CENSORING, SURVIVAL, & HAZARDS

Dr. Aric LaBarr

Institute for Advanced Analytics

INTRODUCTION

What is Survival Analysis?

- In survival analysis, we are interested in the **time until an event occurs**, or **failure time**.
- Event is a qualitative change that can be tied to a specific point in time.
- Originally designed to study the occurrence of death in medical studies – hence *survival analysis*.
- Other names:
  - Time-to-event analysis
  - Duration analysis
  - Failure time analysis

"Time-to-Event" Data?

- In survival analysis, "time" generally refers to **tenure** rather than actual calendar time.
- The "event" is some specific outcome of interest:
  - Customer cancel service
  - Customer make another purchase
  - Patient develops disease

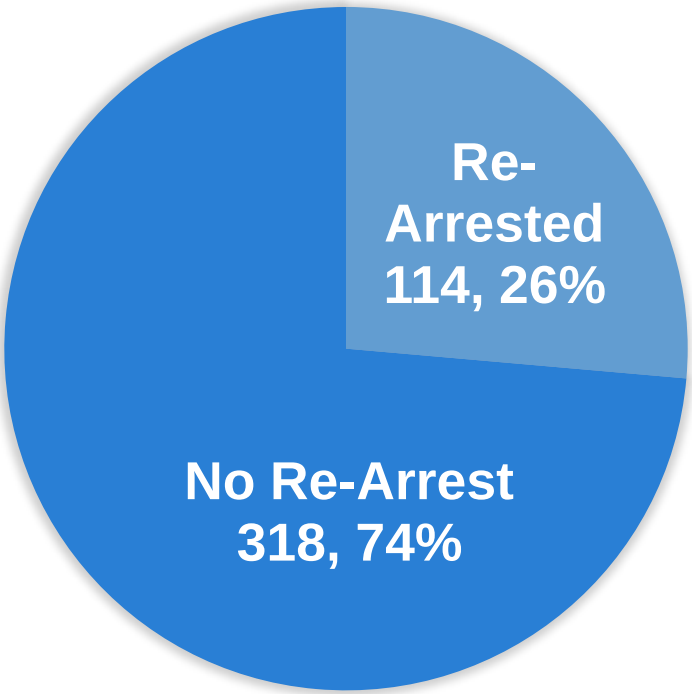
- Logistic regression: "Did it happen?"
- Survival analysis: "How long did it take to happen?"

Numeric Target – Linear Regression?

- Biggest problem with using OLS for time-to-event data is **censoring** – for some observations, the event may never occur (or hasn't occurred yet).
- Other problems with OLS:
  - Tenure is always positive – problem satisfying normality assumption
  - Risk of failure may change over time

Maryland Recidivism Data Set

- Study from 1970's following men for one year after being released from Maryland state prisons
- Of the 432 men, 114 were re-arrested within one year



Data Structure

- In survival analysis, the target variables is actually two pieces – one continuous and one categorical:
  1. **Time:** the tenure for an observation (continuous)
  2. **Event/status:** At the end of that time, what happened? (categorical)

Observation	Time (Week)	Status (Re-Arrest?)
1	20	1
2	17	1
3	25	1
4	52	0
5	52	0
6	52	0
7	23	1

Data Structure

- In survival analysis, the target variables is actually two pieces – one continuous and one categorical:
  1. **Time:** the tenure for an observation (continuous)
  2. **Event/status:** At the end of that time, what happened? (categorical)

Observation	Time (Week)	Status (Re-Arrest?)
1	20	1
2	17	1
3	25	1
4	52	0
5	52	0
6	52	0
7	23	1

Maryland Recidivism Data Set

- Model the association between various factors and length of time before re-arrest.
- Target:
  - **week:** week of arrest – week = 52 if censored (not arrested)
  - **arrest:** indicator for arrest (1 = yes, 0 = no)

Maryland Recidivism Data Set

- Model the association between various factors and length of time before re-arrest.
- Predictors:
  - **fin:** received financial aid upon release (1 = yes, 0 = no)
  - **age:** age at time of release (years)
  - **race:** indicator for *Black* (1 = yes, 0 = no)
  - **wexp:** indicator of prior work full-time work experience prior to incarceration (1 = yes, 0 = no)
  - **mar:** married at time of release (1 = yes, 0 = no)
  - **paro:** released on parole (1 = yes, 0 = no)
  - **prio:** number of prior convictions

TIME & CENSORING

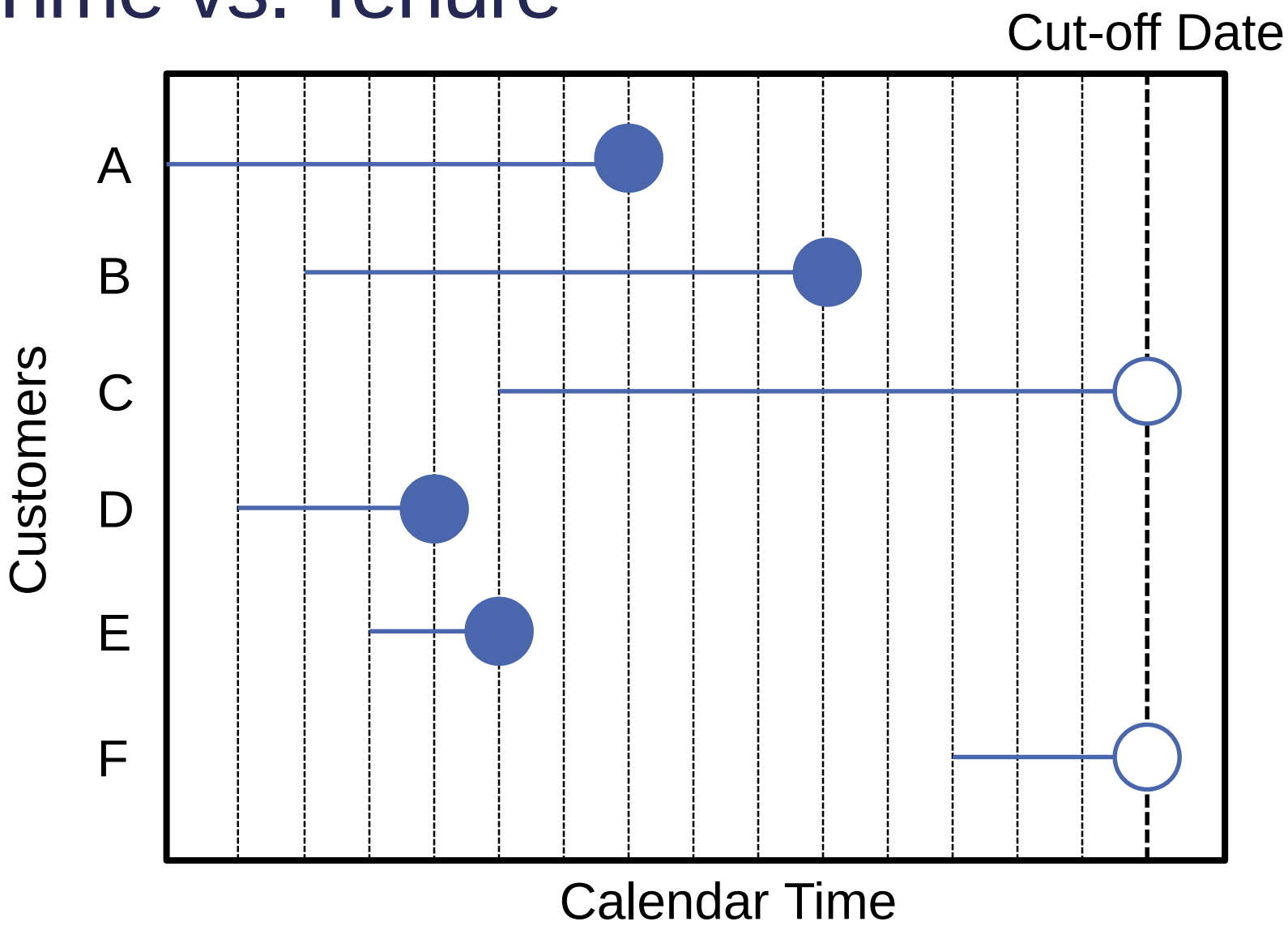
The Meaning of Time

- Survival analysis has a few things that set it apart from any other statistical modeling you've seen in this program so far:
  - Time to an event
  - Censoring
- Just like most models, survival analysis depends on certain assumptions.

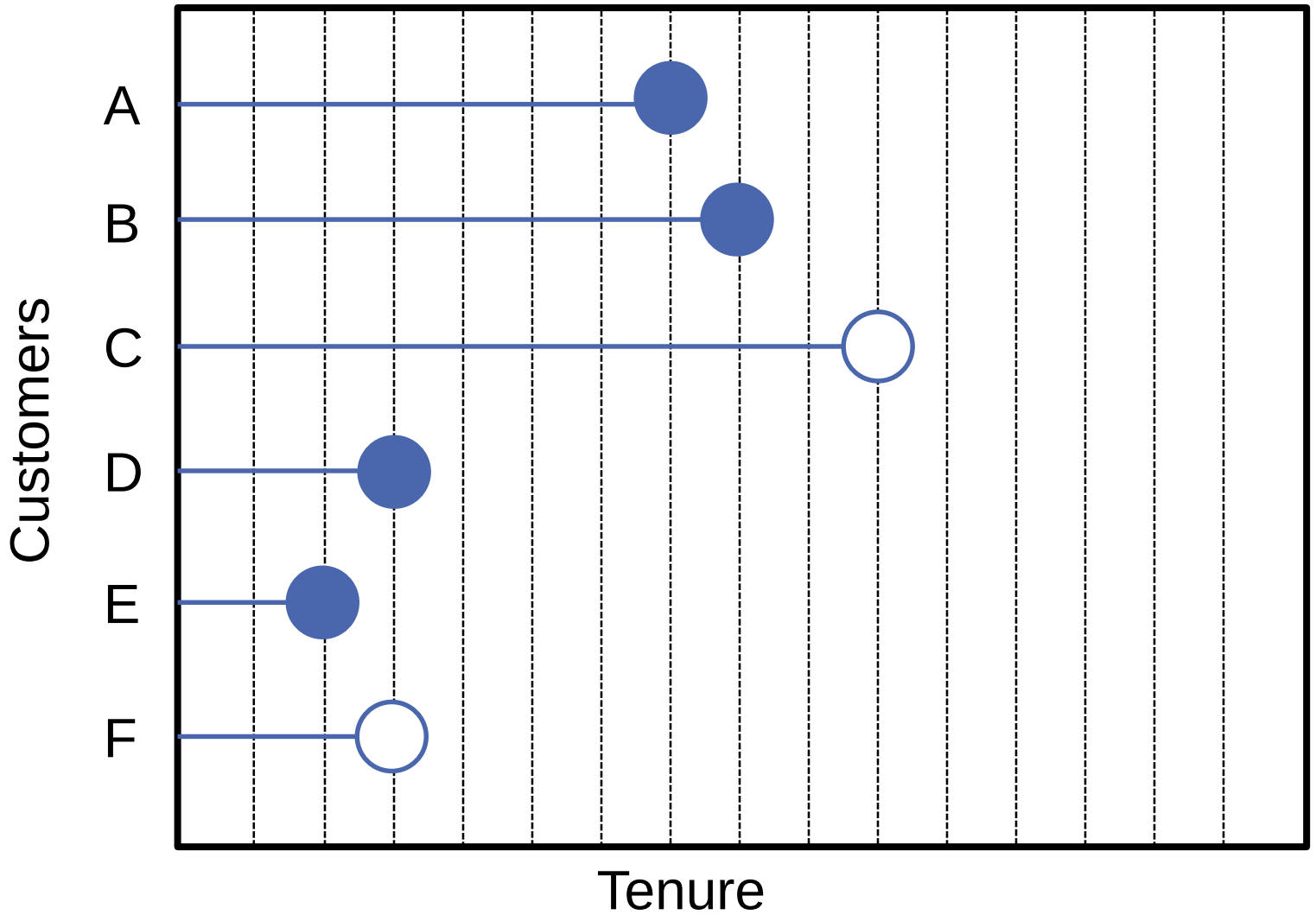
When Does Time Start?

- Create an artificial world in which everyone “starts” at the same time.
  - Not actually interested in time, but **tenure**.

Time vs. Tenure



Time vs. Tenure



When Does Time Start?

- Create an artificial world in which everyone “starts” at the same time.
  - Not actually interested in time, but **tenure**.

- Choice of starting point isn’t always obvious:
  - Time since exposure to disease vs. developing disease
  - Time since diagnosis vs. surgery vs. treatment
  - Time since another event
  - Time until car dies from production vs. purchase vs. last repair

Observed Time & Status

- Interested in time to event *T*, but we can not observe this for all observations.
- These observations are **censored**.
- The "time" we actually observe for each observation *i* is the minimum between T_i and C_i:
  - T_i is the time until the event
  - C_i is the censoring time

- Need another "status" variable to tell us which one we observe for each observation.

Data Structure

- In survival analysis, the target variables is actually two pieces – one continuous and one categorical:
  1. **Time:** the tenure for an observation (continuous)
  2. **Event/status:** At the end of that time, what happened? (categorical)

Observation	Time (Week)	Status (Re-Arrest?)
1	20	1
2	17	1
3	25	1
4	52	0
5	52	0
6	52	0
7	23	1

Censored **IS NOT** Missing

- Do not know the actual time to event T_i for censored observations.
- Do know that for some amount of time – namely, C_i – the event has not occurred.
  - Provides **some** but not all information about T_i.

- Censored data is **incomplete**, but not missing.
- Ignoring censored observations would be falsely acting as if we know nothing about T_i.

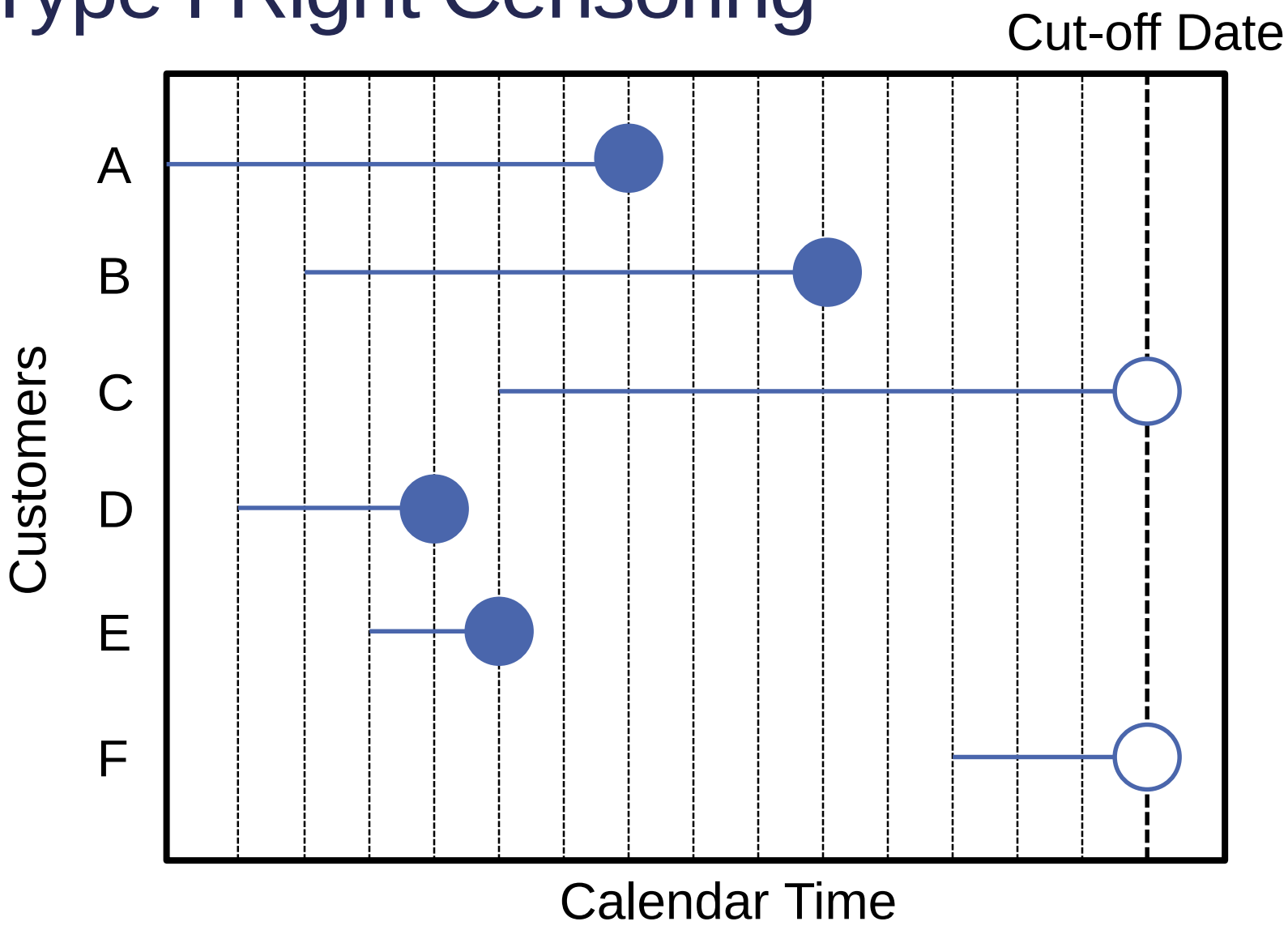
Type I vs. Type II Censoring

- **Type I** – there is a specific end time *c*, and any subject that hasn't had the event by time *c* is censored (most common).
- **Type II** – time goes until a certain (pre-specified) number of events have occurred, and any subjects who haven't had the event by that time are censored.

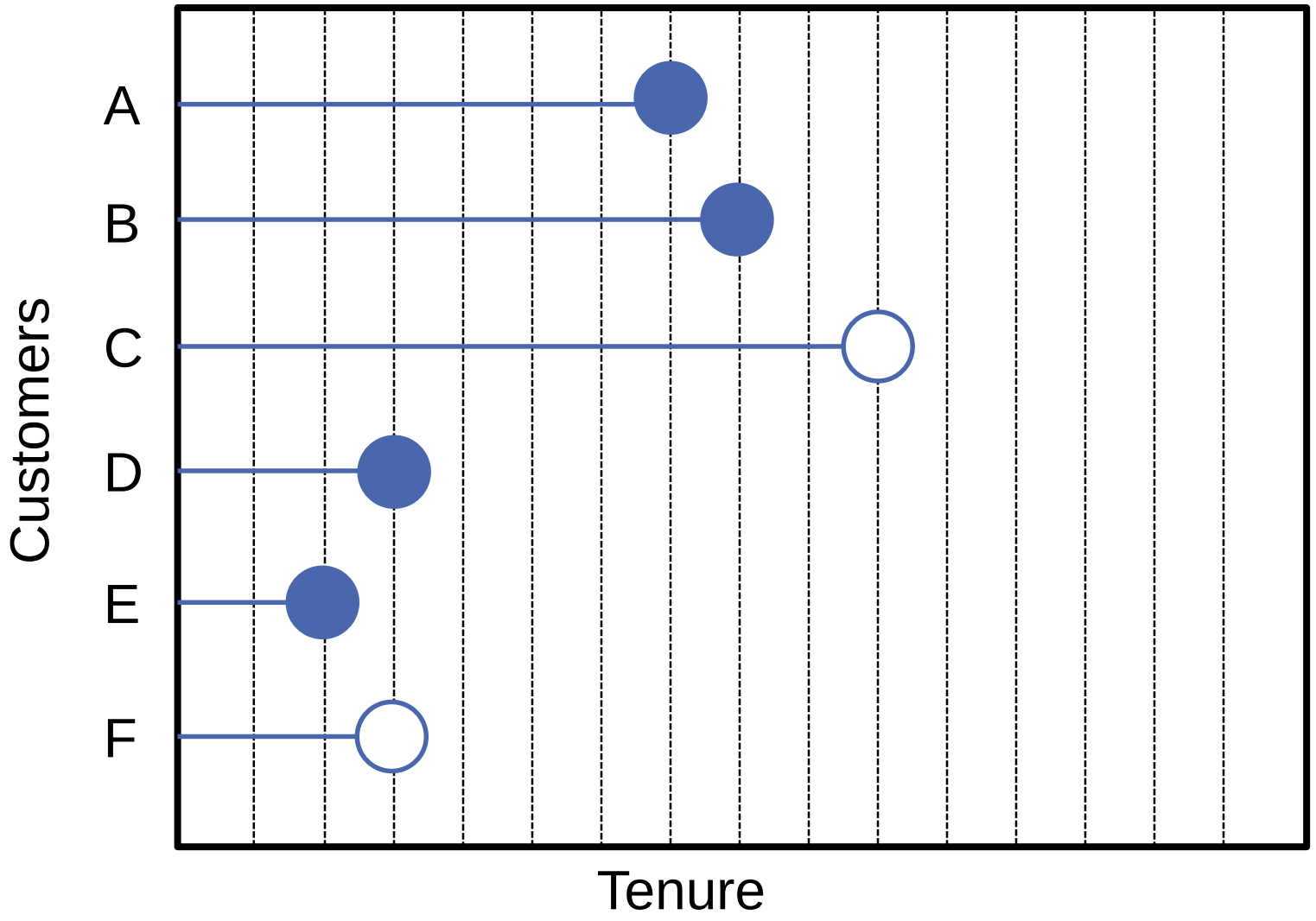
Right Censoring

- If an observation is right censored, then all we know is that $T_i > c$.
- Examples:
  - Person arrested after 52 weeks.
  - Income greater than $75,000.
  - Customer has monthly subscription for at least 3 months.
- Type I right censoring is the most common form.

Type I Right Censoring



Type I Right Censoring



Fixed vs. Random Censoring

- **Fixed censoring** – censoring only occurs at the end of the study ($C_i = c$ is known in advance).
  - Recidivism data: Not arrested in 52 weeks is censored by design because that is when study ended.

- **Random censoring** – C_i may vary between subjects for reasons beyond the investigator's control.
  - Recidivism data: No arrest within first 30 weeks, but lose contact with subject for whatever reason.
  - Recidivism data: Study done only for one year, but people can have delayed entry into the study (as they were released).

Independence

- Assume T_i and C_i are independent – subjects censored at time *t* were randomly selected to be censored from all subjects still in the risk set at *t*.
- **IF** this is true, then fixed vs. random censoring is mathematically equivalent.
- **IF NOT**, then we might need to get more complicated…(later in the course)

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. Implemented new customer tracking system, but current customers were around before.
- **Interval censoring** combines both right and left censoring where $a < T < b$.
- Example:
  - Person tests negative during appointment at *a*, but positive during appointment at *b*. So time developing disease is between *a* and *b.*

SURVIVAL FUNCTION

Summarizing Survival Data

- Interested in the event time *T*.
- Unique challenges to summarizing information about *T*:
  - Are means/variances useful for skewed distributions such as time?
  - In the presence of censoring, can we even estimate means and variances without actually knowing all the true values of *T*?
- Survival analysis described in **two** major quantities:
  - **Survival Function**
  - **Hazard Function**

Survival Function

- **Survival function**: probability of surviving **beyond** time *t*:

$$S(t) = P(T > t)$$

- Properties:
  - Always starts at 1 (or 100%).
  - Never increases.
  - Bounded below by 0 (or 0%).

- Survival curves use to be the only method in survival analysis.

Kaplan-Meier Method

- Estimating the survival function:
  - Want to estimate the proportion of individuals “still alive” at any given time *t*.

$$\hat{S}(t) = \prod_{k \le t} \left(1 - \frac{d_k}{r_k}\right)$$

→ # events occurring at time *t*

Kaplan-Meier Method

- Estimating the survival function:
  - Want to estimate the proportion of individuals “still alive” at any given time *t*.

$$\hat{S}(t) = \prod_{k \le t} \left(1 - \frac{d_k}{r_k}\right)$$

d_k: # events occurring at time *t*

r_k: # observations available right before time *t* (**risk set**)

Kaplan-Meier Method

- Estimating the survival function:
  - Want to estimate the proportion of individuals “still alive” at any given time *t*.

$$\hat{S}(t) = \prod_{k \le t} \left(1 - \frac{d_k}{r_k}\right)$$

d_k → # events occurring at time *t*

r_k → # observations available right before time *t* (**risk set**)

- 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.

Calculating K-M Estimate

- At the beginning ($t = 0$), all observations are at risk ($r_0 = n$) and no events have occurred ($d_0 = 0$):

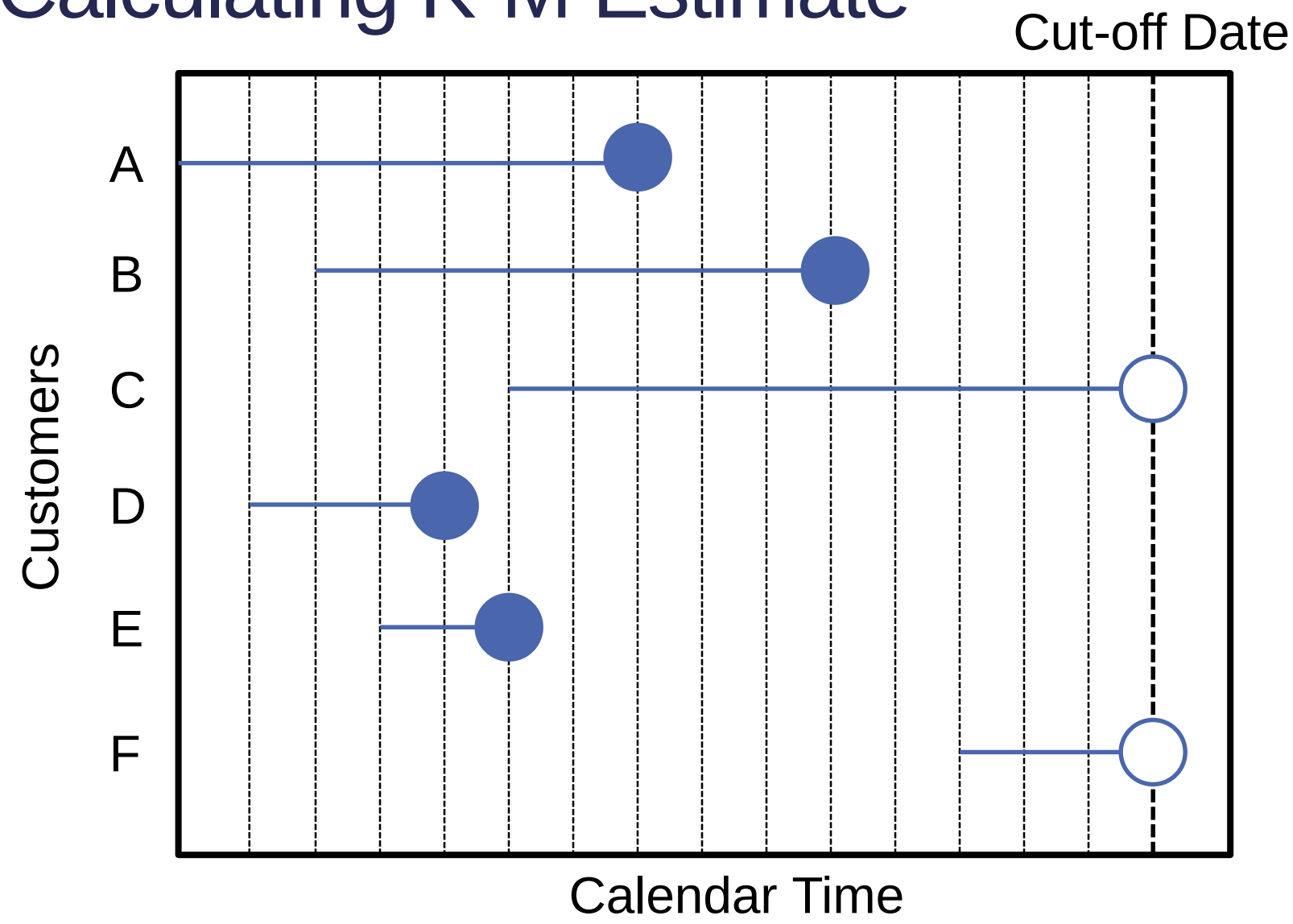
$$\hat{S}(t) = \prod_{k \le t} \left(1 - \frac{d_k}{r_k}\right) = \left(1 - \frac{0}{n}\right) = 1$$

- Start with $S(0) = 1$ and step forward in time, reducing $\hat{S}(t)$ by a factor of $\left(1 - \frac{d_t}{r_t}\right)$ at each time period:

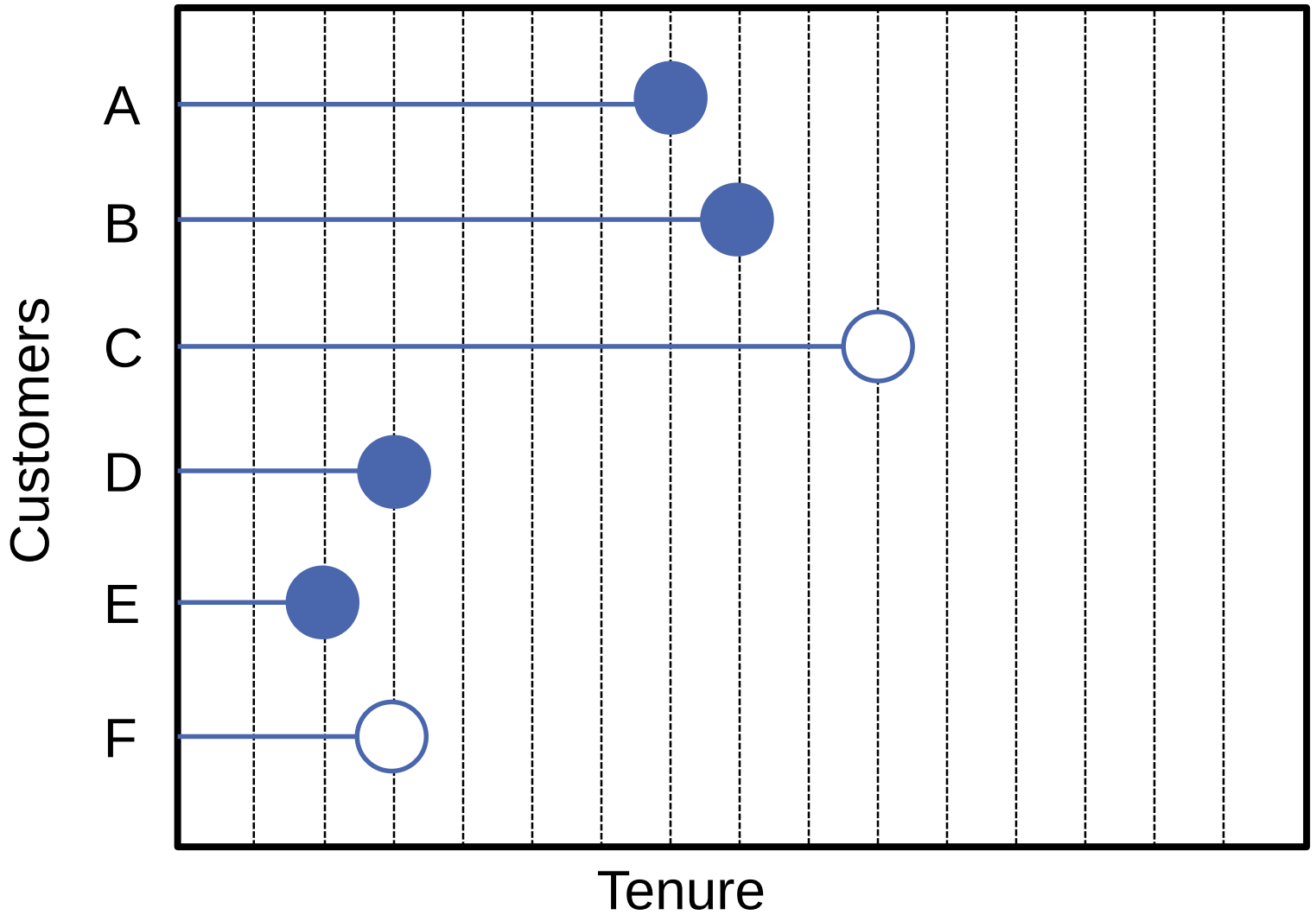
$$\hat{S}(1) = S(0) \times \left(1 - \frac{d_1}{r_1}\right)$$

$$\hat{S}(2) = \hat{S}(1) \times \left(1 - \frac{d_2}{r_2}\right)$$

Calculating K-M Estimate



Calculating K-M Estimate



Calculating K-M Estimate

- Time = 0:

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

- Time = 1:

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

- Time = 2:

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

Calculating K-M Estimate

- 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:

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

Calculating K-M Estimate

- Time = 6:

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

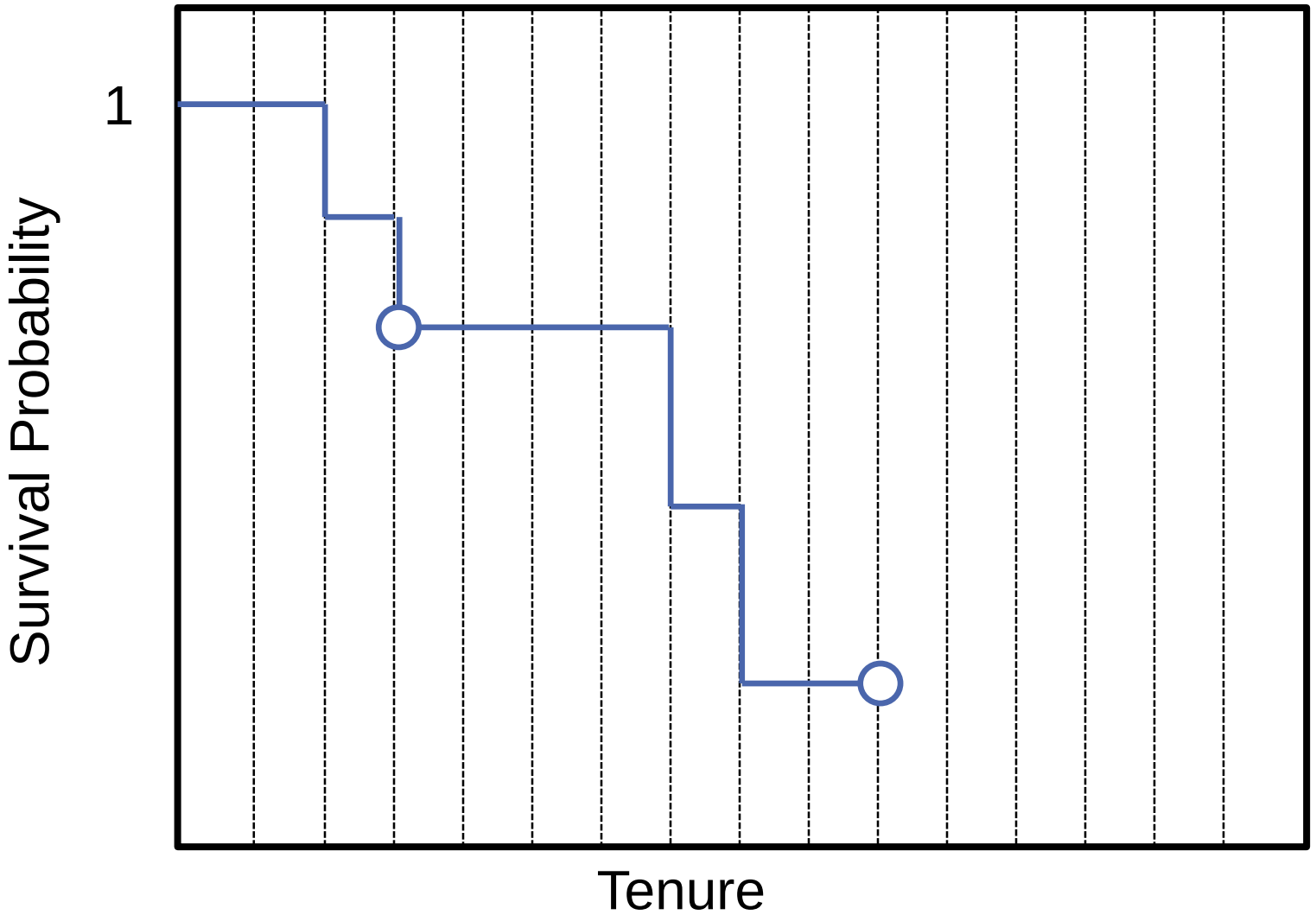
- Time = 7:

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

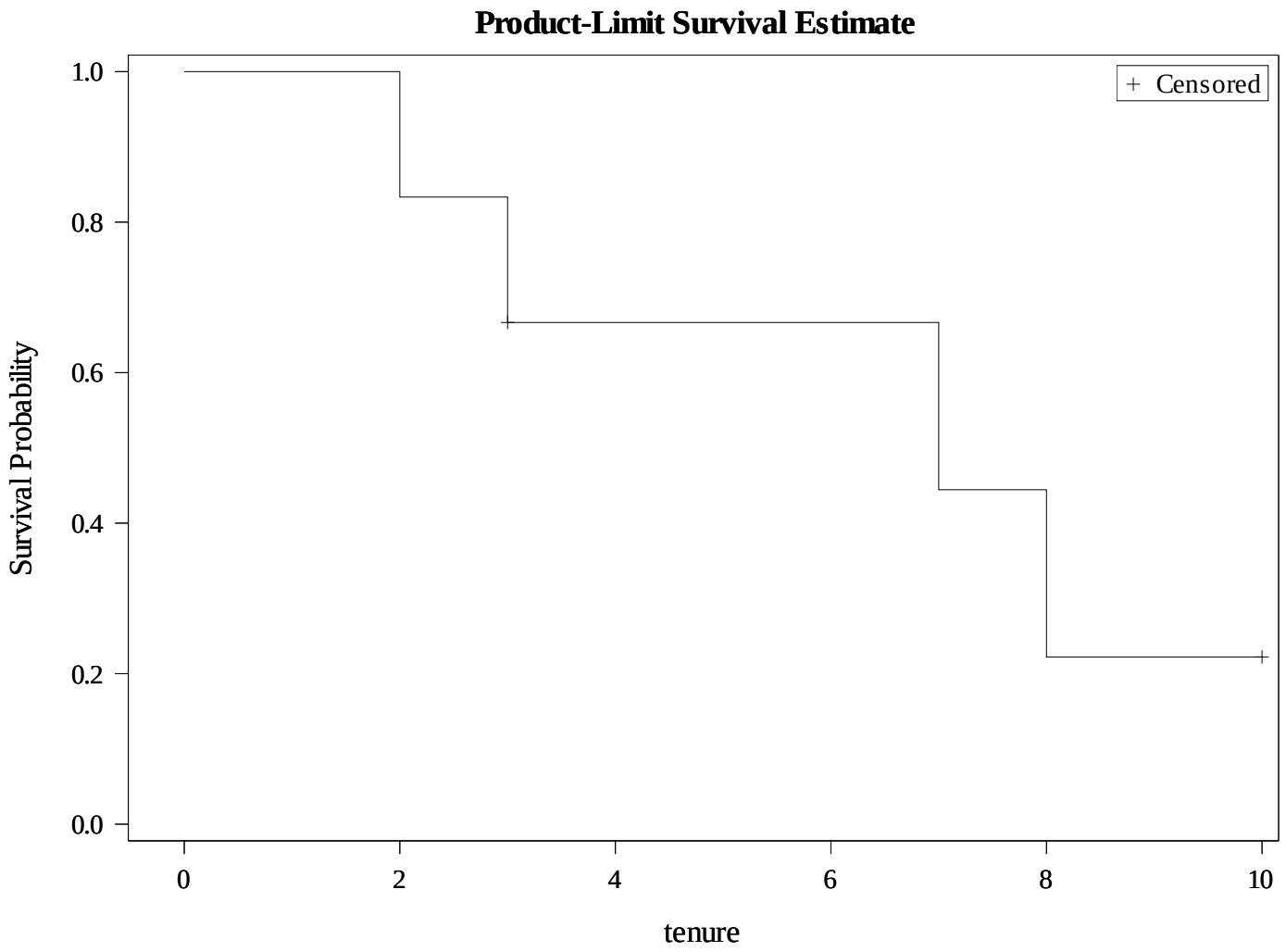
- Time = 8:

$$\hat{S}(5) = \hat{S}(4) \times \left(1 - \frac{1}{2}\right) = 0.444 \times 0.5 = 0.222$$

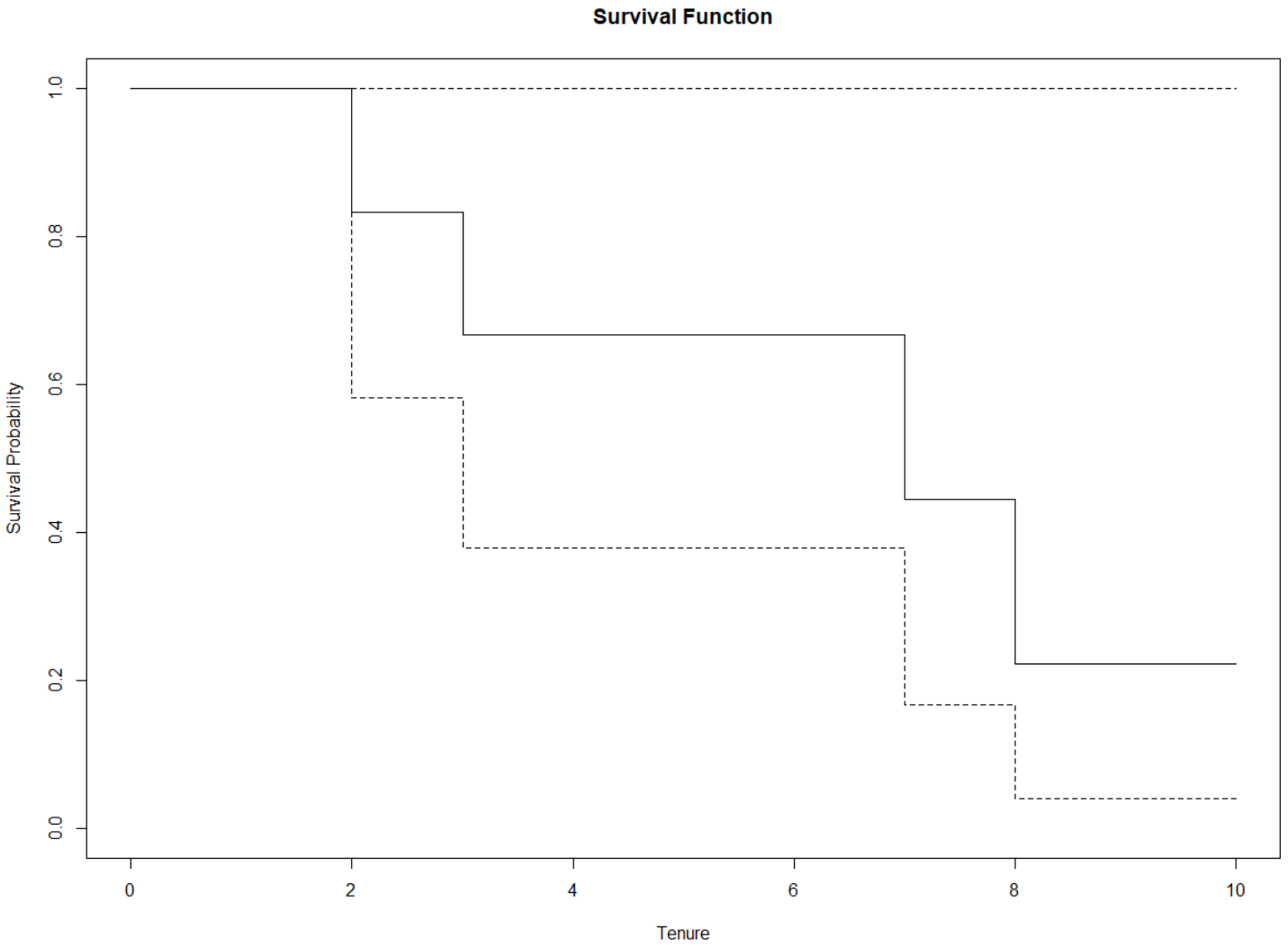
Visualizing K-M Estimate



Visualizing K-M Estimate



Visualizing K-M Estimate



Summary Statistics

- Due to censoring, the mean is impossible to truthfully estimate, but the **median** is still valid as long as the event occurs for at least half of the sample.
- The median (also called **half-life**) is the time t that $\hat{S}(t)$ drops below 0.5 (or 50%).
- **Half-life** interpretation: 50% of observations survive beyond time t.

Survival Function – SAS

```
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;
```

Survival Function – SAS

The LIFETEST Procedure

Product-Limit Survival Estimates						
Tenure		**Survival**	**Failure**	**Survival Standard Error**	**Number Failed**	**Number Left**
0.0000		1.0000	0	0	0	6
2.0000		0.8333	0.1667	0.1521	1	5
3.0000		0.6667	0.3333	0.1925	2	4
3.0000	*	.	.	.	2	3
7.0000		0.4444	0.5556	0.2222	3	2
8.0000		0.2222	0.7778	0.1925	4	1
10.0000	*	.	.	.	4	0

Note: The marked survival times are censored observations.

Survival Function – SAS

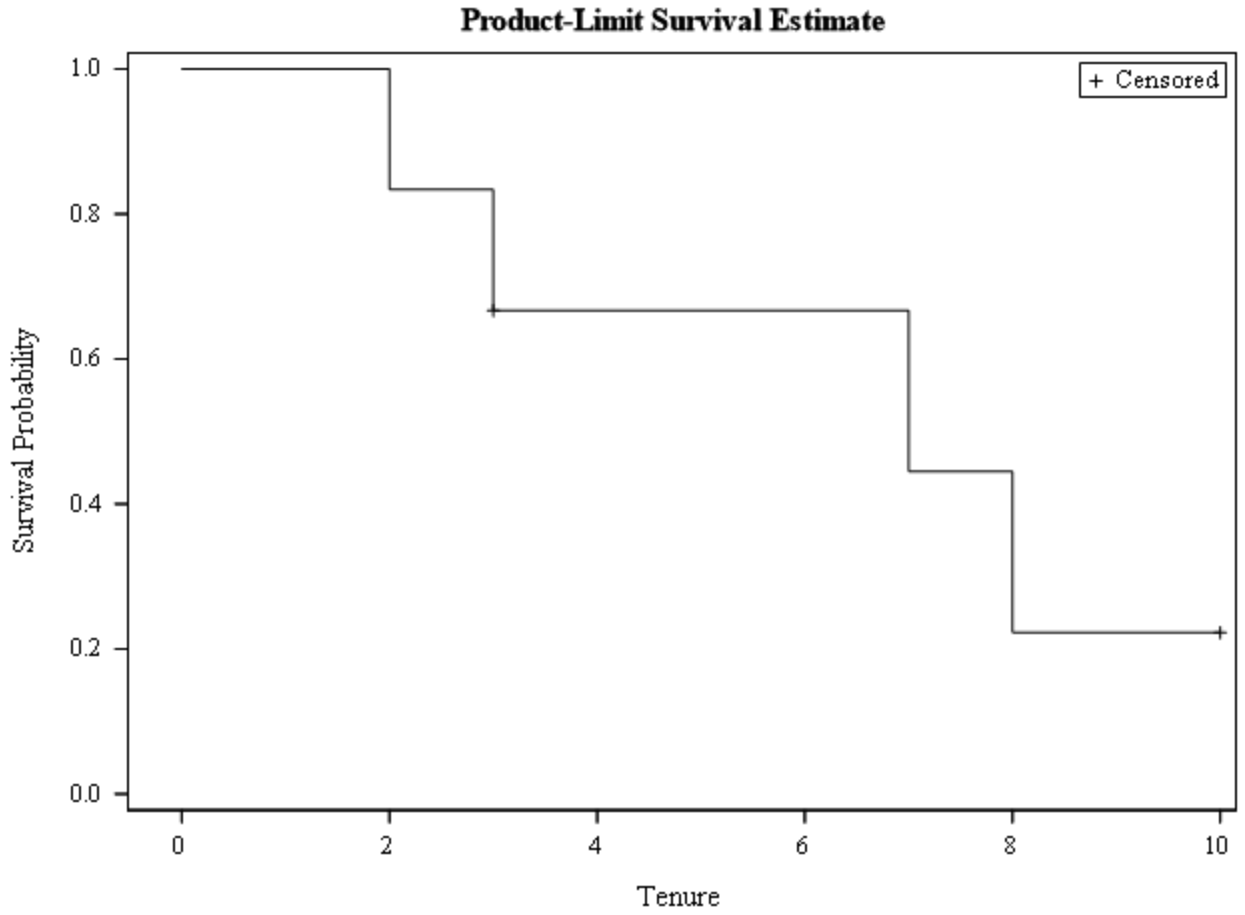
Summary Statistics for Time Variable Tenure

Quartile Estimates				
Percent	Point Estimate	95% Confidence Interval		
		Transform	[Lower	Upper)
75	8.0000	LOGLOG	3.0000	.
50	7.0000	LOGLOG	2.0000	.
25	3.0000	LOGLOG	2.0000	8.0000

Mean	Standard Error
5.9444	1.1750

Note: The mean survival time and its standard error were underestimated because the largest observation was censored and the estimation was restricted to the largest event time.

Survival Function – SAS



Summary of the Number of Censored and Uncensored Values			
Total	**Failed**	**Censored**	**Percent Censored**
6	4	2	33.33

Survival Function – SAS

```
proc lifetest data = Survival.Recid;
    time week*arrest(0);
run;
```

Survival Function – SAS

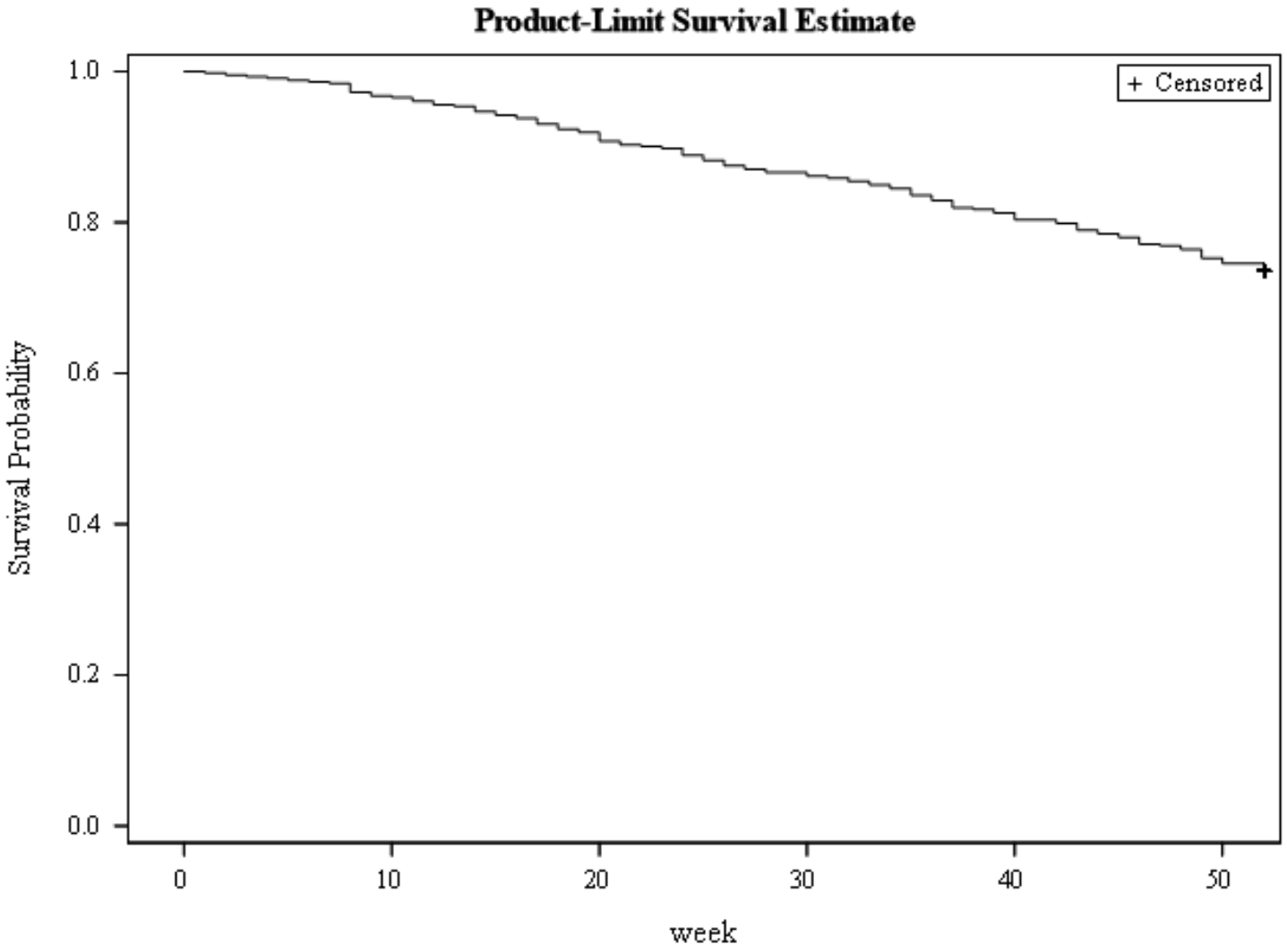
Summary Statistics for Time Variable week

Quartile Estimates				
Percent	Point Estimate	95% Confidence Interval		
		Transform	[Lower	Upper)
75	.	LOGLOG	.	.
50	.	LOGLOG	.	.
25	50.0000	LOGLOG	43.0000	.

Mean	Standard Error
45.8542	0.6112

Note: The mean survival time and its standard error were underestimated because the largest observation was censored and the estimation was restricted to the largest event time.

Survival Function – SAS



Survival Function – R

```
Surv(time = simple$tenure, event = simple$censored == 0)
```

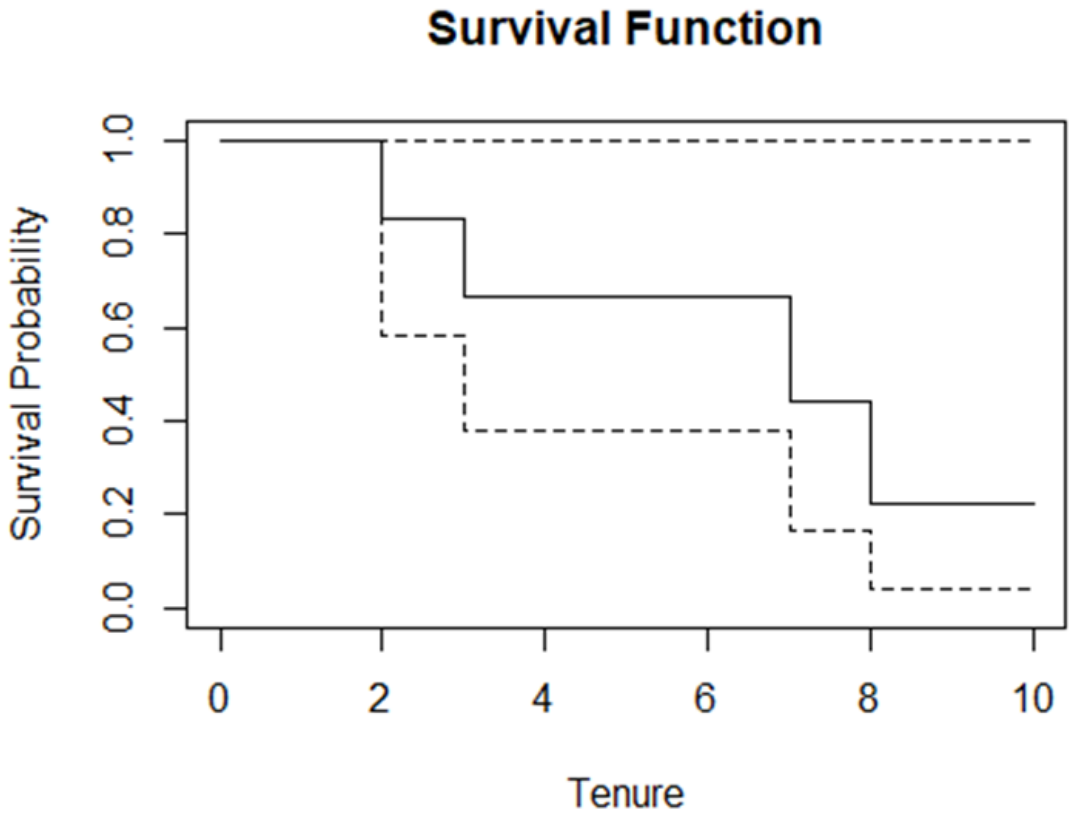
```
## [1]  7   8  10+  3   2   3+
```

```
simple_km <- survfit(Surv(time = tenure, event = (censored == 0)) ~ 1,
                     data = simple)
summary(simple_km)
```

```
## Call: survfit(formula = Surv(time = tenure, event = (censored == 0)) ~
##     1, data = simple)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     2      6       1    0.833   0.152       0.5827            1
##     3      5       1    0.667   0.192       0.3786            1
##     7      3       1    0.444   0.222       0.1668            1
##     8      2       1    0.222   0.192       0.0407            1
```

Survival Function – R

```
plot(simple_km, main = "Survival Function", xlab = "Tenure",
     ylab = "Survival Probability")
```



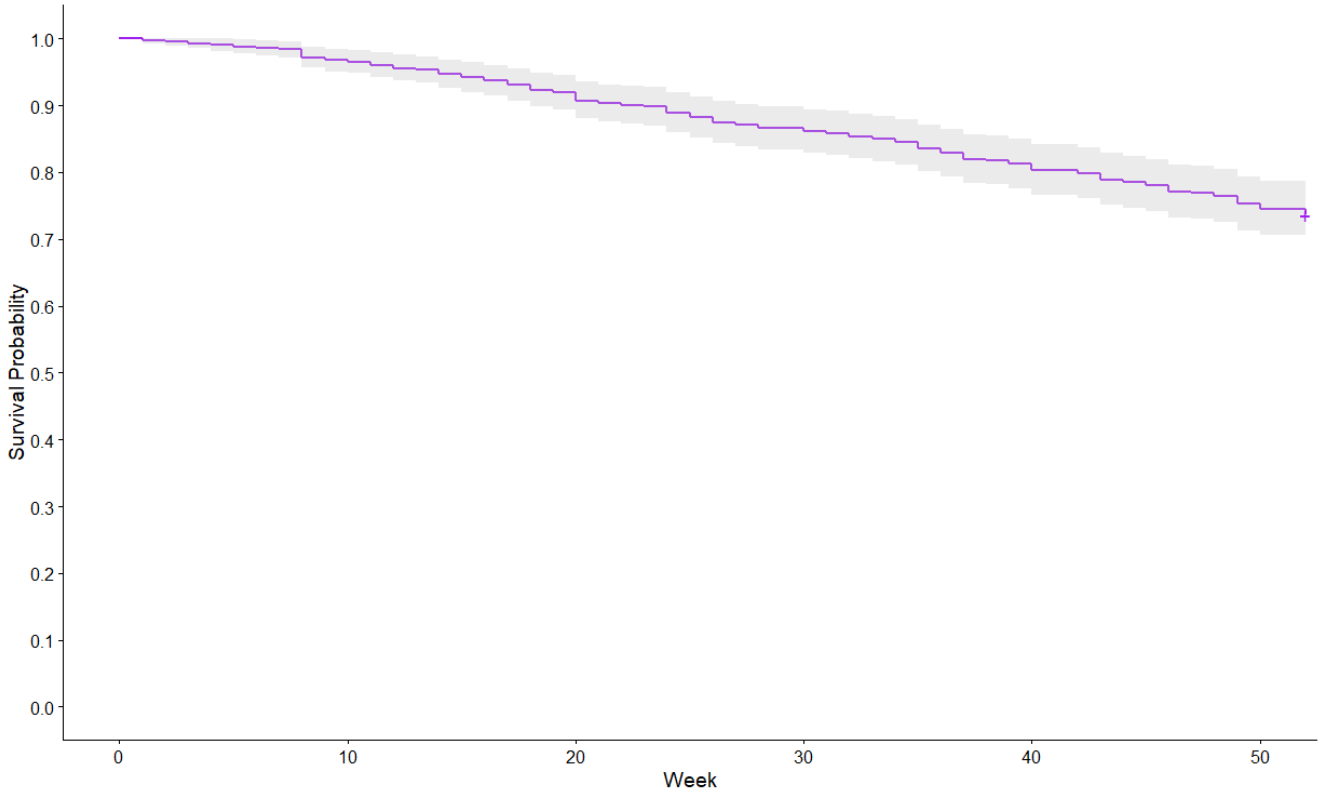
Survival Function – R

```
recid_surv <- Surv(time = recid$week, event = recid$arrest == 1)

recid_km <- survfit(recid_surv ~ 1, data = recid)
summary(recid_km)
```

```
## Call: survfit(formula = recid_surv ~ 1, data = recid)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1    432       1    0.998 0.00231        0.993        1.000
##     2    431       1    0.995 0.00327        0.989        1.000
##     3    430       1    0.993 0.00400        0.985        1.000
##     4    429       1    0.991 0.00461        0.982        1.000
##     5    428       1    0.988 0.00515        0.978        0.999
##     6    427       1    0.986 0.00563        0.975        0.997
##     7    426       1    0.984 0.00607        0.972        0.996
##     8    425       5    0.972 0.00791        0.957        0.988
```

Survival Function – R

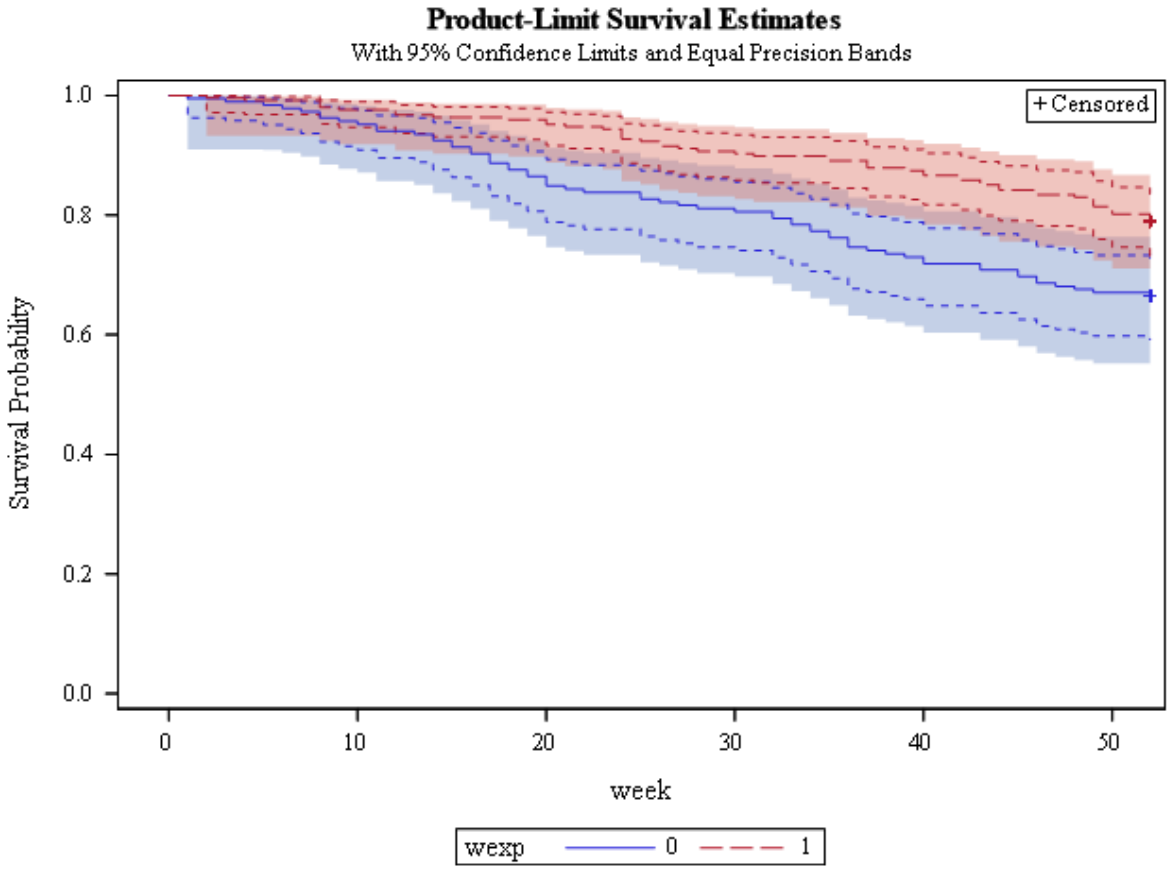
```
ggsurvplot(recid_km, data = recid, conf.int = TRUE, palette = "purple",
           xlab = "Week", ylab = "Survival Probability", legend = "none",
           break.y.by = 0.1)
```



STRATIFIED ANALYSIS

Stratified Analysis

- Can also create separate/stratified curves by group.
- Different curves result in different estimates for each group.



Stratified Analysis

- SAS provides 3 tests that each have the same null hypothesis – all survival curves are **equal**.
  1. Log-rank test (developed by Mantel-Haenszel)
  2. Wilcoxon test
  3. Likelihood Ratio test (exponential distribution!)

- R provides whatever you ask for! ☺

Log-rank Tests

- The **log-rank test** combines all the information from the K-M estimate at times where events occur.
- Similar to the Mantel-Haenszel tests for association from categorical data.

At time *t*	# Events	# Non-events	Total
Group 1	$d_{1,t}$	$r_{1,t} - d_{1,t}$	$r_{1,t}$
Group 2	$d_{2,t}$	$r_{2,t} - d_{2,t}$	$r_{2,t}$
Total	d_t	$r_t - d_t$	r_t

Comparing Survival Function

- Log-Rank test:

$$\text{LogRank} = \frac{1}{\hat{\sigma}^2}\left\{\sum_{j=1}^{r}\left(d_{1,j} - e_{1,j}\right)\right\}^2$$

- Wilcoxon test (places larger emphasis on earlier event times):

$$\text{Wilcoxon} = \frac{1}{\hat{\sigma}^2}\left\{\sum_{j=1}^{r}\left(d_{1,j} - e_{1,j}\right)n_j\right\}^2$$

Stratified Analysis – SAS

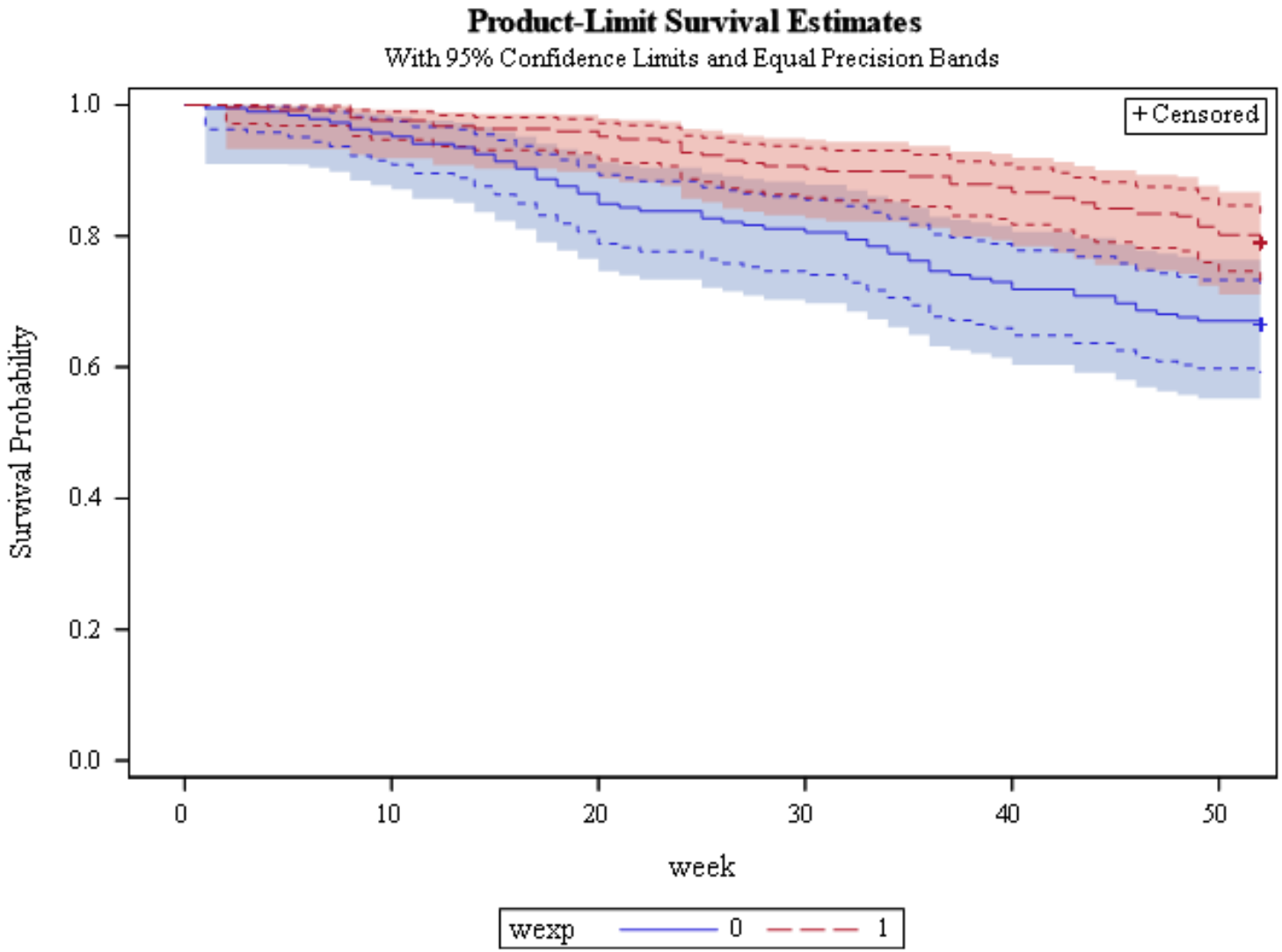
```
proc lifetest data = Survival.Recid plots=s(cl cb=ep);
    time week*arrest(0);
    strata wexp;
run;
```

Stratified Analysis – SAS

Summary of the Number of Censored and Uncensored Values					
Stratum	**wexp**	**Total**	**Failed**	**Censored**	**Percent Censored**
1	0	185	62	123	66.49
2	1	247	52	195	78.95
Total		432	114	318	73.61

Test of Equality over Strata			
Test	**Chi-Square**	**DF**	**Pr > Chi-Square**
Log-Rank	9.9104	1	0.0016
Wilcoxon	10.9815	1	0.0009
-2Log(LR)	9.0608	1	0.0026

Stratified Analysis – SAS



Stratified Analysis – R

```
survdiff(recid_surv ~ wexp, rho = 0, data = recid)
```

```
## Call:
## survdiff(formula = recid_surv ~ wexp, data = recid, rho = 0)
##
##          N Observed Expected (O-E)^2/E (O-E)^2/V
## wexp=0 185       62     45.6      5.91      9.91
## wexp=1 247       52     68.4      3.94      9.91
##
##  Chisq= 9.9  on 1 degrees of freedom, p= 0.002
```

Stratified Analysis – R

```
recid_strat <- survfit(recid_surv ~ wexp, data = recid)
summary(recid_strat)
```

```
## Call: survfit(formula = recid_surv ~ wexp, data = recid)
##
##                 wexp=0
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     1    185       1    0.995 0.00539        0.984        1.000
##     3    184       1    0.989 0.00760        0.974        1.000
##     5    183       1    0.984 0.00929        0.966        1.000
```

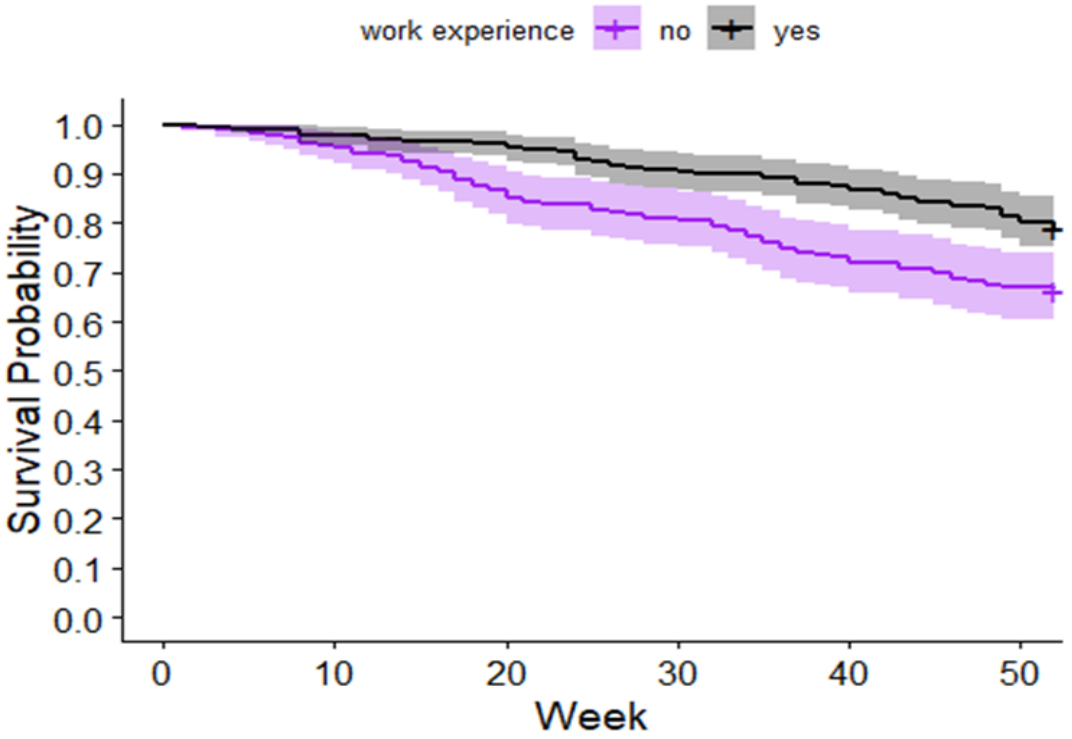
⋮

```
##                 wexp=1
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     2    247       1    0.996 0.00404        0.988        1.000
##     4    246       1    0.992 0.00570        0.981        1.000
##     8    245       3    0.980 0.00896        0.962        0.997
##     9    242       1    0.976 0.00980        0.957        0.995
```

⋮

Stratified Analysis – R

```
ggsurvplot(recid_strat, data = recid, conf.int = TRUE,
           palette = c("purple", "black"),
           xlab = "Week", ylab = "Survival Probability", break.y.by = 0.1,
           legend.title = "work experience", legend.labs = c("no", "yes"))
```



HAZARD FUNCTION

Hazard Function

- In survival analysis we also use the **hazard function** to summarize the data.
- There are two common types of hazard functions:
  1. Hazard Probabilities:

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

  2. Hazard Rates:

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

Hazard Function

- In survival analysis we also use the **hazard function** to summarize the data.
- There are two common types of hazard functions:
  1. Hazard Probabilities:

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

  2. Hazard Rates:

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

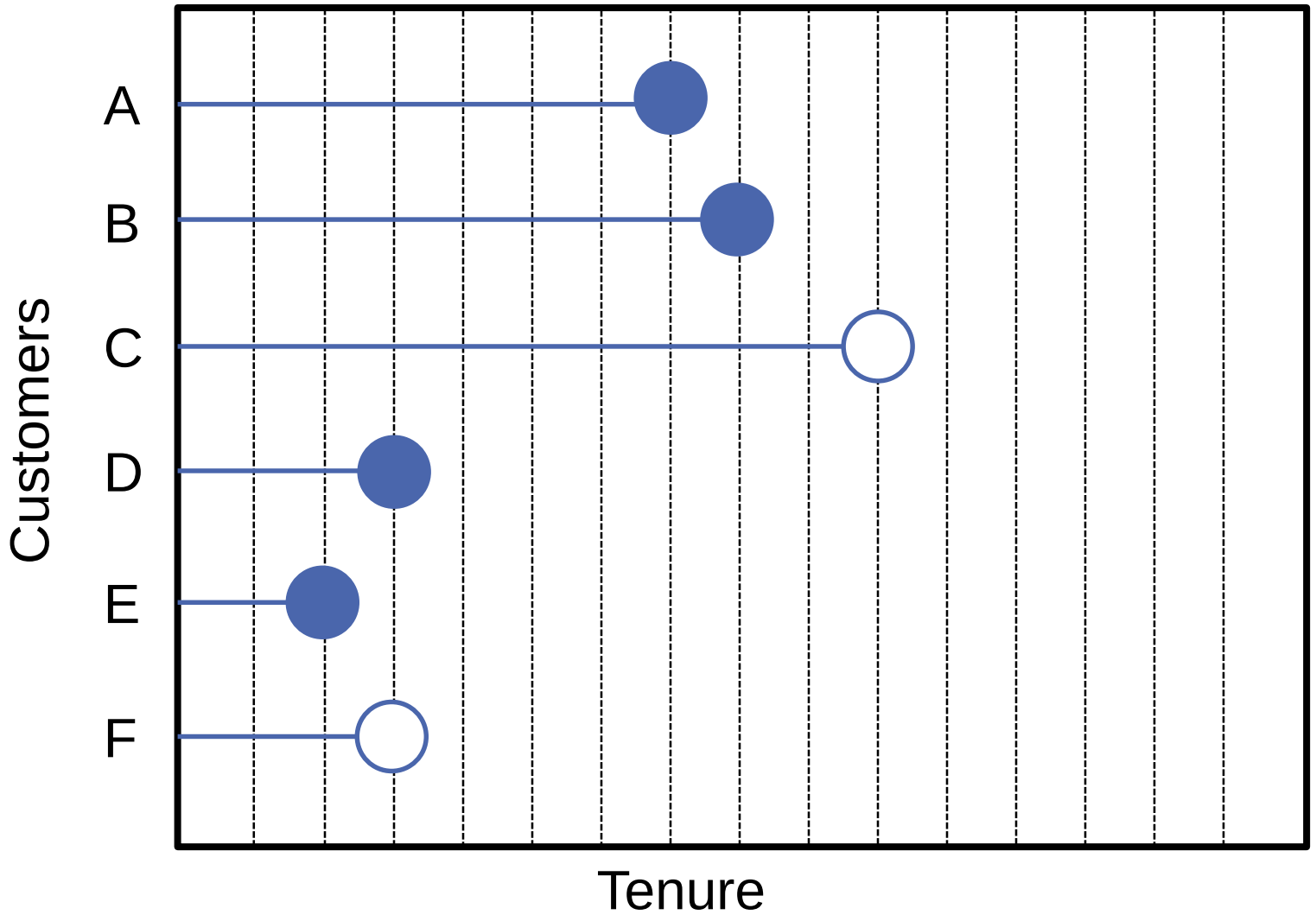
Both are denoted the same way in different texts!

Hazard Probabilities

- Hazard probabilities are very useful and common is 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*?

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

Calculating Hazard Probabilities



Calculating Hazard Probabilities

- Time = 0:

$$h(0) = 0$$

- Time = 1:

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

- Time = 2:

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

Calculating Hazard Probabilities

- Time = 3:

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

- Time = 4:

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

- Time = 5:

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

Calculating Hazard Probabilities

- Time = 3:

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

- Time = 4:

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

- Time = 5:

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

Calculating Hazard Probabilities

- Time = 6:

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

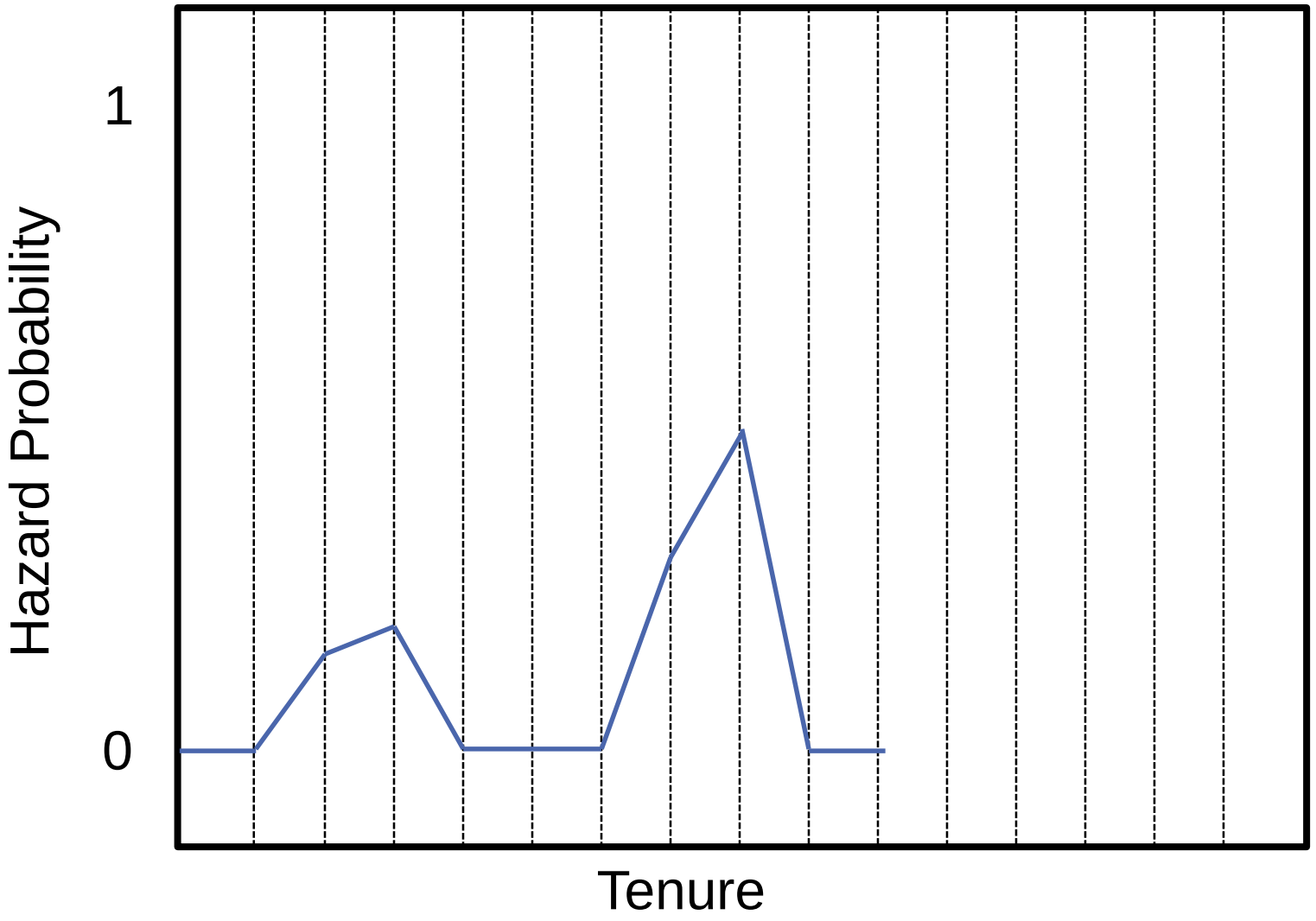
- Time = 7:

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

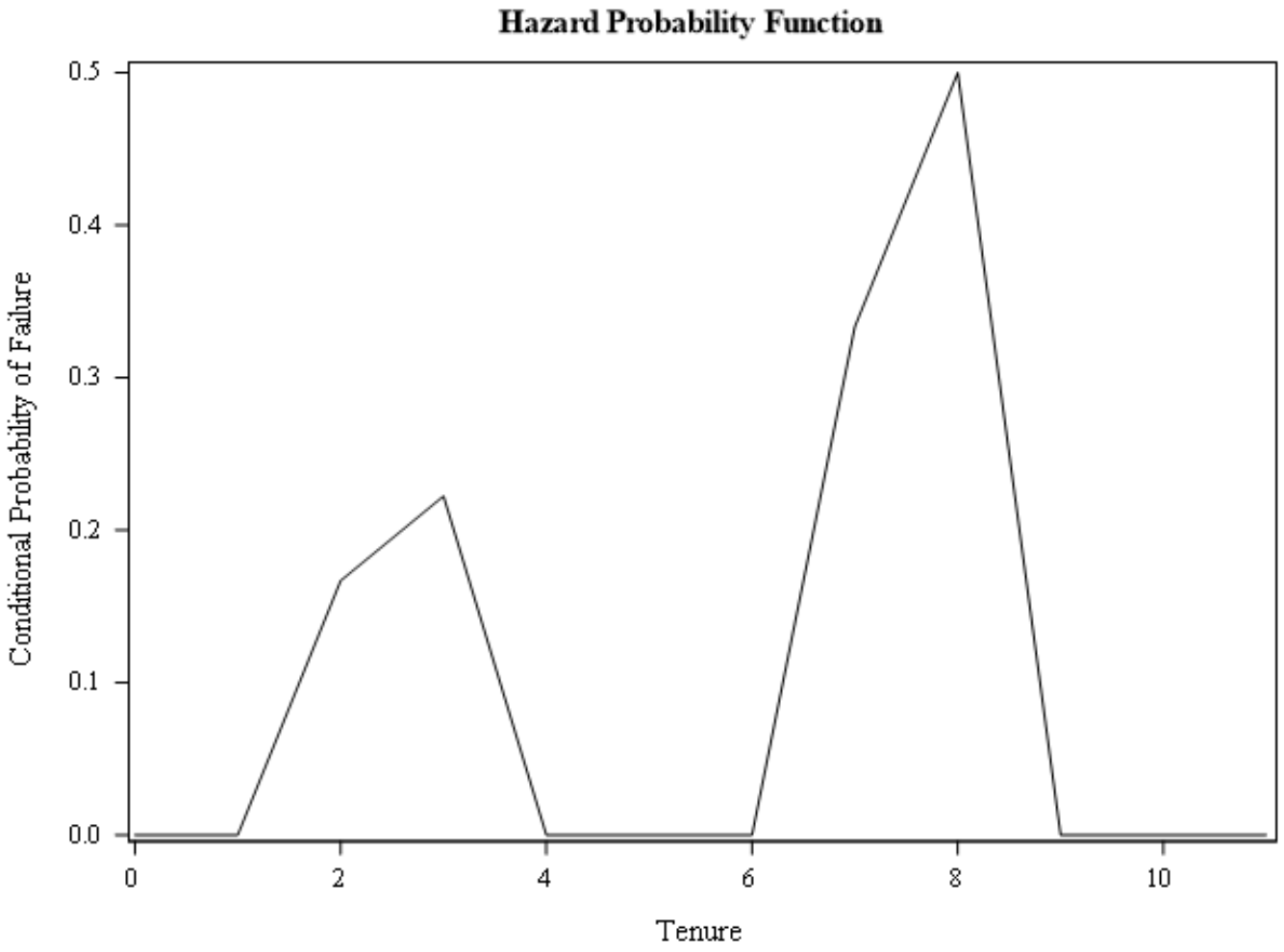
- Time = 8:

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

Visualizing Hazard Probabilities



Visualizing Hazard Probabilities



Hazard Rates

- Hazard rates have a slightly different interpretation than the hazard probabilities because they are limits of conditional probabilities.

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

- The hazard rate is the **instantaneous event rate** for the risk set at time *t*.
  - Given survival up until time *t*, it is the rate of events in the interval $[t, t + \Delta t)$.

Hazard Rates

- Hazard rates have a slightly different interpretation than the hazard probabilities because they are limits of conditional probabilities.

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

- The hazard rate is the **instantaneous event rate** for the risk set at time *t*.
- Bounded below by 0, but are NOT bounded above by 1!

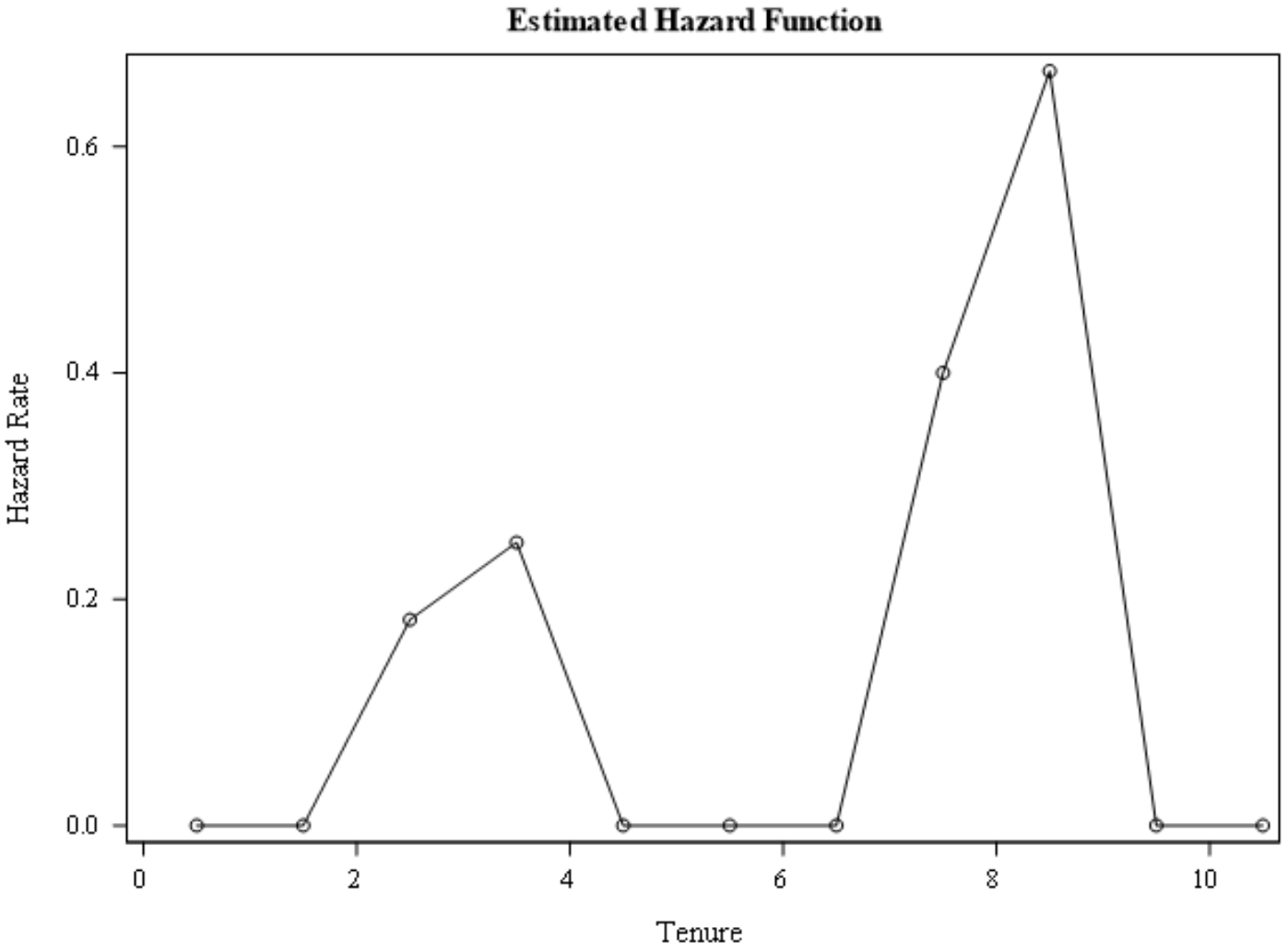
Hazard Rates

- 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

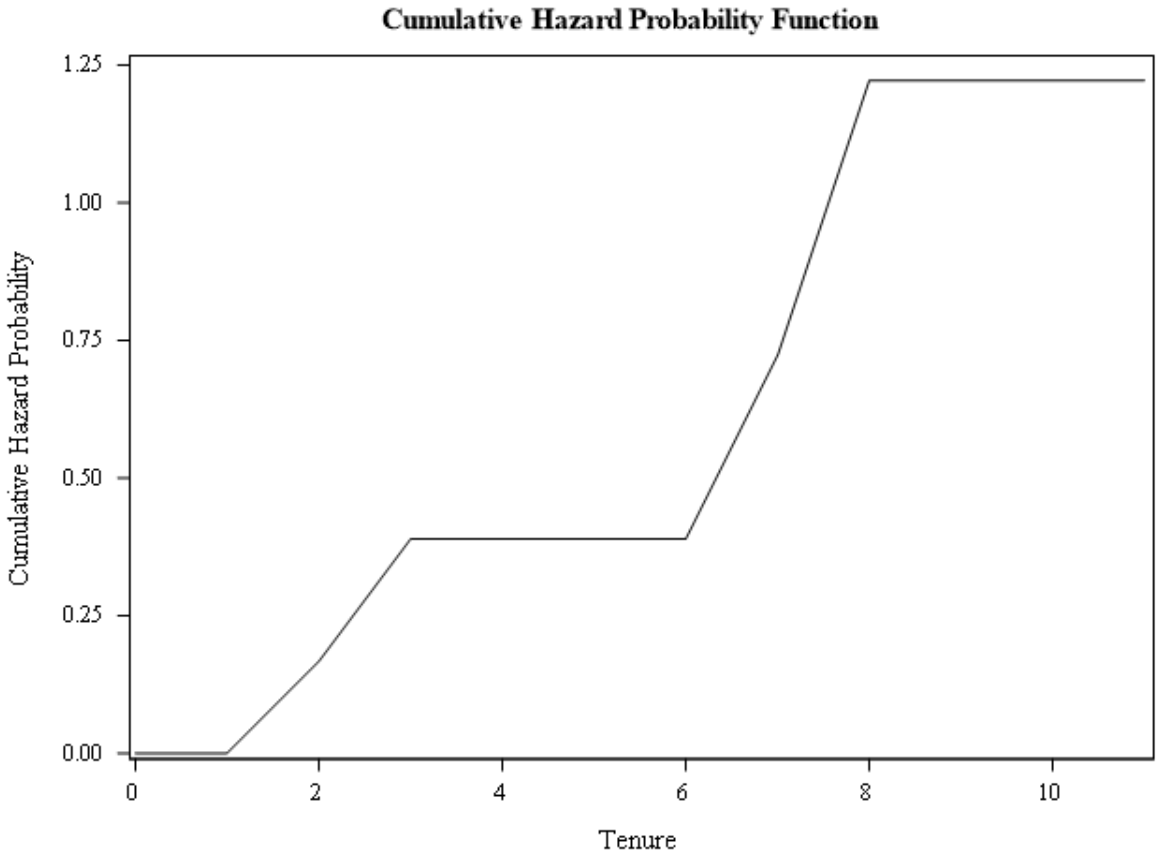
- The interpretation of the inverse of the hazard function is the length of time before the next occurrence.
- Examples:
  - Hazard for some point in time for contracting a sinus infection is 0.2 with a time measured in months.
  - “I am expected to make it 5 (= 1/0.2) months before contracting my next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

Visualizing Hazard Rate



Cumulative Hazard Probability

- The **cumulative hazard probability** is just the total hazard rate up until time t – denoted $\Lambda(t)$.



Life-Table (Actuarial) Method

- Life-table method groups event times into intervals.
- Handles large data sets better than Kaplan-Meier.
- Life-table method allows for estimation of the *hazard function* in SAS.
- Calculation is based on conditional probabilities.

Hazard Functions – SAS

```
proc lifetest data = Simple method = life width = 1
              plots=hazard;
   time tenure*censored(1);
   ods output LifetableEstimates = condprob;
run;

proc sgplot data = condprob;
   series x = lowertime y = condprobfail;
   xaxis label='Tenure';
   title 'Hazard Probability Function';
run;
quit;
```

Hazard Functions – SAS

The LIFETEST Procedure

Life Table Survival Estimates

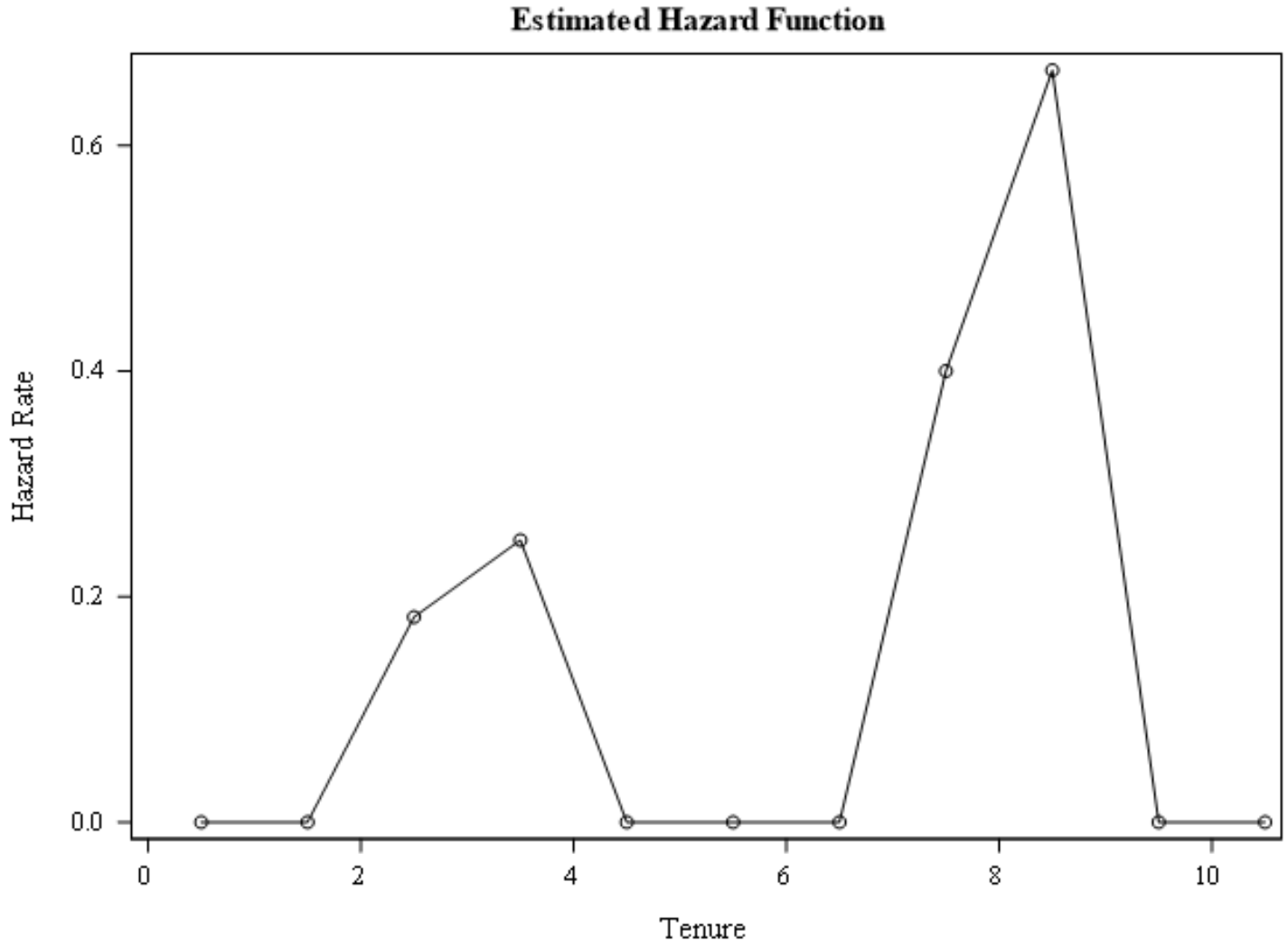
Interval		Number Failed	Number Censored	Effective Sample Size	Conditional Probability of Failure	Conditional Probability Standard Error	Survival
[Lower,	Upper)						
0	1	0	0	6.0	0	0	1.0000
1	2	0	0	6.0	0	0	1.0000
2	3	1	0	6.0	0.1667	0.1521	1.0000
3	4	1	1	4.5	0.2222	0.1960	0.8333
4	5	0	0	3.0	0	0	0.6481
5	6	0	0	3.0	0	0	0.6481
6	7	0	0	3.0	0	0	0.6481
7	8	1	0	3.0	0.3333	0.2722	0.6481
8	9	1	0	2.0	0.5000	0.3536	0.4321
9	10	0	0	1.0	0	0	0.2160
10	11	0	1	0.5	0	0	0.2160
11	.	0	0	0.0	0	0	0.2160

Hazard Functions – SAS

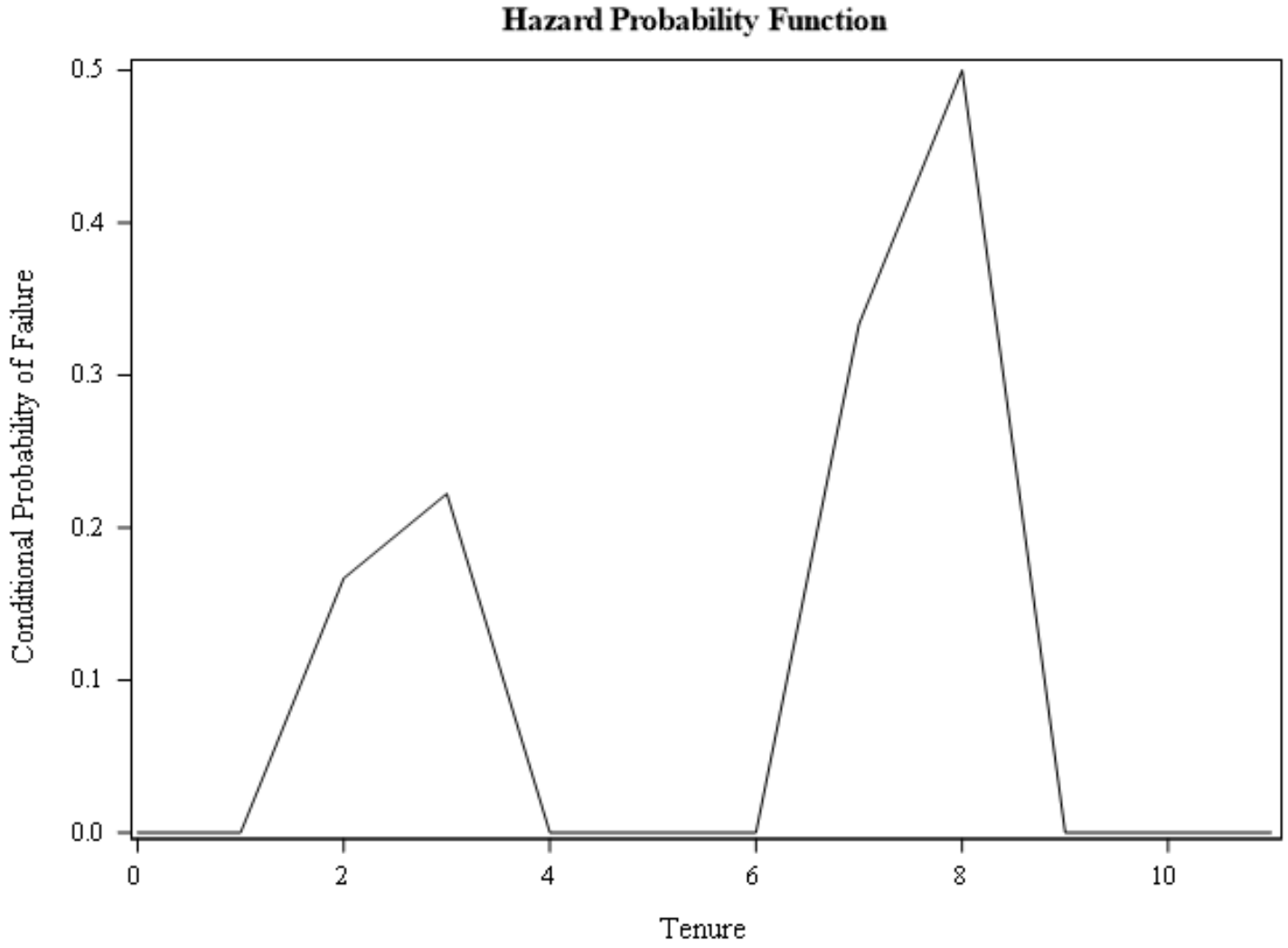
The LIFETEST Procedure

Life Table Survival Estimates							
Failure	Survival Standard Error	Median Residual Lifetime	Median Standard Error	Evaluated at the Midpoint of the Interval			
				PDF	PDF Standard Error	Hazard	Hazard Standard Error
0	0	7.6857	0.9448	0	.	0	.
0	0	6.6857	0.9448	0	.	0	.
0	0	5.6857	0.9448	0.1667	0.1521	0.181818	0.181065
0.1667	0.1521	5.0714	0.9091	0.1852	0.1668	0.25	0.248039
0.3519	0.2017	4.5000	0.8660	0	.	0	.
0.3519	0.2017	3.5000	0.8660	0	.	0	.
0.3519	0.2017	2.5000	0.8660	0	.	0	.
0.3519	0.2017	1.5000	0.8660	0.2160	0.1888	0.4	0.391918
0.5679	0.2218	1.0000	0.7071	0.2160	0.1888	0.666667	0.628539
0.7840	0.1888	.	.	0	.	0	.
0.7840	0.1888	.	.	0	.	0	.
0.7840	0.1888	.	.	.	.	.	.

Hazard Functions – SAS



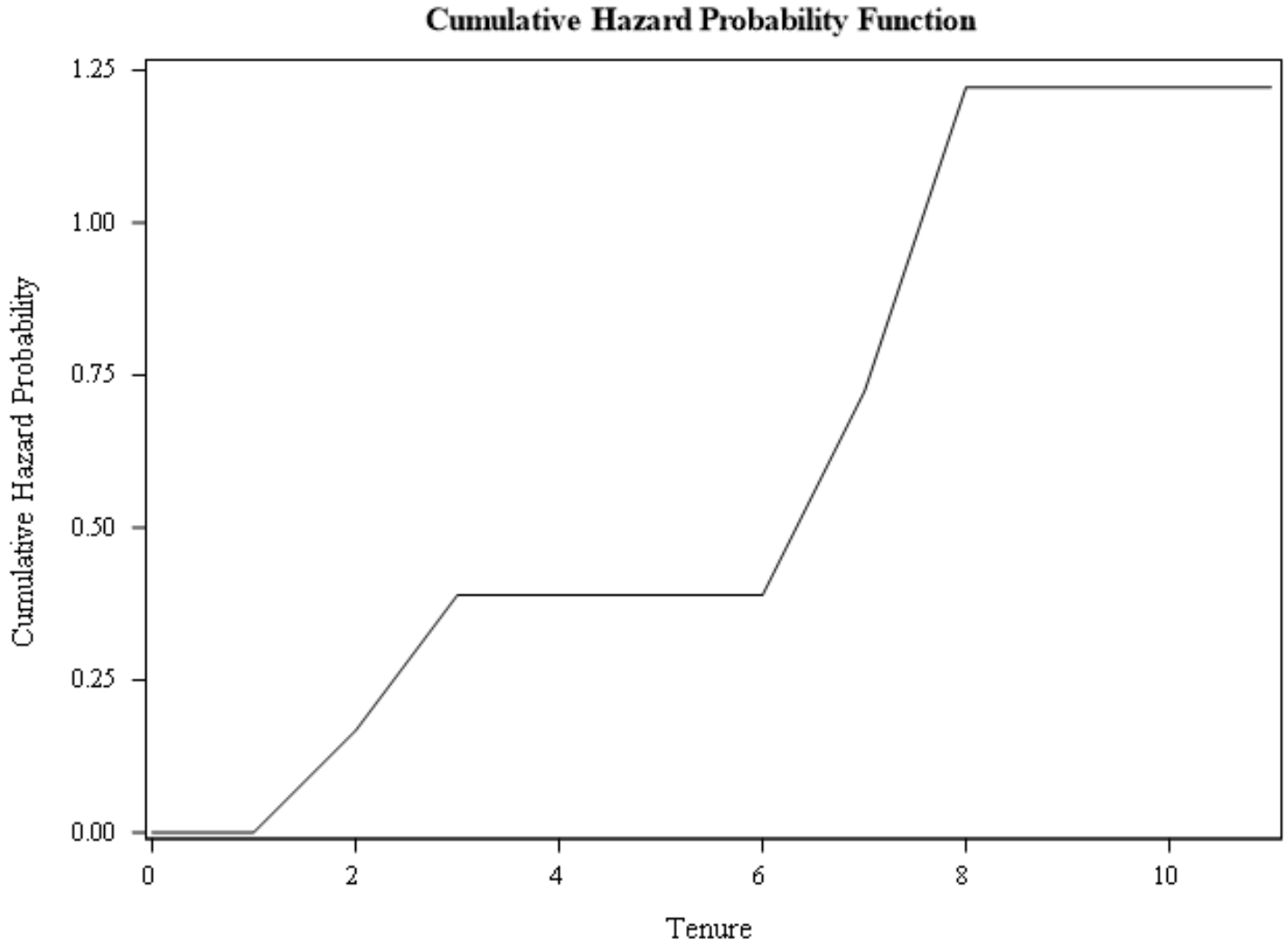
Hazard Functions – SAS



Hazard Functions – SAS

```
data condprob;
    set condprob;
    retain cum_sum;
    cum_sum + condprobfail;
run;

proc sgplot data = condprob;
    series x = lowertime y = cum_sum;
    xaxis label='Tenure';
    yaxis label='Cumulative Hazard Probability';
    title 'Cumulative Hazard Probability Function';
run;
quit;
```

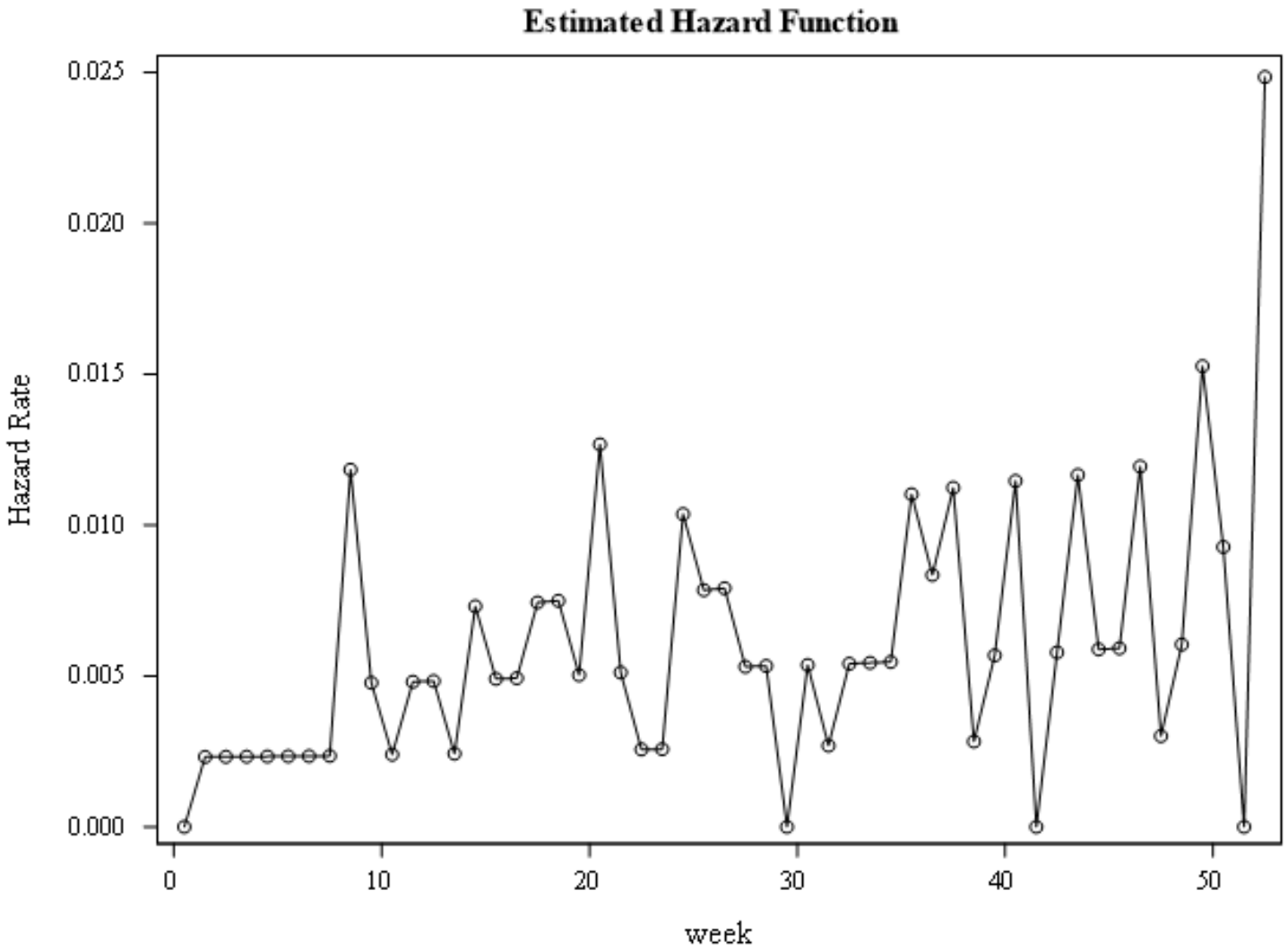
Hazard Functions – SAS



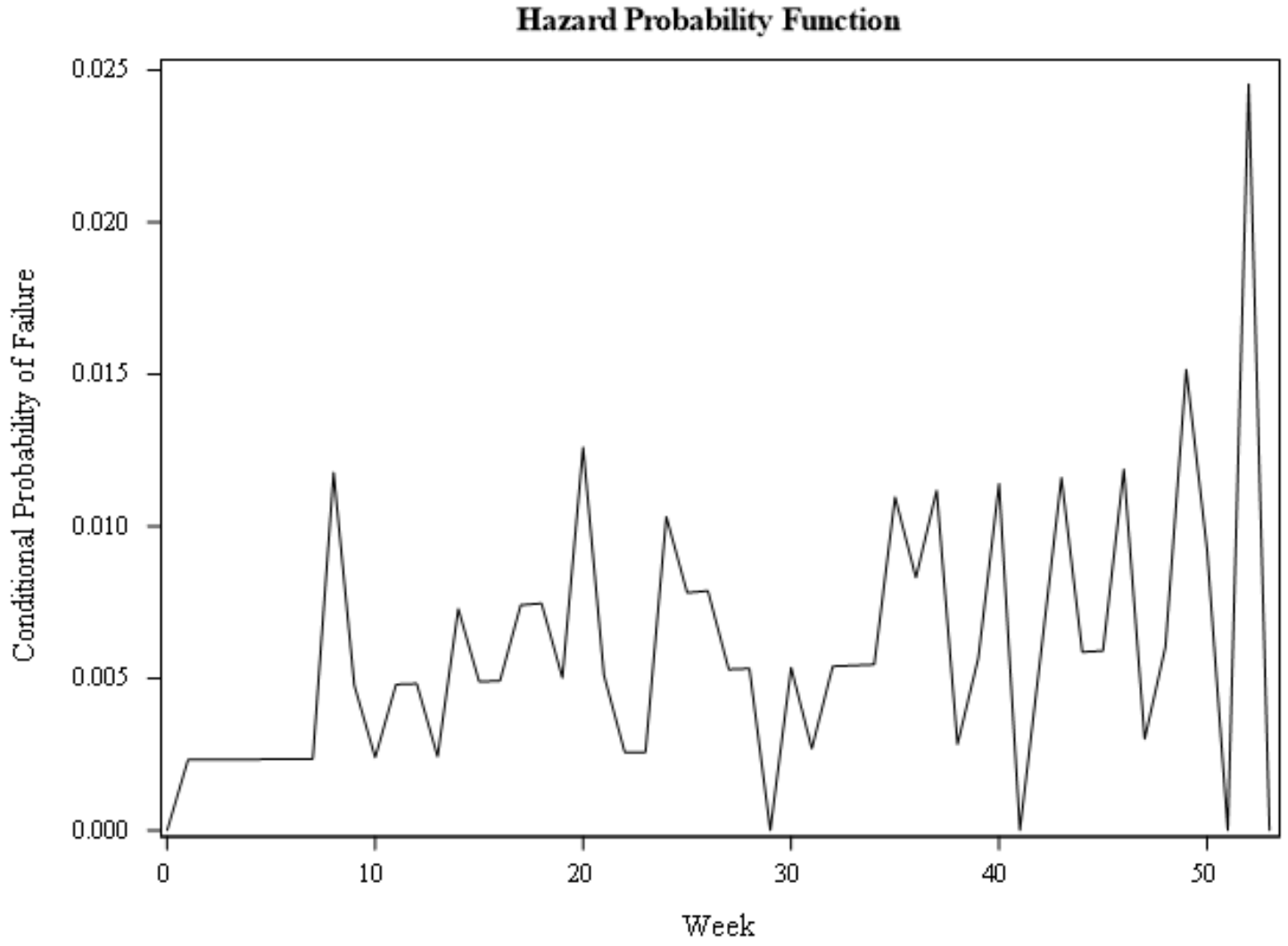
Hazard Functions – SAS

```
proc lifetest data = Survival.Recid method = life
              width = 1 plots=hazard;
   time week*arrest(0);
   ods output LifetableEstimates = condprob;
run;
```

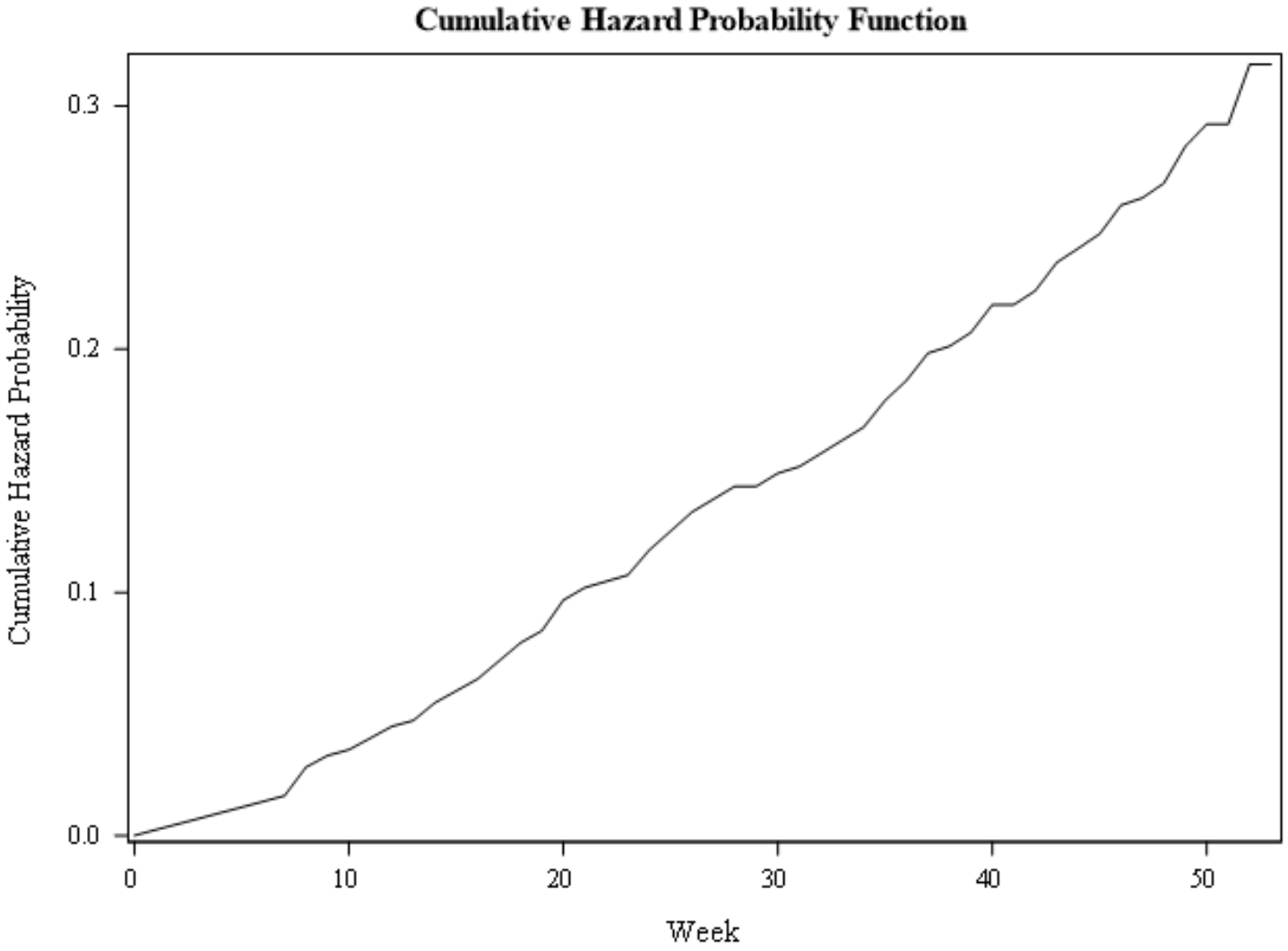
Hazard Functions – SAS



Hazard Functions – SAS



Hazard Functions – SAS



Hazard Functions – R

```
summary(simple_km)
```

```
## Call: survfit(formula = Surv(time = tenure, event = (censored == 0))
~
##     1, data = simple)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     2      6       1    0.833   0.152       0.5827            1
##     3      5       1    0.667   0.192       0.3786            1
##     7      3       1    0.444   0.222       0.1668            1
##     8      2       1    0.222   0.192       0.0407            1
```

```
simple_km$hp <- simple_km$n.event/simple_km$n.risk
print(simple_km$hp)
```

```
## [1] 0.1666667 0.2000000 0.3333333 0.5000000 0.0000000
```

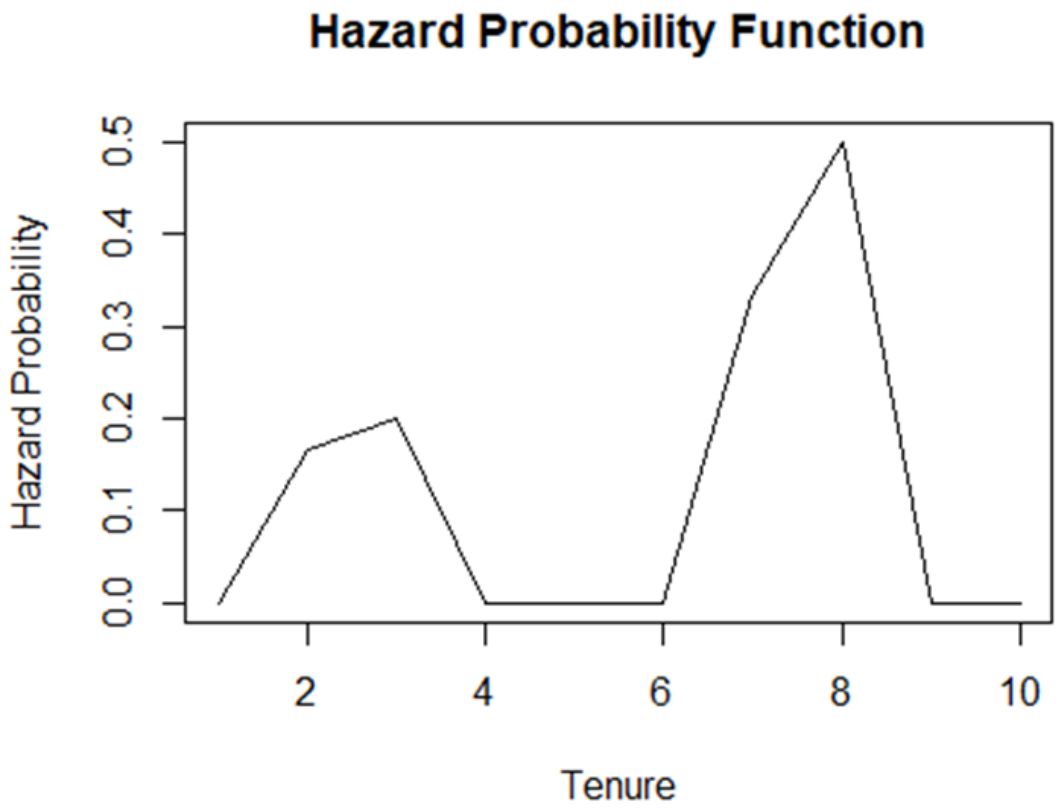
Hazard Functions – R

```
simple_haz <- merge(data.frame(time = seq(1,10,1)),
                    data.frame(time = simple_km$time, hp = simple_km$hp),
                    by = "time", all = TRUE)
simple_haz[is.na(simple_haz) == TRUE] <- 0
print(simple_haz)
```

```
##    time        hp
## 1     1 0.0000000
## 2     2 0.1666667
## 3     3 0.2000000
## 4     4 0.0000000
## 5     5 0.0000000
## 6     6 0.0000000
## 7     7 0.3333333
## 8     8 0.5000000
## 9     9 0.0000000
## 10   10 0.0000000
```

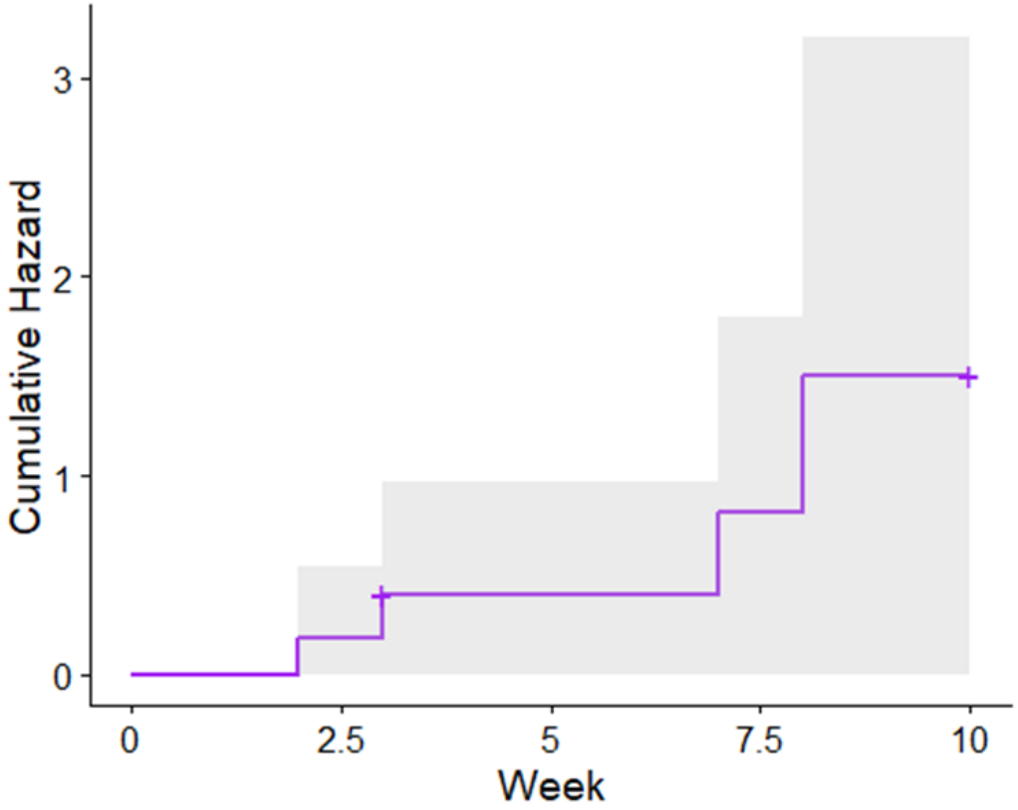
Hazard Functions – R

```
plot(y = simple_haz$hp, x = simple_haz$time,
     main = "Hazard Probability Function", xlab = "Tenure",
     ylab = "Hazard Probability", type = 'l')
```



Hazard Functions – R

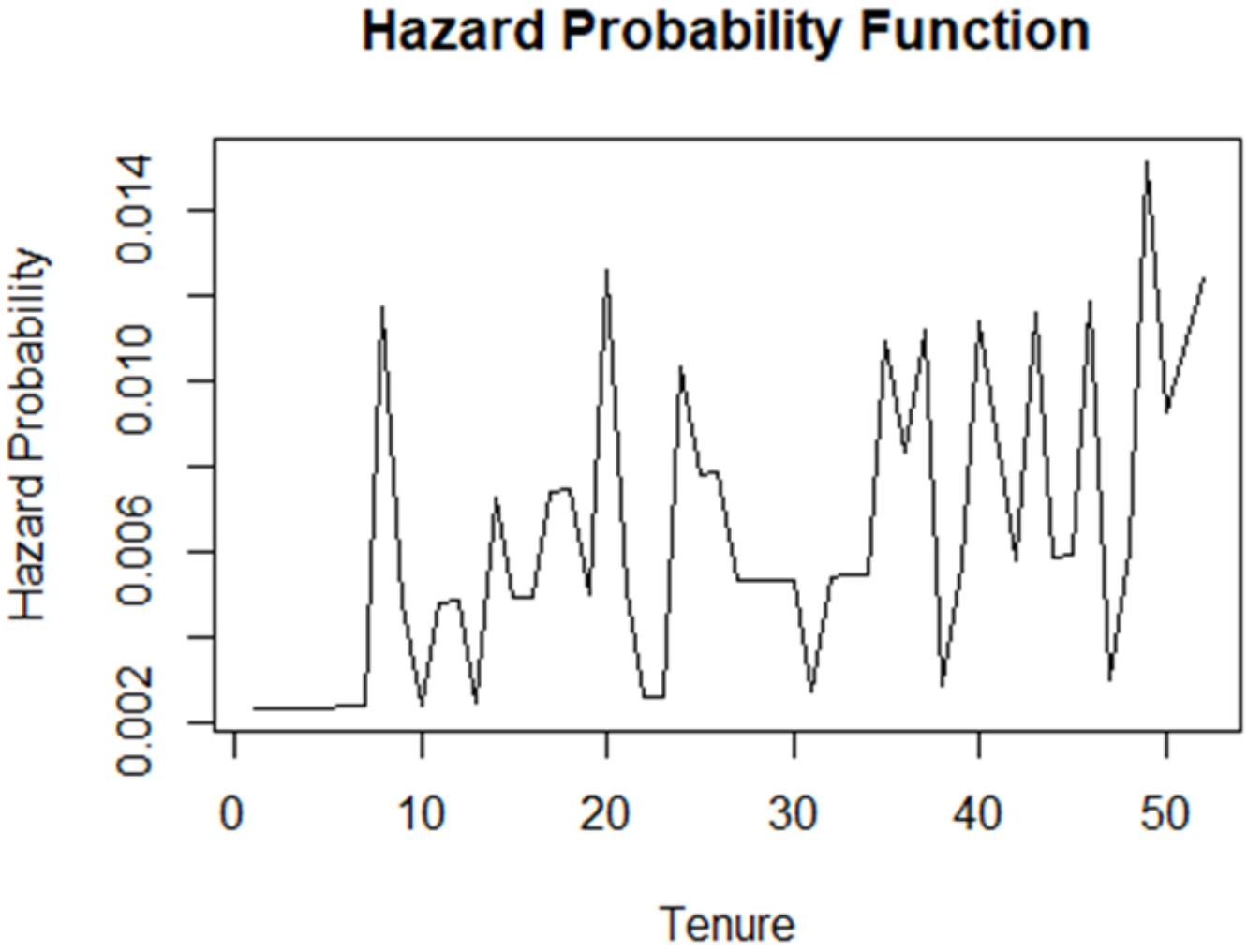
```
ggsurvplot(simple_km, data = simple, fun = "cumhaz", conf.int = TRUE,
           palette = "purple", xlab = "Week",
           ylab = "Cumulative Hazard", legend = "none")
```



Hazard Functions – R

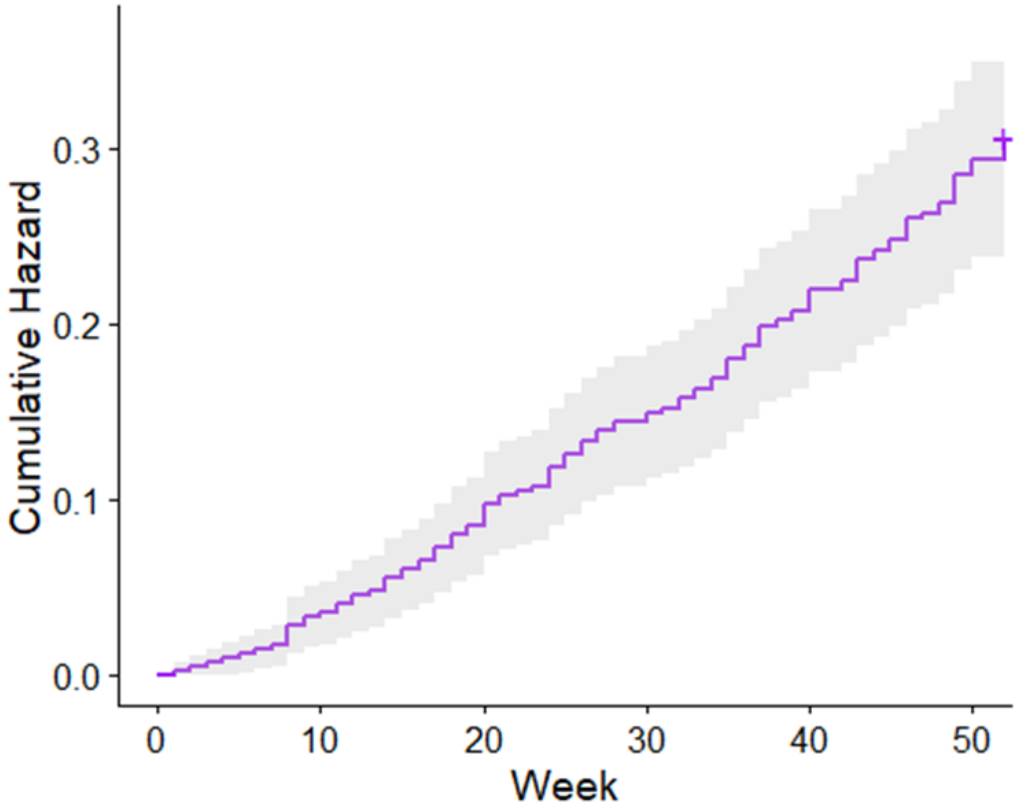
```
recid_km$hp <- recid_km$n.event/recid_km$n.risk
recid_haz <- merge(data.frame(time = seq(1,10,1)),
                   data.frame(time = recid_km$time, hp = recid_km$hp),
                   by = "time", all = TRUE)
recid_haz[is.na(recid_haz) == TRUE] <- 0

plot(y = recid_haz$hp, x = recid_haz$time,
     main = "Hazard Probability Function", xlab = "Tenure",
     ylab = "Hazard Probability", type = 'l')
```

Hazard Functions – R



Hazard Functions – R

```
ggsurvplot(recid_km, data = simple, fun = "cumhaz", conf.int = TRUE,
           palette = "purple", xlab = "Week",
           ylab = "Cumulative Hazard", legend = "none")
```



Survival and Hazard Relationship

- The survival, hazard, and cumulative hazard functions are all directly related:
  - $\Lambda(t) = -\log S(t)$
  - $S(t) = e^{-\Lambda(t)}$
  - $h(t) = -\frac{d}{dt}\log S(t) = \frac{f(t)}{S(t)}$

- These three quantities are all different ways of describing the same distribution; if you know one of them, you can compute the others.