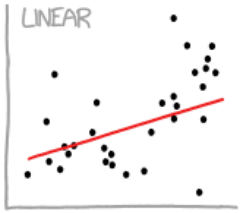
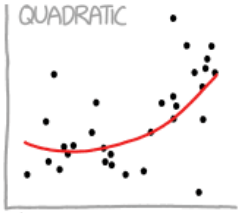
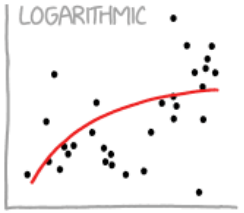
CURVE-FITTING METHODS
AND THE MESSAGES THEY SEND



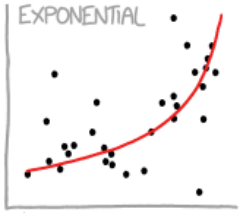
"HEY, I DID A REGRESSION."



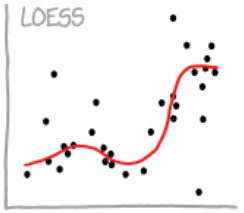
"I WANTED A CURVED LINE, SO I MADE ONE WITH MATH."



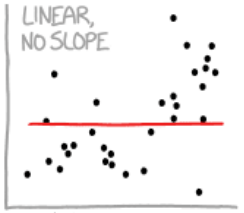
"LOOK, IT'S TAPERING OFF!"



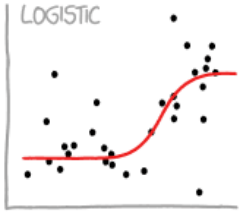
"LOOK, IT'S GROWING UNCONTROLLABLY!"



"I'M SOPHISTICATED, NOT LIKE THOSE BUMBLING POLYNOMIAL PEOPLE."

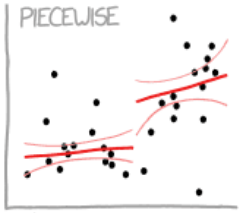


"I'M MAKING A SCATTER PLOT BUT I DON'T WANT TO."

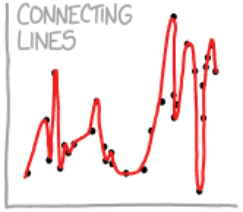


"I NEED TO CONNECT THESE TWO LINES, BUT MY FIRST IDEA DIDN'T HAVE ENOUGH MATH."

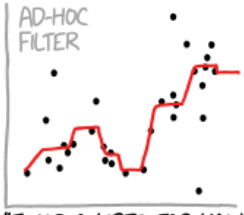
"LISTEN, SCIENCE IS HARD. BUT I'M A SERIOUS PERSON DOING MY BEST."



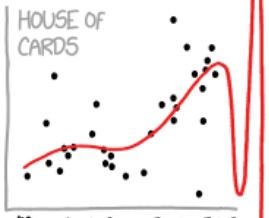
"I HAVE A THEORY, AND THIS IS THE ONLY DATA I COULD FIND."



"I CLICKED 'SMOOTH LINES' IN EXCEL."



"I HAD AN IDEA FOR HOW TO CLEAN UP THE DATA. WHAT DO YOU THINK?"



"AS YOU CAN SEE, THIS MODEL SMOOTHLY FITS THE– *WAIT NO NO DON'T EXTEND IT AAAAAA!!*"

Censoring, Survival, & Hazards

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
- Reliability 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).

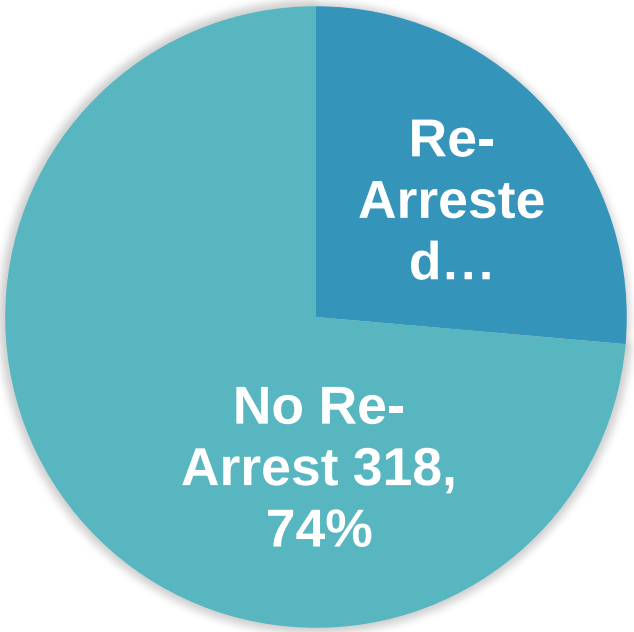
Other problems with OLS:

- 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)
- **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

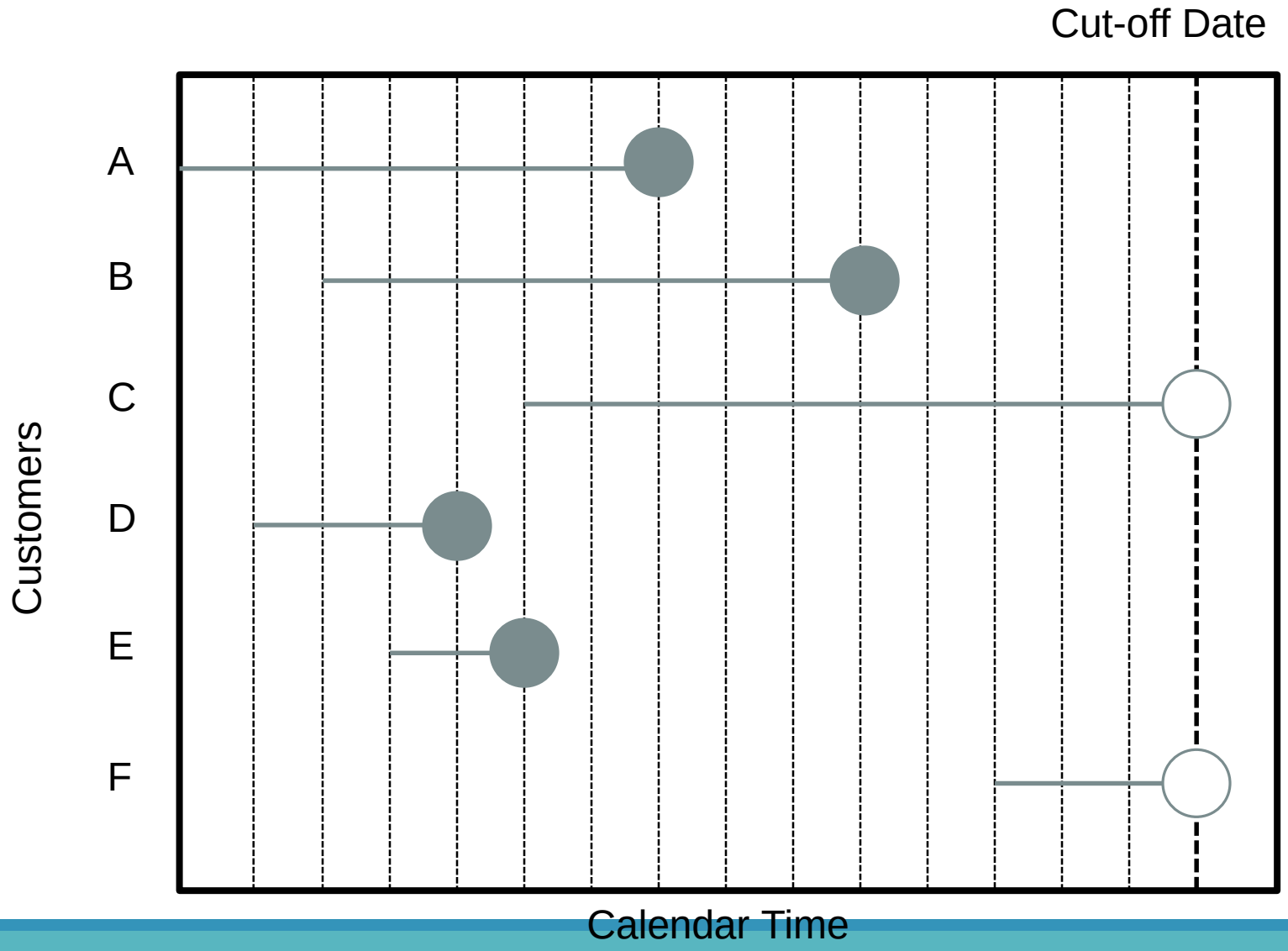
Just like most models, survival analysis depends on certain assumptions.

When Does Time Start?

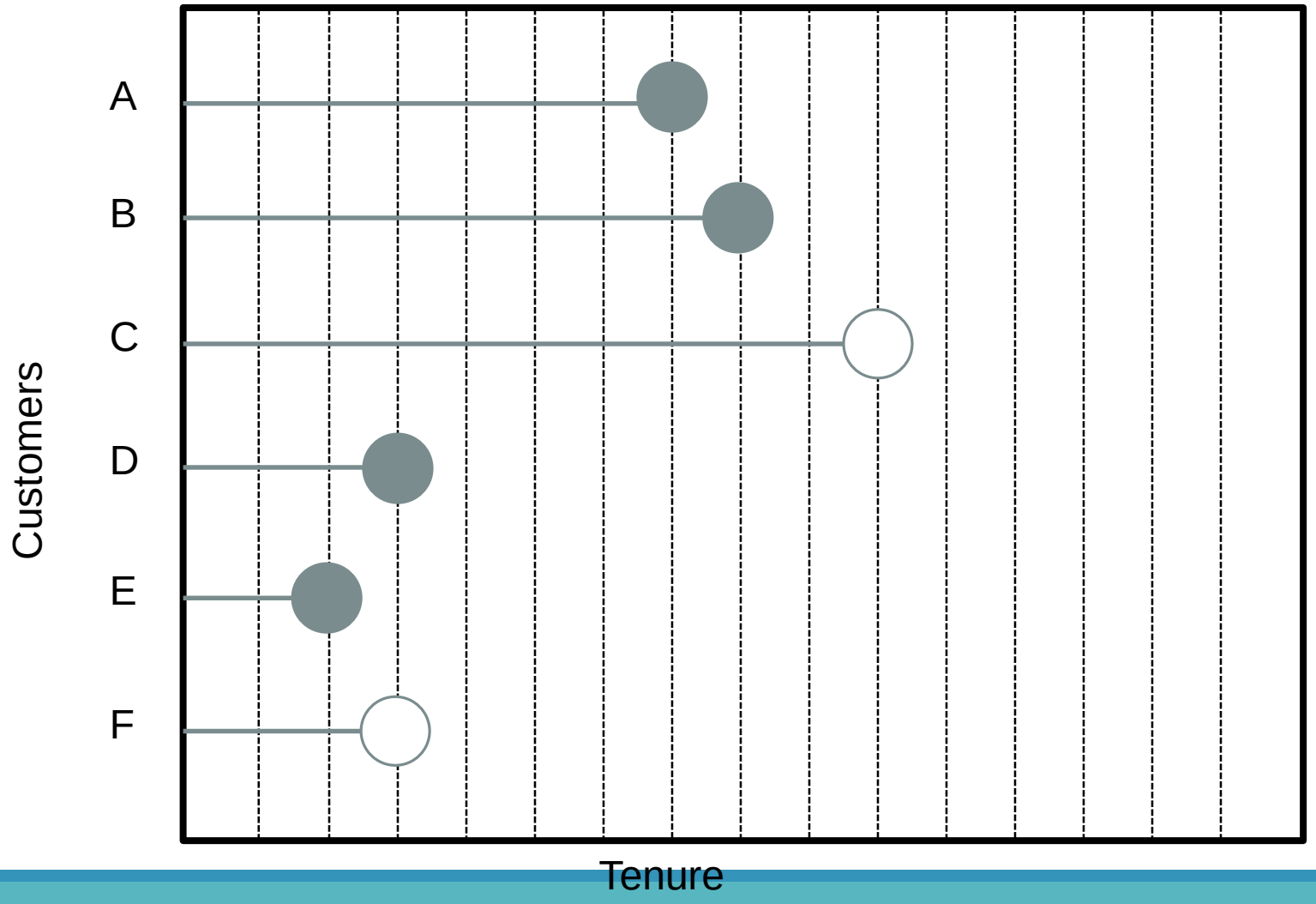
Create an artificial world in which everyone “starts” at the same time.

- 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, Left and Interval Censoring

If an observation is **right censored**, then $T > c$. This is what normally happens.

Example:

- Clinical trial ends, and patient is still alive.

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.

Right, Left and Interval Censoring

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.*

Assumption for Right Censoring

Censoring is NONINFORMATIVE

- This means that patients who are censored should have the same future risk for the occurrence of the event of interest, conditional on exposure, as those who continue to be followed (“assumption of independent censoring”)

Survival 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 used to be the only method in survival analysis(1958: Edward L. Kaplan and Paul Meier).

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*.

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

→ # events occurring at time *t*

→ # 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*.

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

→ # events occurring at time *t*

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

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 ($d_0 = 0$):

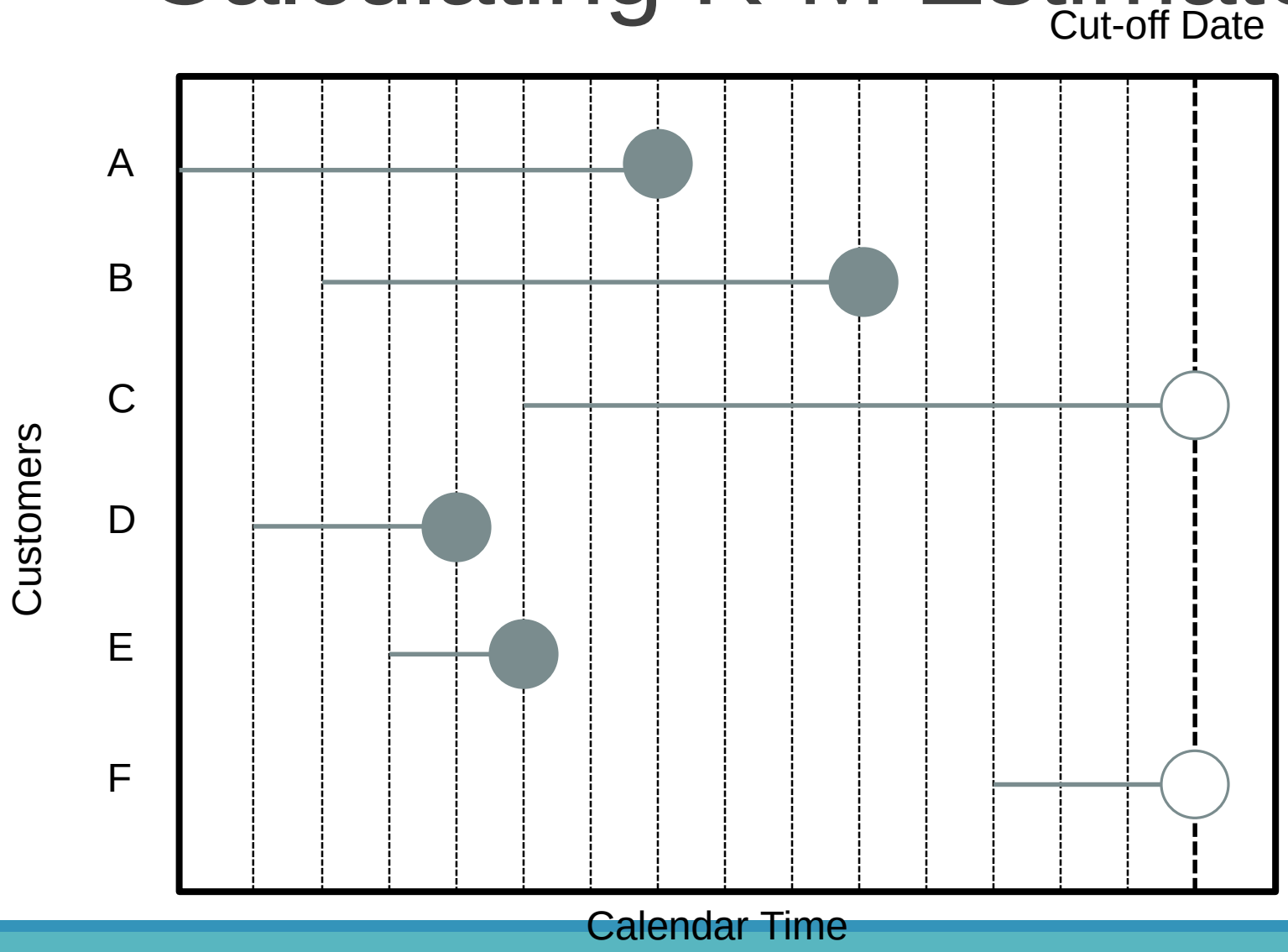
$$\hat{S}(t) = \prod_{k \le t} \left(1 - \frac{d_k}{r_k}\right) = \left(1 - \frac{0}{n}\right) = 1$$

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:

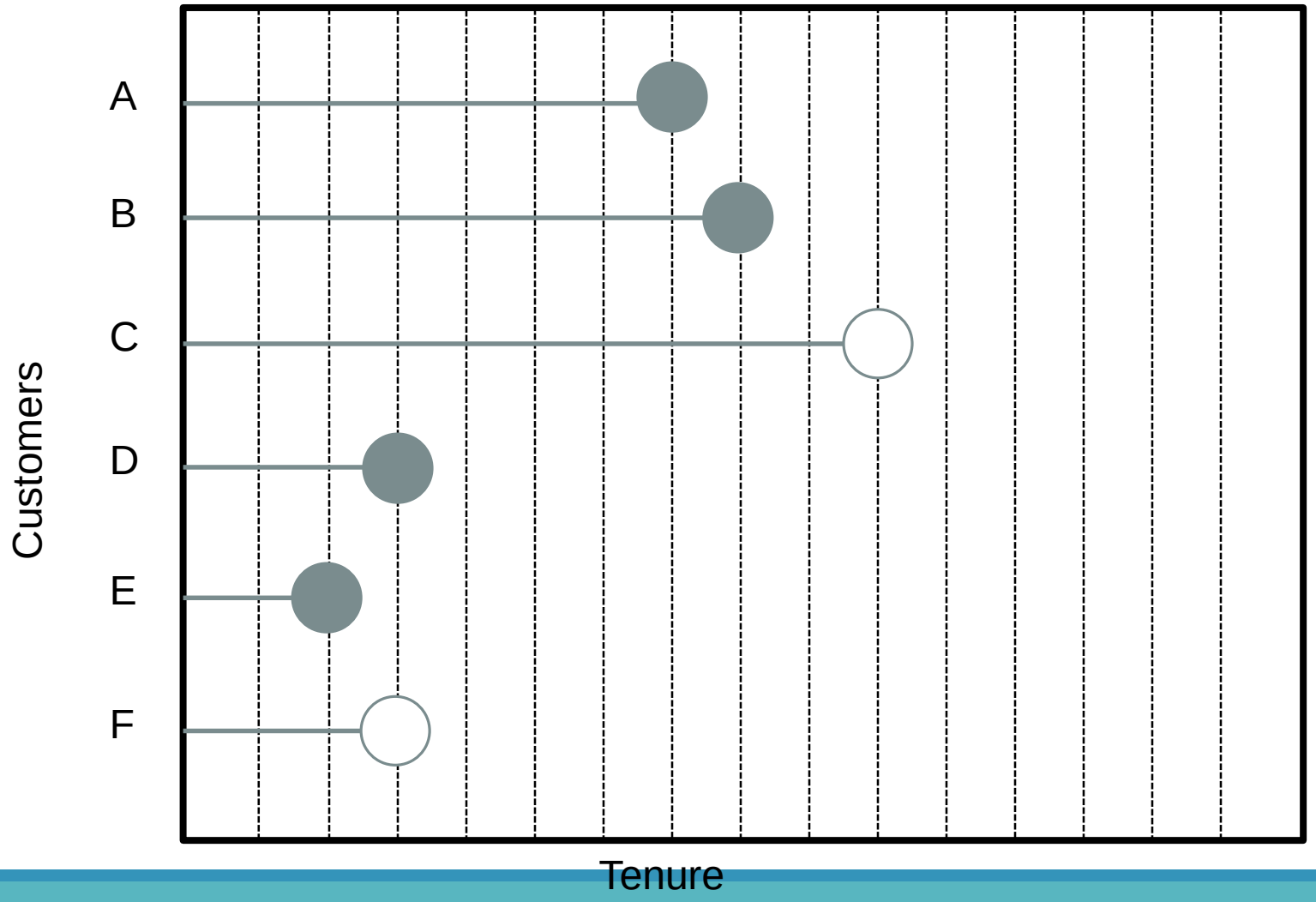
$$\hat{S}(1) = S(0) \times \left(1 - \frac{d_1}{r_1}\right)$$

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

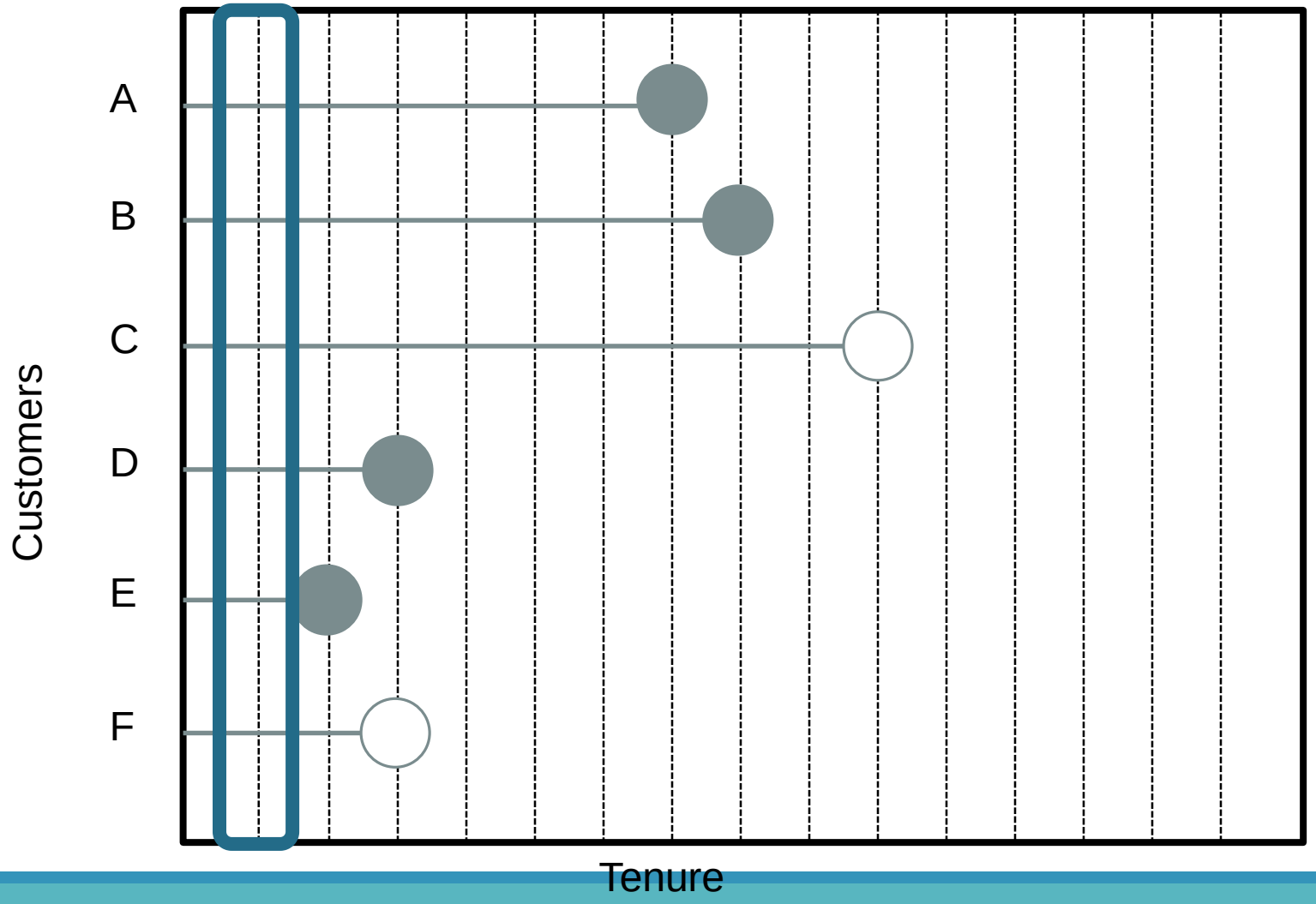
Calculating K-M Estimate



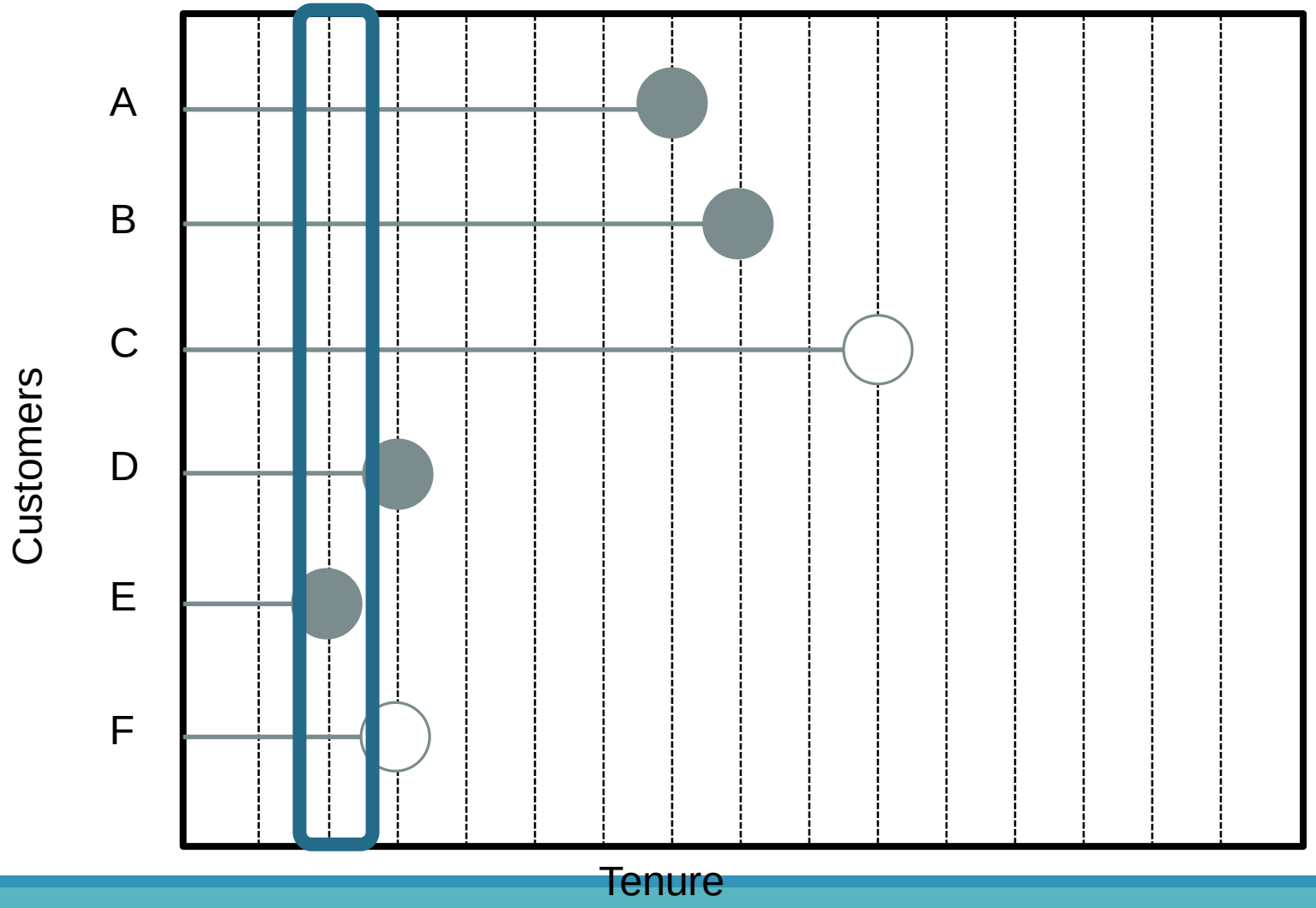
Calculating K-M Estimate



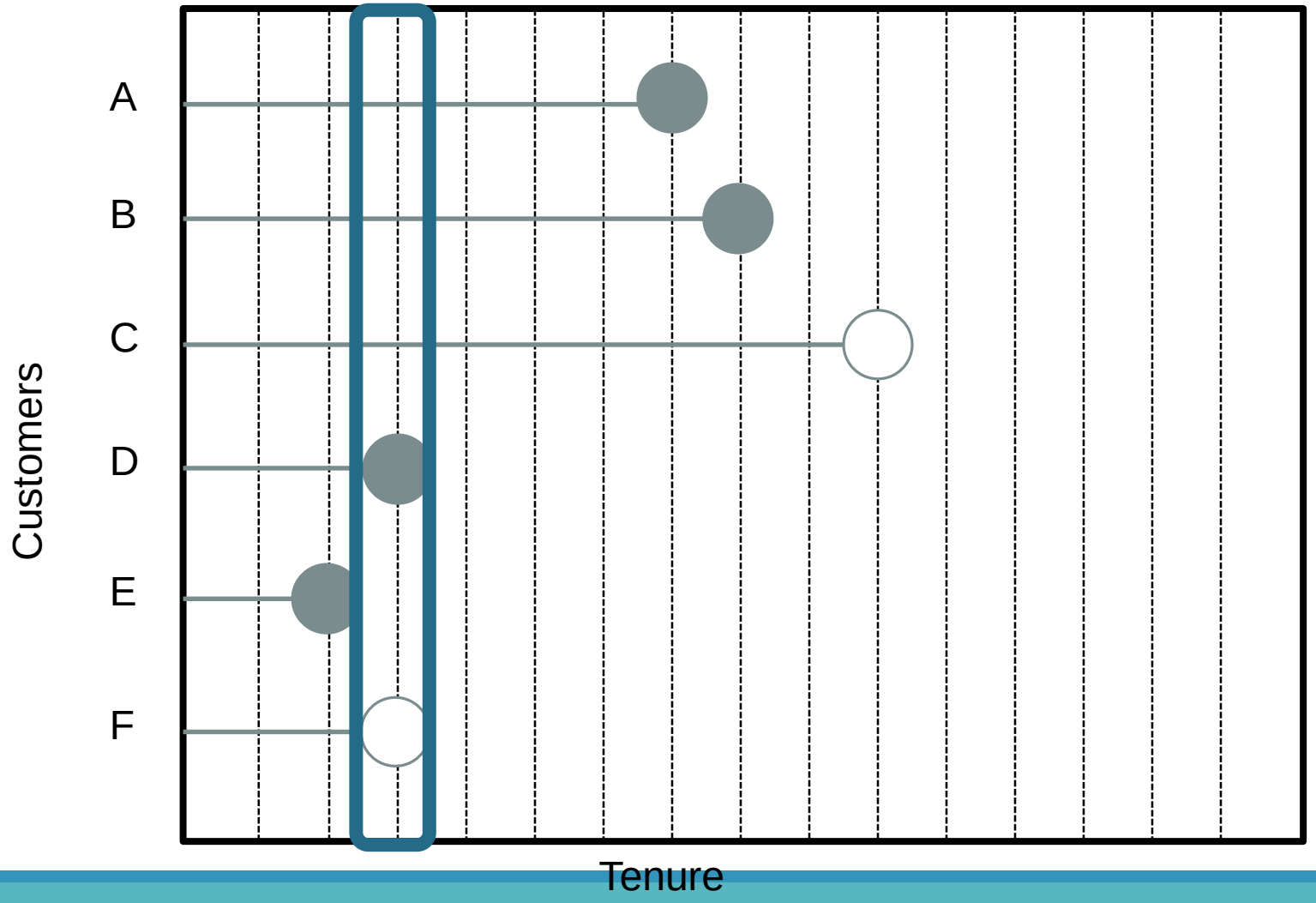
Calculating K-M Estimate



Calculating K-M Estimate



Calculating K-M Estimate



Calculating K-M Estimate

Time = 0:

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

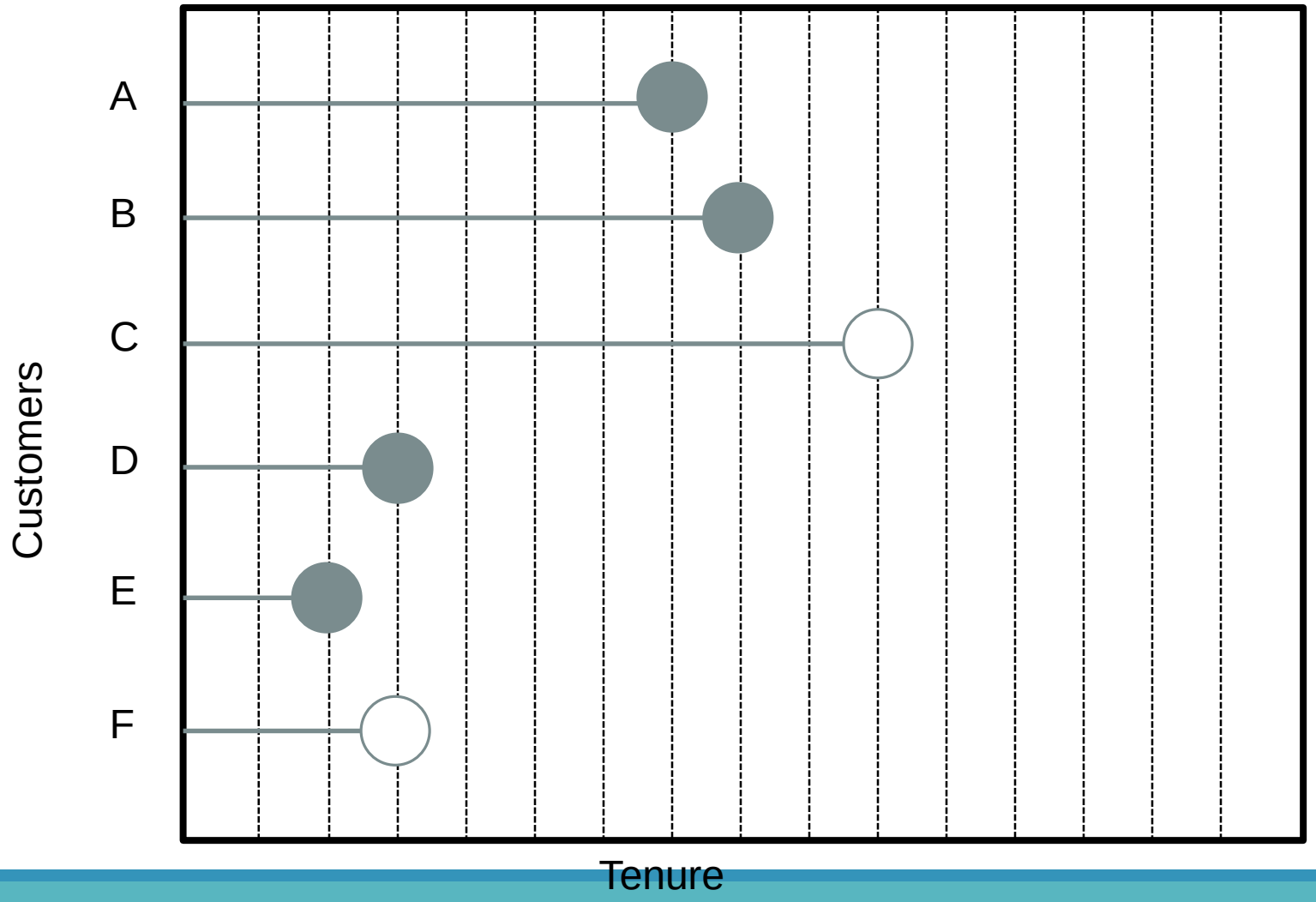
Time = 1:

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

Time = 2:

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

Calculating K-M Estimate



Calculating K-M Estimate

Time = 3:

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

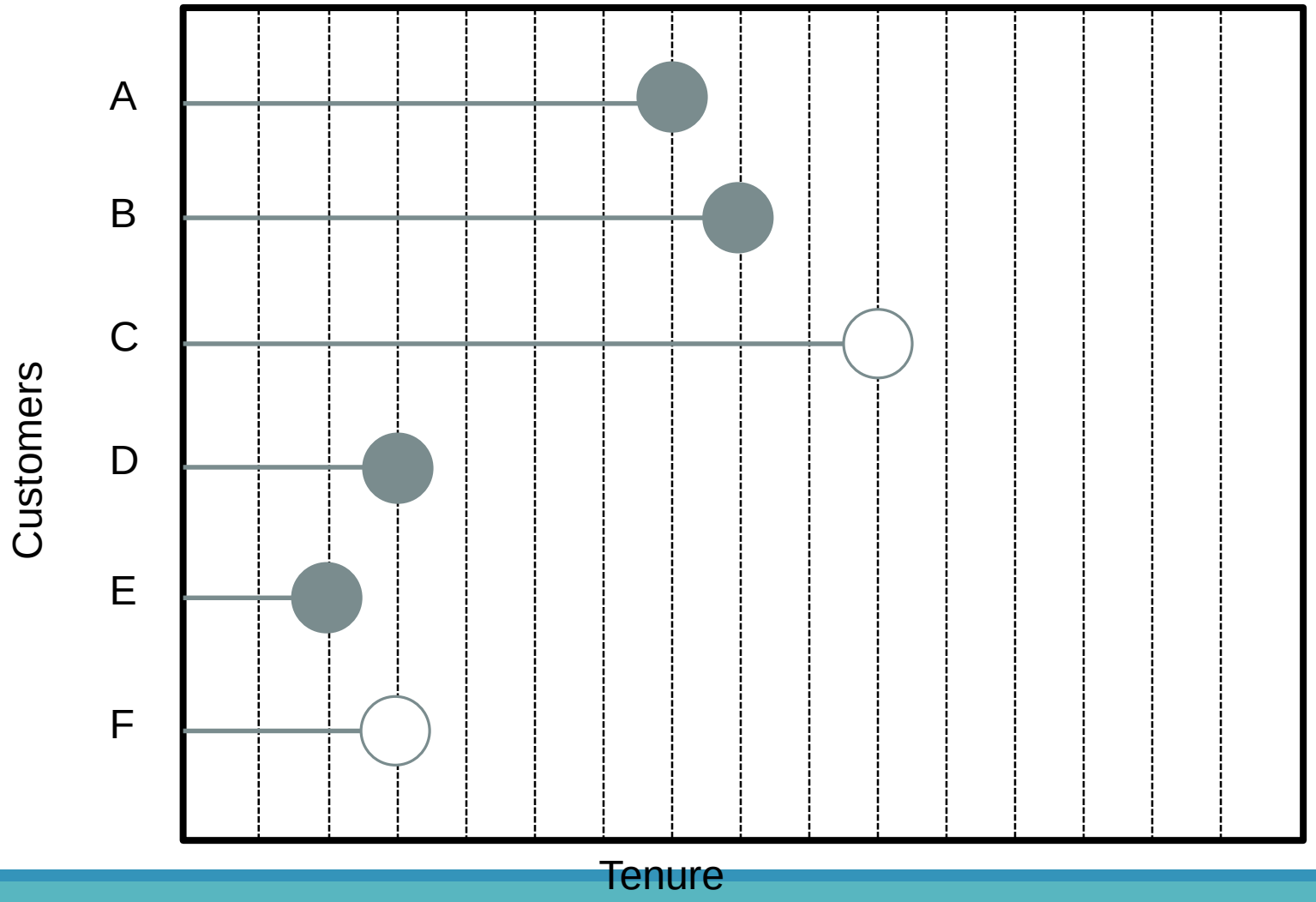
Time = 4:

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

Time = 5:

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

Calculating K-M Estimate



Calculating K-M Estimate

Time = 6:

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

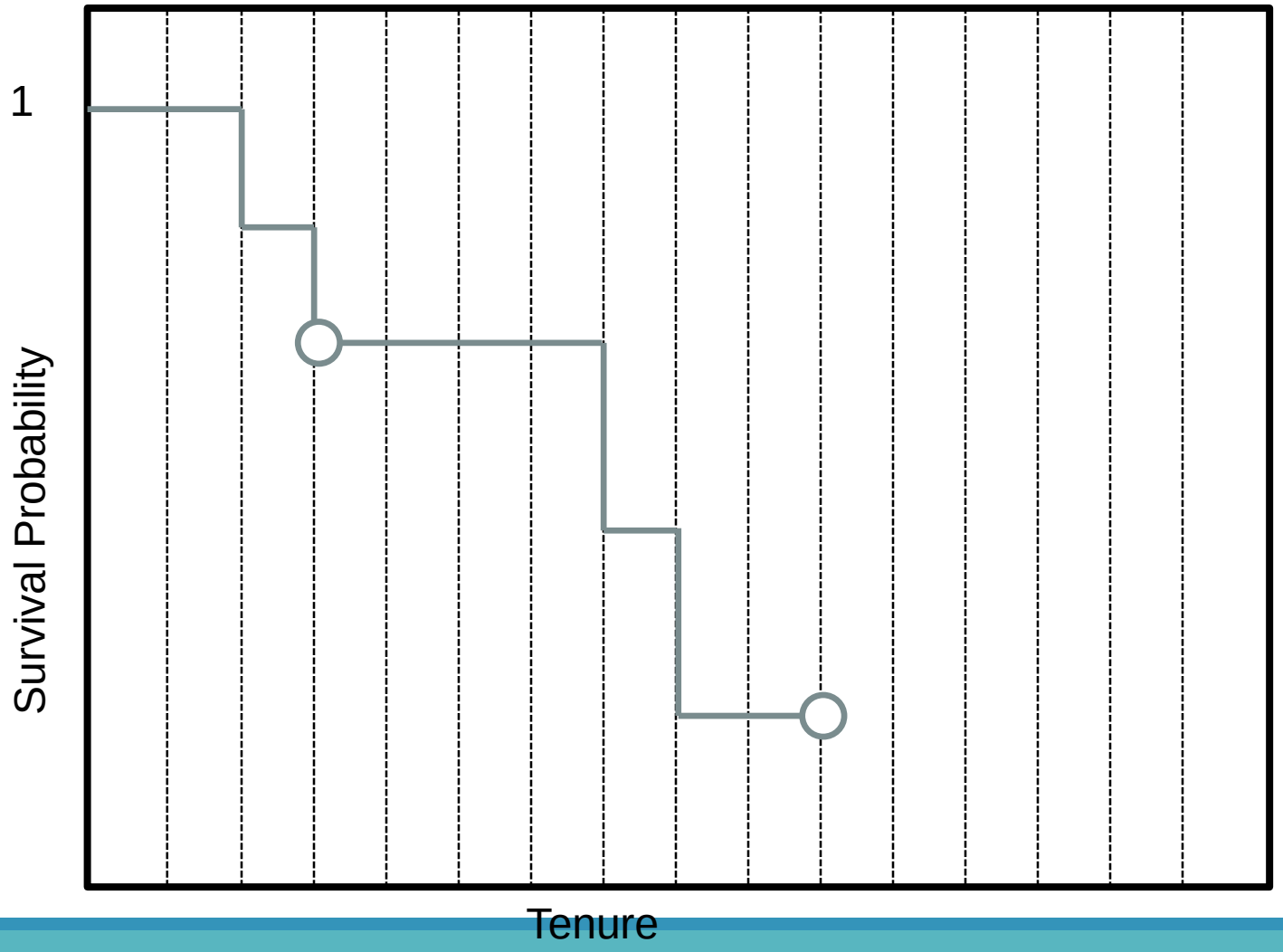
Time = 7:

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

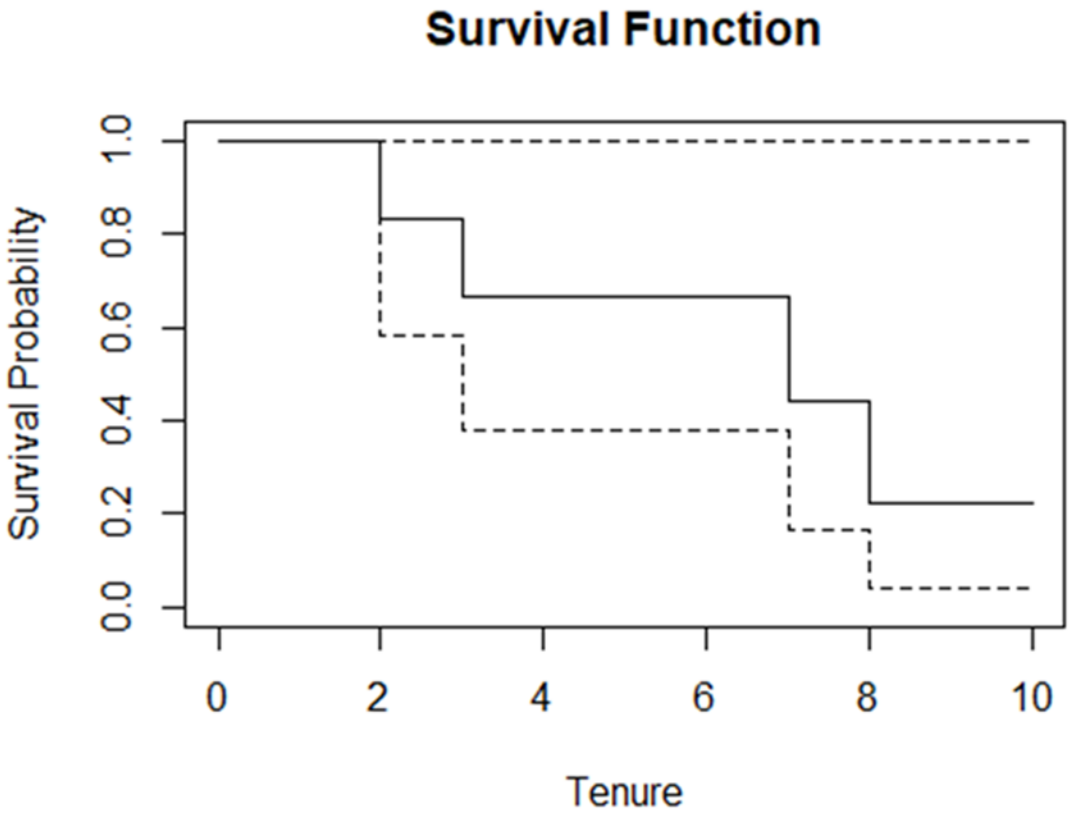
Time = 8:

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

Visualizing K-M Estimate



Visualizing K-M Estimate



Summary Statistics

Due to censoring, the mean is difficult to 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 – R

Data set:

tenure	censored
7	1
8	1
10	0
3	1
2	1
3	0

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

simple.s
[1]  7   8  10+  3   2   3+
```

Kaplan-Meier

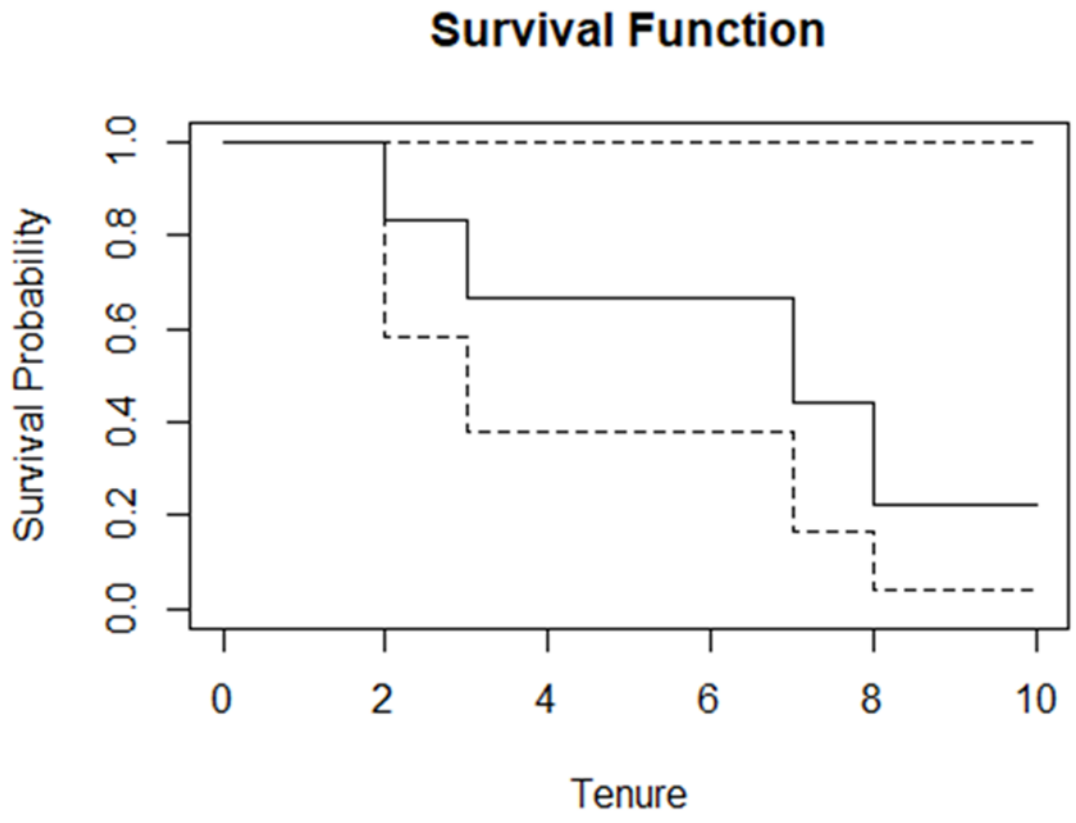
```
simple_km=survfit(Surv(time = tenure, event = censored)~1,
data = simple)

summary(simple_km)
```

time	n.risk	n.event	survival	std.err	lower 95% CI
2	6	1	0.833	0.152	0.5827
3	5	1	0.667	0.192	0.3786
7	3	1	0.444	0.222	0.1668
8	2	1	0.222	0.192	0.0407

Survival Curve

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



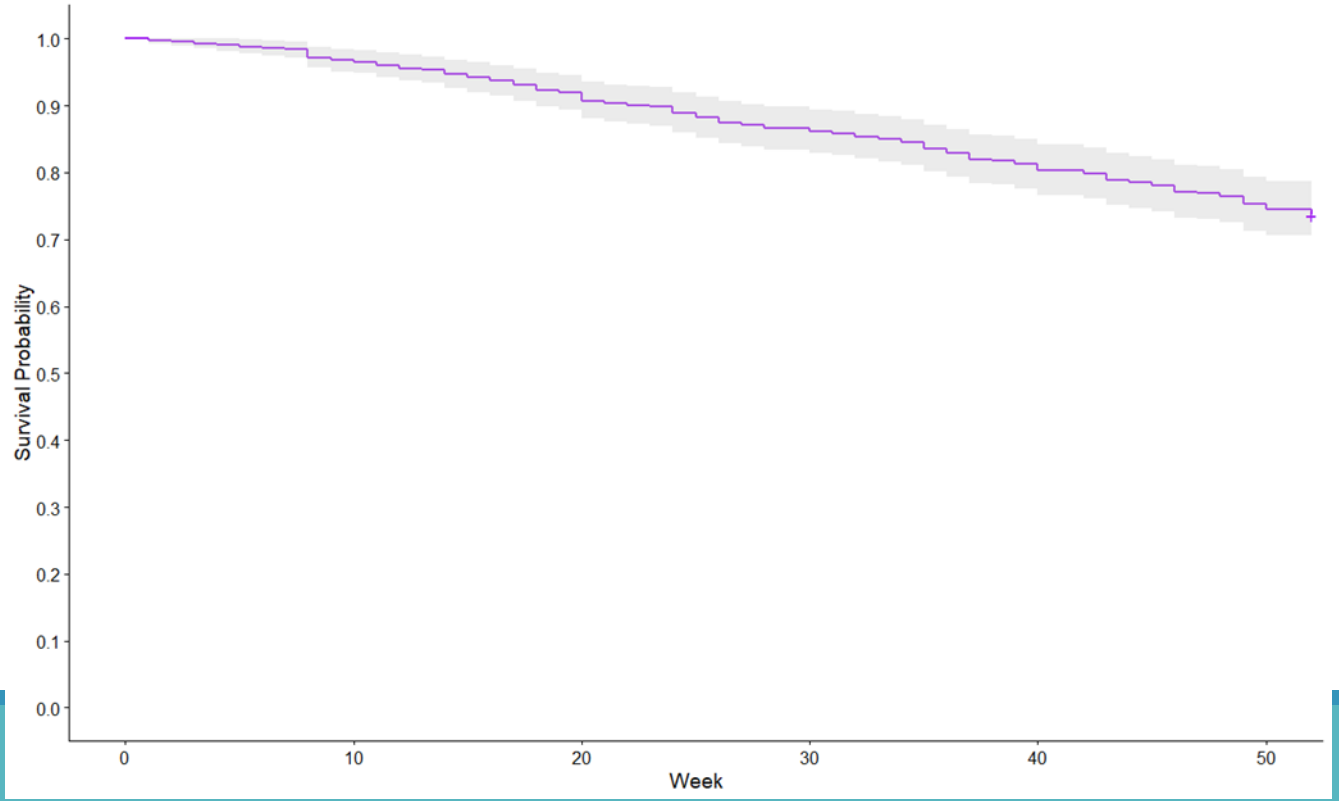
K-M for Recid

```
Recid.fit = survfit(Surv(time = week, event = arrest) ~ 1,
data = recid)
summary(recid.fit)
```

```
##  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
                                    ⋮
##    52    322       4    0.736 0.02121        0.696        0.779
```

Survival Function – R

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

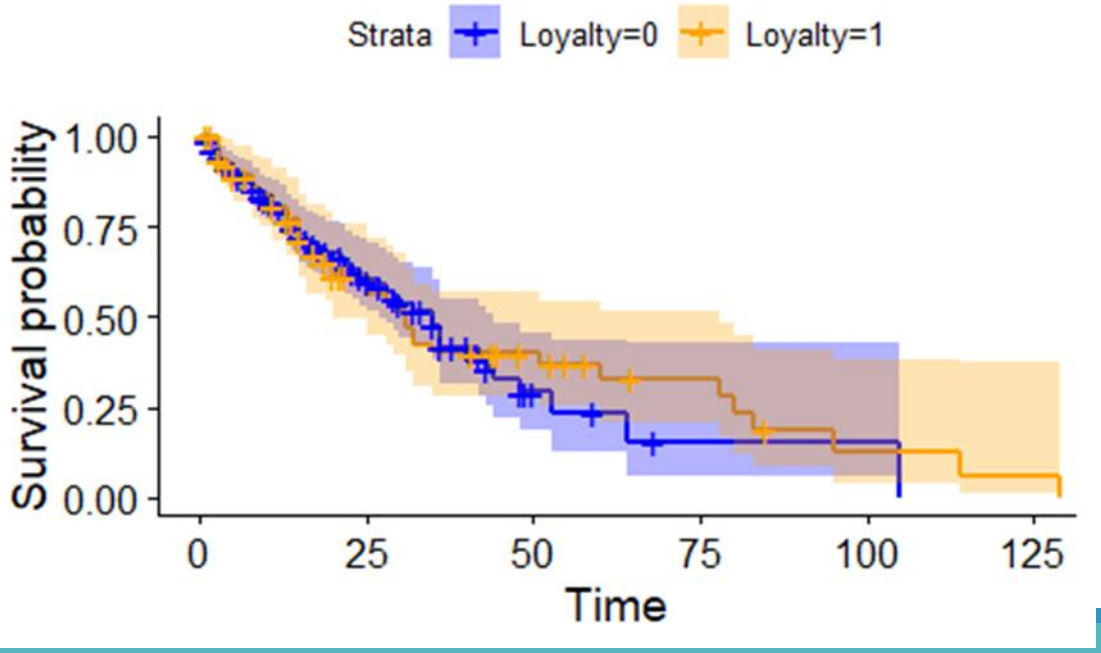


Stratified Analysis

Stratified Analysis

Can also create separate/stratified curves by group.

Different curves result in different estimates for each group.



Stratified Analysis

R provides 2 tests that each have the same null hypothesis – all survival curves are **equal** (alternative is that at least one curve is different).

1. Log-rank test (developed by Mantel-Haenszel)
2. Wilcoxon test

Comparing Survival Function

Log-Rank test:

For each group, calculate expected events and compare to observe events ($(O-E)^2/E$). This is a χ^2 statistic with k-1 df (k is the number of groups being compared!). This is the statistic when we set “Rho = 0”)

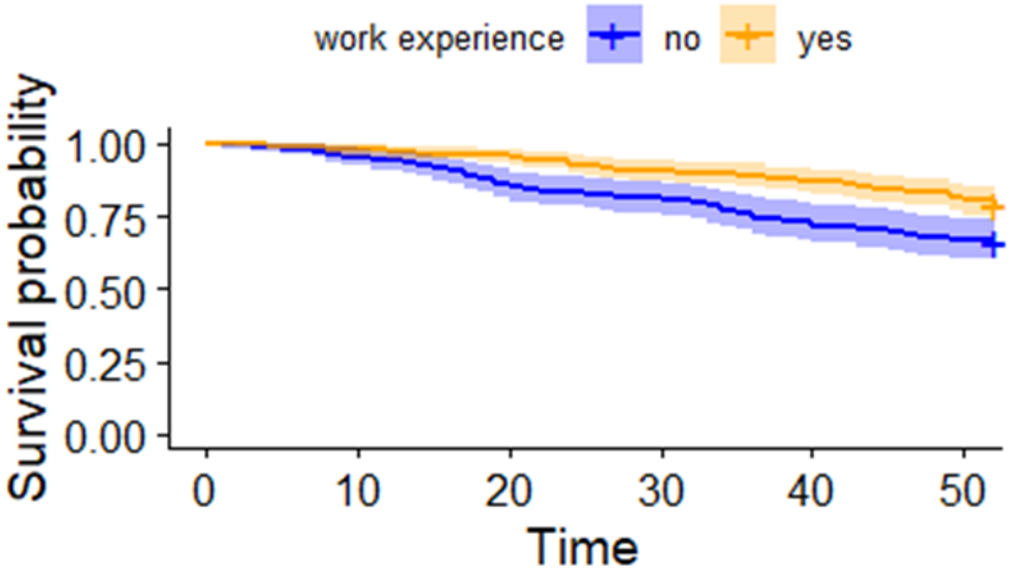
Wilcoxon test (places larger emphasis on earlier event times):

Similar to Log-Rank test except that we now use weights. This is what happens when “Rho = 1”).

Stratified Analysis – R

```
Recid.KP = survfit(Surv(week, arrest) ~ wexp,data=recid)

ggsurvplot(Recid.KP,data=recid,palette = c("blue","orange"),conf.int = T,
legend.title = "work experience", legend.labs = c("no", "yes"))
```

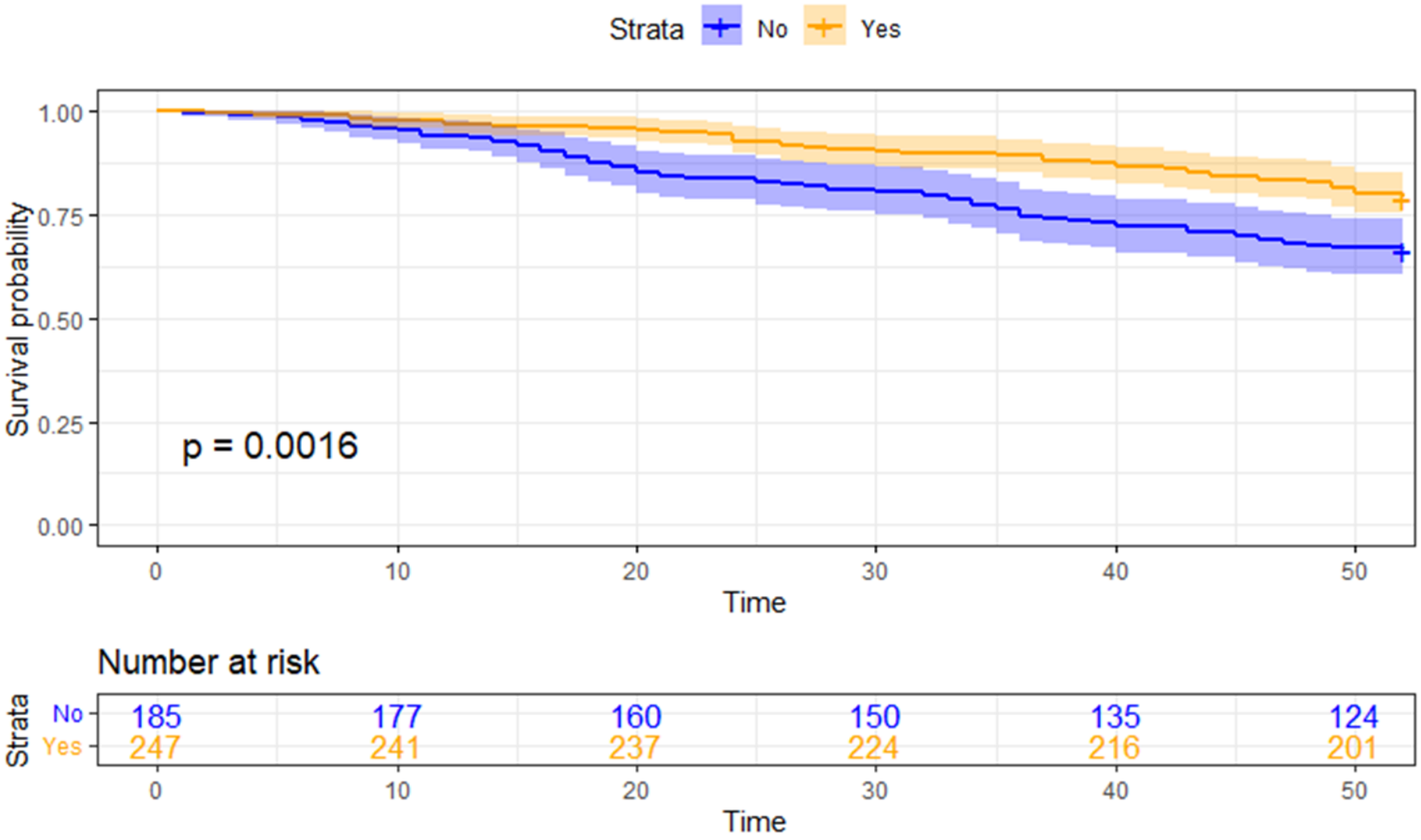


Stratified Analysis – R

```
survdiff(Surv(week, arrest) ~ wexp, data=recid, rho=0)
```

```
##  Chisq= 9.9  on 1 degrees of freedom, p= 0.002
```

```
ggsurvplot(Recid.KP, data = recid, size = 1, palette =c("blue","orange"),
  conf.int = TRUE,  pval = TRUE,  risk.table = TRUE,  risk.table.col =
"wexp", legend.labs =c("No", "Yes"),  risk.table.height = 0.25,
  ggtheme = theme_bw() )
```


Strata
No
Yes
1.00
0.75
0.50
0.25
0.00
Survival probability
p = 0.0016
0
10
20
30
40
50
Time
Number at risk
Strata
No
Yes
185
177
160
150
135
124
247
241
237
224
216
201
0
10
20
30
40
50
Time

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:

$$h(t) = P(T = t | T \geq t)$$

2. Hazard Rates:

$$h(t) = \frac{P(T = t)}{1 - P(T \leq t)} \; or \; \frac{f(t)}{1 - F(t)}$$

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:

$$h(t) = P(T = t | T \geq t)$$

2. Hazard Rates:

$$h(t) = \frac{P(T = t)}{1 - P(T \leq t)} \; or \; \frac{f(t)}{1 - F(t)}$$

Both are denoted the same way in different texts!

Hazard Probabilities

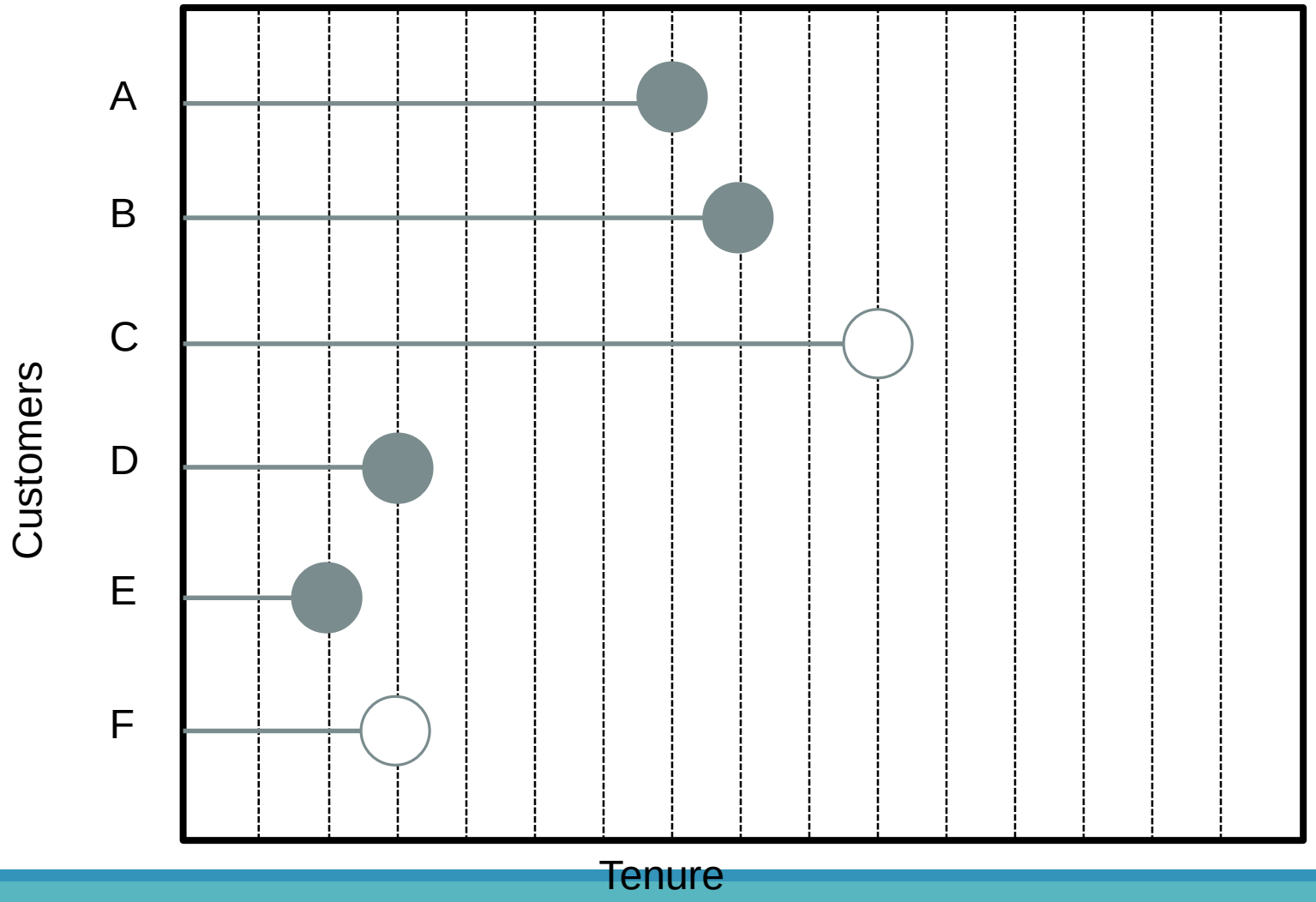
Hazard probabilities are very useful and common in business settings.

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 at time *t* given they lasted until time *t*?

$$h(t) = P(T = t | T \geq t) = \frac{d_t}{r_t}$$

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:

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

OR

$$h(3) = \frac{1}{4.5} = 0.222$$

Time = 4:

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

Time = 5:

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

Calculating Hazard Probabilities

Time = 6:

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

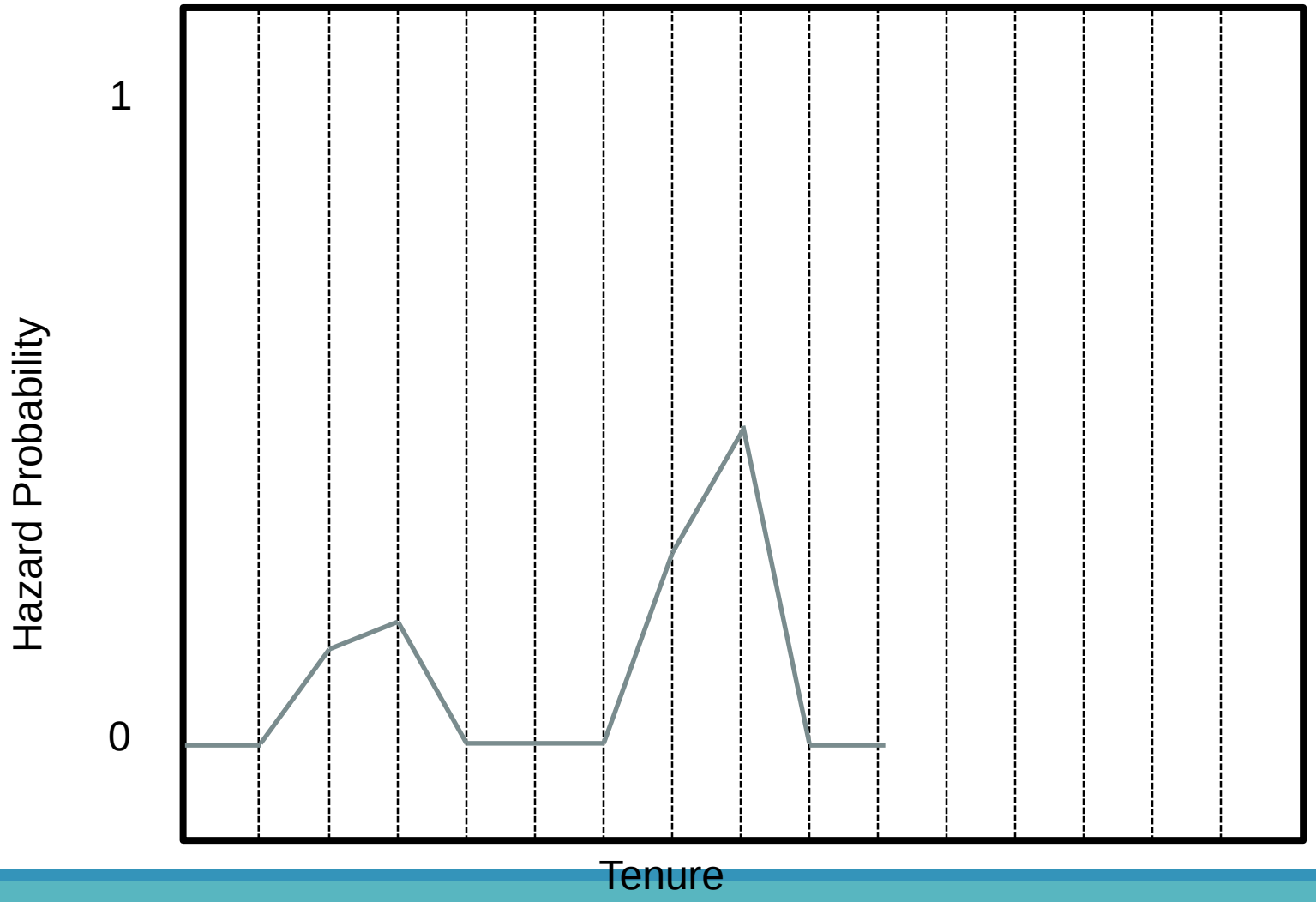
Time = 7:

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

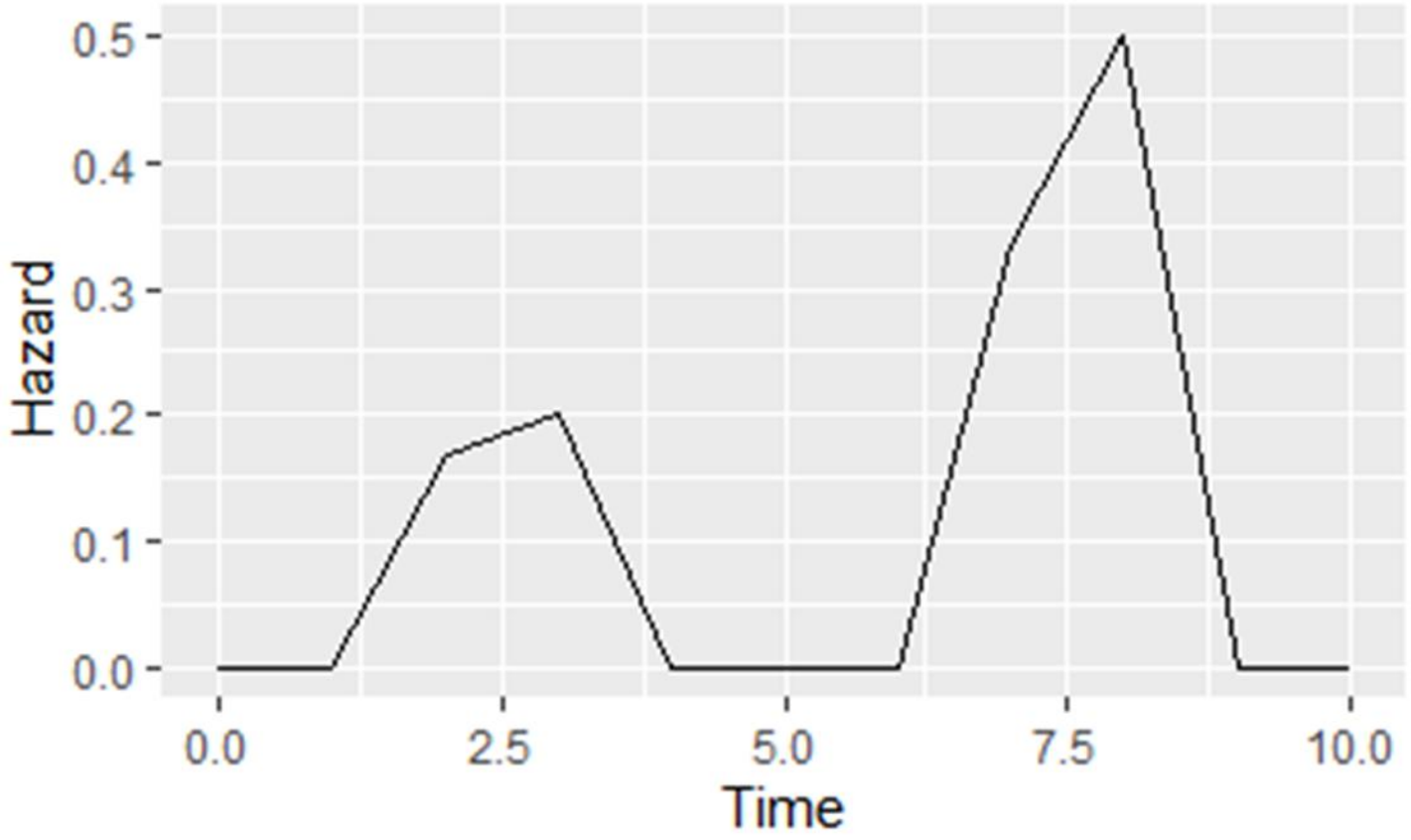
Time = 8:

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

Visualizing Hazard Probabilities



Visualizing Hazard Probabilities



Hazard Probability

Hazard probability can be interpreted as “Assuming the event has not occurred yet, this is the probability the event occurs at this time point”.

Examples:

- Say the hazard probability for a given point in time for contracting a sinus infection is 0.2 with a time measured in months.
- “Assuming the individual has not contracted a sinus infection, the probability s/he contracts a sinus infection this month is 0.2.”

Hazard Probability – Inverse

The interpretation of the inverse of the hazard function is the length of time until the expected next occurrence.

Examples:

- Hazard for some point in time for contracting a sinus infection is 0.2 with a time measured in months.
- “Assuming an individual has not yet contracted a sinus infection, we expect this individual to go 5 (= 1/0.2) months before contracting a sinus infection (assuming the hazard stays constant).”

Hazard Functions – R

```
summary(simple_km)
```

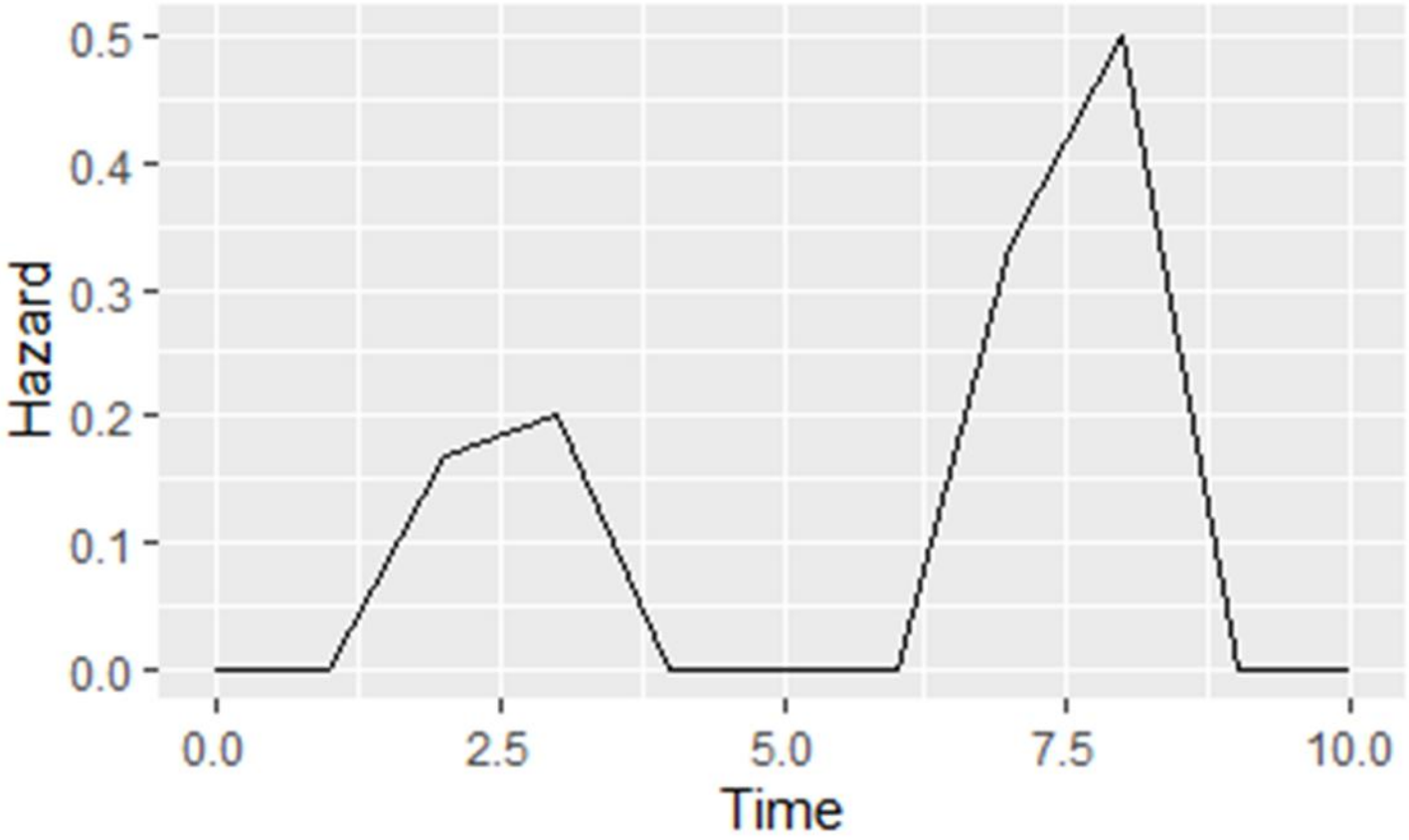
```
## Call: survfit(formula = Surv(time = tenure, event = (censored == 0))
~
##     1, data = simple)
##
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     2      6       1    0.833   0.152       0.5827            1
##     3      5       1    0.667   0.192       0.3786       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

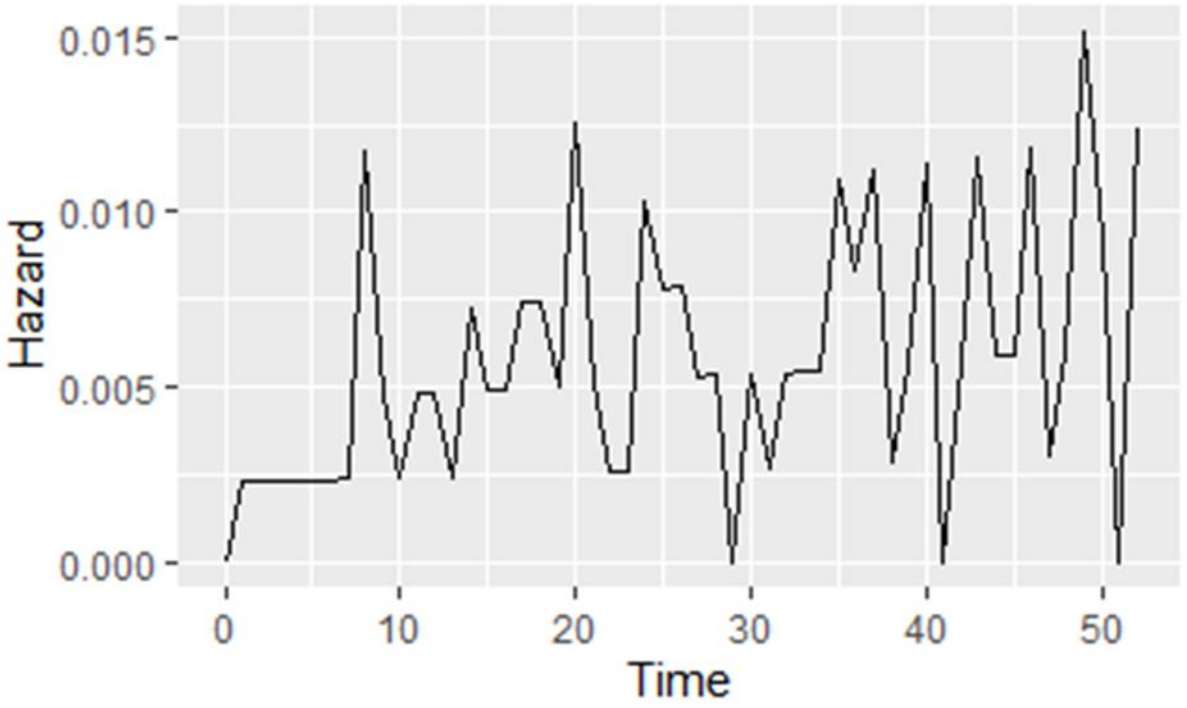
```
## [1] 0.1666667 0.2000000 0.3333333 0.5000000 0.0000000
```

Hazard Functions – R

```
h= simple_km$n.event/simple_km$n.risk
index.h=rep(0,length=(max(simple$tenure)+1)) #Need to add 0
index.h[(simple_km$time)+1]=h #Because of 0
haz.plot=data.frame(cbind(seq(0,max(simple$tenure)), index.h))
colnames(haz.plot)=c("Time","Hazard")
ggplot(haz.plot,aes(x=Time,y=Hazard))+geom_line()
```


Hazard
0.5
0.4
0.3
0.2
0.1
0.0
0.0
2.5
5.0
7.5
10.0
Time

Hazard Functions – Recid data



Hazard Rates

Hazard rates have a slightly different interpretation than the hazard probabilities because they based on continuous distributions.

$$h(t) = \frac{f(t)}{1 - F(t)}$$

The hazard rate is the **instantaneous event rate** for the risk set at time *t*.

- Given survival up until time *t*, it is the instantaneous rate of the event happening at time *t*.

Hazard Rates

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

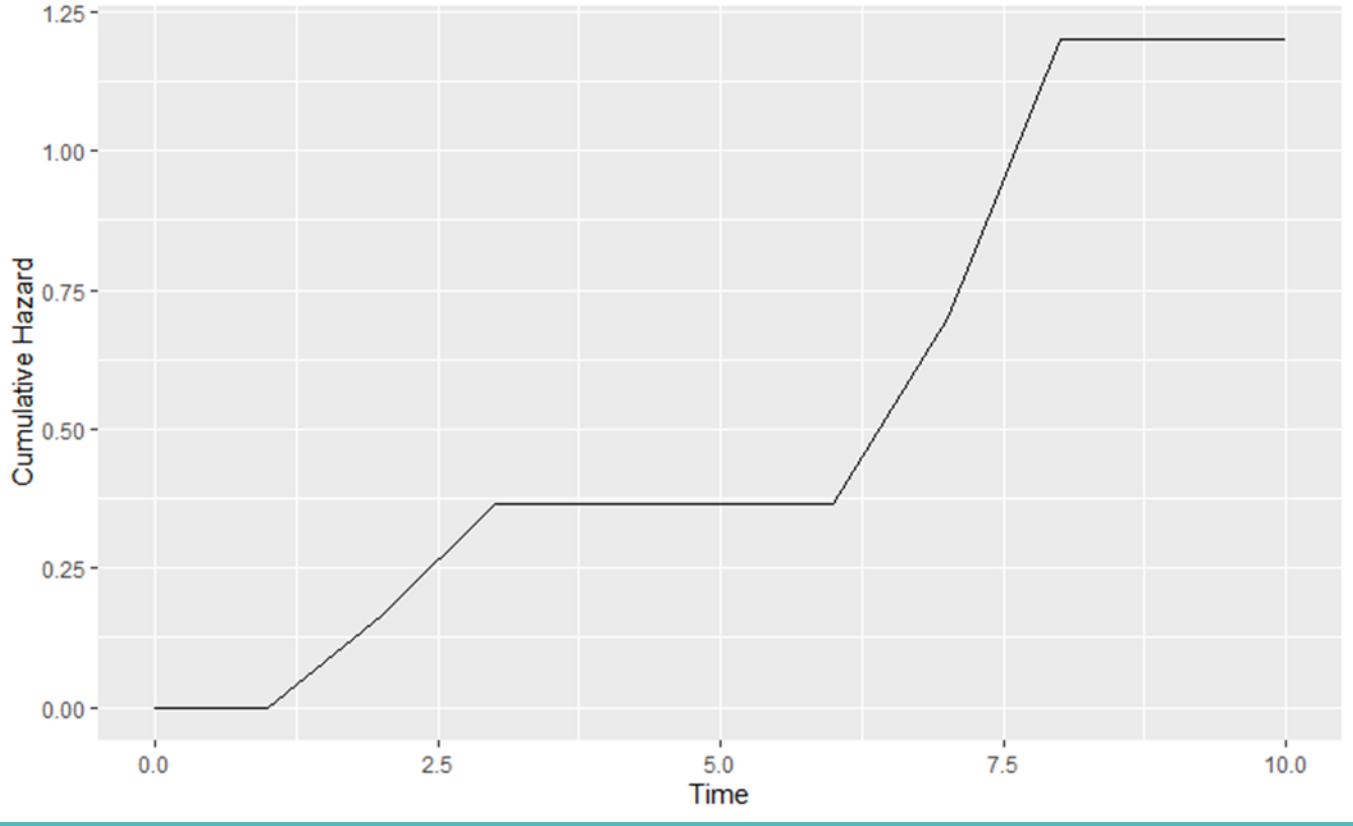
$$h(t) = \frac{P(T = t)}{1 - P(T \leq t)} \; or \; \frac{f(t)}{1 - F(t)}$$

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!

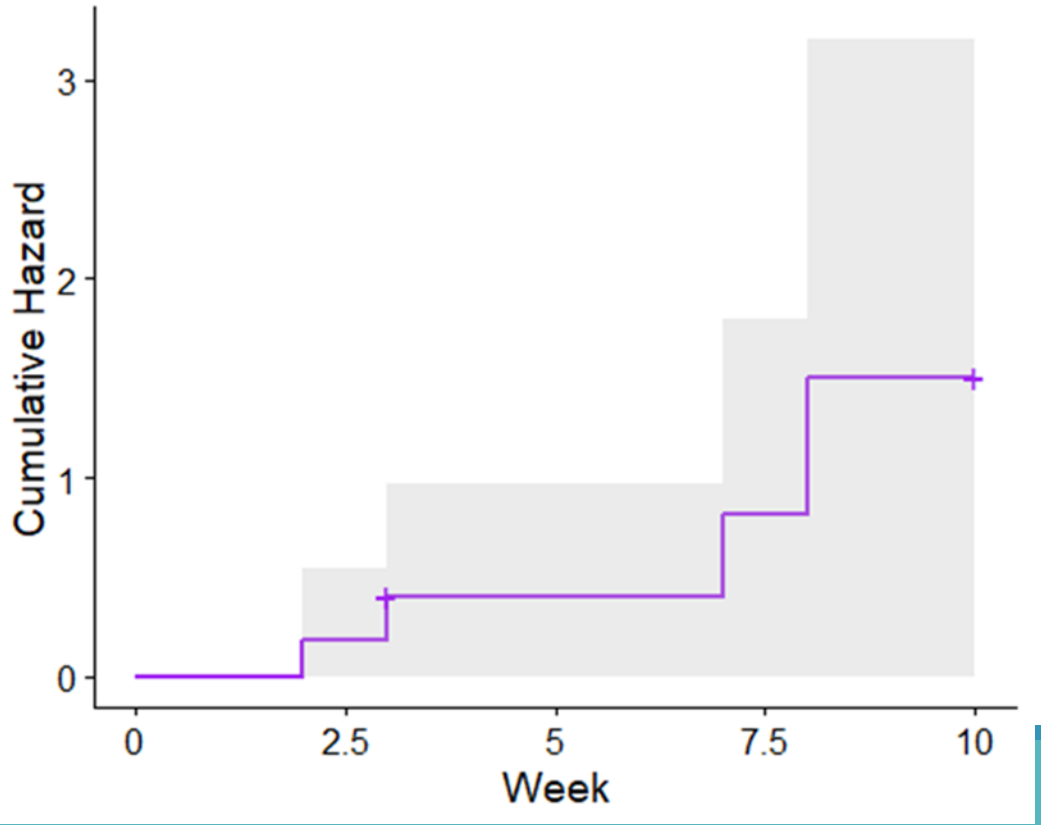
Cumulative Hazard Probability

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



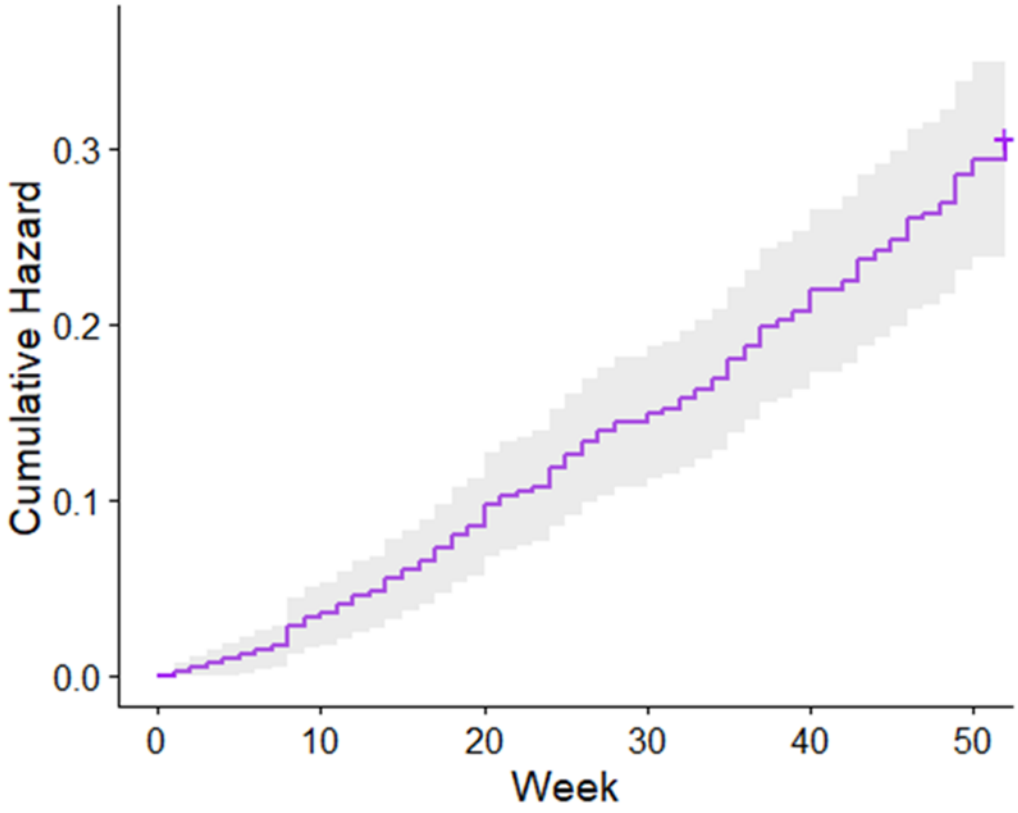
Hazard Functions – R

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



Hazard Functions – R

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



Survival and Hazard Relationship

The survival, hazard rate, 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.