USING CONTROL CHARTS IN SURVEYS

Weimin Sun
Business Survey Methods Division, Statistics Canada, 3-D R. H. Coats Bldg., Ottawa, ON, K1A 0T6

KEY WORDS: CATI, Control Chart, Coverage Probability, Detection Power.

Abstract: Mudryk, Burgess and Xiao (1996) propose to use statistical process control charts for computer aided telephone interview surveys at Statistics Canada. In this paper, we discuss two alternative charts to be used in the same application. By borrowing the concepts from hypothesis testing, we introduce criteria of coverage probability and detection power to compare the performance of these two charts. Both heuristic argument and empirical results show the equivalence of these two charts. Therefore, a choice is made based on relization convenience. In addition, we discuss how to adopt the ANOVA approach to testing the stability of an survey operation. The idea is tested with the weekly keying error rates observed during the twenty-one week data capture operation for the 1991 Census of Agriculture in Canada.

1. Introduction

Statistics Canada, a leading official statistical agency in Canada, conducts numerous surveys and publishes the results. To publish high quality data, the agency adopts an extensive quality assurance program. Included in the program are statistical quality control and quality checking. Statistical quality control (QC) means quality improvement with the aid of statistical theory. A hundred percent sampling, acceptance sampling with rectifying inspection (Hald 1981) and skip lot sampling with rectifying inspection (Schilling 1982), in the descending order of sampling intensity, are the most popular statistical QC methods at Statistics Canada. A less commonly used method is the statistical process control. Quality checking refers to a collection of ad hoc quality assurance methods.

With respect to the acceptance sampling methods, statistical process control requires fewer sampled units, resulting in less QC cost. Furthermore, there are some applications for which implementation of acceptance sampling plans is difficult and control charts are preferable. A computer aided telephone interview (CATI) survey provides such an example. Mudryk, Burgess and Xiao (1996, hereafter as MBX) use process control chart to enhance the quality of CATI surveys.

MBX's process control chart is not a statistical chart, since they use a user-specified upper control limit (UCL) line. In this paper, we propose two alternative statistical charts. One chart is based on the Poisson distribution and the other on the Bernoulli. In either chart, the UCL is three standard deviations above the centre line. Our heuristic argument and simulation show that there is no significant difference between these two charts. Therefore, we suggest selecting the Bernoulli-based chart because the chart is easier to implement. The need to estimate the standard deviations and centre lines of our charts with historical data leads us to the discussion of how to identify the breaking points.

Throughout the paper, we assume serially uncorrelated data. Control charts for temporally dependent survey observations are discussed, for example, by Spisak (1995). This paper is structured as follows. Section 2 introduces the two control charts. Simulation results regarding the performance of these two charts are presented in the same section. In section 3, we discuss the stability of a survey operation. The final section is devoted to concluding remarks.

2. Two Alternatives to MBX's Control Charts

CATI is an integrated data collection and capture system. It allows interviewing, editing and data capture to be carried out on a unified system, thereby reducing survey operation steps and costs. During a CATI survey, a QC technician randomly monitors part of a live conversation between an interviewer and respondent. The technician records errors made by the interviewer during the

monitoring period. To measure the quality of telephone interview quantitatively, MBX propose a demerits rate, which is a weighted sum of error counts. More specifically, let C_{ijlh} $i = 1, ..., t, j = 1, ..., n, l = 1, ..., m$, be the number of class h errors committed by the j^{th} interviewer, during the l^{th} monitoring session, at time i. The demerits rate is

$$C^{*}_{ijl} = \sum_{h=1}^{3} W_h C_{ijlh},$$

where W_h is the weight assigned to error class h. MBX define three classes of error severity: minor, major and critical.

For a fixed interviewer j, MBX plot C^{*}_{ijl} on a weekly control chart over l and i. They set a lower limit line zero and let UCL be specified by the user. An example of the user-specified UCL is the weight assigned to the class of critical error, if the user's objective is to control critical errors. The centre line equals the previous week's grand mean of demerits rates over n interviewers and m monitoring sessions.

To simplify the presentation below, we only discuss the case of one error severity class and replace the notation C^{*}_{ijl} with P_{ijl}. The extension to more than one class is straightforward. We, also, assume P_{ijl} follows a Poisson distribution with mean λ, and is independent over index j and l. Define

$$B_{ijl} = \varphi(P_{ijl}) = \begin{cases} 1 & \text{if } P_{ijl} > 0 \\ 0 & \text{if } P_{ijl} = 0, \end{cases}$$

and $\bar{P}_{ij} = \frac{1}{m}\sum_{l=1}^{m} P_{ijl}$, $\bar{B}_{ij} = \frac{1}{m}\sum_{l=1}^{m} B_{ijl}$. Two alternatives to the MBX chart are to plot $\bar{P}_{ij}$ and $\bar{B}_{ij}$ on a control chart respectively. Call the first chart P-chart and the second B-chart.

In both charts, we take LCL = 0, since a low error rate is good in a survey operation. The centre line is the grand mean of $\bar{P}_{ij}$ (or $\bar{B}_{ij}$) over i and j. Note that when m is large enough, both $\bar{B}_{ij}$ and $\bar{P}_{ij}$ are close to normal. UCL is, then, taken to be three standard deviations above the centre line. Both the grand mean and standard deviation are estimated with historical data, which will be discussed further in the next section.

Since there are two alternative control charts available, selecting one becomes the focus of the section. To this end, we need to compare the performance of two statistical control charts. It is unclear, however, how to compare statistical control charts directly. The link between a control chart and a sequence of hypothesis tests provides a solution to this comparison problem, as the tools from the hypothesis testing theory can be borrowed for the context of statistical process control chart.

To explain the link, let us look at the case of a standard statistical control chart, where the upper/lower limit line is three standard deviations above/below the centre line. At a fixed time, when a sample point is plotted on the chart, it is equivalent to testing whether the sample point falls inside a three standard deviation confident interval (CI). If the sample point follows a normal distribution, which usually is the implicit assumption for a standard statistical control chart, the confidence level of this CI is almost 100%. The connection between a CI and a hypothesis test is a well-known fact. Thus, a control chart corresponds to a sequence of tests.

Suppose a survey operation is at a stable status with λ_0 being the commonly targeted mean error count for every interviewer. The P-chart of an interviewer j is approximately identical to a sequence of tests for the hypothesis H_0: $\lambda=\lambda_0$ vs H_1: $\lambda>\lambda_0$. The approximation (instead of exactness) is due to the fact that λ_0 is estimated with historical data. Similarly, the B-chart identifies to tests for H_0: $p=p_0$ vs H_1: $p>p_0$, where $\lambda_0 = -\log_e(1-p_0)$ as $P(B_{ijl} = 1) = 1 - P(P_{ijl} = 0)$. Because of the monotonic relationship between λ_0 and p_0, the above two hypotheses are equivalent. Based on these arguments, we can treat the two control charts as two test statistics designed to test for the same hypothesis. Thus, we have justified borrowing the tools of comparing two test statistics in the hypothesis test to the context of control charts.

According to the hypothesis testing theory, a good test statistic should assume a small type I error and a big power. A small type I error means a high value of one minus type I error. By translating these terms in hypothesis test into the language of control chart, we introduce coverage probability (CP) and detection power (DP). Intuitively, when a CATI survey process is in control, a good control chart should, with a high probability, cover a sample point, i.e. the sample point should fall under the UCL. On the other hand, when the process is out of control, the chart should, with a high probability, detect the anomaly by letting the sample point lie above UCL. We call the first probability the CP of a control chart and the second the DP of a control chart. A control chart with a low CP sends out excessive number of false alarms and a chart with a low DP creates a false sense of security. The counterpart of CP in test theory is one minus type I error and DP the power of a test statistic. Formally, CP refers to $P(\bar{P}_{ij} \le UCL \mid \lambda_0)$ and DP $P(\bar{P}_{ij} > UCL \mid \lambda)$, where $\lambda > \lambda_0$.

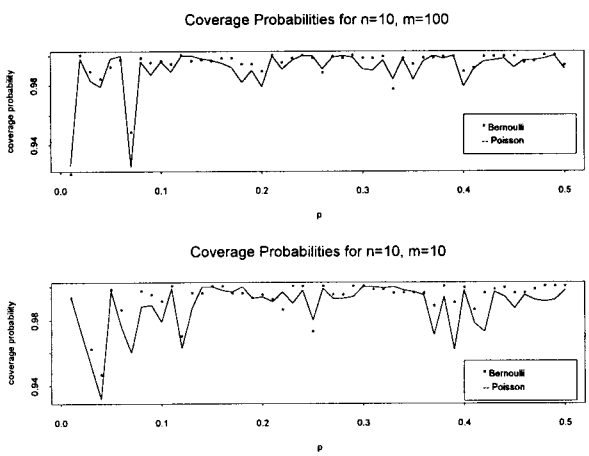

Figure 1: Coverage probabilities of $\bar{B}_{ij}$ and $\bar{P}_{ij}$ charts.

Figure 1 displays the simulated CP's of the *P*-chart and the *B*-chart, where p_0 changes from 0.01 to 0.5, or equivalently λ_0 from 0.1 to 0.69. The CP of the *B*-chart corresponding to a fixed p_0 is computed in the following way. An estimate of p_0 is calculated as the mean of *nm* randomly generated numbers from *Bernoulli*(p_0). UCL is taken to be $3*\sqrt{\hat{p}_0*(1-\hat{p}_0)/m}$. A large number of independently observed $B_{ij|}$ are generated from the same distribution and CP is the proportion of $\bar{B}_{ij}$ lying below UCL. CP's of the *P*-chart is computed similarly, but with random numbers generated from $Poisson(-\log_e(1-p_0))$.

Figure 2 gives detection powers. We compute the simulated DP's of the *B*-chart as follows. p_0 is taken to be 0.1 and the drifted p equals to $(1+k\%)*p_0$, with k ranging from 0 to 500. The UCL is $3\sqrt{p_0(1-p_0)/m}$. Random observations $B_{ij|}$ are then generated from *Bernoulli(p)* and DP is the proportion of $\bar{B}_{ij}$ lying above the UCL. For the *P*-chart, the calculation is similar.

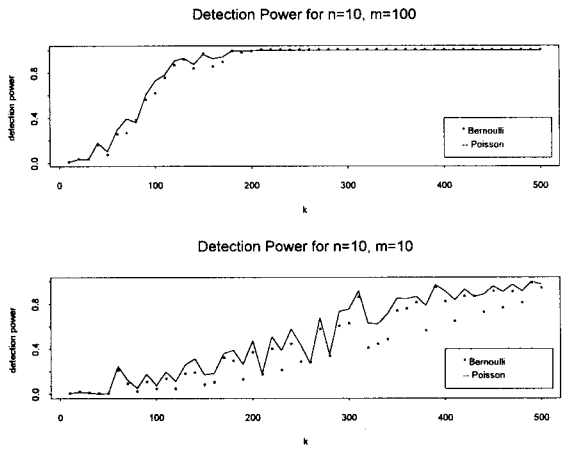

Figure 2: Detection power of $\bar{B}_{ij}$ and $\bar{P}_{ij}$ charts.

We have heuristically showed that the *P*- and *B*-charts are equivalent. The simulated CP's and DP's in Figure 1 and 2 display no significant difference, particularly when *m* is large. Thus, heuristically the two charts are not different: either one serves the purpose of controlling a survey operation (e.g. a CATI survey) equally well.

In terms of implementation, the *B*-chart enjoys certain advantages over its rival, since it requires collecting less information. More specifically, in an actual CATI survey operation a QC technician only needs to decide whether an interviewer commits an error or not during a monitoring session. This reduces the work load of the technician. A QC technician commonly works in a demanding environment while monitoring a live conversation. The reduction on the workload of data collection decreases their stress level. The *B*-chart helps the technician make less measurement errors as well. Judging whether an interviewer makes a mistake during a monitoring session is much more likely error-free

than counting the exact number of errors.

3. Stability of an Operation

The above section showed how to construct two alternative charts and concluded implementational superiority of *B*-chart over *P*-chart. Yet, the discussion is incomplete, since for either chart the UCL and center line are unknown. Historical data will be used to estimate these lines. The question of using what historical data for the estimation motivates us to discuss the stability of an operation.

In addition, it is worth pointing out that the discussion in the previous section centred on the level of an individual interviewer's error count. The following discussion will be carried at the level of all interviewers' error counts.

For either *P*-chart or *B*-chart, the sampling distribution of an interviewer's error count at a fixed time is uniquely determined by its parameter. Suppose the commonly targeted mean of each interviewer's error count is λ_0 and H_0: $p_0 = 1 - e^{-\lambda_0}$. When a CATI operation is stable, p_0 or λ_0 remains constant. An operation may have several stable periods separated by breaking points, where a breaking point refers to a point in time when p_0 or λ_0 changes. Thus, testing whether an operation is stable is equivalent to testing whether p_0 or λ_0 is constant.

To define a *stable operation* formally, let X_{ij} be some error measurement of j^{th} interviewers at time i. For instance, X_{ij} can be $\bar{P}_{ij}$ or $\bar{B}_{ij}$. Suppose X_{ij} follows a sampling distribution with density function $f(x_{ij}|\theta_i)$. Stating that the operation is stable from time $i+1$ to $i+k$, in a strict sense, we mean that θ_i does not change; that is, $\theta_{i+1} = \theta_{i+2} = \cdots = \theta_{i+k}$. A wide sense definition will be $\|\theta_{i+r} - \theta_{i+s}\| < c_{rs}$ for some pre-determined constants c_{rs}, $r, s = 1, \ldots, k$. Here, k serves as the moving window size. In this paper, we only use the strict definition.

When m is sufficiently large, $\bar{B}_{ij}$ and $\bar{P}_{ij}$ are close to normal. In this case, $f(x_{ij}|\theta_i)$ is normal with $\theta_i = (\mu_i, \sigma_i)$, where μ_i is the mean of X_{ij} and σ_i the standard deviation. The strict definition implies to show $\mu_{i+1} = \mu_{i+2} = \cdots = \mu_{i+k}$ and $\sigma_{i+1} = \sigma_{i+2} = \cdots = \sigma_{i+k}$. Bartlett's test (Montgomery 1991) is a widely used method for checking homogeneity. This test, however, is sensitive to non-normality. Another test which is less known, but less sensitive to normality assumption is proposed by Burr and Foster (refer to Anderson *et al* 1974). To test for $H_0: \mu_{i+1} = \mu_{i+2} = \cdots = \mu_{i+k}$, we can use one-way ANOVA.

Since appropriate CATI survey data are not available, we test the above idea with weekly keying error rates. The richness of this data set provides a proper test ground to our method. The error rates were observed during the twenty-one week data capture operation for the 1991 Census of Agriculture in Canada. During the twenty-one week operation, 131 keyers captured information from about 280,000 questionnaires (Duddek 1996). After data-capture, a portion of questionnaires were sampled and keying errors were identified by verifiers. Here, an error means a discrepancy in entry between a keyer and a QC verifier. The weekly error rate is the average number of errors per hundred questionnaires (phu) committed by a keyer within a week.

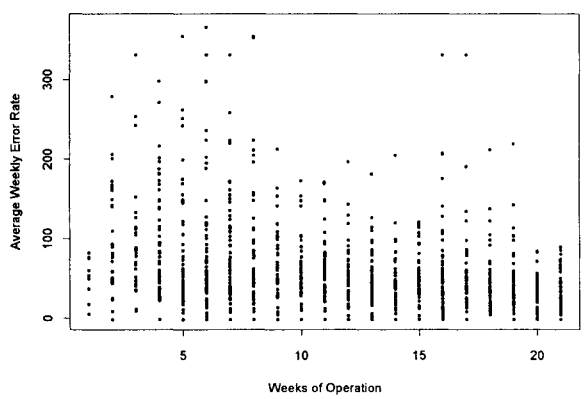

Figure 3: Dot plot of keyers' weekly error rates in *phu* from week 1 to 21.

Figure 3 is the dot plot of error rates. Each dot represents a keyer's weekly error rate. Note that not all 131 keyers worked in every week. For example, there were only 10 keyers worked in week 1 and 31 keyers in week 2. The plot shows a general downward trend in variability and mean

of error rates.

Initial data analysis uncovers no serial correlation and histograms at each fixed time show the weekly error rate distributions are highly skewed to the right. A log(data+a) transformation is necessary to make the data normal. The value a should be big enough so that the logarithm of zero plus a does not blow up the sample variance. Figures 4 and 5 show the weekly sample variances and means at the original scale and the log-transformed scale (a=100). The plots show that this drifted logarithm transformation keeps temporal trends of the raw data's sample variances and means.

When the moving window size, k, equals 2, Bartlett's test shows that there is significant change of variability between week 1 and 2, 8 and 9, 15 and 16, and 19 and 20. When the size is increased to 3, same test reveals a similar story. We then conclude that week 2, 9, 16 and 20 are breaking points for the operation. Applying F-test of ANOVA to the sample mean series for k=*2* and *3*, we identify the breaking points at week 2 and 20. Table 1 gives the period of weeks, where either Bartlett's test or F test is significant. By combining these two tests, we conclude that the operation have four breaking points: week 2, 9, 16 and 20.

Table 1: period of weeks where either a Bartlett's or F test is significant

	Bartlett's Test, Significant level=0.1
k=2	(1,2) (8, 9) (15, 16) (19, 20)
k=3	(1, 2, 3) (8, 9, 10) (14, 15, 16) (15, 16, 17) (18, 19, 20)
	ANOVA Test, Significant level=0.1
k=2	(1, 2) (19, 20)
k=3	(1, 2, 3) (18, 19, 20) (19, 20, 21)

As long as the breaking points of an operation are identified, we should use historical data from the most recent breaking point up to the previous time point to estimate p_0 or λ_0. Basing on p_0 or λ_0, we can easily find the centre line and UCL for the control chart of an individual interviewer. For example, suppose the current time is week 14. If we want to create a control chart for a keyer, only historical data between week 9 and 13 should be used to estimate the UCL or centre line.

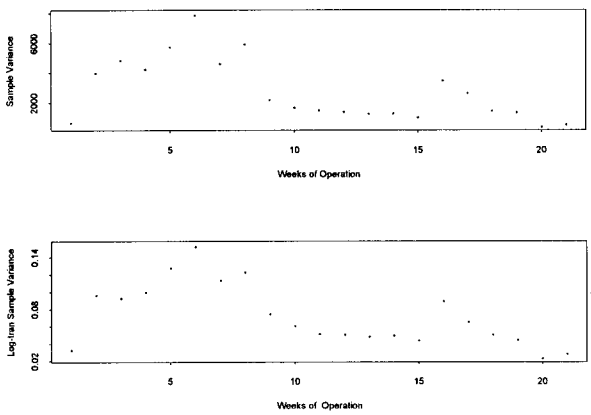

Figure 4: Sample variances of original and log-transformed weekly error rates from week 1 to 21.

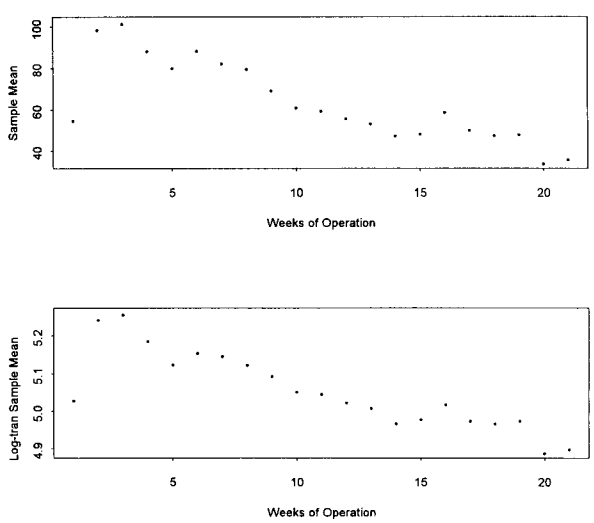

Figure 5: Sample means of original and log-transformed weekly error rates from week 1 to 21.

4. Concluding Remarks

P_{ijl} contains more information than B_{ijl}, since B_{ijl} only indicates whether an interviewer commits at least one error where P_{ijl} actually tells us how many errors the interviewer commits during monitoring session l. Yet, this additional information does not help P-chart performing better.

In this paper, no comparison, in terms of CP and DP, between MBX's chart and its two alternatives is made. There are two reasons for that. First, the MBX chart's UCL is specified by a user. Statistical characteristics like CP and DP cannot be computed with such a chart. Second, the main purpose of MBX's chart is to control severe errors, while the two alternatives control the average number of errors committed by an interviewer. The two alternatives enjoy an advantage over MBX's chart. In the case of having several error severity classes, if the error counts of an interviewer among classes are independent, then the extension of our charts to the d-chart is straightforward. This is because the average, over monitoring sessions, of demerits rates is still normal. This is not true for the MBX's chart.

Although the above discussion on two alternative charts takes the context of a CATI survey, it can be easily extended to other surveys. The idea of mapping control charts to a sequence of tests, and then comparing them with the tools of hypothesis test theory can be valuable to the general comparison of several statistical process control charts. The discussion of testing stability of an operation could be useful for other QC sampling plans as well. For example, when an 100% sampling plan is replaced with a QC plan at a less intensively sampling level, *e.g.* an acceptance sampling plan, the general assumption is that the operation has become more stable. Finding when this shift-towards-being-more-stable occurs helps us to timely switch an intensively sampled QC plan to a less intensive one.

References

Anderson, V.L., and McLean, R. A. (1974). *Design of Experiments: A Realistic Approach*. New York: Marcel Dekker, Inc.

Duddek, C. (1996). "Modelling Quality Control Strategies for the 1996 Census of Agriculture." *Proceedings of the Section on Government Statistics, American Statistical Association*.

Hald, A. (1981). *Statistical Theory of Sampling Inspection by Attribution*. London: Academic Press.

Montgomery, D.C. (1991). *Design and Analysis of Experiments*. 3rd Ed. New York: John Wiley & Sons.

Mudryk, W., Burgess, M.J., and Xiao, P. (1996). "Quality Control of CATI Operations in Statistics Canada." *Proceedings of the Section on Survey Research Methods, American Statistical Association*.

Schilling, E. (1982). *Acceptance Sampling in Quality Control*. New York: Marcel Dekker.

Spisak, A. W. (1995). "Statistical Process Control of Sampling Frames." *Survey Methodology*, Vol. 21, No.2, p185-190