BenchmarkXPRT

BenchmarkXPRT Development Community

WebXPRT 2013 results calculation and confidence interval

WebXPRT 2013 uses scenarios created to mirror the tasks you do every day to compare the performance of almost any Web-enabled device.

April 8, 2013

WebXPRT

Last Revision: April 8, 2013

TABLE OF CONTENTS

There has been some confusion about WebXPRT's use of a confidence interval. This white paper explains what a confidence interval is, how it differs from the run-to-run variability we frequently talk about with benchmark results, and how WebXPRT computes its confidence interval.

To supplement this overview, we have provided a spreadsheet[1] that reproduces the exact calculations WebXPRT makes when computing its results.

1 What is a confidence interval?

A confidence interval is a measure of how precise a measurement is. As an example, consider a poll taken by Poll by Truven Health Analytics and NPR in the period from August 1–12, 2012.[2] That poll found that found 52 percent of respondents believe wild-caught fish and seafood has more health benefits than other types of fish and seafood.

However, most people understand that there is a limit to how precise a poll can be. We have become used to hearing about the margin of error, which was plus or minus 1.8 percent for the seafood poll in question. What does that mean?

In this instance, the report does not state the confidence interval for the 1.8 percent margin of error. However, the most commonly used margin of error is the 95 percent confidence interval.[3] In the simplest terms, this means that there is a 95 percent chance that between 50.2 percent and 53.8 percent of consumers believe wild caught fish and seafood has more health benefits than other types of fish and seafood (52 percent plus or minus 1.8 percent). Conversely, there is a 5 percent chance that fewer than 50.2 percent of consumers or more than 53.8 percent of consumers do not believe this.[4] While 95 percent is the most common confidence interval, you can report at other levels of confidence, such as the looser 90 percent level.

WebXPRT uses a 95 percent confidence interval. In concrete terms, if you see an overall score of 1,180 +/- 14, it means that if you ran the test 100 more times under identical conditions, 95 of the scores would fall between 1,166 and 1,194.

2 What is variability?

Although variability is less statistically rigorous than the confidence interval, it has stood the test of time. Ziff Davis used it in its benchmark testing 20 years ago. When using a well-tested, released benchmark, highly variable results usually point to a problem with the tests or the test systems. Having a quick, easily calculated way to check can be very useful.

---

[1] http://www.principledtechnologies.com/webxprt.com/WebXPRT%202013%20result%20calculation.xlsx
[2] http://media.npr.org/documents/2013/feb/sustainablefishing.pdf
[3] http://www.itl.nist.gov/div898/handbook/eda/section3/eda352.htm
[4] *Ibid*.

We use variability as a measure of how consistent the benchmark results are. For WebXPRT, runs executed under identical conditions should have overall scores within 10 percent of each other. The formula we use for this is (Maximum_result – Minimum_result)/Maximum_result ≤ 10%.

Note that, unlike the confidence interval, this is **not** plus or minus 10 percent. Using the example above, if 1,180 were the bottom of the range, scores could range from 1,180 to 1,311. If 1,180 were the top of the range, scores could range from 1,062 to 1,180. If 1,180 were in the middle of the range, the results could range from 1,118 to 1,242.

3 How does WebXPRT calculate its scores and confidence intervals?

In this section, we show exactly how WebXPRT computes its scores, starting with the raw data from a run. The data for this example comes from a Dell XPS 15 with an Intel Core i7-3632QM processor and 16 GB of RAM. It was running Windows 8 and IE 10. The results are as follows:

Face Detection (ms): 484.9 +/-4.96 (1.02%)
Stocks Dashboard (ms): 217 +/-5.51 (2.54%)
Photo Effects (ms): 407.4 +/-5.89 (1.45%)
Offline Notes (ms): 357.9 +/-13.56 (3.79%)
Overall score 1180 +/- 14

WebXPRT comprises four scenarios. During a test, it repeats those tests seven times. Table 1 presents the raw data from the test.

Iteration	Face Detection	Stocks Dashboard	Photo Effects	Offline Notes
1	488	**244**	410	346
2	484	219	408	370
3	488	221	397	360
4	483	213	410	372
5	487	210	416	374
6	474	215	401	341
7	490	224	410	342

Table 1. The time, in milliseconds, for each iteration of each scenario. The value in red is an outlier.

The value in red (244 at the top of the Stocks Dashboard column) is an outlier. While a small amount of variation is normal for any benchmark, there is always the potential for outliers to distort the result. In an attempt to avoid any distortion, WebXPRT excludes outliers from the result calculation. WebXPRT defines outliers as any value greater than the 75th percentile[5] plus 1.5 times the interquartile range.[6] Table 2 shows the results of the calculations for determining the outlier cutoff.

[5] For an explanation of percentiles, see http://en.wikipedia.org/wiki/Quartile.

	Face Detection	Stocks Dashboard	Photo Effects	Offline Notes
First quartile	483	213	401	342
Third quartile	488	224	410	372
Inter quartile range	5	11	9	30
Outlier cutoff	495.5	240.5	423.5	417

Table 2: The calculations for determining the outlier cutoff.

As we stated above, there is only one outlier in this test data, the value of 244 in the Stocks Dashboard data. All of the calculations below exclude the iteration that includes that value, and the calculations for the Stocks Dashboard scenario are based on a sample size of six rather than seven.

The scores for the scenarios are simply the mean of the iterations, excluding outliers. The plus or minus value is the confidence interval, as we have discussed. Because computing the confidence interval depends on the standard deviation, the table below shows the standard deviation for completeness. Table 3 shows the means and confidence intervals for the scenarios.

	Face Detection	Stocks Dashboard	Photo Effects	Offline Notes
Standard deviation	5.37	5.25	6.37	14.66
Mean average	484.9	217.0	407.4	357.9
95% confidence interval	4.96	5.51	5.89	13.55

Table 3: The mean and confidence interval for each scenario, which correspond to the score for each category.

Note: There is more than one way of computing a confidence interval. WebXPRT computes the confidence interval using the Student's T-distribution.[7] If you replicate these calculations in Excel, use the confidence.t function, not the confidence.norm function.

To calculate the overall score and its associated Confidence Interval, we begin by calculating the mean, variance, and standard error for each scenario. As with the other calculations, we exclude any outliers. We calculated the mean above, but Table 4 repeats it for your convenience.

	Face Detection	Stocks Dashboard	Photo Effects	Offline Notes
Mean	484.9	217.0	407.4	357.9
Variance	28.81	27.60	40.62	214.81
Standard Error	2.03	2.14	2.41	5.54

Table 4: The mean, variance, and standard error for each scenario.

---

[6] For an explanation of the interquartile range, see http://en.wikipedia.org/wiki/Interquartile_range.

[7] For an explanation of the Student's T-distribution see http://en.wikipedia.org/wiki/Student%27s_t-distribution.

WebXPRT uses these values to generate a normal distribution based on the data from the run and combines them to give the distribution for the run. The results of these calculations are too involved to reproduce here, but you will find them in the associated spreadsheet. Once it has the normal distribution for the run, WebXPRT can derive the overall score. In this case, the calculations yield 1,180, which matches the overall score WebXPRT gave for the run. Similarly, The 2.5 and 97.5 percentiles for the distribution give us the bounds of the 95 percent confidence interval, 1180 +/- 14.

4 Conclusion

We hope this paper and the associated spreadsheet have answered any questions you may have about how WebXPRT computes its scores. If you have suggestions about ways to improve the statistics in the benchmark, or if you have any other questions, please post on the community forum or e-mail us at benchmarkxprtsupport@principledtechnologies.com. For more information, visit us at www.benchmarkxprt.com and www.webxprt.com.