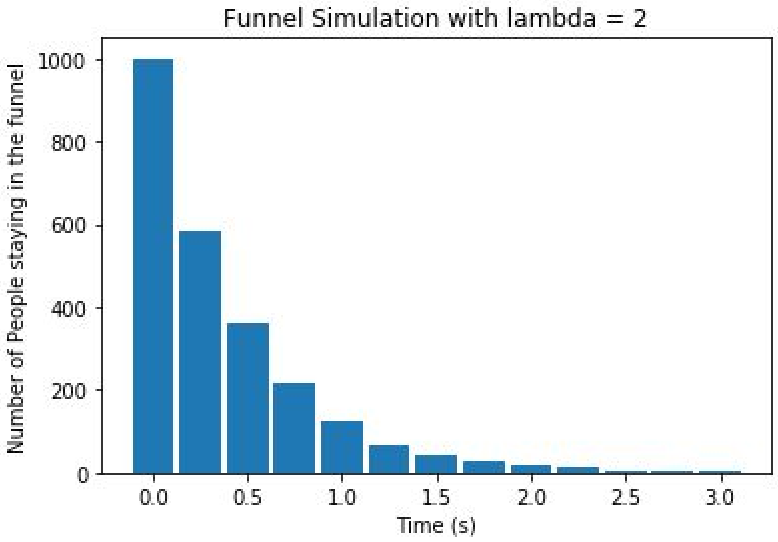
Funnel Assignment

Chenxi Ge, Dixin Yan, Yue Lan, Xiaowen (Sarah) Zhang

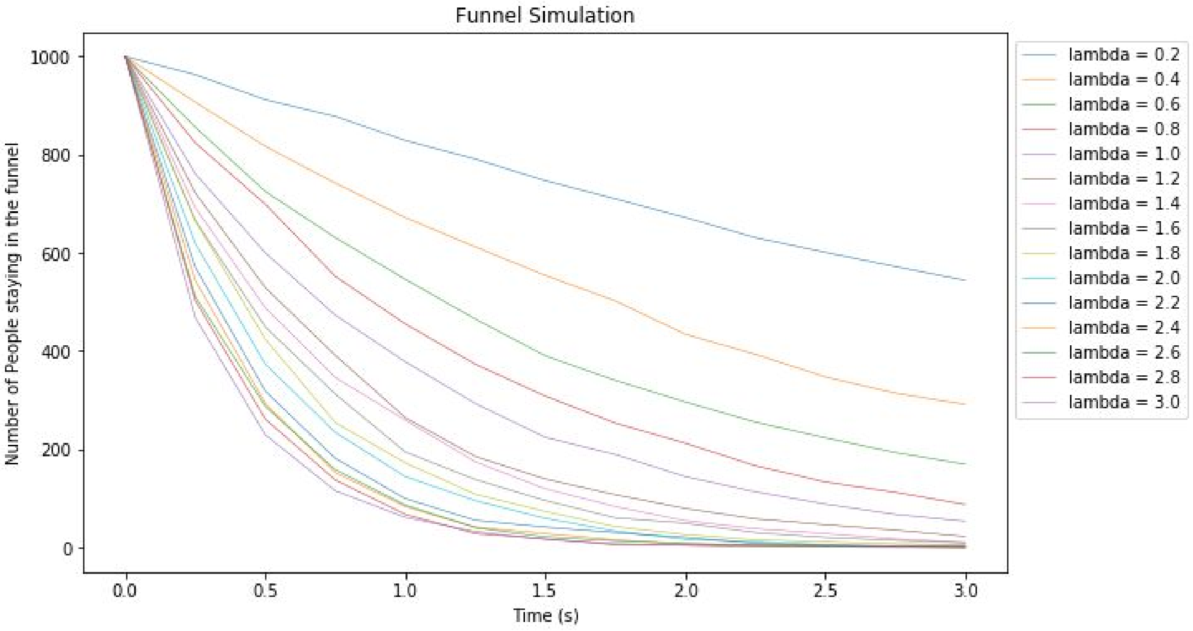
1.

(a) Create a funnel visualization with 1,000 users and a parameter of 2 with stops every 0.25 to 3.



The above visualization reflects a funnel simulation with 1000 users and λ = 2.

(b) Repeat the previous assignment, but with λ equal to the values of 0.2 to 3.0 in step sizes of 0.2, making sure to plot each.



The above visualization reflects a funnel simulation with 1000 users and λ ranging from 0.2 to 3.0 with step size 0.2.

(c) What is the relationship between λ and survival times?

We define the survival time as the value of each exponential random variable simulation, and it reflects the time a user spent in the funnel.
With larger λ the simulation will on average return a smaller survival time.

2.
(a) EstLam1 is an unbiased estimator. What does this mean?

It means that the expected value of the estimated λ is equal to the real λ.

(b) Generate a sample of 1,000 users using UserSim with λ equal to 1 and estimate λ using EstLam1.

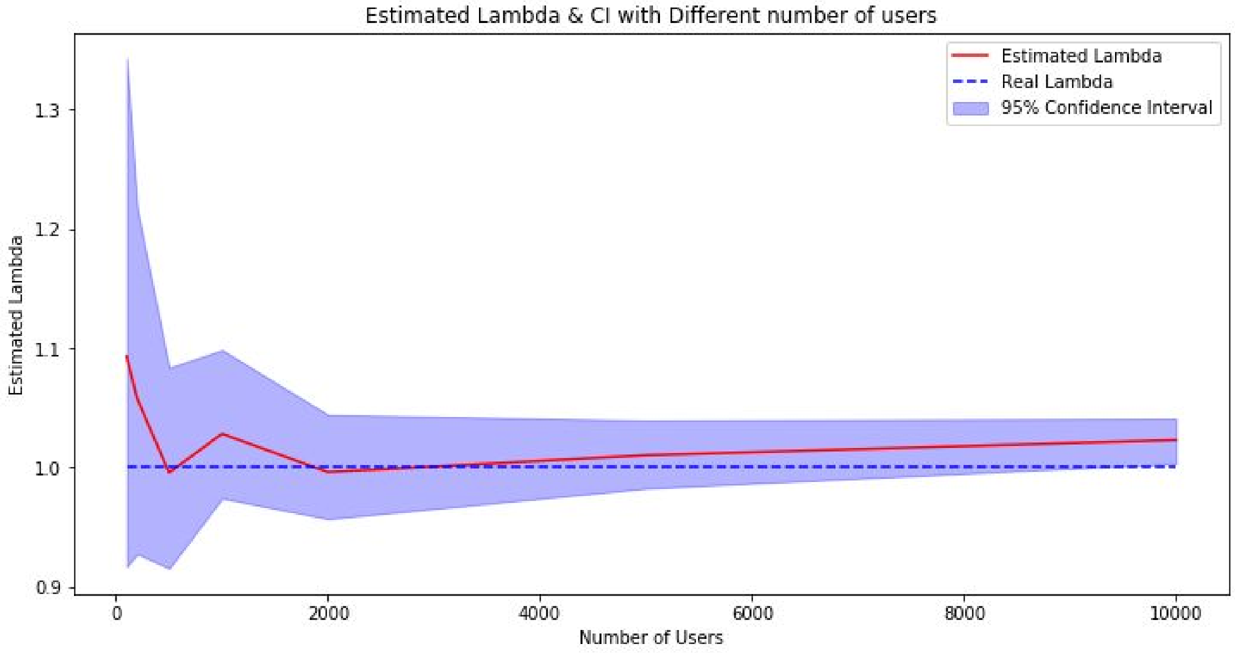
The estimated λ is: 1.02827.

(c) Using that same sample of 1,000 users, bootstrap a 95% confidence interval for the estimate (use 500 bootstraps).

The 95% confidence interval of lambda is (0.97412, 1.09652)

(d) Repeat the above process with number of users equal to 100, 200, 500, 1,000, 2,000, 5,000 and 10,000. Create both a table and a visualization which includes both the estimated λ and the confidence intervals. How do the results change as the number of users increase?

Number of Users	Estimated Lambda	Lower CI	Upper CI
100	1.09320	0.91702	1.34298
200	1.05766	0.92769	1.21666
500	0.99571	0.91572	1.08390
1000	1.02827	0.97434	1.09859
2000	0.99620	0.95711	1.04432
5000	1.01035	0.98265	1.03975
10000	1.02302	1.00335	1.04110



The above table and visualization reflect estimated λ. We can see that with the increase of number of users, the 95% confidence interval gets narrower.

3.

Equation 2.1 is for estimating λ using Maximum Likelihood Estimation ("MLE")

$$\lambda = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}} \tag{2.1}$$

Equation 2.2 is the likelihood function for event data.

$$l = \left\{ \prod_{i=1}^{m_0} F(BP_1|\lambda) \right\} \cdot \left\{ \prod_{i=m_0+1}^{m_0+m_1} F(BP_{U_i+1}|\lambda) - F(BP_{U_i}|\lambda) \right\} \cdot \left\{ \prod_{i=m_0+m_1+1}^{n} 1 - F(BP_t|\lambda) \right\} \tag{2.2}$$

(a) Carefully explain why equation 2.1 will not work using event data.

The equation 2.1 does not work with event data, since we do not know the exact time a user quits. With event data, we only know if a user has made it past a certain hurdle, thus regardless the actual time spent, users that made it past BP_{ui} but not BP_{ui+1} will be treated the same. While equation 2.1 is using the expected time spent in the funnel, it is not suitable for event data.

(b) Take the log of equation 2.2 and, using the distribution function, simplify it by turning it into sums and removing unnecessary expressions.

$$L = log(l) = \sum_{i=1}^{m_0} log(F(BP_1|\lambda)) + \sum_{i=m_0+1}^{m_0+m_1} log(F(BP_{u_i}|\lambda) - F(BP_{u_i+1}|\lambda)) + \sum_{i=m_0+m_1+1}^{n} log(1 - F(BP_t|\lambda))$$

$$= m_0 * log(1 - e^{-\lambda BP_1}) + \sum_{i=m_0+1}^{m_0+m_1} log(e^{-\lambda BP_{u_i}} - e^{-\lambda BP_{u_i+1}}) + (m_0 + m_1 + 1)(-\lambda BP_t)$$

4.

(a) Using the functions defined above, run 1,000 simulations of 100 users with the following sets of break points: [0.25, 0.75], [0.25, 3], [0.25, 10]. For each simulation calculate the difference between the estimated λ using EstLam1 and EstLam2. What is the average difference? How does moving the second breakpoint affect the estimate?

Breaks	Average Difference
[0.25, 0.75]	0.09417
[0.25, 3]	0.17398
[0.25, 10]	0.20166

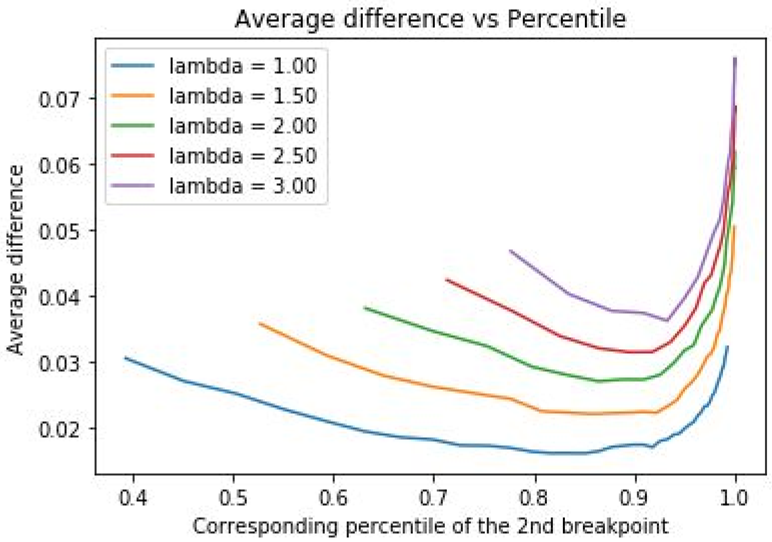
From the simulations above, we conclude that in the scenario where we only move the second breakpoint, we will get a smaller difference if the second breakpoint is closer to the real λ.

(b) Using what you learned from doing the above, how should you design breakpoints? What trends can you identify from the above data. Feel free to run additional simulations to identify other trends.

We tried other designs of break points and calculated the average difference. Specifically, we fixed the first breakpoint, and only explored what would happen if we move the second breakpoint.

Our approach is that we set the first breakpoint as 0.25, and tested the second breakpoint from 0.5 to 5 with step size 0.1. We simulated 100 different groups with 1000 users in each group, and calculated the mean absolute difference across the group. We repeated the whole process using λ from 1 to 3 with step size 0.5.

We visualized our experiment results with the following graph. Specifically, the y-axis is the absolute difference, and the x-axis is the percentile of each second breakpoint. For different λ, the same value of the second breakpoint does not represent the same percentile, thus we normalized the axis for each distribution using percentile. We found that for different λ, the 90-percentile point is almost always the point that will result in the smallest difference.



From what we can observe, to get a smaller difference with the actual λ, the best practice is to pick the 90-percentile of the observed value, then use that value as the second breakpoint to estimate the λ.