Performance evaluation: Homework 2

Vaubien Jimi 226977, Zimmermann Timon 223720

February 13, 2019

1 Statistics warmup

1.1 Problem 1

The different generated histograms for each number of *iid* samples can be found on Figure 1.

1.2 Problem 2

1. The plots of the different density functions can be found in Figure 2, we plotted identical distributions on the same graph with distinctive red and black colors.

2. The standard normal QQ plots can be found in Figure 3, each plot represent the result with respect to one of the different distributions.

3. When the QQ-plot has an S-shape, i.e. its tails are rotating clockwise, it means that the distribution has light tails (Platykurtic kurtosis) and such distributions are often called *sub-Gaussian*.
On the other hand, if the QQ-plot has an "inverted" S-shape, i.e. its tails are rotating counter-clockwise, it means that the distribution has heavy tails (Leptokurtic kurtosis) and such distributions are often called *super-Gaussian*.
Now if the QQ-plot has an U-shape, i.e. the left tail rotates clockwise and the right one counter-clockwise, it is said that the distribution has *right skew*.
More generally, when a specific tail (right or left) twists off *counter-clockwise* (respectively *clockwise*) compared to the reference line, it means that the given distribution has *more* (respectively *less*) data on this specific tail than a theoretical normal distribution.

2 Simulate random waypoint and look at what you have done

2.1 Problem 3

1.

Mean*$_{\text{number of waypoints}}$	152.35
Minimum*$_{\text{number of waypoints}}$	131
Maximum*$_{\text{number of waypoints}}$	178

*among 100 users running for 1 day

2. The trajectory of the users 5 different are displayed in Figure 4. To give the reader a good grasp of how the trajectory evolves over time, we use a gradient to indicate the different timestamps at which the waypoints have been generated.

3 Different viewpoints

3.1 Event average (palm) viewpoint

a) The two histograms (for a single mobile and all the different mobiles) can be found in Figure 5, as it can be seen the value of the speed is correctly sampled uniformly between the two values $(0.7 + 0.01 * (226977\%21))$ and $(1.9 + 0.1 * (226977\%21))$. It is even more clear among all the mobiles, indeed the histogram converges clearly to the uniform density distribution.
Both histograms are sampled at each transition time T_n.

b) The two position histograms of discretized area is shown in Figure 6, here we divided the square in 20x20 bins. We clearly see the visit frequency proportional to the intensity of green. The passage frequency is distributed uniformly over the whole area, as can be see on the *right* histogram where we used 100 mobiles.
Both histograms are sampled at each transition time T_n.

3.2 Time average viewpoint

a) The two speed histograms corresponding to one mobile and all mobiles are shown in Figure 7. This histogram results from the sampling the speed every 12 seconds.

b) The two position histograms corresponding to one mobile and all mobiles are shown in Figure 8. This histogram results from the sampling the position every 12 seconds, discretized in 20x20 bins. We clearly see that the mobiles frequencies are concentrated in the center of the map, whether it is for 1 or 100 mobiles.

c) To begin with, the speed seems to be uniformly distributed for the event-based and time-based sampling. We find this result weird but could not find any error in our implementation of the simulator or the algorithm. Indeed, the speed should not be distributed uniformly when sampled on a time interval since the lower speed should be observed more frequently than the higher ones. This comes from the fact that we spend more time at lower speed for the same distance, and hence we should obtain an imbalanced distribution with more weight on lower speeds.
On the other hand, the position samplings make sense. With the event based sampling the distribution is uniform from the simulator construction and for an observation taken at an arbitrary transition time T_n. Respectively, the time-based sampling seems to be distributed with more weight around the center area, which makes sense since you have a higher chance to be observed through the center at an arbitrary point in time.

4 Confidence intervals

4.1 Confidence intervals for medians and means

a) The plot with $N = 100$ for the median, the mean and their respective confidence intervals can be found in Figure 9 for the sampling done at transition steps T_n for event n, and in Figure 10 for the time-based sampling. The assumptions are that the draws are done independently and they are identically distributed. The distribution also must not be heavy-tailed in order for the *Law of Large Numbers* to apply to the mean and its confidence interval. These assumptions have been verified by the aforementioned answers to the exercise 3.

b) The event-based transition sampling offers a much more concentrated mean (blue) and median (red) values among all the simulations, the confidence interval for the mean is much tighter than the one for the median which is spread at its border (v_{max} and v_{min}).
For the time-based sampling, the mean and median are not concentrated at all among different simulations, they vary by a large margin but the respective confidence interval is very tight. It is also important to note that they are combined and the same, whether it be the median/mean confidence interval or the mean/media in itself.

c) The plot with $N = 25$ for the median, the mean and their respective confidence intervals can be found in Figure 11 for the sampling done at transition steps T_n, and in Figure 12 for the time-based sampling.

d) We notice the same results than the one with $N = 100$ simulations.

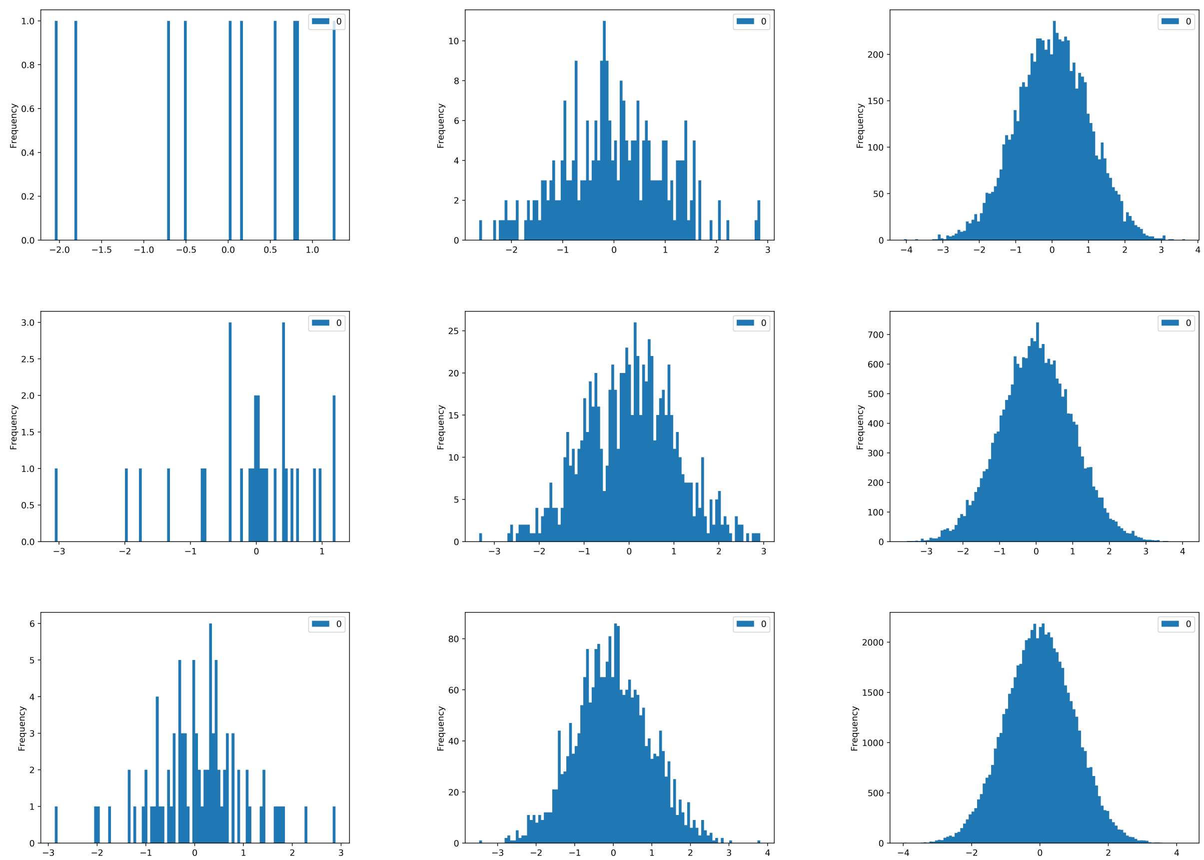

Figure 1: Histograms for different numbers of *iid* samples $n = 10, 30, 90, 270, 810, 2430, 7290, 21870, 65610$ (from top to bottom and then left to right).

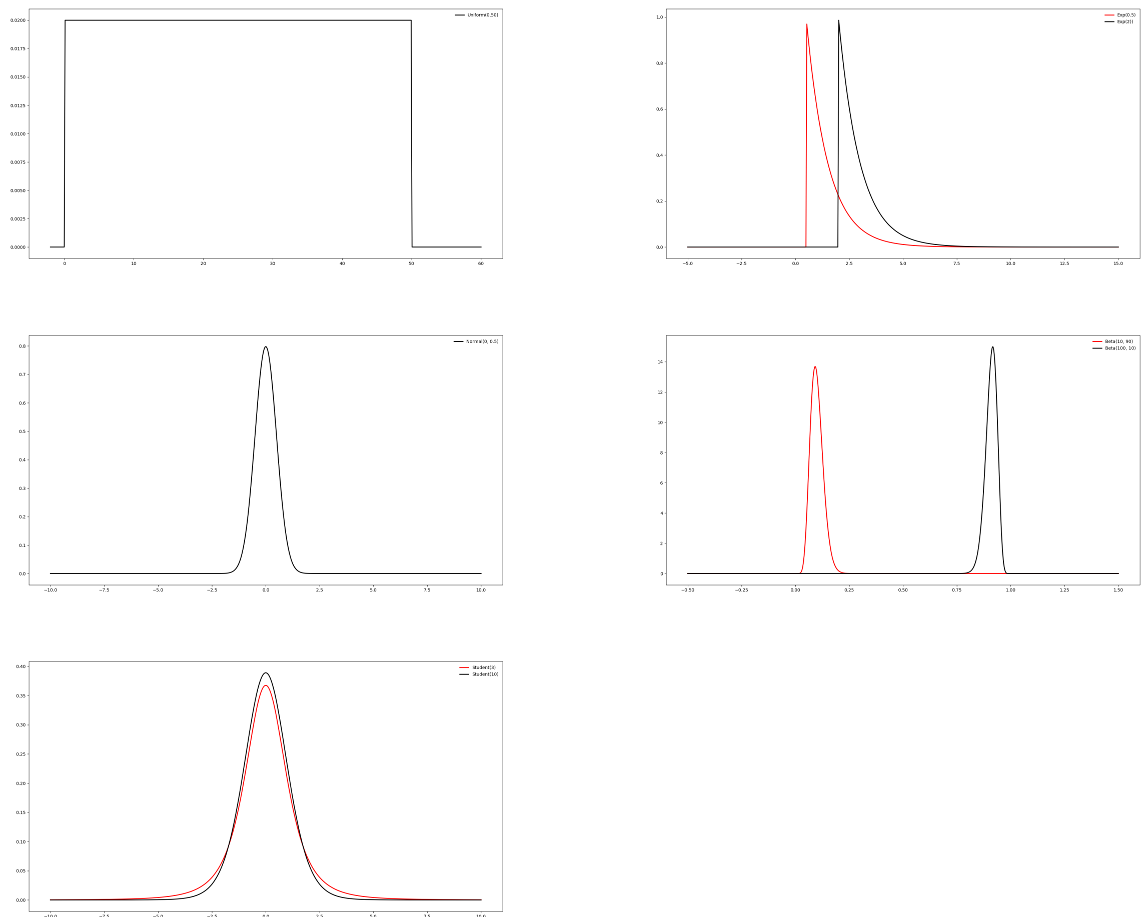

Figure 2: Density function of the distribution (from top to bottom and then left to right) Uniform(0, 50) - Normal(0, 0.5) - Student(3) and Student(10) - Exp(0.5) and Exp(2) - Beta(10, 90) and Beta(100, 10).

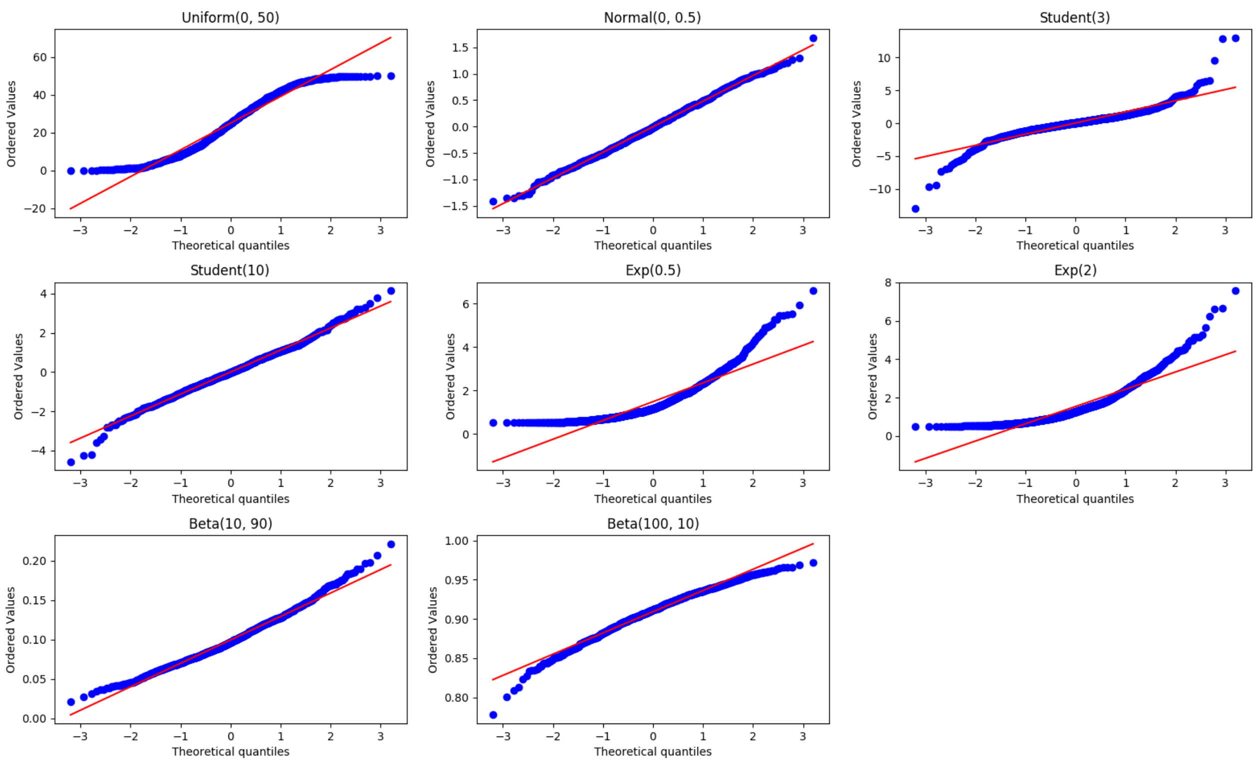

Figure 3: Standard normal Q-Q plots with respect to different distributions.

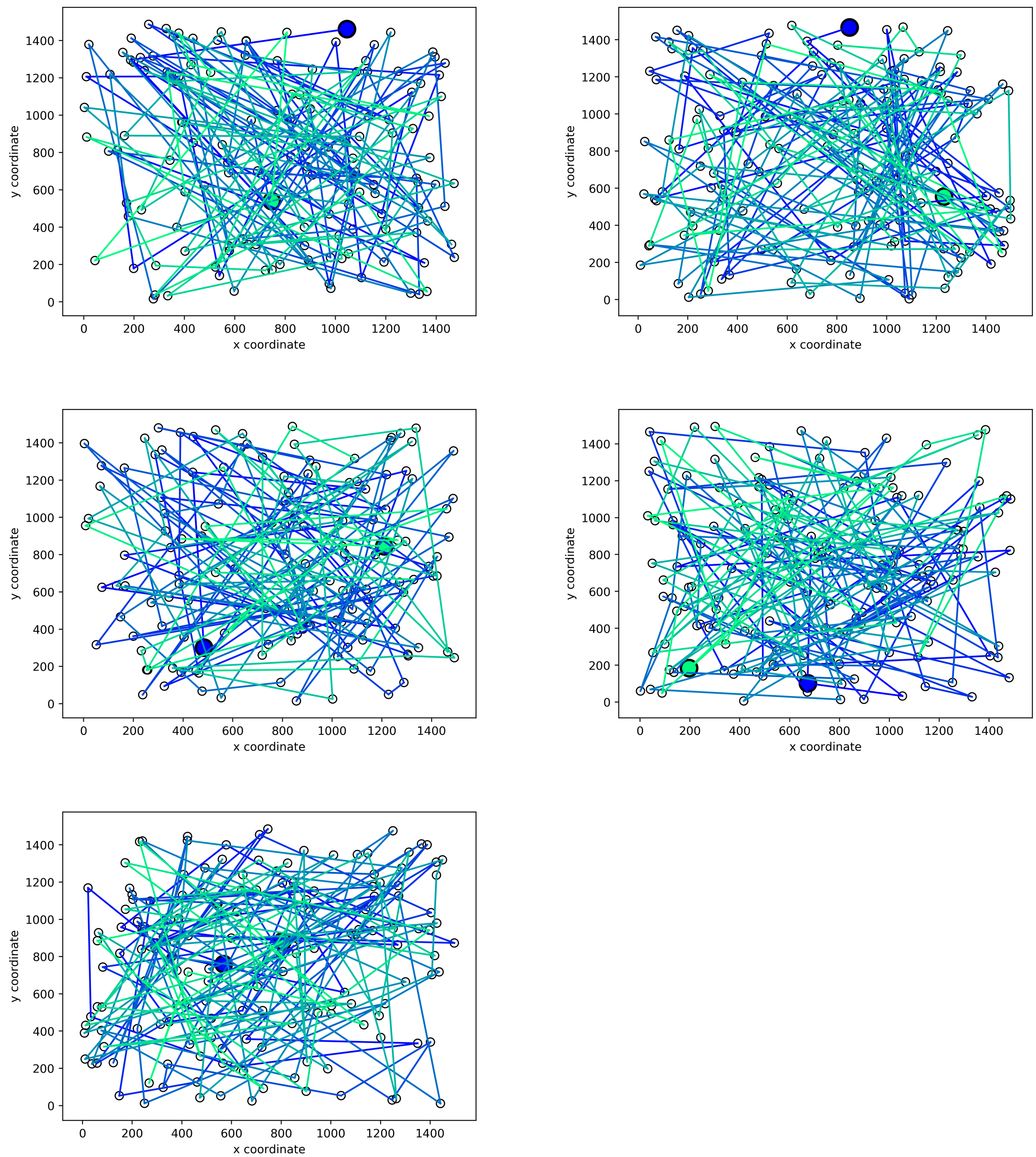

Figure 4: Trajectories and waypoints of 5 different users, the blue dot is the first waypoint and the green one is the last one, the gradient has a time meaning. Waypoints closer to T_0 are closer to blue (respectively, waypoints closer T_{last} are closer to green).

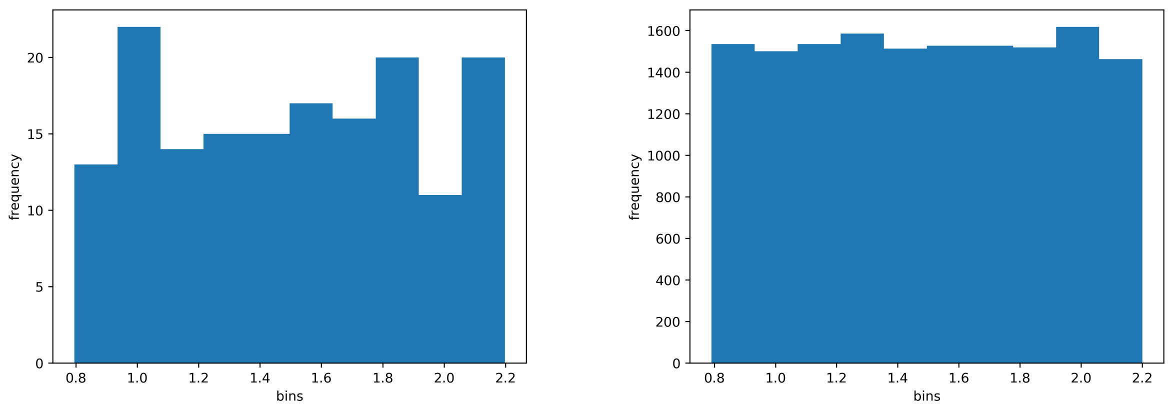

Figure 5: Histograms of the frequency for the speed, for one mobile on the *left* and all mobiles (100) on the *right*. Sampled every T_n events.

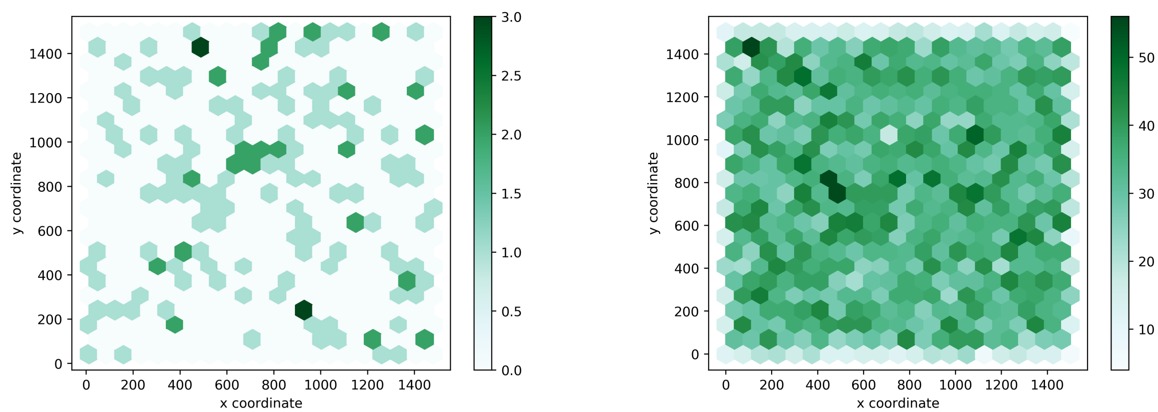

Figure 6: Histograms of the frequency for the position, for one mobile on the *left* and all mobiles (100) on the *right*. Sampled every T_n events.

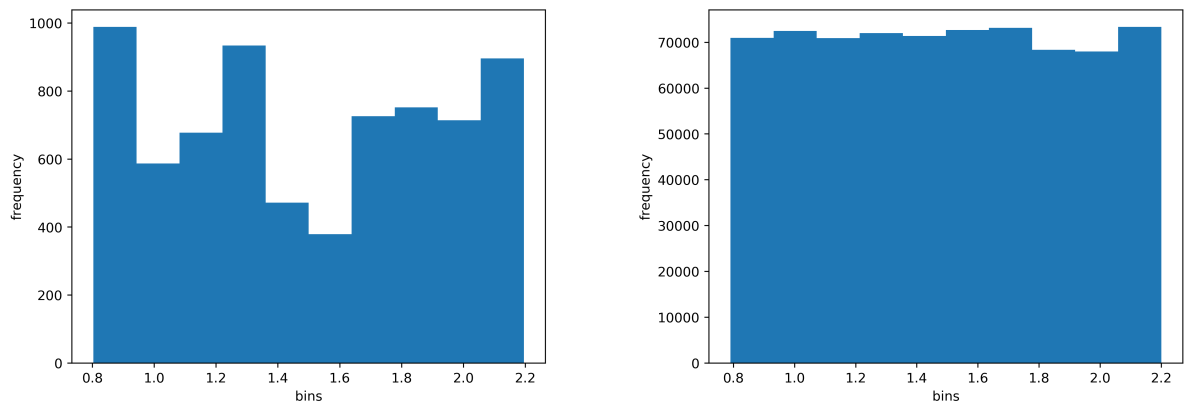

Figure 7: Histograms of the frequency for the speed, for one mobile on the *left* and all mobiles (100) on the *right*. Sampled every 12 seconds.

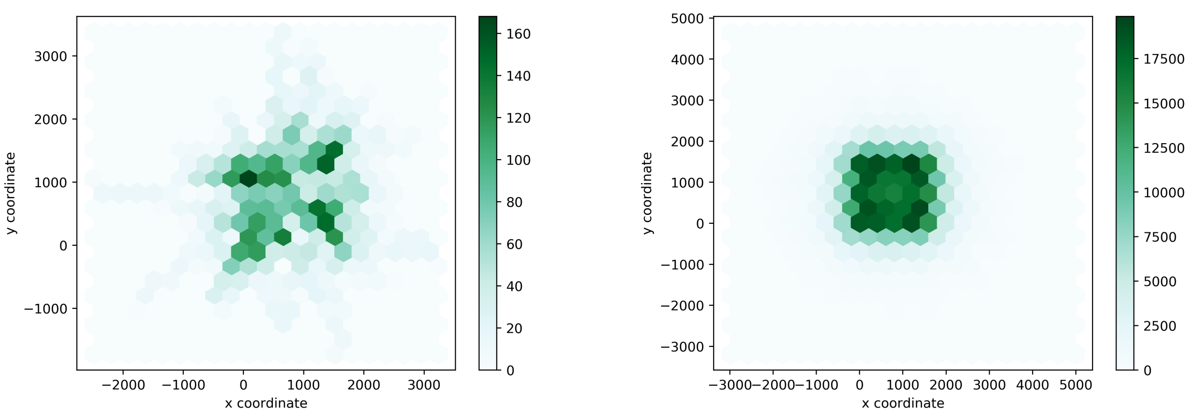

Figure 8: Histograms of the frequency for the position, for one mobile on the *left* and all mobiles (100) on the *right*. Sampled every 12 seconds.

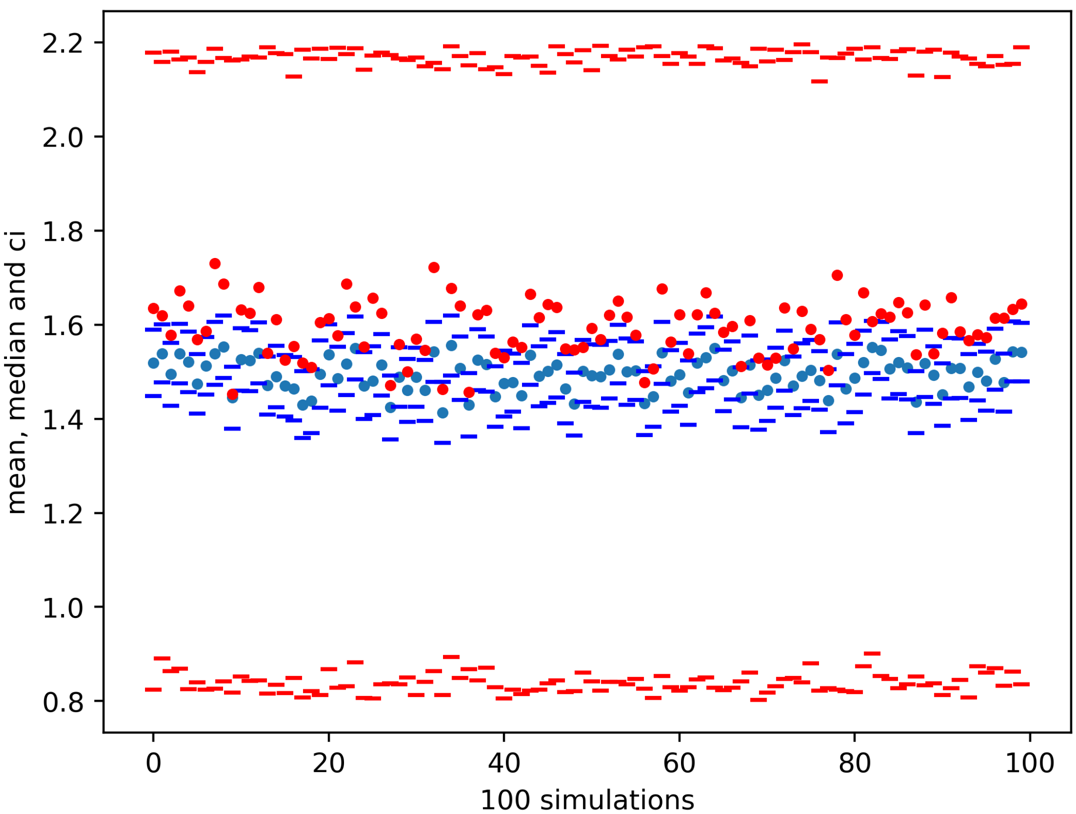

Figure 9: The blue dot corresponds to the mean, and its corresponding interval in a blue dash. On the other hand, the red dot corresponds to the median, and its corresponding interval in a red dash. For an event-based sampling and $N = 100$.

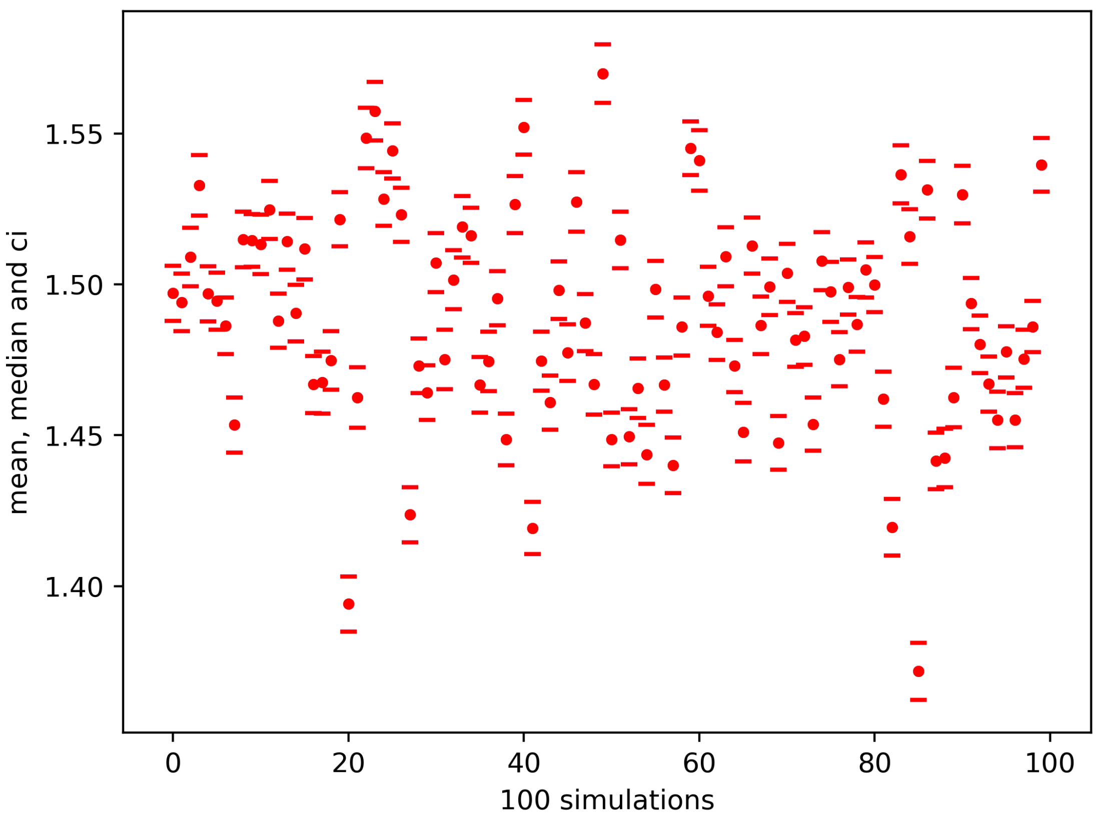

Figure 10: The blue dot corresponds to the mean, and its corresponding interval in a blue dash. On the other hand, the red dot corresponds to the median, and its corresponding interval in a red dash. For an time-based sampling and $N = 100$.

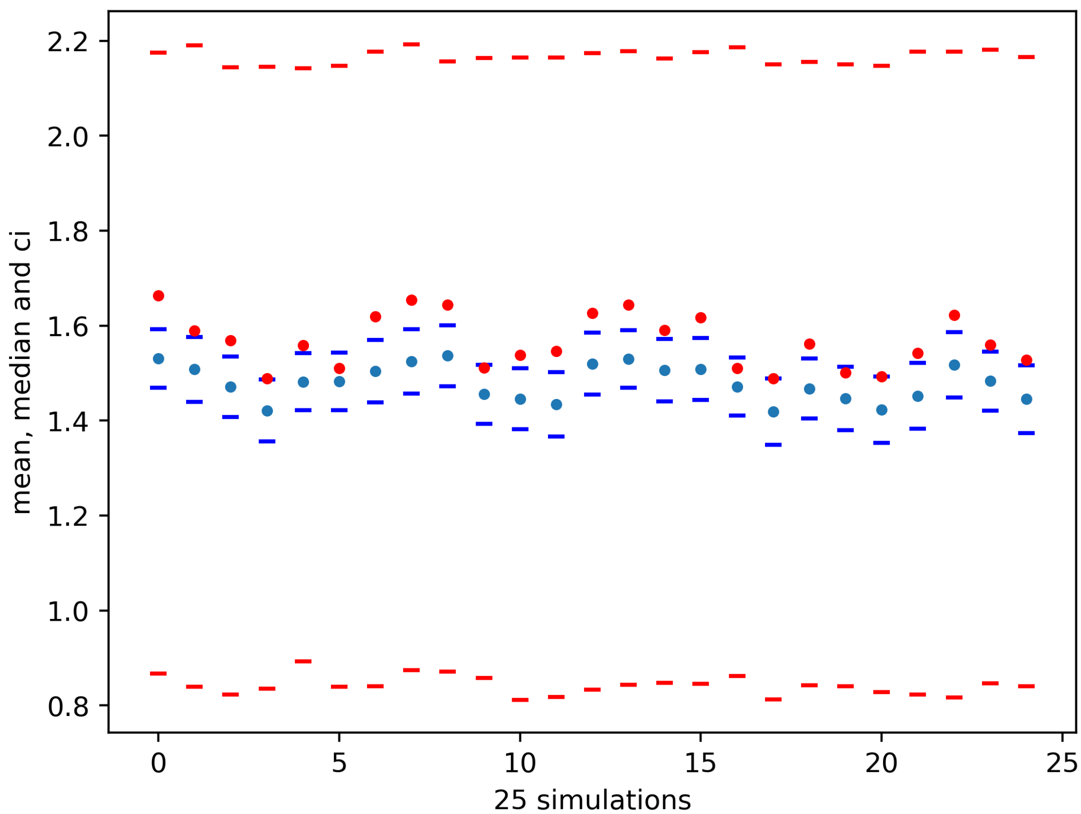

Figure 11: The blue dot corresponds to the mean, and its corresponding interval in a blue dash. On the other hand, the red dot corresponds to the median, and its corresponding interval in a red dash. For an event-based sampling and $N = 25$.

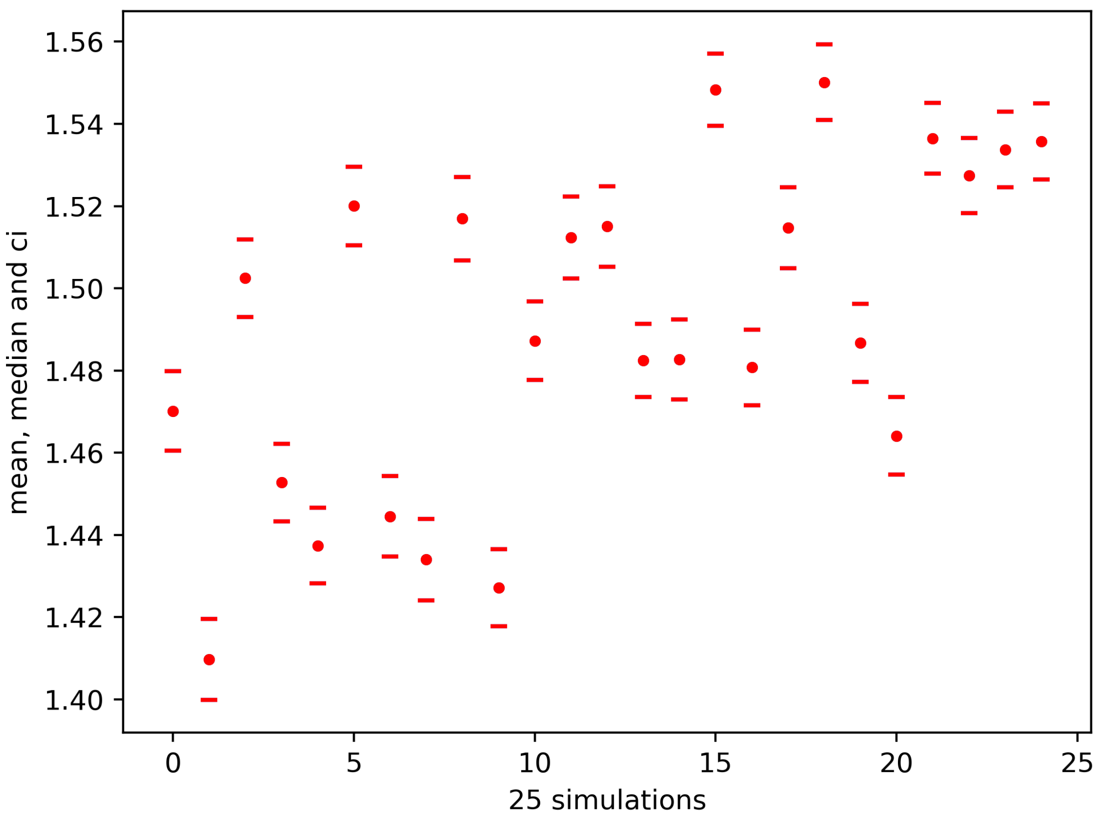

Figure 12: The blue dot corresponds to the mean, and its corresponding interval in a blue dash. On the other hand, the red dot corresponds to the median, and its corresponding interval in a red dash. For an time-based sampling and $N = 25$.