Length Bias Estimation of Small Business Lifetime

Honor Thesis Presentation

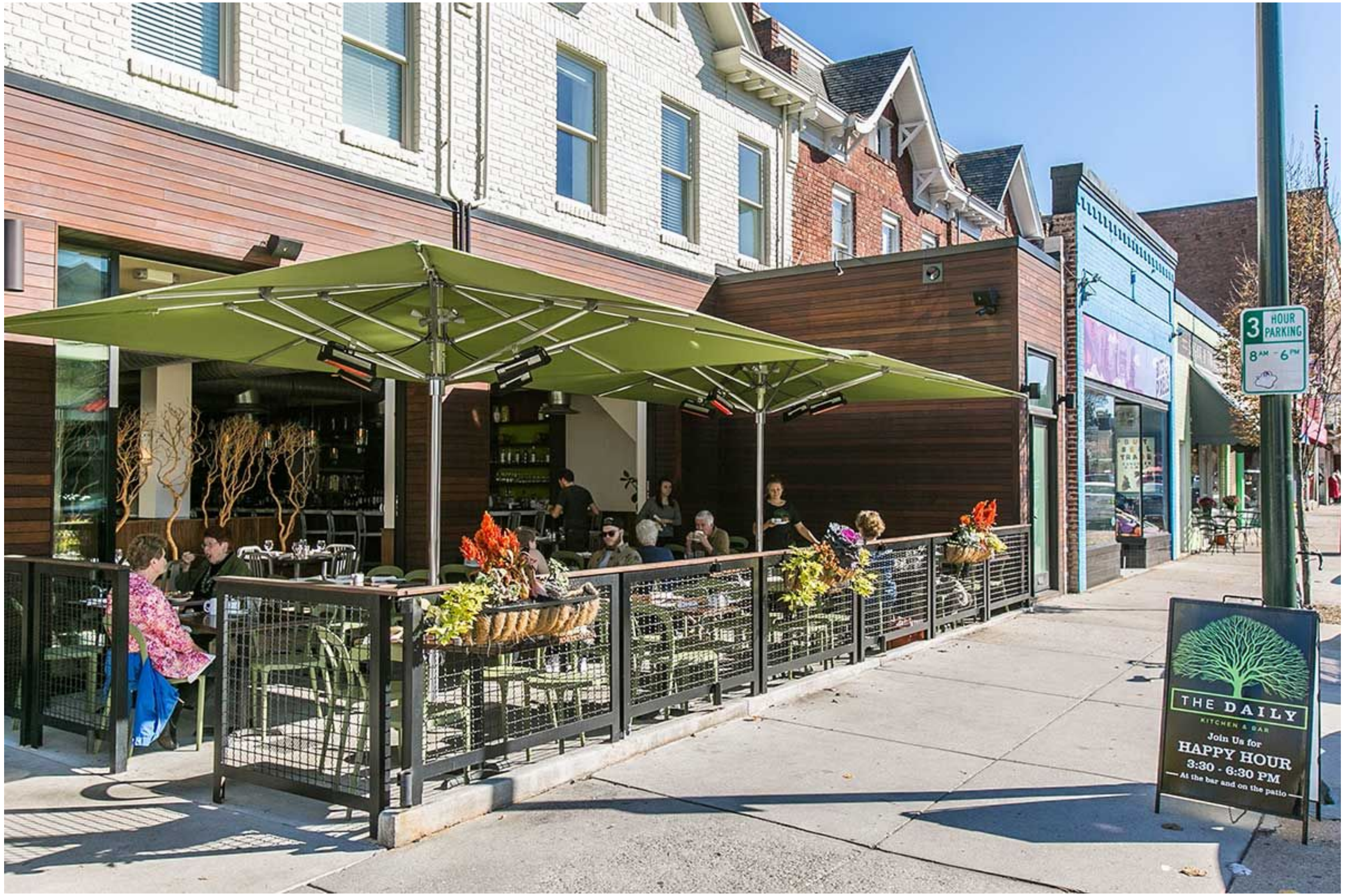
Simeng Li
Primary Advisor: Dr. Paul Kvam
Spring 2023

- **What is the goal of this study?**
- **Why are we interested in small businesses like restaurants?**
- **What is length bias?**
- **What is right censoring?**
- **What problems will length bias / right censoring cause for researchers?**
- **Regarding length bias, how can we better estimate our data?**

Why We Use Restaurants Lifetimes?

- Job Creation
- Tourism
- Find Potential Problems
- Help with Tourism & City Planning

Carytown, RVA



The Dataset (In Operation & Already Closed 12/2022)

Pho Luca’s RVA	**Baker’s Crust Carytown**	**Carytown Burger and Fries**	**The Daily Kitchen and Bar**	**Citizen Burger Bar**
Mom’s Siam Restaurant	**New York Deli**	**Ginger Thai Taste**	**Galaxy Diner**	**Les Crepe**
Home Sweet Home	**Mellow Mushroom Carytown**	**The Mantu**	**Carytown Sushi**	**Don’t Look Back**
Stella’s	**CanCan**	**Tulsi**	**East Coast Provision**	

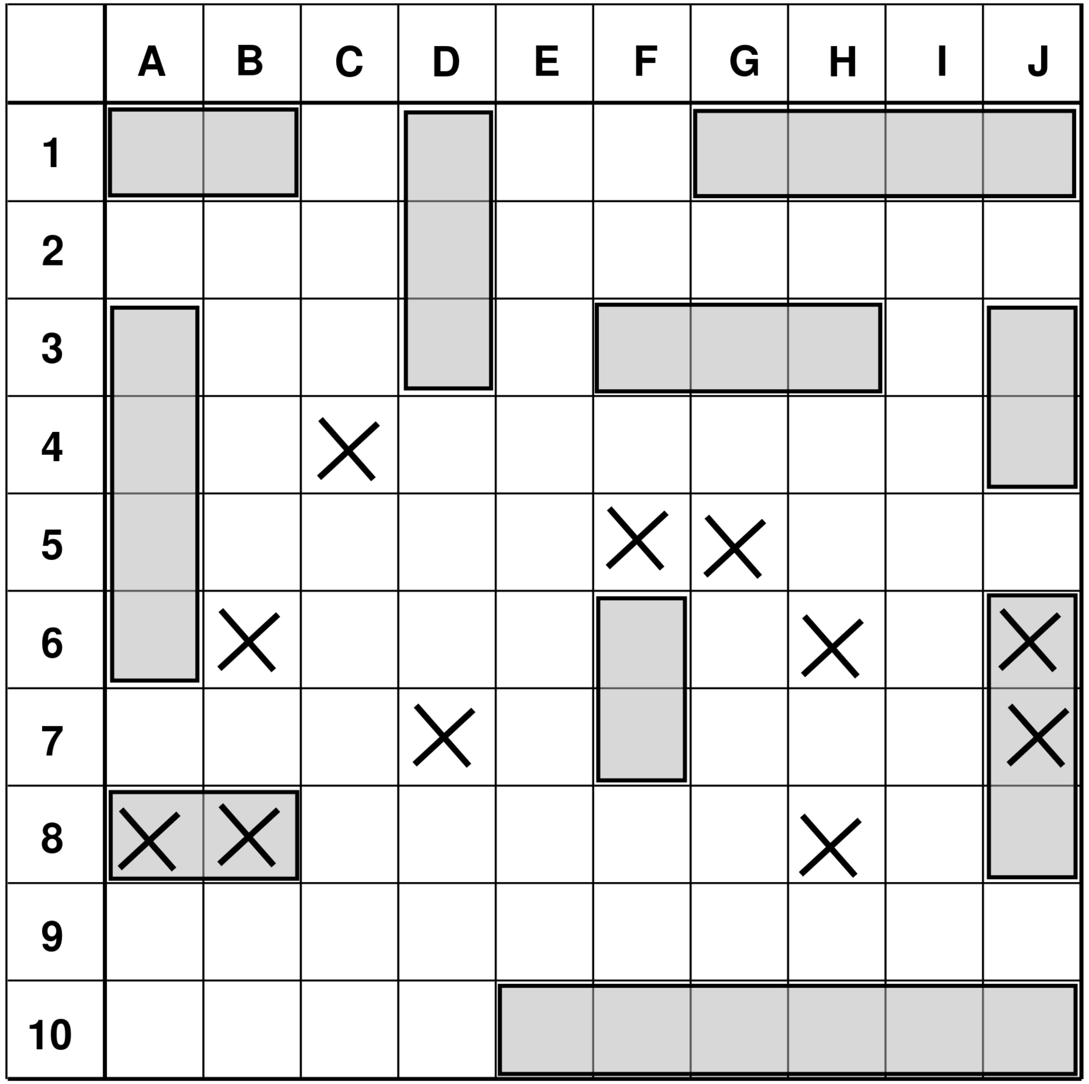
Weezie’s Kitchen	**Portrait House**	**Mezzanine**	**Broken Tulip**	**Xtra’s Café**

Why don’t we just collect restaurants lifetimes and calculate their mean?

Length Bias!

Length bias occurs when the probability of detecting a case of a failure depends on the duration of the event. In other words, a longer duration of a specific event is more likely to be detected than a shorter duration. This can lead to a bias in the estimated survival or failure rates.

Battleship Game

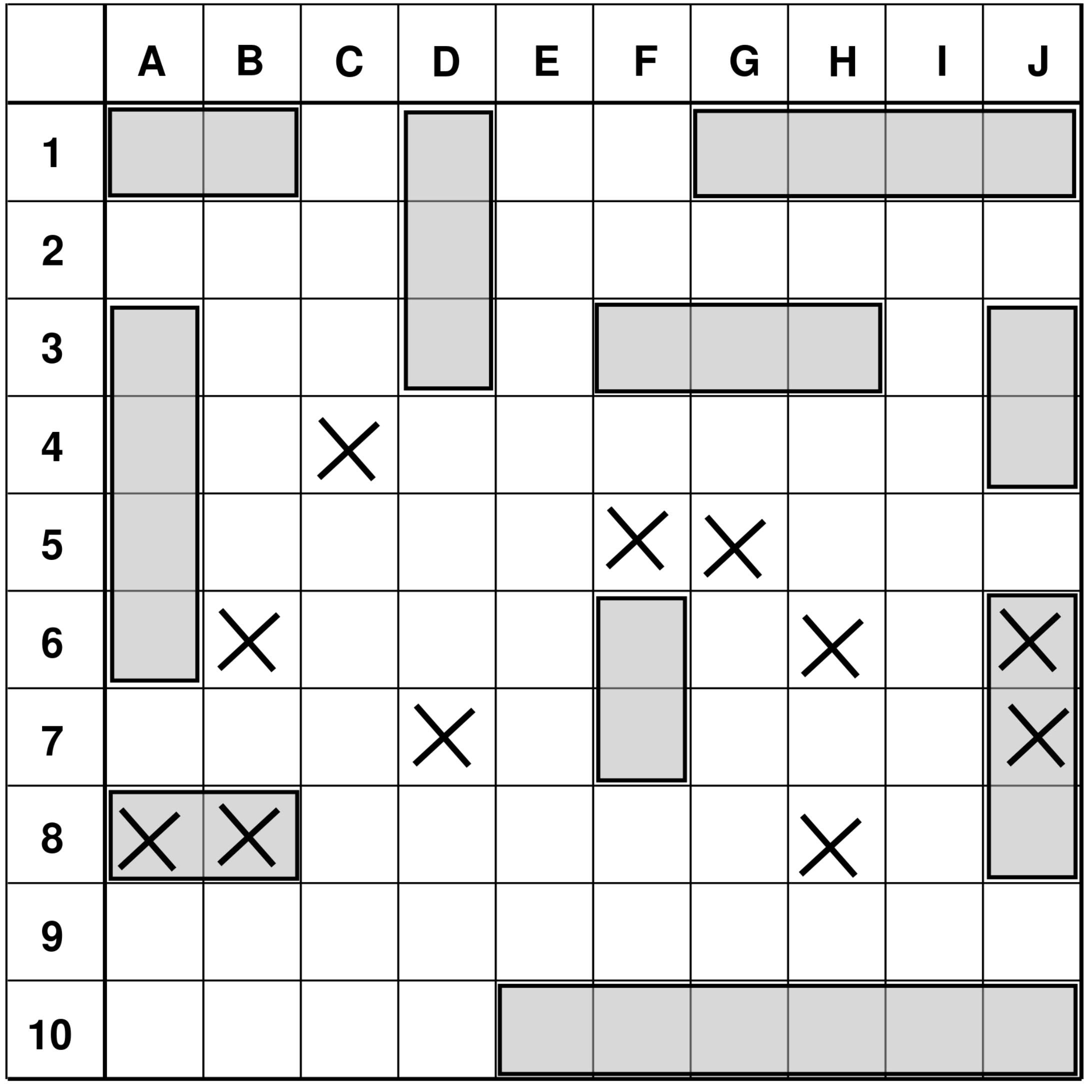


	A	B	C	D	E	F	G	H	I	J
1	▨	▨		▨			▨	▨	▨	▨
2				▨						
3	▨			▨		▨	▨	▨		▨
4	▨		X							▨
5	▨					X	X			
6	▨	X				▨		X		☒
7				X		▨				☒
8	☒	☒						X		▨
9										
10					▨	▨	▨	▨	▨	▨

- Each player must secretly place their ships on a grid of ten columns by ten rows. These represent the location of the ships on a battlefield.

- Once it is determined who will go first, that person will pick a square at random, calling it by its reference of column reference, row number (C3 for example). This represents their firing a missile directly at that square.

Battleship Game



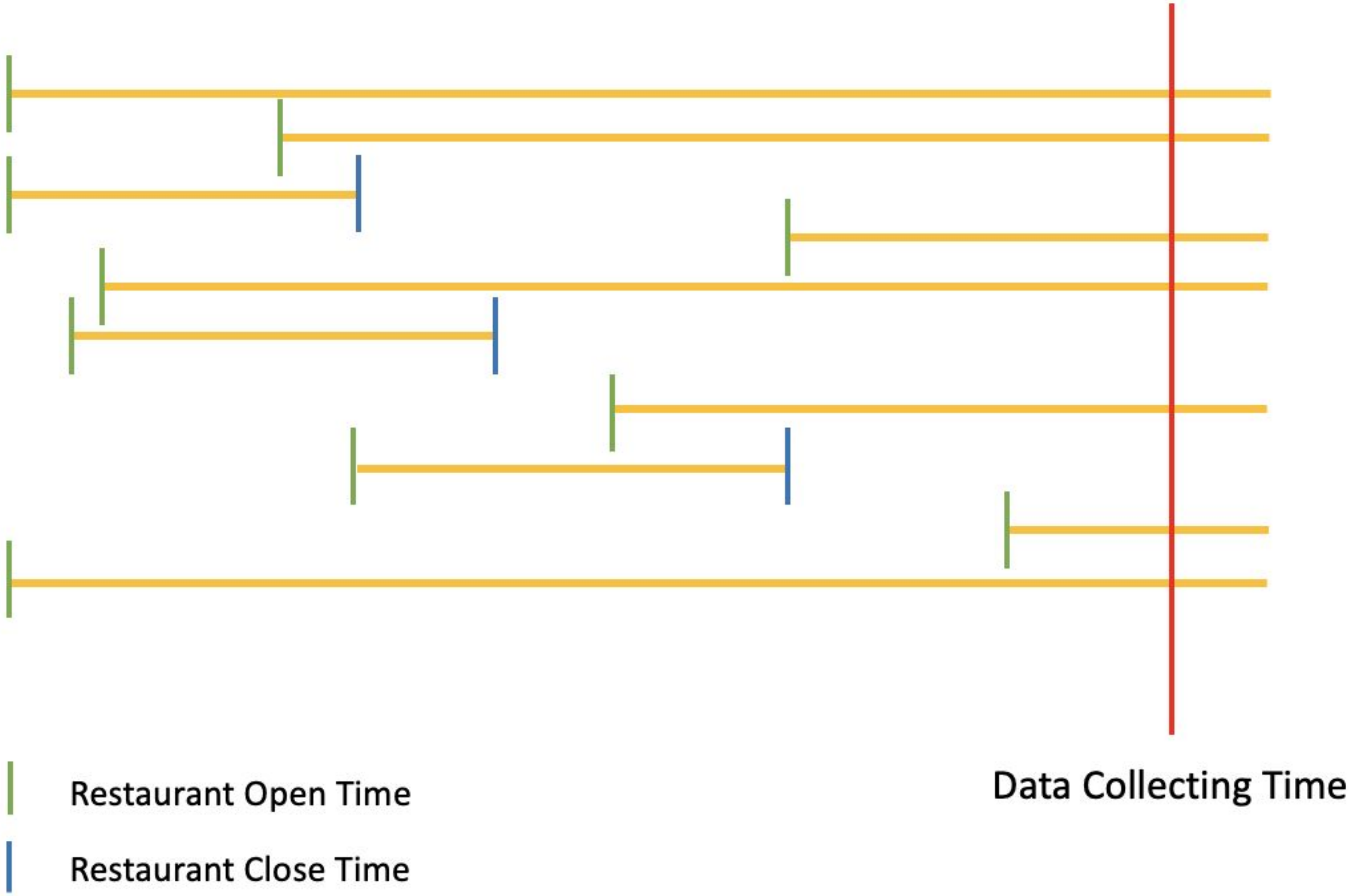
	A	B	C	D	E	F	G	H	I	J
1	▨	▨		▨			▨	▨	▨	▨
2				▨						
3	▨			▨		▨	▨	▨		▨
4	▨		✕							▨
5	▨					✕	✕			
6	▨	✕				▨		✕		☒
7				✕		▨				☒
8	☒	☒						✕		▨
9										
10					▨	▨	▨	▨	▨	▨

- It is always more likely to hit larger ships first (greater length, higher probability).

- If there are equal number of large ships and smaller ships, it is more likely to see larger ships first.

- If the rival doesn't know there are equal numbers of ships, it is reasonable for them to say: "You have more larger ships than smaller ones!"

In Terms of Restaurants…



Restaurant Open Time

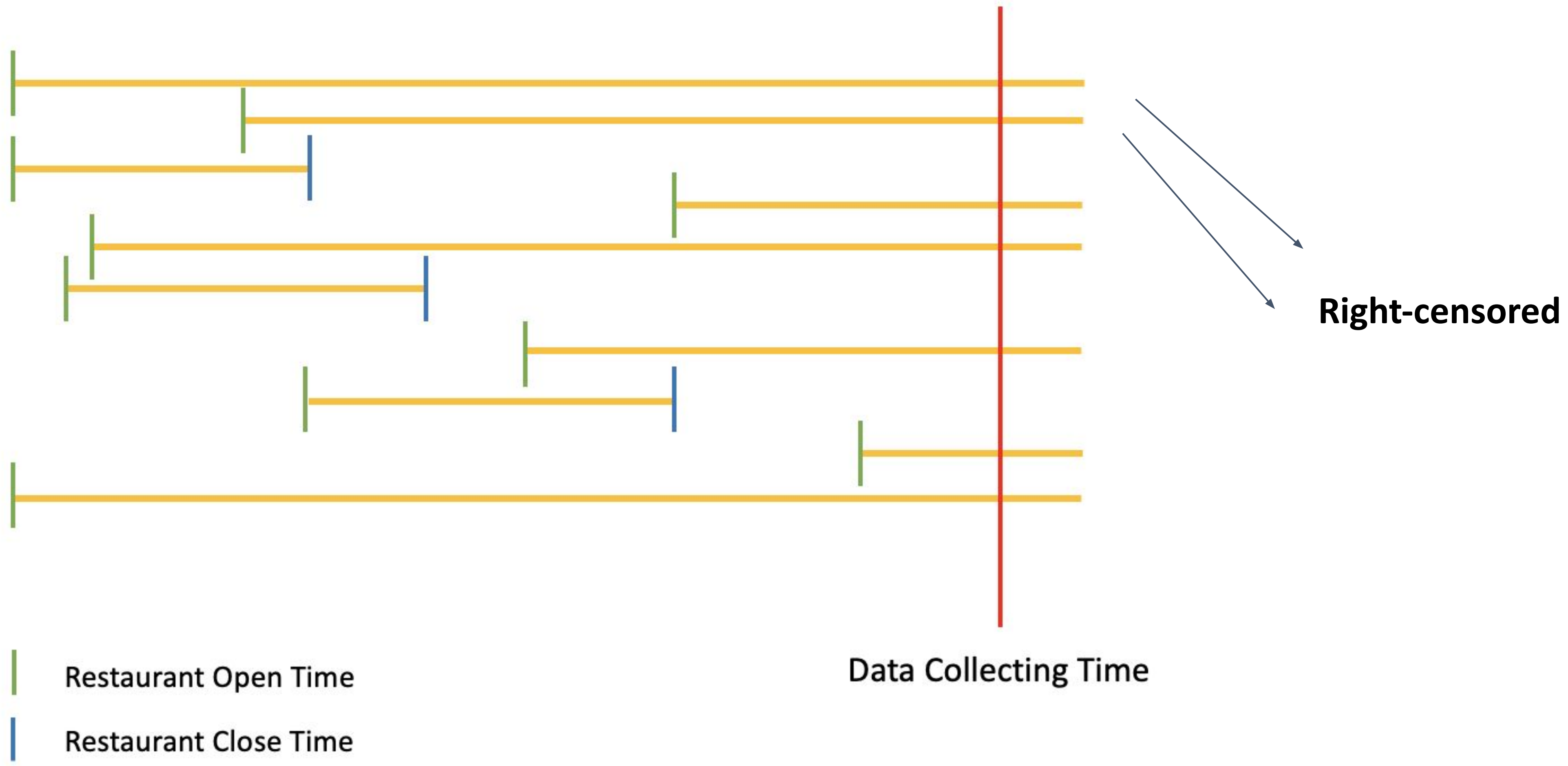
Restaurant Close Time

Data Collecting Time

Right-Censoring

- The data we have are collected at a specific time t and at this point, the event we are looking for (the closure of restaurant) has not yet occurred.

In Terms of Restaurants…



Right-censored

Data Collecting Time

Restaurant Open Time

Restaurant Close Time

Right-Censoring

- The data we have are collected at a specific time t and at this point, the event we are looking for (the closure of restaurant) has not yet occurred.
- If we ignore the censoring problem, it will lead to an underestimation of the lifespan data, and may result in potential useful data being misused or thrown away.
- In this study, most restaurant data are right-censored.

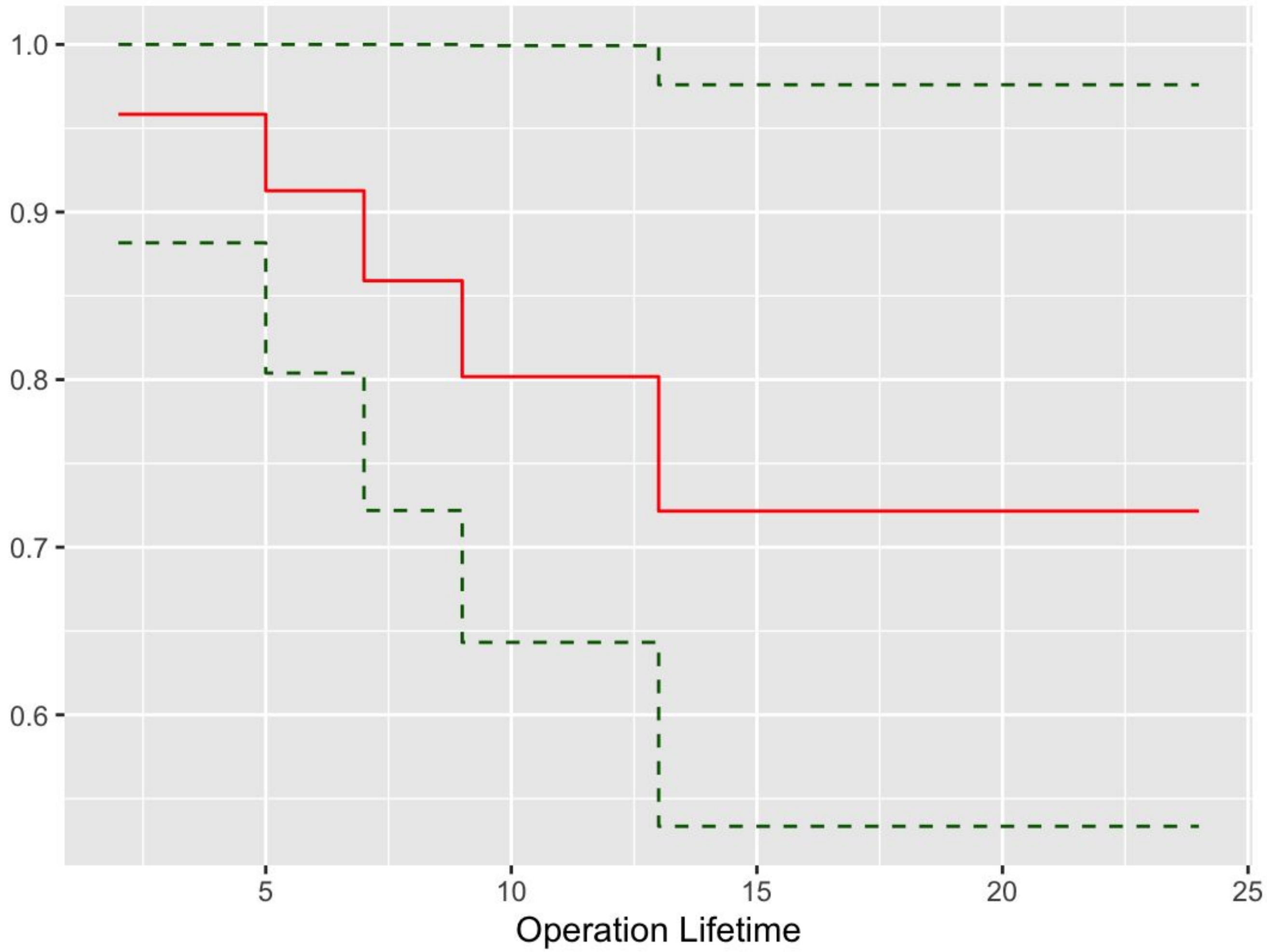
Kaplan-Meier Estimator

- Kaplan-Meier **nonparametric** estimator is used for **censored** lifetime data analysis.
- Using the Kaplan-Meier method, we can take care of the right-censoring and get a more accurate estimate of the average lifetime (rather than just use the mean of our data which is an underestimation).

$$F_{KM}(t) = 1 - \prod_{x_j \leq t} (1 - \frac{d_j}{m_j})$$

where d_j = number of failures at x_j, m_j = number of observations that had survived up to x_j.

Kaplan-Meier Estimator



- Specify whether each restaurant has failed or not (censored or not)
- We can calculate the area under the survival curve to estimate the average lifetime.
- 17.56 vs 11.21 (underestimation)

Now It’s Time to Deal with the Length Bias!

Exponential Fit

- We want to know the distribution of restaurants lifetime data so that we can know its behavior (“real” distribution without length bias)
- Most other distributions have increasing failure rates (not the case for restaurants lifetimes)
- “Memoryless” property: Probability of an event occurring is independent of how long it has been since the previous event occurred. Thus, our goodness-of-fit test for the exponential distribution are not negatively affected by right-censoring.

Why We Can’t Observe Exponential

$$g(x) = \frac{xf(x)}{\mu}$$

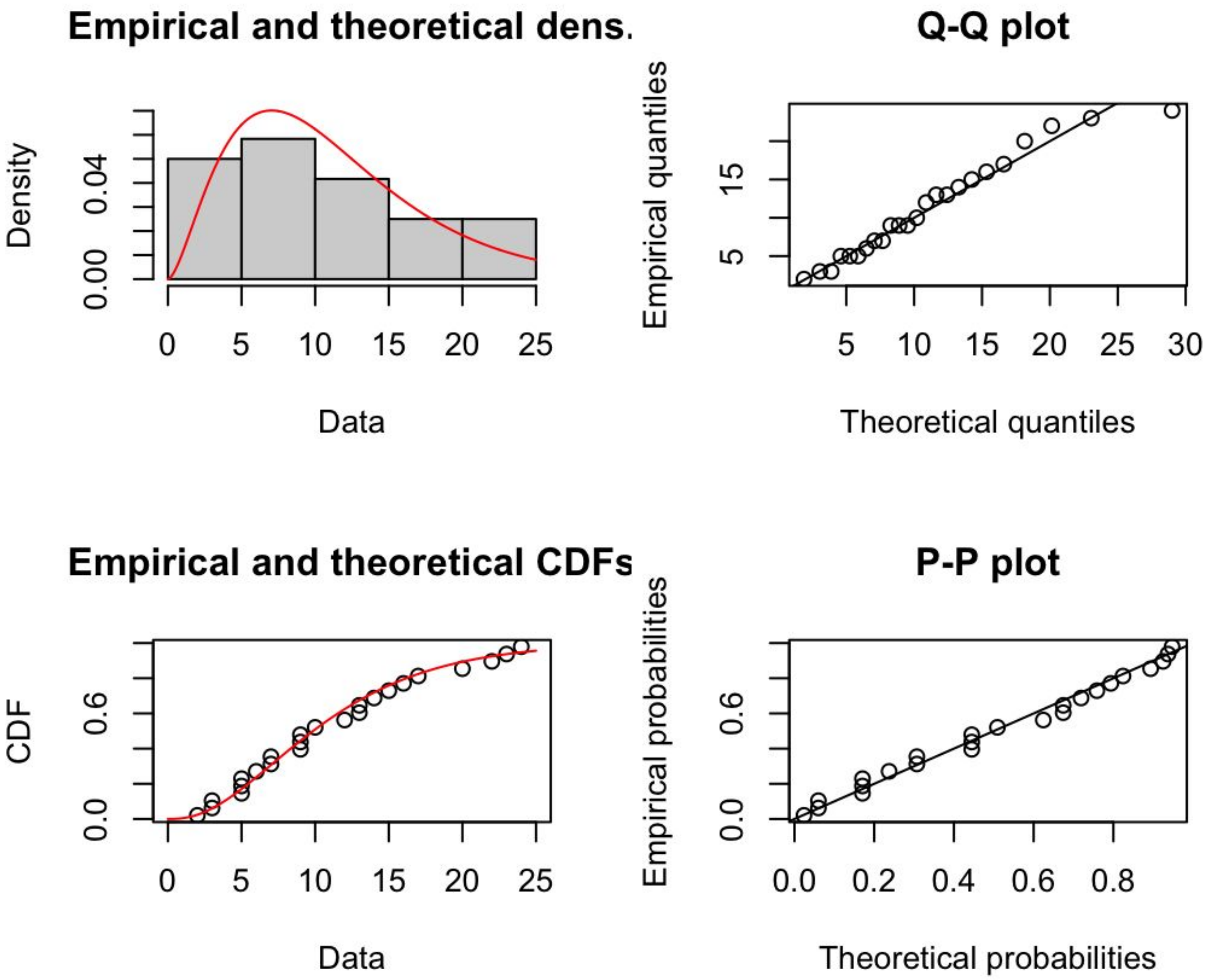
(Asgharian and Wolfson, 2005)

$$f(x) = \lambda e^{-\lambda x} \longrightarrow g(x) = \frac{x\lambda e^{-\lambda x}}{1/\mu} = x\lambda^2 e^{-2\lambda} \longrightarrow \Gamma(2, 1/\boldsymbol{\lambda})$$

- For f(x), E(x)=1/$\boldsymbol{\lambda}$
- For g(x), E(x) = $\alpha\boldsymbol{\beta}$ = 2(1/$\boldsymbol{\lambda}$)
- Overestimate f(x) by 100% if not considering length bias

Exponential Fit

- For Gamma distribution we know E(x) = $\alpha\beta$ = 2β, and from KM we get an estimate of 17.56, thus we know β = 17.56/2 = 8.78.
- So, we want to fit restaurants data with $\Gamma(2, 8.78)$
- Cramer-von Mises Goodness-of-Fit Test (CvM Test) -> p-value=0.97



Empirical and theoretical dens.

Q-Q plot

Empirical and theoretical CDFs

P-P plot

Exponential Fit

- For Gamma distribution we know E(x) = $\alpha\beta$ = 2β, and from KM we get an estimate of 17.56, thus we know β = 17.56/2.
- So, we want to fit restaurants data with Γ(2, 8.78)
- Cramer-von Mises Goodness-of-Fit Test (CvM Test) -> p-value=0.97
- Once we know our data fit exponential distribution, we know even if we observe a gamma distribution, the data are actually distributed exponentially (“real” distribution without length bias)

The End