A METHODOLOGICAL APPROACH TO CONTROLLING INTER-MONITOR VARIABILITY IN COMPUTER-ASSISTED TELEPHONE INTERVIEWING OPERATIONS

Hansheng Xie[1] and Walter Mudryk

ABSTRACT

Quality Control (QC) monitoring has been implemented in Computer Assisted Telephone Interviewing operations of most social surveys at Statistics Canada. The QC monitoring has been demonstrated to be instrumental in lowering interviewer errors. However, the consistency and reliability of the monitors who evaluate and assess the interviewers remains a general concern of this technique. In this paper, two methods are proposed to objectively evaluate inter-monitor variability over time and determine monitors' consistency in assessing interviewers.

KEY WORDS: Computer assisted telephone interviewing; control chart; inter-monitor reliability; statistical process control.

RÉSUMÉ

Statistique Canada a mis en place la surveillance du Contrôle de la Qualité dans les opérations de l'Interview Téléphonique Assistée par Ordinateur de la plupart des enquêtes sociales. Cet exercice a contribué à diminuer les erreurs commises par les intervieweurs. Cependant, cette technique cause une certaine inquiétude quant à la cohérence et la fiabilité des surveillants qui évaluent les intervieweurs. Dans notre article, nous proposons deux méthodes pour évaluer objectivement la variabilité entre les surveillants au cours du temps ainsi que pour déterminer objectivement la cohérence des surveillants lors de leurs évaluations des intervieweurs.

MOTS CLÉS : Contrôle statistique du processus; fiabilité entre les surveillants; graphique de contrôle; interview téléphonique assistée par ordinateur.

1. INTRODUCTION

Computer Assisted Telephone Interviewing (CATI) systems are used widely in survey data collection. This automated system offers a good opportunity for monitoring interviewer performance during data collection and training interviewers using a Quality Control (QC) process. Statistics Canada has implemented CATI QC monitoring in all regional offices for most social surveys. The QC monitoring has demonstrated to be instrumental in lowering interviewer errors during data collection, which achieved better quality interviews for these surveys. However, a general concern of this technique arises from the consistency and reliability of the monitors' assessments. When the interviewer performance criteria are the same for all interviewers, which is the case for the social surveys at Statistics Canada, the primary cause of inconsistent evaluations and assessments could be the result of the variation among the monitor assessments. If the monitoring results are not consistent and reliable, the QC feedback on interviewer performance could be both meaningless and misleading. The monitoring process should ideally have its own control system.

It is the purpose of this paper to describe the QC methodology that has been proposed to evaluate inter-monitor variability over time, using specially designed *d control charts* as well as a key indicator of monitors' evaluations and assessments – the *inter-monitor reliability coefficient*, to determine their consistency in assessing interviewers within and across regional offices in

Hansheng Xie (hansheng.xie@statcan.ca), Statistics Canada, 11th Floor, R. H. Coats Building, Tunney's Pasture, Ottawa, Canada, K1A 0T6

Canada. This approach will quickly identify which QC monitors require additional training or upgrading of their monitoring skills and objectively evaluate how reliable monitor assessments are on interviewer performances. The information on inter-monitor variability will also be used extensively by management for purposes of feedback to individual monitor or groups. It enables them to better manage the monitoring process, thus resulting in a higher quality monitoring for CATI operations.

2. OVERVIEW OF QC MONITORING PROCEDURE

The overall quality objective for QC monitoring is to measure, control and improve the quality of the entire CATI operation in Statistics Canada on a continuous basis in terms of interviewer training, operational procedures, instrument design, survey response rates, interviewer performance and data processing.

Standardized training packages for interviewers and monitors have been developed to set training standards that not only reduce training time for multiple surveys but also establish a good standard base for QC monitoring. Interviewers are trained on various aspects of CATI procedures before they initiate telephone interviews. They are required to take courses to improve their interviewing skills and knowledge of subject matter concepts, and to ensure that they understand the CATI procedures well. Monitors also receive this training along with topics such as QC monitoring procedures, techniques of proper interviewing behaviour and QC feedback procedures so that all the monitoring should be conducted in an impartial and consistent manner.

The QC monitoring involves a quantitative approach to measuring quality that results in objective measures of interviewer performance. There are six major interviewing functions being monitored: Question Delivery, Subject Matter, Interviewing Environment, Respondent Relations, Data Processing and Procedures.

The sample unit is a complete computer ‘screen display’, that is, any piece of interviewing activity involving a respondent contains a question and expected response. This unit would allow for the selection of a fixed number of observations (screen displays) for each interviewer. Some CATI interviewing errors are more serious than others. Therefore, a seriousness classification system was established to assign a relative weight to each different type of error condition, such as Critical, Major, Minor and Other errors, according to its relative seriousness. The seriousness of the errors made is also taken into account when determining an interviewer’s or group of interviewers’ performance.

The required number of monitoring samples per interviewer per production period is based on a sampling plan. Monitors select random samples for each interviewer at regular intervals of time from the list of active interviewers. The monitor will observe 20 consecutive screens for each selected sample for that interviewer, and records the errors with detailed comments for positive feedback relating to the interview on a QC Monitoring Form. The monitor also provides immediate feedback to the interviewer and/or supervisor concerning any critical errors observed during or after the sample monitoring session.

The QC Monitoring Forms are then compiled and processed periodically using the CATI: QC Feedback System which generates the appropriate control charts, Pareto analysis and various operational summaries required for feedback to interviewers, supervisors and managers of the operation. This provides the timely feedback to each group and enables them to track and statistically analyze interviewer performance over time. Subsequently this process provides an overall framework for controlling the quality of the entire operation.

3. METHODOLOGY OF MEASURING AND CONTROLLING INTER-MONITOR VARIABILITY

3.1 Quality Measure

For the monitoring process, a quality measure should represent the monitoring ability of a monitor or group of monitors and be effective in measuring and expressing the desired quality characteristics of the process. Since an inaccurately weighted error could result from the monitor who allocated an incorrect weight to an error, the quality measure is defined as the *Average Allocation Errors per sample.* It is defined as the average number of the weighted errors per sample (across all weight classes) allocated by a monitor or group of monitors for a group of samples.

For each monitor or group of monitors within a certain production period (e.g., a week or month), the mathematical expression of *average allocation errors per sample* is defined as follows:

$$d_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \sum_{h=0}^{3} w_h c_{ijh} \qquad (1)$$

where $i = i_{th}$ monitor or group of monitors;
$j = j_{th}$ sample;
$h = h_{th}$ weight class;
w_h = weight for the h_{th} weight class (in our case, Critical error weight $w_3 = 4$, Major error weight $w_2 = 2$, Minor error weight $w_1 = 1$, Other error weight $w_0 = 0$);
c_{ijh} = number of errors in the h_{th} weight class in the j_{th} sample of the i_{th} monitor or group of monitors;
and n_i = number of samples assessed by the i_{th} monitor or group of monitors.

3.2 *d* Control Charts

A major objective of Statistical Process Control (SPC) is to monitor an ongoing process and detect quickly the occurrence of process shifts or distinctive patterns so that corrective action may be taken for quality improvement. As a result, the process can be kept in a state of statistical control for a long period of time. For this purpose the control chart is one of most effective tool and has been widely used in quality control.

In CATI operations, four different weight classes or demerits can be assigned to each error by a monitor according to his/her judgement. The *average allocation errors per sample* are the average demerit counts per sample that measure a monitor's assessment of the interviewer performance. Since the monitoring process is expected to be relatively stable, using the SPC *d* control chart is appropriate to control this monitoring process. The *d* control chart is applicable in a situation when the quality measure is an attributive count of demerits according to severity (see Montgomery, D. C. (1996)), which is the case in this monitoring process. It should be noted that each class of error is independent and the occurrence of errors in each class is well modeled by a Poisson distribution. Therefore, the *average allocation errors per sample* are a linear combination of independent Poisson random variables, where there are four classes of error with different demerit weights.

The *d* control chart can be constructed with control limits of Center Line (CL), Upper Control Limit (UCL) and Lower Control Limit (LCL) set at:

$$CL = \frac{\sum_{h=0}^{3} w_h \sum_{i=1}^{I} \sum_{j=1}^{n_i} c_{ijh}}{\sum_{i=1}^{I} n_i} \tag{2}$$

$$UCL = CL + \frac{3}{\sqrt{n_i}} \sqrt{\sum_{h=0}^{3} {w_h}^2 \lambda_h} \tag{3}$$

$$LCL = max(0, CL - \frac{3}{\sqrt{n_i}} \sqrt{\sum_{h=0}^{3} {w_h}^2 \lambda_h} \quad) \tag{4}$$

where $w_0 = 0$, $w_1 = 1$, $w_2 = 2$, $w_3 = 4$, I is the number of monitors and

$$\lambda_h = \frac{\sum_{i=1}^{I} n_i \sum_{j=1}^{n_i} c_{ijh}}{\sum_{i=1}^{I} n_i} . \tag{5}$$

The point d_i is then plotted on the *d control chart* by monitor (or by group of monitors) to determine if this monitoring process was operating in a state of statistical control. As long as the points plotted are within the control limits and don't behave in a systematic or non-random manner, the process is assumed to be in control and no action is necessary. Otherwise, there are out-of-control conditions. Investigation is required to identify the special causes responsible for these conditions and corrective action should be taken to eliminate the cause. Please note that control limits should be revised when special causes are found and eliminated (or implemented).

3.3 Inter-Monitor Reliability

Inter-monitor reliability (or inter-rater reliability) has been recognized as an important source of error in survey research and other disciplines. Several measures of monitor variance have been developed, among which the *inter-monitor reliability coefficient* is a widely used measure of reliability for the continuous variable case. This coefficient is based on the assumption of an analysis of variance (ANOVA) random effects model and it can be used to estimate monitor variance with some conditions required. Biemer and Forsman (2003) developed a general model for the practical use in survey data. In the following paragraphs, this model is introduced and adapted to our design.

Suppose that there are J interviewers and I monitors in a survey. Each interviewer is usually monitored by one monitor at a time although multiple monitors may observe the same interviewer at different points of time over the course of the survey. Assume that the assignment of interviewers to monitors is done randomly. In actual practice, this assumption is tenable if monitor assignments (i.e., monitoring sessions) are made without regard to interviewer performance. Suppose that there are n_j elements (i.e., total number of potential errors) to be assessed in the j^{th} interviewer's assignment. Let m_{jk} denote the recorded result of the assessment for element k in interviewer j's assignment for $k = 1,2, \ldots, n_j$ and $j = 1,2,\ldots,J$. Let m_{ijk} denote that assessment m_{jk} was determined by monitor i and assume that

$$m_{ijk} = \mu + \theta_j + \mu_{ijk} \tag{6}$$

where

$\mu = E\ (m_{ijk})$, the expected performance for all interviewers over all elements and monitors; $\theta_j = E\ (m_{ijk}) - \mu \sim (0, \sigma_\theta^2)$, the deviation from expected performance for the j^{th} interviewer; and μ_{ijk} is random error term associated with the monitor, the interaction of the monitor and the interviewer, and the element.

Further assuming that, $v_i \sim (0, \sigma_v^2)$, the allocated error that is common to all interviewers monitored by the i^{th} monitor; $(\theta\, \mathrm{v})_{ij} \sim (0, \sigma_{\theta v}^2)$, the error interaction of the i^{th} monitor and j^{th} interviewer; $e_{ijk} \sim (0, \sigma_e^2)$, a random error associated with the k^{th} element of the j^{th} interviewer, and all errors are independent, μ_{ijk} can be decomposed as

$$\mu_{ijk} = v_i + (\theta v)_{ij} + e_{ijk}\,. \tag{7}$$

In practice, the model for $\bar{m}_{ij} = \sum_k \frac{m_{ijk}}{n_{ij}}$, the mean of the $(ij)^{th}$ assignment, is particularly of interest since it satisfies the assumptions for mixed ANOVA models more closely than Model (6). From the above two formulas, this model can be obtained as

$$\bar{m}_{ij} = \mu + \theta_j + v_i + (\theta v)_{ij} + \bar{e}_{ij} \tag{8}$$

where $\bar{e}_{ij} = \sum_k \frac{e_{ijk}}{n_j} \sim (0, \sigma_e^2 / n_j)$.

The *inter-monitor reliability coefficient* is defined as

$$R_e = 1 - \frac{E_{j,k}[Var(m_{ijk} \mid j,k)]}{E_k[Var(m_{ijk} \mid k)]} \tag{9}$$

where $E_{j,k}$ is the expectation over interviewers and elements, and $Var(\cdot \mid j,k)$ is the conditional variance given the j^{th} interviewer and k^{th} element; E_k is the expectation over elements, and $Var(\cdot \mid k)$ is the conditional variance over interviewers and monitors given the k^{th} element.
In terms of model (6), this coefficient is given by

$$R_e = 1 - \frac{\sigma_v^2 + \sigma_{\theta v}^2}{\sigma_\theta^2 + \sigma_v^2 + \sigma_{\theta v}^2} = \frac{\sigma_\theta^2}{\sigma_\theta^2 + \sigma_v^2 + \sigma_{\theta v}^2}\,. \tag{10}$$

Thus, this coefficient is the ratio of the variance of interviewer performance to the total variance. A large value of R_e indicates that the monitor assessments are reliable due to a small amount of monitor variance in the data, while a small value of R_e indicates that the monitor assessments are unreliable.

In practice, these three estimated variance components are often obtained based on Model (8) using weighted least squares.

Because $\overline{m}_{ij}$ is actually the μ statistic for the μ control chart and our *d* statistic is a linear combination of several μ statistics, Model (8) is easily adapted to our case without further assumptions. In our case, each sample has the same number of elements to be assessed and thus a sample can be considered as an assessment unit. Model (8) can be rewritten as

$$d_{ij} = \mu + \theta_j + v_i + (\theta v)_{ij} + \bar{e}_{ij} \qquad (8')$$

where $d_{ij} = \frac{1}{n_{ij}} \sum_{j=1}^{n_{ij}} \sum_{h=0}^{3} w_h c_{ijh}$, is the *Average Allocation Errors per sample* associated with the i^{th} monitor and the j^{th} interviewer, $\bar{e}_{ij} = \sum_{k=1}^{n_{ij}} e_{ijk} \sim (0, \sigma_e^2 / n_{ij})$ is the average random error associated with the i^{th} monitor and the j^{th} interviewer, and n_{ij} is the number of samples associated with the i^{th} monitor and the j^{th} interviewer.

5. CONCLUSIONS

Due to the increasing awareness and concern for monitor errors, the assessment of interviewer performance has become an important part of survey quality work. In this paper, two statistical methods are proposed as a methodological approach to controlling the inter-monitor variability in the CATI operations at Statistics Canada.

The primary goal of the *d control chart* is to get the monitoring process in control by eliminating all of the special causes of variation among monitors in a timely manner. Using this chart enables us to quickly identify which monitors have poor monitoring skills and need retraining. Using results of SPC (control chart) analysis and making use of QC feedback, actions will then be taken to improve the required monitoring skills, which results in consistent assessments on interviewers. Hence, this control chart is not just for process surveillance, and it should also be used as an active and on-line method for reduction of process variability.

The *inter-monitor reliability coefficient* provides a summary measure of monitor variance for evaluating how reliable monitor assessments are on interviewer performances. Using this coefficient it can help us to determine if monitors in a group are consistent in their application of the interviewer assessments. For example, the low inter-monitor reliability could indicate that some monitors in the group have low monitoring skills and/or show inequity toward some interviewers, and the control chart records can be used to identify which monitors have these problems.

REFERENCES

Biemer, P. P. and Forsman, G. (2003). "Evaluator Error in the Assessment of Interviewer Performances," draft paper obtained from the US Bureau of Census.

Montgomery, D. C. (1996). *Introduction to Statistical Quality Control*, John Wiley & Sons, Inc., New York.