

**WP. 38**  
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and  
ECONOMIC COMMISSION FOR EUROPE  
CONFERENCE OF EUROPEAN STATISTICIANS**

**EUROPEAN COMMISSION  
STATISTICAL OFFICE OF THE  
EUROPEAN COMMUNITIES (EUROSTAT)**

**Joint UNECE/Eurostat work session on statistical data confidentiality**  
(Bilbao, Spain, 2-4 December 2009)

Topic (vi): Case studies

**EXAMPLES OF APPLICATION OF LINEAR SENSITIVITY RULES IN  
KOREAN STATISTICAL DATA**

**Supporting Paper**

Prepared by Kim, KyungMi (Statistics Korea), Republic of Korea

# Example of application of linear sensitivity rules in Korean statistical data

Kim, KyungMi\*

\* Statistical Research Institute, Statistics Korea, Narakeyum 282-1, Wolpyeong, Seo-gu, Daejeon, Korea, 27kyung@korea.kr

**Abstract:** Research on suitable methods for guaranteeing the safety of private information while ensuring the diversity and detail of collected data is both necessary and inevitable. This paper focuses on disclosure control methods for macrodata: specifically, on ways of identifying sensitive cells. This paper introduces the methodologies currently being used to locate sensitive cells, examines their feasibility through a case study involving the macrodata for Statistics Korea's survey on industrial structure.

## 1 Introduction

An issue that is becoming prominent alongside the question of collecting diverse and highly utilizable data is disclosure control on the part of individual respondents or responding organizations. Individuals and businesses are becoming increasingly averse to providing information for reasons of privacy protection, while the rendering of various types of information into databases heightens the risk of exposure during the process of compiling or analyzing statistical data. Therefore, research on suitable methods for guaranteeing the safety of private information while ensuring the diversity and detail of collected data is both necessary and inevitable.

Studies on disclosure control methods are actively being conducted at present, especially in nations with advanced statistical technologies, and the results of these studies are being applied via various methodological approaches to the safeguarding of provided data. Similar research is also underway at the Statistical Research Institute (SRI), under Statistics Korea. The majority of research on disclosure control has focused thus far on the protection of microdata. By contrast, methodologies for the protection of macrodata constitute an area that remains relatively neglected.

Therefore, this paper focuses on disclosure control methods for macrodata: specifically, on ways of identifying sensitive cells. The second paragraph introduces the methodologies currently being used to locate sensitive cells, while the third paragraph examines their feasibility through a case study involving the macrodata for Statistics Korea's survey on industrial structure.

## 2 Methodology

Macrodata can be divided largely into count data and magnitude data. Count data is a form of data whose significance revolves around the number of units, such as the number of establishment. The significance of magnitude data, by contrast, stems from matters of scale, such as total revenue, wages, value of shipments, major production cost, etc. The method for determining the sensitivity of cells differs according to the type of data involved, and in the following each method is treated separately.

### 2.1 Count data

In the case of count data, the method for identifying sensitive cells is rather simple. First, one determines a base value (B), and then defines a cell with a value less than B as a sensitive cell. The value of B is set subjectively in context. In Table 1, sensitive cells determined to be vulnerable to exposure when B=3 are signified by shading.

**Table 1** Number of establishments by group of industry and employment size of establishment (KSIC: C31)

Group of industry	C311	C312	C313	C319
Employment size of establishment				
10 ~ 19	263	26	20	38
20 ~ 49	342	25	21	17
50 ~ 99	226	3	10	7
100 ~ 199	154	5	4	0
200 ~ 299	28	0	1	0
300 ~ 499	4	0	0	2
500 ~	11	1	3	3

### 2.2 Magnitude data

Alternately, when the information in question is in the form of magnitude data, one of three representative methods ((n, k)-dominance rule, p% rule, p/q ambiguity rule) can be used to determine which cells are sensitive to disclosure. Cells that constitute macrodata are composed of data sums. Thus, the basic concept behind the aforementioned methods is that if a particular piece of higher data, or several smaller pieces of data, account for a large portion of the value for an entire cell, the exposure

risk for the cell in question becomes greater. Now let us examine the criteria for determining sensitivity in each of the three methods.

### 2.2.1 (n, k) – dominance rule

This method works by defining a cell as sensitive when the upper n pieces of data for the cell account for k% or more of the overall cell value. Expressed as a formula, it is given as shown in Formula (1). In this case, the individual pieces of data comprising the cell in question are arranged in descending order, as shown in Formula (2).

$$S_{(n,k)}(X) = \sum_{i=1}^N x_i - \frac{k}{100-k} \sum_{i=n+1}^N x_i \quad (1)$$

$$x_1 \geq x_2 \geq \dots \geq x_n \geq x_{n+1} \geq \dots \geq x_N \quad (2)$$

This is called “linear sensitivity,” and when this value is greater than zero, the corresponding cell is identified as sensitive, i.e. having a high risk of exposure. The criteria for determining n and k are somewhat subjective, but statistical agencies frequently use 1, 2, or 3 as the value for n. This method applies a relatively more stringent set of criteria than p% rule, which is described below. Or, phrased another way, this method works more sensitively in judging the risk of exposure, identifying a cell as sensitive if it has even the slightest tendency toward exposure. For this reason, this method is frequently used for large-scale data tabulation.

### 2.2.2 p% rule

In this method, sensitivity is determined by the degree to which a higher value can be predicted when one particular value is known. Linear sensitivity for this method is as shown in Formula (3).

$$S_{(p\%)}(X) = x_i - \frac{100}{p} \sum_{i=c+2}^N x_i \quad (3)$$

In this instance, c stands for the size of the group of combined data used to predict the highest response value. Generally, in this method, the value of c is set at 1. In other words, exposure risk is determined by how well the data carrying the largest value can be predicted using the data carrying the second-largest value.

$$S_{(p\%)}(X) = x_1 - \frac{100}{p} \sum_{i=3}^N x_i \quad (4)$$

### 2.2.3 p/q ambiguity rule

This method is a more advanced version of p% rule, developed by Statistics Canada for cases in which individual raw data values can be partially known. Linear sensitivity for this method is expressed as shown in Formula (5).

$$S_{(p/q)}(X) = x_1 - \frac{q}{p} \sum_{i=3}^N x_i \quad (5)$$

When the value of q is 100, this method becomes identical with p% rule. The smaller the value of q, the more sensitive (and, hence, conservative) this method becomes in judging exposure risk.

However, when cell X contains less than three responses, i.e. when a cell is composed of two pieces of data or less, its linear sensitivity is calculated to be larger than zero for all three methods, and the cell is thus uniformly identified as a sensitive cell.

### 2.3 Example

To better illustrate how the various methods define sensitive cells, we may apply each method to the macrodata contained in the box marked in Table (2), to determine whether any of the cells in question may be defined as sensitive.

**Table 2** Statistics by employment size of establishment

Employment size of establishment	Number of Establishments	Number of Workers	Wages & Salaries	Value of Shipments
Sub total	44	61,997	3,698,793	41,071,317
10 ~ 19	5	63	835	5,756
20 ~ 49	8	297	9,133	282,198
50 ~ 99	10	676	18,819	561,235
100 ~ 199	5	755	22,282	527,687

200 ~ 299	4	925	22,039	810,100
300 ~ 499	2	728	23,902	663,067
500 ~	10	58,553	3,601,783	38,221,274

To assess exposure risk, the cells' linear sensitivity must be calculated for each of the three methods. To do this, the microdata composing each cell was obtained and arranged in descending order.

Based on the data in the table, the linear sensitivity for each method was calculated using Formulas (2), (3), and (4) as specified above; the results are displayed in Table (3).

**Table 3** Linear sensitivity of each method

Employment size of establishment / Linear sensitivity	Number of Workers	Wages & Salaries	Value of Shipments
10 ~ 19	<b>63</b>	<b>835</b>	<b>5,756</b>
$S_{(2,80)}(X)$	-69	283	2,216
$S_{20\%}(X)$	-147	-188	-343
$S_{20/50}(X)$	-65	157	1,870
20 ~ 49	<b>297</b>	<b>9,133</b>	<b>282,198</b>
$S_{(2,80)}(X)$	-539	-11,815	212,542
$S_{20\%}(X)$	-1,001	-23,868	162,930
$S_{20/50}(X)$	-479	-10,776	206,465
50 ~ 99	<b>676</b>	<b>18,819</b>	<b>561,235</b>
$S_{(2,80)}(X)$	-1,280	-28,537	-323,833
$S_{20\%}(X)$	-2,351	-55,596	-893,912
$S_{20/50}(X)$	-1,129	-25,999	-340,745
100 ~ 199	<b>755</b>	<b>22,282</b>	<b>527,687</b>
$S_{(2,80)}(X)$	-857	-21,358	101,183
$S_{20\%}(X)$	-1,825	-48,694	-219,681
$S_{20/50}(X)$	-818	-21,419	468,84

200 ~ 299	<b>925</b>	<b>22,039</b>	<b>810,100</b>
$S_{(2,80)}(X)$	-899	-6,417	194,348
$S_{20\%}(X)$	-2,045	-26,379	-429,000
$S_{20/50}(X)$	-905	-8,594	-44,155
300 ~ 499	<b>728</b>	<b>23,902</b>	<b>663,067</b>
$S_{(2,80)}(X)$	728	23,902	663,067
$S_{20\%}(X)$	379	13,688	342,086
$S_{20/50}(X)$	379	13,688	342,086
500 ~	<b>58,553</b>	<b>3,601,783</b>	<b>38,221,274</b>
$S_{(2,80)}(X)$	-29,355	-1,573,641	-21,383,866
$S_{20\%}(X)$	-84,609	-4,843,424	-59,022,681
$S_{20/50}(X)$	-29,667	-1,608,784	-21,769,469

When the linear sensitivity of a particular cell is greater than zero, that cell is defined as a sensitive cell. When this principle is applied to each method, the results are as shown in Tables (4) through (6).

**Table 4** Result of (2,80) – dominance rule

Employment size of establishment	Number of Establishments	Number of Workers	Wages & Salaries	Value of Shipments
Sub total	44	61,997	3,698,793	41,071,317
10 ~ 19	5	63	835	5,756
20 ~ 49	8	297	9,133	282,198
50 ~ 99	10	676	18,819	561,235
100 ~ 199	5	755	22,282	527,687
200 ~ 299	4	925	22,039	810,100
300 ~ 499	2	728	23,902	663,067
500 ~	10	58,553	3,601,783	38,221,274

**Table 5** Result of 20% rule

Employment size of establishment	Number of Establishments	Number of Workers	Wages & Salaries	Value of Shipments
Sub total	44	61,997	3,698,793	41,071,317
10 ~ 19	5	63	835	5,756
20 ~ 49	8	297	9,133	282,198
50 ~ 99	10	676	18,819	561,235
100 ~ 199	5	755	22,282	527,687
200 ~ 299	4	925	22,039	810,100
300 ~ 499	2	728	23,902	663,067
500 ~	10	58,553	3,601,783	38,221,274

**Table 6** Result of 20/50 ambiguity rule

Employment size of establishment	Number of Establishments	Number of Workers	Wages & Salaries	Value of Shipments
Sub total	44	61,997	3,698,793	41,071,317
10 ~ 19	5	63	835	5,756
20 ~ 49	8	297	9,133	282,198
50 ~ 99	10	676	18,819	561,235
100 ~ 199	5	755	22,282	527,687
200 ~ 299	4	925	22,039	810,100
300 ~ 499	2	728	23,902	663,067
500 ~	10	58,553	3,601,783	38,221,274

In each table, the shaded cells are those determined to be sensitive. The number of sensitive cells is eight for (2,80)-rule, four for 20% rule, and seven for 20/50 rule. As indicated by these results, method A has the most conservative criteria for identifying sensitive cells, while 20% rule provides data that has the most utility from a user's perspective. The tables also show that any cell composed of just two pieces of data is determined to be sensitive for all three methods; this is owing to the nature of the formulas used to derive linear sensitivity, as noted above. Of course, when the data in question is zero, it is treated as an exception, because a data of zero has great significance in itself.

### **3 Korean Status, Example of application**

To apply the abovementioned criteria and methods for identifying sensitive cells to actual data, we use the macrodata from the Mining and Manufacturing Statistics Survey, one of the representative statistical surveys on Korea's industrial sector, conducted by Statistics Korea.

#### **3.1 Target of analysis**

At present, Statistics Korea compiles 10 different sets of statistics regarding Korea's industrial structure; the macrodata corresponding to the various statistical surveys are provided in virtually identical formats. For the purposes of this paper, the macrodata from the Mining and Manufacturing Statistics Survey was used. The Mining and Manufacturing Statistics Survey is conducted annually among establishments with a staff of five or more in Sectors B and C of the Korean Standard Industrial Classification (KSIC). The publicized results are organized as follows: a) by industry, and b) by industry x number of workers. The main indexes publicized include the number of establishments and the monthly average number of workers. In this survey, Sector B comprises four Divisions, while Sector C comprises 24. For the purposes of this paper, one of these Divisions was selected for analysis. The selection was based on the relative importance of the key indexes and the distribution of the Groups, Classes, and Subclasses. Among the Divisions for Sectors B and C, those showing 10 or more workers were examined in terms of their key indexes and the number of subordinate classifications. Based on this examination, Division C26 was chosen for analysis, because its key indexes, including the total value of shipments, accounted for the largest portion (17%) of the entire mining and manufacturing industry, and it had a relatively even distribution of subordinate classifications (6 Groups, 12 Classes, 27 Subclasses).

#### **3.2 Status**

Currently, Statistics Korea uses the number of establishments, which is a type of count data, to determine the exposure risk of the other indexes when compiling macrodata for the Mining and Manufacturing Statistics Survey. The criterion used in this case is 2: that is, a cell is defined as sensitive if the number of establishments that make up the cell data is two or less. Table (7) is a portion of the survey report showing the key indexes by Subclass and by workforce size, and represents the key index macrodata by workforce size for industrial classification code C26219. In this table, the parts marked with an X are sensitive cells. As the table shows, when the number of corresponding establishments is two or less, the key indexes for the entire row are identified as sensitive cells. However, determining the exposure risk of the key indexes, which are types of magnitude data, simply by the number of establishments can be considered risky; setting the criterion as two or less further increases the risk of data disclosure.

**Table 7** Korean Status of definition if sensitive cells

2. 산업세분류 및 종사자규모별 주요지표(10인 이상) Statistics by Sub-Class of Industry and Employment Size of Establishment(10 or More Workers)								
단위: 개, 명, 백만원			In each, person, million won					
부 호 Code	종 사 자 규 모 Employment Size of Establishment	사업체수 Number of Establish- ments	일평균 종사자수 Number of Workers	급여액 (퇴직금제 외) Wages & Salaries	출하액 Value of Shipments	주요생산비 Major Production Cost	부가가치 Census Value-added	유형자산연 말 (인출액 제외) Value of Tangible Assets at end
C26219	물류서비스 및 기타 운반 서비스업의 지원업	25	8 269	245 302	4 077 243	2 892 594	1 037 518	1 899 295
	10 ~ 19	5	58	1 294	19 551	15 731	3 502	4 956
	20 ~ 49	9	280	7 784	31 428	20 173	11 860	52 183
	50 ~ 99	5	322	8 901	188 105	100 087	87 666	117 311
	100 ~ 199	1	X	X	X	X	X	X
	200 ~ 299	2	X	X	X	X	X	X
	300 ~ 499	-	-	-	-	-	-	-
	500명 이상	3	6 951	207 157	3 627 448	2 598 311	984 288	1 843 497

### 3.3 Example of application

Thus, all three of the abovementioned methods for identifying sensitive cells of magnitude data may be applied to Division C26 to determine their relative feasibility. Among the key indexes, major production costs and census value added have a high correlation to the value of shipments, and thus the analysis was limited to the following four variables: number of establishments, monthly average number of workers, wages and salaries, and the value of shipments. To identify sensitive cells, (n,k) dominance rule, p% rule, and p/q ambiguity rule were applied using general criteria and the linear sensitivity of the cells was calculated for each method. Table (8) shows C2651 and C2652, which are part of the results derived using p% rule. As can be seen, several cells with just three establishments are actually identified as non-sensitive. This indicates that the amount of information provided can be greater in comparison to the method used when the number of establishments is three or less, or five or less. Therefore, rather than relying on the number of establishments to determine sensitivity, it would be more desirable to assess the degree of exposure risk by calculating the cells' linear sensitivity.

**Table 8** Result of 20% rule (C2651, C2652)

Class of Industry / Employment size of establishment	Number of Establishments	Number of Workers	Wages & Salaries	Value of Shipments
C2651	211	32,197	1,171,125	6,967,940
10 ~ 19	93	1,244	23,201	223,097

20 ~ 49	65	1,937	39,869	342,562
50 ~ 99	27	1,819	38,253	337,162
100 ~ 199	19	2,678	73,077	694,342
200 ~ 299	3	687	17,778	92,221
300 ~ 499	1	X	X	X
500 ~	3	23,385	958,352	4,879,559
C2652	280	12,968	358,376	2,551,366
10 ~ 19	150	1,985	37,045	259,687
20 ~ 49	84	2,348	48,020	388,565
50 ~ 99	22	1,448	32,937	315,612
100 ~ 199	14	1,845	42,719	280,306
200 ~ 299	3	699	19,270	77,555
300 ~ 499	4	1,340	37,111	139,601
500 ~	3	3,303	141,274	1,090,040

As already noted in conjunction with the example above, all three methods judge a cell to be sensitive when the number of data comprising the cell in question is two or less. Statistics Korea currently defines cells with data from two establishments or less as sensitive cells. The number of additional cells determined to be sensitive by applying the three methods is 38 for (2,80)-dominance rule, 23 for 20% rule, and 40 for 20/50 ambiguity rule. (2,80)-dominance rule and 20/50 ambiguity rule, which are more conservative in their criteria for determining sensitivity, tend to restrict too much data. The amount of data restricted as sensitive is the smallest in the case of 20% rule.

**Table 9** Comparison of the amount of information

	(2,80) rule	20% rule	20/50 rule	Base(number of establishment)	
				Base = 3	Base = 5
Number of Sensitive Cells	38	23	40	63	117

In addition to (2,80)-dominance rule, 20% rule, and 20/50 ambiguity rule, if a modified version of the current method used by Statistics Korea is posited by adjusting the criterion for the number of establishments to three or five, the amount of restricted data for each methodology is as shown in Table (9). These results show that identifying sensitive cells using the sensitivity  $S(X)$  value for (2,80)-dominance rule, 20% rule, and 20/50 ambiguity rule is more effective than adjusting the criterion for the number of establishments, both in terms of a more logical determination of exposure risk and of the greater amount of information thus provided.

#### 4 Conclusion

One of three representative methods ((n, k)-dominance rule, p% rule, p/q ambiguity rule) can be used to determine which cells are sensitive to disclosure. Thus, all three of the abovementioned methods for identifying sensitive cells of magnitude data applied to Division C26 to determine their relative feasibility. In addition to each method, if a modified version of the current method used by Statistics Korea is posited by adjusting the criterion for the number of establishments to three or five, the amount of restricted data for each methodology is as shown. These results show that identifying sensitive cells using the sensitivity  $S(X)$  value for each method is more effective than adjusting the criterion for the number of establishments, both in terms of a more logical determination of exposure risk and of the greater amount of information thus provided.

#### References

- Cox, L.H. (2008). “ Statistical Disclosure Limitation Methods For Tabular Data and Effects on Data Quality” .
- Cox, L.H. (2000). “ Models and Algorithms for Optimizing Cell Suppression in Tabular Data with Linear Constraints,” *Journal of the American statistical Association* 95, 916-928.
- Cox, L.H. (1981). “ Linear Sensitivity Measures in Statistical Disclosure Control,” *Journal of Statistical Planning and Inference* 5, 153-164.
- Cox, L.H. (1980). “ Suppression Methodology and Statistical Disclosure Control,” *Journal of the American statistical Association* 75, 377-385.
- Natalie Shlomo (2007). “ Assessing the Impact of SDC Methods on Census Frequency Tables” .
- Gonzalez, J.F. and L.H. Cox (2005). “ Software for Tabular Data Protection,” *Statistics in Medicine* 24(4), 659-669.
- Christine and Philip Lowthian (2005). “ A process for writing Standards and Guidance for Tabular outputs from ONS” .