1 Exercise 1

In this exercise we are given a Poisson process of rate λ yielding N arrivals in the interval $[0, T]$ and we are asked to draw some arrivals.

This can be done following two distinct methods (drawing N *arrival* times uniformly in $[0, T]$ or drawing N *inter-arrival* times exponentially).

Let's delve into our work, proving the equivalence of the two methods: we will prove that the inter-arrival times computed from the arrivals of the first method are exponentially distributed, while the arrival times computed from the inter-arrivals of the second method are uniformly distributed.

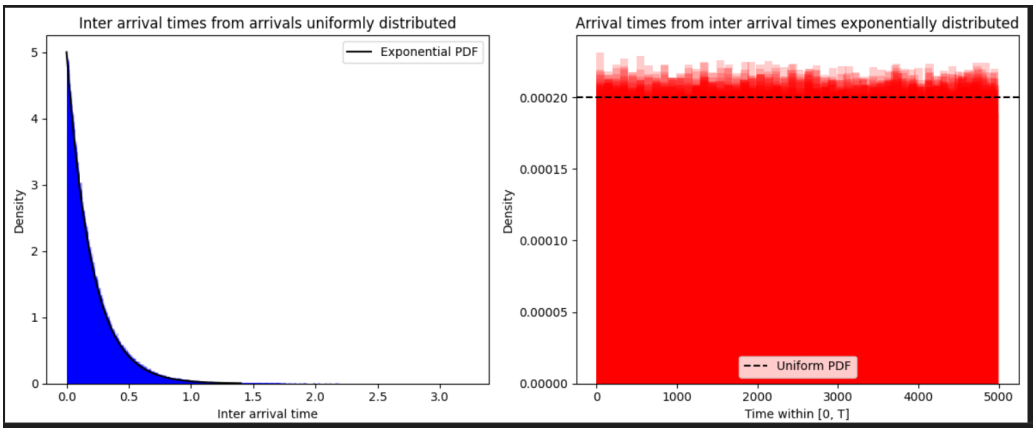
1.1 Checking the theoretical implications

We have proceeded with the sampling of N arrival times from *Uniform(0, T)* (we'll call this sample U) and N inter-arrival times from *Exp(λ)* (we'll call this sample E).

From the arrival times in U then we compute the inter arrival times between each pair of adjacent point, while from the inter arrival times E we compute the various arrival times.

Observation!
In order to draw samples for method 2, we needed to implement an additional "rejection check". Specifically, after generating a sequence of event inter-arrival times E, we verified that the last event occurred before the time limit T. This ensures that all events happen within the interval $[0, T]$.

Results
The results we obtained meets the theoretical intuition:



On the left we see that the inter-arrival times computed from U are actually matching the theoretical exponential PDF.

On the right we observe that the arrival times computed from E are actually matching the theoretical uniform PDF.

Our simulation setting was the following: $\lambda = 5$, $T = 5\,000$ and $N = 25\,000$. Additionally the experiment has been repeated 50 times to achieve smoother histograms.

1.2 Further considerations

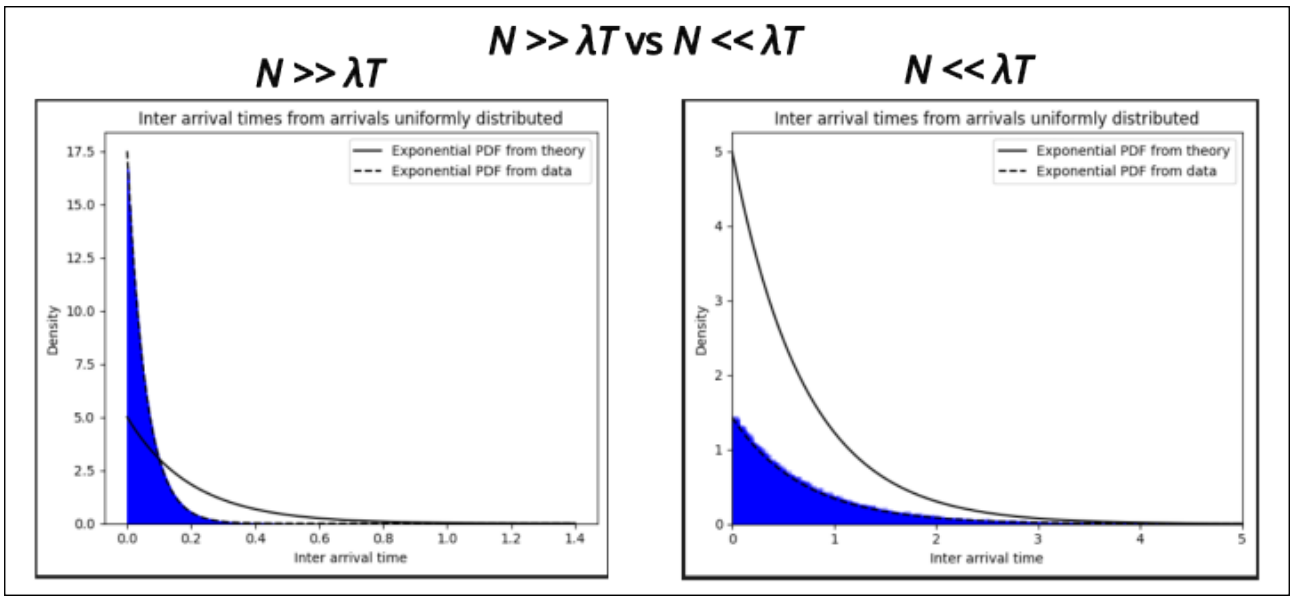
We set the number of points sampled $N = \lambda T$. This is intuitive, since we are in an interval of T time units and λ arrivals per time unit on average.

But... "What happens if change the value of the points sampled N?".

In case of $N >> \lambda T$ we are actually observing *more* events than expected in our interval. This means that the actual inter-arrival time of the data points is *shorter* respect to the one we originally set.

This can be confirmed by plotting the inter-arrival times for a case where $N = 3.5\,\lambda T$.

We can see (in the left plot) as the empirical distribution is no more following the theoretical one, but an exponential with a bigger parameter (arrival rate approximately at 17.5)!



In case of $N << \lambda T$ we are actually observing *fewer* events than expected in our interval. This means that the actual inter-arrival time of the events is *longer* respect to the one we originally set.

This can be confirmed by plotting the inter-arrival times for a case where $N = \lambda T \frac{1}{3.5}$.

We can see (in the right plot) as the empirical distribution is no more following the theoretical one, but an exponential with a smaller parameter (arrival rate approximately at 1.43)!

2 Exercise 2

Here our task was to consider a "weird" probability density function and perform some analysis on it.

Our PDF was: $f(x) = \frac{1}{A}x^2 \sin^2(\pi x)$ for $-3 \leq x \leq 3$ where $A = 8.8480182$ is a normalization factor such that $\int_{-3}^{3} f(x) = 1$.

2.1 Part 1 and part 2

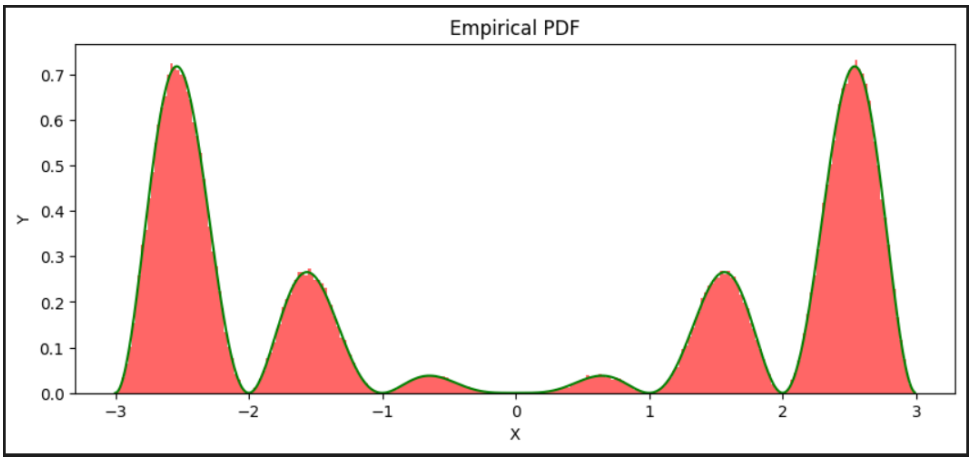
First parts of the exercise required employing rejection sampling to draw a sufficient number of samples and plot them to convince us that our empirical PDF is similar to the theoretical PDF.

A critical part of the method involved the choices of the constant c and the function g from which we have to sample.

For the latter, we simply choose $g(x) = 1$, a natural choice due to its simplicity in sampling it.

The constant c has to be chosen in a way that $\frac{f(y)}{g(y)} \leq c, \forall y$.
We have to choose a value of c such that the function g is an upper bound of our PDF f. In our case $g(x) = 1$ is already an upper bound of f, so we take c as the maximum point of the PDF ($c = 0.717705$), in order to have the highest efficiency in sampling and reducing the number of points to reject.

We sample $N = 300\,000$ points and plot the results in the red histogram. We then enhance the theoretical PDF with the green plot to show the correctness of the distribution followed by our points.



Was the value for A mandatory?
The normalization value A was not strictly mandatory since the rejection method only requires to compute the ratio between the two functions and to be sure that the one from which we are sampling is "above" the PDF we are interested in.

However, knowing the value of A allows us to have full understanding of the bounds of the PDF so as to have a tighter upper bound and reject fewer samples.

2.2 Part 3

In this part, we had to estimate the confidence interval for the mean, the median (0.5-quantile) and the the 0.9-quantile, using two different methods. The confidence level was set at 95%.

We draw a sample of 20 000 variates and consider only the first 200.

First of all we use the **asymptotic method** for the estimation since its assumptions were met: we have a large enough dataset ($n = 200$), distribution not heavily tailed and IID (independent

and identically distributed) points.

To make a comparison of our results, we have used also the **Bootstrap Percentile Method**, a very simple procedure to compute confidence intervals (since it requires only to have IID points), that however guarantees robust results, comparable to the previous ones.

Results
We can summarize in this table the results obtained by the two methods and compare them with the actual value.

	Asymptotic method	Bootstrap Percentile Method	Actual values
Mean	$[-0.512, 0.125]$	$[-0.334, 0.118]$	0.0
Median (0.5-quantile)	$[-1.687, 1.394]$	$[-1.518, 1.370]$	0.0
0.9-quantile	$[2.550, 2.775]$	$[2.596, 2.727]$	2.641

Both the methods allowed us to get intervals very close to each other, indicating that both methods are reliable and accurate for our dataset.

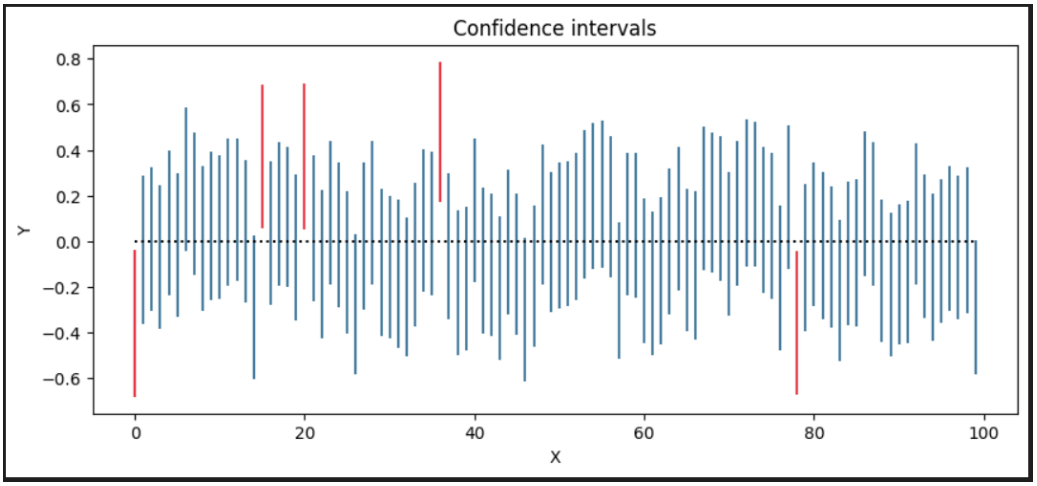
This result supports the robustness and reliability of the Bootstrap method, confirming that it can be an effective alternative to the asymptotic method when its assumptions might be questionable (for instance when we have a very wild distribution).

2.3 Part 4

In the last part we take 100 disjoint sets of 200 samples each and compute the confidence interval for the mean in each of the set, in order to ask "How many confidence intervals will contain the true mean?".

In order to solve this problem, we can just intrinsically think about the meaning of a confidence interval of confidence 95%. A confidence interval of confidence 95% means that with a probability of 0.95 it will contain the true value of the metric we are measuring. This means that, on average, 5 intervals (5%) will not contain the actual value of the mean!

To illustrate this concept graphically, we have plotted our confidence intervals, emphasizing those that do not include the true mean value (marked in red):



To conclude, it is important to notice that is not always guaranteed to have at every run 5 intervals not containing the mean. This result is just observable on the long term.