ENABLING LEARNERS TO DISCOVER REAL STORIES IN OFFICIAL STATISTICS WITH A NEW SYNTHETIC UNIT RECORD FILE OF THE NEW ZEALAND INCOME SURVEY 2011

KEEGAN, Alan and TIDESWELL, Andrew

Statistics New Zealand
Wellington
New Zealand

Contact email: alan.keegan@stats.govt.nz

ABSTRACT
Learners of statistics need datasets reflecting real life contexts. Unit record datasets have interesting properties and contain stories that could engage learners. However, NSOs have legal and ethical duties to protect unit records. To enable access to these stories, Statistics New Zealand has published Synthetic Unit Record Files (SURFs) produced using several methods. With a new SURF based on the New Zealand Income Survey 2011 (NZIS 2011), we enable learners to access a new unit record dataset. Learners have an opportunity to discover and tell the stories about their region or country that are in the actual sample dataset. Learners can experience for themselves interesting aspects, such as: disaggregated data, semi-continuous distributions, and formal classifications. We hope learners will welcome the value of official statistics, as contributors of data and consumers of information from it.

INTRODUCTION – WHY USE A SURF TO TEACH STATISTICS?

Many learners are just interested in summary statistics. However, without unit records, there is no opportunity to create summary statistics, thus denying the chance to learn how to create them. Official statistics unit records introduce learners to official statistics contexts, which may not be taught in statistics courses at school or tertiary level.

Education networks, such as NAOS (Network of Academics in Official Statistics) and the NZSA (New Zealand Statistical Association) Education Committee, have interest in the creation of more New Zealand datasets. Their motivation is to increase the number of New Zealand stories told in the teaching and learning of statistics, to create more relevant learning contexts that other sources may lack. Official statistics datasets conform to formal classifications, standards, protocols, and methods set on a national (eg ethnicity, region, industry) or international level (eg employment). These may not be the same as those commonly used by researchers in various fields or data users in industry.

Statistics has increased status in the New Zealand education system, as detailed in the learning area now called Mathematics and Statistics in the New Zealand Curriculum (Ministry of Education, 2007). For example at senior school levels, inference includes resampling or randomisation methods. This increased the interest in suitable datasets for teaching and learning. These methods require unit level data to work, and are assessed in credits for the National Certificate of Educational Achievement (NCEA). NCEA is the state qualification learners study for in their final years of secondary schooling in New Zealand. Software enabling students to use these methods, such as iNZight (University of Auckland, n.d.), has been developed. The Ministry of Education and Statistics New Zealand have provided additional support to iNZight via the CensusAtSchool programme. Figure 1 shows iNZight used for bootstrapping medians.

Statistics NZ has published Synthetic Unit Record Files (SURFs) to enable learners to access the stories in unit record data. Rubin (1993) proposed synthetic data as a Statistical Disclosure Limitation (SDL) method. In a sampling context, he suggested treating the unsampled units in the population as missing. Then the missing values could be imputed using statistical models built on the survey data. These imputed units would comprise a synthetic dataset. Analysis of these data would create similar statistics to those from the real unit records, but without releasing real data about individuals.

Figure 1: bootstrap inference for median incomes of males and females

Statistics NZ SURFs are available from the School's Corner section of the Statistics NZ website. There are also resources using these datasets to support the Mathematics and Statistics curriculum. No permission, notification or sign up is needed to obtain these datasets. Forbes, Camden, Pihama, Bucknall and Pfannkuch (2011) have written about School's Corner and other collaborations for statistical education between universities and Statistics NZ.

THE NEW ZEALAND INCOME SURVEY 2011 SURF

The New Zealand Income Survey (NZIS) is an annual supplement, measuring income, to the Household Labour Force Survey (HLFS). The HLFS is the official measure of New Zealand's employment and unemployment. The supplement is carried out with the HLFS in the June quarter each year. Respondents participating in the HLFS are asked about the sources and amount of income from wages/salaries, self-employment, investments, benefits, superannuation, and hours.

We now describe a new SURF based on the NZIS 2011 (Statistics NZ, 2013). The dataset contains the variables in table 1. The first six variables are categorical, the last two numerical.

SURFs of the NZIS have been produced before. The first SURF published by Statistics NZ was based on the NZIS 2004 using simple perturbation SDL methods on a small sample of a subset of wage/salary earners. Official Statistics System (OSS) research by Lee (2007) underpins the methods used to create 2006 Census and Household Savings Survey (HSS) SURFs.

OSS research by Graham, Rodnyanskiy and Henley (2009) led to a larger, higher utility dataset based on the NZIS 2003. Their dataset has 100 multiply imputed sets of 11,000 synthetic records of people aged 25 to 64 in paid work.

What differentiates the NZIS 2011 SURF from previous SURFs is:

- coverage of the general New Zealand population aged 15 and over, not just subsets such as wage/salary earners or particular age groups
- a large sample size replicating the actual sample size of the real survey ($n = 29471$) rather than small sample sizes of a few hundred compared to the NZIS 2004 and HSS SURFs
- the addition of the region and occupation variables
- up to two responses for ethnicity

LEARNING OPPORTUNITIES

The differentiating features of the NZIS 2011 SURF expose learners to a dataset that better reflects the reality faced by data users in real life. These features were put in the new SURF because it is targeted to learners at Year 13 (final year of schooling) and undergraduate level. A scatter plot of income and hours (Figure 2) illustrates this reality. A simple linear regression analysis for example is not going to be sufficient. The features visible in Figure 2 are very similar to those in the real dataset.

A MORE COMPLETE PICTURE OF NEW ZEALAND

No subsets were excluded from the NZIS 2011 data in creating the new SURF. This makes the new SURF resemble aspects of the real NZIS dataset more. The new SURF is more diverse in representing New Zealanders at different life stages; New Zealanders at school, in tertiary study, in work, at home, unemployed, self-employed, in retirement are all represented – not just wage/salary earners. Even if only a subset is required, this empowers educators to make decisions appropriate for the needs of learners.

The large number of observations in this dataset permits learners to disaggregate national estimates by demographic variables that interest them. The large sample makes the structural and repeated zeroes in the dataset more visible. Larger samples enable learning opportunities on sampling variation and comparisons between groups across multiple variables.

The addition of region allows for some geographical analysis. It will enhance the ability for learners to disaggregate the dataset. Learners are not just limited to the national picture – they can find out about their own region too.

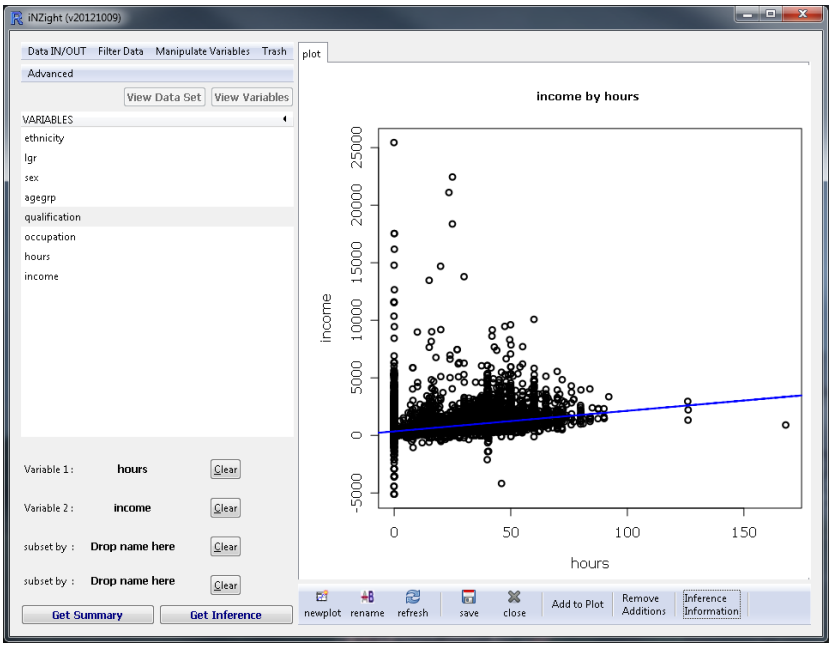

Figure 2: scatterplot of total gross weekly income and weekly hours worked in all jobs

Variable	Description	Categories or Numeric Ranges
age	five year age groups from 15 up to 65 and older	15-19, 20-24, … , 55-59, 60-64, 65+
sex	male or female	male, female
ethnicity	as per Level 1 (top level) of the Statistics NZ (2005) ethnicity classification – limited to two responses for confidentiality reasons	one or two of the following: European Māori Pacific Peoples Asian Middle Eastern/Latin American/African Other Ethnicity Residual Categories
region	local government regions (LGR)	Northland Auckland Waikato Bay of Plenty Gisborne/Hawke's Bay Taranaki Manawatu-Wanganui Wellington Nelson/Tasman/Marlborough/West Coast Canterbury Otago Southland
qualification	highest educational qualification	None School Vocational/Trade Bachelor or Higher Other
occupation	as per Level 1 (top level) of the Australia New Zealand Standard Classification of Occupations (ANZSCO) in main job	Managers Professionals Technicians and Trades Workers Community and Personal Service Workers Clerical and Administrative Workers Sales Workers Machinery Operators and Drivers Labourers Residual Categories
hours	weekly hours worked in all wage and salary jobs combined	0 to 168 hours rounded to half hours
income	gross weekly income from all sources	-5100 to 25443 New Zealand dollars rounded to whole dollars (losses are negative)

Table 1: description of variables

USING OFFICIAL CLASSIFICATIONS

The occupation variable introduces learners to the Australia New Zealand Standard Classification of Occupations (ANZSCO). ANZSCO is an example of an official classification used for a number of purposes (Statistics NZ, n.d.). ANZSCO is a hierarchical classification with four levels. The SURF uses the top level, with nine categories. The introduction of occupation in the new SURF allows a conversation about its use by employers, recruiters, insurers, immigration authorities, market researchers amongst others. And also students, potential future people in these occupations.

ANZSCO is also an example of international collaboration on statistical standards. The use of a joint classification by Statistics NZ and the Australian Bureau of Statistics for occupation

opens up a conversation about the comparability of occupations with other countries. The NZIS itself, as a supplement to the HLFS, uses standards recommended by the International Labour Organisation (ILO).

Ethnicity has not had multiple responses in Statistics NZ SURFs before. Up to four responses were in the NZIS unit record data, but the SURF only has up to two responses for confidentiality reasons. However, learners are confronted with a real life issues for data users. For example ethnicity is differentiated from race, descent or nationality. This lies at the heart of the debate about New Zealander ethnicity responses in the census (Statistics NZ, 2009).

To handle ethnicity data, learners have to decide how they will handle multiple responses. The inclusion of multiple responses for ethnicity provides interesting stories for learners. There are now choices to be made for handling this data. Whether to use total response, single/combination output, prioritised ethnicity, or something else are options learners must consider. They can now explore the issues outlined in Statistics NZ (2005a) for handling this kind of data.

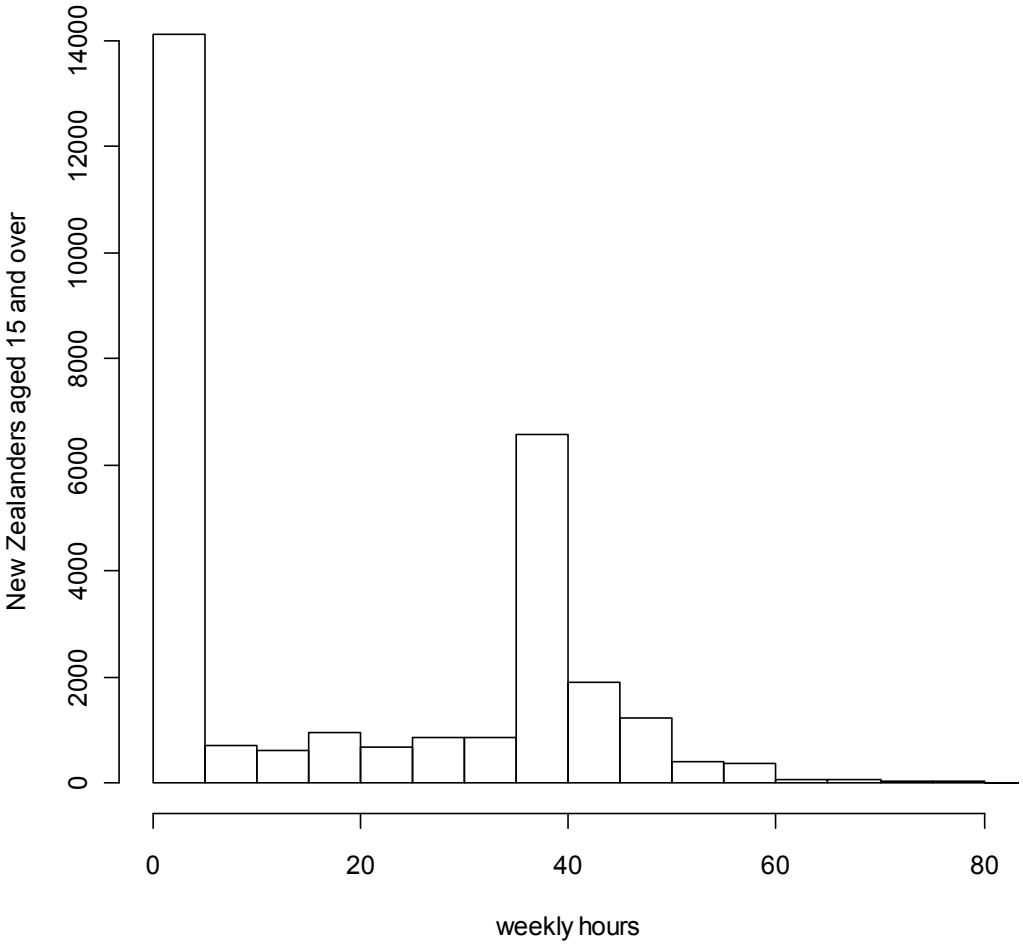

Figure 3 - distribution of total weekly hours for all wage/salary jobs from SURF

CHARACTERISTICS OF A LARGER MORE REALISTIC DATASET

The scatter plot of income against hours (Figure 2) reveals interesting properties of the data. There are repeated zeroes, skewness, and clusters visible in the dataset. Some insights can be gained from examining these variables in detail.

The distribution of hours (Figure 3) is strongly bimodal. It has a large frequency at one discrete point, zero hours, and frequencies at a range of points, centred on 40 hours. The bi-modal nature of this variable poses challenges for analysis. The addition of new variables like occupation mean learners can explore this behaviour, navigating this challenge. The distribution of hours by occupation (excluding those with no occupation) is in Figure 4.

Learners might question why there are still peaks for zero hours in Figure 4 even after people with no occupation were excluded. It seems counterintuitive that there are several managers working zero hours per week.

These discrepancies and others are explained by the structure of the NZIS 2011. There is large variability in income for people on zero hours – these people can be seen in Figure 2. As there are no excluded subsets in the SURF, some people who are recorded as working zero hours are self-employed. The NZIS 2011 did not collect hours worked for self-employment. Some self-employed made losses so are recorded as working zero hours with a negative income.

Having a dataset that provides a more complete picture of the New Zealand population enables the study of the stories underlying the numbers. The multiple reasons for having zero recorded hours of work is one introduction to the nature of official statistics survey datasets.

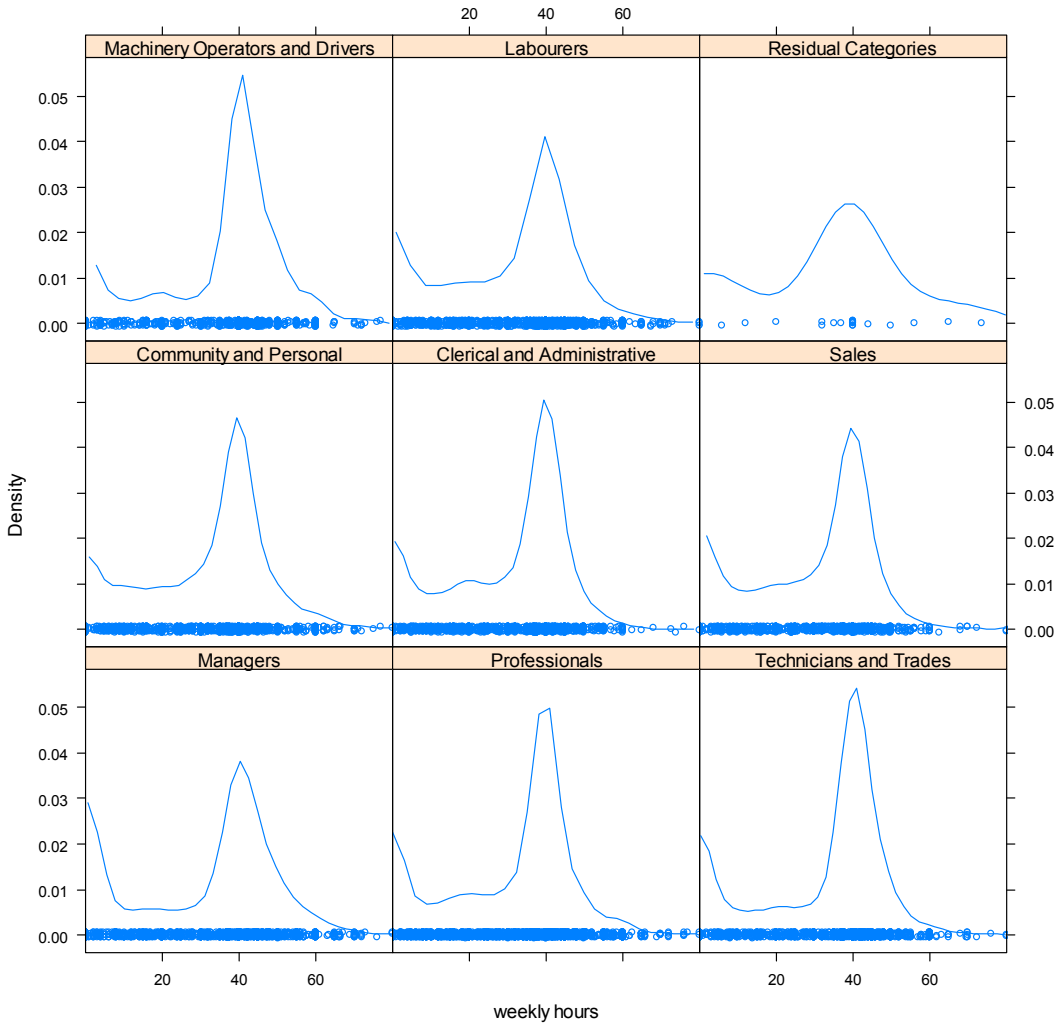

Figure 4 – density plots of weekly hours by occupation from SURF

HOW WE MADE THE SURF

As a synthetic dataset, the NZIS 2011 SURF balances the tradeoff between data utility and safety (disclosure risk). The dataset was created using Classification and Regression Tree (CART) models applied to the NZIS 2011 microdata. CART was chosen due to its ability to handle semi-continuous data, and its ease of application. Only a brief summary of the SURF creation process is outlined. Further details are available from the corresponding author.

A synthetic sample of New Zealanders aged 15 and over was created by sampling from benchmark population tables of age, sex, ethnicity and region. Then qualification was imputed based on distributions for groups from a classification tree on the variables from the benchmark table. Occupation (including no occupation) was imputed in a similar way.

Hours and income were imputed by first predicting whether a synthetic person would make a loss, work zero hours, work less than forty hours, work exactly forty hours or work more than forty hours. Then conditional on that prediction, hours and income were imputed. Their imputation was based on inverting uniform random numbers on empirical cumulative distribution functions of the observations assigned to terminal CART nodes in the NZIS 2011 dataset.

LIMITATIONS

The synthetic dataset does not have sampling weights or household information. Breakdowns of weekly income by wages/salary, government transfers, investments, self-employment income were excluded. These decisions were taken to avoid the confidentiality checking needed to manage attribute disclosure and subtraction-based data hacking.

Ethnicity and occupation were limited to Level 1 output. Region was the only geographic information used. These were done to reduce the disclosure risk and reduce confidentiality checking also.

The new SURF reproduced the overall skewness of income and semi-continuous nature of hours worked well. In creating the SURF, we considered the balance between analytical utility and safety. Higher dimensional precision is limited, but learners can use the SURF to study two-way or three-way relationships, such as income by ethnicity and region. At this level of aggregation, the SURF tells a similar story to the survey's story.

CONTRIBUTION TO STATISTICAL EDUCATION

The new NZIS 2011 SURF is a rich resource for statistics educators. It provides a recent update to previous SURFs published by Statistics NZ. The new SURF is a large dataset with more variables, observations and more detail, each of these elements enhances the storytelling ability of official statistics. The dataset presents a fuller picture, in the interesting context of income data.

In building the SURF, we were addressing one of the core problems NSOs all have; How to provide data products that are as useful as possible (in this case to statistics teachers and learners), and are safe in terms of avoiding disclosure about real individuals. If we have further time resources in future, we could progress the method we used here to achieve higher utility whilst maintaining safety.

We hope that statistics educators will be able to engage learners with stories they can discover about their region and their country with a freely available official statistics dataset.

ACKNOWLEDGEMENTS

The authors wish to thank Mike Camden, for his confidentiality expertise. And also Ann Ball, Phillip Marshall, and Sarah Anastasiadis for their subject knowledge of the NZIS dataset.

REFERENCES

Forbes, S., Camden, M., Pihama, N., Bucknall, P. & Pfannkuch, M. (2011). Official Statistics and statistical literacy: They need each other. *Statistical Journal of the IAOS: Journal of the International Association for Official Statistics*, *27*(3), 113-123. doi:10.3233/SJI-2011-0729

Graham, P., Rodnyanskiy, L., & Henley, L. (2009). *Confidentialising microdata using multiple imputation: Development and evaluation of a non-parametric hierarchical Bayesian imputation model for numerical data*. Official Statistics Research Series, *5*.

Lee, A. (2007). *Generating synthetic unit-record data from published marginal tables*. Official Statistics Research Series, *1*.

Ministry of Education. (2007). *The New Zealand Curriculum*. Wellington, New Zealand: Learning Media.

Rubin, D. B. (1993). *Statistical Disclosure Limitation*. Journal of Official Statistics, *9*(2).

Statistics NZ. (2005). *Ethnicity*. Retrieved March 21, 2013 from http://www.stats.govt.nz/surveys_and_methods/methods/classifications-and-standards/classification-related-stats-standards/ethnicity.aspx

Statistics NZ. (2005a). *Understanding and Working with Ethnicity Data.* Retrieved March 22, 2013, from http://www.stats.govt.nz/~/media/Statistics/surveys-and-methods/methods/class-stnd/ethnicity/Understanding%20and%20Working%20with%20Ethnicity%20Data.pdf

Statistics NZ. (2009). *Review of the Official Ethnicity Statistical Standard*. Retrieved March 22, 2013, from http://www.stats.govt.nz/browse_for_stats/population/census_counts/review-measurement-of-ethnicity/~/media/Statistics/Census/2011-Census/final-report-review-official-ethnicity-statistical-standard-2009.pdf

Statistics NZ. (2013). *New Zealand Income Survey 2011 CART SURF*. Retrieved September 12, 2013 from http://www.stats.govt.nz/tools_and_services/microdata-access/nzis-2011-cart-surf.aspx

Statistics NZ. (n.d.). *Occupation*. Retrieved March 22, 2013 from http://www.stats.govt.nz/surveys_and_methods/methods/classifications-and-standards/classification-related-stats-standards/occupation.aspx

University of Auckland, Department of Statistics. (n. d.). *iNZight.* Retrieved August 8, 2013, from https://www.stat.auckland.ac.nz/~wild/iNZight/index.html

APPENDIX: COMPARISON OF SURVEY AND SURF STATISTICS

BASIC SUMMARY STATISTICS OF AVERAGE WEEKLY INCOME

	Average: all sources collected	Median: all sources collected	Number of people (000)	Average: all sources collected	Median: all sources collected	Number of people (000)
	SURF			NZIS 2011 Release		
Sex						
Male	779	642	1,674.0	850	700	1,683.3
Female	611	461	1,787.1	563	432	1,777.8
Total	692	536	3,461.1	703	550	3,461.1
Age group (years)						
15–19	198	0	309.1	116	0	316.3
20–24	567	501	321.7	465	404	321.9
25–29	715	660	301.1	692	683	291.5
30–34	796	717	275.9	819	740	270.9
35–39	899	774	286.8	899	779	290.2
40–44	883	726	308.3	923	777	311.4
45–49	871	725	322.4	957	804	318.6
50–54	911	733	296.2	944	776	297.0
55–59	844	690	251.3	878	722	255.7
60–64	816	662	234.2	717	525	235.4
65+	421	315	554.2	533	385	552.2
Total	692	536	3,461.1	703	550	3,461.1
Ethnic group[1]						
European	706	552	2,585.0	752	580	2,586.5
Māori	590	443	428.9	562	459	436.5
Pacific peoples	606	432	183.9	479	390	188.7
Asian	678	531	365.2	560	405	367.1
MELAA[2]	658	495	40.3	618	414	41.1
Other ethnicity[3]	731	552	81.5	700	520	72.8
Total[4]	692	536	3,461.1	703	550	3,461.1
Highest qualification						
No qualification	565	397	809.3	459	359	798.6
School Certificate / NCEA level 1	564	429	829.6	524	397	265.4
Sixth form / NCEA level 2				539	410	217.5
Higher school / NCEA level 3				545	361	223.2
Other school				487	372	130.5
Vocational or trade	734	612	990.6	789	680	999.9
Bachelor's or higher degree	955	796	613.4	1,097	948	603.4
Other post-school	725	571	218.2	804	614	174.2
Total[4]	692	536	3,461.1	703	550	3,461.1

1. People who reported more than one ethnic group are counted once in each group reported. This means that the total number of responses for all ethnic groups can be greater than the total number of people who stated their ethnicities.
2. The MELAA category contains all Middle Eastern, Latin American, and African ethnicity responses.
3. The category 'other ethnicity' includes the 'New Zealander' responses.
4. Totals include the 'not specified' category.

NCEA level 1 is typically attempted in the final year of compulsory schooling (Year 11), with levels 2 and 3 attempted in Years 12 and 13 respectively. NCEA replaced other New Zealand state secondary school qualifications in the mid-2000s.