Network Analysis and Simulation - Homework 1

Michele Scapinello

1 Exercise 1

Using the data provided on moodle, reproduce Figures 2.1, 2.2, 2.3 (no need to draw the boxes, just the values), 2.7, 2.8, 2.10 (skip (b)).

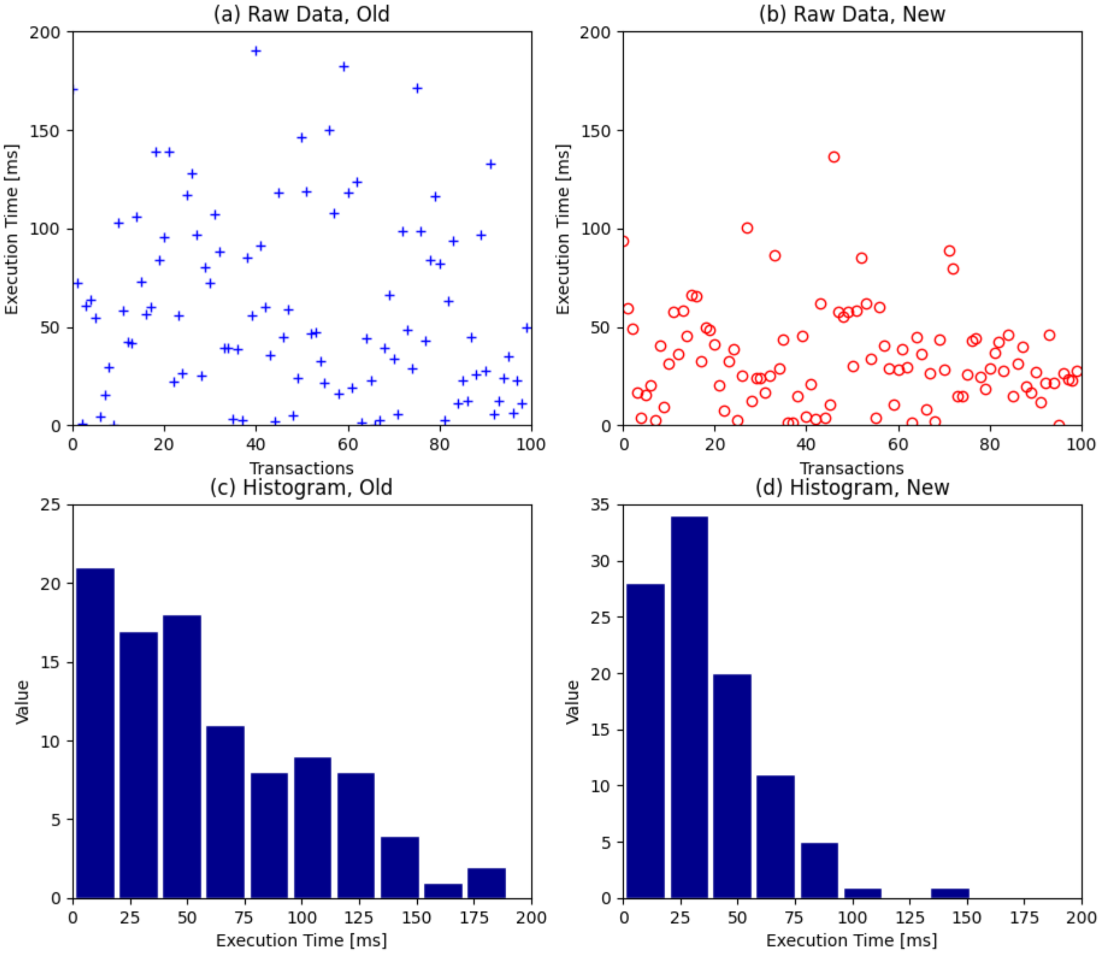

Figure 1.1: Reproduction of Figure 2.1, data for Example 2.1. Measured execution times, in ms, for 100 transactions with the old and new code and visualized also with histograms.

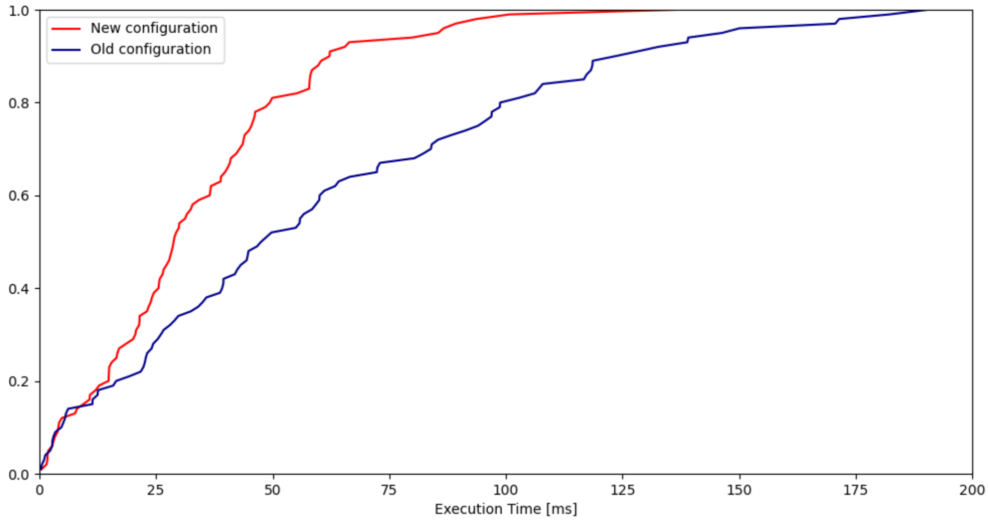

Figure 1.2: Reproduction of Figure 2.2, data of Example 2.1. Empirical distribution functions for the old code (right curve) and the new one (left curve). The new outperforms the old, the improvement is significant at the tail of the distribution.

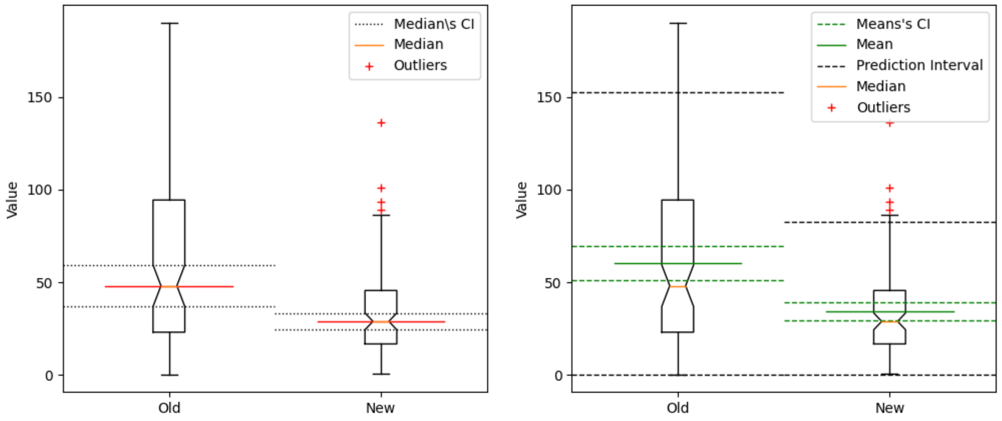

Figure 1.3: Reproduction of Figure 2.3, box Plots for the data for Example 2.1. Left: standard box plot showing median, median's confidence interval and quartiles (top and bottom of boxes); Right: box plot showing mean, mean's confidence interval (C.I = mean $\pm 1.96\sigma/\sqrt{n}$, where σ is the standard deviation and n is the number of sample) and prediction interval (P.I = mean $\pm 1.96\sigma$).

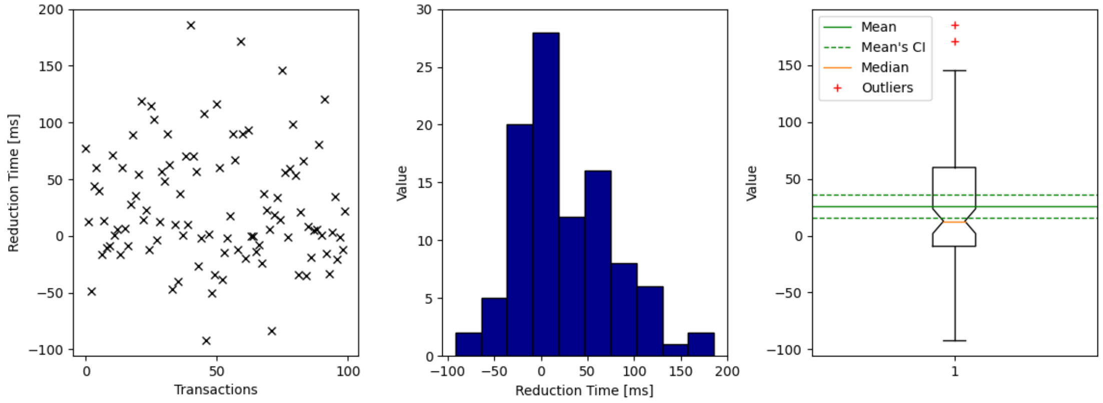

Figure 1.4: Reproduction of Figure 2.7, data for Example 2.2: reduction in run time (in ms). Left: normal plot of the data values; center: histogram of the data values Right: box plot with mean, median and respective confidence interval (median's confidence interval is denoted by the notch of the box).

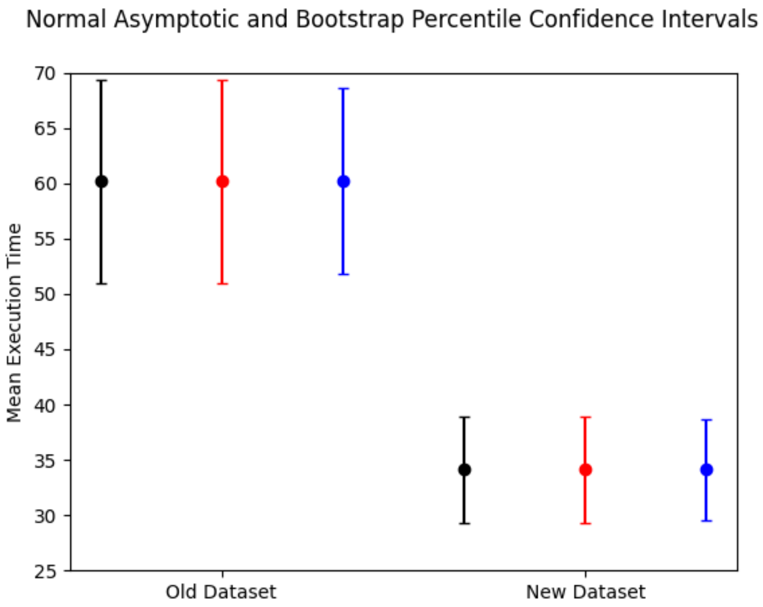

Figure 1.5: Reproduction of Figure 2.8, confidence intervals for both compiler options of Example 2.1 computed with three different methods: assuming data would be normal (left); the general method (center); bootstrap method (right)

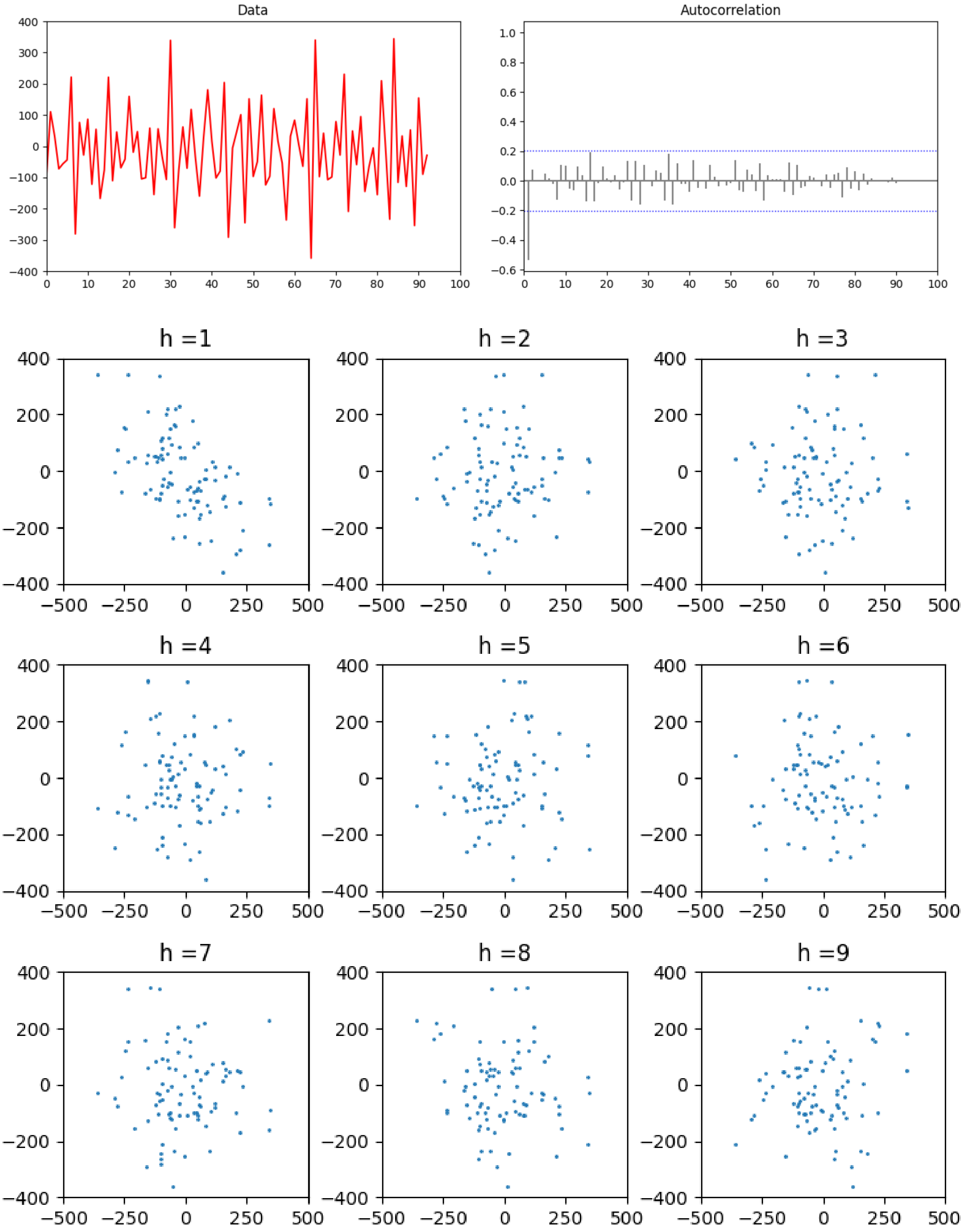

Figure 1.6: Reproduction of Figure 2.10 (b skipped), daily balance at Joe's wireless access shop over 93 days. The lag plots show y(t) versus y(t + h) where y(t) is the time series in (a). The data appears to have some correlation at lag 1 and is thus clearly not iid.

2 Exercise 2

Using MATLAB's random number generator, run the following experiment:

a. Generate n=48 iid U(0,1) r.v.'s

b. Find sample mean, sample std dev and 95%-confidence interval for the mean

c. Repeat the experiment independently for 1000 times, and find how many times the confidence interval does not contain the true value of the mean; plot the results ordering the intervals by increasing lower extreme of the CI; comment

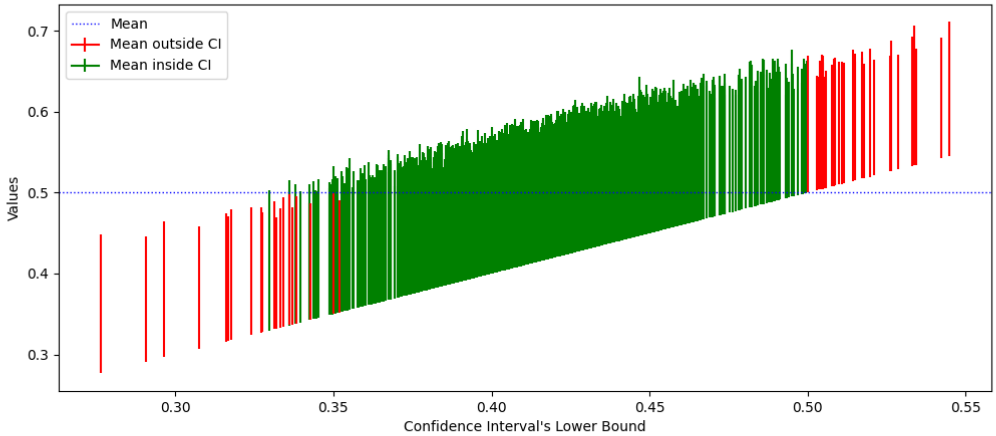

Figure 2.1: Results of the simulation for U(0,1) plotted by increasing order of the lower extreme of the CI

Confidence interval, in statistics, is a range of values that expresses the uncertainty of an estimated unknown parameter due to the randomness of the sampled data. The uncertainty is bounded by a confidence level, γ, and in b is set to $\gamma = 95$, meaning that with at least 95% probability the true value of the unknown parameter will fall inside the confidence interval. Figure 2.1 shows the results obtained from the experiment, performed independently 1000 times with U(0,1) (we assume that each experiment is independent from the other and that the variables are iid). In red are marked the simulations in which the confidence interval did not contain the true value of the mean, which is equal to $\frac{a+b}{2}$ for a uniform distribution, thus making our true value of the mean equal to 0.5. In green instead are marked the simulations where the confidence interval contained the true value of the mean. It is also easy to see from Figure 2.1 how confidence intervals are marked in red when either the upper bound is less than 0.5 or the lower bound is greater than 0.5. In the particular case of Figure 2.1 55 simulations, which corresponds to 5.5% of the runs, failed to contain the true value of the mean. This is expected as a 95%-confidence intervals means that we would expect to have 5% of the computed intervals to not contain the true mean. However, the number of failures may vary depending on how the data is generated, the number of samples and the number of experiments. To compute the confidence interval the following equation have been used:

$$\hat{\mu}_n \pm \eta \frac{s_n}{\sqrt{n}}$$

where $\hat{\mu}_n$ is the sample mean computed on n samples, $\eta = 1.96$ for $\gamma = 0.95$ and n is the size of the sample data-set.

3 Exercise 3

(difficult and optional)Prove that, for n U(0,1) r.v.'s, we have $\mathbb{E}(U_j) = \frac{j}{n+1}$

4 Exercise 4

Using MATLAB's random number generator, run the following experiment

a. Generate n iid U(0,1) r.v.'s, and compute sample mean and sample variance

b. Study the accuracy of the estimate with respect to the true value vs. n

c. Find confidence intervals for the variance vs. n

d. Find 95% prediction interval using theory and using bootstrap

To perform this exercise a series of simulations were carried out with a different number of samples generated, to show that with increasing number of samples, the estimates of mean and variance tend to converge to their true value. More precisely, simulations were carried out with 10,100,1000,10000,100000 samples.

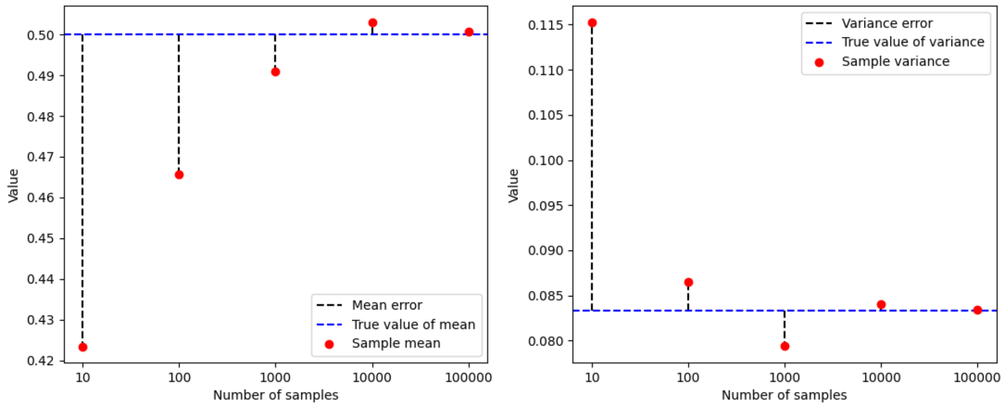

Figure 4.1: Plot of sample mean and sample variance (dots), true mean and true variance (blue line), and error on the estimation (black line)

In Figure 4.1 is plotted the error of the estimation as the distance (black line) between the mean/variance computed on the samples (red dots) and the value of the true mean/variance (blue line). It is possible to note, as mentioned above, that as the number of variables increases the absolute error between the true value of the mean/variance and the value estimated for the latter through the iid generated variables tends to converge towards zero, demonstrating that by increasing the data set size, it's possible to obtain an increasing accurate estimate, closer to the true value of the mean/variance. However it's important to notice that it could be possible to obtain some different results due to the randomness of the generated data. Figure 4.2 shows how the confidence intervals for the variance, computed using the bootstrap algorithm, tend to be more accurate as the data set size increases.

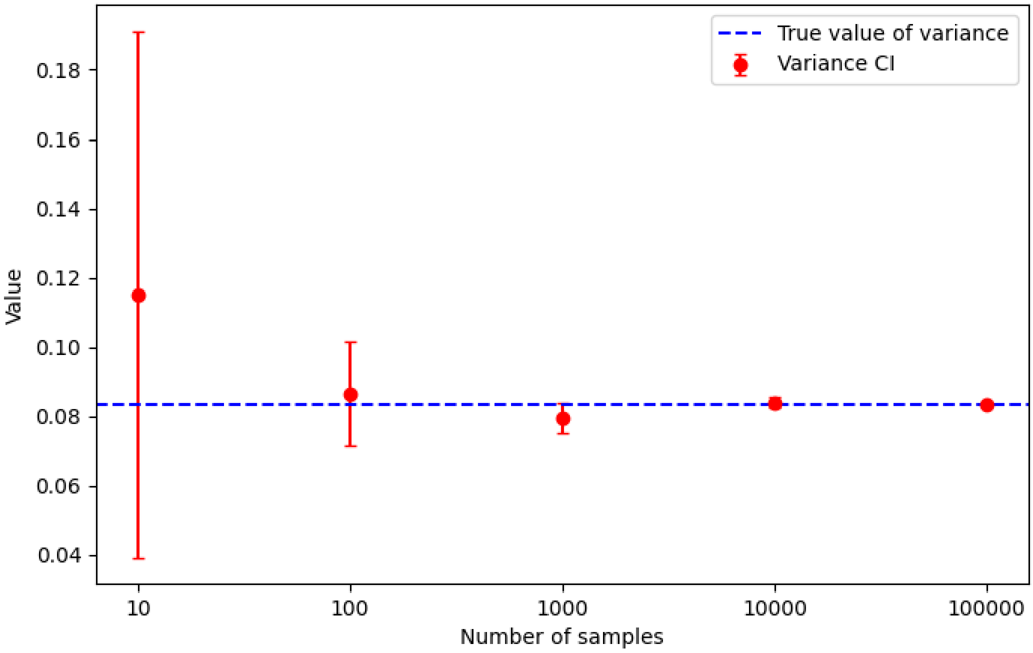

Figure 4.2: Confidence intervals for the variance vs n

Prediction interval allows to give an estimation of the variability of the samples. An interval at level γ gives a γ-% probability that the next observation will fall within the prediction interval range. Figure 4.3 shows the prediction interval computed over 1000 sample using theory and bootstrapping technique. The prediction interval are computed with $\gamma = 95$ using theory, by ordering in statistical order the variables and by taking, for $\alpha \geq \frac{2}{n+1}$ the interval bounded by the variables in position

$$[X^n_{(\lfloor (n+1)\frac{\alpha}{2} \rfloor)}, X^n_{(\lceil (n+1)(1-\frac{\alpha}{2} \rceil))}]$$

where α is a value equal to $(1 - \gamma)$. In order to fulfill the condition on α it is necessary to have a data set size greater than 39. For small n (lower than 39) the prediction interval offers a poor prediction much lower than 100% and the prediction interval is computed by ordering in statistical order the samples and by taking the X_{min} and X_{max} values as bound.

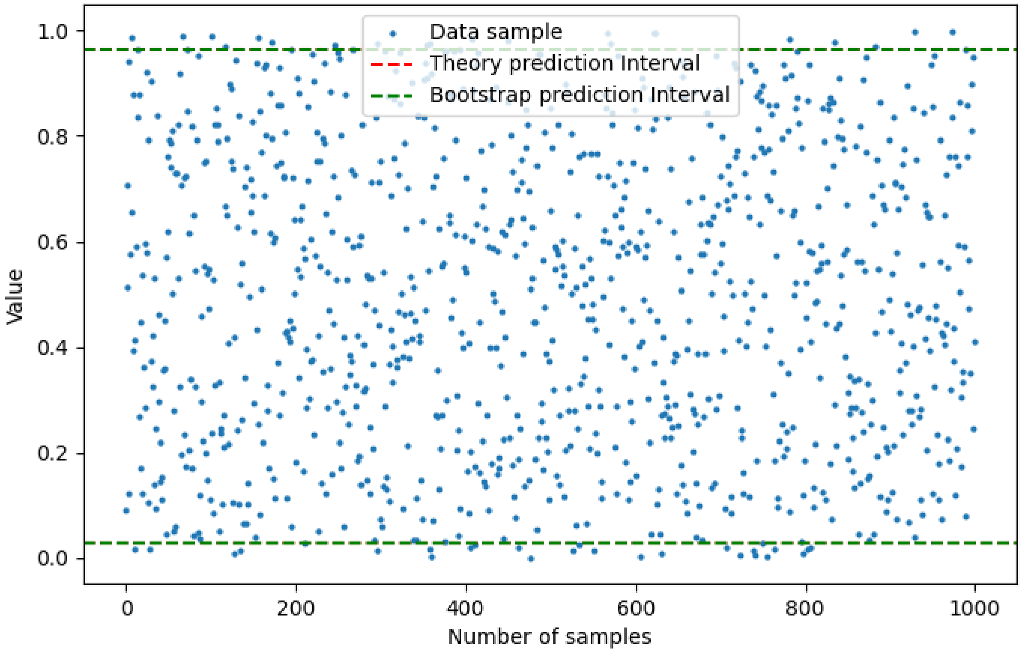

Figure 4.3: Prediction intervals using bootstrap algorithm and theory

5 Exercise 5

5.1 Redo exercise number 2 for N(0,1)

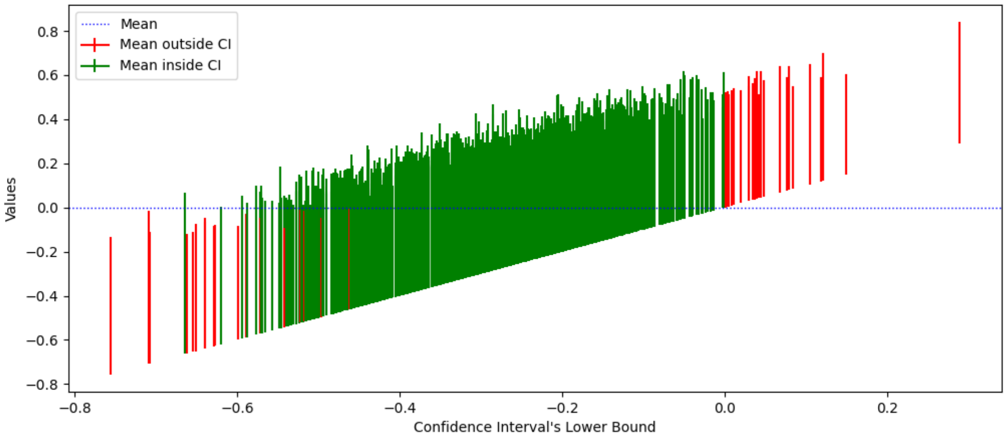

Figure 5.1: Results of the simulation for N(0,1) plotted by increasing order of the lower extreme of the CI

In red are marked the simulations in which the confidence interval did not contain the true value of the mean, which is equal to 0 because N(0,1) means that samples are drawn from a normal distribution with mean 0 and variance 1. In green instead are marked the simulations where the confidence interval actually contained the true value of the mean. It is also easy to see from Figure 5.1 that confidence intervals are marked in red when either the upper bound is less than 0 or the lower bound is greater than 0. By repeating the experiment 1000 times we find out in the particular case showed in Figure 5.1, that 44 simulations, which corresponds to 4.4% of the cases, failed to contain the true value of the mean. This is expected as a 95%-confidence interval means that we would have 5% of such intervals to not contains the true mean. However, the number of failures may vary depending on how the data is generated, the data-set size and the number of experiments. To compute the confidence interval the following equation have been applied:

$$\hat{\mu}_n \pm \eta \frac{\hat{\sigma}_n}{\sqrt{n}}$$

where $\hat{\mu}_n$, $\hat{\sigma}_n$ are respectively the sample mean and sample standard deviation computed on n samples, $\eta = 1.96$ for $\gamma = 0.95$ and n is the size of the sample data set.

5.2 Redo exercise number 4 for N(0,1)

The consideration for this exercise are the same for the exercise number 4 with the difference that here the samples are generated from a normal distribution with mean 0 and variance 1. As stated in exercise 4, with an increasing number of samples the estimated values for the mean and the variance tend to converge to their true value. Figure 5.3 shows how the confidence interval becomes much smaller and precise each time the number of samples increase and Figure 5.2, instead, shows the difference, computed as a distance, between the estimated value of mean/variance their true value. The prediction intervals shown in Figure 5.4 are computed using theory and bootstrapping technique. In this case the theory prediction interval is computed as

$$\hat{\mu} \pm \eta\hat{\sigma}$$

where η is the $1 - \frac{\alpha}{2}$ quantile of the normal distribution $N_{0,1}$. In our case, considering that $\alpha = 0.05$, η is equal to the 97.5% quantile of the distribution

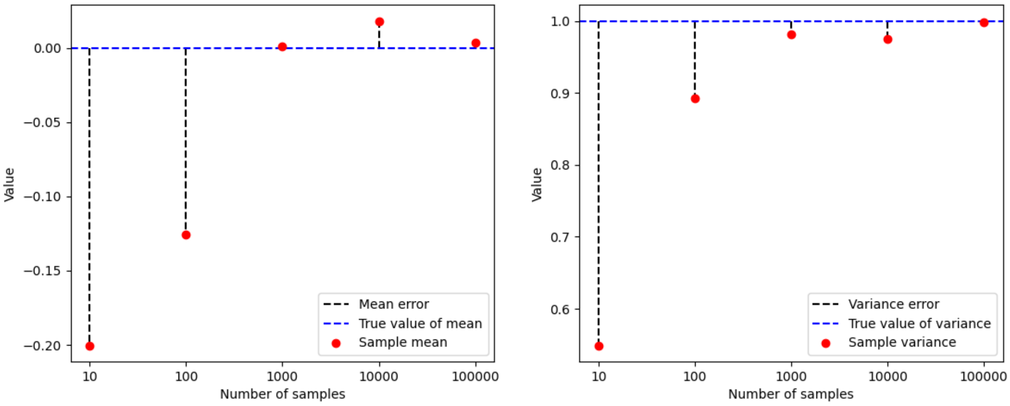

Figure 5.2: Plot of mean and variance value (dots), true mean and true variance (blue line), and error on the estimation (black line)

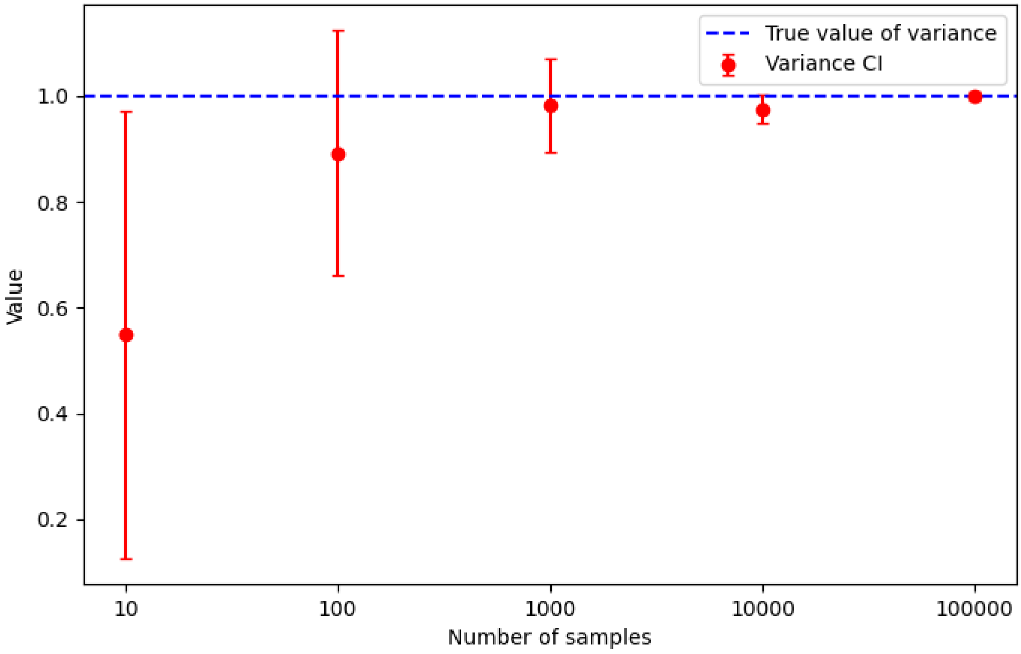

Figure 5.3: Confidence interval for the variance

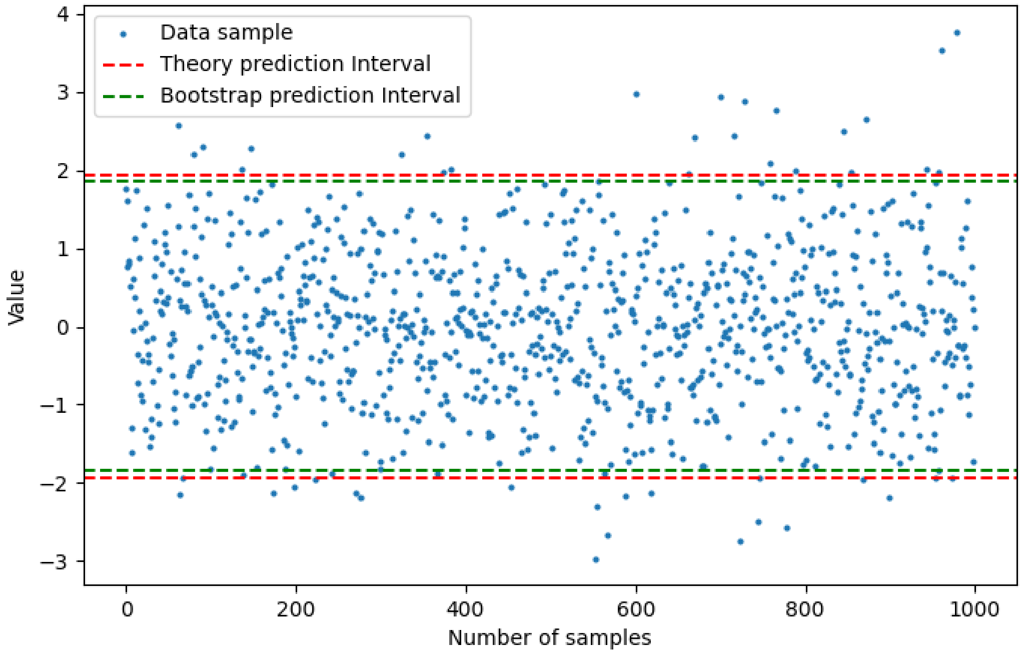

Figure 5.4: Prediction intervals using bootstrap algorithm and theory