Statistics New Zealand
TATAURANGA AOTEAROA

Microdata output guide

Fourth edition

New Zealand Government

Citation
Statistics New Zealand (2016). *Microdata output guide (Fourth edition)*. Available from www.stats.govt.nz.

ISBN 978-0-908350-68-1 (online)

This guide replaces *Microdata access output guide (Third edition)*, published in August 2015 by Statistics New Zealand.

Published in September 2016 by
Statistics New Zealand
Tatauranga Aotearoa
Wellington, New Zealand

Contact
Please direct any queries or feedback about this guide to: access2microdata@stats.govt.nz.

Contents

1 Purpose of this guide

Microdata output guide describes methods and rules that researchers must use for confidentialising output produced from Statistics New Zealand's microdata. If you follow these rules, your output has a high chance of being released.

Note that these rules do not cover every eventuality, and should be applied as 'rules-of-thumb'.

If you produce an output that violates a confidentiality rule, or is not covered by any of the rules in this guide, you will need to explain why the output contains no disclosure risks. Submit your explanation using the output submission form on the virtual machine desktop.

See chapter 8: Checking microdata output for release for instructions on how to get your output released.

If you are producing phase 2 output, you need to include the appropriate disclaimer.

See chapter 7: Disclaimers for phase 2 microdata output.

This guide contains output rules for all microdata datasets except the Census of Population and Dwellings and longitudinal census.

See 2013 Census confidentiality rules and how they are applied for output rules relating to the Census of Population and Dwellings. For longitudinal census output rules, contact the microdata access team.

If you have any questions about your output, email the microdata access team at access2microdata@stats.govt.nz or phone +64 4 931 4253.

2 Principles of confidentiality for microdata access

Researchers who access our microdata must comply with output rules designed to maintain the confidentiality of information entrusted to us. This chapter covers:

- microdata access for researchers
- why confidentiality is important
- confidentiality, privacy, and security
- goals for confidentiality
- legislative requirements for protecting information.

Microdata access for researchers

We provide approved researchers with access to anonymised unit record datasets, which are known as microdata. Microdata contains information about specific people, households, and businesses.

These datasets are rich sources of information, which allow you to perform advanced statistical analysis and answer complex questions. We treat microdata datasets with extreme care, and only allow access to the data under specific conditions that meet the requirements of section 37C of the Statistics Act 1975.

You can access data from social and business surveys collected by Statistics NZ, as well as administrative data. These datasets contain information about people, households, businesses, education providers, and other entities.

Examples of datasets available to you include:

- Census of Population and Dwellings, and longitudinal census
- Integrated Data Infrastructure
- Household Economic Survey
- Survey of Working Life
- New Zealand General Social Survey
- Disability Survey
- Te Kupenga.

The Integrated Data Infrastructure (IDI) is one of the more commonly used microdata datasets. The IDI is a linked longitudinal dataset comprising a series of datasets from different source agencies. You can produce statistical outputs on the pathways, transitions, and outcomes of people.

See Integrated Data Infrastructure for more information.

Our microdata access service also includes confidentialised unit record files (CURFs).

See Confidentialised unit record files for more information, as CURFS are not covered in this guide.

Why confidentiality is important

We collect a diverse range of information to produce official statistics. Much of the data we collect and use is about individuals, households, and businesses, and contains personal and commercially sensitive information.

We rely on people's trust and goodwill to continue supplying us with high quality information, so we can produce the data that New Zealand needs to grow and prosper. Therefore, maintaining privacy, confidentiality, and data security is one of our core values as leader of the Official Statistics System.

Confidentiality, privacy, and security

We are committed to ensuring the privacy, security, and confidentiality of all our information. This includes the collection, use, storage, and distribution of the information we collect from and about individuals, households, businesses, and administrative sources.

The terms privacy, security, and confidentiality are often used interchangeably, but each term has a different meaning:

- **Privacy** refers to the ability of a person to control the availability of information about themselves.
- **Security** refers to how the agency stores and controls access to the data it holds.
- **Confidentiality** refers to the protection of information from, and about, individuals and organisations, and ensuring that the information is not made available or disclosed to unauthorised individuals or entities.

To protect confidential information, we have policies and protocols to control statistical disclosure. This protection is applied when we process data and publishing outputs. It also extends to our microdata access service, and the output researchers produce from microdata.

Goals for confidentiality

We operate a risk management framework for the microdata access service, which protects against the disclosure of confidential information. The output rules and checking processes covered in this guide are part of the risk management framework.

A disclosure may occur when a person recognises or learns something they did not already know about an individual or organisation, through microdata or output produced from microdata. For a disclosure to occur, this information must enter the public domain. The Statistics Act 1975 states that this type of disclosure must be prevented.

Microdata output rules are designed to meet the following four goals:

- **Utility** – we want the research output to be as rich, detailed, and unmodified as possible.
- **Safety** – we manage the risk of disclosure of particulars about data subjects, down to the level required by the Statistics Act 1975, ethical obligations, and the preservation of trust.
- **Simplicity** – we want the rules to be as simple to apply and check as possible.
- **Consistency** – we aim to maximise consistency across output produced by different channels (microdata access, Statistics NZ production), and across similar output from different source collections.

The first two goals (utility and safety) lead to rules that aim to release as much detail as possible, and to protect the entities that need to be protected.

We aim to maximise the potential utility of research outputs by ensuring that the unit record data provided through the Statistics NZ microdata access service is as rich and detailed as possible, and by not adding confidentiality perturbation to this data.

Legislative requirements for protecting information

We are required by law to protect the information we collect. These requirements are outlined in the Statistics Act 1975 and the Privacy Act 1993.

Sections 37 and 37C of the Statistics Act 1975 govern confidentiality protection of personal information and access to microdata.

See the Statistics Act 1975, section 37: Security of information on www.legislation.govt.nz.

See the Statistics Act 1975, section 37C: Disclosure of individual schedules for bona fide research or statistical purposes on www.legislation.govt.nz.

3 Summary of microdata output rules

The tables below show the microdata output rules that apply to different types of output produced from Statistics NZ surveys and the Integrated Data Infrastructure. The next two chapters describe the output rules you need to apply to each type of output.

Statistics NZ surveys

Type of statistic	Type of output	Output rule
Descriptive statistics	Unweighted counts	4.1
	Weighted counts	4.2
	Totals and means (value magnitudes)	4.3
	Medians and other quantiles	4.4
	Percentages, proportions, and ratios	4.5
	Maximum/minimum values	4.6
	Aggregation	4.10
	Suppression	4.11
Regression output	Regression models	4.7
Graphical output	Graphs	4.8
Program code	Programming code and logs	4.9

Integrated Data Infrastructure

Type of statistic	Type of output	Output rule
Descriptive statistics	Unweighted counts	5.1
	Weighted counts	5.2
	Count magnitudes	5.3
	Totals and means (value magnitudes)	5.4
	Medians and other quantiles	5.5
	Percentages, proportions, and ratios	5.6
	Measures of spread	5.7
	Maximum/minimum values	5.8
	Aggregation	5.12
	Suppression	5.13
	Underlying entities (eg businesses)	5.14
Output from specific datasets	Suppression under 6	5.15
	Census data	5.16
	Annual Enterprise Survey data	5.17
	Overseas Merchandise Trade data	5.18
Analytical output	Regression models	5.9
Graphical output	Graphs	5.10
Program code	Programming code and logs	5.11

4 Microdata output rules for Statistics NZ surveys

Apply the following rules to output produced from social surveys, apart from those in the Integrated Data Infrastructure (IDI):

- 4.1 Unweighted counts
- 4.2 Weighted counts
- 4.3 Value magnitudes (cell totals and means)
- 4.4 Medians and other quantiles
- 4.5 Percentages, proportions, and ratios
- 4.6 Maximum and minimum values
- 4.7 Regression models
- 4.8 Graphs
- 4.9 Programming code and logs
- 4.10 Aggregation
- 4.11 Suppression

The output rules apply to the following surveys:

- New Zealand General Social Survey
- Survey of Working Life
- Survey of Dynamics and Motivation for Migration
- Childcare Survey
- Disability Survey
- Time Use Survey
- Te Kupenga.

Note: The following surveys are included in the IDI: Household Economic Survey (HES); Survey of Family, Income and Employment (SoFIE); the Household Labour Force Survey (HLFS); Longitudinal Immigration Survey (LisNZ); New Zealand Income Survey (NZIS); and Census 2013. The surveys are also available as 'stand-alone' datasets outside the IDI. You must apply the IDI output rules to output produced from these surveys.

See chapter 5: Microdata output rules for the Integrated Data Infrastructure.

For output produced from 'stand-alone' SoFIE data, project-specific rules may apply. Please email access2microdata@stats.govt.nz for more information.

4.1 Unweighted counts

Unweighted counts refer to the number of observations that possess certain characteristics before any weighting has been applied to the data. You can produce unweighted counts from full-coverage datasets and sample surveys. Unweighted counts from survey data are often requested to assess data quality and the reliability of results.

See Appendix: Random rounding to base 3 for more details and examples.

4.1.1 Randomly round all counts to base 3.

4.1.2 Show all empty cells as zero.

4.1.3 When producing the same count in the same cell, apply the rounding to the count in the same direction, even if it is the same cell but in a different set of output. This check must be done manually. Alternatively, phase 1 output may be released but the researcher must destroy all earlier versions of the output.

See Appendix: Random rounding to base 3 for an example.

4.2 Weighted counts

Weighted counts refer to the number of observations that possess certain characteristics after weighting has been applied to the data. Statistics NZ gives weights to survey respondents to represent the population they characterise, and to allow publication of population estimates.

See Appendix: Weighted counts for more details and examples.

The usual procedure for weighted counts is as follows:

- Suppress below a specified threshold and round to a specified base.
- Suppress all zeros.
- Secondary suppression is not required.

See Appendix: Suppression for more details.

4.2.1 For output produced from the following datasets, suppress weighted counts that are below the described threshold, and conventionally round using the described rounding base:

Survey	Threshold	Base
New Zealand General Social Survey	1,000	1,000
Survey of Working Life	1,000	100
Survey of Dynamics and Motivation for Migration	1,000	100
Childcare Survey	1,000	100
Disability Survey	1,000	1,000
Time Use Survey	1,000	1,000
Te Kupenga	500	500

4.2.2 For output produced from Te Kupenga, apply the following rules:

- Suppress estimates with a relative sampling error (RSE) of 100 percent or greater.
- Identify estimates with an RSE between 30 percent and less than 50 percent with one asterisk (*).
- Identify estimates with an RSE between 50 percent and less than 100 percent with two asterisks (**).

4.2.3 For output produced from the Disability Survey, apply the following rules:

- Estimates with very few contributors are deemed a risk to respondents' confidentiality.
- Estimates based on an estimated population of less than 1,000 are suppressed. This is indicated in tables by an 'S'.

- Estimates with high RSE are suspect in quality. Therefore all estimates with an RSE of 50 percent or greater are suppressed.
- Estimates with an RSE of 30 percent to 49.9 percent should be viewed with caution (indicated in tables by an asterisk *), and an error of 50 percent or greater will be indicated by an ‘S’.

4.3 Value magnitudes (cell totals and means)

Value magnitudes refer to measures (cell totals and means) from a numerical variable, which is usually a financial variable. For example, average income by age group and sex in Wellington for 2015, total hours spent attending a cinema by sex, and labour force status in Christchurch between June and August.

4.3.1 Suppress cell totals and means if the unrounded count is less than 5.

4.4 Medians and other quantiles

4.4.1 Suppress medians if the unrounded cell count is less than 10. For other quantiles, use the table below to find how many observations are needed for each quantile. Where a median or quantile is equal to the minimum or maximum value, apply rule 4.6.

Quantile	Number of observations needed overall
0.01	500
0.05	100
0.10	50
0.25	20
0.50	10
0.75	20
0.90	50
0.95	100
0.99	500

4.5 Percentages, proportions, and ratios

4.5.1 You may calculate percentages, proportions, and ratios using the unrounded counts. However, suppress percentages, proportions, or ratios where either, or both, of the counts used to calculate the percentage, proportion, or ratio have been suppressed. Round percentages calculated from unweighted counts to 1 decimal place.

4.6 Maximum and minimum values

4.6.1 Suppress maximum and minimum values. Where a maximum or minimum value is not identifying, it may be considered for release.

See Appendix: Maximum and minimum values for more details and examples.

4.7 Regression models

Regression output does not usually have confidentiality issues. However, you need to check that pieces of output are not equivalent to, or based on, small counts.

Output from regression models does not have confidentiality issues, except in the following circumstances:

4.7.1 Regression output may contain counts or lead to the calculation of counts. If this occurs, suppress unweighted counts smaller than 5 (including 0).

4.7.2 Degrees of freedom are sometimes equivalent to unweighted counts. If this occurs, suppress unweighted counts smaller than 5 (including 0).

4.7.3 Classification and regression tree models may produce the equivalent of detailed count tables. If this occurs, suppress unweighted counts smaller than 5 (including 0).

4.7.4 Regression outputs equivalent to other forms of output need to have the relevant rules applied. For example, coefficients produced by ordinary least squares regressions with binary (0/1) right-hand-side variables are equivalent to cell means and, therefore, need to comply with the means rule. In contrast, the inclusion of continuous independent variables in such models negates this requirement, as the coefficients on the binary variables are no longer raw means.

4.8 Graphs

There are four main types of graph:

- Type A: Graphs produced from aggregated data, or tables that have been confidentialised (eg frequency histograms, bar charts of magnitudes).
- Type B: Graphs produced directly from the unit record data, but aggregated in the process by the software (eg frequency histograms, kernel density plots).
- Type C: Graphs produced directly from the unit record data, and displaying unit record values (eg scatterplots, residual plots).
- Type D: Graphs produced from the results of modelling or derivation that use the unit record data (eg regression curves).

You can format graphs in the following ways:

- Static – the graph is simply a picture with no data attached.[1]
- Interactive – can be modified by the software that contains the data.

4.8.1 Release type A graphs – either static or interactive format.

4.8.2 Release type B graphs – static format, only if the graph provides a high level of uncertainty.[2]

[1] When graphs are released in this format, you need to ensure that the points on the graph cannot be recalculated in some way (eg by counting the pixels).

[2] The level of uncertainty is high if the level of uncertainty about the data values is equal to or larger than that in the confidentialised tables. For graph types B and C you do not need to provide the underlying data, but you do need to justify in your output submission form why the graph has a high enough level of uncertainty to be released.

4.8.3 Release type C graphs – static format after further processing, and at the discretion of the output checker. For this type of graph to be released, you need to ensure that individuals cannot be recognised and that values can only be estimated with a high level of uncertainty.[2] Further processing can include, but is not restricted to: cutting off the tails of a distribution, removing outliers, jittering the actual values, and removing or modifying axis values.

4.8.4 Release type D graphs – either static or interactive format, but only if the values shown in the graph cannot be used to find the original data values (ie where the modelling or derivation cannot be reversed to find the original data values for each individual).

4.9 Programming code and logs

4.9.1 Apply rules as you would to any other type of output to programming code and logs. Do not include unit record data or counts in comments; however, it is acceptable to state the general size of a count. For example, the count is ‘too small’ or ‘sufficient for analysis’.

4.10 Aggregation

Aggregation is a method for protecting sensitive cells by collapsing categories.

4.10.1 If a table contains sensitive cells, a method you can use to protect these cells is to aggregate (collapse) those categories. If the produced tables are the same as other tables released by Statistics NZ (eg information release tables, NZ.Stat tables) then the same aggregation must be applied. These published tables are available on the Statistics NZ website.

See Appendix: Aggregation for more details and examples.

4.11 Suppression

Suppression is the removal of a cell’s value when it has been deemed sensitive.

4.11.1 When sensitive cells still occur and no further grouping is appropriate, suppress the cell (remove its value), then suppress other cells to stop the first cell from being determined. This later stage is called secondary suppression.

4.11.2 Secondary suppression is the suppression of other cells or marginal totals in the table so that the suppressed cell cannot be recalculated. There are no universal guidelines for applying secondary suppression, except there must be enough secondary suppression to ensure primary suppressed values cannot be derived. If the produced tables are the same as other tables released by Statistics NZ (eg information release tables, NZ.Stat tables), then the same secondary suppression must be applied. These published tables are available from the Statistics NZ website.

See Appendix: Suppression for more details and examples.

5 Microdata output rules for the Integrated Data Infrastructure

The Integrated Data Infrastructure (IDI) is a linked longitudinal dataset composed of administrative datasets and surveys from Statistics NZ and other agencies. Apply the following rules to output produced from the IDI:

- 5.1 Unweighted counts
- 5.2 Weighted counts
- 5.3 Count magnitudes
- 5.4 Value magnitudes (cell totals and means)
- 5.5 Medians and other quantiles
- 5.6 Percentages, proportions, and ratios
- 5.7 Measures of spread
- 5.8 Maximum and minimum values
- 5.9 Regression models
- 5.10 Graphs
- 5.11 Programming code and logs
- 5.12 Aggregation
- 5.13 Suppression
- 5.14 Output relating to business, education, and other underlying entities
- 5.15 Suppression under 6
- 5.16 Census data
- 5.17 Annual Enterprise Survey (AES) data
- 5.18 Overseas merchandise trade data.

The IDI contains the following Statistics NZ datasets:

- Household Economic Survey (HES)
- Household Labour Force Survey (HLFS)
- New Zealand Income Survey (NZIS)
- Longitudinal Immigration Survey (LisNZ)
- Longitudinal Business Database (LBD)
- Survey of Family, Income, and Employment (SoFIE)
- Census 2013.

Note: The Statistics NZ datasets are also available as ‘stand-alone’ datasets outside the IDI. You must apply the IDI output rules to output produced from these datasets.

For any new datasets that have been linked to the IDI after the June 2016 refresh, please contact access2microdata@stats.govt.nz as dataset-specific rules may apply.

5.1 Unweighted counts

Unweighted counts refer to the number of observations that possess certain characteristics before any weighting has been applied to the data. You can produce unweighted counts from full-coverage datasets and sample surveys. Unweighted counts from survey data are often requested to assess data quality and the reliability of results.

See Appendix: Random rounding to base 3 for more details and examples.

5.1.1 Randomly round all counts to base 3.

5.1.2 Show all empty cells as zero.

5.1.3 When producing the same count in the same cell, apply the rounding to the count in the same direction, even if it is the same cell but in a different set of output. This check must be done manually. Alternatively, phase 1 output may be released but the researcher must destroy all earlier versions of the output.

See Appendix: Random rounding to base 3 for an example.

5.2 Weighted counts

Weighted counts refer to the number of observations that possess certain characteristics after weighting has been applied to the data. Statistics NZ gives weights to survey respondents to represent the population they characterise, and to allow publication of population estimates.

See Appendix: Weighted counts for an example.

The usual procedure for weighted counts is as follows:

- Suppress below a specified threshold and round to a specified base.
- Suppress all zeros.
- Secondary suppression is not required.

See Appendix: Suppression for more details.

5.2.1 For output produced with the LBD datasets, treat weighted firm counts by randomly rounding all counts to base 3 (as per rule 5.1).

5.2.2 For output produced from the following datasets, suppress weighted counts that are below the described threshold and conventionally round using the described rounding base:

Survey	Threshold	Base
Household Economic Survey	3,000	1,000
Household Labour Force Survey	1,000	100
New Zealand Income Survey	1,000	100
Survey of Family, Income and Employment	1,000	100
Longitudinal Immigration Survey	20	20

5.2.3 For output produced from the Household Economic Survey, also apply the following quality rules (due to the small sample size and under-reporting):

- Estimates with a relative sample error in the range 21–50 percent are flagged with a warning that they are unreliable.

- Estimates with a relative sample error of over 50 percent are suppressed.
- Cross tabulation – only estimates with a relative sample error of 20 percent and under are cross-tabulated with another variable.
- Weighted counts with a corresponding unweighted count of less than 5 are suppressed.

5.3 Count magnitudes

Count magnitudes are cell totals of the contributed values of the businesses in the cells. The contributed values come from a numerical variable which is a count of individuals. For example, number of employees in the retail industry in Auckland for 2015, number of sheep in Wellington in 2015.

5.3.1 Apply graduated random rounding to all count magnitudes.

See Appendix: Graduated random rounding for more details and examples.

5.4 Value magnitudes (cell totals and means)

Value magnitudes refer to measures (cell totals and means) from a numerical variable, which is usually a financial variable. For value magnitudes produced from the IDI, apply either rule 5.4.1 or 5.4.2. Which rule to apply depends on whether the tabular output contains information about businesses, or information about individuals or households.

If the tabular output contains information about **businesses**, then the businesses need protection to ensure a business's contribution to a value magnitude cannot be estimated with accuracy. To protect businesses, apply the p% rule.

5.4.1 For tabular output containing business information, apply the p% rule to value magnitudes. If a cell fails this test (ie is deemed to be sensitive), protect the cell by either aggregation or suppression. For means, apply the p% rule to cell totals and calculate means from rounded counts.

See Appendix: The p% rule for more details and examples.

If the tabular output contains information about **individuals or households**, the individuals or households need protection to ensure an individual or household's contribution to a value magnitude cannot be estimated with accuracy. To protect individuals and households, you must calculate value magnitudes from at least 20 observations. For example, average income of PhD graduates in Auckland for 2014, average number of hours worked by migrants in Otago between September and February.

5.4.2 For tabular output containing social information, suppress cell totals and means if the unrounded count is less than 20. Calculate means from rounded counts.

5.4.3 For output containing means of log transformed variables and means of growth rates, suppress means if the unrounded count is less than 10.

5.5 Medians and other quantiles

5.5.1 Suppress medians if the unrounded cell count is less than 10. For other quantiles, use the table below to find how many observations are needed for each quantile. Where a median or quantile is equal to the minimum or maximum value, apply rule 5.8.

Quantile	Number of observations needed overall
0.01	500
0.05	100
0.10	50
0.25	20
0.50	10
0.75	20
0.90	50
0.95	100
0.99	500

5.6 Percentages, proportions, and ratios

5.6.1 Derive all percentages, proportions, and ratios from the rounded counts.

5.7 Measures of spread

5.7.1 Suppress all measures of spread for a cell if the values in the cell have a coefficient of variation of less than 10 percent. Measures of spread include variance, standard deviation, range, and interquartile range.

5.8 Maximum and minimum values

5.8.1 Suppress maximum and minimum values. Where a maximum or minimum value is not identifying, it may be considered for release.

See Appendix: Maximum and minimum values for more details and examples.

5.9 Regression models

Regression output does not usually have confidentiality issues. However, you must ensure that pieces of output are not equivalent to statistics subject to other confidentiality rules, particularly small counts or statistics based on small counts.

Output from regression models does not have confidentiality issues, except in the following circumstances:

5.9.1 Regression output may contain counts or lead to the calculation of counts. If this occurs, suppress unweighted counts smaller than 5 (including 0).

5.9.2 Degrees of freedom are sometimes equivalent to unweighted counts. If this occurs, suppress unweighted counts smaller than 5 (including 0).

5.9.3 Classification and regression tree models may produce the equivalent of detailed count tables. If this occurs, suppress unweighted counts smaller than 5 (including 0).

5.9.4 Regression outputs equivalent to other forms of output need to have the relevant rules applied. For example, coefficients produced by ordinary least squares regressions with binary (0/1) right-hand-side variables are equivalent to cell means and, therefore, need to comply with the means rule. In contrast, the inclusion of continuous independent variables in such models negates this requirement, as the coefficients on the binary variables are no longer raw.

5.10 Graphs

There are four main types of graphs:

- Type A: Graphs produced from aggregated data or tables that have been confidentialised (eg frequency histograms, bar charts of magnitudes).
- Type B: Graphs produced directly from the unit record data but aggregated in the process by the software (eg frequency histograms, kernel density plots).
- Type C: Graphs produced directly from the unit record data and displaying unit record values (eg scatterplots, residual plots).
- Type D: Graphs produced from the results of modelling or derivation that use the unit record data (eg regression curves).

You can format graphs in the following ways:

- Static – the graph is simply a picture with no data attached.[3]
- Interactive – the graph contains the data, and can be modified by software.

5.10.1 Release type A graphs – either static or interactive format.

5.10.2 Release type B graphs – static format, only if the graph provides a high level of uncertainty.[4]

5.10.3 Release type C graphs – static format after further processing, and at the discretion of the output checker. For this type of graph to be released, you need to ensure that individuals cannot be recognised and that values can only be estimated with a high level of uncertainty.[4] Further processing can include, but is not restricted to: cutting off the tails of a distribution, removing outliers, jittering the actual values, and removing or modifying axis values.

5.10.4 Release type D graphs – either static or interactive format, but only if the values shown in the graph cannot be used to find the original data values (ie where the modelling or derivation cannot be reversed to find the original data values for each individual).

5.11 Programming code and logs

5.11.1 Apply rules as you would to any other type of output to programming code and logs. Do not include unit record data or counts in comments; however, it is

[3] When graphs are released in this format, you need to ensure that the points on the graph cannot be recalculated in some way (eg by counting the pixels).

[4] The level of uncertainty is high if the level of uncertainty about the data values is equal to or larger than that in the confidentialised tables. For graph types B and C you do not need to provide the underlying data, but you do need to include a justification in your output submission form explaining why the graph has a high enough level of uncertainty to be released.

acceptable to state the general size of a count. For example, the count is ‘too small’ or ‘sufficient for analysis’.

5.12 Aggregation

Aggregation is a method for protecting sensitive cells by collapsing categories.

5.12.1 If a table contains sensitive cells, a method you can use to protect these cells is to aggregate (collapse) those categories. If the produced tables are the same as other tables released by Statistics NZ (eg information release tables, NZ.Stat tables) then the same aggregation must be applied. These published tables are available on the Statistics NZ website.

See Appendix: Aggregation for more details and examples.

5.13 Suppression

Suppression (or primary suppression) is the removal of a cell’s value when it has been deemed sensitive.

5.13.1 When sensitive cells still occur and no further grouping is appropriate, the procedure is to suppress the cell (remove its value), then suppress other cells to stop the first cell from being determined. This later stage is called secondary suppression.

5.13.2 Secondary suppression is the suppression of other cells or marginal totals in the table so that the suppressed cell cannot be recalculated. There are no universal guidelines for applying secondary suppression, except that there has to be enough secondary suppression to ensure that primary suppressed values cannot be derived. If the produced tables are the same as other tables released by Statistics NZ (eg information release tables, NZ.Stat tables) then the same secondary suppression must be applied. These published tables are available from the Statistics NZ website.

See Appendix: Suppression for more details and examples

5.14 Output relating to business, education, and other underlying entities

5.14.1 For output containing business or justice information, suppress data if the underlying count of entities is fewer than 3. This rule applies to businesses and courts. Use both the permanent business number (PBN) and the enterprise (ENT) variable as employer variables.

5.14.2 For output containing education information, suppress data if the underlying count of entities is fewer than 2. This rule applies to education providers and industry training organisations.

5.14.3 For output containing mental health information, suppress data if the underlying count of entities is fewer than 2.

5.14.4 For output containing youth services information, suppress data if the underlying count of entities is fewer than 2.

5.15 Suppression under 6

5.15.1 For output produced from the following datasets, suppress output if the underlying unrounded count is fewer than 6:

- Ministry of Justice
- Department of Corrections
- Police data (eg Recorded Crime Victims Statistics (RCVS) and Recorded Crime Offender Statistics (RCOS))
- Ministry of Health (including Mental Health)
- Meshblock level (eg counts of individuals, families, and households. Any geographical unit lower than meshblock should be suppressed and/or aggregated to at least meshblock level.)
- Births, Deaths, Marriages, and Civil Unions
- Child, Youth, and Family
- Youth Services
- Housing New Zealand
- Census (see rule 5.16 for additional census rules)
- Auckland City Mission.

5.16 Census data

5.16.1 For output produced from census data, suppress output if the underlying unrounded count is fewer than 6.

5.16.2 All measures have simple conventional rounding applied. Different variables require different level of rounding:

- Measures from annual income are rounded to the nearest $100
- Measures from weekly rent paid are rounded to the nearest $10
- Measures for age are rounded to one decimal place
- Measures from whole number count variables are rounded to one decimal place.

5.17 Annual Enterprise Survey (AES) data

5.17.1 When producing industry-level output using AES data alone, ensure that the aggregation and secondary suppression rules have been followed. In particular, if the produced tables are the same as other tables released by Statistics NZ then the same aggregation or secondary suppression must be applied. You are welcome to email the microdata access team at access2microdata@stats.govt.nz to check whether your planned output will meet this rule.

5.17.2 Suppress output for the following ANZSIC96 industries:

- G5110 – Supermarkets and grocery stores
- G511010 – Supermarkets

Outputs can be released for one (but not both) of the following ANZSIC96 and ANZSIC06 industries subject to the general IDI rules:

- G51 – Food retailing
- G511020 – Groceries and dairies

Suppress output for the following ANZSIC06 industries:

- G411 – Supermarkets and grocery stores

- G411000 – Supermarkets and grocery stores

Outputs can be released for one (but not both) of the following ANZSIC06 industries subject to the general IDI rules:

- G41 – Food retailing
- G412 – Specialised food retailing

This requirement is due to a long-standing confidentiality agreement between Statistics NZ and organisations within the industry.

5.18 Overseas merchandise trade data

The overseas merchandise trade dataset contains confidential items. For these items, an exporter or importer has requested suppression and Statistics NZ has accepted their request. Outputs produced from the IDI that contain overseas merchandise trade data must not lead to the disclosure of these confidential items.

Generally, aggregated totals and counts based on Harmonised System (HS) groupings that include confidential items will not be released, due to the need to protect these items. Outputs of this nature will only be considered for release on a case-by-case, discretionary basis. To be considered for release, you must clearly demonstrate that the output does not risk disclosure of the confidential item(s).

See Trade confidentiality for the list of confidential export and import items.

6 Guidance for sharing microdata output

This chapter provides guidance on situations when it is acceptable, or not, to share your microdata output. Note: the guidance below is currently under review:

- Phase 1 output
- Phase 2 output.

Phase 1 output

Phase 1 output (eg data contained in a table or regression model) is confidentialised output used for planning further research or writing up findings. You can share the output, but it must not be made public.

You can only share phase 1 output with your research team, and/or those researchers who were listed on the microdata access application.

Further guidance:

- Do not share phase 1 output with ministers.
- Only share phase 1 output in a safe environment, for example a trusted recipient's office. The output should not be shared in public places, or on social media (eg Facebook).
- Consider the confidentiality training you have received, and assess whether the people who you are sharing the output with understand the confidentiality requirements.

Phase 2 output

Phase 2 output is confidentialised output that is usually in the form of a publication, paper, or presentation. This output can be released into the public domain, but must include the appropriate Statistics NZ disclaimer.

See Disclaimers for phase 2 microdata output.

Any phase 2 output that includes data produced from microdata access must be reviewed by Statistics NZ before it is released. The purpose of this review is to ensure that all confidentiality measures have been applied.

If you 're-package' phase 2 output, you do not need to submit the re-packaged output to Statistics NZ for checking unless new data has been added. For example, you have submitted a paper to a statistical conference, which has been checked and released by Statistics NZ. You then prepare a presentation (eg PowerPoint) using the same material (and no new information). In this situation, the presentation material does not need to be checked by Statistics NZ.

7 Disclaimers for phase 2 microdata output

Phase 2 microdata output must include the appropriate disclaimer, depending on whether the output is produced from:

- Statistics NZ surveys
- Integrated Data Infrastructure.

Disclaimer for output produced from Statistics NZ surveys

All phase 2 output produced from Statistics NZ surveys must include an acknowledgement and disclaimer.

Acknowledgement – stating that Statistics NZ is the source for any tables, graphs, or data (supplied by Statistics NZ) that are quoted in the paper or presentation.

Disclaimer – stating that the researcher takes full responsibility for the paper, that Statistics NZ will not be held accountable for any error or inaccurate findings within the paper or presentation, and acknowledgement that access to data is in accordance with the Statistics Act 1975. Use the following wording:

> Access to the data used in this study was provided by Statistics New Zealand under conditions designed to give effect to the security and confidentiality provisions of the Statistics Act 1975. The results presented in this study are the work of the author, not Statistics NZ.

Disclaimer for output produced from the Integrated Data Infrastructure

Phase 2 output produced from the Integrated Data Infrastructure (IDI) must include the following disclaimer. The first four paragraphs must always be used:

> The results in this [report, paper] are not official statistics They have been created for research purposes from the Integrated Data Infrastructure (IDI), managed by Statistics New Zealand.
>
> The opinions, findings, recommendations, and conclusions expressed in this [report, paper etc] are those of the author(s), not Statistics NZ, [Department X, or Organisation Y].
>
> Access to the anonymised data used in this study was provided by Statistics NZ under the security and confidentiality provisions of the Statistics Act 1975. Only people authorised by the Statistics Act 1975 are allowed to see data about a particular person, household, business, or organisation, and the results in this [report, paper] have been confidentialised to protect these groups from identification and to keep their data safe.
>
> Careful consideration has been given to the privacy, security, and confidentiality issues associated with using administrative and survey data in the IDI. Further detail can be found in the Privacy impact assessment for the Integrated Data Infrastructure available from www.stats.govt.nz.

The **additional paragraphs** below can be ignored if they do not apply to the data that you used.

Inland Revenue tax data

You must also include the following paragraphs in your disclaimer if your phase 2 output uses Inland Revenue tax data:

> The results are based in part on tax data supplied by Inland Revenue to Statistics NZ under the Tax Administration Act 1994. This tax data must be used only for statistical purposes, and no individual information may be published or disclosed in any other form, or provided to Inland Revenue for administrative or regulatory purposes.
>
> Any person who has had access to the unit record data has certified that they have been shown, have read, and have understood section 81 of the Tax Administration Act 1994, which relates to secrecy. Any discussion of data limitations or weaknesses is in the context of using the IDI for statistical purposes, and is not related to the data's ability to support Inland Revenue's core operational requirements.

Overseas merchandise trade or government assistance data from the LBD

You must also include the following paragraph in your disclaimer if your phase 2 output uses overseas merchandise trade or government-assistance data from the Longitudinal Business Database (LBD):

> Statistics NZ confidentiality protocols were applied to the data sourced from the {*/^ include only the agencies from which data has been sourced}. Any discussion of data limitations is not related to the data's ability to support these government agencies' core operational requirements.

Symbols:
{*} New Zealand Customs Service [if overseas merchandise trade data is used].
{^} Ministry of Social Development; the Ministry of Business, Innovation and Employment; New Zealand Trade and Enterprise; and Te Puni Kōkiri [if government-assistance data is used].

Publishing on the Statistics NZ website

You must also add the following paragraphs if your phase 2 output will be published on the Statistics NZ website (aligned with standard Statistics NZ disclaimers):

> **Copyright:** This work is licensed under the Creative Commons Attribution 4.0 International licence. You are free to copy, distribute, and adapt the work, as long as you attribute the work to [*insert name of research owner*] and abide by the other licence terms.
>
> **Liability:** While all care and diligence has been used in processing, analysing, and extracting data and information in this publication, [Statistics New Zealand, *insert name of research owner*] gives no warranty it is error free and will not be liable for any loss or damage suffered by the use directly, or indirectly, of the information in this publication.

Short-form disclaimer for briefing, ministerial paper, journal article, conference poster, or presentation

Use the following short-form disclaimer for phase 2 outputs that take the form of a briefing, ministerial paper, journal article, conference poster, or presentation.

The short-form disclaimer can only be used for briefings/ministerial papers that are **not intended for wider release**, or for journal articles **where a prior (full disclaimer) working paper has been published**.

> Access to the anonymised data used in this study was provided by Statistics New Zealand in accordance with security and confidentiality provisions of the Statistics Act 1975[, and secrecy provisions of the Tax Administration Act 1994][A]. The findings are not Official Statistics. The results in this paper are the work of the authors, not Statistics NZ [nor Department X, Organisation Y,…], and have been confidentialised to protect [individuals, households, businesses, and other organisations] from identification. [See {original publication reference} for the full disclaimer.][B]

Criteria for clauses [A–B]:

> A: Include if Inland Revenue data used.
>
> B: Include if this is a reuse of a prior phase 2 release. This citation is mandatory for a journal article, which must be associated with a prior (full disclaimer) working paper to be eligible for short-form disclaimer use.

Any phase 2 output that will be given to your colleagues to write briefings/ministerial papers must include the standard (full) disclaimer for output produced from the IDI. Since Statistics NZ cannot bind downstream users (ie, your colleagues) to always include the appropriate disclaimer, it will depend on your colleagues' judgement and responsibility as users of the statistics as to whether the full or short-form disclaimer appears in the final briefing/ministerial paper.

8 Checking microdata output for release

Follow these steps to get your microdata output checked and released by Statistics NZ.

Use the standard checking process, unless you are an accredited researcher who needs phase 1 output released:

- Standard checking process – phase 1 output
- Standard checking process – phase 2 output
- Accredited researchers – phase 1 output

Standard checking process – phase 1 output

Follow these steps when you access microdata at Statistics NZ premises:

1. Carefully apply the appropriate confidentiality methods to your output.

2. Place your output files in the assigned folder for checking. Put each set of output in a new sub-folder, labelled with the date the folder was created.

3. Select the output submission application from the virtual machine desktop and complete the necessary details. Please provide detailed explanations, and include as much supporting evidence as possible (eg underlying counts) so we can complete the checking process quickly and efficiently.

   You will need to include:

   - all output rules that were applied
   - definitions for all new variables
   - the type and format of each graph.

   Note that large quantities of output, and/or more complex output, may require more than three working days to check. The checker will inform you if this is the case. You may wish to indicate which output files are highest priority, so the checker can release the more urgent files first.

   If you need the output checked urgently, let us know on the submission form and a checker will review your output as quickly as possible.

4. The output checker reviews your phase 1 output, which may take up to three working days to complete.

   The output checker may need to email or phone you to clarify the output confidentialisation techniques that you used, if this is not clear on the output submission form.

   If the output needs modification before we can release it, the output will be withheld and you will be notified. You will then need to modify the output and submit it to be checked again.

5. If there are no confidentiality issues, the checker will release the output to you using the email address that you provided.

Standard checking process – phase 2 output

Follow these steps to get your phase 2 microdata output released:

1. Email any phase 2 output directly to the microdata access team at access2microdata@stats.govt.nz.

2. The microdata access team informs the output checker there is output for checking.

3. The output checker checks your phase 2 output for information that could constitute a disclosure, which may take up to 10 working days to complete. If you require an urgent check, email the microdata access team and a checker will be organised to perform the check as quickly as possible.

4. If a confidentiality issue is found in your output, the checker will negotiate a mutually acceptable solution with you. If there are no confidentiality issues, the checker will notify you that the output is safe for release.

Accredited researchers – phase 1 output

Follow these steps if you are an accredited researcher who needs phase 1 output released.

Please note that an output checker audits a sample of these outputs on a regular basis. If a problem is identified, sanctions may be put in place.

1. Carefully apply the appropriate confidentiality methods to your output.

2. Place the output into the assigned self-release folder. Note that once the file is copied into this folder, the file cannot be edited or removed.

3. Select the output submission application from the virtual machine desktop and complete the necessary details.

4. If your output qualifies for automatic release – meaning it is not produced from a data source that must be checked by Statistics NZ before it can be released (eg the IDI), then the microdata access team is automatically notified about the submission, and will email the output to you at the address you provided.

5. If your output does not qualify for automatic release, a checker is automatically notified that a check is required and the output will be processed as for standard phase 1 output.

Glossary

Aggregation: A method for protecting sensitive cells is to collapse (aggregate) categories.

Anonymised data: Data with direct identifiers (eg name, address, ID number, phone number) removed.

Attribute disclosure: This occurs when confidential information is revealed about an individual or organisation.

Confidential data: Data to be protected from disclosure.

Confidentialised data: Data modified or with suppressions in order to protect individuals' and organisations' information.

Confidentiality: The protection of individuals' and organisations' information, and ensuring that the information is not made available or disclosed to unauthorised individuals or entities.

Conventional rounding to base x: A method that rounds numbers to the nearest multiple of x.

Count magnitudes: Count magnitudes are cell totals of the contributed values of the businesses in the cells. The contributed values are counts. For example, number of employees in the food service industry.

Data Lab: A secure facility that has been established on Statistics NZ premises. It is a place where external researchers can be permitted access to microdata under contractual agreements that cover the maintenance of confidentiality, and that place strict controls on the uses of the data.

Data utility: A property of data products that enables them to meet the information needs of users.

Disclosure: The inappropriate attribution of information to an individual or organisation.

Family: Two or more people living in the same household who are either a couple, with or without children, or one parent and their children. A child in a family can be of any age.

Graduated random rounding (GRR): Random rounding where the size of the added uncertainty (ie the rounding base) increases with the value being rounded. This method can be used for count magnitudes.

Household: A household is either one person, or one or more families, or a family plus other people, or a group of people living together who are not a family.

Identity disclosure: This occurs when an individual or organisation is revealed as a respondent of a data collection.

Microdata: Unit-record level data, or data corresponding to information at the respondent level.

Output checker: A Statistics NZ staff member who checks output for confidentiality issues.

The p% rule: This rule determines whether a cell is sensitive. A cell is considered sensitive if the value for any contributor can be calculated to within p percent.

Perturbation: Disclosure control methods that add uncertainty to data by changing some values.

Privacy: The ability of a person to control the availability of information about themselves.

Random rounding to base x: A method that randomly, and in an unbiased way, rounds values either up or down to the nearest multiple of x. For example, random rounding to base 3 (RR3) rounds values to multiples of 3.

Secure Microdata Access Environment: The physical environment where researchers can be permitted access to microdata under contractual agreements that cover the maintenance of confidentiality, and that place strict controls on the uses of the data. This environment is specified in the Microdata Access Agreement signed between the research institution and Statistics NZ.

Security: This refers to how the agency stores and controls access to the data it holds.

Sensitive cell: A cell for which knowledge of the value would permit an unduly accurate estimate of the contribution of an individual or organisation, or that reveals a small count.

Suppression: The removal of a cell's value when it has been deemed sensitive (also referred to as primary suppression). Secondary suppression is the suppression of other non-sensitive cells or marginal totals in the table so that the (primary) suppressed cell cannot be recalculated.

Threshold rule: A rule that defines a cell as sensitive, based on the number of observations contributing to the cell.

Unique: An individual or organisation that can be distinguished from all other members in the sample (sample unique) or population (population unique), by using a set of identifying variables.

Unweighted counts: Unweighted counts refer to the number of observations that possess certain characteristics before any weighting has been applied to the data. Unweighted counts can be produced from full-coverage datasets and sample surveys. Unweighted counts from survey data are often requested to assess data quality and the reliability of results.

Value magnitudes: Value magnitudes are cell totals or means from a (non-count) numerical variable, which is usually a financial variable. For example, personal income, household expenditure, business revenue (or income), hours worked, etc.

Weighted counts: Weighted counts refer to the number of observations that possess certain characteristics after weighting has been applied to the data. Statistics NZ gives weights to survey respondents to represent the population they characterise, and to allow publication of population estimates.

References and further reading

References

International Statistical Institute (2010). Declaration on professional ethics. Retrieved 19 September 2013, available from http://isi-web.org.

Statistics New Zealand (1997). New Zealand Standard Institutional Sector Classification 1996. Retrieved 19 September 2013, available from www.stats.govt.nz.

Statistics New Zealand (2007). Principles and protocols for producers of tier 1 statistics. Wellington: Statistics New Zealand. Retrieved 19 September 2013, available from www.statisphere.govt.nz.

Statistics New Zealand (nd). *Methodological standard for confidentiality in business collections*. Unpublished document, Statistics NZ, Wellington.

Statistics New Zealand (nd). *Methodological standard for confidentiality in social collections*. Unpublished document, Statistics NZ, Wellington.

Further reading

Duncan, GT, Elliot, M, & Salazar-Gonzalez, J (2011). *Statistical confidentiality: principles and practice*. New York: Springer.

ESSNet SDC (2009). Glossary on statistical disclosure control. Retrieved 19 September 2013, available from http://neon.vb.cbs.nl/casc/.

ESSNet SDC (2009). Guidelines for the checking of output based on microdata research. Retrieved 19 September 2013 from http://neon.vb.cbs.nl/casc/.

ESSNet SDC (2010). Handbook on statistical disclosure control, version 1.2. Retrieved 19 September 2013 from http://neon.vb.cbs.nl/casc/

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

Statistics New Zealand (2006). *Confidentiality best practice manual (First edition)*. Unpublished document, Statistics NZ, Wellington.

Statistics New Zealand (2009). *Methodological standard for confidentiality in census*. Unpublished document, Statistics NZ, Wellington.

Statistics New Zealand (2009). *Methodological standard for Confidentiality Standard for Microdata Access*. Unpublished document, Statistics NZ, Wellington.

Appendix: Output rules – extra details and examples

Refer to this appendix for additional details and examples to supplement the microdata output rules in chapters 4 and 5:

- Random rounding to base 3 (RR3)
- Weighted counts
- Graduated random rounding
- The p% rule
- Maximum and minimum values
- Aggregation
- Suppression.

Random rounding to base 3 (RR3)

Unweighted counts are randomly rounded to base 3. Marginal totals of these counts can be independently and randomly rounded to base 3. Alternatively, you can calculate marginal totals by summing the rounded counts, but this introduces avoidable noise.

Random rounding to base 3 (RR3) involves randomly changing each count in a table to a multiple of 3. Apply RR3 by rounding values to:

- the nearest multiple of 3 with a probability of 2/3
- the second nearest multiple of 3 with a probability of 1/3.

Values that are already multiples of 3 are left unchanged.

Justification:

Small counts are sensitive. This rule protects counts of 0, 1, and 2. It also protects small counts from being revealed when differencing occurs, as all counts (large and small) are rounded.

Macros:

Macros that perform RR3 are available from Statistics NZ.

Examples:

An original (unrounded) count of 17 would be rounded to 15 with a probability of 1/3, and rounded to 18 with a probability of 2/3. Since 15 x 1/3 + 18 x 2/3 = 17, the expected value is unchanged and, over a table, bias is avoided.

A researcher performs some analysis and produces two tables of counts, which are disaggregations (by different demographics) of the same population. Suppose one cell in each of the tables represents the same count and this unrounded count is 11. This count could be rounded to 9 or 12. The researcher must ensure this count is rounded in the same direction for the two tables. So if the count is rounded to 9 in the first table, then it must also be 9 in the second table.

A researcher performs some analysis on Monday (for example) and produces an unrounded count of 8, which could be randomly rounded to 6 or 9. Suppose it is rounded to 9. If the researcher re-runs this analysis on Tuesday and produces the same unrounded count of 8, then this count must be rounded to 9.

Weighted counts

The usual procedure for weighted counts is as follows:

- Suppress below a specified threshold, which is usually three times the mean weight.
- Conventionally round all other weighted counts to a specified base, which is usually three times the mean weight.
- Suppress all zeros.
- Secondary suppression is not required.

Example:

The following table contains unrounded weighted counts produced from the Household Labour Force Survey (HLFS):

Number of people in part-time employment in Wellington		
Age	Male	Female
15–19	7,707	5,408
20–24	13,310	15,601
25–29	24,548	25,123
30–34	32,353	34,021
35–39	21,134	11,346
40–44	5,603	3,017
45–49	2,450	874
50+	1,789	902

After rounding and suppression:

Number of people in part-time employment in Wellington		
Age	Male	Female
15–19	7,700	5,400
20–24	13,300	15,600
25–29	24,500	25,100
30–34	32,400	34,000
35–39	21,100	11,300
40–44	5,600	3,000
45–49	2,500	S
50+	1,800	S
Symbol: S suppressed		

Graduated random rounding

Graduated random rounding (GRR) rounds the total number of individuals (eg employees, cows, sheep) within a cell by adding protection that depends on the size of the number. In this way, the noise added forms a proportion that increases with the size of the number. Apply the following GRR procedure to tables containing count magnitudes:

Count magnitude	Rounded to base
0–18	3
19	2
20–99	5
100–999	10
1,000+	100

Justification:

GRR prevents the derivation of the exact value of a contributor, even if the number of contributors is small.

Macro:

A macro that performs GRR is available from Statistics NZ.

Example:

Before confidentialising:

Number of employees in the retail industry				
Industry	**Invercargill**	**Queenstown**	**Dunedin**	**Christchurch**
Food retailing	384	992	1,226	3,156
Fuel retailing	77	24	71	98
Other	2	34	284	555

After rounding:

Number of employees in the retail industry				
Industry	**Invercargill**	**Queenstown**	**Dunedin**	**Christchurch**
Food retailing	390	990	1,200	3,100
Fuel retailing	75	20	75	100
Other	0	35	280	550

The p% rule

A table of value magnitudes usually contains totals of the contributed values of the businesses in the cells of the table. The contributed values usually come from a financial variable, measured in dollars. Statistics NZ's confidentiality standards state that any estimate for the contribution of a business to a cell total needs to have a sufficient level of uncertainty attached. The p% rule provides a measure of this level of uncertainty, and therefore the sensitivity of the cell.

The p% rule states that a cell, contained in a table of magnitudes, is sensitive if the value for any contributor can be calculated to within a given percentage. Cells that are identified as sensitive must be suppressed or otherwise avoided (through table design).

The p% rule calculates the distance (as a percentage) between the estimated value and the true value for the largest contributor in a table, as follows:

$$p = \left(\frac{\hat{X} - X}{X}\right) \times 100\%$$

X is the value for the largest contributor, $\hat{X} = Total - Y$ is the estimate of X, and Y is the value for the second largest contributor. If the value of p is less than the p% threshold then the cell is deemed sensitive. If a cell is sensitive, you must apply aggregation or suppression (with secondary suppression) to the table. If a contributor's value is negative, take the absolute value before calculating Total, X and Y.

The value of the threshold is confidential and must not be made public, but it is built into the macro and visible to researchers.

Justification:

The justification for the p% rule follows directly from its description. If a cell is deemed not sensitive, then a business can estimate the value contributed by a competitor, but only with a sufficient level of uncertainty. The p value estimates this level of uncertainty.

Macro:

A macro that calculates the p value for each cell is available from Statistics NZ.

Example:

A cell in a table contains five businesses with the following income values:

Business	BP	Z	Caltex	Mobil	Total
Income ($millions)	50	100	150	200	500

Assume that Caltex knows its own income and the total income for all businesses. If Caltex would like to estimate the income for Mobil, then the value of p is calculated as follows:

$$p = \left(\frac{(500 - 150) - 200}{200}\right) \times 100\% = 75\%$$

Therefore Caltex will overestimate the income for Mobil by 75 percent. If 75 percent was less than the p% threshold, then the cell total would be deemed sensitive and must be suppressed.

Maximum and minimum values

Maximum and minimum values are normally suppressed. Where a maximum or minimum value is not identifying, it may be considered for release.

Justification:

Maximum and minimum values are respondent values and may be outliers that pose a high risk of disclosure.

Examples:

Income, expenditure, and revenue are examples of sensitive variables. Maximum and minimum values for these variables are normally suppressed.

A researcher produces output about adults from a Statistics NZ dataset, including maximum and minimum ages. The minimum age may be released as this is likely to be defined by Statistics NZ. The maximum age will be suppressed as this may be an outlier.

A researcher creates derived scores from the Integrated Data Infrastructure (IDI) and produces maximum and minimum values for these scores. These maximum and minimum values may be considered for release on a case-by-case basis. The researcher will need to provide justification that the values are safe, and are not identifying.

Aggregation

A method for protecting sensitive cells, or cells containing small counts, is to aggregate (collapse) those categories.

Justification:

Tables may contain counts, magnitudes, or measures. Small counts need protection. Magnitudes may allow a contributor to be estimated with insufficient uncertainty. Aggregation is a way of avoiding these cells.

Example:

Before confidentialising:

Total turnover in the retail industry ($million)				
Industry	**Invercargill**	**Queenstown**	**Dunedin**	**Christchurch**
Food retailing	11	47	58	116
Fuel retailing	2	32	33	66
Other	1*	31	20	53
Total	14	110	111	235
Note: * This cell is deemed sensitive (by the p% rule) and needs protection.				

After aggregation:

Total turnover in the retail industry ($million)				
Industry	**Invercargill**	**Queenstown**	**Dunedin**	**Christchurch**
Food retailing	11	47	58	116
Other	3**	63	53	119
Total	14	110	111	235
Note: **This cell is now safe (passes the p% rule).				

Suppression

Suppression (or primary suppression) is the removal of a cell's value when it has been deemed sensitive. Secondary suppression is the suppression of other cells or marginal totals in the table so that the suppressed cell cannot be recalculated.

Justification:

If secondary suppression is not applied when appropriate, a user can recalculate the suppressed data.

Example:

Before confidentialising:

Total turnover in the retail industry ($million)				
Industry	**Invercargill**	**Queenstown**	**Dunedin**	**Christchurch**
Food retailing	11	47	58	116
Fuel retailing	2	32	33	66
Other	1*	31	20	53
Total	14	110	111	235
Note: * This cell is deemed sensitive (by the p% rule) and needs protection.				

After suppression has been applied:

Total turnover in the retail industry ($million)				
Industry	**Invercargill**	**Queenstown**	**Dunedin**	**Christchurch**
Food retailing	11	47	58	116
Fuel retailing	S	S	33	66
Other	S*	S	20	53
Total	14	110	111	235
Symbols: S suppressed **Note:** * This cell is primary suppressed. The other three suppressed cells are secondary suppressed.				

In this two-dimensional example, there are four suppressions altogether. Tables with other structures will need other patterns of suppression.