It's All About Variation: Improving Your Business Process with Statistical Thinking — Revised 2012

Robert N. Rodriguez
SAS Institute Inc., Cary, NC

ABSTRACT

This paper explains how statistical thinking and statistical process monitoring, which have been practiced in manufacturing for the past thirty years, are proving valuable for process improvement in business environments that range from health care to financial services. Basic examples drawn from real scenarios introduce the statistical concepts and show how to get started with SAS/QC® software. The concepts also apply to complex systems that involve large volumes of multivariate data with multiple sources of variation. The examples demonstrate the use of graphical displays, created with ODS Statistical Graphics, for visualizing and analyzing the variation in a process and for explaining results to clients and management.

THE AMERICAN QUALITY REVOLUTION

Thirty years ago, an NBC television documentary ignited a revolution in American industry with a provocative title, "If Japan Can ... Why Can't We?" The documentary explained how Japanese manufacturers in the automotive and electronics industries had overtaken their American competitors by using a management system that emphasized continuous improvement of quality. Viewers were surprised to learn that it was an American statistician, Dr. W. Edwards Deming, who had taught Japanese companies to apply statistical thinking and statistical methods as the foundation for a systematic approach to manufacturing. In an interview during the documentary, Dr. Deming explained, "Statistical thinking and statistical methods are to Japanese production workers, foremen, and all the way through the company, a second language. In statistical control you have a reproducible product hour after hour, day after day. And see how comforting that is to management: they now know what they can produce, they know what their costs are going to be."

The documentary launched Dr. Deming into national prominence as a widely sought-after management consultant and speaker. At his seminars, the message he preached was that top management—not the work force—is directly responsible for 85% of all problems, and he presented 14 principles for management, which he later expanded in his book *Out of the Crisis*. Deming was once asked whether he could reduce his 14 principles to a single sentence, and his response was, "It's all about understanding variation."

Continuous improvement requires that management measure, understand, and act upon the variability in business processes. "Statistical thinking"—a term that Deming often used—starts with the recognition that all processes are subject to variability and that improvement comes about through understanding and reducing variability. Wheeler and Polling (1998) explain that "instead of focusing on outcomes, such as expenses and profits, this better way focuses on the processes and systems that generate the outcomes. Rather than trying to directly manipulate the results, it works to improve the system that causes the results. Rather than distorting . . . the data, it seeks to use the data to understand the system as a basis for improving the system." More recently, Hoerl and Snee (2002) have defined statistical thinking as the integration of process thinking, understanding of variation, and data-based decision making.

Figure 1 Beads Used by Deming to Demonstrate Variability
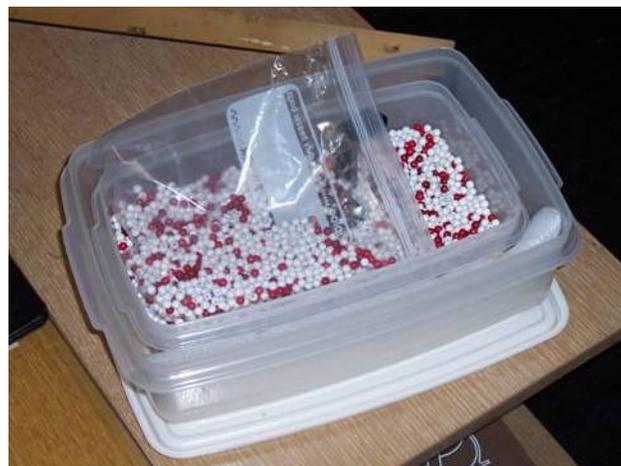
UNDERSTANDING AND ACTING ON PROCESS VARIATION

Although Deming did not teach statistical methods in his seminars, his management principles were based on the work of Walter A. Shewhart (1891–1967), a pioneer in the field of industrial statistics, who made fundamental contributions to the understanding of variation in manufacturing processes. Shewhart recognized that every process displays variation, and he distinguished two types of variation: chance cause variation, which is naturally present in all processes, and assignable cause variation, which is sporadic and not present at all times. By eliminating assignable cause variation, the process can be brought into a state of statistical control, which provides the stability that is essential for predicting future output and assessing improvements. In 1924, Shewhart introduced the control chart as a statistical technique for deciding when assignable causes are present.

In order to emphasize the levels of accountability for acting on the two types of variation, Deming renamed chance cause variation as common cause variation and assignable cause variation as special cause variation. Deming stressed that it is the responsibility of top management to address common cause variation because only at this level can the entire system be improved. On the other hand, workers, supervisors, and middle managers have direct knowledge of special causes of variation and are best equipped to fix these problems.

Interest in Deming's management philosophy declined following his death in 1993, but his emphasis on statistical process control (SPC) has had a lasting influence on manufacturing practice. This is especially the case in the automotive, electronics, and semiconductor industries, where control charts and process capability analysis are standard tools. At some companies the Six Sigma movement, which began in the 1980s, has succeeded in spreading the implementation of SPC and related methods through a systematic approach to quality management which relies on executive sponsorship, training, consulting, and teamwork.

IS THE QUALITY REVOLUTION OVER?

In recent years, the Six Sigma approach has been criticized for promoting process improvement at the expense of innovation and productivity. The cover article of the June 6, 2007, issue of *BusinessWeek* faulted the introduction of Six Sigma at 3M for disrupting the company's culture of innovation (Hindo 2007). While the article clearly acknowledged the role of statistical analysis in helping to "produce better quality, lower costs, and more efficiency," it questioned the relevance of the Six Sigma approach in the current economy, where new ideas and designs, rather than quality, are driving competition. The article concluded that "while process excellence demands precision, consistency, and repetition, innovation calls for variation, failure, and serendipity."

Has the quality revolution come to an end? Recent headlines about safety recalls by a Japanese automotive manufacturer demonstrate once again that quality and reliability require continuous improvement of internal processes, and that the responsibility for leading this effort lies with top management (Maynard 2010). In many industrial settings, the need for basic statistical thinking and SPC is growing, and more advanced statistical techniques are being developed to deal with increasingly complex manufacturing processes (Ramirez and Tobias 2007), massive volumes of process data that are collected automatically, and new regulatory requirements (Peterson et al. 2009).

At the same time, the power of statistical thinking and statistical process improvement is being demonstrated in business environments that range from health care to financial services. This paper discusses several of these applications using examples from actual scenarios.

HEALTH CARE: A NEW FRONTIER FOR QUALITY IMPROVEMENT

Improving the quality of patient care is a major component of health care reform in the United States. Hospitals and other health care providers face the challenges of retaining qualified staff and containing costs. Institute of Medicine studies show that over half of medical deaths in hospitals are preventable, and statewide data reveal variability in hospital quality. In 2000, a landmark publication titled *To Err Is Human* (Institute of Medicine 2000) exposed the impact of hospital errors. A follow-up report, *Crossing the Quality Chasm* (Institute of Medicine 2001), listed six areas in which health care systems perform at low levels:

- avoiding unintended injuries to patients

- providing evidence-based services where needed

- ensuring that patients' values are respected

- reducing harmful delays that affect both patients and providers

- avoiding waste of materials and time

- providing a consistent level of care to all patients

In order to improve performance in these areas, many hospitals have turned to the quality management philosophies of W. Edwards Deming, Joseph Juran, and others.

Although the health care industry generates large amounts of patient-specific data, few organizations are able to use these data to identify unusual variability in staff and physician performance, cost of care, and preventable incidents that affect the outcome of a patient's care. SAS® provides many solutions for accessing multiple data sources and data management. Statistical process control can then be used to identify variability due to special causes and focus further study to reduce variability. These techniques lead to improvements in quality of care, reduction of costs, opportunities to grow market share, and negotiation of better third-party payment.

In their 1990 book *Curing Health Care*, Berwick, Godfrey, and Roessner illustrated the basic tools of quality improvement and especially the use of SPC for improving hospital processes. Since then these applications have grown rapidly. Although an overview of the literature is outside the scope of this paper, several publications are especially relevant to the topics considered here. Benneyan (2001a, b) introduced specialized control charts for monitoring adverse effects. Woodall (2006) surveyed the use of control charts in health care. The 20th anniversary issue of *Quality Engineering* is dedicated to statistical quality control in health care: some of the articles in this issue focus on control chart methods (Limaye, Mastrangelo, and Zerr 2008) and other articles cover topics in syndromic surveillance (Tsui et al. 2008).

Rodriguez and Ransdell (2010) provide examples that explain the use of SAS/QC software to analyze health care data with a variety of SPC methods. The two examples that follow are borrowed from this paper and illustrate how to construct a basic control chart.

Example 1: Basic u Chart for Rate of CAT Scans

This example introduces the use of the SHEWHART procedure in SAS/QC software to construct a u chart, which is one of several standard control charts for count data. In manufacturing, u charts are typically used to analyze the number of defects per inspection unit in samples that contain arbitrary numbers of units. In general, the event that is counted need not be a defect. A u chart is applicable when the counts can be scaled by some measure of opportunity for the event to occur, and when the counts can be modeled statistically by the Poisson distribution. The SHEWHART syntax for this example is described in detail since it extends to other charts that can be constructed with the procedure, as indicated in subsequent examples.

A health care provider uses a u chart to analyze the rate of CAT scans performed each month by each of its clinics. Figure 2 shows data collected for Clinic B and saved in a SAS data set named ClinicB.

Figure 2 SAS Data Set ClinicB

```
                    CAT Scan Data for Clinic B

        month     nscanb      mmsb      days      nyrsb

        JAN04        50       26838       31      2.31105
        FEB04        44       26903       28      2.09246
        MAR04        71       26895       31      2.31596
        APR04        53       26289       30      2.19075
        MAY04        53       26149       31      2.25172
        JUN04        40       26185       30      2.18208
        JUL04        41       26142       31      2.25112
        AUG04        57       26092       31      2.24681
        SEP04        49       25958       30      2.16317
        OCT04        63       25957       31      2.23519
        NOV04        64       25920       30      2.16000
        DEC04        62       25907       31      2.23088
        JAN05        67       26754       31      2.30382
        FEB05        58       26696       28      2.07636
        MAR05        89       26565       31      2.28754
```

The variable nscanb is the number of CAT scans performed each month, and the variable mmsb is the number of members enrolled each month (in units of "member months"). The variable days is the number of days in each month. The following SAS statements compute the variable nyrsb, which converts mmsb to units of "thousand members per year":

```
data ClinicB; set ClinicB;
   nyrsb = mmsb * ( days / 30 ) / 12000;
run;
```

Note that nyrsb provides the measure of opportunity.

The following statements create the u chart in Figure 3:

```
ods graphics on;
ods listing style=statistical;

title 'U Chart for CAT Scans per 1,000 Members: Clinic B';
proc shewhart data=ClinicB;
   uchart nscanb * month / subgroupn = nyrsb  tests = 1 to 4  nohlabel
                           testnmethod = standardize nolegend
                           odstitle = title;
   label nscanb = 'Rate per 1,000 Member-Years';
run;
```

The ODS GRAPHICS ON statement enables ODS Statistical Graphics, causing SAS/QC procedures to produce ODS Graphics output instead of traditional graphics. This statement remains in effect for subsequent examples in this paper. (You can disable ODS Graphics with the ODS GRAPHICS OFF statement.) For an introduction to ODS Graphics, refer to Rodriguez (2011), and refer to Chapter 3, "SAS/QC Graphics" (*SAS/QC User's Guide*), to learn more about the use of ODS graphics in SAS/QC software. The STYLE= option in the ODS LISTING statement specifies that the graphs be produced with the STATISTICAL style.

The PROC SHEWHART statement invokes the SHEWHART procedure, and the DATA= option specifies the input data set. The UCHART statement requests a u chart. After the keyword UCHART, you specify the count variable (in this case, nscanb), followed by an asterisk and the variable that identifies the sample (in this case, month).

You can specify options for analysis and graphical presentation after the slash (/) in the UCHART statement. Refer to the *SAS/QC User's Guide* for details on syntax and statistical methods. The SUBGROUPN= option specifies the number of "opportunity" units per sample. You can use this option to specify a fixed number of units or (as in this case) a variable that provides the number of units for each sample. The TESTS= option requests tests for special causes, also referred to as runs tests, pattern tests, and Western Electric rules. For example, Test 1 flags points outside of the control limits. The TESTNMETHOD=STANDARDIZE option applies a standardization method to adjust for the fact that the number of units varies from sample to sample.

The NOHLABEL option suppresses the label for the horizontal axis (which is unnecessary since month has a datetime format), and the NOLEGEND option suppresses the default sample size legend. The LABEL statement assigns a label to the variable nscanb, which is displayed on the vertical axis. The ODSTITLE= option incorporates the title specified in the TITLE statement into the graph.

Figure 3 Basic u Chart
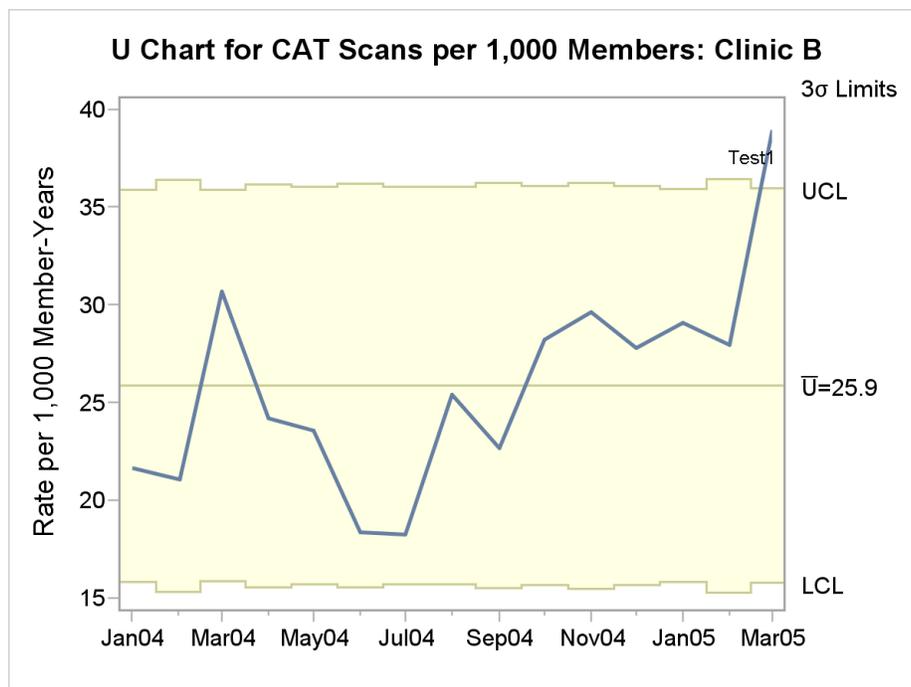
In Figure 3, the upper and lower control limits are 3σ limits estimated by default from the data; the limits vary because the number of opportunity units changes from month to month. The increase in the rate of CAT scans for March 2004 is interpreted as common cause variation because it lies within the control limits, whereas the increase for March 2005 should be investigated.

You can use the SHEWHART procedure to create a wide variety of control charts. Each of the standard chart types is created with a different chart statement (for example, you use the PCHART statement to create p charts). Once you have learned the basic syntax for a particular chart statement, you can use the same syntax for all the other chart statements.

Example 2: Control Limits for a u Chart with a Known Shift in Rate

This example illustrates the construction of a u chart in situations where the process rate is known to have shifted, requiring multiple sets of control limits.

A health care provider uses a u chart to report the rate of office visits performed each month by each of its clinics. The rate is computed by dividing the number of visits by the membership expressed in thousand-member years. Figure 4 shows data collected for Clinic E and saved in a SAS data set named ClinicE.

Figure 4 SAS Data Set ClinicE

```
                    Office Visit Data for Clinic E

          month    _phase_     nvisite     nyrse     days     mmse

          JAN04    Phase 1      1421      0.66099     31      7676
          FEB04    Phase 1      1303      0.59718     28      7678
          MAR04    Phase 1      1569      0.66219     31      7690
          APR04    Phase 1      1576      0.64608     30      7753
          MAY04    Phase 1      1567      0.66779     31      7755
          JUN04    Phase 1      1450      0.65575     30      7869
          JUL04    Phase 1      1532      0.68105     31      7909
          AUG04    Phase 1      1694      0.68820     31      7992
          SEP04    Phase 2      1721      0.66717     30      8006
          OCT04    Phase 2      1762      0.69612     31      8084
          NOV04    Phase 2      1853      0.68233     30      8188
          DEC04    Phase 2      1770      0.70809     31      8223
          JAN05    Phase 2      2024      0.78215     31      9083
          FEB05    Phase 2      1975      0.70684     28      9088
          MAR05    Phase 2      2097      0.78947     31      9168
```

The variable nvisite is the number of visits each month, and the variable mmse is the number of members enrolled each month (in units of "member months"). The variable days is the number of days in each month. The variable nyrse expresses mmse in units of thousand members per year. The variable _phase_ separates the data into two time phases because a change in the system is known to have occurred in September 2004 at the beginning of Phase 2.

The following statements create a u chart with a single set of default limits. The chart is shown in Figure 5.

```
title 'U Chart for Office Visits per 1,000 Members: Clinic E';
proc shewhart data=ClinicE;
   uchart nvisite * month / subgroupn = nyrse nohlabel nolegend
                   odstitle = title;
   label nvisite = 'Rate per 1,000 Member-Years';
run;
```

The default control limits are clearly inappropriate because they do not allow for the shift in the average rate that occurred in September 2004.

Figure 5 u Chart with Single Set of Limits

U Chart for Office Visits per 1,000 Members: Clinic E
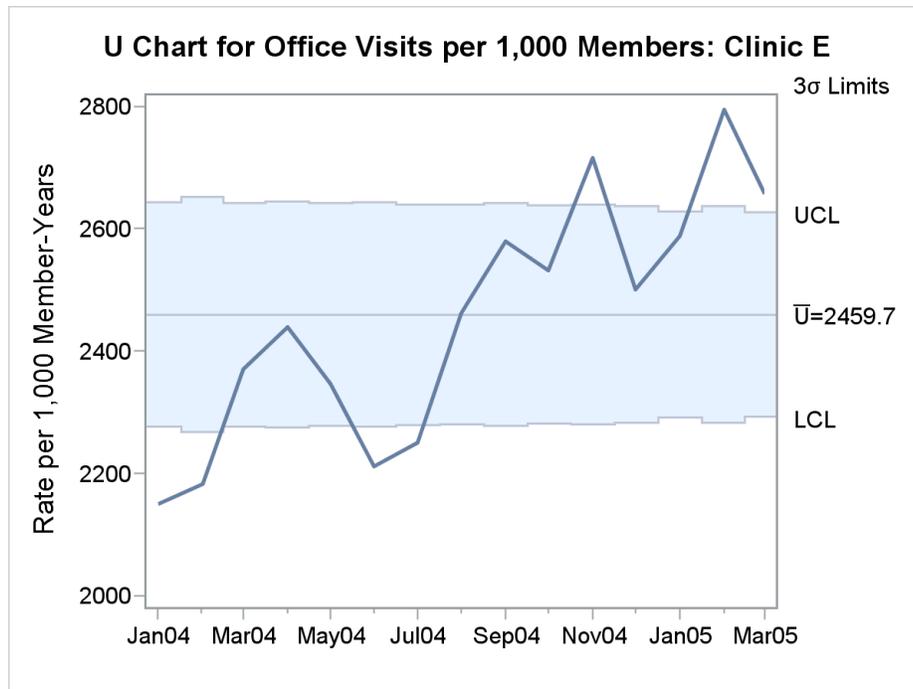
The following statements use BY processing to compute distinct sets of control limits from the data in each phase and save the control limit information in a SAS data set named Vislimit. The NOCHART option suppresses the display of separate control charts for each phase.

```
proc shewhart data=ClinicE; by _phase_;
   uchart nvisite * month / subgroupn = nyrse  nochart
                            outlimits = Vislimit (rename=(_phase_=_index_));
run;
```

Figure 6 shows a listing of Vislimit. The values of the lower and upper control limit variables _LCLU_ and _UCLU_ are equal to the special missing value V; this indicates that these limits are varying. The variable _index_ identifies the control limits in the same way that the variable _phase_ identifies the time phases in the data.

Figure 6 SAS Data Set Vislimit

Control Limits for Office Visit Data									
index	_VAR_	_SUBGRP_	_TYPE_	_LIMITN_	_ALPHA_	_SIGMAS_	_LCLU_	_U_	_UCLU_
Phase 1	nvisite	month	ESTIMATE	V	V	3	V	2302.99	V
Phase 2	nvisite	month	ESTIMATE	V	V	3	V	2623.52	V

The following statements combine the data and control limits for both phases in a single u chart, shown in Figure 7:

```
title 'U Chart for Office Visits per 1,000 Members: Clinic E';
proc shewhart data=ClinicE limits=Vislimit;
   uchart nvisite * month / subgroupn   = nyrse
                            readindex   = all
                            readphase   = all
                            nohlabel     nolegend
                            phaselegend  nolimitslegend
                            odstitle    = title;
   label nvisite = 'Rate per 1,000 Member-Years';
run;
```

The READINDEX= and READPHASE= options match the control limits in Vislimit with observations in ClinicE by the values of the variables _index_ and _phase_, respectively.

6

In Figure 7, no points are out of control, indicating that the variation is due to common causes after adjusting for the shift in September 2004. Both sets of control limits in Figure 7 were estimated from the data with which they are displayed. You can, however, apply pre-established control limits from a LIMITS= data set to new data.

Figure 7 u Chart with Multiple Sets of Control Limits
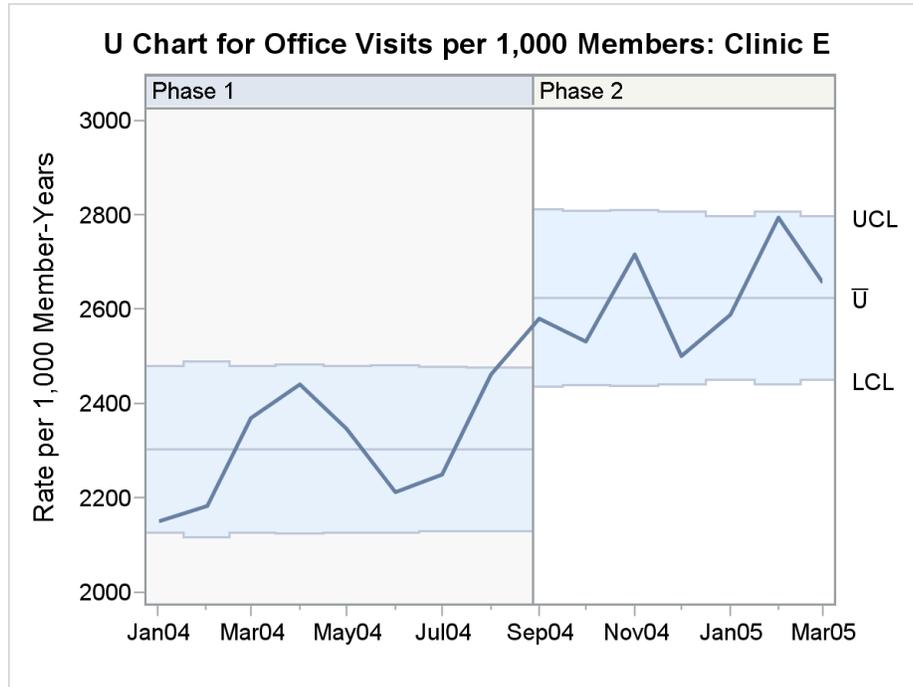
In applications that involve count data, control charts for individual measurements can sometimes be used in place of u charts and c charts, which are based on a Poisson model, or in place of p charts and np charts, which are based on a binomial model. Wheeler (1995) makes the point that charts based on a theoretical model "allow one to detect departures from the theoretical model," but they require verification of the assumptions required by the model. On the other hand, charts for individual measurements often provide reasonably approximate empirical control limits, as illustrated in the next example.

STATISTICAL THINKING FOR BUSINESS PROCESS IMPROVEMENT AT A BANK

Because variation exists in all processes, statistical thinking can be used to monitor and improve the processes in any business environment. However, in contrast to manufacturing environments, where engineers and operators are trained to work with processes, the steps that create products or services in a business such as a bank are not generally analyzed as a process by managers and analysts. Consequently, a helpful preliminary step in statistical thinking is to diagram the various steps in creating a product or a service. Hoerl and Snee (2002) provide a generic framework for any process, which they refer to as the SIPOC model, an acronym for its components, which are as follows:

- suppliers are individuals or groups who provide the inputs to the process

- inputs are materials or information that flow from the supplier

- process steps are activities that transform inputs into outputs

- outputs are the product or service that is produced by the process

- customers are the individuals or groups who use the outputs

In any specific situation only a subset of these elements might be present, as illustrated in the next example.

Example 3: SPC in a Bank Lending Application

Schleicher (2008) described the application of statistical process control to the process of making automated lending decisions in a large bank, where the ability to understand and react to the variability in this process is essential for risk management. Figure 8 illustrates the steps in the process.

Figure 8 Process for Lending Decisions
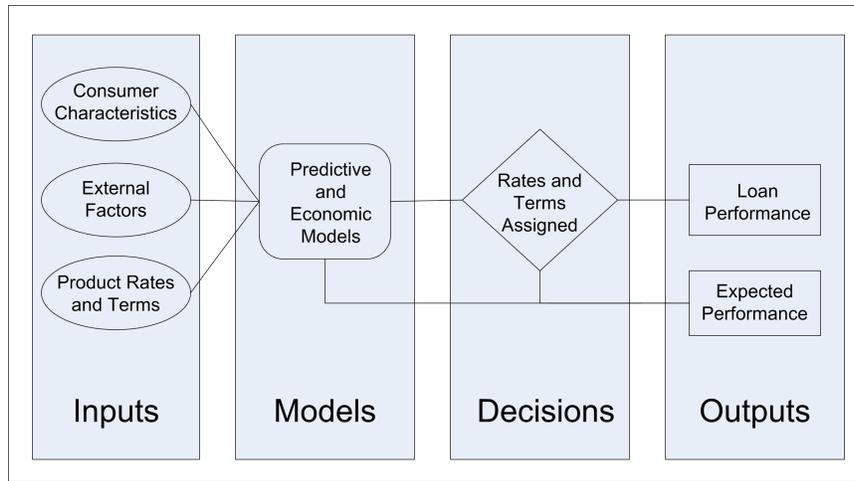
Variability is present at each step. At the input step, the consumers who apply for credit vary in their characteristics and in their options for obtaining credit elsewhere. While product rates are controlled by the lender, external factors are not, and these include rates available at other banks. The lender uses information about the consumer along with models, such as credit scores, to determine the rate for the loan. This decision is subject to variability because the models are based on historic data, which are an imprecise predictor of the future. The output of the process is the performance of the loan, which can vary due to changes in the economy, industry rates and terms, and legislation.

Statistical process control can be used to monitor for unusual variation in the inputs and outputs of the process and the intermediate steps. A noteworthy distinction between this use of SPC and manufacturing applications of SPC is that special causes of variation in business environments can be detected but not necessarily eliminated.

Unusual variation in the inputs provides a leading indicator of shifts downstream in the process. One example of an input process measurement is the reported income of applicants for an auto loan. Figure 9 displays a control chart for the average monthly incomes of loan applicants for a particular vehicle (the averages have been standardized to protect data confidentiality). Although one average exceeds the upper 3σ limit, this is not unusual with small samples from skewed income distributions. Based on experience with other charts of this type, this chart does not show an unusual pattern of variation.

The following statements produce the control chart in Figure 9. The XCHART statement requests a control chart for the monthly averages, which are treated as individual measurements.

```
title "Means Chart for Reported Income";
proc shewhart history=AvgIncome;
   xchart income*month / odstitle=title nolegend outfill;
   label income = "Avg Reported Income (Standardized)";
         month  = "Month";
run;
```

8

Figure 9 Control Chart for Reported Income
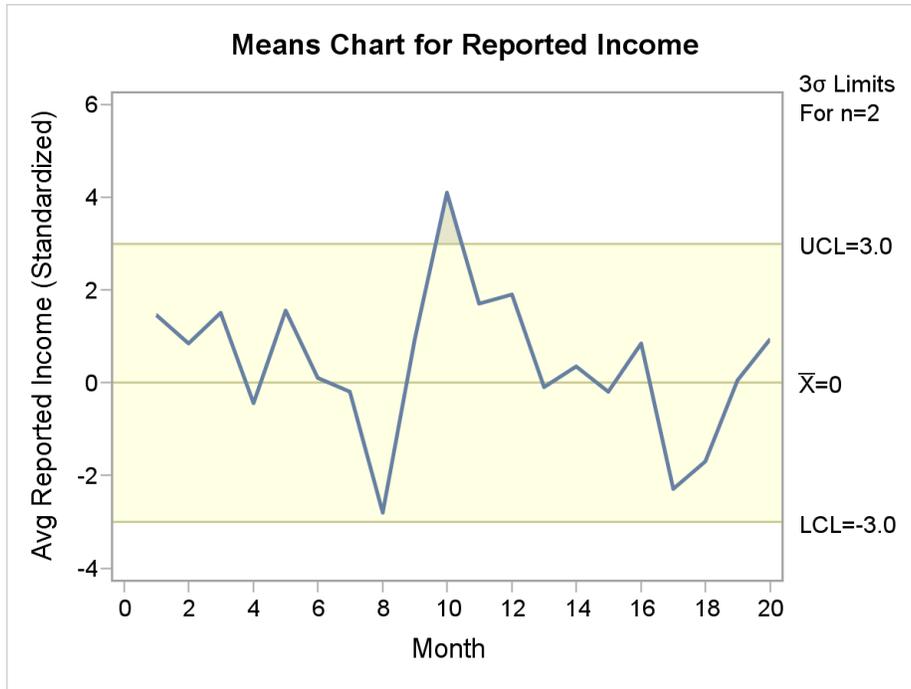
Another input process measurement is the percent of auto loan applicants who do not have a home mortgage. Figure 10 displays a p chart for the proportions of these applicants at the months when the loans were booked. This chart shows a downward trend followed by a sudden spike, which calls for an investigation to determine the cause. Possibilities to consider are a change in lending practice, a data definition issue, or a data quality problem.

Figure 10 Control Chart for Applicants without a Mortgage
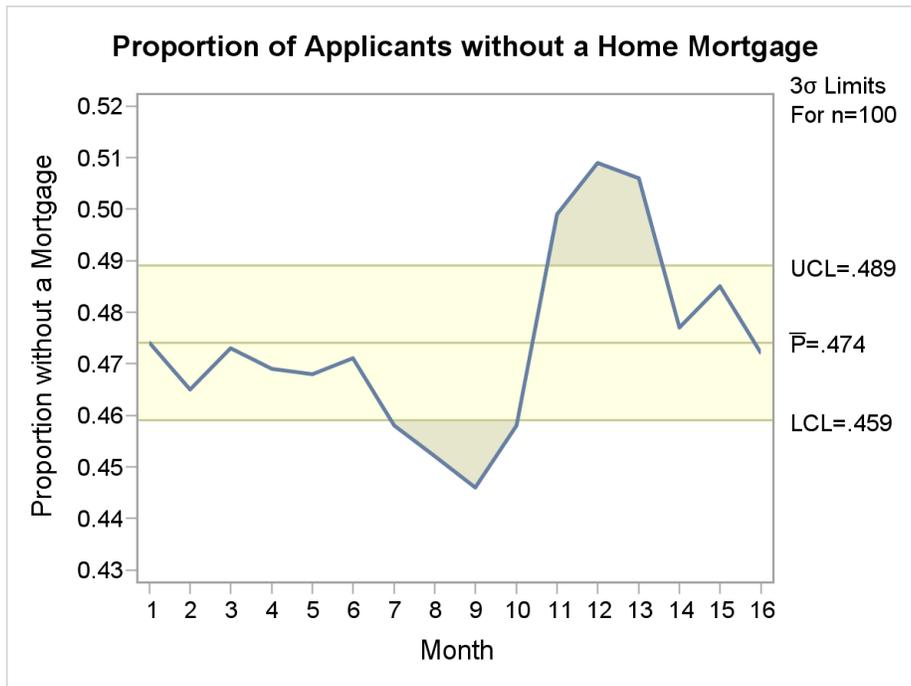
The following statements produce the control chart in Figure 10. The PCHART statement requests a control chart for monthly proportions.

```
title "Proportion of Applicants without a Home Mortgage";
proc shewhart data=missprop limits=misslim;
    pchart pmiss*id / subgroupn=100 nolegend odstitle=title outfill;
    label pmiss = "Proportion without a Mortgage";
    label month = "Month";
run;
```

For the analysis of process outputs, Schleicher (2008) pointed out that monitoring residuals (the differences between actual results and expected results based on a model) is helpful for assessing the performance of the model by detecting variability that is not adequately captured by the model. Sources of such variability include consumer behavior, the banking industry and the economy, decisions made by the bank, and seasonal patterns.

Figure 11 shows a control chart for monthly averages of the residual amounts of credit card purchases (actual minus expected) for a group of customers. For the first eleven months, the model underestimated the volume of purchases, as indicated by a series of consecutive positive residuals. This discovery led to an investigation which identified changes in consumer behavior.

Figure 11 Control Chart for Average Card Purchase Residuals
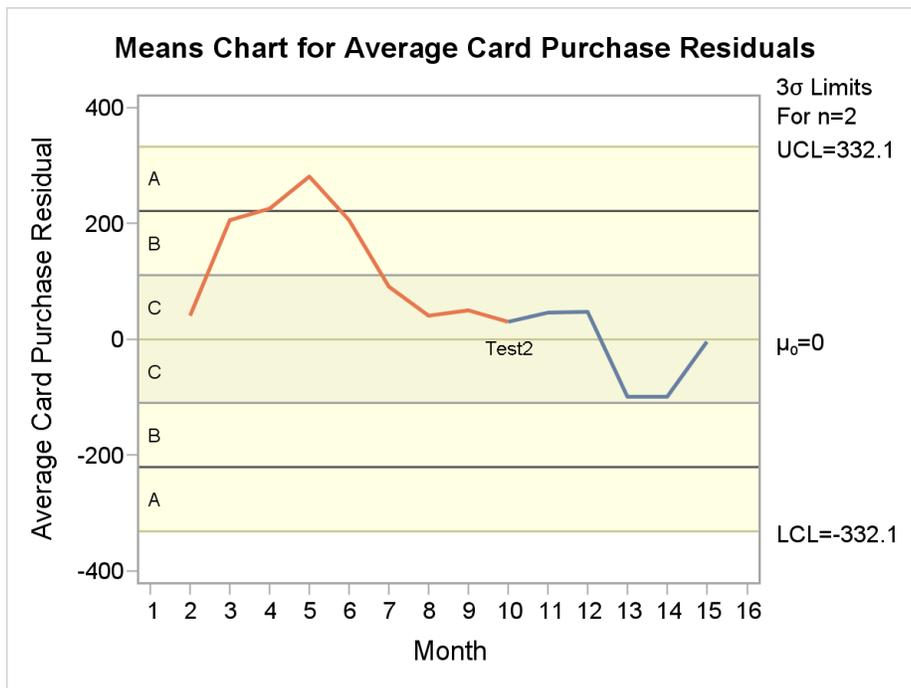
The following statements produce the chart in Figure 11:

```
title "Means Chart for Average Card Purchase Residuals";
proc shewhart data=cardpurchdat limits=cardpurchlim;
    irchart resavg*month / nochart2
                           zones zonelabels tests = 1 2 3 4
                           xsymbol = mu0
                           odstitle = title
                           odsfootnote = ".";
    label resavg = "Average Card Purchase Residual"
          month  = "Month";
run;
```

Here, the IRCHART statement is used because the residuals are treated as individual measurements. Test 2 is signaled because it detects nine points in a row in Zone C or beyond on one side of the central line.

More powerful techniques are available for analyzing residuals from predictive models and forecasts, For example, variance component estimation methods can be applied to decompose residual variability into components that correspond to product differences, consumer differences, and month-to-month variation for individual consumers. A variety of statistical procedures can be used for this purpose, including the VARCOMP, MIXED, and GLIMMIX procedures. Variance component analysis can also be used to adjust the limits for a control chart to appropriately reflect the variation in the data. Refer to *SAS/QC User's Guide* for an example of this approach (see "Specialized Control Charts").

Example 4: Exploring Process Variation in a Call Center Application

The managers of a call center operation at a bank are interested in handling calls more effectively in order to improve customer satisfaction and reduce costs. One measure of the call-handling process is the time (in seconds) required to handle the call.

Histograms are basic tools for examining the distributions of process measurements, and they are especially useful when the measurements are skewed and cannot be adequately summarized by their mean and standard deviation. The following statements use the CAPABILITY procedure in SAS/QC software to create a histogram for handling times. The histogram is shown in Figure 12.

```
ods graphics on;

proc capability data=callcenter noprint;
   var time;
   histogram time / kernel
                   curvelegend=none
                   midpoints=100 to 4100 by 200 ;
   inset n / position=ne;
   label time = "Handling Time (seconds)";
run;
```

This display shows that the distribution has a long tail, with most calls handled in less than 22 minutes.

Figure 12 Handling-Time Distribution
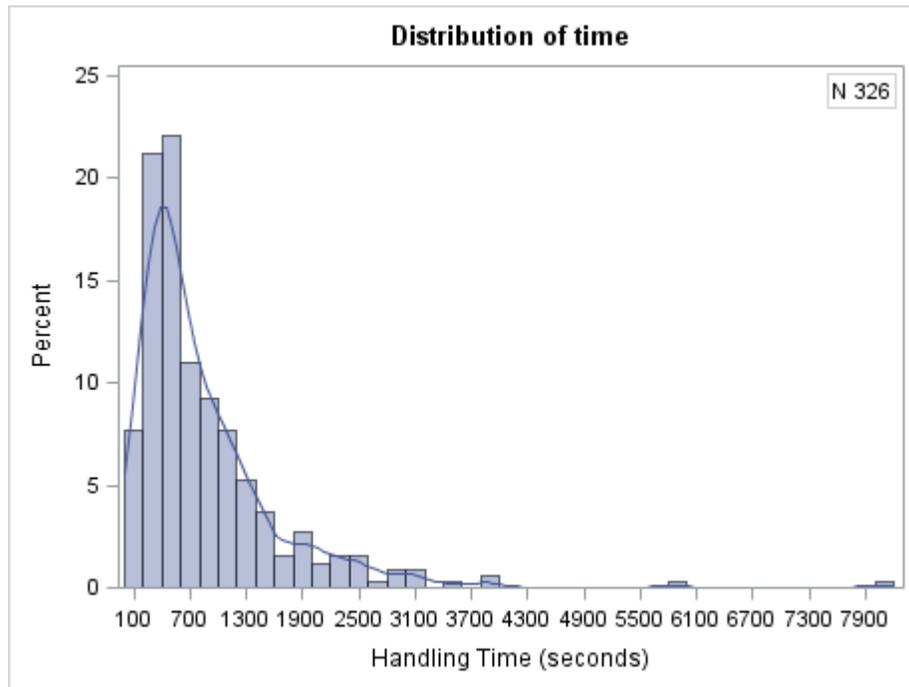
The next statements create a comparative histogram for time stratified by the classification variable unit, which designates the business unit that corresponds to each call.

```
proc capability data=callcenter noprint;
   var time;
   class businessunit;
   histogram time / kernel
                   curvelegend=none
                   nrows=3
                   intertile = 0.5
                   midpoints=100 to 4100 by 200 ;
   inset n / position=ne;
   label time = "Handling Time (seconds)";
   label businessunit = "Unit";
run;
```

11

The display, shown in Figure 13, reveals that the distribution for the Key Accounts unit has the most dispersion, and the distribution for the Service unit has the least dispersion. This difference had gone unnoticed by management until it was discovered in the graph.

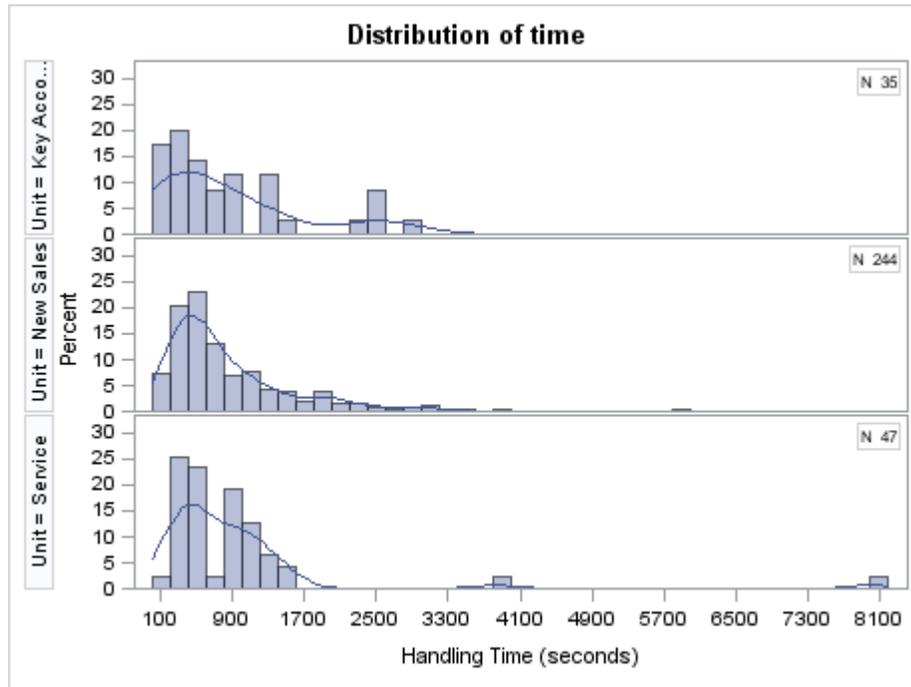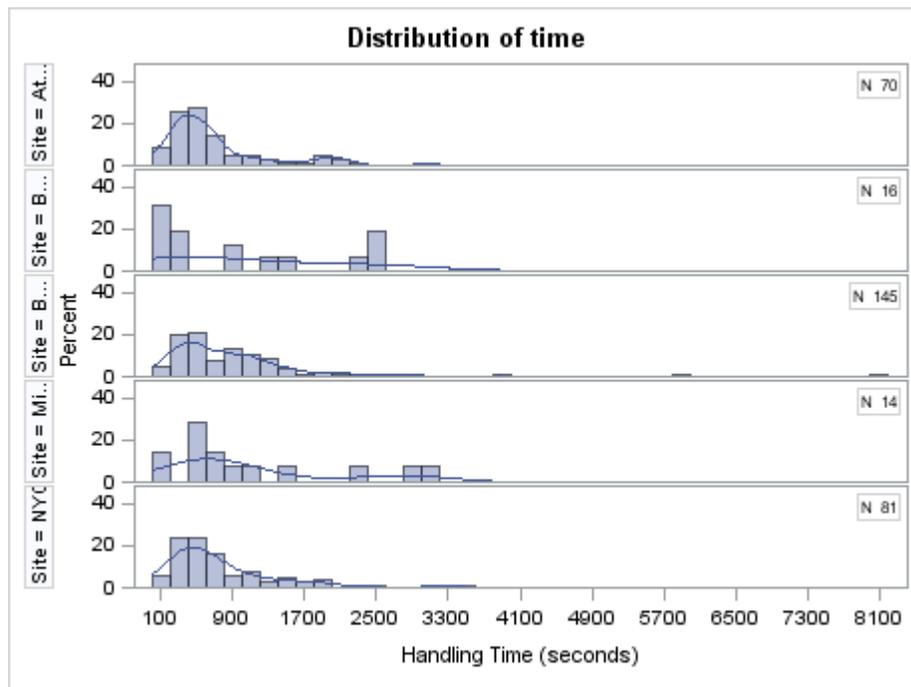**Figure 13** Handling-Time Distribution by Business Unit

Figure 14 displays a comparative histogram stratified by the classification variable site, which designates the location of the center that handled each call.

Figure 14 Handling-Time Distribution by Site

This display also revealed differences that were previously unknown, raising questions that led to a better understanding of the process.

12

MULTIVARIATE PROCESS MONITORING WITH COMPLEX SYSTEMS

The control charts constructed in the previous sections are based on simple statistical models for the common cause variation in a single process variable, such as the Poisson distribution for the u chart in Example 1, the normal distribution for the $\bar{X}$ chart in Example 2, and the binomial distribution for the p chart in Example 2. These models work well in Shewhart's framework, where process measurements are univariate and where common cause variation is viewed as a series of chance disturbances away from an average level that is assumed to be constant over short periods of time.

In contrast, the application of statistical process monitoring to complex systems—both in manufacturing applications and in business environments—requires methods that work well with large volumes of multivariate data collected over time, often from disparate sources and databases. In addition, a much richer set of statistical models is needed to characterize the expected variation in the data, which, in addition to common cause variation, can include effects such as correlation among process variables, time-dependent behavior, and multiple sources of variation. More extensive models are also needed in situations where a set of internal process variables is used to construct a predictive model for a set of external measures, such as product quality, reliability, and customer satisfaction.

Example 5: Multivariate Process Monitoring in a Regulatory System

To illustrate the modeling issues that arise with a large number of process variables observed over months, consider the problem facing a regulatory organization responsible for monitoring 6,000 securities firms for unusual activities. Each firm submits data on more than 100 variables at regular time intervals, including the number of sales representatives, revenue per representative, customer complaints per representative, and customer complaints per sales volume. Examiners use this data to investigate financial problems and questionable sales practices, but because of the large volume of the data, statistical process monitoring is used to screen the data for unusual patterns that occur over time measured in days.

Multivariate process monitoring based on a principal components model is an effective approach for dealing with hundreds or thousands of correlated process variables. This approach is one of many techniques that were introduced during the mid 1990s by chemometricians for applications within the chemical process and pharmaceutical industries, which range from modeling the quality of paper pulp from digester process variables to the development of structure-activity relationships for new drug compounds. The multivariate monitoring approach outlined here is patterned after the work of Kourti and MacGregor (1995, 1996).

The following statements use the MVPMODEL procedure in SAS/QC software to build a principal components model for the firm data (for simplicity, only seven variables are analyzed here).

```
ods graphics on;
proc mvpmodel data=peerfirms
      nfactors=3
      out=mvpnasd
      plots=(score(ellipse));
   var CompPerSales CompPerRep RevPerRep MeritSalesComp
      MeritTerms NReps ExcessCapital;
run;
```

For details covering the syntax of the MVPMODEL procedure, see Christian and Ransdell (2012). The model information is stored in the MVPNASD data set, and is used by the MVPMONITOR and MVPDIAGNOSE procedures.

The firms in the system are grouped into peer groups. For one of these groups, as shown in Figure 15, three principal components account for 82% of the variation in the variables reported during a period of 40 months.

Figure 15 Eigenvalue and Variance Information

```
              The MVPMODEL Procedure

         Eigenvalues of the Correlation Matrix

          Eigenvalue    Difference    Proportion    Cumulative

      1   2.88992536    1.10169223       0.4128        0.4128
      2   1.78823313    0.74983408       0.2555        0.6683
      3   1.03839905                     0.1483        0.8167
```

The PLOTS= option generates score plots for all components in the model. The ELLIPSE option specifies that a 95% prediction ellipse is displayed, and the scatter plot matrix is shown in Figure 16.

Figure 16 Score Plots for Principal Components
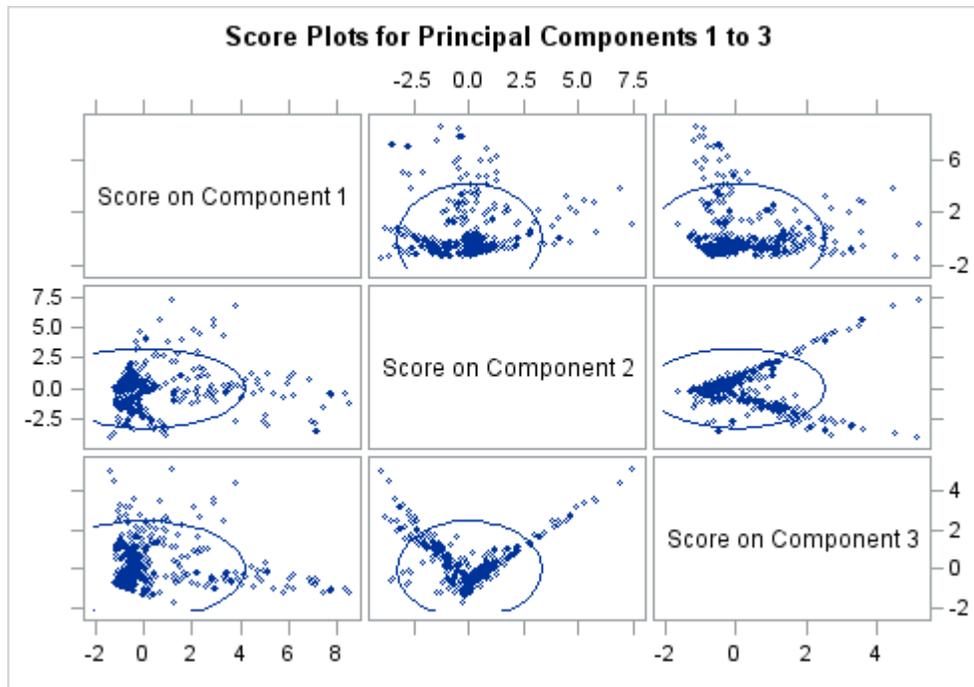
In order to monitor the variation for one particular firm in the peer group (call it Firm 7870), the multivariate approach computes the statistic T_i^2 for the ith day as follows:

$$T_i^2 = \frac{t_1^2(i)}{s_{t_1}^2} + \frac{t_2^2(i)}{s_{t_2}^2} + \frac{t_3^2(i)}{s_{t_3}^2}$$

T_i^2 is the sum of squares of the scores for the first three principal components, scaled by their variances.

The following statements use the MVPMONITOR procedure in SAS/QC software to create T^2 and SPE charts using the output data set created by the MVPMODEL procedure. That data set contains the original data, the principal component scores, the T^2 and SPE statistics and more.

```
proc mvpmonitor history=mvpnasd;
   time timekey;
   series firmkey;
   tsquarechart / seriesvalue=7870 totpanels=1;
   spechart / seriesvalue=7870 totpanels=1;
run;
```

Control limits for T_i^2 are computed from a reference distribution based on data from the entire peer group. The chart, which is shown in Figure 17, reveals unusual variation for Firm 7870 in the three most recent quarters.

14

Figure 17 Multivariate Control Chart for T^2 Statistics
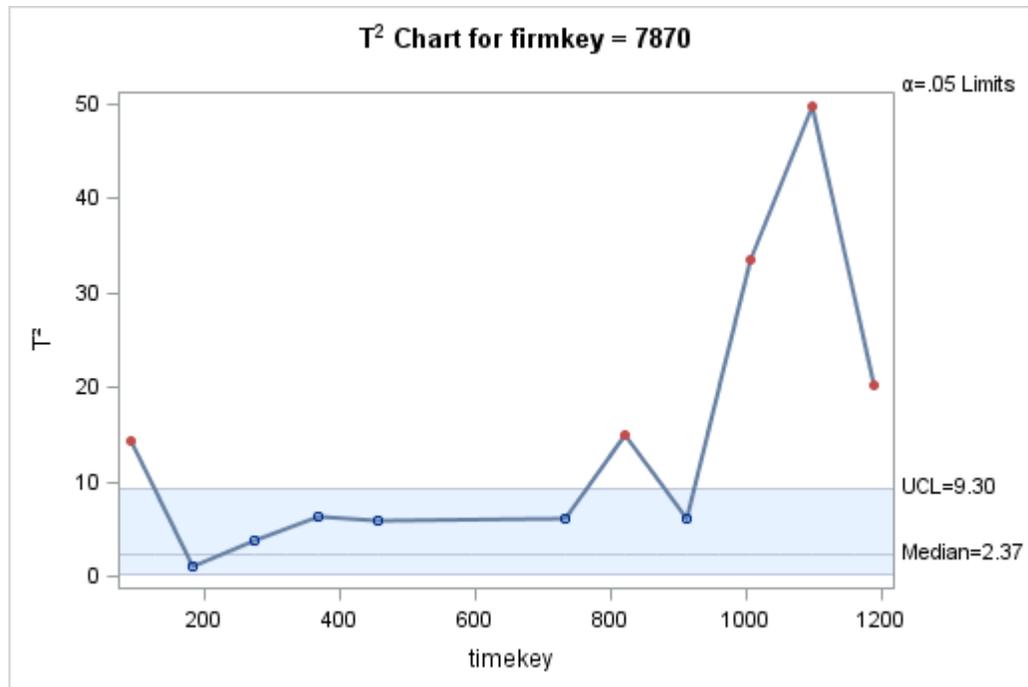
You can also visualize how the variation for Firm 7870 departs from the pattern of variation for its peer firms by making pairwise scatter plots of the three scores t_1, t_2, and t_3 for all the firms, and adding 95% prediction ellipses. The following statements generate a dummy variable for Firm 7870, which is subsequently used to diagnose unusual variation in Firm 7870.

```
data mvpnasdNew;
   set mvpnasd;
   firmOfInterest='Other';
   if firmkey=7870 then firmOfInterest='7870';
   label firmofinterest="Firm of Interest";
run;
```

The following statements use the MVPDIAGNOSE procedure to produce score plots from the principal components model computed previously by the MVPMODEL procedure.

```
proc mvpdiagnose history=mvpnasdNew;
   scoreplot /  ellipse group=firmOfInterest labels=outside;
   scoreplot /  ellipse group=firmOfInterest labels=outside xcomp=1 ycomp=3;
   id firmkey;
run;
```

The score plots are shown in Figure 18 and Figure 19. These score plots highlight the points for Firm 7870 in red using the GROUP= option, and labels points lying outside the prediction ellipse with the firm number using the LABELS= option and the ID statement. There is evidence that several other firms are wandering away from the expected region of variability for their peers.

**Figure 18** Score Plot for Principal Components One and Two
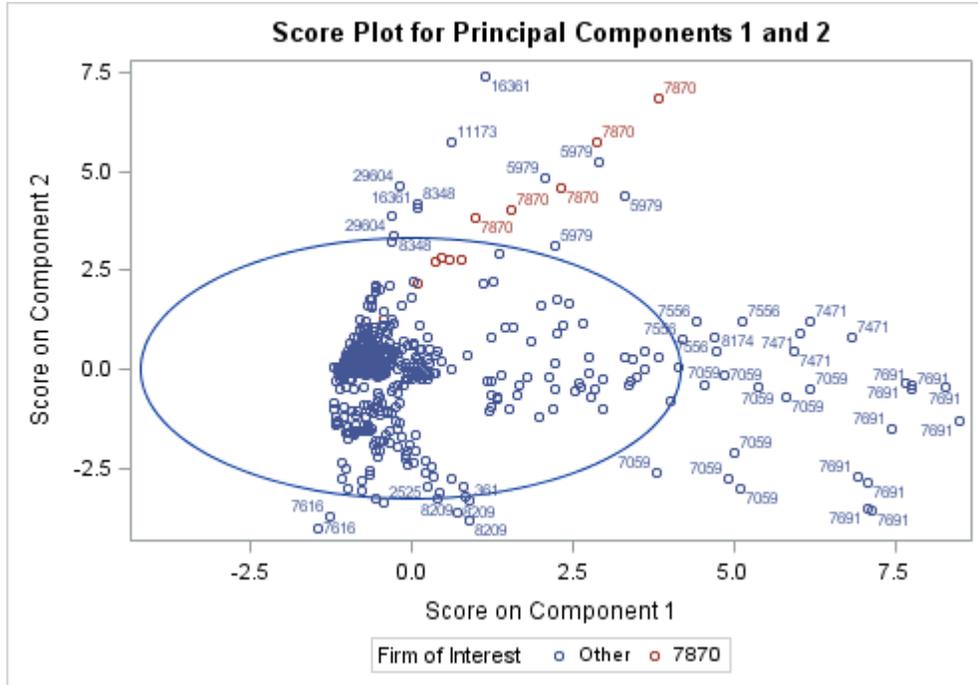
Figure 19 Score Plot for Principal Components One and Three
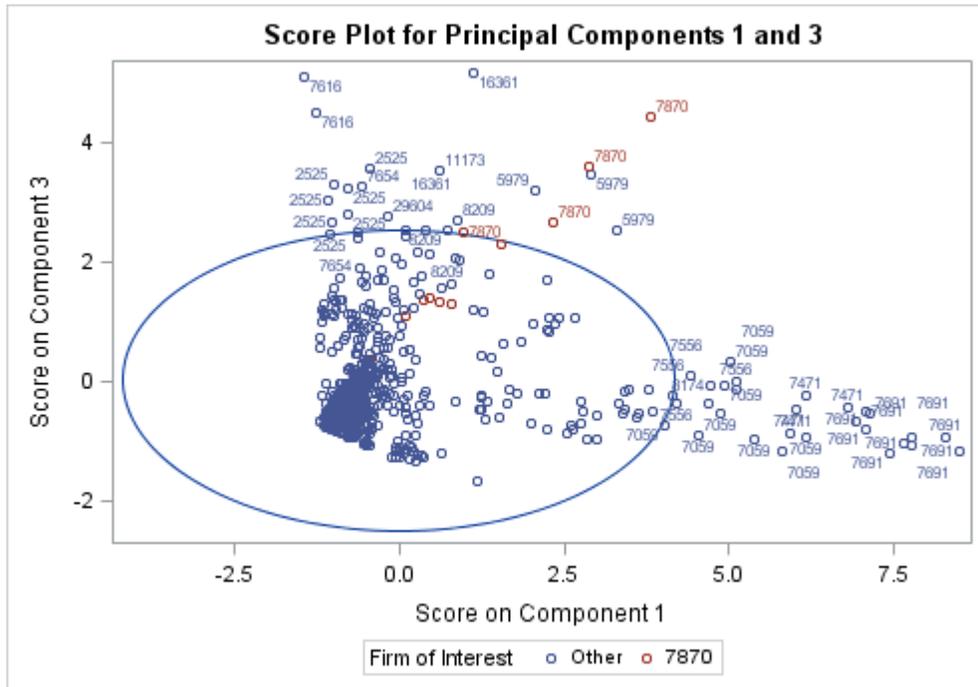
There are two explanations for this behavior. The first is that the process has moved outside the control ellipse but is still within the hyperplane defined by the principal components model. The second is that the process has moved off the hyperplane and has changed in a way that is not captured by the model. You can check for the second possibility by plotting the distance of each time point to the model hyperplane, or equivalently, the squared prediction error (SPE), as shown in Figure 20.

**Figure 20** Multivariate Control Chart for SPE Statistics
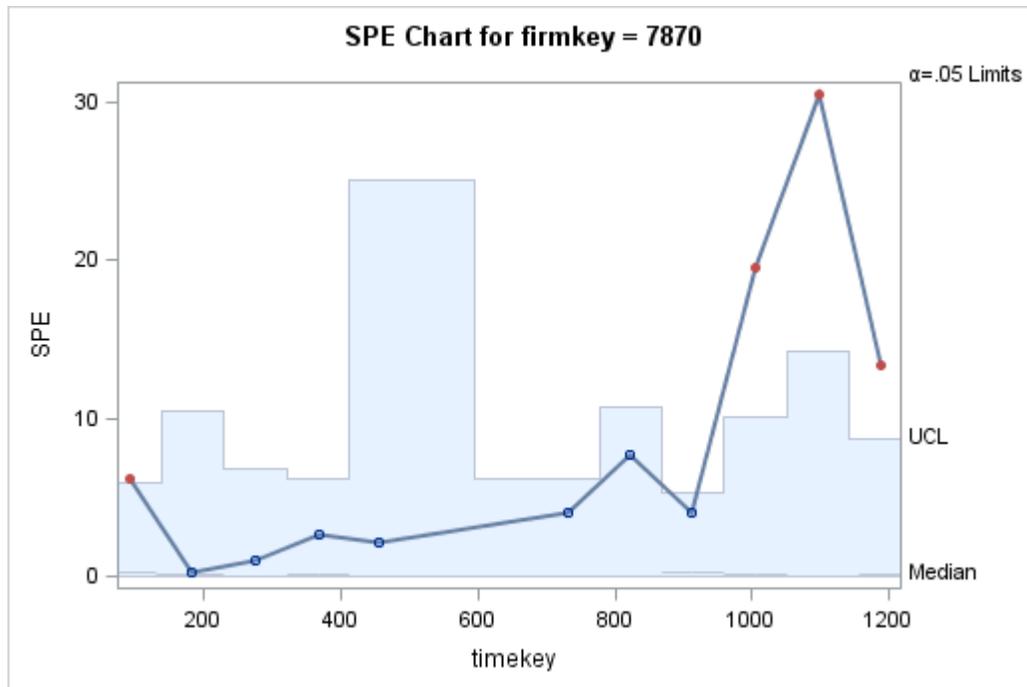
Here the control limits for SPE are computed using all the firms in the peer family as a reference normal data set (Nomikos and MacGregor 1995). The chart indicates that the process has moved off the model plane. Not only is this firm's behavior diverging from the common cause variation exhibited by its peers, it is also diverging in a new way with variation not observed in the data that were used to develop the model. This indicates the need for constructing a new model.

Another way to diagnose the behavior in Figure 17 is with a contribution plot, which tells you which variables contribute to the "gap" between a point, such as the one at Day 1097, and the center of the ellipse. For each variable in the model, the contribution plot displays a root sum of squares of weighted residuals. The following statements use the MVPDIAGNOSE procedure in SAS/QC 12.1 software to create a contribution plot for the observation of firm 7870 when the time variable is 1097.

```
proc mvpdiagnose history=mvpnasd;
    where timekey=1097 and firmkey=7870;
    time timekey;
    contributionplot;
run;
```

The plot, displayed in Figure 21, shows that two complaint variables and a variable that measures merit terminations contributed to the jump at Day 1097. Events related to these variables should be examined to determine the causes for this variation.

Figure 21 Contribution Plot
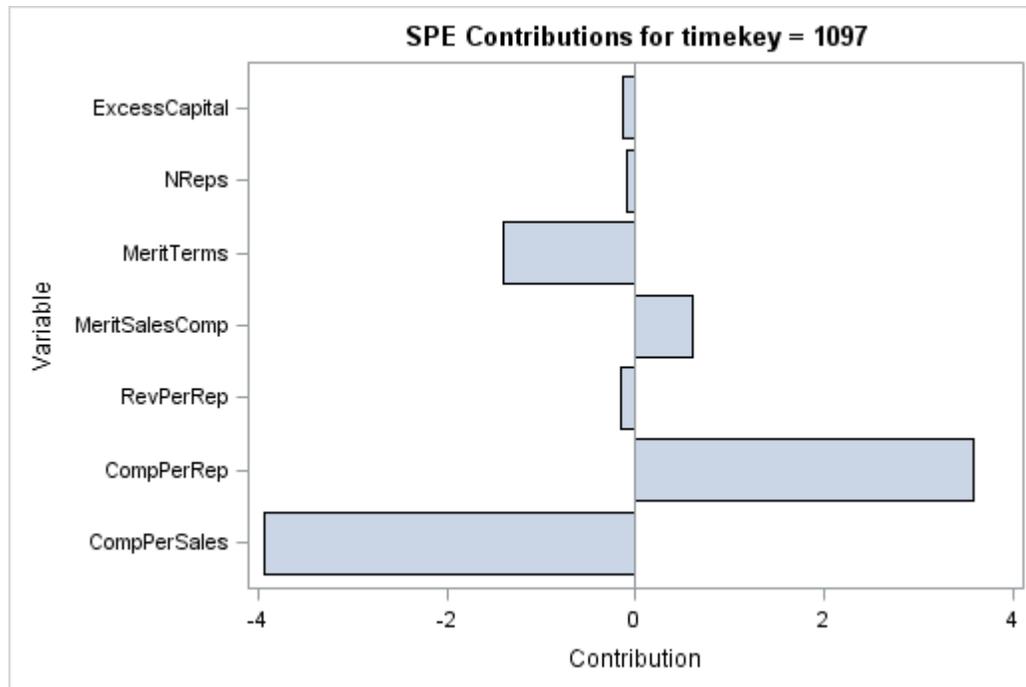
NOTE: The MVPMODEL and MVPMONITOR procedures are experimental in SAS 9.3. The examples in this paper reflect the production versions of these procedures along with the new MVPDIAGNOSE procedure, which will be included with SAS/QC 12.1.

CONCLUSION: IT REALLY IS ALL ABOUT VARIATION!

The examples presented in this paper demonstrate how the principles of statistical thinking and statistical process monitoring apply in a variety of business settings. A common theme is that variation is present in all processes and that the key to process knowledge and improvement is analyzing and acting on the sources of variation.

In manufacturing applications, the goals of statistical process control are to create a stable, predictable process by eliminating special causes of variation and then to achieve a process that is capable of meeting customer requirements by reducing common cause variation (Wheeler 2000). In health care applications, special causes can often be eliminated through corrective action such as redesigning procedures and training staff. On the other hand, this form of statistical process control is not feasible in many business environments. Instead, statistical process monitoring can be used for improved understanding of variability and for early detection of problems, which reduces costs and minimizes risk.

Successful application of process monitoring in complex systems requires statistical models that adequately describe the variation in hundreds or thousands of variables that are correlated or subject to multiple sources of variation, including seasonality and other time-dependent behavior. SAS provides a wealth of modeling techniques for this purpose.

ACKNOWLEDGMENTS

REFERENCES

Benneyan, J. (2001a), "Number-Between g-Type Statistical Control Charts," *Health Care Management Science*, 4, 305–318.

Benneyan, J. (2001b), "Performance of Number-Between g-Type Statistical Control Charts for Monitoring Adverse Events," *Health Care Management Science*, 4, 319–336.

Berwick, D. M., Godfrey, A. B., and Roessner, J. (1990), *Curing Health Care, New Strategies for Quality Improvement*, San Francisco, CA: Jossey—Bass.

Christian, J. B. and Ransdell, B. (2012), "Monitoring the Variation in Your Multivariate Process: An Introduction to the MVP Procedures," in *Proceedings of the SAS Global Forum 2012 Conference*, Cary, NC: SAS Institute Inc., available at http://www2.sas.com/proceedings/forum2012/318-2012.pdf.

Deming, W. E. (1982), *Out of the Crisis*, Cambridge, MA: Massachusetts Institute of Technology, Center for Advanced Engineering Study.

Hindo, B. (2007), "At 3M, A Struggle between Efficiency and Creativity," *Business Week*, last accessed February 13, 2010.
URL http://www.businessweek.com/magazine/content/07_24/b4038406.htm?chan=top+news_top+news+index_best+of+bw

Hoerl, R. W. and Snee, R. D. (2002), *Statistical Thinking: Improving Business Performance*, Belmont, CA: Brooks/Cole.

Institute of Medicine (2000), *To Err Is Human: Building a Safer Health System*, Washington, DC: National Academy Press.

Institute of Medicine (2001), *Crossing the Quality Chasm: A New Health System for the 21st Century*, Washington, DC: National Academy Press.

Kourti, T. and MacGregor, J. F. (1995), "Process Analysis, Monitoring and Diagnosis, Using Multivariate Projection Methods," *Chemometrics and Intelligent Laboratory Systems*, 28, 3–21.

Kourti, T. and MacGregor, J. F. (1996), "Multivariate SPC Methods for Process and Product Monitoring," *Journal of Quality Technology*, 28, 409–428.

Limaye, S. S., Mastrangelo, C. M., and Zerr, D. M. (2008), "A Case Study in Monitoring Hospital-Associated Infections with Count Control Charts," *Quality Engineering*, 20, 404–413.

Maynard, M. (2010), "An Apology From Toyota's Leader," New York Times, http://www.nytimes.com/2010/02/25/business/global/25toyota.html, last accessed February 25, 2010.

Nomikos, P. and MacGregor, J. F. (1995), "Multivariate SPC Charts for Monitoring Batch Processes," *Technometrics*, 37, 41–59.

Peterson, J. J., Snee, R. D., McAllister, P. R., Schofield, T. L., and Carella, A. J. (2009), "Statistics in Pharmaceutical Development and Manufacturing," *Journal of Quality Technology*, 41(2), 111–147.

Ramirez, J. G. and Tobias, R. D. (2007), "Split and Conquer! Using SAS/QC to Design Quality into Complex Manufacturing," in *Proceedings of the SAS Global Forum 2007 Conference*, Cary, NC: SAS Institute Inc.

Rodriguez, R. N. (2011), *An Overview of ODS Statistical Graphics in SAS® 9.3*, Technical report, SAS Institute Inc.
URL http://support.sas.com/rnd/app/papers/papers_ods.html

Rodriguez, R. N. and Ransdell, B. (2010), "Statistical Process Control for Health Care Quality Improvement Using SAS/QC Software," http://support.sas.com/rnd/app/papers/papers_qc.html.

Schleicher, K. (2008), "SPC for Lending," Presentation at the 2008 Fall Technical Conference sponsored by the American Statistical Association and the American Society for Quality.

Tsui, K. L., Chiu, W., Gierlich, P., Goldman, D., Liu, X., and Maschek, T. (2008), "A Review of Healthcare, Public Health, and Syndromic Surveillance," *Quality Engineering*, 20, 435–450.

Wheeler, D. J. (1995), *Advanced Topics in Statistical Process Control*, Knoxville, TN: SPC Press.

Wheeler, D. J. (2000), *Understanding Variation: The Key to Managing Chaos*, Second Edition, Knoxville, TN: SPC Press.

Wheeler, D. J. and Polling, S. R. (1998), *Building Continual Improvement: A Guide for Business*, Knoxville, TN: SPC Press.

Wikipedia (2010), "If Japan Can... Why Can't We?" http://en.wikipedia.org/wiki/If_Japan_Can..._Why_Can't_We%3F, last accessed February 13, 2010.

Woodall, W. H. (2006), "The Use of Control Charts in Health-Care and Public-Health Surveillance," *Journal of Quality Technology*, 38(2), 89–104.

Contact Information

Robert N. Rodriguez
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
(919) 531-7650
Bob.Rodriguez@sas.com