Biostatistics

Mats J. Stensrud

Spring 2021

Section 1

Structure of the course

Structure

- 90 minutes lectures every Tuesday 10h15.[1]
- Moodle is our platform.
    - Announcements
    - Links to relevant literature
    - Link to Piazza (Password: SURVIVAL).
- Slides, video recordings and problem sheets will be uploaded every Tuesday.

---

[1] However, if people do not meet up, I will pre-record the lectures.

Exam

- Written exam.
- Two graded homeworks.

Features of the course

- This is a *statistics* course.
- We will study theory and methods that is relevant to solve common practical problems.
- The course will contain proofs, but *all* the results we are using will *not* be shown. That said, I will strive to motivate all the results.
- I will also spend time on discussing the interpretation of the results.

Goal

After the course, you should be able to:

- Understand mathematical and statistical theory for event history analysis and longitudinal data analysis.
- Furthermore, understand the concepts and ideas that this theory expresses.
- Apply these methods to data (there are ubiquitous applications!).
- Critically evaluate how these methods are used in practice.
- Build on the material in this course to derive new results yourself.

Outline of the course

- **Time-to-event outcomes** ("survival analysis")
- Longitudinal data analysis
- Research synthesis

Section 2

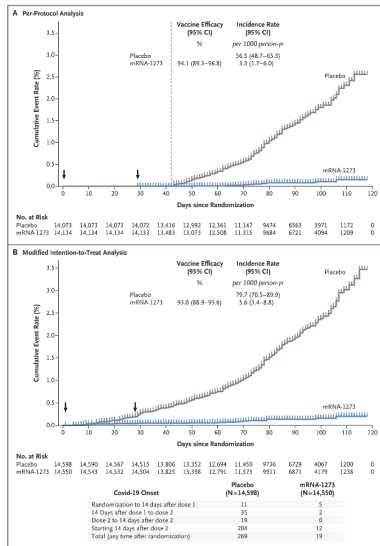## Time-to-events and survival analysis

Figure 1: Survival analysis is e.g. used to present results from vaccine trials.

Time to events are all over the place

- Time from birth to death.
- Time from birth to cancer diagnosis.
- Time from disease onset to death.
- Time from entry to a study to relapse.
- Time from marriage to divorce.
- Time from production until a machine is broken.
- Time from origin of the coronavirus until a stock (marked) crashes.

NEWS FEATURE

THE TOP
100
PAPERS

Nature *explores the most-cited research of all time.*

STATISTICS

Although the top-100 list has a rich seam of papers on statistics, says Stephen Stigler, a statistician at the University of Chicago in Illinois and an expert on the history of the field, "these papers are not at all those that have been most important to us statisticians". Rather, they are the ones that have proved to be most useful to the vastly larger population of practising scientists.

Much of this crossover success stems from the ever-expanding stream of data coming out of biomedical labs. For example, the most frequently cited statistics paper (number 11) is a 1958 publication[15] by US statisticians Edward Kaplan and Paul Meier that helps researchers to find survival patterns for a population, such as participants in clinical trials. That introduced what is now known as the Kaplan–Meier estimate. The second (number 24) was British statistician David Cox's 1972 paper[16] that expanded these survival analyses to include factors such as gender and age.

Figure 2: The two most cited statistics papers concern survival analysis

Some common questions

- What is survival under treatment A vs B?
- What is the duration of a certain component in the machine?
- How long does it take before a stock marked crashes?

PS: These questions are very often about causal effects....

An overview of the data structure 2

- We follow units of over time;
  humans, animals, engines, etc.
- The events of interest may be **the time to** deaths, cancer diagnoses,
  divorces, child births, engine failures, etc.
- We often stop the study before everyone has experienced the event of
  interest.

Censored survival times (illustration)

Consider 10 patients with newly diagnosed cancer. Let $T \in (0, \tau]$ be a survival time.
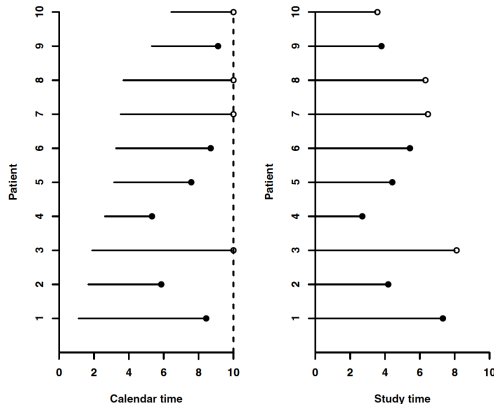
7.32, 4.19, 8.11, 2.70, 4.42, 5.43, 6.46, 6.32, 3.80, 3.50.
How do you estimate $\mathbb{E}(T)$, that is, the mean survival?

Why not use "standard methods"?

- We have incomplete observations.
- Instead of observing T_i we observe $(\tilde{T}_i, D_i)$,

$$\tilde{T}_i = T_i \ \text{ if } \ D_i = 1,$$
$$\tilde{T}_i < T_i \ \text{ if } \ D_i = 0.$$

  where D_i is a censoring indicator.

  We want to use our information on $\tilde{T}_i$ to make inference on T.

- There is a strong link to causal inference and "what if" questions: What would happen if we observed T_i instead of $\tilde{T}_i$.
- We must make assumptions about the censoring, similarly to assumptions in causal inference.

Let's start with a single outcome process

Assume $T > 0$ is an absolutely continuous random variable.

Definition (Survival function)

The survival function is $S(t) = P(T > t)$, that is, the probability that the survival time T exceeds t.

Definition (Hazard rate)

The hazard rate $\alpha(t) = \lim_{dt \to 0} \frac{1}{dt} P(t + dt > T > t \mid T \geq t)$ is the rate of events per unit of time.
That is,.

Informally, $\alpha(t)dt = P(t + dt > T > t \mid T \geq t)$ is the probability that the event will happen between time t and time $t + dt$ given that it has not happened earlier.[2]

---

[2]PS: We are going to extend this to multiple events later.

Cumulative hazard and some relations

Define the cumulative hazard,

$$A(t) = \int_0^t \alpha(s)ds.$$

Then,

$$A'(t) = \alpha(t) = \lim_{dt \to 0} \frac{1}{dt} \frac{S(t) - S(t + dt)}{S(t)} = -\frac{S'(t)}{S(t)} = \frac{f(t)}{S(t)}.$$

By integration

$$\int_0^t \alpha(s)ds = -\log\{S(t)\},$$

and thus

$$S(t) = \exp\{-\int_0^t \alpha(s)ds\}.$$

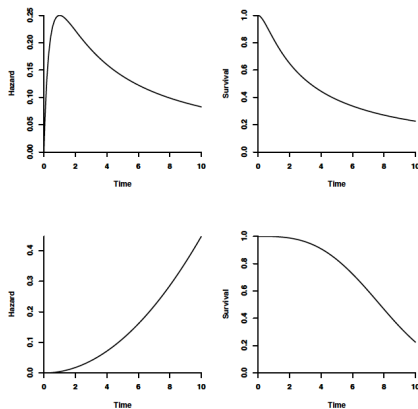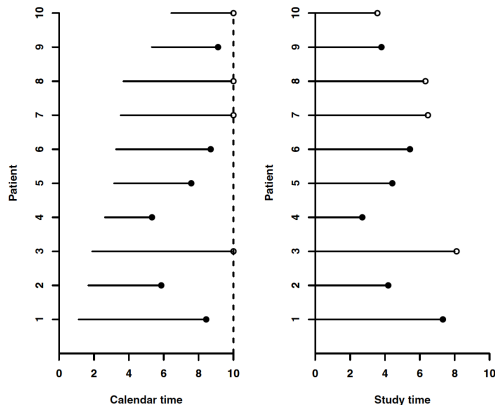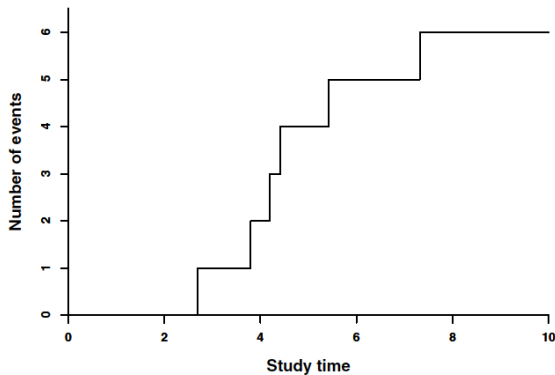$\alpha(t)$ completely determines the distribution of survival times T.

Fig. 1.2 *Illustrating hazard rates and survival curves. The hazard rates on the left correspond to the survival curves on the right.*
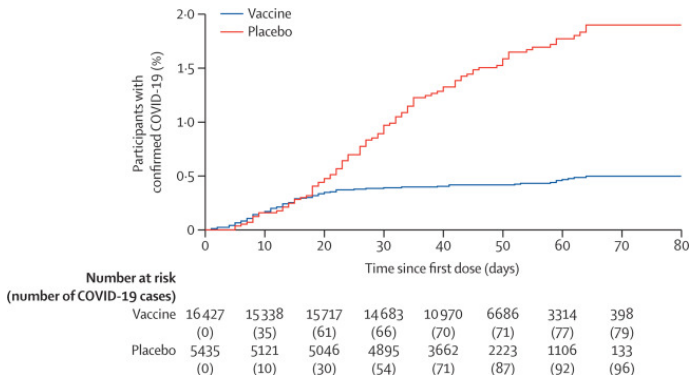
7.32, 4.19, 8.11, 2.70, 4.42, 5.43, 6.46, 6.32, 3.80, 3.50.

Example: Counting process description

7.32, 4.19, 8.11, 2.70, 4.42, 5.43, 6.46, 6.32, 3.80, 3.50.

The Sputnik Vaccine.

Section 3

Processes

Many of you are familiar with stochastic processes

Here I will review *basic* concepts and results on stochastic processes.
I will give definitions and proceed at a "working technical" level.
We will focus on counting processes and martingales.
There are rigorous courses on stochastic processes at EPFL, such as:

- MATH-330 Martingales et mouvement brownien (Prof. Aru).
- MATH-332 Stochastic processes (Prof. Mountford).

Stochastic process

- A stochastic process is a time-indexed collection of random variables, say, $\{X(t) : t \in [0, \tau]\}$.
- Consider a probability space $(\Omega, \mathcal{F}, P)$.
- A filtration $\{\mathcal{F}_t\}_{t>0}$ is an increasing right-continuous family of sub-σ-algebras of $\mathcal{F}$ such that $\mathcal{F}_s \subset \mathcal{F}_t$ whenever $s \leq t$.
  Think about the filtration as representing the **past**, that is, the history.
- We denote $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t>0}, P)$ a filtered probability space.

Example
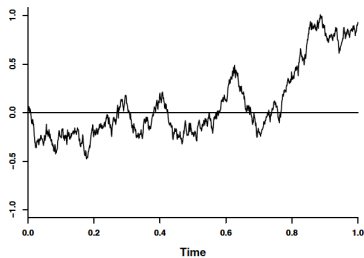
Wiener process (Brownian motion)

Definition (Wiener process)

The $W = \{W(t) : t \in [0, \tau]\}$ is a process satisfying

- $W(0) = 0$,
- independent increments, that is, $W(t + u) - W(t)$ $u \geq 0$ are independent of $W(s)$, for all $s \leq t$,
- Gaussian increments, that is, $W(t + u) - W(t) \sim \mathcal{N}(0, u)$,
- continuous sample paths, that is, $W(t)$ is continuous t.

Adapted process

Definition (Adapted process)

A stochastic process $X = \{X(t); t \in [0, \tau]\}$ is adapted to $\{\mathcal{F}_t\}$ if $X(t)$ is $\mathcal{F}_t$ measurable for each t.

```
Intuitively, the value of X(t) is known at t.
```
PS: We will also consider the stronger notion of a *predictable* processes. We omit a formal definition of *predictable* but state the sufficient conditions that a process $X = \{X(t); t \in [0, \tau]\}$ is predictable if

- X is adapted to $\{\mathcal{F}_t\}$, and
- the sample paths of X are left-continuous.[3]

```
Intuitively, the value of X(t) is known just before t.
```

---

[3]A sample path is a realization of X, which is a function of t.

Martingale

Definition (Martingale)

A stochastic process $M = \{M(t); t \in [0, \tau]\}$ is a martingale relative to $\{\mathcal{F}_t\}$ if X is adapted to $\{\mathcal{F}_t\}$ and $\mathbb{E}(M(t) \mid \mathcal{F}_s) = M(s)$ for all $t > s$.

Informally, $\mathbb{E}(dM(t) \mid \mathcal{F}_{t-}) = 0$, where $\mathcal{F}_{t-}$ is the filtration just before t. $\mathcal{F}_{t-}$ is the smallest σ algebra containing all $\mathcal{F}_s$, $s < t$.

Martingale intuition

Definition (Discrete martingale)

Let $M = \{M_0, M_1, M_2, \dots\}$ be a *discrete* stochastic process adapted to $\{\mathcal{F}_n\}$.
The discrete process M is a martingale if

$$\mathbb{E}(M_n \mid \mathcal{F}_{n-1}) = M_{n-1}$$

.

- Heuristic: Think about the Martingale as cumulative noise,
  similar to random errors in "standard" statistical models.

Some features of (discrete) Martingales

- The definition is equivalent to saying that $\mathbb{E}(M_n \mid \mathcal{F}_m) = M_m$ for $m < n$.
- Suppose $M_0 = 0$. Then $\mathbb{E}(M_n) = 0$
  because $\mathbb{E}(M_n) = \mathbb{E}(\mathbb{E}(M_n \mid \mathcal{F}_0)) = \mathbb{E}(M_0) = 0$.
- It also follows that

$$\text{Cov}(M_m, M_n - M_m) = 0, \forall n > m$$

Some features of (discrete) Martingales

- The *predictable variation process* $\langle M \rangle_n$ for $n > 0$ is the sum of conditional variances of martingale differences,

$$\langle M \rangle_n = \sum_{i=1}^{n} \mathbb{E}\{(M_i - M_{i-1})^2 \mid \mathcal{F}_{i-1}\} = \sum_{i=1}^{n} \mathsf{Var}(\Delta M_i \mid \mathcal{F}_{i-1}),$$

where $\Delta M_i := M_i - M_{i-1}$. and $\langle M \rangle_0 = 0$.

- The *optional variation process* $[M]_n$ for $n > 0$ is

$$[M]_n = \sum_{i=1}^{n} (M_i - M_{i-1})^2 = \sum_{i=1}^{n} (\Delta M_i)^2,$$

where $[M]_0 = 0$.