# Biost/Epi 537 – Survival Analysis

## Discussion section – Jan 14, 2025

**TA: Anand Hemmady**

# Welcome to Survival Analysis!

- These discussion sections are meant to help you!
  - Let me know what would be most helpful (anandh@uw.edu) so I can make sure that these sections are a good use of your time.
- Plan is to go over key concepts from lectures/assignments.
- Reminder: TA office hours are over Zoom.
  - 3-4pm on Wednesdays, 10-11am on Thursdays.

# What is survival analysis?

- The analysis of "time-to-event" outcomes – often denoted by $T$.
  - How long does it take until some event of interest occurs?
  - $T =$ Time until death after brain surgery
  - $T =$ Time until a bridge collapses
- Important to be very specific about what the event is.
  - E.g. "death" and "death due to lung cancer" are different!

# What can we do with survival analysis?

- Compare survival time between groups.
  - E.g. COVID vaccine trial – event is diagnosis with COVID 19.
    - Do vaccinated participants “survive” longer (i.e. have longer time until diagnosis with COVID) than unvaccinated participants?
- Estimate the probability of “failure” (the event occurring) or “survival” (the event not occurring) by a certain point in time.
  - E.g. what is the probability that a patient remains cancer-free six months after having a brain tumor removed?

# Features of survival data

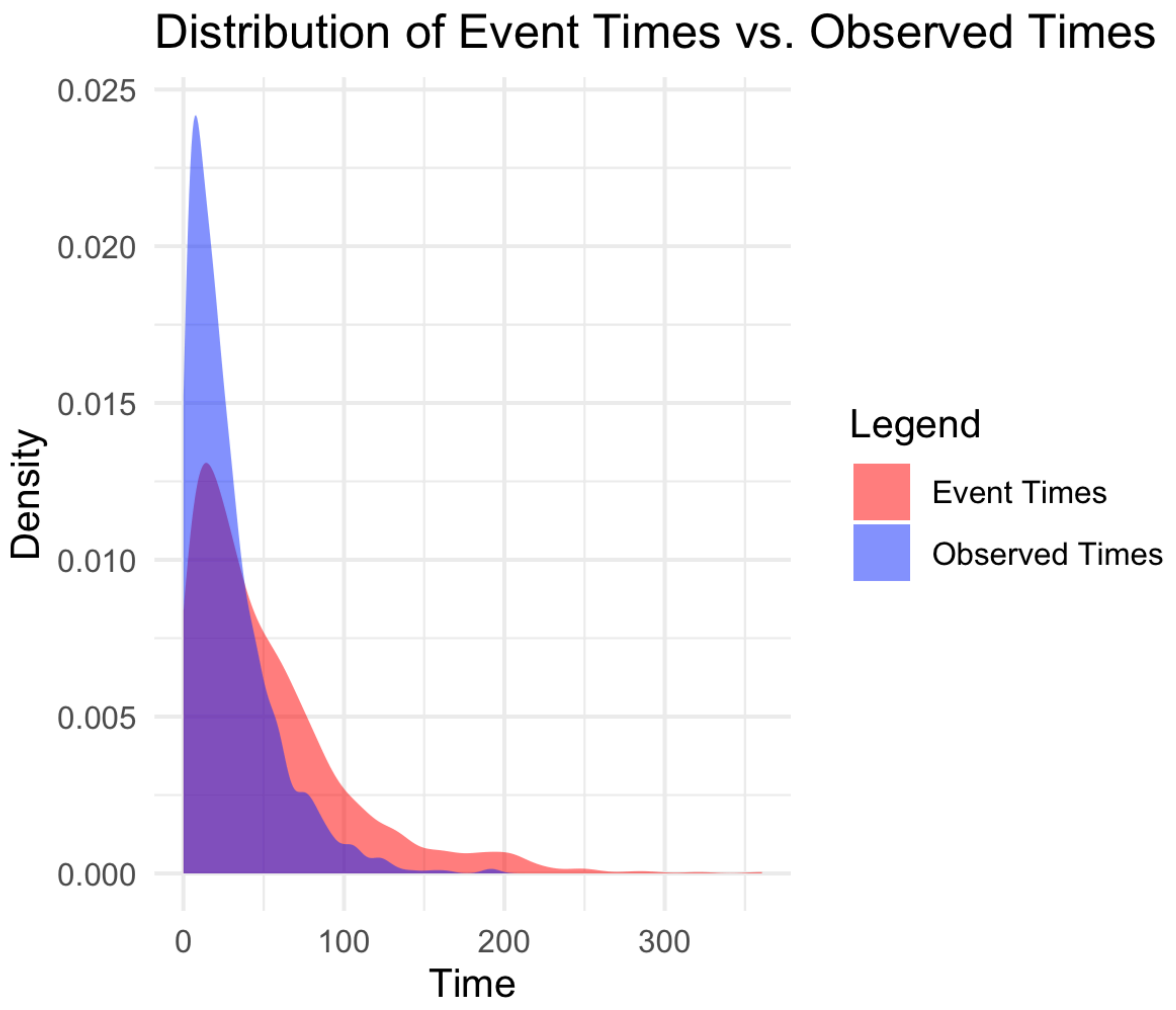
- Survival times are positive: $T > 0$.
- Prone to two kinds of missingness.
    - Censoring – when the value of $T$ is not precisely known.
    - Truncation – when individuals with certain values of $T$ are excluded from the study.
- We can't ignore censoring or truncation!

# More on censoring

- Two types of censoring we are most concerned with
- Right-censoring: When we know that a patient has survived at least until a certain time
    - I.e. we know that $T > t$ for some time $t$.
- Interval censoring: When we know that a patient had the event between two times, but we don't know exactly when
    - I.e. we know that $t_0 < T < t_1$ for times $t_0 < t_1$.

# Example: ignoring left-censoring

- Simulate 1,000 event times from an exponential$(1/50)$ distribution.
  - True mean is 50.
- Simulate 1,000 censoring times from an exponential$(1/60)$ distribution.
- The observed times are the minimum of the event and censoring times.
- What happens if we:
  - 1. Pretend that the observed times (which we see for everyone) are event times?
  - 2. Only focus on uncensored observations (for whom we get to see event times)?




- Using observed times under-estimate the mean.
- Restricting to uncensored observations also underestimates the mean.

```
> mean(dat$survival_times)
[1] 51.49896
> mean(dat$observed_times)
[1] 27.80783
> mean(dat[dat$event == 1,]$survival_times)
[1] 28.04431
```

# So, how do we deal with censoring/truncation?

- Censored data are incomplete but still contain information.
  - Ignoring censored data leads to bias.
  - Pretending that observed times are event times leads to bias.
- Truncation is trickier – while censored data give us clues, truncation means certain people are omitted entirely from the study!
- This quarter is all about ways to deal with these two wrinkles.

# Independent censoring

- Risk set at time $t$ – individuals who have survived until time $t$ (i.e. haven't experienced the event) *and* are not censored at time $t$.
    - The set of people for whom, at time $t$, we can hope to know their actual survival times.
- Many methods rely on the *independent censoring* assumption.
    - The survival experience of individuals in the risk set at time $t$ is the same as the survival experience of censored individuals who haven't yet experienced the event.
    - Often restrict this to subgroups defined by covariates rather than at the population level.

# Why independent censoring?

- We can use those in the risk set to make predictions about those who were censored.
- Example: simulate 1,000 event times from an exponential$(1/50)$ distribution.
- Around 30% of individuals are censored at $t = 5$, remaining are uncensored.
- Want to estimate $P(T > 10)$.
  - Among those who aren't censored, we simply count how many people have $T > 10$.
  - Among those who are censored, we can use independent censoring! The proportion the censored individuals who survived until $t = 10$ is equal to the proportion of individuals in the risk set at $t = 5$ who survived until $t = 10$.
- In R…

# Functions to know

- The time-to-event outcome is denoted $T$. In this class, we will assume that $T$ is continuous. Two functions you might be familiar from usual statistics:
- Denote by $F$ its cumulative distribution function (cdf)
  - $F(t) = P(T \leq t)$.
- Denote by f its probability distribution function (pdf)
  - $f(t) = \frac{d}{dt} P(T \leq t) = \lim_{h \to 0} \frac{P(T \leq t + h) - P(T \leq t)}{h}$.

# Functions we care about in survival analysis

- Survival function: $S(t) = P(T > t)$ .
  - The probability that an individual does not experience an event by time $t$
- Hazard function: $h(t) = \lim_{h \to 0} \dfrac{P(t \leq T < t + h \,|\, T \geq t)}{h}$.
- Usually, we are most concerned with estimating one of the above two quantities. They are equivalent to knowing $F(t)$ or $f(t)$; they all let us completely understand the distribution of $T$.

# Relationships between the functions

- When $T$ is continuous, the following relationships hold:
  - $h(t) = \dfrac{f(t)}{S(t)}$;
  - $S(t) = \exp\left\{-\int_0^t h(s)ds\right\}$;
  - $S(t) = 1 - F(t)$.
- Takeaway – all of these functions are related. Knowing one gives you the others.

# Summary

- Interested in time-to-event data.
  - Comparing survival times between two groups.
  - Estimating probability of survival at a certain point in time.
- Time-to-event data are prone to censoring and truncation. Failing to account for these can lead to substantial bias in the above (and other!) tasks.
- Usually attempt to estimate/model the survival or hazard functions, often using the assumption of independent censoring (among others).