Performance Testing Plus: Do the Math!


Scott Barber

If you're like me, you barely squeaked by in whatever math class you took last. If you're like two of the best programmers I've ever worked with, you failed your last math class miserably, dropped out of school and got a job writing code. Or maybe you enjoy math and statistics, in which case I'm happy for you and encourage you to put that enjoyment to practical use when designing and reporting on software tests.

Whatever your particular situation is, I'm starting to feel like a math and statistics teacher. As a whole, it seems to me that members of software development teams, developers, testers, administrators and managers alike have an insufficient grasp on how to apply mathematics or interpret statistical data on the job.

As an example, I just finished another several-hour discussion with someone claiming to understand statistical principles who believed that a data set including five response-time measurements and a standard deviation roughly equal to the mean was statistically significant. The discussion reminded me that as performance testers, we not only must know and be able to apply certain mathematical and statistical concepts, we must also be able to teach them. Worse, we often have to teach these concepts to people who like math even less than we do. Over the years I've stumbled upon some relatively effective explanations for the mathematical and statistical principles I most often use as a performance tester. I'd like to share them with you.

Averages

Also known as *arithmetic mean*, or *mean* for short, the *average* is probably the most commonly used and most commonly misunderstood statistic of them all. Just add up all the numbers and divide by how many numbers you just added—what could be simpler? What most folks don't realize is that if the average of 100 measurements is 4, that could imply one quarter of those measurements are 3, half are 4 and another quarter are 5 (we'll call this data set A)—or it could mean that 80 of those measurements are 1 and the rest are 16 (data set B). If we're talking about response times, those two sets of data have extremely different meanings. Given these two data sets and a response time goal of 5 seconds for all users, looking at only the average, both seem to meet the goal. Looking at the data, however, shows us that data set B not only doesn't meet the goal, it also probably demonstrates some kind of performance anomaly. Use caution when using averages to discuss response times, and, if at all possible, avoid using averages as your only reported statistic.

Percentiles

Not everyone is familiar with what *percentiles* represent. It's a straightforward concept easier to demonstrate than define, so I'll explain here using the 95th percentile as an example. If you have 100 measurements ordered from greatest to least, and you count down the five largest measurements, the next largest measurement represents the 95th percentile of those measurements. For the purposes of response times, this statistic is read "Ninety-five percent of the simulated users experienced a response time of this value or less under the same conditions as the test execution."

The 95th percentile of data set B above is 16 seconds. Obviously this does not give the impression of achieving our five-second response-time goal. Interestingly, this can be misleading as well: If we were to look at the 80th percentile on the same data set, it would be one second. Despite this possibility, percentiles remain the statistic that I find to be the most effective most often. That said, percentile statistics can stand alone only when used to represent data that's uniformly or normally distributed and has an acceptable number of outliers.

Uniform Distributions

Uniform distribution is a term that represents a collection of data roughly equivalent to a set of random numbers that are evenly distributed between the upper and lower bounds of the data set. The key is that every number in the data set is represented approximately the same number of times. Uniform distributions are frequently used when modeling user delays, but aren't particularly common results in actual response-time data. I'd go so far as to say that uniformly distributed results in response-time data are a pretty good indicator that someone should probably double-check the test or take a hard look at the application.

Normal Distributions

Also called a *bell curve*, a data set whose member data are weighted toward the center (or median value) is a *normal distribution*. When graphed, the shape of the "bell" of normally distributed data can vary from tall and narrow to short and squat, depending on the standard deviation of the data set; the smaller the standard deviation, the taller and more narrow the bell. Quantifiable human activities often result in normally distributed data. Normally distributed data is also common for response time data.

Standard Deviations

By definition, one *standard deviation* is the amount of variance within a set of

Scott Barber is the CTO at PerfTestPlus. His specialty is context-driven performance testing and analysis for distributed multi-user systems. Contact him at sbarber@perftestplus.com.

measurements that encompasses approximately the top 68 percent of all measurements in the set; what that means in English is that knowing the standard deviation of your data set tells you how densely the data points are clustered around the mean. Simply put, the smaller the standard deviation, the more consistent the data. To illustrate, the standard deviation of data set A is approximately .7, while the standard deviation of data set B is approximately 6. Another rule of thumb is this: Data with a standard deviation greater than half of its mean should be treated as suspect.

Statistical Significance

Mathematically calculating *statistical significance*, also known as *reliability*, based on sample size, is not only beyond the scope of this column, it's just plain complicated. Luckily, you can get usually get away with skipping the math by applying some common sense. Since it's typically fairly easy to add iterations to your tests to increase the total number of measurements collected, the best way to ensure statistical significance is simply to collect additional data if you have any doubt about whether or not the collected data represents reality. Whenever possible, ensure that you collect at least 100 measurements from at least two independent tests. In support of the common-sense approach described below, check out this excerpt from a StatSoft, Inc. (www.statsoftinc.com) discussion on the topic from StatSoft, (www.statsoftinc.com), a company that provides analytic software:

There is no way to avoid arbitrariness in the final decision as to what level of significance will be treated as really 'significant.' That is, the selection of some level of significance, up to which the results will be rejected as invalid, is arbitrary. In practice, the final decision usually depends on whether the outcome was predicted a priori or only found post hoc in the course of many analyses and comparisons performed on the data set, on the total amount of consistent supportive evidence in the entire data set, and on 'traditions' existing in the particular area of research... But remember that those classifications represent nothing else but arbitrary conventions that are only informally based on general research experience.

While there's no hard-and-fast rule about how to decide which results are statistically similar without complex equations that call for volumes of data, try comparing results from at least five test executions and apply these rules to help you determine whether or not test results are similar enough to be considered reliable if you're not sure after your first two tests:

1. If more than 20 percent (or one out of five) of the test execution results appear *not* to be similar to the rest, something is generally wrong with either the test environment, the application or the test itself.
2. If a 95th percentile value for any test execution is greater than the maximum or less than the minimum value for any of the other test executions, it's probably not statistically similar.
3. If measurement from a test is noticeably higher or lower, when charted side-by-side, than the results of the rest of the test executions, it's probably not statistically similar.
4. If a single measurement category (for example, the response time for a specific object) in a test is noticeably higher or lower, when charted side-by-side with all the rest of the test execution results, but the results for all the rest of the measurements in that test are not, the test itself is probably statistically similar.

Usually, you can get away with skipping the math by applying some common sense.

Statistical Outliers

If we were to ask statisticians what an *outlier* is, they would tell us that it's any measurement that falls outside of three standard deviations, or 99 percent, of all collected measurements. The problem with this definition in our case is that it assumes that our collected measurements are statistically significant and are distributed normally—which is not nearly as common as we'd like for response times.

A more applicable definition of an outlier can be found in StatSoft's glossary:

Outliers are atypical (by definition), infrequent observations; data points which do not appear to follow the characteristic distribution of the rest of the data. These may reflect genuine properties of the underlying phenomenon (variable), or be due to measurement errors or other anomalies which should not be modeled.

Based on this definition, I recommend that if you see evidence of outliers—occasional data points that just don't seem to belong—re-execute the tests and compare them to your first set. If the majority of the measurements are the same, plus or minus the potential outliers, the results are likely to contain genuine outliers that can be disregarded, but if the results show similar potential outliers, these are probably valid measurements that deserve consideration.

The next question is, "How many outliers can we dismiss as 'atypical infrequent observations'?"

Assuming we've made the determination that we have collected a statistically significant sample of measurements, we can address this question. I submit that there is no set number of outliers that can be unilaterally dismissed, but a maximum percentage of the total observations should do as a rule of thumb.

If we apply the spirit of the two definitions that we have discussed, we come to the conclusion that up to 1 percent of the total measurements beyond the third standard deviation are significantly outside the rest of the measurements and can be considered outliers.

I hope you find this useful in educating the folks who view your results as to what those results truly represent, so they can make informed decisions about the application's performance.