



**The Association of System
Performance Professionals**

The **Computer Measurement Group**, commonly called **CMG**, is a not for profit, worldwide organization of data processing professionals committed to the measurement and management of computer systems. CMG members are primarily concerned with performance evaluation of existing systems to maximize performance (eg. response time, throughput, etc.) and with capacity management where planned enhancements to existing systems or the design of new systems are evaluated to find the necessary resources required to provide adequate performance at a reasonable cost.

This paper was originally published in the Proceedings of the Computer Measurement Group's 2006 International Conference.

For more information on CMG please visit <http://www.cmq.org>

Copyright 2006 by The Computer Measurement Group, Inc. All Rights Reserved

Published by The Computer Measurement Group, Inc., a non-profit Illinois membership corporation. Permission to reprint in whole or in any part may be granted for educational and scientific purposes upon written application to the Editor, CMG Headquarters, 151 Fries Mill Road, Suite 104, Turnersville, NJ 08012. Permission is hereby granted to CMG members to reproduce this publication in whole or in part solely for internal distribution with the member's organization provided the copyright notice above is set forth in full text on the title page of each item reproduced. The ideas and concepts set forth in this publication are solely those of the respective authors, and not of CMG, and CMG does not endorse, guarantee or otherwise certify any such ideas or concepts in any application or usage. Printed in the United States of America.

Did Something Change?

Using Statistical Techniques to Interpret Service and Resource Metrics.

Frank Berezney
Kaiser Permanente

In a perfect world, one would always know the answer to that question. Unfortunately, nobody works in a perfect world. This paper will explore statistical techniques used to look for deviations in metrics that are due to assignable causes as opposed to the period to period variation that is normally present. Hypothesis Testing, Statistical Process Control, Multivariate Adaptive Statistical Filtering, and Analysis of Variance will be compared and contrasted. SAS code will be used to perform the analysis. Exploratory analysis techniques will be used to build populations for analysis purposes.

Introduction

This paper attempts to address two questions: (1) Are Statistical Techniques an effective method to interpret instrumentation data? and (2) How should instrumentation data be organized for this analysis and interpretation?

Statistical techniques are not new to CMG. Starting in the early 1990's there have been numerous papers addressing this subject, [Brey90], [Chu92], [Lipner92] and [Schwartz93]. This body of work seemed to cumulate with Jeff Buzen and Annie Schum's 1995 CMG Paper introducing Multivariate Adaptive Statistical Filtering (MASF) as a new statistical technique [Buzen95]. Interest in the subject seemed to decline from that point on, with the notable exception of Igor Trubin's set of papers on the application of MASF to many measurement and management areas [Trubin01], [Trubin02], [Trubin03], [Trubin04], and [Trubin05]. All of these papers are excellent treatments of the subject and are recommended reading.

Papers leading up to the MASF paper used existing statistical techniques and pretty much took the body of data to be analyzed as a given. The MASF paper is unique in that it introduces a new statistical technique as an alternative to Statistical Process Control (SPC). A two step method is introduced to overcome the conceptual problems of using interval based data with the SPC technique. Instrumentation data is organized into a reference set (a purposed subset or filtering of a data population) from which control limits are established. The reference set is used to analyze subsequent data populations to look for exceptions. MASF also provides a visual oriented reporting scheme for displaying the results of the analysis.

Four statistical techniques will be reviewed. A discussion about the complexity of time series data follows the presentation of the statistical techniques. These two subjects are then merged to provide an example. The paper finishes with a summary and conclusions.

Statistical Techniques

This section will describe four statistical detection techniques: (1) Hypothesis testing, (2) SPC, (3) MASF and (4) Analysis of Variance (ANOVA). But before this is done a couple of brief comments on what statistics is are in order. Many formal definitions exist, but for this author a three point statement captures the objective of the science:

- Populations have Parameters.
- Samples have Statistics.
- The science of Statistics is all about estimating Population Parameters by taking Samples and calculating Statistics.

The result of virtually all statistical methods is an estimate or a statement about a population parameter with a corresponding level of certainty. Detection techniques that are going to be described in this section all attempt to validate an assertion that a population mean has changed or is different from other population means based on an observed data sample.

A key, often overlooked, component of the analysis is a clear definition of the population you are trying to estimate. Many times in a Computer Performance Evaluation (CPE) study, the complete data population is present and it is incorrectly treated as a sample. The analysis is further complicated by the fact that CPE studies typically involve Time Series data and that provides additional issues when it comes to taking random samples and making predictions. Time Series data is discussed in a later section.

Hypothesis Testing

Hypothesis testing is a technique to determine if a population parameter is different from what it is expected to be. For example, assume the daily average arrival rate of a particular message is 15 per minute. The population in this case is a 24 hour period. Hypothesis testing can be used to test if the actual message rate during this 24 hour period is different from the expected rate.

In this example, the null hypothesis or expected rate is 15 messages per minute and the alternative hypothesis is anything greater or less than 15 messages per minute. The test would be stated in the following manner:

$$H_0 : \mu = 15$$

$$H_A : \mu \neq 15$$

A probability or confidence level is stated at 95%. It is also assumed the standard deviation of the population is not known and will be estimated from the sample data.

We will collect a random sample of 10 message rates throughout the 24 hour period, calculate the sample mean and the sample standard deviation, then compare these statistics against established limits for this particular case. If our sample data falls outside the established limit, then we can reject the null hypothesis and say with an accuracy of 95% the mean of the population is not 15. If we can't accept the alternative hypothesis, it does NOT mean we are proving in any way the average message volume is 15. We state the data or evidence is insufficient to reject the null hypothesis. In actuality, the real population mean could be 14.9, 15.1 or some other value real close to 15.

So to repeat this very important point, when we reject the alternative hypothesis, it does not mean we are accepting the null hypothesis as being true.

A random sample of 10 message volumes are collected for one minute intervals (13,14,16,11,16,15,12,16,12,14) randomly during the 24 hour period. These samples are combined to create an overall sample mean (13.9) and sample standard deviation (1.85).

The estimate of the population mean from the null hypothesis (15) is combined with the sample mean (13.9) and sample standard deviation (1.85). These values are used to create a t value of $-1.86 = (13.9 - 15)/(1.85/\text{sqrt}(10))$. The cutoff value for a two tailed 95% test with 9 Degrees of Freedom is 2.262. Since the calculated t value is inside the acceptance region, we can't reject the null hypothesis and conclude there is insufficient evidence to state the mean message volume is not 15. Figure 1 shows the t-Distribution for nine degrees of freedom and a 5% two tailed test.

t-Distribution with Nine Degrees of Freedom

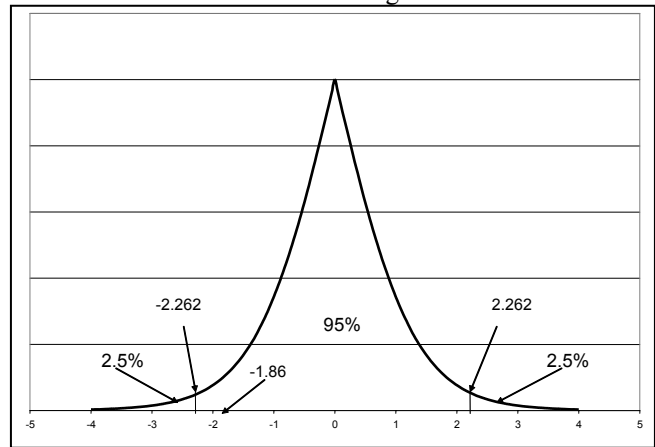


Figure 1

This type of a detection technique is useful if you have mandated SLA metrics or other situations where there is an expected value for a quality metric. It is very easy to calculate and only requires you have an a priori value for the null hypothesis and an agreed upon confidence level.

Statistical Process Control

When one thinks about SPC, two names immediately come to mind: Dr. Walter Shewhart and Dr. W Edwards Deming. Shewhart originally developed the concept of 'common' and 'assignable' forms of variation while he was working at Western Electric Company in the 1920's. He introduced the control chart as a method to visually track the behavior of a manufacturing process to proactively address defects. Deming was a colleague of Shewhart at Western Electric where learned and built upon Shewhart's concepts. Deming demonstrated the value of SPC during World War II while consulting with munitions factories, but he is far better known for his work in Japan after the war where he is credited for transforming the Japanese manufacturing industry into a world class model for quality and consistency. It was in Japan where Deming expanded the role of SPC to become a broader management strategy called the "Deming Cycle". SPC is occasionally used interchangeably with hypothesis testing and this creates considerable debate. William Woodhall's 2004 article, "Controversies and Contradictions in Statistical Process Control" [Woodhall00] is an excellent summarization of these issues.

SPC is conceptually similar to hypothesis testing, but computationally different. Both techniques use degrees of variation as a detection technique, but SPC does not require an a priori value null hypothesis value like hypothesis testing does. More importantly, SPC groups sampled data into rational subgroups and calculates metrics for the subgroups before calculating population level statistics. This organization of sampled data into subgroups is a key element of the process and permits SPC to effectively analyze complex manufacturing processes by analyzing

variation within and across subgroups. Creating rational subgroups also adds to the complexity of SPC and if done wrong can invalidate any conclusions drawn from a study. In addition to a standard deviation, range values are commonly used to create subgroup level metrics and overall thresholds. Generally speaking, SPC will create tighter control limits than a hypothesis test on the same body of data.

The following example will demonstrate the computational difference between a control chart and a hypothesis test. Hourly samples are taken across three shifts to measure the output rate (orders per minute) of an order processing application. Rational subgroups are formed for each 8 hour shift. For the purposes of this example assume an infinite backlog of orders and a uniform order size. Variation in output reflects the performance of the automated warehouse and / or the individuals operating this equipment. Figure 2 contains the hourly samples.

Sample Order Output										
Shift	Hour								Mean	Range
	1	2	3	4	5	6	7	8		
1	14	15	15	14	18	14	13	17	15	5
2	10	12	11	15	13	12	10	13	12	5
3	16	19	19	17	18	17	19	19	18	3

Figure 2

The overall average for the entire set of samples is 15 and the average range is 4.3. Using the formulas and factors from a textbook on Statistical Process Control [Wheeler92], the following chart, Figure 3, was constructed. The upper control limit is 16.62, the process average is 15 and the lower control limit is 13.38. Notice only three points are plotted, one for each subgroup.

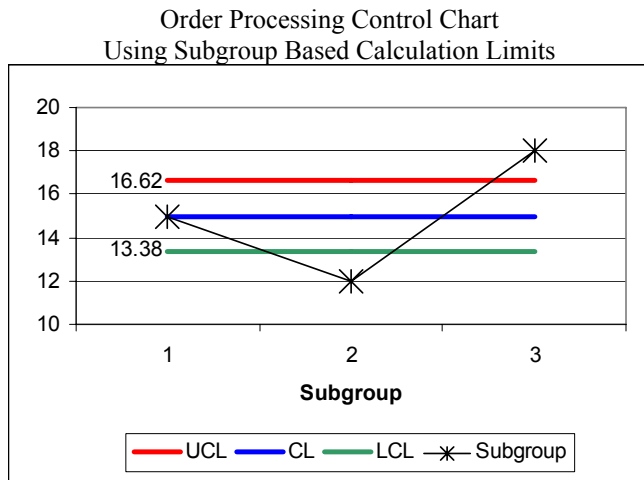


Figure 3

The example highlights the power and simplicity of an SPC control chart. Without any a priori information, this detection technique has identified two subgroups that are exhibiting assignable variation, i.e. they are out of control. This detection was made even though these outliers are part of the dataset that established the control limits.

Now the control limits will be calculated by treating the samples as a single dataset and a standard deviation will be used instead of the range value to measure variation. It is also assumed that the sample mean is the desired population mean. With this calculation technique the upper control limit increased to 18.08 and the lower control limit is reduced to 11.92. The two subgroups that were previously considered out of control are now considered in control. Figure 4 shows the revised control chart.

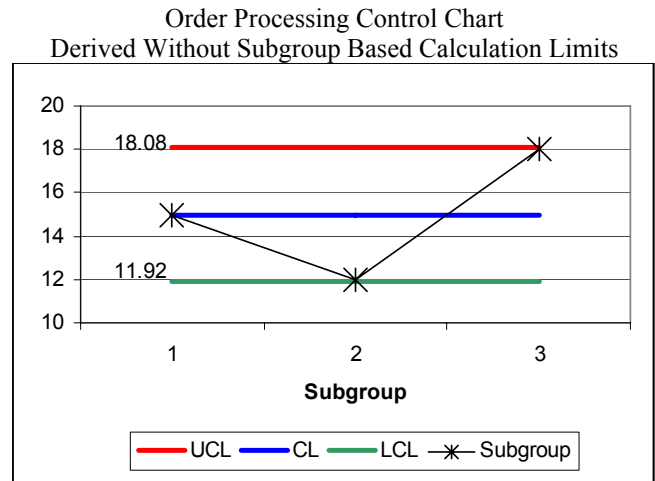


Figure 4

One can easily see from this example how the differences between hypothesis testing and SPC could fuel many papers in the academic community as to which one is the correct technique to detect change. Again, Woodhall's [Woodhall00] paper is an excellent treatment of this subject.

While this author is by no means an expert in this field, he tends to prefer SPC over hypothesis testing. The ability to take a body of data and subgroup it into rational subsets while at the same time calculating metrics for the overall body of data is superior to the method of treating it as a single body of data. Hypothesis testing could be used to analyze each shift, but then one could not make any statements about the overall body of data because you would be dealing with three distinct data populations in this case. Making statistical statements across data populations will be covered later with the Analysis of Variance technique.

As attractive as SPC is for detection testing, it is, unfortunately, an inappropriate technique for the interval based sampling data CPE people typically work with. This is one of the many findings of the MASF paper (SPC is discussed in Appendix A of the MASF paper) and a reason for the development of the MASF statistical technique which can be considered a hybrid of hypothesis testing and SPC.

MASF

In 1995, Jeff Buzen and Annie Shum introduced a new statistical technique for the detection of change due to assignable causes, MASF. It has similarities with hypothesis testing and SPC, but is based on a different conceptual framework.

MASF uses a period of ‘normal’ operation, instead of a random sample, to collect a set of measurements. Individual point measurements or rational subgroups can be used. The mean and variability of the point measurements or rational subgroups becomes the basis for filtering criteria to determine if subsequent ‘normal’ periods are the same or different.

What constitutes a ‘normal’ period? Workloads vary by time of day, day of week, week and month of year. To address this variability the concept of Adaptive Statistical Filtering was introduced. The authors used the term reference set to identify a period of repeatable behavior, i.e. a ‘normal’ period. In a statistical sense, they were defining a data population to be estimated by taking samples.

A 24 hour day, 7 day week creates a table of 168 separate hourly reference sets. Each cell in the table potentially becomes a separate process or data population to be estimated by cutting through 10 to 20 weeks of data. For example, Monday 8:00 AM data is accumulated for N weeks to become a period of ‘normal’ operation. A similar process is completed for Monday 9:00 AM data. Figure 5 provides a visual of what the full 24x7 table would look like. The numbering scheme for the process is arbitrary and could follow other patterns. A finer or coarser granularity could be also used depending on the nature of the workloads.

Table of 24x7 Hourly Data Populations

Day	Hour																							
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Mon	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
Tue	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48
Wed	49	50	51	52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69	70	71	72
Thur	73	74	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96
Fri	97	98	99	100	101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120
Sat	121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140	141	142	143	144
Sun	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160	161	162	163	164	165	166	167	168

Figure 5

Each measurement cycle (a week) provides one data point for each data population. To collect the recommended ten data points for estimation purposes would take ten weeks. Spanning ten weeks before making an estimation introduces other potential problems that will be discussed in the section on Time Series Data. Data sparseness and other reasons motivated the authors to combine cells within a measurement cycle. This creates more data points per measurement cycle and allows for quicker estimations to take place.

The 24x7 table can be reduced by clustering techniques that would identify different hours of a day or hours of the day across days of the week that have common characteristics.

This compression of the basic hourly, daily data would allow for a sample collection to be accumulated in a shorter timeframe. The MASF paper referred to the consolidation of data as Aggregation Policies. Figure 6 is a sample consolidation of 50 prime shift data populations into 14 reference sets. Note the consolidation can cross hour and day cells. This table is color coded to highlight the 14 reference sets.

Sample Prime Shift Reference Sets

Day	Hour									
	8	9	10	11	12	13	14	15	16	17
Mon	1	2	2	2	3	4	4	4	5	5
Tue	6	7	7	7	3	8	8	8	9	9
Wed	6	7	7	7	3	8	8	8	9	9
Thur	6	7	7	7	3	8	8	8	9	9
Fri	10	11	11	11	12	13	13	13	14	14

Figure 6

MASF does not impose any constraints on the definition of a reference set. In theory, one can have as many different reference sets as one wants: typically to reflect time periods, but also for holidays, end of a quarter, end of a year, etc. Also, you can even have a different set of reference sets for different metrics or workloads. But for pragmatic reasons, there are a limited number of reference sets for each environment one can practically work with

Once the ‘normal’ periods have been established, data collection can proceed and statistics can be derived. For detection purposes three standard deviations from the mean are used to establish the upper and lower limits. Note in the prime shift example above, the number of data points collected will vary by reference set or data population. For example, Monday at 8:00 AM (#1) is a unique hour for the entire week and one observation or rational subgroup would be collected per week. Tuesday through Thursday from 9:00 AM to 11:00 AM (#7) are considered common hours and in a single week nine data points or rational subgroups will be collected. Figure 7 shows the count of weekly data points by consolidated reference set (data population) and the calculated metrics for each. In this example, 15 minute response time values are averaged into hourly rational subgroups.

Plotting the mean and detection limits for the consolidated reference sets creates three separate control charts: (1) Monday – Figure 8, (2) Tuesday thru Thursday – Figure 9, and (3) Friday – Figure 10. These charts are presented to show how the mean and detection limits can change hour by hour and by day since we are treating each hour, or hour group, as a separate population to be estimated.

Weekly Response Time Metrics
Summarized by Reference Set / Data Population

Period	# of Obs per Week	Mean	Std Dev.
1	1	0.04	0.02
2	3	0.85	0.06
3	4	0.25	0.09
4	3	0.81	0.06
5	2	0.73	0.04
6	3	0.35	0.02
7	9	0.66	0.03
8	9	0.62	0.03
9	6	0.55	0.02
10	1	0.34	0.02
11	3	0.58	0.03
12	1	0.28	0.06
13	3	0.45	0.04
14	2	0.35	0.04
Total	50		

Figure 7

Response Time Mean and Detection Limits
Monday

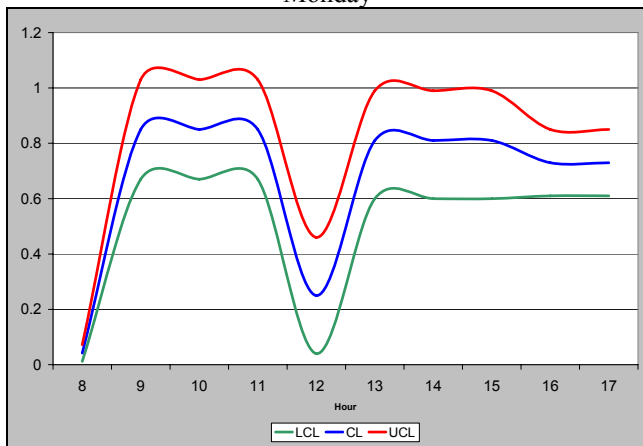


Figure 8

Response Time Mean and Detection Limits
Tuesday thru Thursday

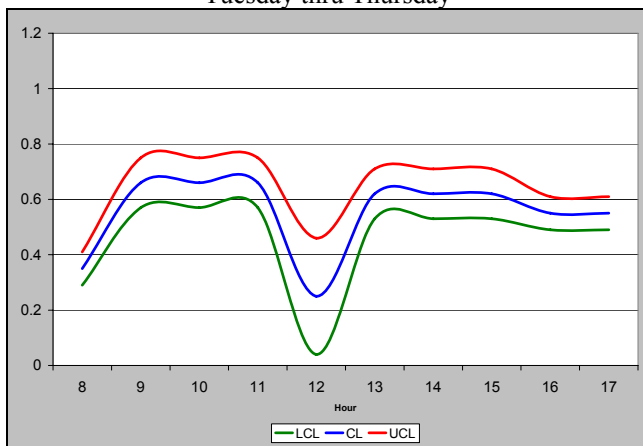


Figure 9

Response Time Mean and Detection Limits
Friday

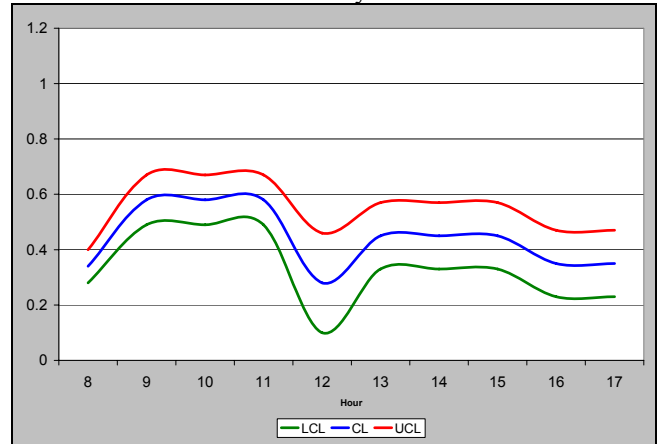


Figure 10

These three line plots are in essence SPC control charts for the respective days of the week. In fact, there are actually fourteen control charts embedded in the three line plots, one for each reference set or data population. Each hour, or hour group, creates a separate mean and set of detection limits.

Data collected in subsequent intervals can be compared to the detection limits to determine if it falls outside the range of normal variability and should be examined for assignable causes. This is the essence of statistical detection control. We have a certain amount of variability that is considered normal day to day fluctuations and is inherent in the operation of the process. When we exceed or go outside this range, there is a specific assignable cause which is responsible for the variation and that cause needs to be identified and corrected.

In addition to the detection capability, examining the charts allows one to make some overall observations about the characteristics of the workload. Monday experiences the highest response times and the greatest range of variability. The range between low and high detection limits Monday afternoon is .4 seconds. Response time improves Tuesday thru Thursday and variability is reduced to a .2 second range. Friday experiences further improvements in response time, but variability increases on that day.

This measurement framework is intended to be an N period rolling average process. The ideal collection interval will be between 10 to 20 points per reference set or data population. The aggregation policies determine how many measurement intervals are needed to collect the required sample size. It is not a good idea to span a lot of time intervals in a collection interval because other Time Series based factors will be influencing the data points and that distorts the detection process. Data used for this type of detection processing has a very short useful life, unless the workload being analyzed is not subject to any of the types of variability typically associated with Time Series data.

MASF provides a very robust statistical detection technique that addresses the variability and Time Series data challenges that are inherent in commercial computing workloads. It is more of a framework that a specific procedure to follow. For example, the definition of 'normal' periods can be date / time or possibly workload specific and are up to the user to determine. The choice of the measurement type, point estimate or rational subgroup and the measure of variability, either standard deviation or range are also a user prerogatives. This detection technique will require some programming on the part of the user, but it is well worth the effort.

ANOVA

The best way to describe ANOVA, or Analysis of Variance, is to explain why the technique was developed by Sir Ronald Fisher. In the late 1800's farmers were trying to determine the best combination of factors: watering, fertilizers, seed type, planting period, etc. to improve crop yield. They would take a large area of land and divide it up into multiple homogenous plots to conduct these tests. Each plot of land became a separate data population and was subject to a different set of factors called treatments. Crop samples were taken from each plot of land and statistics were created for each sample as an estimate of their population parameters. The farmer's question was, "Did any treatment create a significant difference in the yield of my crop?" Since each plot of land was a separate data population, the question was restated as, "Is there a statistically significant difference in the means of this set of data populations?" If there was, then the treatment associated with the unique highest producing plot of land would become the treatment used for future growing cycles.

ANOVA is a comparison across multiple data populations to make a determination if any of the means are different. The test would be stated in the following manner:

$$H_0 : \mu_1 = \mu_2 = \mu_3 \dots = \mu_n$$

$$H_A : \text{Not_all_}\mu_n\text{_are_equal}$$

A set of equal means represents the tentative assumption or null hypothesis. The alternative hypothesis is the statement we are trying to prove. The earlier discussion about hypothesis testing is relevant here. The purpose of the test is not to validate the means are equal. It is an attempt to prove that one or more are not equal at a stated level of confidence.

The calculation of the test statistic is a laborious effort and beyond the scope of this paper. Fortunately, many software products, including SAS and EXCEL have a basic ANOVA function and will generate it for you. The resultant test statistic is compared against an F distribution to accept or reject the alternative hypothesis. This is typically done by the tool and you get the result of this test as a probability value, p-value, that the means are equal. You reject the null hypothesis when the p-value is less than one minus your

stated confidence level. For example, if the confidence level is 95%, you accept the alternative hypothesis (i.e. reject the null hypothesis) whenever the p-value is less than 5%

One weakness of the ANOVA technique is the outcome of the test. If you accept the null hypothesis, end of discussion. The data is insufficient to prove a difference in this set of means. If you reject the null hypothesis and accept the alternative hypothesis, you can state one or more of the means are different, but you don't know which ones are different. Fortunately, some follow up work by John Tukey created a technique to place the means in groups that were statistically different from each other. This is known as the Tukey test. Combining ANOVA and the Tukey test allows an analyst to examine multiple data populations and group them into sets that are statistically different from each other.

The combination of these two tests is well suited to the interval based instrumentation data associated with computer workloads. Most traditional statistical techniques are not intended for longitudinal or time series based data because it can't be randomly sampled until the last interval has been captured. ANOVA answers this shortcoming by treating each interval as a separate data population and makes comparisons across data populations.

The following example will present a technique to look for changes in SMF recording options and possibly changes in the workload itself by using the volume of SMF dumps as a proxy. First there are a couple of questions that need to be answered regarding the volume of instrumentation data: (1) Are the days of the week the same? and (2) Is a day of the week repeatable across weeks?

While the interpretation of the meaning of this data is somewhat unorthodox, it does provide an excellent example of the difference between sampling for hypothesis testing and ANOVA analysis across data population. Recall that a data population is any body of data you wish to make an estimate about. In the case of this data, we are saying each day's logging data will be a population and we will count the number of SMF dump events as a proxy for the data volume. At the end of the day we have a number of dumps and we know with certainty what this number is. There is no sample or statistic being generated. We are starting with a population and we have its parameter. Now we want to ask the question, is the next period the same as this period? We could take today's count as an a prior mean value for a hypothesis test and sample tomorrow's rate to perform a hypothesis test and then repeat that process for each day of the week. This would provide a day to day comparison, but it could not be extended to any day other than the immediate neighbor day. One could also perform all possible combinations of hypothesis tests and draw some conclusions, but this technique increases the probability of error because each test introduces its own error margin and the errors would have to be aggregated when stating the

confidence of the test. ANOVA addresses both issues. A single test is performed and the margin of error for all tests combined is the specified confidence level.

Figure 11 is a table of the prime shift dump count frequency for the month of June for the LPAR being analyzed. The days of the week (Mon to Fri) will be compared to each other to see which ones should be considered as the same and which ones as different. Think of each day of the week as a separate treatment or data population and the question is stated as, "Is there a statistically significant difference between the mean dump values for the days of the week?" This will permit a potential consolidation of the days and then the reference sets can be used across weeks to detect any change. The individual days of the week will then be compared across weeks to look for same day in subsequent weeks that are considered statistically significantly different.

Daily Count of SMF Dump Events

Date	Day	Count
1-Jun	Wed	120
2-Jun	Thur	118
3-Jun	Fri	104
6-Jun	Mon	146
7-Jun	Tue	124
8-Jun	Wed	113
9-Jun	Thur	119
10-Jun	Fri	118
13-Jun	Mon	138
14-Jun	Tue	119
15-Jun	Wed	118
16-Jun	Thur	112
17-Jun	Fri	112
20-Jun	Mon	137
21-Jun	Tue	118
22-Jun	Wed	114
23-Jun	Thur	116
24-Jun	Fri	112
27-Jun	Mon	134
28-Jun	Tue	115
29-Jun	Wed	114
30-Jun	Thur	118

Figure 11

SAS software will be used to perform these tests. The following statements are used to invoke PROC ANOVA to perform the initial analysis to look for differences between the days of the week.

```
Proc ANOVA;
  Class Day;
  Model Count = Day;
  Means Day / Tukey;
Run;
```

The output of the procedure is included as Appendix A. It is divided into three sections: (1) Data, (2) Analysis of Variance and (3) Means.

The first section provides a summary of the class values and the number of data elements being used for the analysis. A quick review of this is in order to make sure the test is properly setup. The Day variable was listed as the Class or value to segment the data into populations. Five values, corresponding to the days of the week, were listed and there were a total of 44 observations read and processed. The SAS convention of using numeric values for the day of the week was used with Sunday = 1, etc.

The second section contains the result of the ANOVA analysis. The format is very similar to the sum of squares decomposition that is used for most linear regression packages. Many values are reported and have the same meaning they do in a least squares analysis. One of the key values is the Pr > F value on the right hand side. This is the p-value mentioned earlier. If this value is less than one minus the confidence level then the null hypothesis is rejected and you accept the alternative hypothesis. The p-value is .0425, which is less than 5% and we conclude at a 95% confidence level that one or more of the population means are statistically different from the rest. But which one(s) is / are different?

Section three contains the segmentation or grouping of the means. This is where the Tukey test is performed. All possible day to day comparisons are presented. Day pairs that are statistically significantly different are highlighted by three asterisks. In this example, Monday was determined to be statistically significantly different from Friday and all other day to day combinations are not considered different.

The days of the week were treated as separate populations and were compared to each other. The initial test concluded with a 95% level of confidence that all five means are not the same. Further examination of the means revealed that Monday and Friday were the pair of populations that have different means. The following table is typically used to display these results.

Mon	Tue	Wed	Thur	Fri

There is a certain degree of ambiguity here. It would be appropriate to group Tue to Fri as a group for the purposes aggregating populations and it would also be appropriate to aggregate Mon to Thur as a group of populations. Tue to Thur as considered in the same pool as Mon and Fri, but the differences between Mon and Fri are large enough to state they should not be in the same population pool. Even with

this ambiguity, ANOVA provides a very valuable technique to classify data populations.

A second test was performed by grouping the data into multiple populations which contain the same day of the week across multiple weeks. This time ANOVA is invoked five times, once for each day of the week. (by using a By Statement). The following SAS code was used to perform the second test.

```
Proc ANOVA;
  Format Date Date8.;
  Class Date;
  Model Count = Date;
  By Day;
Run;
```

This test compared a day of the week to the same day in subsequent weeks. All five tests accepted the null hypothesis. The p-values were: Mon - .9813, Tue - .9792, Wed - .9957, Thur - .9962 and Fri - .9403. There was insufficient evidence to conclude any of the population means were statistically different week over week for the same day of the week. We conclude there is insufficient evidence to say there are any changes week to week in the volume of SMF data being generated for the same day of the week. The result of one day's (Mon (2)) ANOVA test is included as Appendix B.

This is a very powerful analytic capability for classifying and segmenting recurring data populations. It can be used to combine multiple data populations into reference sets for MASF analysis and it can even detect change across time for a particular set of recurring populations. It should be part of every CPE analyst's toolkit for analyzing and classifying data.

Time Series Data

The instrumentation data used to perform statistical detection tests has a very short useful life. Most computer workloads exhibit repeatable weekly or monthly patterns that can be effectively analyzed by the detection techniques described in this paper. However, there are other factors influencing the workloads that fall into the category of Time Series factors and they tend to distort the patterns being used. This is especially true of resource metrics (processor utilization or I/O rates) which tend to mirror workload patterns more than service related metrics (response time or message delivery time).

Analysts who work with Time Series data decompose the data into four components: (1) Trend, (2) Cycle, (3) Seasonal variations and (4) Irregular fluctuations. The Trend component is a constant growth rate over the period being analyzed. The Cycle component is a multi-year business cycle which has a beginning and end or an increasing and decreasing phase. Seasonal variations are cycle like events that take place within a single year, such

as summer peak for vacation air travel. Irregular fluctuations are what the name implies, random events. One method to deal with these four influences is to remove three of the four Time Series components by decomposing the data. This subject is beyond the scope of this paper, but the interested reader is referred to Brian Barnett's 1991 CMG paper, "An Introduction to Time Series Forecasting for CPE" [Barnett91]. Tools are available to help with the decomposition and it is a good idea to be aware of these influences that affect Business Capacity Management decisions, but they should be removed from the more tactical data used for detection testing.

The recommendation is to keep your reference set within the 10 to 20 observations and try to stay within one quarter of the year with the data. That should tend to minimize Time Series influences for most commercial workloads. However, like all things in the CPE world, your result may vary.

Resource Utilization Example

The following processor usage data is for a midrange server that is dedicated to an OLTP workload. The CPU metric is collected every fifteen minutes and these point values are summarized into hourly rational subgroups. One month of data has been collected. Figure 12 is a table of the hourly data.

CPU Usage Metrics
Midrange OLTP Server

Date	Day	Hour								Daily Mean
		8.0	9.0	10.0	11.0	12.0	13.0	14.0	15.0	
3/1	Wed	34.7	35.1	29.4	27.9	27.9	27.9	27.9	27.8	29.8
3/2	Thur	29.0	28.7	28.1	28.2	27.9	27.9	28.4	27.9	28.2
3/3	Fri	36.0	30.5	29.6	29.4	28.4	27.9	28.4	28.1	29.7
3/6	Mon	29.4	28.9	30.2	28.7	27.9	28.1	28.0	27.9	28.6
3/7	Tue	35.2	29.3	28.2	28.2	28.6	27.9	28.1	28.5	29.2
3/8	Wed	36.5	31.9	30.6	32.4	31.9	36.1	30.3	30.3	32.4
3/9	Thur	37.2	31.5	29.9	31.0	36.6	31.3	28.4	29.7	31.8
3/10	Fri	27.6	27.9	27.9	27.7	28.0	27.7	27.7	27.8	27.8
3/13	Mon	30.9	30.2	29.9	30.3	29.6	29.3	29.0	27.9	29.6
3/14	Tue	34.9	32.2	33.0	33.9	31.9	31.5	33.3	31.5	32.8
3/15	Wed	35.9	33.3	35.8	36.2	34.6	30.6	37.9	28.6	34.0
3/16	Thur	34.6	33.8	28.0	38.9	33.9	30.5	28.6	33.7	32.7
3/17	Fri	31.2	33.3	31.8	31.0	30.7	30.7	29.6	28.3	30.8
3/20	Mon	30.1	28.3	28.4	28.4	28.7	29.4	28.4	28.8	28.8
3/21	Tue	28.1	28.1	28.8	28.2	28.0	30.1	30.6	28.1	28.8
3/22	Wed	37.8	28.9	28.1	28.1	28.8	27.9	28.6	28.0	29.5
3/23	Thur	38.2	28.7	28.7	33.5	42.0	30.6	27.9	28.0	32.1
3/24	Fri	27.9	28.5	28.1	33.3	27.9	28.0	28.1	27.9	28.6
3/27	Mon	30.8	33.4	32.3	28.9	28.8	28.2	28.8	28.3	30.0
3/28	Tue	32.3	29.0	28.2	30.0	34.9	33.6	32.4	43.8	33.2
3/29	Wed	28.5	29.3	28.8	28.7	28.8	28.2	28.3	28.3	28.6
3/30	Thur	33.2	31.4	34.8	31.3	28.2	28.2	28.3	28.3	30.4
3/31	Fri	28.4	28.4	28.2	29.4	28.4	28.2	28.4	28.3	28.4
Hourly Mean		32.4	30.4	29.9	30.6	30.4	29.5	29.3	29.5	30.3

Figure 12

The first three weeks (up to Friday March 24) will be used to create a reference set that will be used to test hourly values for the fourth week to see if they fall inside or outside the detection limits. PROC ANOVA was used to identify differences in the hourly mean averages to look for aggregation opportunities. The test reported there was

sufficient evidence to conclude the mean values are different at a 95% level of confidence (p-value = .0005 from the test).

The Tukey test was then used to look for differences. The output from the Tukey test is included as Appendix C and summarized here.

CPUAVE	33.1	30.8	30.7	30.5	29.7	29.6	29.4	28.8
Hour	8	11	12	9	10	13	14	15
	-----			-----				
	-----			-----				

Again, there is some ambiguity in the test. Hours 11, 12 and 9 can be grouped with hour 8 or they can be grouped with the other hours of the day. The 8 hour is 2.3 % greater than its neighbor and that is a much larger spread than any remaining neighbors. Based on the spread criteria, hour 8 will be treated as a unique hour and all other hours will be combined into a single reference set. The spread between the high to low value in the consolidated group is 2%, less than the spread between hour 8 and its neighbor. A similar test was performed on the days of the week and it was concluded that Monday and Friday would be grouped together and Tuesday, Wednesday and Thursday would be a second group. Combining the day and hour summarization criteria provided the following reference set construction for this example which is shown as Figure 13.

Reference Set Construction

	Hour								
Day	8	9	10	11	12	13	14	15	
Mon	1	3	3	3	3	3	3	3	3
Tue	2	4	4	4	4	4	4	4	4
Wed	2	4	4	4	4	4	4	4	4
Thur	2	4	4	4	4	4	4	4	4
Fri	1	3	3	3	3	3	3	3	3

For Midrange Server Utilization
Figure 13

These reference sets were used to summarize the usage data and create a table of detection limits, Figure 14.

Detection Limits
For Midrange Server Utilization

Reference Set	Days	Hours	Std Deviation	Mean	Upper Control Limit	Lower Control Limit
1	Mon & Fri	8	2.1	31.1	37.4	24.8
2	Tue,Wed,Thur	8	4.6	34.6	48.5	20.7
3	Mon & Fri	9 to 15	1.5	28.9	33.5	24.3
4	Tue,Wed,Thur	9 to 15	3.3	30.3	40.2	20.4

Figure 14

Figure 15 plots the actual values for Monday March 27 against the established detection limits. In this particular example, there are two separate processes being plotted, one for the 8:00 hour and one for the remainder of the day. This is why the reference lines are much more like a

traditional SPC control chart instead of the M shaped plots in the previous example. Based on this data, the 9:00 hour should be investigated for possible assignable causes for the increase in utilization. Similar plots can be created for the remaining days in the week.

This example was intentionally constructed to show aggregation possibilities. Most commercial workloads are much more volatile and it is very unlikely that eight hours of the day would be considered statistically similar. In practice, reference sets will typically have five to eight separate summarization groups per day.

Processor Utilization Mean and Detection Limits
Monday

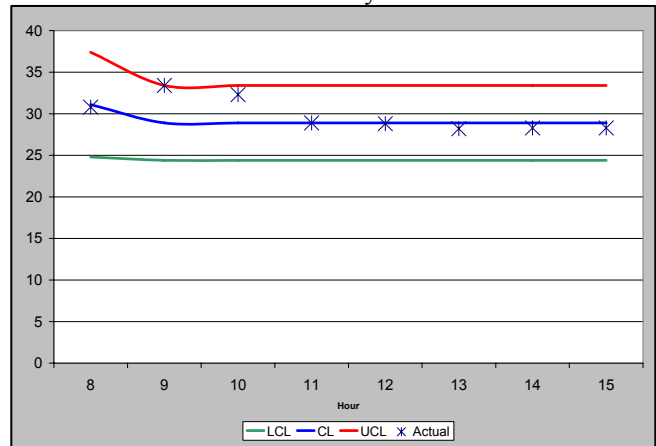


Figure 15

Summary and Conclusions

Statistical tools provide very power analytical capabilities to evaluate the behavior of repeatable processes. These same tools can help CPE analysts better understand and organize data. They should be part of every analyst’s toolkit. The best part is they are generally available as part of commonly used products, like Microsoft EXCEL and SAS. They just take some getting used to before they can be effectively and efficiently used.

The biggest challenge in this area is the complex and dynamic nature of the instrumentation data. The underlying assumption for most of the tests is the process is repeatable with some degree of normal variation. Considerable effort is needed to reformat most commercial instrumentation data to meet this requirement. There is no simple answer or procedure that one can run their data through to determine this. However, one of the ancillary benefits of this discovery work is a much better understanding of the workload being studied and this synergistically improves the ability of the analyst to manage it.

The evolution of SPC from analyst tool to management paradigm is history worth studying. Right now CPE analysts use statistical technique to log exceptions, much like SPC was used during WWII and at Bell Labs. The true value and benefit of these statistical techniques will not be

fully realized until it becomes a fundamental part of the Systems Management Framework and every detection exception becomes an incident record that is studied until a root cause analysis is completed. In order to reach that level of excellence innovative new techniques for workload characterization need to be developed.

Acknowledgements

In May of 2005, this material was presented at the Northern California Regional CMG meeting. The author is indebted to the members of that region, especially Cathy Nolan, Bill Jouris and Denise Kalm for their constructive feedback and suggestions for improvement. The author would also like to recognize the help of Brian Barnett to keep the statistics accurate and especially Annie Shum for making sure this discussion of MASF accurately describes the technique she and Jeff Buzen developed. All of these contributions made a very positive impact to the material in this paper.

References

[Brey90] J. Brey and R. Sironi, "Managing at the Knee of the Curve (The Use of SPC in Managing a Data Center)", *Proceedings of the Computer Measurement Group, 1990*.

[Chu92] L. Chu, "A Three Sigma Quality Target for 100 Percent SLA", *Proceedings of the Computer Measurement Group, 1992*.

[Lipner92], L. Lipner, "Zero-defect Capacity and Performance Management", *Proceedings of the Computer Measurement Group, 1992*.

[Schwartz93], R. Schwartz, "Adapting Statistical to the Management of Computer Performance", *Proceedings of the Computer Measurement Group, 1993*.

[Buzen95], J. Buzen and A. Schum, "Multivariate Adaptive Statistical Filtering (MASF)", *Proceedings of the Computer Measurement Group, 1995*

[Trubin01], I. Trubin and K. McLaughlin, "Exception Detection System, Based on the Statistical Process Control Concept", *Proceedings of the Computer Measurement Group, 2001*.

[Trubin02], I. Trubin, "Global and Application Levels Exception Detection System, Based on MASF Technique", *Proceedings of the Computer Measurement Group, 2002*

[Trubin03], I. Trubin and L. Merritt, "Disk Subsystem Capacity Management, Based on Business Drivers, I/O Performance Metrics and MASF", *Proceedings of the Computer Measurement Group, 2003*.

[Trubin04], I. Trubin and L. Merritt, "Mainframe Global and Workload Levels Statistical Exception Detection

System, Based on MASF", *Proceedings of the Computer Measurement Group, 2004*.

[Trubin05], I. Trubin, "Capturing Workload Pathology by Statistical Exception Detection System", *Proceedings of the Computer Measurement Group, 2005*.

[Woodhall00], W. Woodhall, "Controversies and Contradictions in Statistical Process Control", *Journal of Quality Technology, Volume 32, No. 4, October 2000*

[Wheeler92], D. Wheeler and D. Chambers, "Understanding Statistical Process Control", Second Edition SPC Press Knoxville, Tennessee 1992.

[Barnett91], B. Barnett, "AN Introduction to Time Series Forecasting for CPE", *Proceedings of the Computer Measurement Group, 1991*

Appendix A

The ANOVA Procedure

Dependent Variable: count

Class Level Information		
Class	Levels	Values
day	5	2 3 4 5 6

Number of Observations Read	44
Number of Observations Used	44

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2905.58182	726.39545	2.73	0.0425
Error	39	10358.60000	265.60513		
Corrected Total	43	13264.18182			

R-Square	Coeff Var	Root MSE	count Mean
0.219055	15.46776	16.29740	105.3636

Source	DF	Anova SS	Mean Square	F Value	Pr > F
day	4	2905.581818	726.395455	2.73	0.0425

Appendix A

The ANOVA Procedure

Dependent Variable: count

Note: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	39
Error Mean Square	265.6051
Critical Value of Studentized Range	4.04392

Comparisons significant at the 0.05 level are indicated by ***.				
day Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
2 - 3	12.750	-10.551	36.051	
2 - 4	16.000	-6.105	38.105	
2 - 5	17.200	-4.905	39.305	
2 - 6	26.250	2.949	49.551	***
3 - 2	-12.750	-36.051	10.551	
3 - 4	3.250	-18.855	25.355	
3 - 5	4.450	-17.655	26.555	
3 - 6	13.500	-9.801	36.801	
4 - 2	-16.000	-38.105	6.105	
4 - 3	-3.250	-25.355	18.855	
4 - 5	1.200	-19.641	22.041	
4 - 6	10.250	-11.855	32.355	
5 - 2	-17.200	-39.305	4.905	
5 - 3	-4.450	-26.555	17.655	
5 - 4	-1.200	-22.041	19.641	
5 - 6	9.050	-13.055	31.155	
6 - 2	-26.250	-49.551	-2.949	***
6 - 3	-13.500	-36.801	9.801	
6 - 4	-10.250	-32.355	11.855	
6 - 5	-9.050	-31.155	13.055	

Appendix B

The ANOVA Procedure

Day=Mon (2)

Class Level Information		
Class	Levels	Values
date	4	06JUN05 13JUN05 20JUN05 27JUN05

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	115.000000	38.333333	0.05	0.9813
Error	4	2845.000000	711.250000		
Corrected Total	7	2960.000000			

Number of Observations Read	
	8
Number of Observations Used	
	8

R-Square	Coeff Var	Root MSE	count Mean
0.038851	22.22439	26.66927	120.0000

Source	DF	Anova SS	Mean Square	F Value	Pr > F
date	3	115.0000000	38.3333333	0.05	0.9813

Appendix C

The ANOVA Procedure Tukey's Studentized Range (HSD) Test for CpuAve For Hourly Data Values from March 1 to March 24

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	136
Error Mean Square	7.759871
Critical Value of Studentized Range	4.35392
Minimum Significant Difference	2.8587

Means with the same letter are not significantly different.				
Tukey Grouping		Mean	N	Hour
	A	33.06	18	8
	A			
B	A	30.84	18	11
B	A			
B	A	30.73	18	12
B	A			
B	A	30.50	18	9
B				
B		29.67	18	10
B				
B		29.62	18	13
B				
B		29.39	18	14
B				
B		28.81	18	15