Journal of Official Statistics, Vol. 28, No. 2, 2012, pp. 221–237

Response Rates in Business Surveys: Going Beyond the Usual Performance Measure

Katherine Jenny Thompson[1] *and Broderick E. Oliver*[1]

Many ongoing programs compute response rates for usage both as performance measures and as quality indicators. There is extensive literature on the computation and analysis of response rates for demographic surveys, which are generally characterized by multi-stage designs with heterogeneous populations within selected clusters. In contrast, business surveys are characterized by single-stage designs with highly skewed populations. Consequently, business surveys in the Economic Directorate of the U.S. Census Bureau compute two "flavors" of response rates: the unit response rate (URR), defined as the rate of the total unweighted number of "responding" units to the total number of sampled units eligible for tabulation; and a total quantity response rate (TQRR), which is the weighted proportion of a key estimate reported by responding units or obtained from equivalent quality sources (Lineback and Thompson 2010). Thus, for each statistical period, a survey produces one unit response rate and several total quantity response rates – one per key item. In this article, we describe how these two rates are computed, then introduce a statistical process control analysis perspective for monitoring them. We illustrate this approach with examples from ongoing economic programs conducted by the U.S. Census Bureau.

Key words: Performance measure; quality indicator; *p*-chart; general linear hypothesis test.

1. Introduction

Business surveys in the Economic Directorate of the U.S. Census Bureau compute two "flavors" of response rates: the unit response rate (URR), defined as the rate of the total unweighted number of "responding" units to the total number of sampled units eligible for tabulation; and a total quantity response rate (TQRR), which is the weighted proportion of key estimates reported by responding units or obtained from equivalent quality sources (Lineback and Thompson 2010). Thus, for each statistical period, a survey produces one URR and several TQRRs – one per key item. Beginning in 2006, the Economic Directorate conducted a large scale project to determine standard formulae for the computation of all response rates produced by economic programs. By fall 2007, these formulae were implemented in the U.S. Census Bureau Standard Economic Processing System (StEPS) (see Ahmed and Tasky 2000 for more details on StEPS). The release of these metrics was

[1] Office of Statistical Methods and Research for Economic Programs, U.S. Census Bureau, Washington, DC 20233, U.S.A. Emails: katherine.j.thompson@census.gov and broderick.e.oliver@census.gov
Acknowledgments: We thank Rita Petroni, Xijian Liu, the Associate Editor, and three referees for their useful review of earlier versions of this manuscript and acknowledge the contributions of William Davie, Jr., Miriam Rosenthal, and Justin Z. Smith in the cited nonresponse bias studies. This report is released to inform interested parties of ongoing research and to encourage discussion. Any views expressed on statistical or methodological issues are those of the authors and not necessarily those of the U.S. Census Bureau.

accompanied by extensive classroom training on concepts and usage, and implementation was performed in partnership with program managers, statisticians, and programmers.

Response rates are performance measures and are compared against benchmarked targets. These target goals (e.g., a required 80% response rate) are not necessarily useful indicators of data quality, of the stability of the survey process being evaluated (e.g., data collection, data review), or of the potential level of nonresponse bias (Peytcheva and Groves 2009). However, since many of our economic surveys and censuses already compute response rates, it would benefit these programs if they evaluate these response rates within the statistical process control framework. Current procedures typically compare the most recent rates to those obtained in a prior period. With such limited comparisons, the program managers are unable to detect upward or downward trends in the response process. Simply put, the focus on short term comparison does not provide an adequate framework for analyzing the response process. Instead, we propose using control charts to monitor and study program-level URR and TQRR processes. Examining response rates over time with control charts is useful because they help survey managers distinguish between variation that is inherent in the process and expected and variation that is unusual and in many cases unexpected. Moreover, control charts are useful because they show what a process is capable of – for example, providing evidence that a given program's current procedures yield a stable process that may or may not hover around a target value. In the latter case, some form of modification of the process would be required to meet this target.

Section 2 describes the "unique" characteristics of economic data, introducing some of the challenges of obtaining useful performance metrics. Section 3 provides the formulae for computation of two response rates used in the Economic Directorate at the U.S. Census Bureau. Section 4 provides background on our two case studies. In Section 5, we discuss retrospective analysis procedures used to analyze response rates in the Economic Directorate, which motivates the proposed control chart framework also introduced in Section 5. We conclude in Section 6 by describing our ongoing plans for implementation and providing a few ideas for future research.

2. Characteristics of Economic Data

Economic data generally have very different characteristics from their household counterparts. First, business populations are highly skewed, with a large proportion of the estimated totals originating from a small set of cases. Consequently, the majority of economic programs administered at the U.S. Census Bureau utilize stratified designs that: (1) include "large" cases with certainty; (2) may sample "medium sized" cases with high sampling rates; and (3) sample the remaining cases with very low sampling rates (usually less than 0.01). As a result, sampled cases with large design weights often contribute very little to the overall tabulated totals.

To avoid overrepresentation of such small cases in computation, the URR is computed without using design weights. Doing this, however, tends to downplay the importance of the certainty or large noncertainty cases, making URR an inconsistent indicator of data quality. Consider the fictional business survey example presented below in Figures 1 and 2. The sample consists of 30 units, 26 of which are noncertainty units and account for

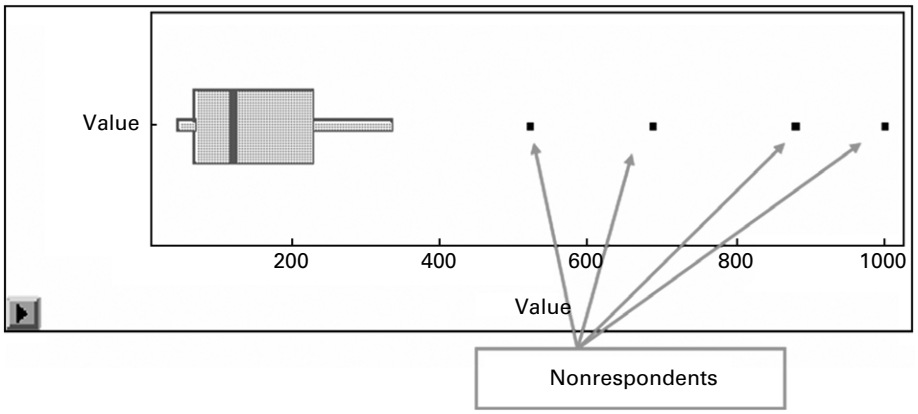

Fig. 1. Fictional business sample with unit nonresponse (Scenario 1)

53 percent of the total estimate (value). The remaining four units, which are selected with certainty, account for 47 percent of the total estimate.

In the scenario presented in Figure 1, this fictional survey has four nonrespondents: the four certainty cases. The URR is approximately 87% (26/30), but the TQRR value is approximately 53%.

The situation changes considerably in the scenario presented in Figure 2. Here, the survey also has four unit nonrespondents, and the URR is still 87%. However, these small nonresponding cases contribute very little to the total tabulation so that the TQRR value is approximately 93%. These examples demonstrate how the more consistent measures of data quality computed from a skewed population include a measure of size (e.g., payroll, capital expenditures) to account for the unit's relative importance in the estimates (Tucker et al. 2007), as is done with the TQRR.

A second concept in business surveys is the distinction between the "reporting unit" and the "tabulation unit." A reporting unit is one that has been established for the purpose of collecting survey data. A tabulation unit is one used for estimation. For example,

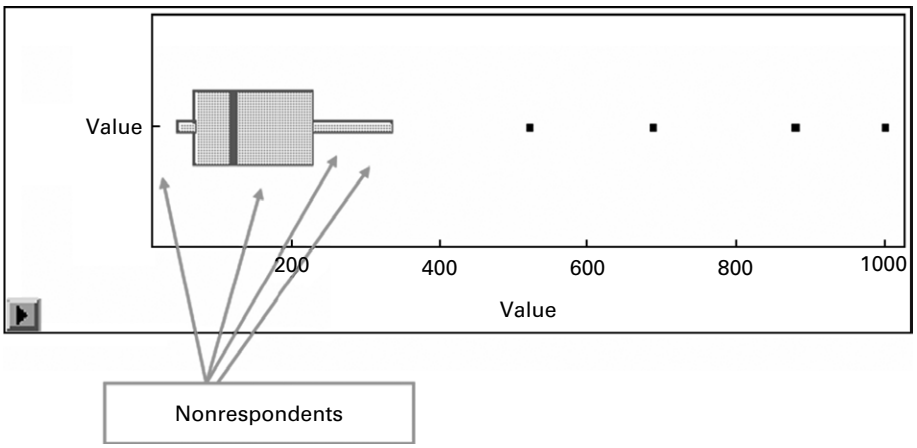

Fig. 2. Fictional business sample with unit nonresponse (Scenario 2)

a company would be assigned to at most one industry on the sampling frame, but may operate in several different industries. To deal with this, the data received from the reporting unit (the company) may be split into "tabulation units." In other cases, a program may consolidate establishment or plant level data to the company level to create a tabulation unit. For business surveys, URRs are based on the disposition of the reporting unit and TQRRs are based on the tabulation units. (It is possible for the reporting unit and the tabulation unit to be the same.)

Finally, often "equivalent quality" auxiliary data are available for some key items collected by business programs. For example, the U.S. Census Bureau conducts an economic census every five years and maintains an "up-to-date" business register of administrative data. Frame variables may be timely and highly correlated with survey characteristics of interest. Moreover, in contrast to household surveys, in some cases it is possible to obtain a valid value of a characteristic from an alternative source: for example, a published company report might contain quarterly sales figures that could be effectively substituted for missing response data.

3. Computation of Response Rates (Unit and Total Quantity)

In the Economic Directorate of the U.S. Census Bureau, the URR is computed as

$$[R/(E + U)]^* 100 \tag{3.1}$$

where

- R is the number of *responding* reporting units that were eligible for data collection in the statistical period;
- E is the number of reporting units eligible for data collection in the statistical period (including chronic refusals); and
- U is the number of reporting units for which eligibility for data collection in the statistical period could not be determined.

This is a conservative calculation, since an unknown percentage of the "U" cases will be out-of-scope for the program. The URR denominator excludes the cases from which there was no attempt to collect data because of planned "imputation" from auxiliary data. URRs are reported without standard errors as performance measures.

The URR is a measure of combined controllable and uncontrollable processes. The program managers have control over the contact component of the response process in the sense that they can direct personnel to conduct follow-up interviews with nonrespondents. However, the U.S. Census Bureau standards define a respondent as an eligible reporting unit for which: (1) an attempt was made to collect data; (2) the unit belongs to the target population; (3) and *the unit provided sufficient data to be classified as a response*. To satisfy the latter requirement, each program determines which collected data items are required *in advance* of data collection: response status for each unit is determined after all data processing – including analyst review and editing – is completed. As the data quality restrictions on required data items for the program increase, the greater the likelihood that the URR will decrease because it is more difficult for reporting units returning the questionnaire to qualify as a respondent. Thus, the additional restrictions on the quality of

the data received add a random (uncontrollable) element. Consequently, it is possible to offer specific protocols designed to improve the amount of contact with nonrespondents that do not improve the unit response rates.

In-house procedures for analyst review and follow-up of survey data are designed to improve the quality of the estimates. Analysts strive to reduce imputation rates for all key items. This is usually best accomplished by unit nonresponse follow-up of the large cases expected to contribute substantially to the estimate, followed by intensive analyst research for "large impute" cases comprised of more phone calls (targeted questions) and searches for auxiliary data sources (e.g., financial reports) to replace imputed values with equivalent data. The cumulative impact of these effects on the data quality input to tabulations is assessed via the TQRR.

In the Economic Directorate of the U.S. Census Bureau, the TQRR for item y is computed as

$$[(R_y + A_y)/Y]*100 \tag{3.2}$$

where

- R_y is the weighted estimate obtained by summing reported data for item y in the statistical period;
- A_y is the weighted estimate obtained by summing equivalent source data (auxiliary data from the same unit) for item y in the statistical period;
- Y is the estimated (weighted) total of the item in the statistical period. It includes all data used to develop the publication estimate, including imputed data and (nonmailed) auxiliary data imputation cases.

Both numerator and denominator cases are weighted by "unbiased" weights, which include subsampling and outlier adjustment factors. In addition, denominator weights include unit nonresponse adjustment factors if the program uses weight adjustment instead of imputation to account for unit nonresponse.

When reported with survey totals as performance measures, TQRRs provide exact measures at that point in time of the proportion of the estimate that was not imputed. In this context, it would be inappropriate to report the measure along with confidence limits. However, the analysis of the TQRR *process* must take into account that the TQRR is a ratio of correlated random variables and that the point estimates in the time series are not independent in a cross-sectional or longitudinal survey. Furthermore, the TQRR variance and covariance estimates used in analysis must account for the complex survey design, if applicable. We discuss this further in Section 5.1.

The TQRR measures controllable processes, so that increasing an item's TQRR generally leads to improved data quality for the estimate. For programs that publish information on one or two items, the TQRRs for each item are clearly superior performance measures over the URR. However, if the survey publishes several key items, it may be unwise to measure *performance* by setting target TQRR values for all published items, since meeting all target values may be difficult or even impossible depending on the number of collected items and the processing cycle time allotment. It can be equally challenging to monitor the levels of several different items' TQRRs over time. And, of course, with a large number of items, it is difficult to develop a simple course of corrective action.

Throughout the remainder of the article, we use "real-life" examples culled from two economic programs conducted by the U.S. Census Bureau: the Annual Capital Expenditures Survey (ACES) and the Quarterly Services Survey (QSS). Each example is extracted from a larger cited nonresponse bias analysis study. Both studies were mandated internally and the statistical periods for both were limited to validated data stored in the StEPS database. The retrospective analysis examples provided in Section 5.1 are presented in Smith and Thompson (2009); the control chart examples provided in Section 5.2 use statistics presented in Rosenthal and Davie (2008). We use the ACES example for our retrospective analysis in Section 5.1 for presentation reasons: the ACES methodologists restricted their analysis to three subpopulations, making it easy to fully illustrate the presented techniques. We use the QSS data to illustrate the statistical process control procedures in Section 5.2 for similar reasons: during the time period considered, the QSS published one key item, so the example is self-contained.

4. Case Studies

4.1. The Annual Capital Expenditures Survey (ACES)

The ACES examples use data collected from the 2003 through 2008 survey years[2]. The ACES is an annual survey of companies that collects data about the nature and level of capital expenditures by nonfarm businesses operating within the United States. Respondents report capital expenditures for the calendar year in all subsidiaries and divisions for all operations within the United States. ACES respondents report total capital expenditures, broken down by type (expenditures on Structures and expenditures on Equipment).

The ACES universe contains two subpopulations: employer companies and nonemployer companies. Different forms are mailed to sample units depending on whether they are employer (ACE-1) companies or nonemployer (ACE-2) companies. New ACE-1 and ACE-2 samples are selected each year, so that ACES estimates are based on independent samples. The ACE-1 sample comprises approximately 75 percent of the total ACES sample.

The ACES has a stratified simple random sample design. The ACE-1 frame is developed from administrative payroll data. The ACE-1 survey strata are defined by five company size class categories – each based on payroll – within industry: one certainty stratum per industry, and four noncertainty strata. The majority of the capital expenditures estimate in a given industry is usually obtained from the certainty and large noncertainty strata; reported zero values for capital expenditures are quite frequent with units from other strata. The ACE-2 sampling frame is comprised of businesses without paid employees or payroll, sole proprietors, and companies for which no administrative data have been received. From the ACE-2 frame, four substrata are formed based on legal form of organization and available administrative data.

Annual estimates of totals are Horvitz-Thompson estimates, with standard errors computed using the delete-a-group jackknife method. The ACES uses weight adjustment

[2] Smith and Thompson (2009) present similar ACES results from the 2002 through 2006 survey years in the context of a nonresponse bias analysis study.

to account for unit nonresponse, using "adjustment-to-sample" procedures (Kalton and Flores-Cervantes 2003). To do this, sampling weights for each unit (computed as the inverse probability of selection) are multiplied by a weighting-cell specific adjustment factor that is based on data known for both respondents and nonrespondents. The ACE-1 component employs a ratio adjustment procedure to account for unit nonresponse, using administrative payroll values obtained from the sampling frame. The ACE-2 component inflates the sampling weights by the inverse response rate to account for unit nonresponse. Variances are estimated using the delete-a-group jackknife with fifteen random groups. More details concerning the ACES survey design, methodology, and data limitations are available online at www.census.gov/econ/aces/.

4.2. The Quarterly Services Survey (QSS)

The QSS examples use data collected from the first quarter of 2004 through the fourth quarter of 2005; Rosenthal and Davie (2008) contains the complete report. The QSS provides Horvitz-Thompson estimates of total and change in quarterly receipts (published about 75 days after the end of the reference quarter) and early estimates of calendar year receipts for selected service sectors. Standard errors are computed using the method of random groups. Sampling units for the QSS are groups of establishments under common ownership – generally companies or administratively convenient parts of companies, including Employer Identification Numbers (EINs). The QSS sample comprises approximately 6,000 units and is subsampled from the Services Annual Survey (SAS).

A new QSS sample is selected every five years. During the five-year cycle, sample maintenance activities are performed each quarter. During this process, out-of-business units are identified and removed from mailing; and newly formed businesses are identified, subjected to a two-phase sampling process, and selected units are added to the sample. These procedures are designed to alleviate undercoverage. Sample units are interviewed each quarter. Thus, QSS estimates are repeated measures estimates.

QSS uses ratio imputation to account for unit nonresponse. Imputation cells are defined by six-digit NAICS code cross-classified by tax status unless the imputation cell contains fewer than ten respondents. In this case, the imputation cell is collapsed to the three-digit NAICS code cross-classified by tax status for all six-digit NAICS contained within the three-digit NAICS code. Within each NAICS by tax status cell, separate imputation cells are created for large companies (mainly consists of large businesses selected with certainty) and EINs (primarily consists of small and medium-sized businesses selected with a weight greater than one). Variances are estimated using the random group estimator with sixteen random groups. Further details about QSS are available at http://www.census.gov/indicator/qss/qsstecdoc.pdf.

5. Analysis of Response Rates

5.1. Retrospective Analyses

Taking a holistic approach, we begin by jointly considering both the URR and TQRRs of key items at a program level over time. In an ideal world, these performance measures will

consistently meet their targets, and the investigation ends before it starts. In our experience, however, target values for performance measures are rarely consistently met when response rate analysis is introduced to a program. Instead, the program computes the performance measures, determines that at least one fails to meet its target, and takes immediate action to increase the rate, such as increasing the nonrespondent contact.

Many programs will also conduct nonresponse bias analyses, which at a minimum include a study of response rates at program and subpopulation level, as suggested by the 2006 Federal Register Notice. Such analyses compute historical measures from prior data and examine:

- The URR and TQRR patterns over time; and
- The average URR and TQRR over the time period.

The program level analysis provides the bottom line, but rarely gives any indication of potential areas of concern (e.g., a downward trend) or an assignable cause for a lower-than-expected value. Thus, a logical next step is to compute response rates by selected subpopulations and examine these rates over time. Once the causes have been identified, corrective actions can be taken. Examining the TQRR for each key item by subpopulation also helps determine the effect of nonresponse on the estimate. The higher the TQRR for an item, the lower proportion of imputed or adjusted data used for estimation.

The following example demonstrates this structured detective work. Table 1 presents the URR for the ACES, both overall and by subpopulations. In Table 1, we see that the program-level (ACES) URR over the time period studied ranges from 71% to 76% (rounded values). Starting in 2004, the rates are monotonically increasing. The ACE-1 (employer) population is further split into two subpopulations, whereas the ACE-2 subpopulation comprises entirely noncertainty cases. To determine whether URR was consistently low for all reporting units or if the URR was disproportionately affected by one sector of the sample, the ACES subject matter experts computed the same measures in each subpopulation.

The URR values for each subpopulation are fairly stable. Moreover, the ACES (and ACE-1 component) URRs increase slightly each year. This is not a coincidence: since 2002, the ACES has participated in an outside agency review that requires performance benchmark targets for URR. This target URR value is determined by the prior year's level, and the program measure must meet or exceed the target. Consequently, analyst respondent contact efforts do not cease until the performance targets are met.

Table 1. ACES URR (2003–2008)

	ACES (All combined)	ACE-1 (Certainty)	ACE-1 (Noncertainty)	ACE-2
2003	72.0	84.4	70.6	58.4
2004	71.1	83.9	74.2	48.1
2005	73.9	84.3	74.0	59.1
2006	74.8	85.3	74.8	59.6
2007	75.2	86.6	75.6	55.1
2008	75.8	86.6	76.6	56.8
Average	73.8	85.2	74.3	56.2

Notice that the (program level) ACES and ACE-1 noncertainty URRs have consistently close – almost indistinguishable – values. Also, notice that the ACE-2 URR values does not appear to "influence" the program level rates and are not in fact increasing along with the ACE-1 rates; the ACE-2 unit response process appears to be stable, however. Thus, any improvement to the ACES URR must concentrate primarily on the ACE-1 noncertainty subpopulation.

Total capital expenditures is ACES's key item. Since capital expenditures are generally not statistically associated with other collected items such as payroll, this item is difficult to impute and is not indirectly obtainable from other data sources. Consequently, the TQRR is an important measure of data quality for the ACES. Table 2 presents the TQRR values for total capital expenditures over the studied time period.

Since the TQRR is a random variable and the point estimates are subject to sampling and nonsampling variability, it is often useful to examine the mean value of the TQRR over a studied time period. This is easily accomplished by constructing a linear combination of the point estimates and using a general linear hypothesis test approach to determine a range of values (potential rates) for this average. To perform a general linear hypothesis test on the mean value of the TQRR, let

- $\underline{\mu}^{\mathrm{T}}$ = the $1 \times \mathrm{T}$ vector of TQRR values for item i computed for the specified population or subpopulation (T refers to the number of available time periods)
- $\underline{K} = d \times T$ matrix of known constants (d = no. of parameters being tested);
- $\underline{\mathrm{K}}_0 = d \times 1$ matrix of known constants representing *potential values* of the TQRR e.g., 0.65, 0.66, . . . ,0.70, . . . , 0.90
- $\Sigma = T \times T$ covariance matrix of the TQRR values. The point estimates in the matrix should incorporate all complex survey design features as applicable.

Our hypothesis of interest is

- $\mathrm{H}_0 : \bar{\mu} = K_0$ (i.e., the mean TQRR for item i over the studied time period is equal to K_0)
- $\mathrm{H}_\mathrm{A} : \bar{\mu} \neq K_0$

The test statistic is given by $(\underline{K\mu} - \underline{K_0})^{\mathrm{T}}(\underline{K}\,\Sigma\,\underline{K}^{T})^{-1}(\underline{K\mu} - \underline{K_0})$ which is distributed as χ^2_1 under H_0.

Table 2. ACES TQRR for capital expenditure (2003–2008)

	ACES (All combined)	ACE-1 (Certainty)	ACE-1 (Noncertainty)	ACE-2
2003	86.02	95.64	72.85	60.53
2004	86.24	95.08	77.94	52.58
2005	88.14	95.80	77.07	61.96
2006	88.86	96.36	78.19	64.74
2007	90.22	96.67	80.32	60.46
2008	90.38	96.65	81.16	59.53
Average	88.31	96.03	77.92	59.97
(Potential rates)	(87.56–89.09)	(*n*/a)	(77.31–78.56)	(58.42–61.53)

We obtain the elements of μ directly using the formula provided in Section 3. Assembling Σ is more involved. For example, programs can use replication methods or Taylor linearization methods to obtain point estimates of the variance and autocovariance elements of the matrix from the survey data or might use averaged variance estimates and autocorrelation estimates if the survey has more than one collection period within each sample to mitigate the effects of sampling error on the variance estimates.

When examining the *mean* TQRR for a given item, we iteratively perform general linear hypothesis tests provided above to determine a range for which H_0 will not be rejected. For these tests, we use $K = 1/T[1. \ . \ .1]$ and $K_0 = 0.01, 0.02, \ . \ . \ . \ , 0.99$. Table 2 provides the values of K_0 for which we fail to reject the null hypothesis at the 10-percent significance level using the ACES data. The final row of the table presents the mean TQRR from 2003–2008, along with these potential rates in parenthesis. Note that the potential rates are not confidence intervals. The potential rates provide information about the *mean* TQRR, and individual point estimate values may not be contained in these intervals. The general linear hypothesis test can be easily modified to test individual point estimates or to test contrasts by using alternative values of K.

At a program level, the average TQRR for total capital expenditures is well above the expected target value of 70 percent, given the lower bound on the average potential rates of 87.56. Moreover, the corresponding average TQRR for the noncertainty ACE-1 component is also above the target value, with the potential rates ranging from 77.31 to 78.56 (although the low point estimate value at 2003 was of concern to the survey managers). At first glance, the low ACE-2 TQRR component is unsettling, with the potential rates falling well below the target. However, the majority of nonzero capital expenditures reported to the ACES are provided by the ACE-1 certainty units and "large" noncertainty companies. Thus, the high URR for the certainty cases translate into high TQRR at the total program level. Clearly, there is little room for improvement in TQRR for the ACE-1 certainty cases. Any improvements to the TQRR would need to be made for the ACE-1 noncertainty or ACE-2 cases, and it is doubtful that improvements in TQRR would be seen with improved ACE-2 response rates, given the high reported zero rate for that subpopulation.

This example demonstrates why both URR and TQRR need to be evaluated simultaneously. Moreover, it illustrates the utility of these analyses by subpopulation over time. In the ACES example, the high TQRR from the certainty component offset the low URR. It should be noted that this is often the exception, rather than the rule. The ACES results are an artifact of the highly skewed population of capital expenditures, where most nonzero capital expenditures are obtained from large companies that are included with certainty or sampled with probability greater than 20% and the remaining units tending to report zero capital expenditures. In general, TQRR's for key items tend to be higher than the program URR, but the degree of difference is generally not as large. Indeed, the distinction between the two measures can be quite negligible when the survey has several key items and certain items are poorly reported (e.g., TQRR for sales could be quite high whereas the TQRR for inventories could be much lower).

5.2. *Process Control Measures*

All processes – including the processes that produce the URR and the TQRR – have some form of variation. Regardless of how well a production process operates, there will always

be an inherent or natural variability that exists. This natural variability or "background noise" is considered normal and acceptable so long as the data points fall within determined limits or do not exhibit a trend. In the framework of statistical quality control, this variability is often referred to as "a stable system of chance causes." A process that is operating with only chance causes of variation present is said to be in statistical control.

In practice, most processes do not remain stable forever. User error, equipment error, or differences due to the input material can lead the process to be unstable. In a business survey environment, response processes could be affected by changes in reporting unit personnel, accounting practices, revised questionnaires, the introduction of new collection methodology, or a change in reporting units, for example. These sources of variability that are not part of the chance causes of variation are called "assignable causes" or "special causes." A process that is operating in the presence of special causes is said to be out of control or unstable (Montgomery 2005, pp. 148–149).

Control charts can be used in many production environments to determine when a process is no longer stable, at which point intervention is needed to find the special cause(s). If the special cause(s) is found and corrected, the system has now stabilized and is once again in a state of control. In finding the special cause(s) and correcting it (them), one has a certain amount of control over the process.

With response rates, the concerns arise when a decrease occurs, especially when there appears to be a trend. In this situation, it is helpful first to analyze the control chart to determine whether there is in fact a trend and whether there is a special cause(s) for the trend. In such cases, an investigation of the response rates on a subpopulation level such as industry code and/or stratum often proves to be beneficial.

The StEPS software mentioned in Section 1 is undergoing a redesign, with a planned 2015 implementation. The redesigned StEPS will have enhanced Management Information System (MIS) capabilities for response rates analysis, including control charts. In Sections 5.2.1 and 5.2.2 below, we discuss techniques for producing the control charts for the URR and TQRR rates.

5.2.1. The Unit Response Rate Control Chart

The URR process can be easily monitored by a p-chart. The statistical principles underlying the p-chart are based on the binomial distribution. Consider a process that consists of n independent trials where the outcome of each trial is dichotomous: "success" or "failure" (i.e., conforming or not conforming) and the probability of success on any trial (called Bernoulli trials) is p. With respect to the URR, we chose to monitor the proportion of sampled units that "respond" (i.e., conform) to a survey (see Section 3).

Assumptions

- The production process is operating in a stable manner across T consecutive statistical periods.
- The URRs calculated in each of the T statistical periods are approximately equal.
- Conditioning on the statistical period, the URR is an exact value.
- Each reporting unit selected for sample and eligible for data collection in the tth statistical period has the same probability, p, of "responding" to the survey.
- The response to the survey is binomially distributed, i.e., $R \sim b(n, p)$.

In the following formulae, a t subscript indicates the URR measure computed at time period t, i indexes the URR at time period t within the studied time interval $[i, \mathrm{I}]$, $\mathrm{T} =$ number of adjacent point estimates used in computation, $\mathrm{T} = (\mathrm{I} - i + 1)$.

Formulae:

$$\hat{p} = \frac{1}{\mathrm{T}} \sum_{t=i}^{I} \frac{R_t}{n_t} = \frac{1}{T} \sum_{t=1}^{I} \hat{p}_t = \text{average value of URR over the studied time interval}$$

$$\hat{\sigma}^2 = \frac{\hat{p}(1-\hat{p})}{\bar{n}}$$

$$\bar{n} = \frac{\sum_{t=i}^{I} n_t}{T}$$

In a control chart, the centerline is given by $\hat{p}$, a rolling average, and the Upper (UCL) and Lower Control Limits (LCL) are given respectively as $\hat{p} + 3\hat{\sigma}$ and $\hat{p} - 3\hat{\sigma}$.

Tague (2004) provides the following indicators for a process that is not stable:

1. A single point plots outside the control limits
2. Two out of three consecutive points that plot on the same side of the centerline exceed 2σ
3. Four out of five consecutive points plot on the same side of the centerline exceed 2σ
4. Eight consecutive points plot on the same side of the centerline
5. An obvious consistent, nonrandom pattern (e.g., an upward or downward monotone trend).

Figure 3 presents a control chart for the QSS URR using measures from the first quarter of 2004 (2004Q1) through the fourth quarter of 2005 (2005Q4), using all eight point estimates to obtain the centerline[3]. The UCL and LCL are presented as solid lines, as is the centerline. The 2σ control limits are dashed lines, and the 1σ control limit is a dotted line. The URR values are plotted as diamonds.

Notice that the URR value at point 4 (2004Q4) is outside of the control limits, providing an indication of an unstable process. The atypically low URR in 2004Q4 corresponds with the introduction of data collection in two additional sectors for the QSS, specifically the Hospitals and Nursing and Residential Care Facilities, so there does appear to be an assignable cause for this atypically low point estimate. The URR values at 2005Q2 and 2005Q3 also provide limited evidence that the process could be unstable by criteria (2) from Tague. However, since the last three point estimates in this plotted series are consistently above the centerline, the evidence that the process is "out of control" is generally considered an improvement, as the goal is always to improve (increase) response rates. This is promising, but three consecutive points above the centerline are indicators of *potential* improvement, not substantive evidence of a change in the process. The *p*-chart does not provide any other indicators of instability with respect to criterion 3 through 5.

[3] Assuming that the size of the rolling average is $\mathrm{T} = 8$, the *p*-chart that would be created for 2006Q1 would have a (slightly) differently located centerline and control limits, since the rolling average would use measures from the 2nd quarter of 2004 through the 1st quarter of 2006.

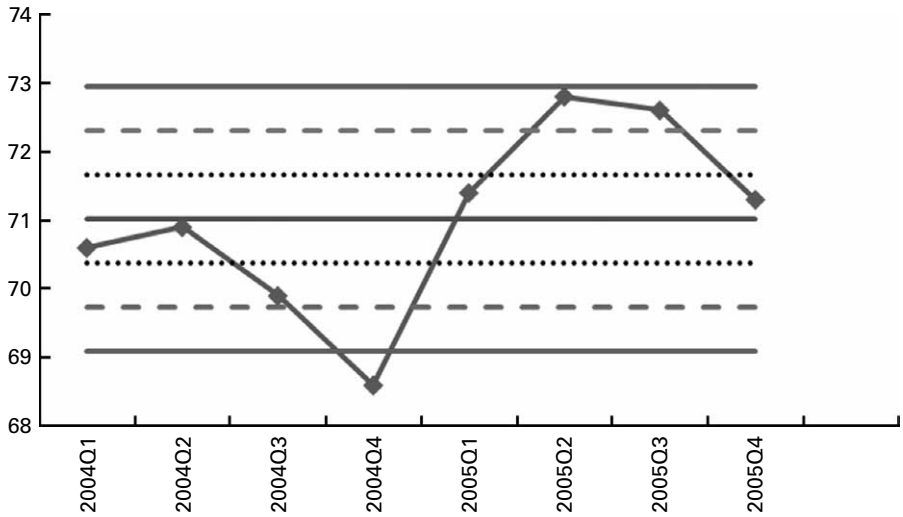

Fig. 3. Unit Response Rate (URR) control chart for the QSS

Our example uses historic data, and we were limited to eight collection periods for analysis. In an ongoing survey, the time series of URR values continuously increases. The control chart procedure needs to determine the length of the interval T that best captures the *current* state of the process. In other words, a predetermined number of earlier values need to be dropped from the computation of the centerline, and the same number of time periods should be used for each computation of the centerline and the upper and lower control limits.

When the process is stable, the rolling averages should not differ greatly between adjacent time periods. If the process is unstable, however, that will not be the case. Therefore, changes in the mean (increases or decreases) would mirror changes in the process. Tague (2004) recommends that $T \geq 20$, to provide a reasonable estimate of the true mean when creating a new control chart. However, this lower threshold was developed in a manufacturing context, where one would expect to produce thousands of measurements in a short time period.

With survey data, we have fewer measures and need to correct unstable processes as soon as they are detected. Consequently, we want to develop rolling averages that are sensitive to process changes, particularly trends. This implies that we should strive to find the fewest number of observations in our interval T that provide accurate means that are not overly affected by an outlier but also quickly reflect changes in the process. Therefore, the objective is to find the smallest value of T that allows the centerline to reflect a process change but minimizes the probability of artificially increasing or decreasing the centerline because of a single outlying observation.

Means are notoriously sensitive to outlying observations, with a breakdown point of one observation. With a small number of observations in the interval T, it may be beneficial to use a median[4] instead of a mean to reduce the computations' sensitivity to

[4] That is, compute $\hat{p}_T = \text{median}(p_t, \ldots p_I)$.

outliers but still allow a shift in the process to gradually be incorporated into the centerline and control limits.

At the U.S. Census Bureau, the majority of business surveys select a new sample every five years. Thus, for a quarterly survey, there will be twenty values of URR for a given sample (one per quarter); there are sixty values of URR for a given monthly sample. Even though the program analysts may remain constant and the collection procedures may remain the same, a large outlying value at the introduction of a new sample (or even as the sample seasons) is not atypical.

It would not be unreasonable for a quarterly survey to use the median value of eight consecutive URR observations as a rolling average. The choice of $T = 8$ dampens the influence of the new sample URR on the rolling average and allows the effects of the new sample on the response process to phase in gradually. The median has a breakdown point of 50 percent, i.e., the outlier-resistance breaks down (exceeds the breakdown point) when the actual number of outliers exceeds the expected number of outliers ($4 = 8 \; x \; 0.50$). Using the median would dampen the influence of the new sample until the fourth quarter of the new sample, when half of the values in T were obtained from the old sample. Thus, there would be an expected one-year lag for a genuine process shift due to the new sample to be incorporated into the control chart limits for a quarterly survey.

The same considerations should be taken into account in developing T for a monthly survey. For a monthly survey that uses a median to obtain a rolling average, a lower threshold of $T = 12$ months would be expected to incorporate a process shift due to the new sample approximately six months after the new sample's introduction. The program manager might, however, wish to consider a longer time interval, especially if monthly response rates are highly volatile.

Determining an interval T for an annual survey is a more challenging problem, especially for a five-year sample. In theory, the process is not changing, just the data. However, the economy does change quite a bit over time, and prior samples reflect the economic environment at the time of selection. Annual programs need to examine their historic data to determine if the control chart approach is even feasible, and if so, will need to determine program-specific values of T.

This discussion above is limited to programs that draw samples that last five years. It is possible that our discussion cannot be extrapolated to programs with different frequencies of sampling. However, there is some anecdotal evidence that program response processes do not change much over time, if the same data collectors (analysts) are used. That said, if the new sample contains a very different set of reporting units than the replaced sample, it is quite possible that the response process would also be substantively impacted as would the control chart limits.

5.2.2. Total Quantity Response Rates

The TQRR is a ratio of two correlated random variables. Consequently, a p-chart with fixed upper and lower control limits does not adequately display the sources of expected process variability and can lead to misleading inferences about the variability due to sampling error. Instead, we propose a variation of the p-chart successfully employed by the National Highway Traffic Safety Administration (Pierchala and Surti 2009) where the control limits vary from one sample to the next (i.e., "stairstep" control limits).

This requires two statistics: (1) μ_t = TQRR estimate for the specified item during statistical period t; and (2) $\hat{\sigma}_t^2$ the estimated variance of the TQRR estimate at time t, i.e., the diagonal elements of the variance-covariance matrix Σ in Section 4.1. Here, the stairstep upper and lower control limits at time t are given respectively as $\hat{\mu} + 3\hat{\sigma}_t^2$ and $\hat{\mu} - 3\hat{\sigma}_t^2$.

Figure 4 presents the modified p-chart for the QSS during the studied quarters. The individual TQRR values are marked by squares, and the individual end-points are indicated by circles. Notice that not only does the process appear to be very stable, but the limits actually narrow slightly as the time period increases. This conforms to expectations: the survey was first introduced in 2003, and response protocols became more refined with the passage of time.

We are investigating the usage of canned software routines to produce viable variance estimates of TQRR for a variety of survey designs, but do not have any definitive answers. For now, we recommend developing program-specific variance estimates.

6. Conclusion

Many ongoing programs compute response rates as performance measures. At the U.S. Census Bureau, we took on the challenge of developing response rates that were also quality indicators. With highly skewed populations, it is not reasonable to develop one single measure of response. Using an unweighted rate for unit response tends to overemphasize the "importance" of the small units in the sample. Using a measure that combines survey weights with a measure of size variable such as sales may understate systematic response process issues with small sampled units by overemphasizing the contribution to totals from the large units. By computing both the URR and the TQRR for key variables and by studying both measures at a program level and by subpopulation, programs get a more complete picture of each individual collection period's data composition and quality.

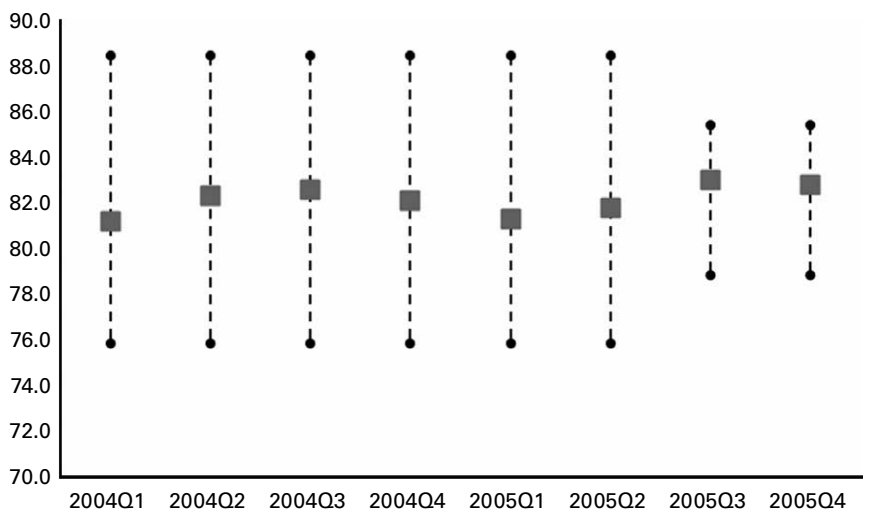

Fig. 4. Modified P-Chart of TQRR (Receipts) from the QSS

When URR and TQRR were first introduced to our economic programs, we held extensive discussions with program managers and statisticians at the design/development stage (determining the formulae) and at the implementation stage. At all times, the objective was to develop a set of measures that were meaningful for the program analysts that could be included in publications as performance measures. Foremost in all discussions was the requirement that the computed measure should increase with data quality improvements, such as improved response contact protocols. This is particularly important in a business survey environment, where the respondent contacts tend to be limited to the larger reporting units. Indeed, program analysts monitor these larger units over time, often developing personal contacts. Smaller cases with larger weights may not be targeted for nonresponse follow-up, depending on the survey periodicity. In other settings, where units are more homogeneous in size or data values provided, it could be worthwhile to explore producing response process control charts at a finer level such as the individual reporting unit, as suggested by a referee.

During the metric-development process, it became evident that the response rates measures could be viewed in a statistical process control framework. Even now, typical analyses of response rates are limited to comparison to the previously obtained value or perhaps to the value obtained one year ago. Having standard measures in place and systematic analyses procedures such as those presented in Section 5.1 is an advance over ad hoc procedures. However, these types of comparisons still treat response rates as performance measures, with the underlying assumption that a measure that fails to meet or exceed a benchmark value is a failure.

In the Economic Directorate at the U.S. Census Bureau, we are moving towards standardizing this statistical process control framework of response rates. Both measures are computed in the StEPS system, and we have provided extensive training on their implementation to our program managers. For now, we have introduced control charts to selected programs on a case-by-case basis usually as a part of a nonresponse bias analysis. The redesigned StEPS will include automatically generated time series plots of URR and TQRR, and we expect these charts to include the control chart limits for URR and TQRR when the revised system is introduced in 2015. This approach to response rate analysis is a new paradigm, and we will need to provide training to methodologists and program analysts. In the meantime, as we continue to develop requirements for these useful visual tools, we have additional questions that need to be answered, such as optimal interval size for URR and TQRR centerlines, advantages of mean or median statistics for the centerline, and possible "shortcuts" for standard error computation for TQRR.

The statistical process control framework presented in our article allows one to see the how the process is operating in the past and currently. This allows the program manager to estimate process capability and to determine when a genuine shift takes place. If properly implemented, it can demonstrate when an intervention is necessary to bring the process into control or to improve the process. However, our process control setting is unconventional with respect to the standard literature. Surveys do not have thousands of observations in a relatively short period of time (cf. a factoring manufacturing process) and corrective actions must be made fairly quickly. Several factors need to be taken into account, such as frequency of collection, duration of sample, and relationship of mandated performance benchmarks to data-determined control chart limits. Dealing with these

factors brings us one step closer to our ultimate goal: total quality management. A total quality management approach would continue to monitor the response rate process measures on an ongoing basis to ensure that once acceptable rates are achieved, the process remains in control. Such an approach should be undertaken jointly by survey methodologists and subject-matter expert program managers.

7. References

Ahmed, S.A. and Tasky, D.L. (2000). An Overview of the Standard Economic Processing System (StEPS). Proceedings of the Second International Conference on Establishment Surveys.

Federal Register Notice (2006). OMB Standards and Guidelines for Statistical Surveys.

Kalton, G. and Flores-Cervantes, I. (2003). Weighting Methods. Journal of Official Statistics, 19, 81–97.

Lineback, J.F. and Thompson, K.J. (2010). Conducting Nonresponse Bias Analysis for Business Surveys. Proceedings of the American Statistical Association, Section on Government Statistics.

Montgomery, D.C. (2005). Introduction to Statistical Quality Control. New York: John Wiley.

Pierchala, C.E. and Surti, J. (2009). Control Charts as a Tool for Data Quality Control. Journal of Official Statistics, 25, 167–191.

Peytcheva, E. and Groves, R.M. (2009). Using Variation in Response Rates of Demographic Subgroups as Evidence of Nonresponse Bias in Survey Estimates. Journal of Official Statistics, 25, 193–201.

Rosenthal, M. and Davie, J. (2008). Nonresponse Bias Analysis for the Quarterly Services Survey. Internal unpublished U.S. Census Bureau Memorandum, available upon request.

Smith, J.Z. and Thompson, K.J. (2009). Nonresponse Bias Study for the Annual Capital Expenditures Survey. Proceedings of the American Statistical Association, Section on Government Statistics.

Tucker, C., Dixon, J., and Cantor, D. (2007). Measuring the Effects of Unit Nonresponse in Establishment Surveys. Introductory Overview Lecture. Third International Conference on Establishment Surveys, the American Statistical Association.

Tague, N.R. (2004). The Quality Toolbox (Second Edition). ASQ Quality Press.

Received January 2011
Revised December 2011