Component of Statistics Canada Catalogue no. 11-522-X
Statistics Canada's International Symposium Series: Proceedings

Article

Symposium 2008:
Data Collection: Challenges, Achievements and New Directions

Interaction Between Data Collection and Sample Design: A Plus in a Complex Selection Process?

by Marie-Eve Tremblay and Karine Moisan

2009

Statistics Canada Statistique Canada

Canada

Proceedings of Statistics Canada Symposium 2008
Data Collection: Challenges, Achievements and New Directions

Interaction Between Data Collection and Sample Design: A Plus in a Complex Selection Process?

Marie-Eve Tremblay and Karine Moisan [1]

Abstract

The purpose of the Quebec Health and Social Services User Satisfaction Survey was to provide estimates of user satisfaction for three types of health care institutions (hospitals, medical clinics and CLSCs). Since a user could have visited one, two or all three types, and since the questionnaire could cover only one type, a procedure was established to select the type of institution at random. The selection procedure, which required variable selection probabilities, was unusual in that it was adjusted during the collection process to adapt increasingly to regional disparities in the use of health and social services.

Key Words: Sample design, Variable probabilities, Collection waves.

1. Introduction

1.1 Overview of the survey and its objectives

The Quebec Health and Social Services User Satisfaction Survey, 2006-2007, was conducted by the Institut de la statistique du Québec at the request of the Ministère de la Santé et des Services sociaux du Québec and the health and social service agencies of the various health regions in Quebec. The survey's methodology is described in full in Neill et al. (2007). The first of its kind, the survey was part of a survey program to provide the various parties involved in organizing health care services with planning data by developing a statistical portrait of users' satisfaction and expectations regarding health and social services. Specifically, the survey's main objectives were as follows:

- develop a profile of persons who consulted a health and social services professional at a hospital,[2] a medical clinic[3] or a Centre local de services communautaires[4] (CLSC) at least once in the 12 months preceding the survey;
- provide estimates of user satisfaction for the types of health care institution covered by the survey (hospitals, clinics and CLSCs) and region of residence;
- measure users' expectations according to region of residence;
- identify the most and least satisfactory elements based on a combined analysis of user satisfaction and expectations;
- collect complementary data to assist in the study of connections between user satisfaction and certain broad socio-demographic and health characteristics.

[1] Marie-Eve Tremblay, Institut de la statistique du Québec, 200, chemin Sainte-Foy, Québec City, Quebec, G1R 5T4, Canada (marie-eve.tremblay@stat.gouv.qc.ca); Karine Moisan, Institut de la statistique du Québec, 200, chemin Sainte-Foy, Québec City, Quebec, G1R 5T4, Canada (karine.moisan@stat.gouv.qc.ca).

[2] Persons who consulted a health and social services professional in a hospital's out-patient clinic or emergency ward and persons who were hospitalized were in scope, while persons who consulted a professional in a psychiatric hospital were not.

[3] Private medical clinics,– i.e., clinics where patients are charged for services, such as dental clinics, chiropractors' offices, physiotherapy and massage therapy clinics, osteopaths' offices and acupuncture clinics – were excluded from the survey.

[4] Home care services were included, and Info-Santé services were excluded.

1.2 Data collection and questionnaire

The data were collected between November 2006 and June 2007 from 38,389 respondents in 16 health regions in Quebec (this represents a weighted response rate of 56.9%). Respondents were asked to complete a questionnaire with three separate parts. Part 1 documented users' consultation profile in the 12 months preceding the survey, i.e. the number of times they consulted a health and social services professional, for themselves or a dependant,[5] at a hospital, medical clinic or CLSC. Part 2 of the questionnaire dealt with respondents' satisfaction and expectations. A total of 43 questions were used to measure satisfaction: 41 of these related to specific aspects of service quality (competence of the professional, reasonable waiting time for an appointment, cleanliness of the institution, etc.), which were grouped into three major dimensions and 12 sub-dimensions, while the other two questions were about the respondents' overall satisfaction and general opinion of Quebec's health care system. Part 3 dealt with selected health and socio-demographic characteristics of the respondents.

2. Determining the institution type's selection probabilities

2.1 Description of the sample design and the selection procedure

The survey's target population consists of all persons aged 15 and over living in private households in Quebec who consulted a health and social services professional at a hospital, medical clinic or CLSC, for themselves or a dependant, in the 12 months preceding the survey.

The survey's sample was selected through random generation of telephone numbers belonging to eligible private households. For each household contacted, a person aged 15 or over was selected at random from the household members who had consulted a health and social services professional in the previous 12 months at one of the types of health care institutions included in the survey.

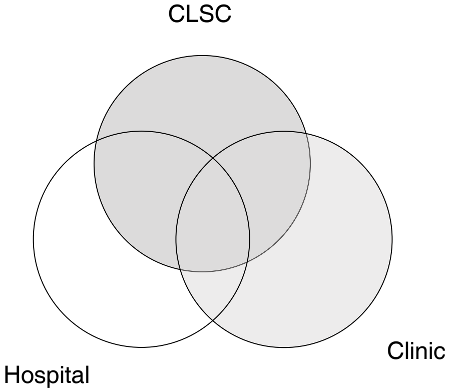
As noted above, one of the survey's objectives was to provide reliable estimates by health region and type of institution (hospital, medical clinic and CLSC). The sample design was developed in such a way that it would be possible to estimate a proportion of 15% or more with a coefficient of variation of no more than 10% for each type of institution in a region. While the region of each telephone number (and therefore of each respondent) was known in advance, the institution type was not. As a result, it was impossible to select a sample by institution type *a priori*. Moreover, since we were asking respondents to report their satisfaction level for a specific experience, that experience had to be selected at random. Response burden and time constraints made it impossible to collect information about more than one institution type per respondent. The average length of an interview for a single reference consultation was nearly 20 minutes. To identify that consultation, a supplementary selection procedure was required.

First, for a user who had more than one visit in the previous 12 months, the type of institution at which the reference consultation took place had to be determined (hospital, CLSC or clinic). Second, for the selected institution type, if the user had visits for both himself or herself and a dependant, the consultation type was selected at random with equal probabilities. Third, for the institution type and consultation type selected, the user was asked about his or her most recent consultation experience. The selection procedure described above ensured that we would have a probabilistic sample of the various institution types and consultation types.

Each respondent's consultation profile had to be established before the institution type could be selected. Since a user could have consulted a health and social services professional at one, two or all three of the survey's institution types in the previous 12 months, seven different consultation profiles were possible. All of them are represented in Figure 2.1-1.

[5] A dependant is a person aged 14 or under or a person with a disability that prevents him or her from making decisions about health care.

Figure 2.1-1
Summary of the possible consultation profiles



Source: Institut de la statistique du Québec, Quebec Health and Social Services User Satisfaction Survey, 2006-2007.

To achieve reasonably equal precision in the estimates for all three institution types in a region, the number of respondents would have to be roughly equal as well. Ideally, each institution type should be covered by one third of the questionnaires in each region. Since the proportion of persons who consulted a professional was different for each institution type, and since the proportion who visited a CLSC was much smaller than the proportions who visited a clinic or a hospital, the institution type had to be selected with varying probabilities to ensure balanced representation of the three institution types in the sample.

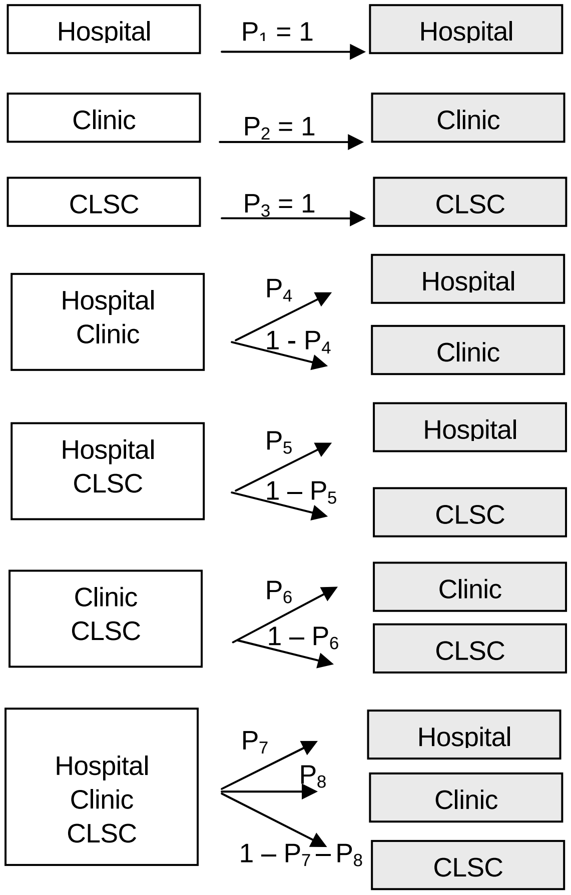
2.2 Determining the institution type's selection probabilities

The procedure for selecting the institution type is shown in Figure 2.2-1. The possible consultation profiles are in the left-hand column, and the institution type selected is in the right-hand column. Each institution type appears in four consultation profiles. The selection probabilities associated with each profile (P_1 to P_8) had to be determined before collection began.

As mentioned previously, users visited CLSCs less than the other two institution types. Consequently, CLSCs had to have a higher selection probability when they were not the only institution type in the consultation profile. By changing the selection probabilities, we can increase the number of respondents (*n*) for a particular institution type (CLSCs in this case), but the variability of the selection probabilities can result in a loss of precision. These two phenomena can be controlled with effective sample sizes ($n_{effective}$). Effective sample size is defined as $n_{effective} = \frac{n}{deff}$,where *deff* is the design effect that measures the loss of precision relative to a simple random sample design. The proportion of respondents for a given institution type from each of the four possible consultation profiles is altered when the selection probabilities are changed. The greater the difference between that distribution and the distribution in the population, the greater the design effect, which reduces the effective sample size for producing estimates. The ultimate goal of our selection procedure is to equalize, to the extent possible, the projected effective sample sizes for the three institution types for 15 health regions and thereby obtain similar levels of precision for all institution types and all regions. There was no selection process for institution type in the Nord-du-Québec region since health care services are organized differently in that region.

The selection probabilities can be determined using various criteria, which we want to optimize. Specifically, we attempt to minimize the sum of the missing $n_{effective}$ for the three institution types, minimize the differences between these missing $n_{effective}$, and maximize the number of region-institution type combinations for which the desired precision is attained.

Figure 2.2-1
Institution type selection procedure



Source: Institut de la statistique du Québec, Quebec Health and Social Services User Satisfaction Survey, 2006-2007.

Furthermore, the probabilities must be optimized under certain constraints. First, for all users who visited only one institution type in the previous 12 months, no selection could be made, since that institution was selected automatically. This can have a significant impact. For example, in the Bas-Saint-Laurent region, we found that in the population, only 17% of users who had a consultation at a hospital had not visited any other institution type, while the proportion of respondents whose reference consultation was at a hospital and who had not visited any other institution type was 54%. Second, when the probabilities were altered too much in favour of CLSCs, the effective sample sizes for clinics and hospitals became too small, and it was better to attain the precision targets for two of the three institution types than to have effective sample sizes that were similar but too small for all three types. Third, there always had to be some chance of sampling each institution type if we wanted to maintain the probabilistic character of the selection. Thus, P_4 through P_8 could not have the value 0 or 1.

Once the methodologists finished optimizing the selection probabilities, the information was passed to the collection managers so that it could be programmed into the CATI software. Table 2.2-1 shows the selection probabilities used in the first collection wave.

Table 2.2-1
Selection probabilities used in collection wave 1

	Selection probabilities (%)		
Institution type	Hospital	Clinic	CLSC
Hospital	100	0	0
Clinic	0	100	0
CLSC	0	0	100
Hospital/clinic	55	45	0
Hospital/CLSC	30	0	70
Clinic/CLSC	0	20	80
Hospital/clinic/CLSC	22	18	60

2.3 Regionalization of the selection probabilities

To facilitate collection management and field follow-up, the sample was split randomly into three batches, one for each of three collection waves, which began at different times (November 2006, January 2007 and April 2007). Wave 1 contained about 25% of the initial sample of telephone numbers, wave 2 about 45% and wave 3 about 30%.

Splitting the sample into several waves makes it possible to track various collection parameters (eligibility rate, productivity rate and response rate) and adjust them if some initial assumptions prove incorrect. In this survey, the eligibility rate – the proportion of households with at least one member aged 15 or over who had had a health care consultation in the previous 12 months – was not known precisely, and we had to be able to adjust our initial assumptions if necessary. Wave 1 was thus kept smaller than the other two so that adjustments could be made at the earliest possible stage.

Following the survey's pre-test, we had only provincial estimates of the proportions of users who had had consultations at each institution type. We knew that the proportions varied from health region to health region, and the collection waves helped us refine the parameters for selecting the institution type as the survey proceeded. For wave 1, the selection probabilities (presented in Table 2.2-1) programmed into the CATI questionnaire were based on the provincial probabilities. By analyzing the data from wave 1, we were able to develop five sets of selection probabilities (P_1 through P_8), each of which was used for between one and seven health regions with similar consultation profiles. These sets of probabilities were used in wave 2. Analysis of the wave 2 data led to the production of a set of specific probabilities for each region, which were then used for data collection in wave 3. However, this approach imposed an additional constraint on the optimization of selection probabilities, since in the preparatory work for waves 2 and 3, we had to take account of the results from the previous waves, the selection probabilities of which could no longer be modified.

By progressively adjusting the selection probabilities for regional disparities, we were better able to control the number of effective interviews completed for each region. This is a major advantage of using collection waves. We refer to this process as the regionalization of selection probabilities.

2.4 Interaction between data collection and sample design

Throughout this process of regionalizing the selection probabilities, continual interaction between the sample design and the data collection results was required. Thus, to forecast the number of respondents for each institution type, we first had to forecast the number of respondents per region (all institution types combined). The initial assumptions had to be continually adjusted on the basis of the results obtained. This was done by the Institut's collection managers, who developed tools for estimating the expected number of respondents by dividing the files into various categories by response code at the time of projection. In particular, some response codes had a greater chance of being converted than others. For example, it is reasonable to expect that by the end of the collection period, there will be proportionally fewer respondents in files coded "refusal due to household unavailability" than in files coded "firm appointment with selected person" or "agreement reached with third party". By making assumptions about the eligibility, productivity and response rates for each intermediate response code, collection managers were able to estimate the expected number of respondents for each collection wave in each health region.

An iterative process was therefore implemented so that those projections could be fully exploited in determining the institution types' selection probabilities. In wave 1, projections of the number of respondents were made as collection proceeded. Using estimated design effects for each institution type, the projections were turned into expected effective sample sizes for each region and institution type. On this basis, we were able to regionalize the selection probabilities for wave 2. The process was then repeated from beginning to end for wave 3. The projections produced at the end of wave 3 could not be used to refine selection probabilities since collection was coming to a close; instead, they were used to make decisions about managing the end of the collection process. Through this approach, regions for which the targets for all three institution types seemed attainable with a little extra effort were given priority over other regions where one of the institution types was so underrepresented that it would have taken a very large number of respondents to reach the targets. To minimize their impact and avoid introducing biases, such decisions were made only in the last few weeks of the nearly seven-month collection process. The entire exercise required close cooperation between methodologists and collection managers.

3. Results

3.1 Results obtained

The average design effect obtained at the regional level by using variable selection probabilities was 1.9 for clinics, 1.7 for hospitals and 1.3 for CLSCs. The target was about 500 effective respondents for each institution type in each health region, and that target was achieved in 38 out of 45 cases. In some regions, such as Saguenay-Lac-Saint-Jean, the proportions of CLSC users were so low (about 27%) that the targets for CLSCs could not be attained with variable selection probabilities and regionalization of those probabilities. In Saguenay-Lac-Saint-Jean, there were only 370 effective respondents for CLSCs, compared with 648 for hospitals and 608 for clinics. In regions with higher proportions of CLSC users, the results were much more useful. In the Laval region, for example, the proportion of CLSC users was 44%, and the number of effective respondents was 502 for hospitals, 544 for clinics and 543 for CLSCs.

3.2 Benefit of regionalizing the selection probabilities

A simulation was carried out to determine whether adjusting the selection probabilities during collection was beneficial. It consisted in estimating the effective sample sizes that would have obtained if the provincial selection probabilities had been used throughout the collection process, i.e. if the parameters had not been regionalized. The results of the simulation are presented in Figures 3.2-1, 3.2-2 and 3.2-3. The histograms provide a comparison of the actual effective sample sizes and the ones obtained in the simulation.

Figure 3.2-1
Effective sample sizes obtained in the survey and by maintaining the provincial selection probabilities throughout collection for CLSCs

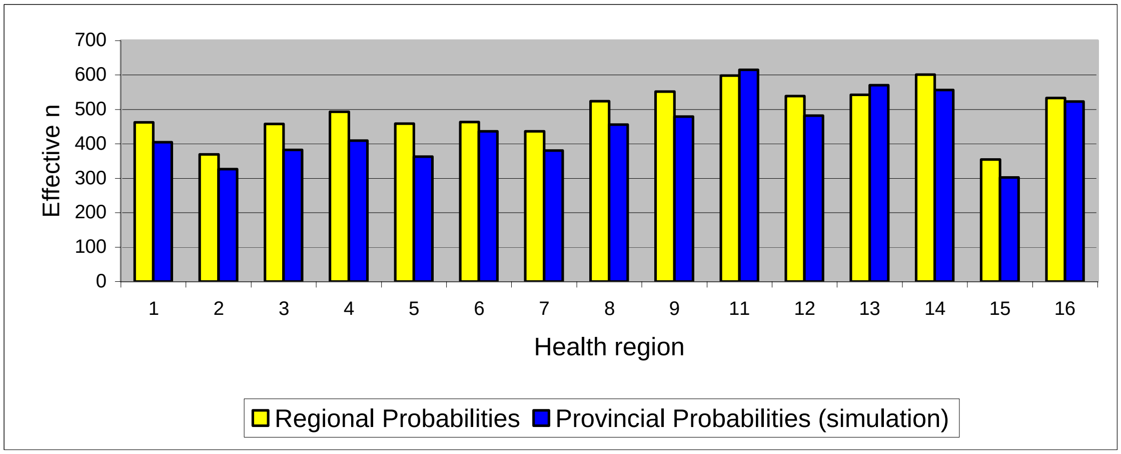


Figure 3.2-2
Effective sample sizes obtained in the survey and by maintaining the provincial selection probabilities throughout collection for hospitals

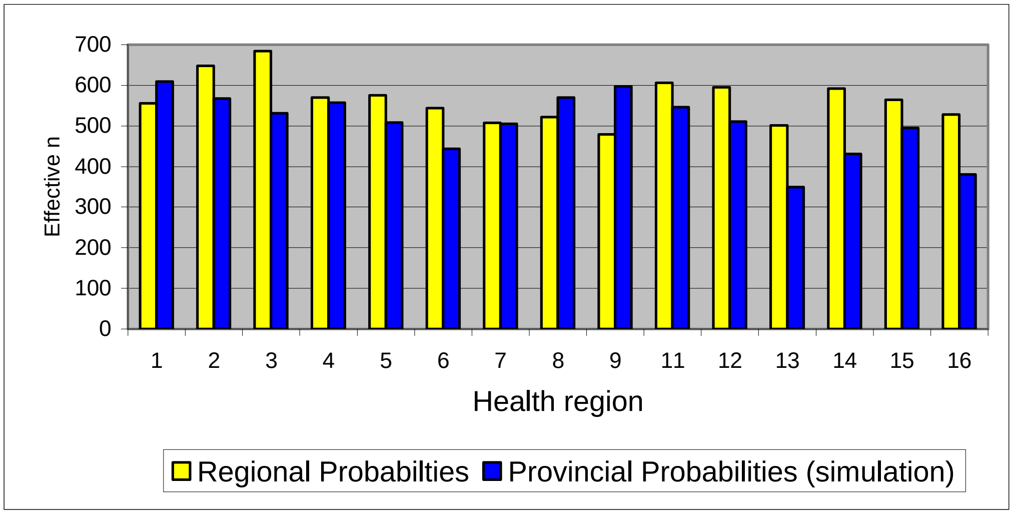
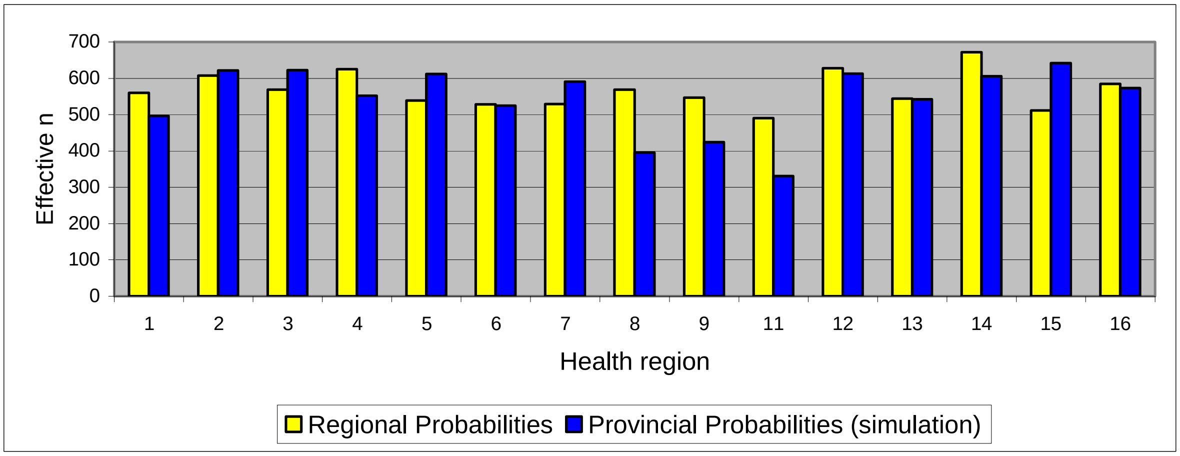


Figure 3.2-3
Effective sample sizes obtained in the survey and by maintaining the provincial selection probabilities throughout collection for clinics



As noted previously, one of the objectives was to secure enough respondents whose reference consultation was at a CLSC and obtain effective sample sizes that were as similar as possible for the three institution types, so that estimates of similar precision could be produced. Table 3.2-1 confirms that, on average, regional effective sample sizes increased for all institution types, but especially for CLSCs, which was the main goal of the selection procedure. The table also shows that regionalizing the selection probabilities helped bring the effective sample sizes for the regions closer together, since the standard deviations are lower for the regional selection probabilities. This leads to the conclusion that while the gains were not spectacular, they made a significant contribution to meeting the objectives.

Table 3.2-1
Average and standard deviation of effective sample sizes by institution type and regional and provincial selection probabilities

	Average (standard deviation) of the $n_{effective}$		
Selection probabilities	CLSC	Hospital	Clinic
Regional	493 (73)	565 (55)	567 (49)
Provincial (simulation)	446 (92)	507 (76)	543 (94)

4. Conclusion

More generally, the approach described in this paper can be used in various situations, in particular when a survey targets various subpopulations with different prevalences in the population. The use of variable selection probabilities makes it possible to obtain enough respondents who belong to each subpopulation to produce estimates for each subpopulation. In addition, using collection data to regionalize certain parameters may prove to be a very useful technique when there is insufficient reliable information at the outset for a detailed geography. With the interaction between sample design and data collection that is necessary in such a process, methodologists become more heavily involved in data collection and collection managers make a greater contribution to the sample design.

References

Neill, G., Tremblay, M.-E., Végiard, S., Lavoie, A. and Moisan, K. (2007). Enquête sur la satisfaction des usagers à l’égard des services de santé et des services sociaux du Québec, 2006-2007: description et méthodologie. Québec, Canada : Institut de la statistique du Québec.