Measuring Risk in Tables Where the Study Variable May be Negative

Ann-Marie Flygare[1], Ingegerd Jansson[2], Tiina Orusild[3]

[1]Statistics Sweden, Klostergatan 23, 701 89 Örebro, Sweden
[2]Statistics Sweden, Karlavägen 100, 104 51 Stockholm, Sweden
[3] Statistics Sweden, Karlavägen 100, 104 51 Stockholm, Sweden

Abstract

Disclosure control of tables where the study variable may take both positive and negative values poses a particular problem. The common sensitivity rules like the dominance rule and the *p*% rule do not apply. To solve this, it has been suggested to transform cell values (e.g. add a constant or take absolute values) in order to make all values positive and facilitate the use of the common sensitivity rules. With this approach, it is assumed that the risk scenario for variables that may take negative values is similar to variables that only take positive values. We use empirical and simulated data to illustrate how the common sensitivity rules perform in different situations and we initiate a discussion of the sensitivity of data that may take both positive and negative values and discuss the need for a different approach to determining the sensitive values. Ideas for modified measures of risk are presented.

Key Words: statistical disclosure control, sensitivity measures, magnitude tables, negative contributions

1. Introduction

In statistical disclosure control, determining the sensitivity of the cells in a table is the first step towards producing a properly protected table that will be safe enough to publish. A widely used and often recommended measure to determine the sensitivity of cells in a magnitude table is the *p*%-rule. Another well-known rule is the (*n*,*k*)-rule. These concentration rules will work when the variables that the table is based on can take only non-negative values, but there are a number of economic variables that can take also negative values. In order to apply sensitivity measures to tables with such data, some approaches have been suggested in the literature.

Giessing (2008) proposes two methods for dealing with variables that can take both positive and negative values. The first one is to relax the parameters of the common concentration rules by either reducing *p* or increasing *k*. It might even be adequate to replace the concentration rules by a frequency rule, since data taking both negative and positive values are likely to be less sensitive than data taking only nonnegative values. The same approach is suggested in Hundepool et al (2012). The second proposed method is to transform micro data by adding to the original values a constant M. Each original value is at least of size M, and M is assumed to be known. The sensitivity of data is then determined by applying the common concentration rules on the transformed data. With this rule, secondary suppression can be carried out using the methods available in the τ-Argus software (Hundepool et al 2011).

Hundepool and de Wolf (2011) describe improvements to the modular approach in τ-Argus, including how to deal with negative cell values, but they only consider secondary suppression, assuming that the sensitive cells are already determined. The approach is to add the smallest cell value plus 1 to all interior cells in the table, and to their lower and upper a priori bounds, and then adjust the marginal cells accordingly.

Daalmans and de Waal (2011) discuss the problem that when a (p,q)-sensitivity rule is applied to single cells in a table and a safe suppression pattern is determined for the table based on suppressions intervals, the resulting pattern might still not offer sufficient protection. They suggest to solve the problem by using an extended (p,q)-sensitivity rule and apply the rule to aggregations of primary suppressed cells. With the extended sensitivity rule, the absolute contributions of the respondents are used, and thus negative contributions are allowed.

Tambay and Fillion (2013) describe the methodology for cell suppression as used in Statistics Canada´s G-Confid system, including how negative values can be handled. The authors state that using the absolute value of the contributions to a cell will often suffice, but if a very large unit has a small absolute value on the variable to be protected, they suggest the use of a proxy variable. With a proxy variable of the form discussed in the paper, a safe suppression pattern can be determined.

The methods described in the literature referred to above suggest to somehow change the variable values, i.e. add a constant or take the absolute value, or to use a proxy (or shadow) variable. It is implied that the risk scenario is the same as with a study variable taking only nonnegative values, and the suggested approaches are to apply traditional (or extended) sensitivity measures to the altered or replaced data. With these approaches, the sensitivity measure is not applied to the same data that a possible intruder has access to, that is, the risk is not measured under the true scenario. (Giessing (2008) assumes in the second scenario above that the constant M is known, but does not discuss if this would be a realistic assumption.) We mean to investigate a different approach with alternative measures focusing on the dispersion of the data. To focus on dispersion has also been suggested by Domingo-Ferrer and Torra (2002), as a solution to specific cases when the traditional sensitivity measures are insufficient. They suggest basing the sensitivity assessment on the concentration of the relative contributions to a cell, i.e. the distribution of values instead of the actual values, and propose to use the entropy of the relative contributions.

2. Alternative Measures of Sensitivity

In the following, we assume that the surveyed population is totally enumerated. Before measuring the risk in a cell, we need to consider the division between negative and positive values within a cell. A measure that seems reasonable to use is

$$T_R = \frac{\min\left(\left|T_-\right|, T_+\right)}{\max\left(\left|T_-\right|, T_+\right)}$$

where T_+ is the sum of all positive contributions to a cell value and T_- is the sum of all negative contributions. We thus have that the cell value, the total, equals $T_+ + T_-$. The ratio will be zero if there are only positive or only negative contributions. In this case, the traditional sensitivity measures can be used. If the ratio is close to zero, the contribution of either negative or positive values might be negligible, suggesting that traditional methods can still be used. However, if the ratio is larger than a constant *k* (where k is a small value), there are non-negligible contributions of both positive and negative values and we need an alternative measure that can be calculated irrespective of the sign of the values in the cell.

Consider

$$S_{R,h} = \frac{S_{n-h}^2}{S_n^2}, \qquad (1)$$

where S_n^2 is the variance in the cell, calculated using all objects in the cell, and S_{n-h}^2 is the variance when the *h* largest (in terms of absolute value) objects are omitted. If there are no extreme values, the variance is not much affected when the largest absolute values are excluded from the calculation. On the other hand, if the largest values differ very much from the other values, the drop in the variance ratio will be large when an extreme value is excluded. The cell is considered as safe if the ratio S_R is larger than some constant *c*. First results indicate that *c* should be small, maybe below 0.05.

Another measure taking the dispersion into account is the following ratio,

$$Q_{R,h} = 1 - \frac{\sum_{h \geq 1} \left(\max_{i \in A^{h-1}} \left\{ \left| y_i \right| \right\} - \tilde{y} \right)^2}{\sum_{i=1}^{n} \left(y_i - \tilde{y} \right)^2}, \qquad (2)$$

where $\tilde{y}$ is the median and A^0 is the set of all observations, A^1 is the set of all observations excluding the largest (in terms of absolute value), A^2 is the set of observations excluding the two largest, etc. If Q is close to zero, the cell is sensitive.

The two different measures give similar results, except that the variance ratio in (1) can take values larger than 1, while the quantile ratio in (2) is always smaller than or equal to 1.

3. Empirical illustration

The following example is an empirical illustration of the behavior of the traditional rules and the suggested measures. Data are from a real survey where the variable can take both positive and negative values. The original data is shown in the first column of Table 1. The second and third columns show a transformation of the data and the absolute values, respectively. All the values in a column of Table 1 belong to the same cell. Looking at the observations, there is reason to suspect that two of the values are so much larger than the rest that the cell total might be too sensitive to publish.

Table 1: Original data, transformed data and absolute values of the original data

Original data, y_i	*Transformed data,* y_i *+ 19303*	*Sorted absolute values,* $\lvert y_i \rvert$
-19 302	1	0
-18 599	704	0
-1 409	17 894	0
-582	18 721	1
-485	18 818	3
-463	18 840	3
-11	19 292	6
-3	19 300	11
-3	19 300	11
0	19 303	11
0	19 303	32
0	19 303	236
1	19 304	391
6	19 309	463
11	19 314	485
11	19 314	582
32	19 335	715
236	19 539	1 356
391	19 694	1 409
715	20 018	18 599
1 356	20 659	19 302

The dominance rule and the *p*%-rule were applied to the transformed data. The alternative measures, the variance ratio $S_{R,h}$ and the quantile ratio $Q_{R,h}$ were calculated using the original data.

For the transformed data the largest observation contribute with 6 percent to the total and the two largest observations contribute with 11 percent to the total. Thus according to the dominance rule the cell is far from sensitive. The *p*%-rule gives the same result, the cell is not sensitive.

For the absolute values of the original data the largest observation contribute with 44 percent to the total and the two largest observations contribute with 87 percent to the total. The dominance rule with commonly used parameters (*n*,*k*) = (1,50) and (2,90) indicates that the cell is safe. With the *p%*-rule, the cell is safe if *p* < 30%.

In this example the T-ratio is $T_R \approx 0.07$, which is small but probably far enough from zero to indicate that the standard measures should not be used. This is also justified by looking at the data: significant parts of the observations in the cell are negative or positive, respectively.

In Table 2, results for the alternative measures are shown. For *h*=2 the values of both ratios are small, indicating that there is a risk of disclosure.

Table 2: Results for alternative measures using original data.

h	*Variance ratio* $S_{R,h}$	*Quantile ratio* $Q_{R,h}$
1	0.536	0.485
2	0.009	0.007

4. Simulations

To further compare the variance ratio with the combined dominance rule, (1,50) and (2,*90*), we performed the simulations described below.

Figures 1a, 1b, and 1c show the variance ratio for *h*=0, 1,…,5 with only positive values, in order to facilitate comparison with the dominance rule. The following cases are considered:

1. Simulation, Uniform(0,10), 1000 observations, no extreme value
2. Simulation, Uniform(0,10), 1000 observations, one extreme value, the largest observation contribute with 50% to the total
3. Simulation, Uniform(0,10), 1000 observations, two extreme values, the largest observation contribute with 50% and the two largest contribute with 90% to the total

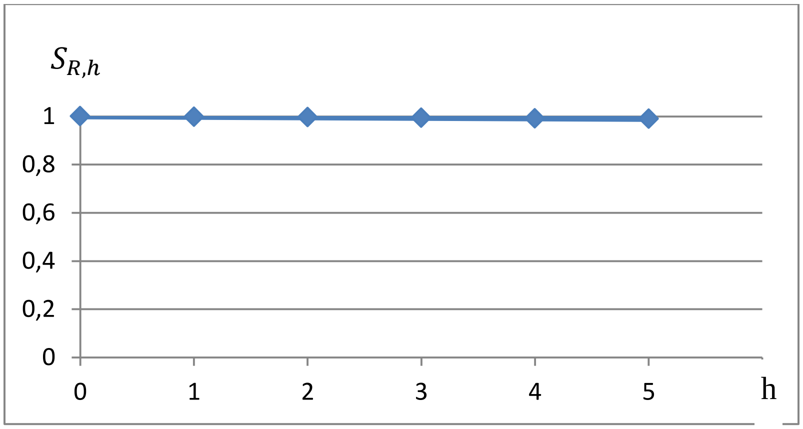

Figure 1a: Variance ratio for case 1 for *h*=0, 1, …,5.

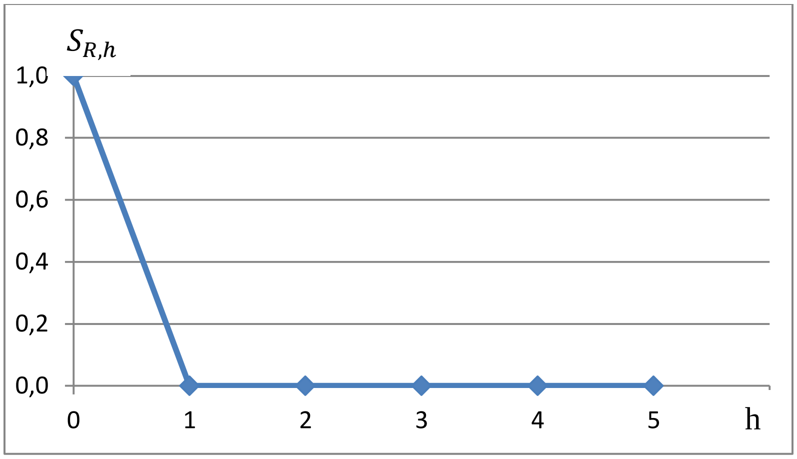

Figure 1b: Variance ratio for case 2 for *h*=0, 1, …,5.

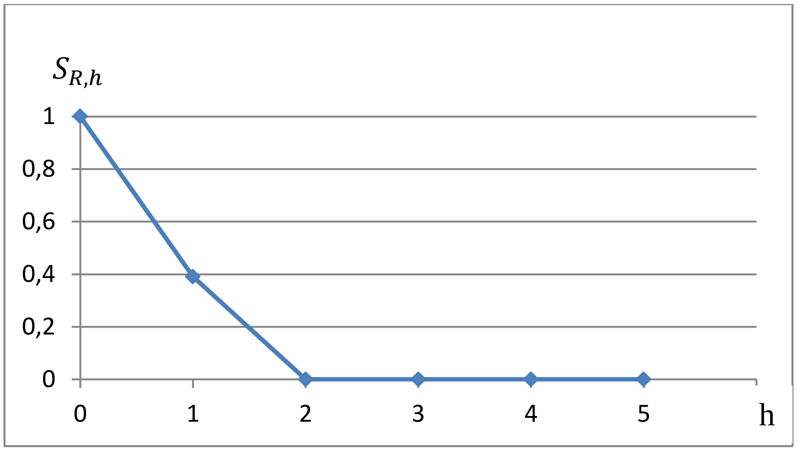

Figure 1c: Variance ratio for case 3 for *h*=0, 1, …,5.

From Figure 1c we see that the variance ratio with *h* = 2 is close to zero and this corresponds to the dominance rule (2,90). That is, both rules indicate sensitivity.

Figures 2a, 2b, and 2c show the variance ratio for *h*=0, 1,…,5 with both positive and negative values. The following cases are considered:

I. Simulation, Normal(1000,1000), 1000 observations, no extreme value
II. Simulation, Normal(1000,1000), 1000 observations, one extreme value, the largest observation contribute with 50% to the total
III. Simulation Normal(1000,1000), 1000 observations, two extreme values, the largest observation contribute with 50% and the two largest contribute with 90% to the total

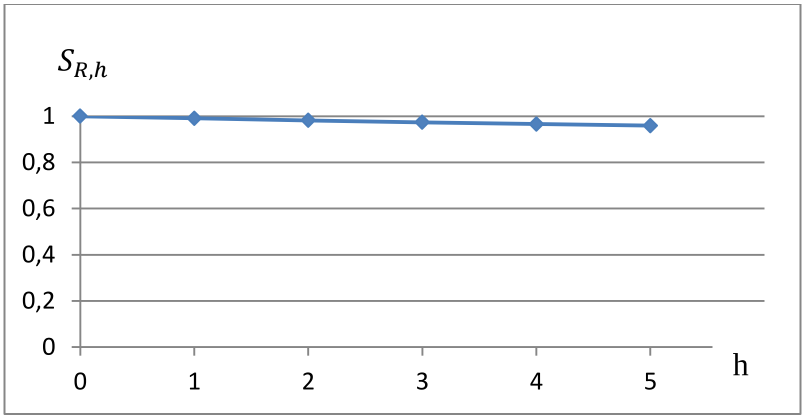

Figure 2a: Variance ratio for case I for *h*=0, 1, …,5.

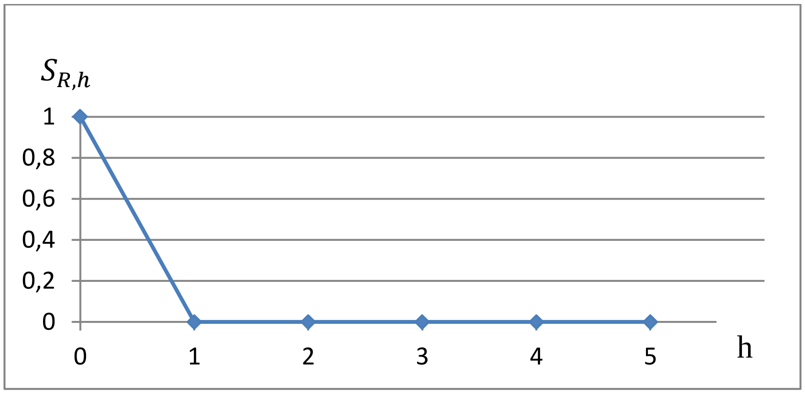

Figure 2b: Variance ratio for case II for *h*=0, 1, …,5

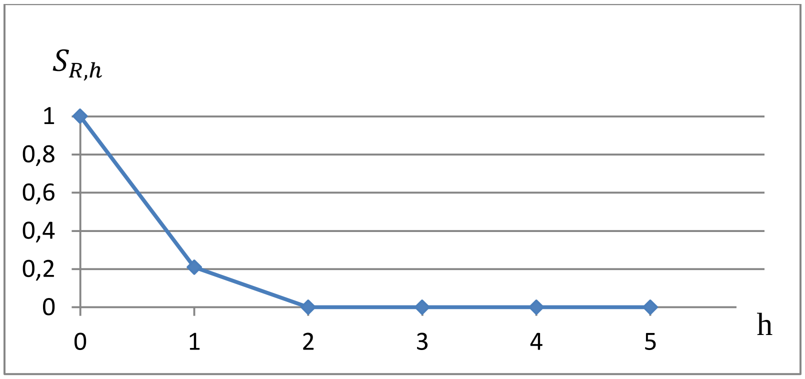

Figure 2c: Variance ratio for case III for *h*=0, 1, …,5.

From the simulations, we conclude that when all values are nonnegative, the variance ratio rule performs similar to the dominance rule, but in contrast to the dominance rule, the variance ratio can be calculated irrespective of the sign of the contributions to the cell.

5. Discussion

We argue that for variables that can take negative values, the standard traditional measures do not always apply. These variables will require special treatment if the negative contributions are not ignorable. It is probably better if sensitivity measures are used on original data and not on transformations, and thus there is a need for alternative measures. We suggest using the dispersion of values in a cell, however the measures discussed in this paper should only be seen as first attempts.

In Figure 3 we give an outline of a possible decision rule when a study variable can take both positive and negative values.

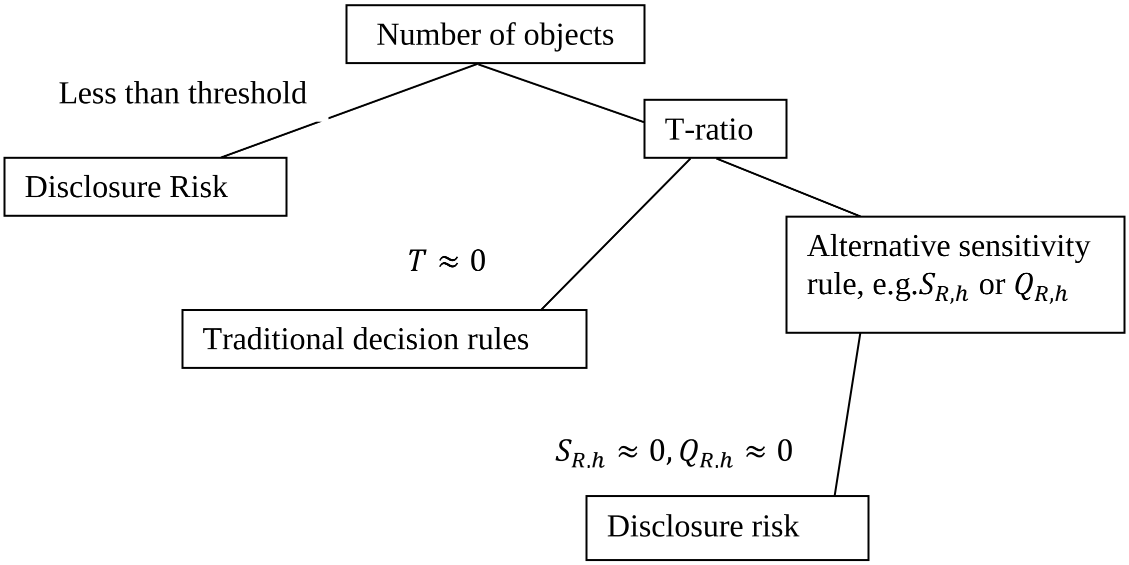
Number of objects
Less than threshold
T-ratio
Disclosure Risk
$T \approx 0$
Alternative sensitivity rule, e.g.$S_{R,h}$ or $Q_{R,h}$
Traditional decision rules
$S_{R.h} \approx 0, Q_{R.h} \approx 0$
Disclosure risk

Figure 3: Outline of an alternative decision rule

Several issues remain to be further investigated. As mentioned above, the variance ratio can take values larger than 1, and then the rule is insufficient. This might happen when we have two or more about equally extreme observations and the other observations are small and few. This will cause the mean value to change drastically when the most extreme observation is excluded, resulting in a larger variance than when the extreme value is included. This will not happen with the quantile rule since it can only take values on [0, 1].

The variance ratio and the quantile ratio will both be zero if all but one of the observations are equal and *h*=1. The most extreme value can however be only slightly larger or smaller than the other observations. In this situation, there is no real risk of disclosure, but the variance rule and the quantile rule will suggest so. Therefore this must be checked. It can be noted that for the special case when all values but one are zero, the dominance rule will also indicate that there is a risk even if the value differing from zero is very small.

When a study variable can take only nonnegative values, it can be shown that it is the largest contributor to the cell value who is most at risk to be disclosed, and that if the largest contributor cannot be disclosed by the second largest contributor, the cell is safe. Our comparisons of the alternative measures with the traditional measures assume that the scenario is the same when negative values are present, but the scenario might be different when the variable under study can take both positive and negative values. For example, Giessing (2008) and Hundepool et al (2012) suggests that such data are less sensitive and that a frequency rule might be enough. In the future, we mean to further investigate if an alternative scenario would be better for data with negative values. A proper sensitivity measure could then be tailored to this alternative scenario and does not have to follow the traditional approaches for measuring sensitivity.

References

Daalmans, J., and de Waal, T. (2010). A general formulation of the secondary cell suppression problem. Discussion paper (10009). Statistics Netherlands, The Hague/Heerlen. http://www.cbs.nl/NR/rdonlyres/993654A5-C5BC-4469-ACC1-6760D5F67AE7/0/201009x10pub.pdf

Domingo-Ferrer, J., and Torra, V. (2002). A Critique of the Sensitivity Rules Usually Employed for Statistical Table Protection. International Journal of Uncertainty, Fuzziness and Knowledge Based Systems. 10: 545-556.

Giessing, S. (2008). Protection of tables with negative values. http://neon.vb.cbs.nl/casc/..%5Ccasc%5CESSnet%5CPosNegReport.pdf

Hundepool, A., and de Wolf, P.-P. (2011). Negative cell values, singletons and linked tables in Tau-Argus. Joint UNECE/Eurostat work session on statistical data confidentiality, Tarragona, Spain, 26-28 October 2011. http://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.46/2011/11_Netherlands.pdf

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Schulte Nordholt, E., Spicer, K., and de Wolf, P.-P. (2012). Statistical Disclosure Control. John Wiley & Sons.

Hundepool, A., Castro, J., Fischetti, M., Giessing, S., Lowthian, P., Ramaswamy, R., Salazar, J.-J., van de Wetering, A. and de Wolf , P.-P. (2011). τ-Argus User´s Manual, Version 3.5. Statistics Netherlands. http://neon.vb.cbs.nl/casc/Software/TauManualV3.5_rev.pdf

Tambay, J.-L., and Fillion, J.-M. (2013). Strategies for Processing Tabular Data Using the G-Confid Cell Suppression Software. Joint Statistical meetings, Montréal, Canada, 3-8 August 2013.