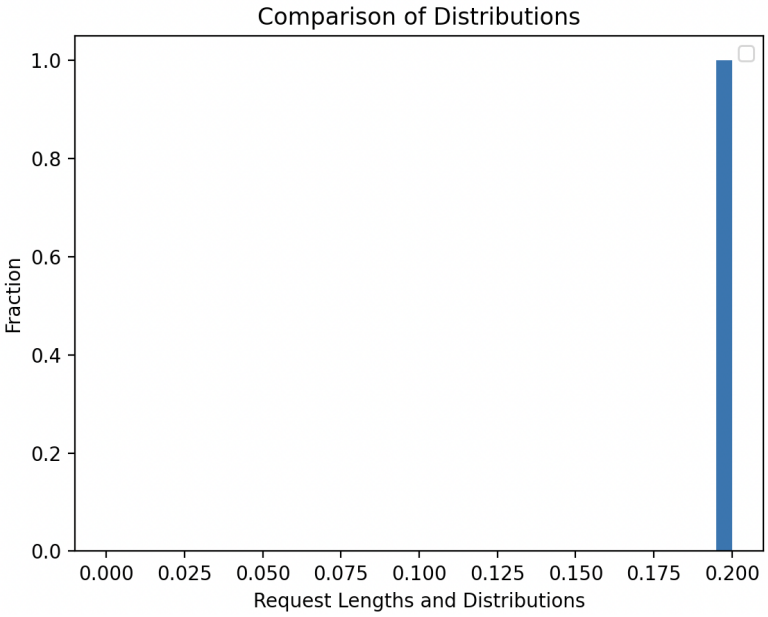
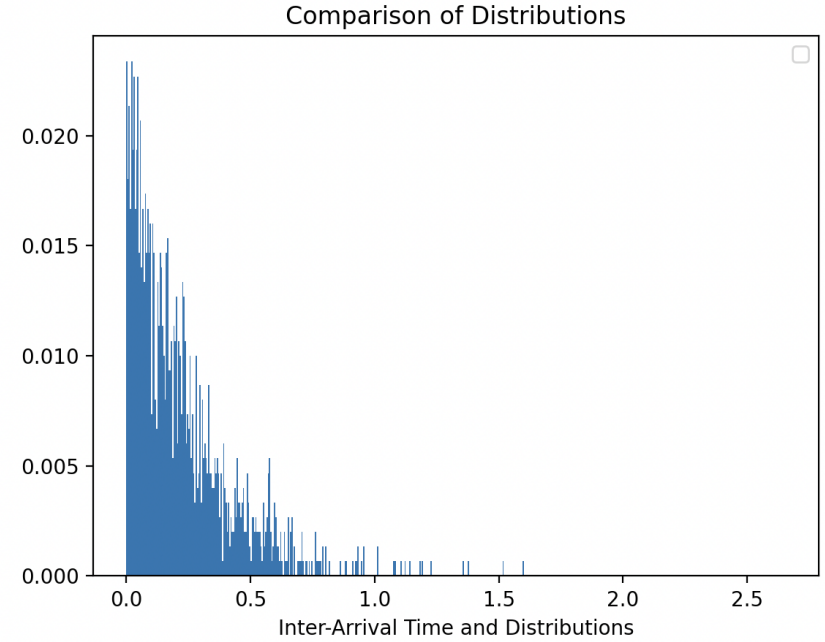
HW3 EVAL

1.

a.



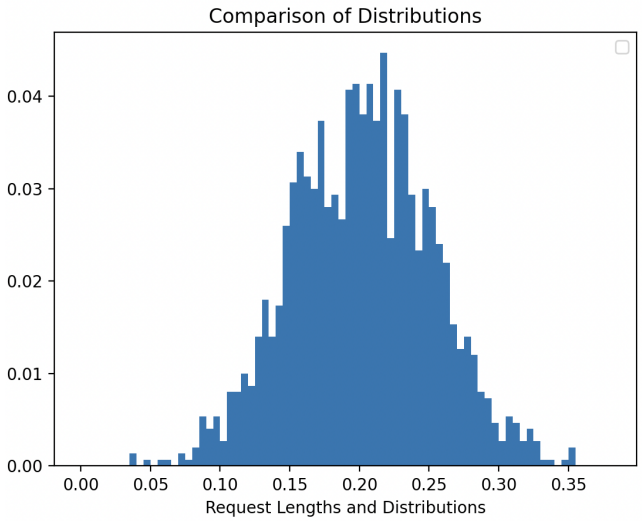


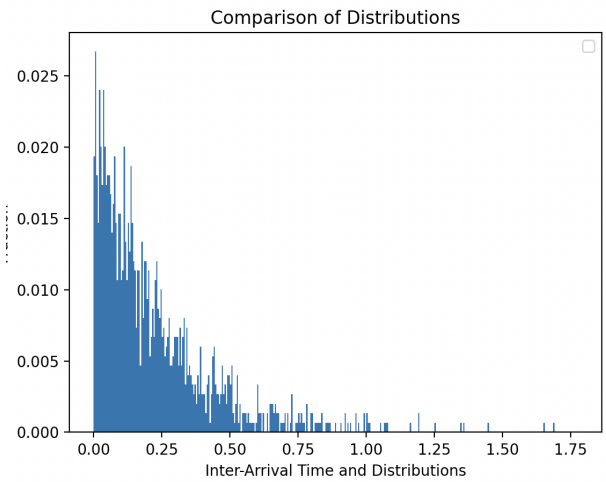
The distribution of length when -d is set to 1 follows point mass distribution (deterministic service time or M/D/1) because all the request lengths are the same (0.2 second). mean = 0.2, standard deviation = 0.

The distribution of inter-arrival time when -d is set to 1 is still exponential distribution with mean and standard deviation equals 1/4.5.

As a result, -d only controls distribution of request lengths when set to 1.

b.

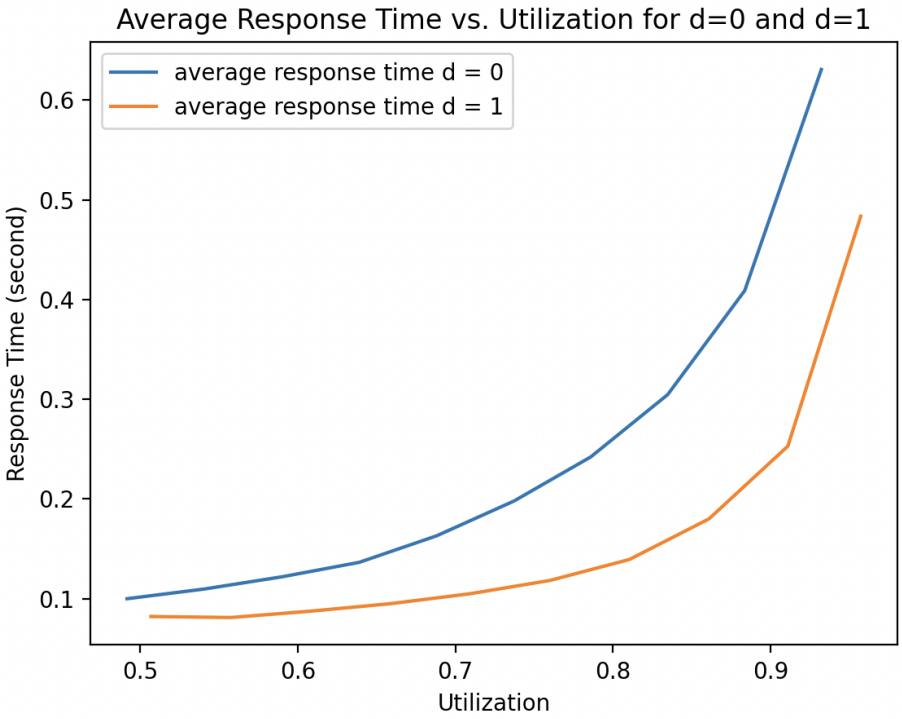




- The distribution of request lengths when -d = 2 follows normal distribution. mean = 0.5, standard deviation = 0.0504.
- The distribution of inter-arrival time when -d = 2 follows exponential distribution. mean = standard deviation = 1/4.5.

As a result from the two cases above, -d controls only distribution of request lengths when set to 1, not controlling the inter-arrival time.

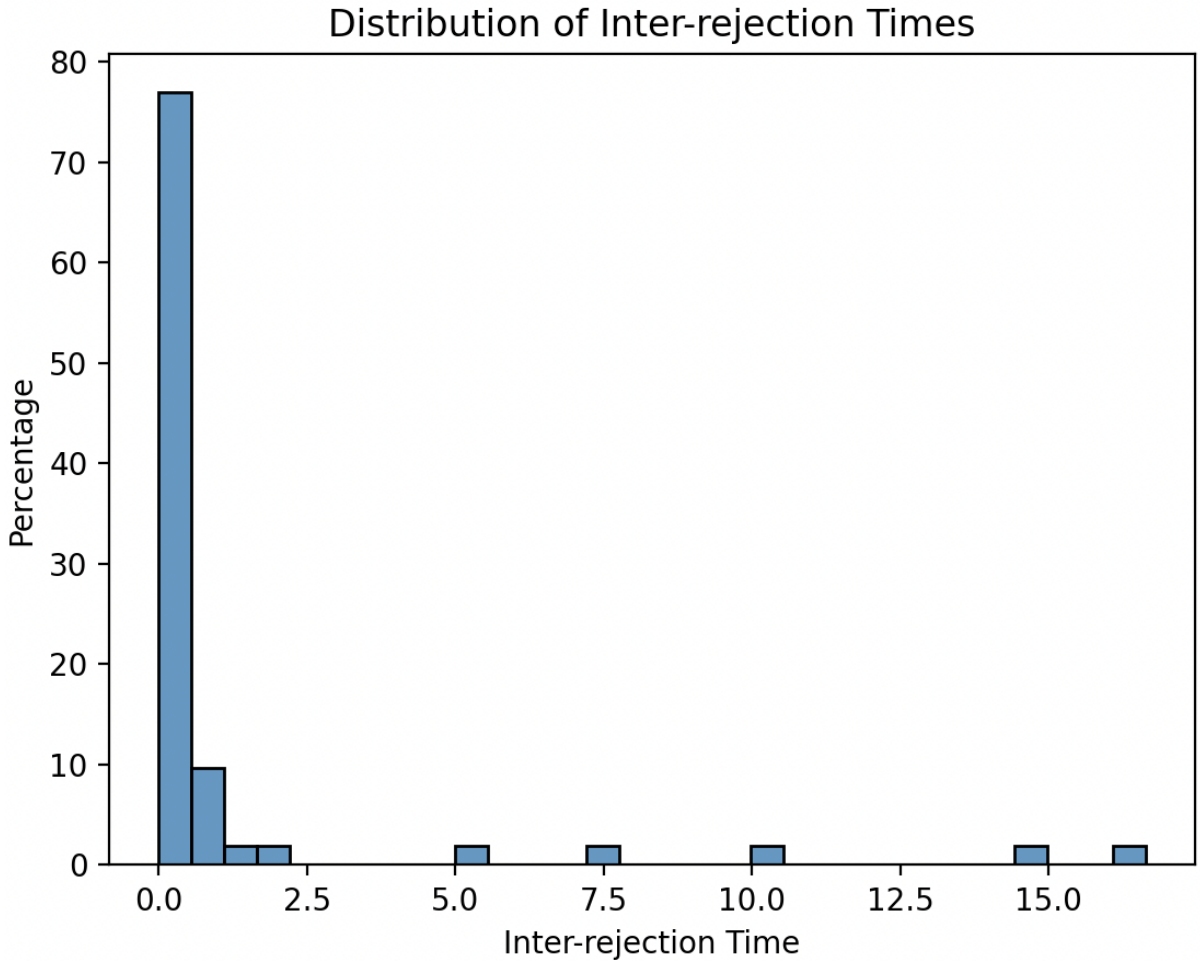
c.



Average response time T_q when -d is set to 0 (request lengths follow exponential distribution) is generally higher than T_q when -d is set to 1 (request lengths are deterministic and unchanged) on same utilization of the server. Back to the question, quality of service perceived by clients are better in M/D/1 (-d = 1).

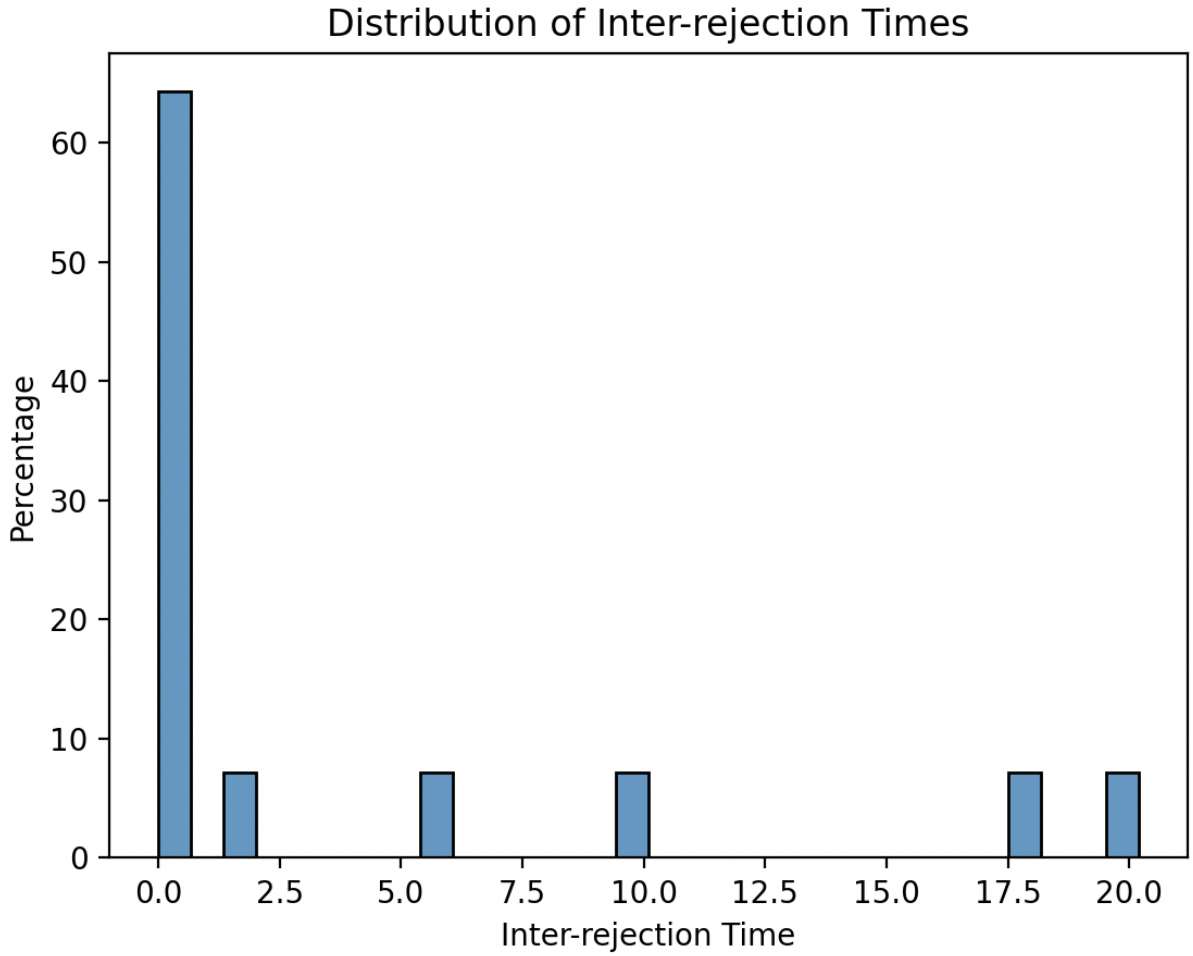
d.

ratio of rejected requests over the total: 0.035 as calculated by Python (using Regex to extract the number of rejected requests)



Inter-rejection time mostly clustered around a small value near zero (at a percentage of around 77), with some sparsely distributed around higher value. This means that a great portion of rejected queue is crowded inside a short period of time. During the time period, the queue keeps at a high load, frequently rejecting incoming requests.

e.
rejection rate when d = 1 is 0.01, using the same procedure as above to calculate.



Compared to exponentially distributed service time (d=0), deterministic service time behaves better since the portion of crowd surrounding 0-1 inter-rejection time is less than that in the above. This means that there is fewer times that the queue is at a intense load. From the user perspective, fewer times when server is in high load lead to better perception of service as wait time is shorter.