Statistics Netherlands
Division of Technology and Facilities
Methods and informatics department
P.O. Box 4000
2270 JM Voorburg
The Netherlands

Notes on sensitivity measures and protection levels

J.A. Loeve

Projectnumber: *TMO-102966*
BPAnumber: *01892-01-S-TMO*
Date: *10 September 2001*

Contents

Notes on sensitivity measures and protection levels

At Statistics Netherlands, a heuristic approach for cell suppression in non-negative hierarchical tables, HiTaS, is developed. HiTaS itself makes use of ARGUS, a cell suppression method for the protection of non-hierarchical tables. HiTaS protects a hierarchical table in a top-down approach by finding protection patterns for non-hierarchical sub-tables using ARGUS. ARGUS is based on a method by Fischetti and Salazar. To use this method, the cells which have to be protected are assigned the status primary unsafe. For these cells, non-negative protection levels can be chosen. These protection levels define a protection interval around the cell value. If a cell is primary unsafe, it will be contained in the final protection pattern, along with secondary suppressed cells such that the value of the first cell cannot be calculated to lie within the protection interval. For the determination of the primary unsafe cells, some kind of safety rules are used. Often a combination of a linear sensitivity measure and a number rule will be employed. It would be nice if the choice of the protection levels is related to the rules used to determine the safety of the cells. In this note, some safety rules and their relation with the choice of protection levels are discussed.

Keywords: HiTaS, ARGUS, hierarchical tables, cell suppression, (linear) sensitivity measures, protection levels.

1. Introduction

In the scope of TMO-research on protection of non-negative tables by cell suppression, a project to improve HiTaS has started. HiTaS is a heuristic approach used for cell suppression in hierarchical tables, where a hierarchical table is protected in a top-down approach by the protection of non-hierarchical sub-tables (see De Wolf, 1999). For the protection of the sub-tables, ARGUS is used. ARGUS is based on a method and corresponding program developed by Fischetti and Salazar (see Fischetti & Salazar-González, 1998). For this method, two protection levels have to be defined for cells which are primary unsafe: an upper and a lower protection level. In principle, also a sliding protection level can be defined, but this is not yet implemented. These levels define the extent to which the primary unsafe cell may be calculated from the final protected table. Furthermore, non-negative weights have to be assigned to all cells. Then, the method finds secondary suppressions in the sub-table such that the protection levels of the unsafe cells are satisfied. These secondary suppressions are chosen such that the total weight of the suppressed cells is minimized.

A part of the project concerns the use of protection levels for the unsafe cells. Furthermore, because of the hierarchical structure, secondary suppres-

sions found in some sub-table have to be protected in other tables where they occur. In these other tables, they play the role of primary unsafe cells for HiTaS and therefore, protection levels for these cells have to be defined too. In many cases, primary unsafe cells are found using a linear sensitivity measure. For example, a p-percent rule, which defines a cell as primary unsafe when a lower or upper bound of a contribution can be found within p percent of its true value. The protection levels should be such that this is no longer possible, otherwise the cell may remain unsafe after finding a protection pattern. On the other hand, cells may be defined primary unsafe when there are only a few respondents contributing to the cell value. This rule is often used in combination with a linear sensitivity measure. The idea is that these few respondents can estimate each others contributions fairly accurately if the cell value is known. Thus, the cell has to be suppressed. Also, in this case, protection levels have to be defined. Furthermore, it would be nice if protection levels can be chosen such that there is some relation between the choice of the levels and the method used to find primary unsafe cells.

In this note some sensitivity measures are considered. Furthermore, questions about the choice of protection levels for primary and secondary unsafe cells in relation to these sensitivity measures are discussed.

2. Linear sensitivity measures

Different criteria can be used to determine which cells are primary unsafe. Often, linear sensitivity measures are used. A linear sensitivity measure has the property of sub-additivity. This means, that the union of two safe cells will always be safe. Thus, aggregation of cells never leads to an increase in the degree of unsafety. The most common linear sensitivity measures are the prior posterior or p-q rule, the p-percent rule and the dominance rule. The prior posterior rule and the p-percent rule define a cell unsafe if a lower or an upper bound of one of its contributions can be calculated lying within p percent of the true value. The difference between the two rules is that the prior posterior rule assumes that possible attackers have some kind of prior knowledge about the values of the contributions. The p-percent rule has no assumption of prior knowledge, but for the knowledge of non-negativity of all contributions. According to the dominance or (n, k)-rule, a cell is primary unsafe when the total of the n largest contributions is greater than k percent of the cell value. This rule is based on the idea that a coalition of $n - 1$ respondents in a cell may not be able to estimate an n-th contribution with a certain accuracy. This idea might also be used with the p-q and the p-percent rule, leading to versions with coalitions.

Of course, other choices for linear sensitivity measures are possible, but only the ones mentioned above will be considered in this note.

3. Safety check

3.1. Prior posterior rule

The prior posterior or p-q rule is defined by two (fixed) parameters p and q with $p < q$. The q is used to model prior knowledge: all respondents in a cell are assumed to know the values of the other contributions to that cell within at most q percent. This also implies that each individual respondent knows which other respondents contribute to the total cell value. A cell is safe according to the p-q rule if, after publication of the cell value, no lower or upper bound of a contribution can be found within p percent of its true value.

In fact, the q can also be used to model prior knowledge by assuming that every respondent to an arbitrary cell knows the values of all other contributions to all other cells within at most q percent. This means that a respondent to a cell also knows something about the contributions to other cells. This is a stronger assumption, but this does not influence the formulation of the sensitivity measure.

Of course, an estimate of a contribution to a cell by a fellow respondent in the same cell will be better than an estimate by someone only knowing the total cell value. Furthermore, this situation is the most dangerous case of possible disclosure, because the immediate competitors are likely to be found among the fellow respondents. Thus, this worst case situation is used to derive a safety check.

Now, let the ordered contributions of a cell be denoted by $x_1 \geq x_2 \geq \cdots \geq x_N > 0$, where x_i is the contribution of respondent i. If respondent j knows the total cell value, he can calculate a lower and a upper bound of contribution x_i as

$$x_i \geq \sum_{t=1}^{N} x_t - x_j - (1 + \frac{q}{100}) \sum_{\substack{t=1 \\ t \neq i,j}}^{N} x_t,$$

and

$$x_i \leq \sum_{t=1}^{N} x_t - x_j - (1 - \frac{q}{100}) \sum_{\substack{t=1 \\ t \neq i,j}}^{N} x_t.$$

In fact, publication of the cell value offers each respondent j the possibility to calculate an interval containing contribution x_i.

The requirements that the lower and upper bounds are not within p percent of the true value, lead to the same inequality

$$p\ x_i - q \sum_{\substack{t=1 \\ t \neq i,j}}^{N} x_t \leq 0. \tag{1}$$

Because of the ordering of the contributions, the left-hand side is always lower than or equal to $p\ x_1 - q \sum_{t=3}^{N} x_t$. This means that if the lower and upper

bounds for x_1 by respondent 2 (worst case) are not within p percent of the true value, the bounds for other contributions will not be within p percent of their true values either. Thus, the cell is safe according to the p-q rule if

$$p\ x_1 - q \sum_{t=3}^{N} x_t \leq 0. \tag{2}$$

With the notation T for the total cell value, this is equivalent to

$$(q+p)\ x_1 + q\ x_2 - q\ T \leq 0.$$

By viewing the cell from the point of respondent 1, a 'rest' cell can be seen with total value $\sum_{t=2}^{N} x_t$. This rest cell may seem unsafe, because it need not satisfy $p\ x_2 - q \sum_{t=4}^{N} x_t \leq 0$. However, this test is valid for a real cell consisting of the contributions $x_2, \ldots, x_N$, but not for the rest cell because respondents 2 to N do not know the corresponding cell value $\sum_{t=2}^{N} x_t$. They only know the cell value of the whole cell including x_1.

As an illustration, consider the cell with contributions 40, 15, 4 and 2. This cell is safe according to the p-q rule with $p = 7$ and $q = 50$. Namely, $p\ x_1 - q \sum_{t=3}^{N} x_t = 7*40-50*(4+2) = -20 < 0$. In fact, the upper and lower bounds for the largest contribution calculated by the respondent of contribution 15 are equal to $61-15-0.5*(4+2) = 43$ and $61-15-1.5*(4+2) = 37$, respectively, and these are not within 7 percent of the true value of 40. Now, assume that the respondent with contribution 4 wants to estimate the contribution with value 15. His best upper bound is equal to $61 - 4 - 0.5 * (40 + 2) = 36$, which does not come close to the real value. However, a cell consisting of contributions 15, 4 and 2 is not safe, because $7 * 15 - 50 * 2 = 5 > 0$. This means that the respondent with contribution 4 can estimate the value of contribution 15 within p percent. For this estimate, the cell total $15 + 4 + 2 = 21$ would be used. This is not possible in the larger cell.

The prior posterior rule can be extended to situations with more complicated requirements on safety or with other forms of a priori information. For example, an extension of the p-q rule is the requirement that a cell is safe if for none of its contributions a lower bound within p_l percent or an upper bound within p_u percent of the true value can be found. Here, a 'lower p' and an 'upper p' are used instead of one fixed p. Furthermore, 'upper' and 'lower' q's can also be used to model a priori information (each contribution x is known to lie in the interval $[(1 - \frac{q_l}{100})\ x, (1 + \frac{q_u}{100})\ x]$) which is asymmetric around the values of the contributions. Of course, $q_l > p_l$ and $q_u > p_u$. This extension of the p-q rule leads to the following safety checks

$$p_l\ x_1 - q_u \sum_{t=3}^{N} x_t \leq 0 \quad \text{AND} \quad p_u\ x_1 - q_l \sum_{t=3}^{N} x_t \leq 0. \tag{3}$$

Even more complicated p-q rules can be created by using different p's and q's for different contributions, but still fixed per contribution. Also, for each different respondent, different p's and q's for all other contributions can be chosen. That way, q's which differ according to the distance between respondents in the hierarchy can be used. Still, it is fairly easy to derive the corresponding formulas for these rules, but instead of checking one or two requirements, many requirements per cell may be needed (for each bound of a contribution calculated by a respondent of another contribution). Therefore, these kinds of rules will almost certainly never be used in practice. Also, there would be a lot of work involved in choosing and storing all these different values.

3.2. p-percent rule

The p-percent rule is defined by one parameter, p. According to this rule, a cell is safe if no lower or upper bounds of the contributions can be found within p percent of the true values.

Again, it suffices to consider the best bounds of a contribution to a cell found by a fellow respondent in the same cell to derive a safety check.

With $x_1 \geq x_2 \geq \cdots \geq x_N > 0$ the ordered contributions of respondents $1, \ldots, N$ respectively, the best upper bound of contribution x_i found by respondent j is given by $\hat{x}_i = \sum_{t=1}^{N} x_t - x_j$. This $\hat{x}_i$ is not within p percent of the true value x_i if $\hat{x}_i - x_i \geq \frac{p}{100}\, x_i$, or equivalently, if

$$\frac{p}{100}\, x_i - \sum_{\substack{t=1 \\ t \neq i,j}}^{N} x_t \leq 0. \tag{4}$$

The best lower bound of contribution x_i by respondent j is equal to zero, because respondent j has no prior knowledge about the size of x_i. The restriction that this bound is not within p percent of the true value of x_i is always satisfied, so no extra inequality is found.

The worst case of inequality (4) is assumed when the left-hand side is largest. Because of the ordering of the contributions, this is the case for $i = 1$ and $j = 2$, where x_1 is estimated by respondent 2. This means that the cell is safe according to the p-percent rule if

$$\frac{p}{100}\, x_1 - \sum_{t=3}^{N} x_t \leq 0. \tag{5}$$

The rest cell seen by respondent 1 (of a safe cell) might seem unsafe, but just as is the case with the p-q rule, this is not true because the corresponding cell total is not known to its respondents. If the cell satisfies (5), the cell is safe from the viewpoints of all respondents.

At first sight, the p-percent rule is a kind of prior posterior rule, without the assumption of prior knowledge (so with $q = \infty$). However, it is a priori known that all contributions are non-negative. In fact, the p-percent rule is a special case of a p-q rule with an upper and a lower q parameter. The a priori information about every contribution is that it lies between zero and infinity. This corresponds to $q_l = 100$ and $q_u = \infty$. This means that the safety check for the p-percent rule can be derived immediately from (3) for the extended p-q rule.

Note that an extension with different p's is useless for the p-percent rule. The lower p is redundant because the best lower bound of a contribution x is equal to zero and for every p this is smaller than $(1 - \frac{p}{100})\ x$.

3.3. Dominance rule

A cell is primary unsafe according to the dominance or (n, k)-rule, if the largest n contributions together make up more than k percent of the total cell value. Let the ordered contributions of a cell be denoted by $x_1 \geq x_2 \geq \cdots \geq x_N > 0$ by respondents $1, 2, \ldots, N$, respectively. Then, the cell is safe according to the dominance rule, if

$$\sum_{t=1}^{n} x_t \leq \frac{k}{100} \sum_{t=1}^{N} x_t, \tag{6}$$

or equivalently,

$$(1 - \frac{k}{100}) \sum_{t=1}^{n} x_t - \frac{k}{100} \sum_{t=n+1}^{N} x_t \leq 0. \tag{7}$$

Viewing the cell from the point of respondent 1, a rest cell with total value $\sum_{t=2}^{N} x_t$ can be seen. This rest cell on its own may be unsafe, because it does not satisfy equation (7). However, as is the case with the p-q rule, the respondents do not know the cell value of the rest cell. Therefore, testing a rest cell for safety is a pointless exercise.

The dominance rule is a somewhat peculiar rule. It is based on the idea that a coalition of $n - 1$ respondents in a cell may not be able to make an accurate estimate of an n-th contribution. This seems to imply that the error in a bound for the biggest contribution made by respondents 2 to n has to exceed a certain threshold. This would seem equivalent to some p-percent rule. However, for $n \geq 2$ this is not the case. Observe therefore the following example with a safe and an unsafe cell, where the relative error in the upper bound for the biggest contribution is the same.

Consider a cell with contributions $x_1, \ldots, x_5$ with values 25, 19, 13, 8 and 2, respectively, and suppose that the $(3, 85)$-dominance rule is used. Now, $\sum_{t=1}^{3} x_t = 57 > 0.85 \sum_{t=1}^{5} x_t = 0.85 * 67 = 56.95$. Thus, the cell is primary

unsafe according to the $(3, 85)$-dominance rule. The best upper bound of x_1 which can be calculated by the contributors of x_2 and x_3 is equal to $\hat{x}_1 = 67 - (19 + 13) = 35$ with relative error

$$\frac{\hat{x}_1 - x_1}{x_1} = \frac{10}{25} = 0.4.$$

Now, consider another cell with contributions $y_1, \ldots, y_5$ equal to 25, 19, 12, 8 and 2, respectively. Then, $\sum_{t=1}^{3} y_t = 56 \leq 0.85 \sum_{t=1}^{5} y_t = 0.85 * 66 = 56.1$. Therefore, the cell is safe. However, the best upper bound of y_1 found by the contributors of y_2 and y_3 is equal to $\hat{y}_1 = 66 - (19 + 12) = 35$ with relative error

$$\frac{\hat{y}_1 - y_1}{y_1} = \frac{10}{25} = 0.4,$$

which is the same relative error as in the upper bound of x_1 in the unsafe cell.

Note, that for a specific cell, the absolute error in the best upper bound for each one of the n biggest contributions is the same. Namely, with $\hat{x}_i$ the best upper bound for contribution x_i (with $i \leq n$) by the other $n - 1$ largest contributors, the absolute error is equal to

$$\hat{x}_i - x_i = \sum_{t=1}^{N} x_t - \sum_{\substack{t=1 \\ t \neq i}}^{n} x_t - x_i = \sum_{t=n+1}^{N} x_t.$$

Just as is the case with the p-percent rule, the best lower bound of a contribution by a coalition of respondents is equal to zero, which does not give any extra information.

This kind of example can be found for all $n \geq 2$. However, if $n = 1$, the dominance rule is kind of similar to some p-percent rule. In fact, the $(1, k)$-dominance rule is equivalent to the requirement that the relative error in the upper bound for x_1 is greater or equal to $\frac{100}{k} - 1$, if the bound is calculated by someone outside the cell. Then, this bound will be equal to the cell value T and the relative error is equal to

$$\frac{\hat{x}_1 - x_1}{x_1} = \frac{T - x_1}{x_1} = \frac{T}{x_1} - 1.$$

Now, using (6), the cell is safe if and only if

$$\begin{aligned} x_1 &\leq \frac{k}{100} \sum_{t=1}^{N} x_t \iff \\ \frac{100}{k}\, x_1 &\leq T \iff \\ \frac{T}{x_1} - 1 &\geq \frac{100}{k} - 1. \end{aligned} \tag{8}$$

Starting from a dominance rule with $n = 1$, T is indeed the best possible upper bound for x_1 because no coalitions (not even a coalition of one) from respondents within the cell are allowed. In fact, this is the best upper bound

by an attacker who does not know anything about the internal structure of the cell. However, in case of a p-percent rule, the best upper bound for x_1 would be equal to $T - x_2$. Thus, the $(1, k)$-dominance rule is not really equivalent to a p-percent rule, but the relative error in an upper bound of x_1 can be used. Consequently, there is a kind of discontinuity between the dominance rule with $n = 1$ and with $n \geq 2$.

3.4. Prior posterior and p-percent rules with n-coalitions

The prior posterior and p-percent rules are easily extended to incorporate the use of coalitions. In this note, the following definitions for safety will be used.

> A cell is safe according to a prior posterior rule or a p-percent rule with n-coalitions if no lower or upper bound of a contribution can be calculated by a coalition of n other respondents within p percent of the true value.

This means that the 'normal' rules correspond to the rules with 1-coalitions. Note, that the use of n is not quite the same as in the dominance rule. In fact, the n used here plays a similar role as $n - 1$ in an (n, k)-rule. However, by choosing the definitions this way, there is no discontinuity between cases with $n = 1$ and $n \geq 2$. Furthermore, the interpretation is more straightforward.

The safety checks for both the prior posterior and the p-percent rules are easily extended to the cases with n-coalitions. With the notation of sections 3.1 and 3.2, a cell is safe according to the prior posterior rule with n-coalitions if

$$p\ x_1 - q \sum_{t=n+2}^{N} x_t \leq 0, \tag{9}$$

and a cell is safe according to the p-percent rule with n-coalitions if

$$\frac{p}{100}\ x_1 - \sum_{t=n+2}^{N} x_t \leq 0. \tag{10}$$

4. Authorization

If a cell is not safe according to the sensitivity measure used, the cell is suppressed. However, suppression may be prevented if respondents with large contributions authorize the release of information about these contributions. In that case, a cell value which is primary unsafe may still be published. Of course, authorization by one respondent does not extend to contributions of other respondents. Otherwise, the respondent with the largest contribution might give an authorization with the aim to disclose the value of the next largest contribution. This means that an authorization for the publication of

the largest contribution may not be sufficient to publish the cell value. This raises the question what criterion has to be used to determine if a cell can be published in the case of authorizations.

4.1. The prior posterior and p-percent rules with authorization

Consider a cell with contributions $x_1 \geq x_2 \geq \cdots \geq x_N$ by respondents $1, \ldots, N$, respectively. The cell is safe according to the prior posterior rule if equation (2) is satisfied. Analogously, for safety according to the p-percent rule, equation (5) has to be satisfied. For both rules, this means that the best bounds of x_1 computed by respondent 2 are not within p percent of the true value of x_1 and all other bounds will be worse in terms of percentages.

Now, consider the case where the equation for the prior posterior rule is not satisfied. This means that $p\ x_1 - q \sum_{k=3}^{N} x_k > 0$. If respondent 1 has given an authorization, it may still be possible to publish the cell. This depends on how well the other contributions can be estimated. Therefore, observe equation (1) and ignore the case with $i = 1$, because respondent 1 has given an authorization. The left hand side of the equation is largest for $i = 2$ and $j = 1$, because of the ordering of the contributions. If $p\ x_2 - q \sum_{k=3}^{N} x_k > 0$, it means that respondent 1 can find an upper or lower bound of x_2 within p percent of x_2. Without authorization by respondent 2, this means that the cell still has to be suppressed. If, on the other hand, $p\ x_2 - q \sum_{k=3}^{N} x_k \leq 0$, for none of the contributions $x_2, \ldots, x_N$ an upper or lower bound can be calculated within p percent of its true value and the cell can be published.

For example, observe the p-q rule with $(p, q) = (10, 50)$ for the cell with contributions 100, 90, 10, and 6. The respondent with contribution 100 has given an authorization. Now, $p\ x_2 - q \sum_{k=3}^{N} x_k = 10 * 90 - 50 * (10 + 6) = 100 > 0$ and thus, the cell has to be suppressed. Respondent 1 can calculate an upper bound of x_2 as $T - x_1 - \frac{q}{100}(x_3 + x_4) = 206 - 100 - 0.5 * (10 + 6) = 98$, which is within 10 percent of the true value of 90.

Of course, the argument can be generalized to a cell with several authorizations. The cell may be published if

$$p\ x_s - q \sum_{\substack{k=2 \\ k \neq s}}^{N} x_k \leq 0 \quad \text{(bounds of } x_s \text{ calculated by respondent 1)},$$

where x_s is the largest contribution without authorization. A nice property of this test is that if it is not satisfied, the original safety test (2) for the cell without authorizations was not satisfied either.

Furthermore, this test is easily extended to the p-q rule with n-coalitions. A cell with n-coalitions and several authorizations may be published if, with x_s

the largest contribution without authorization,

$$p\ x_s - q \sum_{k=n+2}^{N} x_k \leq 0 \quad \text{for } s < n+1,$$

or

$$p\ x_s - q \sum_{\substack{k=n+1 \\ k\neq s}}^{N} x_k \leq 0 \quad \text{for } s \geq n+1.$$

The same arguments can be used with the p-percent rule to find that a cell with authorizations may be published if

$$\frac{p}{100}\ x_s - \sum_{\substack{k=2 \\ k\neq s}}^{N} x_k \leq 0,$$

if x_s is the largest contribution without authorization, and analogously, for the p-percent rule with n-coalitions if

$$\frac{p}{100}\ x_s - \sum_{k=n+2}^{N} x_k \leq 0 \quad \text{for } s < n+1,$$

or

$$\frac{p}{100}\ x_s - \sum_{\substack{k=n+1 \\ k\neq s}}^{N} x_k \leq 0 \quad \text{for } s \geq n+1.$$

4.2. The dominance rule with authorization

A cell with contributions $x_1 \geq x_2 \geq \cdots \geq x_N$ is safe according to the dominance rule if equation (7) is satisfied. Otherwise, $(1-\frac{k}{100})\sum_{t=1}^{n} x_t - \frac{k}{100}\ \sum_{t=n+1}^{N} x_t > 0$. Now, suppose that the respondent with contribution x_1 in an unsafe cell has given an authorization. What test is needed now to see if the cell can be published?

For $n = 1$, the (n,k)-rule is equivalent to the requirement that the relative error in the best upper bound for x_1 is greater or equal to $\frac{100}{k} - 1$, if the bound is calculated by someone outside the cell (see section 3.3). This interpretation can be used to determine if the cell can be published. The respondent of x_1 can be viewed as an outsider to the rest cell, consisting of the contributions $x_2 \geq \cdots \geq x_N$. The best upper bound for x_2 by this respondent is equal to $T - x_1$. Analogously to (8), the cell may be published if

$$\frac{(T-x_1)}{x_2} - 1 \geq \frac{100}{k}\ - 1 \iff \frac{T-x_1}{x_2} \geq \frac{100}{k}.$$

Even if there are several authorizations, the best upper bound by an 'outsider' for the largest contribution without authorization x_s, is equal to $T - x_1$. In that case, publication of the cell is allowed if

$$\frac{(T-x_1)}{x_s} - 1 \geq \frac{100}{k}\ - 1 \iff \frac{T-x_1}{x_s} \geq \frac{100}{k}.$$

However, for $n \geq 2$, the example in section 3.3 shows that the relative error in an upper bound cannot be used. Thus, the previously mentioned tests are not possible. It also means that a natural choice as using a $(n-1,k)$-dominance rule on the rest cell with contributions $x_2, \ldots, x_N$ is impossible, because this choice is also based on the relative error in the upper bound. Namely, the absolute error in the upper bound for the second largest contribution found by a (worst case) coalition of $n-1$ other respondents in the original cell is equal to $\sum_{t=1}^{N} x_t - \sum_{t=1}^{n} x_t = \sum_{t=n+1}^{N} x_t$. In the rest cell with only the contributions $x_2 \geq x_3 \geq \cdots \geq x_N$, the absolute error in the upper bound for x_2 by a (worst case) coalition of $n-2$ other respondents is also equal to $\sum_{t=2}^{N} x_t - \sum_{t=2}^{n} x_t = \sum_{t=n+1}^{N} x_t$. A clear example why this rule is not a good choice, is given by a cell with the contributions 10, 10, 5, 3 and 2. Under the $(3, 80)$-dominance rule, this cell is not safe. Now, suppose that the respondent of one of the contributions with value 10 has given an authorization. The $(2, 80)$ rule applied to the rest cell yields a safe cell. However, why should one contribution of 10 in a cell be safe, while another contribution of 10 in the same cell would be unsafe?

So far, no good test has been found when authorizations are used with the (n,k)-rule for $n \geq 2$.

5. Protection levels for primary unsafe cells

5.1. Protection levels in relation to linear sensitivity measures

The linear sensitivity measures can be used to find primary unsafe cells. Then, to find feasible protection patterns with HiTaS or ARGUS, protection levels have to be defined. Of course, it would be best if the protection levels are chosen such that they agree in some way with the sensitivity measure used. In this section, possible ways to define the protection levels corresponding to different sensitivity measures are discussed. For a specific cell with total value T, the upper, lower, and sliding protection levels will be denoted by U, L, and S, respectively. This means that in the final protected table, the feasibility interval[1] should contain the protection interval $[T-L, T+U]$ with the length of the feasibility interval greater or equal to S.

5.1.1. Protection levels for the prior posterior rule

Consider the cell with ordered contributions $x_1 \geq x_2 \geq \cdots \geq x_N > 0$ by respondents 1, ... , N, respectively. Assume that a possible attacker tries to estimate the value of x_i using the feasibility interval from the final protected table. Then, the attacker can calculate an interval containing x_i with certainty.

[1]The feasibility interval of a suppressed cell is the interval of possible values of this cell with the given protection pattern.

Of course, this interval will be best if the actual feasibility interval is at its smallest, and thus equal to the protection interval. Again, the worst case will occur if respondent 2 calculates an interval for x_1. The upper bound of x_1 calculated by respondent 2 is equal to $\overline{x}_1 = T + U - x_2 - (1 - \frac{q}{100}) \sum_{t=3}^{N} x_t$. This bound is not within p percent of the true value of x_1 if $\overline{x}_1 \geq (1 + \frac{p}{100})x_1$. This is the case if

$$U \geq \frac{p}{100}\ x_1 - \frac{q}{100}\ \sum_{t=3}^{N} x_t, \tag{11}$$

which is equivalent to

$$U \geq (\frac{q}{100} + \frac{p}{100})\ x_1 + \frac{q}{100}\ x_2 - \frac{q}{100}\ T.$$

Similarly, requiring the calculated lower bound of x_1, equal to $\underline{x}_1 = T - L - x_2 - (1 + \frac{q}{100}) \sum_{t=3}^{N} x_t$, to be smaller than $(1 - \frac{p}{100})\ x_1$, leads to the following requirement for L

$$L \geq \frac{p}{100}\ x_1 - \frac{q}{100}\ \sum_{t=3}^{N} x_t. \tag{12}$$

Requirements for U and L in case of a p-q rule with n-coalitions or with varying p's and q's can be derived similarly.

The preceding inequalities have been derived on the assumption that the a priori knowlegde of respondents (modelled by parameter q) is restricted to other contributions of the cell. However, if the stronger assumption is used where a respondent also knows the values of contributions in other cells within at most q percent, the upper and lower protection levels should be chosen higher. Namely, the amount U which is the required upwards variability for the primary unsafe cell under consideration in the final protection pattern, is originating from another suppressed cell. Using the prior knowledge about the true value of that cell (known within at most q percent of its true value), it is possible that a respondent can calculate a better upper bound of x_1 than without this information. For example, the requirement of an upper protection level U could be met in the final protection pattern by (among others) choosing a secondary suppression with cell value U in the same row/column/... as the primary unsafe cell. It is possible that in this direction of the table, the secondary and primary suppression are the only two suppressed cells. Then, the best upper bound for x_1 by the respondent of contribution x_2 is equal to $T + U - x_2 - (1 - \frac{q}{100}) \sum_{t=3}^{N} x_t - (1 - \frac{q}{100})\ U$. Requiring that this bound is not within p percent of the true value x_1, results in the following requirement for U,

$$U \geq \frac{p}{q}\ x_1 - \sum_{t=3}^{N} x_t. \tag{13}$$

This is equivalent to

$$U \geq (1 + \frac{p}{q})\ x_1 + x_2 - T.$$

Note, that the required U is $\frac{100}{q}$ (which is larger than or equal to 1) times the amount which is required under the assumption that the a priori information of a respondent only concerns contributions in his own cell.

Of course, the same arguments hold for the lower protection level L, leading to the requirement

$$L \geq \frac{p}{q}\, x_1 - \sum_{t=3}^{N} x_t, \tag{14}$$

under the stronger form of a priori information.

For the sake of convenience, the weaker assumption of a priori knowledge is used in the rest of this note.

Note, that the aim of the protection levels, is to enlarge the intervals that can be calculated for the contributions in an unsafe cell, such that no interval can be found lying within p percent of the true value of the contribution. From the definition used for the p-q rule, this means that protection levels U and L have to be chosen, satisfying at least (11) and (12), respectively (or satisfying (13) and (14) under the stronger assumption of a priori knowledge). However, if the view is adopted that a suppressed cell is safe if for none of its contributions a lower *and* an upper bound within p percent of the true value can be found, it is not necessary to require both an upper and a lower protection level. Then, a feasibility interval containing values at least as large as T plus the righthand side of (11) or at least as small as T minus the righthand side of (12), will be safe. Thus, it is sufficient to require only an upper protection level or a lower protection level. Moreover, it is also sufficient if only a sliding protection level is chosen with S at least as large as the sum of the righthand sides of (11) and (12). Other possibilities are using a combination of an upper and a sliding protection level or a lower and a sliding protection level. In this note, suppressed cells will be considered safe if both the lower and upper bounds of each contribution are not within p percent of the true value.

The protection levels can be chosen such that the estimation interval for the largest contribution will contain a predefined length with values that will give estimates larger or smaller than p percent of its true value. Of course, if the estimation interval contains values outside q percent of the value of x_1, the attacker will know that they are not true. Thus, it is useless to choose the protection levels too large.

5.1.2. Protection levels for the p-percent rule

Protection levels for the p-percent rule can be found in a similar way to those found for the p-q rule. This leads to the following requirement for the upper

protection level U

$$U \geq \frac{p}{100}\ x_1 - \sum_{t=3}^{N} x_t = (1 + \frac{p}{100})\ x_1 + x_2 - T. \tag{15}$$

For the lower protection level, no requirement can be derived. This is because the lower bound of a feasibility interval cannot be used to find a lower bound of the largest contribution (except for zero, but this bound is already known). Namely, in the worst case where the feasibility interval is equal to the protection interval, the respondent of the second largest contribution will calculate a lower bound of the largest contribution as: the lowest possible value of the cell $(T-L)$ minus his own contribution (x_2) minus an upper bound of the sum of the remaining contributions. However, because there is no a priori information, the best value for the upper bound of $\sum_{t=3}^{N} x_t$ is equal to $T + U - x_2$. Then, the lower bound $\underline{x}_1$ of the largest contribution is equal to

$$\underline{x}_1 = T - L - x_2 - (T + U - x_2) = -(L + U) \leq 0,$$

and this was already known.

In this situation, the restrictions on the upper protection level should be used. No choice of only lower and/or sliding protection levels will guarantee a feasibility interval containing values leading to an upper bound of the largest contribution not within p percent of its true value.

Note, that the formulas for the p-percent rule also follow directly from those for the p-q rule by substituting $q = 100$ for the upper protection level and $q = \infty$ for the lower protection level.

5.1.3. Protection levels for the dominance rule

For the dominance rule, a requirement on the total cell value can be derived from equation (6). A cell is safe according to the dominance rule if

$$\sum_{t=1}^{N} x_t \geq \frac{100}{k}\ \sum_{t=1}^{n} x_t.$$

This inequality will be used to derive requirements on the protection levels. Note that the relative error in the estimate for x_1 can not be used, because the dominance rule is not consistent with respect to this error (see the example in section 3.3).

From the definition of safety according to the (n, k)-rule, the difference between the sum of the n largest contributions and the cell value should be at least $100-k$ percent of the cell value. Thus, the upper bound of the protection interval $T+U$ should be at least equal to the sum of the n largest contributions plus $100 - k$ percent of $T + U$. This leads to the following requirement for U

$$U \geq \frac{100}{k}\ \sum_{t=1}^{n} x_t - T. \tag{16}$$

Restrictions on the lower protection level can not be found.

5.2. Protection for individual cells compared to cell combinations

In the previous section, possible ways to define protection levels for individual cells are described. Using the method of Fischetti and Salazar, a protection pattern for each sub-table of a hierarchical table can be found. This pattern will be such that the primary unsafe cells can not be calculated within the given protection levels. However, the value of various aggregations of suppressed cells can be calculated exactly. For example, when only two cells in the same row of a table are suppressed, their total value can be computed from the subtotal and the other known values in the row. These aggregations which can be calculated exactly, are not restricted to cells in the same direction of a table but can occur through the whole table. In fact, such aggregations may be distributed over several sub-tables of one hierarchical table, with several parts of the aggregation located in different sub-tables (note, however, that the protection method is used per sub-table). If an aggregation whose value can be calculated exactly, does not contain a smaller aggregation with this property, it is called an elementary aggregation. Note that these elementary aggregations depend on the protection pattern.

Now, the question can be asked if these aggregations are safe, or if additional protection is needed. A suggestion (see T. de Waal, 2000) is to define a suppression pattern sufficient if and only if all elementary aggregations of the pattern satisfy the sensitivity measure used[2]. For example, if an (n, k)-rule is used, the biggest n contributions in an elementary aggregation should not exceed k percent of its total value. However, such a requirement need not correspond to the approach to obtain safe tables through suppression patterns based on protection levels.

Even more general, the use of protection levels need not correspond with the sensitivity measure used. This means that a table which is protected using protection levels, may contain a cell for which the protection interval does not contain one value for which the cell satisfies the sensitivity measure (note that this is not possible if the protection levels are chosen in a 'right' relation to the sensitivity measure). This may happen for example, when a dominance rule is used to obtain the primary unsafe cells, which are then protected with protection levels equal to a certain percentage of the cell value.

Also, in the case where the protection levels are derived from the sensitivity measure used, elementary aggregations do not alway satisfy this sensitivity

[2]Note, that the combination of cells of an elementary aggregation may be spread over different (parts of the) subtables. Therefore, this may lead to (in practice) very unlikely combinations of cells and respondents. Thus, the choice can be made to restrict additional protection to only a part of the elementary aggregations.

measure. For example, consider the left hand table in Figure 1 with only two primary unsafe cells, the ones with value 43 and 49 in the first row. The cell with value 43 has contributions 32, 4, 4, 2 and 1, and the cell with value 49 has contributions 33, 6, 5 and 5. According to a $(2, 70)$-dominance rule, the cells are primary unsafe. According to requirement (16), the upper protection level of the first cell can be chosen equal to 9 and of the second cell equal to 7. Then, the protection pattern in the second part of the figure protects both cells sufficiently. The feasibility interval for both unsafe cells is equal to $[T - 9, T + 9]$ with T the cell value of the cell considered. However, the total value of the combination of the suppressed cells in the first row is known and equal to 92. Furthermore, the sum of the two largest contributions in this combination is equal to $32+33 = 65$, which makes the combination cell primary unsafe according to the $(2, 70)$-dominance rule.
Note that in this example, the part of each unsafe cell needed to protect the other unsafe cell does not involve the two largest contributions.

10	**43**	**49**	102
11	9	9	29
21	52	58	131

$\Longrightarrow$

10	X	X	102
11	X	X	29
21	52	58	131

Figure 1 Unsafe elementary aggregations with two unsafe cells

In the previous example, two unsafe cells were used to protect each other. However, even if there is only one unsafe cell, there may arise an elementary aggregation which does not satisfy the dominance rule. Therefore, consider the table in Figure 2 with only one primary unsafe cell, the one with value 43. The cell has contributions 32, 4, 4, 2 and 1. According to a $(2, 70)$-dominance rule, the cell is primary unsafe. Again using (16), the upper protection level can be chosen equal to 9. With the same protection pattern as in the previous example, the feasibility interval for the unsafe cell is equal to $[34, 52]$. Now, suppose that the contributions to the cell in the first row with value 9 are equal to 5, 1, 1, 1, and 1, respectively. Then, this cell is safe. However, with the known total value of the suppressed cells in the first row equal to 52 and the sum of the two largest contributions in this combination equal to $32 + 5 = 37$, the combination cell is again primary unsafe according to the $(2, 70)$-dominance rule.

10	**43**	9	62
11	9	9	29
21	52	18	91

Figure 2 Unsafe elementary aggregation with one unsafe cell

This problem is not restricted to the use of the dominance rule with $n \geq 2$ (note that this problem does not occur when a $(1, k)$-rule is used). For example, for the p-q and p-percent rules, the same kind of examples can be constructed.

Thus, elementary aggregations may need extra protection, even if the protection levels are derived from the sensitivity measure used. The problem is even worse, if the same respondent can occur in different cells. This has to do with the holding problem, which is discussed in section 7.1. In that section, a possible solution is suggested which will also work in the case of unsafe aggregations. It is based on restrictions on the amount of protection that cells can give each other. Only the part of the cell which is extra in relation to the minimal cell value for which the cell would be safe, may be used for protection of other cells. This is called the protection capacity of the cell. Unfortunately, the solution will lead to overprotection almost certainly.

In the next paragraph a possible solution for the problem of suppressed cell combinations with too few contributions in the same direction of the table is proposed. This approach could be adapted to the problem of elementary aggregations in a specific direction violating the linear sensitivity measure. Then, the protection levels are chosen such that elementary aggregations of cells occurring in the same direction (row/column/...) of a table will satisfy the desired linear sensitivity measure. In fact, for each unsafe cell, all combinations with other cells in the same direction are tested for compliance with the linear sensitivity measure. Then the protection levels are chosen higher than the protection given by the largest combination which would be unsafe. This approach will require very much extra time and effort and is expected to cause substantial overprotection. In practice, this solution does not seem suitable to enforce the linear sensitivity measure for elementary aggregations.

5.3. Protection levels for cells with too few contributions

Linear sensitivity measures are often used in combination with a rule defining a cell primary unsafe when the number of contributions is lower than a certain number. To use this kind of 'number rule' with HiTaS or ARGUS, these cells are assigned the status primary unsafe. Some positive protection levels for these cells have to be defined, to prevent the values of these cells to be calculated from the protected table.

Furthermore, the protection patterns are preferred to be such that combinations of suppressed cells in the same direction of a table, also satisfy the number rule and the sensitivity measure. In fact, this could also be required for every elementary aggregation. This means that the protection levels at least should be chosen in relation to the linear sensitivity measure (see section 5.1). Unfortunately, as shown in the previous section, it is difficult to guarantee the compliance with the linear sensitivity measure. Also, the most straightforward ways to define protection levels will not guarantee that the number rule is satisfied.

For example, when a number-3 rule is used, cells with less than four contribu-

tions have to be protected. Assume that this rule is used in combination with a $(2, 80)$-dominance rule. For all cells with four or more contributions, the dominance rule defines which cells are sensitive. To protect these cells, protection levels such as suggested in section 5.1.3 can be used. However, all cells with less than four contributions are sensitive because of the number rule. Cells with three contributions might also be sensitive because of the dominance rule. In this last situation, again protection levels from section 5.1.3 can be used. For cells with three contributions satisfying the dominance rule and for cells with less than three contributions, at least one positive protection level has to be chosen. Given the final protection pattern, it will then be impossible to calculate the value of the cell exactly. However, this does not mean that requirements concerning the combination of suppressed cells are met. The protection pattern can be such that two cells, each with one contribution, are used to protect each other in a row of the table. However, the value of the combination 'cell' can be calculated and is composed of only two contributions. In the original table, this kind of cell would be primary unsafe (each respondent can calculate the other contribution exactly).

In this example, the problem that suppressed combinations do not satisfy the number rule could be prevented by requiring that only cells with at least two contributions may be used to protect other cells. Of course, the example can be changed easily by requiring a number-5 rule, such that this solution does not work any longer.

Remark. *Even when no number rule is used in combination with a sensitivity measure, it is still a good idea to require at least two contributions for cells used to protect others. This is because a single respondent to a specific cell is often able to calculate the values of the other suppressed cells (assuming that the respondent knows that he is the only one in the cell). In principle, this requirement could be used in HiTaS/ARGUS. It is achieved by choosing the a priori bounds*[3] *of the cell equal to zero if there is only one contribution (the protection capacity is set to zero). Then, the cell is automatically excluded for the protection of other cells. However, in the current implementation of the suppression algorithm, the a priori bounds of a cell need to be larger than the protection levels. That means that unsafe cells (with some positive protection levels and thus, positive a priori bounds) may be used to protect other unsafe cells. This holds in particular for primary unsafe cells with only one contribution.*

Thus, the choice of protection levels based on the structure of the cells involved, will not suffice to ensure that elementary aggregations of suppressed cells whose values can be calculated, satisfy the number rule. Still, the cells with too few

[3] The a priori bounds l_i and u_i for cell i define an interval $[T_i - l_i, T_i + u_i]$ around the total cell value T_i. This interval contains the possible values that a cell may assume to protect other cells. It is based on the idea that attackers will know certain bounds for the cell values. See also section 7.

contributions have to be protected, so protection levels have to be chosen. The question is what levels then to use?

Instead of only using the internal structure of a cell, the structure of other cells can be taken into account. Of course, this means that this information has to be available. If that is the case, the following approach will guarantee the compliance with the number rule for combinations of suppressed cells in the same direction of the table, at least if the holding problem (section 7.1) is ignored.

Consider the case where a number-N rule is used in combination with a linear sensitivity measure. Choose a cell i which is unsafe because of the number rule (possibly not safe because of the sensitivity measure either). Determine the upper protection level $U_i(ls)$ for the cell according to the linear sensitivity measure (see section 5.1). Consider all combinations of the cell with other cells in a specific direction d, with at most $N-1$ contributions (note that the other cells in the combinations are also unsafe because of the number rule). Find the combination where the total value of the extra cells (without the value of the unsafe cell under consideration) is maximal, say $V_i(d)$. Let $U_i(d) = V_i(d) + \varepsilon$ for some small positive ε. Then, choose the upper protection level of the cell larger or equal to $\max\{U_i(ls), \max_d U_i(d)\}$. Thus, only combinations of suppressed cells with more than N contributions in a specific direction will be able to satisfy the protection level.

Of course, this approach will lead to overprotection, because the 'worst' combination of cells imposes a necessary restriction in a specific direction. However, it also imposes this restriction on the other directions where it could be too large.

The approach can be improved somewhat by considering for a specific direction d (row/column/...) all $U(d)$'s of the primary unsafe cells with too few contributions in that direction at once. Then, it is sufficient to assign only one of these cells a protection level of at least its corresponding $U_i(d)$ to ensure that the combination of suppressed cells in that direction will satisfy the number rule. To avoid as much overprotection as possible, the smallest $U_i(d)$ can be chosen. The $U(d)$'s of the other cells in that direction can be set to zero. It is possible that one of these other cells already had a positive protection level caused by a $U(d')$ from another direction d'. If this $U(d')$ is larger than the one needed in direction d for this cell, this cell can be chosen to assign a positive $U(d)$ instead of the cell with the smallest $U(d)$. After assigning each cell in the table a $U(d)$ for every direction d, the actual upper protection levels are chosen as $\max\{U_i(ls), \max_d U_i(d)\}$ for every cell i. To ensure suppressed combinations which are safe according to the number rule, it is not necessary to assign positive lower protection levels. Of course, positive lower protection levels may be needed for the linear sensitivity measure used.

6. Protection levels for secondary suppressions

The HiTaS approach is used for the protection of hierarchical tables by a top-down approach. A hierarchical table is divided into sub-tables, which are protected separately. However, secondary suppressions found in some sub-table need to be protected in all other sub-tables where they occur. This means that they play the role of primary unsafe cells in the other sub-tables. Note, that it is possible that they occur in sub-tables which were already protected. Then, these sub-tables have to be protected anew, taking into account the extra suppressions.

To force HiTaS to protect cells sufficiently which are secondary suppressed in another sub-table, some positive protection levels have to be chosen. On first sight, these protection levels should be chosen such that they guarantee the level of protection for the primary unsafe cells for whose protection they are suppressed.

For example, consider a two-dimensional table with only one primary unsafe cell with upper protection level equal to 10 and lower protection level equal to 5. Let the final suppression pattern consist of the four corner points of a rectangle, with the primary unsafe cell as corner 1, and the three secondary suppressions clockwise numbered as 2, 3, and 4. Then, suppressions 2 and 4 will be able to vary between $[T_i - 10, T_i + 5]$ with T_i the total value of suppression i. A good choice for their upper and lower protection levels would be 5 and 10 respectively. Suppression 3 however, needs to vary between $[T_3 - 5, T_3 + 10]$, so the upper and lower protection levels can be chosen as 10 and 5 respectively.

Unfortunately, the information used to choose the secondary protection levels in the example is not available within HiTaS. The method used to find the secondary suppressions does not give information about the amount of variability needed in the secondary suppressions. Even worse, there is no information available about which primary unsafe cells have caused the secondary suppressions. It is even possible that the same secondary suppression is used for the protection of different primary unsafe cells. Thus, it is not possible to choose protection levels in this way.

A possible approach to choose protection levels could be based on the feasibility intervals of the secondary suppressions. These intervals can be calculated by solving two linear programming problems for each secondary suppression. Apart from the amount of time this will take, the feasibility intervals might be much larger than really needed for the protection of the primary unsafe cells. Using these bounds for protection levels would lead to overprotection and thus, this is no good solution either.

Still, some choice for the protection levels has to be made. A possible heuristic iterative approach is the following. After protecting a sub-table, all secondary

suppressions are assigned an upper protection level equal to the maximum of the upper and lower protection levels of the suppressed cells (primary and secondary) located in the same direction(row/column/...) of the table. Then, the lower protection level of each secondary suppression is taken equal to its upper protection level. This procedure is repeated until no more changes take place. With this procedure, the marginal entries in the table are seen as part of the row/column/... under consideration.

The heuristic can be accelerated by simplifying it even further. This can be done by assigning each secondary suppressed cell an upper protection level equal to the maximum of the upper and lower protection levels of all the primary unsafe cells in the table. Then, the lower protection level is taken equal to its upper protection level.

These heuristic procedures are very fast (compared to the solving of linear programming problems), but probably lead to overprotection. In the first place, this is caused by choosing the lower and upper protection levels symmetrically around the cell value. In the second place, secondary suppressions can get assigned protection levels from primary unsafe cells for which they are not used in the protection pattern. For example, the only reason that a secondary suppression is suppressed, might be for the protection of a particular primary unsafe cell with upper and lower protection levels U and L. However, if this secondary suppression is located in a row/column/..., where other secondary or primary suppressions can be found with larger upper or lower protection levels, the protection levels of the secondary suppression will get assigned these larger values.

7. Use of a priori bounds

For the protection of unsafe cells, protection intervals are chosen. The method used in HiTaS/ARGUS then finds patterns such that the feasibility intervals contain these protection intervals. This means that other suppressed (safe) cells must have a possible variability such that the desired protection levels for the unsafe cells are met. In principle, it is possible that a suppressed safe cell must be able to assume values between zero and a large positive value. These values could be very improbable. Often, outsiders know for example, which cells in a table are not equal to zero. Therefore, the method of HiTaS/ARGUS uses a priori bounds on the cell values. These are used to model a realistic interval of variability of a suppressed cell. Another interpretation is that these bounds model the amount of protection that this cell can give another cell. Thus, for each cell i, the non-negative a priori bounds l_i and u_i define an interval $[T_i - l_i, T_i + u_i]$ around the total cell value T_i. This interval contains the possible values that a cell may assume to protect other cells.

This last interpretation leads to the wish to set these bounds equal to zero for cells with only one respondent. Namely, if such a cell is used in a protection pattern and the respondent knows he is the only respondent in the cell, he could possibly unravel a part of or the whole protection pattern. Furthermore, it can be assumed that for primary unsafe cells the a priori bounds (at least the lower bounds) are desired to be small. However, the current implementation of the method in HiTaS requires that the protection interval is contained in the a priori interval. This means that for the time being, this choice is not possible. However, in the future this restriction will probably be removed.

If this restriction is indeed eliminated, there is still a point of interest. By choosing a priori intervals with length zero for cells with one respondent, these cells cannot be used to protect other cells. However, if in each dimension of the table, at least two other cells are suppressed, the respondent will not be able to unravel the protection pattern. By the choice of the zero a priori bounds, this possible protection pattern is excluded and thus, some overprotection may result.

Apart from the desire to choose small or zero bounds for unsafe cells, a priori bounds must be assigned to the safe cells. If the holding problem is ignored (see section 7.1), several choices are possible. For example, bounds defining a 50 to 150 percent interval around the cell value seem reasonable. However, for the choice of the lower bound it may be wise to take the size of the largest contribution into account. The a priori lower bound should be such that the cell value cannot become smaller than its largest contribution. Otherwise, if the lower bound is really necessary in a protection pattern, the pattern will not be safe with respect to the respondent of this largest contribution. This kind of problem does not occur for the choice of a priori upper bounds.

7.1. Holding problem

In this note, an important problem in the protection of tables is ignored up till now. This issue is the so-called holding problem, and is caused by respondents contributing to different cells. In some cases, these contributions can be seen as different ones, in other cases, these contributions have to be added and are seen as one contribution. For example, a company with different branches in different regions may contribute to different cells in a sub-table corresponding to these different regions. However, in a marginal cell of the sub-table, these contributions can be seen as coming from the same company, so they have to be added and counted as one contribution. This is important, because most rules to determine what cells are primary unsafe use information about the number and the size of the contributions in a cell. This means that a certain amount of information about the microdata underlying a table has to be known. It is not sufficient to know the sizes of the contributions, but the respondents themselves

must be known too. If this information is not available, it is not possible to take the holding problem into account.

If the desired information about the microdata is available, the table can be built using this data, while at the same time determining what cells are primary unsafe. Thus, for the determination of the primary unsafe cells, there need not be a problem. However, there is no provision in the protection method in Hi-TaS/ARGUS to check if candidates for secondary suppression have common respondents with the primary unsafe cells. Nevertheless, in finding safe protection patterns, the information is still very important.

To illustrate the need for this information, consider the following example. Suppose there is a two-dimensional table with the $(1, 75)$-dominance rule. There is only one primary unsafe cell with contributions 75, 15 and 5. According to (16), this cell needs an upper protection level of at least 5. Suppose that the final protection pattern contains only one secondary cell in the same row as the primary unsafe one. The contributions of this cell are 20, 5 and 5. Obviously, this cell is safe and when the holding problem is ignored, the combination of these two cells would be safe too according to the dominance rule used. However, not ignoring the holding problem and assuming that the largest contributions in both cells originated from the same respondent, the combination cell is not safe. Namely, the largest contribution is equal to $75 + 20 = 95$ and this is equal to 76 percent of the total value of the combined cell. Thus, this is a suppression pattern to be prevented.

Considering the fact that the protection method cannot use information about particular respondents, another approach is needed. It might be possible to obtain a protection pattern without the problems as in the previous example by choosing the right a priori bounds. This is done in the following way. If a cell is safe according to the linear sensitivity measure used, the smallest value of the cell total is computed for which the cell would still be safe. For the p-q rule, this would be equal to $T + \frac{p}{100}\ x_1 - \frac{q}{100}\ \sum_{t=3}^{N} x_t$[4], for the p-percent rule $T + \frac{p}{100}\ x_1 - \sum_{t=3}^{N} x_t$, and for the (n, k)-rule this would be equal to $\frac{100}{k}\ \sum_{t=1}^{n} x_t$. This value is subtracted from the total cell value. The result R is called the protection capacity and will be positive, because the cell is safe. The amount R can be used for protection of other cells, without causing problems because of joint respondents. Namely, consider the case of a combination of two cells, one primary unsafe and one safe. Then, the worst case for the safety of the combination cell occurs if the respondent of the i-th largest contribution of each cell is the same for all i. Let R be defined as above for the safe cell and define R_p and R_c analoguously for the unsafe cell (note that R_p will be negative) and the combination cel. Then, $R_c \geq R + R_p$ for all possibilities of

[4]This is independent of the interpretation of q for a weak or strong assumption of a priori information.

respondents, with equality in the worst case with the same respondents for the ordered contributions. Of course, the same argument can be used with more than one safe cell. Note, that this approach works because of the subadditivity of linear sensitivity measures.

To restrict the amount of the cell values to be used for protection of other cells, the lower a priori bounds of all cells can be chosen equal to or smaller than the R involved. Of course, these bounds should be non-negative, thus if R is negative, a lower bound of zero is chosen. This means that unsafe cells receive lower a priori bounds equal to zero. In the previous example, for the safe cell $R = 30 - \left(\frac{100}{75} * 20\right) = 3\frac{1}{3}$. By setting the lower a priori bound of the secondary cell equal to R, the cell on its own can not give enough protection to the primary unsafe cell (because the secondary suppression should be able to decrease at least 5) and another protection pattern will be found.

Note that with this approach, cells with only one respondent receive lower a priori bounds equal to zero. Thus, these cells will never be used to protect other cells. However, to implement this approach in HiTaS, the method for the protection of sub-tables needs to be adapted to allow cells with smaller a priori bounds than protection levels (see also the remark in section 5.3). Moreover, this approach focuses on the linear sensitivity measure used and does not solve problems originating from using a number rule.

The drawback of this approach is that it may lead to overprotection. In most cases, the respondents in two different cells will be different from each other. However, the restriction on the a priori lower bounds is based on the worst case situation where the t-th largest contributions in different cells are provided by the same respondent for all t involved in the sensitivity measure used. Thus, the a priori bounds may be too restrictive. This can be improved if the lower bounds are calculated at the time that the table is built. Then, it would be possible to check if the same respondents occur in different cells of a table. If that is the case, than only the a priori lower bounds of those cells are restricted in the way described.

7.2. Suppression of 'safe' cells and publication of 'unsafe' cells

For some reasons, it is possible that cells which are safe according to the rules used, may not be published for another reason. For example, because this value is suppressed in another table (not a part of the same hierarchical table, then the tables are linked) for the protection of other cells. With HiTaS/ARGUS, this can be achieved by assigning this cell the status primary unsafe and choosing at least one positive protection level. Every choice of a positive value will suffice to include the cell in the suppression pattern and ensuring that it cannot be calculated exactly. However, if a very small protection level is chosen, it might

be possible to calculate the cell almost exactly. Therefore, the choice of the protection levels should depend on the reason for the suppression.

On the other hand, some cells which are unsafe according to the rules used, still have to be published. This is easily achieved because HiTaS/ARGUS provides the possibility to assign cells the status 'published', which will ensure that these cells are not suppressed. Alternatively, publication of a cell can be achieved by assigning the status 'safe' and setting all protection levels and the a priori bounds equal to zero. The last action is needed to ensure that the cell is not used as secondary suppression.

8. Weighted cell contributions

A final subject in this note is concerned with cells consisting of contributions with weights. Up till now, each cell is assumed to be composed of a number of (possibly one) positive contributions by different respondents. The total cell value is obtained by summing the individual contributions. The size and number of these contributions are used in determining the safety of the cells. However, in practice, cells are often calculated as the sum of weighted contributions. For example, in the case where the information is gathered by a survey. Then, one large contribution may be representative for the contributions of several other respondents which did not occur in the sample and this one contribution gets a weight larger than one.

It is possible to transform the case with weighted contributions to the 'normal' situation if all weights are positive. Then, most of the topics discussed in this note are still valid. If all weights are integer, the transformation is simple. Each contribution is duplicated according to its weight, such that different contributions with the same value are created. For example, a contribution with value 100 and weight 5 transforms into five different contributions with value 100. This way, the number and size of contributions is known. However, the respondents of the different contributions are no longer known. The respondent of the weighted contribution can be linked to only one of the created contributions. The respondents of the other contributions are unknown.

In the case with non-integer weights, the same kind of approach could be used. By truncating each non-integer weight to the nearest lower integer, the contributions can be duplicated in the way above. The decimal part of the weight is transformed into one contribution with the value equal to the original value times this decimal part. For example, a contribution with value 100 and weight 5.25 is transformed into five contributions with value 100 and one contribution with value 25 ($= 100 * 0.25$).

References

Fischetti, M. & Salazar-González, J.J. (1998). *Models and algorithms for optimizing cell suppression in tabular data with linear constraints*. Report, University of La Laguna, Tenerife.

Waal, T. de (2000). *Voorstel tot tabelbeveiliging gebaseerd op CONFID en ACS*. Concept nota, CBS, Voorburg.

Wolf, P.-P. de (1999). A heuristic approach to cell-suppression in hierarchical tables. Research paper no. 9913, CBS, Voorburg.