Network Analysis and Simulation

Homework 1

Aynur Cemre Aka

Matricola: 2071493

Submission Date: March 31, 2023

TABLE OF CONTENTS

Question 1

Figures are reproduced from "Performance Evaluation of Computer and Communication Systems" by Jean-Yves Le Boudec, which is an educational resource provided by EPFL (École Polytechnique Fédérale de Lausanne). The version used for reference is Version 2.2.β, dated January 27, 2011.

Figure 2.1

Python has been used to compile and assess the set of results for a performance metric. Using histograms, which use bins for the data values and plot on the y-axis the proportion of data samples that fall in the bin on the x-axis, the distribution of the data has been fully described. The study's figures, in particular Figure 2.1, provide scatter plots of the old and new data on the left and corresponding histograms on the right. The scatter plots were generated in Python with the help of the Matplotlib package, and they show the measured execution times, in milliseconds, for 100 transactions using the old and new data:

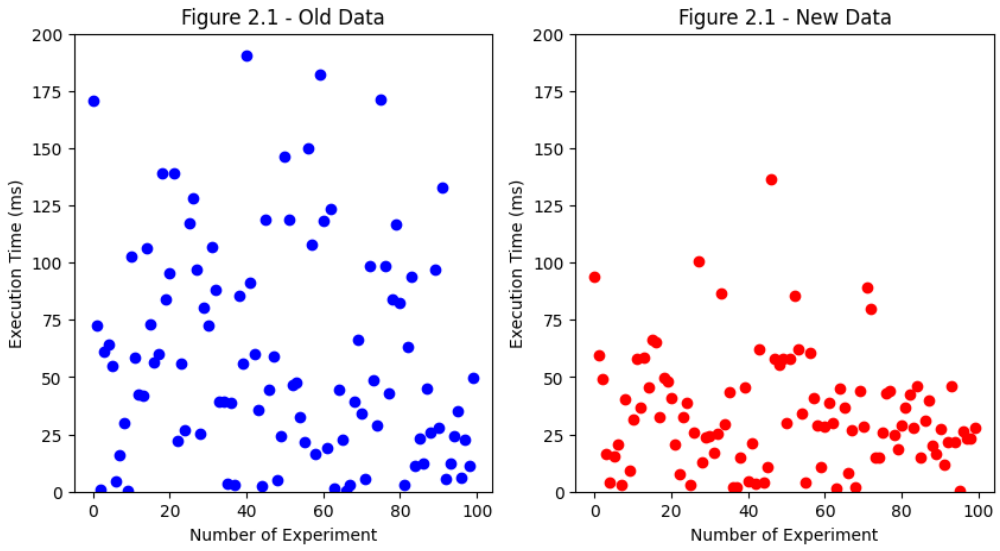

Figure 1. Scatter plots of old and new data (Figure 2.1).

The histograms, which show the frequency of the data samples in each bin, were likewise produced using the Matplotlib package. Using the characteristics of the data and an inspection of the resulting plots, the number of bins used for the histograms in this study was chosen as 10:

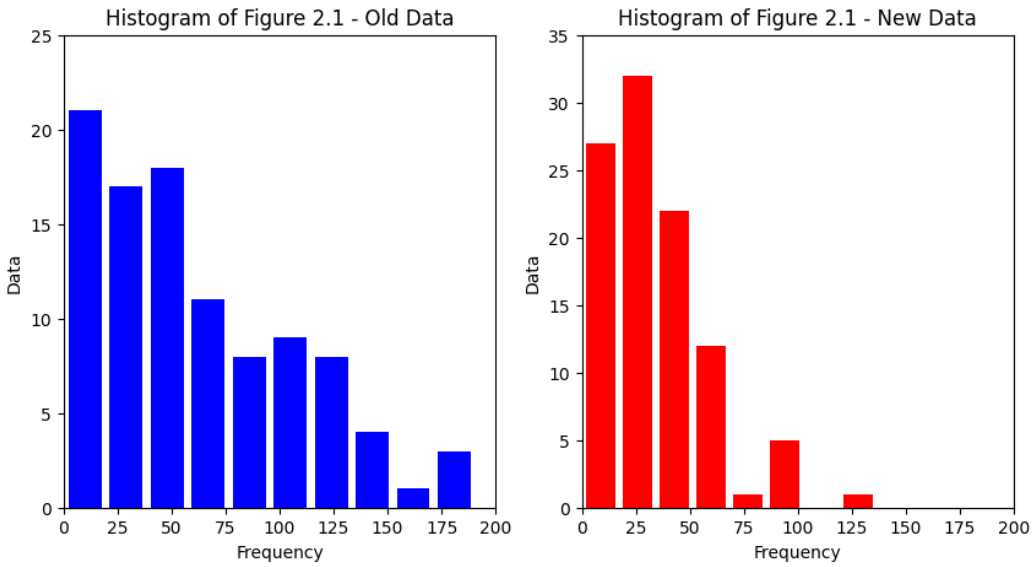

Figure 2. Histograms of old and new data (Figure 2.1).

Figure 2.2

Histograms can occasionally be replaced with the empirical cumulative distribution function (ECDF), which simplifies comparisons. A dataset's ECDF is the function F specified by:

$$F(x) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i \leq x\}}$$

F(x) is the proportion of data samples that do not exceed x as a result.

Even if some individual data points are less favorable, we can determine which dataset performs better by comparing the ECDFs of the different datasets.

Matplotlib package is used again to plot the ECDFs of the old and new datasets in Figure 2.2 to demonstrate this idea:

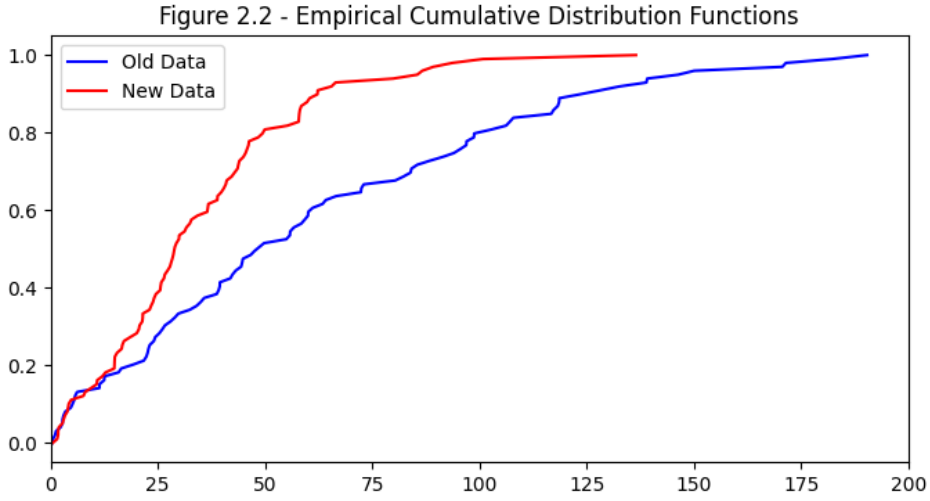

Figure 3. Empirical cumulative distribution functions of old and new data (Figure 2.2).

The new dataset performs better than the old one, as seen by the figure, where its ECDF is always higher than the other.

Figure 2.3

A box plot is a visual depiction of a dataset's distribution along its quartiles. It consists of a box that extends to both the lower and upper quartiles of the data's interquartile range (IQR). The whiskers reach the smallest and largest data points within 1.5 times the IQR, while the line inside the box denotes the median. Individual points outside the whiskers are plotted to represent outliers.

In our case, as shown in Figure 2.3, we were able to produce a box plot for the measured execution times of 100 transactions using the old and new code:

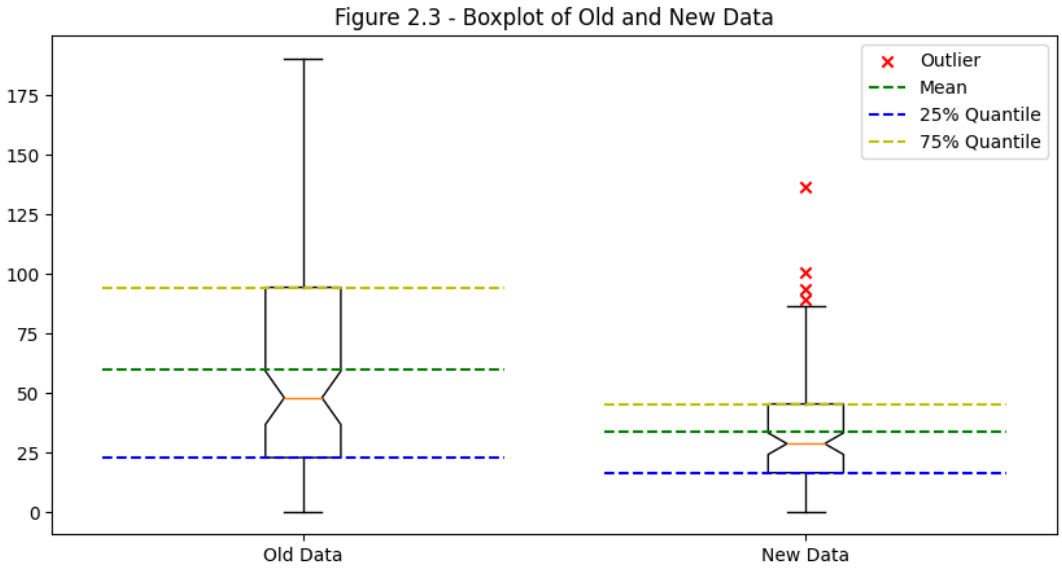

Figure 4. Box plots of old and new data (Figure 2.3).

The box plot demonstrates that the new system has a smaller range and a lower median than the old system. The new system's box plot also has a shorter upper whisker than the old systems, which suggests that the new system has fewer data points in the upper range. The individual points outside the box plot's whiskers, however, emphasize the fact that the new system also has more outliers.

The box plots typically show that the latest system has a more compact distribution of data, with a lower average and range, but with more extreme values.

Figure 2.7

Reduction in run time (in ms) is found by subtracting new data from the old data. Then, the confidence interval for mean is calculated as (15.76, 36.36) indicating that we can be 95% confident that the true difference in means falls between these values.

Figure 2.7 contains 3 subplots:

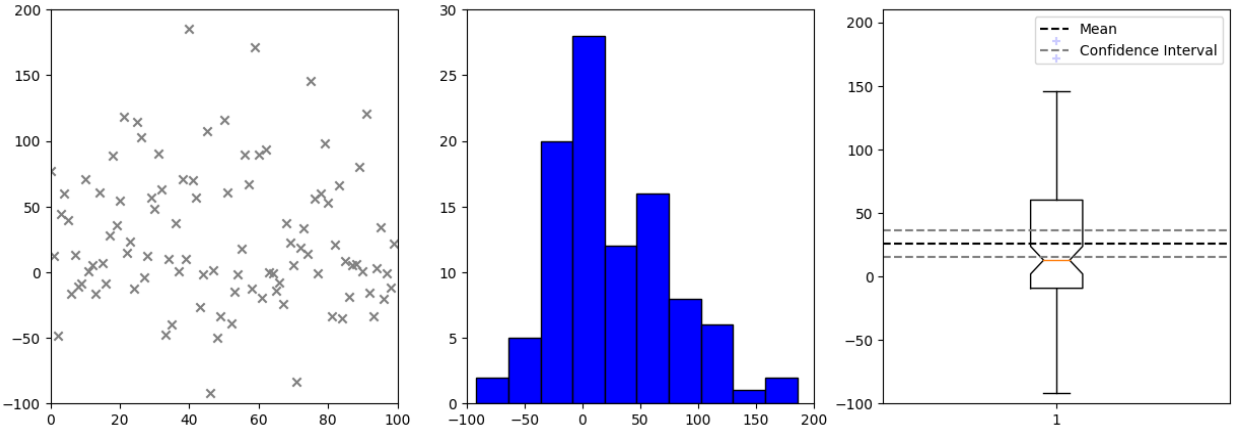

Figure 5. Data, histogram, and box plot with mean and confidence interval for mean (Figure 2.7).

The first subplot is a scatter plot with the values of the independent variable on the x-axis and the differences in run times between the old and new systems on the y-axis. The data points have a gray "x" symbol to identify them. The same data is split into 10 bins in the second subplot's histogram, which has blue bars and black borders. The y-axis displays the frequency of those differences in each bin, while the x-axis shows the range of run time variations. The third subplot is a box plot of the same data with notches, showing the median and confidence interval with horizontal lines.

Figure 2.8

A non-parametric method for estimating the confidence interval for a statistic is the bootstrap percentile confidence interval. The method involves generating several replacements resamples from the original sample, computing the mean of each resample, and then calculating the percentiles of the distribution of the resample means in order to get the confidence interval. The lower and upper limits of the confidence interval are calculated using the 2.5^{th} and 97.5^{th} percentiles of the distribution of the resample means:

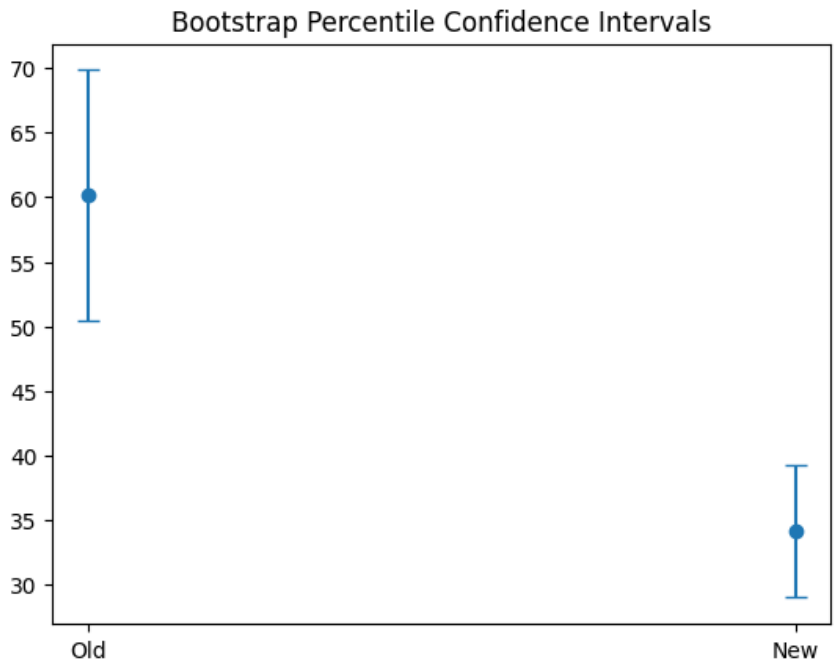

Figure 6. Bootstrap Percentile Confidence Intervals for old and new data (Figure 2.8).

At 95% of confidence, the resulting confidence interval provides an estimate of the range of values that the real population statistic is likely to lie within. The resulting confidence interval for old data is (51.32, 68.23), and for the new data is (29.51, 34.47).

Figure 2.10

The visualization of the data is as follows:

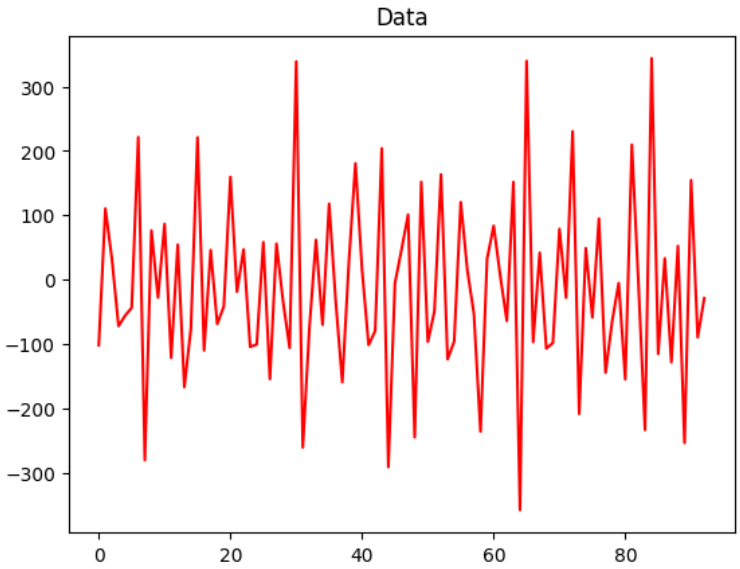

Figure 7. Visualizing the data (Figure 2.10 a).

The quantile-quantile plot, or QQ plot for short, is a graphical method for comparing the distribution of a dataset with a theoretical distribution, often the normal distribution. The quantiles of the sample data are compared to the corresponding quantiles of the theoretical distribution using the QQ plot. The QQ plot's points will fall along a straight line if the data is regularly distributed. Any deviations from this line are signs that something is out of the ordinary. The QQ plot of quantiles of input sample versus standard normal quantiles is as follows:

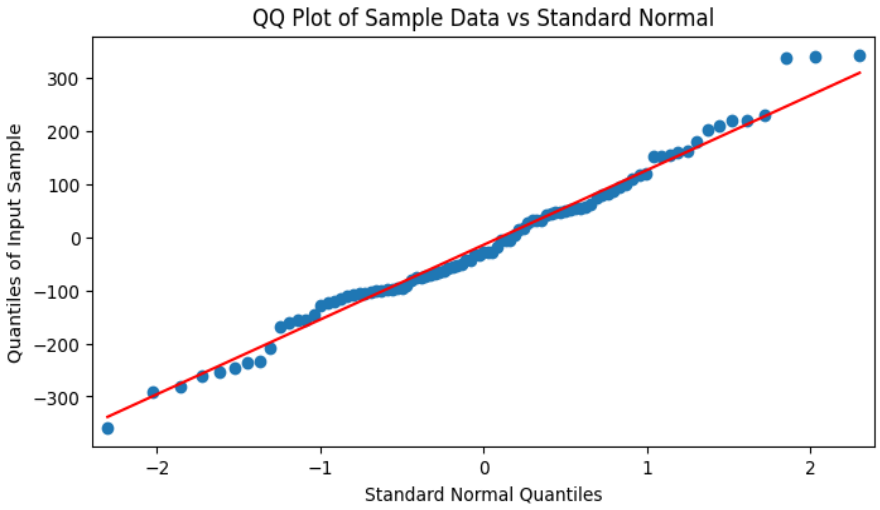

Figure 8. QQ plot of quantiles of input sample versus standard normal quantiles (Figure 2.10 b).

A statistical technique used to show the association between a time series and its lagged versions is the autocorrelation plot. It is produced by graphing the time series' lag against the autocorrelation function (ACF). The ACF, which ranges from -1 to 1, calculates the correlation between a time series and its lagged version at various time lags. Perfect positive correlation is represented by a value of 1, perfect negative correlation by a value of 1, and no correlation by a value of 0:

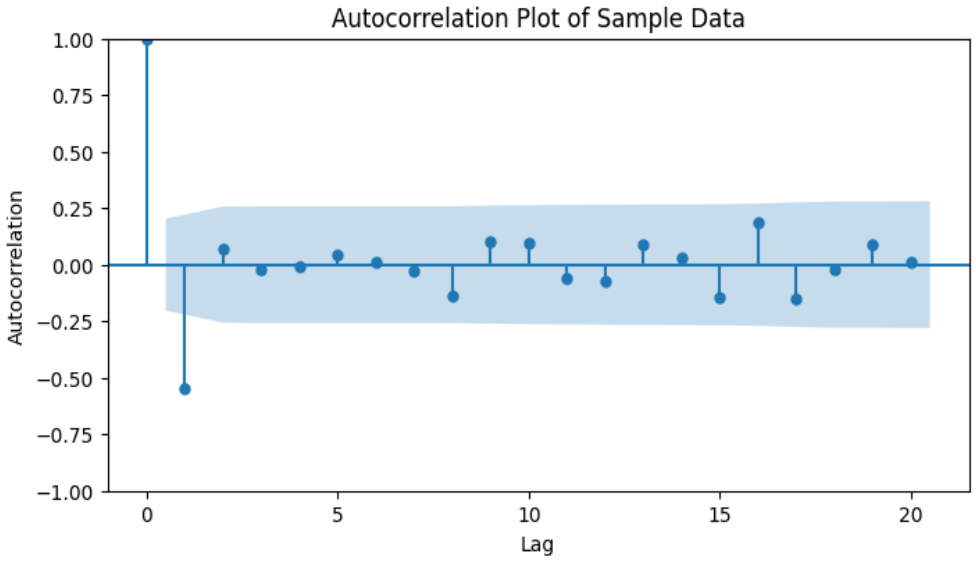

Figure 9. Autocorrelation Plot of Sample Data (Figure 2.10 c).

A time series' randomness and autocorrelation can be visually assessed using lag plots. The lag plots of the given data are as follows:

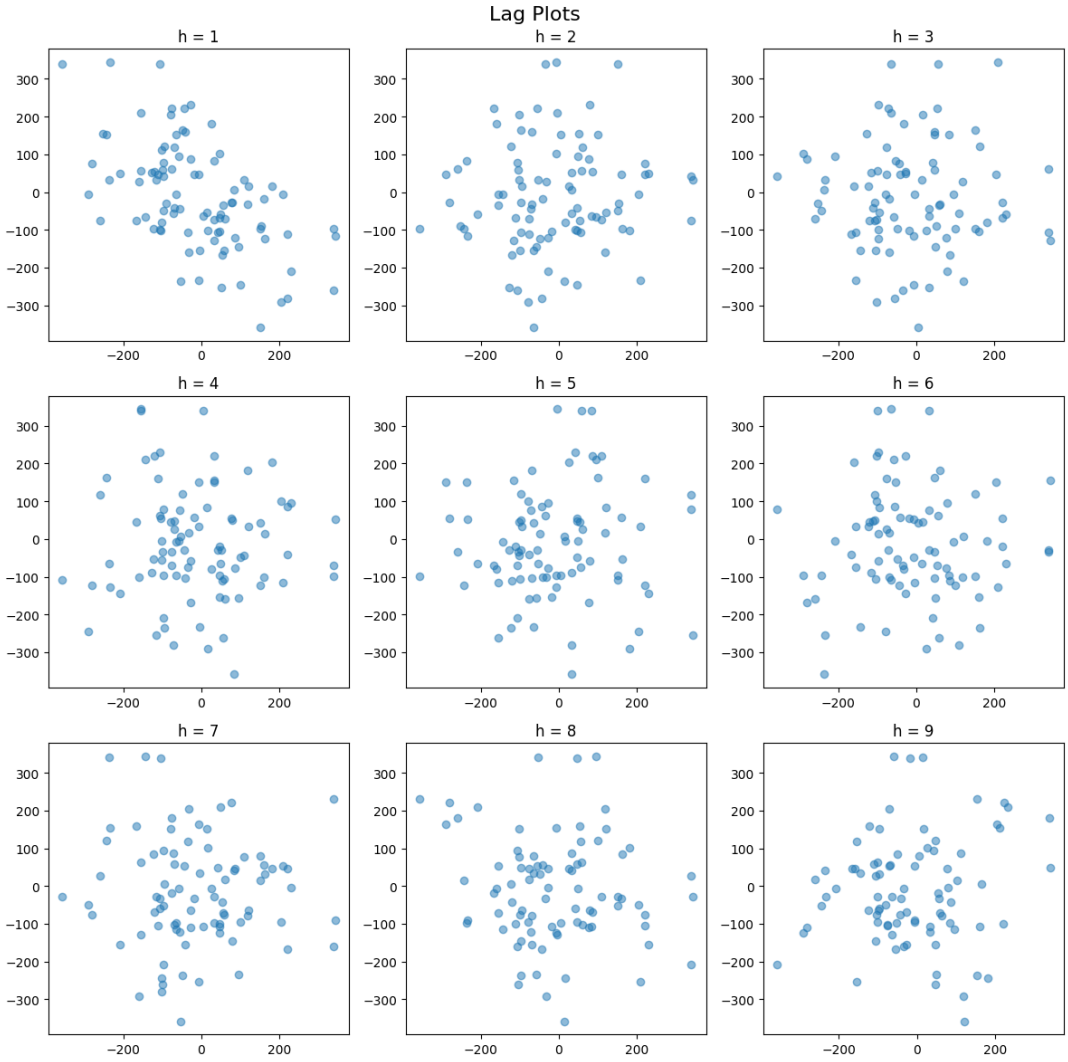

Figure 10. Lag Plots (Figure 2.10 d).

From the lag plots, it is observed that the data is not iid since it appears to have some correlation.

Question 2

The experiment is performed using MATLAB's random number generator. Firstly, 48 iid uniformly distributed random variables are created between 0 and 1. Then, sample mean, sample standard deviation, and 95% confidence interval are calculated. Obtained results are as follows:

- Sample Mean: 0.56
- Sample STD: 0.32
- 95% Confidence Interval for Mean: (0.47, 0.65)

As the sample size increases, it is expected the sample mean to approach the population mean (0.5) and the sample standard deviation to approach the population standard deviation (0.29). Then, the experiment is repeated independently 1000 times and sorted. The resulting plot is as follows:

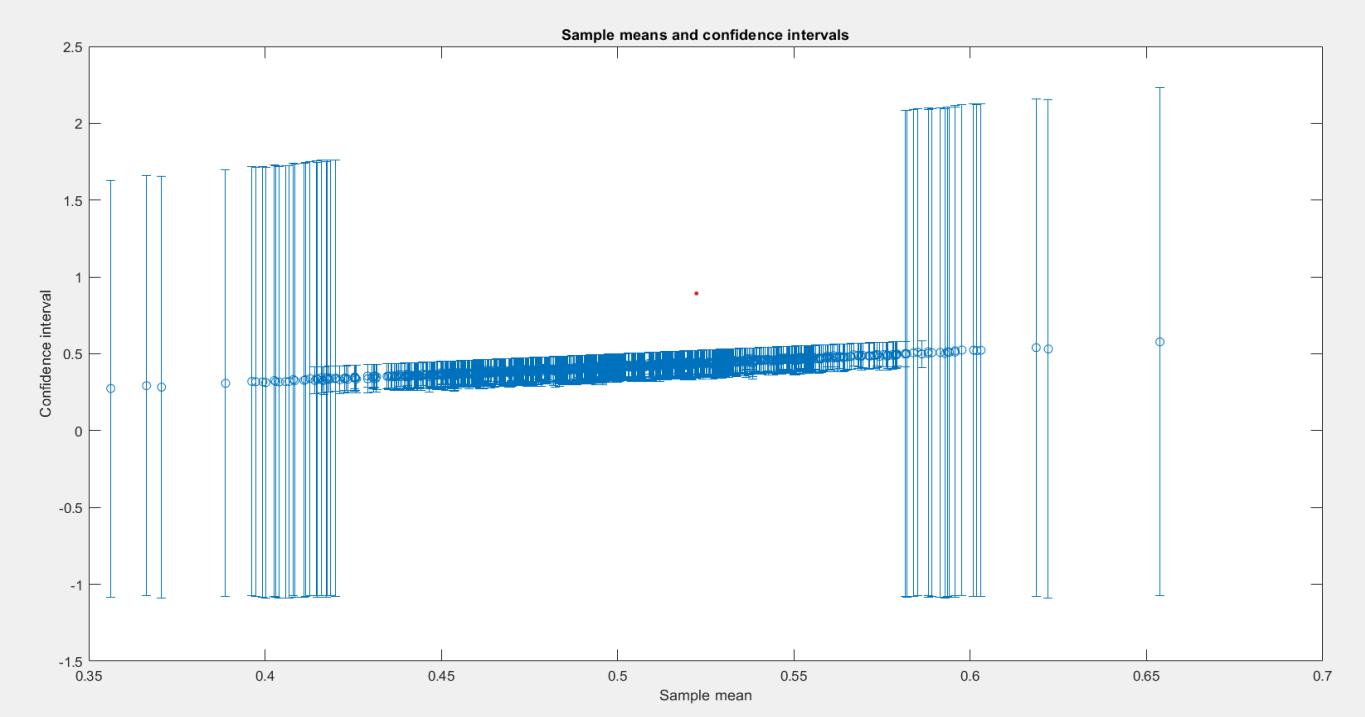

Figure 11. Sample means and confidence intervals.

As a result, 46 times the confidence interval does not contain the true value of the mean, out of 1000 times, which corresponds to 4.6%. It shows that the result is accurate.

Question 4

The experiment is performed using MATLAB's random number generator, like question 2. The aim of this experiment is to investigate the accuracy of the estimate with respect to the true value versus sample size. In order to achieve this, iid uniformly distributed random variables between 0 and 1 of varying sizes, ranging from 1 to 1000, are generated. Then, their mean and variance are computed. The resulting estimates are then compared with the corresponding theoretical values of 0.5 for mean and 1/12 for standard deviation. The absolute difference between the estimated and theoretical values is used to determine the mean and variance errors. The graph of errors for mean and variance is as follows:

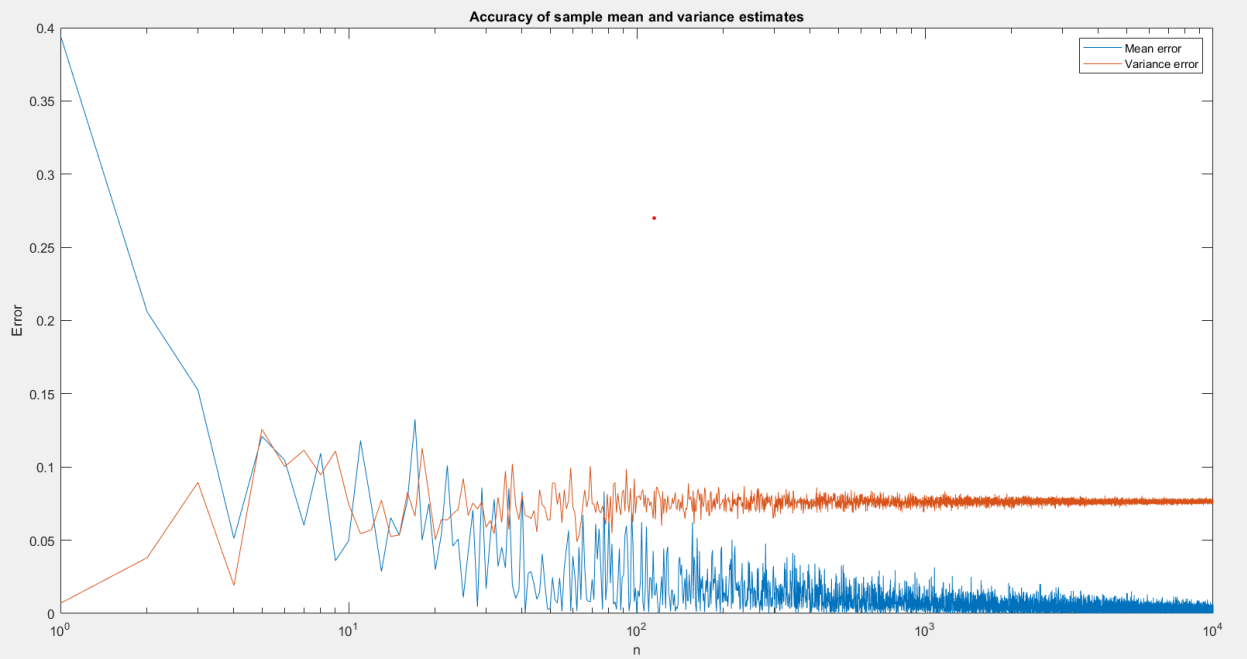

Figure 12. Accuracy of sample mean and variance estimates.

As a result, it is obtained that mean and variance error decreases, as the sample size increases. According to the Law of Large Numbers, as the sample size increases, the sample mean and variance converge to the population mean and variance, respectively. In other words, the accuracy of the mean and variance estimations grows with sample size, but the error between the estimated and real values decreases.

The confidence intervals for the variance versus sample size is as follows:

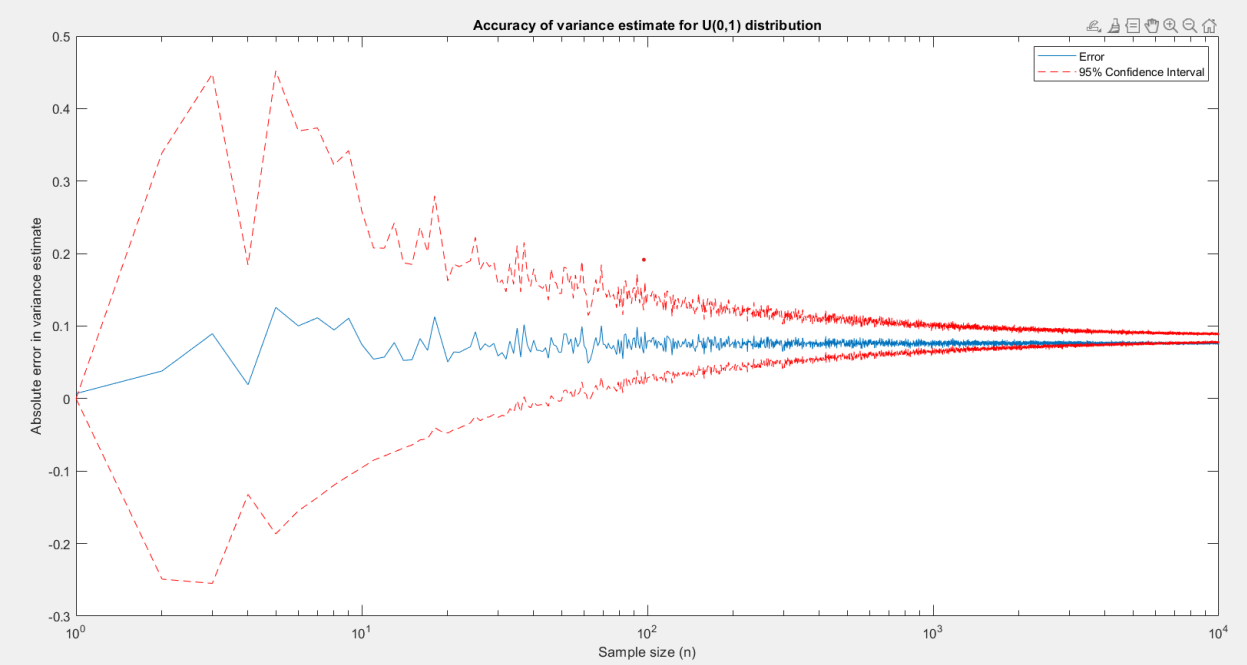

Figure 13. Confidence intervals for the variance versus sample size.

It is observed from the graph that the confidence interval narrows, and the estimate gets more accurate as the sample size decreases.

The confidence interval for the variance of a uniform distribution on the interval [0, 1] was estimated using the bootstrap method. The bootstrap approach includes generating new samples by randomly selecting with replacement from the original dataset, then computing the relevant statistic for each of these new samples. In order to get a distribution of the statistic, this method is iterated over several times. An estimation of the confidence interval may be made using the statistic's distribution. The sample size in this study ranged from 1 to 10,000. For each sample size, the estimated variance and its 95% confidence interval were determined.

The confidence intervals calculated with bootstrap versus sample size is as follows:

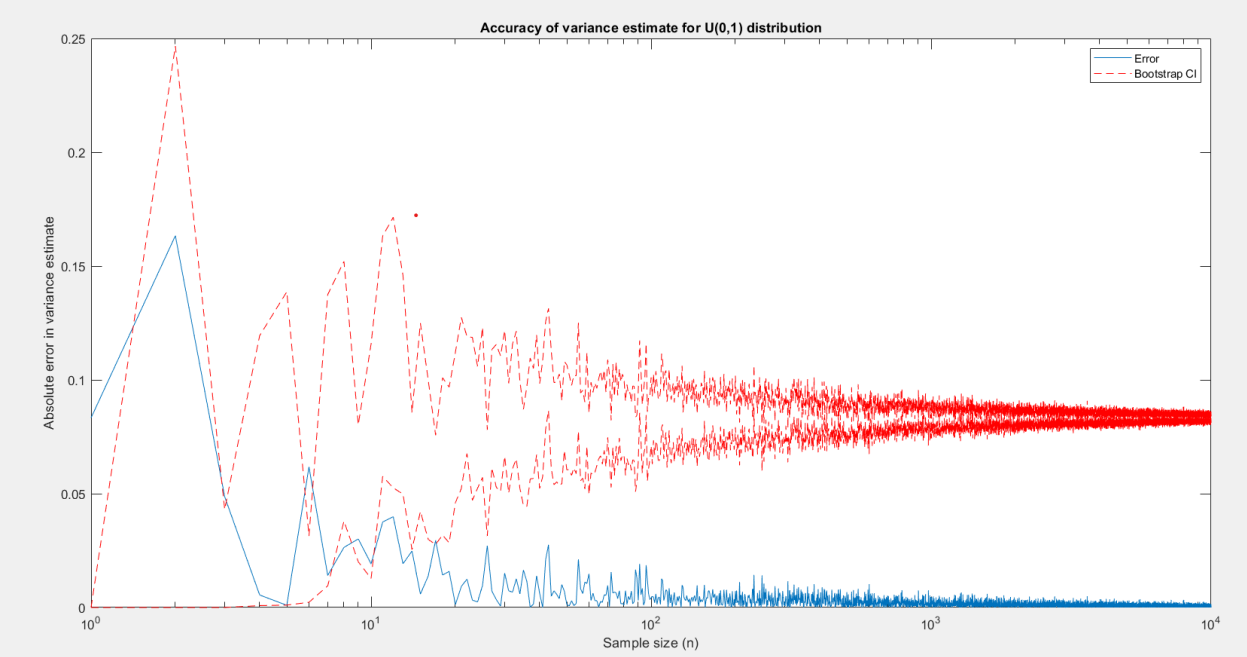

Figure 14. Confidence intervals with bootstrap versus sample size.

Similar results are obtained with the previous one, bootstrap method is noisier, because creating new samples using the bootstrap approach requires repeatedly sampling from the original dataset. This resampling procedure adds more randomness to the estimation, which might cause more variability in the confidence intervals that are produced.

Question 5

Question 2 for N(0, 1)

The experiment is performed again using standard normal distribution MATLAB's random number generator. Firstly, 48 normal distributed random variables are created between 0 and 1. Then, sample mean, sample standard deviation, and 95% confidence interval are calculated. Obtained results are as follows:

- Sample Mean: 0.34
- Sample STD: 0.44
- 95% Confidence Interval for Mean: (0.21, 0.47)

As the sample size increases, it is expected the sample mean to tend to approach the true population mean and the sample standard deviation to approach the true population standard deviation. Then, the experiment is repeated independently 1000 times and sorted. The resulting plot is as follows:

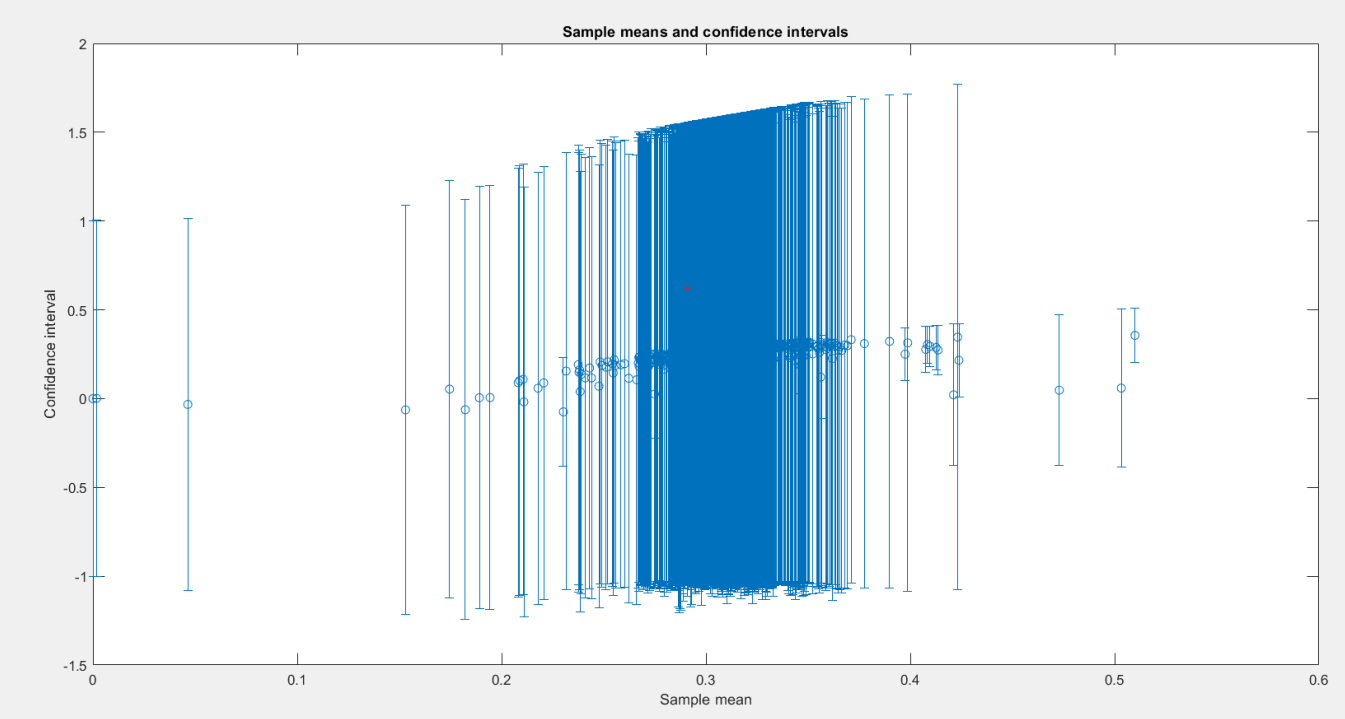

Figure 15. Sample means and confidence intervals.

As a result, 984 times the confidence interval does not contain the true value of the mean, out of 1000 times, which corresponds to 98.4%.

Question 4 for N(0, 1)

The aim of this experiment is the same as question 4, to investigate the accuracy of the estimate with respect to the true value versus sample size. In order to achieve this, normally distributed random variables between 0 and 1 of varying sizes, ranging from 1 to 1000, are generated. Then, their mean and variance are computed. The graph of errors for mean and variance is as follows:

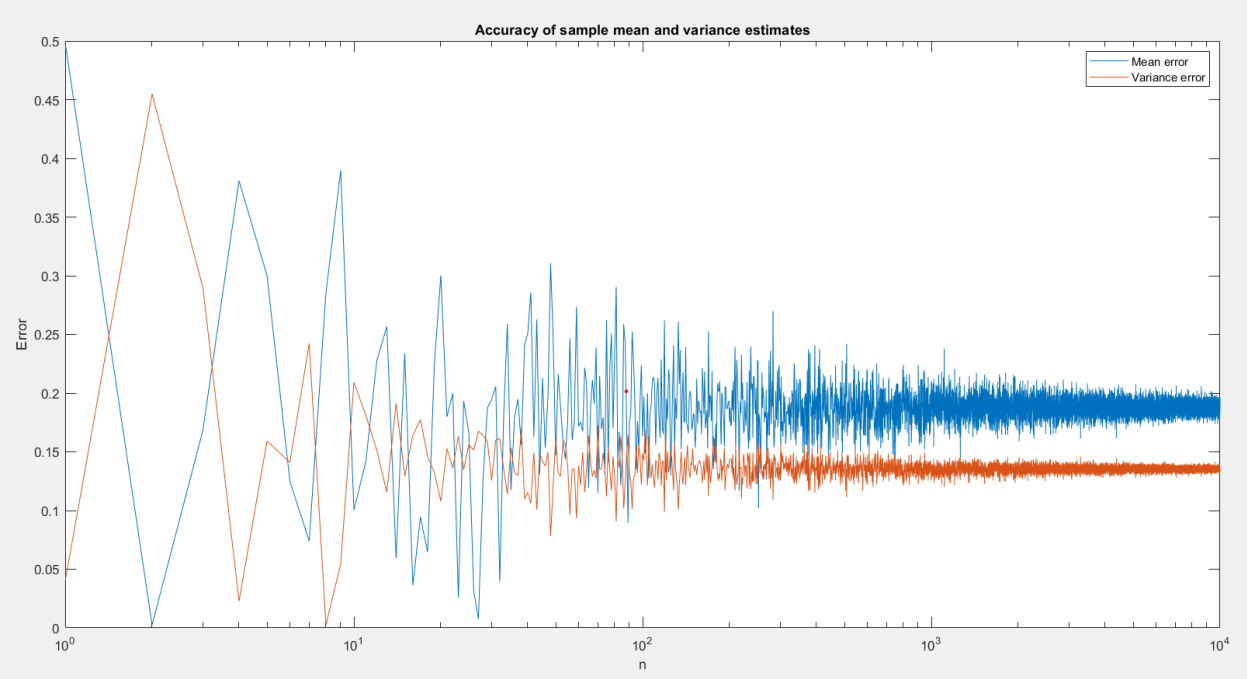

Figure 16. Accuracy of sample mean and variance estimates.

As a result, it is obtained that mean and variance error decreases and converges, as the sample size increases, but never becomes 0.

The confidence intervals for the variance versus sample size is as follows:

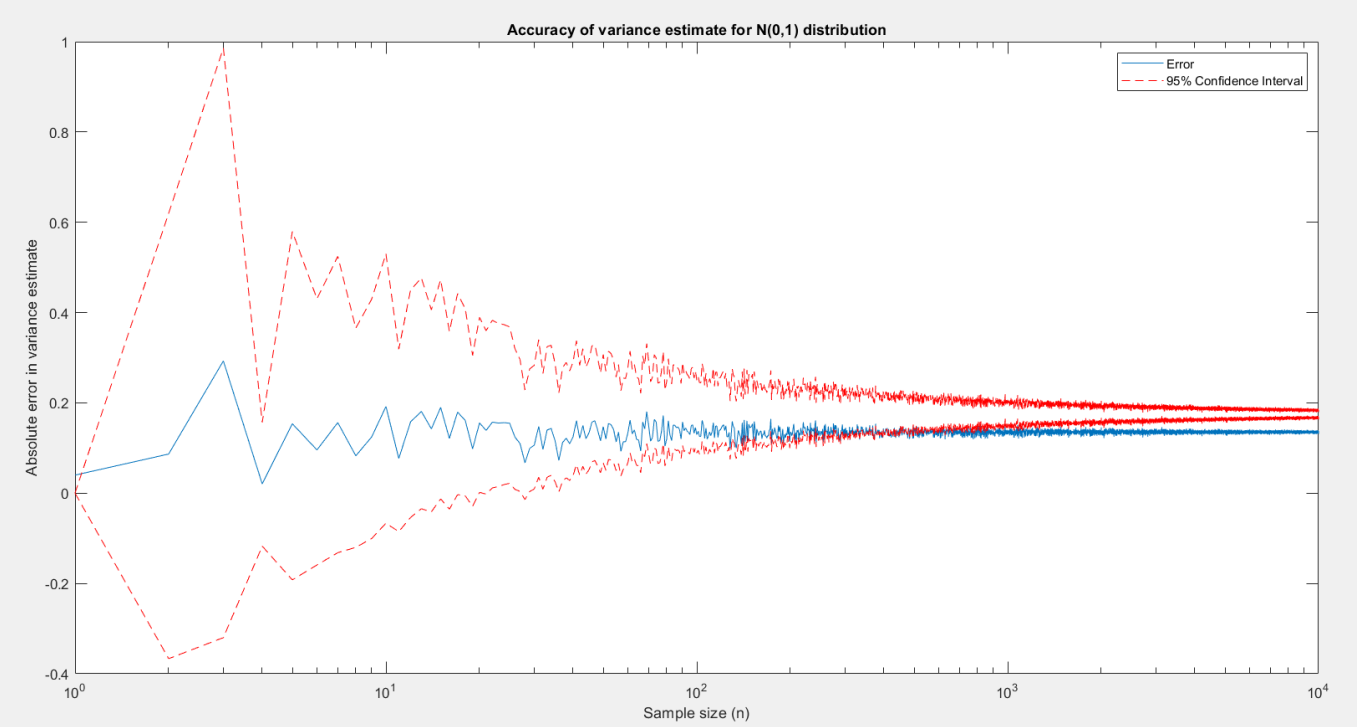

Figure 17. Confidence intervals for the variance versus sample size.

Similar results as question 4 is obtained for normally distributed random variables for varying sizes.

The confidence intervals calculated with bootstrap versus sample size is as follows:

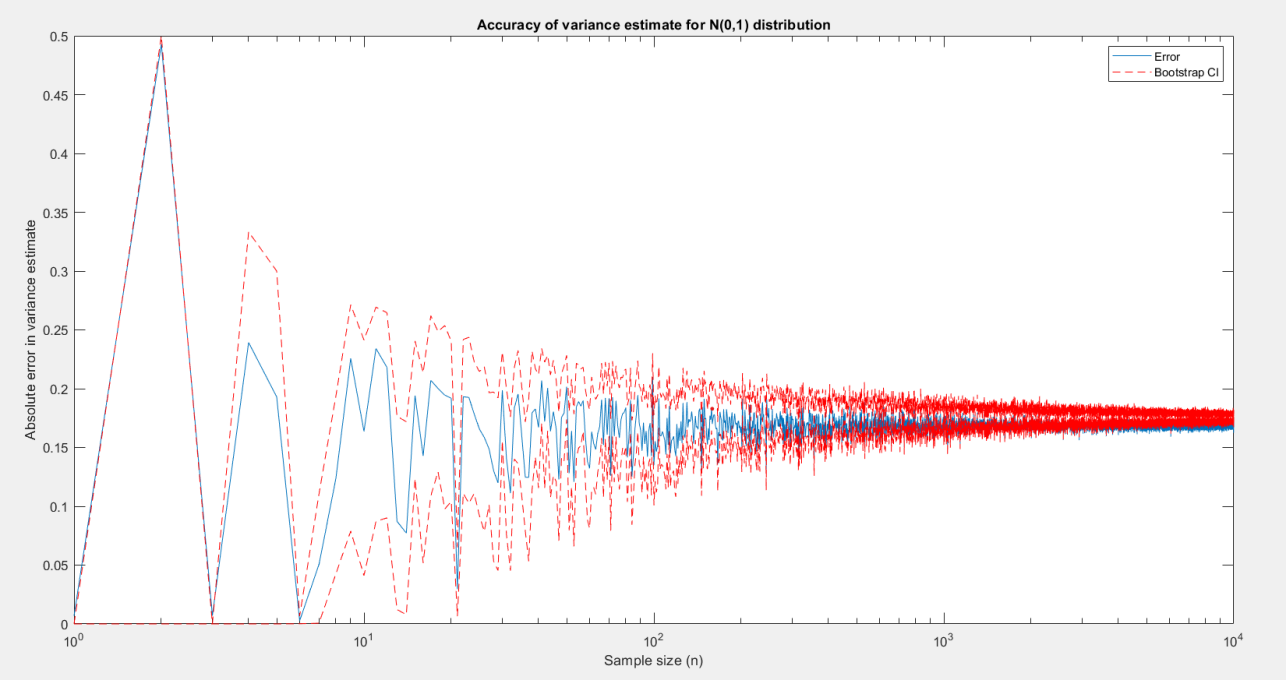

Figure 17. Confidence intervals with bootstrap versus sample size.

Like question 4, Similar results are obtained with the previous one, bootstrap method is noisier, because creating new samples using the bootstrap approach requires repeatedly sampling from the original dataset. This resampling procedure adds more randomness to the estimation, which might cause more variability in the confidence intervals that are produced.

Appendix

All code files can be found from the following link:

- https://github.com/cemreaka/Network-Analysis-and-Simulation/tree/main/HW1