

Safeguarding Confidentiality

Contents

1. Introduction
2. Count tables
3. Magnitude tables
4. Microdata
5. Overview of methods used

Note: All examples presented in this document are entirely fictional.

1. Introduction

When publishing data, Statistics New Zealand has an obligation to protect the information of individuals and businesses who have been surveyed. At the same time, the point of collecting the information is to make use of it for statistical purposes, so the aim is to make available as much useful information as possible while maintaining the confidentiality of respondents. This report provides an overview of how data can be changed in the application of confidentiality techniques.

The term **confidentiality** refers to protecting data that is accessed by anyone other than the *publishing agency*. Confidentiality can be defined as "the agreement, explicit or implicit, made between the data subject and the data collector regarding the extent to which access by others to personal information is allowed" (National Research Council and Social Science Research Council, 1993:22). This ranges from protecting output tables that are published, to modifying microdata for access by other government departments or researchers. Confidentiality methods are applied to reduce the risk of disclosures*.

Statistics New Zealand data originates from sample surveys, population censuses and the collection of administrative data. All three produce unit record datasets, otherwise known as microdata datasets. In these, every individual (person, enterprise, event, etc) has one record in the dataset.

Statistics New Zealand releases information for microdata datasets in two distinct ways. The first is by publishing tables; the second is by removing identifiers and granting access to researchers, under strict conditions, to modified versions of the datasets.

The confidentiality techniques applied to tables will depend on the type of data in the table. A table consists of cells defined by categories, containing (usually) aggregated (combined) responses. The two main types of output tables, each with their own specific confidentiality risks and particular modification methods, are:

- Count data tables
- Magnitude data tables.

***Disclosure** Recognition of confidential information.

2. Count Tables

Outline

Count tables contain counts of the individual records that possess certain properties.

Examples of count tables could include:

- the number of people who fall into a series of age groups, sorted by the region they live in
- the number of businesses in each industry, sorted by the number of employees they have.

Table 1 is an example of a count table giving income ranges for various age groups.

Table 1 **Age by Income Bracket**
People aged 15 years and older from town X

Age (years)	Income			
	Low	Medium	High	Total
15 – 29	0	3	0	3
30 – 39	1	0	1	2
40 – 49	1	0	8	9
50 – 59	3	2	2	7
60+	0	4	0	4
Total	5	9	11	25

Detection of sensitive cells in a count table

In count tables, small frequencies like ones and twos are a problem because they could disclose information about particular respondents. Zeros can also disclose information, as they show that no-one in the corresponding row or column has the corresponding attribute. Consider, for example, the 15–29 year age group, where everyone responded to being in the medium income range - this discloses that anyone who is between the ages of 15 and 29 has not got a high income, or a low income.

Confidentiality applied to count tables

Small values in count tables are generally protected by:

- collapsing/aggregating table rows or columns
- modifying the cell values – random rounding
- suppression of cells or tables

Census tables are a very special case of count tables as they are based on the entire population, and therefore sampling does not provide any protection. Suppression of small area data and sparsely populated tables is also used to protect census data.

Data may be collected from populations, or samples of populations. Both situations can produce count tables, but their confidentiality needs differ. For populations, small counts (0, 1, 2) are a disclosure risk and need to be dealt with using the methods listed above. For samples, small counts contain a large amount of sampling error. They are less risky, and are often suppressed for quality reasons. Counts from

samples are usually 'weighted up' to represent population numbers, and this means that small counts are harder to associate with individuals.

Collapsing categories / aggregation

What does this method do?

This is a fairly obvious solution, but one that can be very effective, in the long run, in ensuring a good balance between releasing as much information as possible and restricting the work involved in production of tables. As the name implies, this method involves collapsing two or more groups into one new group representing all the original categories. Aggregation is often used for industries with very few businesses.

Tables 2a and 2b show an example of aggregation.

Table 2a Number of businesses by Turnover and Industry – Karori

Industry	Turnover		
	< \$500,000	\$500,000–\$1,000,000	> \$1,000,000
Carrot farming	1	3	1
Beetroot farming	3	1	2
Corn farming	10	5	3

Table 2b Number of businesses by Turnover and Industry – Karori
After aggregation

Industry	Turnover		
	< \$500,000	\$500,000–\$1,000,000	> \$1,000,000
Carrot & Beetroot farming	4	4	3
Corn farming	10	5	3

How does this affect the final output data?

Aggregation lowers the amount of detail in the final output data. This may affect the researchers' ability to get detailed data for sparsely populated areas.

Random rounding

What does this method do?

Counts are modified using random rounding to protect the respondents, for example in the census count tables. In the census, random rounding to base three (RR3) is used. Counts are randomly rounded to one of the two nearest multiples of three. Counts that are already multiples of three are left unchanged. For instance, a one will be rounded to either a zero or a three. The procedure is unbiased, so the statistical properties of the table are retained. Counts that are already multiples of three are left unchanged. This is illustrated in Table 3.

Random rounding does add some error to the count values, but this error is only significant for very small counts. However, other non-sampling errors from response or data capture in the data. For this reason, any statistical analysis or conclusions drawn from small values like ones and twos is not very sound, regardless of any rounding.

Subtotals and totals are rounded independently and in just the same way as other cells. This means tables will not be additive* but does ensure that totals are, at most, two away from the original total.

Table 3 **Random Rounding Base 3**

Value in the rounded table	Original value could have been
0	0, 1, 2
3	1, 2, 3, 4, 5
6	4, 5, 6, 7, 8
21	19, 20, 21, 22, 23

Tables 4a and 4b show an example of random rounding.

Table 4a **Before Random Rounding**

Industry	Tax income		
	< \$500,000	> \$500,000	Total
Airports	2	4	6
Seaports	0	2	2
Cable cars	7	1	8

Table 4b **After Random Rounding**

Industry	Tax income		
	< \$500,000	> \$500,000	Total
Airports	3	6	6
Seaports	0	3	3
Cable cars	6	3	9

How does this affect the final output data?

Small numbers can be changed by a large percentage, while the change of larger numbers reflects a smaller proportion of the cell value. For example, a cell with a one changed to a three has been changed by 200 percent, but a cell with 1,001 changed to 1,002 has been changed by only 0.2 percent.

In addition, as tables are no longer additive, percentages may not add up to 100 percent.

For censuses, small counts are very sensitive so random rounding needs to be applied. For surveys, the sampling process gives protection to the population, and so random rounding is not necessary.

Cell suppression

What does this method do?

Primary cell suppression protects sensitive cells by blanking them out. If it is still possible to indirectly work out the content of some sensitive cells after all reasonable collapsing and primary cell suppressions have been made, secondary suppression – the deleting of non-sensitive cells to protect sensitive cells – must be employed. Because of row and column totals existing in tables (referred to as marginals),

deleting the sensitive cell alone will not protect the value. For this reason, secondary suppression is applied to any table with sensitive cells.

Table 5a Industry and Number of Farms

Original table

Industry	Number of farms
Banana farms	23
Ant farms	2
Chocolate farms	17
Cardboard farms	30
Total	72

Table 5b Industry and Number of Farms

After primary suppression

Industry	Number of farms
Banana farms	23
Ant farms	S
Chocolate farms	17
Cardboard farms	30
Total	72

Symbol: S = suppressed/confidential

The sensitive cell is now suppressed, but it is still possible to work out the value of this cell by using the remaining values. For example, $72 - 23 - 17 - 30 = 2$. To remove this sensitivity, secondary suppression is required.

Table 5c Industry and Number of Farms

After secondary suppression

Industry	Number of farms
Banana farms	23
Ant farms	S
Chocolate farms	S
Cardboard farms	30
Total	72

Symbol: S = suppressed/confidential

Now it is not possible to work out the exact numbers in the suppressed cells.

The choice of which cells to secondary-suppress also incorporates judgement about minimising the loss of information. So, for example, there might be preference given to suppressing a cell corresponding to fewer respondents, or a smaller magnitude cell. This judgement would take into account the potential uses of the data.

Census rules

As the census is a full coverage survey and everyone is required to respond, there are some special requirements for output in order to protect individuals' responses in small areas and in sparsely populated tables. The overarching principle is that we are meant to provide information on groups, not individuals, therefore tables in which most of the cells have very small counts (ie 0, 1, 2) and would reveal individuals should be avoided.

The 2006 Census rules can be found via this link:

<http://www.stats.govt.nz/census/2006-census/methodology-papers/confidentiality-rules.htm>

***Additive tables** Interior cells add to row/column totals.

3. Magnitude Tables

Outline

Magnitude tables group members of the population into different cells and then sum up some numerical property across all members of the population that fall into that cell. An example might be a table of the total number of employees across all businesses, by region and industry. The key property is that each business that falls into a cell can contribute a different amount. This is an example of a magnitude table, such as the example in Table 6.

Table 6 **Net Profit Before Tax for Region Y**
By industry and number of employees

Industry	No. of employees					Total Profit \$(000)
	0–9	10–19	20–49	50–99	100+	
Industry A	1	23	0	0	120	144
Industry B	0	340	5	0	0	345
Industry C	1	1	45	1	0	48
Industry D	0	4	12	0	150	166
Total	2	368	62	1	270	703

There are two sorts of magnitude tables: numerical measures such as turnover, and numerical counts such as number of sheep.

It is not always easy to distinguish the difference between count and magnitude tables. For example, consider a table counting the total number of sheep across all farms, broken down by region. It might seem that because you are counting sheep, this is a count table. However, the unit being sampled is not sheep, it is farms. Because each farm will contribute a different number of sheep to relevant cells, and therefore the sheep are a numerical property of the farms, this is in fact a magnitude table/magnitude count.

A table can also consist of both magnitude and count data.

Example:

Figure 1 **List of Business Type, Employee Count, and Turnover**

Business type	Employee count	Turnover
Individual proprietorship	1	\$5,000
Partnership	5	\$8,000
Partnership	8	\$20,000
Individual proprietorship	0	\$1,000
Partnership	3	\$2,000
Individual proprietorship	2	\$10,000 etc....

A condensed version:

Table 7 **Business Type by Employee Count**

Business Type (classification variable)	Employ		
	Enterprises (count)	Employee Count (magnitude count)	Total Turnover (magnitude value)
Individual Proprietorship	79,407	46,010	\$9,000,010
Partnership	52,607	77,805	\$22,000,000

Detection of sensitive cells in a magnitude table

A dominance rule, known as the (n,k) rule, is used to detect sensitive cells that are dominated by a few large contributors. A cell is said to be sensitive if n entities make up k percent of a cell's value.

Another method to determine sensitive cells that is likely to be used more in the future is a concentration rule, known as the p percent rule. A cell is sensitive if any cell contributor can estimate another contributor's value to within p percent. For instance for a 15 percent rule, no business should be able to estimate a competitor's contribution to within 15 percent.

Sometimes permission is sought from the largest entities to release cells that are sensitive without applying confidentiality methods.

Confidentiality applied to magnitude tables

Magnitude cells in tables that are at risk of disclosure are generally protected by:

- collapsing/aggregating table rows or columns
- suppression of cell contents
- modifying the cell values using rounding.

Collapsing categories / aggregation

The same techniques are used as for collapsing cells in count tables. See the previous explanation for collapsing categories for count tables.

Cell suppression

What does this method do?

Suppression means blanking out sensitive cells and suppressing other values so that the sensitive values cannot be recalculated. Tables 8a, 8b, and 8c below demonstrate cell suppression.

Table 8a Original Table

Industry	Tax income				Total
	< \$0.25m	\$0.25m–\$0.5m	\$0.5m–\$1m	> \$1m	
Banana farms	8	5	7	3	23
Ant farms	5	6	3	1	15
Chocolate farms	4	5	5	3	17
Cardboard farms	12	10	4	4	30
Total	29	26	19	11	85

There is a single ant farm with more than \$1 million in tax income. This cell is blanked out with 'S' (for 'suppressed').

Table 8b Cell Suppression on Ant Farms with Tax Income Greater Than \$1 Million

Industry	Tax income				Total
	< \$0.25m	\$0.25m–\$0.5m	\$0.5m–\$1m	> \$1m	
Banana farms	8	5	7	3	23
Ant farms	5	6	3	S	15
Chocolate farms	4	5	5	3	17
Cardboard farms	12	10	4	4	30
Total	29	26	19	11	85

Symbol: S = suppressed/confidential

However, it's quite easy to work out the value of the suppressed cell using the remaining values (eg $15 - 5 - 6 - 3 = 1$). Some more cells are now suppressed so that the suppressed values cannot be worked out.

Table 8c Complete Cell Suppression to Prevent Working out the Suppressed Values

Industry	Tax income				Total
	< \$0.25m	\$0.25m–\$0.5m	\$0.5m–\$1m	> \$1m	
Banana farms	8	5	7	3	23
Ant farms	5	6	S	S	15
Chocolate farms	4	5	S	S	17
Cardboard farms	12	10	4	4	30
Total	29	26	19	11	85

Symbol: S = suppressed

How does this affect the final output data?

You will not be able to see the values in some cells and you cannot work out what the sensitive value is. This may reduce the usefulness of the data.

Graduated random rounding

What does this method do?

This method is similar to random rounding to base 3, which is used on count tables. With graduated random rounding (GRR), the base increases to ensure the relative protection offered does not diminish as the number in the cell rises.

In magnitude data, the size of the cell generally has no bearing on how disclosive* it is. A count of one may not be disclosive at all, while a count of 1,000 may be extremely sensitive. GRR is intended to ensure that protection from rounding stays constant, in an approximate way, as the cell size increases.

A small number is rounded to a different base than a large number is rounded to.

Statistics New Zealand uses the following GRR standard:

Table 9 **Statistics New Zealand Graduated Random Rounding Standard**

Original cell size	Rounding base
0 - 19	3
20 - 99	5
100 - 1,000	10
1,000 - 10,000	100 etc

How does this affect the final output data?

There is a slight reduction in the accuracy of the reported data. Because subtotals and totals are rounded independently and in just the same way as other cells, they may cease to be additive.

***Disclosive** Risks individuals information being released to the public.

4. Microdata

Outline

Statistics New Zealand treats microdata (unit record data) datasets with extreme care and allows access only under specific conditions that adhere to the requirements of the Statistics Act 1975. There are currently two main sources of access for researchers: the Data Laboratory and Confidentialised Unit Record Files. Any requested access to microdata undergoes a rigorous application process to ensure no other alternatives for the research exist.

All microdata is anonymised by removing all direct identifiers. A range of additional confidentiality techniques are applied to datasets, when they are needed, to ensure the confidentiality of respondents.

Data laboratory (datalab)

Three secure datalabs are located in Statistics New Zealand offices in Wellington, Auckland and Christchurch. These are tightly controlled environments where researchers can perform analysis on datasets they have been allowed to access. Any physical or electronic material going into or out of the datalab is reviewed by the organisation. Any output produced is assessed against the normal Statistics New Zealand output rules for that dataset and is released from the datalab only if no confidentiality risk is posed.

Researchers must follow a strict process to enable them access to microdata in one of the three datalab offices. A proposal specifying the intended research, the methods of analysis, the outputs, and the variables required is assessed against the

Microdata Access Protocols, with a range of sections being asked for input, before a final decision is made by the Government Statistician. All researchers are required to sign a declaration of secrecy, as specified in the Statistics Act 1975, before being given access to any microdata.

The Statistics Act 1975 allows other government departments to use microdata in their own secure environment, but Statistics New Zealand must be satisfied that this environment offers the same level of security as the datalabs.

Confidentialised Unit Record File (CURF)

Since 2004, Statistics New Zealand has been producing CURFs. A CURF is a heavily confidentialised version of a microdata dataset that is provided to a researcher on a compact disc. Although CURFs have been significantly modified to protect respondents, researchers must still apply for access to any CURF, to ensure their use is monitored.

Some of the techniques employed in the production of CURFs involve removing variables, collapsing the categories of remaining variables, removing unusual records and swapping the values of variables between remaining records.

Currently, CURFs are available for the Income Supplement in 2003 and 2004. Other CURFs being considered are for the Household Savings Survey and one based on a 2–3 percent sample of the census dataset.

The conditions of use by the researcher are:

- A CURF can only be used for bona fide statistical purposes.
- Statistics New Zealand must be assured that the researcher and his/her institution has a proven track record in keeping data secure and complying with agreements.
- CURFs will be available for public, commercial, and overseas use.

CURFs are intended to become a key teaching tool for universities, allowing students access to actual data, as well as researchers who are not able to access the datalab.

Future methods

Additional methods are currently being developed for use in facilitating quality research on microdata. A pilot of the Remote Access Data Laboratory (RADL) is under way which, rather than giving a researcher direct access to the microdata, would allow a researcher to submit code via a computer-based portal. If the output of this code passed a series of checks, the researcher would then be able to access the output via the same portal. This process has been successfully used by the Australian Bureau of Statistics.

Another option is to use synthetic datasets. Synthetic datasets involve microdata that has been statistically generated based on the structure and properties of the original microdata, but containing very little or none of the actual records. There are a range of methods possible for developing this sort of confidentialised microdata.

As with existing methods, access to any form of microdata would still require a stringent application process before access could be obtained.

5. Overview of methods used

The current methods used at Statistics New Zealand are summarised below. For more information on confidentiality, please contact Mike.Camden@stats.govt.nz.

Detection methods

- (n,k) rule – used on magnitude data.
- p percent rule – used on magnitude data.

Data reduction methods

- Collapsing – used on magnitude and sometimes count data.
- Cell suppression – used on magnitude data (and in rare cases count data).

Data perturbation methods

- RR – used on count data.
- GRR – used on magnitude data.