The importance of the Census to Canada

David A. Green and Kevin Milligan UBC Economics
August, 2010

Intro

In June, the federal government published plans to replace the mandatory 'long form' census with a new National Housing Survey for the 2011 census cycle. The new NHS is to be circulated to more households, but will be voluntary rather than mandatory. The announcement generated a response unique both for its breadth across civil society and its near uniformity. The breadth of response reflects the plethora of uses for census data, stretching into so many important decisions made by businesses, municipal and provincial governments, and non-profits. The near uniformity of response reflects the certainty of science on the statistical nature of voluntary versus mandatory sampling techniques. In this piece, we begin by showing the statistical importance of the distinction between voluntary and mandatory sampling techniques. We then proceed to explain how the ubiquity of the census in Canada's national statistics is even greater than many appreciate due to the role of the census as the ultimate benchmark for other surveys. Finally, we close with thoughts on the role of government in the realm of statistics.

Why voluntary sampling fails

To understand the issues in choosing between the mandatory long form census and the NHS, it is useful to consider an example. Suppose that we are interested in knowing the value of a parameter for the whole adult population of Canada---mean income, for example. Clearly, the most accurate way to proceed would be to get the income for every adult and then simply average them. Mandatory censuses, which have been used in Canada and around the world, aim to get responses to some set of questions for everyone in the country. But if it is too costly to survey everyone in the country then we can get an estimate of adult mean income by surveying a sample of the population. The key issue of concern is the conditions under which a given sample will provide an accurate, or unbiased, estimate of the population mean.

To continue with our example, suppose we selected two adults at random from the whole population and got their income. By basic laws of statistics, the average of these two incomes would be an unbiased estimate of the true population mean. But for any specific sample, we wouldn’t expect the sample mean to be the same as the population mean, as the mean of the small two-person sample would be highly dependent on which people were drawn into the sample. The difference between the sample and population mean, however, would not be systematically too high or too low--- the positive errors would be expected to exactly balance the negative errors. If we drew many, many such samples and then took the average of the mean incomes for each of the two-person samples, that average would equal the mean income for the entire adult population. This, in fact, is the definition of an unbiased sample: it is one that delivers estimates that are not systematically off in one direction or the other.

Of course, in most surveys we sample much more than two people. The larger we make the sample size, the closer we expect the sample mean in any specific sample to be to the population mean (to see this, compare what we would expect with a sample of just two people – where the sample mean could be very high if we happened to sample two very rich people – versus a sample consisting of the whole population minus one person – where we expect the sample mean to be very, very close to the population mean). This is called the Law of Large Numbers and it is what underlies the long form Census. With a sample of 20% of the population chosen at random, we can expect that the sample mean income will be very close to the true population mean.

What happens with a voluntary survey? As we will discuss in a moment, we know that a sizeable proportion of people do not answer such surveys. If people refuse to answer at random then non-response would not cause a problem. We just wouldn't get as large a sample. We would expect more variation in the mean incomes we calculate but they wouldn't be systematically different from the population mean. But if there is something systematic in non-response (if, say, poor people respond less than rich people) then there will be biases. It is as if we are sampling from a different population: rather than sampling from the population of all adults in Canada, we are sampling from the population of all adults who are ready to respond to a survey. Increasing the sample size won't fix the fact that we are sampling from a systematically different population. In fact, we could send out the voluntary survey to everyone in Canada and it wouldn't fix this problem.

So if there is systematic bias in the response rates across groups in Canadian society then the claim made by Industry Minister Tony Clement that the accuracy of a voluntary survey is preserved by sampling more households is wrong. It is an incorrect application of the Law of Large Numbers. Former Chief Statistician Munir Sheikh made this point famously in his resignation letter with his definitive statement about the claim that voluntary surveys can substitute for a mandatory census. "It can not."(Chase and Grant(2010)).[1]

How pervasive is non-response to voluntary surveys? Response rates in voluntary surveys conducted by Statistics Canada are presently in the range of 60 to 70 percent. To take one example, the Survey of Household Spending, which is used as part of the construction of the Consumer Price Index, had a response rate of 63.4% in 2008. Importantly, as is shown in Figure 1, this response rate has fallen substantially since the 1990s, and this trend shows no sign of abating. This example is representative of the broad patterns in response rate levels and trends--both in Canada and abroad.

Crucially, there is strong evidence that survey non-response is non-random. That is, certain groups systematically are less likely to respond. There are many examples of this fact, but we focus here on one example coming out of our own research. In two studies (Frenette et al(2006) and Frenette et al(2009)), we examined trends in family income inequality over the last three decades in Canada. In Frenette et al(2006), we compared measures of the income distribution based on three different sources. The first was a combination of the Survey of Consumer

[1] This is a statement from Munir Sheikh's resignation letter as Chief Statistician. The letter was originally posted on the Statistics Canada website but was subsequently removed.

Finances (SCF), which ran annually up to 1997, and the Survey of Labour and Income Dynamics (SLID) , which took over as the flagship income and labour survey after 1997. These surveys were both special surveys sent to a subset of those surveyed for the Labour Force Survey and were voluntary. Their response rates changed over time but were roughly near 80% for both. We also examined incomes using Census data and tax data. We found very substantial differences between the SCF/SLID and the other two data sources, with a tendency for the SCF/SLID to overstate incomes at the bottom of the distribution and, to some extent, under-state incomes at the top. Table 1 recreates part of Table 3.5 from Frenette et al(2006), showing the ratio of either the mean income from the SCF/SLID or from tax data to the mean income from Census data for a variety of vingtiles for 1995 (a year in which we have data from all three sources).[2]

Table 1: Ratios of Mean Income by Vingtile from Various Data Sources, 1995

Vingtile	SCF/Census	Tax Data/Census
Bottom	2.33	0.87
2nd	1.26	0.94
3rd	1.13	0.89
4th	1.08	0.87
8th	1.01	0.90
10th	1.00	0.91
12th	1.00	0.93
17th	0.98	0.97
18th	0.98	0.97
19th	0.97	0.98
Top	0.93	1.02

Source: Table 3.5, Frenette et al(1995)

The column showing the relative means obtained from Census and tax data show a relatively close agreement across the income distribution.[3] In contrast, for the poorest 5% of families, the mean income reported in the SCF is over double that in the Census (and the tax data). This problem decreases at higher incomes, such that the SCF and Census provide nearly identical mean incomes for those in the middle of the distribution. A less severe problem emerges at the top, where average income in the SCF is about 7% lower than that found in the Census. We tried to investigate the sources of these discrepancies and came to the conclusion that "a likely explanation for the discrepancy between the SCF and the other data sources is relative under-coverage at the very bottom of the income distribution in SCF (and SLID)." (Frenette et

[2] If we sort the sample by income from lowest to highest and divide it into 20 equally sized groups, each such group is called a vingtile.

[3] As a side point, our investigations in these papers focus on household income and there is some evidence that attempts to impute families in the tax data has been problematic at times. For that reason, we favoured Census over the tax data. The tax data also has issues related to incentives not to report income that are not present in the Census.

al(2006), p. 89).

The hidden ubiquity of the census

The public debate on the census has made clear the multitude of direct uses for data coming from the census long form. Housing information is used by the Canadian Mortgage and Housing Corporation to fulfill its legislative mandate, and also by local governments and private sector actors to learn about trends in housing. Information on languages is used to determine local linguistic service levels--again both by governments and others in Canadian society. Less attention has been focused on the more indirect--yet in ways even more vital--role the census plays in the Canadian statistical system.

To make this discussion more concrete, we take the example of the Labour Force Survey (LFS). The LFS is a monthly survey used to obtain key information for constructing unemployment, employment and participation rates. It is the basis for all the unemployment rate statistics reported by Statistics Canada and, so, is the basis for many policy considerations. The unemployment rates constructed from it are also used at the regional level as part of calculating eligibility for Employment Insurance. Beyond its direct uses, the LFS is made even more important because it is used as a starting point for a large proportion of voluntary surveys conducted by Statistics Canada. For example, the SLID, mentioned earlier as the primary source for labour and income data, is based on a sample drawn from the LFS.

The creation of the LFS sample starts with choosing a sampling frame.[4] A stratified sampling design is used, which means that the survey designers first cut up Canada into strata or geographically and demographically defined groups. Clusters within each stratum are then chosen, and random dwellings within these chosen clusters are contacted. Importantly, some 'special strata' are used to target specific populations of interest. These populations of interest include aboriginal peoples, immigrants, and those with high income. As our discussion in the previous sections indicated, to obtain estimates for these groups that have low variance (i.e., to take advantage of the power of the Law of Large Numbers), we need sample sizes that are as large as possible. If the group of interest is small within the population then simple random sampling will imply small numbers of sample members for that group. In response, Statistics Canada uses Long Form Census counts to determine which geographic areas have large concentrations of the targeted group (e.g., immigrants) and those areas are then over-sampled.

Beyond the stratification, further corrections to the responses can be made using population benchmarks. Once survey responses are collected, a further correction for non-response bias is made. Households are effectively sorted into groups by certain demographics. The proportion of responders in these groups is compared to a benchmark, and corrections calculated so that the responses from the survey can reflect the whole population. This approach is called “weighting” and can be understood with a simple example: imagine that all native born Canadians responded to a survey but only half of the immigrants contacted did so. We could re-create a picture of the overall population using this data if we counted each immigrant as if he or she were two people

---

[4] This discussion follows the description in Statistics Canada (2008).

(giving them a weight of two) while counting each native born person as one person (giving them a weight of one).

In the LFS, an initial set of weights are created related to groups who are known (based on comparisons to the Census) to respond to the survey at low rates. These groups include people in remote rural locations, aboriginals, and high income households. There is also a final exercise to generate weights that insure that weighted counts in the LFS for groups defined by age, gender and province match those in the most recent Census. The latter weights can be generated using just Short Form Census information but the other weights (relating, for example, to aboriginals) are based on the Long Form.

All voluntary Statistics Canada surveys come with a set of weights of this type that researchers are required to use. But constructing those weights requires having a "true" population benchmark and the Census is that benchmark. Thus, without the Census, both the stratification and weighting stages of all other surveys would be affected. For the LFS, this would mean inferior statistics on unemployment and employment. Beyond the set of surveys collected by Statistics Canada, privately collected surveys (e.g., by polling firms) must also be compared to some standard to insure they are providing unbiased statistics. Comparing them to some other voluntary survey (such as the NHS) which has its own, unknown, response biases is obviously of limited usefulness. Thus, to insure the quality of these surveys, the mandatory Census (both short and long form) is important.

Finally, it is worth emphasizing the point raised by Veall(this issue) that re-weighting is useful but is still not nearly as good as having a true census. With re-weighting, we assume that we can make up for any non-response biases by, for example, counting each immigrant as if he or she is two people. But if non-responding immigrants are fundamentally different from immigrants who respond to a survey in ways that relate to measures of interest (e.g., if non-respondent immigrants are just out of the house working when survey takers come around and so are likely higher income than respondents) then even a carefully re-weighted survey will not give a completely accurate estimate of the population. Only a Census, where we talk to a random sample of everyone will provide complete accuracy.

Government and statistics

To this point, our discussion has focused on the somewhat narrow question of whether we should be concerned about the proposal to replace the mandatory long form with the NHS. But we believe the broad public attention given to this topic represents an opportunity to open a public debate about the role of government in statistics gathering in general and the institutional form for Statistics Canada in particular. We end with some thoughts on these topics.

At the broadest level, one might reasonably ask why the government is in the business of statistics colleting at all. Why not have a market in statistical information? We can think of three main reasons for government provision of statistics:

1) Externalities. Information is a non-rivalrous good (that is, my use of information does not, in itself, reduce your ability to use it). Further, once information has been published

in the media or on-line, it is difficult to keep others from using it (i.e., it has non-excludable features). In these circumstances, it is well known that a standard market will tend not to take account of positive externalities and, hence, will under-provide the good (Samuelson(1954)).[5]

2) Economies of scale. Just as it would have been inefficient to have many firms building parallel railways, it would be inefficient to have many firms trying to survey the whole population. The economies of scale generated by the armies of survey takers that are needed would imply that something like a natural monopoly could arise. We could then regulate the monopoly or simply have government provision.
3) Co-ordination efficiencies. A market for statistical information would generate an asymmetric information problem. In particular, firms looking to purchase statistics as inputs for a contract might shop around to find the most advantageous numbers. The firm on the other side of the bargain would, similarly, find a firm to support its position, resulting in problems for contracting that would lead to inefficiencies for the economy as a whole. The fact that there are polling firms that are seen as "Conservative" or "Liberal" suggests this point is not just fanciful. One simple solution to this problem is to have the government act as a disinterested provider of statistics that everyone agrees to use and to trust. It is probably the recognition that impartially provided statistics are so important for the functioning of modern economies that has led business groups to weigh into the Census debate on the side of preserving the long form.

In addition to these general reasons, there is an added reason for government provision of censuses. As we argued earlier, the real value of a census lies in its obtaining universal response and, therefore, ultimately in its being mandatory. We suspect that most people would rather have government than a private firm wield the power to insist on compliance. Government can both draw on the 'carrot' of civic feelings and also the 'stick' of penalties for non-compliance. Because the imposition of penalties requires democratic oversight, the touch of government on the collection of census-quality data is necessary. Of course, part of the current debate is about whether we should give even the government this power. We believe that it is very reasonable to have a debate about the content of any government mandated survey – weighing benefits such as externalities and co-ordination efficiencies against the cost of intrusion.

But regardless of the reasonableness of mandating responses, the debate over the Census has highlighted our third point in favour of government provision. A very wide range of individuals and organizations in Canadian society use statistics as the basis for making contracts or for arguing their cases to their fellow citizens. Those statistics must be seen as beyond reproach for our economy and our society to function effectively. We believe that there should be a debate about the form of the relationship between the government and Statistics Canada and propose a few points to promote that discussion.

---

[5] Note that we aren't thinking of the spill-overs from using the Census as a benchmark for other surveys mentioned earlier. It seems possible this is something a private firm could charge for and Statistics Canada does charge private firms for use of its data.

First, the Chief Statistician should be more secure in his job and more arms-length than is currently the case. Thus, the Statistics Act could stipulate measures like those applying to the Governor of the Bank of Canada: that s/he be appointed by a board of directors (with these directors ultimately appointed by the government) rather than directly by the government; and that s/he serve for a fixed term and not be fireable by the minister of the day. This alone would make it more difficult for the government to tamper with the daily business of Statistics Canada. Second, that Statistics Act should be revised to state that Statistics Canada has a quasi-arms-length status (that it is not, as Minister Clement indicated, just another department reporting to a Minister). This could involve a statement that the government cannot direct day to day practices such as how statistics are collected but can demand that specific surveys needed for policy making in other departments be done. Third, following from the second point, there should be core funding for the core functions that we need Statistics Canada to carry out: the Census (since other statistics work off it); inflation; and unemployment statistics. This would make sure these key functions are completely beyond manipulation.

Conclusion

The Census is a vital, even pivotal, component of our statistical infrastructure. If the government announced ill-advised technical changes to the power grid or road system, few Canadians might notice or care initially. It is only when their electricity blacks out or a bridge fails, that the folly of the changes become clear. Similarly, the degradation of the Canadian census has impacts on Canadian society that, while perhaps not immediately clear to all Canadians, will eventually have a large impact on the quality of Canadian society. There can be tensions at times in our democracy between expert technical advice and populist instincts. We feel that an important task of experts is to inform those populist instincts with reasoned arguments and facts. Ultimately, we have faith in the good judgment of Canadians, but we can only hope their elected representatives are listening.

References:

Chase, Steven and Tavia Grant (21 July 2010). "Statistics Canada chief falls on sword over census". Globe and Mail. http://www.theglobeandmail.com/news/politics/statistics-canada-chief-falls-on-sword-over-census/article1647348/.

Coase, R. H. (1974), "The lighthouse in economics," *Journal of Law and Economics*, vol. 17, No. 2, pp. 357-376.

Frenette, Marc, David A. Green, and Garnett Picot (2004). “Rising income inequality in the 1990s: An exploration of three data sources,” Statistics Canada, Analytical Studies Research Paper Series, No. 219.

Frenette, Marc, David A. Green and Garnett Picot (2006). “Rising Income Inequality in the 1990s: An Exploration of Three Data Sources,” in David A. Green and Jonathan R. Kesselman,

eds., Dimensions of Inequality in Canada, Vancouver: University of British Columbia Press, pp. 65-100.

Frenette, Marc, David A. Green and Kevin Milligan (2007). “The tale of the tails: Canadian income inequality in the 1980s and 1990s,” Canadian Journal of Economics, Vol. 40, No. 3, pp. 734-764.

Samuelson, Paul A. (1954) “The Pure Theory of Public Expenditure,” *Review of Economics and Statistics,* Vol. 36, No. 4, pp. 387-389.

Statistics Canada, (2008) "Methodology of the Canadian Labour Force Survey," Catalogue No. 71-526-GIE. www.statcan.gc.ca/pub/**71-526-x**/**71-526-**x2007001-eng.pdf