Mathematics of the Data Center

A Comprehensive Look at the Mathematics Used to Analyze Data Center Infrastructure

Data Center DNA Series

Jim Adams, Sirius Computer Solutions, Inc.

June 1, 2020

Contents

Introduction

This whitepaper is the first in a series of comprehensive guides directed at various infrastructure analytical methods that I collectively call ***Data Center DNA***. There are many areas that can be, or should be, analyzed within a data center. The obvious areas are storage, networks, and computing infrastructure but there are other areas that are important such as energy cost, cooling costs, total cost of ownership, availability and reliability measurements and usage consumption models.

This installment covers the statistical methods needed to analyze various areas of the Data Center. We will be taking a deeper look at data variability to gain valuable and useful insight from an infrastructure analysis. Understanding this variability and its associated measurements will allow us to determine future states of our infrastructure components. More importantly, the techniques imparted in this whitepaper can be transferred to any area of business or economics.

The material presented in this paper is meant to be light-hearted and entertaining all the while informative and useful. So, before we dive into the Mathematics of the Data Center let us first turn our attention to why this is important by understanding the Operational Definition.

Operational Definitions

One overarching concept that I am trying to portray in this paper is the notion of the Operational Definition[1]. To convey a thought or to describe some phenomenon we generally use empirical data which means whatever we are studying must be observable and measurable. If it is not observable and measurable, it is considered to be an opinion or a philosophy. Observable and measurable data points are called operational definitions. An infrastructure analysis will bring a considerable number of operational definitions to the conversation.

So many times, I see reports, emails and presentations that illustrate an idea or concept but there is no numerical data to support these ideas or concepts. This paper brings considerable light to the notion of defining things statistically in a manner that is understandable and genuine. The best example I can give here is a statement like, “I want to lose weight”. Well, to define this operationally, we should be more specific and say, “I want to lose 15 pounds in 45 days”. Now that means something by being more specific and certainly measurable. It still may not be achievable, but that is a different article altogether.

We see this type of problem a lot in conversations regarding data center systems and components. How many times have you heard someone say that they can reduce your operational cost or business risk? Well, those are nice assertions, but how much and when? 5%, 10%, 50%, and over how many months, years or even decades. Or the statement “buy this product and it will improve your I/O performance”. Okay, again, by how much? These are hopeful opinions and need to be operationally defined. Or perhaps it is a very broad comment like “our user community says their applications are running slow”. These are all examples where we need more operational definition to make sense of the statements.

We will extend this notion of operational definitions by applying specific statistical measures to the performance and capacity characteristics of systems in the data center[2]. These statistical measures are not difficult and will give your analysis more depth and credibility if applied appropriately. And these same specific statistical measures can be applied to all sorts of business areas such as Total Cost of Ownership and Business Economics. So, if you are ready to learn, let us dig into the *Mathematics of the Data Center*.

[1] *Thinking*. Gary Kirby and Jeffery R Goodpaster. (1995). Prentice Hall. P. 177

[2] *Probability and Statistics for Engineers*. 3rd Ed. Jay L. Devore. Brooks/Cole. (1991)

Mathematics of the Data Center

Any form of business or technology has its own numerical representations and vernacular. The industry of Information Technology and Data Centers are no exception. We use all sorts of numbers and acronyms to represent the state of many different systems and applications. Most of the numerical representations used in the Data Center are simple counting numbers and time-based numbers and there are many. Here are a few of the numerical measures that are used frequently in the IT industry.

- kVA (kilovolt-Amps)
- IOPS (I/O per Second)
- BTU (British Thermal Units)
- TCO (Total Cost of Ownership)
- NPV (Net present Value)
- ROI (Return on Investment)
- Gb/s (Gigabits per Second)
- TB/day (Terabytes per day)
- $/TB (Cost per TB)
- $/IOPS
- Latency
- Throughput
- Averages
- Growth Rates
- Cache Hit Rates
- MTBF

I am going to give a shout-out to MS Excel™ right here. It will be your friend when doing all the calculations in this paper. I will refer to it as Excel and reference its various functions from here on.

Descriptive Statistics

Statistical measurements come in two flavors, specifically, descriptive statistics and inferential statistics. We can use both to learn how our infrastructure is performing. Descriptive statistics are those pesky means, modes and medians and some others we will discuss here. They are used primarily to describe past events. Inferential statistics are those calculations that are used to infer something based upon past values. These statistical measures can be used to describe events in the data center and can operationally define the current and future states of your data center infrastructure.

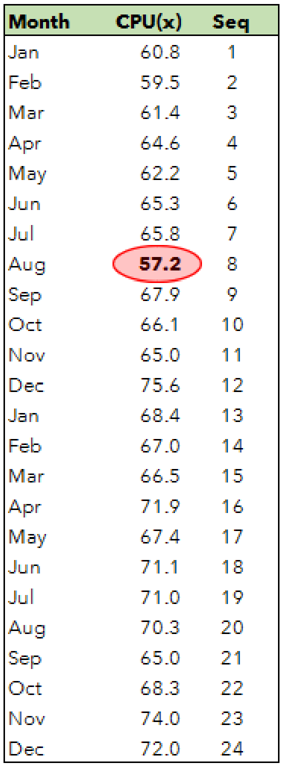
There are some basic statistical measures that I use a lot when analyzing data center infrastructure components. They are the mean, mode, median the standard deviation and the standard error. So, let's see how these fundamental statistical measures work and how we can leverage them in our data center analyses.

The overarching concept of what I want to explain in this section is understanding the variability in a set of data. Not understanding the variability of a data set can trip up any analysis process.

Most of us know what an average is and can generally calculate that on a set of numbers. But we want to understand a bit more about that type of measurement.

The Average

If you are reading this you more than likely know what an "average" is because we learned about it in high school and it is used daily in reports, newscasts, business and technology discussions across the planet.



Month	CPU(x)	Seq
Jan	60.8	1
Feb	59.5	2
Mar	61.4	3
Apr	64.6	4
May	62.2	5
Jun	65.3	6
Jul	65.8	7
Aug	**57.2**	8
Sep	67.9	9
Oct	66.1	10
Nov	65.0	11
Dec	75.6	12
Jan	68.4	13
Feb	67.0	14
Mar	66.5	15
Apr	71.9	16
May	67.4	17
Jun	71.1	18
Jul	71.0	19
Aug	70.3	20
Sep	65.0	21
Oct	68.3	22
Nov	74.0	23
Dec	72.0	24

Figure 1 – Data Set-1

The average is also known as the "mean" of a set of numerical data points. This is where one sums the data points and then divides the sum by the number of data points. In our example shown in Figure-1, we have 24 data points each representing the average CPU utilization per month for two years. But what does the average really describe about this data set. Additionally, can we better understand this set of data.

The average, or mean, of a data set, it is a commonly used statistical measure but there are supplementary ways to describe a data set. These supplementary measures help us further describe and understand what a data set truly is saying.

I used CPU Utilization in the data for Figure-1 since it is related to the data center but it could literally be anything such as miles per gallon, temperatures, distances, velocities, prices, economic values, or IOPS to name but a few. Also, this particular set of data is referred to as a Time Series since it is spread over a series of months.

To start with, I will graph the data points. This is helpful especially when there is a lot of data points. It is difficult for humans to look at hundreds or thousands of numerical data points and extract anything meaningful, unless you are Rain Man, of course. The rest of us need some other abstract medium to analyze many data points. The graphical view of a data set will show outliers and trends and perhaps clusters. Shown in Figure-2 is a line graph of the data points in Figure-1.

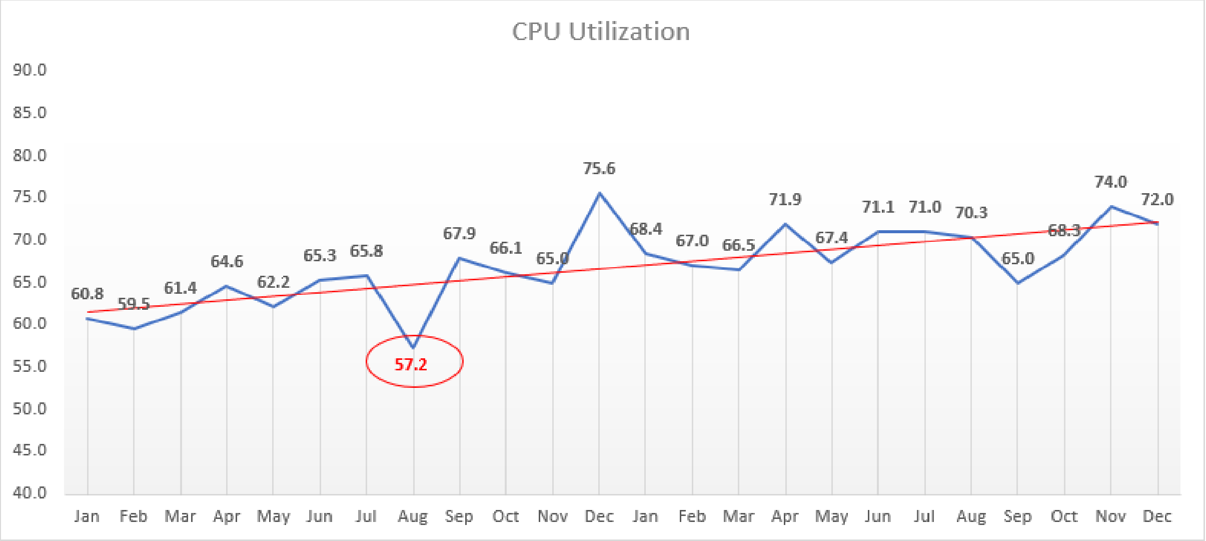


Figure 2 - Graph of Data Points

What can we see from this graph? Well, for starters there is a definite increase in the data values over time. This is illustrated by the red trendline plotted through the data points. This is a Linear Regression line which we will cover later. Most of the data points seem reasonable but wait, what is that 57.2 value doing there? Is this a mis-typed value or some legitimate data point we should consider? This is where further questions need to be asked or an understanding of the data points and business you are analyzing. In our case, we had a business shutdown for two weeks in August and therefore a reduction in overall computer utilization. We can choose to disregard this value or retain it. Let us retain it for now.

Most of our data points all seem to float nicely around that red trendline. What we want to know is how nicely are they floating around that trendline and can we use that trendline to predict some future state. Well yes, we can. But first, we need to understand the descriptive statistics about this data set.

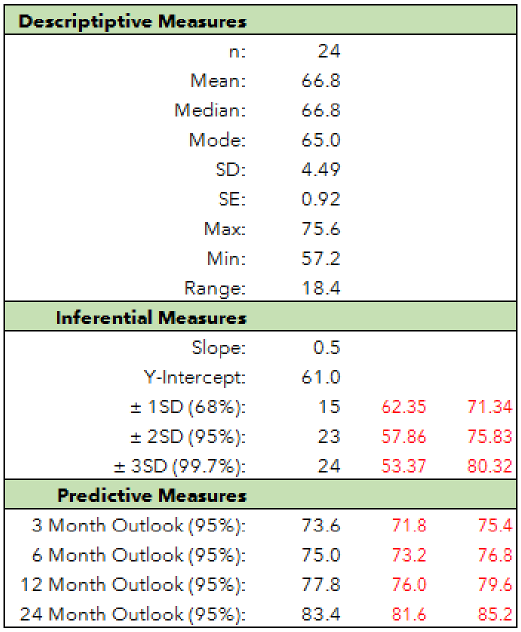
The basic statistical descriptive measures are going to be the Mean, the Median, the Mode, the Max, the Min, the Range and then the Standard Deviation and Standard Error are used to understand the variability of the data.

Shown in Figure-3 are all the measurements that I use, and we will cover them now with the Median as our next measure.

The Median

The "median" is the "middle" value in a sorted list of numbers. To find the median of a set of numbers, your numbers need to be listed in numerical order from smallest to largest, so you may have to rewrite your list before you can find the median. Excel works great for this. If we have an odd set of numbers, such as 25, we can simply pick the middle value, the 13th value, of this sorted set. If we have an even set of data, as in our case, we need to add the middle two values and divide by two the get the median value. Excel will do all this, splendidly, for you with no sorting and worrying about odd or even counts. Use the `MEDIAN()` function in Excel.

Referring to Figure-3, we can see how the median value compares to the mean value. Notice that the mean and the median values in our data set are equal at 66.8. This tells us something about the data. When the mean and the media are equal, or very close, that means the spread, or variability of the data is not large.



Descriptiptive Measures			
n:	24		
Mean:	66.8		
Median:	66.8		
Mode:	65.0		
SD:	4.49		
SE:	0.92		
Max:	75.6		
Min:	57.2		
Range:	18.4		
Inferential Measures			
Slope:	0.5		
Y-Intercept:	61.0		
± 1SD (68%):	15	62.35	71.34
± 2SD (95%):	23	57.86	75.83
± 3SD (99.7%):	24	53.37	80.32
Predictive Measures			
3 Month Outlook (95%):	73.6	71.8	75.4
6 Month Outlook (95%):	75.0	73.2	76.8
12 Month Outlook (95%):	77.8	76.0	79.6
24 Month Outlook (95%):	83.4	81.6	85.2

Figure 3 - Descriptive Measures

The Mode

The "mode" of a set of numerical values is the value that occurs most often in that data set. If no number in the list is repeated, then there is no mode for the list. Sometimes there are two modes in the case where there are pairs of similar values, which is called bimodal. Or trimodal if there are three. Excel will give you the mode of a set of numerical data points using the `MODE()` function.

> *Note: as you can see, it would be exceedingly difficult to pick out the mode of thousands of data points. Excel is your friend here.*

In our data set in Figure-1 we have a mode of 65.0.

So, what does this tell us? Look at the Histogram in Figure-4. It contains a distribution of all the data points in groups with a range of three. From the Histogram one can see that most data points are in the range of 65.2 and 67.2 and the remainder of the data points are spread out similarly almost like a bell curve. This is a good thing. In many cases we do not get a bell curve but some other curve. Nevertheless, the curve tells us something. In our case, the mean, median and mode all being close indicates a tendency towards a bell curve. In fact, the mean and median of 66.8 and the mode of 65 all fall in the tallest bar with the 5 at the top.

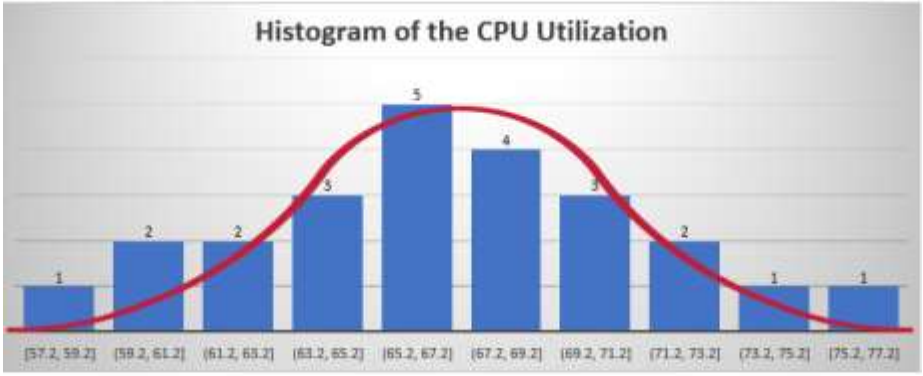


Figure 4 - Histogram of Data Set 1

Our outlier data point of 57.2 will affect the mean but not the median or the mode. So, we need a way to numerically represent how much variability is in our data set. The mean, the median and the mode tell us a lot about the variability but not the entire story. This is where the Standard Deviation and Standard Error come into play.

Median Home Prices. You have probably read where the median home price in Phoenix is $269,175. They use the median because that number is the middle-most value in the sorted set of hundreds of thousands of home prices. Scottsdale and Paradise Valley home prices skew the average home price, but the median more accurately reflects the most common home value.

Standard Deviation

The Standard Deviation (SD) is a statistical measure that helps explains the variability in your data set. It is the measure of how far each measure is, on average, from the mean of the entire data set. While there are many ways to measure variability within a set of data, one of the most popular is the *standard deviation*.

Looking back at Figure-1 and Figure-3, we have n=24 data points. The mean and median are both 66.8 and the mode is 65.0. But what does this really mean. Enter the Standard Deviation or SD. The Standard Deviation (SD) of our data set is σ = 4.49, where σ is called sigma. The Standard Deviation is cumbersome to calculate by hand but here again, Excel to the rescue. Use the Excel `STDEV.P()` function to get the standard deviation of a set of data. If the SD is zero, then all the data points are equal which would be a perfect case. I have never seen this happen but that is our baseline. Consequently, the larger the SD the more variability there is in the data set.

In Figure-5 below, I show our original data set with a blue shaded box overlapping the area of plus/minus one Standard Deviation. In this case the area between 62.3 and 71.3 represents 63% of our data points. It is generally expected that 68% of the data points will fall within plus/minus one SD but in our case, we don't have many data points to begin with at n=24 points. The expectation is that 95% of the data points will fall between plus/minus two SDs and that 99.7% will fall between plus/minus three SDs.

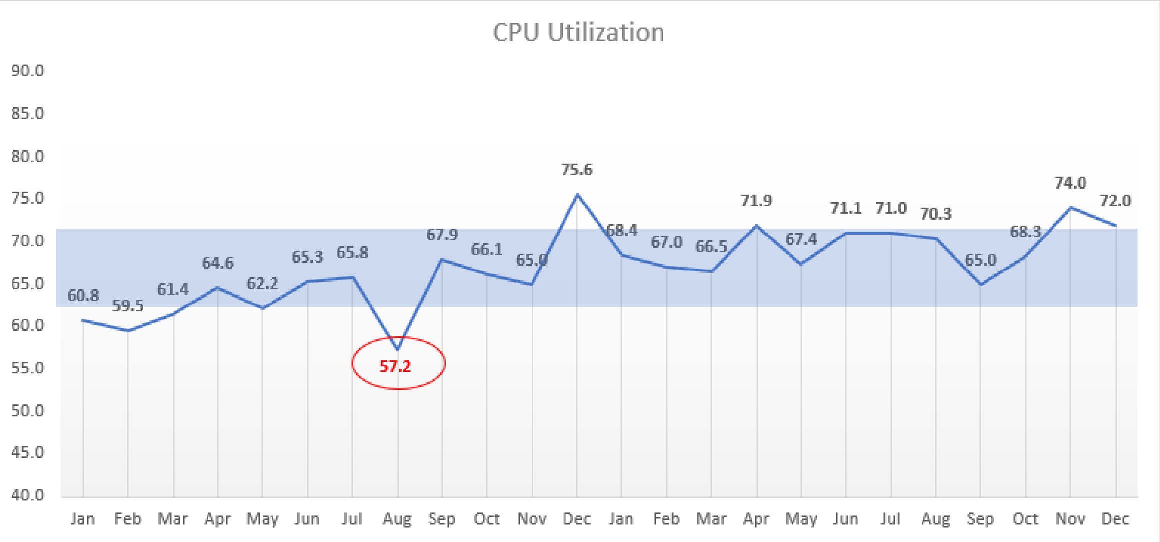


Figure 5 - Standard Deviation Analysis

Understanding the variability in the data is critical when doing predictive analytics on infrastructure components or when determining the confidence level of the variability within a data set. If there is a great deal of variability it is more difficult to predict with any accuracy a future value. We will look at this in a moment when predicting future CPU Utilization from our original data set.

Some may ask why not use the Average Deviation (AD). The Average Deviation is used in some industries such as the stock market, but it is not as popular as the Standard Deviation[3].

Range

Another key indicator of variability is the *range* of a data set. The range is simply the difference between the highest value and the lowest value in your data set. In our case, the range is 18.4. So, even if the average and the median are the same the range indicates a spread in our data set. To put a different spin on the range, if all the values in our data set were the same, the range would be zero, indicating no variability in the data set. There is no range function in Excel but you can get the range of a data set by subtracting the Excel `MIN()` value from the `MAX()` value. Easy as pie.

Six Sigma. A six-sigma process is one in which 99.99966% of all parts manufactured are statistically expected to be free of defects. Six Sigma represents six standard deviations.

Linear Regression

If you have been reading up to this point, we are getting to the heart of the statistical measure, as it pertains to analyzing data center infrastructure. We now will develop an equation of a line that will express the relationship between the dependent variable and the independent variable, CPU Utilization and Time in our case. The dependent variable can be nearly anything; IOPS, Utilization, Cost, TB, Throughput, BTU, MTBF, etc. The independent variable will usually be time in a data center analysis but in other applications can be anything. In other life-related applications these variables could be temperature and pressure, height and weight, test scores and ages, etc. We will continue with our Time Series analysis using CPU Utilization and Months, since that is the most popular pair of variables analyzed in an infrastructure assessment.

The objective in any infrastructure analysis is to determine when a particular asset will be at its limits in terms of I/O, CPU Utilization, TB used or Memory usage. Most infrastructure assets (storage, network and compute) have upper limits where performance may be degraded once that threshold is reached. Typically, it is 80% but each manufacture has its own specifications for upper limits. Most modern infrastructure assets will self-limit by throttling the data streams and I/O to keep the asset working but under-performing. The objective with a high-quality infrastructure analysis is to predict when this wall will be reached so one can either add hardware or reduce the data flows.

Figure-6 below shows our original data set that has the dashed red line illustrating the trendline through our data set. This is the Linear Regression line. There are non-Linear Regression lines, but they are complicated. Linear Regression works just peachy for what we are trying to accomplish.

The Linear Regression involves performing a lot of pedantic arithmetic, and Excel, once again, is our friend on this one. What we need is the equation of the regression line. The regression line is a line that best expresses the relationship between two variables, namely, the CPU values with the Month values, in our example.

Equation of the Line. When you add a trendline to a chart within Excel, there is an option to display the regression equation. This is extremely handy especially of you are wanting non-linear equations.

There are two ways to obtain the regression equation and Excel can help greatly with both. One way is to use the `SLOPE()` and the `INTERCEPT()` functions within Excel to get the slope-intercept equation for our trendline. See the sidebar titled "Equation of the Line" for the second method. The slope and the y-intercept are identified in Figure-

[3] https://www.statisticshowto.com/average-deviation/

3 and again in Figure-6. Once we have the slope and the y-intercept we can construct the equation of the line. The line equation we will use is called the slope-intercept and its formula is,

$$y = mx + b$$

where *m* is the slope and *b* is the y-intercept. Our x value will be a month (in numeric form of course) and y will be our predicted CPU utilization value. You can try this with the values shown in Figure-6. Take the first value for January from our data set, which would be 1.

$$y = 0.4673(1) + 61.004$$

$$y = 0.4673 + 61.004$$

$$y = 61.47$$

This is very close to the 60.8 illustrated in Figure-6 for January. Aiming now at month 24 on Figure-6 we get the following.

$$y = 0.4673(24) + 61.004$$

$$y = 11.21 + 61.004$$

$$y = 72.2$$

Which is very close to our 72.0 shown on the graph for December, in Figure-6.

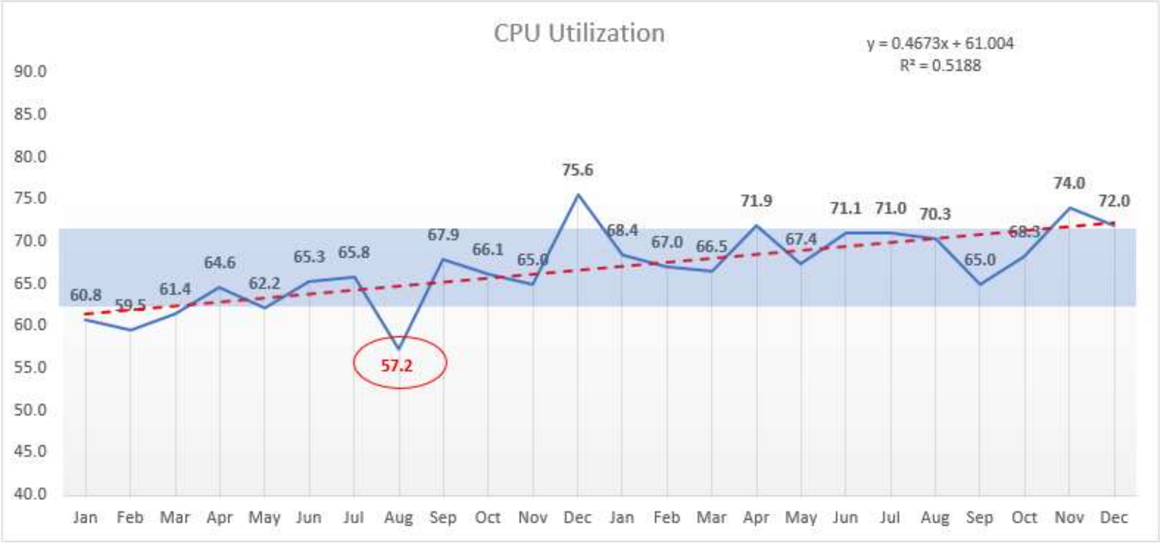


Figure 6 – Linear Regression Line

Looking Forward

The regression line can be used to estimate future values, say, one year beyond our last data point. We would add 12 months to the existing 24 months and use 36 in our line equation, like this.

$$y = 0.4673(36) + 61.004$$

$$y = 16.82 + 61.004$$

$$y = 77.8$$

In other words, we originally had 24 months of data and looking out another year means we are looking at the 36th month. So, we can expect CPU utilization to be around 77.8% in one year. I did this for you in Figure-3 for six months, 12 months and 24 months. One word of caution here. We cannot predict years into the future as we lose precision the farther into the future we look. There is a saying in the math and engineering community that goes like this, "We are statisticians, not psychics". Have you ever watched the weather channel and seen how the prediction interval widens when they predict the path of a hurricane? See Figure-7. This is because we can be more certain looking out one or two days but looking forward weeks has a lot of uncertainty, in terms of hurricanes. In terms of data center infrastructure, six months is fine or even a year. Anything more distant is looked upon with skepticism. We will discuss this prediction interval next.

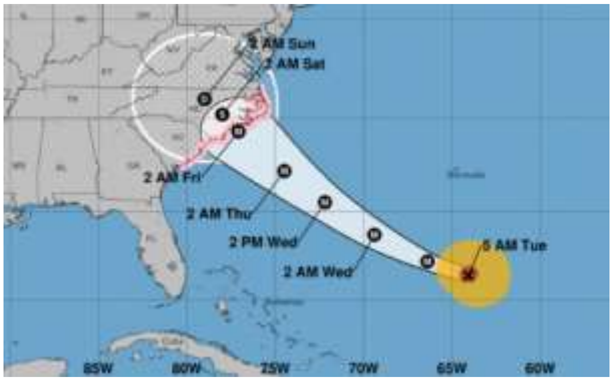


Figure 7 - Hurricane Prediction path

The newer versions of Excel have an option to display the regression equation right in the chart itself. Figure-6 shows this along with what we call r^2 (r-squared). The official terminology of r^2 is the Coefficient of Determination and it can be used to determine Goodness of Fit for our regression line.

Goodness of Fit

We need a way to determine how well the regression equation fits to our data set. Are the data points in close proximity to the regression line or are they spread out wildly around the trendline. The nearer the data points are to the line the more accuracy we have when predicting future values. The more variability we have in the data the less accurate will be our predictions. Let's turn our attention to the Coefficient of Determination and then the Coefficient of Variation.

Coefficient of Determination

R-squared, or the coefficient of determination, represents the strength of the relationship between two variables. It is a statistical measure of how well the regression line approximates the data values[4].

To obtain the coefficient of determination (r^2), we need first the correlation coefficient *r*. In our example in Figure-1, the correlation coefficient is 0.72031 (expressed as 72%) and can be found using the Excel `CORREL()` function. The correlation is a numerical measure of the strength of the relationship between two sets of data. Values near zero imply no relationship between the data sets and values near 100% imply a strong relationship. We can have both positive and negative correlations. Then we simply square this *r* value as r^2 which is equal to 0.5188 or 51.9% in our case. Usually, the closer the r^2 value is to 100% the better the regression model fits your observations. The closer you are to 100% the better. A perfectly straight line would have an r^2 value of 100% but I have never seen that in an infrastructure analysis. What the r^2 value is telling us is that 51.9% of the CPU Utilization can be explained by the Time value. This almost makes no sense on a Time Series but that is what it means. It generally makes more sense on a two-variable set like temperature and pressure or age and IQ, as two of many examples not using a time factor[5].

[4] *Probability and Statistics for Engineers*. 3rd Ed. Jay L. Devore. Brooks/Cole. (1991). p 453

[5] We could write an entire paper on correlation and its effectiveness and meaning. A strong correlation does not mean causation. It simply means a statistical relationship exists between two variables. I mention this as there are many silly stories of correlations than make no sense at all. Examples such as ice cream consumption leads to murder or a pirate shortage causes global warming are two of thousands. Common sense needs to be applied.

Coefficient of Variation

The coefficient of variation (CV) is the ratio of the standard deviation to the mean. The higher the coefficient of variation, the greater the level of dispersion around the mean. It is generally expressed as a percentage. Without units, it allows for comparison between distributions of values whose scales of measurement are not comparable, such as CPU Utilization and Time.

The Coefficient of Variation is useful when we want to compare results from two different data sets or that have different measures or values[6]. For example, if we are comparing the results from two analyses. If analysis A has a CV of 12% and analysis B has a CV of 25%, we would say that sample B has more variation in the data, relative to its mean. The lower the value of the coefficient of variation, the more precise our estimates will be.

The formula for the coefficient of variation is:

CV = (Standard Deviation / Mean) * 100.

In symbols: CV = $(\frac{SD}{mean})$ * 100.

Multiplying the coefficient by 100 is an optional step to get a percentage, as opposed to a decimal. There is no Excel formula for this calculation, so you need to do the division within a cell using the formula above.

I only use the Coefficient of Variation when doing a secondary analysis. Say, we do an infrastructure assessment in January and make some recommendations. The customer makes the recommended changes and we do another assessment. We want to know how the variability in assessment one compares to the variability on assessment two. We use the CV to accomplish this comparison.

The Accuracy of Averages

I forget where I read it, but I learned a few years ago that the mean and the linear regression explain the majority of data sets very nicely. In other words, the process of applying statistics to a set of data is a fairly wiggly process anyway and if we can apply the average and a linear equation to the data set, we are likely to see a good result. Sometimes we will get crazy data sets with extraordinarily wild outliers but for an infrastructure analysis, we don’t usually see that type of data fluctuation.

With that being said, then, just how accurate is an average? First, we need to find what is called the Standard Error of the average (SE)[7]. The Standard Error (SE) of the average is found by multiplying the square root of the total count of the data set (n=24 in our case) by the Standard Deviation of that same set, then that quantity is divided by the original count, noted like this:

$$SE_{average} = [\sqrt{count}\ x\ (SD)]/count$$

$$SE_{average} = [\sqrt{24}\ x\ (4.49)]/24$$

$$SE_{average} = 21.99/24$$

$$SE_{average} = 0.916$$

What this means is that our average for our data set is 66.8 plus or minus 0.92. So, we now have a bit of confidence on this average, from 65.9 to 67.8, with 66.8 landing right in the middle. This is within the 68% confidence interval.

[6] *Statistical Techniques for Business and Economics*. Robert Mason and Douglas Lind 8th Ed. (1993). Irwin. P. 137

[7] *Statistics*. 2nd Ed. David Freedman, Robert Pisani, Roger Purves, Ani Adhikara. Norton. (1991). p. 396.

If we wanted to be 95.7% confident o the average, we could use plus/minus 2SE's to get this interval of, 64.9 to 68.6. This is the average ± (2SE), thus we can say that we are 95.7% confident that our average is between 64.9 to 68.6.

Extenuating Factors

All this math is exciting, at least it is for me. But keep in mind there are many factors in the data center especially as we load up servers, networks and storage. The workload on these devices changes literally by the second. There is user think time, applications start and stop based on business demand. There are daily, weekly and month-end workloads to consider and even seasonal issues to consider. And, with the way hardware is dispersed, sometimes there are outages on a server or network component and transactional workloads are shifted elsewhere, automatically, to diverse and redundant components. Recently, we were plagued with the COVID-19 mess and that shifted workloads to different networks and changed the way some businesses operate. This caused all sorts of workload shifting to other networks, other times of day and workload scaling up and down depending on the nature of the business. We need to keep these business processing changes in mind as we analyze the data center infrastructure.

Limitations of Excel

I use MS Excel on a daily basis and for all practical purposes it does a splendid job. The only area where we get into trouble is if there are literally millions of data points. I don't see this problem often as most servers and other infrastructure components will overlay data after so many pre-configured days, usually 180 days or some other configurable time duration. If you find yourself staring at hundreds of thousands, or millions of data points, then you will need to use software such as R, C# or Python to analyze the data. The statistical topics discussed above still apply but a different tool set will be needed.

I recently had to write a C# program to process RV Tools output[8]. This tool creates hundreds of thousands of data points about VMware. Excel was choking on the sheer magnitude of data points, but software handled it with ease.

A Lesson Learned

Many years ago, when I was a young whippersnapper working in the data center, I thought I was being clever by making changes to improve performance on a large server complex. Our average response time for the online transaction system was 0.52 seconds per transaction with a SD of 0.06 seconds. This small SD deceived me. When I see 0.06 for a standard deviation it seemed super small compared to larger whole numbers. But the average transaction time was only 0.52 seconds so the 0.06 was a whopping 11.5% of the data set. That is a huge SD relative to the mean of the data itself.

My then recent configuration changes showed me we were getting 0.49 second response time the next few days. I was flying high and the ready to celebrate. In my statistically undeveloped brain, I thought I had made a huge improvement in the average response time. But wait. Once I looked at the original mean of 0.52 seconds plus and minus the SD, my 0.49 was still within the original variability of the data. Dang. In other words, there was no difference, statistically, in those two mean values. Therefore, I had not made a material change to anything. Swing and a miss!!

What happened was that I failed to check the Coefficient of Variation (CV). That would have showed me the 11.5%, easily. Lesson Learned.

[8] RV Tools Web Site. https://www.robware.net/rvtools/

Summary

I hope you enjoyed this first installment in my series of *Data Center DNA* articles, titled *Mathematics of the Data Center*. I wanted to lay a solid foundation for the remaining articles in this series. Mathematics - statistics in this case - is vital to understanding how data is presented and what data is really telling us. Whether it is infrastructure, particle physics, gaming, political surveys or economics, these statistical measures work the same. We need to understand the variability of the data before we can understand what the numbers are telling us.

The statistical measurements described in this report are not terribly difficult to perform and will give your analysis more depth and credibility if applied appropriately. I mean they are not terribly difficult if you leverage a tool such as Excel. If you want further information, I have several references in the last section that are helpful.

And, to close, being able to describe something accurately with numbers in such a way as to illustrate the uncertainty is always good, especially when it comes to technology, engineering and data center infrastructure. The methods presented in this article will get you one step closer to operationally defining your infrastructure. I hope you enjoyed this article.

Further Reading and Research

I would like to offer up some excellent books on Statistics and Systems Analysis. There are hundreds of books and web sites devoted to the subject of systems analysis and statistics and the ones that I used are listed alphabetically below. These are the books that I have examined extensively for my day to day work life and these are referenced throughout this report. I have enjoyed these books for many years.

Books Worth Reading

- *An Introduction to Error Analysis*. 2nd Ed. John R. Taylor. (1997). University Science Books.
- *Handbook of Statistical Methods for Engineers and Scientists*. Harrison M. Wadsworth. (1990). McGraw Hill.
- *Probability and Statistics for Engineers*. 3rd Ed. Jay L. Devore. (1991). Brooks/Cole.
- *Statistics*. 2nd Ed. David Freedman, Robert Pisani, Roger Purves, Ani Adhikara. (1991). Norton Publishing.
- *Statistics for Scientists and Engineers*. William Navidi. (2006). McGraw Hill.
- *Statistical Techniques for Business and Economics*. Robert Mason and Douglas Lind 8th Ed. (1993). Irwin.
- *Systems Engineering and Analysis*. 3rd Edition. Benjamin S. Blanchard and Woltor J Fabrycky. (1981). Prentis Hall.
- *Thinking*. Gary Kirby and Jeffery R. Goodpaster. (1995). Prentice Hall