Practical Considerations in Applying the *pq*-Rule for Primary Disclosure Suppressions

Meghan O'Malley, Lawrence R. Ernst
O'Malley.Meghan@bls.gov, Ernst.Lawrence@bls.gov
Bureau of Labor Statistics, 2 Massachusetts Ave., N.E., Room 1950, Washington, DC 20212-0001

Abstract

As statistical agencies try to "move as far as possible toward the use of a small number of standardized disclosure limitation methods whose effectiveness has been demonstrated" (Working Paper 22), many practical difficulties arise in applying these standard sensitivity measures to particular data sets. This paper discusses possible ways of handling several common complications which arise when applying the *p*-percent rule and, more generally, the *pq*-rule. This includes: analysis of increases and decreases in disclosure risk due to imputation, suggestions for use of weights as protection, and discussion of handling final weights less than one which generally arise from controlling sampled statistics to independent universe values. Ideas in this paper come mainly from considering how the *pq*-Rule and similar methods could be applied to the Occupational Employment Statistics Program conducted by the U.S. Bureau of Labor Statistics in partnership with the 50 States and the District of Columbia.

Keywords: Confidentiality, *p*-Percent Rule, *pq*-Rule, Sample Weights, Imputation, Primary Suppressions

1. Introduction and Background

1.1 Introduction

This paper was motivated by a study done using data from the Occupational Employment Statistics (OES) Program at the Bureau of Labor Statistics. OES is a survey of approximately 1.2 million business establishments for the purpose of gaining information on employment and wages by occupation. The program is a combined effort of individual states, territories, and the Bureau of Labor Statistics. The OES Program was tremendously kind during the course of this project, helping us to understand the system and the data, identify key issues, and evaluate available and developed methods.

1.2 Disclaimer

Keeping with the topic of this paper, it should be noted that, the methods discussed here are meant to assist persons maintaining complex surveys in providing reliable disclosure protection to their respondents. They are not a description of the procedures used by the OES Program. Furthermore, to uphold the confidentiality pledges given to BLS respondents, detailed examples included in this paper have been fabricated to illustrate the statistical issues without disclosing respondent information.

1.3 Background: The *pq*-Rule

The basic form of the *pq*-Rule assumes that there is a census of N units, with characteristic values x_i, and that these units sum to T, which is the cell value that would be published, directly or implicitly.

$$x_1 \geq x_2 \geq \cdots \geq x_N$$
$$x_1 + x_2 + \cdots + x_N = T$$

The *pq*-Rule indicates that the cell value should be suppressed if the following holds:

$$\frac{p}{q} x_1 > \sum_{i=3}^{N} x_i \qquad (1.3.1)$$

Equivalently:

$$\frac{p}{q} x_1 > T - x_1 - x_2 \qquad (1.3.2)$$

where p and q are set parameters with $p<q$.

Large scale establishment surveys such as OES face challenging tasks of identifying and concealing sensitive information across an enormous number of cells whose contributions are often extremely uneven in size. The *pq*-rule is particularly well suited to accommodate both these issues.

We think it is helpful to consider this rule in two ways: mathematically and intuitively. Mathematically, the *pq*-rule can be thought of as a sensitivity measure which identifies cells dominated in a particular manner by one or two units. Intuitively, the *pq*-rule can be thought of as checking the worst case scenario for possible disclosure. The value of the largest unit, x_1, is the value most at risk of unintended disclosure, and the second largest unit, with value x_2, has the most information that can be used to estimate x_1. If the owner of x_2 can estimate x_1 within some set limit, p%, then the cell is considered at risk. If x_1 is safe from the owner of x_2's attempt, then every unit in the cell is safe

from a similar attempt by any other unit. The intuitive viewpoint provides practical meaning to the rule and inspires sensible extensions of it. The mathematical viewpoint provides a sturdier platform where we can work with situations where the assumptions needed for the steps of the intuitive meaning are known to be faulty and where the actual bound around reported values needed to protect them (ideally, parameter p) and the actual bound within which information on the rest of the cell is known (ideally, parameter q) may not be constant or measurable.

A more thorough description of the intuitive explanation is given below:

The second largest respondent could create an equation with the value of the largest establishment, x_1, their own value, x_2, the rest of the cell, R, and the cell total, T. From the perspective of the second largest respondent, x_2 and T are known, and x_1 and R are unknown.

$$x_1 + x_2 + R = T$$

or, equivalently,

$$x_1 + R = T - x_2$$

The second largest unit is assumed to know R within q% of its true value. In a census, this could be regarded in two ways: as knowledge of each unit's presence and value within q%, or as knowledge of the remainder of the industry as a single quantity, within q%. Either viewpoint results in the following bound.

$$\frac{100-q}{100}\sum_{i=3}^{N} x_i \le \hat{R} \le \frac{100+q}{100}\sum_{i=3}^{N} x_i$$

This information can be used to make an upper and lower bound for x_1:

$$\hat{x}_1 \le T - x_2 - \frac{100-q}{100}\sum_{i=3}^{N} x_i$$

$$\hat{x}_1 \ge T - x_2 - \frac{100+q}{100}\sum_{i=3}^{N} x_i$$

Simplified:

$$x_1 - \frac{q}{100}\sum_{i=3}^{N} x_i \le \hat{x}_1 \le x_1 + \frac{q}{100}\sum_{i=3}^{N} x_i$$

That is, the second largest respondent can estimate the value of the largest respondent within:

$$\frac{q\sum_{i=3}^{N} x_i}{x_1}\%$$

So the cell should be flagged for suppression if:

$$\frac{q\sum_{i=3}^{N} x_i}{x_1} < p\,,$$

or, equivalently, if

$$\frac{p}{q}x_1 > \sum_{i=3}^{N} x_i\,.$$

The above description is helpful as a starting point. Before the rule can be applied to survey data, an extension must be made for sampling weights and further extensions are likely to be needed for coalitions, other weights, non-response procedures, and special bypass procedures.

2. Collapsing Coalitions

In many situations, it would be naïve to assume that units act alone. Several units may share values to attempt to estimate the values of another unit. An alternate form of the pq-Rule is available from Working Paper 22 to account for colluding units.

$$\frac{p}{q}x_1 > \sum_{i=c+2}^{N} x_i$$

where c is the size of a coalition.

In practice, however, the above rule is only feasible when units tend to be roughly similar in size. Because establishment data varies so much in size, using a formula which assumes any unit may collude with any other unit, regardless of company, industry, location, or other status can result in an inordinately large number of suppressions. In establishment surveys, it is common to aggregate establishments in a cell belonging to a particular company before testing for confidentiality. This is done to reflect the prior knowledge that a data user from a company is likely to have access to the values of all establishments reported for that company. Collapsing establishments within each company accounts for known coalitions in a reasonable manner without creating an impossibly large number of suppressions based on implausible situations. After collapsing same-company units, an assumption that units are non-colluding, although imperfect, is more reasonable.

3. Weights

3.1 Added Protection from Weighting

Moving from a census to a survey requires a fundamental change in the way the *pq*-rule is interpreted. When a sample is drawn, the units which sum to T are:

$$w_1x_1 + w_2x_2 + \cdots + w_nx_n = T\ .$$

In many situations, several of which are described in more detail below, weighting provides additional protection to respondents. This additional protection can be quantified in part and incorporated as protection by using a modified version of the *pq*-rule. Working Paper 22 suggests the following modified version:

$$\frac{p}{q}x_1 > T - x_1 - x_2 \qquad (3.1.1)$$

This formula differs from (1.3.2) in that here, T is the weighted total and x_1, x_2 are un-weighted values.

The intuitive explanation proceeds as follows:

With little to no knowledge about weights or presence in the sample, particularly for smaller units, the second largest unit might create the following equation:

$$x_1 + x_2 + R = T \qquad (3.1.2)$$

where x_1 and x_2 are the un-weighted values of the largest and second largest units, T is the weighted total for the cell, and R is the value of the rest of the cell. If the cell were published, the second largest unit would know x_2 and T.

Note that the publishing agency's value of R may be different from the data user's concept of R. The agency estimates the value of the remainder of the cell excluding x_1 and x_2 using:

$$R = T - x_1 - x_2 \qquad (3.1.3)$$

For simplicity, we then assume that the data user knows this estimated R within q% of its value.

$$\frac{100-q}{100}\left(T - x_1 - x_2\right) \le \hat{R} \le \frac{100+q}{100}\left(T - x_1 - x_2\right) \qquad (3.1.4)$$

This information can be used to make an upper and lower bound for x_1.

$$x_1 - \frac{q}{100}\left(T - x_1 - x_2\right) \le \hat{x}_1 \le x_1 + \frac{q}{100}\left(T - x_1 - x_2\right) \qquad (3.1.5)$$

That is, the second largest respondent can estimate the value of the largest respondent within:

$$\frac{q\left(T - x_1 - x_2\right)}{x_1}\%$$

So the cell should be flagged for suppression if:

$$\frac{q\left(T - x_1 - x_2\right)}{x_1} < p\ ,$$

or, equivalently, if

$$\frac{p}{q}x_1 > T - x_1 - x_2 \qquad (3.1.6)$$

Practically, this formula is very intuitive in both form and performance and, assuming weights are greater than one, has several desirable properties. Low chances of suppression are desirable for cells where there is uncertainty regarding which units are in the sample. Under this form of the *pq*-Rule, suppressions are decreased when x_1 and x_2 have weights greater than one, and impossible when x_1 has a weight greater than two. Subadditivity is an additional desirable property that can be easily verified for this form of the *pq*-Rule, even when same-company units are collapsed. Subadditivity refers to the property of a sensitivity measure, such as the *pq*-rule, that the union of any two non-sensitive cells remains non-sensitive. Cox (1980) describes subadditivity as a minimum requirement for a reasonable sensitivity measure.

However, there is questionable logic in the steps of the intuitive explanation above in a several places. Setting up (3.1.2) is only reasonable is the owner of x_2 can identify, from general knowledge, the largest unit, and be reasonably sure that this unit is in the sample. Moving from 3.1.2 to 3.1.3 we are estimating a quantity, R, which, when x_1 and x_2 are not certainties, depends on the sample. Moving from 3.1.3 to 3.1.4, we are assuming either that R is the true parameter value, or that the data user knows the agency's estimate of R within q%, rather than the true value of R within q%. Furthermore, q, for most applications, is neither consistent nor measurable.

Fortunately, we can lean back on the mathematical understanding of the *pq*-Rule as providing a sensitivity measure which identifies cells dominated in some way by one or two units. This is certainly done by the form of the *pq*-Rule above. The intuitive explanation, with its flaws, still serves its purpose of providing useful, practical meaning to the rule.

3.2 Many Weighting Factors

In practice, surveys often include a whole array of weights which may include: base sample weights, over or under-reporter magnitude adjustments, non-response adjustments, panel combination weights, population controls, and more. The ways in which many of these can and cannot be used as added protection are described in more detail below. Because of possible additional sensitivity issues for and from non-responding units, non-response is discussed in section 4 as imputation rather than here as a weighting procedure. For the other weights, it may be helpful to

consider the weights in two groups, those that provide protection, w, and those that do not, v. The partially weighted values, $v_i x_i$, would be treated as the un-weighted values.

$$w_1(v_1 x_1) + w_2(v_2 x_2) + \cdots + w_n(v_n x_n) = T$$

$$\frac{p}{q}(v_1 x_1) > T - (v_1 x_1) - (v_2 x_2)$$

3.2.1 Base Sampling Weights

As long as public knowledge of sampling methods, size, and frame are limited such that users cannot closely estimate sampling weights, sampling weights can be viewed as added protection to reported values. The value of the sampling weight is typically the inverse of the probability of selecting a particular unit. Sampling weights of one indicate that the unit would be in any sample, that is, no protection is added from sampling. Larger sampling weights indicate smaller chances of selecting particular units in the sample, hence, larger amounts of protection from sampling.

3.2.2 Guidelines for Other Weights

Panel combination weights, population control weights, and other weights may or may not be able to be used for protection in this manner. The following two guidelines can be used to determine whether or not a specific set of weights adds protection. In order to add protection, the weights cannot be known or easily predicted by data users, and they cannot be intended to adjust reported values to better reflect the units represented in the cell.

3.2.3 Magnitude Adjustments

Weights from adjustments designed to match the magnitude of the respondent with the unit represented in the cell, such as adjustments for over-reporters and some types of population controls, may not be knowable or predictable by data users but the amount of protection added is not necessarily related to the magnitude of the weight. Two examples are shown below:

Example 1) Suppose you are collecting data on number of cooks at fast food restaurants in a particular state, and an extremely large chain of restaurants gave employment counts for the whole country, rather than that particular state:

Estimated Total Fast Food Cooks in the State: 75,000
Over-Reported Cooks for One Company: x_1=50,000
Over-Reporter Adjustment: w_1=0.3

$$w_1 x_1 = 0.3 \times 50{,}000 = 15{,}000$$

Example 2) Suppose you were collecting data on number of cooks at fast food restaurants in a particular state, and an extremely large chain of restaurants gave employment counts for a particular establishment rather than all establishments in the state.

Estimated Total Fast Food Cooks in the State: 75,000
Under-Reported Cooks for One Company: x_k=10
Under-Reporter Adjustment: w_k =1,500

$$w_k x_k = 1{,}500 \times 10 = 15{,}000$$

In both examples above, the company's un-weighted values do not represent the number of cooks employed by the company in the state. Data users looking for that company's value know something about the range of the value. The un-adjusted value for this company appears either more sensitive than it is, as in example one, or less sensitive than it is, as in example two. The magnitude adjustment is not providing protection as sampling weights do. It is more sensible to apply the magnitude adjustment and work with the resulting value. In both examples, 15,000 is an estimate of the company's contribution to the cell. If any value for that company should be tested for confidentiality, it is this value. The following decision would then be to determine if disclosing this value would be improper or if this value should bypass confidentiality testing altogether.

3.2.4 Population Controls

Population controls, depending on their purpose, may or may not provide protection quantified by the weights. Population controls used primarily to adjust for sampling and non-response errors could be used as extra protection; these errors are not predictable and population controls of this sort are intended to make adjustments to overall totals, not to adjust the values of the units in the cell to better reflect the units represented in the cell. But population controls used primarily to account for changes over time or other overall value shifts cannot be used in the same manner as extra protection. Because changes over time are predictable, these weights may be predictable. Additionally, they are intended to adjust the reported values in the cell to better reflect the units represented in the cell. Mostly predictable weights and value adjustments add some sort of protection, but in these cases, the value of the weight is not a good indicator of the amount of protection added.

3.3 Protecting Weights Less Than One

Protecting weights which are less than one can arise due to population controls. Depending on the extent to

which values are controlled and on incoming weights, weights less than one can be sparse or frequent, close to one or close to zero. Larger units (x_1 or x_2) are often more likely to have weights less than one since larger units often have sampling weights equal to or close to one. When x_1 or x_2 have weights less than one, the *pq* rule involves subtracting corresponding un-weighted values from the cell total.

$$\frac{p}{q}x_1 > T - x_1 - x_2$$

In this situation, the un-weighted values are larger than their contribution to the cell total. Depending on the magnitude of the weights and un-weighted values, the quantity $R=T-x_1-x_2$, could be extremely small in magnitude or even negative. When R is small, suppressions are very likely, and when R is negative, the cell will always be suppressed.

When magnitude is large and negative, under the intuitive meaning of the *pq*-rule, the data user cannot accurately estimate the value of x_1. Labelling these cells as sensitive is somewhat counter to our usual concept of sensitivity.

Two possible methods came to mind to avoid labelling these cells as sensitive, but both detracted from the main strengths of the *pq*-rule. First, consider subtracting the un-weighted value when weights are greater than one, and the weighted value otherwise. Before applying the pq-rule, we would need to reorder the values:

$$y_i = \begin{cases} w_i x_i \text{ for } w_i < 1 \\ x_i \text{ for } w_i \geq 1 \end{cases}$$

$$y_1 \geq y_2 \geq \cdots \geq y_n$$

$$\frac{p}{q}y_1 > T - y_1 - y_2$$

Although sensitive cells due to weights less than one are avoided and most other sensitive cells remain the same, much of the intuitive meaning of the rule is lost. The rule could no longer be viewed as a reasonable procedure set up by a respondent, using information known by the respondent, to estimate the values of another unit. Although removing these suppressions seems desirable, the cost of losing the intuitive meaning may outweigh the benefit.

Next consider taking the absolute value of R in the sensitivity measure. This is a more conservative change. The intuitive meaning of the *pq*-rule is left intact but subadditivity problems arise. The union of two non-sensitive cells, one with a negative value of R and one with a positive value, may be sensitive. An example is shown below.

Example) let $p = 40, q = 80$

Cell A :

$w_1 = 0.3, x_1 = 100, w_2 = 0.5, x_2 = 80, \sum_{i=3}^{n} w_i x_i = 20$

calculate : $T = 90, R = -90$

nonsensitive : $\frac{40}{80} \times 100 < |90 - 100 - 80|$

Cell B :

$w_1 = 1, x_1 = 20, w_2 = 1, x_2 = 15, \sum_{i=3}^{n} w_i x_i = 15$

calculate : $T = 50, R = 15$

nonsensitive : $\frac{40}{80} \times 20 < |50 - 20 - 15|$

Cell A $\cup$ B :

$w_1 = 0.3, x_1 = 100, w_2 = 0.5, x_2 = 80, \sum_{i=3}^{n} w_i x_i = 70$

calculate : $T = 140, R = 0$

sensitive : $\frac{40}{80} \times 100 > |140 - 100 - 80|$

The disclosure concern for these cells is not the units in the cell, but the units in possible unions of cells. If it can be verified that no reasonable unions of cells are sensitive because the cell in question is included, then labelling the cell as non-sensitive may be a viable option. Otherwise, although counter intuitive on the surface, these cells should be labelled as sensitive.

A more fundamental issue regarding whether or not these cells are considered sensitive may arise. On one hand, under the assumption that users do not know their weights, the fact that x_2 arrives so close to the true value of x_1 is coincidental. Subtracting an un-weighted value from the total is never guaranteed to give a result close any particular response or quantity and generally does not give a result close to any particular response or quantity. On the other hand, if the *pq*-rule truly mimics a user's attempt to estimate a respondent, then this procedure correctly identifies cells where a user could and may accurately estimate a respondent's information. If the fact that there is an element of chance in the owner of x_2 estimating x_1 disqualifies the sensitivity measure when the weights of the largest units are less than one, it would also disqualify the sensitivity measure when the weights of the largest units are greater than one.

4. Imputation

4.1 Disclosure Risks Involving Imputation

Imputation or non-response adjustment does not necessarily eliminate all disclosure concerns. When considering methods for handling imputed values, we

think it is helpful to recall the possible types of disclosure relevant to the survey: *identity disclosure*, which occurs when a respondent is linked to a particular record, and *attribute disclosure*, which occurs when a data user obtains additional information about some unit's values. Imputation procedures should be carefully reviewed to determine which, if any, forms of disclosure may be possible for imputed units and for responding units because of other imputed units in the cell. Possible disclosure concerns from imputed values include disclosure of response statuses of certainty units, of confidential frame information of imputed units, of values for accurately imputed units, and of values of donors used for imputing other values in the cell.

4.1.1 Disclosure of Response Status

For many establishment surveys, although individual records are not released, identity disclosure cannot be entirely avoided. Data users who are very familiar with the units represented in the cell may know who the major contributors are as well as some of their recent behaviour. For sample designs where units are selected with probability proportional to size and or method with a similar component, for some cells, data users can know for sure that certain establishments are in the sample. For these establishments, the only part of identity disclosure that can be protected is whether or not that establishment responded.

Depending on agency policies and uses of the survey data, response status for these units may or may not be a disclosure concern. If it is determined that response status needs to be protected for these units, public information about publishing criteria should be examined carefully. For cells where one unit is heavily dominating, simply publishing or suppressing may be revealing information to data users about the dominating unit. For example, if publishing a cell is only possible if that unit responds (or does not respond), then simply publishing or suppressing discloses the unit's response status. Fortunately, disclosing response status can be prevented by carefully limiting public information about publishing criteria.

4.1.2 Sensitive Imputed Values

Imputed values are generally not considered a disclosure concern. However, imputed values are not necessarily non-sensitive. Sensitive imputed values may include those imputed from sensitive frame data, those imputed using very accurate procedures, and those which dominate particular cells.

Care should be taken when sensitive frames are used and imputation is performed based on values from the frame. Imputed values which are taken directly from sensitive frame information would remain sensitive. Similarly, imputed values which are close to the sensitive frame values from which they are imputed, would be sensitive.

Additionally, there are situations where data users arriving at close estimates for the values of a particular unit can be harmful for that unit regardless of how close the imputed value is to the actual value. For example, if it was perceived that a major manufacturer's percent mark-up for a popular product could be estimated closely from the agency's published values and from general outside knowledge, that estimate, whether or not it is close to the actual mark-up, could greatly impact public perception of the manufacturer and its future business. In these situations, whether a unit's values are reported or imputed makes no difference on its sensitivity.

The accuracy of the imputation procedures can also cause imputed values to be sensitive. If the imputation procedures are very accurate, then, even though the actual value was not reported, disclosing it may still be harmful to the unit, hence, of concern for the survey. On the other hand, if the imputation procedure is not particularly accurate, then attribute disclosure is not a concern for imputed units. If the cell is dominated by a unit whose values have been inaccurately imputed, it is a data quality concern rather than a confidentiality concern. Although this distinction does not affect the number of cells suppressed for primary disclosure, it may have a substantial impact on the number of cells suppressed for secondary disclosure. Cells unnecessarily labelled as sensitive for primary suppression can cause further unnecessary suppressions when secondary suppression methods are applied. For this reason, when possible, it is wise to separate flags for sensitivity, data quality, and other publishing criteria.

4.1.3 Disclosure through Donor Identification

A less common situation where imputation may cause disclosure concerns is when non-respondents may be able to identify their donor. If imputation procedures are straightforward and sufficient information about the imputation procedures is available to data users, for some cells, it may be possible for non-responders to identify their potential donor or donors. This situation is particularly risky when the values are imputed from donors in the same cell, such as in a nearest neighbour imputation method. Take, for example, a survey where data users are given information about straightforward

imputation procedures and a cell in which the values of the second largest unit are imputed from the largest unit or vice versa. The non-responding second unit then knows that the contributions from both the largest and second largest units come from the largest unit. In some cases, this may allow the non-responding unit to make a more accurate estimate of the largest unit. Fortunately, as with disclosure of response status, this can be prevented by carefully limiting public information about imputation methods.

4.3 Some Options for Handling Imputed Values

Treatment of imputed values can have a large impact on the number of cells suppressed even for surveys with high response rates. Three possibilities for handling the sensitive and non-sensitive imputed values described above are outlined below. Each involves re-ordering the units so that the values of certain units are or are not eligible to be x_1 and x_2. These methods can be coded directly, or flagged for bypassing confidentiality testing and dealt with along with other non-standard cases (section 5).

$$\frac{p}{q}x_1 > T - x_1 - x_2 \qquad (3.1.1)$$

The underlying principle behind each method is derived from the intuitive meaning of the pq-rule. The unit whose values are most at risk of disclosure is assigned to x_1 and the unit with the most information which can be used to estimate those values is assigned to be x_2.

4.3.1 Imputed Values are Sensitive and Accurate

For this method, no distinction is made between units which responded and units which were imputed. The largest value used in the estimates, whether a response or imputed, is assigned to x_1 and the corresponding second largest value is assigned to x_2.

This method may be appropriate for situations where disclosure of imputed values is a concern and where imputation procedures are considered to be very accurate. However, a large number of cells are labelled as sensitive and in most situations, we think this method is overly cautious.

4.3.2 Imputed Values are Sensitive but Not Exact

For this method, since imputed values are at risk of being disclosed, x_1 is assigned to the value of the largest unit regardless of response status. However, since imputed values are not exact, the unit with the most information is not necessarily the second largest unit. The value of the next largest responding unit is assigned to x_2. This can be viewed as the owner of that unit attempting to estimate x_1 or, if the second largest unit in the cell was imputed, as quantifying the protection added through the imputation procedure to the second largest unit's attempt to estimate x_1. The units are ordered as follows before the *pq*-rule is applied:

$$x_1 \geq \max\{x_2, \ldots, x_n\}$$
$$\text{responding: } x_2 \geq x_3 \geq \ldots \geq x_k$$
$$\text{imputed: } x_{k+1} \geq x_{k+2} \geq \ldots \geq x_n$$

By subtracting an equal or lesser value in place of x_2 in (3.1.1), the number of cells labelled as sensitive will be less than or equal to the number of cells labelled as sensitive under (4.3.1). However, when the values of x_2 and smaller responding values are fairly close, the impact of this change on the number of cells labelled as sensitive is slight. This method may be appropriate for surveys where disclosing imputed values is a concern or where response status is considered confidential.

4.3.3 Bypassing Imputed Values

When disclosing imputed values is not a concern, then it is reasonable to exclude imputed values from being assigned to x_1. Similarly, when units whose values have been imputed do not know the values assigned, it is also reasonable to assign the nearest responding value to x_2. This can be viewed as quantifying the protection added to the imputed unit. The units are ordered as follows before the *pq*-rule is applied:

$$\text{responding: } x_1 \geq x_2 \geq \ldots \geq x_k$$
$$\text{imputed: } x_{k+1} \geq x_{k+2} \geq \ldots \geq x_n$$

Here, in addition to an equal or lesser value in place of x_2, we exchange x_1 for an equal or lesser value, making both the left hand side of (3.1.1) smaller and the right hand side of (3.1.1) larger. The number of cells labelled as sensitive will be less than or equal to the number of cells labelled as sensitive under (4.3.2). Furthermore, when the number of cells with dominating units imputed is substantial, the impact of this change on the number of cells labelled as sensitive will be substantial. In almost every case, cells where the largest unit has been imputed will not be labelled as sensitive.

5. Bypasses

Frequently in practice, disclosure may be acceptable for some groups of sampled units. For example, certain information about government establishments is accessible to the public, and therefore, does not need to

be protected. Additionally, for non-sensitive questions, special permission may be sought from respondents which dominate certain cells so that the corresponding estimates can be published.

Once permission has been obtained for the largest or several largest units to bypass confidentiality testing, publishing the cell without further testing is a strong temptation. Many algorithms handle this scenario by re-ordering units so that the largest unit is defined to be the largest non-bypassed unit. This results in impossible or extremely infrequent suppressions for these cells. In many cases this procedure may be appropriate, however, it should be noted that with this approach, there seems to be an unspoken assumption that the units which bypassed confidentiality do not openly share their values.

Units may allow an agency to bypass confidentiality testing because the values collected are already publicly available, as with the government example above. In this situation, the bypass is actually indicating that a value is known or knowable to any data user.

Publicly known values do not add protection to other units. In fact, publicly known values may put other, private value at greater risk of disclosure. For example, imagine data on the number of detectives working in a certain area. If all but one data-providing agency are government agencies with publicly available data then any data user could simply subtract those values out to obtain the number of detectives at the private agency. A sensible way to handle publicly known values is to treat them as colluding with any data user.

$$\text{private}: x_1 \geq x_2 \geq \ldots \geq x_k$$

$$\text{public}: x_{k+1} \geq x_{k+2} \geq \ldots \geq x_n$$

$$\frac{p}{q} x_1 > T - x_1 - \left(x_2 + \sum_{i=k+1}^{n} x_i \right)$$

Unfortunately, it may not be possible to distinguish between bypassed units whose values are publicly known and bypassed units whose values are not publicly known. In these cases a judgement call must be made about which, if any, should be considered public knowledge.

References

Cox, Lawrence H.. *Journal of the American Statistical Association.* Vol. 75, No. 370. 1980. pp. 377-385. “Suppression Methodology and Statistical Disclosure Control.”

Lambert, Diane. *Journal of Official Statistics.* Vol. 9 No. 2, 1993. pp.313-331. “Measures of Disclosure Risk and Harm.”

Federal Committee on Statistical Methodology. Statistical Policy Working Paper 22. V2. 2005. “Report on Statistical Disclosure Limitation Methodology.”

NISS/NCES Seminar on Statistical Disclosure Limitation 12/7/2006