Report HW2

Anonymous Submission

Exercise 1

We aim to simulate a Poisson process with a fixed number of arrivals N. Given the time interval $[0, T]$, with $T = 100$ the rate parameter is set as $\lambda = \frac{N}{T}$.

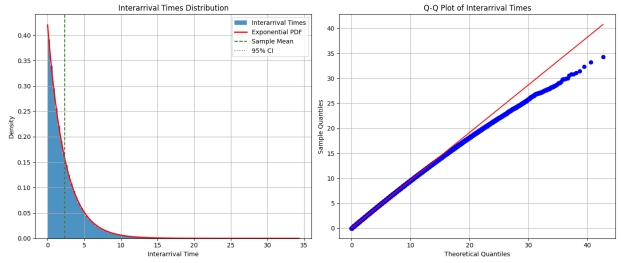
Part 1: Uniform Arrival Times

In this experiment, we generate k realizations of N arrival times sampled uniformly over $[0, T]$. Then we sort them to compute the inter-arrival times. We aggregate the inter-arrival times from all k to assess whether they follow an exponential distribution. Specifically, we:

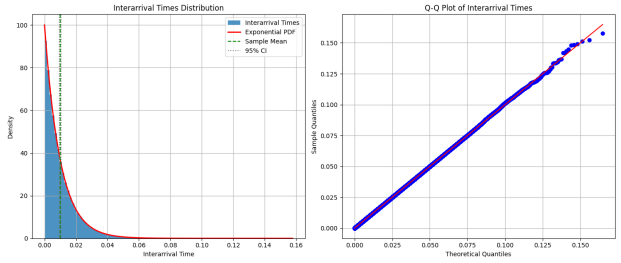
- Plot the histogram of all inter-arrival times across the k realizations, overlaid with the corresponding exponential probability density function (PDF), as shown in Figure 1.
- Construct a Q-Q plot comparing the empirical distribution of inter-arrival times with the theoretical exponential distribution (Figure 1).
- Compute empirical mean and variance, and compare them to theoretical values. (Table 1).

λ	k	Source	Mean	Var	Mean 95% CI	Mean 99% CI	Var 95% CI	Var 99% CI
100	10^3	Theoretical	0.01	0.00	–	–	–	–
100	10^3	Empirical	0.01	0.00	[0.01,0.01]	[0.01,0.01]	[0.00,0.00]	[0.00,0.00]
0.42	10^6	Theoretical	2.38	5.67	–	–	–	–
0.42	10^6	Empirical	2.33	5.16	[2.32,2.33]	[2.32,2.33]	[5.14,5.19]	[5.13,5.20]

Table 1: Summary of inter-arrival time statistics: mean, variance, and confidence intervals caluclated with the bootstrap method.



(a) $k = 10^6$ and $\lambda = 0, 42$.



(b) $k = 10^3$ and $\lambda = 100$.

Figure 1: Inter-arrival time histograms and Q-Q plots obtained from uniform distribution sampling.

Part 2: Exponential Inter-Arrivals

In this experiment, we generate k realizations of Poisson processes from N inter-arrival times drawn independently from an exponential distribution with mean $\frac{1}{\lambda}$. To sample valid inter-arrival times, we sample $N + 1$ inter-arrivals and check the following condition:

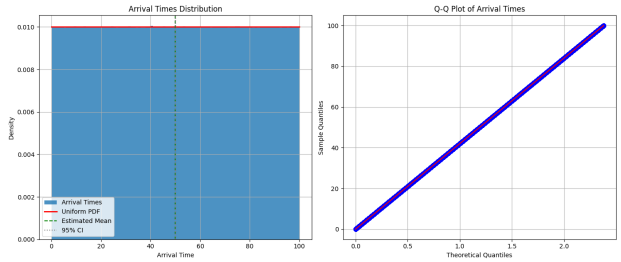
1. That the sum of the first N inter-arrival times does not exceed T.
2. That the sum of all $N + 1$ inter-arrival times was greater than T.

Condition 2 ensures that there is no bias towards small inter-arrivals. Arrival times are then obtained via cumulative summation.
We then analyze the resulting arrival times by:

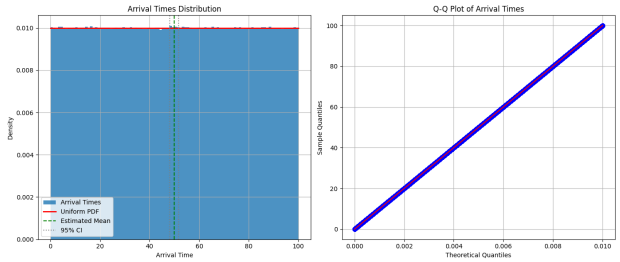
- Plotting their histogram and overlaying the theoretical uniform PDF on $[0, T]$, as illustrated in Figure 2.
- Generating a Q-Q plot to compare the empirical distribution with the theoretical uniform distribution.
- Computing the empirical mean and variance of the aggregated arrival times and assessing their agreement with uniformity assumptions. (Table 2).

λ	k	Source	Mean	Var	Mean 95% CI	Mean 99% CI	Var 95% CI	Var 99% CI
100	10^3	Theoretical	50.00	833.33	–	–	–	–
100	10^3	Empirical	50.00	833.12	[48.17,51.75]	[47.64,52.22]	[783.70,880.25]	[763.66,907.33]
0.42	10^6	Theoretical	50.00	833.33	–	–	–	–
0.42	10^6	Empirical	50.01	833.37	[49.95,50.06]	[49.93,50.08]	[831.91,834.74]	[831.52,835.37]

Table 2: Summary of arrival time statistics: mean, variance, and confidence intervals caluclated with the bootstrap method.



(a) $k = 10^6$ and $\lambda = 0, 42$.



(b) $k = 10^3$ and $\lambda = 100$.

Figure 2: Arrival time histograms and Q-Q plots obtained from exponential inter-arrival times..

Exercise 2

Rejection Sampling

To apply the rejection sampling, we need to find a distribution such that we can bound the probability $f(x)$ such that $\frac{f(x)}{g(x)} \leq c$. A valid proposal distribution can be:

$$g(x) = \begin{cases} k\,x^2 & \text{if} \quad -3 \leq x \leq 3 \\ 0 & \text{otherwise} \end{cases}$$

where k is such that $\int_{-3}^{3} g(x)\,dx = 1$. Such k is given by:

$$\int_{-3}^{3} g(x)\,dx = \int_{-3}^{3} kx^2\,dx = 18k \overset{!}{=} 1 \Rightarrow k = \frac{1}{18}$$

The proposal distribution is then $g(x) = \frac{x^2}{18}$ and the bound is:

$$\frac{\frac{1}{A}x^2\sin^2(\pi x)}{\frac{x^2}{18}} = \frac{18}{A}\sin^2(\pi x) \leq c \Rightarrow c = \frac{18}{A} = 2.03435386. \tag{1}$$

To draw a sample from the distribution $g(x)$ we can apply the CDF-Inversion method, so we derive the inverse of the CDF $G(x)$:

$$G(x) = \int_{-3}^{x} g(t)\,dt = \int_{-3}^{x} \frac{t^2}{18}\,dt = \frac{1}{2} + \frac{x^3}{54} \Rightarrow G^{-1}(u) = 3\sqrt[3]{2(u - \frac{1}{2})} \quad \text{for} \quad u \in [0, 1].$$

We can now choose 2 ways of applying the rejection sampling, the first with the knowledge of A, which allow us to have an higher acceptance rate, or without knowledge of A, which allow us to know the target distribution up to a constant but it does not let us to bound it so tightly, having a drop in the acceptance rate. We report both methods in the following.

Algorithm with the full knowledge of $f(x)$

In this case we assume to know the value of A, so that we know also the value of c from Equation 1, so we can bound $f(x)$ with the function $cg(x) = \frac{18}{A}\frac{x^2}{18} = \frac{x^2}{A}$ which is basically the envelope of $f(x)$. Thus, the algorithm is the one reported in Algorithm 1.

The distribution resulting from Algorithm 1 is shown in Figure 3a. The acceptance rate of this algorithm in 10^8 iterations is $\frac{10^8}{49155544} \approx 0.49$. The distribution of the drawn samples is shown against the theoretical one in Figure 3a, while the accepted samples and the rejected samples (from the proposal) are shown in 4a.

Algorithm 1 Rejection sampling with full knowledge of $f(x)$

1: Draw $u_1 \sim \mathcal{U}[0,1]$
2: Compute $x = G^{-1}(u_1)$
3: Draw $u_2 \sim \mathcal{U}[0, c \cdot g(x)]$
4: **if** $u_2 < f(x)$ **then**
5: Accept x
6: **else**
7: Go back to Step 1
8: **end if**

Algorithm with knowledge of $f(x)$ up to a constant

Given that $f(x) = \frac{1}{A}x^2 \sin^2(\pi x) = \frac{1}{A}f^n(x)$, if we do not know the value of A, we can still bound the non-normalized distribution $f^n(x)$:

$$f^n(x) = x^2 \sin^2(\pi x) \leq x^2 \leq 9 = M \quad \text{in} \quad -3 \leq x \leq 3 \tag{2}$$

We can use this upper bound do apply rejection sampling as in Algorithm 2.

Algorithm 2 Rejection sampling with knowledge of $f(x)$ up to a constant

1: Draw $x \sim \mathcal{U}[-3,3]$
2: Draw $u \sim \mathcal{U}[0,M]$
3: **if** $u \leq f^n(x)$ **then**
4: Accept x
5: **else**
6: Go back to Step 1
7: **end if**

We see that in this algorithm we do not employ A at all, and the proposal distribution is a uniform distribution. We build a "bounding box" around the scaled distribution $f^n(x)$ and accept only those points that fall under the plot of $f^n(x)$. This approach works, but, clearly, being less precise in the bounding, the acceptance rate drops: the acceptance rate on 10^8 iterations is $\frac{10^8}{16381697} \approx 0.16$. The distribution of the drawn samples is shown against the theoretical one in Figure 3b, while the accepted samples and the rejected samples (from the proposal) are shown in Figure 4b.

Confidence intervals

For the computation of the confidence intervals, since the data are i.i.d., we use the order statistics and the binomial distribution (approximated with a normal, since $n = 200$ is large enough) for the CIs of the median and the 0.9-quantile. For a quantile q, the formula for CI $[X_{(j)}, X_{(k)}]$ is (using the normal approximation of the binomial):

$$j \approx \lfloor nq - 1.96\sqrt{nq(q-q)} \rfloor, \quad k \approx \lfloor nq + 1.96\sqrt{nq(q-q)} \rfloor + 1.$$

Instead, for the mean, we exploit the central limit Theorem (since we have enough data points and the distribution is symmetric) and obtain the confidence interval:

$$\hat{\mu_n} \pm \eta \frac{s_n}{\sqrt{n}}$$

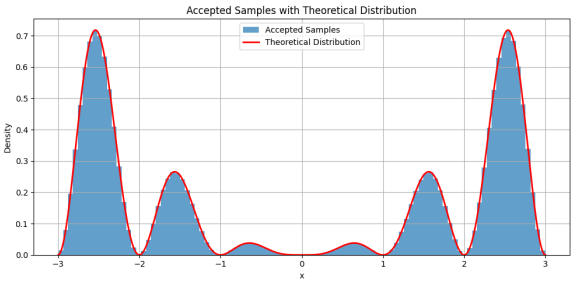
For the bootstrap procedure, the only assumption is to have i.i.d. data, so we can also apply it.

The plots of the median and the 0.9-quantile computed with the two different approaches are shown in Figure 5.

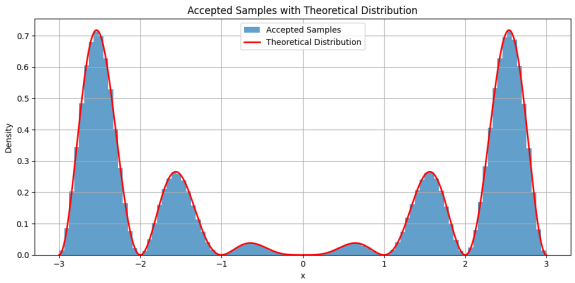
The plots for the mean are shown in Figure 6.

Statistical significance of the mean's confidence intervals

We subdivided 20000 variates into 100 disjoint sets and for each of them we computed the 95% confidence intervals, both with the Gaussian approximation approach and the Bootstrap approach. As expected, the number of confidence intervals containing the true mean is 95 in the case of the Gaussian approximation and 94 in the case of the bootstrap CIs (we can see that the bootstrap method slightly understimates the CI, so we find less accurate CIs). Figure 7 shows the distribution of the sample mean computed in the 100 sets, showing that the Central Limit Theorem holds over this statistic.

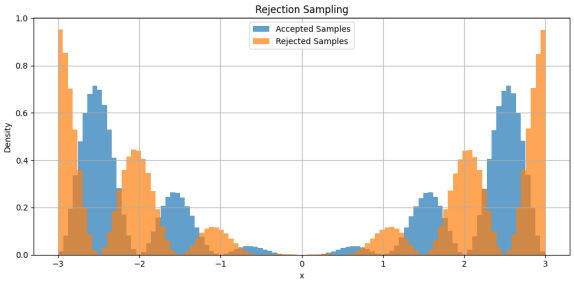


(a) Rejection sampling with Algorithm 1.

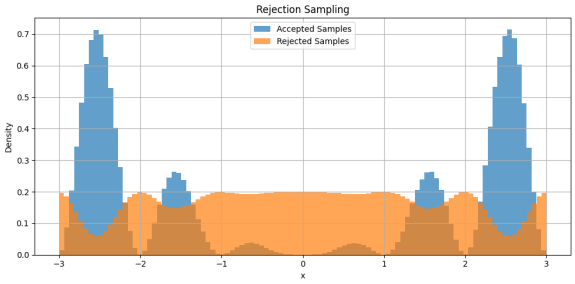


(b) Rejection sampling with Algorithm 2.

Figure 3: Rejection sampling of $f(x)$

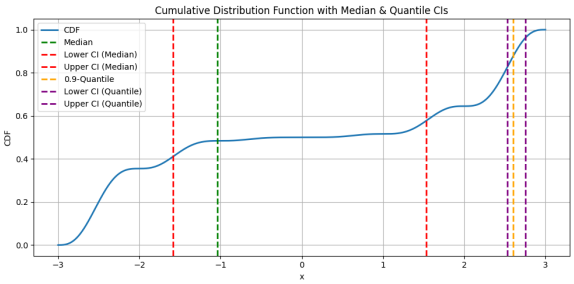


(a) Rejected/accepted samples with Algorithm 1.

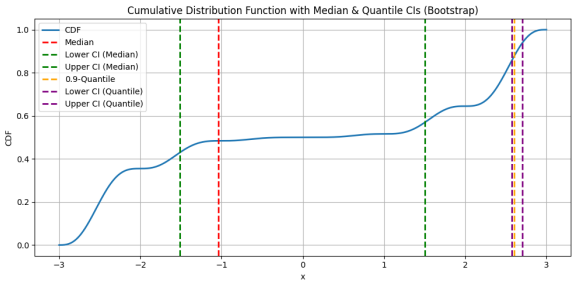


(b) Rejected/accepted samples with Algorithm 2.

Figure 4: Accepted/rejected samples distributions.

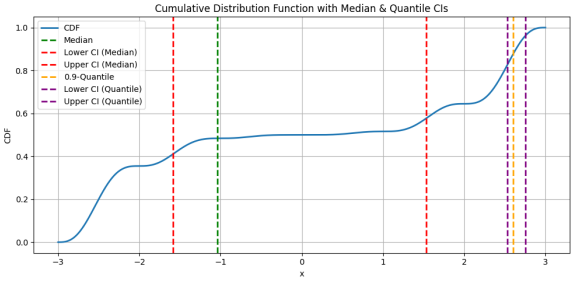


(a) Median and 0.9-Quantile CIs computed with the Binomial formula.

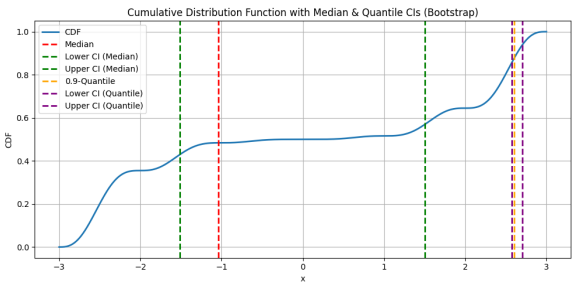


(b) Median and 0.9-Quantile CIs computed with the Bootstrap heuristic.

Figure 5: CDF with median and 0.9-quantile CIs.



(a) Mean CI computed with the Gaussian approximation.



(b) Mean CI computed with the Bootstrap heuristic.

Figure 6: PDF and mean CI.

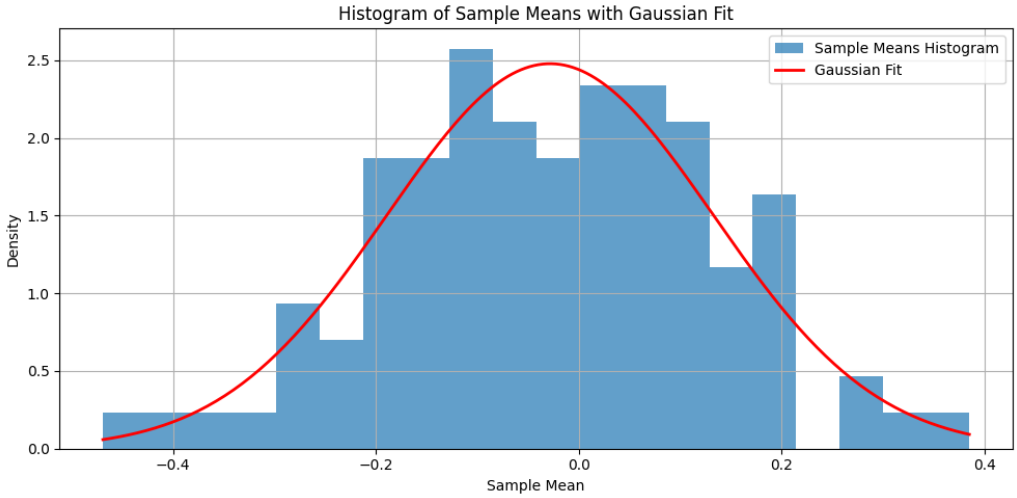


Figure 7: Empirical PDF of the sample mean.