conceptual problem thus present.

4. A framework for defining "statistical disclosure"

As stated in section 1, "statistical disclosure" is used here in accord with the use of this term in the context of releasing statistics from a survey. In line with this notion of disclosure, the following four components are used to provide the conceptual framework called for:

i. A frame comprising certain objects
ii. Data associated with these objects
iii. Statistics released from a survey
iv. Extra-objective data

4.1. The frame

Consider a set of identifiable objects, to be referred to as the total population and denoted by $\{O\}_T$. In a typical case, $\{O\}_T$ may be "all Swedish citizens".

The survey concerns a subset of this total population, viz. that subset which is accessible by means of a certain frame F; for convenience, this subset will be denoted by $\{O\}_F$. In a specific case, $\{O\}_F$ may be "Swedish citizens living in Sweden".

The complimentary subset—i.e., the subset made up by objects in $\{O\}_T$ which are not in $\{O\}_F$—is denoted by $\{O\}_{\bar{F}}$. Thus $\{O\}_T$ is the "union" of $\{O\}_F$ and $\{O\}_{\bar{F}}$:

$$\{O\}_T = \{O\}_F \cup \{O\}_{\bar{F}} = \{O\}_F + \{O\}_{\bar{F}}$$

In the case of a *sample* survey, it may prove useful to make an additional distinction, viz. between objects selected for the sample, say $\{O\}_{F,s}$ and those not selected, $\{O\}_{F,\bar{s}}$.

4.2. Data associated with the objects in the frame

With each object in $\{O\}_F$, we associate data, which serves three different *functions*:

i. *Identifying function:*
We will denote the data serving this function by the identifier I. In a specific case, I may appear as a (registration) number, or as name and street address.

ii. *Classifying function:*
For purposes of presenting the "details" of the statistics to be released, the objects in $\{O\}_F$ will be associated with certain classes, defined by reference to some classifier C. In a specific case, C may appear as a "code" identifying a subset of $\{O\}_F$, for example a subset defined with reference to the sex and age of the objects in $\{O\}_F$.

iii. *Information function:*
The survey is carried out in order to provide information in terms of certain "survey characteristics" $X, Y, \ldots, Z$. For the object O_J in $\{O\}_F$, $J = 1, \ldots, N$, the values of these characteristics are denoted by $X_J, Y_J, \ldots, Z_J$. Typically (but not exclusively), these values may be in the nature of *attributes* or *magnitudes*.

It may be worth noting that some data may serve more than one function in one and the same survey.

4.3. The statistics released from the survey

The *objective* of a survey is expressed in terms of some population and some data C and $X, Y, \ldots, Z$. In order to achieve this objective, the statistics S is released.

We will focus on two different kinds of statistics:

i. statistics for sets of objects—"*macrostatistics*"; typically, the format of a report is used as the means of releasing the statistics.
ii. statistics for individual objects—"*microstatistics*"; typically, the format of microdata tape is used as the means of releasing the statistics.

In view of the role that the above distinction plays in part C, we will elaborate upon it in sections 4.3.1 and 4.3.2.

4.3.1. Macrostatistics

In the case of macrostatistics, the statistics—counts, magnitudes, etc. as the case may be—concern aggregates of the individual values of the survey characteristics belonging to the respective sets.

The follwing tables are two cases in kind:

Number of beneficiaries by county and age

County	Age class				Total
	Under 65	65—69	70—74	75 & over	
A	3	15	11	8	37
B	7	60	34	20	121
C	—	4	—	—	4

Average benefit amount (in $) by county and age

County	Age class			
	Under 65	65—69	70—74	75 & over
A	63.30	94.30	85.20	79.60
B	62.40	89.90	81.80	72.40
C	59.80	92.40	80.40	77.60

These tables—while featuring the characteristics of real-life statistics—are admittedly "small". In the interest of making clearer the discussion in part C, we will further reduce the size by focusing on some detail.

4.3.2. Microstatistics

In this kind of statistics, the individual values observed with respect to the characteristics $X, Y, \ldots, Z$ (possibly in conjunction with the associated classifiers) are released. The identifiers, however, are *not* released.

The following excerpt — slightly edited — from U.S. Bureau of the Census (1976) is illustrative:

Household data:	State of residence	Urban/ Rural	Size of household	Telephone	Plumbing	Rent	No. of cars	Household type
Household No. 1	Virginia[1]	Urban	3	Yes	Yes	$125	2	H-W family
Individual Data:	Relationship	Sex	Age	Race	Place of Birth	Years of School	Occupation	Earnings
Person a	Husband	M	37	W	Kansas	12	Plumber	$13,000
b	Wife	F	35	W	Virginia	12		
c	Child	M	6	W	Virginia	1		
Household No. 2	Virginia	Rural	1	Yes	No	$30	0	Primary Indiv.
Person a	Primary	F	68	NW	Alabama	6	Service	$1,400
Household No. 3	Virginia	Urban	6	Yes	Yes	$205	2	H-W family
Person a	Husband	. . . . .						
.	.							
.	.							
f	.							

[1] Public Use Sample tapes do not actually contain alphabetic information, but represent the characteristics in the form of numeric codes.

4.4. Extra-objective data

In section 4.3, we related the *objective* of a survey to two kinds of data: C, and $X, Y, \ldots, Z$, respectively. It is characteristic of the design of a survey that it provides a source of this data.

We will use the term "extra-objective data" to denote any kind of *additional* data; for convenience, this data will be denoted by E. It is characteristic of E that it is not part of the objective of the survey; thus, the design does not explicitly provide a source of this data.

4.5. Summary

Thus, the four components of the framework may now be stated as:

(1) The frame: $\{O\}_F$

(2) The data associated with the objects in the frame: $I; C; X, Y, \ldots, Z$

(3) The statistics released from the survey: S

(4) The extra-objective data: E

5. Statistical disclosure defined

We will now suggest a definition of disclosure within the conceptual framework presented in section 4.

Thus, consider an object O_K in $\{O\}_T$. This object may be a member of $\{O\}_F$, or it may be a member of $\{O\}_{\bar{F}}$. We introduce a characteristic D, which may be one of the survey characteristics $X, Y, \ldots, Z$; or it may be some other characteristic. For the object O_K, this characteristic assumes the value D_K. It is helpful to consider two special cases:

i. $D_K = 1$ if O_K has a certain property, otherwise $D_K = 0$.

ii. D_K is measured on a ratio scale: it is expressed as a magnitude.

If the release of the statistics S makes it possible to determine the value D_K more accurately than is possible without access to S, a disclosure has taken place; more exactly, a D-disclosure has taken place. In a specific case, this D-disclosure may be an X-disclosure, or an Y-disclosure, etc.

The definition just given applies to both release of micro-statistics and release of macro-statistics.

C. A THEORY OF STATISTICAL DISCLOSURE

6. The basic approach

In order to be able to develop a methodology for SDC, it is necessary to understand the disclosure phenomenon. More specifically, it is necessary to construct a theory which

further reduce the size by focusing on some detail.

4.3.2. Microstatistics

In this kind of statistics, the individual values observed with respect to the characteristics $X, Y, \ldots, Z$ (possibly in conjunction with the associated classifiers) are released. The identifiers, however, are *not* released.

The following excerpt — slightly edited — from U.S. Bureau of the Census (1976) is illustrative:

le-one	Plumbing	Rent	No. of cars	Household type
es	Yes	$125	2	H-W family
ace	Place of Birth	Years of School	Occupation	Earnings
'	Kansas	12	Plumber	$13,000
'	Virginia	12		
'	Virginia	1		
es	No	$30	0	Primary Indiv.
W	Alabama	6	Service	$1,400
s	Yes	$205	2	H-W family
. . .				

ain alphabetic information, but represent the

E. It is characteristic of E that it is not part of the objective of the survey; thus, the design does not explicitly provide a source of this data.

4.5. Summary

Thus, the four components of the framework may now be stated as:

(1) The frame: $\{O\}_F$

(2) The data associated with the objects in the frame: $I; C; X, Y, \ldots, Z$

(3) The statistics released from the survey: S

(4) The extra-objective data: E

5. Statistical disclosure defined

We will now suggest a definition of disclosure within the conceptual framework presented in section 4.

Thus, consider an object O_K in $\{O\}_T$. This object may be a member of $\{O\}_F$, or it may be a member of $\{O\}_{\bar{F}}$. We introduce a characteristic D, which may be one of the survey characteristics $X, Y, \ldots, Z$; or it may be some other characteristic. For the object O_K, this characteristic assumes the value D_K. It is helpful to consider two special cases:

i. $D_K = 1$ if O_K has a certain property, otherwise $D_K = 0$.

ii. D_K is measured on a ratio scale: it is expressed as a magnitude.

If the release of the statistics S makes it possible to determine the value D_K more accurately than is possible without access to S, a disclosure has taken place; more exactly, a D-disclosure has taken place. In a specific case, this D-disclosure may be an X-disclosure, or an Y-disclosure, etc.

The definition just given applies to both release of micro-statistics and release of macro-statistics.

C. A THEORY OF STATISTICAL DISCLOSURE

6. The basic approach

In order to be able to develop a methodology for SDC, it is necessary to understand the disclosure phenomenon. More specifically, it is necessary to construct a theory which reflects the underlying cause system.[6] This actualizes a basic question, which we will try to answer in this section, viz.: what kind of approach is one to use?

In looking for a feasible approach, it is natural to try to identify some characteristic of the disclosure phenomenon which somehow makes the underlying cause system accessible, and then give this characteristic a pivotal role in the choice of *the* approach. In the case under consideration, the considerable *complexity* of the disclosure phenomenon is indeed such a characteristic.

As elaborated in Ashby (1970), considerations of operational feasibility often make it imperative to tackle complex problems by first breaking them up into minor, less complex problems, and then dealing with these problems in turn. In this context, such a "scientific" approach calls for two major steps:

i. the development of a *typology* of statistical disclosure which provides a basis for dividing the overall disclosure problem into a set of sub-problems which can more easily be understood; and

ii. the analysis of each such sub-problem from a *causal* point of view.

7. A typology of statistical disclosure

In what follows, we will present a typology of disclosure which makes use of 6 dimensions:

i. kinds of statistics S released: micro-statistics, or macro-statistics;

ii. the measurement scale used to express S; in what follows, we will focus on scales yielding attributes/counts and yielding magnitudes respectively;

[6] This is the appropriate place to quote the saying: "There is nothing as practical as a good theory."

iii. accessibility of disclosure: direct or indirect disclosure;
iv. accuracy of disclosure: exact disclosure or approximate disclosure;
v. scope of disclosure: external or internal disclosure;
vi. the disclosing entities: S- or $S \times E$-disclosure.

A typology based on these 6 dimensions yields a classification with (at least) $2^6 = 64$ categories. The two first-mentioned dimensions (kinds of statistics, and measurement scale) have already been introduced in section 4. The four remaining dimensions will be discussed in section 7.1—7.4, respectively. The discussion will be tied to D_K as defined in section 5. The disclosure will be denoted by D_K^*.

7.1. Accessibility of disclosure

If D_K^* is explicitly given by the released statistics, the disclosure will be called *direct*.

If, however, the computation of D_K^* calls for carrying out certain operations on S in order to generate an additional statistics S', then D_K^* will be called *indirect*.

7.2. Accuracy of disclosure

The term "accuracy of disclosure" will be used here in a sense which is best described by means of an example.

Consider the characteristic "age" X_K for the object O_K. We adopt the convention of denoting X_K the *exact* age if X_K equals the age of O_K as of his last birthday.

Now, if $D_K^* = X_K$, we will refer to D_K^* as *exact* disclosure; in the case where X denotes age, we will refer to D_K^* as exact age disclosure. Otherwise, we will refer to D_K^* as *approximate* disclosure.

There may be several *types* of approximation, for example:

i. approximation by means of *interval*, as exemplified by

$$X_L \leqslant D_K^* \leqslant X_U$$

for the characteristic X.

ii. approximation in terms of a *category*, as exemplified by:

$$D_K^* = 0$$

for "living in the urban area", and

$$D_K^* = 1$$

for "living in the rural area".

An interval/category approximation may be labeled "*certain*" if the object concerned does in fact belong to the interval/category involved; otherwise it will be labeled "*uncertain*". An important special case of an uncertain approximation is provided by a probabilistic approximation when the statement "O_K belongs to the interval [;] or the category [.]" is associated with a probability P that the statement is true.[7]

7.3. External v. internal disclosure

Consider two objects, O_J and O_K, with D-values D_J and D_K respectively; without loss of generality, it may be assumed that O_J and O_K are members of a set of objects, for which S has been released.

If D_K^* can be computed without information about D_J, then we have a case of *external* disclosure. This designation reflects the fact that D_K^* can be computed by someone who is not a member of the same set as O_K.

If information about D_J makes it possible to compute an approximation D_K^{**} which is 'closer' to D_K than D_K^*, then D_K^{**} represents *internal* disclosure. The choice of this designation reflects that fact that the disclosure will be restricted to members of the set to which O_J and O_K belong. Clearly, internal disclosure (and only internal disclosure) can take place *by itself* (i.e. without external disclosure taking place).

[7] This formulation may be made rigorous as follows. Assume that it is known that O_K belongs to a set of objects of which a proportion P belongs to the interval [;] or the category [.]; then the probability is P that an object selected at random from this set will be an object which does in fact belong to [;] or [.]. For a further elaboration of this notion, reference is given to Cassel (1976).

7.4. S- v. S×E-based disclosure

In section 4, E was introduced to denote "extra-objective" data.

If the computation of D_K^* makes use of S only (but *not* E), the disclosure will be designated as *S-based* disclosure.

If access to E makes it possible to compute an approximation which is closer to D_K than is an approximation which uses S only, then we have a case of $S \times E$-*based* disclosure.

8. The subsequent discussion

In the following sections, we will discuss in more detail a few of the 64 types of disclosure identified according to the typology just outlined.

The discussion will—as mentioned in section 4.3—make use of oversimplified cases of statistics released; the emphasis of the discussion is on *mechanisms* of disclosure.

D. EXAMPLES—MACROSTATISTICS: COUNTS[8]

9. External, S-based disclosure

In this section, we will consider both exact and approximate disclosures, as well as direct and indirect disclosures.

9.1. Exact direct disclosure

Example No. 1

Consider the following table of people

[8] The *possibility* of extracting additional information from count statistics is well known; for illustrations, reference is given to Bishop et al. (1975), pp. 107—11, Fisher (1935), pp. 94—95, and Yule and Kendall (1950), chapter 1.

exemplified by

$$X_L \leqslant D_K^* \leqslant X_U$$

for the characteristic X.

ii. approximation in terms of a *category*, as exemplified by:

$$D_K^* = 0$$

for "living in the urban area", and

$$D_K^* = 1$$

for "living in the rural area".

An interval/category approximation may be labeled "*certain*" if the object concerned does in fact belong to the interval/category involved; otherwise it will be labeled "*uncertain*". An important special case of an uncertain approximation is provided by a probabilistic approximation when the statement "O_K belongs to the interval [;] or the category [.]" is associated with a probability P that the statement is true.[7]

7.3. External v. internal disclosure

Consider two objects, O_J and O_K, with D-values D_J and D_K respectively; without loss of generality, it may be assumed that O_J and O_K are members of a set of objects, for which S has been released.

If D_K^* can be computed without information about D_J, then we have a case of *external* disclosure. This designation reflects the fact that D_K^* can be computed by someone who is not a member of the same set as O_K.

If information about D_J makes it possible to compute an approximation D_K^{**} which is 'closer' to D_K than D_K^*, then D_K^{**} represents *internal* disclosure. The choice of this designation reflects that fact that the disclosure will be restricted to members of the set to which O_J and O_K belong. Clearly, internal disclosure (and only internal disclosure) can take place *by itself* (i.e. without external disclosure taking place).

7.4. S- v. S× E-based disclosure

In section 4, E was introduced to denote "extra-objective" data.

If the computation of D_K^* makes use of S only (but *not* E), the disclosure will be designated as *S-based* disclosure.

If access to E makes it possible to compute an approximation which is closer to D_K than is an approximation which uses S only, then we have a case of $S\times$ *E-based* disclosure.

8. The subsequent discussion

In the following sections, we will discuss in more detail a few of the 64 types of disclosure identified according to the typology just outlined.

The discussion will—as mentioned in section 4.3—make use of oversimplified cases of statistics released; the emphasis of the discussion is on *mechanisms* of disclosure.

D. EXAMPLES—MACROSTATISTICS: COUNTS[8]

9. External, S-based disclosure

In this section, we will consider both exact and approximate disclosures, as well as direct and indirect disclosures.

9.1. Exact direct disclosure

Example No. 1

Consider the following table of people classified by place of living and health status; the assumption is made that "health status" is meaningfully measured only by a nominal scale.

County	Health status class				Total
	1	2	3	4	
C	—	4	—	—	4

This is an example of exact, direct disclosure: the table shows immediately that all people in county C have a health status X expressed by the value $X = 2$. The "cause" of the disclosure is obvious: the margin for county C equals one of its health status cells.

Example No. 2

This example illustrates an important point: each one of a set of tables $T_1, \ldots, T_k$ may by itself be (relatively) harmless; in combination, however, they may be seriously disclosing.

Consider a survey which for county C has yielded the following statistics:

County	Age group	No. of men	No. of women
C	0	5	—
	1	—	50

To be sure, this table discloses that all men in county C are in age group "0".

Now suppose that, in addition, the following table is released:

County	Age group	No. with criminal record	No. without criminal record
C	0	5	—
	1	—	50

In summary:

i. all persons in age group "0" have a criminal record

ii. all persons in age group "0" are men

[7] This formulation may be made rigorous as follows. Assume that it is known that O_K belongs to a set of objects of which a proportion P belongs to the interval [;] or the category [.]; then the probability is P that an object selected at random from this set will be an object which does in fact belong to [;] or [.]. For a further elaboration of this notion, reference is given to Cassel (1976).

[8] The *possibility* of extracting additional information from count statistics is well known; for illustrations, reference is given to Bishop et al. (1975), pp. 107—11, Fisher (1935), pp. 94—95, and Yule and Kendall (1950), chapter 1.

iii. no man is in age group "1"

Thus, it follows that all men in county C have a criminal record!

9.2. Exact indirect disclosure

Example No. 3

Consider the following table; it is analogous to that used in example No. 1:

County	Urban area Health status class				Total Health status class			
	1	2	3	4	1	2	3	4
C	15	11	6	5	15	12	6	5

This table makes it possible to compute the following table:

County	Rural area Health status class			
	1	2	3	4
C	—	1	—	—

The derived table discloses that the only person in the rural area in county C belongs to health status class 2.

The need to derive this table justifies the term "*indirect*" disclosure.

9.3. Approximate direct disclosure

Example No. 4

This is a variation of example No. 1: four geographic categories take the place of the health status classes:

County	Geographic category				Total
	NW	NE	SE	SW	
C	—	4	—	—	4

The disclosure is in the nature of approximation in terms of a *category*.

Example No. 5

Assume the following table is published from a survey dealing with "tax cheating":

County	Per cent tax-cheaters
C	95

If we select a person at random from county C and state that he is a tax-cheater (as defined for this survey), there is a prior probability $P = .95$ that the statement turns out to be true!

9.4. Approximate indirect disclosure

The construction of an example is straightforward, but will not be undertaken here.

10. External, S × E-based disclosure

The "extra-objective" data E may be of a variety of kinds. We will give two examples.

10.1. Exact direct disclosure

Example No. 6

Consider the total set of objects:

$$\{O\}_T = \{O\}_F + \{O\}_{\bar{F}}$$

as defined in section 4, and the object O_K, which in fact belongs to $\{O\}_{\bar{F}}$.

A survey is made of the objects in $\{O\}_F$. If the documentation of the survey design discloses that O_K is not in $\{O\}_F$ (= E), O_K must be in $\{O\}_{\bar{F}}$. Depending upon the kind of survey, this may be stigmatizing. For example, $\{O\}_F$ may be the set of objects which have filed an "acceptable" income tax return form; this implies that those who have *not* filed such a form constitute the set $\{O\}_{\bar{F}}$, to which O_K belongs.

An analogous example may be formulated in terms of the subsets $\{O\}_{F,S}$ and $\{O\}_{F,\bar{S}}$ defined in section 4.1.

10.2. Approximate direct disclosure

Example No. 7

This example considers again the total set of objects

$$\{O\}_T = \{O\}_F + \{O\}_{\bar{F}}$$

and an object O_J, which in fact belongs to $\{O\}_F$.

The survey of $\{O\}_F$ shows that a proportion P has some characteristic, say being "tax-cheater". If the documentation of the survey design discloses that O_J is in $\{O\}_F$, a disclosure of the kind discussed in example No. 5 has occurred.

11. Internal disclosure

We will be satisfied with one example.

Example No. 8

Assume the following table has been released for number of persons on welfare:

County	No. of persons		Total
	on welfare	not on welfare	
C	48	2	50

O_J and O_K are both not on welfare; if they know each other, they can conclude that *all* remaining 48 persons in county C are on welfare!

E. EXAMPLES—MACROSTATISTICS: MAGNITUDES

12. External, S-based disclosure

The discussion will be parallel to that in section 9.

12.1. Exact direct disclosure

Example No. 9

This is a parallel to example No. 1:

County	Number of establishments in industry				Total sales
	1	2	3	4	
C	—	—	—	1	$100,000

a survey dealing with "tax cheating":

County	Per cent tax-cheaters
C	95

If we select a person at random from county C and state that he is a tax-cheater (as defined for this survey), there is a prior probability $P = .95$ that the statement turns out to be true!

9.4. Approximate indirect disclosure

The construction of an example is straightforward, but will not be undertaken here.

10. External, S × E-based disclosure

The "extra-objective" data E may be of a variety of kinds. We will give two examples.

10.1. Exact direct disclosure

Example No. 6

Consider the total set of objects:

$$\{O\}_T = \{O\}_F + \{O\}_{\bar{F}}$$

as defined in section 4, and the object O_K, which in fact belongs to $\{O\}_{\bar{F}}$.

A survey is made of the objects in $\{O\}_F$. If the documentation of the survey design discloses that O_K is not in $\{O\}_F$ $(= E)$, O_K must be in $\{O\}_{\bar{F}}$. Depending upon the kind of survey, this may be stigmatizing. For example, $\{O\}_F$ may be the set of objects which have filed an "acceptable" income tax return form; this implies that those who have *not* filed such a form constitute the set $\{O\}_{\bar{F}}$, to which O_K belongs.

An analogous example may be formulated in terms of the subsets $\{O\}_{F,S}$ and $\{O\}_{F,\bar{S}}$ defined in section 4.1.

10.2. Approximate direct disclosure

Example No. 7

This example considers again the total set of objects

$$\{O\}_T = \{O\}_F + \{O\}_{\bar{F}}$$

and an object O_J, which in fact belongs to $\{O\}_F$.

The survey of $\{O\}_F$ shows that a proportion P has some characteristic, say being "tax-cheater". If the documentation of the survey design discloses that O_J is in $\{O\}_F$, a disclosure of the kind discussed in example No. 5 has occurred.

11. Internal disclosure

We will be satisfied with one example.

Example No. 8

Assume the following table has been released for number of persons on welfare:

County	No. of persons		Total
	on welfare	not on welfare	
C	48	2	50

O_J and O_K are both not on welfare; if they know each other, they can conclude that *all* remaining 48 persons in county C are on welfare!

E. EXAMPLES—MACROSTATISTICS: MAGNITUDES

12. External, S-based disclosure

The discussion will be parallel to that in section 9.

12.1. Exact direct disclosure

Example No. 9

This is a parallel to example No. 1:

County	Number of establishments in industry				Total sales
	1	2	3	4	
C	—	—	—	1	$100,000

5–770001. Statistisk tidskrift 1977: 5

12.2. Exact indirect disclosure

This type of disclosure parallels that discussed in section 9.2; as its meaning is obvious, we will not elaborate on it by way of an example.

12.3. Approximate direct disclosure

Example No. 10

Consider the following income statistics:

County	Income class, $			Total
	—1,999	2,000—4,999	5,000—	
C	—	4	—	4

12.4. Approximate indirect disclosure

The construction of an example is straightforward but will not be carried out here.

13. External, S × E disclosure

This type of disclosure parallels that discussed in section 10.

13.1. Exact direct disclosure

We will be satisfied here with one example.

Example No. 11

The following table has been released:

County	No. of beneficiaries	Total amount of benefits, $
C	15	3,000

Thus, in county C, the 15 beneficiaries receive all together $3,000 or $200 per beneficiary. If it is known that the maximum amount given to any beneficiary is $200 $(= E)$, then obviously *every* beneficiary in C receives $200!

13.2. Approximate direct disclosure

Example No. 12

A survey of firms deals with two characteristics:

Y: the value of production
X: the number of workers

The statistics released may be a $r \times c$ table, giving the number of firms in each cell defined in terms of Y and X.

If the structure of the relation between Y and X is known, for example by way of a regression function $Y = f(X)$, and the X-value is known for some firm then it is possible to estimate the corresponding Y-value by means of the regression function.

14. Internal disclosure

We will be satisfied with one example.

Example No. 13
Consider the following table:

County	No. of companies	Total sales, $
C	5	100,000

Company O_J in county C has a total sale of $80,000. Thus the remaining 4 companies account for $20,000; for any other company O_K, it can be safely stated that:

$$D_K^* \leqslant \$20{,}000$$

F. EXAMPLES—MICROSTATISTICS

15. A model of microstatistics

The notion of microstatistics was illustrated in section 4.3.2 by means of an example from the U.S. Bureau of the Census. In order to make our discussion general in scope, we will use the following representation.

Consider a case with r data: C; X, Y, . . ., Z. For simplicity, we will assume that these data assume two values[9] only: 1 and 0.

[9] The generalization to the case where some or all data assume more than two values is straightforward.

Thus, the values associated with the object O_J may be given by a vector:

$$V_J\text{: } 1, 0, 0, \ldots, 1$$

By definition, two vectors are equal:

$$V_J = V_K$$

if *all* elements of these vectors are pairwise equal (both 1 or both 0); otherwise,

$$V_J \neq V_K$$

If $V_J \neq V_K$ for all $K \neq J$, then the vector V_J is said to be *unique*.

16. The disclosure problem

The statistics released for an individual object O_J does not contain the identifier I_J as defined in section 4.2. This does not, however, mean that O_J cannot be identified—the possibility of identifying O_J is exactly the disclosure problem[10] in the context of microstatistics. In order to illuminate this point, we will return to the illustration given in section 4.3.2.

Example No. 14
For this example, $r = 16$, corresponding to

$$2^{16} = 65{,}536$$

different vectors; not all of these vectors may be associated with some object.

Let us consider the problem of identifying (by name and adress) the husband in household No. 1. We already know the following:

	Datum	Value of datum
(1)	State of residence	Virginia
(2)	Urban/rural	Urban
(3)	Occupation	Plumber

Consider now some additional values:

[10] The disclosure problem as discussed here may be viewed as a special case of the disclosure problem associated with the release of macrostatistics. We will not elaborate on this aspect here.

	Datum	Value of datum
(4)	Telephone	Yes
(5)	Automobiles	Yes
(6)	Sex	Male
(7)	Age	37
(8)	Race	White

These 5 data (4)—(8) may reasonably be assumed to be "public data", or at least easily accessible. It seems reasonable to assume that the constellation of values corresponding to data (1)—(8) is unique, or at any rate that there are very few objects having this specific constellation. Consequently, an effort to identify the husband in household No. 1 may appear to be feasible.

Disclosure in the context of microstatistics may be defined with reference to the notion of "equal vectors" in section 16. If the microstatistics released for some area comprises one or more unique vectors, a disclosure has taken place.

A word of caution is called for here. Assume that V_J *is* unique; thus, a disclosure *has* taken place. This fact says *nothing* about the physical/economic effort necessary in order to 'pin-point' O_J. In fact, trying to 'pin-point' O_J may prove to be *operationally infeasible* in a real-life situation.

It is close at hand to ask a question such as the following one: "Which frequency of 'equal vectors' can one expect in a given release of microstatistics?" The answer will clearly depend upon such factors as:

i. the number N of objects in an area with identical vectors;
ii. the number r of data released for each object; increasing r will tend to reduce the frequency;
iii. the dependence between the data; this factor may be discussed in terms of the frequency of objects with $X = 1$, $Y = 1$, etc.

Thus, the values associated with the object O_J may be given by a vector:

$$V_J: 1, 0, 0, \ldots, 1$$

By definition, two vectors are equal:

$$V_J = V_K$$

if *all* elements of these vectors are pairwise equal (both 1 or both 0); otherwise,

$$V_J \neq V_K$$

If $V_J \neq V_K$ for all $K \neq J$, then the vector V_J is said to be *unique*.

16. The disclosure problem

The statistics released for an individual object O_J does not contain the identifier I, as defined in section 4.2. This does not, however, mean that O_J cannot be identified—the possibility of identifying O_J is exactly the disclosure problem[10] in the context of microstatistics. In order to illuminate this point, we will return to the illustration given in section 4.3.2.

Example No. 14

For this example, $r = 16$, corresponding to

$$2^{16} = 65{,}536$$

different vectors; not all of these vectors may be associated with some object.

Let us consider the problem of identifying (by name and adress) the husband in household No. 1. We already know the following:

Datum	Value of datum
(1) State of residence	Virginia
(2) Urban/rural	Urban
(3) Occupation	Plumber

Consider now some additional values:

Datum	Value of datum
(4) Telephone	Yes
(5) Automobiles	Yes
(6) Sex	Male
(7) Age	37
(8) Race	White

These 5 data (4)—(8) may reasonably be assumed to be "public data", or at least easily accessible. It seems reasonable to assume that the constellation of values corresponding to data (1)—(8) is unique, or at any rate that there are very few objects having this specific constellation. Consequently, an effort to identify the husband in household No. 1 may appear to be feasible.

Disclosure in the context of microstatistics may be defined with reference to the notion of "equal vectors" in section 16. If the microstatistics released for some area comprises one or more unique vectors, a disclosure has taken place.

A word of caution is called for here. Assume that V_J *is* unique; thus, a disclosure *has* taken place. This fact says *nothing* about the physical/economic effort necessary in order to 'pin-point' O_J. In fact, trying to 'pin-point' O_J may prove to be *operationally infeasible* in a real-life situation.

It is close at hand to ask a question such as the following one: "Which frequency of 'equal vectors' can one expect in a given release of microstatistics?" The answer will clearly depend upon such factors as:

i. the number N of objects in an area with identical vectors;
ii. the number r of data released for each object; increasing r will tend to reduce the frequency;
iii. the dependence between the data; this factor may be discussed in terms of the frequency of objects with $X = 1$, $Y = 1$, etc.

G. TOWARDS A METHODOLOGY FOR SDC

17. Retrospect and prospect

In part B of this report, we have suggested a definition of statistical disclosure.

In part C, we have presented a theory of statistical disclosure.

In parts D—F we have presented some examples.

It remains to answer the question if these elements can be integrated into a methodology for SDC which—while operationally feasible—is compatible with the overall objective of striking a reasoned balance between "the right to privacy" and "the need to know", to use the succinct wording in Barabba (1975).

In the present part G we will give a partial answer to the question just cited. More specifically, we will discuss two important aspects of a methodology for SDC, viz.:

i. the *criterion* problem; and
ii. the *techniques* for control.

The discussion will show that while considerable progress has been made towards the development of an SDC-methodology, much more remains to be done.

18. The criterion problem

The use of any methodology for SDC must be guided by considerations of what is desirable with respect to both focus and level of the control.

A reasonable starting point is to discard the notion of *elimination* of disclosure. Two arguments for doing so are:

i. it would be unrealistic to aim at elimination: such a goal is not operationally feasible;
ii. it would place unreasonable restrictions on the kind of statistics that can be released; it may be argued that elimina-

[10] The disclosure problem as discussed here may be viewed as a special case of the disclosure problem associated with the release of macrostatistics. We will not elaborate on this aspect here.

tion of disclosure is possible only by elimination of statistics.

What has just been said is in fact the reason for our use of the term "statistical disclosure *control*" rather than "prevention" or "avoidance", which have also been suggested.

Next, it seems necessary to make a minor concession with respect to "disclosure by collusion". The case with collusion involving only two objects may clearly be dealt with by some such rule as "do not release results for cells with 3 or less objects"; the case with collusion involving a 'small' number of objects may be dealt with in a similar manner. Collusion involving 'many' objects may, however, be operationally untractable. As pointed out in Hansen (1971), p. 52, the U.S. Bureau of the Census accepts the view that "it is not feasible to protect against disclosure by collusion".

The problem of developing a criterion function for use in the context of SDC may now be formulated as the problem of specifying a function, which depends on the amount of disclosure and the benefits of the release. More specifically, we need two measures:

i. $M = M(S_i, E)$, the amount of disclosure associated with the release of some statistics S_i $(i = 1, 2, \ldots, k)$ and the extra-objective data E; and
ii. $B = B(S_i)$, the benefit associated with the statistics S_i.

It would then be possible—in principle—to use a criterion of the following type:

$$\text{Maximize } B \text{ for } M = M_0$$

where M_0 is some accepted level of disclosure.

The construction of realistic measures M and B will for sure not be easy. The difficulties may, however, not be insurmountable: they are, by and large, similar to the difficulties in the realm of total survey design, an area where significant progress has indeed been made in the last two decades. We will be satisfied here to point to two areas, where some preliminary work has been done.

(1) In Bing (1972) and Turn (1976), the notion of "sensitivity" of data is discussed; what is an "acceptable" disclosure should depend on the "sensitivity" of the data involved.
(2) Very little is as yet known about the public's attitudes to a variety of issues in this area. Some efforts are, however, being made to remedy this situation; a study under the auspices of the Committee on National Statistics, the National Academy of Sciences, is worth special mention (Goldfield (1976)).

19. The means of control—systematics

Experience has shown that the survey statistician has some options when looking for means of control. It is helpful to consider two classes of options:

i. general-purpose means; and
ii. special-purpose means.

We will discuss these classes in section 20 and 21 respectively.

20. General-purpose means for SDC

This is a broad and heterogeneous class of means for SDC, among which we will briefly focus on two:

i. training of the statisticians; and
ii. use of sampling rather than complete coverage.

20.1. Training of the statisticians

Some statistical agencies have today special training programs, which aim at improving the statisticians' skill in identifying potential disclosures and coping with such cases. Making 'disclosure analysis' of the tabulations etc. a routine is likely to achieve a similar effect.

20.2. Use of sampling

There are several reasons why use of sampling rather than complete coverage may have a beneficial impact. We will list four specific reasons:

1. Use of sampling may release personal and other resources for SDC, which else would be used up for other purposes.
2. Use of sampling leads typically to the release of less detailed statistics.
3. Use of sampling means by necessity the release of estimates instead of results from a complete coverage thus adding an element of approximation to the statistics released.
4. Use of sampling reduces the options for 'disclosure by collusion'.

21. Special-purpose means for SDC

'Special-purpose means' are means which are tailored to one or a few of the types of disclosure identified in part C, section 7.

It is beyond the scope of this paper to discuss the availibility of means for each one of the (at least) $2^6 = 64$ types of disclosure identified; such a discussion will not be undertaken for two reasons:

i. making the inventory of the technical literature called for would represent a volume of work which is beyond the scope of this paper; it is better carried out in the context of developing a *manual* for SDC;
ii. assessing if a certain technique is or is not "suitable" is premature: too little is as yet known in this respect from the field of applications.

We will be satisfied by giving an overview of techniques available. This overview will

able: they are, by and large, similar to the difficulties in the realm of total survey design, an area where significant progress has indeed been made in the last two decades. We will be satisfied here to point to two areas, where some preliminary work has been done.

(1) In Bing (1972) and Turn (1976), the notion of "sensitivity" of data is discussed; what is an "acceptable" disclosure should depend on the "sensitivity" of the data involved.

(2) Very little is as yet known about the public's attitudes to a variety of issues in this area. Some efforts are, however, being made to remedy this situation; a study under the auspices of the Committee on National Statistics, the National Academy of Sciences, is worth special mention (Goldfield (1976)).

19. The means of control—systematics

Experience has shown that the survey statistician has some options when looking for means of control. It is helpful to consider two classes of options:

i. general-purpose means; and
ii. special-purpose means.

We will discuss these classes in section 20 and 21 respectively.

20. General-purpose means for SDC

This is a broad and heterogeneous class of means for SDC, among which we will briefly focus on two:

i. training of the statisticians; and
ii. use of sampling rather than complete coverage.

20.1. Training of the statisticians

Some statistical agencies have today special training programs, which aim at improving the statisticians' skill in identifying potential disclosures and coping with such cases. Making 'disclosure analysis' of the tabulations etc. a routine is likely to achieve a similar effect.

20.2. Use of sampling

There are several reasons why use of sampling rather than complete coverage may have a beneficial impact. We will list four specific reasons:

1. Use of sampling may release personal and other resources for SDC, which else would be used up for other purposes.
2. Use of sampling leads typically to the release of less detailed statistics.
3. Use of sampling means by necessity the release of estimates instead of results from a complete coverage thus adding an element of approximation to the statistics released.
4. Use of sampling reduces the options for 'disclosure by collusion'.

21. Special-purpose means for SDC

'Special-purpose means' are means which are tailored to one or a few of the types of disclosure identified in part C, section 7.

It is beyond the scope of this paper to discuss the availibility of means for each one of the (at least) $2^6 = 64$ types of disclosure identified; such a discussion will not be undertaken for two reasons:

i. making the inventory of the technical literature called for would represent a volume of work which is beyond the scope of this paper; it is better carried out in the context of developing a *manual* for SDC;
ii. assessing if a certain technique is or is not "suitable" is premature: too little is as yet known in this respect from the field of applications.

We will be satisfied by giving an overview of techniques available. This overview will be carried out in sections 21.1—21.3, parallelling the organization of the discussion in part C.

21.1. Macrostatistics: Counts

In example No. 1, the disclosure occurs because the count in the "detail cell" (health status 2, county C) equals the count in the "total cell".

A disclosure problem of this kind may be dealt with in various ways:

i. by combining health status classes; combining classes 1 and 2 would not eliminate the disclosure but would have a beneficial impact on the accuracy of disclosure;
ii. by combining two or more counties, "rolling-up".

The disclosure problem in example No. 3 may be dealt with in a similar fashion. Other possibilities are:

iii. by adding "noise" to some/all cells, by way of "random rounding", "random perturbation", etc.;
iv. by cell suppression.

The discussion above concerns exact disclosure. Problems of *approximate* disclosure must clearly be dealt with in a way which takes into account the type of approximation, as discussed in section 7.2.

If, for example, we are facing a problem of probabilistic disclosure, as in example No. 5, we may use a criterion according to which the disclosure is acceptable if

$$P_L < P < P_U$$

where P_L is a lower limit and P_U is an upper limit. If P does not fall in this interval, we may adhere to the use of some technique discussed above.

$S \times E$-based disclosure may easily prove to be a much more serious problem than S-based disclosure: the statistician may not

know about E, or—if he does—he may not have the authority to control it.

Example No. 6 offers a challenge. The straightforward means of control would be to suppress (or 'dilute') those parts of the documentation of the survey, which play an instrumental role for the disclosure. Doing so would, however, in many instances limit the usefulness of the survey, and perhaps seriously so.

Closely related to this issue is the question: What to do about the rules used to suppress information, should they be published or not? Many statisticians do in fact recommend that the rules are not being published, as part of the disclosure control.

The problem of internal disclosure, finally, as discussed in example No. 8, may be dealt with by means of a rule according to which the count in any cell must not be less than a critical number. In example No. 8, this number might be at least 4, corresponding to a "rule of 4".

21.2. Macrostatistics: Magnitudes

Many of the disclosure problems in this area may be dealt with along the same lines as those discussed in section 21.1. We will therefore be satisfied here by considering some specific cases.

The problem of approximate disclosure in example No. 10 may be dealt with by changing the income classes: if the range of the class "2,000—4,999" is too short, it may be made wider, for example "1,000—4,999" or "2,000—5,999", or a combination such as "1,500—5,499".

The U.S. Bureau of the Census has used an interesting "rule of thumb" when dealing with magnitudes: the size classes used for releasing a magnitude X are determined as follows:

$$X_L = .75X;\ X_U = 1.5X$$

that is, the upper limit is twice the lower limit.

The problem of internal disclosure illustrated by example No. 13 may be dealt with by a rule according to which no object must account for more than a fraction Q of the total; it remains, of course, to choose the "proper" Q-value!

21.3. Microstatistics

We venture the judgement that this is the area where it will prove to be most difficult to get the disclosure problem under control; this is due to the structure of the disclosure problem, as discussed in section 16 with reference to three factors governing the frequency of "equal vectors".

21.4. Some additional references

The discussion in sections 21.1—21.3 has drawn upon several references not explicitly given, but listed in section 23, references.

H. CLOSING SECTIONS

22. Acknowledgements

In the course of writing this paper, I have had the privilege of working with and for the Interagency Committee on Statistical Methodology of the Statistical Policy Division, Office of Management and Budget, and particularly with its Subcommittee on Confidentiality Issues. This contact has played an instrumental role in forming my ideas about how to develop a methodology for statistical disclosure control, the main theme of this paper. I want especially to acknowledge the constructive help that I have received from Mr. Thomas B. Jabine, Chairman of the Subcommitte, and Mr. Richard A. Bell, one of its other members.

23. References

23.1. References cited in the paper

Ashby, W. R., Analysis of the system to be modeled. In: Stogdice, R. M., Ed., The process of model-building. Ohio State University Press, Columbus, Ohio, 1970.

Barabba, V. P., The right to privacy and the need to know. In: U.S. Bureau of the Census: A numerator and denominator for measuring change. Technical Paper 37. Government Printing Office, Washington, D.C. 1975.

Bing, J., Classification of personal information with respect to sensitivity aspect. In: Selmer, K. S., Ed., Data banks and society. Proceedings of the first international Oslo symposium on data banks and society. Universitetsförlaget, Oslo-Bergen-Tromsø, 1972.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W., Discrete Multivariate Analysis: Theory and Practice. The MIT Press, Cambridge, Massachusetts, and London, England, 1975.

Cassel, C.-M., On probability based disclosures in frequency tables. Mim. report. National Central Bureau of Statistics, Stockholm, Sweden, 1975.

Dalenius, T., The invasion of privacy problem and statistics production—an overview. Statistisk tidskrift, 1974, 213—225.

Fisher, R. A., The design of experiments. Oliver and Boyd, Edinburgh and London, 1935.

Goldfield, E., Personal information. 1976.

Hansen, M. H., The role and feasibility of a national data bank, based on matched records, and alternatives. Chapter 1 in Federal Statistics, Report of the President's Commission, Vol. II. Government Printing Office, Washington, D.C., 1971.

that is, the upper limit is twice the lower limit.

The problem of internal disclosure illustrated by example No. 13 may be dealt with by a rule according to which no object must account for more than a fraction Q of the total; it remains, of course, to choose the "proper" Q-value!

21.3. Microstatistics

We venture the judgement that this is the area where it will prove to be most difficult to get the disclosure problem under control; this is due to the structure of the disclosure problem, as discussed in section 16 with reference to three factors governing the frequency of "equal vectors".

21.4. Some additional references

The discussion in sections 21.1—21.3 has drawn upon several references not explicitly given, but listed in section 23, references.

H. CLOSING SECTIONS

22. Acknowledgements

In the course of writing this paper, I have had the privilege of working with and for the Interagency Committee on Statistical Methodology of the Statistical Policy Division, Office of Management and Budget, and particularly with its Subcommittee on Confidentiality Issues. This contact has played an instrumental role in forming my ideas about how to develop a methodology for statistical disclosure control, the main theme of this paper. I want especially to acknowledge the constructive help that I have received from Mr. Thomas B. Jabine, Chairman of the Subcommitte, and Mr. Richard A. Bell, one of its other members.

23. References

23.1. References cited in the paper

Ashby, W. R., Analysis of the system to be modeled. In: Stogdice, R. M., Ed., The process of model-building. Ohio State University Press, Columbus, Ohio, 1970.

Barabba, V. P., The right to privacy and the need to know. In: U.S. Bureau of the Census: A numerator and denominator for measuring change. Technical Paper 37. Government Printing Office, Washington, D.C. 1975.

Bing, J., Classification of personal information with respect to sensitivity aspect. In: Selmer, K. S., Ed., Data banks and society. Proceedings of the first international Oslo symposium on data banks and society. Universitetsförlaget, Oslo-Bergen-Tromsø, 1972.

Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W., Discrete Multivariate Analysis: Theory and Practice. The MIT Press, Cambridge, Massachusetts, and London, England, 1975.

Cassel, C.-M., On probability based disclosures in frequency tables. Mim. report. National Central Bureau of Statistics, Stockholm, Sweden, 1975.

Dalenius, T., The invasion of privacy problem and statistics production—an overview. Statistisk tidskrift, 1974, 213—225.

Fisher, R. A., The design of experiments. Oliver and Boyd, Edinburgh and London, 1935.

Goldfield, E., Personal information. 1976.

Hansen, M. H., The role and feasibility of a national data bank, based on matched records, and alternatives. Chapter 1 in Federal Statistics, Report of the President's Commission, Vol. II. Government Printing Office, Washington, D.C., 1971.

Miller, A. R., The assault on privacy—computers, data banks and dossiers. University of Michigan Press, Ann Arbor, Michigan, 1971.

Turn, R., Classification of personal information for privacy protection purposes. P-5652 The Rand Corporation, Santa Monica, California, 1976. Also in: AFIPS Conference Proceedings, National Computer Conference—1976 AFIPS Press, Montvale, N. J., 1976, 301—307.

U.S. Bureau of the Census, Background paper: Policy on public-use microdata. Prepared in the Data Uses Service Division for discussion at the September 30, 1976 meeting of the Census Advisory Committee of the American Statistical Association. Bureau of the Census, Washington, D.C., 1976.

Yule, G. U. and Kendall, M. G., An introduction to the theory of statistics. Charles Griffin and Co., Ltd., London, 1950.

23.2. Some additional references

Cox, L. H., Statistical disclosure in publication hierarchies. Report No. 14 of the research project Confidentiality in Surveys. Department of Statistics, University of Stockholm, Stockholm, 1976.

Fellegi, I. P., On the question of statistical confidentiality. Journal of the American Statistical Association, 1972, 7—18.

Frank, O., Reconstruction of individual data from classificational frequency distributions, Mim. report. Department of Statistics, University of Uppsala, 1973.

Hoffman, L. J. and Miller, W. F., Getting a personal dossier from a statistical data bank. Datamation, May 1970, 74—75. Also in: Hoffman, L. J., Ed., Security and privacy in computer systems. Melville Publishing Company, Los Angeles, California 1973 289—293.

Jabine, T. B. and Bell, R. A., Guidelines for preventing disclosure in tabulations of program data. Mimeographed draft. Social Security Administration, Office of Research and Statistics, Washington, D.C., 1976.

Sweden, National Central Bureau of Statistics, Confidentiality in statistical tables. National Central Bureau of Statistics, Stockholm, 1974.

Korstabulering med utnyt
inform

av avdelningsdirekt

Denna uppsats ger en kort beskrivning av vad som avses med "utnyttjande av supplementär information" vid korstabulering i frekvenstabeller. Vidare ges några referenser till aktuella metoder. Slutligen ges några exempel på variansvinster, då tilläggs-(supplementär) information utnyttjas.

1. Inledning

Korstabulering av två variabler x och y är en vanlig metod för att i en statistisk undersökning beskriva sambandet mellan dessa variabler. Ofta görs korstabuleringen utifrån sådana urvalsobjekt, för vilka värden på båda variablerna är kända. I vissa situationer har man emellertid tillgång till ytterligare information om de variabler som skall korstabuleras. Denna supplementära information kan ibland utnyttjas vid korstabuleringen, dels för att få vissa tabellvärden att överensstämma med annan statistik från t. ex. en totalundersökning och dels i syfte att öka precisionen i de erhållna tabellvärdena.

2. Exempel på situationer med supplementär information

Ett exempel på en situation med supplementär information är en undersökning, i vilken värdet på variabeln x registreras hos alla objekt i en population, medan värdet på variabeln y endast registreras för ett urval av objekt. Vid korstabuleringen kan man då inte endast utnyttja informationen från de