Modelling and Simulation of Survey Collection Using Paradata

Kristen Couture[1], Yves Bélanger[2], Elisabeth Neusy[3]
Statistics Canada, R.H. Coats, 18A, Tunney's Pasture, Ottawa, ON, Canada, K1A 0T6

Abstract
In recent years, the cost of survey collection has grown significantly and nonresponse has increased. To counter these trends, strategies are being studied to optimize collection activities, resulting in a more time efficient and cost effective survey collection process. For example, recent initiatives for Computer Assisted Telephone Interviewing surveys at Statistics Canada include experimenting with time slices, limiting the number of calls, and establishing calling priorities. Field testing new procedures however has its drawbacks: it is costly and it is difficult to control, which can render the results difficult to interpret. To address these issues, we describe in this paper the creation of a microsimulation system of the collection process which uses paradata as input. We discuss characteristics of the model as well as results of simulation runs with various parameters.

Key Words: microsimulation, paradata, data collection, time slices

1. Introduction

Of all the activities associated with surveys, data collection represents one of the largest segments of the global budget. For a number of years now, a constant increase in collection costs, combined with the gradual decrease in response rates has been observed at Statistics Canada and elsewhere in the world (Curtin, Presser and Singer 2000, de Leeuw and de Heer 2002). Various strategies have been employed during the collection phase in an attempt to reverse these trends at Statistics Canada. Among them are the adoption of a limit on the number of calls, which is intended to make better use of resources, as well as the establishment of time slices for calls, which leads to better distribution of calls throughout the day. Other strategies being investigated include using the best time to call (provided by the respondent or obtained through modelling), as well as experimenting with calling priorities. The common goal for all these strategies is to achieve a more efficient collection process.

Ideally, adoption of a new measure should take place in the context of a real-time experiment involving a group that is subject to the new measure and a control group that is not (see a practical example in Laflamme and Karaganis 2010 or more theoretical aspects in van den Brakel and Renssen 2005). By comparing the results obtained for the two groups, we could be able to isolate the impact of the new measure. In practice, however, we rarely have the luxury of using a control group, which makes it more difficult to evaluate the impact of any new measure. We can attempt to compare the survey results with those of a preceding cycle of the same survey, but we can never be completely sure we have fully isolated the impact of the new measure being introduced.

And even in cases where the design of the experiment includes a control group, the fact remains that the entire exercise, from design to results analysis, may take a lot of time and be quite costly, without any guarantee of meaningful results.

As an alternative to real-time field experiments, and in order to address their limitations, we are proposing in this article the use of a microsimulation system for Computer Assisted Telephone Interviewing (CATI) survey collection. Microsimulation is a modelling technique that operates at the level of individual units and is used to simulate large representative populations of these units. In the context of CATI surveys, units are sampled telephone numbers, also known as cases. The proposed system involves the following elements: the cases, the servers (interviewers), the waiting queues for cases yet to be interviewed, the calls and their results, the rules governing priorities and the flow of cases in the system.

This paper begins with an overview of the microsimulation system. We then describe the simulation model created to represent the CATI system currently in use at Statistics Canada. Next, the simulation model is validated by comparing simulation output to that of a real survey. Finally, we present the preliminary results of some simulations.

2. Overview of the Microsimulation

There are several components that go into constructing a microsimulation of CATI collection as shown in Figure 1. A central component is the computer simulation model, which requires input parameters in order to replicate the collection process as accurately as possible. As shown in the diagram, two types of parameters are entered into the simulation: model parameters and user defined parameters. Model parameters are determined prior to performing any simulation runs and are calculated from pre-existing survey data. The calculated model parameters are used to assign call outcomes and call duration, and are described in Section 2.2. In comparison, user defined parameters can be changed prior to each simulation run in order to control and manage the collection process as outlined in Section 2.3.

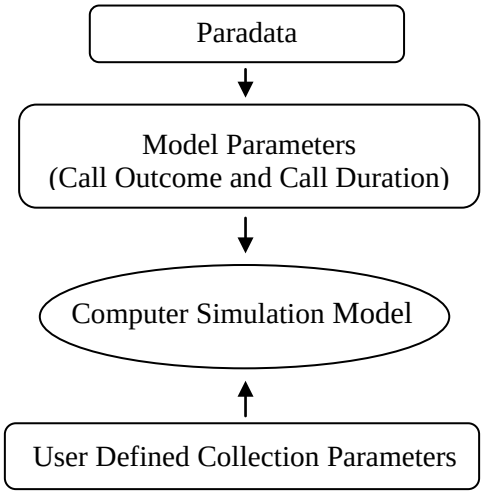

Figure 1: Overview of Building a Microsimulation of Data Collection

2.1 Description of the Simulation Model

As mentioned above, the simulation model is built to replicate CATI collection. The logical sequence of the simulation model is as follows.

At the start of the simulation, a batch of cases is created, representing telephone numbers which need to be called to conduct interviews. Once created, the cases are sent into a "calling" queue where they will wait until an interviewer resource becomes available. When an interviewer resource becomes available, the next available case is taken from the queue and a call outcome is assigned to the case. The interviewer resource will remain busy for a simulated duration of time depending on which call outcome has been assigned to the case.

The outcome assigned to the case determines whether the case receives a "finalized" or "in progress" status. Cases that receive a finalized status are sent out of the system; otherwise, the case receives an "in progress" status and is returned to the "calling" queue. Calls that result in a successfully completed questionnaire or in an out of scope are examples of outcomes that produce a finalized status. Calls that result in no contact or a busy signal are examples of outcomes that produce an in progress status.

The simulation will continue to run until either the collection period has finished or all cases have resulted in a finalized status. At this point, a file is produced containing the transaction history of each call attempt.

A diagram outlining the flow of the simulation model is shown in Figure 2.

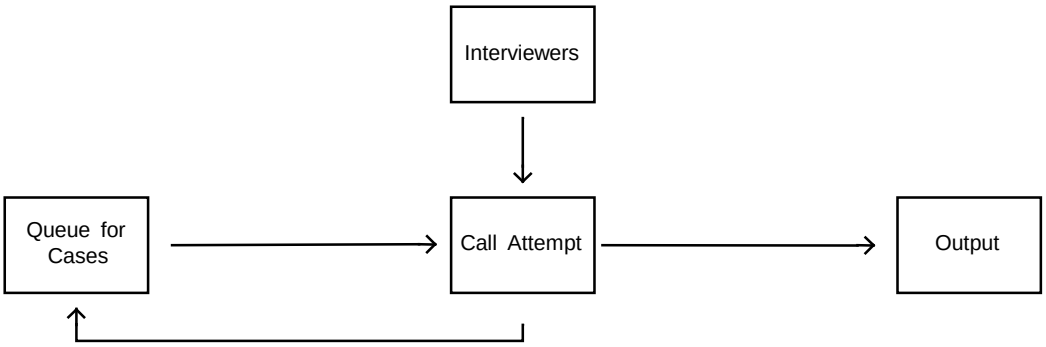

Figure 2: Flow of the Simulation Model

2.2 Model Parameters

Similar to other simulation models (such as Demosim, as described in Statistics Canada 2010), several parameters are calculated prior to any scenarios being run. Our current simulation model requires two sets of model parameters.

The first set of model parameters is used by the simulation to assign outcomes to call attempts. In actual collection, telephone calls result in a variety of possible outcomes, such as completed questionnaire, out of scope, answering machine, or busy signal. The probability of the various call outcomes depend on a number of different factors. In order to simulate data collection as realistically as possible, the probabilities of the various call outcomes are modeled using actual data. As the model's explanatory variables can only

be variables that are available prior to the call being made, we have three types of explanatory variables:

1) Frame variables (e.g. age, sex of the selected respondent, household composition);
2) Characteristics of the call attempt itself (e.g. weekend vs. weekday, time of day);
3) Call history of the case (e.g. number of previous call attempts, previous refusals).

The latter two types of explanatory variables are obtained from a survey's collection process data or paradata.

The probabilities for the call outcomes are modeled using multinomial logistic regression models. For example, suppose that there are k+1 possible outcomes, and that the probability of each outcome is denoted by p_1, p_2, … p_{k+1}. As well, suppose that each call attempt has n characteristics denoted by variables x_1, x_2,…, x_n. The vector of n variables are explanatory variables of the three types described above. The outcome probabilities are modeled using the following multinomial logistic regression model:

$$\log\left(\frac{p_j}{p_{k+1}}\right) = \sum_{i=1}^{n} \beta_{ij} x_i, \quad \text{for } j = 1,...,k, \quad \text{where} \sum_{j=1}^{k+1} p_j = 1.$$

The parameters estimated from the model, $\hat{\beta}_{ij}$, i=1,…,n, and j=1,…,k are used in the simulation. Whenever a call attempt is made for a case during the simulation, the current values of x_1, x_2,…, x_n, and the estimated parameters are used to calculate estimated outcome probabilities, $\hat{p}_1$, $\hat{p}_2$, … $\hat{p}_{k+1}$, as follows:

$$\hat{p}_j = \frac{\exp\left(\sum_{i=1}^{n} \hat{\beta}_{ij} x_i\right)}{1 + \sum_{j=1}^{k} \exp\left(\sum_{i=1}^{n} \hat{\beta}_{ij} x_i\right)}, \quad \text{for } j = 1,...,k, \quad \text{and} \quad \hat{p}_{k+1} = \frac{1}{1 + \sum_{j=1}^{k} \exp\left(\sum_{i=1}^{n} \hat{\beta}_{ij} x_i\right)}.$$

These probabilities are then used to randomly select one of the possible k+1 outcomes for the call. This is done by generating a random number (denoted r) from a uniform distribution between 0 and 1. The call is assigned an outcome of L if $r \le \hat{p}_1$ for L=1 or $\sum_{i=1}^{L-1} \hat{p}_i < r \le \sum_{i=1}^{L} \hat{p}_i$ for 1<L≤k+1.

A second set of model parameters is used by the simulation to assign the duration of the call. In actual collection, the duration depends on the call outcome. For example, an attempt which resulted in a fully completed questionnaire will tend to have a longer duration than an attempt where no contact was made.

Paradata are used to model call duration. The distribution of the call durations obtained during collection for each outcome is fit to a theoretical distribution supported by the simulation software, such as a normal distribution. The parameters of the distribution are imported into the simulation system and used to randomly assign the duration of each call attempt.

2.3 User Defined Parameters

We differentiate user defined parameters from model parameters as follows. User defined parameters can be changed prior to each simulation run, allowing the user to pose and answer "what if" type questions (Wolfson, 1995). For instance, the user is able to set an interviewer agenda for the entire collection period, specifying the number of interviewers available per shift and the duration of each shift. A "what if" question involving the interviewer parameter is "What if the distribution of interviewers was changed?" We can answer this question by running different scenarios where we vary the interviewer agenda to see which one gives the highest response rate, for instance. Overall, user defined parameters allow the user to experiment with combinations of parameters to determine which scenario will help improve collection efficiency.

Other parameters that the user controls are the number of cases created at the beginning of the simulation, the priorities of the cases in the queue and the duration of the collection period. There is also an option to place a limit on the total number of attempts that can be made per case. In addition to this, time slice parameters can be implemented. These control the number of attempts made at different times of the day and days of the week to ensure calls are properly distributed. The user specifies each time slice definition and the maximum number of attempts within each time slice.

3. Implementation of the Simulation Model

The simulation model discussed in this paper was built in SAS Simulation Studio© 1.5 (SAS Institute Inc. 2009) to represent the CATI process at Statistics Canada. Currently, the model has been built to handle Random Digit Dialling (RDD) surveys, but can be altered to handle other types of surveys as well.

A Blaise Transaction History (BTH) file from an RDD survey conducted in 2004 at Statistics Canada, the Canada Survey of Giving, Volunteering and Participating (CSGVP) was used as the paradata to build the models for call outcomes and duration. The BTH file contains one record for each call attempt made during collection with information such as the start and end time of the attempt, and a code which describes the outcome of the call.

An outcome variable was created by classifying the BTH outcome codes into one of five categories: respondent (or fully completed questionnaire), out of scope, unresolved, refusal, and other contact. A multinomial logistic regression model was constructed using the BTH data to model the probabilities of these five outcomes. Seven explanatory variables were entered in the model; these are listed below, grouped into the three categories outlined in Section 2.2:

<u>Group One: Frame variables</u>
1) Residential Status. Since the survey we are simulating is an RDD survey, the telephone numbers are classified as either residential or of unknown status. A residential status means that it is known, prior to collection, that the telephone number is associated with a household. The variable indicating whether the status of the number is residential or unknown is available on the frame for each case.

<u>Group Two: Time of call variables</u>
2) Afternoon. A binary variable indicating if the call is made from 12pm – 5pm

3) Evening. A binary variable indicating if the call is made from 5pm – 9pm
4) Weekend. A binary variable indicating if the call is made on a Saturday or Sunday

Group Three: Call history variables
5) Previous Refusal. A binary variable indicating if there was at least one refusal in previous call attempts
6) Previous Contact. A binary variable indicating if there was at least one contact other than refusal in previous call attempts
7) Previous Unresolved. A binary variable indicating if there were no refusals and no contacts in previous call attempts. In other words, the call history contains only unresolved outcomes.

Note that on the first attempt, the call history variables are zero as there is no call history available.

Once the call outcome model was created, the call duration was modelled by fitting distributions to the data by call outcome as described in Section 2.2. The parameters from the call outcome and the call duration modelling procedures were entered into the simulation model.

Since we are simulating an RDD survey, each case was randomly given a residential status. We attempted to follow the CSGVP BTH data as closely as possible, so that 67% of the cases were initially given a residential status and the other 33% were initialized as unknown status. If at any point during the simulation, any form of contact was made, the case status was changed from unknown to residential, as was done in the CSGVP collection.

As the cases flow through the simulation system, and receive a call outcome, they are grouped into one of two categories: finalized or in progress. As mentioned earlier, finalized cases are sent out of the system and in progress cases are returned to the queue. A case is considered finalized if it results in an outcome of respondent or out of scope or if there are three refusals in the call history. If the user has specified a limit on the number of call attempts, the case would be finalized once it reaches its cap.

4. Validating the Simulation

As mentioned in the previous section, CSGVP paradata was used to build the call outcome and call duration models. To verify that the simulation is functioning as expected, we compared the simulation output to that of the CSGVP survey. The validation was based on the proportion of finalized calls, the response rate, and the distribution of call outcomes and call duration for each outcome. Note that the results shown below are derived from one simulation run. However, at the time of writing, several runs had been completed and yielded similar results.

The simulation user defined parameters were as follows. Three interviewer shifts were implemented each consisting of ten interviewers for a duration of 240 minutes, resulting in 12 hour collection days. Thirty days of collection were simulated. No restrictions were made on the total number of call attempts for each telephone number since we want the simulation to replicate CSGVP as much as possible and for the CSGVP, no limits were in place.

Note that the user defined parameter values (such as length of collection and interviewer agenda) for the simulation runs were not chosen to exactly replicate those of the CSGVP, but were chosen only to ensure that the main results were reasonably comparable. The response rate was calculated as follows:

$$\text{Response Rate} = \frac{\text{\# responses}}{\text{\# cases} - \text{\# out of scope}}$$

Table 1 compares the results from the simulation to the CSGVP survey paradata. The response rate for the simulation at 48% is close to the response rate of the CSGVP survey at 47% with a similar proportion of finalized cases.

Table 1: Comparison of Response Rate

	# of Cases	*Length of Collection*	*% of Finalized Cases*	*Response Rate*
CSGVP	90,718	90 days	66%	47%
Simulation	10,000	30 days	71%	48%

Table 2 compares the distribution of all call attempts made during the collection period for the simulation and the CSGVP. The distribution of calls across all five outcomes for the simulation closely resembles the distribution of the CSGVP validating that the simulation model is effective at simulating these aspects of the collection process.

Table 2: Comparison of All Call Attempts

Outcome	*CSGVP*	*Simulation*
Unresolved	31%	33%
Out of Scope	6%	6%
Refusal	5%	6%
Respondent	6%	6%
Other Contact	52%	50%
Total	100%	100%

It was important to look at the distribution of the final outcome of each call attempt, as shown in Table 3. The five outcomes have been grouped into four categories: response, out of scope, finalized refusals, and cases that are still in progress. When comparing the distribution of final call outcomes from the simulation to the CSGVP, they are very similar, once again reinforcing that our simulation is functioning as expected.

Table 3: Comparison of Final Outcome

Outcome	*CSGVP*	*Simulation*
Response	31%	32%
Out of Scope	33%	34%
Finalized Refusals	2%	5%
In Progress	33%	29%
Total	100%	100%

5. Simulation Results

Now that the simulation model has been validated, we can run some scenarios and study the results. As mentioned earlier, simulation models can be used to test the effect of varying user defined parameters in order to answer 'what if' type questions. Questions that we want to answer are ones that will help to improve collection efficiency such as "What would be the best way to distribute interviewers in order to achieve higher response rates?"

The following two examples are used to demonstrate how, through the use of our simulation model, users can efficiently isolate and test the impact of different collection strategies in order to find the most optimal in terms of efficiency.

Example 1: Changing the distribution of interviewers throughout the day
Example 2: Changing the distribution of interviewers throughout the day combined with different time slices

5.1 Example 1

The set-up for Example 1 is as follows:

1) The collection day was split into three time periods each four hours in duration:

Morning: 9:00am – 1:00pm
Afternoon: 1:00pm – 5:00pm
Evening: 5:00pm – 9:00pm

2) A fixed total of 30 interviewers per day were allocated to the different time periods. One possible allocation of the interviewers is shown in Table 4.

Table 4: Comparison of All Call Attempts

Time period	*Number of Interviewers*
Morning (9am – 1pm)	4
Afternoon (1pm-5pm)	4
Evening (5pm-9pm)	22
Fixed Total	30

3) Scenarios were run to study the impact on response rate when altering the proportion of interviewers available in the evening. We altered the number of interviewers available in the evening from zero to 30 incrementing by two as to keep the number of interviewers in the morning and afternoon shifts equal. Note that the fixed total of 30 interviewers remained constant across all simulation runs.

The results for Example 1 are displayed in Figure 3. From this graph, the trend is that as the number of interviewers in the evening increases, so does the response rate. This trend is what we would expect to observe in real collection, assuming that respondents are

more accessible during this time period. This simple example suggests that more interviewers be allocated to the evening shift.

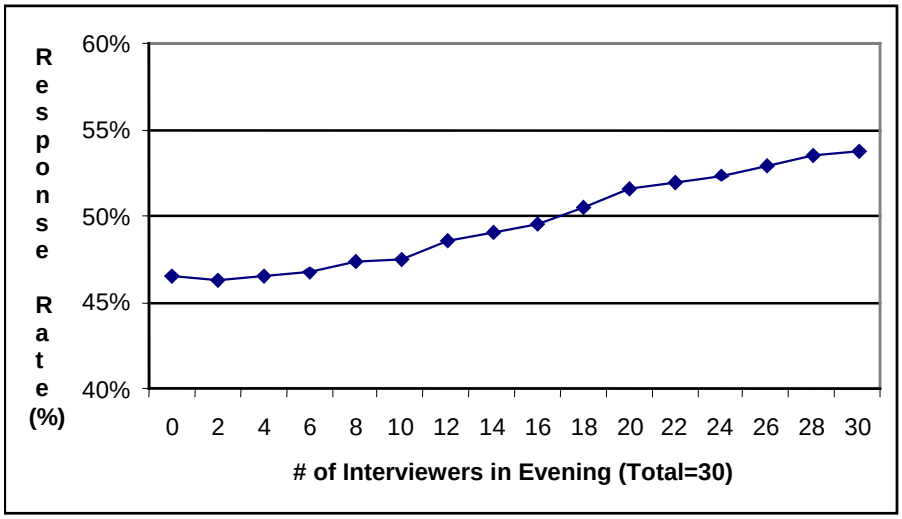

Figure 3: Impact on Response Rate when Changing Concentration of Interviewers in Evening

5.2 Example 2

In this example, we show how a user has the flexibility to change multiple parameters at once to observe the impact on the response rate. The set-up is similar to Example 1 where a fixed number of 30 interviewers per day are allocated to the three time periods. However, we now introduce a limit on the total number of attempts that can be made per case. For telephone numbers of unknown status, there are a total of five call attempts allowed and for known residential numbers there are a total of 20 call attempts allowed, which are the typical call limits now in use at Statistics Canada. The total number of attempts allowed per case is then allocated to the three time periods so that each time period has a maximum number of attempts. Table 5 shows an example of how the interviewers and call attempts could be allocated.

Table 5: One Possible Allocation of Interviewers and Attempts for a Known Residential Telephone Number

Time period	*# of Interviewers*	*Max # of Attempts*
Morning	4	2
Afternoon	4	2
Evening	22	16
Fixed Total	30	20

Four different scenarios were run:

Scenario 1: The majority (~75%) of attempts and interviewers were allocated to the morning/afternoon shifts
Scenario 2: The majority (~75%) of attempts and interviewers were allocated to the evening shift

Scenario 3: The majority (~75%) of attempts were allocated to the morning/afternoon shift, but the majority (~75%) of interviewers were allocated to the evening shift
Scenario 4: The majority (~75%) of attempts were allocated to the evening shift, but the majority (~75%) of interviewers were allocated to the morning/afternoon shift

The results from the simulation runs for Example 2 are shown in Table 6. As expected, the highest response rate (52%) occurs when the majority of attempts and interviewers are placed in the evening as compared to when the majority of the attempts and interviewers are in the morning/afternoon where a response rate of 47% is obtained.

On the other hand, if the distribution of attempts and the distribution of interviewers are not coordinated (i.e. placing the majority of the interviewers in the morning, but allowing the majority of the attempts to be made in the evening), the lowest response rates occur. This result implies that the distribution of attempts and the distribution of interviewers need to be coordinated with each other in order to achieve better response rates.

Table 6: Response Rates Obtained when Changing Distribution of Interviewers and Number of Attempts Simulataneously

Time Period with the Majority of Interviewers	*Time Period with Majority of Attempts Permitted*	
	Morning/Afternoon	*Evening*
Morning/Afternoon	47%	37%
Evening	42%	52%

6. Concluding Remarks

The implementation of strategies to improve collection efficiency involves a considerable amount of money, effort and time. As a result, a microsimulation of data collection was proposed as an alternative to real-time field experiments.

Currently, the simulation model we have built includes a few basic features, which give the user the option to change one or more of the parameters and observe the results. Two examples were presented which demonstrate how, through the use of the simulation model, users can isolate and test the impact of different collection scenarios efficiently and effectively. The results obtained from these examples reflect what we would expect to observe in a real collection process confirming that our simulation model is functioning as expected.

The simulation model presented in this paper is still a preliminary model. The main objective was to build a simple model and verify that it produces reasonable results. This preliminary model provides the basis for the future work outlined in the next section.

7. Future Work

Future work includes improving the multinomial logistic regression model used to model call outcome probabilities by adding more explanatory variables and outcomes. Furthermore, we would like to build models using paradata from different surveys and survey occasions to assess the robustness of the models across surveys. We would also like to add more complicated collection procedures to the simulation model in order to have the simulation model better reflect what occurs in the collection process. Features such as matching cases to interviewers based on their characteristics (e.g., spoken

language), implementing scheduled appointments and best time to call procedures will be considered in future work.

Acknowledgements

The authors would like to thank their colleagues Marcelle Tremblay, Hélène Bérard and Karla Fox for reviewing the paper, and providing valuable comments and suggestions.

References

Curtin, R., Presser, S., and Singer, E. (2000). The effects of response rate changes on the index of consumer sentiment. Public Opinion Quarterly, 64(4), 413-428.

de Leeuw, E., and de Heer, W. (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. In R. Groves, D. Dillman, J. Eltinge, & R. J. A. Little (Eds.), Survey nonresponse (p. 41-54). New York: Wiley.

Laflamme, F. and Karaganis, M. (2010), "Development and Implementation of Responsive Design for CATI Surveys at Statistics Canada", presented at the European Quality Conference, Helsinki, Finland.

Peytchev, A., Riley, S., Rosen, J., Murphy, J. and Lindblad, M (2010). Reduction of Nonresponse Bias in Surveys through Case Prioritization. *Survey Research Methods*, 4-1, 21-29.

SAS Institute Inc. (2009). *SAS© Simulation Studio 1.5: User's Guide.* Cary, NC: SAS Institute Inc.

Statistics Canada (2010). Projections of the Diversity of the Canadian Population 2006 to 2031. Statistics Canada Catalogue no. 91-551-X, Ottawa.

van den Brakel, J. A. and Renssen, R. H. (2005), Analysis of Experiments Embedded in Complex Sampling Designs. *Survey Methodology*, 31-1, 23-40.

Wolfson, Michael C (1995). Socio-Economic Statistics and Public Policy: A New Role for Microsimulation Modelling. *Proceedings of the 50th Session of the International Statistical Institute*.