Response Rates as Process Control Tools: Creative Usage of Existing Performance Measures

Katherine Thompson, Broderick Oliver, Yarissa González[1]
U.S. Census Bureau, 4600 Silver Hill Road, Washington D.C. 20233
katherine.j.thompson@census.gov

Abstract

The Federal Statistical Standards require that programs publish response rates. For business surveys, a single measure does not adequately describe the response process because of the highly skewed nature of business populations. Consequently, business surveys conducted at the U.S. Census Bureau compute two types of response rates: the *unit response rate*, defined as the unweighted proportion of responding units, used as a measure of survey response; and *total quantity response rates*, defined as weighted proportions of estimates obtained from reported or equivalent quality data. Surveys may also compute the *weighted volume response rate*, computed as the weighted proportion of the responding units' measure-of-size. Using empirical data from two surveys, we demonstrate how to use these metrics to understand and monitor the response process via statistical process control methods adapted for usage with complex survey data.

Key words: response rates, statistical process control, control chart

1. Introduction

The Federal Statistical Standards require that programs publish response rates (Federal Register Notice, 2006). For business surveys, a single measure does not adequately describe the quality of the response process. Consequently, business surveys in the Economic Directorate of the U.S. Census Bureau compute two types of response rate. The first measure is a respondent level response rate called the *unit response rate* (URR), defined as the unweighted proportion of the sampled units eligible for data collection that respond to the survey. The second measure is an item level response rate called the *total quantity response rate* (TQRR), defined as the weighted proportion of a key estimate that is obtained from directly reported data or from validated external sources where the data are deemed to be of equivalent quality as the reported data. By definition, at most one program-level URR and one or more program-level TQRRs (one per key estimate) are produced for a given program. Taken jointly, these measures provide indications of the quality of the survey response, under the implicit assumption that the quality of the program data increases as each rate approaches 100 percent.

Why do business surveys produce two separate metrics? Business populations are highly skewed. Often, a relatively small number of the tabulated units contribute significantly to the estimated totals. Consequently, the majority of the economic surveys administered at the U.S. Census Bureau employ stratified designs where the largest units are included with certainty and the remaining units are sampled. To avoid overrepresentation of small

[1] This report is released to inform interested parties of ongoing research and to encourage discussion of work in progress. Any views expressed on statistical, methodological, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

units with large sampling weights in the program-level response measure, the URR is computed without using design weights. Doing this, however, tends to downplay the importance of the certainty or larger sampled cases, making the URR an inconsistent indicator of overall data quality. With skewed sampled data, a more consistent measure of data quality includes a weighted measure-of-size (e.g., payroll, capital expenditures) to account for the unit's relative importance in the estimates (Tucker et al. 2007), such as the TQRR.

Response rates are commonly treated as performance metrics that need to be compared against benchmark measures. In fact, the Office of Management and Budget (OMB) guideline requires a nonresponse bias analysis be conducted for surveys with a URR of less than 80 percent or an item response rate (TQRR) of less than 70 percent (Federal Register Notice, 2006). These benchmark measures are simply guidelines. Unfortunately, it is quite possible for a business survey to have a high URR (close to 100 percent) and still produce biased estimates (Peytcheva and Groves, 2009), particularly if any of the large units have not responded (Thompson and Oliver, 2012).

Often, response rates are compared to prior period values with the objective of increasing the response rate or at least maintaining a constant level. However, such limited comparisons can result in misleading insights into the stability of the response process[2]. Therefore, Thompson and Oliver (2012) propose viewing response rates as process measures instead of performance metrics and reframe their analysis within a statistical process control framework. In particular, they advocate analyzing response rates over time using control charts to help survey managers assess the stability of the response process and to help them distinguish between variation that is inherent in a stable process and variation that is unusual, and in many cases unexpected. Control charts are also useful for determining process capability, which would be useful for programs that have target benchmark values. If a program's current process were stable, but not able to achieve these target values, then a modification of the current process (i.e., an intervention) would be required.

The statistical process control framework generally assumes a large number of measurements are available in a relatively short time. However, this is simply not true for the response rate process of an ongoing survey. At best, we can obtain twelve unit response rate values in a given year for a monthly survey and as few as four for a quarterly survey. Consequently, the analysis procedure needs to reflect actual changes in survey process quickly (after a predetermined "burn-in" period) while remaining insensitive to one-time outlying measurements.

This paper builds on the work presented in Thompson and Oliver (2012), González, Oliver, and Thompson (2013), and González and Oliver (2013). The latter two papers present a modified *p*-chart for the URR and illustrate its usage on two monthly business surveys. Although the presented applications provide useful insight into the surveys' respective response processes, the authors acknowledge the limitations of the URR as a measure of survey quality. The TQRR provides more insight into the quality of a given estimate but tracking the TQRR over time using a *p-chart* presents unique challenges because its numerator and denominator are random variables, and the denominator value

[2] A process is a set of activities that follow some logical flow, where value is added and some output/result is expected (Workbook on Leadership, 2003)

is expected to change each collection period. As a compromise between the URR (which does not take unit size into account) and the TQRR (which takes unit size into account, but differs by variable and is more subject to sampling and nonsampling errors), we propose an additional measure that incorporates a unit's sampling weight and frame measure-of-size value to approximate the proportion of the population size that has responded to the survey. The process described by this *weighted volume response rate* (WVRR) can be monitored likewise with a *p*-chart.

In Section 2, we provide background information on the response processes of business surveys to provide context for how we developed the parameter values for the *p*-charts for the three process metrics (URR, WVRR, and TQRR). In Section 3, we formally introduce each of these process metrics and describe the accompanying control charts. Section 4 presents two case studies that demonstrate how to analyze the different response processes, using empirical data from two monthly business surveys conducted by the U.S. Census Bureau. We provide concluding remarks in Section 5.

2. Response Processes for Business Surveys

Unlike household surveys, in business surveys, not all units are equal. Some units contribute more to the estimates of total for the characteristic(s) of interest than others do. Therefore, business surveys generally employ a stratified sampling design using some measure-of-size value available to all units on the frame. With business surveys, it is also important to define the three types of units that play a role in the survey process: the *survey unit*, the *reporting unit*, and the *tabulation unit*. The *survey unit* represents the entity selected via a probability sample from the frame. *Reporting units* are established by the sampled business to collect survey data. *Tabulation units* house the data for estimation. For example, a firm (company) can comprise several establishments and can operate in more than one industry. A sampled firm might request that a separate questionnaire be mailed to each of its establishments rather than complete a single questionnaire. Each establishment would therefore be a reporting unit. The data provided from the reporting units are allocated to artificial tabulation units, each representing the firm part that operates in a given industry (Thompson and Oliver, 2012).

For business surveys conducted at the U.S. Census Bureau, there are two types of response processes: a unit response process, measured by the *unit response rate* and a tabulation process, measured by the *total quantity response rate*. The *weighted volume response rate* is a hybrid of these two metrics. We formally define each of these metrics, calculated for a given statistical period, *t* below:

The *unit response rate* (URR) is based on reporting units, the units that have the opportunity to respond to the survey. The U.S. Census Bureau Quality Standards (Methodology and Statistics Council, 2012) defines a *respondent* as an eligible unit for which: (1) an attempt was made to collect data; (2) the unit belongs to the target population; (3) the unit provided *sufficient data* to be classified as a response. To provide sufficient data, a reporting unit must provide a valid value for its required data item(s). The validity of a reported value is determined by an automatic editing procedure or by an analyst review. The URR is defined as

$$URR = \frac{R}{n}$$

Where *R* is the number of eligible reporting units that respond to the survey (i.e., classified as a respondent) and *n* is the total number of reporting units (eligible units and units whose eligibility could not be determined -- for example, an undeliverable-as-addressed.

The *weighted volume response rate* (WVRR) modifies the URR by including the design weight and a constant measure-of-size (MOS) for each *reporting unit*[3] in the numerator and the denominator. The WVRR estimates the proportion of the population that responded to the survey, and is given by

$$WVRR = \frac{\hat{N}_R}{\hat{N}}$$

Where $\hat{N}_R$ is the unbiased estimate of the frame population obtained from *R* respondents used in the URR and $\hat{N}$ is an unbiased estimate of the frame population size during the statistical period. As with the URR, the denominator of the WVRR can change each statistical period.

The WVRR is a useful measure for approximating coverage, but can be less useful for assessing quality with respect to amount of non-imputed data in the tabulated estimates. This is especially of concern for survey characteristics that are not highly positively correlated with the MOS, for example, real valued characteristics such as income.

The *total quantity response rate* (TQRR) of a characteristic measures the proportion of the weighted estimate obtained by reported or equivalent quality data. For an item *x*, the TQRR is defined as

$$TQRR^x = \frac{\hat{X}_R}{\hat{X}_N}$$

where $\hat{X}_R$ is the unbiased estimate obtained from units that *responded* to the survey with a valid value or from units whose value was substituted with equivalent quality data, and $\hat{X}_N$ is the estimate of characteristic for the survey obtained by using the final (possibly imputed) value of *x* weighted by the design weight multiplied by a nonresponse adjustment factor as applicable. The denominator of the TQRR may not correspond to the published total for the characteristic if the final survey weights include post-stratification, coverage, or calibration adjustments.

3. Response Rates within a Statistical Process Control Framework

Regardless of how well a production process operates, inherent or natural variability will always exist. When this random variation is "small," the process is said to be *stable* and operating in *statistical control*. Variation that is "large" or non-random results from a *special* or *assignable cause*. A process that is operating in the presence of special

[3] When one survey unit corresponds to more than one reporting unit, the survey methodologists must appropriately allocate the MOS and the design weight from the sampled survey unit to each reporting unit.

cause(s) is said to be *out of control* or *unstable* (Montgomery, 2005). Here, we consider two interdependent but separate response processes: the *reporting unit response process* and the *tabulation unit response process*.

With respect to the URR, the concern occurs when the URR falls below a benchmark measure or if the rates appear to be declining over time. With survey data, we need to correct these unstable processes as soon as they are detected and validated. We propose using a *p*-chart to monitor the URR process. A *p*-chart plots individual process measures (URR) against a centerline (a process average) and control limits. To develop these statistics, we assume:

- The URR value is approximately constant across *T* consecutive statistical periods.
- Conditioning on the statistical period *t*, the URR is an exact value.
- Each reporting unit in the t^{th} statistical period has approximately the same probability (*p*) of responding to the survey.
- The response to the survey is binomially distributed: $R \sim \text{bin}(n, p)$.

The true process average (*p*) is unknown and is estimated using a rolling average of the *T* most recent URR values that fall on the time interval [*i*, I]. Following the recommendation of González and Oliver (2013), we estimate the centerline (the process average) with the *median* value of the five most recent response rate values (*T*=5). Using a small interval for the rolling average retains sensitivity to process changes (particularly trends). Using the median instead of a mean protects against artificially increasing or decreasing the centerline because of a single outlying observation. We estimate the standard error $(\hat{\sigma})$ of the estimated process average $\hat{p}$ as $\hat{\sigma}_{URR} = \sqrt{\frac{\hat{p}(1-\hat{p})}{\bar{n}}}$, where $\bar{n} = \frac{\sum_{j=I-4}^{I} n_j}{5}$. We calculate our control limits at the 3-sigma (3σ) level. For a stable process drawn from a normal or sufficiently large binomial population, approximately 99.7% of the URR values are expected to fall between $\hat{p} - 3\hat{\sigma}_{URR}$ and $\hat{p} + 3\hat{\sigma}_{URR}$. Values that fall outside of this range either occur by chance (with a 0.3% probability) or occur due to an assignable cause (Montgomery, 2005). The justification for development of the upper and lower control limits follows from the usage of a *process mean* to estimate the centerline of a stable process from a binomial distribution, invoking the Central Limit Theorem. However, by using a median instead of the mean to estimate the process average, we can no longer invoke the Central Limit Theorem. That said, the protection against outlying observations offered by the median and the short burn-in period (approximately [T/2] statistical periods) outweigh this theoretical disadvantage, and the empirical results presented in Section 4 are not detrimentally affected.

The number of reporting units used to compute the unit response rate affects the width of the control limits. As the number of reporting units decrease, the control limits widen and may not be able to discern a shift in the process average or detect outliers. Therefore, we set a desired tolerance level, *d* such that if $3\sigma \leq d$ we recommend producing a *p*-chart; otherwise, we suggest only producing a time series plot against the centerline, without control limits. For our research, we chose an arbitrary tolerance level of $d = 0.02$.

To determine when the URR process is becoming unstable, we recommend using the following three guidelines to inspect the *p*-chart (González and Oliver, 2013):

- A single point plots outside the control limits.
- Three consecutive points plot on the same side of the centerline.
- An obvious consistent, nonrandom pattern in the plotted values (e.g., an upward or downward monotone trend).

The guidelines are a subset of control chart guidelines delineated by Tague (2004). Tague's guidelines include two other indicators based on narrower limits than 3σ. We expect our estimated standard deviation to have more variability than similar measures from a manufacturing setting and therefore use only the 3σ limits to reduce the possibility of falsely concluding that the process is out of control. In addition, because of the length of time between consecutive URRs and the need to quickly remediate any identified out-of-control decreasing URR process, we view three points on the same side of the centerline as an indication of a trend instead of the eight suggested by Tague.

We have two objectives when monitoring survey data with a *p*-chart. The first objective is to assess the *current state* of the process. Using the modified Tague guidelines, we inspect the *end* of the series -- the *most recent observations* to assess the stability of the current process. If the process is unstable, the survey manager should conduct research to determine if the instability is due to an assignable cause (e.g., a few industries that have unusually low reporting rate) or is endemic. Such investigation may include producing separate control charts or time-series plots by subdomain. Interventions should be undertaken immediately to bring the process into control. The second objective is to assess the process capability. We do this by inspecting the *complete series* for trends in the data, upward or downward that could provide insight into the current level of the estimated process average. A long-term downward trend, which is problematic, would provide evidence to the survey manager that an intervention is needed to reverse this trend and stabilize the process.

Determining when the TQRR process is becoming unstable is less straightforward than the URR or the WVRR. The TQRR is a ratio of two correlated random variables, so 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. We present a variation of the modified *p*-chart successfully employed by the National Highway Traffic Safety Administration (NHTSA) where the control limits vary from one sample to the next (Pierchala and Surti, 2009). These stairstep control limits reflect the variability in the denominator. In the NHTSA applications, the centerline is estimated as the mean of the process using historic data and is fixed for a calendar year. In our applications, we use the rolling median of the most recent two years' observations (24 points in a monthly survey, eight in a quarterly survey), which allows changes due to a sample redesign to phase in over time. The usage of a rolling median over a longer time interval likewise reduces the effect of expected special reporting arrangements on the centerline. For example, to ease respondent burden, the Census Bureau allows sampled businesses to develop special reporting arrangement such as providing monthly data on a quarterly basis: in these months, the TQRR values are much larger than in other months because of the large influx of data.

The modified *p*-chart requires two statistics: (1) the TQRR estimate for the specified item during statistical period *t*; and (2) the estimated variance of the TQRR estimate at

time *t*, i.e., the $\hat{\sigma}^{X}_{TQRR,t}$. Pierchala and Surti (2009) found that 3-sigma limit control charts flagged an overly high proportion of processes where the sample size varies considerably from one statistical period to the next. We found this to be the case with the TQRRs and therefore followed the authors' suggestion and produced 5-sigma limit stairstep control charts. The upper and lower 5-sigma control limits of the TQRR for item *X* at time *t* are given by $TQRR^{X}_{median} \pm 5\hat{\sigma}^{X}_{TQRR,t}$. We apply the same guidelines for analysis as with the earlier *p*-charts.

In concept, the proposed TQRR control charts are easy to implement. In execution, it is challenging. It is possible to produce sample-based point estimates of the $\hat{\sigma}^{X}_{TQRR}$ using Taylor Linearization or replication. However, these variance estimates can be quite variable or unstable, especially with small sets of respondents and can be burdensome to implement. Instead, we recommend developing a parameterized model for the variances using a generalized variance function (GVF). This approach generates smooth variance estimates that depend upon the statistics needed for TQRR production ($\hat{X}_R$ and $\hat{X}_N$). Our GVF models have two important properties: (1) the relative variances of $\hat{X}_R$ and $\hat{X}_N$ are a decreasing function of the corresponding estimate levels and (2) the level of estimated variance of the $TQRR^{X}$ decreases as the measure approaches 100 percent.

4. Case Studies

In this section, we use our recommended control charts to study the response processes of two business surveys: the Monthly Retail Trade Survey (MRTS) and the Monthly Wholesale Trade Survey (MWTS).

4.1. Monthly Retail Trade Survey

The U.S. Census Bureau's Monthly Retail Trade Survey (MRTS) samples approximately 12,000 retail businesses with paid employees to collect data on sales and inventories. The MRTS is an economic indicator survey, whose monthly estimates are inputs to the Gross Domestic Product (GDP) estimates, and the primary item of interest is receipts (total and month-to-month change). The MRTS is a stratified simple random sample, selected approximately every five years. Companies are stratified by their major kind of business (industry) then are further sub-stratified by estimated annual receipts or revenue. All companies with total receipts above applicable size cutoffs for each kind of business are included in the survey as part of the certainty stratum. Within each noncertainty size stratum, a simple random sample of employer identification numbers (EINs) is selected without replacement. Thus, the sampling units are either companies or EINs. The initial sample is updated quarterly with a sample of births (new businesses) and removal of deaths (businesses no longer in operation). The most recent MRTS sample was selected in 2012. Totals are Horvitz-Thompson estimates and variances are produced using random group variance estimation with 16 random groups. See http://www.census.gov/retail/mrts/how_surveys_are_collected.html for more details on the MRTS estimates and methodology.

We use 48 months of historic data from the 2006 sample, beginning in January 2009 and ending in December 2012. All presented response rates (URR, WVRR, and TQRR for receipts) were obtained from U.S. Census Bureau's Standard Economic Processing

System (StEPS). We developed our own GVF parameters for the TQRR control charts by fitting the relative variance model discussed in Section 3 to *industry* estimates (at the six-digit NAICS level) of total receipts (the TQRR denominator) independently in twelve consecutive statistical periods.

Figure 1 presents the *p*-chart for the URRs. On average, there were approximately 11,500 reporting units per statistical period. The *p*-chart's process average of 60.1 percent is the median of the five most recent URR values (August – December 2012). Examining the six *most recent observations* (July - December 2012), we find two indications that the unit response process is declining and therefore out of control using our modified Tague guidelines: a monotone downward trend from September through December and a single point below the lower control limit in December 2012. Moreover, the control chart shows that this downward trend in the URRs has persisted from (at least) 2009. Except for one rate (in November of 2011), all URRs fall above the upper control limit indicating that the process average has clearly shifted downward. We verified our visual suspicions by calculating the average response rate by years (using last five months of each year): 2009 (65.4 percent); 2010 (64.0 percent); 2011 (62.3 percent). Analysis of the URR at the industry (domain) level showed the same declining pattern, suggesting that the cause of the program-level decline is not confined to a subset of industries.

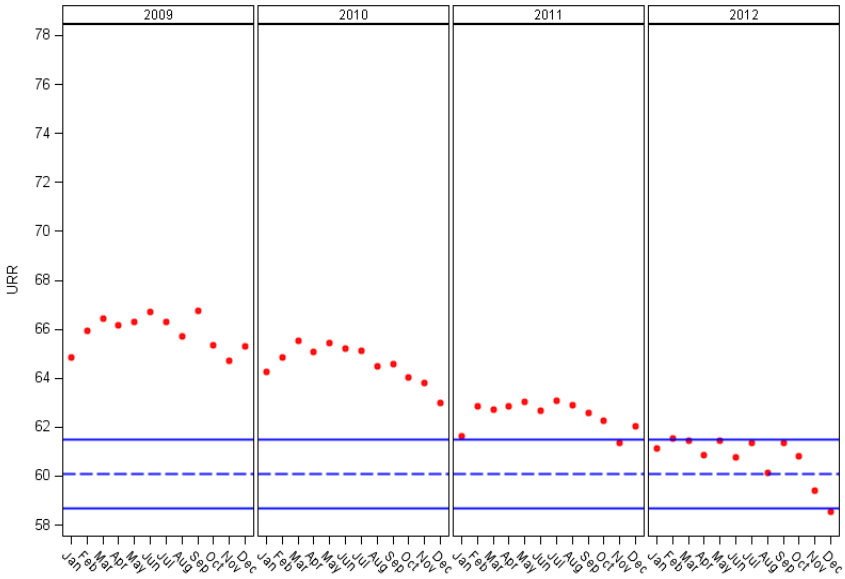

Figure 1: *P*-Chart for URRs for Retail, 2009-2012

The *p*-chart in Figure 1 also reveals a cyclical pattern of the URRs within year, which is consistent with the nature of the retail businesses. Some businesses are busier at certain times of the year, which makes them less likely to respond. Other factors such as filing requirements (e.g., tax filings) may also contribute to the lower response. Although this self-correcting process exists, it is not sufficient to offset the overall downward trend. Without viewing this process via a control chart, a survey manager might dismiss the downward trend, expecting that the natural correction process would reverse it. Nevertheless, the control chart provides strong evidence that an intervention is necessary to bring the response process into control. Examining the *p*-chart for the URR demonstrates that the reporting rate is declining. However, this phenomenon may not

affect the quality of the estimates if the respondent set represents a high proportion of the sample; that is, if the WVRR process average is close to 100-percent.

Figure 2 presents the *p*-chart for the WVRR. First, notice that the process average of 63.4 percent for the WVRR (the median of the five most recent values) is slightly higher than that of the URR (60.1 percent), but is still not close to 100-percent. An inspection of the six most recent points reveals that four of the WVRR rates fall outside of the control limits. The current process is clearly not stable. The extreme outliers (spikes) occurring in January, April, July, and October of 2010 results from some businesses providing their data quarterly instead of monthly (by arrangement). These same spikes occur in the 2011 and 2012 data (for the same reason). An inspection of the series from 2010-2012, the years that have this same arrangement for the late reporters, the estimated WVRR process average is decreasing, meaning the percentage of the frame population represented by the respondent set is decreasing.

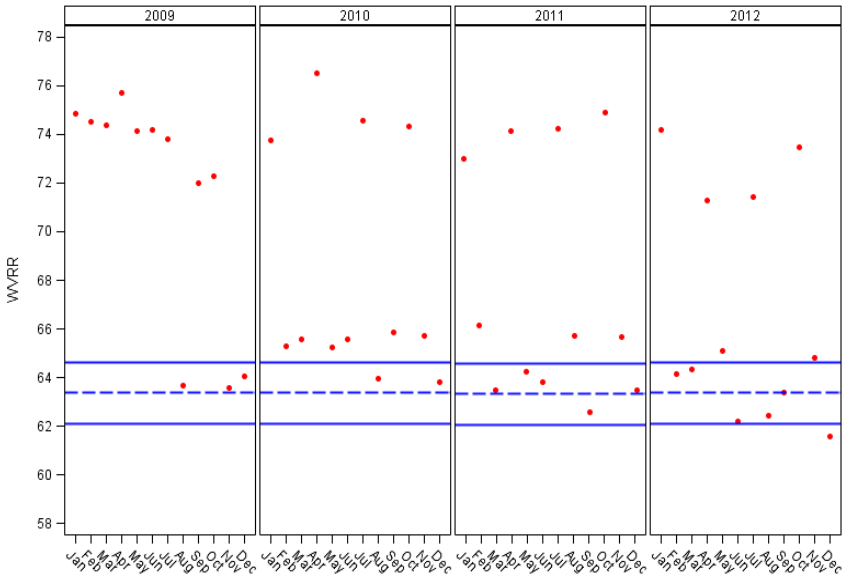

Figure 2: *P*-Chart for WVRRs for Retail, 2009-2012

Neither of these analyses demonstrates the effect of a declining unit response process on the key survey items. Figure 3 presents the modified *p*-chart with stairstep limits of the TQRR for *receipts*, a key item. We begin by examining the most recent points in the series. Of the six most recent points, four fall outside of the control limits. The TQRR process is clearly unstable. Similar to the WVRR, beginning in January 2010, the process experiences spikes in January, April, July, and October (for the same reason mentioned previously)[4]. Again, examining the series from 2010-2012, the years that have a similar arrangement for late reporters, the estimated process average is decreasing. The effects of the sample attrition and decreasing response are demonstrated by increasingly wider stairstep limits.

[4] Beginning in 2012, the production survey mathematical statisticians began producing revised response measures; however, for our case study, we did not have access to these revised files.

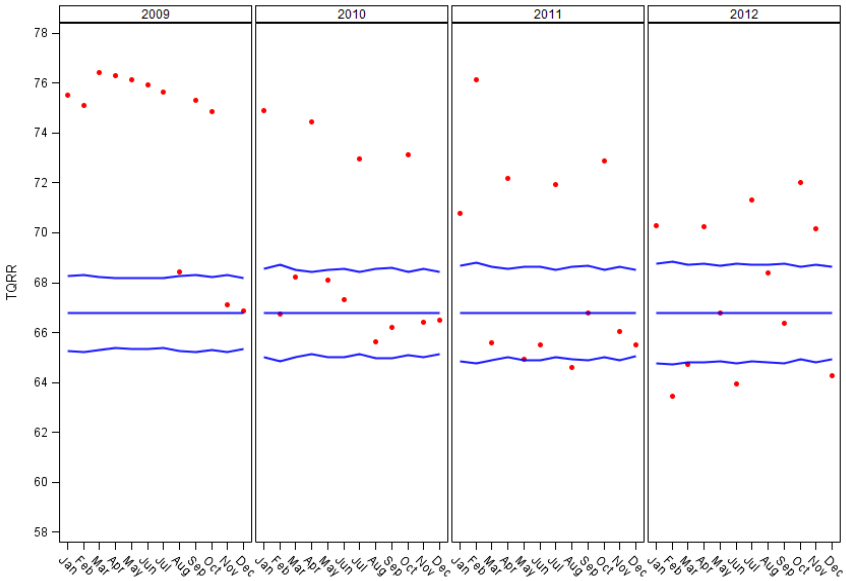

Figure 3: Modified *P*-Chart for TQRRs for Receipts for Retail, 2009 -2012

4.1.1 Summary of MRTS Response Process

Each of these control charts illustrates different facets of the response process. The URR *p*-chart provides the clearest evidence of a declining process over time and provides compelling evidence that the process will continue to decline without an intervention. Unfortunately, the relationship between declining URR and data quality is not linear, therefore, we examine the WVRR process. Like the URR process, the WVRR levels fall far below 100 percent. The respondent set represents a relatively small percent of the frame population. The TQRR *p*-chart provides a measure of the overall impact of these declining response processes on a key estimate, receipts and provides a useful picture of the cumulative impact of the declining response process on this quality measure. Over time, the declining number of respondents leads to an unstable process. If interventions are needed, the URR and WVRR *p*-charts are more useful, as they examine a single source of process variability.

4.2 Monthly Wholesale Trade Survey

The Monthly Wholesale Trade Survey (MWTS) provides monthly estimates of sales and inventories (total and month-to-month change) of wholesale trade industries. The inventory estimate is an input to the GDP estimates. The MWTS samples approximately 4,500 wholesale firms. The sample is selected approximately every five years and updated quarterly to account for births and deaths, with the most recent sample selected in 2006. As with the MRTS, sampling units are either firms or EINS. See http://www.census.gov/wholesale/www/about_the_surveys/monthly_overview.html for more information on the MWTS estimates and methodology.

For this case study, we had three years of historic data to produce URRs, WVRRs, and TQRRs for receipts. All presented response rates (URR, WVRR, and TQRR for receipts) were obtained from StEPS. The modified *p*-chart and time series plot of the TQRR for receipts uses random group variance estimates with 16 random groups.

Figure 4 presents the resulting *p*-chart for the three years of response rate data, January 2010 to December 2012 on the program-level. Since the average sample size for MWTS was not large enough to produce control limits within our desired tolerance level (d = 0.02), we increased the tolerance level to 0.025 so that we could produce a *p*-chart for the MWTS on at least the program-level. As usual, we derived the parameters for this *p*-chart from the five most recent rates (August-December of 2012).

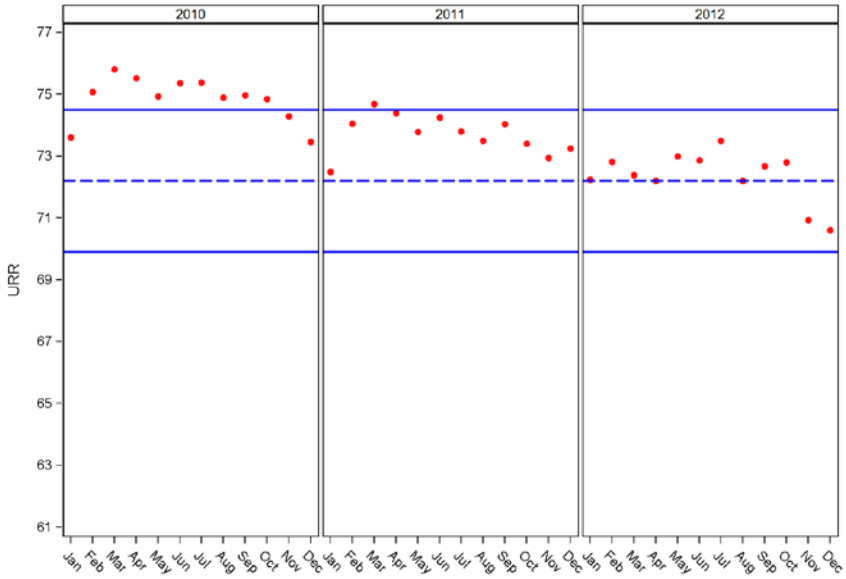

Figure 4: *P*-Chart for URRs for Wholesale, 2010 - 2012

Confining ourselves to the most recent data from 2012 -- the last six months for example, we notice that three of these six points fall above the centerline. In fact, for the entire year, at least 7 of the 12 most recent rates exceed the process average. The response process is clearly out of control (i.e., not stable). Across the three studied years, the URRs exhibit a similar cyclic pattern and declining trend as observed with the MRTS's rates. The process average is clearly declining. To determine if this decrease has an assignable cause we analyzed the URRs by form type, a variable that the MWTS uses to produces URR rates on a domain level.

The MWTS has two separate forms: Form-A for companies that report for multiple EINs and Form-E for companies that report for a single EIN. Form-A reporters are generally larger than Form-E reporters, in terms of volume of sales. Respondent sample sizes in each subdomain (Form) were too small to produce usable control limits, so we produce the time series plot shown in Figure 5.

The high process average of 82.4% for the URR for the Form-A companies greatly contrast with the lower process average of 66.8% for the Form-E companies. Although the rates have been declining for both the Form-A and Form-E companies since (at least) 2010, the decline in the latter is more pronounced. These declining rates, particularly those from the Form-E companies, help to explain the declining rates observed overall on the program-level. To assess the effects of this decline on the respondent sample, we examine the *p*-chart of the WVRR presented in Figure 6.

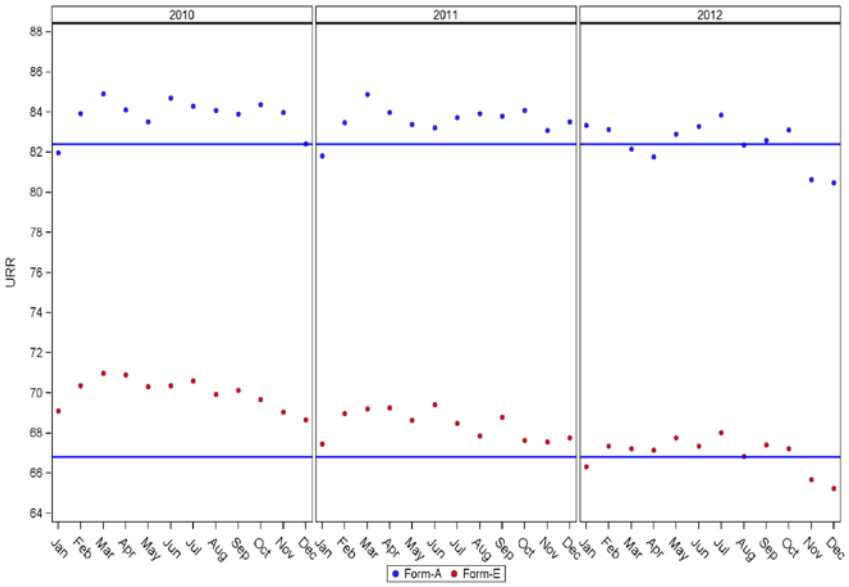

Figure 5: Time Series Plot for URRs by Form Type for Wholesale, 2010-2012

We begin our analysis by examining the six most recent rates in 2012 and observe that besides the low outlier value in November, all other values, including those from the first half of the year, all fall within the control limits and randomly about the estimated process average. The current process is clearly stable. However, an examination of the entire series, from 2010 through 2012 reveals a declining process average. The pattern displayed in the *p*-chart for the WVRR mirrors the URR process. However, the centerline is lower (69.5% versus 72.2%), demonstrating the continuing effect of the missing small single units in the sample. This control chart tells a slightly different story than the URR chart. Here, we see a decline in the proportion of sample obtained from respondents from 2010 to 2011, but a process that has largely stabilized by mid-2011. Furthermore, although there are indications that the process might again be going out of control in 2012 (the November 2012 WVRR falls below the control limits), it appears that the process corrects itself in December 2012. A survey manager might therefore monitor the process but would not necessarily be compelled to perform an intervention.

Finally, we examine the TQRR process for total receipts (see Figure 7). The five most recent rates all fall below the estimated process average, indicative of a declining process average. Overall, 8 of the 12 values in 2012 fall below the process average. The current process is clearly not stable. An examination of the studied years, 2010 through 2012 reveals a declining process average. As with the MRTS, the decreasing response rates processes (URR especially) affects the reliability of the respondent sample-based estimates, as demonstrated by the increase in the stairstep limits from 2010 to 2011.

4.2.1 Summary of MWTS Response Process

In examining the response processes for the MWTS, we clearly see a decline in unit response rate as measured by the URR and WVRR. This decline is largely attributable to a declining response for the small units. Hence, an intervention focused on the small units could prove helpful, although it will not likely have a large impact on the TQRR values since the small cases contribute relatively little to the numerator and denominator.

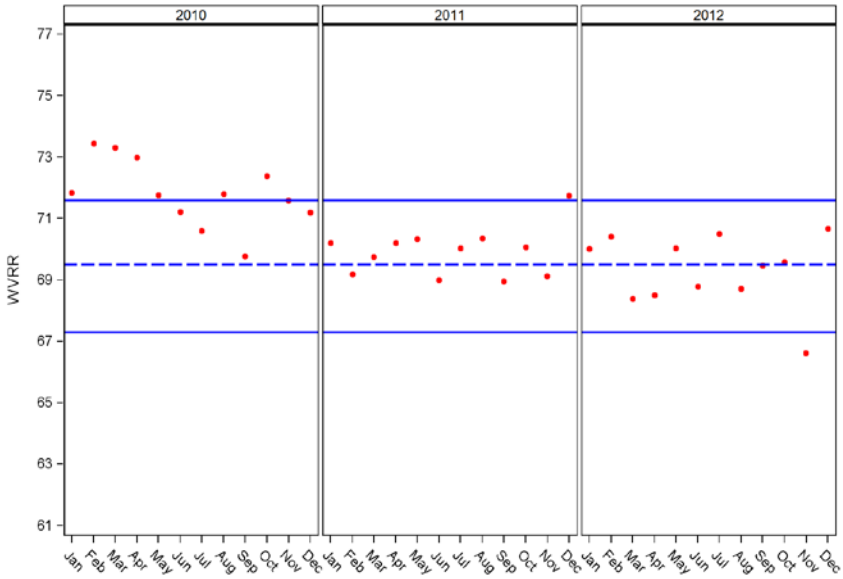

Figure 6: *P*-Chart for WVRRs for Wholesale, 2010-2012

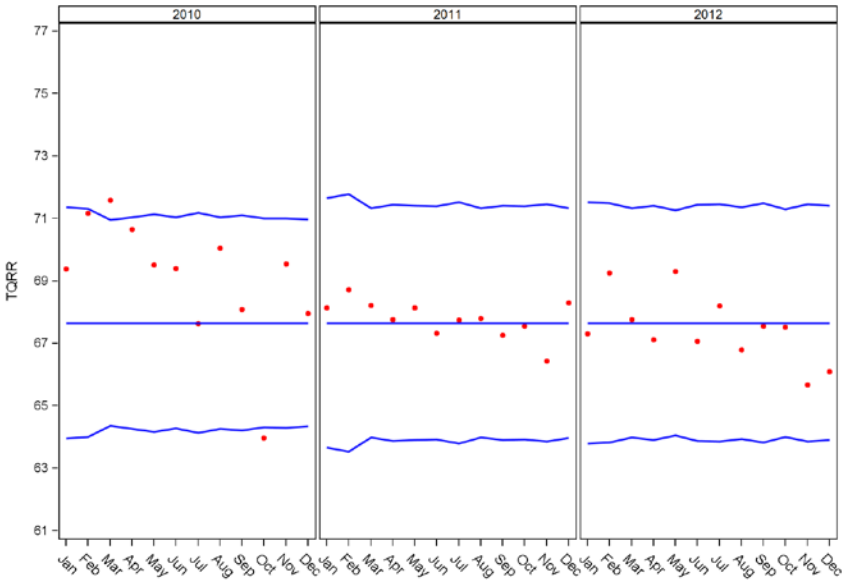

Figure 7: Modified *p*-chart for TQRR for Receipts for Wholesale, 2010–2012

5. Conclusion

In this paper, we demonstrate that when survey managers evaluate the current response rates of their business surveys by comparing them to the previous rates, with the goal of at least maintaining a level of response, important patterns in the response process are missed. Therefore, we advocate that survey managers instead analyze the rates collectively, over time within a statistical process control framework to assess the stability and capability of the response process; specifically, using control charts. For response data, a *p*-chart is appropriate. Developing a *p*-chart for response data provides challenges generally not found in a manufacturing setting, where control charts were developed and first used.

In a manufacturing setting, a *p*-chart is developed from thousands of measurements, collected in a relatively short time. In a monthly or quarterly sample survey, the number of separate observations is limited, and there is a longer lag between each measure. To maintain the "spirit" of the analysis, we must balance the simultaneous objectives of avoiding a false conclusion of a process change while quickly reflecting actual process changes in the *p*-chart limits so that the program managers can assess the process and make timely interventions if necessary. To reduce the probability of a false conclusion of a process change, we provide a modified set of *p*-chart analysis guidelines from those advocated in Tague (2004). Our *p*-charts use a rolling median of consecutive measures to estimate the centerline and as input to the control limits, instead of the mean, thus limiting the impact of a single anomalous measure and allowing a short burn-in period for the charts to reflect a true process change.

In a manufacturing setting, a single outlier that has an assignable cause is usually discarded from the control chart parameter estimation and thus does not affect the process average. Using the median with a relatively small number of measures in effect accomplishes the same thing. With the unit and weighted volume response rates, we retain the assumption of binomially distributed data when estimating the standard errors for the three sigma limits; with the total quantity response rates, we employ measures that incorporate the complex survey design and widen the limits from three standard deviations to five. With the first of our case studies (MRTS), this proved to be quite effective. With the second (MWTS), there was some discussion on whether the number of standard deviations should be reduced because of the survey's smaller sample. This choice is a good topic for future research as more empirical results become available.

As a rule of thumb, it is not reasonable to develop one single measure of response that represents quality. This is especially evident with business surveys with skewed populations. In this context, the URR tends to overemphasize the "importance" of the small units in the sample. In most business surveys, the larger units receive a disproportionate portion of the nonresponse follow-up, and the *p*-charts that present the program-level URR discounts that effort (Thompson and Oliver, 2012). This disadvantage can be offset by producing domain level time-series plots as in the MWTS. Using a measure that combines survey weights with a measure-of-size variable such as the WVRR or the TQRR may understate systematic response process issues with small sampled units by overemphasizing the contribution to totals from the large units. Moreover, the WVRR is useful if the frame measure-of-size is positively correlated with all key items. Lastly, although the TQRR is likely the most direct measure of quality (in terms of non-imputed data in the tabulations), it is also the most difficult measure to adapt to a statistical process control framework, as the statistic itself is not binomially distributed and the small number of observations in a time series render the normal theory assumptions somewhat tenuous. Moreover, the tabulation unit response process could vary greatly by item.

Despite these caveats, studying these process measures with control charts provides valuable insights. In one case study, we demonstrated detrimental effects on quality of sample attrition, explicitly linking the decreasing URR to a decreased TQRR for a key item. Equally important, we demonstrated that these processes are truly out of control, and provided survey managers with evidence that expected periodic corrections are not sufficient to bring the process into control. In another case study, we demonstrated that the response rate processes are stable, but low, and that the process capability will not

change without an intervention targeted at a low-responding subpopulation (Wagner, 2012).

Acknowledgements

We thank William Abriatis, Xijian Liu, Jennifer Reichert, and Ian Thomas for the careful review of earlier versions of the manuscript.

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.* American Statistical Association.

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

González, Y. and Oliver, B. (2012). Producing Control Charts to Monitor Response Rates for Business Surveys in the Economic Directorate of the U.S. Census Bureau. *Proceedings of the Federal Committee on Statistical Methodology Research Conference*.

González, Y., Oliver, B., and Thompson, K.J. (2013). Producing Control Charts to Monitor Response Rates for Selected Business Surveys of the U.S. Census Bureau. *Proceeding of the Joint Statistical Meetings*. American Statistical Association.

Methodology and Statistics Council (2012). *U.S. Census Bureau Statistical Standards.* Washington, DC: U.S. Census Bureau.

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, pp. 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, pp. 193-201.

Tague, N.R. (2004). *The Quality Toolbox (Second Edition).* Milwaukee, WI: ASQ Quality Press.

Thompson, K.J. and Oliver, B. (2012). Response Rates in Business Surveys: Going Beyond the Usual Performance Measures. *Journal of Official Statistics*, 28, pp. 221-237.

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

Wagner, J. (2012). Research Synthesis: A Comparison of Alternative Indicators for the Risk of Nonresponse Bias. *Public Opinion Quarterly*, 76 (3), pp. 555-575.

Workbook on Leadership. (2003). Six Sigma Qualtec Workshop, June 26, 2003, Washington, DC.