STAT 5000 Project:
Kolmogorov-Smirnov Tests on Firefly flash data

Krithikesh Ravishankar Rahul Baikadi

December 16, 2023

Motivation

In this paper, we are comparing the data obtained by simulating the model described in [1] with experimental data, to analyse if they are sampled from the same distribution. The experiment performed in [1], is based on the following: fireflies, when present in a group, flash in a synchronized manner such that over time, most fireflies flash at approximately the same moment. Experiments are conducted where different numbers of fireflies are introduced in a tent, and the time interval between two consecutive flashes of the group are obtained. It is observed this inter-flash time-interval is sampled from a probability density function, that changes as the number of fireflies in the tent changes. A computational model is proposed, which uses the number of fireflies as an input parameter, to simulate the flashing behavior of the fireflies. From these simulations, a probability distribution for the inter-flash interval is calculated. The computational model has a tuning parameter β, which is varied from 0 to 1, and for each value of β, a different probability distribution is generated. These distributions are compared with the distribution obtained from the experimental data, and an optimal value for the tuning parameter β is obtained. In order to compare the distributions, the Kolmogorov-Smirnov test is used, where the value of β that produced the smallest KS-statistic is chosen as the optimal value of β.

Theory

Determining whether two samples came from the same distribution is an old problem with constant relevance. Particularly when two distributions may have the same mean, but differ in other important ways, testing their similarity can be both difficult, and critical. The Kolmogorov-Smirnov test (KS- test) is a test that is used for this purpose to answer the following question - how likely is it that two samples are obtained from the same underlying distribution.

Consider the following situation: there are two independent samples: A and B, of sizes n_a and n_b. Within each sample, all observations are independently drawn from the same distribution: $A_i \overset{iid}{\sim} E$ and $B_i \overset{iid}{\sim} F$. Our null hypothesis is that the two (cumulative) distributions are the same, $H_0 : E = F$, and the alternate hypothesis is that they are not the same $H_1 : E \neq F$ Without making any further assumptions, we would like a valid (and ideally consistent/powerful) test of this hypothesis.

$$A_i \overset{iid}{\sim} E$$
$$B_i \overset{iid}{\sim} F$$
$$H_0 : E = F$$
$$H_1 : E \neq F$$

For the two-sample KS test, the KS test statistic is defined as the maximum value of the difference between the samples A and B's empirical cumulative distribution functions (ECDF).The empirical CDF of a random sample X_i is defined as:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(X_i)$$

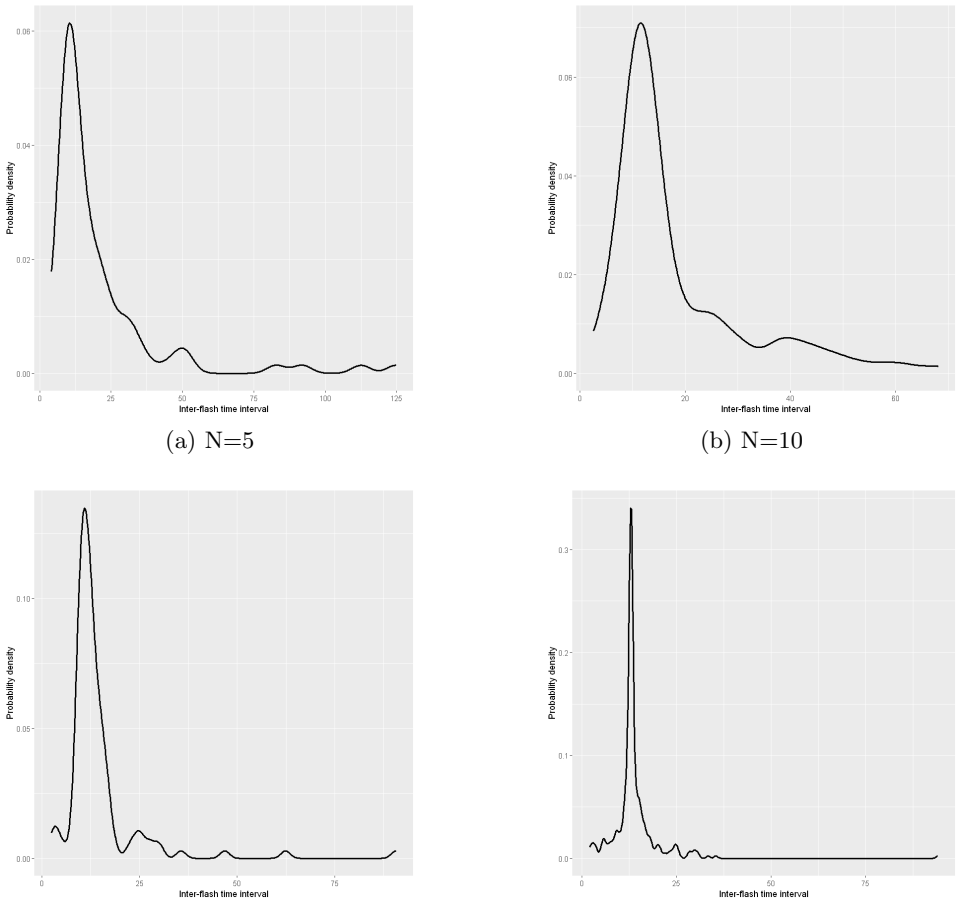
where $I_{(-\infty,x]}(X_i)$ is the indicator function, equal to 1 if $X_i \leq x$ and 0 otherwise. Thus, the term in the numerator counts the number of elements in the sample that are lesser than or equal to x. For the two-sample KS test, the KS statistic is defined as:

$$D = \sup_x |F_1(x) - F_2(x)|$$

where $F_1(x)$ and $F_2(x)$ are the empirical CDF's of the first and second sample respectively, and *sup* is the supremum function. In other words, the empirical distribution functions of the two samples $F_1(x)$ and $F_2(x)$ are generated; for each value of x, the absolute difference between the values of $F_1(x)$ and $F_2(x)$ is computed, and the maximum value of these absolute differences is defined as the KS test statistic.

Results and Discussions

The following probability density functions are obtained for the inter-flash interval from experimental data, for different values of number of fireflies in the tent, N:



(a) N=5 (b) N=10

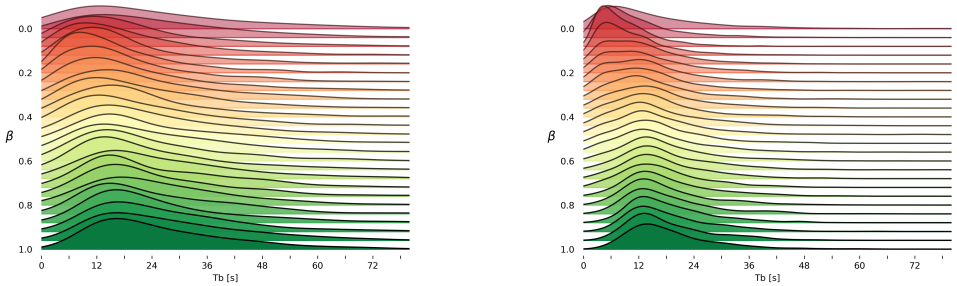
(c) N=15 (d) N=20

Figure 1: Shows the inter-flash interval probability density plots from experimental data based on [1].

The following computational model is used to model the behavior of firefly i:

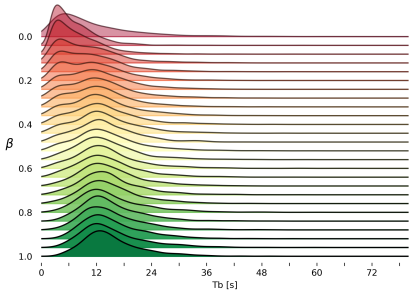
$$\frac{dV_i(t)}{dt} = \frac{1}{T_{si}}\epsilon_i(t) - \frac{1}{T_{di}}[1-\epsilon_i(t)] + \epsilon_i(t)\beta\sum_{i,j}^{N}\delta_{ij}[1-\epsilon_j(t)]$$

For N fireflies, the above model is simulated for each of the fireflies, and the inter-flash interval of the group is computed. In the model, β is a tuning parameter that can go from 0 to 1, such that changing the value of β changes the distribution of the inter-flash interval. Thus, for each possible value of β from 0 to 1, with an increment of 0.01, the inter-flash probability distribution is generated and the following was obtained:

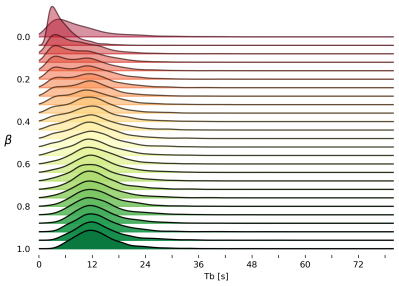


(a) N=5

(b) N=10



(c) N=15



(d) N=20

Figure 2: Shows the inter-flash interval probability density plots from recreated simulations based on the computational model from [1].

.

For each value of N, for each value of β, the KS Statistic is calculated, and the value of β for which the KS statistic is the least is chosen as the optimal β for that value of . For an arbitrary value of $N = 10$, and $\beta = 0.32$, the Empirical Cumulative Distribution Functions looks as follows:

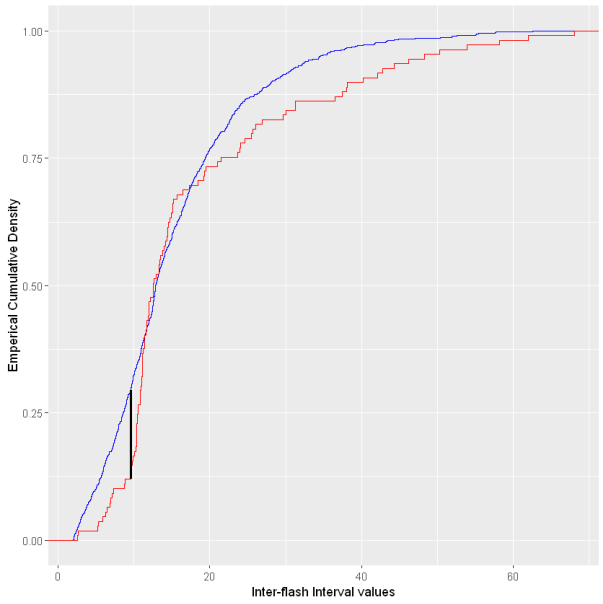
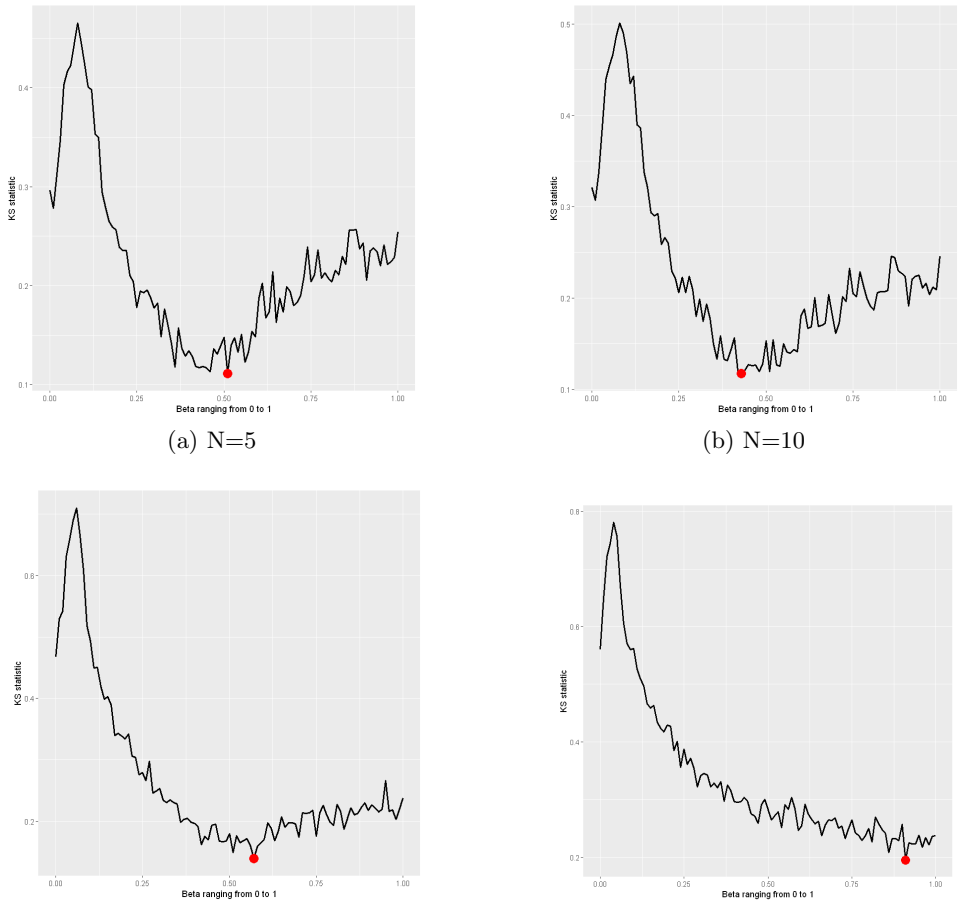


Figure 3: Empirical Cumulative Distribution Functions of the sample from experimental data (blue), and simulated data for $N = 10$ and $\beta = 0.32$. The black line represents the value where the difference the two empirical distributions is maximum. In this case, this value was observed 0.1747

The following KS statistic plots were obtained for each value of N:



(a) N=5 (b) N=10

(c) N=15 (d) N=20

Figure 4: Shows the KS statistic plots for each value of N. The value of β for which the least value of the KS statistic is observed is the optimal value of β for that N. (a) For N = 5, $\beta = 0.51$, (b) For N = 10, $\beta = 0.43$, (c) For N = 15, $\beta = 0.57$, (d) For N = 20, $\beta = 0.91$

.

The optimum values of β for $N = 5, 10, 15, 20$ were observed to be 0.51, 0.43, 0.57 and 0.91 respectively.

Conclusions

The KS test was used effectively to obtain optimum values for the tuning parameter β for different values of N. The KS test was useful for this purpose, since it is a non-parametric test that can be used to compare two samples directly, without having to compare them with an analytic distribution with parameters, such as the normal or gamma distributions.

References

[1] Sarfati, R., Joshi, K., Martin, O., Hayes, J. C., Iyer-Biswas, S., & Peleg, O. (2023). Emergent periodicity in the collective synchronous flashing of fireflies. *eLife*, 12, e78908. Advance online publication. https://doi.org/10.7554/eLife.78908