RESEARCH REPORT SERIES
(Statistics #2006-4)

Using Uncertainty Intervals to Analyze Confidentiality Rules for Magnitude Data in Tables

Paul B. Massell

Statistical Research Division
U.S. Census Bureau
Washington, DC 20233

Report Issued: March 21, 2006

Using Uncertainty Intervals to Analyze Confidentiality Rules for Magnitude Data in Tables

Paul B. Massell
Statistical Research Division
U.S. Census Bureau

Abstract

Protecting the confidentiality of survey respondent data is related to the notion of data user uncertainty in various ways. The source of uncertainty that is most frequently exploited by agencies in formulating protection rules for tabular data is the fact that there is often more than one respondent (e.g., a company) contributing to a given table cell value. Agencies are required to protect these individual contributions. The uncertainty in a data user's mind about how the published cell value is distributed among the contributions is often sufficient to protect them. This "cell value distributional uncertainty" may be the most exploited source of uncertainty, but it is by no means the only one. Data user uncertainty about respondent contributions is created through many of the procedures involved in the design of a survey and in processing the collected data. It is usually possible to express a given data user's uncertainty about a particular respondent's contribution to a particular cell as a finite interval. The interval may be derived from inequalities associated with the table's additivity or it may be based on "knowledge models" that describe, for example, the data user's prior (approximate) knowledge of respondent contributions or sampling weights. We call such intervals "uncertainty intervals". Sometimes the knowledge models may allow a probability distribution to be defined on the uncertainty interval. The major thesis of this paper is that uncertainty intervals can be used as a means of unifying the description of many of these sources of uncertainty. We show how uncertainty intervals can unify the description of several formulas and algorithms that are frequently used during the process of protecting data, e.g., those related to the p% rule, sliding and two-sided protection, cell value rounding, and weights applied to the underlying microdata. In future work, the author hopes to extend this approach to additional sources of uncertainty.

Keywords: confidentiality, disclosure protection, p% rule, midpoint attack, uncertainty interval, uncertainty model, knowledge model

Table of Contents

1. The Role of Uncertainty in Protecting the Confidentiality of Data
2. The Basic p% Rule and Some Extensions
3. Using Uncertainty Intervals to Measure Interval Protection
4. Comparing Mathematical and Statistical Approaches for Protection Rules
5. Conclusions

1. The Role of Uncertainty in Protecting the Confidentiality of Data

1.1 Uncertainty from Survey Errors and from Survey Processing Procedures

The idea of uncertainty plays at least two important roles in the description of data products that are released by statistical offices. One role is the uncertainty associated with survey errors in the survey microdata. Most of these errors are related to the design of the survey and collection of data. The other role is uncertainty that is introduced during various data processing procedures applied to the data prior to release of the data products. A procedure such as rounding will introduce uncertainty about the true values, but this may not be the primary reason for using this procedure. By contrast, disclosure avoidance (or control) procedures have, as their primary purpose, the addition of enough uncertainty to fully protected data elements (e.g.. cell values) from disclosure as defined by the statistical office (SO), (e.g., a government agency).

The outline below attempts to list the major sources of uncertainty during the collection and processing of survey data, including the construction of data products to be released by the statistical office and the protection of data in these products.

Uncertainty Sources for Survey Data

1. Survey Design (generates sampling error)
2. Data Collection (generates measurement error)
3. Data Cleaning (e.g., editing)
4. Data Smoothing of Microdata
   (e.g., top coding, categorization, rounding of microdata values)
5. Data Smoothing involved in Statistical Tables and Models
6. Confidentiality Protection of Data

If a survey error can be estimated well, it should be possible to use the error estimate to express the amount of uncertainty the error introduces and then express the associated amount of protection derived from the uncertainty. For example, we will show how sampling weights are used in the extended version of the p% rule to determine which cells in a table need to be protected. (These weights are also used to calculate sampling error associated with cell estimates.) When using weights in protection rules, one needs to consider if the weight is known by the best informed data users. If a weight is not known by any data users, then it provides some protection. For example, if a given sampling weight is large and is unknown by all data users, it may, by itself, protect the cell values to which it applies. Generally the SO will have reliable estimates of the sampling weights and some other parameters used in the survey processing procedures. For each such parameter, ideally the SO could estimate the amount of uncertainty that the parameter contributes to a data user's estimate of a data element requiring protection. Typically the SO assumes that some users have approximate knowledge of such parameters. The SO also assumes users have approximate knowledge of certain microdata values, e.g., those that can be estimated from a variety of publicly available data sources. These assumptions form what we call a "data knowledge model".

It is usually possible to express a given data user's uncertainty about a particular respondent's contribution to a particular cell as a finite interval. The interval may be derived from inequalities associated with the table's additivity or it may be based on "knowledge models" that describe, for example, the data user's prior (approximate) knowledge of respondent contributions or sampling weights. We call such intervals "uncertainty intervals". We have chosen not to use the term "confidence interval" because as defined in mathematical statistics its meaning is too narrow for our purposes. It is used in the context of estimation of parameter values, not uncertainty about a particular data value. The term "error interval" is sometimes used for expressing uncertainty about a data value, but it is usually used when a statistical distribution has been assumed. By contrast, the term "uncertainty interval" seems appropriate for problems involving either deterministic uncertainty (e.g., inequalities) by itself or statistical uncertainty by itself or some combination. This term seems to be used mainly in some recent papers in engineering and geophysics (see bibliography on uncertainty and uncertainty models below).

As each source of data user uncertainty is incorporated in the analysis, we are able to develop a more precise measure of the amount of additional uncertainty that needs to be created by a disclosure avoidance method to ensure all data elements are protected. Many of the calculations related to these ideas can be performed with uncertainty intervals and with simple operations on them. This paper contains many such examples.

The major thesis of this paper is that uncertainty intervals can be used as a means of unifying the description of a variety of sources of uncertainty that play a role in statistical disclosure avoidance of tabular data. We show how uncertainty intervals can unify the description of several formulas and algorithms that are frequently used during the process of protecting data, e.g., those related to the p% rule, sliding and two-sided protection, cell value rounding, and weights applied to the underlying microdata. In future work, the author hopes to extend this approach to some of the other sources of uncertainty listed in the outline above. This work was motivated by a feeling that certain protection formulas developed by the Federal Committee on Statistical Methodology (FCSM), Greenberg, Zayatz, and Sande (see references) could be developed within a common framework.

Information Aspects of the Data Operations Applied to Confidential Survey Data

PART I. Microdata Acquisition, Data Processing, and Formation of Tables

1. Data that Require Protection from Disclosure

(example: values of specified magnitude variables that appear in individual microdata records, and various specified sums of those values, e.g., those that belong to a single company)

2. Information Reducing Operations at the Microdata Level

(examples: applying sampling weights to microdata if the weights are unknown to all data users; editing data; adding microdata-level noise)

3. Information Reducing Operations during Formation of Tables from Microdata

(example: often more than one microdata record is associated with a given table cell, thus the cell value equals the sum of individual values for a specified magnitude variable and some information about the microdata values is lost)

4. Information Reducing Operations on Table from Initial to Final Version

(examples: rounding table values; suppressing sensitive cells and complementary cells; adding tabular-level noise to cell values)

PART II. Estimation of an Individual Magnitude Value

5. Use Table Additivity to Derive an Estimate for a Specific Value of Interest

(example: Use additivity of the published table or algebraic relationship among a set of linked tables to derive a relationship between an individual magnitude variable value of interest, and quantities either in the table(s) or in the underlying microdata)

6. Apply Estimates of Estimator Quantities from Direct Knowledge or Knowledge Models

(example: if data user is a survey respondent he may know some values with more precision than a typical data user; there may be sharing of data among groups of data users; data users typically have rough estimates of all values and weights and the set of these rough estimates can be expressed as "knowledge models.")

7. Express Estimate of Value of Interest as an Uncertainty Interval Possibly with a Density

(example: typically the estimator will allow the data user to derive a finite interval that the data user knows with confidence contains the value of interest. In special situations it may be possible to derive a density function for this interval that allows the data user to derive a more precise estimate; we call this an "uncertainty model.")

1.2 Uncertainty from the Distribution of Contributions to a Cell Value

There is a simpler type of uncertainty that arises when the SO is trying to protect microdata values that are summed to produce a magnitude value for a cell in a table. In that situation, the fact that there are usually contributions from at least a few different companies may protect the contribution from any one company. However, even a large number of contributions may not be sufficient to fully protect the largest one. The amount of protection that the largest contribution receives depends on the distribution of values of the other contributions. Because of this property, we use the adjective "distributional" to describe this type of protection. (See section 2.1 below to see how this type of protection can be quantified).

If the combined effect of distributional uncertainty and uncertainty from the sources listed in the outline above is not sufficient to protect the cell value, the SO must select a protection method to create additional uncertainty. Protection methods often involve complicated algorithms. In addition, they may be based on general assumptions about user knowledge and the extent of data sharing, i.e., collusion, among data users. The class of data users that are most important to consider in the context of protection are the data contributors. One type of assumption that is often made involves user knowledge of the data related to the table about to be released; we call this 'a priori' knowledge. Since uncertainty about data values varies among data users, in a given calculation, it is helpful to specify the user we are referring to. For example, in the p% rule (defined in section 2.1 below) it is common to compute uncertainties from the point of view of the company that has the 2nd largest contribution to a cell. Some protection software incorporates such assumptions about 'a priori' user knowledge. For example, some cell suppression programs assume that the best informed users know that the value x of company C's sales for a given year lies in the interval [0, α*x] where $\alpha \geq 2$. In such a data knowledge model, α depends on the variable being tabulated (e.g., sales) and possibly also on C and x. To be useful in protection programs, the agency does not need to know the value of α exactly, but only to have a lower bound for it that is greater than one.

1.3 Role of Table Additivity in Protection of Cell Values

Most of the above discussion is general enough to apply to the protection requirements of individual data items in a microdata record or a cell value that represents either a count or a magnitude variable. However, in this paper our focus is on tables whose cells contain the value of some magnitude variable that is the sum of contributions from one or more respondents. Usually, such tables that are released by SO's are additive, e.g., a 2D additive table would have a sum row and a sum column. This property plays a major role in the protection process. It causes each cell value to be at least weakly coupled with all other cells in the table. Specifically, to increase uncertainty about cell k's value requires increasing uncertainty in at least one other cell in the row and in the column of cell k. If no other cells in cell k's row or column were modified, additivity would allow immediate recovery of k's value. The cells which require increased uncertainty based on a cell level protection rule (e.g., the p% rule) are called primary (sensitive) cells. Those cells which require increased uncertainty simply to prevent the primary cell values from being recovered using table additivity are called secondary (or complementary) cells.

1.4 List of Common Sources of Uncertainty for a Magnitude Data Table

For concreteness it is helpful to describe a typical magnitude data table formed from survey data which is rounded when cell values are formed. For such a case, there are several sources of uncertainty that protect cell value contributions from any data user who is trying to estimate them.

(1) User has only approximate knowledge of contributions from establishments or companies based on earlier SO releases and other public sources.
(2) Data processing operations, such as editing, may cause a data contributor to be uncertain about the precise value of his own contribution that is used by the SO in table formation.
(3) Rounded data are less precise than the original (i.e., reported) data.
(4) Sampling introduces uncertainty because the sampling weights are not released by the SO.
(5) If the above sources of uncertainty do not protect all data elements, a protection method will be used to create additional uncertainty. Secondary cells may require modification due to table additivity.

1.5 Goals and Structure of Protection Procedures

The overall goal of uncertainty creation for confidentiality protection may be described as: create enough uncertainty of any individual establishment or company value to protect the confidentiality of each respondent's data as required by the SO's confidentiality policy.
This general goal may involve these sub-goals:
(a) provide adequate "interval width protection"; i.e., create an uncertainty interval for each sensitive data value (may be a microdata value or a cell value) that is wide enough; specifying either minimums for total interval width (which provides "sliding protection" (see below)) or for both right and left sided interval widths (which provides "two-sided" protection).
(b) create densities on the above uncertainty intervals that make it difficult for a data user to construct a reliably good estimator of sensitive data values; i.e., protect against data user attacks that are based on mathematical or statistical calculations using intervals that a data user can compute.

1.6 Classification of Data Values Used in Protection Procedures

In the context of protection procedures, it is useful to divide the set of data values into 3 subsets
(S1) external data values; i.e., data from sources other than the data product being protected
(S2) data values that can be read directly from the data product
(S3) data values estimated from values in S1 and S2.

The data values requiring protection often fall into this S3 (e.g., let x be a contribution to a cell value T; x is a microdata value and would be in S3 whereas the T would be in S2). Specifically, a table user might compute the largest contribution to a cell, denoted x1 using:
$x1 = T - x2 - x3$ where T is the cell value and the estimates of x2 and x3 are gotten from external sources. Clearly the SO cannot increase the uncertainty of values of S1, but it could

decrease it. The SO has direct control over only those values in S2. The SO decides how much uncertainty to create for S2 values as follows. It uses a protection rule (e.g., the p% rule) that relates protection requirements for each S3 value that contributes to an S2 value, to the additional uncertainty needed for an S2 value. For example, the SO must know how protection requirements for microdata relate to uncertainty for cell values. Then the SO must select a protection technique (and software) for creating the amount of uncertainty, if any, required for the S2 value.

2. The Basic p% Rule and Some Extensions

2.1 The Basic p% Rule

Example 2.1
Assume we have a statistical table with magnitude data in each cell. For concreteness, assume the value in the cell represents the sum of sales for a given year for all establishments that "belong to the cell", e.g., all shoe stores in Akron, Ohio. Suppose there are four such establishments, E1, E2, E3, E4, that are components of the companies C1, C2, C3, and C4 respectively. Suppose these establishments have corresponding contributions x1, x2, x3, and x4 that satisfy $x1 \geq x2 \geq x3 \geq x4$. If confidentiality were not a concern, the sum of these four sales values, T, where T=x1+x2+x3+x4 could be published. However, if T were published and if x3 and x4 happened to be very small, then company C2, viz., the company that contributed the value x2 for its establishment that belongs to this cell, could subtract x2 from T and get a very good estimate of x1. Of course, C2 would think that T-x2 is a good estimate of x1 only if he were fairly confident that the difference T - (x1 + x2) was a small percentage of T. This situation might mean that the data provider has not protected the confidentiality of C1's data. To know if it actually implies a failure to protect C1's data, we need to have a specific numerical criterion.

Basic Idea of the p% Rule

The basic idea of the p% rule is that C1's data value is protected (from C2 and all other users) if no user can confidently derive an estimate of x1 that is within p% of the true value.

Sensitivity Measures

Let us introduce the notion of sensitivity measures as used in disclosure avoidance theory. Linear sensitivity measures are a subclass of the class of linear subadditive measures (Willenborg and de Waal, p. 114; FCSM [WP22], p.45) A sensitivity function is a concise way of expressing a confidentiality rule. The confidentiality rule is an expression of a general confidentiality policy that is determined by the SO. The rule itself may have one or more parameters p1, p2, ..., pB. Typically the rule and its parameters are fixed for a given set of tables being processed. Then the sensitivity function corresponding to the given rule is written S(X; p1,p2,...,pB), or S(X), once the rule and its parameters have been defined, where X denotes a typical cell X in the tables being tested for sensitivity. The variable X here represents all information about the cell that is used in determining its sensitivity; e.g., the cell value T and its contributions from the underlying

microdata. Then we define cell X to be "sensitive" if and only if S(X) > 0. If cell X is sensitive it must be altered in some way before it can be published.

Statement of the p% Rule for Two-sided Protection (FCSM [WP #22] formulation, p. 46)

The p% rule may be viewed as a linear sensitivity measure. If T denotes the value of cell X for some magnitude variable (e.g., sales, profits, number of employees), and x(i) represents the ith contribution to T, where the x(i) are given in descending order, then the p-percent rule can be defined as S(X) = x1 - (100/p) * (Sum of x(i) as i goes from c+2 to N). Here c is the size of a coalition, a group of respondents who pool their data in an attempt to estimate the largest contribution and N is the total number of contributions to the cell. The cell is declared sensitive if S(X) > 0. (In our examples, we assume respondents do not share information. In that case c=1 and the sum begins with i=3). Note that the single parameter 'c' allows a simple model of data sharing (i.e., collusion) among data users.

For the p% rule, the sign of S(X) is scale invariant in the following sense. Let xvec be the vector formed from the ordered components of T; xvec = (x1, x2, ..., xN). Let yvec = (λ*x1, λ*x2, ..., λ*xN) where 'λ' is a positive constant and let Y be a cell with components yvec. Then S(X) > 0 if and only if S(Y) > 0. This property is important because it allows us to rewrite S(X) in the form: S(X) = (p/100)*x1 - Sum of x(i) as i goes from c+2 to N

This form has a nice interpretation in terms of uncertainty. The 1st term, (p/100)*x1, is the one-sided perturbation needed to protect cell X and the 2nd term is company C2's amount of uncertainty about x1. C2's uncertainty is due to his lack of knowledge of the 2nd term, often called the remainder, denoted 'rem'. Then the above inequality says that if rem < (p/100)* x1, the cell is sensitive. In more general terms, the inequality says the cell is sensitive if the perturbation needed to protect x1 is greater than C2's uncertainty about x1. To make this interpretation complete, we need to specify a "location rule" to accompany the p% rule. The two common choices are two-sided protection and sliding protection. In each of our examples below, we will specify one of these.

If we now specify that the **location rule is two-sided protection**, we have the following formulation. Let C2 define his least upper estimate for x1 (denoted x1U) and his greatest lower estimate for x1 (denoted x1L) by setting

x1U = T - x2 and x1L = T - x2 - 2*rem ;

then x1 satisfies the p% rule with two-sided protection if both of the inequalities

x1U - x1 > (p/100)*x1 and x1 - x1L > (p/100)*x1

are satisfied. The estimates x1U and x1L use the following "x(i) knowledge model": if contributor i has contribution x(i), C2 will know only that the x(i) lies in the interval [0, α*x(i)] where $\alpha \geq 2$. The exact value of α depends on C2 and x(i), but only a lower bound for α is needed in protection programs that use it.

If the location rule is **sliding protection**, then the total interval length needed for protection is twice the 1-sided protection given above. Thus to have sliding protection with the p% rule, the

uncertainty interval for C2 must have width > 2*(p/100)*x1. For the data knowledge model given above, in which $\alpha \geq 2$, one can show that sliding protection is equivalent to two-sided protection (Massell, 2005a). Of course, sliding protection is useful as an alternative to two-sided protection when it creates an uncertainty interval that is closer to x1 on at least one side of x1 than the corresponding two-sided interval. This can occur if one uses a model of weak knowledge; e.g., x(i) lies in the interval [0, α*x(i)] where $\alpha > 2$. Setting α to be a large value is equivalent to assuming users know only that contributions are non-negative.

Statement of p% rule for two-sided protection (Massell formulation)

Suppose a user applies an algorithm to table values and to a set of publicly known estimates of the magnitude variable of interest to derive an uncertainty interval [a,b] for x1. Suppose the user is quite confident that the true value of x1 lies in the interval [a,b]. Assume the user's confidence can be justified using standard mathematical and statistical methods. That is, assume one (not necessarily the user) can prove using such methods, that x1 lies in the interval [a,b] with high probability (say, Pr $\geq$ 0.95). Then if the interval contains x1 and either of the subwidths (b-x1) or (x1-a) is less than p% of x1, that we say x1 has **not** been given two-sided protection in accordance with the p% rule. If no such interval exists, we say x1 **has** been given two-sided protection in accordance with the p% rule.

Remark: The expression "can be justified using standard mathematical and statistical methods" is inserted because we want to rule out methods for estimating x1, that may, on occasion, yield a very good estimate. For example, a pure guess may sometimes produce such an estimate. To fully protect data we need to consider protection against only mathematical and/or statistical attacks, i.e., attacks that are based on reasonable assumptions and statistical estimates (e.g., expected values).

2.2 The p% Rule for Survey Data with Weights

Let us first consider the case of a single weight. Perhaps the most familiar example of such a weight is a sampling weight. In the most general case, the weight depends on the microdata record that contributes to a cell; thus weights assigned to the various records associated with a given cell may differ. For example, a cell may have a contribution from a "certainty" company, with a sampling weight equal to 1, and another contribution from a randomly selected company, with a sampling weight greater than 1.

In that case, FCSM [WP22], p. 89, says the basic p% rule could be extended as follows. Define y(i) = w(i)*x(i) for each contribution i. Then let T = the sum of weighted contributions, i.e., the sum of the y(i). Then define S(X) for the extended p% rule by:

S(X) = (p/100)*x1 - rem where rem = T - x1 - x2

This extension seems to be based on the assumption that the weights w(i) are not known to the data users. This uncertainty contributes to the protection. It is probably reasonable to assume that users know something about the value of a weight for a given contribution. Various models could

be developed after taking an informal survey among data users in which they are asked to estimate weights. For example, consider the "sampling weight knowledge model" (a term we are introducing here) that posits that the best informed data users know that a weight 'w' lies in the interval [1, $\beta^{(w-1)}$] where $\beta > 1$. It implies that users know the weights of companies that have a weight of 1, i.e., the "certainty companies." As with the data knowledge models discussed above (see p% rule for two-sided protection), in order for a sampling weight knowledge model to be useful in protection programs, the agency has only to use a fairly small lower bound for β, e.g. 2, (assuming this is consistent with the informal survey of data users). See example 3.4.2 below for a calculation that uses both the a data knowledge model and a sampling weight knowledge model to derive an uncertainty interval for x1.

Occasionally values of a weight or scaling factor are known to some data users. For example, a contribution from a respondent may reflect a weekly total which is multiplied by 13 by the agency to produce a quarterly total. It is likely that many users know this scaling factor, so it cannot be assumed to add to the protection of the original contribution. For such weights or scaling factors w(i), one uses the y(i) rather than the x(i) in the expression for rem; i.e.,

rem = T - y(1) - y(2) where T is the sum of the y(i).

3. Using Uncertainty Intervals to Measure Interval Protection

3.1 Interval Protection Based on the p% Rule

Interval protection can be calculated using addition and subtraction of uncertainty intervals. Suppose using T, x2, and publicly known estimates for the other x(i), C2 can determine x1 to within p% on at least one side.

Example 3.1 Let p=10. C2 knows his own contribution x2=50 exactly. Suppose C2 also knew T =150 exactly. Suppose x1= 92, x3 = 5 and x4=3, but C2 does not know these values exactly but did know that there are 4 contributors to the cell; x3 lies in the interval [0,10] and x4 lies in the interval [0,6]. Then using only subtraction, C2 deduces that x1= T-x2-x3-x4 = [150,150]-[50,50]-[0,10]-[0,6]=[84,100]. C2's interval for x1, viz. [84, 100] is symmetric about the true value 92. However, at least one side is less than 10% (9.2 units) of the true value 92 (in fact, here both sides are within 10%; viz., they both equal 8 units). Thus C1's confidentiality would be violated according the p% rule interpretation. Therefore we must somehow make the data available to C2 less precise. Assume we cannot change C2's knowledge of x2 (although it may be possible to change this; see below). Also, C2's knowledge of x3 and x4 are based on public information; i.e., not derived from the table being protected.

What is the minimum possible change to T that would lead to a sufficiently wide uncertainty interval for C2's x1 estimate of x1 ? Suppose we replaced T = [150,150] with [148,152]. Then C2's estimate of x1 equals [148, 152] -[50,50] - [0,10] - [0,6] = [82,102]. This is slightly more than 10% on each side of the true value and thus is sufficiently wide. That is, if we published the interval [148, 152] rather than the single value of 150, we would protect x1.

3.2 Calculating Densities for the Uncertainty Interval; the Midpoint Attack

To further understand the degree of uncertainty created for C2, we need to consider what density respondent C2 is able to assign to the uncertainty interval (denoted 'UI'). The simplest assumption would be a uniform density on each interval, i.e., the ones for T, x3, and x4. However, after the subtraction is performed, the density for x1 would not be uniform but rather a symmetric unimodal density whose mean (and mode) equals the midpoint of the C2's UI for x1 [denoted: midpt(UI(x1,C2))] which equals the true value x1. (This is based on the result that any linear combination of random variables, each of which is uniform on some finite interval, will be unimodal symmetric on the "combined interval" (e.g., [a,b]+[c,d]= [a+c,b+d] or [a,b]-[c,d]=[a-d,b-c] if a,b,c,d > 0). Thus if C2 selects midpt(UI(x1,C2)) he will get the (true) value of x1, though he will not be certain that his estimate is this good. The term "midpoint attack" has been used to refer to the selection of the midpoint of an UI as a point estimate of any unknown quantity. This raises some interesting questions:

(i) is protection according to the p% rule sufficient, or is it providing only one type of protection, which might be called "width interval protection" ? In other words, is creating an uncertainty interval with a width as specified by p% rule, in conjunction with a selected location rule (sliding or two-sided) really protecting the cell contribution x1 ?
(ii) it may be useful here to begin using the term "fully protective" to describe methods that not only provide "width interval protection" but provide protection against the midpoint attack. More generally, one could say a protection procedure is "fully protective" if it protects a given value x1 from data users by creating an uncertainty interval that is sufficiently wide (to meet requirements of the p% rule and the selected location rule), and has an asymmetry w.r.t. x1 and density to protect it against a midpoint attack and any other (to be specified) attacks that a data user could realistically use to estimate x1. Obviously, this is only a fuzzy definition of "fully protective"; to make it more precise one needs to specify and describe the other attacks against which a defense is required.
(iii) in general, if the density of the UI is unimodal symmetric; is the p% rule fully protective ? If the p% rule is not fully protective, how can it most easily be "enhanced" to make it "fully protective" ? For example, is it possible to find data operators (e.g., rounding, noise addition) that modify the UI's for the inputs that are asymmetric and lead to an asymmetric UI for x1 ?
(iv) Under what conditions might the UI for at least one of the inputs be asymmetric. ?

We could directly construct an interval for T that is asymmetric about the true value of T and publish that interval rather than T itself (Sande, 2003). Alternatively, we could use a data operator, such as rounding, that produces a rounded value of T (see below). Let us introduce some more notation for random variables on an UI. Let 'dens(UI(x1,C2); StAs)' denote the density for X1 on UI based on knowledge model assumptions where X1 is the random variable associated with the value x1.

3.3. Rounding

Perhaps the simplest way in which the UI for T can become asymmetric is conventional rounding, which is a protection scheme that operates on one cell at a time; i.e., table additivity which "couples" the cells plays no role in this type of rounding. This asymmetry property is easily seen with an example. Suppose x1=96 but x2, x3, x4 have the same values (50, 5, and 3, respectively) as above. Then T=154 is the unrounded cell value. Assume B=10 is the rounding base, and define the rounding process by requiring any cell value in the range [10*(k-1) + 5, 10*k + 4] to round to 10*k (here we assume published values are integers). (The contributions to the cell value, viz. x1, x2, x3, x4 are not rounded). Then the rounded value of T, r(T), is 150 and this value would be the published value, unless even this rounded value is sensitive according to the p% rule. Given r(T), a table user interested in estimating x1 will probably form the uncertainty interval [145,154] for T and probably assume a uniform density on it. Then the interval subtraction for x1 yields = [145, 154] - [50,50] - [0,10] - [0,6] = [79,104]. The midpoint of this interval is 91.5 which is 4.5 units below the true of x1=96. Thus the rounding leads to protection against the midpoint attack, which was our goal. The sub-interval to the left of 96 has width 96-79=17, whereas the sub-interval to the right of 96 has width 104-96=8. Since 8 is less than 10% of 96 = 9.6 (recall p=10), we do not have two-sided p% protection. However, we certainly have sliding protection, since the width of the full interval is 104-79=25, which is greater than 2*(9.6)=19.2 .

If we rounded to base 20 then r(T) = 160 and UI(x1,C2) = [150,169] - [50,50] - [0,10] - [0,6] = [84,119]. The midpoint of this UI is 101.5, thus est(x1,C2)=101.5 whereas x1=96. The left sub-interval of the UI has width, 96-84=12, and the right sub-interval has width = 119-96=23. Since both widths are greater than 9.6 (10% of 96), we have two-sided protection according the p% rule with p=10.

In summary, conventional rounding of a cell value can create a UI(T,C2) of various widths depending on the rounding base selected. The asymmetry of those UI's protects against the midpoint attack.

Let us discuss in more detail, how one assigns UI's for the basic input quantities. One needs to estimate the density that the most informed data user could place on each uncertainty interval for the input quantities, T, x2, x3, x4 and then compute the density of the difference T-x2-x3-x4. For estimating x1, C2 is the most informed data user (except for C1). We include x2 here because for any data user other than C2 there would be a uncertainty interval (of positive width) for x2. Since adjustments of contributed values are common due to editing or for other reasons, C2 may not know the exact value that is being used; i.e., UI(x2,C2) may have positive width; however its width would likely be less than the width of UI(x2,Ci) for any other table user, i.e., for $i \neq 2$.

3.4 General Formulas for Protection Requirements with Rounded Data

In this section we will develop general formulas for testing sensitivity and calculating protection needs (if any) for rounded data protected using either sliding protection or two-sided protection.

Since the formulas for sliding protection are simpler, we begin with those.

Algorithm Outline: Protection steps according to p% rule with sliding protection.
1. Compute uncertainty interval (UI) for data item x of interest
2. Check that x lies in UI
3. Compare width of UI (denote by W) with 2*(p/100)*x (denote by 2PX)
4. If W ≥ 2PX, item x is not sensitive (and needs no protection)
5. If W < 2PX then cell associated with x is sensitive and needs 2PX - W units of protection
6. If protection is needed, add units using any of various methods

Algorithm Outline: Protection steps according to p% rule with two-sided protection.
1. Compute uncertainty interval (UI) for data item x of interest
2. Check that x lies in UI
3. Let UI = [a,b]. Let [a,x] denote the LHS sub-interval. Let [x,b] denote the RHS sub-interval, let c = min(x-a, b-x), let PX = (p/100)*x
4. If c > PX, then cell associated with x is not sensitive; no protection is needed
5. If c < PX, then if
   5a. (x-a) < PX, need PX-(x-a) units on LHS sub-interval
   5b. (b-x) < PX, need PX-(b-x) units on RHS sub-interval
   {Note it may be simpler to add (PX - c) units to both sub-intervals}
6. Add needed protection units to whichever side needs it using any of various methods

Theorem 3.4 Let x be a contribution to a table cell and let UI denote the uncertainty interval of x from the perspective of some particular table user. If UI is symmetric about x, then protection using the p% rule with sliding protection is equivalent to protection using the p% with two-sided protection. This means that the conditions under which the cell associated with x is sensitive are equivalent and the amount of protection needed (if any) is the same for the two location rules.

Proof. If UI is symmetric about x, then in the algorithm outline for two-sided protection, we have c = min(x-a, b-x) = (W/2), where W is the width of the UI. Then c > PX, is equivalent to W > 2PX in the algorithm outline for sliding protection. Also, when rules 5a and 5b are combined, they are equivalent to adding to (W/2) to both sub-intervals; which in turn is equivalent to adding W to the full interval UI. QED

Example 3.4.1 Symmetric UI based on symmetric knowledge assumptions. In example 2.1, we considered the case where x1=T - x2 - x3 - x4 and assumed specific values for T and x2 and UI's for x3 and x4 that were symmetric about their actual values. This is easily generalized to the case x1 = T - x2 - rem where rem=remainder is the sum of all contributions less than x2. If we assume that C2 (the data user of interest here) knows x2 exactly, then we have UI(x1; C2) = T - x2 - UI(rem) where UI(rem) = [0, 2*rem]. If the actual value of T were to be published, then C2 would use actual value estimates for T and x2 and a symmetric UI for rem. Thus UI(x1; C2) would be symmetric.

Example 3.4.2 Let's generalize the calculation in example 3.4.1. Suppose we have both a data knowledge model (DKM) and a sampling weight knowledge model (SAKM), which need not be specified to illustrate this method. Suppose C2 is trying to estimate x1, where

w1*x1= T - w2*x2 - w3*x3 - w4*x4 (eq.1).

Assume C2 knows x2 exactly but does not know w2. C2 knows only that w(i) is in the interval [1, kw(i)*w(i)] where kw(i) are multipliers derived from the SWKM and each is $\geq$ 1. Likewise for i=1,3,4 the DKM says that x(i) is in some interval [0, kx(i)*x(i)] where each kx(i) is $\geq$ 2. Now we must perform each of the arithmetic operations in (eq.1) to find an UI for x1. C2's UI for the quantity w2*x2 is [x2, kw(2)*x2], for the quantity w3*x3 it is [0, kw(3)*kx(3)*x3], and for the quantity w4*x4, it is [0, kw(4)*kx(4)*x4]. Clearly C2's UI for the r.h.s. of (eq.1) ,

UI(w1*x1; C2) = [T - kw(2)*x2 - kw(3)*kx(3)*x3 - kw(4)*kx(4)*x4 , T - x2].

Let us denote this by [rhs1, rhs2].

UI(w1*x2; C2) = [rhs1, rhs2]

To find C2's UI for x1, we have only to "divide" [rhs1, rhs2] by the UI for w1, viz. [1, kw(1)].

UI(x1; C2) = [rhs1 / kw(1) , rhs2].

If one is using the p% rule to determine protection needs, one easily apply one of the algorithms at the beginning of this section to this UI, to see how much protection, if any, is required.

Example 3.4.3 Asymmetric UI from rounding. One of the simplest operations in which asymmetry arises for a UI, is when conventional rounding for T is used. Let the rounding base be B=2*b and let r(T) be the conventionally rounded value of T; i.e., if T lies in interval [(2k-1)*b, (2k+1)*b) then r(T)=2b*k (We assume published values are real numbers, not necessarily integer). Sliding protection:

UI(x1; C2) = UI(T) - [x2, x2] - UI(rem) =
= [r(T) - b, r(T) + b] - [x2, x2] - [0,2*rem]
= [r(T) -b - x2 - 2*rem, r(T) + b -x2]

Width of UI = W = 2*b + 2*rem
Using the algorithm outline for sliding protection given above, we compare W with 2PX, etc.

Using the algorithm outline for two-sided protection we get:
The calculation of UI(x1; C2) is the same as for sliding protection
The width of the RHS sub-interval = WR = r(T) + b - x2 - x1
The width of the LHS sub-interval = WL = x1 + x2 - r(T) + b + 2*rem
Since x1 + x2 = T - rem, replace x1 + x2 with T - rem
Then WR = T - r(T) + b + rem and WL = r(T) - T + b + rem
Let Wmin = min (WR, WL) = rem + b - abs{r(T) - T} where abs = absolute value
Wmax = max (WR, WL) = rem + b + abs{r(T) - T}
if Wmin < PX, add PX - Wmin units to smaller sub-interval
if Wmax < PX, add PX - Wmax units to larger sub-interval
For convenience, one could add PX - Wmin to both LHS and RHS sub-intervals

This formula generalizes two formulas given in a note (Zayatz, 2000) for calculating protection for rounded data. The formulas there are for rounding bases B=1000 and 1,000,000.

4. Comparing Mathematical and Statistical Approaches for Protection Rules

4.1 Mathematical and Statistical Protection Against Attacks: a General Discussion

Suppose a table user has knowledge that corresponds exactly to the assumptions used in one of our models. In the earlier sections we showed that there are two types of methods for estimating the largest contributor to a cell value, x1; methods that involve using only the width of uncertainty intervals and methods that involve statistical ideas. The latter require use of a density function for the interval. The midpoint attack would typically be attempted if the user felt the density on the uncertainty interval were roughly symmetric (e.g., a uniform density) about x1. We showed that the method of rounding provides some protection against the midpoint attack by creating asymmetry into the uncertainty interval for the cell value, which in turn creates asymmetry in the uncertainty interval for x1. Now we generalize these notions to form two major classes of attacks and associated protection methods.

Types of Protection Methods for Protecting the Largest Contributor of a Cell Value

Type 1: Purely Mathematical Protection Methods

Purely mathematical protection methods are designed to protect against purely mathematical (i.e., non-statistical) attacks. The simplest example of such an attack would be simple algebraic calculations performed by a table user to recover a value to a precision greater than the SO releasing the table intended to make possible. One way of implementing a mathematical protection method is to express the quantity, x, for which protection is sought (e.g., x=x1, the largest contributor to a cell) in terms of other quantities (the 'inputs') which can be read from the table or for which approximate external knowledge is assumed available to all table users. One can often easily calculate the uncertainty interval for x in terms of the uncertainty intervals of the inputs. One then decides which measure of sensitivity (e.g., the standard p% rule) to use and which location rules (e.g., sliding or two-sided protection) to use. Using the uncertainty interval for x and the selected rules, one determines if there is enough inherent uncertainty in x, or whether protection methods must be applied to increase the amount of uncertainty. Purely mathematical methods are designed to either increase the full width of the uncertainty interval or the widths of the right and left sub-intervals. Certain methods, such as conventional rounding and even certain types of perturbation may be viewed as mathematical if no use is made of the effect of these methods on the density of the resulting uncertainty intervals.

Type 2: Statistical Protection Methods

Statistical protection methods are designed to provide protection against attacks that use statistical estimation as well as purely mathematical analysis. The "midpoint attack" is probably the best known attack of this type. This uses an uncertainty interval for the quantity of interest, x, and

implicitly assumes a symmetric density on this interval. The table user simply estimates that the true value of x will equal the midpoint of the uncertainty interval. Let us express this in standard statistical terminology. Since we have introduced a density function, it is appropriate to view x as a value of a random variable X. If the uncertainty interval for X is symmetric (about the true value x_t) and one assumes a symmetric density for X, the mean of X will equal x_t. It is often difficult for the SO to decide if protection against such an attack is needed since it is based on assumptions that often do not hold. Thus the midpoint estimate could be described as an educated guess. Then the question becomes: does the SO need to protect data against educated guesses that are close to the truth some of the time ? One reasonable solution for the SO would be to ensure there is enough asymmetry of the uncertainty interval for X, or its density, so the midpoint attack will not often succeed in producing good estimates of x. To make this idea precise, we need to introduce a quantitative "closeness" rule; i.e., a measure of when the estimate x_{est} is so close to the true value, x_t , that confidentiality is not preserved. One could use a rule that is similar in structure to the p% rule. Let $| x_{est} - x_t |$ represent the error of the estimate produced by some attack method. If the error is small only occasionally, no protection against the attack will be needed. However, protection against a statistical attack is needed if

$$\text{Prob} \{ | x_{est} - x_t | \leq (q/100)*x_t \} \geq \text{delta}$$

where (q/100)*T represents the maximum error that the SO finds problematic and delta is the associated minimum problematic probability. Actually, delta and q are not independent; in general delta would be an increasing function of q. This is because if a somewhat small error (say q=20) poses a problem if it occurs, say, 80% of the time, then it is likely that a very small error (say q=10) would pose a problem, say, if it occurred only 50% of the time. Ideally, the SO should determine what combination of values for q and delta it finds problematic; this is actually a policy decision. This will form some region in the q-delta plane, which might be called the "problematic zone". The SO uses this zone for the next step, which is technical. The SO must determine how to add protection to X to create an uncertainty interval and density for X that protects X from the attack; i.e., with the additional protection the attack estimator falls outside the "problematic zone." Ideally, the SO would have some sense of the most likely attacks and could then do this (possibly complicated) analysis for this set of attacks.

When uncertainty intervals are needed for medium or large tables that have been protected using cell suppression, they are usually computed using audit programs. Experience with such output from audit programs shows that the uncertainty intervals for the cell value are often larger than they would be based on the single cell protection requirements. Also, they are often asymmetric about the true value of x. We might call this type of protection "global process statistical protection." *Global* is used because the audit interval for a given cell depends on the entire suppression pattern. This complexity makes it likely that a table user will fairly often experience significant errors if he estimates a cell value to be midpoint of the audit interval for that cell.

4.2 Measuring the Protection Against the Midpoint Attack Supplied by Rounding

Suppose conventional rounding is used in a cell to create an asymmetric uncertainty interval. Let T be the cell value. Let 'rem' be the sum of the third largest and smaller contributions. Then

the true value, x1 = T - x2 - rem. The best estimate that an attacker could make is $x1_{est}$ = r(T) - x2 -rem. Thus the **error (in the estimate)** = $x1_{est}$ - x1 = r(T) - T. Assume, that given r(T), T, viewed as a random variable, has a uniform density on the uncertainty interval [r(T)-b, r(T)+b) where b =B/2 where B is the rounding base. Then E{ error | r(T) } = 0.

Clearly, the absolute value of the error, |error|, is a good measure of protection afforded by rounding. A straightforward integration shows that E{ |error| | r(T) } = b/2 = B/4. Thus we see that such protection is 1/4 of the rounding base B.

4.3 The p% Rule Used for a Table of Rounded Percentages

Here we use a mathematical approach, namely interval protection as was done for protecting rounded totals. Suppose an exact (i.e., unrounded) grand total G is published for some magnitude variable, and one designs a one-column table that contains the percentages of G to the nearest whole integer for each row category. That is, if Ri represents the total for row I, then R1+R2+R3+...= G and we wish to represent each row total as a rounded form of (Ri/G)*100. How does one apply the p% rule to determine which, if any, of the cells in the column need to be suppressed ? How much protection will the suppressed cells need ? The solution below involves extending the interval protection with the p% rule for the rounding of totals in section 3.4.

For a given row cell, express the contributions in the form: x1 + x2 + rem = exact*G where exper=exact*100 is the exact percentage that the row total represents of the grand total G. Suppose we are using conventional rounding to the nearest whole percent. Let rndper denote the rounded form of exper. Then the exper lies in the interval [rndper - 0.5, rndper + 0.5). Then we can convert the p% rule for rounded totals (as opposed to percentages) to rounded percentages as follows. For this situation, the p% rule with sliding protection is expressed as :

the row cell is sensitive if 2*(rem/G)*100 + 2 *(0.5) < 2*p*(x1/G).

If row total is sensitive, then we need to add (2*p*(x1/G) - (2*(rem/G)*100 + 2 *(0.5))) units of protection. If we used cell suppression, we would suppress the given row cell and then suppress additional cells that together have the required protection units (i.e., percentage points).
With two-sided protection,
Let WR = exper - rndper + 0.5 + (rem/G)*100 and WL = rndper - exper + 0.5 + (rem/G)*100
Let Wmin = min (WR, WL) = (rem/G)*100 + 0.5 - abs{rndper- exper} where abs = absolute value.
Wmax = max (WR, WL) = (rem/G)*100 + 0.5 + abs{rndper-exper}
Let PXper = (p/100)*(x1/G)*100 = p*(x1/G)
if Wmin < PXper, add PXper - Wmin protection units to smaller sub-interval
if Wmax < PXper, add PXper - Wmax protection units to larger sub-interval
For convenience, one could add PX - Wmin units to both LHS and RHS sub-intervals
The protection units are percentage points. If cell suppression is used, we would suppress the sensitive row cell and then suppress additional cells that provide the required protection (Massell, 2005b).

4.4 Various Measures of Intruder Error When Using a Midpoint Attack

Let X be the random variable with associated density denoted by den(UI(x1,C2); StAs) as explained above. In the midpoint attack, the intruder uses E[X] to estimate x1. Thus the user's error is the distance ierr1 = E[X] - x1. If the density of X is symmetric about x1, then the mean(X) = E[X] =x1, thus ierr1=0. Another reasonable measure of error is the 'distance error' ierr2 = E[abs(X-x1)] where 'abs' denotes the absolute value. Assume again that X is symmetric about x1. How does ierr2 compare with ierr1 ?

$$E[abs(X-x1)] = Pr(X < x1) * E[-(X-x1) \mid X < x1] + Pr(X > x1) * E[(X-x1) \mid X > x1] \qquad \text{(Eq.4.1)}$$

The symmetry of X about x1 implies $Pr(X < x1) = Pr(X > x1) = 0.5$. Thus

$$E[abs(X-x1)] = 0.5 * \left(E[-(X-x1) \mid X < x1] + E[(X-x1) \mid X > x1] \right) \qquad \text{(Eq.4.2)}$$

Suppose $E[X \mid X > x1] = x1 + delta$. Then, by symmetry $E[X \mid X < x1] = x1 - delta$.
Note that the more concentrated X is about x1, the smaller delta is. We have

$E[(X-x1) \mid X > x1] = x1 + delta - x1 = delta$ Thus

$E[-(X-x1) \mid X < x1] = -x1 + delta + x1 = delta$.

Using Eq.4.1, since $Pr(X<x1) = Pr(X > x1) = 0.5$, we get $E[abs(X-x1)]= delta$

4.5 Additional Topics

Some of the above analysis could be extended to the case of grouped data which is ranked. An example of such data is a one-row additive table in which the cell values represent totals of ranked values. For example, the left most cell value, denoted T1, could represent the sum of the values of the top 4 companies with respect to some magnitude variable. Similarly the next cell total T2 might represent the sum of the next highest 4 companies. When cells are suppressed, a data intruder would try to use some of the ranking information to compute uncertainty intervals for company values of interest.

At the beginning of the paper, we discussed various sources of survey error that can affect the published cell value, e.g., the effect of sampling weights on the use of the p% rule. However, we did not fully discuss the effect of sampling on UI's. We discussed the effect of an unknown weight or scaling factor on a user's uncertainty of a particular contribution to a cell value but we did not discuss the other important way in which sampling introduces uncertainty, namely making it harder to identify the contributor of the value of interest.

In this paper, we did not discuss the effect of non-sampling errors (e.g., editing errors) on UI's. Statistical models for such errors probably would be complex whenever the editing rules themselves are complex.

5. Conclusions

In this paper, we've shown how uncertainty intervals provide a concrete mathematical tool for computing the protection needs of data elements (e.g., cell values, underlying microdata) related to tables being considered for release by a statistical agency. Such intervals can first be used to measure the existing uncertainty that derives from survey errors, especially sampling errors. They also are useful for describing cell value "distributional uncertainty" which is the type of uncertainty that the p% rule, or similar sensitivity measures, are designed to measure. Some extensions of these sensitivity measures can measure the combined effects of survey processing parameters that are unknown to data users and distributional uncertainty. Such sensitivity measures let the SO know if the existing uncertainty is sufficient to protect all data elements or if not, exactly how much additional uncertainty must be provided by one or more of the applicable disclosure avoidance methods. In addition, uncertainty intervals can be used to analyze the protection provided by rounding and other data processing procedures against certain common statistical attacks, such as the midpoint attack. We included examples of uncertainty interval construction for various protection rules and survey data processing procedures.

Acknowledgments

Larry Cox's seminal work on formalizing the notion of sensitivity measures motivated me to attempt to extend this formalization to closely related topics such as uncertainty intervals using both mathematical and statistical ideas. Over the years, I have discussed many of the ideas in this paper with Laura Zayatz, Brian Greenberg, Bob Jewett, Jim Fagan, and Gordon Sande. In particular, my discussions with Laura about protection derived from weights and from rounding and correspondence with Gordon about how rounding can be used to protect against a midpoint attack have illuminated these aspects of disclosure protection.

References

FCSM [WP22] (2005) Statistical Policy Working Paper 22 : Report on Statistical Disclosure Limitation Methodology (see especially Chapter IV; Methods for Tabular Data; Appendix A: Technical Notes: Extending Primary Suppression Rules to Other Common Situations)

Greenberg, Brian, (1998a) Introducing Cell Variances into Disclosure Avoidance Strategy, unpublished report, April 2, 1998.

Greenberg, Brian, (1998b) Using Induced Noise in the Disclosure Avoidance Strategy for Economic Tabular Data, unpublished report, May 20, 1998.

Massell, Paul B., (2005a) Protecting Sensitive Cells in a Cell Suppression Program Using Sliding Protection, SRD research report SSS2005/02.

Massell, Paul B., (2005b) The p% rule for Tables with Rounded Percentages, informal note, March, 2005.

Sande, Gordon (2003). A Less Intrusive Variant on Cell Suppression to Protect the Confidentiality of Business Statistics, FCSM conference paper. http://www.fcsm.gov/03papers/Sande.pdf http://www.fcsm.gov/03papers/Sande.pdf

Willenborg, Leon, and de Waal, Ton, (1996) Statistical Disclosure in Practice, Lecture Notes in Statistics, v. 111, Springer, 1996.

Zayatz Laura. (2000) Designing Primary Suppressions Calculating the Protection Needed for Rounded Data: Informal note: July, 25, 2000

Bibliography on Uncertainty and Uncertainty Models

http://www.cs.utep.edu/vladik/2004/tr04-20b.pdf

http://physics.nist.gov/Pubs/guidelines/sec6.html

http://www.elsevier.com/wps/find/bookdescription.librarians/500588/description#description

http://en.wikipedia.org/wiki/Uncertainty#Relation_between_uncertainty.2C_probability_and_risk

http://ams.confex.com/ams/Annual2006/techprogram/paper_99724.htm