Survival and longitudinal data analysis
Chapter 1 : introduction and basic concepts

Agathe Guilloux
Professeure au LaMME - Université d'Évry - Paris Saclay

Organization

- Week 1: introduction and basic concepts
- Week 2: tests and the Cox model
- Week 3: Lab 1
- Week 4: Lab 2
- Week 5: longitudinal models and variable selection
- Week 6: Lab 3 (project)

+ 1 final exam

Introduction

What is survival analysis ?

Survival analysis

Survival analysis is the study of survival times, durations, or more generally of **time-to-event(s)**, and of the factors that influence them.

Types of fields where time-to-event(s) outcomes are commonly observed and analyzed:

- **biomedical sciences**, in particular in clinical trials, epidemiology / event of interest: onset of a health condition
- **insurance** / event(s) of interest: time(s) of damage
- **economics** / event(s) of interest: time(s) of employment or unemployment
- etc

When there is only one time of interest, it is denoted by T and called **time-to-event, duration or survival time**, equivalently. We will come back later on cases where several times are observed.

What do we want to analyze ?

The main tasks for the statistician are

- to estimate the time-to-event distributions: **estimation**
- to compare time-to-event distributions in different sub-populations: **test**
- to determine which factors/covariates influence these distributions: **regression**.

Why do we need yet another course ? Because durations or survival times are

- **positive random variables**
- **often "ill-observed."**

Parametric distributions for durations

Exponential distribution

$T \sim \mathcal{E}(\lambda)$ with $\lambda > 0$ when T a the p.d.f

$$\lambda \exp(-\lambda t) \text{ on } \mathbb{R}_+.$$

Weibull distribution

$T \sim \mathcal{W}(\lambda, \alpha)$ with $\lambda > 0$ and $\alpha > 0$ when T a the p.d.f

$$\alpha \lambda^\alpha t^{\alpha-1} \exp(-(\lambda t)^\alpha) \text{ on } \mathbb{R}_+.$$

Log-normal distribution

$T \sim \log \mathcal{N}(\mu, \sigma^2)$ with $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ when $\log(T)$ has the $N(\mu, \sigma^2)$ distribution.

Other distributions: gamma, log-logistique, chi-squared, etc

General case

Let the duration T has the c.d.f. F. It is a positive r.v., hence $F(t) = 0$ if $t < 0$.
We will concentrate on $\mathbb{R}_+$.

Survival function

The survival function $\bar{F}$ is defined as

$$\bar{F}(t) = 1 - F(t) = \mathbb{P}(T > t) \text{ for all } t \in \mathbb{R}_+.$$

It is a decreasing, càdlàg function, with $\bar{F}(t) = 1$ when $t < 0$ and $\bar{F}(\infty) = 0$.

Continuous case

Suppose that T has a p.d.f f (with support on $\mathbb{R}_+$).

Hazard rate / intensity function

The hazard rate (aka intensity function) is defined as

$$\lambda(t) = \lim_{h \to 0} \frac{1}{h} \mathbb{P}(t \leq T \leq t+h | T \geq t) = \lim_{h \to 0} \frac{1}{h} \frac{\mathbb{P}(t \leq T \leq t+h)}{\mathbb{P}(T \geq t)}$$
$$= \frac{f(t)}{\bar{F}(t)}$$

for $t \in \mathbb{R}_+$.

It can be interpreted as the infinitesimal probability of "dying" at time t conditionally to "being alive" at time t.

Cumulative hazard/intensity function

The cumulative hazard/intensity function is defined as

$$\Lambda(t) = \int_0^t \lambda(x) dx \text{ for all } t \in \mathbb{R}_+.$$

Exercise: the Weibull distribution

Suppose that $T \sim \mathcal{W}(\lambda, \alpha)$, as defined on slide 6. Compute its

- survival function
- hazard rate
- cumulative hazard rate.
- In the particular case of the exponential distribution ($\alpha = 1$), what is the shape of the hazard function ?

Discrete case

Suppose that T has a discrete distribution on $\{t_1, t_2, \ldots\}$, given by $\mathbb{P}(T = t_i) = p_i$.

Hazard rate / intensity function

The hazard rate (aka intensity function) is defined as

$$\lambda(t_i) = \lim_{h \to 0} \mathbb{P}(t_i \leq T \leq t_i + h | T \geq t_i) = \lim_{h \to 0} \frac{\mathbb{P}(t_i \leq T \leq t_i + h)}{\mathbb{P}(T \geq t_i)}$$
$$= \frac{p_i}{\bar{F}(t_{i-1})} = \frac{p_i}{\sum_{j: t_j \geq t_i} p_j}$$

for $t \in \mathbb{R}_+$.

It can be interpreted as the probability of "dying" at time t_i conditionally to "being alive" at time t_i.

Cumulative hazard/intensity function

The cumulative hazard/intensity function is defined as

$$\Lambda(t) = \sum_{i: t_i \leq t} \lambda(t_i)$$

Exercise: a key relationship

Suppose that T a discrete distribution on $\{t_1 \leq t_2 \leq \ldots\}$, given by $\mathbb{P}(T = t_i) = p_i$. Show that

$$\bar{F}(t_i) = \prod_{j=1}^{i} \big(1 - \lambda(t_j)\big).$$

Exercise: the discrete uniform distribution

Assume that T has a discrete distribution on $\{t_1 \leq t_2 \leq \ldots \leq t_k\}$, given by $\mathbb{P}(T = t_i) = 1/k$. Compute its

- survival function
- hazard rate
- cumulative hazard rate.

Time-to-event data and censoring

Time-to-event data and censoring

Time-to-event or survival time

This is the time between a starting and a ending event.

Examples:

- time between birth and death
- time between the start of a treatment and the start of the effect
- time between the start and end of a unemployment period
- etc

Censoring

Censoring arises when the starting and/or the ending event are not precisely observed.

Right-censoring I

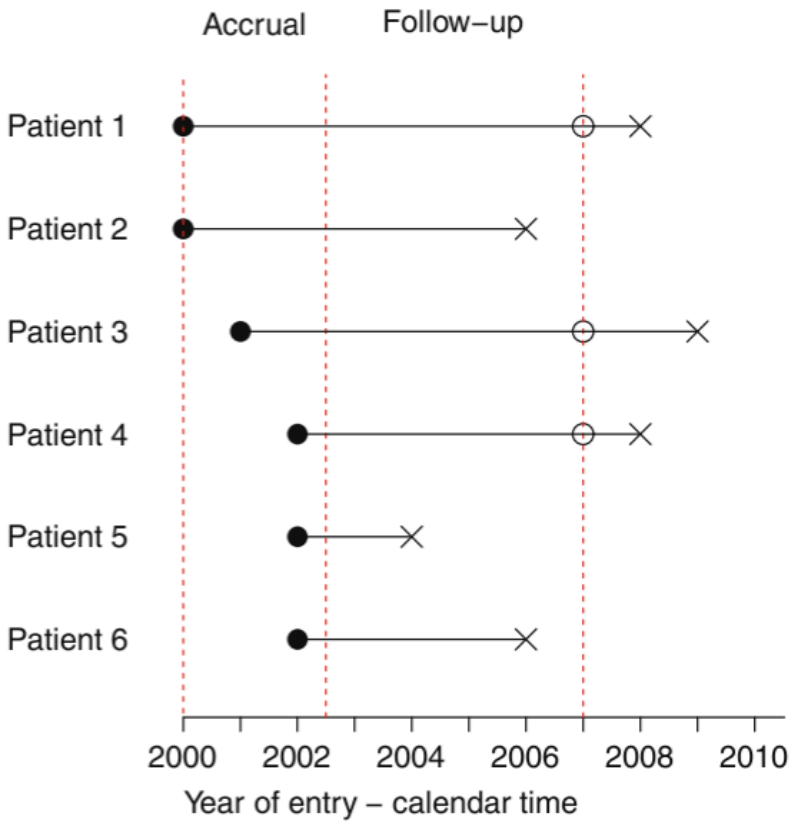


Figure 1: Figure from Moore 2016

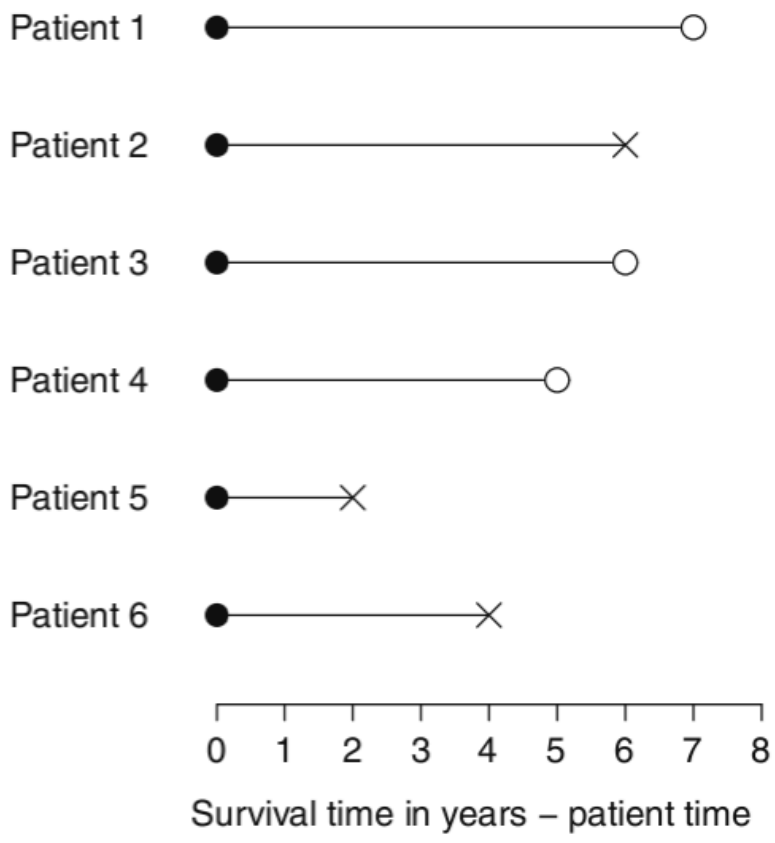


Figure 2: Figure from Moore 2016

Right-censoring II

Independent right-censoring

Let T be the duration and C a positive r.v., independent of T. C right-censors T when we observe

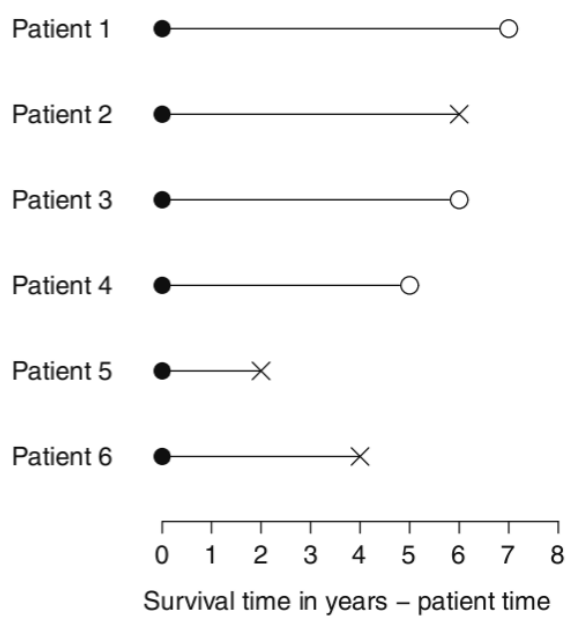
$$T^C = \min(T, C) \text{ and } \delta = \mathbb{1}_{T \leq C}$$

instead of T.

- T^C is the censored time or observed time
- δ is the censoring indicator or status.

Exercise: the form of right-censored data

Fill the tabular on the right.



Patient	Obs. time	Status
1	7	0

Figure 3: Figure from **moore16applied**

The pharmocoSmoking dataset (1)

- Medical therapies to help smokers Randomized trial of triple therapy vs. patch for smoking cessation.
- Data frame with 125 observations and 14 variables:
  - id: patient ID number
  - ttr: Time in days until relapse
  - relapse: Indicator of relapse (return to smoking)
  - grp: Randomly assigned treatment group with levels combination or patchOnly
  - etc

```
##    id ttr relapse         grp
## 1  21 182       0   patchOnly
## 2 113  14       1   patchOnly
## 3  39   5       1 combination
## 4  80  16       1 combination
```

Exercise

- After how many days patient 4 relapsed ?
- After how many days patient 1 relapsed ?

Censoring and quantities of interest (continuous case)

Let

- T be the duration, with survival function $\bar{F}$ and p.d.f. f
- and C a positive r.v., independent of T, with survival function $\bar{G}$ and p.d.f. g

We observe

$$T^C = \min(T, C) \text{ and } \delta = \mathbb{1}_{T \leq C}$$

Key relationships for the likelihood

We have, in the continuous case,

$$\frac{d\mathbb{P}(T^C \leq t, \delta = 1)}{dt} = f(t)\bar{G}(t) \qquad \frac{d\mathbb{P}(T^C \leq t, \delta = 0)}{dt} = g(t)\bar{F}(t)$$

Exercise

Show the two relationships.

Likelihood

Suppose that we observe, for n independent individuals, independently right-censored data:

$$(T_1^C, \delta_1), (T_2^C, \delta_2), \ldots, (T_n^C, \delta_n).$$

Likelihood (continuous case)

The likelihood is defined as:

$$\begin{aligned}\mathcal{L}((T_1^C, \delta_1), (T_2^C, \delta_2), \ldots, (T_n^C, \delta_n)) &= \prod_{i=1}^{n} \left(f(T_i^C) \bar{G}(T_i^C) \right)^{\delta_i} \left(g(T_i^C) \bar{F}(T_i^C) \right)^{1-\delta_i} \\ &= \underbrace{\prod_{i=1}^{n} f(T_i^C)^{\delta_i} \bar{F}(T_i^C)^{1-\delta_i}}_{\text{part for } f} \underbrace{\prod_{i=1}^{n} \bar{G}(T_i^C)^{\delta_i} g(T_i^C)^{1-\delta_i}}_{\text{part for } g}.\end{aligned}$$

The second line implies that we can estimate f or $\bar{F}$ without any knowledge of the distribution of C !

Exercise: the exponential distribution

Suppose that

- the duration (T) has the distribution $\mathcal{E}(\lambda)$ and
- the right censoring is independent.

Based on the data $(T_1^C, \delta_1), (T_2^C, \delta_2), \ldots, (T_n^C, \delta_n)$, find the maximum likelihood estimator of λ.

Other forms of censoring

Left-censoring

Let T be the duration and C a positive r.v., independent of T. C right-censors T when we observe

$$T^{\mathcal{C}} = \max(T, C) \text{ and } \delta = \mathbb{1}_{T \leq C}$$

instead of T.

Baboon descent - example I.3.7 of Andersen et al. 2012

Baboons sleep in a tree and descend at some time of the day. Observers often arrive later in the day that this descent. In this case, they only know that the descent took place before a certain time.

Exercise: left and right-censoring

In a study of time to first marijuana use (example 1.17 of Klein and Moeschberger 2005) 191 high school boys were asked "when did you first use marijuana?".

- Some answers were "I have used it but cannot recall when the first time was".
- Some never used marijuana at the time of the study.
- Some remembered when they first used it

Which observations are left-censored, which are right-censored ?

Other types of problems of observation

Interval censoring, left- and right-truncation

- **Interval censoring**, when the event of interest is only known to take place in an interval.
- **Left truncation**, when the event of interest is only observed if it is greater than a (left) truncation variable.
- **Right truncation**, when the event of interest is only observed if it is less than a (right) truncation variable.

Death times of elderly residents of a retirement community - example 1.16 of Klein and Moeschberger 2005

Exercise

We observe for 462 residents of a retirement home

- death: Death status (1=dead, 0=alive)
- ageentry: Age of entry into retirement home, months
- age: Age of death or left retirement home, months
- etc

From which problem(s) of observation do these data suffer ?

```
##   death ageentry  age
## 1     1     1042 1172
## 2     1      921 1040
## 3     1      885 1003
## 4     1      901 1018
## 5     1      808  932
## 6     1      915 1004
```

Nonparametric estimation

Case without censoring

We consider a duration T and that we have observed the realizations $t_1 < t_2 < \ldots < t_n$ of i.i.d. copies of T.

Exercise: empirical survival function

- Via the moment method, determine an estimator of the survival function.
- Now consider a r.v. U with values in $\{t_1 < t_2 < \ldots < t_n\}$ such that

$$\mathbb{P}\big(U = t_i\big) = \frac{1}{n} \text{ for all } i \in \{1, \ldots, n\}.$$

  what is its survival functions of U ?
- Conclude that the empirical survival function of T constructed from $T_1, \ldots, T_n$ is the survival function of U.

Case with censoring (1)

We consider

- a discrete duration Z
- a censoring time C, independent of Z

Key relation

For all $t \in \mathbb{R}_+$

$$\begin{aligned}
&\lim_{h \to 0} \mathbb{P}(t \leq Z^C \leq t+h, \delta = 1 | Z^C \geq t) \\
&= \lim_{h \to 0} \frac{\mathbb{P}(t \leq Z^C \leq t+h, \delta = 1)}{\mathbb{P}(Z^C \geq t)} \\
&= \lambda(t).
\end{aligned}$$

We need to find empirical counterparts to

$$Z^C \leq t+h, \delta = 1) \text{ and } \mathbb{P}(D^C \geq t).$$

Case with censoring (2)

Now consider that we have access to realizations of n i.i.d. copies of $(T^C = \min(T, C), \delta = \mathbb{1}_{T \leq C})$.

$$(t_1^C, \delta_1), (t_2^C, \delta_2), \ldots, (t_n^C, \delta_n), \text{ where } t_1^C < t_2^C < \ldots < t_n^C.$$

Consider a vector (U^C, D) of r.v. with values in

$$\{(t_1^C, \delta_1), (t_2^C, \delta_2), \ldots, (t_n^C, \delta_n)\}$$

such that

$$\mathbb{P}\big((U^C, D) = (t_i^C, \delta_i)\big) = \frac{1}{n} \text{ for all } i \in \{1, \ldots, n\}.$$

Let us compute

- $\lim_{h \to 0} \mathbb{P}(t \leq U^C \leq t + h, D = 1)$ and
- $\mathbb{P}(U^C \geq t)$.

Case with censoring (3)

$$\lim_{h \to 0} \mathbb{P}(t_i \leq U^C \leq t_i + h, D = 1) = \mathbb{P}(U^C = t_i, D = 1)$$

$$= \begin{cases} 0 & \text{if } t = t_i \text{ but } \delta_i = 0 \\ \frac{1}{n} & \text{if } t = t_i \text{ and } \delta_i = 1 \end{cases}$$

$$= \frac{\delta_i}{n}$$

and

$$\mathbb{P}(U^C \geq t_i) = \frac{n - (i-1)}{n}$$

Case with censoring (4)

we get the hazard function

$$\begin{aligned}
&\lim_{h \to 0} \frac{\mathbb{P}(t_i \leq U^C \leq t_i + h, D = 1)}{\mathbb{P}(U^C \geq t)} \\
&= \frac{\mathbb{P}(U^C = t_i, D = 1)}{\mathbb{P}(U^C \geq t)} \\
&= \frac{\delta_i}{n - (i - 1)}
\end{aligned}$$

Now, with the relations on slides 11 and 10, we can define the Kaplan-Meier estimator of $\bar{F}$ and Nelson-Aalen estimator of Λ.

The Kaplan-Meier estimator

The Kaplan-Meier estimator (continuous case)

We consider

- a duration T, with survival function $\bar{F}$
- a censoring time C, independent of T
- and that we have access to realizations of n i.i.d. copies of $(T^C = \min(T, C), \delta = \mathbb{1}_{T \leq C})$:

$$\{(t_1^C, \delta_1), (t_2^C, \delta_2), \ldots, (t_n^C, \delta_n)\} \text{ where } t_1^C < t_2^C < \ldots < t_n^C.$$

The Kaplan-Meier estimator of $\bar{F}$ is given by

$$\widehat{\bar{F}}(t) = \begin{cases} \prod_{i:t_i \leq t} (1 - \frac{\delta_i}{n-(i-1)}). & \text{for } t \geq t_1^C \\ 1 & \text{for } t < t_1^C. \end{cases}$$

The Kaplan-Meier estimator is the nonparametric maximum likelihood estimator (so we can trust it !!!).

Example on the re-arrest dataset Singer and Willett 2003 (1)

The dataset contains data for 194 inmates released from a medium-security prison to a maximum of 3 years from the day of their release; during the period of the study, 106 of the released prisoners were rearrested.

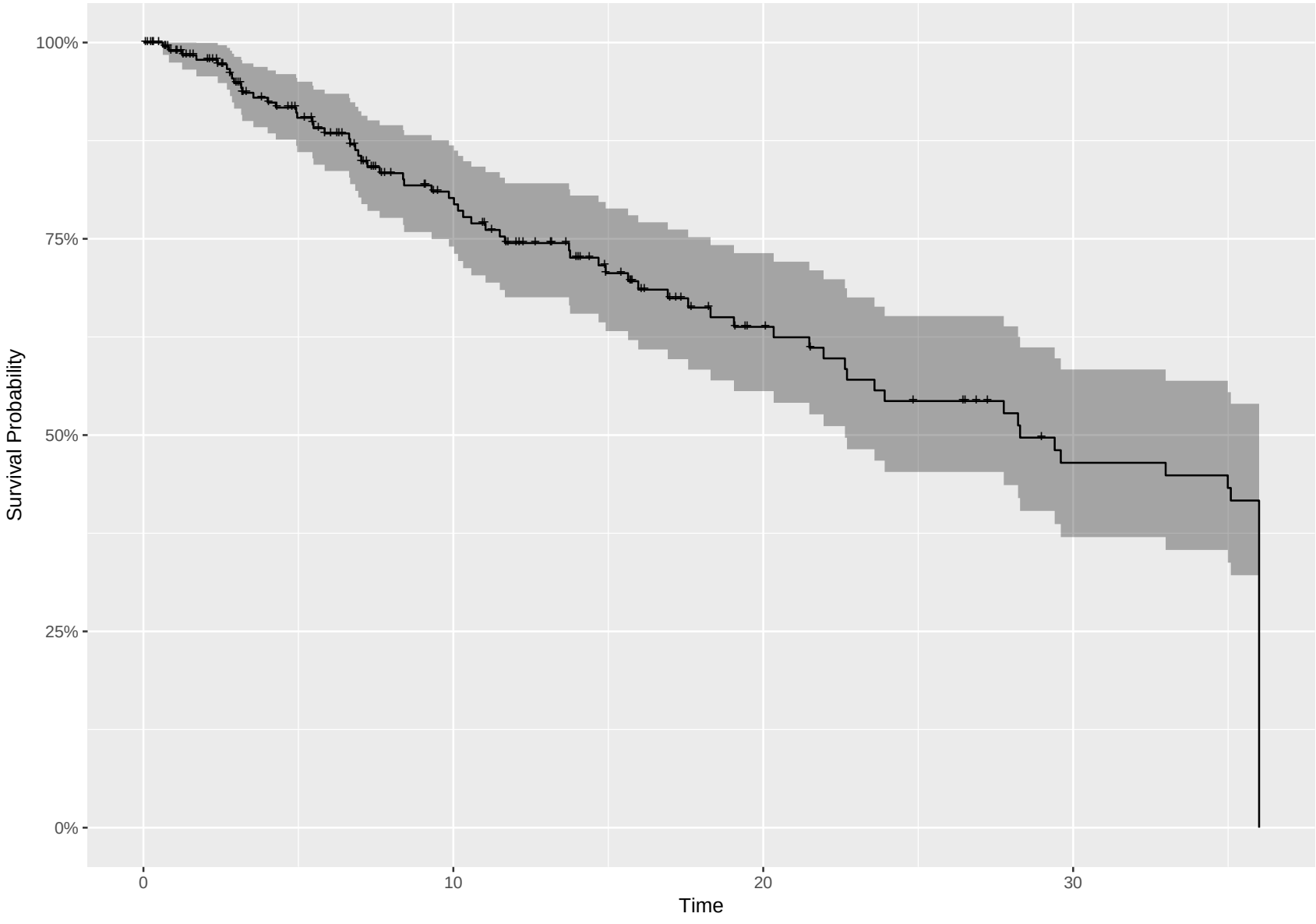
- months: The time of re-arrest in months (but measured to the nearest day).
- censor: A dummy variable coded 1 for censored observations and 0 for uncensored
- etc

```
kmsurvival <- survfit(Surv(months,censor) ~ 1,data=rearrest)

summary(kmsurvival)
```

```
##    time n.risk n.event survival std.err lower 95% CI upper 95% CI
##   0.624    187       1    0.995 0.00533        0.984        1.000
##   0.821    183       1    0.989 0.00758        0.974        1.000
##   1.248    178       1    0.984 0.00936        0.965        1.000
##   1.708    173       1    0.978 0.01090        0.957        1.000
```

Example on the `re-arrest` dataset Singer and Willett 2003 (2)



Example on the pharmocoSmoking dataset of slide 17

```
KM_fit = survfit(Surv(pharmacoSmoking$ttr,pharmacoSmoking$relapse)~1)
summary(KM_fit)
```

```
##  time n.risk n.event survival std.err lower 95% CI upper 95% CI
##     0    125      12    0.904  0.0263        0.854        0.957
##     1    113       5    0.864  0.0307        0.806        0.926
##     2    108       6    0.816  0.0347        0.751        0.887
##     3    102       1    0.808  0.0352        0.742        0.880
##     4    101       3    0.784  0.0368        0.715        0.860
##     5     98       2    0.768  0.0378        0.697        0.846
##     6     96       1    0.760  0.0382        0.689        0.839
```

The Kaplan-Meier estimator (general case)

The Kaplan-Meier estimator

We consider

- a duration T, with survival function $\bar{F}$
- a censoring time C, independent of T
- and that we have access to realizations of n i.i.d. copies of $(T^C = \min(T, C), \delta = \mathbb{1}_{T \leq C})$:

$$\{(t_1^C, \delta_1), (t_2^C, \delta_2), \ldots, (t_n^C, \delta_n)\} \text{ where } t_1^C \leq t_2^C \leq \ldots \leq t_n^C.$$

Let

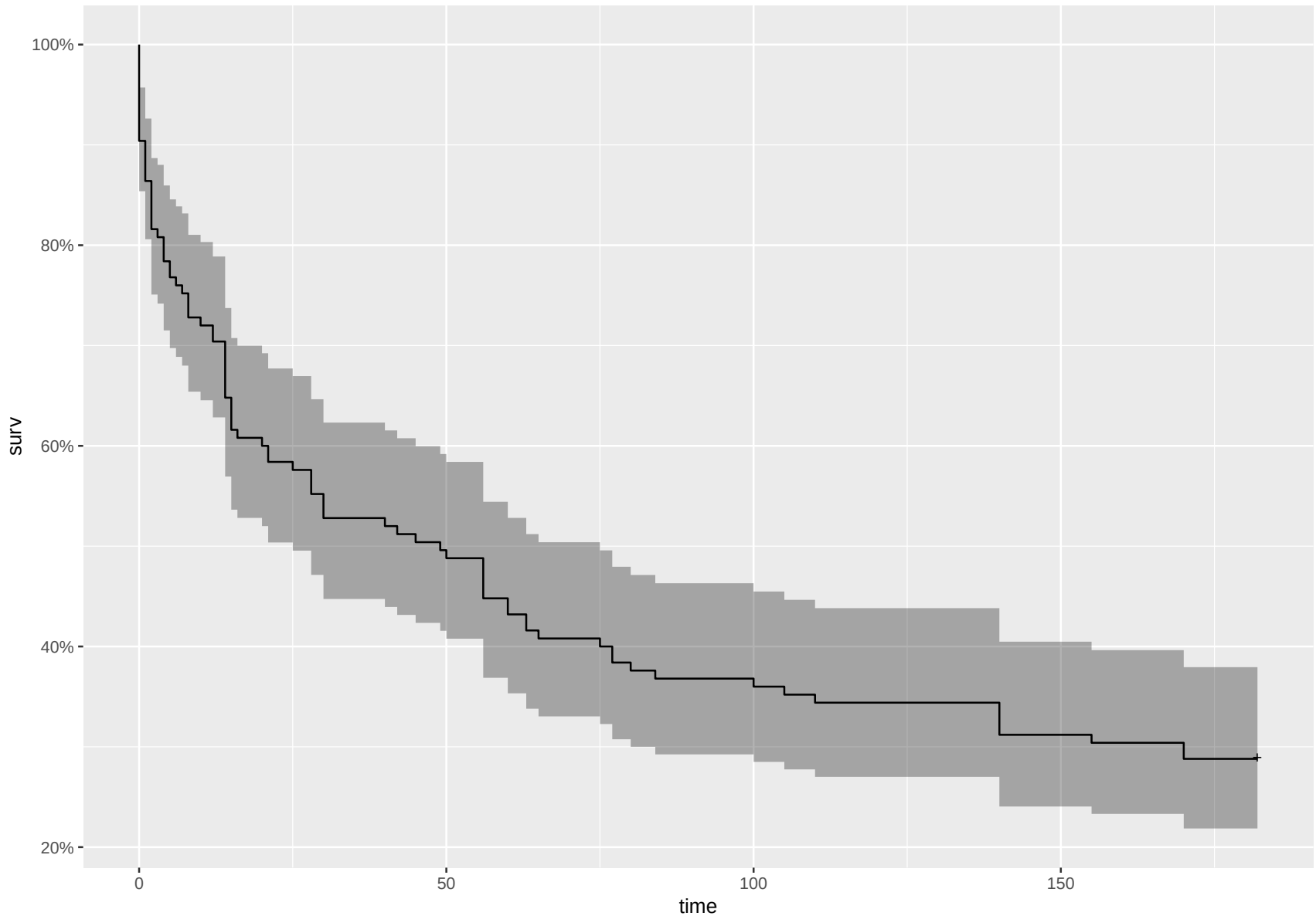
- $\tau_1 < \tau_2 < \ldots < \tau_D$ be the distinct times of event and, for each $k = 1, \ldots, D$
- n_k be the number of observed events at time τ_k
- Y_k be the number of individuals at risk at time τ_k

The Kaplan-Meier estimator of $\bar{F}$ is given by

$$\widehat{\bar{F}}(t) = \begin{cases} \prod_{k:\tau_k \leq t} (1 - \frac{n_k}{Y_k}) & \text{for } t \geq \tau_1 \\ 1 & \text{for } t < \tau_1 \end{cases}$$

Example on the pharmocoSmoking dataset of slide 17

```
autoplot(KM_fit)
```



Variance of the Kaplan-Meier estimator

Greenwood estimator

In the same settings, the Greenwood estimator provides an estimate of the variance of the Kaplan-Meier estimator

$$\widehat{\mathbb{V}}\big(\widehat{\bar{F}}(t)\big) = \widehat{\bar{F}}^2(t) \sum_{k:\tau_k \leq t} \frac{n_k}{Y_k(Y_k - n_k)}$$

The Nelson-Aalen estimator

We consider

- a duration T, with survival function $\bar{F}$ and cumulative intensity function Λ
- a censoring time C, independent of T
- and that we have access to realizations of n i.i.d. copies of $(T^C = \min(T, C), \delta = \mathbb{1}_{T \leq C})$

$$\{(t_1^C, \delta_1), (t_2^C, \delta_2), \ldots, (t_n^C, \delta_n)\} \text{ where } t_1^C \leq t_2^C \leq \ldots \leq t_n^C.$$

Let

- $\tau_1 < \tau_2 < \tau_D$ be the distinct times of event and, for each $k = 1, \ldots, D$
- n_k be the number of observed events at time τ_k
- Y_k be the number of individuals at risk at time τ_k

The Nelson-Aalen estimator of Λ is given by

$$\widehat{\Lambda}(t) = \begin{cases} \sum_{k:\tau_k \leq t} \frac{n_k}{Y_k} & \text{for } t \geq \tau_1 \\ 0 & \text{for } t < \tau_1 \end{cases}$$

Introduction

- What is survival analysis ?
- Parametric distributions for durations
- Quantities of interest

Time-to-event data and censoring

- Definition
- Observations and quantities of interest
- Other forms of censoring

Nonparametric estimation

- Empirical distributions
- The Kaplan-Meier and Nelson-Aalen estimators

References I

Per Kragh Andersen et al. *Statistical models based on counting processes*. Springer Science & Business Media, 2012.

John P Klein and Melvin L Moeschberger. *Survival analysis: techniques for censored and truncated data*. Springer Science & Business Media, 2005.

Dirk F. Moore. "Applied survival analysis using R". In: (2016).

Judith D Singer and John B Willett. *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford university press, 2003.