Homework 1: Summarizing Performance Data

Nicole Zattarin

1. Comparing data: the example of two operating systems

1.1. Display distributions

Let us consider the example of two operating systems: an older one and its new version, whose performances are supposed to be better then the previous one. We aim to compare these two options by studying the execution times of a series of commonly used programs with both options.

In Figure 1 we report both the distributions and the individual data for each of the two versions. It is possible to highlight that the histograms provide informations regarding the statistical distributions of data, while scatterplots allow us only to individuate a qualitative trend in the observed data. A more quantitative way to extrapolate information from the data consist of comparing the cumulative density functions of the two distributions, in Figure 2 we report the CDF for both the versions of the operating system. It is possible to observe that, since the CDF corresponding to the execution times of the new systems is almost constantly above the function of the older version, that the first provide a better alternative to the latter. Finally, it may be useful to compare two distributions is by means of boxplots, swarmplots and violinplots, in Figure 3 we provide both boxplots and swarmplots of each of the two distributions. Such tools are particularly useful to represent quantiles and, for what concerns violinplots and swarmplots, a qualitative shape of the distribution, as it is possible to observe in Figure 3.

1.2. Confidence intervals

Moreover, we could be interested in analyzing the improvement due to the new system. A possible way to quantify it consist of measuring the reduction in run time for the same sequence of tasks. In Figure 4, the first panel shows the scatterplot of reductions in running time, the second exhibits the confidence intervals for the mean and for the median, while the right panel represents the distribution of reductions.

From the theory we know different methods to compute confidence intervals for the mean. We already exploited the standard one to compute the CI of the run time reductions, but two more approaches can be investigated: once relies on the assumption that data are normally distributed, while the second consist of bootstrap method. In Figure 5 we show mean CI for all the methods proposed, both for the new and the old system.

All computations are carried out assuming a confidence level of 0.95.

2. Simulation analysis

In this section we perform a simulation, by means of Python Random Number Generators (RNG), to study the statistical properties of two distributions, uniform and normal, with particular reference to the computation of confidence intervals.

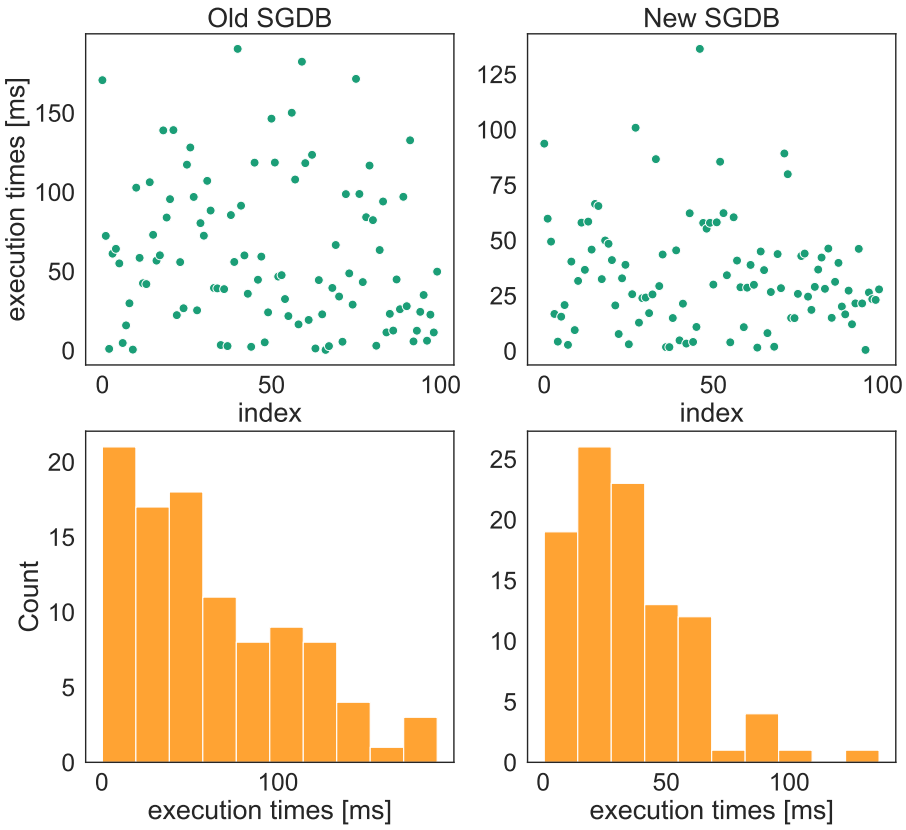


Figure 1

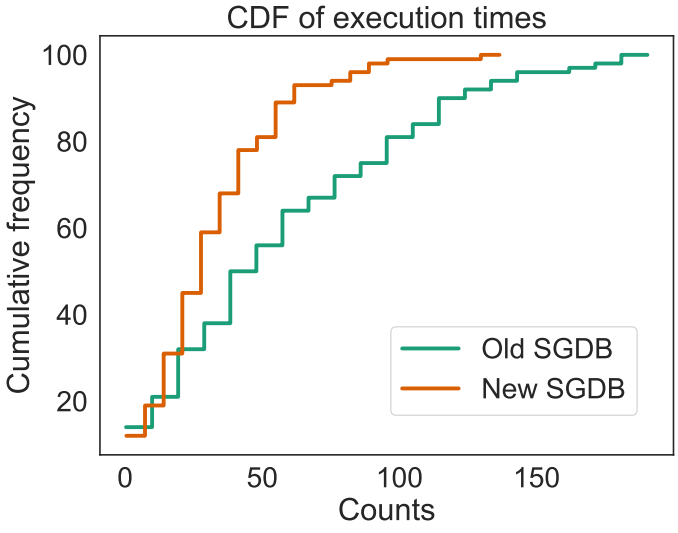


Figure 2

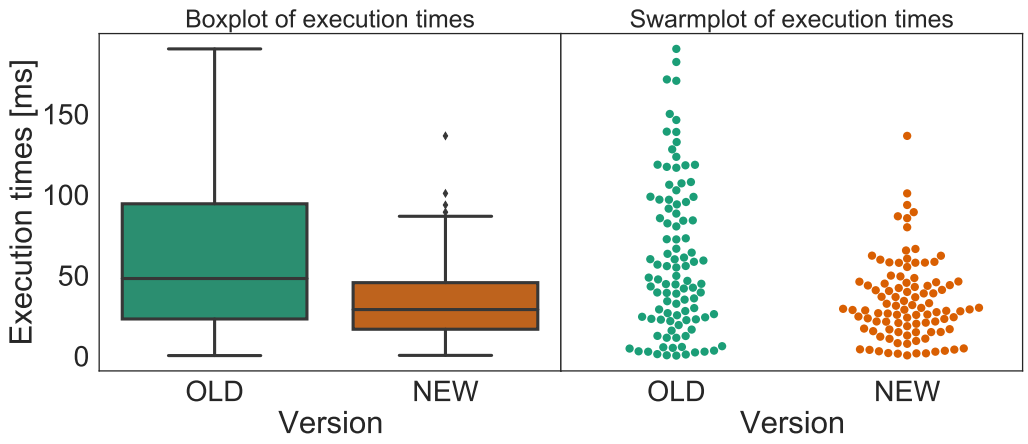


Figure 3

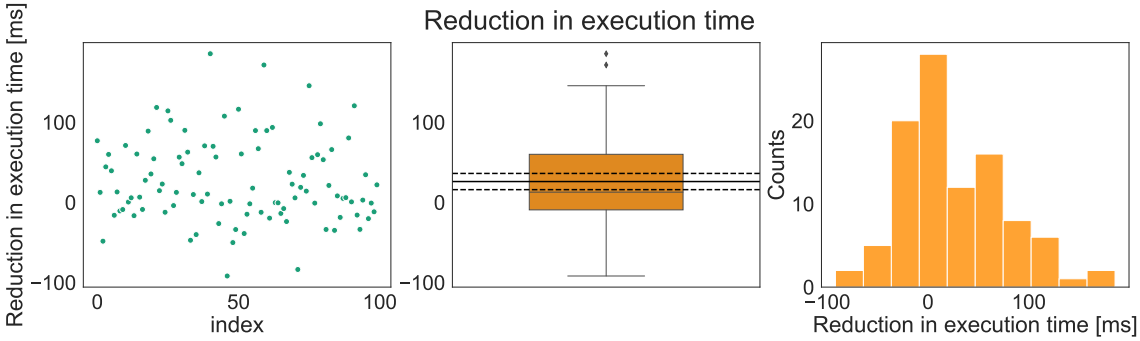


Figure 4

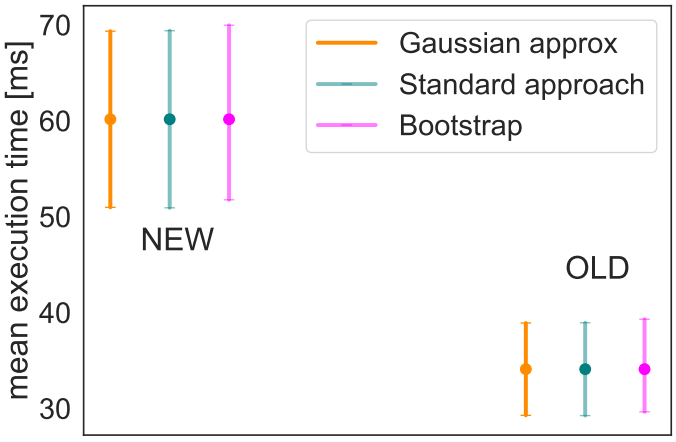
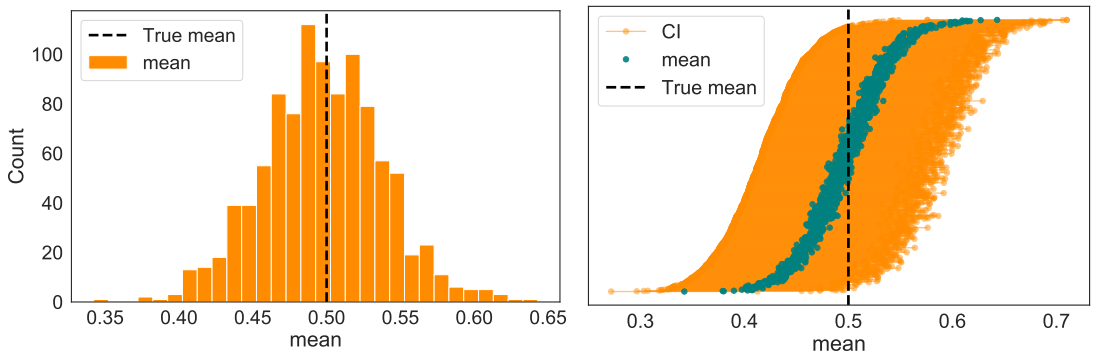
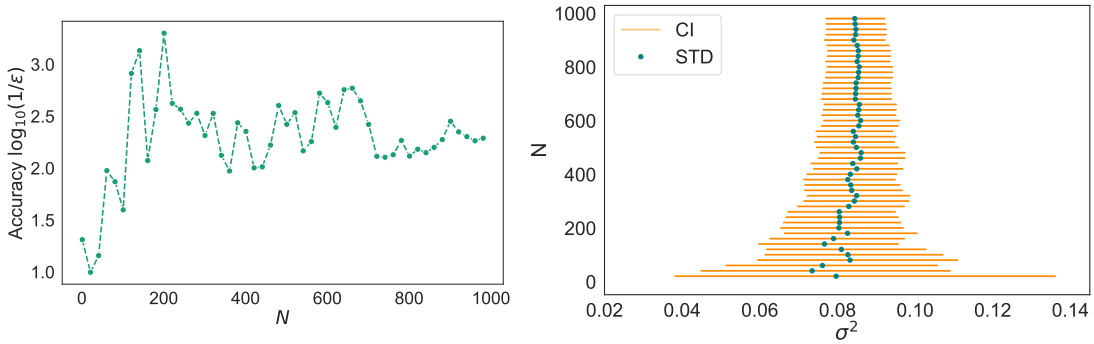


Figure 5



(a) $y = x$ **(b)** $y = x$

Figure 6



(a) $y = x$ **(b)** $y = x$

Figure 7

2.1. Uniform distribution

Let us consider n iid data generated randomly and uniformly distributed: $x_i \sim \mathcal{U}([0, 1])$.

2.1.1 Simulation with n=48

mean: 0.55 std: 0.27 CI: [0.47, 0.62]

2.1.2 Repetition of the experiment 1000 times

Repeat the experiment independently for 1000 times Observed probability of error 0.06.

UNIFORM DISTRIBUTION

Repeat the experiment independently for 1000 times Observed probability of error 0.06.

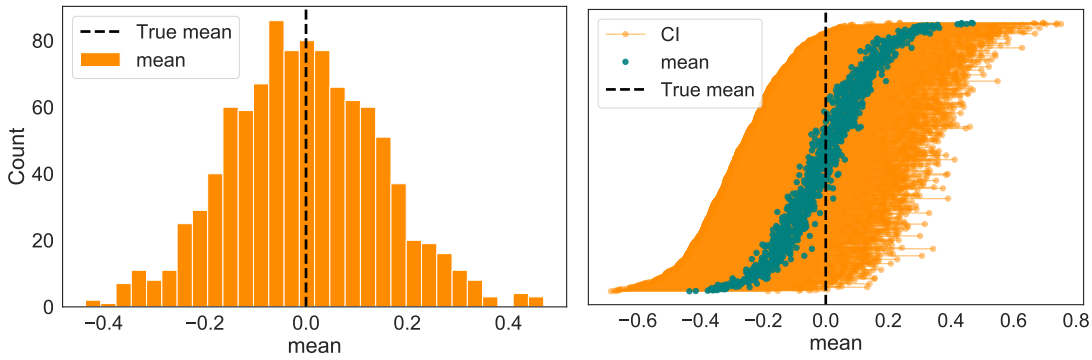
Simulation with n iid U(0,1) r.v.

Find confidence intervals for the VARIANCE vs. n

Find 95% prediction interval using theory

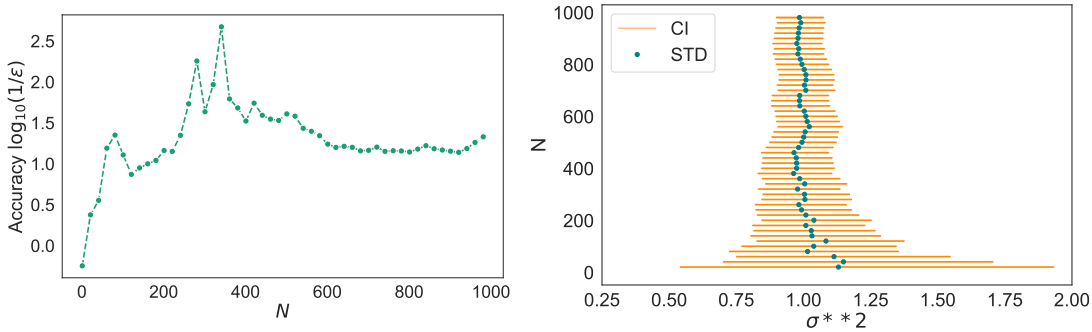
NORMAL DISTRIBUTION simulation with 48 iid norm(0,1) mean: 0.18 std: 1.12 CI: [-0.13, 0.50]

Repeat the experiment independently for 1000 times Observed probability of error 0.07.



(a) $y = x$ **(b)** $y = x$

Figure 8



(a) $y = x$ **(b)** $y = x$

Figure 9

2.2. Normal distribution

Simulation with n iid N(0,1) r.v. Find confidence intervals for the variance vs. n