Benchmarking Methodology

Alfred J. Barchi

ajb@ajbinc.net
http://www.ajbinc.net/

Introduction

Let's say you're a manager who is considering whether or not to invest $500,000 for a network appliance in order to improve the performance of your global network. I come to you and I tell you, "I've measured the performance improvement with and without the appliance installed, and the performance improvement is 64%." You might very well consider investing in the appliance.

However, suppose I said to you, "I've measured the performance improvement with and without the appliance installed, and with 95% confidence the performance improvement is 64%, plus or minus 62%." You would probably think entirely differently about making such an investment.

Yet, for many applications, particularly Internet-based applications, high variability in performance is normal. This can make meaningful comparisons, such as the one just described, extremely problematic. *The key is to understand the quality of the data*.

This is achieved through the use of confidence intervals. We all are exposed to confidence intervals every day. Just open the newspaper and read the results of the latest political poll. By convention, all political polls in the U.S. are conducted at a 95% confidence level, and, unless stated otherwise, have a margin of error of +/- 3%.

In order to perform meaningful benchmarking, it is necessary to understand how much confidence needs to be placed in the results, as well as the level of precision required in the results.

For example, when doing a qualitative analysis to evaluate alternative settings for a network appliance, where the cost of making a mistake is small and the precision required is minimal, very low confidence thresholds may be perfectly acceptable.

On the other hand, when a significant purchase is involved and it is necessary to have a more precise understanding of the cost/benefit tradeoffs, a much higher confidence level with much narrower confidence intervals may be required.

How to Construct a Confidence Interval

Confidence intervals are used to account for the variability in measurements. For example, we might obtain a confidence interval such that we can say with 95% certainty that the average value of a response time is between 5.5 seconds and 7.5 seconds. This is

important, since each time we take a measurement, the results will be different, and we need to understand how the data varies and what the normal range of measurements is.

In order to construct a confidence interval, we need at least a minimal amount of data. For certain types of variables that are independently identically normally distributed, we can construct a confidence interval off of a T distribution. However, many of the things that we measure in the IT world aren't normally distributed. This means that we need to construct the confidence interval off of a normal distribution. *In order for our results to be valid in this case, we need a minimum of 30 samples*. Generally, more samples is better, since more samples allows us to construct smaller confidence intervals at a given level of confidence, or conversely, to assign a higher level of confidence to a particular interval.

To construct a two-sided interval (which is what we would normally use), we calculate the following:

Lower Bound = sample mean – (Zvalue * SD) / sqrt(n)

Upper Bound = sample mean + (Zvalue * SD) / sqrt(n)

The 'Zvalue' is obtained by looking it up in a table (provided below). The 'sample mean' is the average value of our measurements. 'SD' is the sample standard deviation. 'n' is the number of samples. Looking at the equations above, we can see that as the number of samples increases, the size of the interval shrinks. As the variability of the data increases (higher SD), the interval widens. As the confidence level increases, the Zvalue gets larger, causing the interval to widen.

The upper and lower bounds can be calculated using a calculator and the mean and standard deviation of our samples. This can be done, for example, by loading the data into a spreadsheet, and built-in spreadsheet functions can the be used to calculate the mean and sample standard deviation, as well as the confidence interval itself.

How to Interpret Confidence Intervals

Essentially, any sample average that we measure within a confidence interval is statistically the same as any other value that we might measure within that interval. For example, in the example above, an average response time of 5.7 seconds would be indistinguishable from a response time of 7.3 seconds at the 95% level of confidence, since the entire difference could simply be due to normal random variation of our samples

We could reduce the size of the confidence interval until one of the averages lies outside of it, either by reducing the level of confidence or by increasing the number of samples taken. At that point, we could say that there is a statistically significant difference in the two values at that confidence level.

When we are using benchmarking to compare alternatives, there are three things that we want to keep in mind:

- First, if the confidence intervals for two alternatives overlap, then we cannot say that there is any statistical difference between the alternatives at that level of confidence.

- Second, we can only compare confidence intervals constructed at the same level of confidence.

- Third, when we make a comparison between two alternatives, we want to express it as a range of values, from most conservative to most optimistic. For example, if we had two confidence intervals: 5.5 seconds to 7.5 seconds, and 8.5 seconds to 12.5 seconds, we would express the amount of improvement of the first alternative vs. the second alternative as follows:

  8.5 / 7.5 – 1 = 13.3%

  12.5 / 5.5 -1 = 127.3%

And we would say that there was between a 13.3% to 127.3% improvement at that level of confidence. If we want to be more precise in our range, then we need to shrink the size of the confidence intervals, either by reducing the level of confidence or by increasing the number of samples that we measure.

Dealing with Highly Variable Data

There are some situations where the data that we measure is highly variable, with persistent outliers that increase the size of the standard deviation of our sample set. For example, in an informal study of the wireless network latency within a particular facility, a coefficient of variance of 2.0 was measured on one of the 100-sample ping tests. Other tests ranged from a C.O.V. of 1.6 to 1.8 (Coefficient Of Variance = Std. Deviation / Mean). This is extremely high. If we wanted to construct a 95% confidence interval that said that the average latency was X +/- 10%, we would need over 1,600 samples. To say the average is X +/- 5%, we would need 4 times as many samples.

The other thing is that when the range in C.O.V.s is high, and we would need much larger sample sets to get better consistency.

If we use a performance measuring tool to perform the measurements, we could program it to run until the desired interval is obtained. All that is required is to save each sample and re-compute the sample standard deviation periodically, compute the interval bounds, compare them to the sample mean, and stop when the interval is narrow enough.

This could be run as a background task overnight, for example, to periodically take samples until such time as our criteria are reached, or until we decide to terminate it prematurely and settle for less precision.

In this case, we need to keep a running total of the number of samples, a running total of the sum of each sample, and a running total of the sum of each sample squared.

The formula for the sample standard deviation is:

$$SD = \sqrt{\frac{\sum_{i=1}^{n} x_i^2 - n(\bar{x})^2}{(n-1)}}$$

Where *n* is the number of samples.

Zvalue Table For Use In Calculating 2-Sided Confidence Intervals

The following table gives Zvalues for commonly used two-sided confidence intervals. Other confidence levels can be obtained by consulting more extensive tables that may be available on the web, or by computing the quantile function for the normal distribution. For a two-sided 95% confidence interval, the quantile would be 0.975, i.e., 1 – (0.05 / 2).

Confidence Level	Zvalue
20	0.253
40	0.524
60	0.842
80	1.282
90	1.645
95	1.960
99	2.576