Methods for Ensuring that Statistical Information Does Not Reveal Underlying Individual Data

Paul B. Massell
Disclosure Avoidance Research Group
Statistical Research Division
U.S. Census Bureau
paul.b.massell@census.gov

U S C E N S U S B U R E A U
Helping You Make Informed Decisions

Overview: Protecting Statistical Information by Fine-Tuning Uncertainty

General Idea: Statistical Information (summary measures, tables, models) is generated from data on individuals but before releasing the information, the stat office wants to protect (from disclosure) the individuals' data (or certain group data.)

A statistical office needs to measure the disclosure potential in a given information object before releasing it.

Once measured, it may be necessary for office to add uncertainty to object to lower disclosure potential to an acceptable level; hopefully doing only minimal damage to the object's usefulness to data users.

The Census Bureau is required by law to protect all its individual data from disclosure when released to the public. On the other hand, once the Bureau is confident that the data are protected, it is obligated to share as much of this data with the public as is useful.

Census “never” shares unprotected data with other agencies.

Outline:

- Overview
- Background: types of protection methods
- Recovery of a "Value of Interest" by a user
- Types of estimators
- Examples: various estimates of uncertainty of data contributions to table cells
- Generalization of Protection Methods

Background: types of protection methods

Tables: protected with cell suppression

Much research on protection of additive statistical tables. (ref: WP22)

Standard techniques are based on operations research, methods, e.g., mathematical programming (e.g., network flow, LP)

Tables: protected with EZS noise

Some newer protection methods involve addition of noise to the underlying microdata (ref: Massell, Zayatz, Funk)

These methods are purely statistical and depend on the "law of moderate numbers" for their success. They are easy to implement in statistical packages.

Microdata: Eliminating unique records

Ex: demographic or health related info about individual residents or households.
This protection is statistical in nature.
Based on the idea of "unique records" w.r.t. certain key variables known by data users.

Statistical Models

Little research has been done on analysis of disclosure potential for statistical models.

(ref: D. Merrell and A. Reznek, papers at STATCAN and FCSM conferences, 2001)

Conclusions from Merrell & Reznek FCSM paper:
Typical regression models involving cont. variables pose few disclosure risks.
However, categorical input variables create tables that need to be examined.

Much more research is needed.

Recovery of a "Value of Interest" by a user

Consider a continuous variable that is often sensitive, e.g., income.
We wish to protect it in statistical products.
In general, individual values are "not recoverable" from any type of statistical summary that is released (i.e. published).

By statistical summaries we have in mind

(1) simple statistical measures such as a total, average, median, quartiles, etc.

(2) statistical tables (usually additive)

(3) simple statistical models (e.g., a regression equation)

"not recoverable" includes cases in which only a crude approximation can be derived from the data product released.

E.g., with 'income', if we assume the value is non-negative and a total T (based on at least 2 individuals) is released,
a data user can easily derive that $v(i) \leq T$, for any individual 'i'.

This result will generally be so rough that it would not constitute a disclosure.
Also not interesting to a data user.

However, under certain conditions, a data value may be “nearly recoverable”, i.e., a good approximation can be derived with confidence by a well informed data user (defined below).

Below we will give some simple examples of how such good approximations can be derived.

Types of estimators

We need to describe the information and assumptions on which such approximations are based.

Estimate 1, denoted ‘est1', is derived using only statistical info ‘contained’ in product. Typically data products refer to data collected for a specified reference period.

Will know income is non-negative.

We can assume well informed data users will know these basic stat-info properties.

Estimate 2, denoted ‘est2', is derived using all sources (data and metadata) for est1, plus some assumptions based on a (simple) data user knowledge model (‘DKM’).

For example, we may assume that the best informed data users know, by name, all the contributors to a cell and rough approximates of all contributions to product.

Example: suppose there are a small number of companies involved in a certain specialized economic activity in some city.

It may be reasonable to assume that this set of data users would know the contributed values within an order of magnitude (where the ‘base’ may be ‘2', rather than the usual ‘10'). (a Bayesian like model)

e.g., if the true value is 100 for some contributor, the data model yields a prior knowledge interval of [0, 200], but no user knows if this model applies to him.

Estimate 3, denoted ‘est3', is based on information for ‘est2' plus
info derived from independent data sources, e.g., free web & commercial databases.

Examples: various estimates of uncertainty of data contributions to table

We can illustrate est1 and est2 with simple numerical examples. Assume values are from a census and are unweighted.

Example 1:
Data contributor values: 5 and 95.
Agency releases only n =2 and the total 100.
Suppose an independent data user, wishes to estimate the larger contribution ‘xL’ to total.
Using only T, he would know only that est1(xL) lies in [0,100].
Clearly $xL \geq T/n = 50$.
Thus est1(xL) lies in [50,100].

Claim: est2 sometimes improves on est1
Using the DKM “order2 “ above implies that a well informed data user would know that the smaller value, xS, lies in [0,10].
Thus est2(xL) is much more precise than est1(xL);
est2(xL) lies in [90,100].

If a data user were also a data contributor to a particular cell,
and if he knew the final edited version of his value, then he could estimate xL exactly.

A data contributor to a cell has “inside knowledge” and can make better estimates about other contributions to that cell, than a data user who is merely well informed.

Example 2:
Case 1:
Suppose a cell value is the sum of 3 contributions, x1= 2, x2 = 3, and x3 = 95. Agency releases only n =3 and T = 100. Then any data user could deduce that est1(xL) lies in [T/n, T] = [33.33, 100]. An

informed data user, wrt ‘DKM’ would estimate that est2(x1) lies in [0,4] and est2(x2) lies in [0,6]. Then he could deduce that est2(xL) lies in [90, 100]. This is just as precise as the est2 estimate in example 1.

Case 2:

Let n= 3, but set of contributions has a smaller range, i.e. xL - xS is smaller.

Let x1 = 10, x2= 25, and x3 = 65.

Then est1(xL) lies in [33.33, 100].

However, the intervals for est2(x1) and est2(x2) are much wider than in Ex2:case 1;

est2(x1) lies in [0,20] and
est2(x2) lies in [0,50].

Thus est2(xL) lies in [33.33, 100], it is no more precise than est1(xL).

Example 3:
Suppose a cell value has n=10;
values are 2, 4, 6, 8, 10, 12, 14, 16, 18, 20
agency releases only n = 10, and T = 110.

Any data user can show est1(xL) lies in [11, 110].

A data user knowing the ‘order 2' model would know that est2(xL) lies in [0,40]. His estimates of x1, ..., x9 has a cumulative uncertainty of plus or minus 90 units about the true value of 90, so his knowledge of these values cannot be used to improve his direct estimate, est1, of xL.

Generalization of Protection Methods

These simple numerical examples illustrate some basic properties of estimation of a single value of interest (a 'VoI'), from statistical information.

1. For a fixed number of data values, the more the value of interest (VoI) "sticks out"

from the other values, the more we can improve on our (prior) data knowledge model estimate of the VoI.

2. Given a total T, the estimate of a VoI will improve as sum of other values decreases. Assumes that the DKM has the property that the uncertainty created by sum of other values is an increasing function of that sum.

3. Thus, if an agency wishes to release a table in which all VoI's are protected, it should ensure that the VoI's do not dominate T. (i.e. exceed a % threshold of T).

Similar results probably hold for statistical models

4. Protection of Tables; Weights

Most common stat data products released by federal statistical agencies are tables.

For tables, these general rules take the following form:

All individual contributions to a cell value are VoI's.

When a cell value consists of a single contributor it is sensitive and must be modified or suppressed. Similarly if a cell value has only 2 contributions, and one of the contributors is also a data user.

A data contributor can get a better estimate providing he knows the final version of his contribution that is used in data product. May not know because of sampling weights. (Note: sampling weights add uncertainty to recovery process)

But when data users do know the final version of their contribution, and when a cell value is ‘essentially’ the sum of 1 or 2 contributions, i.e., the rem= T - x1 - x2 is small’, then the top 2 values are not protected from the top 2 contributors. (ref: WP22, p% rule)

References

Massell, Paul B., (2006) "Using Uncertainty Intervals to Analyze Confidentiality Rules for Magnitude Data in Tables"; http://www.census.gov/srd/papers/pdf/rrs2006-04.pdf

Paul Massell, Laura Zayatz, Jeremy Funk, "Protecting the Confidentiality of Survey Tabular Data by Adding Noise to the Underlying Microdata: Application to the Commodity Flow Survey", appears in: Josep Domingo-Ferrer, Luisa Franconi (Eds.) :Privacy in Statistical Databases, CENEX-SDS Project International Conference, PSD 2006, Proceedings. Lecture Notes in Computer Science (LNCS) 4302, Springer 2006, ISBN 3-540-49330-1.

David Merrell, Arnold Reznek,
“ On Disclosure Protection for non-Traditional Statistical Outputs”,
www.statcan.ca/english/freepub/11-522-XIE/2001001/session17/s17c.pdf

Federal Committee on Statistical Methodology (FCSM) Working Paper 22 (2005) (definition of p% rule, both standard and extended versions)
http://www.fcsm.gov/working-papers/spwp22.htm

U.S. Census Bureau's webpage on disclosure avoidance
http://www.census.gov/srd/sdc

Appendix 1. Generalizations : Model Development and Disclosure Steps

1. Formulate inputs and outputs for models of interest

2. Collect microdata for model development (sources: survey, admin, criminal records, national security classified documents, etc.)

3. For individual values protection:
decide which variables are sensitive and how much protection is needed for each value (using a formula ?)

4. For group protection:
Decide which groups are sensitive and decide how much uncertainty is required for each value

5. Develop tables or other types of models and determine which components of the model are sensitive. For those components try to add just the right of amount of uncertainty to make the model releasable.