

When Individual Responses Exceed Input Storage - A Procedure For Unbiased Reduction

Technical paper for software developers

Statistics New Zealand

Contents

Introduction	p 2
Background	p 2
Methodology	p 4
Selection methodology	p 5
Example 1 Reducing to six ethnicity responses	p 9
Example 2 Reducing to three ethnicity responses	p 9
Example 3 Reducing to three level one categories	p 10
References	p 12

Introduction

This methodology paper is a technical paper for the development of software systems that support the inputting of large numbers of responses in surveys and administrative data sets. This paper was written by Jo-anne Allan and Robert Didham. It outlines the treatment of responses where the number of ethnic groups given by an individual exceeds the number being output. Scenarios are described for reducing the number of multiple responses to six per individual, and to three responses, and examples are given of each.

As a supporting paper to the Statistical Standard for Ethnicity 2005 it needs to be read in conjunction with the Standard and Understanding and Working with Ethnicity Data (2005). These are available from the Statistics New Zealand website, www.stats.govt.nz. It is essential to use Statistics New Zealand codefile of ethnicity responses for classifying of responses to the Standard Classification of Ethnicity. The codefile is available from Classifications and Standards, Statistics New Zealand.

Background

The number of people reporting more than one ethnicity to a question about ethnicity identification, has been increasing over time. While older people are less likely to report more than one ethnicity, people at the youngest ages are more likely to. More than 20% of people in the youngest ages do so. There is a trend towards a greater number of multiple ethnicity responses particularly among Maori and Pacific People, partly reflecting their age structure. The treatment of multiple responses has been discussed widely in the latest Review of the Measurement of Ethnicity.

The recommendations from the Review of the Measurement of Ethnicity (RME) Report 2004 have led to the development of a method to reduce multiple ethnicity responses for individuals when the number of responses given, exceeds the maximum number being retained. The RME collection recommendations (p.11-12) allow for six responses per person or, where this is not able to be implemented immediately, the provision of three per person. The reasons behind the changes in output are described in the RME report on pages 13 – 14 and in brief are that:

- increasing multiple ethnicity responses mean that prioritised data sets are becoming less useful as a way of presenting the ethnic make up of New Zealand
- when simplifying the data to one response per person, prioritisation biases statistics and
- it is inconsistent with the concept being collected.

When collecting ethnicity information more responses may be given than can be retained in the processing system. A manual method is given in the Statistical Standard for Ethnicity 2005 to reduce to six responses per person, together with a method for reducing to three responses per person, when retaining six is not possible. When information on a large number of people is being collected, the methodology discussed in this paper may be incorporated into software systems. The methodology examples given in this paper relate to collecting 14 fields at the raw data stage. The examples have been based on the 2001 Census question (this question remains the same for 2006). Fourteen fields allows for eight tick box responses and a maximum of six write-in responses on the census form, to sufficiently cover data capture for census.

Classification

The Standard Classification of Ethnicity 2005 categorises ethnicity responses. It is a hierarchical classification with four different levels of detail. Ethnicity categories are able to be aggregated from the most detailed at Level 4 to the broad groupings at Level 1. Responses received for the ethnicity question, range from being very broad descriptions of ethnicity to very detailed descriptions. They are coded to appropriate levels in the classification to reflect this. For example, South East Asian is a broad or less detailed, response and is categorised at Level 2 as code 41, and at Level 4 as code 41000, with the zeros representing this broad response. In contrast, the detailed ethnic group response, Malay, is coded to Level 4, category 41414.

Within in each broad grouping of ethnicities at Level 1, related ethnicities are nested under higher order categories. For example, category code 12116 refers to Irish ethnicity at Level 4 and is a member of the group of British ethnicities which are identified by codes 12100-12199. The British categories belong to another higher order group of European ethnicities at Level 1 which are identified by category codes 10000-12999. Responses that are not able to be classified as an ethnicity are categorised in residual codes, for example, a response of vegetarian is coded to Response Outside Scope, 98888; a response which is repeated, as a tick box and then a write-in response, or responses with the same code, will mean one is coded as Repeated Value 96666.

The codefile used for coding ethnicity responses codes frequently occurring responses to the classification categories. For example, a response of Aussie is coded to the category 12811 Australian. Also included in the codefile are common misspellings and responses not within the scope of the concept being collected.

The Standard Classification of Ethnicity (ETHNIC05) is available in the Statistical Standard for Ethnicity appendices or on the Statistics New Zealand website www.stats.govt.nz.

Methodology

This paper describes the broad methodology for selecting ethnicities when the number of responses needs to be reduced. Because there are many programming languages and each differs in its architecture, no attempt is made to present this in precise programming code. Multiple ethnicity responses are treated according to their classification level, working with the detailed responses at Level 4, after removing residual codes.

The methodology works by successive removal of responses one at a time until the required number of responses are retained. In the programming code, the number of responses required to be kept, should be able to be specified as an input parameter.

While the scenario described relates to creating a database retaining up to six responses, it may sometimes be necessary to reduce this number further for special purposes at output. To cover this situation, two sets of outputs are described, one with six responses and one with three responses.

The six response output can be selected first using the full set of ethnicity codes as input. Then proceed to use the six response output to select the three response codes retained for the three response output. This will ensure that the three response output is compatible with the six

response output, that is, that no ethnicity codes appear in the three response output that do not appear in the six response output.

Importantly, when six or more responses are being retained, no reduction in the number of responses will result in the removal of all responses in any Level 1 category. For example, if any Asian response is present, at least one of the reported Asian responses will be kept. Similarly, if a person identifies themselves as being of Māori ethnicity, this information will be retained because Māori is a Level 1 category. Only when it is necessary to reduce the number of responses to fewer than six, will any Level 1 category potentially be removed. In order to reduce bias in the data of those ethnic groups retained for output, reduction to three responses is accomplished by using a random selection method.

Input data

Each input response consists of a 5-digit ethnicity code. These codes can either be ethnicity categories with a first digit from 1 to 6, a 'residual' code representing some actual stated residual response 94444 to 98888, or a Not Stated residual code of 99999.

Removing a response

Removing an ethnicity response code and specifying precisely how to record that it has been removed could be accomplished by physically removing the value from a list. Alternatively, set up a temporary 0-1 flag variable for each original response code indicating whether or not it has been removed.

Output data

The output data will consist of a set of six selected ethnicity response codes for the six response output, and a set of three ethnicity response codes for the three response output if this is required. Storing the selected response codes can be done by creating a new set of ethnicity variables, `selected_ethnic1` to `selected_ethnic6`.

Another way might be to create a set of 0-1 flag variables indicating which of the input response codes have been selected, where six of the flags have a value of 1 and the remaining have a value of 0. As with the input data, the assumption is that any output responses that are not ethnicity responses or residual codes will have a code of 99999 Not Stated.

Selection Methodology

After removing residual response codes first, the methodology uses random selection of the ethnicity response codes. Response codes identified as the only response coded to a Level 1 category, are retained and removed from the selection process. A number is generated at random for the remaining response codes and the response code with the smallest random number is selected for removal. The procedure successively removes one response code at a time until the required number of response codes is left, at which point the procedure is completed.

Response codes are to be removed in the following order:

- residual response codes of Not Stated 99999
- residual response codes other than Not Stated 94444 – 98888
- broad response codes ending with 00, 000 and 0000 for the five-digit code then
- detailed response codes classified as separate categories at Level 4.

Removing Response Codes of Not Stated

It is common in databases to pack unused fields with residual codes such as 99999. Count the number of valid ethnicity response codes (starting with digits 1 - 6). If the number of codes is less than the number of ethnicities to be retained, then the number of packing codes (99999) is reduced until this number is reached.

For example, in the situation where there are 14 source fields, if the number of responses of 99999 is ≥ 8 remove eight responses of 99999 and the procedure is completed as the maximum number of responses has been retained.

If there are more than six responses remaining proceed to remove the other residual responses codes 94444-98888.

When three responses are required remove response codes of 99999. If more than three responses remain proceed to remove other residual response codes.

Removing Residual Response Codes

Residual response codes are used to ensure classification of ethnicity responses is exhaustive and range in codes from 94444 to 98888. For example, responses may be: 'I object to answering this question', classified to 95555 Refused to Answer; 'don't know' classified to 94444 Don't Know; and a write-in response of 'rugby player', classified to 98888 Response Outside Scope.

If the number of remaining responses is greater than six after removing codes of 99999 Not Stated, proceed to remove residual response codes 94444 to 98888. If removing all the residual response codes will still not reduce the number of responses to ≤ 6 , then simply remove all residual response codes and go on to remove broad response codes.

Otherwise, assign a random number to each of the residual response codes. Remove the residual response code with the lowest random number. If this brings the remaining number of responses down to six then the random selection procedure can stop as the maximum number of responses has been retained. Otherwise, remove the response code with the next lowest random number, and so on until either the number of remaining responses has been reduced to six or all the residual response codes have been removed.

If the number of remaining responses has been reduced to six then the random selection procedure is completed. Otherwise, if there are still more than six responses remaining after removing all the residual response codes, proceed to remove broad response codes.

When three responses are required continue to remove the lowest random number until either the remaining response codes have been reduced to three or all the residual response codes have been removed.

The assumption used above, is that users have no preferential removal order amongst the various residual codes, and residual responses are removed in a purely random order. If users wished to keep, Refused to Answer responses in preference to Response Outside Scope responses, then this step could be modified appropriately.

Removing Broad Response Codes

If the number of remaining response codes is still greater than six, proceed to successively remove broad ethnicity response codes one at a time if there is at least one other response in the same Level 1 category. When reducing to six response codes each Level 1 category with a response coded to it must be represented in the final six ethnicities retained. For example, a broad response of 'Pacific Islander' coded to 30000 cannot be removed if that is the only response recorded for the Level 1 category, Pacific Peoples.

A response is deemed a broad response in this context where there is a more detailed response subordinate to it also present and it is in one of the following forms:

- the first digit is between 1 and 6 inclusive, and
- at least one of digits 2, 3, or 4 is 0 (it is not sufficient that only digit 5 is 0), and
- there exists another response with the same digits as far as the place of the 0 but a non-zero digit in that place.

Note, a code beginning with a digit between 1 and 6 inclusive and ending with only one zero, such as 53120 Eritrean, is a detailed response.

These are examples of broad responses.

10000 European nfd
12000 Other European nfd
12100 British nfd

40000 Asian nfd
41000 Southeast Asian nfd
42100 Chinese nfd

Identify all broad response codes ending in four zeros, for example, code 30000, that have another response in the same Level 1 category (that is, starting with the same first digit). Assign each response code a random number and remove the code with the lowest random number. If this brings the remaining number of responses down to six then the random selection procedure

has finished. Otherwise, remove the response code with the next lowest random number. Repeat until either the number of remaining responses has been reduced to six or there are no more of this type.

Next, identify all broad response codes ending in three zeros, for example, code 12000, that have another response in the same Level 1 category (that is, starting with the same first digit). If there is just one response of this type, remove it. Otherwise, for more than one assign each a random number and remove the response code with the lowest random number. If this brings the remaining number of responses down to six then the random selection procedure has finished. Otherwise, remove the response code with the next lowest random number, and so on until either the number of remaining responses has been reduced to six or there are no more of this type.

Then, identify all broad ethnicity responses ending in two zeros, for example, code 43100, that have another response in the same Level 1 category (that is, starting with the same first digit). If there is just one response then remove it. Otherwise, for more than one assign each a random number and remove the response code with the lowest random number. If this brings the remaining number of responses down to six then the random selection procedure has finished. Otherwise, remove the response code with the next lowest random number, and so on until either the number of remaining responses has been reduced to six or there are no more of this type.

If the number of remaining responses has been reduced to six then the random selection procedure is completed as the maximum number of responses has been retained. Otherwise, if there are more than six responses remaining after removing all the broad responses, proceed to remove detailed responses.

Removing Detailed Response Codes

If the number of remaining responses is still greater than six, proceed to successively remove detailed ethnicity response codes one at a time if there is at least one other response in the same Level 1 category. The level one categories with only one ethnicity representing them are retained.

A detailed ethnicity response is a response code with the 5 digit code beginning with a digit between 1 and 6 inclusive and where zeros are found only at the last digit.

Example of detailed codes

12933 Romanian
37130 Nauruan
44111 Sinhalese
53120 Eritrean

Two or more response codes with the same first digit in each Level 1 category are identified for possible removal. Assign a random number to each response code and remove the code with the lowest random number. Repeat this process until the required number of responses is selected, maintaining representation of each Level 1 category.

Example 1 Reducing to six ethnicity responses

This example uses eight ethnicity responses when six are required to be retained. Two of these responses need to be removed, one at a time.

21111 Māori
 30000 Pacific Peoples nfd
 42111 Hong Kong Chinese
 42116 Taiwanese
 43111 Bengali
 43115 Punjabi
 44111 Sinhalese
 44112 Sri Lankan Tamil

Procedure

There are no residual codes of 999999 Not Stated to remove. Response codes identified as the only response coded to a Level 1 category, are retained and removed from the selection process. These responses are 21111 Māori and 30000 Pacific Peoples nfd.

There are no broad responses to remove. This leaves response codes 42111, 42116, 43111, 43115, 44111, and 44112 to randomly select from for removal of responses, since they aggregate to the same Level 1 category. Assign a random number to each response code and remove the code with the lowest random number. The code randomly chosen to be removed is 44111. Removing this code leaves seven responses and one more needs to be removed.

The remaining codes are 42111, 42116, 43111, and 43115 (44111 has been removed) to randomly select from for removal of responses. Remove the code with the next lowest random number. The code randomly chosen to be removed is 42116. Removing this leaves six responses. The random selection procedure is completed as the maximum number of responses has been retained.

The six retained responses are:

21111 Māori
 30000 Pacific Peoples nfd
 42111 Hong Kong Chinese
 43111 Bengali
 43115 Punjabi
 44112 Sri Lankan Tamil

Example 2 Reducing to three ethnicity responses

When three responses are to be retained, use the method described above to reduce to six responses first. Then proceed to reduce to three responses. This may be necessary, for example, for data comparability purposes.

Procedure

The six responses remaining from the example above need to be reduced to three responses. The codes retained were: 21111; 30000; 42111; 43111; 43115; and 44112. Where possible when reducing multiple responses, each Level 1 category is represented by one or more responses. For the example above, there are three Level 1 categories that are represented by the response codes. This means that each Level 1 category can be represented when reducing to three responses.

Codes 21111 and 30000 are retained as they are the only response to represent their Level 1 categories. One code must be selected from the remaining codes 42111, 43111, 43115 and 44112.

Assign each remaining code a random number. Remove the response code with the lowest random number (43111). Next, remove the response code with the next lowest random number (44112). Finally, remove the response with the next lowest random number (42111). This leaves three responses.

The three retained responses are:

21111 Māori
30000 Pacific Peoples nfd and
43115 Punjabi.

Example 3 Reducing to three level one categories

Where possible when reducing the number of responses, each Level 1 category is represented by one or more responses in the responses retained. Occasionally this may not be possible. This example shows the outcome when there are more than three Level 1 categories represented by the responses, with consequential loss of one or more Level 1 categories.

When three responses are to be retained, use the selection methodology to reduce to six responses first. Then proceed to reduce to three responses. This may be necessary, for example, for data comparability purposes.

The responses including packing codes for a 14 field collection are:

12711 German
12941 Swiss
21111 Māori
30000 Pacific Peoples nfd
32100 Cook Island Maori
32112 Atiu Islander
42100 Chinese nfd
52113 Brazilian
61118 New Zealander
94444 Don't Know
99999 Not Stated
99999 Not Stated
99999 Not Stated

99999 Not Stated

Procedure

Remove the four residual responses 99999 first. As ten responses remain, remove the other residual response, 94444 Don't Know. Nine responses now remain.

Proceed to remove one randomly selected broad responses one at a time, of the type with four 0000's but only if there is another response with the same first digit. There is only one able to be selected, code 30000, and this is removed.

Another broad response of the type 00 is present and as there is another response in this Level 1 category, 32100 is removed retaining seven response codes to this point. Note that this is in the same Level 1 category as the code previously removed but there is still one other code in the same category to represent it (that is, 32112). No other broad response is available for removal at this stage.

Next is the removal of detailed responses within the same Level 1 category. There is one Level 1 category with more than one response. 12711 and 12941 are assigned the random numbers, 631 and 429. The response code 12941 is removed as it has the lowest random number. Six response codes remain representing each Level 1 category:

12711 German
 21111 Māori
 32112 Atiu Islander
 42100 Chinese nfd
 52113 Brazilian
 61118 New Zealander.

These responses would represent the six response output.

Proceed to reduce to three responses. A random number is assigned to each response code and those codes with the three lowest random numbers are removed.

There are many possibilities of what Level 1 categories would be retained. Here are examples of some outcomes:

12711 German
 21111 Māori
 32112 Atiu Islander

OR

42100 Chinese nfd
 52113 Brazilian
 61118 New Zealander

OR

21111 Māori

42100 Chinese nfd

52113 Brazilian

OR

12711 German

32112 Atiu Islander

61118 New Zealander

The output data will contain the least possible bias due to the process of reducing the number of responses because the selection is arrived at randomly.

References

Statistics New Zealand (2004). *Report of the Review of the Measurement of Ethnicity*, Statistics New Zealand, Wellington.

Statistics New Zealand (2005). *The Statistical Standard for Ethnicity 2005*, Statistics New Zealand, Wellington.

Statistics New Zealand (2005). *Understanding and Working with Ethnicity Data*, Statistics New Zealand, Wellington.