User manual of the Speedup-Test software (version 2.0)

Sébastien Briais, Sid-Ahmed-Ali Touati, Julien Worms
University of Versailles ST-Quentin en Yvelines, France

January 7, 2010

1 Introduction

The present software package implements the Speedup-Test version 2.0 presented in [TWB10]. It aims at making a precise and rigorous statistical analysis of the performances of a set of benchmarks.

More precisely, the software determines whether there is a statistically significant speedup between two programs versions P_i (initial program) and P'_i (transformed program). We strongly advise to first learn about the Speedup-Test protocole presented in [TWB10] before using the current software.

The observed *speedup ratio* is computed by the formula

$$\rho_i = \frac{\texttt{ExecutionTime}(P_i, Input)}{\texttt{ExecutionTime}(P'_i, Input)}$$

Execution time measures are presented in Comma-Separated Values files, where each file corresponds to a benchmark and is composed of two columns of data, which correspond to observed time. The *execution time* of a program P is a chosen value within a set of observed execution times: the chosen value can be either the minimum time, or the mean time, or the median time.

The *overall gain factor* of the entire set of benchmarks is defined in [TWB10] as:

$$\begin{aligned} G &= \frac{\sum_{i=1}^{i=b} w(P_i)(\texttt{ExecutionTime}(P_i, Input) - \texttt{ExecutionTime}(P'_i, Input))}{\sum_{i=1}^{i=nb} w(P_i)\texttt{ExecutionTime}(P_i, Input)} \\ &= 1 - \frac{\sum_{i=1}^{i=b} w(P_i)\texttt{ExecutionTime}(P'_i, Input)}{\sum_{i=1}^{i=b} w(P_i)\texttt{ExecutionTime}(P_i, Input)} \end{aligned}$$

where $w(P_i)$ is the *weight* of program P_i.

The weight of program P_i can be any positive number. The only constraint is that the *total weight* $\sum_{i=1}^{i=n} \epsilon_i\, w(P_i)$ is not null.

Similarly, the overall speedup can be computed as follows:

$$S = \frac{\sum_{i=1}^{i=b} w(P_i) \times \texttt{ExecutionTime}(P_i, Input)}{\sum_{i=1}^{i=b} w(P_i) \times \texttt{ExecutionTime}(P'_i, Input)} \tag{1}$$

The execution time of a program (with a fixed input) can be set as the minimum, average or the median of the observed executions times.

If a is the number of benchmarks where a speedup can be observed, then the proportion of *accelerated* benchmarks is $C = \frac{a}{b}$.

As previously shown in [TWB10], it is possible to estimate the number of needed benchmarks b, in order to measure a confidence interval of this proportion with a desired precision r_0.

2 Prerequisites and installation

The present software package needs the R software [R D08] to be installed and executable in the `PATH`.

The present software package is composed of a shell script `SpeedUpTest.sh` and a R script `SpeedUpTest.R`.

These two files may be copied at any place, provided they are put both in the same directory. In the sequel, we assume that both these scripts are located in the directory `/usr/local/SpeedUp`.

Execution rights must be granted to the shell script, e.g. by doing

```
chmod +x /usr/local/SpeedUp/SpeedUpTest.sh
```

It may be convenient to symbolically link the shell script to a directory listed by the `PATH` environment variable, e.g. by doing

```
ln -s /usr/local/SpeedUp/SpeedUpTest.sh /usr/local/bin
```

3 Usage

The shell script may be invoked as follows:

```
SpeedUpTest.sh config.csv
               [--conf-level #value]
               [--weight (custom|equal|fraction)]
               [--precision #value]
               [-o report-file]
```

In the sequel, we explain the precise meaning of these flags.

3.1 Configuring the input of the statistical analysis

The file `config.csv` is a Comma-Separated Values file that describes the set of benchmarks to process. Data must be composed of exactly five fields:

1. the `Name` field refers to a string that is the name of the benchmark;
2. the `Sample1` field refers to the file name containing the measured execution times of the benchmark for the first program; The path of these file names are relative to the current path of the shell command;
3. the `Sample2` field refers to the file name containing the measured execution times of the benchmark for the second program; The path of these file names are relative to the current path of the shell command;
4. the `ConfLevel` field refers to a value (the confidence level) between 0 and 1;
5. the `Coef` field refers to a positive value used to compute the weight of the benchmark. In other words, the weights are the normalised coefficients. The relationship between the coefficients and the weights of a program P_i is simply $W(P_i) = \frac{Coef(P_i)}{\sum_j Coef(P_j)}$

Only the three first fields (`Name`, `Sample1`, `Sample2`) are mandatory. The other two may be left empty, or take the `NA` (Not Available) value. If left empty, the comma must still be present for validity of csv file format.

A valid configuration file is shown below.

```
Name,Sample1,Sample2,ConfLevel,Coef
"First benchmark","bench/bench1.1","bench/bench1.2",0.9,2
"Second benchmark","bench/bench2.1","bench/bench2.2",0.8,1.5
"Third benchmark","bench/bench3.1","bench/bench3.2",0.95,1
```

An other valid configuration file, that omits some values, is shown below.

```
Name,Sample1,Sample2,ConfLevel,Coef
"First benchmark","bench/bench1.1","bench/bench1.2",NA,
"Second benchmark","bench/bench2.1","bench/bench2.2",,
"Third benchmark","bench/bench3.1","bench/bench3.2",0.95,
```

3.2 Setting the confidence level

The confidence level to use is freely adjustable for each benchmark. The user simply needs to specify the desired value in the configuration file. Note that the confidence level is $1 - \alpha$, where α is the risk level used as parameter in all our statistical tests.

However, when the value of a confidence level is omitted or not valid (i.e. not between 0 and 1), a default confidence level may be used instead. This is precisely the purpose of the option flag `--conf-level` which takes in addition a value between 0 and 1.

If no default value is specified for the confidence level, then our software tries to find the best possible confidence level that allows to observe a speedup. Note that the value of the confidence level might be very low if no effective speedup can be statistically declared. To avoid such situation, if no speedup can be found with a confidence level greater than 50%, then a warning is emitted and the statistical test fails.

3.3 Setting the weights of the benchmarks

The weight of benchmarks can be individually set in the configuration file. However, it is possible to override these thanks to:

- `--weight equal` to set all benchmarks weights to be equal;
- `--weight fraction` to set benchmarks weights using the formula

$$w(P_i) = \frac{\texttt{ExecutionTime}(P_i, Input)}{\sum_{i=1}^{i=b} \texttt{ExecutionTime}(P_i, Input)}$$

- `--weight custom` to use custom weights of the configuration file (by default).

3.4 Confidence interval of the proportion of accelerated benchmarks

The confidence level used to compute the confidence interval of the proportion $\frac{a}{b}$ (proportion of the benchmarks with speedups vs. all benchmarks) is either the one specified by the command line, or if none is given, then the default value of 95% is taken.

In order to estimate the minimal number of benchmarks that is necessary to measure the proportion $\frac{a}{b}$ with a precision r, we must give the value of this r parameter by using the option `--precision #value`

By default, the precision is equal to 5%, corresponding to `--precision 0.05`.

3.5 Setting the report file name

The statistical analysis produces four separate files: the status file, the results file, the warnings file and the report file.

The status file logs the possible errors that might have happened during the analysis. If the analysis terminates successfully, it only mentions the elapsed time (of the R process).

The results file is a CSV file which is composed of several columns. These columns are:

1. `Name` which is the descriptive name of the benchmark;
2. `SpeedupMin` which is the speedup ratio, using the minimum execution time of P_i for $\texttt{ExecutionTime}(P_i, Input)$

3. `SpeedupMean` which is the speedup ratio, using the mean execution time of P_i for `ExecutionTime`$(P_i, Input)$

4. `IsMeanSignificant` indicates whether the observed speedup for the mean execution time is statistically significant (the value of this column is TRUE or FALSE or NA).

5. `MeanConfLevel` is the confidence level used for the statistical test of the mean speedup ratio. Its value may be NA if there is a problem with the input data.

6. `SpeedupMedian` which is the speedup ratio, using the median execution time of P_i for `ExecutionTime`$(P_i, Input)$

7. `IsMedianSignificant` indicates whether the observed speedup for the median execution time is statistically significant (the value of this column is TRUE or FALSE or NA).

8. `MedianConfLevel` is the confidence level used for the statistical test of the median speedup ratio. Its value may be NA if there is a problem with the input data.

9. `CoefMin` which is the actual coefficient used to compute the weight of the benchmark, in the gain factor formula, when `ExecutionTime`$(P_i, Input)$ is the minimum execution time of P_i

10. `CoefMean` which is the actual coefficient used to compute the weight of the benchmark, in the gain factor formula, when `ExecutionTime`$(P_i, Input)$ is the mean execution time of P_i

11. `CoefMedian` which is the actual coefficient used to compute the weight of the benchmark, in the gain factor formula, when `ExecutionTime`$(P_i, Input)$ is the median execution time of P_i

The warnings file contains all the warnings that accompany the statistical tests. The warnings can be the followings:

- `File 'xxx' is not readable`

  A sample file is not readable. The benchmark will be ignored.

- `Cannot process benchmark: samples unavailable.` This happens when at least one of the sample file is not readable.

- `Sample? too small for applying the Student's t-test (speedup of the mean).`
  `Please do more than 30 observations of the executions times.`

  One of the sample do not contain enough data to perform the statistical Student's t-test needed to analyse the average execution time.

- `Sample2 data are not normally distributed.`
  `The indicated confidence level for the speedup of the average execution time`
  `may not be accurate.`

  The statistical Student's t-test requires that the two samples follow gaussian distributions. When this is not the case but when the sample contains enough data (more than 30), then the statistical test can be used but the confidence level cannot be guaranteed.

- `Sample? too small for applying the Wilcoxon-Mann-Whitney's test (speedup of the median).`
  `Please do more than 30 observations of the executions times.`

  One of the sample do not contain enough data to perform the statistical Wilcoxon-Mann-Whitney's test needed to analyse the median execution time.

- `The two samples do not fit the location shift model law.`
  `The indicated confidence level for the speedup of the median execution time may not be accurate.`

The statistical Wilcoxon-Mann-Whitney's test requires that the two samples follow the same distribution upto a translation. When this is not the case but when the samples contain enough data (more than 30), then the statistical test can be used but the confidence level cannot be guaranteed.

- `Unable to find a confidence level greater than 50% to guarantee the statistical significance of mean speedup.`

  It was not possible to find a confidence level greater than 50% that guaranteed statistical significance of mean comparison (speedup of the average). It could simply say that no speedup occurs at all or that there is a problem with one (or the two) samples.

- `Unable to find a confidence level greater than 50% to guarantee the statistical significance of median speedup.`

  It was not possible to find a confidence level greater than 50% that guaranteed statistical significance of median comparison (speedup of the median). It could simply say that no speedup occurs at all or that there is a problem with one (or the two) samples.

The report file contains a precise report of the statistical analysis.

By default, if the input csv configuration file is named `inputfilename`, then the status file is named `inputfilename.status`, the results file is named `inputfilename.out`, the report file is named `inputfilename.report` and the warnings file is named `inputfilename.warning`.

To change the name of the used prefix use to generate these files, use

- `-o outputprefix`

4 Example

4.1 Collected data

In the following, we pursue the analysis of four benchmarks. The observed executions times of four are listed in four $\times$ two CSV files (initial an optimised version) named `bench1.data.1`, `bench1.data.2`, `bench2.data.1`, `bench2.data.2`, `bench3.data.1`, `bench3.data.2` and `bench4.data.1`, `bench4.data.2`.

The first benchmark is composed of 2×5 measures (initial an optimised version).

```
> cat bench1.data.1
2.02
2.25
2.30
2.251
2.01
> cat bench1.data.2
1.02
2.05
2.30
2.071
1.05
```

The second benchmark is composed of 2×6 measures (initial an optimised version).

```
> cat bench2.data.1
2.799
2.046
1.259
1.877
```

```
2.244
> cat bench2.data.2
1.046
0.259
0.877
1.244
1.799
```

The third benchmark is composed of respectively 15 and 20 measures. Observe in particular that the number of measures for the two programs to compare need not to be the same.

```
> cat bench3.data.1
6.512692
5.547728
4.171278
5.748114
6.188147
4.860546
6.393239
5.862367
5.724749
7.769651
6.455157
6.975127
5.331494
6.779595
4.839683
> cat bench3.data.2
4.556838
5.491279
5.708276
5.204911
4.454981
5.059760
5.440053
4.780246
4.363734
5.782297
5.195786
5.627607
6.114562
6.552509
3.055505
4.037513
5.445448
3.665237
6.965091
4.396594
```

The fourth benchmark is composed of respectively 4 and 8 measures (initial an optimised version).

```
> cat bench4.data.1
7.308153
```

```
6.891170
6.102855
6.472642
> cat bench4.data.2
6.571750
5.514734
5.705132
7.051386
8.007863
4.187613
6.124584
4.995708
```

The configuration file, named `bench.cfg`, is shown below.

```
> cat bench.cfg
Name,Sample1,Sample2,ConfLevel,Coef
"First sample","bench1.data.1","bench1.data.2",NA,
"Second sample","bench2.data.1","bench2.data.2",NA,NA
"Third sample","bench3.data.1","bench3.data.2",,NA
"Fourth sample","bench4.data.1","bench4.data.2",,
```

Remark that neither confidence levels nor weights are set. So by default the weights are considered equal for all benchmarks unless the command line option specifies something else. For the confidence levels, the default behaviour of Speedup-Test software is that it search for the highest confidence level $> 50\%$ (lowest risk level $< 50\%$) that allows to declare a statistically significant speedup. This behaviour can be modified by the command line option that allows to specify a global confidence level for all the benchmarks.

4.2 Carrying on the Speedup-Test on the collected data

We analyse this set of data thanks to the following command.

```
> SpeedUpTest.sh bench.cfg
Analysis report of  ./bench.cfg

Overall gain (ExecutionTime=min) = 0.371
Overall speedup (ExecutionTime=min) = 1.589

Overall gain (ExecutionTime=mean) = 0.178
Overall speedup (ExecutionTime=mean) = 1.216

Overall gain (ExecutionTime=median) = 0.156
Overall speedup (ExecutionTime=median) = 1.185
```

The above messages print the overall gains and speedups depending on the chosen function to summarise the execution time of a program. Concerning the proportion of the accelerated programs, it is printed in the message below. A program is considered as accelerated if the speedup-test succeeds in declaring a statistically significance of speedup of its average or median execution times, as detailed in [TWB10]. So we have two sorts of proportions, depending if we consider the average of the median execution time.

```
The observed proportion of accelerated benchmarks (speedup of the mean) a/b =
3/4 = 0.75
The confidence level for computing proportion confidence interval is 0.95.
Proportion confidence interval (speedup of the mean) = [0.219; 0.987]
```

```
Warning: this confidence interval of the proportion may not be accurate because
the validity  condition {a(1-a/b)>5} is not satisfied.

The minimal needed number of randomly selected benchmarks is 289 (in order to
have a precision r=0.05).

Remark: The computed confidence interval of the proportion is invalid if b the
experimented  set of benchmarks is not randomly selected among a huge number of
representative benchmarks.

The observed proportion of accelerated benchmarks (speedup of the median) a/b =
4/4 = 1.
The confidence level for computing proportion confidence interval is 0.95.
Proportion confidence interval (speedup of the median) = [0.396; 1]
Warning: this confidence interval of the proportion may not be accurate because
the validity  condition {a(1-a/b)>5} is not satisfied.

Remark: The computed confidence interval of the proportion is invalid if b the
experimented set of benchmarks is not randomly selected among a huge number of
representative benchmarks.
```

We observe that there is a statistically significant speedup for 3 out of the 4 samples for the mean and 4 out of 4 for the median. The Speedup-Test computes a confidence interval for each of the proportions $\frac{a}{b} = \frac{3}{4}$ and $\frac{a}{b} = \frac{4}{4}$. However the printed warning clearly says that the confidence intervals may not be accurate because the condition $a \times (1 - a/b) > 5$ is not satisfied, see [TWB10] for more details. We recall that the confidence intervals of the proportions are invalid if the initial set of b benchmarks is not *randomly* selected among a huge set of benchmarks. In other words, a manual selection of a set of experimented benchmarks (such as from SPEC family or other public benchmarks) is an invalid approach to calculate the confidence interval of the proportion $\frac{a}{b}$.

Now, let us have a look at the detailed results of the performance evaluation of each benchmark.

```
> cat bench.cfg.out
"Name","SpeedupMin","SpeedupMean","IsMeanSignificant","MeanConfLevel",
"SpeedupMedian",
"IsMedianSignificant","MedianConfLevel","CoefMin", "CoefMean","CoefMedian"

"First sample",1.971,1.276,FALSE,NA,1.098,TRUE,0.76,1,1,1
"Second sample",4.861,1.957,TRUE,0.98,1.956,TRUE,0.99,1,1,1
"Third sample",1.365,1.167,TRUE,0.99,1.127,TRUE,0.99,1,1,1
"Fourth sample",1.457,1.112,TRUE,0.83,1.13,TRUE,0.81,1,1,1
```

As example, the first sample has an observed speedup of the mean equal to 1.27 and an observed speedup of the median equal to 1.098. However, we see that the values of the booleans `IsMeanSignificant=FALSE` and `IsMedianSignificant=TRUE`. This means that the observed speedup of the mean is not statistically significant, while the observed speedup of the median is. Note that the observed speedup of the minimum execution time is not analysed statistically because of all the reasons explained in [TWB10]. So we do not advise to rely on this sort of observed speedup because its reproducibility is not guaranteed. The second, third and fourth samples have all statistically significant speedups of the mean and the median because their booleans `IsMeanSignificant=TRUE` and `IsMedianSignificant=TRUE`.

Now, let us have now a look at the warnings that occurred during the Speedup-Test analysis.

```
Warnings regarding analysis of ./bench.cfg
First sample :
```

```
  Sample1 too small for applying the Student's t-test (speedup of the mean).
Please do more than 30 observations of the executions times.
  Sample1 too small for applying the Student's t-test (speedup of the mean).
Please do more than 30 observations of the executions times.
  Unable to find a confidence level greater than 50% to guarantee the
statistical significance of mean speedup.
3 warning(s).
```

These warnings tell us that there are not enough data for the first benchmark in order to analyse the statistical significance of the speedup of the mean execution time. Probably because the executions times of the first sample are not normally distributed, so the Student's t-test require more than 30 values to stay robust.

5 Licence: Copyright UVSQ (2010)

References

[R D08] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[TWB10] Sid-Ahmed-Ali Touati, Julien Worm, and Sébastien Briais. The speedup test. Technical report, University of Versailles Saint-Quentin en Yvelines, January 2010. http://hal.inria.fr/inria-00443839).