RESEARCH REPORT SERIES
(Computing #2008-1)

Stochastic Simulation of Field Operations in Surveys

Bor-Chung Chen*

Statistical Research Division
U.S. Census Bureau
Washington, DC 20233

*Now with the Department of Transportation.

Report Issued: February 19, 2008

Stochastic Simulation of Field Operations in Surveys *

Bor-Chung Chen
Statistical Research Division
U.S. Bureau of the Census

January 16, 2008

Abstract

Discrete-event simulation modeling has become the most commonly used tool for performance evaluation of stochastic dynamic systems in science and engineering. The field operations of surveys can be classified as one of these stochastic dynamic systems. This paper describes the simulation and modeling of simplified field operations for NHIS (National Health Interview Survey). In particular, we apply the simulation and modeling methodology to the field operations. We use the 2004 NHIS CHI (Contact History Instrument) data for the input modeling of the simulation. The input modeling methods are also described in this paper. From this study, we have shown that the simulation model can be used for optimizing the field operations by setting the controllable parameters before a decision is made and implemented. The cost savings might be enormous and would not be at the expense of the response rate.

1 Introduction

Discrete-event simulation modeling has become the most commonly used tool for performance evaluation of stochastic dynamic systems in science and engineering, including such complex systems as manufacturing and material handling systems ([17] and [20]), logistics and transportation systems ([14]), healthcare and service systems ([15] and [11]), computer and communication systems ([9]). These applications of simulation modeling are results of significant achievements in electronic and computer technologies that have led to broad proliferation of powerful computers and computer networks, and significant achievements in software technology, that have resulted in simple but very efficient human-computer interfaces. However, no technological innovation can release simulators from their responsibility of ensuring that their simulation experiments produce credible final results.

In this paper, we will describe how to use the simulation techniques for the field operations application of a national household survey. We will discuss main problems and solutions of quantitative stochastic discrete-event simulation, i.e. the stochastic simulation in which the emphasis is put on statistical correctness of the final results. Whole spectrum of the problems will be covered: from generators of uniformly distributed pseudo-random numbers, which play the role of original sources of randomness in stochastic simulation, to methods of generation of system variables, such as interview length, contact time, in field representative's visits of sample households.

At the U.S. Census Bureau, the mission of the Field Division is to collect quality data at the right time for the lowest cost. Therefore, there is a need to have a valid method of predicting cost, response rates, and timing of new or continuing surveys for the field operations (per discussions with Bitzer ([2]) and others). This project is intended to develop such a method.

In complex field operations for a household survey or census, the scheduling function is typically concerned with determining the starting time and the sequence of visiting the cases assigned to the interviewers in which system performance is to be optimized. The system performance is defined as controlling the cost and timing and maximizing the response rates. The complexity and practical importance of the field operations scheduling problem has motivated the development of models appropriate for a broad range of surveys and censuses, and has focused attention on the impact of scheduling decisions on contact time and travel time. Most importantly, the model would provide a tool for predicting costs, response rates, and timing before the survey begins.

*This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

Currently, regression analysis is used on the data set from the Field Division's CARMN (Cost And Response Management Network) and Population Division's Planning Data Base ([16]) to explore survey-related cost drivers for the CPS (Current Population Survey). Also, Shimizu and Lan [19] use a simplified overall cost model based on the NHIS multistage sample.

For example, one of the system performance measures is the cost. If TDC is the total direct cost of the operation, then

$$TDC = \sum_{i=1}^{r} \sum_{j=1}^{d} C_{ij}, \tag{1}$$

where C_{ij} is the daily cost of FR (Field Representative or interviewer) i on jth day, r is the number of FRs, and d is the number of days.

$$C_{ij} = (H_{ij} \times R_{ij}) + (M_{ij} \times P_{ij}), \tag{2}$$

$$\begin{aligned} H_{ij} &= \text{ Total Hours} \\ R_{ij} &= \text{ Hourly Rate} \\ M_{ij} &= \text{ Total Mileage} \\ P_{ij} &= \text{ Reimbursement Per Mile.} \end{aligned}$$

To accurately predict the total time spent for each FR each day, we further decompose the total hours, H_{ij} into traveling time, contact time, interview time, and so on. Each time segment has its own statistical distribution. For example, the traveling time follows a statistical distribution to be determined given the traveling distance. We will describe how to obtain the statistical distribution for each time segment in the next section.

The objective of the project is to build a model that will take the inputs of workload, staffing, traveling time, productivity, etc., for the field operations of a survey. The output of the model will be the cost, response rates, and timing to complete the operations based on the assumptions given to the input of the model. The performance measures identified are, therefore, the cost, response rates, and timing. The final goal of the field operations is to have Low cost, High response rates, and Short timing; it is referred to as *LHS*.

The current version of the model contains about 1888 lines of C++ code. The compiler used is Microsoft Visual Studio .NET 2003. Section 2 gives a brief description of the simulation modeling concept. Section 3 describes the proposed approaches for the project. Section 4 gives a description of how to perform the input modeling. Section 5 discusses random number and random variate generations. Section 6 briefly describes the simulation model for the simplified field operations. Section 7 describes the output analysis methods. Section 8 provides preliminary results of simulation runs with different seeds and gives an example of sensitivity analysis of the model. Section 9.1 is the conclusion and summary.

2 Preliminary Concept for Simulation Modeling

The concept of simulation modeling for the field operations is described in this section. The reader who is familiar with simulation modeling should skip ahead. Simulation is the use of computations to implement a model of some dynamic system or phenomenon, such as field operations. It is using a model implemented as a computer program, rather than experimenting with a real system. In order to study the field operations scientifically, we need to make a set of assumptions about how the operations work. These assumptions are usually in the form of mathematical or logical relationships, and are therefore called mathematical or logical models. A valid model will help decision makers gain some understanding of how the system behaves. The field operations system is too complex to allow us to obtain a realistic model to be evaluated analytically. Instead, a simulation solution is obtained by using a computer to evaluate the model *numerically* over a time period of interest, and data are collected to *estimate* the desired characteristics of the model, i.e., the operating cost, response rates, timing, etc.

We would like to propose a simulation model for the field operations. Specifically, we would like to use the discrete event simulation technique to model the field operations system. Discrete event simulation concerns the modeling of a system as it evolves over time by a representation in which the variables change only at a countable number of points in time. These points in time are the ones at which an event occurs, where an *event*

is defined as an instantaneous occurrence which may change the state of a system. In this paper, simulation will be used to describe and analyze the behavior of the field operations system, ask what-if questions about the system, and aid in the modification of the field operations when needed. A sensitivity analysis will also be conducted to find out which potential solutions are the most cost effective methods to implement.

Although a discrete event simulation is usually done by computer, we would like to give an example of simulation by hand. Consider the field operations conducted by a group of FRs in a given day (say the jth day). One of the FRs (say FR i) is given 10 cases to conduct interviews for a particular survey. At the beginning of the day, a shortest path to visit each of the 10 households (a case per household) has been determined with the map provided by Geography Division. The distances (in miles) of traveling to each of the households are given in column (3) of Table 1. The average speed for the distances is between 20 and 45 mph, each value determined by road condition. For the purpose of this example, we assume that the average speed is equally likely between 20 and 45 mph. The contact time at each household is between 2 and 12 minutes, each value equally likely. At each household, it is either contacted or not contacted with 50% chance each. If it is contacted, the interviewed time is between 10 and 14 minutes, each value equally likely.

Table 1: Example: Simulation by Hand

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Case	Average Speed	Distance	Time for Traveling	Arrival Time	Time to Contact	Time Interview Begins	Contact	Time for Interview	Time Interview Ends
1	45	50	67	67	5	72	0	0	72
2	30	10	20	92	3	95	1	12	107
3	35	15	28	135	10	145	1	11	156
4	40	20	30	186	6	192	1	14	206
5	35	18	31	237	12	249	0	0	249
6	30	16	32	281	7	288	0	0	288
7	25	8	19	307	4	311	1	10	321
8	30	13	26	347	9	356	1	13	369
9	20	5	15	384	11	395	1	14	409
10	35	12	21	430	2	432	1	11	443
–	30	40	80	523	–	–	–	–	–
Total	–	207	369	–	69	–	7	85	–

The objective is to simulate the field operations, by hand, until the 10 households are visited, and to compute the cost, response rate, and timing of FR i in the jth day. This is just a simple example to illustrate the basic simulation concept that may be applied to the field operations. For a long run, more cases, more FRs, and other factors are needed to draw conclusions about the field operations. To simulate the process, random average speeds, times to contact, contact or no-contact, and times for interview need to be generated. Assume that the average speeds, times to contact, and times for interview are generated using three spinners that have possibilities for the values 20 thru 45, 2 thru 12, and 10 thru 14, respectively. Further assume that the contacts/no-contacts are generated using a fair coin that has possibilities for head (contact) and tail (no-contact).

In Table 1, column (1), Case, lists the 10 cases that need to be visited by the FR for the field operations. The dash after Case 10 is given to indicate the mileage and time needed for the FR to drive back to home or the office. The first row shows that the FR drives 50 miles at average speed of 45 mph to the first household arriving at minute 67, computed from columns (2) and (3). The values of column (2) were generated using the spinner with the possible values of 20 thru 45. Column (3) shows the distance indicated on the map based on the shortest path. Column (5) shows the simulated arrival time (or the simulated clock time of arrival) of the FR at each household. Column (6) shows the time needed to make contact with the respondent in the household. The values for Column (6) were generated using another spinner with possible values of 2 thru 12. Column (7) shows the clock time that the interview began. Column (8), contact or no-contact, shows the binary values by throwing the coin, head for contact = 1; tail for no-contact = 0. The values of column (9), time needed for interview, were generated using the spinner with possible values of 10 thru 14 if there was a contact. Otherwise, there was no interview and the time needed was 0. Finally, column (10) shows the clock time that the interview ended.

Now simulation of the field operations begins. At time 67, the FR arrived at the first household, spent 5

minutes to unsuccessfully make contact with the respondent, and left for the next household at time 72. It took 20 minutes for the FR to get to the second household. The contact time was 3 minutes and the interview began at time 95 and ended at time 107.

This process continues for all 10 cases, and the totals shown in the last row of column (3) (total mileage), column (4) (total time for traveling), column (6) (total time to contact the respondents), column (8) (total number of completed interviews), and column (9) (total time for the interviews) are entered. Assume that the hourly rate of FR i at the jth day is $R_{ij} = \$10.00$, the reimbursable mileage is $P_{ij} = \$0.35$ per mile. The performance measures of interest can now be calulated as follows:

From equation (2), the time needed to visit the 10 cases $= H_{ij} = 523/60 = 8.7$ hours;

the cost of this one-day-one-person field operations
$= C_{ij} = (\$10.00 \times 523/60) + (\$0.35 \times 207) = \$159.62$, where $M_{ij} = 207$;

the response rate $= 7/10 = 70\%$.

This small simulation by hand indicates that the response rate is 70% within the 8.7 working hours for a FR. The average cost for the completed cases is \$159.62/7=\$22.80. The simulation modeling proposed for this study will necessarily be much more complex, as more variables will be taken into account and the distribution functions for those variables will not be simple.

Also, there are constraints to the interviewer's work day at Census Bureau. First, there may be a planning phase prior to any travel. The simulation model described in this paper is assumed that the time of the planning phase (the traveling salesman algorithm) is incorporated into the traveling time. Second, interviewers rarely work more than 8 hours a day. A typical range for a work day is 4 to 6 hours a day. The simulation by hand example shows a work day of 8.7 hours and is used for illustrative purposes only.

3 Proposed Approaches

So far, similar work is rarely found in the literature describing the analytical or simulation modeling of the operations. The field operation is a unique system in the operations research field. Developing an analytical model for the operation requires an extensive investigation of the operation itself as well as the investigation of the operations research techniques. In this project, we will use a computer simulation model based on the concept described in Section 2 and this section.

As indicated before, we would like to propose a simulation model for the field operations. Developing a valid simulation model involves three basic entities: the real system under consideration (the field operations for a particular survey); a theoretical model of the real system; and a computer-based representation of the model, the simulation program. The activities of developing a theoretical model from the real system are referred to as simulation modeling, and the activities of developing a computer-based representation for the theoretical model are referred to as simulation programming. We will use C++ as the programming language. C++ is an object-oriented programming language that can be used to program an FR as an object. FRs are the key persons to the success of the operations. The object-oriented simulation is a technique to view the real system as being composed of various objects ([10]). The FR objects will be the core component of the simulation model of the field operations. Other C++ classes related to the object-oriented simulation, such as random number generation class, will be defined as well.

In the simulation programming, we will concentrate on modeling the behavior of interacting objects, such as FRs, respondents, etc., over time. The behavior of the interaction also involves other important steps in the simulation modeling and programming: random numbers and random variates generation, input data analysis, output data analysis, etc.

The example in Section 2 used input values that were generated by spinners and a coin. In computer simulation, the computer will generate independent random numbers that are distributed continuously and uniformly between 0 and 1 [i.e., $U(0, 1)$]. These random numbers can be converted to the desired statistical distributions, or random variates. The random numbers and random variates generation can be easily implemented with the object-oriented C++ language because the *class* definition in C++ can determine the objects' (random variates') characteristics or properties.

Input data analysis is another important step in the simulation modeling and programming. Input data modeling uses statistical methods to determine the desired statistical distributions needed for the random numbers and random variates generation. We will give a more detailed description of input data analysis in Section 4.

The analysis of simulation output begins with the selection of performance measures. As indicated before, the performance measures of interest in field operations are cost, response rates, and timing. The primary purpose of most simulation studies is the approximation of prescribed system parameters with the objective of identifying parameter values that optimize some system performance measures. Because some of the input processes in the field operations simulation study are random, the output data are also random and runs of the simulation result in *estimates* of performance measures. Unfortunately, a simulation run does not usually provide independent, identically distributed (IID) observations; therefore, "classical" statistical techniques are not directly applicable to the analysis of simulation output. We will briefly discuss the statistical techniques of the simulation output analysis in Section 7.

Finally, we will also perform a sensitivity analysis to determine the impact on the performance measures if some of the input variables (parameters) can be controlled. This analysis is valuable in determining what types of potential solutions are the most cost effective to implement. It will also be a feasibility study to determine the limitations of the simulation modeling applied to the field operations.

4 Input Data Analysis

The most difficult aspect of simulation input modeling is gathering data of sufficient quality, quantity, and variety to perform a reasonable analysis. After a preliminary study, we have identified part of the required and available data sets as described below. If a data set is not available for the project, we will have to make reasonable assumptions with help from the subject matter experts of Field Division. The following is a list of data sets so far identified and required for the project:

1. the average speed distribution for an FR driving between households;
2. the time distribution for an FR to make contact with respondents;
3. the time distribution for an FR to complete an interview if the respondent is contacted;
4. contact histories ([1]).

A random input variable to a simulation model can be viewed as a *stochastic process*. A stochastic process is often defined as a collection of random variables. In simulation modeling, the strongest assumptions of a stochastic process that we can make are: (1) all of the random variables are probabilistically *independent* of one another; (2) all of the random variables follow the same probability distribution and thus are said to be *identically distributed*. In other words, FR's are independent and follow the same rules. Also, the same type of random variables associated with each FR are also independent and follow the same rules. For example, the interview time of respondent A conducted by a particular FR is independent and identically distributed as that of respondent B conducted by the same FR. Therefore, we propose the following methods to perform the input data modeling for the simulation study:

1. Methods for assessing independence: Two simple yet effective heuristic procedures that are available:
   (a) tables and/or plots of estimated lag (linear) correlations: A *lag k correlation* is the correlation between observations k values apart. For example, a lag 2 correlation applies to the observation pairs with indices $(1,3),(2,4),\ldots,(n-2,n)$. For independent samples, we expect the estimated lag correlations to be small in magnitude and clustered about zero, with both negative and positive estimated values.
   (b) scatter diagrams: a scatter diagram is constructed by plotting the sequential pairs of observations with indices $(1,2),(2,3),\ldots,(n-1,n)$. Correlated samples tend to produce plots in which values cluster tightly about one of the diagonals (for positive correlations the lower-left to upper-right diagonal). Independent samples tend to produce plots with points distributed over portions of the plot in accordance with the underlying distribution.
2. Method for assessing stability of distribution: the stability can be evaluated with a simple plot of the observed values against their integer order of occurrence. We would expect the range of plotted points to be roughly the same across the plot. If a process features stability of distribution, the mean of the distribution appropriate to time T will not vary with T. We can use this simple observation to provide a crude, yet effective detector of instability of distribution.

The input data modeling also includes fitting a probability distribution to the data. We will assume that distributions are defined by their distribution functions, or equivalently, by their related density (continuous) or mass (discrete) functions. If the "best" of the fitted distributions provides a reasonable representation of the data, we will use it in the simulation. Otherwise, an empirical distribution will be used to represent the data directly.

4.1 Outcome Frequency Distribution of NHIS

In NHIS (National Health Interview Survey), each sample household is assigned one of the 28 outcomes after the visits of FRs. Table 2 lists the 28 possible outcomes and their frequency distribution from the 2004 interviews.

The following are the definitions used in the NHIS surveys:

- **Eligible cases** = total cases − (Type B's + Type C's);
- **Complete cases** = 201's + 203's;
- **Response rate** = 1 − (**Non-response rate**);
- **Non-response rate**: proportion of eligible cases that were noninterviews (Type A's)

$$\text{Equation} = \frac{\text{Type A's}}{\text{Eligible Cases}} \times 100; \tag{3}$$

- **Interview rate**: proportion of eligible cases that were completed interviews (outcome = 201)

$$\text{Equation} = \frac{\text{201's}}{\text{Eligible Cases}} \times 100; \tag{4}$$

- **Partial rate**: proportion of eligible cases that were sufficiently completed interviews (outcome = 203)

$$\text{Equation} = \frac{\text{203's}}{\text{Eligible Cases}} \times 100; \tag{5}$$

- Therefore, **Response rate** = **Interview rate** + **Partial rate**.

In the simulation study, the outcome frequency distribution needs to be adjusted and taken as the input of the simulation model. Some of the 28 outcomes can be determined when the FR visits only once regardless of the result of contact or no-contact. These outcomes are called one-visit outcomes. We have identified the following outcomes as the one-visit outcomes:
226 (vacant, nonseasonal)
228 (to be demolished)
229 (under construction)
230 (temporarily business or storage)
231 (unoccupied site)
232 (construction not started)
233 (other-Type B)
235 (vacant, seasonal)
240 (demolished)
241 (house-trailer mover)
242 (out of segment bounds)
243 (converted permanent business/storage)
244 (merged)
245 (condemned)
247 (other-Type C)
248 (spawned in error)
The other outcomes, except 216 (no one home), are determined as soon as the respondent is contacted. These outcomes are called contact outcomes and 216 is called no-contact outcome.

In the simulation model, we will assume a zero probability of the no-contact outcome. The final percentage of the no-contact outcome is determined by the contact/no-contact distribution discussed in Section 4.3.

Table 2: The 2004 NHIS Frequency Distribution by Outcome

i	Outcome	Freq.	Original %(f_i)	Original Cumul.	Adjusted %(g_i)	Adjusted Cumul.
1	201 (completed interview)	30992	43.58	43.58	44.48	44.48
2	203 (sufficient partial interview, no follow-up)	5916	8.32	51.90	8.49	52.97
Type A						
3	213 (language problem)	83	0.12	52.01	0.12	53.09
4	215 (insufficient partial interview)	519	0.73	52.74	0.74	53.83
5	216 (no one home, repeated calls)	1224	1.72	54.46	0.00	53.83
6	217 (temporarily absent, no follow-up)	312	0.44	54.90	0.45	54.28
7	218 (refused)	2604	3.66	58.56	3.74	58.02
8	219 (other-Type A)	570	0.80	59.36	0.82	58.83
Type B						
9	223 (all arm force)	122	0.17	59.54	0.18	59.01
10	225 (all URE)	934	1.31	60.85	1.34	60.35
11	226 (vacant, nonseasonal)*	6330	8.90	69.75	8.90	69.25
12	228 (to be demolished)*	243	0.34	70.09	0.34	69.59
13	229 (under construction)*	245	0.34	70.44	0.34	69.94
14	230 (temporarily business or storage)*	203	0.29	70.72	0.29	70.22
15	231 (unoccupied site)*	226	0.32	71.04	0.32	70.54
16	232 (construction not started)*	41	0.06	71.10	0.06	70.60
17	233 (other-Type B)*	138	0.19	71.29	0.19	70.79
18	235 (vacant, seasonal)*	1295	1.82	73.11	1.82	72.61
19	236 (screened out)	13813	19.42	92.53	19.82	92.43
Type C: unit is not there						
20	240 (demolished)*	236	0.33	92.87	0.33	92.77
21	241 (house-trailer mover)*	193	0.27	93.14	0.27	93.04
22	242 (out of segment bounds)*	137	0.19	93.33	0.19	93.23
23	243 (converted permanent business/storage)*	332	0.47	93.80	0.47	93.70
24	244 (merged)*	170	0.24	94.04	0.24	93.94
25	245 (condemned)*	26	0.04	94.07	0.04	93.97
26	246 (built after 4/1/1990)	3418	4.81	98.88	4.91	98.88
27	247 (other-Type C)*	261	0.37	99.24	0.37	99.24
28	248 (spawned in error)*	537	0.76	100.00	0.76	100.00
Total		71120	100.00	100.00	100.00	100.00

Therefore, We need to adjust the frequency distribution of the contact outcomes and make the no-contact outcome 0.0% for the simulation modeling. We will keep the one-visit outcome distribution unchanged and assume that the percentage of the no-contact outcome is redistributed according to the ditribution of the contact outcomes. Let V be the index set of the one-visit outcomes and U be the other outcomes. Also, let f_i be the percentage of the ith outcome, then

$$\sum_i f_i = 100.0$$

and the adjusted frequency distribution of the outcomes is computed as following:

$$g_i = \begin{cases} 0.0, & \text{if } i = 5; \\ f_i \times \dfrac{\sum_{j \in U} f_j}{\sum_{k \in U - \{5\}} f_k}, & \text{if } i \in U - \{5\}; \\ f_i, & \text{if } i \in V. \end{cases}$$

Table 2 also shows the 2004 adjusted frequency distribution by outcome at the national level. In the table, the one-visit outcomes are marked with an asterisk(*). The final percentage of 1.72% for code 216 (no one home) is determined by the distribution of contact/no-contact. Each of the 12 regional offices (and eventually, each PSU) will be handled in the same way. Table 3 shows the 2004 Quarter 2 (Q2) freuqency distribution and its adjusted frequency distribution by outcome for the Denver Regional Office.

Table 3: 2004 Q2 NHIS Frequency Distribution by Outcome (Denver RO)

i	Outcome	Freq.	Original $\%(f_i)$	Original Cumul.	Adjusted $\%(g_i)$	Adjusted Cumul.
1	201 (completed interview)	512	46.80	46.80	47.10	47.10
2	203 (sufficient partial interview, no follow-up)	70	6.40	53.20	6.44	53.54
Type A						
3	213 (language problem)	0	0.00	53.20	0.00	53.54
4	215 (insufficient partial interview)	3	0.27	53.47	0.28	53.81
5	216 (no one home, repeated calls)	6	0.55	54.02	0.00	53.81
6	217 (temporarily absent, no follow-up)	8	0.73	54.75	0.74	54.55
7	218 (refused)	28	2.56	57.31	2.58	57.13
8	219 (other-Type A)	5	0.46	57.77	0.46	57.59
Type B						
9	223 (all arm force)	4	0.37	58.14	0.37	57.95
10	225 (all URE)	32	2.93	61.06	2.94	60.90
11	226 (vacant, nonseasonal)*	86	7.86	68.92	7.86	68.76
12	228 (to be demolished)*	2	0.18	69.10	0.18	68.94
13	229 (under construction)*	5	0.46	69.56	0.46	69.40
14	230 (temporarily business or storage)*	9	0.82	70.38	0.82	70.22
15	231 (unoccupied site)*	2	0.18	70.57	0.18	70.40
16	232 (construction not started)*	2	0.18	70.75	0.18	70.59
17	233 (other-Type B)*	0	0.00	70.75	0.00	70.59
18	235 (vacant, seasonal)*	17	1.55	72.30	1.55	72.14
19	236 (screened out)	198	18.10	90.40	18.21	90.35
Type C: unit is not there						
20	240 (demolished)*	5	0.46	90.86	0.46	90.81
21	241 (house-trailer mover)*	4	0.37	91.22	0.37	91.18
22	242 (out of segment bounds)*	0	0.00	91.22	0.00	91.18
23	243 (converted permanent business/storage)*	2	0.18	91.41	0.18	91.36
24	244 (merged)*	1	0.09	91.50	0.09	91.45
25	245 (condemned)*	2	0.18	91.68	0.18	91.63
26	246 (built after 4/1/1990)	83	7.59	99.27	7.64	99.27
27	247 (other-Type C)*	4	0.37	99.63	0.37	99.63
28	248 (spawned in error)*	4	0.37	100.00	0.37	100.00
Total		1094	100.00	100.00	100.00	100.00

4.2 Interview Length Distributions of Outcomes in NHIS

In this section, we will try to decide what general family of distributions appears to be appropriate for each outcome's interview length. The methods we use for this purpose are scatter diagrams and probability plots. A scatter diagram is constructed for assessing the independence of observations and was described earlier in Section 4. A probability plot is a graphical comparison of an estimate of the distribution function of the interview length data $X_1, X_2, \ldots, X_n$ with the distribution function of one of the standard distributions being considered as a model for the data. Before we perform the input analysis using the probability plots and other methods, we would like to remove the outliers from the observed data. Some of the data are not good and considered as outliers for a variety of reasons. One of the reasons is that some of the observed data gave much longer time than the actual interview time because the computer was kept running without "the end of interview" being entered at the end of interview. Another is that some of the interview lengths are negative values.

There are $m = 30992$ observations for outcome 201 (completed interview) of NHIS in 2004. To remove the outliers, we have truncated the observations that are beyond two standard deviations from the mean. The truncation of removing the outliers has been repeated 4 times ($k = 4$ iterations) for outcome 201. The final number of observations used in the input analysis for outcome 201 is $n = 26741$. Table 4 shows the numbers of m, k, and n for each of the outcomes. The entries with "$-$" indicate that there was no interview time needed even though some observations were still captured.

Table 4: Number of Observations Used for Input Analysis

outcome	m	k	n
201 (completed interview)	30992	4	26741
203 (sufficient partial interview, no follow-up)	5916	4	5055
213 (language problem)	83	2	78
215 (insufficient partial interview)	519	3	460
216 (no one home, repeated calls)	1224	–	–
217 (temporarily absent, no follow-up)	312	–	–
218 (refused)	2604	–	–
219 (other-Type A)	570	3	496
223 (all arm force)	122	3	101
225 (all URE)	934	3	866
226 (vacant, nonseasonal)	6330	–	–
228 (to be demolished)	243	–	–
229 (under construction)	245	–	–
230 (temporarily business or storage)	203	–	–
231 (unoccupied site)	226	–	–
232 (construction not started)	41	–	–
233 (other-Type B)	138	–	–
235 (vacant, seasonal)	1295	–	–
236 (screened out)	13813	3	12470
240 (demolished)	236	–	–
241 (house-trailer mover)	193	–	–
242 (out of segment bounds)	137	–	–
243 (converted permanent business/storage)	332	–	–
244 (merged)	170	–	–
245 (condemned)	26	–	–
246 (built after 4/1/1990)	3418	3	3120
247 (other-Type C)	261	–	–
248 (spawned in error)	537	–	–

4.2.1 Assessing Independence of Interview Length of Outcomes in NHIS

As described in Section 4, a *lag k correlation* plot and a scatter diagram could have been constructed to assess the independence of the interview lengths. However, the interviews were conducted by different interviewers (FRs) at different households and at different time as described earlier in Section 4. Therefore, it is reasonable to assume that the interview lengths conducted by different FRs at the sample households for all outcomes are independent samples.

4.2.2 Probability Plots of Interview Length of Outcomes in NHIS

Let $X_{(i)}$ be the smallest of the X_j's, called the ith *order statistic* of the n X_j's. The distribution function F of a random variable X is defined so that for any x, $F(x) = P\{X \leq x\}$. If X has the same distribution as the X_j data, a reasonable approximation to $F(x)$ is thus the proportion of the X_j's that are less than or equal to x. Therefore, we might want to define an empirical distribution function $\tilde{F}_n(x)$ so that $\tilde{F}_n(X_{(i)}) = i/n$, or $\tilde{F}_n(X_{(i)}) = (i - 0.5)/n$ to avoid an empirical distribution function that is equal to 1 for a finite value of x.

For $0 < q < 1$, the q quantile of a distribution F is a number x_q that satisfies $F(x_q) = q$. Thus, if F^{-1} denotes the inverse of the distribution function F, a formula for the q quantile of F is $x_q = F^{-1}(q)$, where F^{-1} exists if F is continuous and strictly increasing. If F and G are two distribution functions, it is clear that $F = G$ if and only if each of the quantiles of F is the same as the corresponding quantile of G. Thus, if x_q and y_q are the q quantiles of F and G, respectively, a plot of the points (x_q, y_q) for various values of q will produce points along a straight line having slope 1 (a 45° line) and passing thru the origin, since $x_q = y_q$ for all q. Furthermore, if the random variables corresponding to F and G differ only in location and scale, then for some real numbers γ and $\beta > 0$, we have $G(x) = F((x - \gamma)/\beta)$ for all x. In this case, it is easy to see that for all q, $y_q = \gamma + \beta x_q$, so that a plot of the points (x_q, y_q) produces a straight line of points which has a slope

not necessarily 1 and which need not pass thru the origin. Thus, distributions having the same shape (but which may differ in location and scale) have quantiles which are linearly related. A plot of pairs of quantiles such as (x_q, y_q) is called probability plot, or Q-Q plot.

Probability plots provide a way of assessing whether the empirical distribution function $\tilde{F}_n$, defined at the $X_{(i)}$ points, has the same shape as a distribution function from one of the theoretical families. For survey interview length, suppose that we are considering a particular distribution form and that if this distribution has shape parameters, they have already been estimated from the data. Let the resulting distribution function be denoted by F, which represents a trial hypothesized distribution shape, with unspecified location and scale. We would like to compare $\tilde{F}_n$ with F, and we can do so by a (Q-Q) probability plot of the quantile pairs for $q = (i - 0.5)/n$ for $i = 1, 2, \ldots, n$, as follows. By definition, the $(i - 0.5)/n$ quantile of $\tilde{F}_n$ is precisely $X_{(i)}$. The $(i - 0.5)/n$ quantile of F is simply $F^{-1}((i - 0.5)/n)$. Thus, we plot the points

$$\left(X_{(i)}, F^{-1}\left(\frac{i - 0.5}{n} \right) \right)$$

for $i = 1, 2, \ldots, n$, and if the resulting points appear to lie along a straight line, we have informal confirmation that, except for adjustments in location and scale, F is a good distribution function for our interview length data.

Figure 1 shows the beta distribution probability plot for interview length of outcome 201. The plot indeed appears to have a straight line, supporting the beta distribution. To provide an idea of what a probability might look like when an inappropriate distribution is hypothesized, we made probability plots for the Weibull and gamma distributions. The resulting Weibull probability plot in Figure 2 displays obvious nonlinearity at the upper end while the gamma plot in Figure 3 displays nonlinearity at both ends.

4.2.3 Estimation of Parameters for Interview Length of Outcomes in NHIS

After a family of distributions has been hypothesized, we must specify the value(s) of its parameter(s) in order to determine completely the distribution from which we shall sample during the simulation. Our hypothesized distribution is a beta distribution, Beta$(\alpha, \beta, \theta, \lambda)$, where α and β are the shape parameters, θ is the threshold parameter, and λ is the scale parameter. The probability density function is given in Appendix A.1. The density and distribution functions of Weibull, Weibull$(\alpha, \theta, \lambda)$, and gamma, Gamma$(\alpha, \theta, \lambda)$, distributions mentioned in Section 4.2.2 are also given in Appendix A.1. For both distributions, $\alpha > 0$ is the shape parameter, θ is the threshold parameter, and $\lambda > 0$ is the scale parameter.

The parameter estimates using the *maximum-likelihood estimators* (MLEs) for the three distributions and the interview lengths of outcome 201 are given in Table 5. Other distributions used for testing other outcome data are exponential and lognormal distributions. Their density and distribution functions are given in Appendix A.1.

Table 5: Parameter Estimates for the Three Distributions and Outcome 201 Data

		Parameters			
Distribution	Range	Shape(α)	Shape(β)	Threhold(θ)	Scale(λ)
Beta$(\alpha, \beta, \theta, \lambda)$	$\theta < x < \theta + \lambda$	2.127	2.549	8.796	79.258
Weibull$(\alpha, \theta, \lambda)$	$x > \theta$	2.484		6.566	43.030
Gamma$(\alpha, \theta, \lambda)$	$x > \theta$	17.392		-24.802	3.997

4.2.4 Goodness-of-Fit Tests for Interview Length of Outcomes in NHIS

After we have hypothesized a distribution form for our data and have estimated its parameters, we must examine whether the fitted distribution is in agreement with our observed data $X_1, X_2, \ldots, X_n$. If $F(x)$ is the distribution function of the fitted distribution, a hypothesis test is addressed with a null hypotheses of

$$H_0 : \text{The } X_i\text{'s are IID random variables with distribution function } F(x) \tag{6}$$

This is called a *goodness-of-fit test* since it tests how well the fitted distribution "fits" the observed data. We used four goodness-offit test methods to perform the tests. The detailed descriptions of the four methods are given in Appendix A.2.

Figure 1: The Beta Probability Plot of Interview Lengths for Completed Interviews (Outcome 201).

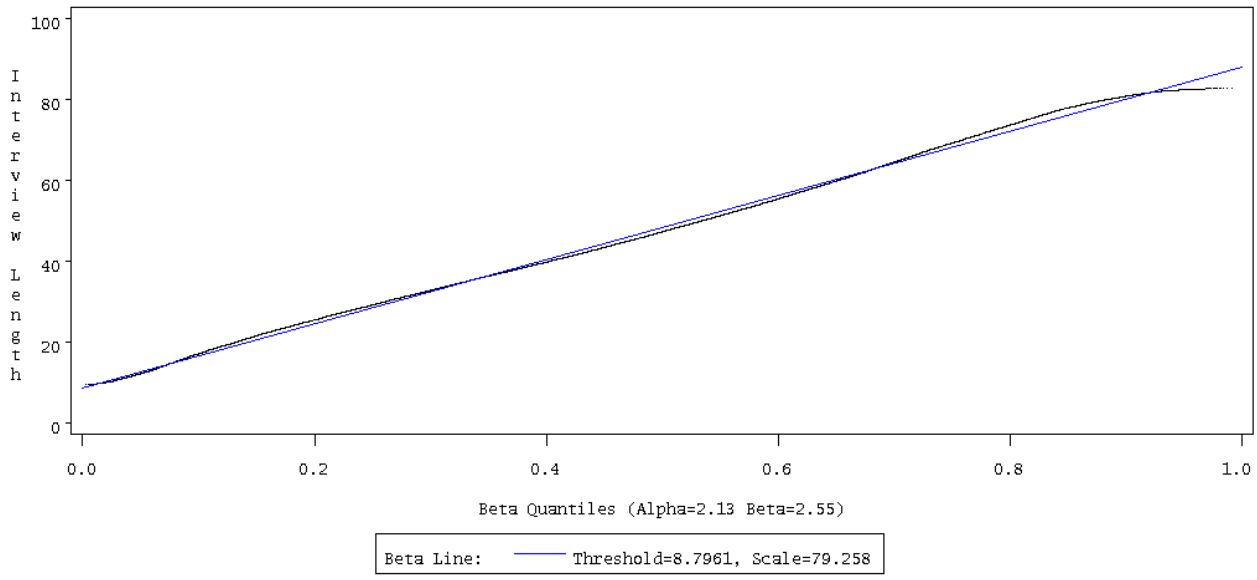


Figure 2: The Weibull Probability Plot of Interview Lengths for Completed Interviews (Outcome 201).

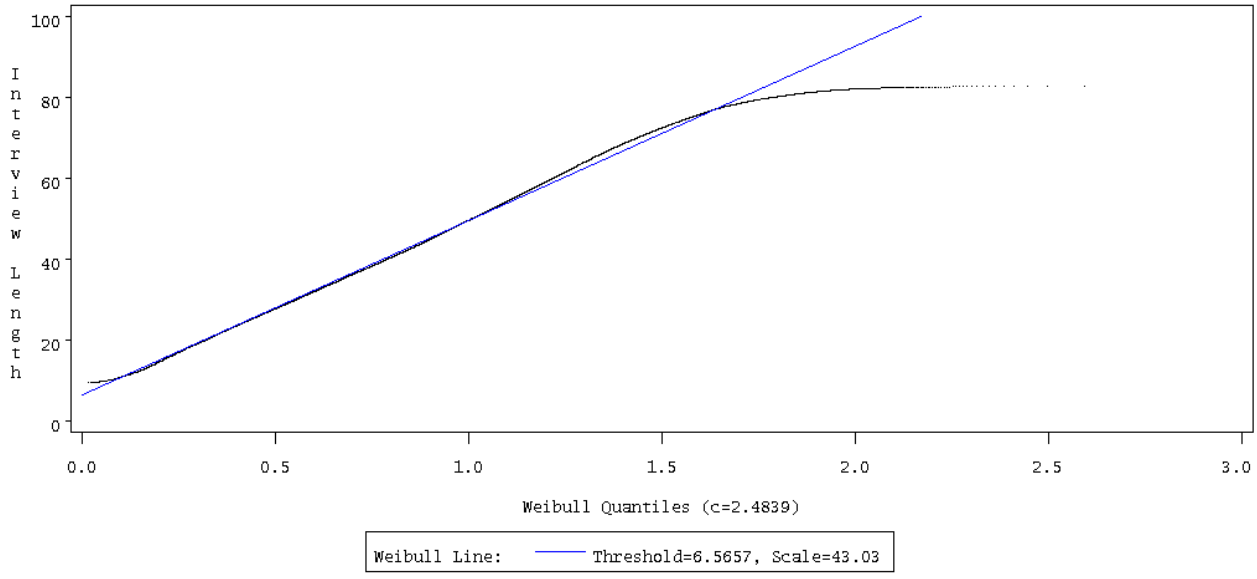


Figure 3: The Gamma Probability Plot of Interview Lengths for Completed Interviews (Outcome 201).

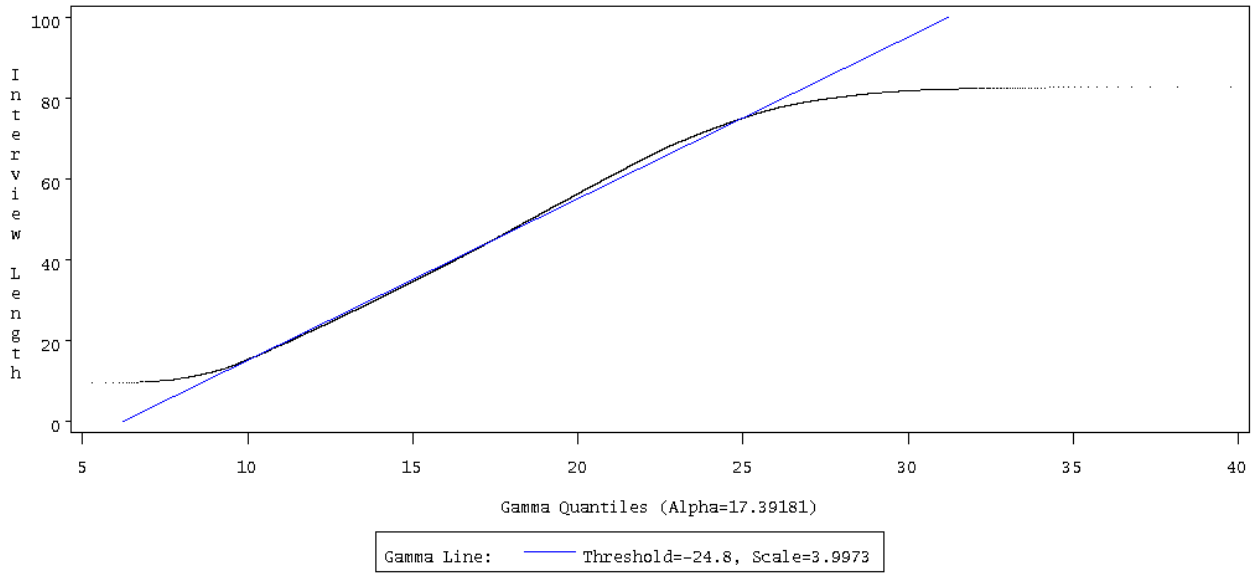


Figure 4: The Histogram Plot of Interview Lengths for Completed Interviews (Outcome 201) with Beta, Weibull, and Gamma Distributions

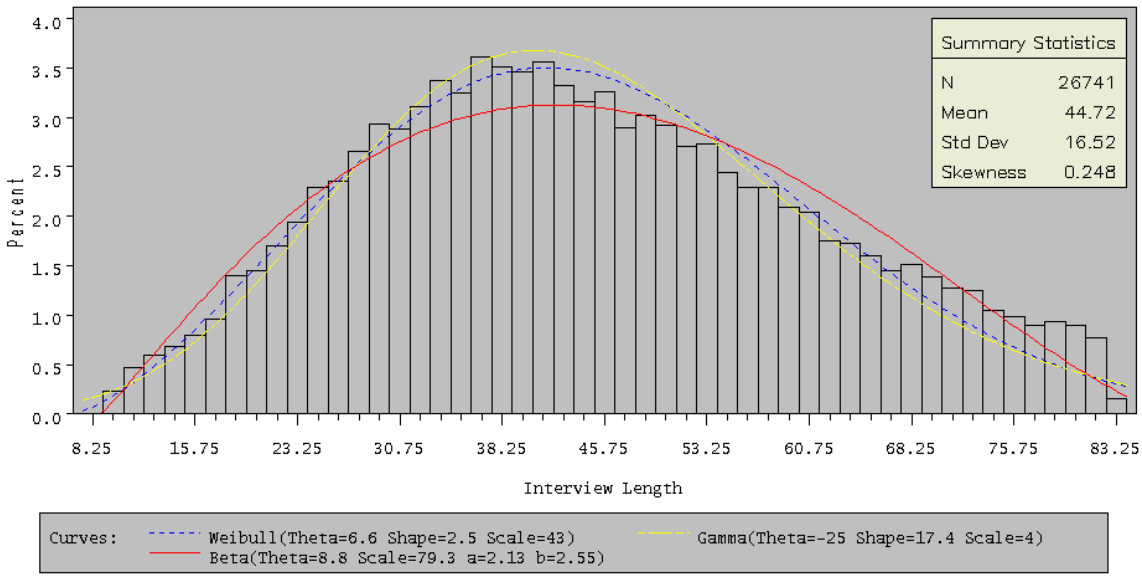


Figure 4 shows the histogram of the interview lengths for outcome 201. Table 6 show the results of the four goodness-of-tests, indicating that the ouctome 201 interview lengths do not fit the three distributions tested. In the simulation, we will use an emprical distribution from which the samples are drawn. Section 4.2.5 will describe how to specify an empirical distribution.

4.2.5 Empirical Distributions for Interview Lengths of Outcomes in NHIS

The description of the previous sections indicated that we simply cannot find a theoretical distribution which fits the outcome 201 data adequately (see Sections 4.2.2 to 4.2.4). Therefore, we need to use the observed data themselves to specify directly a distribution, called an *empirical distribution*, from which samples are drawn during the simulation.

For continuous random variables the type of empirical distribution that can be defined is dependent on whether we have the actual values of the individual original observations $X_1, X_2, \ldots, X_n$ rather than only the number of X_i's which fall into each of several specified intervals, called *grouped data* or *binned data*. If the original data are available, we can define a continuous, piecewise linear distribution function $\widehat{F}$:

$$\widehat{F}(x) = \begin{cases} 0, & \text{if } x < X_{(1)}; \\ \frac{i-1}{n-1} + \frac{x - X_{(i)}}{(n-1)(X_{(i+1)} - X_{(i)})}, & \text{if } X_{(i)} \le x < X_{(i+1)} \text{ for } i = 1, 2, \ldots, n-1; \\ 1, & \text{if } X_{(n)} \le x. \end{cases} \tag{7}$$

Note that for each i, $\widehat{F}(X_{(i)}) = (i-1)/(n-1)$ is approximately (for large n) the proportion of the X_j's that are less than $X_{(i)}$. However, one clear disadvantage of specifying this particular empirical distribution is that random variables generated from it during a simulation run can never be less than $X_{(1)}$ or more than $X_{(n)}$.

If the data are grouped, a different approach must be taken. Suppose that the n X_i's are grouped into k adjacent intervals $[a_0, a_1), [a_1, a_2), \ldots, [a_{k-1}, a_k)$ so that the jth interval contains n_j observations, where $n_1 + n_2 + \cdots + n_k = n$. A reasonable piecewise linear empirical distribution function $\widehat{G}$ could be specified by first letting $G(a_0) = 0$ and $G(a_j) = (n_1 + n_2 + \cdots + n_j)/n$ for $j = 1, 2, \ldots, k$. Then interpolating linearly between the a_j's, we define

$$\widehat{G}(x) = \begin{cases} 0, & \text{if } x < a_0; \\ \widehat{G}(a_{j-1}) + \frac{x - a_{j-1}}{a_j - a_{j-1}} \Big(\widehat{G}(a_j) - \widehat{G}(a_{j-1}) \Big), & \text{if } a_{j-1} \le x < a_j \text{ for } j = 1, 2, \ldots, k; \\ 1, & \text{if } a_k \le x. \end{cases} \tag{8}$$

The random variables generated from this distribution will still be bounded both below by a_0 and above by a_k.

For discrete data, it is simple to define an empirical distribution if the original data values $X_1, X_2, \ldots, X_n$ are available. For each possible value x, an empirical mass function $\widehat{p}(x)$ can be defined to be the proportion of the X_i's that are equal to x. For grouped discrete data, we can define a mass function such that the sum of the $\widehat{p}(x)$'s over all possible values of x in an interval is equal to the proportion of the X_i's in that interval.

4.2.6 Data Analysis for Interview Lengths of Other Outcomes in NHIS

The scatter diagrams, not shown, for the interview lengths of outcomes 203, 213, 215, 219, 223, 225, 236, and 246 indicate that it is reasonable to assume that they are independent samples.

We also performed the Q-Q plots and goodness-of-fit tests for each of those outcome data. Table 7 shows the results of the goodness-of-fit tests of all the outcome data analyzed.

The fitted distributions in Table 7 will be used for the random variate generation described in Section 5.2.

4.3 Analysis of Contact Histories with NHIS

In this section, we describe the analysis of the contact attempt history data collected with the 2004 NHIS. Dahlhamer, Simile, Stussman, and Taylor [5] give a detailed analysis of this CHI (Contact History Instrument) data set. They found weekday evenings and weekends to be the best times to make contact with households in the NHIS, at least for the first four attempts (where prior attempts were no-contacts). In their work, for all analyses involving time of contact attempt, mornings are defined as 12:00 AM to 11:59 AM, afternoons as

Table 6: Results of the Four Goodness of Fit Tests for Outcome 201 Data

Distribution Test	Statistic		DF	p Value			Test Result
Beta$(\alpha, \beta, \theta, \lambda)$							
Chi-Square	$\chi^2 =$	645.802862	47	$\Pr(p > \chi^2)$	$<$	0.001	rejected H_0
Kolmogorov-Smirnov	$D =$	0.024418		$\Pr(p > D)$	$<$	0.001	rejected H_0
Cramér-von Mises	$W^2 =$	5.205604		$\Pr(p > W^2)$	$<$	0.001	rejected H_0
Anderson-Darling	$A^2 =$	31.448265		$\Pr(p > A^2)$	$<$	0.001	rejected H_0
Weibull$(\alpha, \theta, \lambda)$							
Chi-Square	$\chi^2 =$	687.104887	47	$\Pr(p > \chi^2)$	$<$	0.001	rejected H_0
Kolmogorov-Smirnov	$D =$	0.019136		$\Pr(p > D)$	$<$	0.001	rejected H_0
Cramér-von Mises	$W^2 =$	2.413048		$\Pr(p > W^2)$	$<$	0.001	rejected H_0
Anderson-Darling	$A^2 =$	21.969588		$\Pr(p > A^2)$	$<$	0.001	rejected H_0
Gamma$(\alpha, \theta, \lambda)$							
Chi-Square	$\chi^2 =$	834.565941	47	$\Pr(p > \chi^2)$	$<$	0.001	rejected H_0
Kolmogorov-Smirnov	$D =$	0.022515		$\Pr(p > D)$	$<$	0.001	rejected H_0
Cramér-von Mises	$W^2 =$	4.223140		$\Pr(p > W^2)$	$<$	0.001	rejected H_0
Anderson-Darling	$A^2 =$	36.594500		$\Pr(p > A^2)$	$<$	0.001	rejected H_0

Table 7: Results of the Goodness of Fit Tests for Outcome Data Tested

Outcome	Sample Size (n)	Distribution(s) Fitted (Not Rejected)	Distribution Will Be Used in Simulation Runs	Alternative Distribution(s) Will Be Used
201	26741	None	Empirical	Beta$(\alpha, \beta, \theta, \lambda)$ =(2.127,2.549,8.796,79.258)
203	5055	None	Empirical	Beta$(\alpha, \beta, \theta, \lambda)$ =(1.988,3.233,0.969,81.207)
213	78	Weibull Exponential Gamma	Weibull$(\alpha, \theta, \lambda)$ =(1.000,0.367,2.064)	Empirical Exponential(θ, λ) =(0.340,2.090) Gamma$(\alpha, \theta, \lambda)$ =(0.913,0.367,2.261)
215	460	Gamma Exponential Weibull	Gamma$(\alpha, \theta, \lambda)$ =(1.052,1.298,15.257)	Exponential(θ, λ) =(1.265,16.076) Weibull$(\alpha, \theta, \lambda)$ =(1.016,1.299,16.148)
219	496	Gamma Weibull	Gamma$(\alpha, \theta, \lambda)$ =(1.017,0.233,2.965)	Weibull$(\alpha, \theta, \lambda)$ =(1.000,0.233,3.014)
223	101	Gamma Weibull Beta Lognormal	Gamma$(\alpha, \theta, \lambda)$ =(1.708,0.462,2.195)	Weibull$(\alpha, \theta, \lambda)$ =(1.328,0.531,3.999) Beta$(\alpha, \beta, \theta, \lambda)$ =(1.339,6.897,0.533,22.667) Lognormal(ζ, θ, λ) =(1.319,-0.288,0.623)
225	866	Lognormal	Lognormal(ζ, θ, λ) =(0.802,0.194,0.982)	Weibull$(\alpha, \theta, \lambda)$ =(1.031,0.267,3.512)
236	12470	None	Empirical	Lognormal(ζ, θ, λ) =(1.538,-0.086,0.596)
246	3120	Lognormal	Lognormal(ζ, θ, λ) =(0.677,-0.054,0.825)	Gamma$(\alpha, \theta, \lambda)$ =(1.302,0.166,1.910)

Outcome Codes:
201 (completed interview)
203 (sufficient partial interview, no follow-up)
213 (language problem)
215 (insufficient partial interview)
219 (other-Type A)
223 (all arm force)
225 (all URE)
236 (screened out)
246 (built after 4/1/1990)

12:00 PM to 4:49 PM, and evenings as 5:00 PM to 11:59 PM. The input modeling of the simulation is based on their work. However, we are only interested in the contact/no-contact distributions based on the time by the hour of a day and the days of a week. We divide a day by the hour because the starting time of each day by the FRs will be a decision varible when a sensitivity analysis is performed.

Groves and Couper [7] define the *contactability* as the propensity for a household to be contacted by an interviewer at any given moment in time. They also give illustrations of contactability that must be a function of three factors: (1) whether there are any physical impediments like gates or key card entries that increase the efforts for contact by interviewers; (2) at-home patterns of households, i.e. when household members are at home; and (3) the number and timing of contact attempts. In this paper, we only address the last factor of contactability. The other two factors will be considered in the future simulation models.

There are 197607 observations from the 2004 CHI data sets. More than 78% of them (154741 observations) are personal visits, the other 22% (42866 observations) are telephone calls. In our simulation model, we assume that there are no telephone calls in the field operations activities. In reality, phone calls may be used for follow-up interviews after the first contact personal visit. We will add the activities of phone calls into the model in the future.

Table 8: The Frequency Distributions of Contact/No-Contact$^+$

	Hour*	Sun(%)	Mon(%)	Tue(%)	Wed(%)	Thur(%)	Fri(%)	Sat(%)
Contact	08-09	33.96	61.54	50.81	55.01	54.73	57.85	43.31
Contact	09-10	51.96	45.51	55.03	53.70	48.66	53.32	58.82
Contact	10-11	48.88	52.85	51.48	52.84	53.92	52.86	53.58
Contact	11-12	49.11	50.77	48.55	49.15	53.39	49.84	56.25
Contact	12-13	51.96	50.52	50.85	47.73	52.30	49.19	54.56
Contact	13-14	52.90	50.63	52.53	51.49	49.00	52.55	53.61
Contact	14-15	53.61	50.71	50.02	53.77	49.94	53.73	54.60
Contact	15-16	50.47	53.05	53.17	53.02	53.12	54.90	54.87
Contact	16-17	52.78	56.54	55.48	57.10	57.11	54.56	52.94
Contact	17-18	54.40	57.30	57.12	60.18	58.09	54.76	50.43
Contact	18-19	52.10	57.41	59.80	58.09	57.08	53.58	48.42
Contact	19-20	49.93	52.74	53.87	55.50	55.21	51.21	49.06
Contact	20-21	52.04	50.94	50.76	51.66	50.23	47.59	40.79
Contact	21-22	47.03	49.27	50.60	47.23	48.30	41.30	45.82

$^+$ The percentage of No-Contact is 1 - percentage of Contact
* The column of "Hour" shows the local time

Table 8 shows the personal visit frequency distributions of contact/no-contact based on the time (from 8:00 AM to 10:00 PM) of a day and the day of a week. In the table, the column of "Hour" shows the local time of the PSUs that the FRs visit the households. It shows that the best times to make contact with households in the NHIS are between 1:00 PM to 8:00 PM on weekdaysi, 9:00 AM to 5:00 PM on Saturdays, and 12:00 PM to 7:00 PM on Sundays. The table also indicates that the personal visits occurred at any time of a day.

5 Random Number and Random Variate Generations

A random number is a single observation of the continuous uniform distribution on the interval (0, 1). The random number is then transformed as needed to simulate a random variate from different probability distributions, such as the normal, exponential, Poisson, binomial, Weibull, gamma, lognormal, etc. Random number generation is a computational procedure designed to generate a sequence of numbers. In contrast, random variate generation always refers to the generation of variates whose probability distribution is usually different from that of the uniform on the interval (0, 1).

5.1 Random Number Generation

Random numbers are the basic building blocks of simulation study. A random number generator is needed to generate a sequence of independent and identically distributed (iid) $U(0, 1)$ random variables. This sequence of random numbers can be obtained thru deterministic algorithms with a solid mathematical basis. The

numbers produced by these algorithms are in fact not random at all. They should be called pseudorandom. For more detailed description of pseudorandom number generations, see L'Ecuyer [13]. For simplification, the term random is used instead of pseudorandom in the simulation contexts. A *random number* is always meant a unform random variable, denoted by $U(0,1)$ (or `rand()` in our C++ code of the simulation model), whose distribution function is

$$F(u) = \begin{cases} 0, & \text{if } u \leq 0; \\ u, & \text{if } 0 < u < 1; \\ 1, & \text{if } u \geq 1. \end{cases} \tag{9}$$

The algorithm we used to generate a sequence of random numbers is given in Appendix B.1.

5.2 Random Variate Generation

In Section 5.1 the generation of (pseudo) *random numbers* was briefly discussed. In this section, we will briefly discuss the random variate generations, see Cheng [3] for more detailed descriptions.

Random variate generation refers to the generation of variates whose probability distribution is different from that of the uniform on the interval (0,1). The basic concept is to generate a random variable, X, whose distribution function

$$F(x) = \Pr(X \leq x) \quad -\infty < x < \infty \tag{10}$$

is assumed to be completely known, and which is different from that of Equation (9). A list of the random variate generations used in our C++ code of the simulation model is given in Appendix B.2.

6 Description of the Model

First, one thousand and fifty cases (households) are generated for the model. Ten field representatives (FRs) are assumed, each of them is assigned a hundred and five cases and a PSU of 60×60 square miles [1]. Each of the one thousand and fifty cases is identified with its case number and its location (x, y) within its own PSU, where $0 \leq x \leq 60$ and $0 \leq y \leq 60$. The values of x and y come from a *uniform input distribution between* 0 *and* 60, $U(0, 60)$ [2]. The field office and/or the FRs' homes can be located any where in a PSU of 3600 square miles. The simulation results are independent of the locations because the sample households are randomly selected. In this simulation study, they are assumed to be located at $(0, 0)$.

Each FR selects the first n cases (ascending order of case numbers of the incomplete cases) for each day's work. The value of n comes from a *uniform input distribution* $U(8, 16)$. The FR has to visit each of the n selected cases once for that day. The visiting order of the n cases is determined by the following:

1. The direct distance (d_{ij}) between each pair $((x_i, y_i)$ and $(x_j, y_j))$ of the n cases is calculated by

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

2. A distance matrix of the n cases is formed, and therefore, a traveling network is formed with interconnections between the nodes (cases).

3. Starting from the field office or FR's home located at (0, 0) and ending at (0, 0), the FR visits each case only once. This becomes a traveling salesman problem[3].

4. The objective is to minimize the total distance traveled. Instead of this measure of distance, any other measure of effectiveness may be substituted, such as time, likelihood of contact, and so on.

[1] The area of a PSU should not exceed 3,000 square miles except in cases where a single county exceeds the maximum area; we use 3,600 square miles for our experimental runs

[2] All the uniform distributions described in this paper are discrete uniform distributions

[3] The traveling salesman problem can be stated as follows. A salesman, starting from a city, intends to visit each of (n-1) other cities once and only once and return to the start. The problem is to determine the order in which he should visit the cities to minimize the total distance traveled, assuming that the direct distances between all city pairs are known. The structure of the problem shows that there are (n-1)! possible tours, of which one or more should be optimal

5. A branch and bound algorithm[4] is used to determine the visiting order.

With the shortest path to visit each of the n households determined, each of the FRs is to visit each household to conduct an interview for the survey. Table 9 shows the detailed information for conducting interviews of 6 cases by FR 7 at day 5 starting at noon for the simulation run with seed 23. The distances (determined by the two locations (x_i, y_i) and (x_j, y_j)) of traveling to each of the households are given in column (4) of Table 9. The average speed (mph) for the distances is a *uniform distribution* $U(30, 40)$. The contact time (minutes) at each household is a *uniform distribution* $U(3, 7)$. At each household, it is either contacted or not contacted. The contact/no-contact distributions are described in Table 8; no-contact if 0 and contact otherwise (potential refusal is considered contact) so that the probability of contact depends on the time of a day and the day of a week. If it is contacted, the interview length (minutes) is generated from the distributions given in Table 10 (also see Table 7) depending on the outcomes of the visits.

Table 9: Field Representative 7 at Day 5 with seed 23

(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Today's Seq.	Case Number	Average Speed	Distance	Time for Traveling	Arrival Time	Time to Contact	Time Interview Begins	Current Visit	Contact	Time for Interview	Time Interview Ends
1	750	32	13	24	24	5	29	1	1	22	51
2	746	36	38	63	114	6	120	1	1	21	141
3	747	30	51	102	243	5	248	1	1	23	271
4	748	35	12	20	291	3	294	1	0	0	294
5	752	37	27	43	337	3	340	1	0	0	340
6	749	39	3	4	344	5	349	1	1	23	372
–	—	31	15	29	401	–	–	–	–	–	—
Total	–	–	159	285	–	27	–	–	4	89	401

Table 10: Distributions Used in the Model for Interview Lengths

Outcome	Probability Distribution Used in the Model	
201	Beta$(\alpha, \beta, \theta, \lambda)$	= Beta(`2.127,2.549,8.796,79.258`)
203	Beta$(\alpha, \beta, \theta, \lambda)$	= Beta(`1.988,3.233,0.969,81.207`)
213	Weibull$(\alpha, \theta, \lambda)$	= Weibull(`1.000,0.367,2.064`)
215	Gamma$(\alpha, \theta, \lambda)$	= Gamma(`1.052,1.298,15.257`)
219	Gamma$(\alpha, \theta, \lambda)$	= Gamma(`1.017,0.233,2.965`)
223	Gamma$(\alpha, \theta, \lambda)$	= Gamma(`1.708,0.462,2.195`)
225	Lognormal(ζ, θ, λ)	= Lognormal(`0.802,0.194,0.982`)
236	Lognormal(ζ, θ, λ)	= Lognormal(`1.538,-0.086,0.596`)
246	Lognormal(ζ, θ, λ)	= Lognormal(`0.677,-0.054,0.825`)

In Table 9, columns (1) and (2) list the n cases that need to be visited by the FR for the field operations. The dashes after Today's Seq. 6 and Case Number 749 are given to indicate the mileage and time needed for the FR to drive back to the office at (0, 0) at about 6:41 PM (minute 401). The first row shows that the FR drives 13 miles at average speed of 32 mph to the first household (case number 750) arriving at minute 24, computed from columns (3) and (4). Column (6) shows the simulated arrival time of the FR at each household. Column (7) shows the time needed to make contact with the respondent in the household. Column (8) shows the clock time that the interview began. Column (9) indicates that the number of visits to complete the interview so far. Column (10), contact or no-contact, shows the binary values of contact = 1 and no-contact = 0. The values of column (11) are time needed for the interview if there was a contact. Otherwise, there was no interview and the time needed was 0. Finally, column (12) shows the clock time that the interview ended.

[4]The method of the branch and bound algorithm is to first identify a feasible solution and then to decompose the set of all remaining feasible tours into smaller and smaller subsets. At each step of the decomposition, a lower bound on the length of the current best tour is readily available. The bounds provide a guide for the partitioning of the subsets of feasible tours and eventually for the identification of an optimal tour. When a tour with length less than or equal to the minimum lower bound of all other tours is found, this intermediate solution becomes the best available. This process of bounding tours, eliminating suboptimal alternatives, and branching to new (better) tours is the basis of the algorithm.

7 Output Data Analysis

There are two types of simulations with regard to output analysis:

1. *Finite-Horizon Simulations.* In this case the simulation starts in a specific state, such as empty and idle state, and is run until some terminating event occurs. The output process is not expected to achieve any steady-state behavior and any parameter estimated from the output data will be transient in the sense that its value will depend on the initial conditions.

2. *Steady-State Simulations.* The purpose of a steady-state simulation is the study of the long-run behavior of the system of interest. A performance measure of a system is called a *steady-state parameter* if it is a characteristic of the equilibrium distribution of an output stochastic process ([12]).

7.1 Finite-Horizon Simulations

Suppose that one starts in a specific state and simulates a field operation until n output data (such as response rates) $X_1, X_2, \ldots, X_n$ are collected with the objective of estimating $f(X_1, \ldots, X_n)$, where f is a function of the data. For example, $F(X_1, \ldots, X_n) = \overline{X}_n = (1/n)\sum_{i=1}^{n} X_i$ is the average number of visits for the n completed cases. There are two approaches for estimating the output performance measures. We will investigate each of them to determine which is appropriate for the objective of field operations.

1. Estimation of the measure via independent replications: In many real systems, the output data are positively correlated. When this is the case, the traditional variance estimator,

$$\frac{S_n^2(X)}{n} = \frac{1}{n(n-1)} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2, \tag{11}$$

of $\mathrm{Var}(\overline{X}_n)$ is highly biased. To overcome this problem, we can run k independent replications of the system simulation. Each replication starts in the same state and uses a seed of the random number generator that is different from the seeds used to run the other replications. Assume that replication i produces the output data $X_{i1}, X_{i2}, \ldots, X_{in}$. Then the sample means

$$Y_i = \frac{1}{n} \sum_{j=1}^{n} X_{ij} \quad i = 1, 2, \ldots, k \tag{12}$$

are IID random variables,

$$\overline{Y}_k = \frac{1}{k} \sum_{i=1}^{k} Y_i \tag{13}$$

is also an unbiased estimator of the mean μ, and the sample variance of the Y_i's

$$S_k^2(Y) = \frac{1}{k-1} \sum_{i=1}^{k} (Y_i - \overline{Y}_k)^2 \tag{14}$$

is an unbiased estimator of $\mathrm{Var}(\overline{X}_n)$. If, in addition, n and k are sufficiently large, an approximate $1-\alpha$ confidence interval for μ is

$$\overline{Y}_k \pm t_{k-1,1-\alpha/2} \frac{S_k(Y)}{\sqrt{k}}. \tag{15}$$

Denote the half-width of the interval (7) by

$$\delta(k, \alpha) = t_{k-1,1-\alpha/2} \frac{S_k(Y)}{\sqrt{k}}. \tag{16}$$

2. Sequential estimation: We would like to make k runs, so that the estimation of μ is within a tolerance $\pm d$, where d is user specified, i.e.,

$$P(\overline{Y}_k - d \le \mu \le \overline{Y}_k + d) \ge 1 - \alpha \tag{17}$$

where $\alpha \in (0,1)$. The sequential procedure of Chow and Robbins ([4]) can be used for the estimation, which is to run one replication at a time and stop at run k^* such that

$$k^* = \min\left\{ k \;\middle|\; k \geq 2, \delta(k,\alpha) \leq \sqrt{\frac{k}{k-1}d^2 - \frac{t^2_{k-1,1-\alpha/2}}{k(k-1)}} \right\}. \tag{18}$$

Different stopping rules may be used to obtain a confidence interval with coverage close to the specified confidence level.

7.2 Steady-State Simulations

Several methods have been developed for the estimation of steady-state system parameters:

1. Removal of initial bias: Since we are interested in the performance measures of the steady-state simulation system, the initial bias must be removed.

2. Replication-deletion approach: This approach runs k independent replications, each of length n observations, and uses the method of Welch ([21]) or some other method to discard the first l observations from each run. Then, the IID sample means are used to compute point and interval estimators for the steady-state mean μ.

3. Regenerative method: This method assumes the identification of time indices at which the process $\{X_i\}$ probabilistically *starts over* and uses these regeneration epochs for obtaining IID random variables that can be used to compute point and interval estimates for the mean μ. As a result, it eliminates the need to detect the length of the initial transient period.

4. Batch means method: The method of batch means is frequently used to estimate the steady-state mean μ or the variance σ^2 and owes its popularity to its simplicity and effectiveness. This approach divides the output of a long simulation run into a number of contiguous *batches* and uses the sample means of these batches (or *batch means*) to produce point and interval estimators.

5. Standardized time series method: The standardized time series approach of data analysis was proposed by Schruben ([18]). This approach standardizes the entire time series instead of standardizing a single estimator (e.g. the sample mean) as the classical approach does. The standardized time series has as its limit a Brownian bridge stochastic process. Properties of the Brownian bridge are used to develop confidence interval estimators for simulation output analysis.

8 Preliminary Output Analysis Results

FRs are given 17 days, starting with the Monday of the assignment week for each month, to complete each assignment. Therefore, the simulation model starts in a state of no personal visits for all cases assigned each month. We will assume the simulation of field operations is the type of finite-horizon simulations. The estimations of the performance measures via independent replications will be used for the output analysis.

For the one thousand observations, a point estimate and the 95% confidence interval estimation of the mean and variance of the performance measures, such as response rate, average number of visits per case, and cost are analyzed using Equations (13), (14), and (15) with $k = 1000$ and $\alpha = 0.95$. Table 11 shows the estimates and their 95% confidence intervals of the aforementioned performance measures.

Table 11: The Estimation of the Cost, Response Rate, and Number of Personal Visits

Performance Measure	Mean			Variance		
	Estimate	95% Conf.	Limits	Estimate	95% Conf.	Limits
Cost ($)	25,475	25,454	25,495	111870	102673	122367
Response Rate (%)	86.04	85.94	86.15	3.02	2.77	3.30
Average # of Personal Visits	1.74188	1.74020	1.74356	0.0007356	0.0006751	0.0008046

Next, we show an example of how the perfromance measures would change when we change some of the parameters. Those parameters are controllable:

1. The starting time of each day by the field representatives: we assume that all the FRs start at 10:00 AM, 12:00 noon, or 3:00 PM. These parameter settings are based on the contact/no-contact distributions given in Table 8. We repeat part of the distributions in Table 12. Note that we consider *Potential Refusal* as *Contact*. Table 12 shows that the overall contact probability (55.32%) is higher during the hours of 3:00 to 8:00 PM than during the hours of 10:00 AM to 3:00 PM.

Table 12: The Selected Frequency Distributions of Contact/No-Contact

	Hours	Sun(%)	Mon(%)	Tue(%)	Wed(%)	Thur(%)	Fri(%)	Sat(%)	Overall(%)
No-Contact	10:00-12:00	50.98	48.46	50.25	49.33	46.38	48.91	44.83	48.12
Contact		49.02	51.54	49.75	50.67	53.62	51.09	55.17	51.88
No-Contact	12:00-15:00	47.03	49.37	48.90	48.78	49.69	48.04	45.75	48.36
Contact		52.97	50.63	51.10	51.22	50.31	51.96	54.25	51.64
No-Contact	15:00-20:00	48.05	44.50	43.95	43.06	43.74	46.14	48.23	44.68
Contact		51.95	55.50	56.05	56.94	56.26	53.86	51.77	55.32

2. The number of field representatives: we increase the number of FRs from 10 to 15 and keep the same number of cases assigned at 1050. The covered geographical area is changed from 3,600 square miles to 2,401 square miles. The number of cases assigned to each FR is also changed from 105 to 70. The number of working days is reduced from 17 to 11.

Table 13 shows the nine parameter settings discussed above. For each parameter setting, we generate 1000

Table 13: The Nine Parameter Settings for the Experiments

Setting	Starting Time	# of FRs	Days	Area	FR-Days	Adjusted Days
1	10:00	10	17	3600	170	17.00
2	12:00	10	17	3600	170	17.00
3	15:00	10	17	3600	170	17.00
4	10:00	15	11	2401	165	11.33
5	12:00	15	11	2401	165	11.33
6	15:00	15	11	2401	165	11.33
7	10:00	15	17	2401	255	11.33
8	12:00	15	17	2401	255	11.33
9	15:00	15	17	2401	255	11.33

observations, the estimates of the performance measures are given in Table 14. Table 14 also shows the

Table 14: The Estimates of the Performance Measures of the Nine Parameter Settings

Setting	Cost	Response Rate(RR)	Average Visits	Adjusted to 170 FR-Days			Cost Savings
				Cost	Response Rate(RR)	Average Visits	
1	$25,375	86.19%	1.72	$25,375	86.19%	1.72	--
2	$25,238	86.86%	1.71	$25,238	86.86%	1.71	--
3	$25,475	86.04%	1.74	$25,475	86.04%	1.74	--
4	$20,722	82.23%	1.68	$21,349	84.72%	1.73	15.86%
5	$20.575	83.50%	1.66	$21,199	86.03%	1.71	16.00%
6	$20,589	83.88%	1.67	$21,213	86.42%	1.72	16.73%
7	$24,545	89.93%	1.78	RR gain	3.74%	--	3.27%
8	$24.085	89.96%	1.75	RR gain	3.10%	--	4.57%
9	$23,926	89.98%	1.75	RR gain	3.94%	--	6.08%

adjustments of the performance measures to 170 FR-Days for the parameter settings of 4, 5, and 6. By visual inspection, there is no significant difference among the parameter settings of starting time for each day for

all three performance measures. However, there is cost saving if more field representatives are assigned to the 1050 cases as indicated in Table 14 that the settings of 4, 5, and 6 have potential cost savings of 15.86%, 16.00%, and 16.73% over the settings of 1, 2, and 3, respectively. We also examine where the cost saving is coming from. Table 15 shows the cost estimates with seed 169001 for the parameter settings 3 and 6, where five more FRs are assigned to the 1050 cases. The last row labeled `Adjusted` is the adjustments to 170 FR-Days for parameter setting 6. The total traveling distance is 35,012 miles for parameter setting 3 and 27,922 miles for parameter setting 6. It is a saving of 20.25%. Therefore, a smaller PSU area would reduce the traveling time for the FRs, meaning *less time on the roads and more time knocking on the doors.*

Parameter settings 7, 8, and 9 are used to examine the effect of the response rate if we would like the FRs to work 17 days instead of 11 days. The results indicate that these three parameter settings have cost savings of 3.27%, 4.57%, and 6.08% over parameter settings 1 to 3, respectively. The response rates also have increases of 3.74%, 3.10%, and 3.94%, respectively. These are strong evidences that reducing the cost while increasing the response rate is feasible for the field operations if the parameters are properly set.

Table 15: The Cost Estimates of the Replication with Seed 169001

FR	Total time (hours)	Wages ($)	Total distance (miles)	Mileage ($)	Total cost ($)
Parameter Setting 3					
0	129.35	1293.50	3304	1156.40	2449.90
1	135.40	1354.00	3569	1249.15	2603.15
2	138.95	1389.50	3691	1291.85	2681.35
3	136.83	1368.33	3710	1298.50	2666.83
4	127.33	1273.33	3229	1130.15	2403.48
5	130.67	1306.67	3255	1139.25	2445.92
6	134.82	1348.17	3412	1194.20	2542.37
7	133.17	1331.67	3503	1226.05	2557.72
8	132.15	1321.50	3448	1206.80	2528.30
9	150.70	1507.00	3891	1361.85	2868.85
Total	1349.37	13493.67	35012	12254.20	25747.87
Parameter Setting 6					
0	71.83	718.33	1747	611.45	1329.78
1	73.42	734.17	1803	631.05	1365.22
2	71.75	717.50	1772	620.20	1337.70
3	73.85	738.50	1885	659.75	1398.25
4	71.38	713.83	1701	595.35	1309.18
5	73.05	730.50	1835	642.25	1372.75
6	77.12	771.17	1865	652.75	1423.92
7	73.37	733.67	1801	630.35	1364.02
8	79.25	792.50	1918	671.30	1463.80
9	69.75	697.50	1673	585.55	1283.05
10	79.18	791.83	1926	674.10	1465.93
11	75.90	759.00	1861	651.35	1410.35
12	69.17	691.67	1713	599.55	1291.22
13	70.08	700.83	1706	597.10	1297.93
14	78.42	784.17	1895	663.25	1447.42
Total	1107.52	11075.17	27101	9485.35	20560.52
Adjusted	1141.08	11410.78	27922	9772.78	21183.57

9 Summary

9.1 Conclusions

In conclusion, we have shown that the simulation model can be used for optimizing the field operations by setting the controllable parameters before a decision is made and implemented. The cost savings might be enormous as shown in the example (about 16%) of Section 8 and would not be at the expense of the response rate. If more working days are needed by FRs, a cost saving with higher response rate is also feasible.

Figure 5 shows how the optimization of field operations cost can be achieved. In the figure, the solid line shows the direct cost of FRs vs. the number of FRs[5]. The preliminary result indicates that the direct cost is a decreasing function of the number of FRs. If the hiring and training cost (or the overhead) of FRs is an increasing function of the number of FRs, shown in Figure 5 as a dot line, then the minimum total cost can be located by examining the dash line, which is the sum of the solid and dot lines, of the figure.

In summary, the following preliminary tasks have been used for this study:

1. Model conceptulization: The model will begin simply and grow until a model of appropriate complexity has been developed.

2. Data collection: A data set for each variable from the NHIS is collected with help from Demographic Surveys Division.

3. Input data analysis: Determine the distribution function of the data set collected for each variable.

4. Model translation: The conceptual model constructed in Step 1 is coded into a computer-recognizable form, an operational model.

5. Verification and validation: Verification concerns the operational model. Is it performing properly? Validation is the determination that the conceptual model is an accurate representation of the field operations. The process of verification and validation is an iterative one. New details will be added to the model and new results are presented to Field Division (or field operations experts). If the results are not sufficiently accurate, Field Division/experts identify other details that should be included. These details are added, and the cycle starts anew. At some point, we "must" agree that the model is "close enough" to provide useful information. The agreement can be based on the simulation output data and the Field Division historical data.

6. Experimental design: For each scenario that is to be simulated, decisions need to be made concerning the length of the simulation run, the number of runs (also called *replications*), and the manner of initialization, as required.

7. Production runs and output analysis: Production runs and their subsequent output analysis are used to estimate the performance measures for the scenario that are being simulated.

8. Senitivity and feasibility study.

9. Documentation and reporting.

The simulation model will be modified according to the aforementioned Step 5 to make the model valid for a better tool of decision making.

9.2 Future Work

As mentioned in Section 6, the simulation model described in this paper is for the simplified field operations of surveys. For example, we explored the possible simulation models at national level. However, other geographic attributes such as region and MSA (Metropolitan Statistical Area) status should be explored in the future. These attributes have been shown as a measure associated with contact [6]. Other future work that should be included is listed as following:

1. physical impediments, at-home patterns of households as described in Section 4.3;

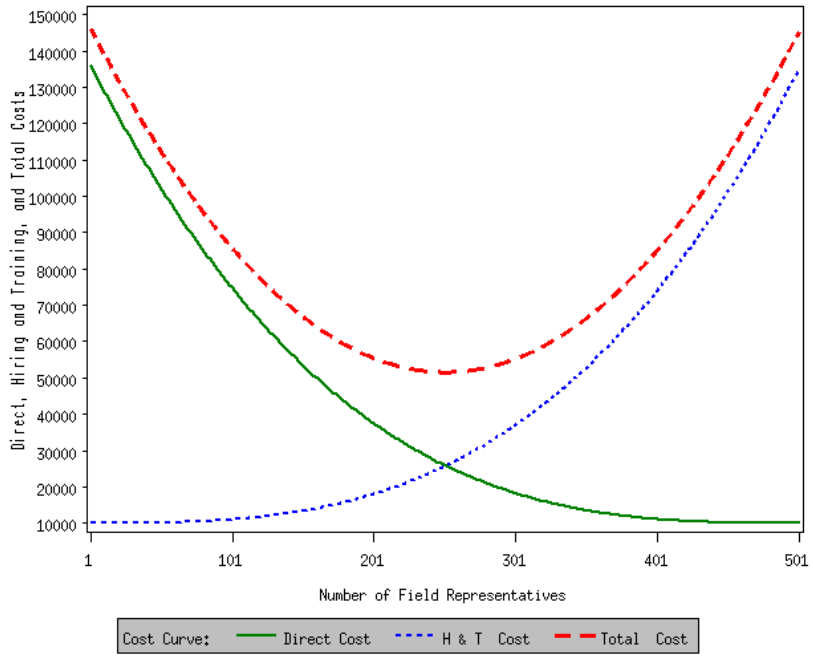
2. interviewer strategies that influence contact such as advance letters and notices of visit [8] and telephone interviews after the first contact;

3. multiple visits of completed interviews (outcome 201), a completed interview may need several visits of the same household in which the interview lengths of the visits may be correlated;

4. a sample household may have several unrelated persons living in the same house, it is required by NHIS to interview each one of them;

[5]We use 501 as the total number of FRs in the figure for illustration purpose only, it is not the actual number of FRs for the NHIS.

Figure 5: Optimization of Field Operations Cost.



5. classification of interviewers (field representatives or supervisory field representatives) based on their experiences and trainings.

Appendix

A Background Information of Input Modeling

A.1 Selected Probability Distributions

Beta Distribution The probability density function of a beta distribution, Beta$(\alpha, \beta, \theta, \lambda)$, is

$$f(x) = \begin{cases} (x-\theta)^{\alpha-1}(\theta+\lambda-x)^{\beta-1}/B(\alpha,\beta)\lambda^{(\alpha+\beta-1)}, & \text{if } \theta < x < \theta+\lambda; \\ 0, & \text{otherwise,} \end{cases} \tag{19}$$

where $B(\alpha, \beta)$ is the *beta function* defined by

$$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt \tag{20}$$

for any real numbers $\alpha > 0$ and $\beta > 0$. Note that:

$$B(\alpha,\beta) = B(\beta,\alpha) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}.$$

The parameters, α and β, are the shape parameters for the beta distribution. The threshold parameter is θ, whose value must be less than the minimum data value. The scale parameter is λ and the sum of the values of scale and threshold must be greater than the maximum data value for the variable being analyzed.

There is no closed form, in general, for the distribution function. If either α or β is a positive integer, a binomial expansion can be used to obtain $F(x)$, which will be a polynomial in x, and the powers of x will be, in general, positive real numbers ranging from 0 through $\alpha + \beta + 1$. The data with the beta distribution can be generated from gamma distributions; see Appendix B.2 for details.

Weibull Distribution The density function of the Weibull distribution is

$$f(x) = \begin{cases} \alpha\lambda^{-\alpha}(x-\theta)^{\alpha-1}e^{-((x-\theta)/\lambda)^{\alpha}}, & \text{if } x > \theta; \\ \\ 0, & \text{otherwise,} \end{cases} \tag{21}$$

and the distribution function is

$$F(x) = \begin{cases} 1 - e^{-((x-\theta)/\lambda)^{\alpha}}, & \text{if } x > \theta; \\ \\ 0, & \text{otherwise.} \end{cases} \tag{22}$$

The shape parameter is $\alpha > 0$ for the Weibull distribution and the scale parameter $\lambda > 0$. The threshold parameter is θ, whose value must be less than the minimum data value.

Gamma Distribution The density function of the gamma distribution is

$$f(x) = \begin{cases} \lambda^{-\alpha}(x-\theta)^{\alpha-1}e^{-(x-\theta)/\lambda}/\Gamma(\alpha), & \text{if } x > \theta; \\ \\ 0, & \text{otherwise,} \end{cases} \tag{23}$$

where $\Gamma(\alpha)$ is the *gamma function* defined by

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1}e^{-t}dt \tag{24}$$

for any real number $\alpha > 0$. Note that:

$$\begin{cases} \Gamma(\alpha+1) = \alpha\Gamma(\alpha) & \text{for any } \alpha > 0, \\ \\ \Gamma(k+1) = k! & \text{for any nonnegative integer } k, \\ \\ \Gamma(k+\frac{1}{2}) = \sqrt{\pi}\cdot 1\cdot 3\cdot 5\cdots(2k-1)/2^k & \text{for any positive integer } k, \\ \\ \Gamma(\frac{1}{2}) = \sqrt{\pi}. \end{cases}$$

If α is not an integer, there is no closed form for the distribution function. If α is a positive integer, then the distribution function is

$$F(x) = \begin{cases} 1 - e^{-(x-\theta)/\lambda}\sum_{j=0}^{\alpha-1}((x-\theta)/\lambda)^j/j!, & \text{if } x > \theta; \\ \\ 0, & \text{otherwise.} \end{cases} \tag{25}$$

The shape parameter is $\alpha > 0$ and the scale parameter is $\lambda > 0$. The threshold parameter is θ, whose value must be less than the minimum data value.

Exponential Distribution The probability density function of exponential distribution is

$$f(x) = \begin{cases} \frac{1}{\lambda}e^{-(\frac{x-\theta}{\lambda})}, & \text{if } x \geq \theta; \\ \\ 0, & \text{otherwise.} \end{cases} \tag{26}$$

The scale parameter of the distribution is $\lambda > 0$. The threshold parameter is θ whose value must be less than the minimum data value. The distribution function is

$$F(x) = \begin{cases} 1 - e^{-(\frac{x-\theta}{\lambda})}, & \text{if } x \geq \theta; \\ \\ 0, & \text{otherwise.} \end{cases} \tag{27}$$

Note that:

1. The exp(λ) distribution is a special case of both the gamma and Weibull distributions (for shape parameter $\alpha = 1$ and scale parameter λ in both cases);

2. If X_1, X_2, ..., X_m are independent exp(λ) random variables, then $X_1 + X_2 + \cdots + X_m \sim$ gamma(m, λ) with shape parameter m and scale parameter λ, also called the m-Erlang distribution;

3. The exponential distribution is the only continuous distribution with the memoryless property. A random variable X is said to have the *memoryless property* if

$$P\{X > t + s \mid X > t\} = P\{X > s\} \text{ for all } t, s \geq 0.$$

Lognormal Distribution The density function of the lognormal distribution is

$$f(x) = \begin{cases} \frac{1}{\lambda\sqrt{2\pi}(x-\theta)} e^{\frac{-(\ln(x-\theta)-\zeta)^2}{2\lambda^2}}, & \text{if } x > \theta; \\ 0, & \text{otherwise.} \end{cases} \tag{28}$$

The shape parameter of the distribution is $\lambda > 0$, $\zeta \in (-\infty, \infty)$ is the scale parameter, and θ is the threshold parameter, whose value must be less than the minimum data value. There is no closed form for the distribution function. Note that:

1. $X \sim LN(\zeta, \lambda^2)$ if and only if $\ln X \sim N(\zeta, \lambda^2)$. Thus, if one has data X_1, X_2, ..., X_n which are thought to be lognormal, the logarithms of the data points, $\ln X_1$, $\ln X_2$, ..., $\ln X_n$ can be treated as normally distributed data for purposes of hypothesizing a distribution, parameter estimation, and goodness-of-fit testing;

2. As $\lambda \to 0$, the lognormal distribution becomes degenerate at e^{ζ}. Thus, lognormal densities for small λ have a sharp peak at the mode;

3. $\lim_{x \to 0} f(x) = 0$, regardless of the parameters.

A.2 Four Goodness-of-Fit Tests

Chi-Square Tests The oldest goodness-of-fit hypothesis test is the *chi-square test*. It is used to test if a sample of data came from a population with a specific distribution. For computing the chi-square statistic in either the continuous or discrete case, the data are divided into k bins (adjacent intervals) $[a_0, a_1), [a_1, a_2), \ldots, [a_{k-1}, a_k)$, where it could be that $a_0 = -\infty$, $a_k = +\infty$, or both, and the test statistic is defined as

$$\chi^2 = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j}, \tag{29}$$

where

N_j = number of X_i's in the jth interval $[a_{j-1}, a_j)$ for $J = 1, 2, \ldots, k$;
(Note that $\sum_{j=1}^{k} N_j = n$.)

p_j = the expected proportion of the X_j's that would fall in the jth interval if we were sampling from the fitted distribution.

In the continuous case,

$$p_j = \int_{a_{j-1}}^{a_j} f(x)dx,$$

where f is the density of the fitted distribution. For discrete data,

$$p_j = \sum_{\{i: a_{j-1} \leq x_i < a_j\}} g(x_i),$$

where g is the mass function of the fitted distribution. Since np_j is the expected number of the n X_i's that would fall in the jth interval if H_0 were true, we would expect χ^2 to be small if the fit is good. Therefore, we reject H_0 if χ^2 is too large. The test statistic follows, approximately, a chi-square distribution with $(k-m-1)$

degrees of freedom where m = the number of estimated parameters (including *location* and scale parameters and *shape* parameters) for the distribution. Therefore, H_0 is rejected if $\chi^2 > \chi^2_{k-m-1,1-\alpha}$, where $\chi^2_{k-m-1,1-\alpha}$ is the upper $1-\alpha$ critical point for a chi-square distribution with $(k-m-1)$ df. An attractive feature of the chi-square test is that it can be applied to any univariate distribution for which you can calculate the cumulative distribution function. The chi-square test is applied to binned data (i.e., data put into classes). For non-binned data we can simply calculate a histogram or frequency table before generating the chi-square test. The value of the chi-square test statistic depends on how the data are binned. Another disadvantage of the chi-square test is that it requires a sufficient sample size in order for the chi-square approximation to be valid. The chi-square test is an alternative to the Anderson-Darling and Kolmogorov-Smirnov tests. The chi-square test can be applied to discrete distributions such as the binomial and the Poisson. The Kolmogorov-Smirnov and Anderson-Darling tests are restricted to continuous distributions.

Kolmogorov-Smirnov Tests The Kolmogorov-Smirnov test is used to decide if a sample comes from a population with a specific distribution. The Kolmogorov-Smirnov test is based on the empirical distribution function (EDF) $F_n(x)$ from the data $X_1, X_2, \ldots, X_n$, defined as

$$F_n(x) = \frac{\text{number of } X_i\text{'s} \leq x}{n} \tag{30}$$

for all real numbers x. Thus, $F_n(x)$ is a right-continuous step function such that $F_n(X_{(i)}) = i/n$ for each $i = 1, 2, \ldots, n$. If $F(x)$ is the fitted distribution function, the Kolmogorov-Smirnov test statistic is defined and computed as

$$D = \max_{1 \leq i \leq n} \Big(F(X_{(i)}) - \frac{i-1}{n}, \frac{i}{n} - F(X_{(i)}) \Big). \tag{31}$$

The hypothesis regarding the distributional form is rejected if the test statistic, D, is greater than the critical value obtained from a table. A serious limitation of the test is that the fitted distribution must be fully specified (i.e. the location, scale, and shape parameters cannot be estimated from the data.)

Cramér-von Mises Tests The Anderson-Darling statistic and the Cramér-von Mises statistic belong to the quadratic class of EDF statistics. This class of statistics is based on the squared difference $(F(x) - F_n(x))^2$. Quadratic statistics have the following general form:

$$Q = n \int_{-\infty}^{\infty} \Big(F(x) - F_n(x) \Big)^2 \psi(x) dF(x) \tag{32}$$

The function $\psi(x)$ weights the squared difference $(F(x) - F_n(x))^2$. Cramér-von Mises statistic W^2 measures the quadratic deviations between the empirical distribution function $F_n(x)$ and the fitted distribution function $F(x)$, multiplied by the weighting function $\psi(x) = 1$:

$$W^2 = n \int_{-\infty}^{\infty} \Big(F(x) - F_n(x) \Big)^2 dF(x) \tag{33}$$

The Cramér-von Mises statistic is computed as

$$W^2 = \sum_{i=1}^{n} \Big(F(X_{(i)}) - \frac{2i-1}{2n} \Big)^2 + \frac{1}{12n} \tag{34}$$

Anderson-Darling Tests The Anderson-Darling test is used to test if a sample of data came from a population with a specific distribution. It is a modification of the Kolmogorov-Smirnov (K-S) test and gives more weight to the tails than does the K-S test. The K-S test is distribution free in the sense that the critical values do not depend on the specific distribution being tested. The Anderson-Darling test makes use of the specific distribution in calculating critical values. This has the advantage of allowing a more sensitive test and the disadvantage that critical values must be calculated for each distribution. The Anderson-Darling test is an alternative to the chi-square and Kolmogorov-Smirnov goodness-of-fit tests. The Anderson-Darling test statistic is defined as

$$A^2 = n \int_{-\infty}^{\infty} \Big(F(x) - F_n(x) \Big)^2 \Big(F(x) \Big(1 - F(x) \Big) \Big)^{-1} dF(x) \tag{35}$$

Here the weight function is $\psi(x) = (F(x)(1 - F(x)))^{-1}$. The Anderson-Darling statistic is computed as

$$A^2 = -n - S \tag{36}$$

where

$$S = \frac{1}{n}\sum_{i=1}^{n}\left((2i-1)\ln F(X_{(i)}) + (2n+1-2i)\ln\Big(1 - F(X_{(i)})\Big)\right). \tag{37}$$

$F(x)$ is the cumulative distribution function of the specified distribution. The critical values for the Anderson-Darling test are dependent on the specific distribution that is being tested. The test is a one-sided test and the hypothesis that the distribution is of a specific form is rejected if the test statistic, A^2, is greater than the critical value. Note that for a given distribution, the Anderson-Darling statistic may be multiplied by a constant (which usually depends on the sample size, n).

B Pseudo Code of Random Number and Variate Generations

B.1 Random Number Generations

The random number generation pseudo code, where `%` is the modulus or remainder operator:
`pseudo(z)` {

`a, b, c, d, h, k, z are integers`

`a = z, b = a/(`2^{16}`-1), c = (a%(`2^{16}`-1))`$\times 7^5$

`d = c/(`2^{16}`-1), h = (b`$\times 7^5$`) + d, k = h/(`2^{15}`-1)`

`a = (c%(`2^{16}`-1)) - (`2^{31}`-1) + ((h%(`2^{15}`-1))`$\times$`(`2^{16}`-1)) + k`

`if(a < 0) z = a + (`2^{31}`-1) else z = a`

`return(z)`

}
Then we have a table of 128 elements:

```
z = table[0] = pseudo(seed), z = table[1] = pseudo(z),

z = table[2] = pseudo(z), z = table[3] = pseudo(z),

z = table[4] = pseudo(z), z = table[5] = pseudo(z),

.................................................

z = table[126] = pseudo(z), z = table[127] = pseudo(z),
```

in whcih each call of `pseudo(z)` generates a different value of `z`. Finally, a random number is obtained by:
`rand()` {

```
z = pseudo(z)

y = table[z%128]/(2^31-1), where y is a number between 0 and 1

table[z%128] = z

return(y)
```

}
A new `z` is obtained and a new random number is generated when `rand()` is called each time. Therefore, the algorithm generates a sequence of random numbers when they are needed in the simulation model.

B.2 Random Variate Generations

The random variate generation pseudo code for various distributions used in the C++ code of the simulation model:

1. *Exponential* `expon(`θ`, `λ`)`, $\lambda > 0$

   Density: see Equation (26)

   Distribution Function: see Equation (27)

   Generation:

   ```
   Generate u = U(0, 1)
   Return X = θ - ((1/λ) × log(u))
   ```

2. *Uniform* `U(a, b)`, a < b

   Density:

$$f(x) = \begin{cases} \frac{1}{b-a,} & \text{if } a < x < b \\ 0, & \text{otherwise} \end{cases} \tag{38}$$

   Distribution Function:

$$F(x) = \begin{cases} 0, & \text{if } x \leq a \\ \frac{x-a}{b-a,} & \text{if } a < x < b \\ 1, & \text{if } x > b \end{cases} \tag{39}$$

   Generation:

   ```
   Generate u = U(0, 1)
   Return X = (b - a) × u + a
   ```

3. *Weibull* `weib(`α`, `θ`, `λ`)`, $\alpha, \lambda > 0$

   Density: see Equation (21)

   Distribution Function: see Equation (22)

   Generation:

   ```
   Generate u = U(0, 1)
   Return X = λ × (-log(u))^(1/α) + θ
   ```

4. *Gamma* `gamma(`α`, `θ`, `λ`)`, $\alpha, \lambda > 0$

   Density: see Equation (23)

   Distribution Function: see Equation (25)

   Generation:

   ```
   If 0 < α < 1
      Set b = (e + α)/e where e = 2.718281828
      While(true) {
         Generate u = U(0, 1), Set w = b × u
         If(w < 1) {
            Set y = w^(1/α), Generate v = U(0, 1)
            If(v ≤ e^(-y)) Return X = λ × y + θ
         }
         Else {
            Set y = -log((b - w)/α), Generate v = U(0, 1)
   ```

```
            If(v ≤ y^(α−1)) Return X = λ × y + θ
        }
    }
If α > 1
    While(true) {
        Generate u = U(0, 1), v = U(0, 1)
        Set w = -log(u), z = -log(v)
        If(z > ((α − 1) × (w - log(w) - 1))) Return X = λ × w + θ
    }
```

5. *Beta* beta(α, β, θ, λ), $\alpha, \beta > 0$

   Density: see Equation (19)

   Distribution Function: No closed form in general.

   Generation:

   Generate u = gamma(α, 0, 1) and v = gamma(β, 0, 1)
   Return X = ((λ × u)/(u + v)) + θ

6. *Normal* norm(μ, σ), $\sigma > 0$

   Density:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{40}$$

   Distribution Function: No closed form expression.

   Generation:

```
While(true) {
    Generate u = U(0, 1) and v = U(0, 1)
    Set a = 2 × u - 1 and b = 2 × v - 1, w = a^2 + b^2
    If(w < 1) {
        Set y = ((-2 × log(w)/w)^(1/2)
        Return X = (σ × a × y) + θ or Return X = (σ × b × y) + θ
    }
}
```

7. *Lognormal* lognorm(ζ, θ, λ), $\lambda > 0$

   Density: see Equation (28)

   Distribution Function: No simple closed form.

   Generation:

   Generate u = norm(ζ, λ)
   Return X = e^u + θ

Acknowledgments

The author would like to thank Mark Gorsak of Field Division, Lynn Weidman and Yves Thibaudeau, both of Statistical Research Division for their reviews and constructive comments on earlier versions of this paper. The author would also like to thank Rob Creecy (Assistant Division Chief) and Tommy Wright (Division Chief) of Statistical Research Division for their inspiration and encouragement while I was working on the simulation project.

References

[1] N. Bates. Contact Histories in Personal Visit Surveys: The Survey of Income and Program Participation (SIPP) Methods Panel. Demographic Surveys Division, U.S. Bureau of the Census, Washington, DC 20233, May 7, 2003.

[2] R. L. Bitzer. Personal Communications, 2003.

[3] Russell C.H. Cheng. Random Variate Generation. In Jerry Banks, editor, *Handbook of Simulation*, pages 139–172. John Wiley & Sons, Inc., New York, 1998.

[4] Y. S. Chow and H. Robbins. On the Asymptotic Theory of Fixed-Width Sequential Confidence Intervals for the Mean. *Annals of Mathematical Statistics*, Vol. 36:457–462, 1965.

[5] James M. Dahlhamer, Catherine M. Simile, Barbara J. Stussman, and Beth Taylor. Determinants and Outcomes of Initial Contact in the National Health Interview Survey, 2004. National Center for Health Statistics, Hyattsville, MD, May 14, 2005.

[6] James M. Dahlhamer, Barbara J. Stussman, Catherine M. Simile, and Beth Taylor. Modeling Survey Contact in the National Health Interview Survey (NHIS). National Center for Health Statistics, Hyattsville, MD, 2005.

[7] Robert M. Groves and Mick P. Couper. *Nonresponse in Household Interview Surveys.* John Wiley & Sons, Inc., New York, 1998.

[8] Robert M. Groves, Floyd J. Fowler, Mick P. Couper, James M. Lepkowski, Eleanor Singer, and Roger Tourangeau. *Survey Methodology.* John Wiley & Sons, Inc., New York, 2004.

[9] Alfred Hartmann and Herb Schwetman. Discrete-Event Simulation of Computer and Communication Systems. In Jerry Banks, editor, *Handbook of Simulation*, pages 659–676. John Wiley & Sons, Inc., New York, 1998.

[10] J. A. Joines and S. D. Roberts. Design of Object-Oriented Simulation in C++. In *Proceedings of the 1995 Winter Simulation Conference*, 1995.

[11] Ron Laughery, Beth Plott, and Shelly Scott-Nash. Simulation of Service Systems. In Jerry Banks, editor, *Handbook of Simulation*, pages 629–644. John Wiley & Sons, Inc., New York, 1998.

[12] A. M. Law and W. D. Kelton. *Simulation Modeling and Analysis.* McGraw-Hill, New York, 2nd edition, 1991.

[13] Pierre L'Ecuyer. Random Number Generation. In Jerry Banks, editor, *Handbook of Simulation*, pages 93–137. John Wiley & Sons, Inc., New York, 1998.

[14] Mani S. Manivannan. Simulation of Logistics and Transportation Systems. In Jerry Banks, editor, *Handbook of Simulation*, pages 571–604. John Wiley & Sons, Inc., New York, 1998.

[15] Frank McGuire. Simulation in Healthcare. In Jerry Banks, editor, *Handbook of Simulation*, pages 605–627. John Wiley & Sons, Inc., New York, 1998.

[16] H. G. Meyers. Estimating Survey Cost Drivers with FLD Administrative Records and POP Demographic Data. Applied to the Current Population Survey, U.S. Bureau of the Census, Field Division Memo, January 31, 2003.

[17] Matthew W. Rohrer. Simulation of Manufacturing and Material Handling Systems. In Jerry Banks, editor, *Handbook of Simulation*, pages 519–545. John Wiley & Sons, Inc., New York, 1998.

[18] L. W. Schruben. Confidence Interval Estimation Using Standardized Time Series. *Operations Research*, Vol. 31:1090–1108, 1983.

[19] I. Shimizu and F. Lan. Approximation of Variable Costs for the National Health Interview Survey. *Proceedings of the Annual Meeting of the American Statistical Association,* August 5-9, 2001.

[20] Onur Ulgen and Ali Gunal. Simulation in the Automobile Industry. In Jerry Banks, editor, *Handbook of Simulation*, pages 547–570. John Wiley & Sons, Inc., New York, 1998.

[21] P. D. Welch. The Statistical Analysis of Simulation Results. In S. Lavenberg, editor, *The Computer Perfromance Modeling Handbook*, pages 268–328. Academic Press, San Diego, CA, 1998.