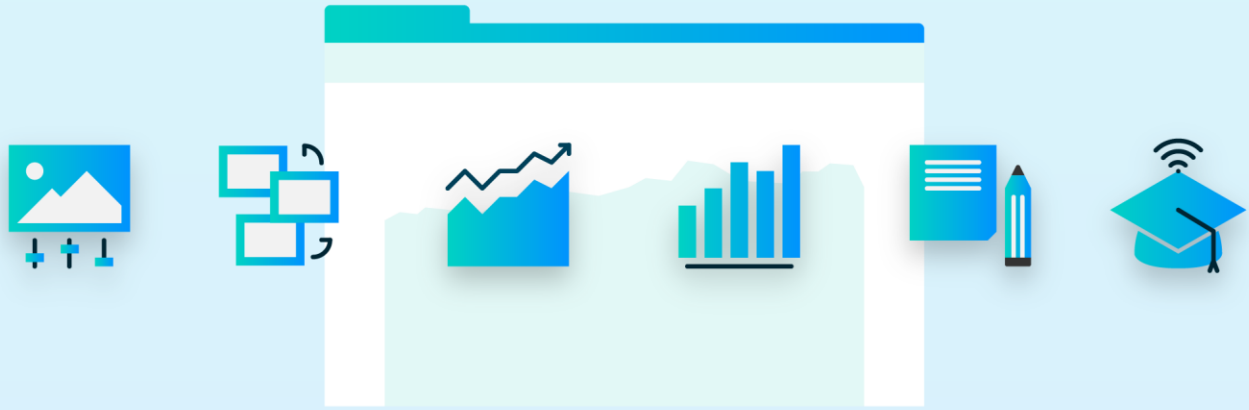
WebXPRT 4 results calculation and confidence interval

September 6, 2022

WebXPRT 4



BenchmarkXPRT

BenchmarkXPRT Development Community

Table of contents

Introduction

This white paper explains calculations for the WebXPRT 4 overall score and individual test scores and describes what a confidence interval is and how WebXPRT 4 computes its confidence interval.

To supplement this overview, we provide a spreadsheet[1] that reproduces the calculations WebXPRT 4 makes when computing its results.

Margins of error and confidence intervals

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 to 29, 2018.[2] That poll found that in the United States, 52 percent of adults would never want to use a driverless car. Most people understand that a poll or other statistic can be only so precise. We often hear about the margin of error, which is one part of measuring the precision of a statistic. The other part is the confidence interval.

For the Gallup poll we mention above, the margin of error is 4.0 percent and the confidence interval is 95 percent (which is the most commonly used confidence interval[3]). In the simplest terms, this means that there is a 95 percent chance that between 48.0 percent and 56.0 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 48.0 percent of drivers or more than 56.0 percent of drivers would never want to use a driverless car.[4] Note that 95 percent is the most common confidence interval,[5] but you can report at looser levels of confidence, such as the 90 percent.

WebXPRT uses a 95 percent confidence interval. In concrete terms, an overall test score of 133 +/- 3 means that if you ran the test 100 more times under identical conditions, there is a 95 percent chance that the mean of the 100 overall scores would fall between 130 and 136.

Variability

Another way of quantifying the accuracy of a statistic is variability, or how consistent results are when we run a test multiple times. 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/whitepapers/webxprt/WebXPRT-4-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.*

[5] http://www.itl.nist.gov/div898/handbook/eda/section3/eda352.htm

For WebXPRT, the overall scores for runs that a tester executes under identical conditions should fall 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 133 were the bottom of the range, scores could range from 133 to 148. If 133 were the top of the range, scores could range from 120 to 133. If 133 were in the middle of the range, the results could range from 126 to 140.

How WebXPRT calculates 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 Pro 6 with an Intel Core i5-8250U processor and 8 GB of RAM, running Windows 10 Home.[6] The results are as follows:

- Photo Enhancement (ms): 577 +/- 5.02%
- Organize Album using AI (ms): 2352 +/- 0.139%
- Stocks Option Pricing (ms): 210 +/- 1.94%
- Encrypt Notes and OCR Scan (ms): 1504 +/- 1.69%
- Sales Graphs (ms): 399 +/- 1.79%
- Online Homework (ms): 4035 +/- 0.64%
- Overall score 133 +/- 3

WebXPRT comprises six scenarios. During a test, it repeats those tests seven times. Table 1 presents the raw data from each workload sorted from better scores (lower numbers, representing less time) to worse scores.

Iteration	Photo Enhancement	Organize Album using AI	Stock Option Pricing	Encrypt Notes and OCR Scan	Sales Graphs	Online Homework
1	542	2,304	203	1,446	386	4,004
2	548	2,327	205	1,498	398	4,004
3	549	2,336	211	1,502	400	4,028
4	577	2,350	211	1,515	402	4,041
5	606	2,361	213	1,517	403	4,043
6	607	2,371	214	1,521	405	4,043
7	613	2,414	214	1,527	436	4,085

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

The value in red (436 on the bottom of the Sales Graphs column) is an outlier, which WebXPRT defines as values greater than the 75th percentile[7] (3rd quartile) plus 1.5 times the interquartile range.[8] While a small

[6] https://www.principledtechnologies.com/benchmarkxprt/webxprt/2021/details.php?resultid=164

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

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

amount of variation is normal for any benchmark, outliers can distort the result. To avoid any distortion, WebXPRT excludes the highest result from the result calculation if it is an outlier. 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	548	2,327	205	1,498	398	4,004
Third quartile	607	2,371	214	1,521	405	4,043
Inter-quartile range	59	44	9	23	7	39
Outlier cutoff	695.5	2,437	227.5	1,555.5	415.5	4,101.5

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 436 in the Sales Graphs data, which exceeds the outlier cutoff of 415.5. All the calculations below exclude the iteration that includes that value, and the calculations for the Sales Graphs scenario draw on a sample size of six rather than seven.

Table 3 shows the standard deviation, mean score excluding outliers, and confidence interval for each scenario. We include the standard deviation because it is necessary to compute the confidence interval.

	Photo Enhancement	Organize Album using AI	Stock Option Pricing	Encrypt Notes and OCR Scan	Sales Graphs	Online Homework
Standard deviation	31.3	35.3	4.4	27.4	6.8	27.8
95% confidence interval (ms)	+/- 28.97	+/- 32.64	+/- 4.08	+/- 25.36	+/- 7.15	+/- 25.70
Mean score	**577**	**2,352**	**210**	**1,504**	**399**	**4,035**
95% confidence interval (%)	**+/- 5.021%**	**+/- 1.388%**	**+/- 1.943%**	**+/- 1.686%**	**+/- 1.792%**	**+/- 0.637%**

Table 3: The standard deviation, mean score, and confidence interval values for each scenario. WebXPRT 4 reports the values in bold above on the end-of-test results screen, below the overall score.

Note: There are multiple ways of computing a confidence interval. WebXPRT uses the Student's T-distribution.[9] If you replicate these calculations in Excel, use the confidence.t function, not the confidence.norm function.

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

Table 4 repeats the mean score and provides the standard error for each scenario, which we use to calculate the overall score and its associated confidence interval. As with the other calculations, we exclude any outliers.

	Photo Enhancement	Organize Album using AI	Stock Option Pricing	Encrypt Notes and OCR Scan	Sales Graphs	Online Homework
Mean score	577	2,352	210	1,504	399	4,035
Standard error	11.8	13.3	1.7	10.4	2.8	10.5

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

WebXPRT uses the values in Table 4 to generate a normal distribution based on the data from the run and combines them to determine the distribution for the run. Once WebXPRT has the normal distribution for the run, it 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 MacBook Pro (13-inch, 2020) with an Apple M1 processor, running macOS Monterey 12.0.1. For the calibration system, we select a device that is popular with users for running the workloads 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 133 +/- 3.

Conclusion

We hope this paper and the associated spreadsheet answer 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.