WebXPRT 3 results calculation and confidence interval

May 25, 2018

WebXPRT 3

BenchmarkXPRT

BenchmarkXPRT Development Community

Table of contents

Introduction 3

What is a confidence interval? 3

What is variability? 3

How does WebXPRT calculate scores and confidence intervals? 4

Conclusion 6

Introduction

This white paper explains calculations for the WebXPRT overall score and individual test scores and describes what a confidence interval is and how WebXPRT computes its confidence interval.

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

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 Gallup in the period from April 23–29, 2018.[2] That poll found that in the United States, 52 percent of adults would never want to use a driverless car.

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 3.0 percent for the poll about driverless cars, but what does that mean?

The Gallup poll states that the confidence interval for the 3.0 percent margin of error is the most commonly used confidence interval of 95 percent.[3] In the simplest terms, this means that there is a 95 percent chance that between 49.0 percent and 55.0 percent of drivers would never want to use a driverless car (52.0 percent plus or minus 3.0 percent). Conversely, there is a 5 percent chance that fewer than 49.0 percent of drivers or more than 55.0 percent of drivers 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, an overall test score of 156 +/- 3 means that if you ran the test 100 more times under identical conditions, 95 of the scores would fall between 153 and 159.

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 25 years ago. When using a well-tested, released benchmark, highly variable results usually point to a problem with the tests or the test systems. This makes it very useful to have a quick, easy way to check variability.

[1] https://www.principledtechnologies.com/benchmarkxprt/webxprt/2018/WebXPRT3_results_calculation_sheet.xlsx

[2] http://news.gallup.com/poll/234416/driverless-cars-tough-sell-americans.aspx

[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 156 were the bottom of the range, scores could range from 156 to 172. If 156 were the top of the range, scores could range from 141 to 156. If 156 were in the middle of the range, the results could range from 148 to 164.

How does WebXPRT calculate scores and confidence intervals?

In this section, we show exactly how WebXPRT computes scores, starting with the raw data from a run. The data for this example comes from a Microsoft Surface Laptop with an Intel Core i5-7200U processor and 4 GB of RAM, running Windows 10 S.[5] The results are as follows:

Photo Enhancement (ms): 346 +/- 5.33%

Organize Album using AI (ms): 2480 +/- 1.89%

Stocks Option Pricing (ms): 360 +/- 6.13%

Encrypt Notes and OCR Scan (ms): 2762 +/- 2.36%

Sales Graphs (ms): 473 +/- 2.37%

Online Homework (ms): 2169 +/- 1.66%

Overall score 156 +/- 3

WebXPRT comprises six scenarios. During a test, it repeats those tests seven times. Table 1 presents the raw data from the test sorted from high to low.

Iteration	Photo Enhancement	Organize Album using AI	Stock Option Pricing	Encrypt Notes and OCR Scan	Sales Graphs	Online Homework
1	375	2,568	462	2,852	492	2,194
2	367	2,510	387	2,843	487	2,193
3	356	2,502	385	2,789	473	2,187
4	340	2,474	351	2,764	472	2,186
5	334	2,445	351	2,717	464	2,183
6	332	2,435	344	2,685	462	2,151
7	321	2,426	339	2,682	462	2,087

Table 1. The time, in milliseconds, for the seven iterations of each scenario, sorted by time. The value in red is an outlier.

The value in red (462 on the top of the Stock Option Pricing 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. To avoid

[5] http://www.principledtechnologies.com/benchmarkxprt/webxprt/2018/details.php?resultid=32

any distortion, WebXPRT excludes the highest result from the result calculation if it is an outlier. WebXPRT defines outliers as any values greater than the 75th percentile[6] (3rd quartile) plus 1.5 times the interquartile range.[7] Table 2 shows the results of the calculations for determining the outlier cutoff.

	Photo Enhancement	Organize Album using AI	Stock Option Pricing	Encrypt Notes and OCR Scan	Sales Graphs	Online Homework
First quartile	333	2,440	348	2,701	463	2,167
Third quartile	367	2,510	387	2,843	487	2,193
Inter-quartile range	34	70	39.5	142	24	26
Outlier cutoff	418	2,615	446	3,056	523	2,232

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 462 in the Stock Option Pricing data. All the calculations below exclude the iteration that includes that value, and the calculations for the Stock Option Pricing scenario draw 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 each scenario.

	Photo Enhancement	Organize Album using AI	Stock Option Pricing	Encrypt Notes and OCR Scan	Sales Graphs	Online Homework
Standard deviation	19.9	50.5	21.0	70.4	12.1	38.9
Mean	346	2,480	360	2,762	473	2,169
95% confidence interval	18.43	46.75	22.07	65.12	11.21	35.93

Table 3: The standard deviation 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.[8] If you replicate these calculations in Excel, use the confidence.t function, not the confidence.norm function.

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

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

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

To calculate the overall score and its associated confidence interval, we use the mean 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.

	Photo Enhancement	Organize Album using AI	Stock Option Pricing	Encrypt Notes and OCR Scan	Sales Graphs	Online Homework
Mean	346	2,480	360	2,762	473	2,169
Standard error	7.5	19.1	8.6	26.6	4.6	14.7

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

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. Once it has the normal distribution for the run, WebXPRT can derive the overall score. The overall score is based on geomeans of ratios of individual scenario scores for the test system relative to those of the calibration system: an Apple iPad Pro 10.5" with an A10X processor, running iOS 11.1. For the calibration system for our benchmarks, we select a device that is popular with users for running the workloads represented in the benchmark. We calculate 2.5 and 97.5 percentiles for the distribution to give us the bounds of the 95 percent confidence interval for the overall score. These calculations are too involved to reproduce here, but you will find them in the associated spreadsheet.

In this case, the calculations yield an overall score and confidence interval of 156 +/- 3.

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 them on the community forum or e-mail us at BenchmarkXPRTsupport@principledtechnologies.com. For more information, visit us at BenchmarkXPRT.com and WebXPRT.com.

Principled Technologies®

Facts matter.®

DISCLAIMER OF WARRANTIES; LIMITATION OF LIABILITY:
Principled Technologies, Inc. has made reasonable efforts to ensure the accuracy and validity of its testing, however, Principled Technologies, Inc. specifically disclaims any warranty, expressed or implied, relating to the test results and analysis, their accuracy, completeness or quality, including any implied warranty of fitness for any particular purpose. All persons or entities relying on the results of any testing do so at their own risk, and agree that Principled Technologies, Inc., its employees and its subcontractors shall have no liability whatsoever from any claim of loss or damage on account of any alleged error or defect in any testing procedure or result.

In no event shall Principled Technologies, Inc. be liable for indirect, special, incidental, or consequential damages in connection with its testing, even if advised of the possibility of such damages. In no event shall Principled Technologies, Inc.'s liability, including for direct damages, exceed the amounts paid in connection with Principled Technologies, Inc.'s testing. Customer's sole and exclusive remedies are as set forth herein.