# Biostat 537: Survival Analysis

## TA Session 1

Ethan Ashby

January 9, 2024

# Presentation Overview

# Hi! I'm your TA!

My name is Ethan Ashby and I'm a third year PhD student in the Biostatistics department at UW.

Typically, discussion section will contain a short lecture followed by time for questions.

Email me (eashby (at) uw dot edu) with any personal questions as you progress through this course! Feel free to post any content-related questions to the "Discussions" page on Canvas so others may benefit!

Office Hours will be Fridays 9-10 AM PST on Zoom!

Lecture slides will also be made available on the course website.

# A quick plug: UAW Union!

(a) Join the Union!

(b) Sign off on the Bargaining Demands!

# Roadmap

# Survival Data

Survival data describes the *time until an event occurs*. Let $T$ denote the positive random variable for a unit's survival/failure time.

Requires specifying a *time origin* ($t = 0$).

Risk Set at time $t$: the set of units *susceptible to experiencing the event* at time $t$.

Some examples:

1. Hardware reliability: time until component fails.
2. Vaccines: time until PCR-confirmed influenza infection.
3. Cancer: time until death/disease progression.

# What do we do with survival data?

1. Estimate probability of failure/survival at a given point(s) in time.
2. Compare survival experience between groups.
3. Assess the relationship of explanatory variables to survival time.

# The unique challenge of censoring

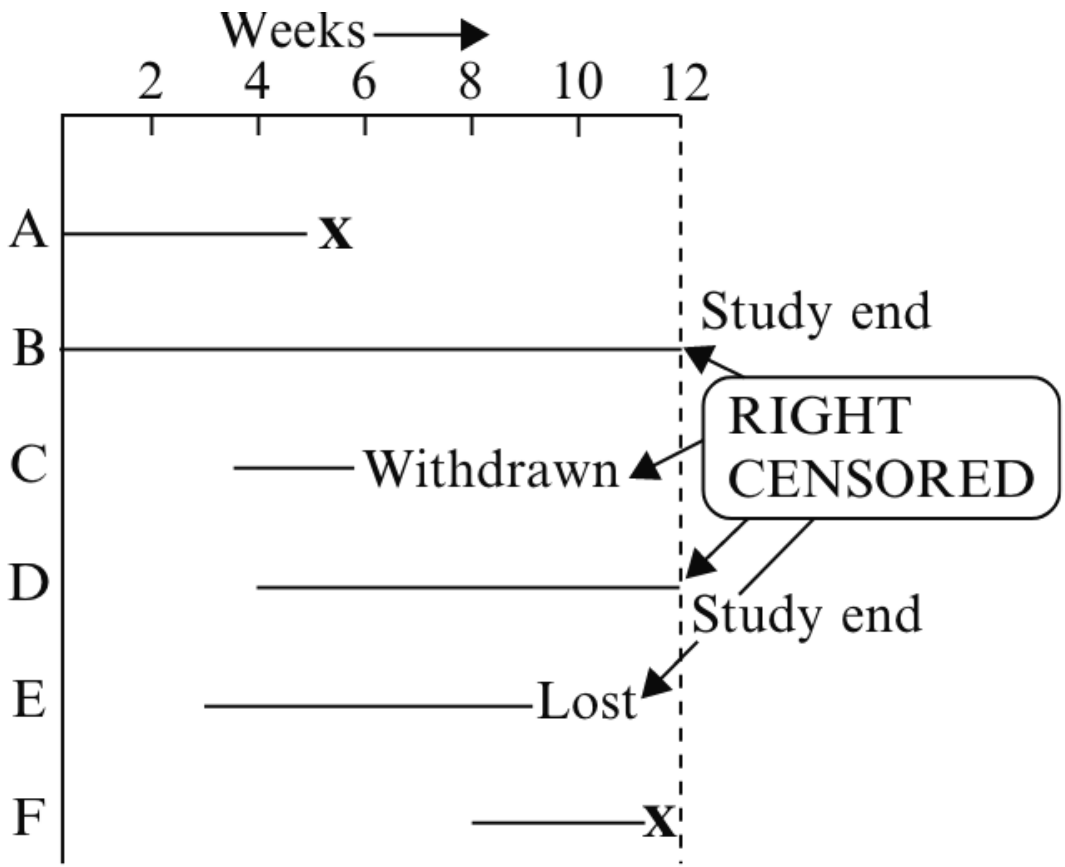
A hallmark of survival data is censoring, or where the survival time is not known precisely, but is known to lie before, after, or between certain values.

Right-censoring: when the true survival time is known to lie *after* a given value; i.e., $T \in [t_1, \infty)$.

Causes of right censoring

1. Administrative censoring: study ends before participant experiences an event.
2. Lost to follow up during study period.
3. Dropout from the study.

# Censoring cont.



# Censoring cont.

Left-censoring: when the true survival time is known to lie *before* a given time point; i.e., $T \in [0, t_1)$.

Interval-censoring: when the true survival time is known to lie *between* two time points; ; i.e., $T \in [t_1, t_2]$.

# Why do we care about censoring?

1. Censored data are incomplete but potentially informative! Throwing out censored data is inefficient!

2. May lead to biased estimates of survival quantities if censoring is informative.

# Necessity of Assumptions

Censoring – when survival times are not exactly known but lie in known ranges – leads to incompletely observed outcome information.

For survival analysis tasks, we must *impose assumptions* on the censoring mechanism.

# Censoring Assumptions

Many methods we will use in this course rely on the following assumption.

Independent censoring: *within any subgroup*, the survival experience of censored participants at any time $t$ can be represented by the survival experience of all subjects in the subgroup who remained in the risk set at time $t$.

One can also interpret this as completely random censoring within each subgroups of participants with shared characteristics.

# An Illustrative Example

Suppose we have a collection of 100 participants in group A and want to estimate their 5-year survival probability.

1. 0-3 years: 20 of 100 experience the event.
2. 3 years: 40 of 80 refuse to continue in the study.
3. 3-5 years: 5 of 40 experience event.

Under *random censoring*, 40 censored participants will have survival experience similar to the uncensored participants who remained at risk at 3 years! Hence, survival probability $= 1 - (20 + 5 + 5)/100 = 0.70$

# An Illustrative Example

Suppose in addition to 100 group A participants, we have 100 group B participants.,

1. 0-3 years: 40 of 100 experience the event.
2. 3 years: 10 of 60 refuse to continue in the study.
3. 3-5 years: 10 of 50 experience event.

Under *random censoring*, 10 censored participants will have survival experience similar to the uncensored participants who remained at risk at 3 years! Hence, survival probability $= 1 - (40 + 10 + 2)/100 = 0.48$ in group B.

# An Illustrative Example

Suppose we wish to estimate the survival in our cohort of 200 group A + group B patients!

1. Group A had 40 of 80 at-risk participants censored at year 3, while Group B only had 10 of 60.
2. Group A had greater 5-year survival probability (0.70) compared to Group B (0.48).

The censored participants at 3-years had higher survival probability than the uncensored participants entirely due to group membership! Hence, censoring is informative, but *within groups* it is random.

# Left truncation

A form of *selection bias* or *length-biased sampling*, where we can only observe an event time if it is *greater* than a certain value. Leads to small event times being unobserved.

Example: suppose we want to estimate time from cancer diagnosis to death and we recruit patients with a diagnosis. We miss data from patients with very aggressive cancers who died shortly after diagnosis.

Critical question to ask: who is my target population and who is included/excluded from my sample?

# Roadmap

# Survival data: Notation

For a given individual $i$, their survival data is summarized by the following vector $(t_i, d_i, X_{i1}, \ldots, X_{ip})$ where

# Survival data: Notation

For a given individual $i$, their survival data is summarized by the following vector $(t_i, d_i, X_{i1}, \ldots, X_{ip})$ where

1. $t_i = T_i \wedge C_i$ (minimum of survival and censoring time)
2. $d_i = \mathbb{I}(T_i < C_i)$ denotes whether the observation was observed or censored.
3. $(X_{i1}, \ldots, X_{ip})$ denote a vector of $p$-covariates.

# Survival data: presentation

Survival data are often summarized like so: $(t_1, t_2, t_3, t_4+, t_5+)$ where "+" denotes censoring at the designated time.

To computers, we often encode survival data like so

| Individual | $t$ | $d$ (failed or censored) |
|---|---|---|
| 1 | $t_1$ | 1 |
| 2 | $t_2$ | 1 |
| 3 | $t_3$ | 1 |
| 4 | $t_4$ | 0 |
| 5 | $t_5$ | 0 |

# Roadmap

# Some machinery

Let $F(t) := P(T \leq t)$ be the *cumulative distribution function (CDF)* of survival times, describing the probability of experiencing an event *at or before* $t$.

Let $f(t) := \frac{d}{dt}F(t)$ be the *density function* of survival times, or the rate of change of the CDF.

# Survivor & Hazard Functions

Survivor function: $S(t) := P(T > t)$ gives the probability of surviving beyond a specified time $t$.

Hazard function: $h(t)$ describes the *instantaneous rate* (per unit time) of experiencing an event at time $t$ *given* the individual survived up to time $t$. We formalize as follows:

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

# Survivor and hazard functions

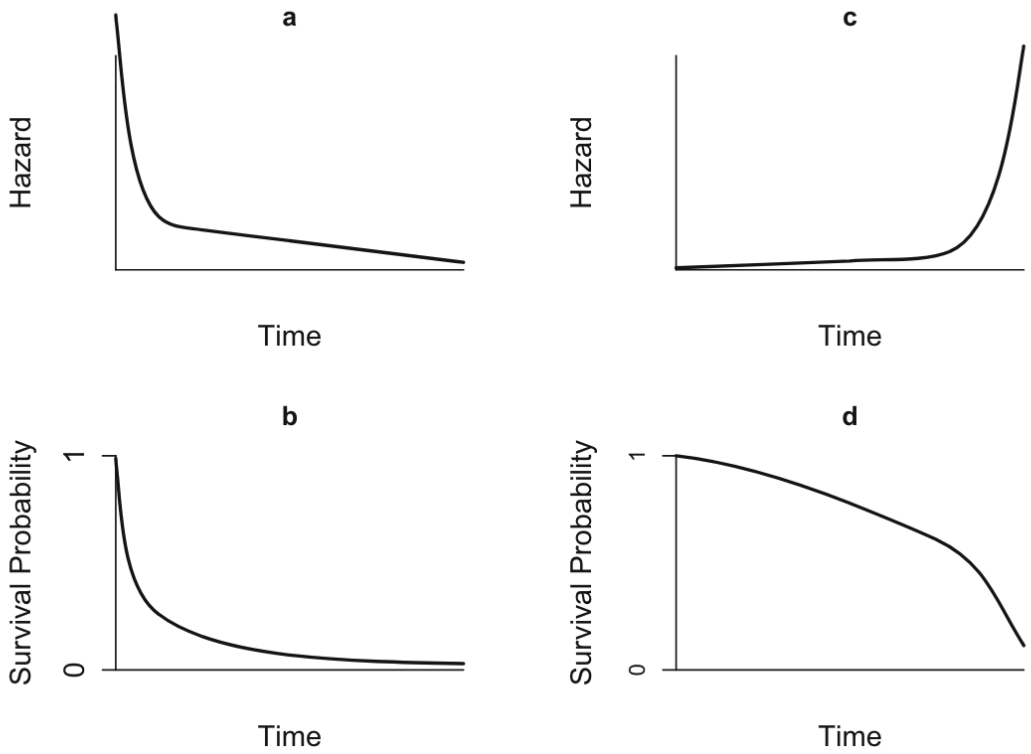



**Fig. 2.1** Hazard and survival functions with high initial hazard (**a** and **b**) and low initial hazard (**c** and **d**)

# Relationship between survivor & hazard functions

There exists a 1-1 relationship between survivor functions $S(t)$ and hazard functions $h(t)$.

$$S(t) = \exp\left[ -\underbrace{\int_0^t h(u)du}_{\star} \right]$$

$$h(t) = -\left[ \frac{dS(t)/dt}{S(t)} \right] \equiv \frac{f(t)}{S(t)}$$

Note: $\star \equiv H(t)$ is the *cumulative hazard function*.

# Relationship between survivor & hazard function

We can derive this relationship

$$
\begin{aligned}
h(t) &= \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t | T > t)}{\Delta t} \\
\implies h(t) \cdot S(t) &= \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t)}{\Delta t} \\
&\equiv \lim_{\Delta t \to 0} \frac{F(t + \Delta t) - F(t)}{\Delta t} = f(t) \\
\implies h(t) &= \frac{f(t)}{S(t)}
\end{aligned}
$$

# Example: parametric survival model

Suppose the hazard function is constant over time with risk $\lambda$:

$$h(t) = \lambda$$

Recalling $S(t) = \exp\left[-\int_0^t h(u)du\right]$, the survivor function is

$$S(t) = \exp\left(-\int_0^t \lambda\right) = \exp(-\lambda t)$$

What is the CDF of the survival times?

$$\begin{aligned} P(T < t) &\equiv 1 - P(T \geq t) \\ &\equiv 1 - S(t) = 1 - \exp(-\lambda t) \end{aligned}$$

The distribution of survival times is *exponential*($\lambda$)!

# Fundamental goals of survival analysis

1. Estimate survivor and/or hazard functions from data $\Longrightarrow$ estimator will depend on your choice of model (what assumptions you impose).
2. Compare survival and hazard functions between groups $\Longrightarrow$ Hypothesis testing!
3. Assess the relationship of explanatory variables to survival time $\Longrightarrow$ Regression modelling!

# Summary

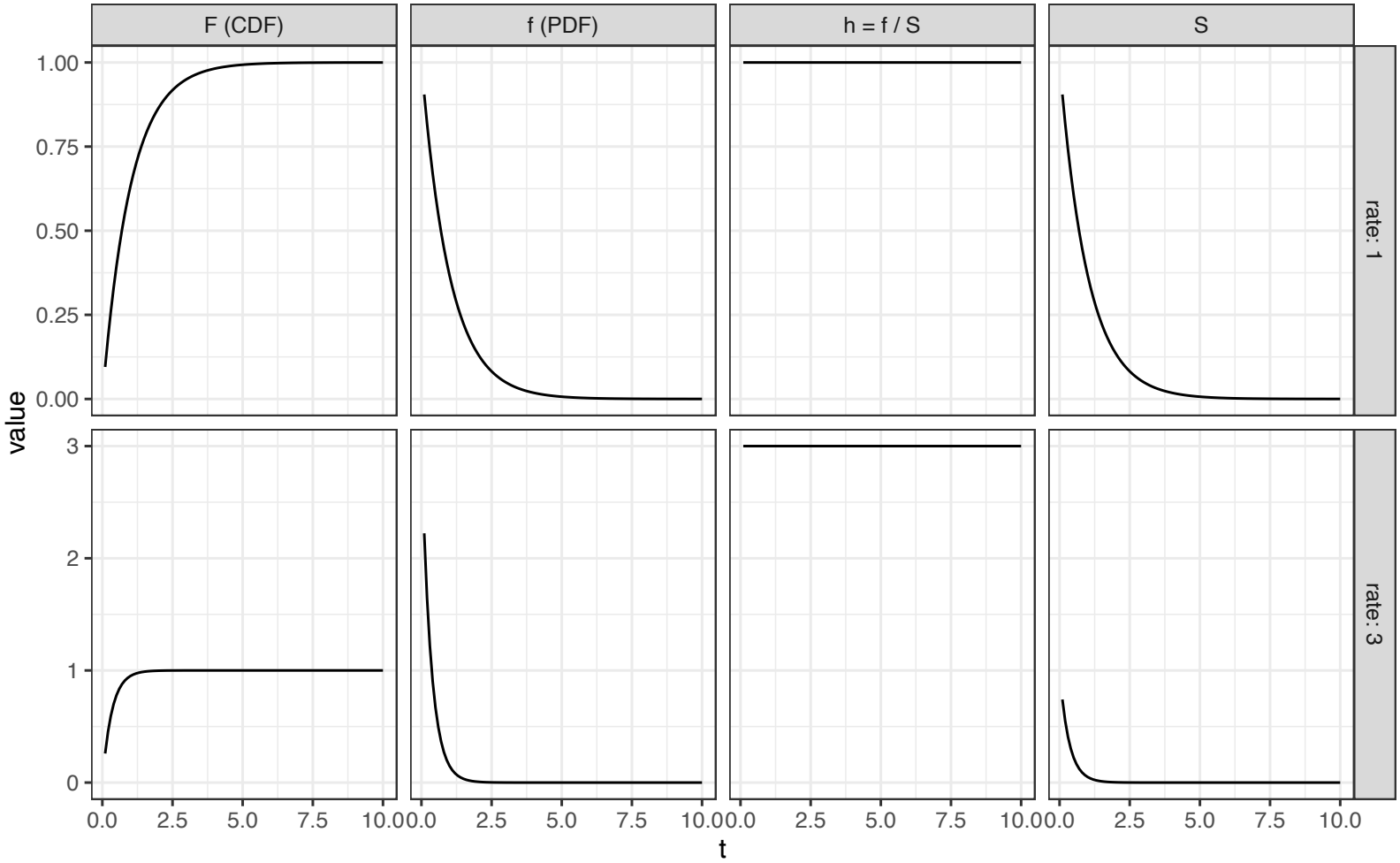
1. Survival data describes the time until the occurrence of an event of interest.
2. Survival data is almost always subject to incompleteness – right censoring is the most common but other forms abound.
3. Survival analysis methods must account for censoring to (a) make efficient use of the available data and (b) avoid bias due to informative censoring.
4. The survivor function and hazard functions are two distinct but related quantities that are central to most survival analysis methods.

# Roadmap

# Useful R tools

```
library(tidyverse);library(survival);library(flexsurv)
expand.grid(
    t = seq(0.1, 10, .1),
    loc = c(1, 3)
) %>%
    mutate(
        `F (CDF)` = pexp(t, rate = loc),
        `f (PDF)` = dexp(t, rate = loc),
        S = 1 - `F (CDF)`,
        `h = f / S` = flexsurv::hexp(t, rate = loc),
        loc = paste("rate:", loc)
    ) %>% pivot_longer(cols=3:6) %>% ggplot(aes(x=t, y=
        value))+geom_line()+facet_grid(loc~name, scales="
        free")+ylab(NULL)+theme_bw()
```

# Useful R tools



# Useful R tools

```
library(flexsurv);library(survival);library(tidyverse)
#Fit exponential survival model
expmodel <- flexsurv::flexsurvreg(Surv(rectime, censrec)
    ~1, data=flexsurv::bc, dist="exponential")

plot(expmodel, type="survival")
plot(expmodel, type="hazard")
plot(expmodel, type="cumhaz")
summary(expmodel, type="median")
summary(expmodel, type="mean")
```

Or you may use the "fitparametric" function developed for this course. Run the code 'source("fitparametric.R")' to load the function into R