Disclosure-Limited Data Dissemination*

GEORGE T. DUNCAN and DIANE LAMBERT **

Statistical agencies use a variety of disclosure control policies with ad hoc justification in disseminating data. The issues involved are clarified here by showing that several of these policies are special cases of a general disclosure-limiting (DL) approach based on predictive distributions and uncertainty functions. A user's information posture regarding a target is represented by one predictive distribution before data release and another predictive distribution after data release. A user's lack of knowledge about the target at any time is measured by an uncertainty function applied either to the current predictive distribution or to the current predictive distribution and the previously held predictive distribution. Common disclosure control policies, such as requiring released cell relative frequencies to be bounded away from both zero and one, are shown to be equivalent to disclosure rules that allow data release only if specific uncertainty functions at particular predictive distributions exceed a limit. Data transformations, such as aggregation and cell suppression, that are intended to reduce the extent of disclosure are analyzed in simple but realistic scenarios.

KEY WORDS: Disclosure control; Confidentiality; Aggregation; Cell suppression; Dominance; Uncertainty functions; Predictive distributions.

1. INTRODUCTION

Free information sharing and confidentiality protection are two major themes guiding data collection and dissemination (e.g., see Flaherty 1979). From the statistical organization's standpoint, without adequate access to data, decision making is poorly based, and without adequate assurance of confidentiality, voluntary reporting is likely lessened. From the individual's standpoint, freedom of information is a mainstay of democracy, and the right of privacy is held dearly. Data dissemination policy should balance the demands of open access and the demands of strict confidentiality through effective disclosure-limiting procedures.

Public and private decisions are often directed from the factual base that data from statistical agencies provide. In a decentralized statistical system (such as the U.S. federal government maintains through some 108 federal offices at a cost in 1983 of about $1.35 billion) the sharing of individual agency data and sampling frames can, it has been argued, reduce duplication of effort (Clark and Coffey 1983; Mugge 1978, 1983). On the other hand, the willingness of individuals and organizations to participate in data collection activities about sensitive issues depends on assurances that the data will not be released with harm to them (Singer 1978).

Concerns of individuals have also received attention. The Freedom of Information Act (FOIA) of 1966 (as amended in 1974 and 1976) is directed toward extending public access to government information. On the other hand, rights of privacy have been well articulated. Article 12 of the United Nations Universal Declaration of Human Rights, for example, asserts that the privacy of no one shall be subjected to arbitrary interference. In the same spirit, a 1970 federal law (Title 20 USC 1232g) requires public schools, colleges, and universities to obtain consent of students or their parents to release student data for nonacademic purposes (such as any form of scientific research). Reynolds (1979), based on an analysis of 24 codes of ethics for the conduct of social science research, stated,

> No norm related to the conduct of research with human subjects is as well established as the norm of maintaining confidentiality of information about individual research participants. (p. 385)

This fact and a regard for rights of privacy have motivated procedures to ensure confidentiality of certain data records under certain circumstances. Informed consent forms, for example, typically contain some statement about confidentiality. For a randomized clinical trial of an interferon ointment for the treatment of recurrent herpes, such a consent form stipulated,

> Although the data will be confidential, we can not guarantee confidentiality to all participants during this study. I do understand that my research records, just like hospital records, may be subpoenaed by court order or may be inspected by the sponsor of this study or federal regulatory authorities. In the event that publishable data were obtained from this study, the identity of the participants will be kept confidential. (Ho et al. 1982, p. 3)

As another example, controversy provoked by an American Council of Education survey on characteristics of college students, which included measures of political orientation and activism, led to the development of techniques for minimizing the possibility that such data can be associated with specific individuals (Boruch 1971).

Disclosure control is a major responsibility for those data collection agencies whose primary mission includes dissemination of statistical information. In the United States, these include the Census Bureau, the Bureau of Labor Statistics, and the National Center for Health Statistics, among others. It is also a major concern to those organizations that routinely collect highly sensitive data that is useful in the research and policy-making functions of other organizations. These include the Social Security Administration (Alexander and Jabine 1978) and the Internal Revenue Service (Wilson and Smith 1983). Indeed, the earliest statute for federal government use of private information was the Internal Revenue Act of 1864.

The design and implementation of statistical disclosure controls by federal agencies is mandated by various legislative acts. The Privacy Act of 1974 (P. L. 93-579) requires federal government agencies to formulate rules for the dissemination of data to protect the confidentiality of respondents to government

*Reprinted with permission from Journal of the American Statistical Association, Copyright 1986, by the American Statistical Association, Vol. 81, No. 393, March 1986, pp. 10-18.

surveys. Volume 5 of the U.S. Code Section 552a(b)(5) of this act permits disclosure only "to a recipient who has provided the agency with advance adequate written assurance that the record will be used solely as statistical research or reporting record, and the record is to be transferred in a form that is not individually identifiable." For business establishments, the 1948 Trade Secrets Act added "confidential statistical data" to the categories of information protected under the criminal code. The tension between these confidentiality acts and FOIA has required court rulings. The Bureau of Labor Statistics' policies, for example, based on the Trade Secrets Act, were found to support an FOIA exemption in a 1981 case (Clark and Coffey 1983). Given the inherent inefficiencies of litigation, it is useful to explore administrative approaches to resolving this tension. Such administrative approaches give guidelines that enable data release while limiting disclosure of confidential information.

Agencies now take a variety of precautions to limit disclosure. For example, in the National Center for Health Statistics,

> Of course, all direct identifiers of study objects, such as name, address, and social security number, are deleted from the public use files. Still, there are so many different items of information about any subject individual or establishment in our typical surveys that the set of information could serve as a unique identifier for each subject, if there were some other public source for many of the survey items. Fortunately there is not. But to minimize the chance of disclosure we take additional precautions: we make sure there are no rare characteristics shown on any case in the files, such as the exact bed-size of a large nursing home, or the exact date of birth of a subject, or the presence of a rare disease, or the exact number of children in a very large family. We either delete or encrypt the code identifying smaller geographic areas—places smaller than 100,000 in population—because anyone trying to identify a respondent will have his task greatly simplified if he knows the respondent's local area. (Mugge 1983, p. 7)

A lack of procedures for protecting confidentiality has precluded data collection in some cases. A study of draft evaders who fled to a neutral country was never conducted because researchers were unable to convince the potential respondents that anonymity could be assured (Sagarin 1973). In West Germany, the Constitutional Court decided in April 1983 to postpone the census after more than 1,000 lawsuits had been brought against the census.

Before ending this brief discussion of the issues surrounding disclosure control, we indicate our own views in the words of Dalenius (1977a):

> It is clearly not satisfactory to aim at "maximum protection of individual privacy". If such an objective could be achieved, it would necessarily mean that we would deprive ourselves of the benefits of statistical programs which may serve as a basis for improving our living conditions. Instead it is imperative to formulate the criterion problem as one of striking a reasoned balance between the individual's right to protection against invasion of privacy and the society's need to know. (p. 207)

2. PROBLEM FORMULATION

Our framework is built on a definition of disclosure proposed by Dalenius (1977b), recommended by the Subcommittee on Disclosure Avoidance Techniques (1978), and discussed in Jabine, Michael, and Mugge (1977):

> If the release of the statistic S makes it possible to determine the (microdata) value more accurately than is possible without access to S, a disclosure has taken place . . . . (p. 6)

There are two important implications of this definition. The first is that a data user can combine the released information with whatever information was previously available to gain information about the microdata value, which we call the target. The second is that, since almost any data release provides new information about the target, total avoidance of disclosure is impossible. At best, the extent of disclosure can be controlled so that it is below some acceptable level.

In order to quantify the extent of disclosure, the information that a data user has before and after data release must be modeled. We choose to express the data user's beliefs about a sensitive target value before data release as a probability distribution. After an agency releases a statistic S without individual identifiers, the user updates the probability distribution. We call these distributions *predictive*, since they can be used to predict the target value.

Although the use of probability distributions may seem unnecessarily mathematical, a sensitive target can best be shielded by assuming that the data user combines the available information about the target optimally. Therefore, as shown in DeGroot (1970), the agencies should act as if the data user's current state of knowledge about the target is expressed as a probability distribution over the possible values of the target and as if it is revised according to Bayes's rule when new data are released.

Since the predictive distribution fully expresses the user's knowledge about the target, characteristics of the predictive distribution before and after data release indicate the extent of disclosure to the user. The characteristics of the predictive distribution that we utilize are expressed as uncertainty functions. Predictive distributions and uncertainty functions are more fully discussed in Section 4. We support disclosure procedures that limit the extent of disclosure by requiring that constraints on the posterior uncertainty functions be satisfied. Generically, we call these procedures *disclosure limiting* (DL). Examples of disclosure-limiting procedures that imply various disclosure control policies discussed by the Subcommittee on Disclosure Avoidance Techniques (1978) are analyzed in Sections 4 and 5. For now we note that, operationally, the constraints would require iterative adjustments, as they were found to be too lax to protect confidentiality or too stringent to allow the inferences needed to guide policy. Finally, some implications of the model for disclosure control techniques such as aggregation are discussed in Section 6.

Probabilistic models of disclosure have been suggested previously by Dalenius (1974) and pursued by Cassel (1976) and Frank (1978, 1979, 1982). Cassel's model is suggestive of ours but more restricted in its applicability. Frank's model is not stated in terms of the predictive distribution of the target but in terms of the set of individuals whose characteristics are known precisely. That is, Frank models the likelihood of exact disclosure. Frank (1982) contrasted models based on the likelihood of exact disclosure with those based on the predictive distribution of the target.

3. THE PREDICTIVE APPROACH

The U.S. Internal Revenue Service (IRS) makes available certain statistical information compiled from tax returns. For

example, the IRS publication *Statistics of Income: Estate Tax Returns 1976* (IRS 1979) supplies data on 1977 estates. For that year, there were 11 individuals with gross estates exceeding $10 million and their gross estates totaled $507.862 million. Suppose an individual was known to have a gross estate over $10 million. To what extent does knowledge that the total for 11 estates was just over $500 million allow a data user to assess too narrowly the individual's gross estate?

An agency concerned with the impact of releasing these data might describe a data user who has no present access to the data in this way: The user is certain that a particular estate declared in 1977 had a value X_1 exceeding $10 million. But the user does not know how many other 1977 estate values also exceeded $10 million, how these estate values might be related to each other, or whether any of these estates have features that distinguish them from the rest of the large estates. The beliefs of such a user can then be expressed probabilistically as follows: For any N, if there are N estates, then their values $X_1, \ldots, X_N$ are independently and identically distributed. The user might more specifically take the X_i to have a Pareto(α) distribution with a truncation point at $10 million, using empirical studies such as Lampman (1962, pp. 210–213) as a guide. [For a contrary view see Stark (1972, p. 73), who suggested that the empirical distribution of wealth data in the United Kingdom during the 1950s is not well fit by a Pareto distribution. This may be due to the inclusion of small estates—say, those below £3,000.] Then the user's prior predictive distribution of the target X_1 is Pareto(α), conditional on α.

Since α is typically unknown, full Bayesian analysis requires specification of a distribution for α. This analysis can be simplified by using an estimate of α.

When the 1977 data are published, the user learns that $N = 11$ and $\Sigma X_i = \$507{,}862{,}000$. The predictive distribution of the target X_1 is then updated as follows: First, conditional on $\Sigma X_i = \$507{,}862{,}000$, the joint density of $X_1, \ldots, X_{11}$ is calculated. Then $X_2, \ldots, X_{11}$ are integrated out of the joint density to give the marginal predictive distribution of the target after data release. Here the posterior predictive density is given by

$$f(x_1 \mid \sum X_i = 507{,}862{,}000) = \frac{y_1^{-\alpha-1} \int \cdots \int_{\mathcal{S}} \prod_{i=2}^{11} y_i^{-\alpha-1} dy_2 \cdots dy_{11}}{\int \cdots \int_{\mathcal{T}} \prod_{i=1}^{11} y_i^{-\alpha-1}\, dy_1 \cdots dy_{11}}, \quad (1)$$

where $y_i = x_i/10{,}000{,}000$, x_1 = target, $\mathcal{S} = \{y_2 \geq 1, \ldots, y_{11} \geq 1 : \Sigma_2^{11} y_i = 50.7862 - y_1\}$, and $\mathcal{T} = \{y_1 \geq 1, \ldots, y_{11} \geq 1 : \Sigma_1^{11} y_1 = 50.7862\}$.

The value of α is now estimated and substituted into Equation (1). This procedure is appropriate if after observing available data the user's posterior distribution for α is peaked at the estimated value. We assume that the data user estimates α from the released data according to $\hat{\alpha} = \Sigma_1^{11} x_i (\Sigma_1^{11} x_i - 110{,}000{,}000)^{-1}$, which implies that $\hat{\alpha} = 1.2765$. [Johnson and Kotz (1970, p. 235) gave this estimator as eq. (10).] Another possibility is to assume that the user, after consulting some empirical literature on wealth distributions, estimates α to be 1.2 [cf. Lampman (1962, p. 210) for U.S. 1953 gross estate data].

It could be argued that no matter what α is, the posterior predictive density (1) is analytically intractable. Specifically, this posterior density is not a Pareto density. The posterior predictive density can, however, be estimated using Monte Carlo techniques. One such technique is described in the Appendix. Figure 1 pictures the corresponding estimated predictive density of the target X_1 after data release for $\hat{\alpha} = 1.2765$.

For a second illustration of the predictive approach, we next consider categorical data. Suppose we are again looking at estates of $10 million or more, but this time for 1983. For returns filed in 1983, there were 224 such estates, as displayed in Table 1 by region and gross estate size (Bentz and Schwartz 1985). Note that the frequencies shown are all sufficiently large to satisfy the current disclosure rules of the Internal Revenue Service (Wilson and Smith 1983).

Take the target X_1 to be the gross estate category of a hypothetical individual—say, Veronica Berry Rich. As before, the data user's knowledge about X_1 is represented probabilistically by a predictive distribution before and after data release.

For example, suppose that the data user cannot distinguish among the 224 individuals described by Table 1 as to who might be more likely than the others to have gross estates in particular categories. In such a case, $X_1, \ldots, X_{224}$ are exchangeable random variables. [For a discussion of exchangeability, see, e.g., DeFinetti (1975, pp. 215 ff) and Johnson and Kotz (1977, pp. 97–105).]

Since the data user believes Veronica Berry Rich's estate value is exchangeable with that of all of the other 223 individuals, the user must believe that any of the 224 gross estates is equally likely to be hers. So the marginal predictive distribution on her three possible categories, regardless of the particular form of the exchangeable prior, is the relative frequency distribution of the "all-regions" margin (see the Appendix for a theorem that justifies this result). Numerically, this is given by $\mathbf{p} = (.728, .138, .134)$. If the user gains more information about Veronica Berry Rich, then the user's predictive distri-

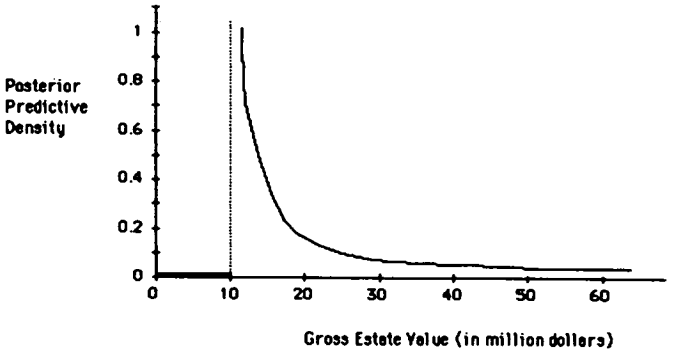

Figure 1. Estimated Predictive Density for Target 1977 Gross Estate Value X_1.

bution is updated. If, for example, the user knows that Veronica Berry Rich maintained primary residence in the Northeast, then the predictive distribution for X_1 is updated to the relative frequency distribution for the Northeast, which is given by $\mathbf{p} = (.732, .179, .089)$.

To extend the example, suppose that the 224 estates are a sample (without replacement) from a population of size N, so Veronica Berry Rich may not be one of the 224 individuals

Table 1. Number of Estates by 1983 Gross Estate Amount and Census Region of Primary Residence (in millions of dollars for estates of $10 million or more)

Census Region	10–19	20–29	30+
Northeast	41	10	5
North Central	40	4	6
South*	47	10	10
West	35	7	9
All regions	163	31	30

*Includes data for returns of citizens and resident aliens living in Puerto Rico, the Panama Canal Zone, the Virgin Islands, or abroad.

whose gross estates are reported in Table 1. Based on historical data or other information, a user has a prior predictive distribution for the categories $(X_1, \ldots, X_N)$ of the N members of the population. Suppose that $X_1, \ldots, X_N$ are exchangeable under this prior predictive distribution. After publication of the all-regions margin of Table 1, the user can update the predictive distribution using simple probabilistic arguments as follows: First, write the predictive distribution for X_1 as

$$P(X_1 = i) = P(X_1 \text{ in sample})P(X_1 = i \mid X_1 \text{ in sample}) + P(X_1 \text{ not in sample})\, P(X_1 = i \mid \text{not in sample}).$$

If the 224 estates in the sample are believed to be exchangeable with the other estates not in the sample, as would be the case if the sample were known to be random, then the conditional predictive distribution of the target X_1 (given that Veronica Berry Rich's estate is in the sample) is, as before, given by $\mathbf{p} = (.728, .138, .134)$. If $\mathbf{q} = (q_1, q_2, q_3)$ is the conditional predictive distribution of the target when X_1 is known not to be in the sample, then the updated predictive distribution for X_1 is $(224/N)\mathbf{p} + (1 - 224/N)\mathbf{q}$.

In general, assessment of prior distributions is difficult [but see, e.g., Chaloner and Duncan (1983) for a discussion of some ways of overcoming these difficulties]. Nevertheless, as the above illustrations show, the basic structure of reasonable predictive distributions for the target may be constructed in certain situations. In other situations, historical data may suggest a plausible prior predictive distribution. In addition, if the objective is to bound the extent of disclosure (see Section 6), only the class of prior predictive distributions that lead to violation of a bound may be important. Precise specification of the prior distribution is then unnecessary.

4. MEASURES OF DISCLOSURE

Much of the discussion of statistical disclosure control within government agencies has focused on administrative procedures and ad hoc rules (Subcommittee on Disclosure Avoidance Techniques 1978). Such procedures are necessary to implement effective disclosure control, but they do not provide a framework that permits the extent of disclosure limitation achieved to be measured. This section builds a theoretical framework for measures of the extent of disclosure based on the following principles.

1. The complete state of a user's uncertainty about a target before and after data release is specified by the user's prior and posterior predictive distributions, respectively.

2. The user's uncertainty about a target can be summarized by applying a nonnegative concave function $U(\cdot)$ to the user's predictive distribution for the target (DeGroot 1962, 1970). Such functions $U(\cdot)$ are called uncertainty functions. The larger the value of U, the more the uncertainty.

Based on the idea that the user's uncertainty about the target determines the extent to which confidentiality has been compromised, these principles lead to three classes of measures of disclosure. We label these classes of disclosure measures *knowledge*, *knowledge gain*, and *relative knowledge gain*. All are based on a choice of an uncertainty function $U(\cdot)$ that is deemed appropriate for a particular application.

Applying the uncertainty function $U(\cdot)$ to the posterior predictive distribution for the target gives the posterior *knowledge* measure $U(\text{posterior})$ of disclosure. If a goal is to limit the user's knowledge about the target, then the data should be released only if $U(\text{posterior})$ exceeds some threshold. We discuss this point in more detail in Section 5.

On the other hand, an informed user or insider may already have a great deal of knowledge about the target before the data are released. In this case, the difference $U(\text{prior}) - U(\text{posterior})$, which describes the increment in the user's knowledge, may be a more appropriate disclosure measure. The difference $U(\text{prior}) - U(\text{posterior})$ is the *knowledge gain* measure of disclosure. The difference may be negative; if so, the user is more uncertain about the target after obtaining the data than before seeing the data.

In some other applications, the *relative knowledge gain* $(U(\text{prior}) - U(\text{posterior}))/U(\text{prior})$ may have more meaning than the unscaled difference. In particular, the relative knowledge gain disclosure measure is scale invariant in the specification of the uncertainty function.

The choice of an appropriate uncertainty function can be guided by specifying in a decision-theoretic framework the information's potential to the user for compromising privacy. This potential is represented by the extent to which the user can infer the target value based on the released data. Given a loss function $\mathcal{L}$ for the user's decision problem "identify the target," the risk of the optimal decision with respect to the predictive distribution $p(\cdot)$ defines an uncertainty $U(p)$ (DeGroot 1962). For example, when the target X_1 is discrete with predictive probability func-

tion $p(\cdot)$, the uncertainty is $U(p) = \inf_d \Sigma \mathcal{L}(x_1, d)p(x_1)$. The advantage of this decision-theoretical approach is that it often suggests appropriate uncertainty functions for complicated situations. For now we consider two simple illustrations.

Illustration 1. Suppose the target is discrete and a user's decision problem is to specify a probability vector $\mathbf{p}^* = (p_1^*, p_2^*, \ldots)$ over the possible values of the target. The user has the option of employing a probability distribution $\mathbf{p}^*$ for the target that is different. But if loss is assessed by $\mathcal{L}$(employ $\mathbf{p}^*$, category r right) $= -\log p_r^*$, then by the fundamental lemma of information theory (Ash 1965) the user's corresponding risk $-\Sigma p_r \log p_r^*$ is minimized by choosing $\mathbf{p}^* = \mathbf{p}$. Consequently, the optimal decision by the user requires no modification of the predictive distribution $\mathbf{p}$, and the risk of the optimal decision with respect to $\mathbf{p}$, which is the uncertainty in $\mathbf{p}$, is Shannon's entropy $-\Sigma p_r \log p_r$. Frank (1979), in his brief discussion of probabilistic disclosure, suggested Shannon's entropy as a measure of disclosure. In our framework, Shannon's entropy is one member of the class of posterior knowledge measures of disclosure.

Illustration 2. Suppose a user must specify a category for discrete target and zero–one loss is to be assessed so that there is no loss for correct identification and a loss of one for any incorrect identification. The risk is then minimized by specifying the category with highest predictive probability, and the minimum risk (or uncertainty) is $1 - \max p_i$. This uncertainty is essentially the measure of disclosure proposed by Cassel (1975). For us, $1 - \max p_i$ is one member of the class of posterior-knowledge measures of disclosure.

It could be argued that uncertainty functions are inadequate for measuring disclosure because they indicate only the smoothness (lack of peaks) of the predictive distribution and not the location of any peaks in the predictive distribution. The substance of this argument is that the uncertainty is low (and the extent of disclosure is high) if the distribution is peaked, even if the peak is not at the correct value of the target. But if the published data are intended to represent the reporting units accurately, and so not mislead the data user, then a posterior predictive distribution that is peaked at an incorrect value for the target is undesirable for reasons of misrepresentation. But furthermore, peaks at the correct value of the target are undesirable for reasons of privacy. So peaks, wherever they may be located, are to be limited and a discussion of statistical disclosure should focus on the smoothness of the posterior distribution.

5. DISCLOSURE-LIMITING RULES

Each measure of the extent of statistical disclosure leads naturally to a disclosure-limiting rule—namely, release a set of data only if the posterior extent of disclosure will be below a given limit. With posterior knowledge measures, data are released only if the uncertainty in the posterior predictive distribution about the target X_1 is at least as large as some chosen limit τ_1; with knowledge-gain disclosure measures, data are released only if the difference—prior predictive uncertainty minus posterior predictive uncertainty—is no larger than some limit τ_2. In theory, these limits τ_1 and τ_2 are chosen to ensure that an acceptable level of disclosure is not exceeded. In practice, the competing demands of privacy protection and valid inference needs would lead to negotiation of these limits between potential data users and the releasing agency.

In the simplest case, the target is the presence or absence of a characteristic of one reporting unit. The predictive distribution is then $\mathbf{p} = (p, 1 - p)$, and a measure of the uncertainty of $\mathbf{p}$ is given by a concave function $U(\cdot)$ of $\mathbf{p}$ satisfying $U(0) = U(1) = 0$. Therefore, for any uncertainty function, posterior knowledge is above a limit τ_1, so the posterior knowledge rule allows data release if and only if $a \leq p \leq b$, where a and b depend on $U(\cdot)$ and the limit τ_1.

A knowledge gain rule with limit τ_2 is equivalent to a knowledge rule with a limit that depends on the prior predictive distribution. To be specific, the knowledge gain rule is equivalent to the knowledge rule that allows data release only if the posterior uncertainty exceeds prior uncertainty minus the knowledge gain limit τ_2. For example, with prior predictive distribution $(\frac{1}{2}, \frac{1}{2})$, posterior predictive distribution $(p, 1 - p)$, and zero–one loss, the equivalent knowledge rule is to release the data only if $\min(p, 1 - p) \geq \frac{1}{2} - \tau_2$.

In the remainder of this section, certain data structures and user-inference strategies are shown to lead to disclosure control rules that were discussed on an ad hoc basis by the Subcommittee on Disclosure Avoidance Techniques (1978). In each case we consider only the posterior knowledge rule. A key advantage to the DL procedure is that since these rules arise in the context of specific decision models, they may be judged to be appropriate if the premises of the model are acceptable.

5.1 Categorical Data

In this section we propose two posterior-knowledge rules that can be applied to categorical data like those in Table 1. We suppose that the target belongs to one of k categories of a table, possibly cross-classified. The first rule allows data release only if the predictive probability p_i that the target falls in the ith category is below a bound that depends on the sensitivity of category i. That is, the user is kept more uncertain about sensitive categories than about nonsensitive categories. The second rule allows data release only if the predictive probability p_i that the target falls in the ith category is within certain upper *and* lower bounds. This second rule arises by viewing specification of a category for the target as equivalent to specification of the categories to which the target does not belong.

By the theorem in the Appendix, in the case that the prior predictive distribution is exchangeable and the target is known to be included in the table, the posterior predictive distribution equals the observed tabular proportions. Therefore, in this case, any posterior-knowledge rule reduces to constraints on the tabular relative frequencies alone. Bounding tabular relative frequencies away from zero and away from one has been discussed by the Subcommittee on Disclosure Avoidance Techniques (1978) in an ad hoc fashion. The DL framework gives formal justi-

fication for such rules—provided the premises that lead to them are believed to be reasonable.

To develop the first disclosure-limiting rule, suppose that some categories are more sensitive than others. The severity of the consequences of incorrectly specifying a category to be i ($1 \leq i \leq k$) may then depend on the sensitivity of category i. If so, a loss function $\mathfrak{L}$ can appropriately quantify this sensitivity through a positive parameter λ_i and be given by $\mathfrak{L}(i$ specified, j correct$) = \lambda_i$ or 0, according to whether $i \neq j$ or $i = j$. The corresponding uncertainty (or risk of the optimal decision) at a predictive distribution specified as $\mathbf{p} = (p_1, \ldots, p_k)$ is $\min_i[\lambda_i(1 - p_i)]$. The posterior-knowledge rule allows data release only if $\min_i[\lambda_i(1 - p_i)] \geq \tau_1$.

In particular, suppose the penalty for incorrectly specifying the target to be i is smaller when category i is common than when it is rare. Then for a nonincreasing function γ the λ_i's satisfy $\lambda_i = \gamma(p_i)$, $i = 1, \ldots, k$. In this case, the posterior-knowledge rule allows data release only if $p_i \leq 1 - \tau_1/\gamma(\max p_i)$ for all i. That is, the data are released when the predictive probability that the target falls in each category is sufficiently small. For illustration, take $\gamma(p) = 1/p$, so this knowledge rule simplifies to $\max p_i \leq 1/(1 + \tau_1)$.

To develop the second disclosure-limiting rule, suppose that the loss for incorrectly specifying a target to belong to category i is decreasing in p_i. That is, incorrectly specifying that a target belongs to category i is more serious when category i is rare than when category i is prevalent. When a prevalent category is specified, the important loss is the missed opportunity to identify an unusual, and typically interesting, feature of the target. Such reasoning suggests the simple loss function

$$\mathfrak{L}(\text{category } i \text{ specified, category } j \text{ correct}) = \min(1 - p_i, bp_i) \text{ for } i \neq j \text{ and } 0 \text{ for } i = j,$$

where b determines the importance of opportunity loss relative to specification error. The corresponding posterior-knowledge rule allows data release only if

$$\min_i \min\{(1 - p_i)^2, bp_i(1 - p_i)\} \geq \tau_1$$

or, equivalently, only if

$$.5 - (.25 - (\tau_1/b))^{1/2} \leq p_i \leq \min\{.5 + (.25 - (\tau_1/b))^{1/2}, 1 - (\tau_1)^{1/2}\}$$

for each $i = 1, \ldots, k$. That is, when the data user is concerned with the categories not specified as well as the category specified, the rule requires that the p_i's be bounded away from both 0 and 1. Such a disclosure control rule has been discussed on an ad hoc basis by the Subcommittee on Disclosure Avoidance Techniques (1978).

5.2 Categorical Data Obtained by Sampling

In many cases the released table of frequencies represents only a sample of the population to which the target unit belongs. If a user believes that the categories $X_1, \ldots, X_N$ of the N members of the population are exchangeable under the prior predictive distribution, then as in the extension of the second illustration in Section 3, the user's updated predictive distribution is $f\mathbf{p} + (1 - f)\mathbf{q}$, where $\mathbf{p}$ is the tabular relative frequency distribution, f is the sampling fraction, and $\mathbf{q}$ is the marginal predictive distribution of X_1 given the sample results and the information that X_1 is not in the sample. A posterior-knowledge rule allows the table to be released only if $U(f\mathbf{p} + (1 - f)\mathbf{q}) \geq \tau_1$. For example, with a loss of $\gamma(fp_i + (1 - f)q_i)$ based on a nonincreasing function γ for specifying category i when some other category is correct, the rule reduces to release the table only if $\max_i (fp_i + (1 - f)q_i) \leq c$, where c depends on γ and τ_1.

A possible disclosure control policy is to release the sample data only if they meet the standards for release of census data. That is, the data are released only if $U(\mathbf{p})$, rather than $U(f\mathbf{p} + (1 - f)\mathbf{q})$, exceeds a limit. Since U is concave,

$$U(f\mathbf{p} + (1 - f)\mathbf{q}) \geq U(\mathbf{p}) + (1 - f)[U(\mathbf{q}) - U(\mathbf{p})].$$

This policy is conservative, in the sense that uncertainty may be held larger than required, for certain users. These users have an uncertainty after data release that would be less if X_1 were known to be in the sample [in which case it is $U(\mathbf{p})$] than it would be if X_1 were known not to be in the sample [in which case it is $U(\mathbf{q})$]. The policy is most conservative for users whose predictive distribution $\mathbf{q}$ is uniform over the possible categories of the target and minimally conservative against users whose predictive distribution $\mathbf{q}$ is identical with the observed sample relative frequencies $\mathbf{p}$. The policy is not guaranteed to be conservative against users for whom $U(\mathbf{q}) < U(\mathbf{p})$, but the risk of generosity to users who have more precise information about the units not included in the table than information about the units included in the table may be outweighed by the advantages of disclosure policies that depend only on the tabular frequencies whenever the joint prior predictive distribution is exchangeable.

5.3 Rates of Incidence Data

For an example with noncategorical data, consider the following scenario. An agency plans to publish industry-wide health statistics about incidence rates of a rare disease (see Table 2). In order to decide whether the proposed data release excessively compromises privacy, the agency focuses on a hypothetical user with a target X_1 of the number of incidents of the disease for a particular company. From medical experts the user has learned that the incidence rate λ per employee exposure year is the same for all workers in the industry and that the disease is not contagious—so workers contract the disease independently of each other. In releasing data, the agency may wish to provide information relevant to λ, which is legitimately the public's business, but in doing so it may provide too much information about X_1. From public records (say, the company's annual report) the user has obtained the approximate number of employee exposure years M_1 for the target company; say that $M_1 = m_2$. The data user may take the numbers of disease incidents for the $N = \Sigma N_i$ companies in the industry to be independently Poisson distributed with means $\lambda M_1, \ldots, \lambda M_N$, where M_i is

the number of employee exposure years in company i.

From the published table, the hypothetical user learns that the total number of disease incidents for the N_2 companies of the same size as the target company is $m_2N_2p_2$. Conditional on this total count, the user's posterior predictive distribution for X_1 is binomial with parameters $m_2N_2p_2$ and $1/N_2$. Note that the user need not specify a distribution for λ in order to determine the posterior predictive distribution.

Finally, in order to evaluate the extent of disclosure to the hypothetical user, the agency must specify an uncertainty function. Suppose squared error loss is chosen. Then the uncertainty in the predictive distribution is its variance, and the posterior knowledge rule allows the data to be released only if $m_2p_2(1 - N_2^{-1})$ exceeds a specified limit τ_1. Therefore it is not possible to specify bounds on only the incidence rates p_i or on only the number of companies N_i to preserve confidentiality universally. Confidentiality could be violated if m_2 is small, if p_2 is small, or if N_2 is small. It is possible, however, to find constants N^* and p^* such that the uncertainty exceeds τ_1 for any $p_2 \geq p^*$ and $N_2 \geq N^*$ for fixed m_2. That is, it is possible to bound the extent of disclosure to the hypothetical user by setting lower bounds on both the number of companies and the incidence rate.

To extend this case, suppose that the data user has inside information about the exact number of incidents x for another company with m_2 employee exposure years. The user's posterior predictive distribution for X_1 is then binomial with parameters $m_2N_2p_2 - x$ and $(N_2 - 1)^{-1}$. The posterior-knowledge rule under squared error loss allows data release only if the binomial variance, $(m_2N_2p_2 - x)(N_2 - 1)^{-1}(1 - (N_2 - 1)^{-1})$, is greater than τ_1. In such situations of insider information, a *dominance* rule allows data release only if (a) the number of companies N_2 is above a number N^* and (b) no company accounts for more than a fraction f of the incidents (for a discussion of this rule see Subcommittee on Disclosure Avoidance Techniques 1978). Since condition (b) implies $x \leq m_2N_2p_2f$ and since $N_2(N_2 - 1)^{-1} > 1$, conditions (a) and (b) imply only that the uncertainty as measured by the variance exceeds $m_2p_2a^*$ for some a^*. Therefore there are incidence rates p_2 for which the dominance rule does not guarantee that the extent of disclosure is controlled. This is an example in which present disclosure control techniques, although intuitively appealing, do not necessarily solve the problem, and in which the uncertainty function approach points out a difficulty.

Table 2. Incidence Rates per Employee Exposure Years

Employee exposure years	$\leq m_0$	m_1	...	$\geq m_k$
Number of companies	N_0	N_1		N_k
Average incidence rate	P_0	P_1		p_k

5.4 Total-Count-With-Total-Value Data

For a final illustration with noncategorical data, we return to the 1977 estate value situation described in Section 2. In this situation, for the open-ended cell of largest estate values, the total count $N = 11$ and the total value \$507,862,000 are released. Concern is with the extent of disclosure about the value of a particular estate. Dalenius and Denning (1982) suggested that release of a mean (hence, total) may compromise privacy if the variance is small relative to the mean. This suggests not releasing the data if the coefficient of variation of estate values is too small. We now examine how this idea fits into the DL framework.

Figure 1 gives the approximate posterior predictive distribution of an individual estate value for a user whose prior predictive distribution is Pareto with parameter α estimated to be 1.2765 from the released data. Again, in order to assess the extent of disclosure from releasing the data, the agency specifies an uncertainty function. Suppose that the uncertainty function is based on squared relative error loss and so defined by

$$\mathfrak{L}(\text{specify value } y, \text{ true estate value } x) = [(y/x) - 1]^2.$$

Then the user's uncertainty is $\text{var}(1/X)/[E(1/X)]^2$, where expectation and variance are calculated with respect to the user's posterior predictive distribution. The posterior-knowledge rule of "release data when uncertainty exceeds τ_1" is then equivalent to "release data when the coefficient of variation of $1/X$ exceeds $\{\tau_1/(1 - \tau_1)\}^{1/2}$." The use of the reciprocal transformation is intuitively reasonable in this context because the Pareto distribution is highly skewed to large values. The coefficient of variation of $1/X$ can be approximated through the simulated density function; details are given in the Appendix.

6. TECHNIQUES FOR DECREASING THE EXTENT OF DISCLOSURE

A statistical table that does not meet standards for release may be modified to reduce its extent of disclosure. For this purpose, categories may be aggregated, cells may be suppressed, counts may be rounded, random variables may be added to summary statistics, or microdata values with similar covariates may be swapped. Cell suppression is discussed in Cox (1980, 1981) and data swapping is discussed in Dalenius and Reiss (1982). The uncertainty function approach to confidentiality can be used to identify which modifications are effective and which are ineffective in a given set of circumstances. The approach is illustrated in a simple setting in this section.

Consider a target value that is one of k categories, a prior predictive distribution $\mathbf{p} = (p_1, \ldots, p_k)$ under which the N population units are exchangeable, and a loss function $\mathfrak{L}$ defined for some $a > 0$, $b > 0$ by

$$\mathfrak{L}(\text{category } i \text{ specified, category } j \text{ correct}) = \min(ap_i, b(1 - p_i)) \quad \text{if } i \neq j$$
$$= 0 \quad \text{if } i = j.$$

Suppose the released table gives only the total number of reporting units in each of the k categories. In this case, the posterior predictive distribution for the target is identical to the

tabulated relative frequencies $\mathbf{r} = (r_1, \ldots, r_k)$, and the posterior-knowledge rule allows data release only if $c_1 \leq r_i \leq c_2$, for all i and some constants c_1, c_2.

Suppose that r_1 violates one of these bounds so that the complete table cannot be published. For simplicity, take $c_1 \leq r_i \leq c_2$ for all $i > 1$. If r_1 is too small, $r_1 < c_1$, then it seems plausible that the extent of disclosure would be reduced if category 1 were aggregated with another category—say, category 2. This is not necessarily the case, however. After the aggregated data are published, the user's posterior predictive distribution is updated to $((r_1 + r_2)p_1/(p_1 + p_2), (r_1 + r_2)p_2/(p_1 + p_2), r_3, \ldots, r_k)$. This result follows from the exchangeability of the population units and can be verified using a two-stage urn model: A category for the target is first chosen randomly from an urn containing labels (1 or 2), 3, . . . , k in the proportions $r_1 + r_2, r_3, \ldots, r_k$. If the label (1 or 2) is chosen, then a category is chosen from an urn containing labels 1 and 2 in proportions $p_1/(p_1 + p_2)$ and $p_2/(p_1 + p_2)$. If p_1 is small, then the aggregated table may also violate the standards for publication. To be specific, if $p_1/p_2 < r_1/r_2$, then the portion of the aggregated cell count attributed by the data user to category 1 is smaller than the relative frequency of category 1 and the aggregated table violates the posterior knowledge rule. Consequently, aggregation does not guarantee a reduction in disclosure, even if the disclosure is due to a cell count being too small.

On the other hand, aggregation is not ruled out as an effective disclosure control technique if the limit violation in the original table was due to r_1 being too large. If $p_1/p_2 < r_1/r_2$, then the posterior predictive probability of category 1 is smaller after aggregation than it was before aggregation. If the aggregation has not greatly increased the posterior predictive probability of category 2, then the aggregated table may satisfy the standards for data release. Plainly, if a user has a prior predictive distribution over the original k categories (or a refinement of them), then it must be taken into account. In particular, the user cannot safely be assumed to have an aggregated prior distribution or an aggregated posterior distribution merely because the data have been aggregated.

Moving now to cell suppression, suppose that instead of combining categories 1 and 2 the count for category 1 is suppressed. The relative frequencies for categories 2 through k, given that the target is not in category 1, can be computed from the modified table. These relative frequencies equal $r_2/(1 - r_1), \ldots, r_k/(1 - r_1)$, where the r_i's are the relative frequencies in the original table. Here we are assuming that the data user does not have external information (e.g., access to the total number of reporting units) that would permit computation of the relative frequency of category 1. The user's posterior distribution for the target based on the modified table is $(p_1, 0, \ldots, 0) + (0, r_2, \ldots, r_k)(1 - p_1)/(1 - r_1)$.

If the probability that the target belongs to category 1 is quite different under the prior distribution from what it is under the posterior distribution based on the original data, then the partially suppressed table may violate the posterior-knowledge rule because of some category $i > 2$ that was not a problem in the original table. The complication is that the user will not abandon or change the previous information about category 1 if no other information is available, and this information affects the posterior probability of all other categories. The fact that the prior information $\mathbf{p}$ is used differently with the complete and the partially suppressed tables must be taken into account in evaluating the effectiveness of category (and by extension, cell) suppression.

The need for considering the prior information $\mathbf{p}$ also arises if the published relative frequencies $\mathbf{r} = (r_1, \ldots, r_k)$ pertain to only a sample fraction f of the population. The table for the sampled units satisfies the posterior-knowledge rule if $a \leq r_i + (1 - f)(q_i - r_i) \leq b$ for $1 \leq i \leq k$, where q_i is, as in the illustration in Section 2, the probability that the target is in category i, given that the target unit is not in the sample and the proportion of sampled units belonging to categories 1, . . . , k is $r_1, \ldots, r_k$. Because $U(f\mathbf{r} + (1 - f)\mathbf{q}) \geq U(\mathbf{r}) + (1 - f)(U(\mathbf{q}) - U(\mathbf{r}))$, the extent of disclosure is reduced if the uncertainty given that the target unit is not in the sample exceeds the uncertainty given that the target unit is in the sample. As would be expected, the magnitude of the reduction increases as the sampling fraction decreases.

7. DISCUSSION

The disclosure-limiting (DL) approach to controlling the extent of disclosure builds directly on the definition of statistical disclosure proposed by Dalenius (1977b) and recommended by the Subcommittee on Disclosure Avoidance Techniques (1978). Consistent with the suggestion of Dalenius, the DL approach recognizes that, typically, data relevant to the target have already been published and will be combined with newly released data. The DL approach requires that the information posture of the data user be expressed in a predictive distribution for the target. The user's predictive distribution can be difficult to assess, but we have shown that the effects of aggregation, for example, cannot be adequately determined without taking into account the data user's prior beliefs about the target. The difficulties involved in assessing these prior distributions are mitigated when the data-releasing agency has some knowledge of the data already available to the user. For example, a reasonable model for a receiving agency's prior distribution can be developed when (a) a group of reporting units with which the target is exchangeable can be identified and (b) the information (e.g., data tables) that the receiving agency has about the exchangeable reporting units is known. Previous work on statistical disclosure has also considered the prior information available to insiders. The DL approach refines the dichotomy of naive user and insider into a continuum from uninformed user to informed user.

The DL approach quantifies the extent of statistical disclosure by means of uncertainty functions applied to predictive distributions. This approach justifies policies that allow data release only if the extent of disclosure is below a cutoff. Indeed, the

DL framework provides a justification for various ad hoc rules. Although specification of uncertainty functions and disclosure limits may appear arbitrary and difficult to justify, it is also difficult to justify rigorously ad hoc rules for releasing data. Furthermore, the approach yields a method for generating new rules—namely, (a) reformulate the issue as a decision problem for the receiving agency of estimating a target value and (b) limit the receiving agency's maximal inferential gain about the exact value of a target.

The DL approach clarifies the issues that are involved in controlling the extent of disclosure. It also lends insight into the behavior of disclosure-controlling techniques, such as aggregation. At this stage we have analyzed simplified scenarios within the DL framework. With these simplifications we have shown, for example, that aggregation does not guarantee a reduction in the extent of disclosure.

There remain important data types, such as microdata, that fit into the context of the DL approach but have not yet been explored. The insights gained from looking at the simple scenarios considered so far suggest the value of extension to these other data types.

APPENDIX

Simulation Calculations

The conditional density given by Equation (1) can be estimated as follows: First, note that with $y_i = x_i/10{,}000{,}000$,

$$f(y_1 \mid \textstyle\sum y_i = t) = \frac{f(y_1)\int\cdots\int_{\mathcal{S}}\prod_{i=2}^{10} f(y_i) f\left(t - y_1 - \sum_2^{10} y_i\right) dy_2 \cdots dy_{10}}{\int\cdots\int_{\mathcal{T}}\prod_{i=1}^{10} f(y_i) f\left(t - \sum_{i=1}^{10} y_i\right) dy_i \cdots dy_{10}}$$

for $y_1 \geq 1$, where $\mathcal{S} = \{y_2 \geq 1, \ldots, y_{10} \geq 1 : t - y_1 - \sum_2^{10} y_i \geq 1\}$, $\mathcal{T} = \{y_1 \geq 1, \ldots, y_{10} \geq 1 : t - \sum_1^{10} y_i \geq 1\}$, and f is a Pareto(α) density with truncation below 1. For each $y_1 \geq 1$, the last equation may be rewritten as

$$f(y_1 \mid \textstyle\sum Y_i = t) = \frac{f(y_1) E\left\{f\left(t - y_1 - \sum_2^{10} Y_i\right) I\left(t - y_1 - \sum_2^{10} Y_i \geq 1\right)\right\}}{E\left\{f\left(t - \sum_1^{10} Y_i\right) I\left(t - \sum_1^{10} Y_i \geq 1\right)\right\}},$$

where $Y_1, \ldots, Y_{10}$ are independent and identically distributed Pareto(α) random variables. It is now straightforward to use Monte Carlo methods to estimate the posterior predictive density using a pseudo-random-number generator such as the International Mathematical and Statistical Libraries (IMSL) subroutine GGEXN [using the fact that if Y has a Pareto(α) distribution, then log Y has an exponential(α) distribution].

In the calculations for Section 3, 20,000 pseudo-random samples of size 10 were generated using the IMSL subroutine GGEXN. To obtain approximations to the moments of $1/X$, the trapezoidal rule with estimated density function values was used. Under the simulated predictive distribution, the coefficient of variation of $1/X$ was estimated to be .529. The accuracy of .529 may be judged as follows: Under the estimated Pareto distribution, the coefficient of variation of $1/X$ was estimated to be .497 using the same 20,000 samples ($Y_1, \ldots, Y_{10}$) that were used to estimate the coefficient of variation of the posterior predictive distribution. Under the Pareto(1.2765) distribution, the coefficient of variation is, in fact, .489. For the actual complete data of 1977 gross estates exceeding $10 million, the coefficient of variation of $1/X$ was .508, as calculated by Bentz and Schwartz (1985).

Posteriors for Categorical Data With Exchangeable Priors

Theorem. Let $X_1, \ldots, X_N$ be exchangeable discrete random variables taking on values $c_1, \ldots, c_k$. Conditional on the frequency counts $n_1, \ldots, n_k$, where $n_j = \#(X_i = c_j)$,

$$P(X_m = c_j \mid n_1, \ldots, n_k) = n_j/N,$$

for $m = 1, \ldots, N$ and $j = 1, \ldots, k$.

Proof. Without loss of generality we need only consider X_1 and c_1. Using an urn model argument with N balls and k urns labeled $c_1, \ldots, c_k$, $P(X_1 = c_1 \mid n_1, \ldots, n_k)$ is the probability that ball 1 is in urn c_1, given that n_1 balls are in urn c_1, $\ldots$, n_k balls are in urn c_k. By exchangeability of $X_1, \ldots, X_N$, all assignments that give $n_1, \ldots, n_k$ balls in the k urns have the same probability, say, p^*. Following the definition of conditional probability, the numerator of $P(X_1 = c_1 \mid n_1, \ldots, n_k)$ is P(ball 1 in c_1, $n_1 - 1$ balls in c_1, n_1 balls in $c_2, \ldots, n_k$ balls in c_k) $= p^*(N-1)!/(n_1 - 1)!\, n_2! \cdots n_k!$ and the denominator is $P(n_1$ balls in $c_1, \ldots, n_k$ balls in $c_k) = p^* N!/n_1! \cdots n_k!$. So $P(X_1 = c_1 \mid n_1, \ldots, n_k) = n_1/N$.

[*Received September 1982. Revised October 1984.*]

** George T. Duncan is Professor, School of Urban and Public Affairs and Department of Statistics, and Diane Lambert is Associate Professor, Department of Statistics, Carnegie-Mellon University, Pittsburgh, PA 15213. Lambert's work was partially supported by National Science Foundation Grant MCS 8301692. The authors thank Frederick Scheuren of the U.S. Internal Revenue Service for providing the data for the second illustration of Section 3 and for his perceptive comments that helped to improve the article's exposition. They also thank Beverley Causey and Lawrence Cox of the U.S. Bureau of the Census, Ramona Trader of the University of Maryland, and J. B. Kadane of Carnegie-Mellon University for discussions on the topic of confidentiality, and the associate editor and three referees for their comments.

REFERENCES

Alexander, L., and Jabine, T. (1978), "Access to Social Security Microdata Files From Research and Statistical Purposes," *Social Security Bulletin*, August 1978.

Ash, R. (1965), *Information Theory*, New York: Interscience Publishers.

Bentz, M., and Schwartz, M. (1985), unpublished IRS tabulation for returns with gross estates of $10,000,000 or more for 1977 and 1983 (Internal Revenue Service Statistics of Income Special Report), Washington, DC: Internal Revenue Service.

Boruch, R. F. (1971), "Education Research and the Confidentiality of Data: A Case Study," *Sociology of Education*, 44, 59–85.

Cassel, Claes (1975), *On Probability Based Disclosures in Frequency Tables*, technical report, Survey Research Institute, National Central Bureau of Statistics, Sweden.

Cassel, C. M. (1976), "Probability Based Disclosures," in *Personal Integrity and the Need for Data in the Social Sciences*, eds. T. Dalenius and A. Klevmarken, Stockholm: Swedish Council for the Social Sciences, pp. 189–193.

Chaloner, Kathryn M., and Duncan, George T. (1983), "Assessment of a Beta Prior Distribution: PM Elicitation," *The Statistician*, 27, 174–180.

Clark, Cynthia Z. F., and Coffey, Jerry L. (1983), "How Many People Can Keep a Secret? Data Interchange Within a Decentralized System," paper presented at the Annual Meeting of the American Statistical Association, Toronto, Canada.

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

——— (1981), "Linear Sensitivity Measures in Statistical Disclosure Control," *Journal of Statistical Planning and Inference*, 5, 153–164.

Dalenius, T. (1974), "The Invasion of Privacy Problem and Statistics Production—An Overview," *Statistisk Tidskrift*, 3, 213–225.

——— (1977a), "Computers and Individual Privacy: Some International Implications," *Bulletin of the International Statistical Institute*, 47, 203–211.

——— (1977b), "Towards a Methodology for Statistical Disclosure Control," *Statistisk Tidskrift*, 5, 429–444.

Dalenius, T., and Reiss, S. P. (1982), "Data Swapping: A Technique for Disclosure Control," *Journal of Statistical Planning and Inference*, 6, 73–85.

Dalenius, Tore, and Denning, Dorothy E. (1982), "A Hybrid Scheme for Release of Statistics," *Statistisk Tidskrift*, 10, 97–102.

De Finetti, Bruno (1975), *Theory of Probability* (Vol. 2), New York: John Wiley.

DeGroot, M. H. (1962), "Uncertainty, Information, and Sequential Experiments," *Annals of Mathematical Statistics*, 33, 404–419.

DeGroot, Morris H. (1970), *Optimal Statistical Decisions*, New York: McGraw-Hill.

Flaherty, D. H. (1979), *Privacy and Government Data Banks: An International Perspective*, London: Mansell.

Frank, O. (1978), "An Application of Information Theory to the Problem of Statistical Disclosure," *Journal of Statistical Planning and Inference*, 2, 143–152.

——— (1979), "Inferring Individual Information From Released Statistics," paper presented at the 42nd Session of the International Statistical Institute, Manila, Philippines.

Frank, Ove (1982), *Statistical Disclosure Control*, Technical Report 108, University of California, Riverside.

Ho, Monto, Pazin, George H., Harger, James H., Armstrong, John A., and Breinig, Mary C. (1982), "Consent Form for Treatment for Recurrent Herpes Genitalis With Topical Application of Human Leukocyte Interferon," University of Pittsburgh and Magee-Womens Hospital.

Internal Revenue Service (1979), *Statistics of Income—1976: Estate Tax Returns*, Washington, DC: Author.

Jabine, Thomas B., Michael, John A., and Mugge, Robert H. (1977), "Federal Agency Practices for Avoiding Statistical Disclosure: Findings and Recommendations," paper presented at the Annual Meeting of the American Statistical Association, Chicago, Illinois.

Johnson, N. L., and Kotz, S. (1970), *Continuous Univariate Distributions-1*, New York: John Wiley.

Johnson, N. L., and Kotz, S. (1977), *Urn Models and Their Applications*, New York: John Wiley.

Lampman, Robert J. (1962), *The Share of Top Wealth-Holders in National Wealth: 1922–1956*, Princeton, NJ: Princeton University Press.

Mugge, Robert H. (1978), "The Experience of the National Center for Health Statistics in the Sale of Public Use Microdata Tapes and Special Tabulations," in *Proceedings of the Social Statistics Section, American Statistical Association*, pp. 512–514.

——— (1983), "Issues in Protecting Confidentiality in National Health Statistics," paper presented at the Annual Meeting of the American Statistical Association, Toronto, Canada.

Reynolds, Paul Davidson (1979), *Ethical Dilemmas and Social Science Research*, San Francisco: Jossey-Boss.

Sagarin, E. (1973), "The Research Setting and the Right Not to Be Researched," *Social Problems*, 21, 52–64.

Singer, E. (1978), "Informed Consent: Consequences for Response Rate and Response Quality in Social Surveys," *American Sociological Review*, 1978, 43, 144–162.

Stark, Thomas (1972), *The Distribution of Personal Income in the United Kingdom, 1949–1963*, Cambridge, U.K.: Cambridge University Press.

Subcommittee on Disclosure Avoidance Techniques (Federal Committee on Statistical Methodology) (1978), Statistical Working Paper 2 (Federal Statistical Policy and Standards), Washington, DC: U.S. Dept. of Commerce.

Wilson, O., and Smith, W. (1983), "Access to Tax Records for Statistical Purposes," in *Proceedings of the Survey Research Methods Section, American Statistical Association*, pp. 595–600.