Res. Lett. Inf. Math. Sci.. (2001) **2**, 55-61
Available online at http://www.massey.ac.nz/wwiims/~rlims

Issues in Data Collection: Missing Data and the 2001 New Zealand Census.

Judi Scheffer

IIMS, Massey University Albany Campus, Auckland, New Zealand
J.Scheffer@massey.ac.nz

Abstract

Missing data plagues all surveys, and to a degree the New Zealand Census suffers from the same malaise. While it is not a high level of missingness, it is present. If not correctly dealt with; just deleting cases with missing data will lead to biased conclusions, particularly if the missingness mechanism is NMAR. Some missing data may be inevitable; sometimes a respondent may be incapable of answering a question. This is usually MAR. If however the respondent refuses to answer a question because of say having a high income, then the results of the income question will be biased. Over time there have been a growing number of people employing avoidance tactics so as not to be classified as a refusal, but to make enumeration just too difficult. Anecdotal evidence among enumerators shows that this accounts for about 5% of respondents.

Introduction.

Statistics New Zealand prides itself on a 98.8% response rate for its '96 census. Is this achievable in 2001? At a 98.8% response, the NZ Census has probably the highest response rate in the world, and Statistics New Zealand would like to keep it that way.

What is Missing data?

Missing data occurs when a cell in a data table remains empty due to nonresponse in the census.

Consequences of Missing Data

The consequence of having missing data is to generally throw away cases that contain any missing data. This is wasteful of data, which may be expensive to collect. In the case of sample surveys the sample may not match the target population, and therefore leading to bias.

MCAR, MAR and NMAR

MCAR occurs when the data is completely randomly missing, i.e. the data is missing randomly within the variable, and the data has been completely observed at random. In the case of the census, this means there is no underlying reason for the missingness, and each case has an equal probability of being missing as another. An example of this might be the address not filled in on the front of the dwelling questionnaire, as the respondent did not realise the question was there.

MAR occurs when the missingness occurs randomly within a variable (it does not depend on the actual response), but the chance of a particular answer being missing is dependent upon another variable (or set of variables), which are observed. An example might be the education question for the elderly, or new immigrants not sure about how to answer the Maori genealogy question.

NMAR (or informatively missing) occurs when the missingness mechanism is dependent on the actual value of the missing data, if it were to have been fully observed. An example of this in the census would be the income question not being answered by those on a higher income. A nonresponse rate of 10% with a NMAR mechanism would definitely bias results.

Bias

Missing data occurs for all kinds of reasons, and the assumptions made about missing values are often incorrect, for example MCAR, or MAR, when the underlying cause may be NMAR. The resulting bias can be quite large, especially when the analysis requires complete case data, where cases are only used if each variable is completely observed. To choose an appropriate method for dealing with missing values requires knowledge of why the data are missing in the first place.

To investigate the reason for the missing data, validation studies should be carried out to understand whether missing values are random across the study population, or occur more frequently in specific subgroups. In case control studies, very often the case information is more complete than that of the controls. With regard to the 2001 census, a post enumeration survey will pick up 'incorrect data', but not necessarily missing data.

Depending on the reason for the missing data, we may decide to use imputed data, or stick with complete case data. Complete case data is only unbiased if the missingness is MCAR. In the case of categorical data a test can be performed by comparing the odds ratio for some factor for the category of missing values, as against the odds ratio for the non-missing category. The other factors can show whether a variable is indeed MCAR. If the odds ratios are different, then a complete case analysis will be biased. If the missingness mechanism is MAR, then the resulting analysis may be biased, if NMAR, the resulting analysis will be biased.

Omitting covariates

Sometimes covariates containing missing data are omitted, as a means of producing a rectangular complete case data matrix (case deletion): again this may be throwing away valuable information. It is well known in the regression literature that omitting variables may result in biased regression parameters and predicted fits.

Forms of Non-Response:

Non-response may be by design, or unit non-response, or item non-response. What actually causes non-response? Unit or Item non-response can occur in many forms.

Unit Nonresponse.

Unit Nonresponse - The entire case data is missing. These are the non-contactable cases (non-coverage). Unit Non Response can occur even when a census is properly designed. The enumerator can have great difficulty making contact with the respondents. It may not be because they do not want to participate (refusal), but just that they are uncontactable.

1. The Non-Contactables. This form of item non-response occurs when the occupants are unable to be reached. This can occur if the dwelling is overlooked by the enumerator, or a person or persons moved house on census day, and the dwelling is incorrectly labeled as empty, or the census forms simply don’t reach the respondents (they may always be out when the enumerator calls, then only to be 'lost' when placed in the letter box. People travelling the country side in vehicles such as campervans which do not use camping grounds, but stay on the side of the road, also are at risk of being non-contactable, as do 'street-kids' and others who sleep 'rough'. Non-contactables are always unit non-response.
2. Refusals. Refusals interestingly can be item non-response or unit non-response. What causes refusals? In the case of unit non-response these include people with an attitude that is against the census (or any survey) for whatever reason. People in this category are often older (for example > 70), who have a great mistrust of younger people asking questions or they may be from immigrant groups, who have survived wars etc., and are very careful about whom they tell what. Careful explanation of what the census is about can go some way to alleviating this problem. Refusal can be due to the respondent taking offence at some sensitive questions (the prime one is personal income), which can lead to either item non-response (refusal to answer individual questions), or unit non-response (refusal to answer any questions at all).

Item Non response

Item non-response occurs when the respondent does not answer a question (or questions) within the census form. This may occur because the respondent is unable to answer the question (does not understand the question), or cannot answer the question (does not know the answer), or more likely, they are offended by a question (or questions). This occurs most often with the income question. It is perceived a 'nosy' question.

Item nonresponse occurs when just one, or more questions are left out, possibly due to a partial refusal of sensitive questions or an inability to answer the question (they don't know). For example, do the 'don't knows' really not know? One way it can come about is through poorly organised questionnaires, where the respondent may be incapable of answering the questions, or may not know the answers. For example, if data is gathered by interview, and if the respondent is deaf, or has some other disability, the respondent may say, "I don't know" because it is easier. If a value is unobserved, say in the census, due to the respondent's inability to choose, this is more correctly a 'don't know' response. A 'don't know' may or may not be inevitable, but if it really is 'don't know' then it will inevitably be missing. According to Rubin, Stern, Vehovar (1995) a 'don't know' response can be considered a true missing value, (but if it is a refusal to answer, it has an underlying response). Don't Knows are ignorable nonresponses. (Cochran, 1977)

Sensitive Questions

These questions cause resentment by the general public filling in forms. Reactions to this type of questioning ranges from: "Why do they need to know this?" to a refusal to answer. Once a respondent becomes annoyed by a question such as this, very often they will refuse to fill in the remainder of the questionnaire. The individual Questionnaire for the 2001 census has lost the question about how many children a woman has given birth to. (Interestingly this question is asked routinely in the Australian census, and not perceived as a sensitive question) Previously this created resentment and distress for women who have had children adopted out, or have had a child die in infancy, or for what ever reason they did not wish to disclose to other household members an otherwise unknown pregnancy. A privacy envelope in this case could be used, but it would often create (within a household), more difficulties than it solved.

Other Questions previously asked were that of 'social' habits, such as smoking drinking and so on. The general public as perhaps a means of denying health care later on perceived this. The income question is still there. Some respondents have blatantly stated that, "they have this information on the Inland Revenue computer, why can't they get it there"? Others state, "They don't need this information to count people and dwellings", and will often understate their income. Those on a high income are very reluctant to part with such information, even when it is 'coarsened' into ordinal bands, rather than an actual number. The problem with that is when successive governments attempt to set threshold levels for access to assistance for families and individuals; the attempt may be to assist the lowest quartile of the population. If the income is underreported the figure used may well be the 15^{th} or 20^{th} quartile, and this may suit the Minister of Finance very well. This occurred some years ago when Family Benefit was abolished, and Family Support was phased in, a higher level of assistance targeting those most in need. News broadcasts at the time stated that the response by families applying for assistance was considerably lower than expected, largely due to the estimates of income being taken from census data.

Persuading the public that census data is <u>not</u> used to feed information to Inland Revenue, or to health boards, or to Work and Income NZ, or Immigration Services, the police, (or any other Government agency) is a very difficult task placed on the enumerator. The linking of all these agencies via the community services card, and Inland Revenue numbers, whilst cutting down on benefit fraud, has done the census a huge disservice. Statistics NZ does not disclose this information, but convincing respondents of this is not easy. "Its all the Government, isn't it?" Beneficiaries are particularly wary of divulging information. Education in this area, and gaining the trust of the respondents, cannot be overstressed. Statistics NZ, for its 2001 Census has engaged in a fairly extensive publicity campaign on national television to rectify these problems, largely based on the benefits of the census for forward planning for future infrastructure.

The giving of false information is a very attractive proposition to some individuals who wish their circumstances to remain private. The Dutch experience of this was to make the census voluntary. This

was somewhat disastrous, as the response rate there now is little more than 50%. But presumably because it is voluntary, therefore the data collected would not be subject to deliberately inaccurate replies.

The New Zealand Census

In NZ, The Statistics Act (1975) states " there will be a census on the first Tuesday in March, 1976, and every 5 years thereafter".

Questions asked include

1. Questions which must be asked each time (mandatory questions), which include: (a) name, address, sex, age (usually date of birth), ethnic origin (on the individual form); (b) Particulars of the dwelling, number of rooms, ownership, number of occupants on census night.
2. Questions that may be asked each time (standard questions), but will turn up quite frequently. These include: (a) occupation (profession), industry in which employed, nationality, citizenship, health, marital condition, religion, birthplace, duration of residence in NZ, address where living in NZ at the previous census, number of children, number of hours worked per week, for salary or wages or financial reward, mode of transport to/from work, name and address of employer, address of usual residence, service in the armed forces of every occupant in the dwelling; (b) particulars of the type and tenure of the dwelling, nature of materials of structure, household amenities, rent paid and details of livestock.

Public interest questions. The general public makes submissions on what they would like to be included in the census. Usually these are public interest groups applying to have their pet topic included. Whether or not these are included depends entirely on which are selected (if any) by Statistics New Zealand for inclusion. Very few of these actually make it onto a census form. In the 2001 Census a question has been asked about telephone, cell-phone, Internet and Fax access, which really is the household amenities question. This will become the pseudo 'poverty question'. This is used to determine access to current modern technology. In the past it may have been 'Do you own a washing machine, or a video recorder'. On the individual form, "Have you looked for work in the last four weeks, and if so how?"

History of the census in New Zealand

New Zealand's first (non-Maori) census was in 1842, and has run mostly every five years since then. Two notable exceptions are the 1931 census (cancelled due to depression) and the 1941 census; delayed due to World War 2; and then held in 1945, with the 1946 census being cancelled. Abandoning the census in 1931 was particularly unfortunate, as this prevented an economic snapshot being taken of New Zealand society at the time of the worst economic depression, to date. The 1951 census was to first to incorporate both Maori and non-Maori, using the same questions. The census has continued every 5 years since then.

Problems with enumeration

Call Backs

Callbacks can be time consuming and frustrating to the enumerator. This can be partially alleviated by allowing enumerators a reasonable amount of time to do the job. Insisting on starting the delivery phase on the Friday 11 days before the census is counterproductive. It is better to make the first sweep of each area during the day, midweek, and have tried the entire area before the first weekend. This way the valuable night and weekend hours can be spent on the more difficult second and third callbacks. The elderly do not like being visited at night, and a number of recent immigrants simply refuse to open the door to anyone knocking, turning out lights and pretending to not be at home. The collection phase has the advantage of having already been at each house and knowing exactly when people were at home for the delivery. Enumerator protocol on the delivery phase is to ask the respondents "when during the day are you likely to be home"? (Statistics NZ, 2001). This implies 'the enumerator will be around at a particular time, or day to collect the forms', and with around 350 dwellings to visit in a 10-day period, this is impractical.

Ways in which people may refuse

Refusals may in fact be reasonably creative. The respondent who does not wish to fill in their forms may well refuse outright as do the Libertarians (a political group) who advocate the burning of their Census forms. Refusals such as these are unit non-response, and are quite straight forward, and will often lead to a prosecution. The $250 fine allowed for in the Statistics act 1975 (this was upped to $500 with the amendment to the act in 1985) for non-compliance may have been a deterrent then, but certainly is not sufficient of one after a further 16 years of inflation.

Many of the refusals are subtler than that, giving the name of Fred Flintstone (a cartoon character) or Rob Muldoon (an ex-Prime Minister of NZ). Clearly their returned forms will be facetiously filled out, and such data will not really be worth anything to Statistics NZ. Others have shown their contempt for the census by throwing the forms on the ground, as soon as the enumerator's back is turned. Still others will state that they are not going to be at that address on census night (they are going away for the night), and wish to have their dwelling enumerated as residents away. Others will legitimately leave the country to avoid the census, but not every objector has these resources. Still others say that their house really is a business, and that they don't live there, even though they quite clearly do, as a late evening visit will prove.

Another favourite is to attack the enumerator, whether by dogs, or by verbal abuse, or by complaining to the local talkback radio station, or ringing the census help line, and complaining to head office (to create embarrassment, so therefore the respondent is left alone). Other forms of harassment include men coming to the door; answering women enumerators scantily dressed, or vice versa. One male enumerator couldn't make himself heard, so he went around the side of a house to speak to the occupants, and met the lady of the house stark naked. He pulled a cap down over his face somewhat and proceeded to deliver the papers. Two houses further on the hapless enumerator fell down a set of stairs as reaction set in. Fortunately these cases are relatively rare, and would probably only constitute one percent of all cases. While all of these do not constitute refusals in the known sense of the word, they are in fact avoidance tactics. It is not helpful that people who have found an avoidance tactic that works in one census, will try that again in another census, and very often get away with it.

The instructions to enumerators are very clear (Statistics NZ, (1996; 2001)), that a refusal is not a refusal until the non respondent has refused to fill in the form directly to the enumerator; that is a third party cannot refuse on behalf of another person. (An enumerator may not raise a refusal report, unless there has been a direct face-to-face refusal) This has not changed over the last fifteen years, but perhaps refusals are smarter today than fifteen years ago when a refusal was obvious. Today a refusal is far subtler, and a reluctant responder will use avoidance tactics instead.

Statistics New Zealand is probably helping the situation considerably with the 2001 census; by only requiring the enumerator to check questions 2,3,4,5,8,11(the L check), which constitute the mandatory questions (as listed above). These are on the front page of the individual questionnaire, and as long as these are filled in no further checks are done (at the door). This is a departure from the past, where the enumerator had to scrutinise the complete form. They were then required to grill the respondent at the door about missing responses. This is a very time consuming exercise for the enumerator, and perceived as 'nosy' by the respondent (particularly by the many who had correctly filled in their forms).

This departure of only checking the mandatory questions (under the Statistics act 1975) may well bear fruit in later censuses, as this will be seen as less invasive (but will undoubtedly lead to more missing data). Presumably the quality of what is there will be improved.

Integrity of the data:

Anecdotal evidence talking to enumerators in the past showed that enumerators needed to have complete data to be paid (or were led to believe this was the case). Certainly they were encouraged to submit complete data. The enumerator would be required to call back on the respondent, to collect the missing data. Very often if there was item refusal (just one or two questions) this became burdensome if they had already called back several times to 'catch' the respondent at home, unsuccessfully. Some unscrupulous enumerators admitted to just 'Ticking the box' presumably at random, or with whatever they presumed were the correct answer. Another case discussed by former enumerators was that of telling respondents to 'just tick anything as long as it is ticked'. The problem with this is (quite apart from the obvious lack of integrity of the data) that respondents actually remember instructions like this, and will use this ploy if they are offended by census questions, in later censuses, causing a flow-on effect.

Misclassification rates are estimated by a post enumeration survey. This is achieved using complex sampling methods, with primary units selected are clusters (sub districts, and within that mesh blocks), secondary units may involve stratification by age; and within that systematic sampling is used (with a random start figure). The Maori language survey is selected in the same way. These provide useful 'checks' on the data. Edit checks are used by Statistics NZ to check on things such aged 0-14, and married. These are logical checks.

Changing attitudes to the census

The census is an invasion of an individual's privacy, and this fact is becoming more apparent with each successive census. Methods to ensure privacy include withholding small area information where there are insufficient respondents, and also to 'mask' data prior to publication.

Questions should be tested on a range of individuals prior to the census to avoid ambiguous or overly sensitive questions. A fault with the 1996 census, which reappeared in the 2001 census, was the secondary educational attainment question. This was offensive to the elderly and continues to be so. Another case of a question not fully thought out, or tested prior to the census was the question "Did you look for work in the past four weeks", if you were not in paid employment in the week immediately prior to the census. For various members of the community, such as mothers of the very young, or the elderly, this question is inappropriate. One elderly gentleman aged 96 felt distressed, because he wasn't aware that he should have (been looking for work). In this day of automatic tax returns, employers are no longer required to give their employees a record of their income for the year, and to that end many people simply do not know their annual income (particularly those who work part time or have casual employment).

Assistance in 2001 has been via an automated help-line. However this system proved inadequate around census time, with many callers receiving the message "This mailbox is full, Please try again later". On the second, third, fourth try, many would give up and leave the question blank.

Sometime political statements made around the time of the census are very unfortunate, and have a detrimental effect on the census. In 1976, immediately prior to the census the government of the day announced it would deport all illegal overstayers. That particular census had a problem in that the count for ethnic minorities was well below the expected figure.

When The Statistics Act, 1975 was passed, people's attitudes to the census were reasonably compliant. This act made the filling in of the form (both dwelling and individual) compulsory. Prior to this public opinion was such that if the government (or any Government Department) required you to fill in a form, you did so without question. There would always have been the odd dissenting voice around, but most people complied without question. With every successive census since then until today, the public at large are now questioning the need for their privacy to be invaded, and will not blindly give out private information on themselves, without good reason.

Education of the general public on why this information is needed is vital, as is further reassurance that this information will not be passed on to other government agencies. Without this, the perception "Because my name is taken down (and address), therefore, as computers are capable of sharing information, other Government agencies are able to access this information."

Some respondents simply will not give out information on their name, and address citing, "If it is going to be anonymous on the computer files, why do you need me to fill in my name and address?" Trying to persuade these people that it (name and address) is removed before being made available for use, or that really it is a check on whether or not the enumerator is doing their job, is not easy when faced with these preconceived prejudices.

Anecdotal evidence shows that a typical objector is usually male, of European (Caucasian) descent, aged between 40-60, and poorly educated (has little or no secondary education, and has not attempted any tertiary education). Often there will be a history of difficulty dealing with local government, or with central government departments. There is not only a perception that all census information will be available to other government agencies, but that this is an insidious means to track down where each person is actually living. Many people who fall in this category will be in rental accommodation, and will frequently move house, that is will have had more than two moves in the last six months. There is little point in the enumerator trying to persuade these people that their information is confidential, that other Government agency computers cannot access it. They feel the enumerator who knows this to be true is

extremely naive. ("Only one press of a computer button and it's all there.") This type of missingness will not be MCAR; at best it will be MAR, it is far more likely to be NMAR.

To get around this problem, before the 2006 census a publicity campaign is needed, aimed specifically at this group of people, stating that the information is confidential to other Government departments. This was done on a small scale, but did not go far enough. What is needed is perhaps an "ad campaign" with the other government agencies specifically stating that they cannot access Census information at an individual level.

Conclusion

The Statistics Act (1975) and its amendment in 1985 compel every person and dwelling to be enumerated in the census, and every person to co-operate. Instead of Direct refusal, as in the past, which could occur a $500 fine (and $20 per day thereafter that a respondent did not comply, a growing number of people are resorting to avoidance tactics. This is unfortunate as unit nonresponse is undoubtedly more serious than item nonresponse. There are perfectly valid imputation techniques available, which can be used to 'plug the gaps', produced by item nonresponse. The best way to prevent item nonresponse, is to have improved public education about why a census is necessary, and to involve other Government Agencies, having them explicitly state that 'they cannot get their hands on an individuals Census data'.

Imputation techniques are certainly an improvement on an enumerator 'grilling' a respondent at the door, as happened in the past -this can lead to a respondent answering facetiously if 'forced into a corner'.

In the past Statistics NZ only imputed gender and ethnicity. With a great improvement in imputation techniques over the years, this has been able to be used, therefore reducing the need to Case Delete.

References

NZ Government (1985) *Amendment to The Statistics Act (1975).*

Cochran, W.G., (1977) *Sampling Techniques*. Wiley.

Department of Statistics (1986) Enumerator *handbook, 1986 census*. NZ Government Press.

Department of Statistics (1991) Enumerator *handbook, 1991 census*. NZ Government Press.

NZ Government (1975) *The Statistics Act (1975).*

Rubin, D.B., Stern, H.S., and Vehovar, V. (1995) Handling "Don't Know" Survey Responses: The Case of the Slovenian Plebiscite. *Journal of the American Statistical Association*. **90**(431) 822-828.

Statistics New Zealand (1996) *Census 1996 Enumerator Handbook*. Government Press

Statistics New Zealand (2001) *Census 2001 Enumerator Handbook*. Government Press