

Statistical Process Control of Sampling Frames

A.W. SPISAK¹

ABSTRACT

Statistical process control can be used as a quality tool to assure the accuracy of sampling frames that are constructed periodically. Sampling frame sizes are plotted in a control chart to detect special causes of variation. Procedures to identify the appropriate time series (ARIMA) model for serially correlated observations are described. Applications of time series analysis to the construction of control charts are discussed. Data from the United States Department of Labor's Unemployment Insurance Benefits Quality Control Program is used to illustrate the technique.

KEY WORDS: Autocorrelation; ARIMA models; Control charts; Quality assurance.

1. INTRODUCTION

The integrity of the sampling frame is of paramount importance in survey research. Frame imperfections include missing elements (incomplete frame), element clusters (more than one element in a single listing), blank or foreign elements, and duplicate listings. These imperfections can cause several difficulties by contributing to nonsampling error, reducing the number of sample cases from subclasses of the population, and requiring the use of complex weights to estimate population characteristics. Techniques to minimize frame problems or reduce their impact on the survey are discussed in detail in most textbooks on statistical surveys.

This article focuses on the statistical process control of sampling frames which are constructed periodically (daily, weekly, or monthly, for example) and which consist of elements that are generated by a continuous process. Because of the variation inherent to any dynamic process, the sizes of the sampling frames will vary. How do we know that the changes in the sizes of the sampling frames reflect the random variation of the process and not errors in the construction of the frames? Statistical process control allows survey managers to distinguish between the variation inherent in the process (common causes) and variation which signals a possible problem with frame construction (special causes).

2. PROCESS VARIATION AND STATISTICAL PROCESS CONTROL

Over the last several years managers in the manufacturing, service, and public sectors of the economy increasingly have adopted the quality philosophies developed by W. Edwards Deming, J.M. Juran, Philip B. Crosby, Kaoru Ishikawa, and others. Quality management comprises an

array of tools and techniques, including the use of control charts to determine if a process is in statistical control. According to Deming (1982), statistical control is achieved by eliminating special causes of variation, leaving only the random variation of a stable process. The behavior of a process that is in statistical control is predictable.

The distinction between common and special causes of variation is a key principle of statistical process control. Deming (1982) credits Dr. Walter A. Shewhart, who developed many of the principles of statistical process control in the 1920s and 1930s, with originating the concept of special or assignable causes. Special causes are usually attributable to one part of the process, such as a worker, machine, or office. They will reoccur unless they are identified and eliminated. Special causes are signaled by data points that fall outside of the control limits, by consecutive points that fall above or below the process average, or by runs of increasing or decreasing points.

Common causes of variation are inherent to the process; they are present at all times and effect the entire process. Common causes are reduced or eliminated through management actions that change the process.

3. STATISTICAL PROCESS CONTROL APPLICATION TO THE CONSTRUCTION OF SAMPLING FRAMES FOR PERIODIC SURVEYS

3.1 United States Unemployment Insurance Benefits Quality Control

The use of statistical process control as a quality management tool for sampling frames is illustrated by an example from the United States Department of Labor's Unemployment Insurance Benefits Quality Control program. Since 1987, the 50 states, the District of Columbia, and

¹ A.W. Spisak, Mathematical Statistician, Unemployment Insurance Service, U.S. Department of Labor, Washington, DC 20210, U.S.A.

Puerto Rico have conducted the Benefits Quality Control program in cooperation with the United States Department of Labor. The goal of the program is to reduce the overpayment and underpayment of Unemployment Insurance benefits by identifying the causes of payment errors and initiating measures to improve the benefit payment process.

When an individual files a claim for Unemployment Insurance benefits, Unemployment Insurance staff determine whether the claimant has met all of the eligibility requirements – for example, the claimant earned sufficient wages in his or her previous employment to qualify for benefits; the claimant is involuntarily unemployed; and the claimant is able and available to work and is actively seeking employment. If all of the eligibility requirements are satisfied, the state Unemployment Insurance agency issues a benefits check for the week of unemployment claimed.

3.2 Benefits Quality Control Sampling Procedures and Sources of Error

Each state selects weekly random samples of Unemployment Insurance payments that are examined to determine if the correct amount was paid to the claimant. If the amount paid was incorrect, the investigator identifies the types and causes of the errors so that program managers can initiate corrective measures. The sampling frames are constructed each week from the universe of Unemployment Insurance payments that were issued between 12:00 am Sunday and 11:59 pm the following Saturday. A computer program edits the state's database to insure that only payments that meet the program's operational definition of the target population are included in the frame. For example, payments for some temporary or small Unemployment Insurance programs are excluded from the frame.

The volume of Unemployment Insurance checks issued each week (and therefore the size of the sampling frames) varies in response to the number of individuals who claim and receive benefits during that week. However, there are several sources of potential errors which can affect the integrity of the frame. Some of the most serious of these errors are:

- The payments made from some of the local Unemployment Insurance offices might not be picked up for inclusion in the state's central database, due to telecommunication or ADP problems.
- If the state builds a separate file for each day's transactions, the transactions for one or more days might be erroneously omitted from the final cumulative file.
- Incorrect coding of transactions could result in either foreign elements being included in the frame or the editing out of transactions that should be included.

4. DATA ANALYSIS AND MODEL DEVELOPMENT

Figure 1 is a time series plot of sampling frame sizes for a 52 week period. Each week's sampling frame consists of the previous week's Unemployment Insurance benefit recipients who continue to receive benefits, minus the previous week's Unemployment Insurance recipients who have returned to work, exhausted their benefits, or failed to file a claim, plus newly eligible claimants and eligible claimants who did not file a claim or were not compensated for a claim the previous week.

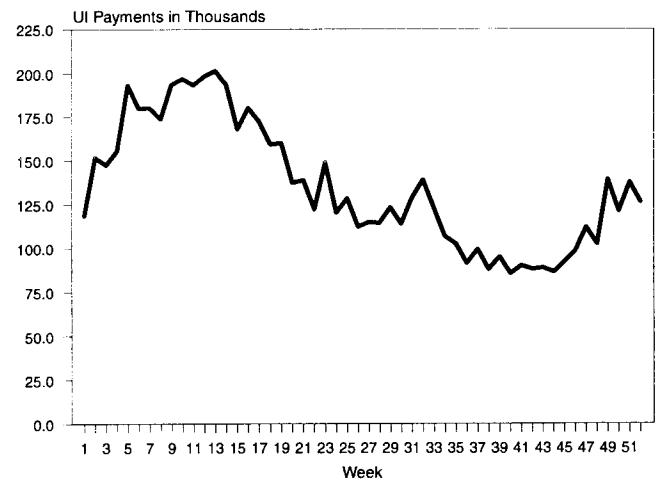


Figure 1. Number of UI payments per week.

Control charts for individual observations assume that the data are independent and identically distributed (i.i.d.). However, if the data are serially correlated, the estimates of the process variance (and therefore the control limits) could be seriously in error. So, before control charts for the Unemployment Insurance sampling frame data can be constructed, we have to determine if the observations are serially correlated.

The plot of the time series in Figure 1 provides *visual* evidence that the observations are not independent. The sampling frame data display distinct trends of increasing values during the first 13-week quarter, decreasing values over the next two quarters, and increasing values during the final 13-week quarter. The serial correlation suggested by the plot of the data in Figure 1 can be tested using methods developed to analyze time series. Although a detailed discussion of the analysis of time series data is beyond the scope of this article, the concepts of stationarity and autocorrelation will be examined, in order to explain the procedures used to identify the appropriate model. Readers who are unfamiliar with the basic principles of time series analysis should consult one of the many texts on the subject, in particular Box and Jenkins (1976).

4.1 Stationarity

We can think of the individual observations that constitute a time series as a collection of jointly distributed random variables – $p(z_1, \dots, z_n)$ – where p is a probability density function and z_1, \dots, z_n are random variables. If the joint distribution of the random variables does not vary with respect to time, that is, $p(z_t, \dots, z_{t+n}) = p(z_{t+m}, \dots, z_{t+n+m})$, the process is said to be *strictly* stationary. In practice strict stationarity is difficult to establish. In this application, the time series is assumed to be *weakly* stationary. This is also referred to as second-order stationarity, because the first and second moments of the process are invariant with respect to time – $E(z_t) = E(z_{t+m})$, $\text{VAR}(z_t) = \text{VAR}(z_{t+m})$, and $\text{COV}(z_t, z_{t+k}) = \text{COV}(z_{t+m}, z_{t+k+m})$.

Throughout the rest of this article, the terms *stationary* or *stationarity* refer to a process that satisfies the conditions of weak stationarity.

4.2 Autocorrelation

In a stationary time series the covariance between any two observations depends only on the number of time periods (lags) that separate them – $\text{COV}(z_t, z_{t+k}) = \text{COV}(z_{t+m}, z_{t+k+m})$. The correlation of z_t and z_{t+k} equals $\text{COV}(z_t, z_{t+k}) / \text{VAR}(z_t)$ and is denoted ρ_k , where k is the number of periods between observations. For example, ρ_1 is the correlation of observations in the time series separated by one period and equals $\text{COV}(z_t, z_{t+1}) / \text{VAR}(z_t)$. A correlation for period k is referred to as an autocorrelation, because it is the correlation for observations which constitute a time series. The autocorrelations for the various lags can be displayed in a graph called a correlogram, which is useful in identifying the appropriate model for a time series.

4.3 Time Series Model Identification

Figure 2 is the correlogram for the 52 week time series of the number of Unemployment Insurance payments in the sample frames. The autocorrelations decrease or “die out” very slowly, which is characteristic of a nonstationary process. (Again, the reader is referred to Box (1976) and other texts on time series for a complete discussion of model identification.)

One method to transform a nonstationary series to a stationary series is *differencing*. The symbol B is the backshift operator, which when applied to z_t shifts the subscript back one period. Thus, the first difference of z_t is $(1 - B)z_t = z_t - z_{t-1}$.

Figure 3 is the time series of the differences $z_t - z_{t-1}$ of the Unemployment Insurance sampling frame data. This series appears stationary around a mean of zero. (The estimated sample mean of the differences is 150.8, with a standard error of 2064.0. The test statistic $t = (150.8 - 0) / 2064$ equals .07, and the hypothesis that $\mu = 0$ cannot be

rejected). First differences might not be sufficient to achieve stationarity for other time series, and transformations such as second differences – $(1 - B)^2 z_t = (z_t - z_{t-1}) - (z_{t-1} - z_{t-2})$, seasonal differences, or logarithmic or other variance stabilizing procedures may be required.

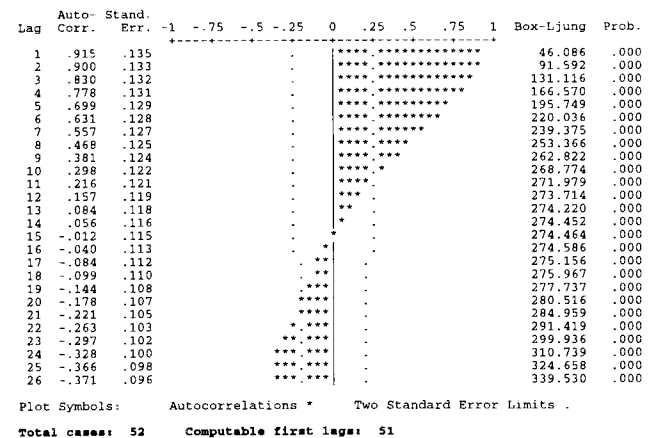


Figure 2. Autocorrelations for UI weeks paid time series.

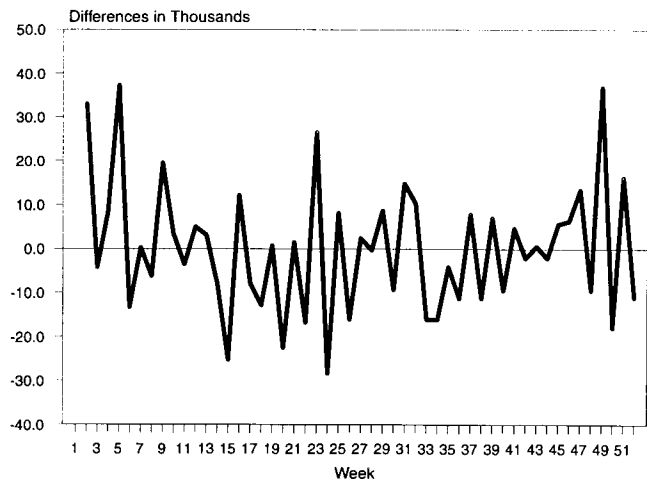


Figure 3. First differences of UI payments.

The autocorrelations of the first differences of the time series, which are displayed in Figure 4, are consistent with a stationary process. The autocorrelations decrease rapidly, while the partial autocorrelations (not displayed) die off after lag 1. This suggests that the data can be modelled with a first-order integrated autoregressive process, $\text{ARI}(1,1)$. The AR term indicates that a single autoregressive parameter will be estimated, and the integration term (I) shows that the original time series has been transformed using first differences.

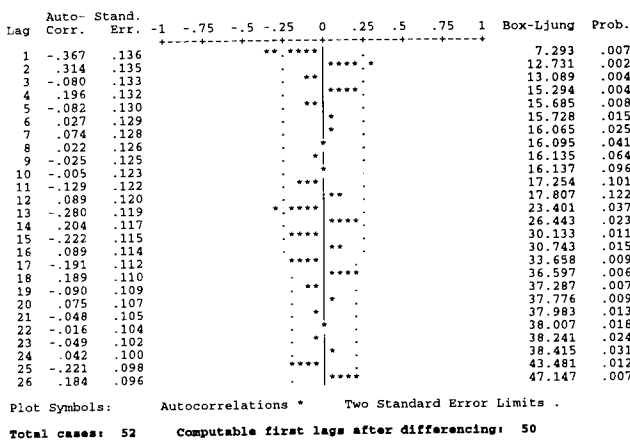


Figure 4. Autocorrelations for first differences of UI weeks paid.

4.4 Model Estimation

The model was estimated using the ARIMA procedure of the SPSS Trends software (release 4.0), which is based on the work of Box and Jenkins.

The tentative model is:

$$z_t = (1 + \phi_1)z_{t-1} - \phi_1 z_{t-2} + e_t, \text{ or}$$

$$z_t - z_{t-1} = \phi_1(z_{t-1} - z_{t-2}) + e_t,$$

where ϕ_1 is the first-order autoregressive parameter, and e is the error term, which is assumed to be normally distributed with a mean of 0 and variance σ_e^2 . The estimated autoregressive parameter, ϕ'_1 is $-.4045$, and the estimated residual variance, $\sigma_e'^2$, is 184,275,853 (with 50 degrees of freedom). The negative sign on the AR parameter is consistent with the alternating signs of the autocorrelations in Figure 4. The model does not include a constant term, because the estimated process mean was not significantly different than zero.

4.5 Model Diagnostics

The adequacy of the estimated model for the observed data can be assessed by examining the model residuals. If the model adequately fits the data, the residuals (e_t) should be "white noise", that is, uncorrelated. Figure 5 displays the autocorrelations of the model residuals. Although the autocorrelation at lag 13 in Figure 5 is significant, the Box-Ljung Q statistic through lag 13 is not significant. (The Q statistic tests the significance of autocorrelations for lags 1 through k . For a detailed discussion, see Box and Pierce (1970)). In addition, none of the partial autocorrelations (not displayed) are significant. These results indicate that the residuals are not serially correlated.

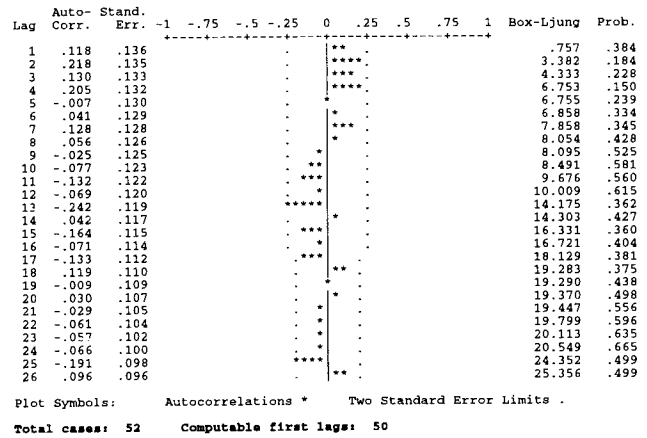


Figure 5. Autocorrelations for time series model residuals.

To test the assumption that the model residuals are normally distributed, $N(0, \sigma_e^2)$, a Kolmogorov-Smirnov ($K-S$) goodness of fit test was conducted. For the estimated variance of 184,275,853, the $K-S$ test statistic equals .591 ($p = .876$), and the hypothesis that the differences are normally distributed cannot be rejected.

For a stationary AR (1) process, the absolute value of the autoregressive parameter must be less than one. To test the hypothesis that $|\phi_1| \geq 1$ for the model, we compute: $t = (|\phi'_1| - 1)/SE(\phi'_1)$, where $|\phi'_1|$ is the absolute value of the estimated autoregressive parameter, and $SE(\phi'_1)$ is the standard error of ϕ'_1 . The model statistics result in $t = (.4045 - 1)/.1295$ or $t = -4.6$. The chance of observing an absolute value of ϕ'_1 as small as .4045 if the true absolute value of $\phi_1 \geq 1$ is very small ($< .00001$). The hypothesis that $|\phi_1| \geq 1$ is rejected, and we can conclude that the series of first differences is stationary.

5. USE OF THE ARIMA MODEL IN A CONTROL CHART

5.1 Control Charts for Individual Observations

The control limits for a chart of individual observations are set at $\bar{x} \pm 3\sigma'$, where \bar{x} is the average of observation values and σ' is the estimated standard deviation of the process. Ryan (1989) discusses alternative procedures to estimate the process standard deviation either by computing the average of the moving ranges (the mean of the absolute differences of successive observations) or using the standard deviation (s) of the sample observations, $\sigma' = s/c$, where c is an adjustment constant which depends on the sample size.

When data are serially correlated, the use of either the sample standard deviation or the average moving range can result in poor estimates of σ . The control limits constructed from these estimates can produce seriously

misleading results by either generating false signals that the process is out of control or failing to detect special causes of process variation. The moving range can underestimate σ , because the differences of successive values will tend to be small if the successive observations are highly correlated. The underestimation of σ will result in control limits that are too narrow and an increase in the number of signals of special causes. Ryan notes that using the sample standard deviation to estimate the process standard deviation will result in a better estimate of σ than the average moving range when the data are correlated, provided the sample consists of at least 50 observations. However, the sample standard deviation is an unbiased estimator of σ only when the observations are independent.

Vasilopoulos and Stamboulis (1978) analyzed the effect of serially correlated data on the control limits of \bar{x} and s (standard deviation) charts and developed equations for factors that can be used to adjust the control limits for data generated by an autoregressive process. Alternatively, a time series model can be identified for the correlated data, and a control chart can be constructed using the model residuals to monitor the process. This approach is described by Berthouex, Hunter, and Pallesen (1978) for subgroups of measurements of environmental data collected at water treatment plants. Alwan and Roberts (1988) use the residuals of exponentially weighted moving average (EWMA) models for both stationary and nonstationary time series. Montgomery and Mastrangelo (1991) use the residuals of an autoregressive model in an EWMA chart and contend that EWMA charts can be used to approximate many autocorrelated models, particularly if the observations are positively correlated and the mean does not drift too quickly. The reader is also referred to Maragah and Woodall (1992) and Woodall and Faltin (1993) for additional discussion of the effects of autocorrelation on statistical process control procedures.

5.2 Control Charts for the Unemployment Insurance Data

Figure 6 is a control chart of the residuals ($e_t = z_t - z'_t$) of the ARI (1,1) model identified for the Unemployment Insurance sampling frame data. Since the model diagnostics support the conclusion that the residuals are independent and identically distributed (i.i.d.) $N(0, \sigma_e^2)$, the residuals are standardized, so that the chart's center line is 0 and the control limits are set at ± 3 . The chart includes model residuals for the sampling frame sizes in the 52 week baseline period and subsequent calendar quarter. The difference between the size of the sampling frame for week 56 and the value predicted by the model falls outside the upper control limit, signaling a special cause.

As an alternative to charting the model residuals, control charts for the Unemployment Insurance sampling frame

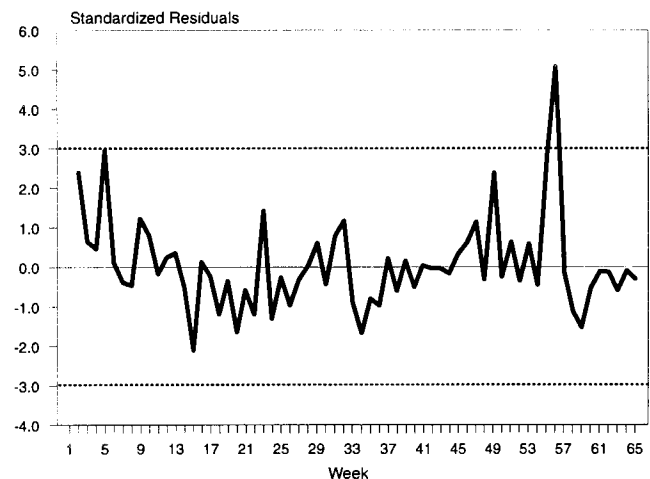


Figure 6. Control chart for model residuals (baseline data + next quarter).

sizes can be constructed. The original observations must be transformed to achieve stationarity, if necessary. The estimated parameters of the time series model are used to construct the mean and control limits of the chart. The variance of an AR(1) process is $\sigma^2 = \sigma_e^2 / (1 - \phi_1^2)$. For the time series model of first differences, ϕ_1' is $-.4045$, and the estimated residual variance, $\sigma_e'^2$, is $184,275,853$. The estimated process variance is $184,275,853 / (1 - .1636)$ or $220,325,579.4$, and the process standard deviation is $14,843.4$. The upper and lower control limits are set at $\pm 3\sigma'$ from the estimated mean difference of zero: $\pm 44,530.2$. The control chart is shown in Figure 7 and signals a special cause for observation 56, like the control chart for the residuals in Figure 6.

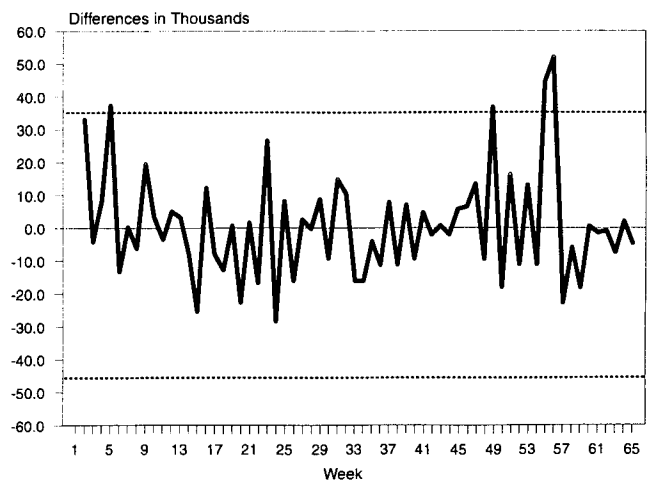


Figure 7. Control chart for UI payments (first differences - baseline + next quarter).

6. CONCLUSIONS

Statistical process control is a useful quality assurance tool for surveys in which samples are selected from frames that are constructed for specified periods from a continuous process. Because the frame sizes constitute a time series, the data may be serially correlated and may have to be transformed in order to achieve stationarity. If the observations are correlated, the appropriate time series (ARIMA) model must be identified in order to estimate the process variance used in setting the control limits. The time series in the preceding example was fitted by a first-order autoregressive integrated (differenced) model – ARI (1,1). More generally, time series may be described by other ARIMA (p, d, q) models, where p is the number of autoregressive terms in the model, d is the degree of differencing to achieve stationarity, and q is the number of moving average terms in the model. Seasonal time series models include additional AR, MA, and differencing parameters for the appropriate lag(s).

Once the model has been identified from baseline data, observations from subsequent periods can be plotted in the control chart. In the control charts in Figures 6 and 7, one calendar quarter (13 weeks) of observations are plotted following the observations from the 52 week baseline. The time series model should be checked periodically, depending on the data collection interval, to determine if the model parameters have changed.

If the statistical process control procedures signal a special cause of variation, survey managers must use other quality management tools to determine the root causes of the frame problems and then implement corrective actions to improve survey procedures. Survey managers can move from troubleshooting and error correction to continuous improvement of the survey process by systematically removing the assignable causes of variation identified through statistical process control.

In the case of the Unemployment Insurance sampling frame data, the special cause was not preventable: the volume of Unemployment Insurance payments spiked during a week which followed a short work week due to a holiday and which coincided with a layoff at a large establishment. The large sampling frame was not the result of a technical problem with the construction of the frame. In other states, at different time periods, statistical process control has detected errors as diverse as data entry mistakes (a frame of 558,432 reported instead of 5,558,432), omission of the Unemployment Insurance transactions for one of five work days, resulting in an approximate 20 percent decrease in the frame size, and the failure to

update edits in the sample selection software, which caused 'foreign elements to enter the frame.

The procedure described in this article is applicable to other areas of survey and information management in addition to the integrity of sampling frames. The procedure can be used to reduce nonsampling error attributable to data recording or data entry for surveys conducted daily, monthly, *etc.* More generally, statistical process control can be used to assure the integrity of databases or management information systems whenever information is collected or reported in subgroups, such as data collected at multiple sites or by several researchers or auditors.

ACKNOWLEDGEMENT

The author wishes to thank the reviewers for their helpful comments and suggestions.

REFERENCES

- ALWAN, L.C., and ROBERTS, H.V. (1988). Time series modeling for statistical process control. *Journal of Business and Economic Statistics*, 6, 87-95.
- BERTHOUEX, P.M., HUNTER, W.G., and PALLESEN, L. (1978). Monitoring sewage treatment plants: some quality control aspects. *Journal of Quality Technology*, 10, 139-149.
- BOX, G.E.P., and PIERCE, D.A. (1970). Distribution of residual autocorrelations in autoregressive moving average time series models. *Journal of the American Statistical Association*, 65, 1509-1526.
- BOX, G.E.P., and JENKINS, G.M. (1976). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden-Day.
- DEMING, W.E. (1982). *Quality, Productivity, and Competitive Position*. Cambridge: Massachusetts Institute of Technology Center for Advanced Engineering Study.
- MARAGAH, H.D., and WOODALL, W.H. (1992). The effect of autocorrelation on the retrospective \bar{X} -chart. *Journal of Statistical Computation and Simulation*, 40, 29-42.
- MONTGOMERY, D.C., and MASTRANGELO C.M. (1991). Some statistical process control methods for autocorrelated data. *Journal of Quality Technology*, 23, 179-204.
- RYAN, T.P. (1989). *Statistical Methods for Quality Improvement*, New York: John Wiley and Sons.
- VASILOPOULOS, A.V., and STAMBOULIS, A.P. (1978). Modification of control chart limits in the presence of data correlation. *Journal of Quality Technology*, 10, 20-30.
- WOODALL, W.H., and FALTIN, F.W. (1993). Autocorrelated data and SPC. *American Society for Quality Control Statistics Division Newsletter*, 13, 18-21.