Does Your Survey Go Well?: A Practical Approach for Monitoring Data Quality on the Macroscopic Aspects in Interviewer-Administered Surveys

Survey & Health Policy Research Center
Technical Report

January 2024

dongguk UNIVERSITY SHPRC

Does Your Survey Go Well?: A Practical Approach for Monitoring Data Quality on the Macroscopic Aspects in Interviewer-Administered Surveys

Jaehoon Kim[1] and Sunwoong Kim[2]

[1]Survey and Health Policy Research Center, Dongguk University, Seoul, South Korea
[2]Survey and Health Policy Research Center and Department of Statistics, Dongguk University, Seoul, South Korea

Corresponding Author:
Sunwoong Kim, Dongguk University, 30 Pildong-ro, 1-gil, Jung-gu, Seoul, South Korea 04620
Email: sunwk@dongguk.edu

Abstract
The quality of survey data is determined during the survey. Thus, during the early or middle stage of data collection, survey statisticians or survey methodologists who design and conduct interviewer-administered surveys are often eager to obtain relevant information regarding their surveys' progress toward their goals. In particular, key survey variables including the demographics of respondents are their primary concern on macroscopic aspects for the capacity of the survey. If reliable provisional estimates of those variables can be obtained in the early stage, it would enable the active implementation of strategies to increase quality or reduce costs, along with a comprehensive judgment of survey qualities. In reality, since a target or a sufficiently large sample size is not feasible at an early stage of a survey, and more non-sampling errors are expected to occur going forward, it is common to utilize microscopic indicators to evaluate data quality instead. Diverse indicators of microscopic aspects (e.g., daily response rates, number of released reserve samples, hours per interview) exist to assess data quality during survey operations. Unfortunately, it is unclear how and to what extent each microscopic indicator or a group of such indicators will affect the quality of key survey variables (e.g., response rates and quality). This is especially true in responsive survey designs that simultaneously use multiple microscopic indicators based on abundant paradata (process data). In this paper, we suggest some cumulative sample estimates, reflecting the live flow of data, as a macroscopic indicator for monitoring data quality. Generally converging to final estimates (before weighting), they can be easily presented with a graph or a table at any point in the survey data collection process and can be used at an early stage of a survey as provisional estimates of key survey variables. We illustrate how to use them in a national CATI survey and a local CAPI survey, given the sample design and data collection protocol in each survey. This approach would help the researchers quickly and proactively check data quality to ensure the survey is on track regarding key survey variables.

Keywords
data collection process, data quality indicator, cumulative sample estimates, provisional estimates, CATI, CAPI, social demographics, survey variables

1. Introduction

The quality of survey data is determined during the survey. Thus, quality control is essential. The practice of the quality control of interviewer-administered surveys is mainly in charge of the quality monitors/team leaders in survey organizations. If computer-assisted interviewing (CAI) is used, they monitor data in real-time through the reports (study-level, case-level, interviewer-level, etc.) produced by sample management systems and evaluate the performance and productivity of interviewers. Apart from this quality management process, during the early or middle stage of data collection, survey statisticians or survey methodologists responsible for designing and conducting surveys are often eager to obtain relevant information regarding their surveys' progress toward their goals. In particular, key survey variables including the demographics of respondents are their primary concern on macroscopic aspects for the capacity of the survey. If reliable provisional estimates of those variables can be obtained in the early stage, it would enable the active implementation of strategies to increase quality or reduce costs, along with a comprehensive judgment of survey qualities.

In reality, since a target or a sufficiently large sample size is not feasible at an early stage of a survey, and more non-sampling errors are expected to occur going forward, it is common to utilize microscopic indicators to evaluate data quality instead. Diverse indicators of microscopic aspects such as the number of completed interviews, the number of call attempts, response rates, eligibility rates, interview duration, refusal conversions, cost per interview, the number of released reserve samples as well as misreporting, incompleteness, inconsistency of response, outliers, and skipping of questions exist to assess data quality during survey operations (see, e.g., Lepkowski et al., 2008, PART IV OPERATIONS, pp.317-422). Unfortunately, it is unclear how and to what extent each microscopic indicator or a group of such indicators will affect the quality of key survey variables (e.g., response rates and quality). This is especially true in responsive survey designs that simultaneously use multiple microscopic indicators based on abundant paradata (process data) (see, e.g., Couper & Lyberg, 2005; Groves & Heeringa, 2006; Wagner et al., 2012). Therefore, the uncertainty about the results of key survey variables rather tends to increase over time.

In this paper, to help address this uncertainty, we suggest some cumulative sample estimates along with theoretical explanations. They can be easily presented with a graph or a table at any point in the survey data collection process. Also, at the early or middle stage of a survey, they can be used as provisional estimates of key survey variables, which are the macroscopic indicators for monitoring data quality. They generally converge to final estimates (before weighting), as described later, and could reflect the live flow of data. We illustrate how to use them in a national CATI survey and a local CAPI survey, given the sample design and data collection protocol in each survey.

2. Cumulative Sample Estimates

The population characteristics most frequently estimated for key survey variables would be the population mean $\overline{Y}$ and population proportion P. The responses obtained from a total of n respondents (completed interviews) for any survey variable can be denoted by $y_1, y_2, \cdots, y_n$.

When a sample of respondents is selected using equal probability of sampling methods (EPSEM) such as simple random sampling, the population mean can be estimated by the (simple) sample mean $\bar{y}$ defined by

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \tag{2.1}$$

It is well-known that the sample mean $\bar{y}$ of (2.1) is an unbiased estimate of the population mean $\bar{Y}$ (see Cochran, 1977, p22). If we define y_i as 1 if the respondent possesses some attribute and as 0 if he or she does not possess it, (2.1) indicates the sample proportion p.

Based on (2.1), we suggest the cumulative sample mean $\bar{y}_{\le g}$ of (2.2) to be used as a macroscopic indicator to assess data quality during the early or middle stage of data collection.

$$\bar{y}_{\le g} = \frac{1}{\sum_{i=1}^{g} n_i}\sum_{i=1}^{g} y_i = \frac{1}{\sum_{i=1}^{g} n_i}\sum_{i=1}^{g}\sum_{j=1}^{n_i} y_{ij}, \quad i = 1, 2, \cdots, g, \cdots, G \tag{2.2}$$

where i indicates each group when distinguishing the respondents into G groups by some criterion for monitoring, g is an arbitrary group, $\le g$ represents the accumulation from the first group to the g group, and n_i is the size (number of respondents) of each group.

It is noted that n_i is a random variable dependent on the data collection process, $\sum_{i=1}^{G} n_i = n$, and accordingly n is also a random variable, if it is not fixed. For instance, the criterion for distinguishing respondents can be a survey date across the survey period, a number of call attempts, or survey interviewers.

If we define y_{ij} as 1 if the respondent has some attribute and as 0 if he or she does not have it, (2.2) denotes the cumulative sample proportion $p_{\le g}$. From $y_i = \sum_{j=1}^{n_i} y_{ij}$ in (2.2), we have the sample mean $\bar{y}_i$ for a group given by

$$\bar{y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}, \quad i = 1, 2, \cdots, g, \cdots, G \tag{2.3}$$

If we define y_{ij} as 1 or 0, (2.3) denotes the sample proportion p_i.

The cumulative sample mean over subpopulations (e.g., age groups) from (2.2) is denoted by

$$\bar{y}_{\le g,k} = \frac{1}{\sum_{i=1}^{g} n_{ik}} \sum_{i=1}^{g} y_{ik} = \frac{1}{\sum_{i=1}^{g} n_{ik}} \sum_{i=1}^{g} \sum_{j=1}^{n_{ik}} y_{ij}, \quad i = 1, 2, \cdots, g, \cdots, G \text{ and } k = 1, 2, \cdots, K \tag{2.4}$$

where k denotes a subpopulation and n_{ik} is the number of subpopulation respondents in each group. The cumulative sample proportion over subpopulations is denoted by $p_{\le g,k}$.

3. Provisional Estimates

In this section, we present the theory for obtaining provisional estimates, denoted as $\bar{y}_{pro}$ (or p_{pro}), of survey variables from the cumulative sample means (or proportions) at an early or middle stage of data collection. The distribution of the sample mean (or sample proportion) $\bar{y}_i$ for a group given by (2.3) would vary seriously for a survey variable since it is highly dependent on the data collection protocol as well as the sample design. In contrast, the distribution of the cumulative sample mean (or proportion) $\bar{y}_{\le g}$ for a survey variable given by (2.2) would not change significantly and converge to (2.1) as the number of groups cumulative is increased. This can be expressed as

$$\bar{y}_{\le g} \to \bar{y} \text{ when } g \to G \tag{3.1}$$

The rate (or order) of the convergence of the cumulative sample mean (or proportion) for a survey variable relies on both n_i (the size of each group) and the number of groups G. The larger the group size and the smaller the number of groups, the sooner the cumulative sample mean (or proportion) approaches (2.1).

The distribution of the cumulative sample mean (or proportion) can be easily graphed or tabled across the survey period. The key to obtaining a provisional estimate of a survey variable is to first identify the rough point at which the cumulative sample means (or proportions) begin to stabilize through this graph or table and then to decide the exact point. For the latter, one can find any two cumulative estimates ($\bar{y}_{\le g}$ and $\bar{y}_{\le g+1}$) that the rate of convergence (R) of (3.2) lies between 0 and 1. But $\bar{y}_{\le G} (= \bar{y})$ in (3.2) is not known until the survey is complete.

$$\frac{\left|\bar{y}_{\le g+1}-\bar{y}_{\le G}\right|}{\left|\bar{y}_{\le g}-\bar{y}_{\le G}\right|}=\frac{\left|\bar{y}_{\le g+1}-\bar{y}\right|}{\left|\bar{y}_{\le g}-\bar{y}\right|}=R\in(0,1) \tag{3.2}$$

Alternatively, another rate of convergence (r) of (3.3) using three cumulative estimates ($\bar{y}_{\le g}$, $\bar{y}_{\le g+1}$, and $\bar{y}_{\le g+2}$) without $\bar{y}_{\le G}(=\bar{y})$ can be adopted.

$$\frac{\left|\bar{y}_{\le g+2}-\bar{y}_{\le g+1}\right|}{\left|\bar{y}_{\le g+1}-\bar{y}_{\le g}\right|}=r\in(0,1) \tag{3.3}$$

In the case of using (3.3), it is recommended to obtain a provisional estimate $\bar{y}_{pro}$ of a survey variable by using the formula (3.4) rather than using one of the three cumulative estimates.

$$\bar{y}_{pro}=\frac{1}{3}\left(\bar{y}_{\le g}+\bar{y}_{\le g+1}+\bar{y}_{\le g+2}\right) \tag{3.4}$$

Alternatively, a different formula (3.5), which relies on n_i, can be used.

$$\bar{y}_{pro}=\frac{\sum_{i=1}^{g} n_i\bar{y}_{\le g}+\sum_{i=1}^{g+1} n_i\bar{y}_{\le g+1}+\sum_{i=1}^{g+2} n_i\bar{y}_{\le g+2}}{\sum_{i=1}^{g} n_i+\sum_{i=1}^{g+1} n_i+\sum_{i=1}^{g+2} n_i} \tag{3.5}$$

The higher the rate of convergence, that is, the sooner the cumulative sample mean (or proportion) begins to stabilize, the earlier we can judge the quality of the data during the survey. But it should be noted that the stability of the cumulative sample mean (or proportion) does not indicate its accuracy. It is, however, useful to forecast an approximation or limit of an unweighted estimate for a variable even during an early or middle stage of data collection. This is because the asymptotic behavior of the cumulative sample mean (or proportion) under appropriately large $\sum_i n_i$ (e.g., hundreds of respondents), which is supported by sampling theory in a finite population, is highly likely to give conclusive information about n respondents, even if the survey is not completed. This can be also applied to the cumulative sample mean (or proportion) for subpopulations.

4. Illustrations

We illustrate how one can obtain the provisional estimates of survey variables from the cumulative sample means (or proportions) at an early or middle stage of data collection using hypothetical examples created from two actual surveys conducted by the Survey & Health Policy Research Center (SHPRC) at Dongguk University, South Korea.

The short introduction, sample design, and data collection protocol for those surveys are as follows. One is a cell phone CATI survey called the 2018 National Survey of Smoking and Health (NSSH) that collected national data on tobacco use. It was implemented over 29 days except for one day off from March 12 to April 10 with an RDD sample of 15,000 cell phone numbers selected by a single-stage EPSEM from a frame. The target population was about 43,000,000 adults. Call attempts were made on weekday evenings and weekends, using specified calling schemes (e.g., days and times of the calls, number of callbacks, days and times of callbacks, time lag between calls, setting appointments, and refusal conversion). At least 10 call attempts were made to sample numbers that had not yet been contacted. Incentives were offered to the respondents. The questionnaire consisted of 119 questions that covered demographics, smoking and tobacco use, attitudes, and perceptions of smoking or cigarettes. There was a total of 968 completed cases. The response rate was 10.6% (using RR1; AAPOR, 2016). See Kim & Couper (2021) for details of this survey.

The other is a local CAPI survey called the 2012 Metropolitan Household Survey of Environment Health (MHSEH) conducted over 30 days from July 14 to August 12. The main purpose of this survey was to investigate the general awareness of environmental health and the incidence or prevalence of environmental disease among the residents who live around the industrial complexes in Incheon, which is the third-largest city in South Korea. There were 154 questions on residential exposures, occupational exposures, environmental diseases, and demographics. The target population was about 100,000 children aged 4-12 and 470,000 adults living in approximately 200,000 households. A sample of households was selected by using stratified four-stage area sampling. Each sample household within the same stratum was selected with equal probability. One or two respondents within a household were randomly selected. In South Korea, many housing units have “access impediments” that prevent strangers from contacting them. For example, about 50% of all households live in high-rise apartment buildings with locked central entrances or security guards. Moreover, the proportion of not-at-homes during the day or evening is very high and nearly a fourth of households have just one resident. It is essential, thus, to make a special effort to bring a good response rate. Therefore, the sample households were contacted based on a new administrative cooperation strategy connecting the interviewers, members of the lowest-level local government agencies, and the sample households to maximize response rates. 783 people who live in 606 households completed the survey. The maximum number of call attempts was 10. The response rate was 36.1% (RR1). See Choi, Kim, Hong, Lee, & Lee (2013) and Woo, Kim, Park, Lee, & Choi (2013) for details on the MHSEH.

In illustrating hypothetically how to obtain the provisional estimates of survey variables from the cumulative sample means (or proportions), two monitoring approaches are introduced: 1) survey date-based monitoring in the CATI survey and 2) call-based monitoring in the CAPI survey. The ‘survey date’ or ‘call’ indicates the criterion distinguishing the whole respondents

into G groups for monitoring, as described in (2.2). There is a reason why a different monitoring approach is used in each survey. It is because, for the NSSH, CATI survey, of the 968 completed cases, almost all (94%) were completed on the first or second call, so monitoring by the call is not meaningful, whereas for the MHSEH, the CAPI survey, 60% of 606 households were completed in the first or second visit and the rest were done in the other number of visits, so monitoring by call (visit) is meaningful.

4.1 Survey Date-Based Monitoring in CATI Survey

Before explaining the monitoring process using the cumulative sample mean (or proportion), we need to look at the actual distribution of 968 completed interviews in the NSSH. Figure 1 shows the distribution of the number of completed interviews each day over 29 days except for one day off from March 12 to April 10. The average number of interviews per day was 33.4 cases and the median was 26 cases. As shown in Figure 1, there is no pattern in the distribution of completed interviews by survey day (the first day, the second day, the third day, etc.). Especially during those days since the 21st day, the number of completed interviews each day continues to be low, compared to the other days. Table 1 presents the distribution of completed interviews by the number of days taken to complete the survey. Although 68% of the interviews were completed on the same day the survey started (Num. of Days Taken = 0), the remaining interviews took a varying number of days to complete the survey. Thus, given the sample design and data collection protocol in the NSSH, and assuming we have just started this survey, since the number of completed interviews each day or the number of days taken to complete the survey would be practically unpredictable, one can never know in advance what data collection would occur daily concerning any survey variables.

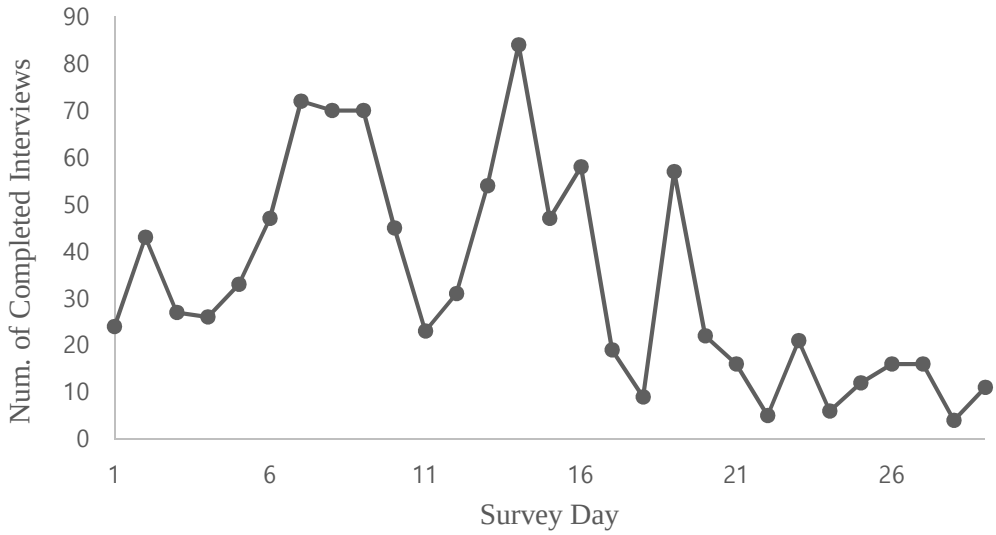

Figure 1. Distribution of the Number of Completed Interviews by Survey Days

Table 1. Distribution of Completed Interviews by the Number of Days Taken to Complete Survey

Num. of Days Taken	Num. of Completed Interviews	%
0	658	68.0
1	51	5.3
2	20	2.1
3	21	2.2
4	27	2.8
5	27	2.8
6	16	1.7
7	12	1.2
8	13	1.3
9	16	1.7
10	13	1.3
11	16	1.7
12	6	0.6
13	16	1.7
14	14	1.4
15 or more	42	4.3
Total	968	100.0

Note. '0' in the number of days taken indicates that it was completed on the same day that the survey for a sample cell phone number started

Now, let us explain the monitoring process. Assume that the NSSH already began, and during the early stage of data collection, one would like to obtain relevant information regarding the progress of the survey toward their goals. They, in particular, would like to know the proportion of respondents by gender, one of the key survey variables on macroscopic aspects of data quality for the capacity of the survey. One knows the gender distribution (49.8% for males, 50.2% for females) of the target population (43,000,000 adults) from the Census. Thus, one wonders what the gender distribution of respondents will look like eventually when the survey is completed. It would be best if a reliable provisional estimate of the gender proportion could be obtained in advance.

On the other hand, Figure 2 shows the actual sample proportion (percentage) of males and females of respondents on each survey date (a group i) in the NSSH, which starts on March 12 and ends on April 10. This sample proportion is calculated by (2.3). For example, the male sample percentage is 73.9% and the female sample percentage is 26.1% (=100.0% - 73.9%) on March 22. Although the male sample percentages on the survey dates are fitted by a regression line going down slowly, as in the Figure, the daily percentage change is very large on some of the survey dates (e.g., 46.7% on March 21 and 73.9% on March 22). Thus, at the beginning of data collection, it seems impossible to accurately estimate the proportion of male (female) respondents.

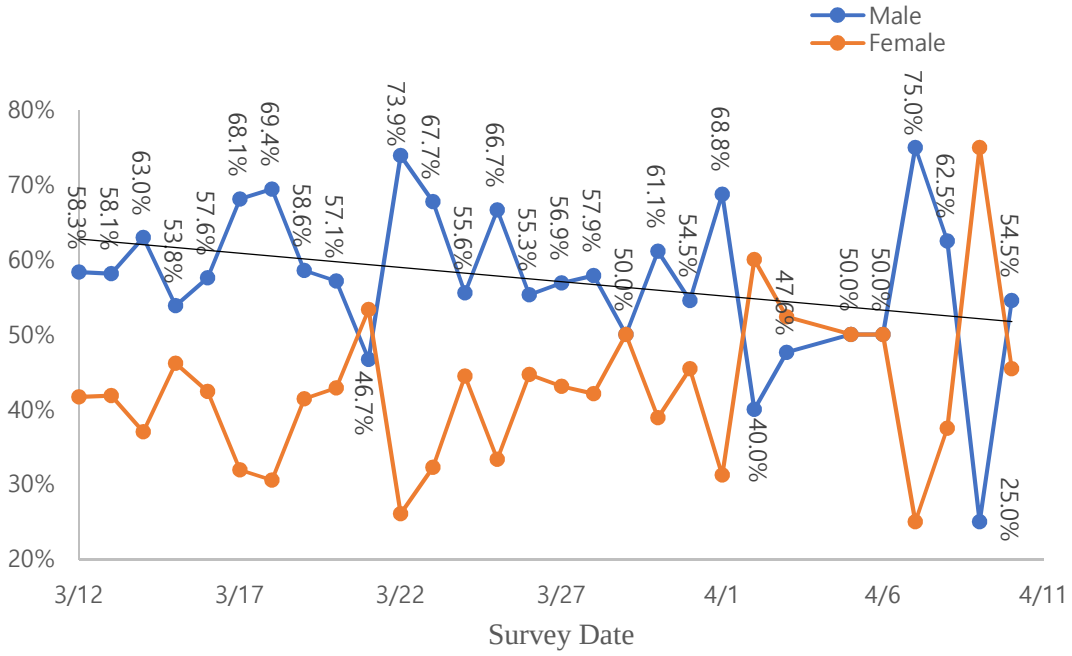

Figure 2. Actual Gender Distribution by Survey Date during the Entire Survey Period (NSSH)

Under this uncertain situation, we describe how to use the cumulative sample proportion to obtain a provisional estimate of the gender proportion. Let us assume that Figure 3 shows the cumulative sample proportions (%) of males and females calculated by (2.2) at the early stage of the daily monitoring process since the survey began on March 12. Let us say that it was found on March 21 through March 23, 9 days after the start of the survey, that the rate of convergence of (3.3) for the cumulative sample proportions (%) of males is 0.7 ($=|60.9-60.4|/|60.4-59.7| \in (0,1)$). It is noted that the cumulative completed cases (n_i) are 457, 480, and 511 (about half of the total completed cases 968) on these three survey dates, respectively. The provisional estimate p_{pro} of (3.4) on these dates is 60.3 (=(59.7+60.4+60.9)/3). Another provisional estimate p_{pro} of (3.5) is 60.4, which is very close to 60.3.

At this time, is this provisional estimate of male proportion (60.3% or 60.4%) reliable? In other words, can one be sure that the male proportion (female proportion) will be about 60% (40%) when the survey is completed? As shown in Figure 4, which presents the cumulative sample proportions for gender during the actual entire survey period, the cumulative sample proportions of males and females obtained by (2.2) are almost kept constant during the other dates since March 23. Moreover, the cumulative sample proportions of males on the final survey date (April 10) is 59.6% ($= p_{\leq G} = p$), which is very close to the provisional estimate of males (60.3% or 60.4%). Therefore, there is no doubt that the provisional estimate of males (females) is reliable.

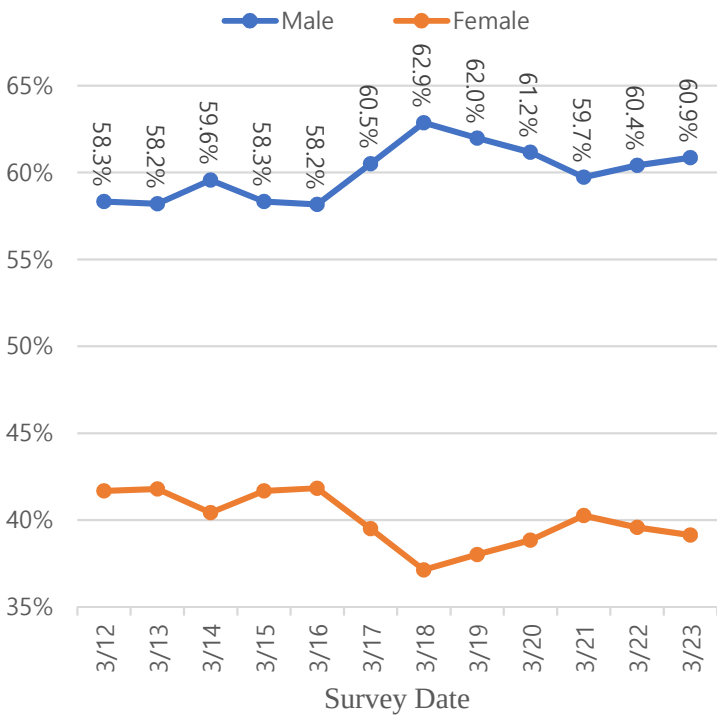

Figure 3. Cumulative Sample Proportion (%) for Gender by March 23 (NSSH)

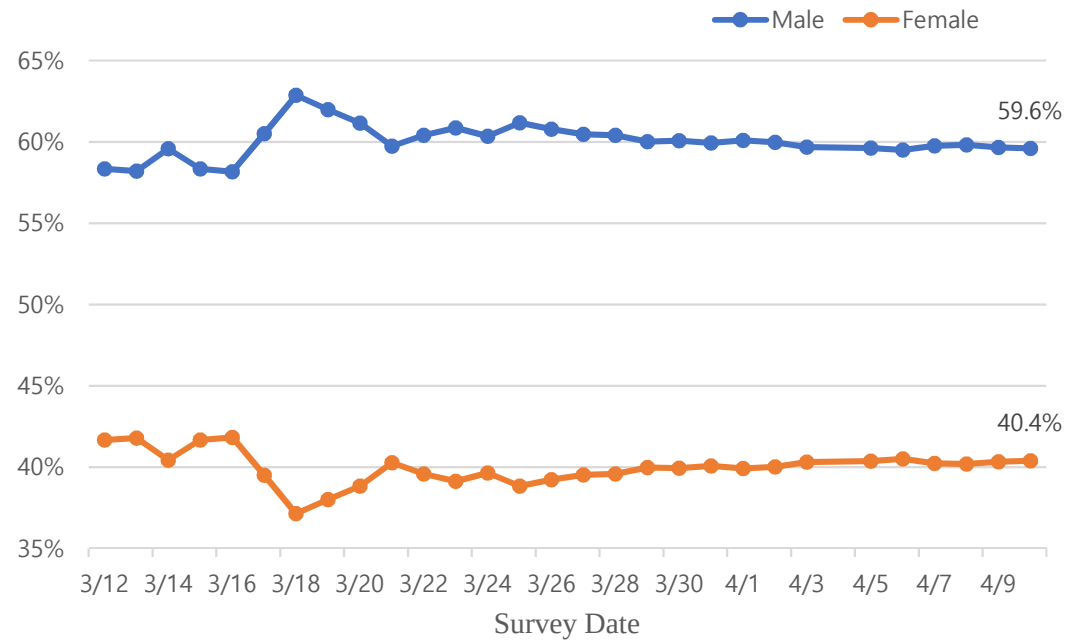

Figure 4. Cumulative Sample Proportion for Gender during the Entire Survey Period (NSSH)

What if survey statisticians or survey methodologists obtained this reliable provisional estimate of males (females), especially during the early stage of data collection? Since he or she knows the population gender distribution (male 49.8%, female 50.2%) from the Census, if necessary, it would be possible to implement responsive designs or strategies to decrease (increase) the proportion of male (female) respondents during the remaining survey period (about 20 days).

Next, we describe how to use the cumulative sample mean to obtain a provisional estimate of the mean age of respondents, another key survey variable on macroscopic aspects. Figure 5 shows the actual sample mean age (sample average age) of (2.3) on each survey date (a group i) in the NSSH. For example, the mean age of respondents is 42.0 on March 22. As can be seen in the figures with the fitted regression line, the mean age fluctuates greatly depending on the survey date. It seems to be difficult to obtain a reliable provisional estimate.

Let us assume that Figure 6 shows the cumulative mean ages of (2.2) at an early stage of the monitoring process since the survey began. When applying the survey dates (March 20 through March 22) similar to above (March 21 through March 23), the three cumulative mean ages are the same (41.0). The cumulative completed cases (n_i) are 412, 457, and 480 on these survey dates, respectively. The provisional estimate $\bar{y}_{pro}$ of (3.4) or (3.5) on these survey dates is 41.0. Is this provisional estimate reliable as well? As shown in Figure 7, the cumulative mean age at the final date (April 10) is 40.6 ($=\bar{y}_{\leq G}=\bar{y}$), which is not far from 41.0. Thus, the reliable estimation for the mean age of the respondents is feasible at an early stage of a survey.

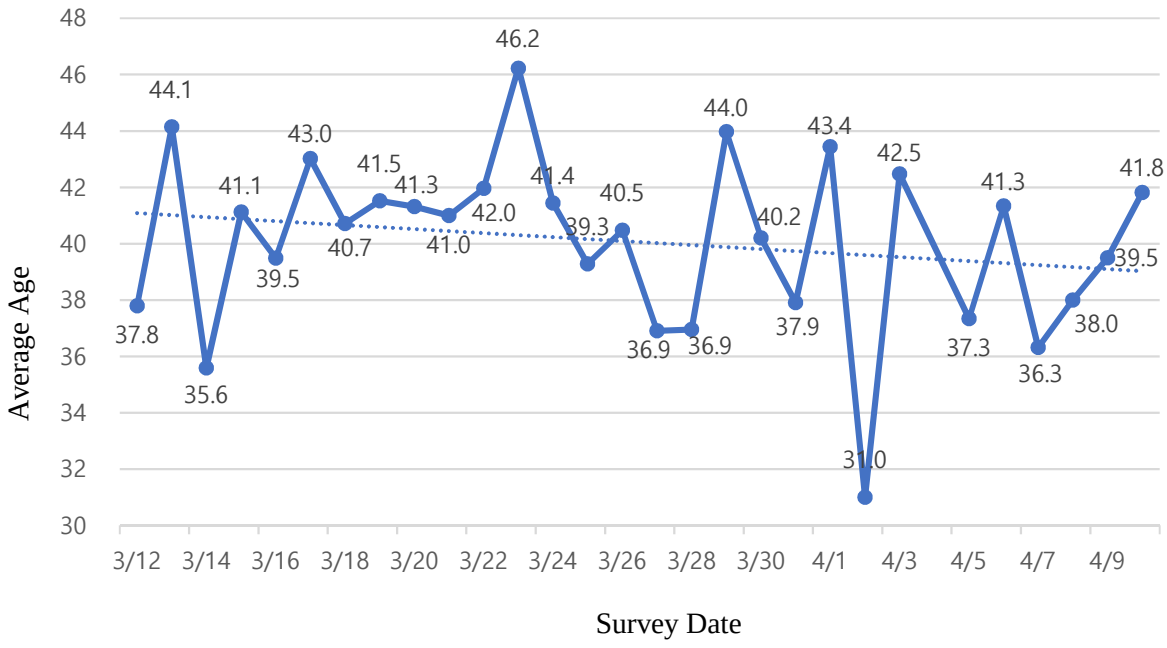

Figure 5. Mean Age Distribution by Survey Date during the Entire Survey Period (NSSH)

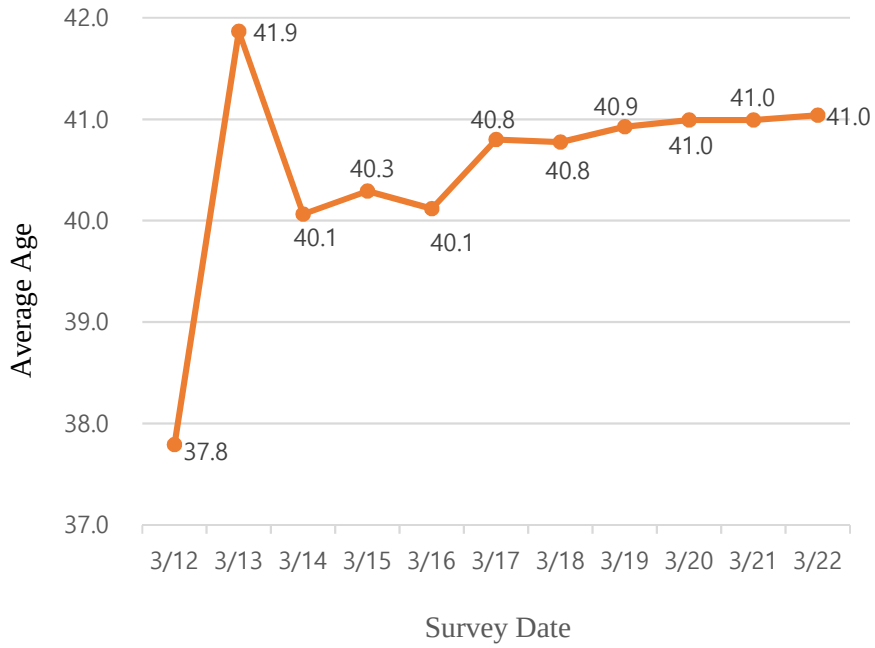

Figure 6. Cumulative Sample Mean Age by March 22 (NSSH)

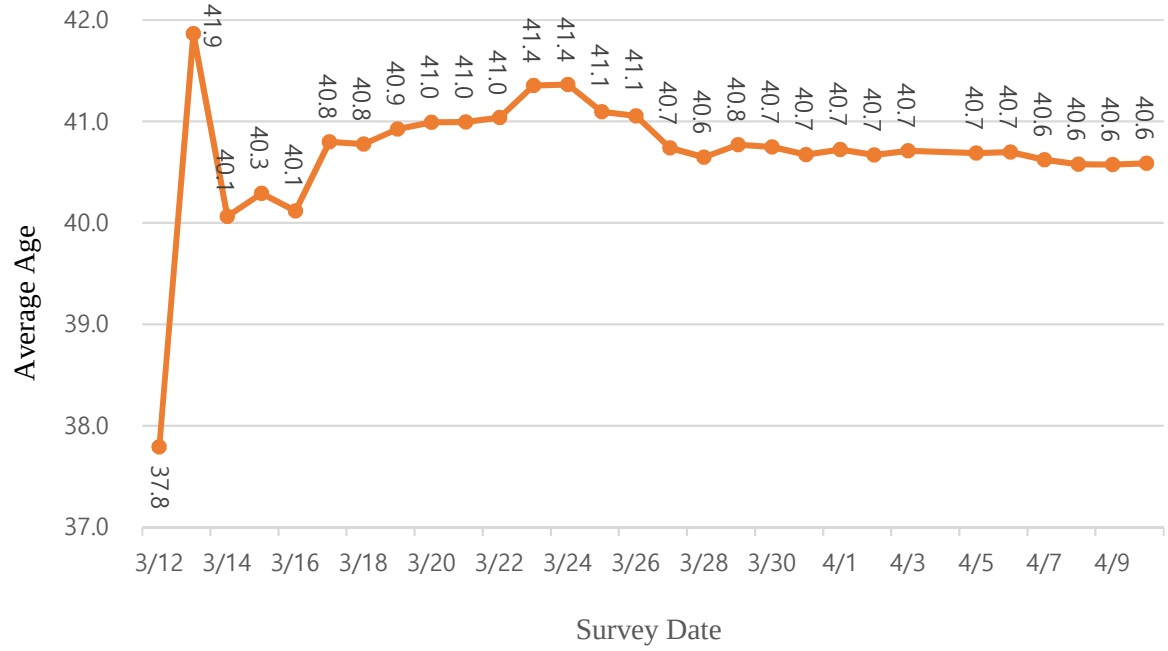

Figure 7. Cumulative Sample Mean Age during the Entire Survey Period (NSSH)

From now on, let us take a look at how the cumulative sample proportion for subpopulations of (2.4) can be used to monitor data quality for the age groups (19 or 20s, 30s, 40s, etc.). The gender described above is also a subpopulation. Considering the high stability from March 21 to March 23 in Figure 8, which presents the cumulative sample proportions (%) for age groups by March 23, we can obtain that as a provisional estimate of (3.4), 33.4% for age 19-29, 17.9% for age 30-39, 15.6% for age 40-49, 16.4% for age 50-59, 10.6% for age 60-69, and 6.1% for age 70 or over. Are these provisional estimates reliable? As shown in Figure 9, the cumulative sample proportion at each age group at the final survey date (April 10) is the same or very similar to these provisional estimates, respectively (e.g., 33.4% vs. 32.7 for age 19-29).

These reliable provisional estimates for subpopulations available at an early stage of a survey would be useful for survey statisticians or survey methodologists. For example, considering the Census results for those age groups (17.7% for age 19-29, 18.3% for age 30-39, 21.0% for age 40-49, 19.7% for age 50-59, 12.6% for age 60-69, and 10.7% for age 70 or over), he or she would like to lower the proportion, especially for age 19-29, which has a large difference (+ 15.7%) with the Census result, by using some strategies during the remaining data collection period.

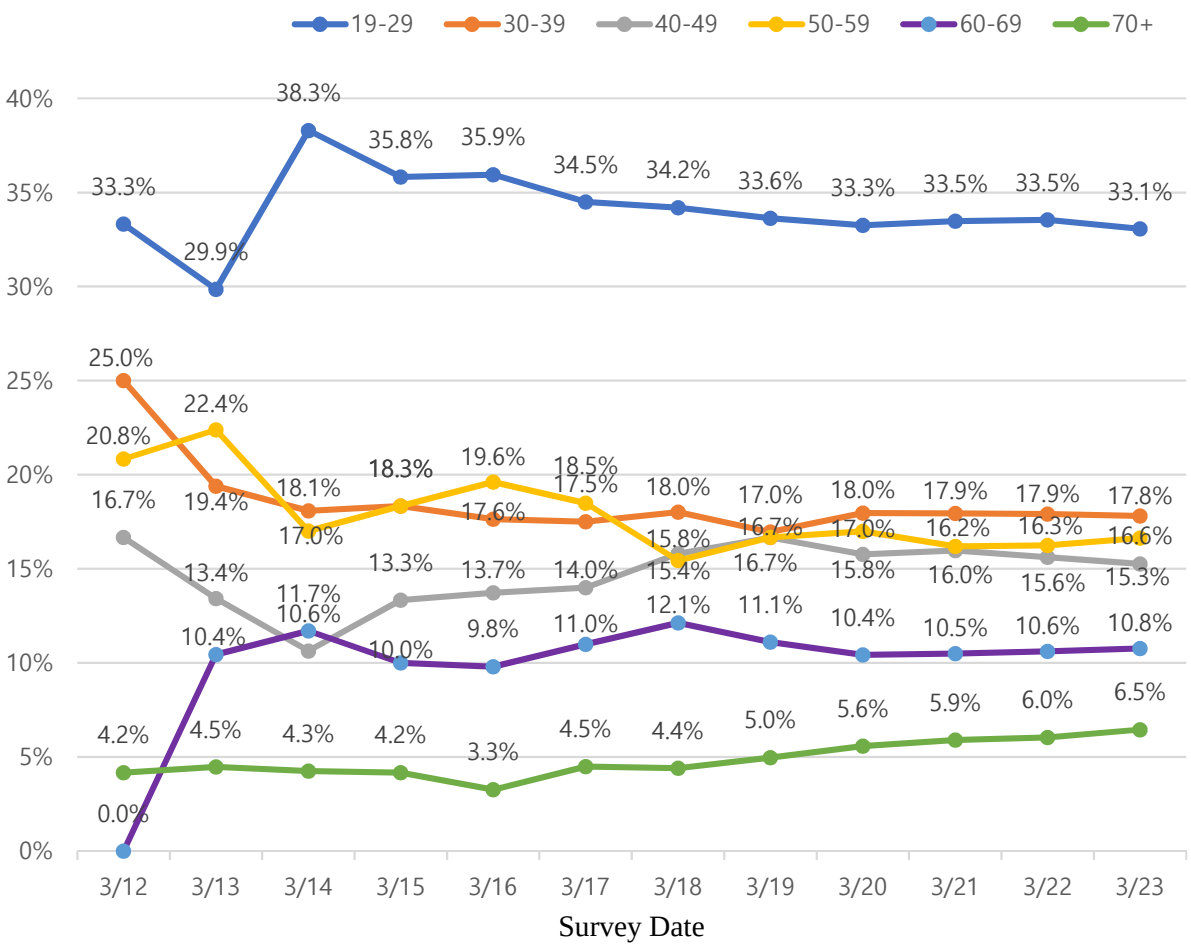

Figure 8. Cumulative Sample Proportion (%) for Age Groups by March 23 (NSSH)

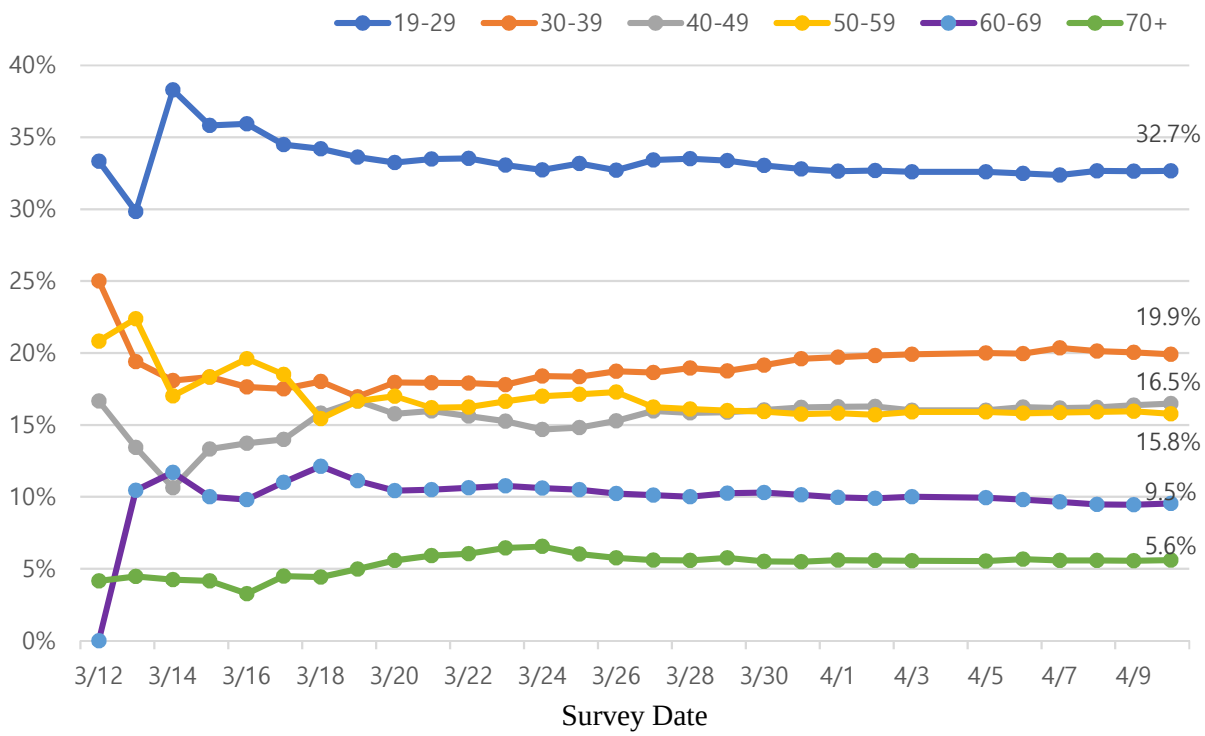

Figure 9. Cumulative Sample Proportion (%) for Age Groups during the Entire Survey Period (NSSH)

Finally, we describe how to use the cumulative sample proportion of (2.2) for a key survey variable we chose: "Have you ever smoked cigarettes in your lifetime? (Yes, No)". As shown in Figure 10, the sample percentage (2.3) of 'Yes' fluctuates greatly over the survey period. Is it possible to obtain a reliable provisional estimate of the proportion of 'Yes'?

When monitoring as shown in Figure 11, the cumulative sample proportions (52.5%, 53.4%, and 53.8%) on March 22 through March 24 yield the provisional estimates of 53.2% ('Yes') from (3.4) and 53.3% from (3.5) with $r = 0.4\,(\in (0,1))$ from (3.3). The cumulative completed cases (n_i) are 480, 511, and 565 on these survey dates, respectively. Are these estimates reliable like the demographics (gender, age) described above? Compared to the final cumulative sample proportion (51.7%) on April 10 in Figure 12, these provisional estimates (53.2% or 53.3%) are not significantly different.

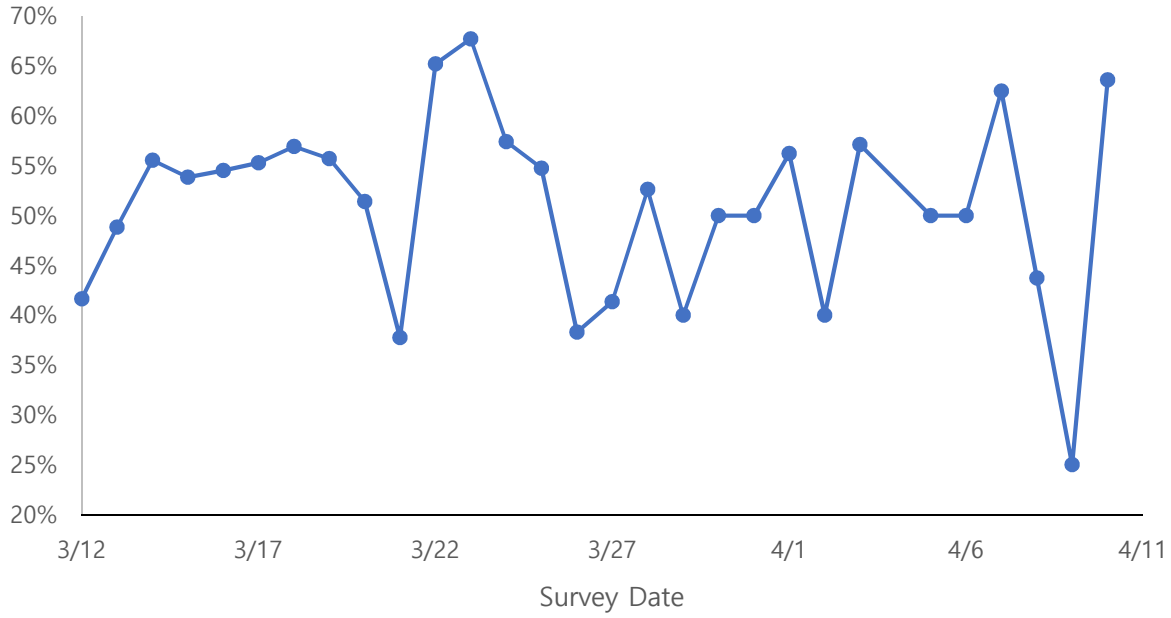

Figure 10. The Sample Proportion of 'Yes' to "Have you ever smoked cigarettes in your lifetime?" by Survey Date during the Entire Survey Period (NSSH)

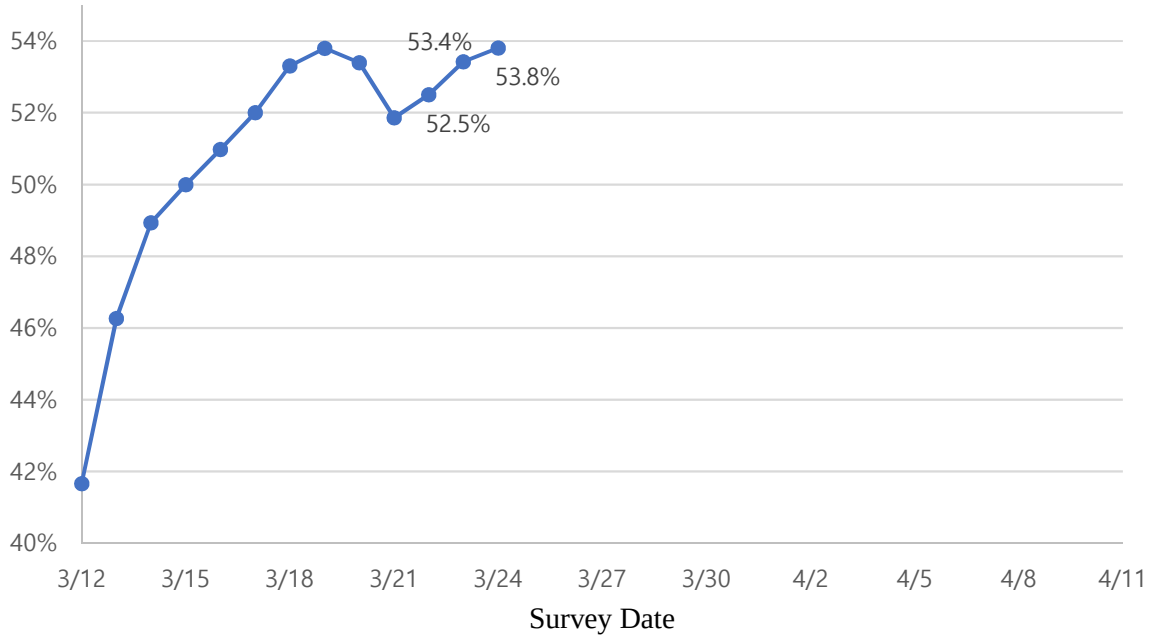

Figure 11. The Cumulative Sample Proportion of 'Yes' to "Have you ever smoked cigarettes in your lifetime?" by March 24 (NSSH)

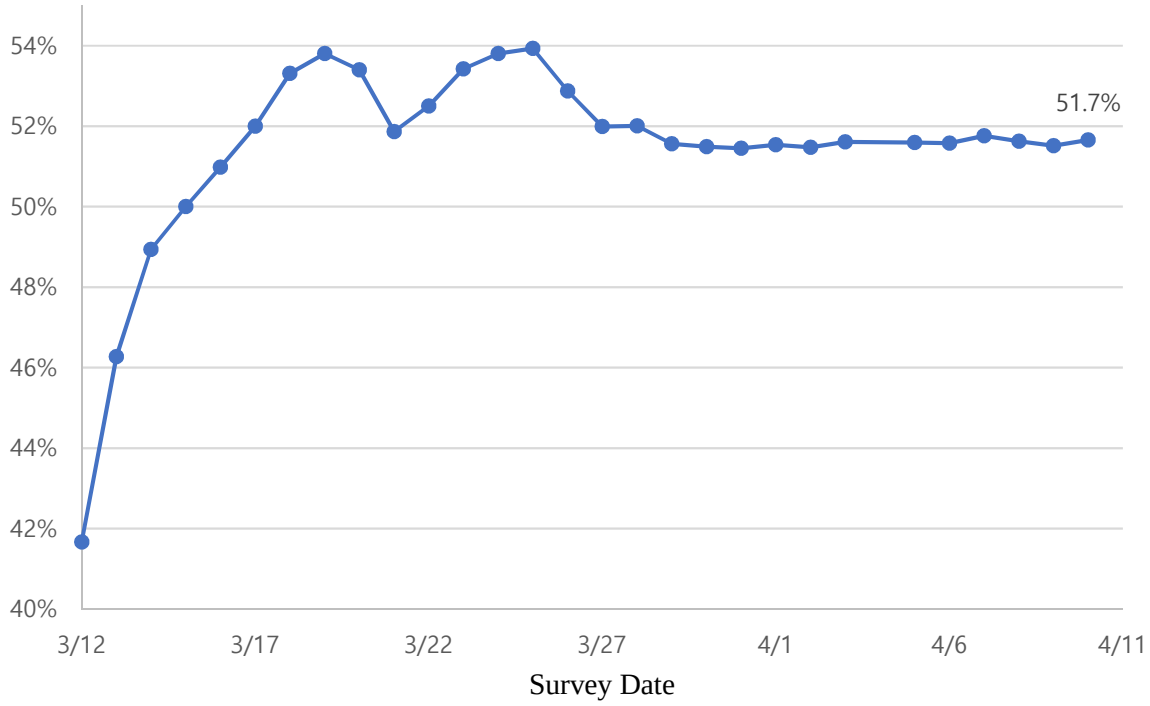

Figure 12. The Cumulative Sample Proportion of 'Yes' to "Have you ever smoked cigarettes in your lifetime?" over the Entire Survey Period (NSSH)

4.2 Call-Based Monitoring in CAPI Survey

For the MHSEH, among sampled households, 606 households completed the CAPI survey through a total of 1,478 visits (calls) and 1,082 households did not complete the survey through a total of 3,911 visits (calls). In other words, the interviewers visited a total of 5,389 times for 606 completed interviews and 1,082 uncompleted interviews and the average number of visits per completed interview was 8.9 times. We illustrate how to conduct a call-based monitoring by using the cumulative sample mean (or proportion) in the data collection process.

Table 2 shows the distribution of completed interviews according to a different number of calls (visits). The number of completed interviews gradually decreases as the number of calls made increases. Let us assume that one would like to decide whether to continue a callback after a certain number of calls. Recall that the sample households were contacted based on a new administrative cooperation strategy to maximize response rates, as explained above.

Table 3, Table 4, Table 5, and Table 6 present the cumulative sample proportion or mean by calls for gender, housing type, diseases treated in the last 12 Months, and years of residence, respectively, which are key survey variables chosen for monitoring data quality. Since the number of calls is limited to 10, we used a table instead of a graph. Using the rate of convergence (r) of (3.3) and provisional estimates of (3.4) or (3.5) is not recommended because of the limited number of calls. Instead, simultaneously considering the net differences of the cumulative sample proportion or mean between calls across different survey variables, we may find that those between the fourth and fifth calls especially began to reduce across

most survey variables in the monitoring process. Thus, we may regard the cumulative sample proportions or means at the fifth call as the provisional estimates, and one could determine whether to stop callbacks from the sixth call or not.

On the other hand, as shown in those tables, we should note that the cumulative sample proportion or mean does not much change by calls, especially earlier calls versus later calls. This implies that if call-based monitoring focuses on whether there are large net differences in cumulative sample estimates between calls, we may conclude that the MHSEH goes well since the net changes are small between calls. Also, this monitoring method, which was used in the MHSEH conducted in a large city, could be extensively applied to a national household survey involving many cities and counties.

Table 2. Distribution of Calls for 606 Completed Interviews (MHSEH)

	Calls						
	1	2	3	4	5	6	7+
%	37.0	23.3	15.5	12.5	8.3	2.6	0.9

Table 3. Cumulative Sample Proportion (%) for Gender by Calls (MHSEH)

	Calls						
	1	2	3	4	5	6	7+
Male	47.2	47.5	48.4	49.4	49.5	49.5	49.7
Female	52.8	52.5	51.6	50.6	50.5	50.5	50.3
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 4. Cumulative Sample Proportion (%) for Housing Type by Calls (MHSEH)

	Calls						
	1	2	3	4	5	6	7+
Detached house	9.8	7.1	7.2	7.6	7.3	7.3	7.4
Detached house (2+ HHs)	12.0	13.0	11.5	11.2	10.5	10.3	10.2
Multiplex house	37.8	39.8	38.8	37.0	37.3	37.4	37.4
Apartment	39.8	38.5	41.1	42.1	42.9	43.1	42.9
Other buildings	0.6	1.6	1.6	2.1	2.0	2.0	2.1
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Table 5. Cumulative Sample Proportion (%) for Diseases Treated in the Last 12 Months by Calls (MHSEH)

	Calls						
	1	2	3	4	5	6	7+
Asthma	0.92	0.98	1.08	1.21	1.12	1.09	1.08
Allergic Rhinitis	5.51	6.11	5.78	5.78	5.56	5.46	5.42
Allergic Conjunctivitis	4.09	3.44	3.18	2.82	2.61	2.54	2.52
Cardiovascular Disease	1.41	1.07	1.22	1.41	1.36	1.43	1.42
Atopic Dermatitis	1.57	1.31	1.47	1.51	1.56	1.59	1.58
Thyroid disease	0.90	0.69	0.72	0.61	0.89	0.95	0.95

Table 6. Cumulative Sample Mean for Years of Residence (MHSEH)

	Calls						
	1	2	3	4	5	6	7+
Average	10.7	10.3	9.9	9.5	9.4	9.4	9.4

5. Discussion

In this paper, we suggested some cumulative sample estimates, reflecting the live flow of data, as a macroscopic indicator for monitoring data quality. Generally converging to final estimates (before weighting), they can be easily presented with a graph or a table at any point in the survey data collection process and can be used at an early stage as provisional estimates of key survey variables. We illustrated how to use them in a national CATI survey and a local CAPI survey, given the sample design and data collection protocol in each survey. This approach based on cumulative sample estimates would help the researcher quickly and proactively check data quality in the early stage of data collection to ensure the survey is on track regarding key survey variables.

If survey researchers could view the cumulative sample estimates in real-time or regularly on their computer monitors, it would contribute to reducing uncertainty about resulting key survey statistics as well as the performance of a given survey design and would help them make faster decisions about maintaining or changing the data collection protocol or survey design. Therefore, the suggested cumulative sample estimates (mean or proportion) would be greatly beneficial to survey researchers monitoring data quality. In the case of well-planned, well-organized, and regularly conducted surveys, it will be much easier to make more informed judgments if experience in using cumulative sample estimates is accumulated.

Data Availability

The data for the CATI and CAPI surveys are available from an author at a given email address.

Software Information

The analyses were conducted using SAS 9.4 and R.

Funding

The authors have no support or funding to report. The SHPRC's own funds were used for this study.

References

American Association for Public Opinion Research. (2016). Standard definitions: Final dispositions of case codes and outcome rates for surveys (9th ed.). https://aapor.org/wp-content/uploads/2022/11/Standard-Definitions20169theditionfinal.pdf

Choi, E. H., Kim, S. W., Hong, S. J., Lee, S. Y., & Lee, H. N. (2013). Reducing survey nonresponse through enhanced administrative cooperation: an experience in Korea, Paper presented at *the Joint Statistical Meetings*, Montréal, Québec, Canada.

Cochran, W. G. (1977). *Sampling Techniques*, The Third Edition, New York: Wiley.

Couper, M. P., & Lyberg. L. (2005). The use of paradata in survey research. In *Proceedings of the 55th Session of the International Statistical Institute*.

Groves, R. M., & Heeringa, S. G. (2006). Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 169(3), 439–457.

Kim, S. W., & Couper, M. P. (2021). Feasibility and quality of a national RDD smartphone web survey: comparison with a cell phone CATI survey. *Social Science Computer Review*, 39, 1218-1236.

Lepkowski, J. M., Tucker, C., Brick, J. M., de Leeuw, E. D., Japec, L., Lavrakas, P. L., Link, M. W., & Sangster, R. L. (2008). *Advances in Telephone Survey Methodology*, New York: Wiley.

Wagner, J., West, B. T., Kirgis N., Lepkowski, J. M., Axinn, W. G., & Ndiaye, S. K. (2012). Use of paradata in a responsive design framework to manage a field data collection. *Journal of Official Statistics*, 28 (4), 477–499.

Woo, Y. J., Kim, S. W., Park, S. H., Lee, S. E., & Choi, B. Y. (2013). Using new IT for area sampling in a metropolitan household survey, Paper presented at *the Joint Statistical Meetings*, Montréal, Québec, Canada.

Author Biographies

Jaehoon Kim is a research associate at the Survey & Health Policy Research Center, Dongguk University. He is currently involved in several research projects of survey and data science methodology. E-mail: kimx4591@umn.edu

Sunwoong Kim is a professor at the Department of Statistics and the director of the Survey & Health Policy Research Center, Dongguk University. E-mail: sunwk@dongguk.edu