

HAZARD PROBABILITIES & CENSORING

Dr. Aric LaBarr

Institute for Advanced Analytics

MSA Class of 2014

CENSORING

Censoring

- A common occurrence in survival analysis is censored data.
- Censored data can occur outside of survival analysis, but not all statistical models can handle this.
- Example: Customer Loyalty Program
 - Want to look at all of our customers between January 2011 and December 2011.
 - How long before customers cancel their subscription?
 - What if the customer has not yet cancelled their subscription by the end of last year?

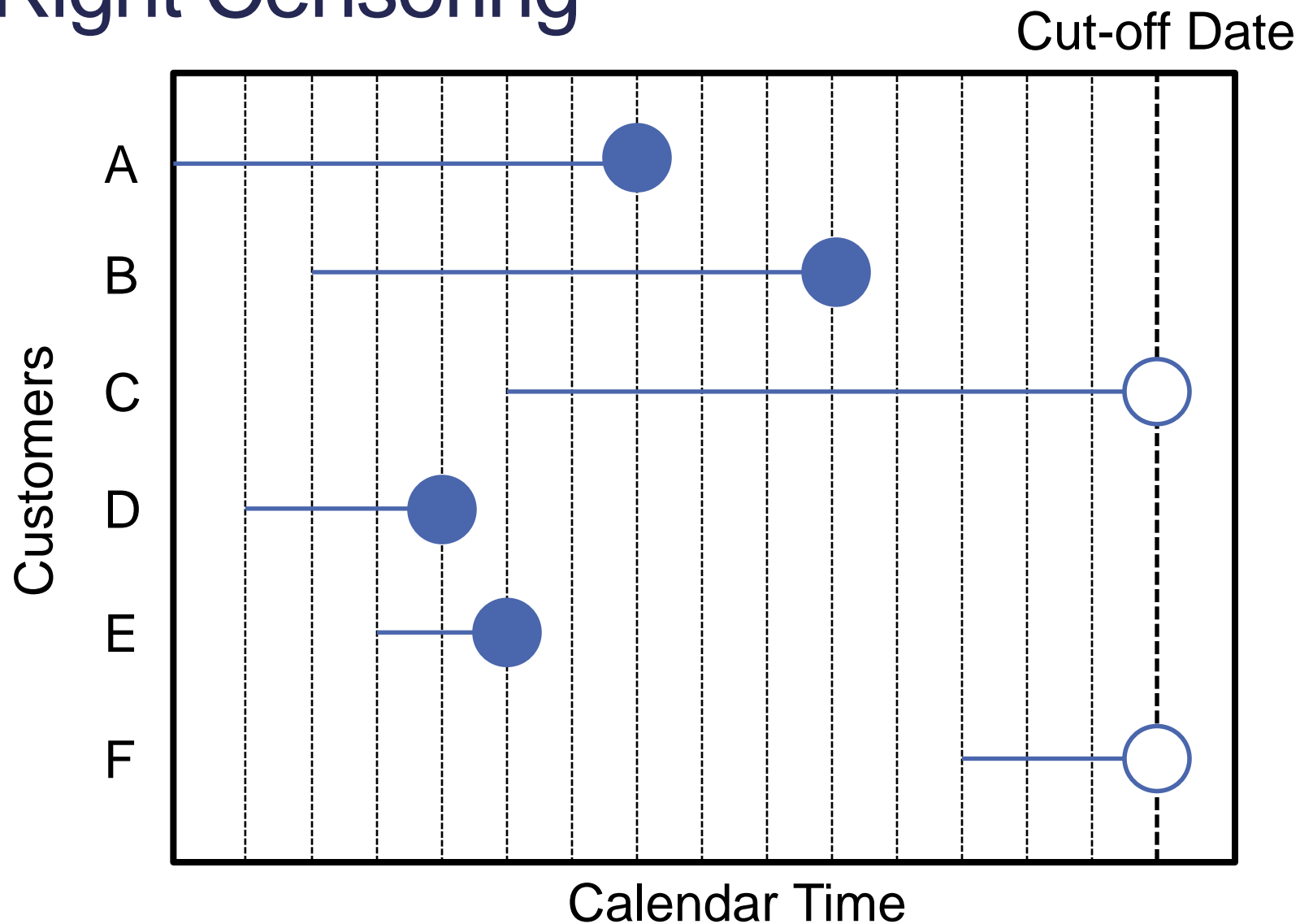
Censoring

- A common occurrence in survival analysis is censored data.
- Censored data can occur outside of survival analysis, but not all statistical models can handle this.
- Example: Customer Loyalty Program
- Example: Prison Recidivism Rates
 - Want to follow recently released inmates for one year.
 - How long before they are arrested again?
 - What if they are not arrested again within the 52 weeks of the study?

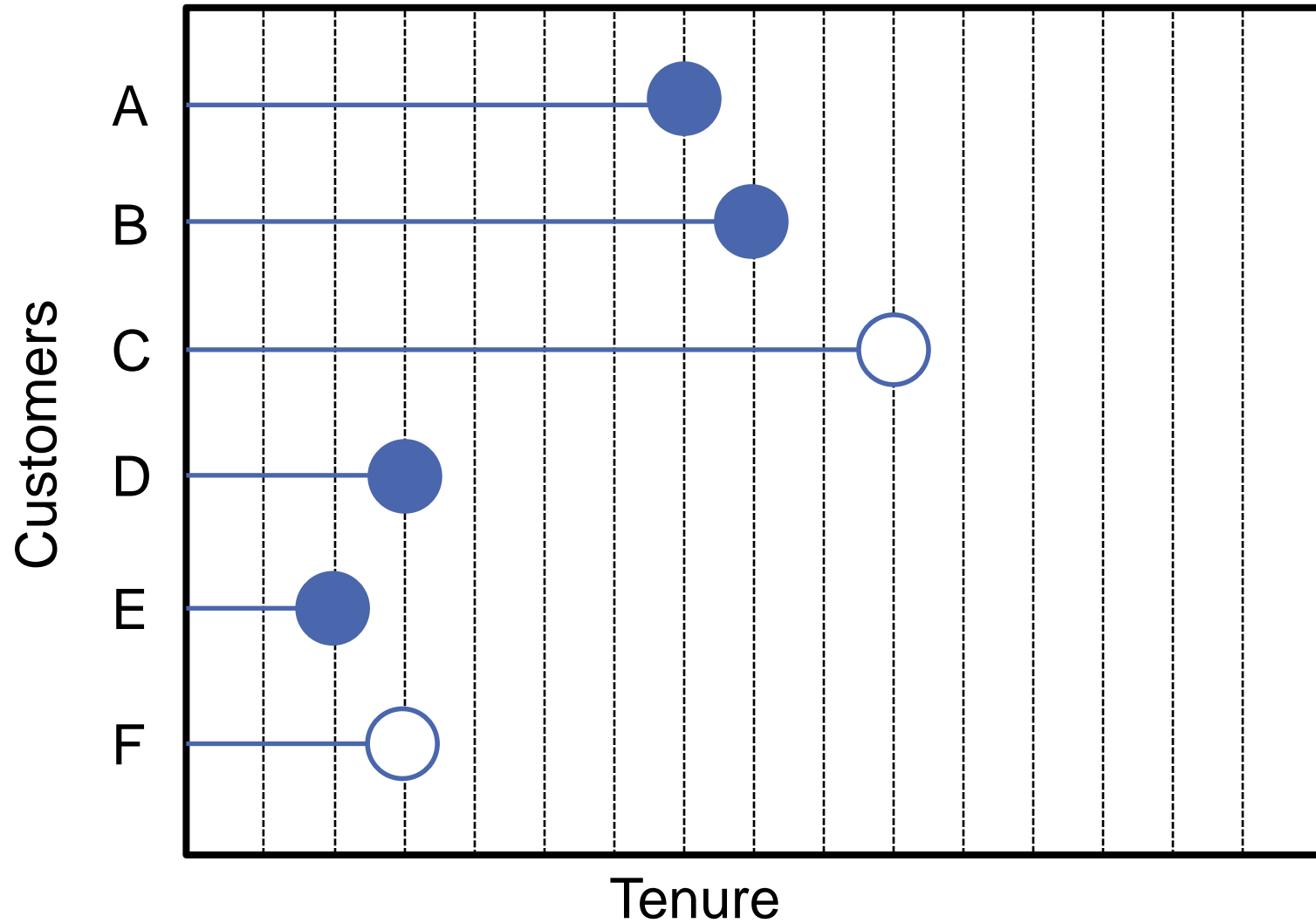
Right Censoring

- T is the time of occurrence of the desired event.
- If an observation is right censored, then all we know is that $T > c$.
- Examples:
 - Death occurs sometime past 50 years old.
 - Person arrested after 52 weeks.
 - Income greater than \$75,000.
 - Customer has monthly subscription for at least 3 months.

Right Censoring



Right Censoring



Survival Curves – Censoring

- The calculation of survival curves changes slightly in the presence of censoring.
- We do not want to exclude observations just because they are censored.

Kaplan-Meier Method

- The Kaplan-Meier method existed long before Kaplan and Meier.
- Kaplan and Meier showed it was the maximum likelihood estimate for the nonparametric estimation of the survival curve.
- 3 situations for calculating survival curve:
 1. No censoring (very intuitive)
 2. Right Censored (somewhat intuitive)
 3. Complicated Censoring (not very intuitive)

Kaplan-Meier Method

- The Kaplan-Meier method existed long before Kaplan and Meier.
- Kaplan and Meier showed it was the maximum likelihood estimate for the nonparametric estimation of the survival curve.
- 3 situations for calculating survival curve:
 1. No censoring (very intuitive)
 2. Right Censored (somewhat intuitive)

$$\hat{S}(t) = \prod_{t=1}^k \left(1 - \frac{d_t}{n_{t-1}} \right)$$

3. Complicated Censoring (not very intuitive)

Kaplan-Meier Method

- 3 situations for calculating survival curve:
 1. No censoring (very intuitive)
 2. Right Censored (somewhat intuitive)

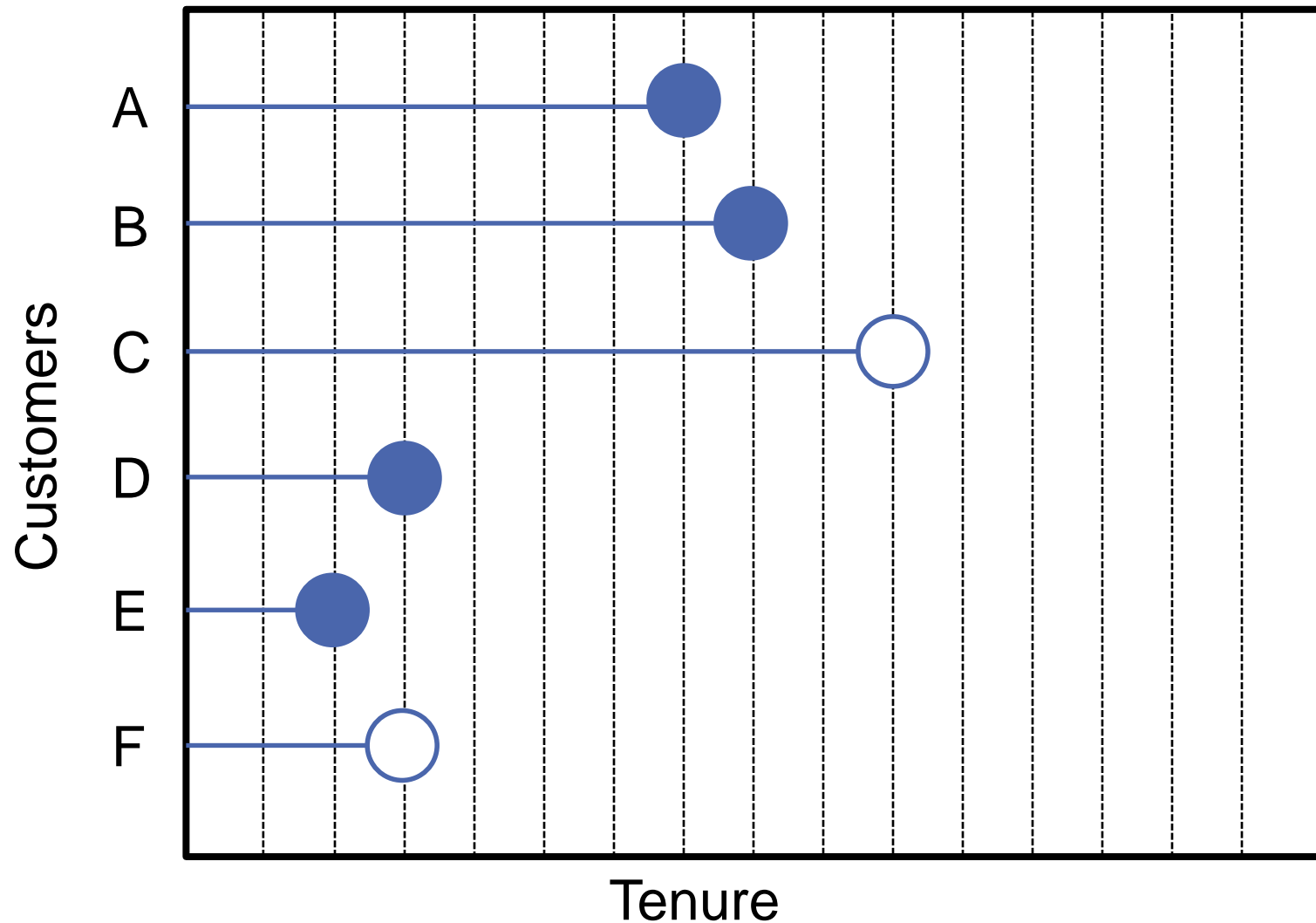
$$\hat{S}(t) = \prod_{t=1}^k \left(1 - \frac{d_t}{n_{t-1}} \right)$$

Customers leaving at t

Customers remaining at $t - 1$ (Customer at risk at t)

3. Complicated Censoring (not very intuitive)

Survival Curve – Censoring



Survival Curve – Censoring

- Time = 0:

$$\hat{S}(0) = 1$$

- Time = 1:

$$\hat{S}(1) = \hat{S}(0) \times \left(1 - \frac{0}{6}\right) = 1$$

- Time = 2:

$$\hat{S}(2) = \hat{S}(1) \times \left(1 - \frac{1}{6}\right) = 0.8333$$

Survival Curve – Censoring

- Time = 3:

$$\hat{S}(3) = \hat{S}(2) \times \left(1 - \frac{1}{5}\right) = 0.833 \times 0.80 = 0.667$$

- Time = 4:

$$\hat{S}(4) = \hat{S}(3) \times \left(1 - \frac{0}{3}\right) = 0.667$$

- Time = 5:

$$\hat{S}(5) = \hat{S}(4) \times \left(1 - \frac{0}{3}\right) = 0.667$$

Survival Curve – Censoring

- Time = 6:

$$\hat{S}(6) = \hat{S}(5) \times \left(1 - \frac{0}{3}\right) = 0.667$$

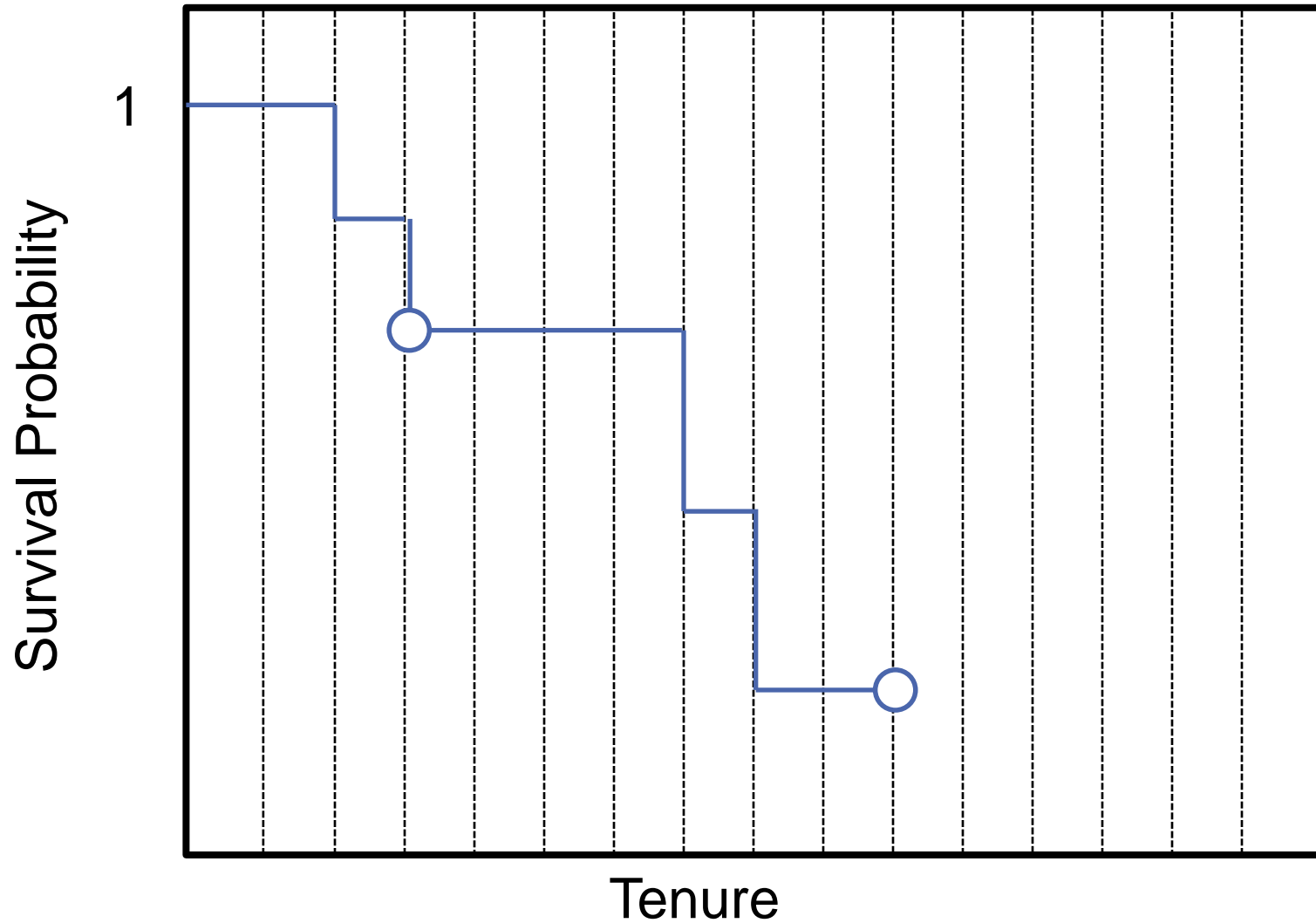
- Time = 7:

$$\hat{S}(7) = \hat{S}(6) \times \left(1 - \frac{1}{3}\right) = 0.667 \times 0.667 = 0.444$$

- Time = 8:

$$\hat{S}(8) = \hat{S}(7) \times \left(1 - \frac{1}{2}\right) = 0.444 \times 0.5 = 0.222$$

Survival Curve – Censoring



Survival Curve - Censoring

```
data Simple;  
    input Customer Tenure censored;  
datalines;  
A 7 0  
B 8 0  
C 10 1  
D 3 0  
E 2 0  
F 3 1  
;  
  
proc lifetest data=Simple;  
    time Tenure*censored(1);  
run;
```

Other Forms of Right Censoring

- The most common form of right censoring is where there is a specific cut-off date (Type I censoring).
- Other forms include the following:
 - Censor after a certain number of events occur (Type II censoring).
 - Censoring occurs due to other reason outside of investigator controls (random censoring).

Other Forms of Right Censoring

- The most common form of right censoring is where there is a specific cut-off date (Type I censoring).
- Other forms include the following:
 - Censor after a certain number of events occur (Type II censoring).
 - Censoring occurs due to other reason outside of investigator controls (random censoring).

POTENTIAL PROBLEM!!!



Type I, Type II, Random Censoring

- Standard survival analysis methods do not distinguish between these types of censoring.
- This could lead to potential bias with random censoring!
 - Are the individuals different if they are randomly censored by outside causes?

Left and Interval Censoring

- If an observation is left censored, then all we know is that $T < c$.
- Example:
 - Became a customer more than 3 years ago.
- Interval censoring combines both right and left censoring where $a < T < b$.
- Example:
 - Onset of disease measured annually.



HAZARD FUNCTIONS

Hazard Function

- Hazards are one of the most important concepts to survival analysis.
- There are two common types of hazard functions:
 1. Hazard Probabilities:


$$h(t) = P(t < T < t + 1 \mid T > t)$$

2. Hazard Rates:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t < T < t + \Delta t \mid T > t)}{\Delta t}$$

Hazard Function

- Hazards are one of the most important concepts to survival analysis.
- There are two common types of hazard functions:
 1. Hazard Probabilities:


$$h(t) = P(t < T < t + 1 \mid T > t)$$

2. Hazard Rates:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t < T < t + \Delta t \mid T > t)}{\Delta t}$$

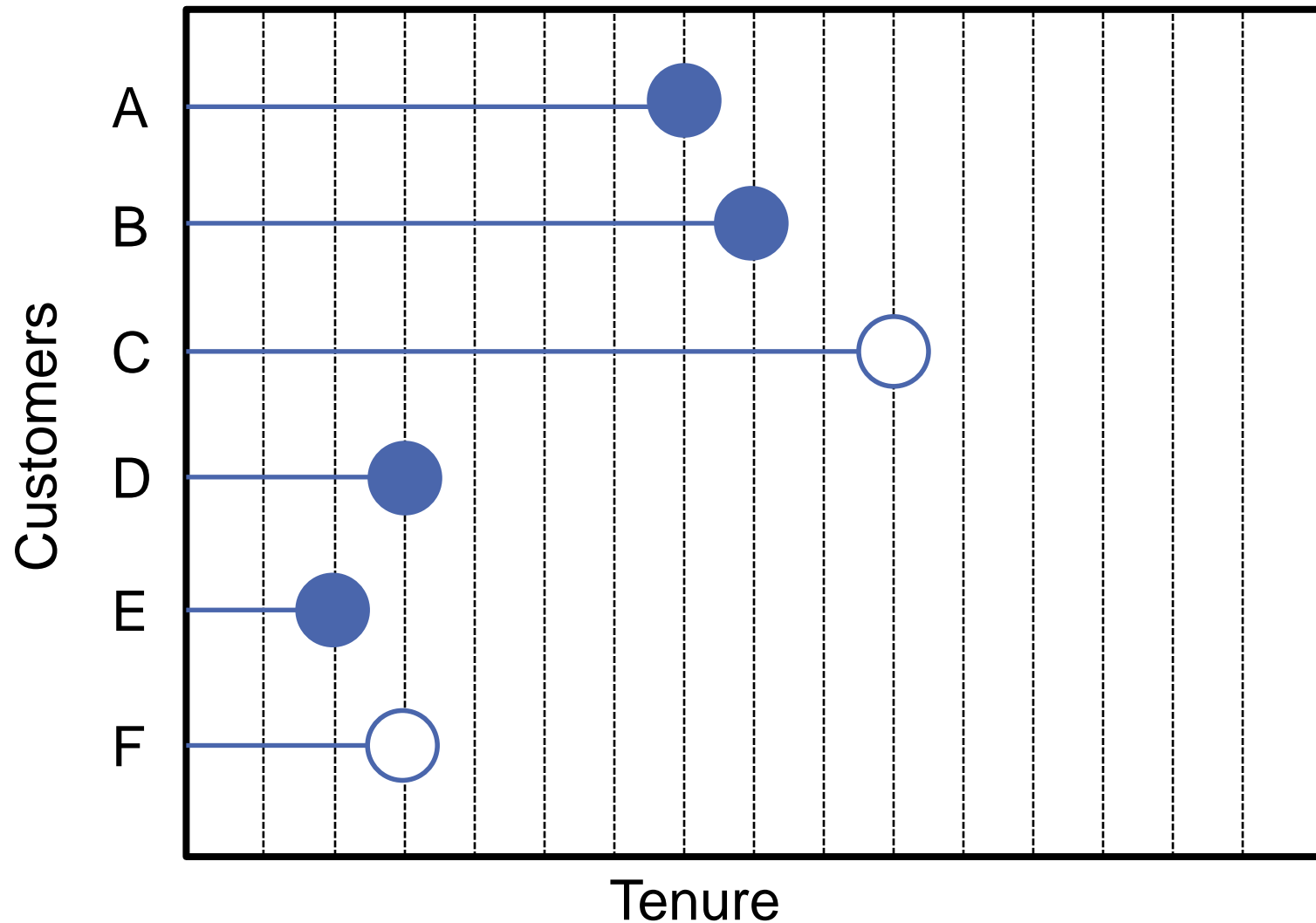
→ Both are denoted the same way in different texts!

Hazard Probabilities

- Hazard probabilities are very useful and common in business settings.
- Example:
 - A customer has survived for a certain length of time, so the customer's tenure is t .
 - What is the probability that the customer leaves before $t+1$?

$$h(t) = P(t < T < t + 1 \mid T > t) = \frac{d_t}{n_{t-1}}$$

Hazard Probabilities – Censoring



Hazard Probabilities – Censoring

- Time = 0:

$$h(0) = 0$$

- Time = 1:

$$h(1) = \frac{0}{6} = 0$$

- Time = 2:

$$h(2) = \frac{1}{6} = 0.1667$$

Hazard Probabilities – Censoring

- Time = 3:

$$h(3) = \frac{1}{5} = 0.2$$

- Time = 4:

$$h(4) = \frac{0}{3} = 0$$

- Time = 5:

$$h(5) = \frac{0}{3} = 0$$

Hazard Probabilities – Censoring

- Time = 3:

$$h(3) = \frac{1}{5} = 0.2 \quad \text{OR} \quad h(3) = \frac{1}{4.5} = 0.222$$

- Time = 4:

$$h(4) = \frac{0}{3} = 0$$

- Time = 5:

$$h(5) = \frac{0}{3} = 0$$

Hazard Probabilities – Censoring

- Time = 6:

$$h(6) = \frac{0}{3} = 0$$

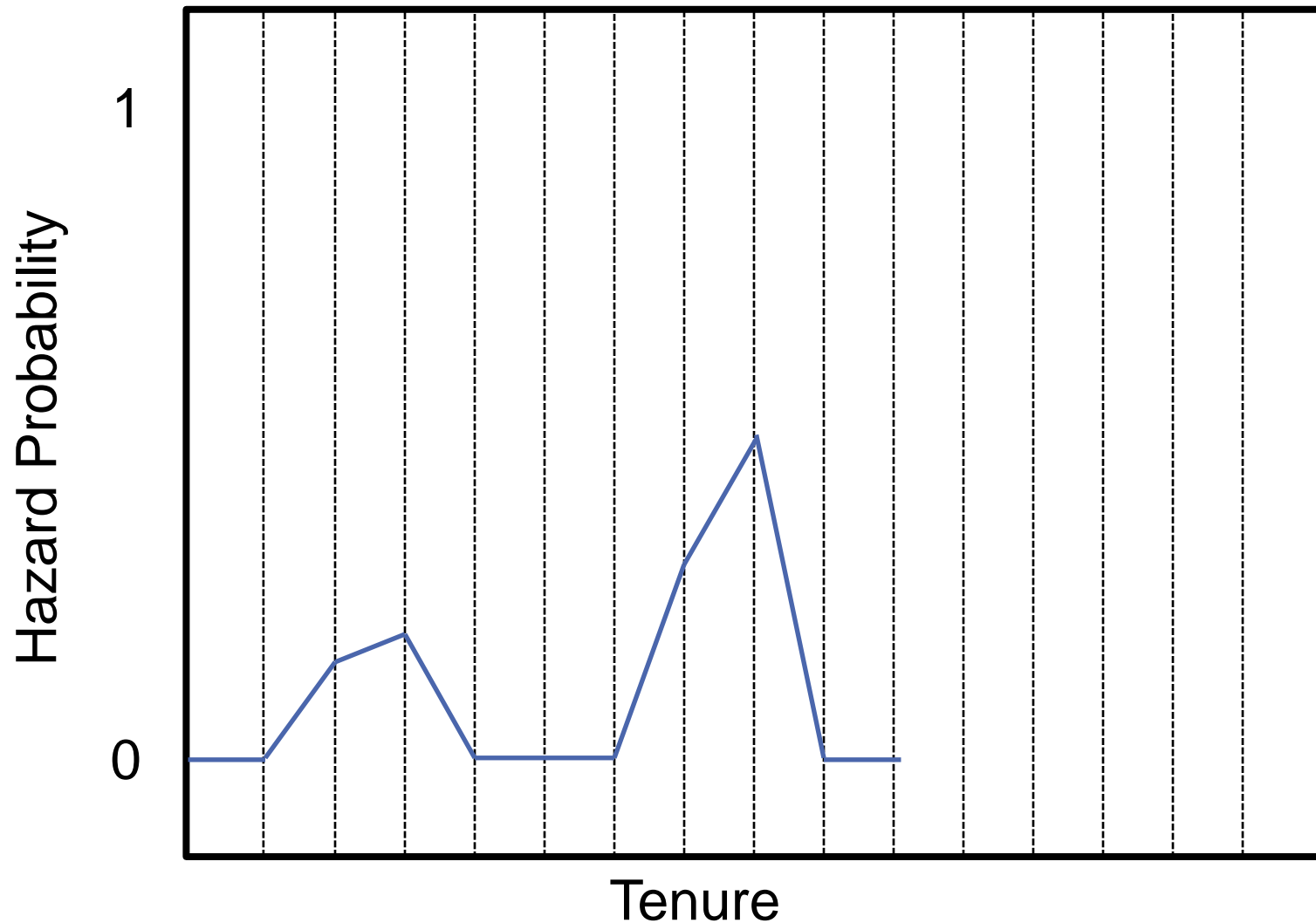
- Time = 7:

$$h(7) = \frac{1}{3} = 0.333$$

- Time = 8:

$$h(8) = \frac{1}{2} = 0.5$$

Hazard Probabilities – Censoring



Hazard Probabilities – Censoring

```
proc lifetest data=Simple method=life width=1  
              plots=h;  
  time Tenure*censored(1);  
run;
```

Hazard Rates

- Hazard rates have a slightly different interpretation than the hazard probabilities because they are limits of conditional probabilities.
- They are bounded below by 0, but are NOT bounded above by 1!

Hazard Rates

- Hazard rates are the rate of occurrence of an event.
- Examples:
 - Hazard for some point in time for contracting a sinus infection is 0.2 with a time measured in months.
 - “I am expected to contract a sinus infection 0.2 times in the next month (assuming the hazard stays constant).”

Hazard Rates – Inverse

- The interpretation of the inverse of the hazard function is the length of time before the next occurrence.
- Examples:
 - Hazard for some point in time for contracting a sinus infection is 0.2 with a time measured in months.
 - “I am expected to make it 5 ($= 1/0.2$) months before contracting my next sinus infection (assuming the hazard stays constant).”

Hazard Rates

```
proc lifetest data=Simple method=life width=1  
              plots=h;  
  time Tenure*censored(1);  
run;
```

