Survival Analysis, Lecture 1
Introduction to survival analysis

Ruth Keogh

Department of Medical Statistics
London School of Hygiene and Tropical Medicine

Aims

Broadly

To introduce some of the main concepts which underlie the analysis of survival data and set the scene for future lectures

Part 1

- What do we mean by survival analysis
- Features of survival data that mean special methods are needed

Part 2

- Describing distributions of survival times
  - hazard function, survivor function

Part 3

- Outline some particular distributions for survival times with different features
  - exponential distribution, Weibull distribution

What is survival analysis?

Background

What is survival analysis?

Survival analysis is the study of observations which are times at which some outcome or event of interest occurs.

Examples of outcomes of interest in survival analysis:

- Death (all causes)
- Death following a disease diagnosis or following a procedure
- A woman conceiving
- Diagnosis with a disease

Terminology

The time at which the event occurs is referred to as a survival time. The terms failure time and event time are also used.

Studying the patterns of survival in a given population over a particular time scale

For a person born in the UK in a particular year, what is the probability that the person lives to age 5, 40, 100?

Comparing survival times for individuals in two groups, or more generally in several groups

Following a disease diagnosis, do individuals receiving a new treatment have better survival prospects than individuals receiving a standard treatment?

Studying the effects of several continuous and categorical variables on survival times, taking into account possible confounding

How is adult body mass index associated with time to disease diagnosis after controlling for potential confounders?

Predicting future survival based on features of an individual

What is the probability that an individual with features x,y,z will survive 5 years following a particular disease diagnosis?

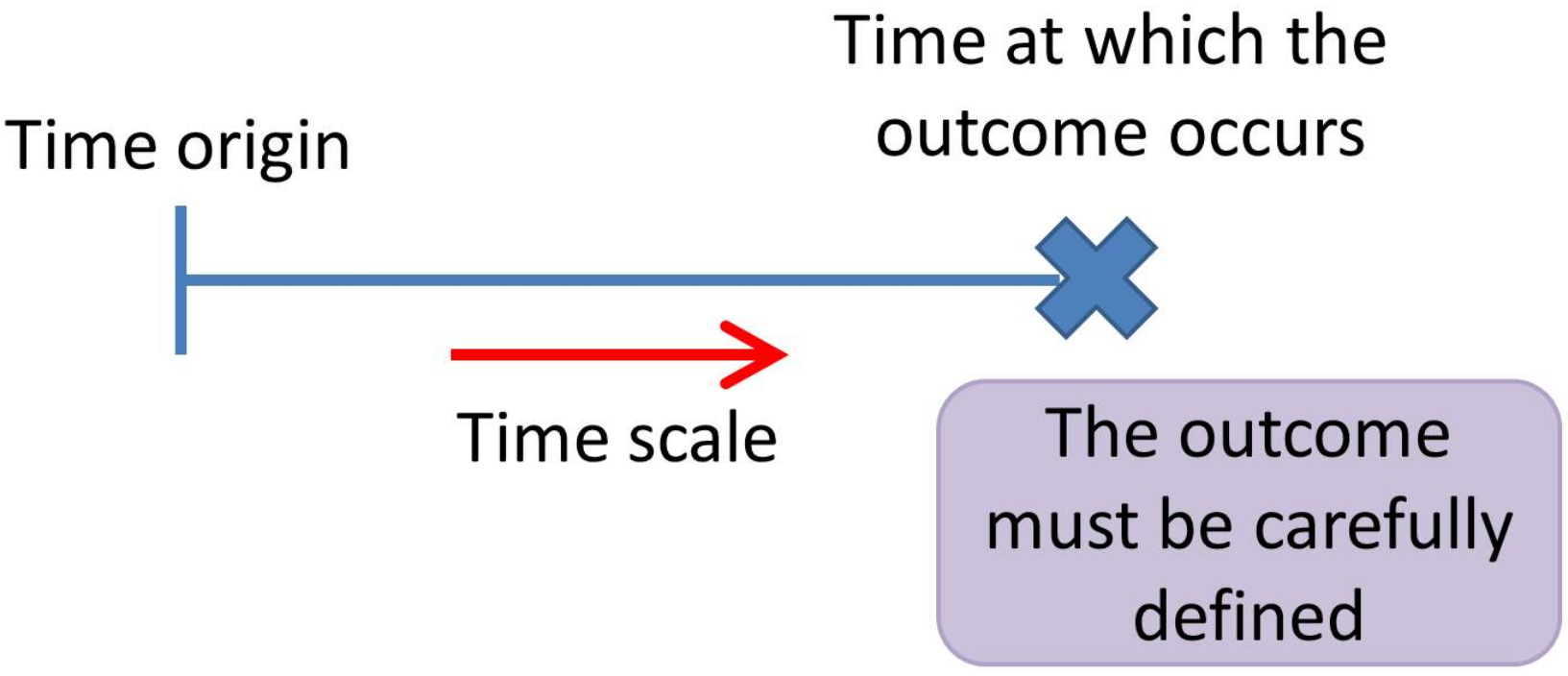
Use of survival analysis in Covid-19 research

- To investigate efficacy of new treatments (Recovery trial)
- To predict risk of severe outcomes according to individual features, e.g. age, comorbidities (QCOVID)
- To study whether new variants are associated with a higher risk of mortality
- To estimate how long patients stay in hospital and ICU

Where do we find survival data?

- In national registers of births and deaths
- In randomized controlled trials, where it is of interest to study whether individuals with a particular medical condition who were randomized to a new treatment tended to live longer than individuals randomized to the standard treatment.
- In prospective observational studies (cohort studies), in which individuals are recruited to a cohort and followed-up for a range of different outcomes.
- In routinely collected data, e.g. electronic health records
- Survival data also arise in a number of non-medical settings
  - in industrial studies, in studies of occurrence of geological or weather-related events.

Essential features of survival data



Example

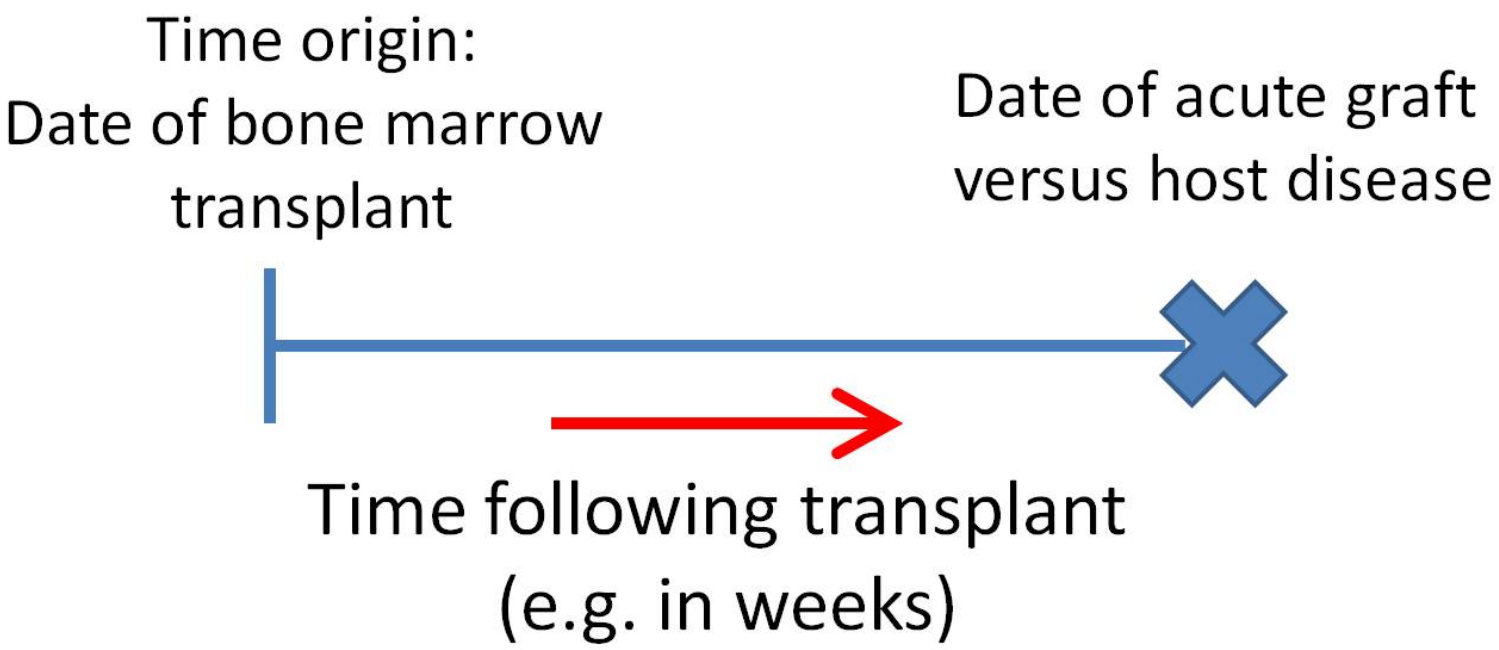
Outcome: Death from all causes

Considering all causes of death in a general population, such as the UK population, the time origin is date of birth and the survival time may be measured by age in years.

Example

Outcome: Acute graft versus host disease following bone marrow transplantation

Individuals who receive a bone marrow transplant are followed up from the date of transplant to the time of acute graft versus host disease



Feature of survival data: censoring

- A particular feature that nearly always arises in survival data from studies of human health is that not all individuals are observed to have the outcome of interest.
- We say that their survival time is ‘censored’.

Administrative censoring

- If individuals in a cohort study are followed up for the outcome ‘death’, then we may wish to perform some analysis of the data before the point at which all members of the cohort have died.

Loss to follow-up

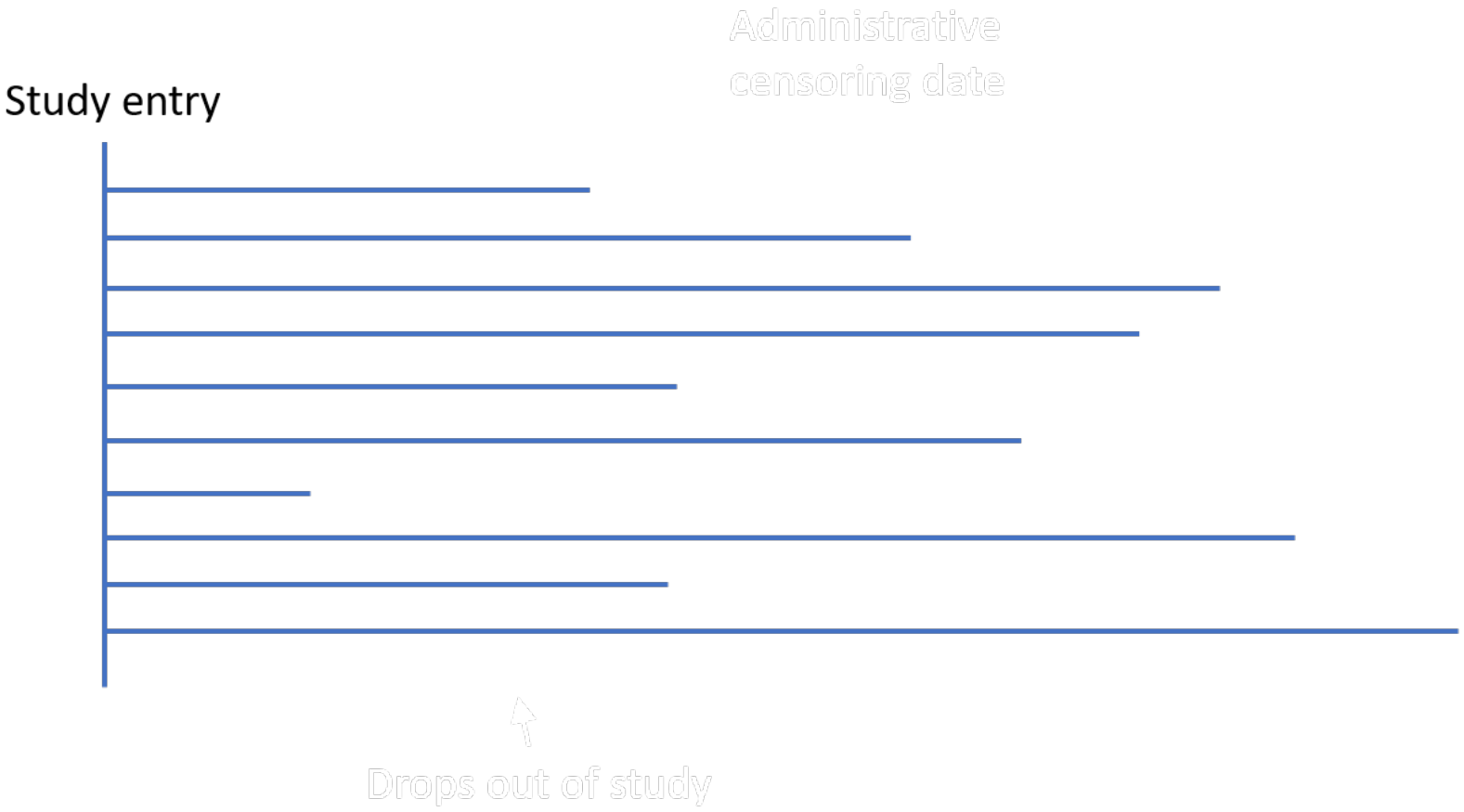
In both intervention studies and observational studies some individuals may be lost to follow-up, meaning that the investigators lose contact with them.

Death from other causes

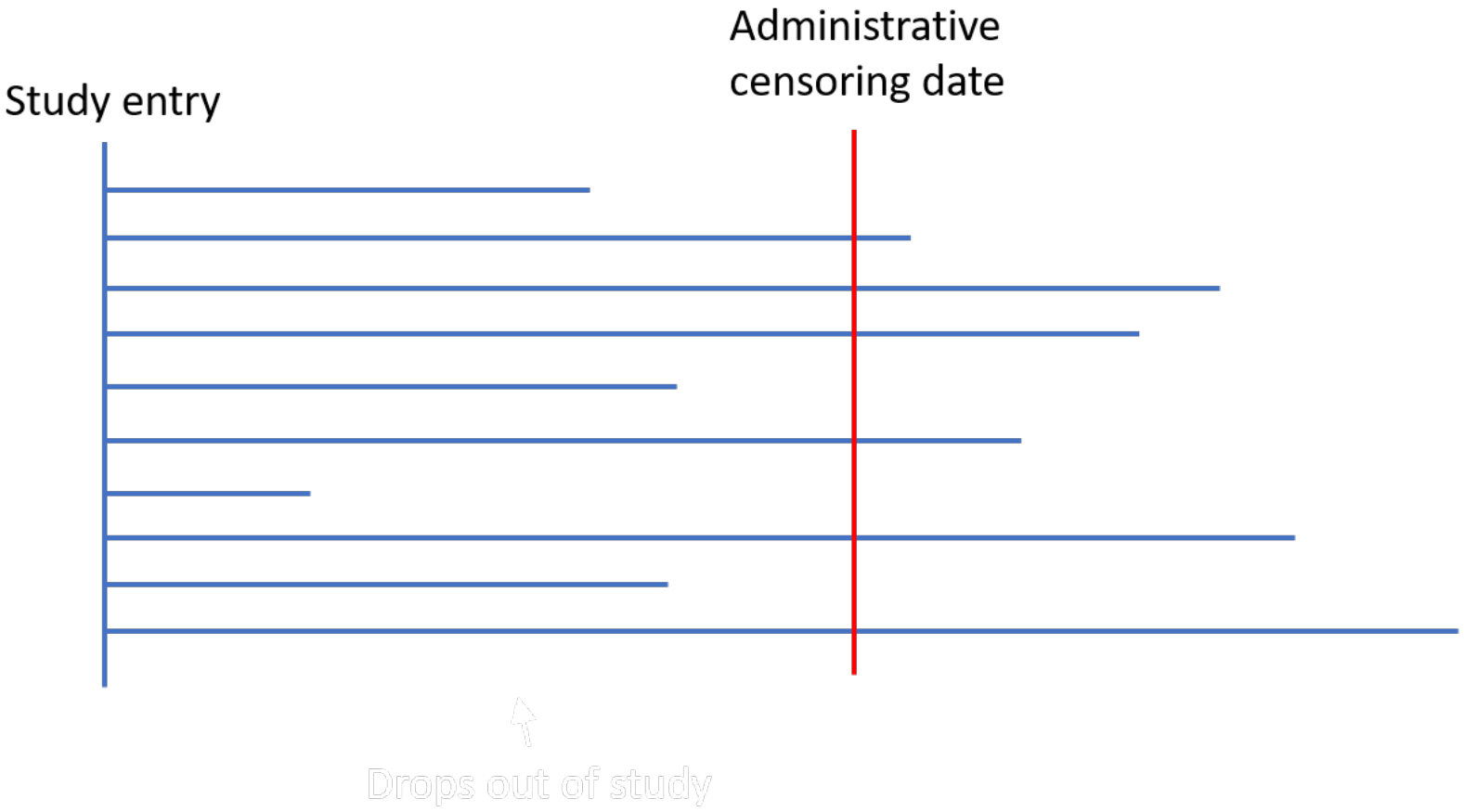
If we are interested in time to disease diagnosis then some (in fact, usually most) individuals in our study population will never be diagnosed with the disease in question and will eventually die of another cause.

In some studies we will have all of these types of censoring

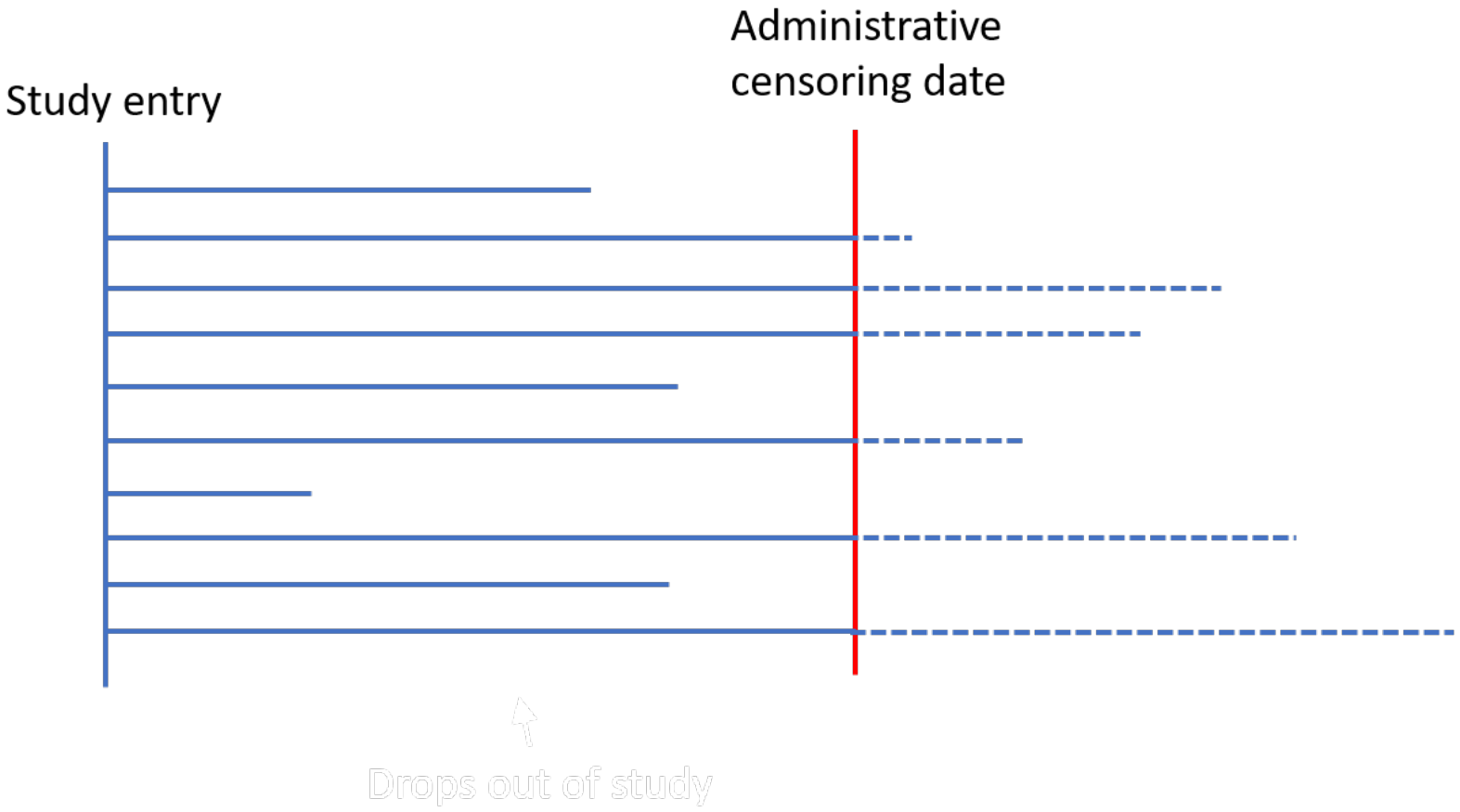
Feature of survival data: censoring



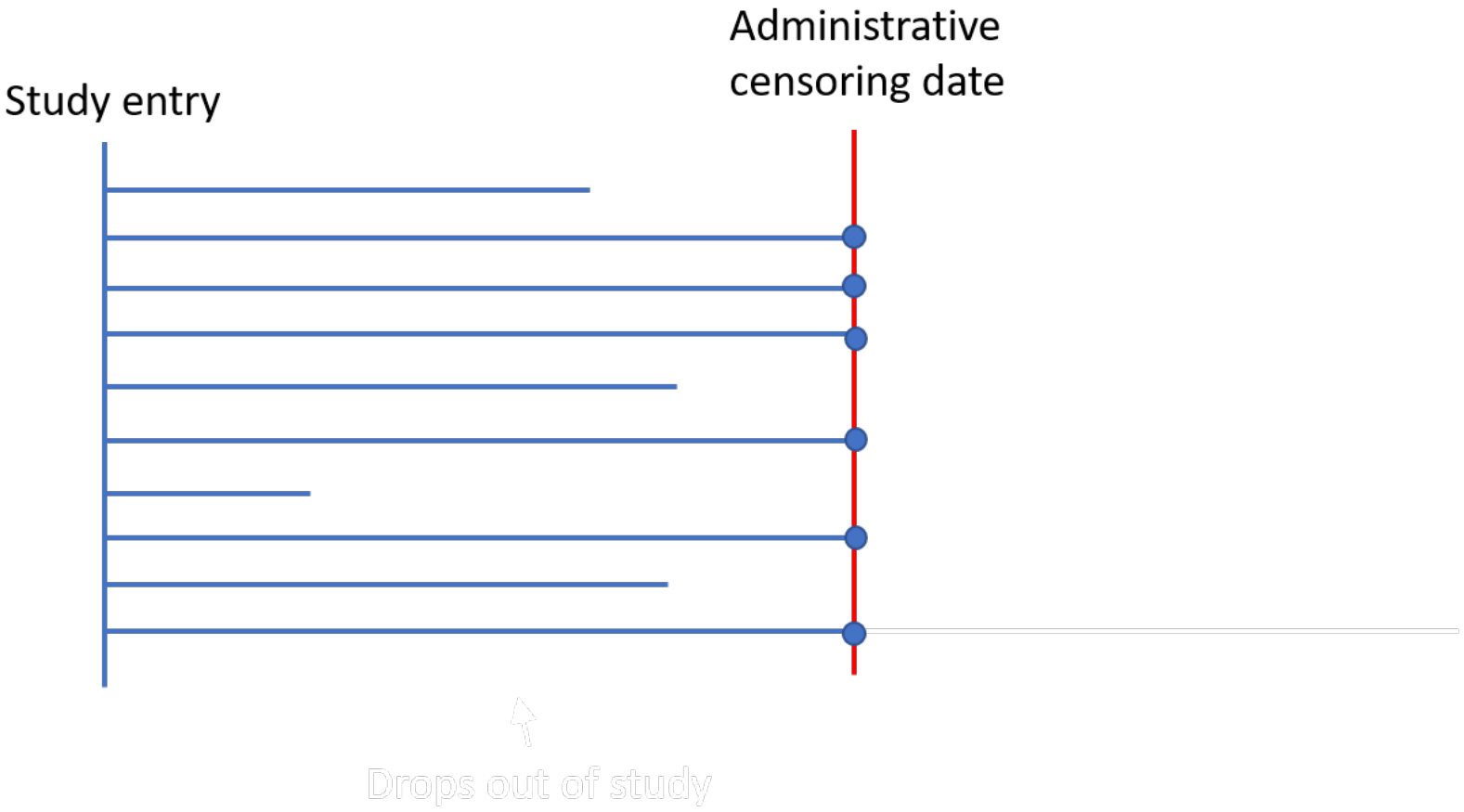
Feature of survival data: censoring



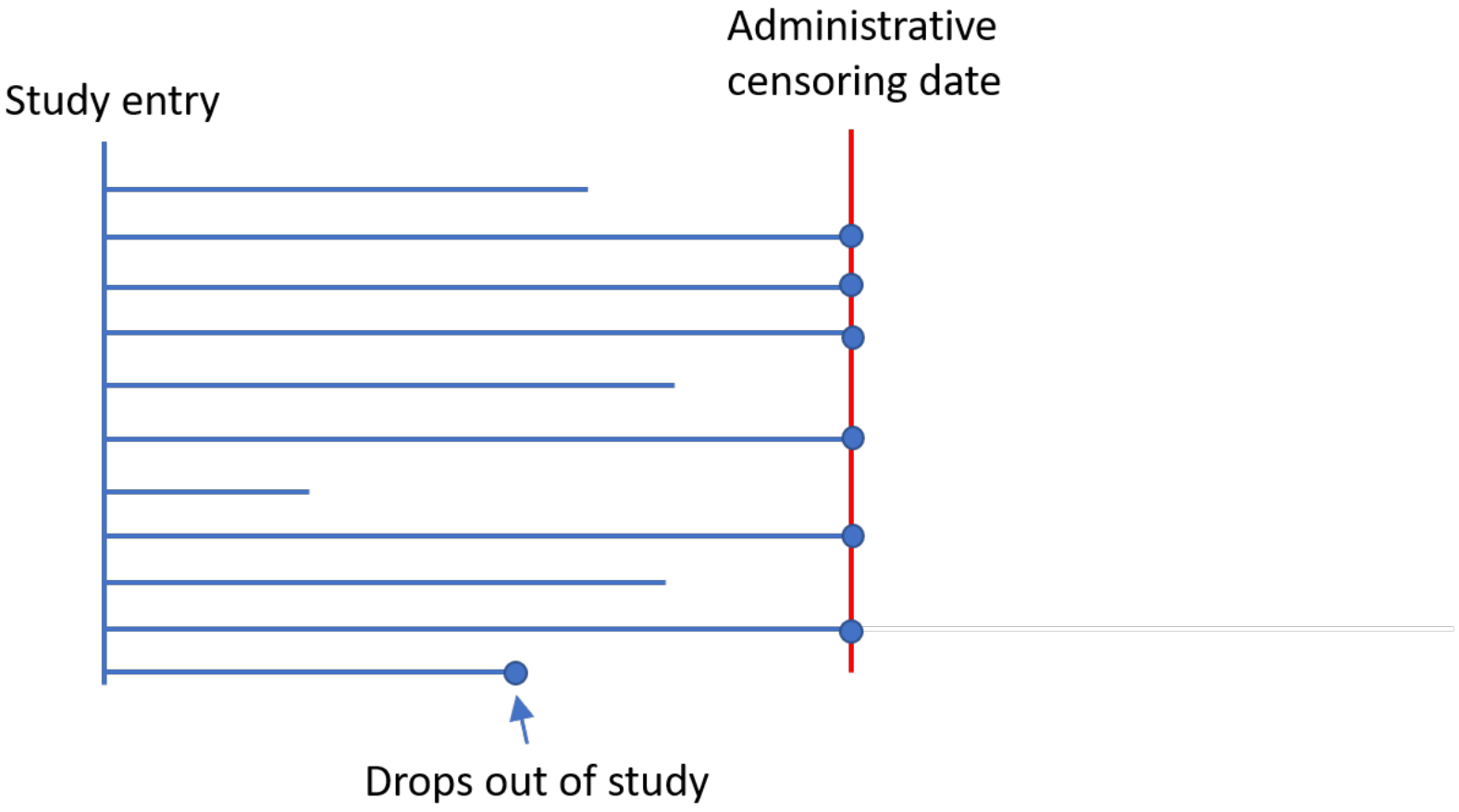
Feature of survival data: censoring



Feature of survival data: censoring



Feature of survival data: censoring



Feature of survival data: censoring

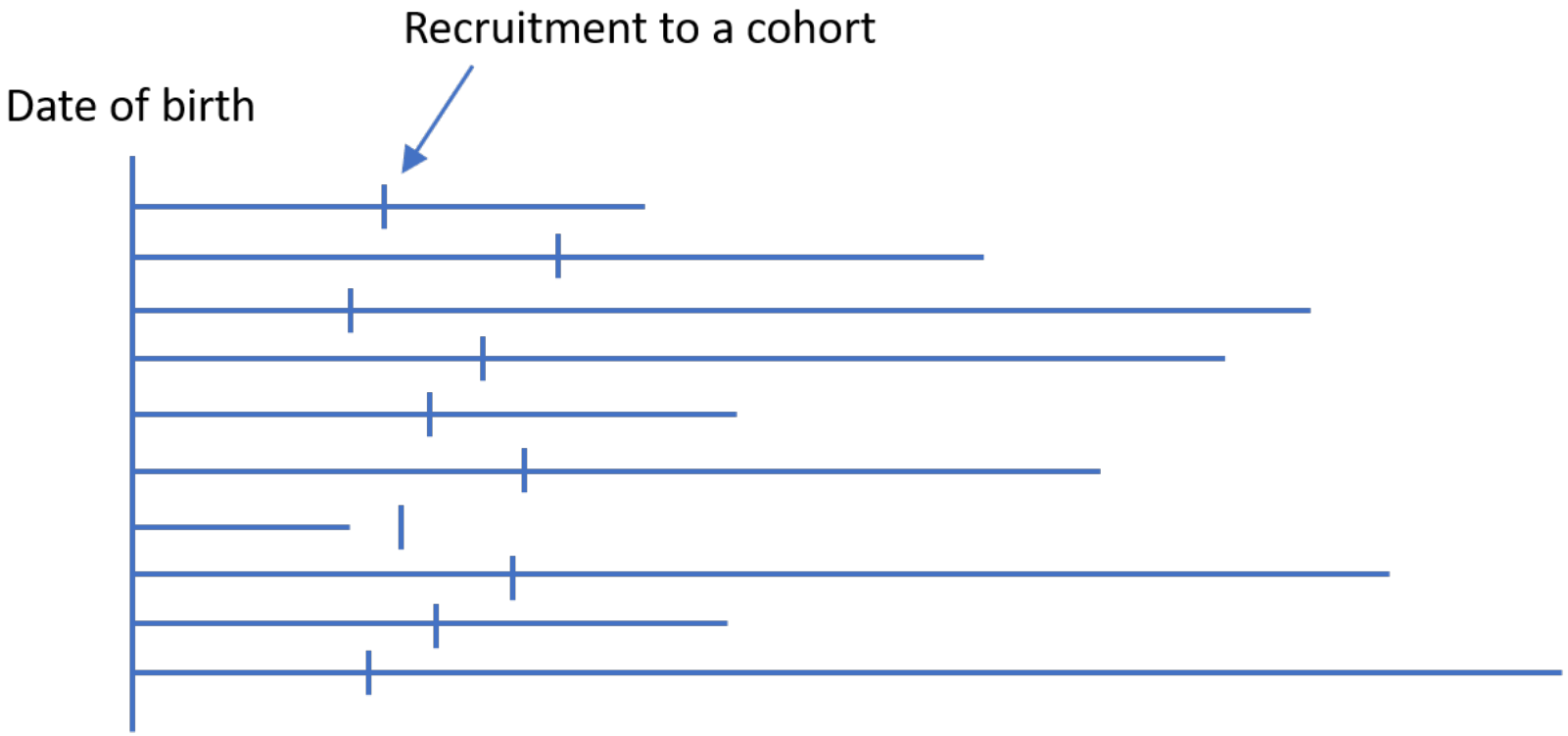
- Instead of observing the time of the outcome for each individual (the survival time), for some individuals we only observe a time up to which we know they have not had the outcome.
- This is referred to as right censoring and it must be accounted for in our analyses.
- The information from censored survival times is still informative - it tells us a person lived at least that long

Important assumption

- It will be assumed throughout this module that censoring is uninformative about event times.
- This means that the time at which an individual is censored, or the fact that they are censored, does not give us any information about when that person may have the event.

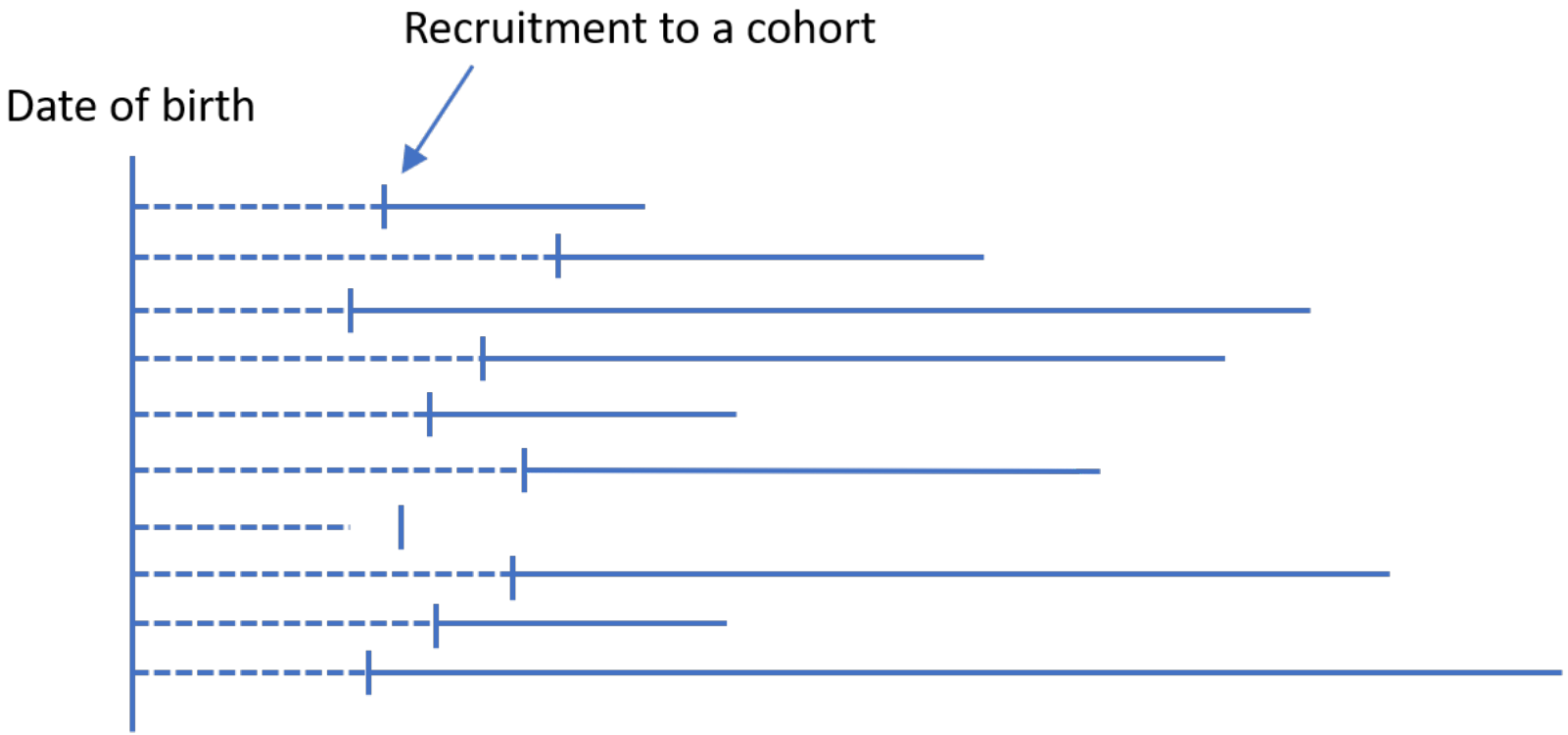
Left-truncation

- Left-truncation is another feature of some survival data
- It is also called delayed entry



Left-truncation

- Left-truncation is another feature of some survival data
- It is also called delayed entry

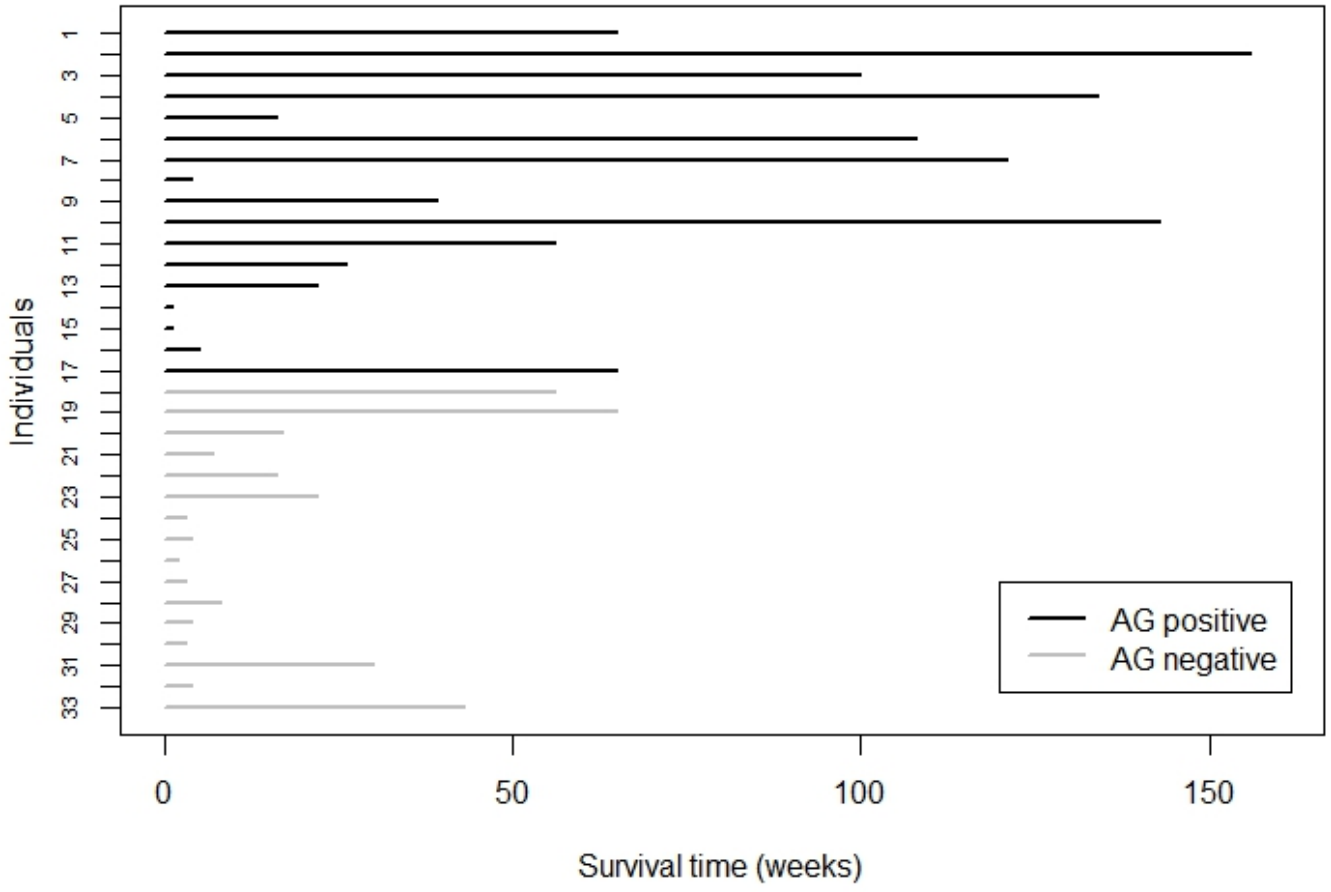


Analysis of survival data

Simple analyses

We can study patterns of observed survival times in a very simple way

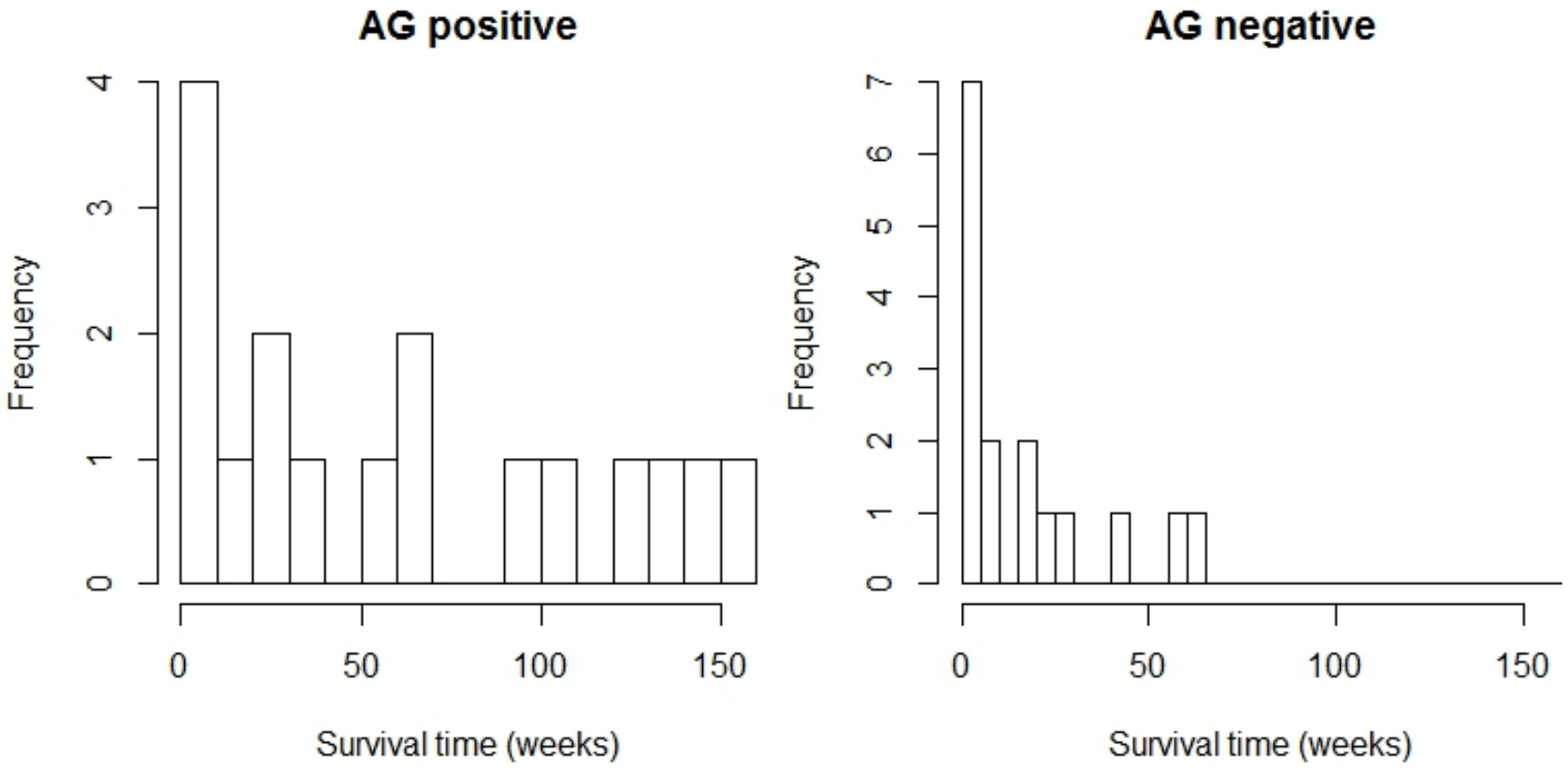
- ...by looking at histograms of survival times
- ... by producing summary statistics, e.g. median survival



Simple analyses

We can study patterns of observed survival times in a very simple way

- ...by looking at histograms of survival times
- ... by producing summary statistics, e.g. median survival



Simple analyses

- But how do we accommodate censoring?
- Perhaps by summarising survival times and censoring times separately
  - range of event and censoring times
  - median event time, median censoring time
- To make progress, we need to be able to describe survival times in a formalised way (coming later)

Simple analyses: incorporating information on individual features

- Simple summaries could be extended to look separately within two or more treatment or exposure groups
- But we want to do better than this
- In general, and in particular to incorporate continuous variables and adjustment for confounders, we want to use regression-based approaches to analysing survival data
- Methods of analysis need to handle censoring and need to allow for the fact that survival times are strictly non-negative.

Analysing survival data

- Non-parametric methods [introduced in Lecture 2]
- Fully parametric methods [introduced in Lecture 3]
- Semi-parametric methods [introduced in Lecture 4]
  - including the Cox proportional hazards model

Describing survival data

Preliminaries

- All of the methods we consider depend on us being able to describe survival data in way which is meaningful for the questions of interest
- We define a random variable T, which represents survival time

Three ways in which the distribution of the random variable T can be described

1. The survivor function.
2. The hazard function.
3. The probability density function.

These are all related.

The survivor function

Definition

The survivor function at a time t is the probability that the survival time T exceeds a value t:

$$S(t) = \Pr(T > t)$$

Example

If T denotes age at death in the UK population, then the survivor function at age 80 ($t = 80$) is

$$S(80) = \Pr(T > 80)$$

Relation to the cumulative distribution function

$$\begin{aligned} F(t) &= \Pr(T \leq t) \\ &= 1 - \Pr(T > t)) = 1 - S(t) \end{aligned}$$

The hazard function

Hazard at time t: Definition in discrete time

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

- The probability that the outcome occurs at time t conditional on survival to time t

Example

- Given that I am alive now at age t what is the probability I will be dead 1 year from now? This is the death rate at age t.

Cumulative hazard

$$H(t) = \sum_{u=0}^{t} h(u)$$

The hazard function

Hazard function: Definition

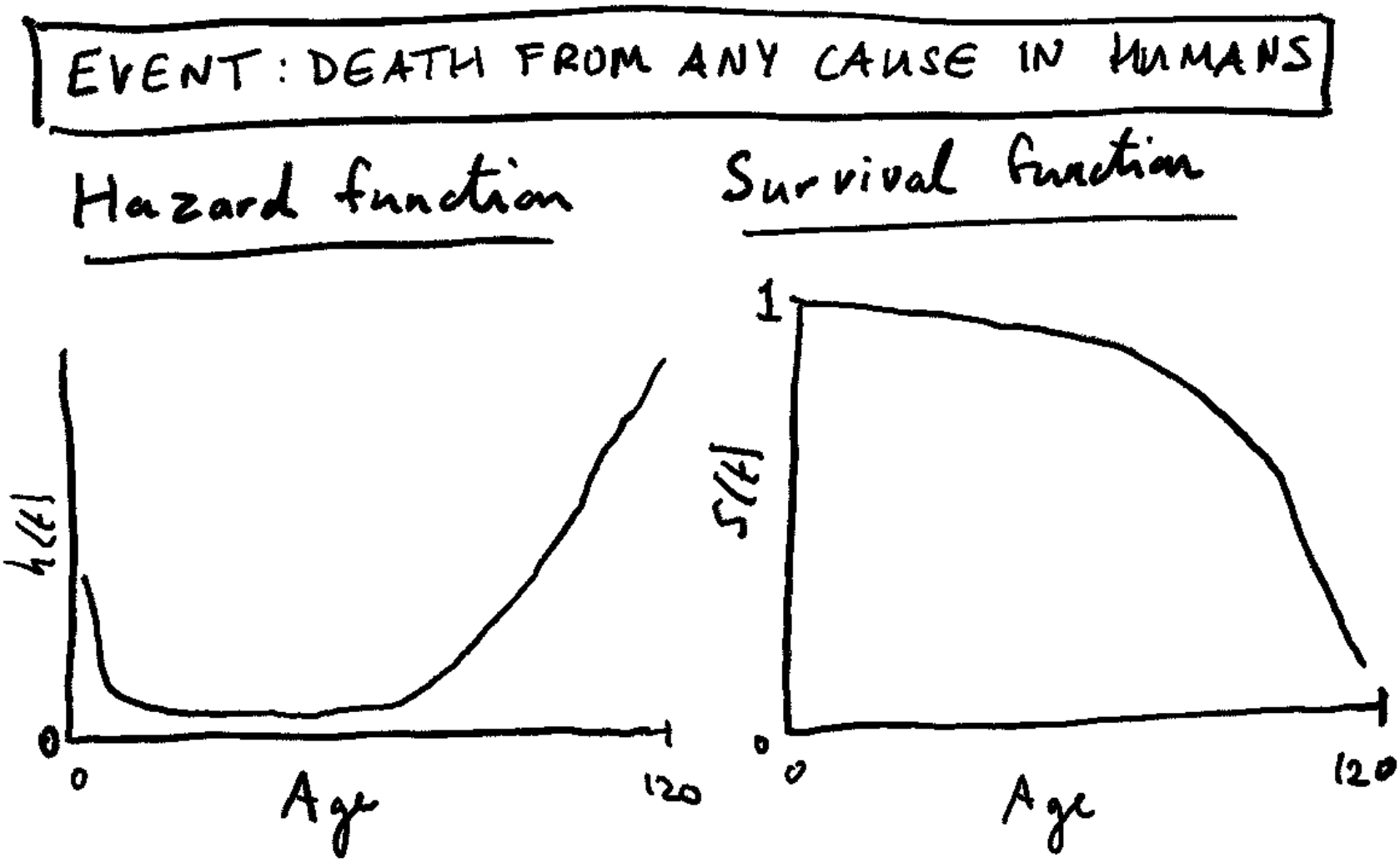
$$h(t) = \lim_{\delta \to 0} \frac{1}{\delta} \Pr(t \leq T < t + \delta | T \geq t)$$

- The hazard function at time t is the probability that the outcome occurs in a short time instant after t given that the outcome did not occur up to time t, divided by the length of time δ, therefore giving a rate.
- The hazard function is defined as δ gets very small so that it becomes as smooth function over time.

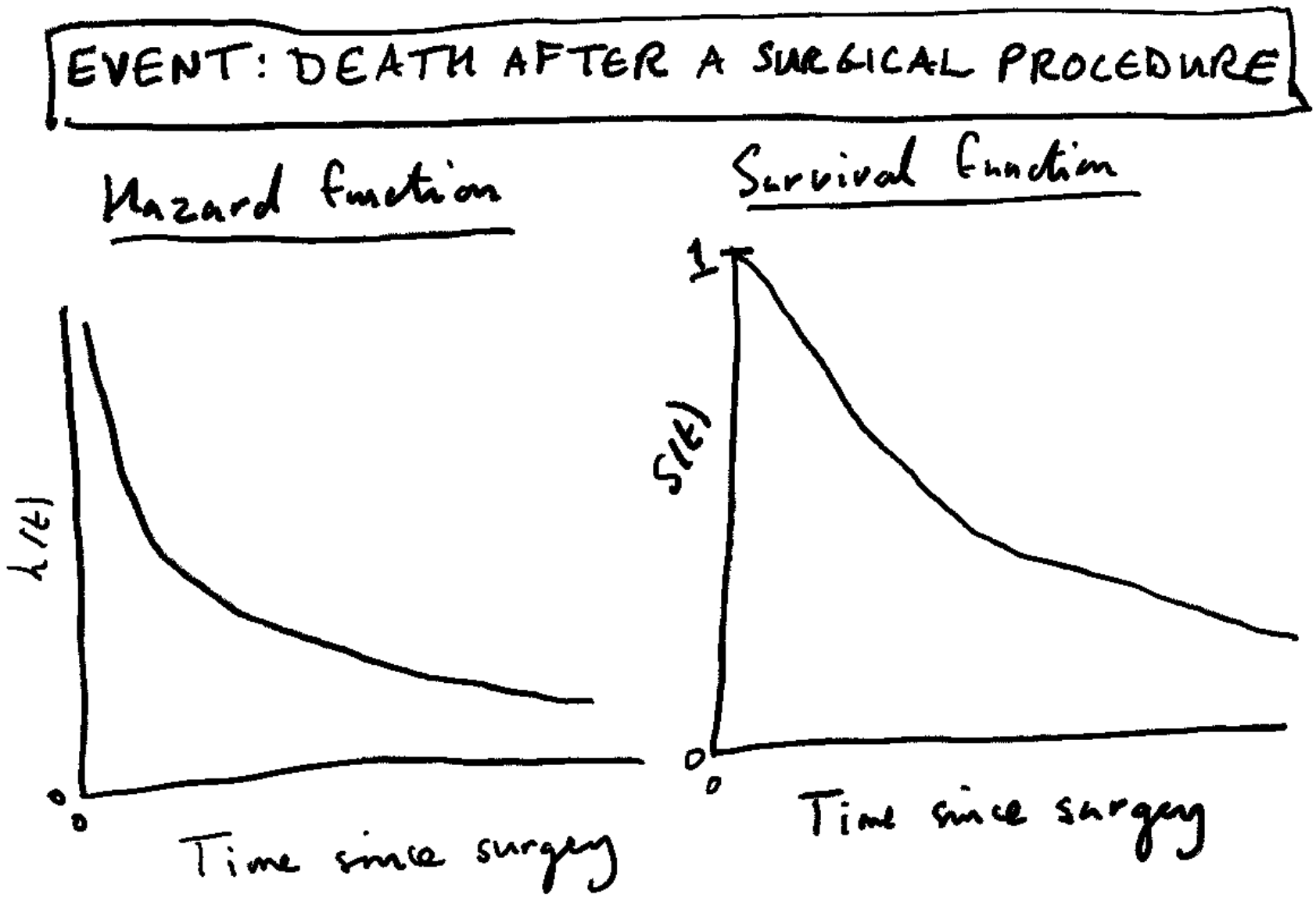
Cumulative hazard

$$H(t) = \int_0^t h(u)du$$

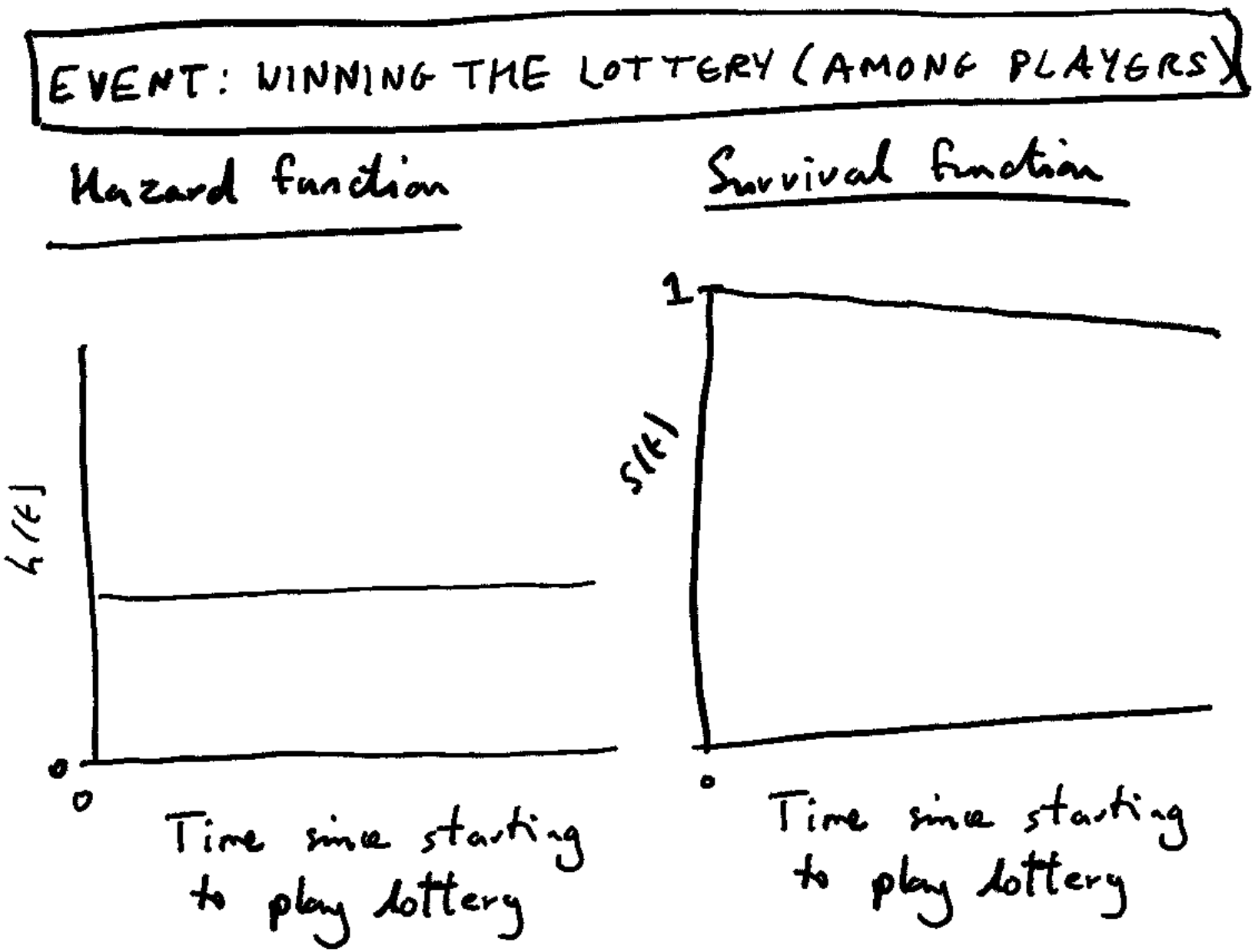
Example


EVENT: DEATH FROM ANY CAUSE IN HUMANS
Hazard function
Survival function
h(t)
0
0
Age
120
S(t)
1
0
0
Age
120

Example


EVENT: DEATH AFTER A SURGICAL PROCEDURE
Hazard function
λ(t)
0
Time since surgery
Survival function
1
S(t)
0
Time since surgery

Example


EVENT: WINNING THE LOTTERY (AMONG PLAYERS)
Hazard function
Survival function
h(t)
0
Time since starting to play lottery
S(t)
1
0
Time since starting to play lottery

The probability density function

Definition

The probability density function at time t is defined as

$$f(t) = \frac{d}{dt}F(t) = \lim_{\delta \to 0} \frac{1}{\delta} \Pr(t \leq T < t + \delta)$$

Relationships between $S(t)$, $h(t)$, $H(t)$, $f(t)$

Describing survival distributions

$$S(t) = \Pr(T > t)$$

$$h(t) = \lim_{\delta \to 0} \frac{1}{\delta} \Pr(t \leq T < t + \delta | T \geq t)$$

$$f(t) = \lim_{\delta \to 0} \frac{1}{\delta} \Pr(t \leq T < t + \delta)$$

In future lectures we will see that it is particularly useful in the analysis of survival data to focus on the form of the hazard function.

$$f(t) = -\frac{d}{dt} S(t), \qquad S(t) = \int_0^t f(u) du$$

$$h(t) = f(t)/S(t), \qquad h(t) = -\frac{d}{dt} \log S(t)$$

It is important to get to grips with these relationships as we will be using them quite a lot during the module

Parametric distributions of survival times

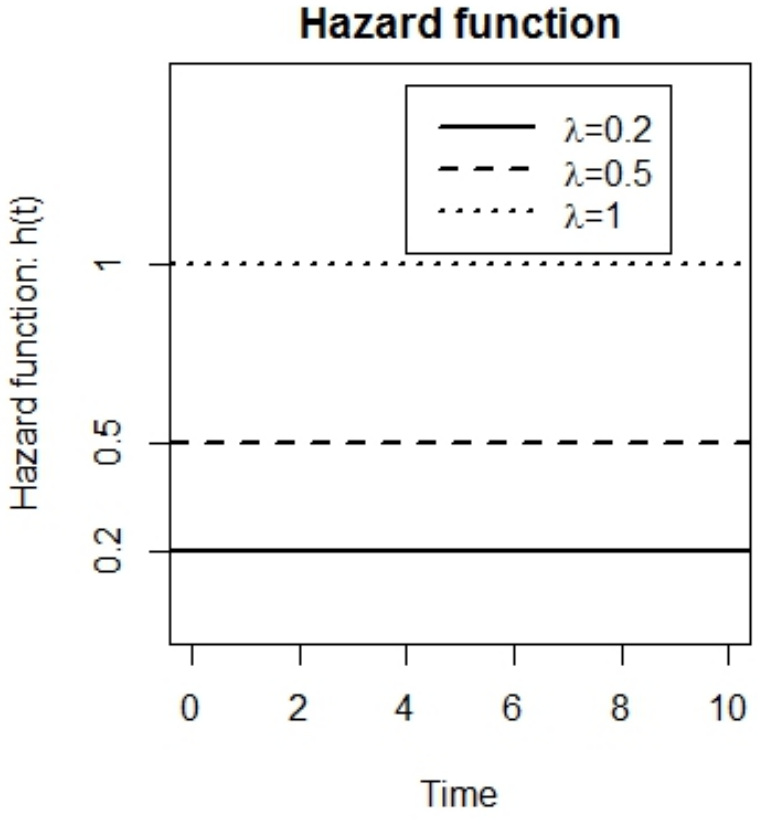
Next, we learn about some functions that can be attached to the survivor, hazard and probability density functions.

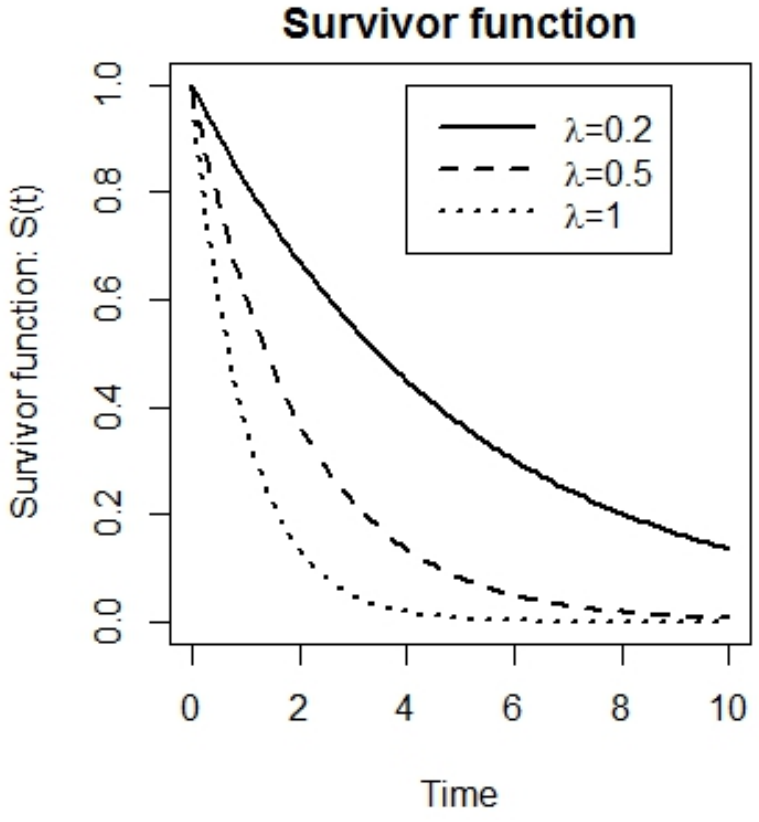
The simplest distribution for survival data

The exponential distribution

Under the exponential distribution the hazard rate is constant over time: λ.

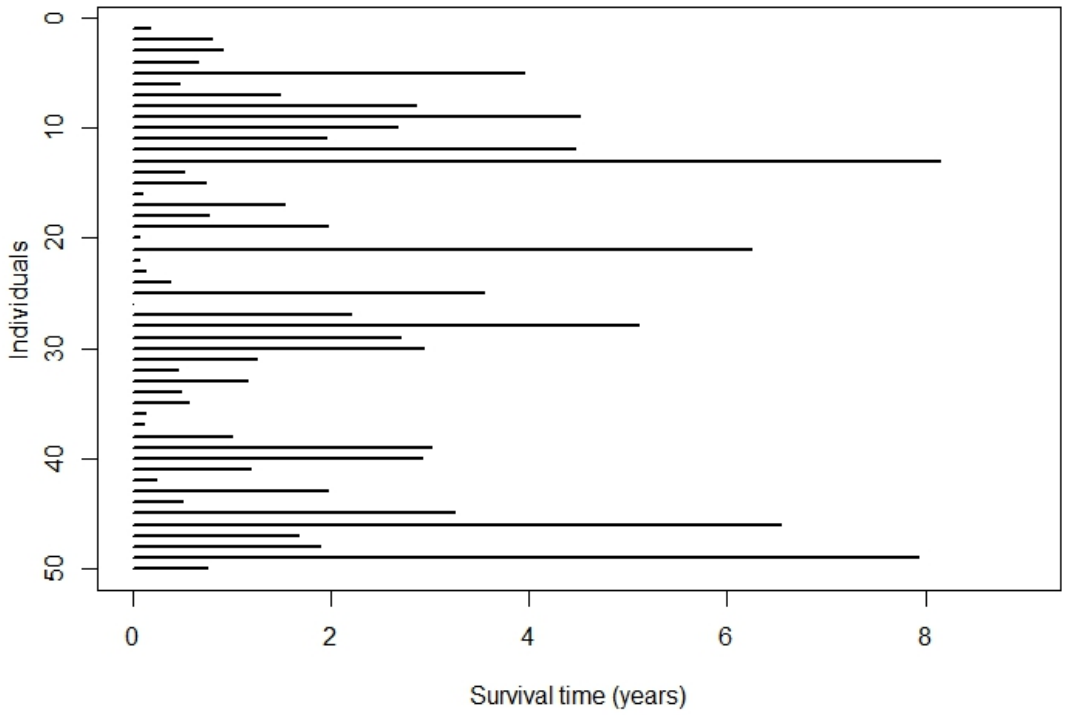
$$h(t) = \lambda, \qquad S(t) = e^{-\lambda t}, \qquad f(t) = \lambda e^{-\lambda t}$$





The exponential distribution

Suppose we assume that the survival times in some particular data set come from an exponential distribution.



We would want to estimate the hazard rate λ and therefore gain knowledge about the hazard and also about the probability of survival to a particular time, via $S(t)$.

Another distribution for survival times

- The exponential distribution will often be unsuitable for studies of human health because we usually expect that the rate at which the outcome occurs will change over time,
- i.e. the hazard function $h(t)$ is not constant
- e.g. the death rate increases with increasing age

The Weibull distribution

$$h(t) = \lambda\kappa t^{\kappa-1}$$

$$S(t) = exp\{-\lambda t^{\kappa}\}$$

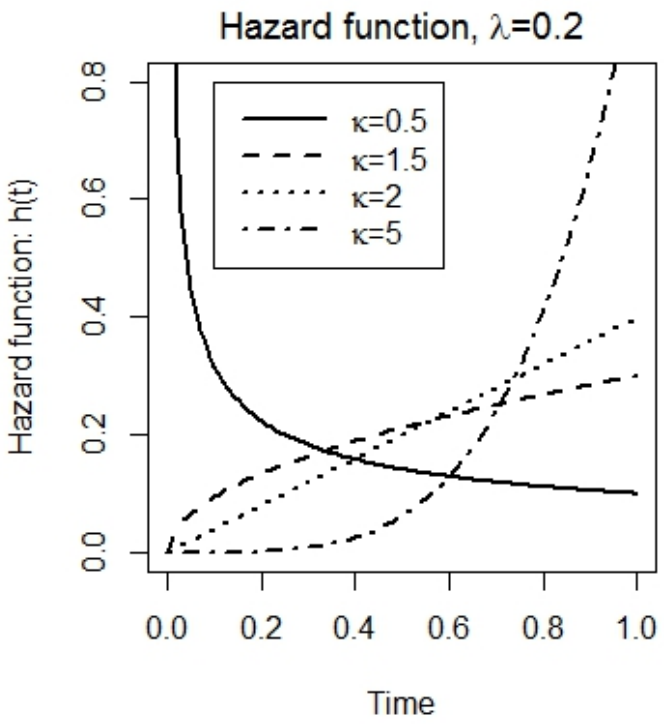
$$f(t) = \lambda\kappa t^{\kappa-1} exp\{-\lambda t^{\kappa}\}$$

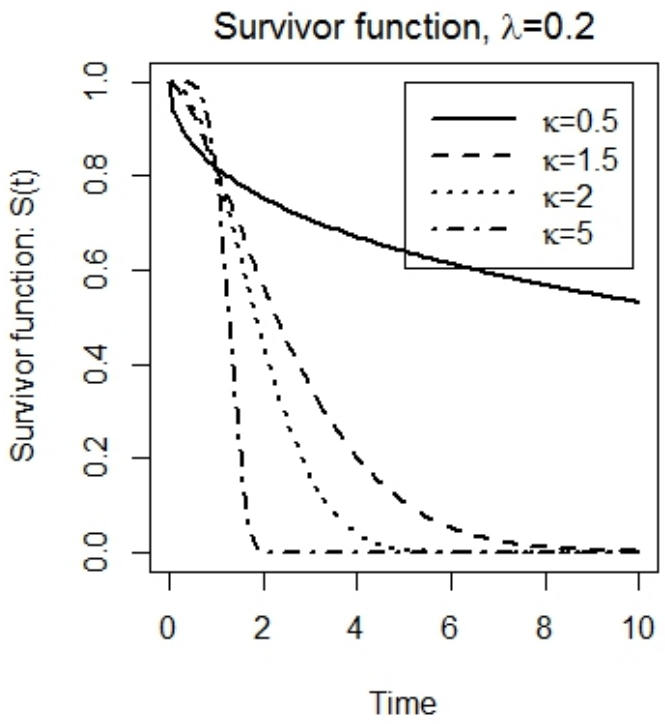
The Weibull distribution

The Weibull distribution

$$h(t) = \lambda\kappa t^{\kappa-1}, \qquad S(t) = exp\{-\lambda t^{\kappa}\}, \qquad f(t) = \lambda\kappa t^{\kappa-1} exp\{-\lambda t^{\kappa}\}$$

The Weibull distribution has 2 parameters, allowing flexibility in the shape of the hazard function





Other distributions

- There are many other distributions for survival data which have different features
- In the practical you will also look at the log-logistic distribution

Estimating the parameters of survival distributions

- Suppose we have decided on a suitable distribution for our survival data, e.g. exponential, Weibull.
- We can estimate the parameters of the survival distribution by maximum likelihood estimation.

Notation: Population of n individuals ($i = 1 \ldots, n$)

- Some individuals have the outcome of interest and a survival time is observed: t_{Ei}
- Some individuals are censored, and for them we observe the time of censoring: t_{Ci}.

$$\delta_i = \begin{cases} 1 & \text{if } t_{Ei} \text{ observed} \\ 0 & \text{if } t_{Ci} \text{ observed} \end{cases}$$

$$t_i = \begin{cases} t_{Ei} & \text{if } \delta_i = 1 \\ t_{Ci} & \text{if } \delta_i = 0 \end{cases}$$

The data

Individual	Survival or censoring time	Indicator of outcome or censoring
1	t_1	δ_1
2	t_2	δ_2
3	t_3	δ_3
.	.	.
.	.	.
.	.	.
n	t_n	δ_n

The data

Individual	Date of time origin	Date of event or censoring	Indicator of outcome or censoring
1	20Jan2012	05Oct2012	1
2	04Nov2012	31Dec2013	0
.	.	.	.
.	.	.	.
.	.	.	.
n	27Feb2012	31Dec2013	0

Likelihoods for survival data

If we had no censoring to worry about then the likelihood for survival data would take the same form as the likelihoods for continuous data that you are already familiar with:

$$L = \prod_{i=1}^{n} f(t_i)$$

where $f(t_i)$ is the probability density function.

Example: Exponential distribution

$$L = \prod_{i=1}^{n} \lambda e^{-\lambda t_i}$$

Incorporating censoring into the likelihood

- For a censored individual i we observe that the person survived at least up until time t_i
- The probability that their unobserved survival time is beyond t_i is the survivor function

$$S(t_i) = \Pr(T > t_i)$$

Full likelihood for survival data with censoring

$$L = \prod_{\text{survival times}} f(t_{E_i}) \prod_{\text{censoring times}} S(t_{C_i})$$

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

Finding maximum likelihood estimates

MLEs and their estimated variances are found in the usual way

Likelihood:

$$L = \prod_{i=1}^{n} f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

Log likelihood:

$$l = \sum_{i=1}^{n} \delta_i \log f(t_i) + \sum_{i=1}^{n} (1 - \delta_i) \log S(t_i)$$

MLEs:

$$\frac{dl}{d\theta} = 0$$

Future directions for parametric models

- In the practical we will learn how to fit these models in Stata and R
- In future lectures we will learn how to extend these models to incorporate explanatory variables