IE 6200

Engineering probability and statistics SEC 01

Fall2016

Professor: Rajesh Jugulum

Technical Paper Review :

Statistical process control and its relevance in data quality monitoring and reporting

Team Members:

Abhishek Thosar
Sanket Kulkarni
Saish Pai
Swenal Fargose

1) Please review and analyze the paper titled “Statistical process control and its relevance in data quality monitoring and reporting.

The purpose of the effort:

The main purpose of this paper is to learn about the Statistical process control and its benefits in data quality monitoring and reporting. Statistical Process Control is a mathematically based control technique which can be used to automate the identification of variations in the data. The paper helps us understand ways in which SPC is beneficial, especially as a tool which provides a scalable means of monitoring data quality level and reporting on trends over time, and in turn recognize the variations in data elements . It also provides us with tools that enable us to reduce these variations in cases where data is expected to be consistent. It aims at explaining how data quality monitoring and control activities can be performed at different data levels. By explaining how standardization and automation of Statistical Process Control charts can enable monitoring and controlling large volumes of data, the paper identifies techniques for successful implementation of data quality operations. The use of SPC in quantitatively determining thresholds, control levels and monitoring activities associated with data elements is explicated in this paper.

The paper highlights use of SPC in effectively and efficiently managing data quality activities. This is done by using an analytical framework with SPC enabling us to identify the focus areas for improvement and conducting drill down analysis by identifying the causes for anomalies. SPC helps us in predicting the process by clearly distinguishing the causes of variations (common causes and special causes) within the process. In addition, Statistical Process Control is a technique by which we can learn about the stability, predictability and capability of a process.

Methodology used in the technical paper:

Statistical Process Control is used as a technique for the purposes of data quality monitoring and process control. Implementing SPC requires an understanding of the process and analysis of the risk that any component will introduce variations. The methodology used by Statistical Process Control in data quality begins by defining the expectations or specifications gathered from the customers or clients. This is typically called as the assess phase in a data quality methodology. Once the baseline is set, the data elements are plotted against a time scale, their graph is analysed and the variation is measured against customer expectations. The critical data elements(CDE), which are our focus, should lie within 3 sigma levels either above or below the mean of the graph for a stable system. These are also called as the control limits, which help us understand if the system is under control or not. The plots which help us visualize this data are called control charts. For simplicity, we could categorize the data or CDE’s in specific segments, in which it is collected, as a service or product offered to the customer. While evaluating this data we only have to look at the segment where the variations exist rather than going through the entire data set. This helps us to troubleshoot and mitigate the variations at a category-specific level. After identifying the outliers from the control charts, their causes and subsequent impact on the system is analyzed. Modifications in the system are done to either incorporate or eliminate them according to the desired/expected results. These modified data are plotted and then 3 sigma limits are re-calculated. These new limits can be considered as the thresholds for the system. These threshold limits are verified against historic data (if available)or with the subject

matter expert inputs, and modified if required. These new limits serve as a model for controlling the system.

Tools and techniques used in the paper:

The following tools are used for Data Quality Monitoring purposes using Statistical Process Control:

Parameter Diagram - It is a typical process representation depicting the inputs, outputs and sources of variations (noise factors) for any data. This is a visual tool helping us understand the factors affecting the process. These factors are the primary sources of variation in the system. P-diagram helps us identify these.

Control Charts – It is a time-series run plot of the set of measurements along with historical statistical data and upper and lower control limits. Control limits are 3 sigma values above and below the mean respectively. Control charts help us identifying and classifying the variation in data. Variations can be attributed to one of the two causes: special causes and common causes. Special causes are typically caused by external factors and cause the data to vary beyond the control limits. Once identified, steps can be taken to make the system robust. Common causes are those that induce variations in the system within the control limits. These are typically system parameters, and their effect can be mitigated by re-designing the system. These help in depicting the predictability and stability of the process in a graphical form.

I-Chart - A type of control chart which deals with numeric data. These are used to determine the thresholds when CDE's are defined in numeric values, for example count data.

P-Chart - Represents the proportion of a particular parameter of the system on a time-scale. These are type of control charts which are used to determine the thresholds when CDE's are represented in percentages or proportions.

Thresholds - Thresholds are defined as the acceptable limits of measurement. Thresholds help identify CDE's that are outside the required specifications corresponding to one or more dimensions. The values beyond the threshold limits render the process out-of-control, and thus are looked into to make the process predictable and more stable. The limits are calculated using 3 sigma variations above and below the historical mean values. I-charts or P-charts can be used to take into account the variations in the data and thus redefine the new threshold limits.

Heat Map – It is a graphical method of representation of data where the individual data is depicted in cells with different color scheme. The color scheme helps in visually signifying the different levels of data quality. The cells may also contain different parameters of the data which will help in identifying the causes for deviation.

How subject of statistics has been utilized in this effort:

Statistics, as a subject, is sub-divided into 2 major parts - descriptive statistics and inferential statistics. Descriptive statistics deals with describing the acquired data in either graphical or analytical form. Inferential statistics deals with predicting data about the population based on sample observations, historical knowledge and Subject Matter Experts' inputs. Here, in this paper, both sub-divisions of statistics and their corresponding techniques are utilized for data quality methodology.

Descriptive statistics are used in the form of parameters of distribution of data, namely measures of central tendency like mean and measures of spread like range and standard deviation. Also, data are graphically represented with the help of control charts, p-charts, i-chart and heat maps. Inferential Statistics are utilized by the way of estimating the thresholds of a system by analyzing historical data or consulting Subject Matter Experts. We also assume the process to be under control after eliminating the out-of-control anomalies, and thus the data to be normally distributed.

How differently would you solve the problem given in the paper?

Statistical process control is a powerful tool for data quality monitoring, but can be made more robust when used in conjunction with techniques like pareto charts, cause-and-effect diagrams and defect concentration diagrams.

Pareto Charts are a visualization tool depicting the proportion of defective data caused by different types of errors. The type of error with the highest proportion can be prioritized as our immediate point of focus for process improvement. In our case, we have multiple types of possible errors such as demographic information, eligibility for different insurance types and the eligibility status of the policy. Therefore, from a pareto chart we can narrow down our focus into the specific type of error in our system and analyze accordingly.

Subsequently, after identifying the point of focus from a pareto chart, we can use a defect concentration diagram in order to ascertain the area that is most severely affected by the defect. Also, we can construct a cause-and-effect diagram to identify the potential causes for the defect and their interrelationship.

In our case study, for example we are getting eligibility errors as our major area of focus, then defect concentration charts will help us identify the time period (either start date or end date) of the insurance claim which are most severely affected. Furthermore, a cause-and-effect diagram can be utilized to understand the most-likely causes for this particular defect. This helps us in understanding the root cause of the error and thus leads to faster resolution of the problem and process improvement.

Combining SPC with the above stated tools, we can further hasten the process improvement time and make the SPC process even more efficient and powerful.

Lessons Learned:

The paper helps us understand the topic of statistical process control and its applications to data quality measurements. The subject can also be extended to predict the data behaviour, thus helping us to automate the process. It puts light on data measurement tools such as I-charts and P-charts which in turn are used to determine the variable thresholds depending on the 3 sigma limits about the mean of the historical data or after consulting with the subject matter experts. Also this method can be used to measure the data quality which is served as inputs for deploying Artificial intelligence using self learning algorithms.We have identified the causes of variations in two categories. A common cause and special cause. We learnt that common causes are typically attributed to system parameters where as special causes are generally external factors affecting the system. Use of control charts goes a long way in identifying these anomalies and thus redesigning the system to become more robust.

2) Review and summarize non-parametric hypothesis test procedures for "single sample" and "two sample" cases. Give practical examples wherever possible

Two-Sample Nonparametric Hypothesis Testing

If we have a case where we cannot for certain say that our populations are (approximately) normally distributed, then we have a nonparametric hypothesis test case.

In the case of two sample nonparametric hypothesis testing, we use the Wilcoxon Rank-Sum test. We assume that the two samples have identical and independent distributions with the same spread but different locations.

Let X11, X12,...., X1n1 and X21, X22,....., X2n2 be two independent random samples of sizes n1 ≤ n2 from the continuous populations X1 and X2. We wish to test the hypotheses H0: μ1 = μ2, and Ha: μ1 ≠ μ2.

The test procedure is as follows. Arrange all n1 + n2 observations in ascending order of magnitude and assign ranks to them. If two or more observations are tied (identical), use the mean of the ranks that would have been assigned if the observations differed. Let W1 be the sum of the ranks in the smaller sample, and define W2 to be the sum of the ranks in the other sample. Then,

$W2$ = [(n1 + n2) * (n1 + n2 + 1) / 2] - $W1$

Now if the sample means do not differ, we will expect the sum of the ranks to be nearly equal for both samples after adjusting for the difference in sample size. Consequently, if the sums of the ranks differ greatly, we will conclude that the means are not equal. The null H0:μ1 = μ2 is rejected in favor of H1:μ1 < μ2, if either of the observed values W1 or W2 is less than or equal to the tabulated critical value W α (obtained from the table of the Critical Values for Wilcoxon Rank-Sum Test).

The procedure can also be used for one-sided alternatives. If the alternative is H1 : μ1 < μ2, reject H0 if W1 ≤ W α ; for H1:μ1 > μ2, reject H0 if W2 ≤ W α .

One-Sample Nonparametric Hypothesis Testing

For the case of one sample nonparametric hypothesis testing, we can use the Wilcoxon Rank-Sum test under the assumption that the data is symmetrically distributed about its mean. If we can't ascertain that the data is symmetric about the mean, and there is significant skewness in the data, then we use the 1 sample Sign Test to test our hypothesis. Under the hypothesis that the sample median (μ) is equal to some hypothesized value (μo , so H0: μ = μo), we would expect half the data set S of sample size n to be greater than the hypothesized value η o. If S > 0.5n then μ > μo, and if S < 0.5n then μ < μo. The SIGN TEST simply computes whether there is a significant deviation from this assumption, and gives you a p value based on a binomial distribution. If we are only interested in whether the hypothesized value is greater or lesser than the sample median (H0: μ > or < μo), the test uses the corresponding upper or lower tail of the distribution. The workings for calculating the confidence intervals (CI) get complicated to do manually as you have to use a technique called non-linear interpolation, therefore, we use MINITAB for the same. Basically, it is doing the same thing as in the parametric tests, setting a CI around

a measure of central tendency, however, because we're now dealing with discrete data it isn't always possible to calculate exactly the CI corresponding to our assigned α level: Non-linear interpolation is simply a method of getting as close to that value as possible.