Lecture 1: Time-to-event analysis

Graeme L. Hickey

Department of Epidemiology and Population Health, Institute of Infection and Global Health, University of Liverpool, UK

27th November 2014

Spatial and Temporal Statistical Modelling for Population Health Sciences

Outline of the day

- Morning: 1-hour lecture + 2-hour computer-lab session on general principles and estimation
- Afternoon: 1-hour lecture + 2-hour computer-lab session on regression

What are time-to-event data?

- Data from any study in which the response from each subject is the time at which an event of interest occurs
- Relevant for analysing data of the sort:
  - Survival time following surgery
  - The length of time from birth to development of calf pneumonia
  - The time taken for a cow to conceive following fertility treatment

What are time-to-event data?

- Data from any study in which the response from each subject is the time at which an event of interest occurs
- Relevant for analysing data of the sort:
  - Survival time following surgery
  - The length of time from birth to development of calf pneumonia
  - The time taken for a cow to conceive following fertility treatment

Analytical approaches

Analytical methods fall into two categories:

Survival analysis

Each subject provides at most one event-time

Recurrent event analysis

Each subject provides a (possibly empty) sequence of event-times, which can be considered ordered or unordered, for different or similar events

This course will focus mainly on survival analysis

Analytical approaches

Analytical methods fall into two categories:

Survival analysis

Each subject provides at most one event-time

Recurrent event analysis

Each subject provides a (possibly empty) sequence of event-times, which can be considered ordered or unordered, for different or similar events

This course will focus mainly on survival analysis

Analytical approaches

- The endpoint for survival analysis does not have to be death or 'failure', nor does it have to be a negative event — it can be a positive event (e.g. time to winning the lottery)[1]

- Always be clear about what the time origin is, e.g.
  - Time of treatment
  - Date of birth

[1] We use terms 'survival' and 'failure' interchangeably regardless of outcome

Example dataset

Observational dataset on survival of 10 renal failure patients receiving peritoneal dialysis (PD). Each row denotes a unique patient. The follow-up time column is the number of days since starting treatment to either death (Status = 1) or censoring (Status = 0). Also shown is the treatment type (CAPD = Continuous Ambulatory PD, APD = Automated PD) and age at the start of dialysis.

Patient ID	Follow-up time (days)	Status	Treatment	Age (years)
1	3444	0	APD	41
2	3499	0	APD	35
3	6230	0	APD	41
4	1324	1	APD	67
5	6230	0	APD	29
6	147	1	CAPD	55
7	709	1	APD	54
8	6230	0	APD	42
9	422	1	CAPD	45
10	5096	0	CAPD	46

Probability

Random variable

Let T denote the time to event, with cumulative distribution function $F(t) = P(T \leq t)$

Survival function

The probability the event occurs **after** time t is $S(t) = 1 - F(t) = P(T > t)$

Lifetime distribution

The probability density function is $f(t) = dF(t)/dt$

Probability

Random variable

Let T denote the time to event, with cumulative distribution function $F(t) = P(T \leq t)$

Survival function

The probability the event occurs **after** time t is $S(t) = 1 - F(t) = P(T > t)$

Lifetime distribution

The probability density function is $f(t) = dF(t)/dt$

Probability

Random variable

Let T denote the time to event, with cumulative distribution function $F(t) = P(T \leq t)$

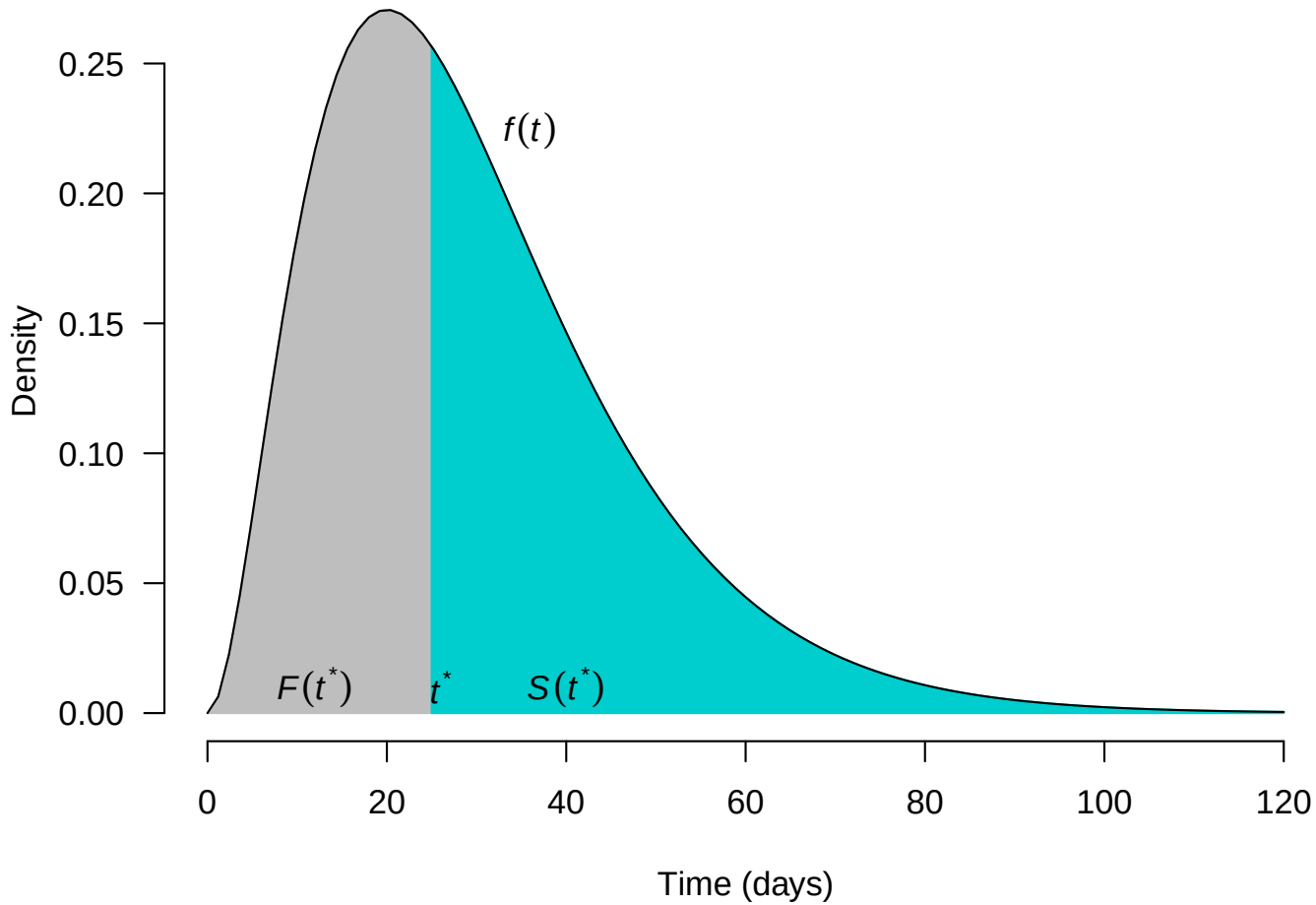
Survival function

The probability the event occurs **after** time t is $S(t) = 1 - F(t) = P(T > t)$

Lifetime distribution

The probability density function is $f(t) = dF(t)/dt$

A hypothetical lifetime distribution [density] function, $f(t)$. The grey shaded area to the left of t^* denotes $F(t^*)$ — the proportion of subjects who experience an event before time t^*. The cyan shaded area to the right of t^* denotes $S(t^*)$ — the proportion of subjects who have survived to time t^*.



Hazard

Conditional on the subject having survived up until time t, we consider $P(t < T \leq t + \Delta t | T > t)/\Delta t$

Hazard function

If we let $\Delta t \rightarrow 0$, then we get the instantaneous hazard rate

$$h(t) = \frac{f(t)}{S(t)}$$

Cumulative hazard

The area under the hazard function up until time t gives a measure for the risk of failure

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

Hazard

Conditional on the subject having survived up until time t, we consider $P(t < T \leq t + \Delta t | T > t)/\Delta t$

Hazard function

If we let $\Delta t \rightarrow 0$, then we get the instantaneous hazard rate

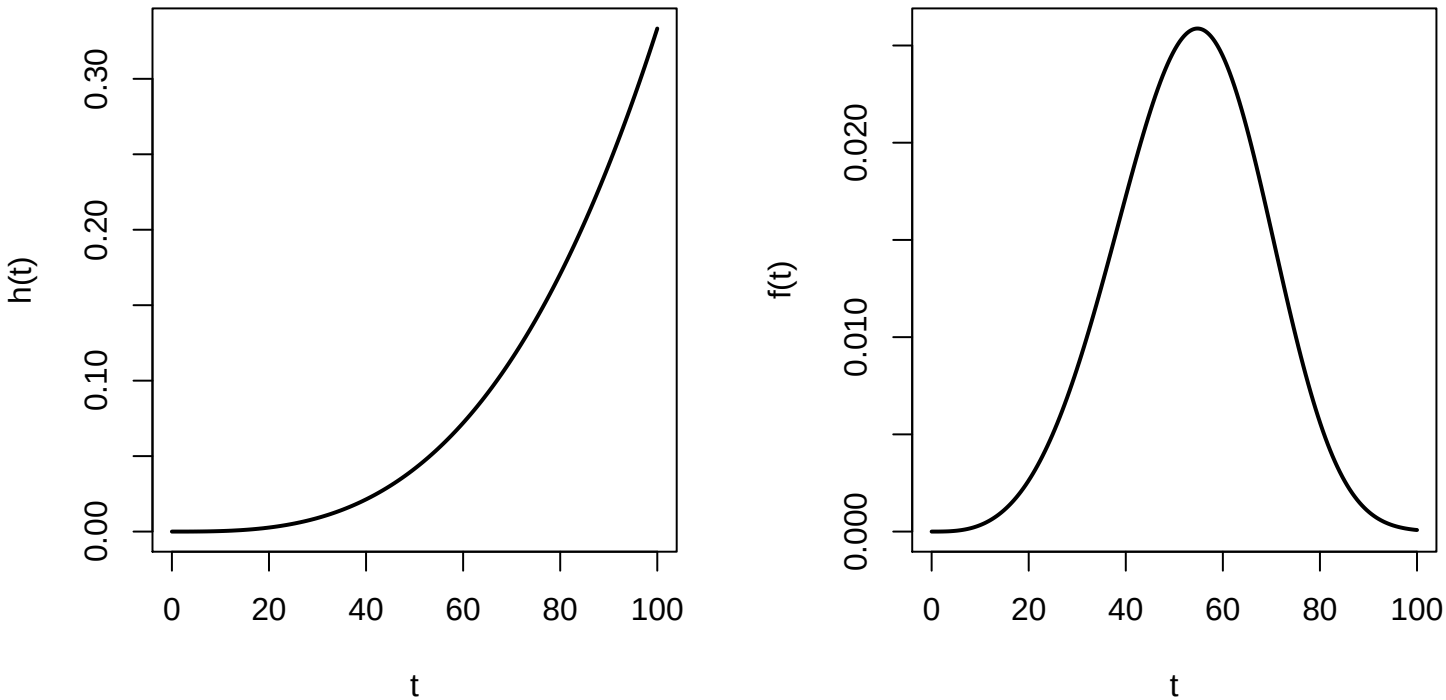
$$h(t) = \frac{f(t)}{S(t)}$$

Cumulative hazard

The area under the hazard function up until time t gives a measure for the risk of failure

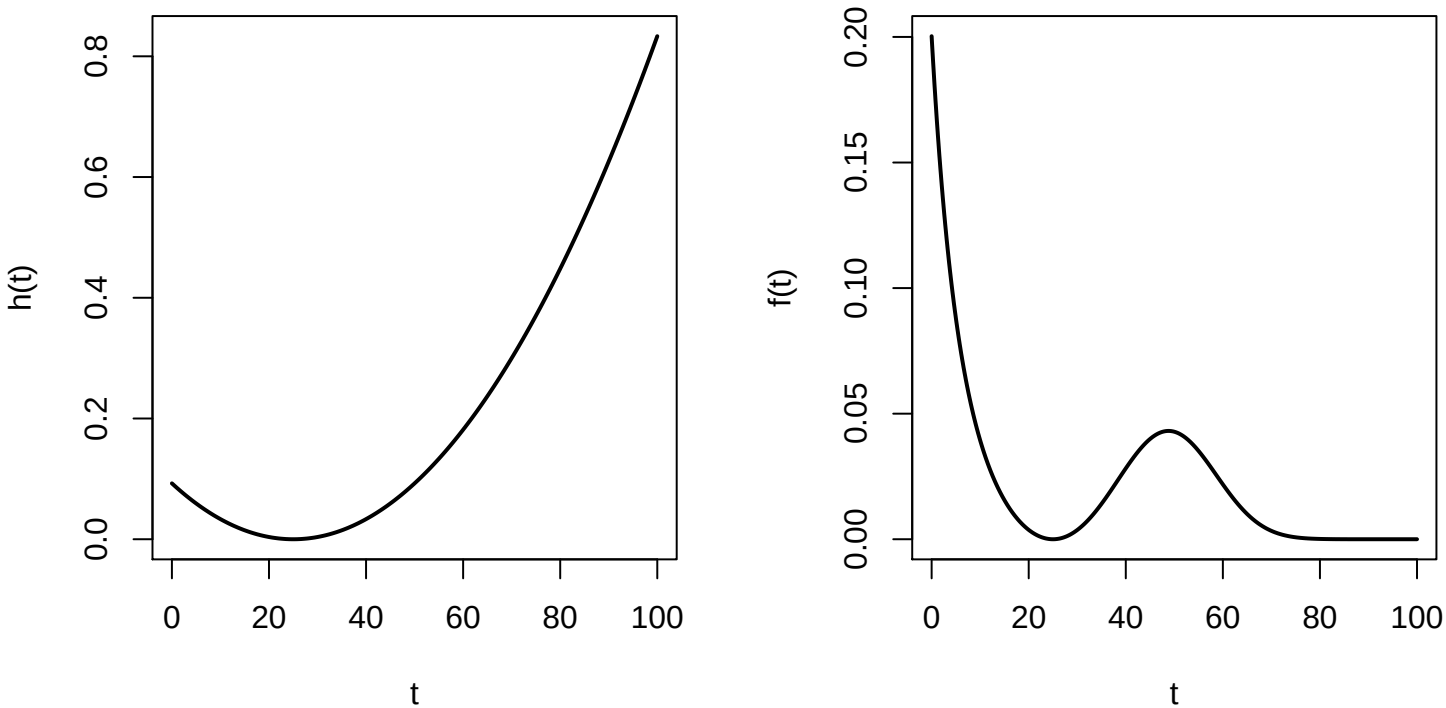
$$H(t) = \int_0^t h(u)du$$

An increasing hazard function $h(t)$ (left panel) and its corresponding lifetime distribution $f(t)$ (right panel)



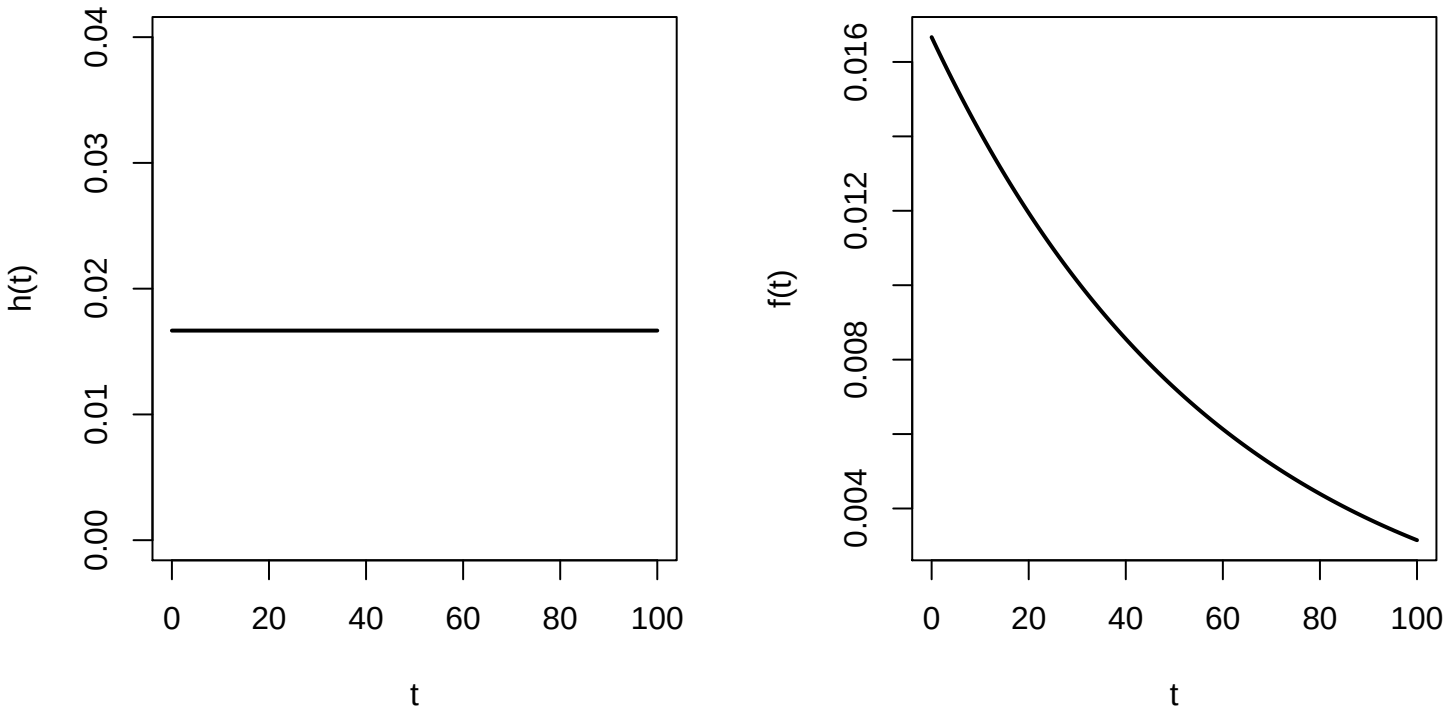
Appropriate for: death rates among adult animals

A decreasing then increasing hazard function $h(t)$ (left panel) and its corresponding lifetime distribution $f(t)$ (right panel)



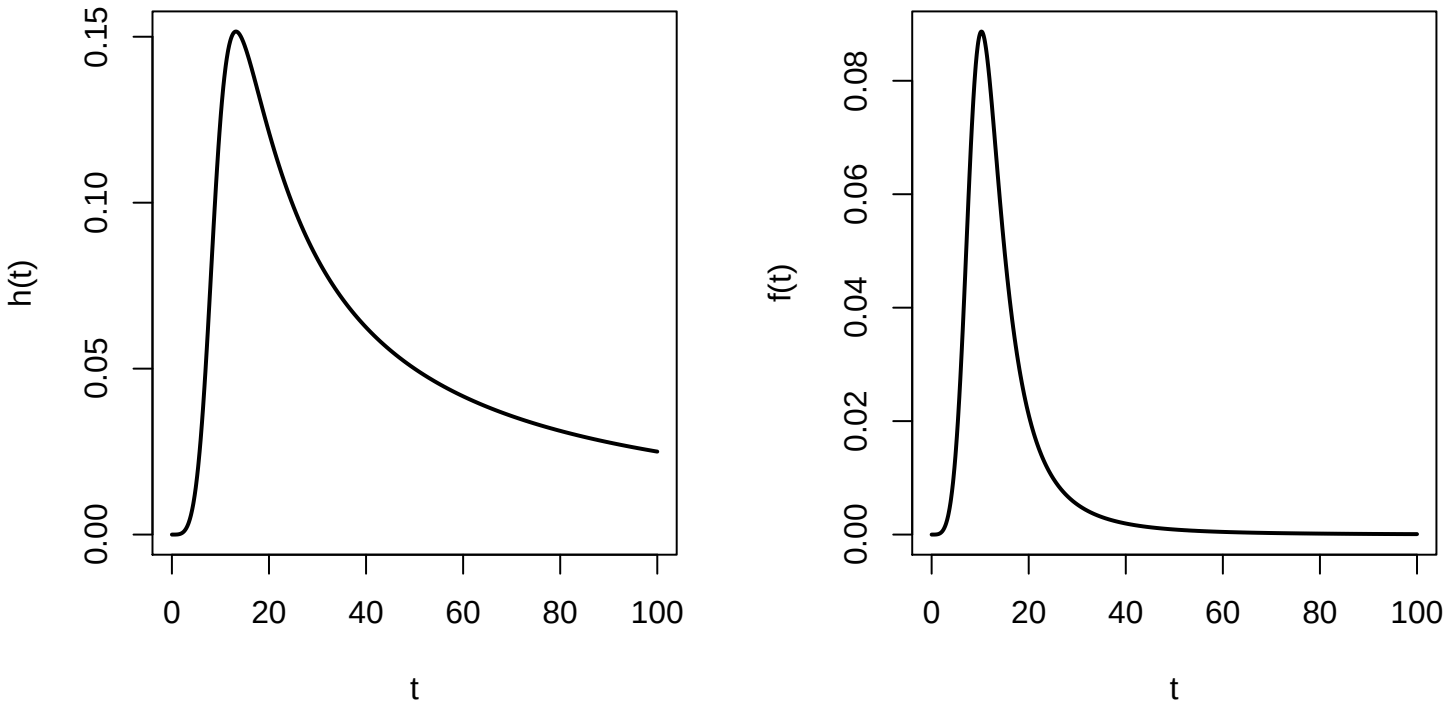
Appropriate for: lifespan of animals ("force of mortality")

A constant hazard function $h(t)$ (left panel) and its corresponding lifetime distribution $f(t)$ (right panel)



Appropriate for: time until next case of influenza in a non-seasonal country

An increasing then decreasing hazard function $h(t)$ (left panel) and its corresponding lifetime distribution $f(t)$ (right panel)



Appropriate for: survival following tuberculosis infection

What if the event does not occur?

Censoring

The time to event is only partially known

Example: time to death following surgery is > 15.4 years

Truncation

A variant of censoring whereby if the time is censored, we do not observe it at all

Example: patients with AIDS are enrolled onto a study to model the time from infection with HIV to development of AIDS, but not everyone infected with HIV has yet developed symptoms of AIDS

What if the event does not occur?

Censoring

The time to event is only partially known

Example: time to death following surgery is > 15.4 years

Truncation

A variant of censoring whereby if the time is censored, we do not observe it at all

Example: patients with AIDS are enrolled onto a study to model the time from infection with HIV to development of AIDS, but not everyone infected with HIV has yet developed symptoms of AIDS

Censoring types

Observed

$T = t_0$: We get to observe the time to the event

Right-censored

$T > t_0$: E.g. when a study specifies a maximum follow-up time

Left-censored

$T < t_0$: E.g. unsure when HIV was contracted

Interval-censored

$t_0 < T < t_1$: E.g. a patient seroconverted between hospital visits

Censoring types

Observed

$T = t_0$: We get to observe the time to the event

Right-censored

$T > t_0$: E.g. when a study specifies a maximum follow-up time

Left-censored

$T < t_0$: E.g. unsure when HIV was contracted

Interval-censored

$t_0 < T < t_1$: E.g. a patient seroconverted between hospital visits

Censoring types

Observed

$T = t_0$: We get to observe the time to the event

Right-censored

$T > t_0$: E.g. when a study specifies a maximum follow-up time

Left-censored

$T < t_0$: E.g. unsure when HIV was contracted

Interval-censored

$t_0 < T < t_1$: E.g. a patient seroconverted between hospital visits

Censoring types

Observed

$T = t_0$: We get to observe the time to the event

Right-censored

$T > t_0$: E.g. when a study specifies a maximum follow-up time

Left-censored

$T < t_0$: E.g. unsure when HIV was contracted

Interval-censored

$t_0 < T < t_1$: E.g. a patient seroconverted between hospital visits

Relevance to epidemiological studies

- Most survival analysis studies specify a maximum follow-up time: subjects still alive at the end of follow-up are right-censored
- Most statistical methods assume that censoring is independent of survival time

Estimation

Methods are classified into two categories:

1. Non-parametric methods
   - Kaplan-Meier estimator
   - Actuarial (life tables) method[2]
   - Nelson-Aalen estimator
2. Parametric modelling

[2] Not discussed in this lecture

Estimation

Methods are classified into two categories:

1. Non-parametric methods
   - Kaplan-Meier estimator
   - Actuarial (life tables) method[2]
   - Nelson-Aalen estimator
2. Parametric modelling

[2] Not discussed in this lecture

Kaplan-Meier method

Kaplan-Meier estimator

An estimator for the survival function is given by

$$\hat{S}(t) = \prod_{i:t_i \leq t} \frac{(n_i - d_i)}{n_i}, \text{ for } 0 < t \leq t_N$$

Where

- $t_1 < t_2 < \cdots < t_N$ are the ordered unique **failure** times (i.e. times for subjects who experienced the event)
- d_i is the number of failures at failure time t_i
- n_i is the number of subjects at risk (i.e. have not yet experienced a failure) just before time t_i

Example

Consider the following fake time-to-event dataset, which we have ordered by time

Subject	Follow-up time	Status
4	9	1
1	13	0
3	15	1
5	35	0
2	49	1

Example

- There are 5 times to consider: 9, 13^+, 15, 35^+, 49
- Only 3 (red) are failure times — we calculate the Kaplan-Meier at these points
- 2 are right-censored — note the use of $+$ superscripts

Time of event (t_i)	Number at risk just before event (n_i)	Number of failures (d_i)	Survival probability $P_i = (n_i - d_i)/n_i$	Cumulative survival $S_i = P_i \times P_{i-1}$
9	5	1	$4/5 = 0.80$	$0.80 \times 1.00 = 0.80$

Example

- There are 5 times to consider: 9, 13^+, 15, 35^+, 49
- Only 3 (red) are failure times — we calculate the Kaplan-Meier at these points
- 2 are right-censored — note the use of $+$ superscripts

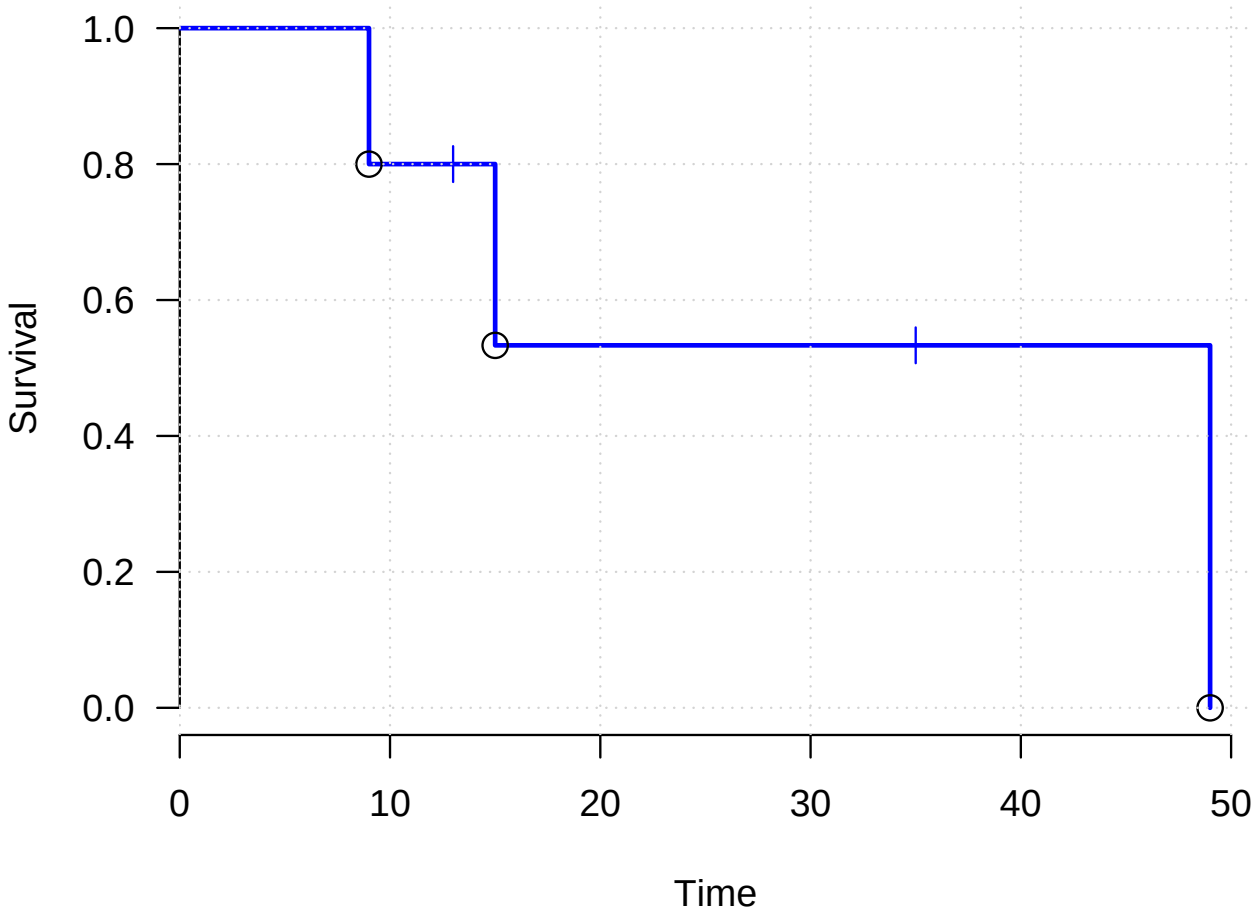
Time of event (t_i)	Number at risk just before event (n_i)	Number of failures (d_i)	Survival probability $P_i = (n_i - d_i)/n_i$	Cumulative survival $S_i = P_i \times P_{i-1}$
9	5	1	$4/5 = 0.80$	$0.80 \times 1.00 = 0.80$
15	3	1	$2/3 = 0.67$	$0.67 \times 0.80 = 0.53$

Example

- There are 5 times to consider: 9, 13^+, 15, 35^+, 49
- Only 3 (red) are failure times — we calculate the Kaplan-Meier at these points
- 2 are right-censored — note the use of $+$ superscripts

Time of event (t_i)	Number at risk just before event (n_i)	Number of failures (d_i)	Survival probability $P_i = (n_i - d_i)/n_i$	Cumulative survival $S_i = P_i \times P_{i-1}$
9	5	1	$4/5 = 0.80$	$0.80 \times 1.00 = 0.80$
15	3	1	$2/3 = 0.67$	$0.67 \times 0.80 = 0.53$
49	1	1	$0/1 = 0.00$	$0.00 \times 0.53 = 0.00$

A Kaplan-Meier curve for our fake data: points denote failure times; ticks denote censoring times



Nelson-Aalen estimator

An estimator of the cumulative hazard is given by the Nelson-Aalen estimator

Nelson-Aalen estimator

$$\hat{H}(t) = \sum_{i:t_i \leq t} \frac{(d_i)}{n_i}, \text{ for } 0 < t \leq t_N$$

We can combine this with the relationship that

$$S(t) = e^{-H(t)}$$

to yield the so-called Flemington-Harrington survival curve estimator

Parametric methods

- As an alternative to non-parametric methods, we might model the survival distribution according to some known parametric model
- Pros ☺
  - Survivorship pattern might be dictated by laws of nature
  - Can be used for prediction
  - Can flexibly incorporate complex structures
- Cons ☹
  - Might mis-specify the model
  - Requires us to think

Parametric methods

- As an alternative to non-parametric methods, we might model the survival distribution according to some known parametric model
- Pros ☺
  - Survivorship pattern might be dictated by laws of nature
  - Can be used for prediction
  - Can flexibly incorporate complex structures
- Cons ☹
  - Might mis-specify the model
  - Requires us to think

Parametric methods

- As an alternative to non-parametric methods, we might model the survival distribution according to some known parametric model
- Pros ☺
  - Survivorship pattern might be dictated by laws of nature
  - Can be used for prediction
  - Can flexibly incorporate complex structures
- Cons ☹
  - Might mis-specify the model
  - Requires us to think

Exponential distribution

Properties

- Lifetime distribution: $f(t) = \lambda \exp(-\lambda t), \ \ \lambda > 0$
- Hazard function: $h(t) = \lambda$
- Survival function: $S(t) = \exp(-\lambda t)$
- Cumulative hazard: $H(t) = \lambda t$

Key features

- Hazard rate is constant
- Memoryless

Exponential distribution

Properties

- Lifetime distribution: $f(t) = \lambda \exp(-\lambda t), \ \ \lambda > 0$
- Hazard function: $h(t) = \lambda$
- Survival function: $S(t) = \exp(-\lambda t)$
- Cumulative hazard: $H(t) = \lambda t$

Key features

- Hazard rate is constant
- Memoryless

Weibull distribution

Properties

- Lifetime distribution: $f(t) = \lambda p t^{p-1} \exp(-\lambda t^p), \ \lambda > 0, \ p > 0$
- Hazard function: $h(t) = \lambda p t^{p-1}$
- Survival function: $S(t) = \exp(-\lambda t^p)$
- Cumulative hazard: $H(t) = \lambda t^p$

Key features

- Parameterised by rate, λ, and shape, p
- Flexible: $p = 1 \rightarrow$ exponential; $p < 1 \rightarrow$ monotonically decreasing hazard with time; $p > 1 \rightarrow$ monotonically increasing hazard with time

Weibull distribution

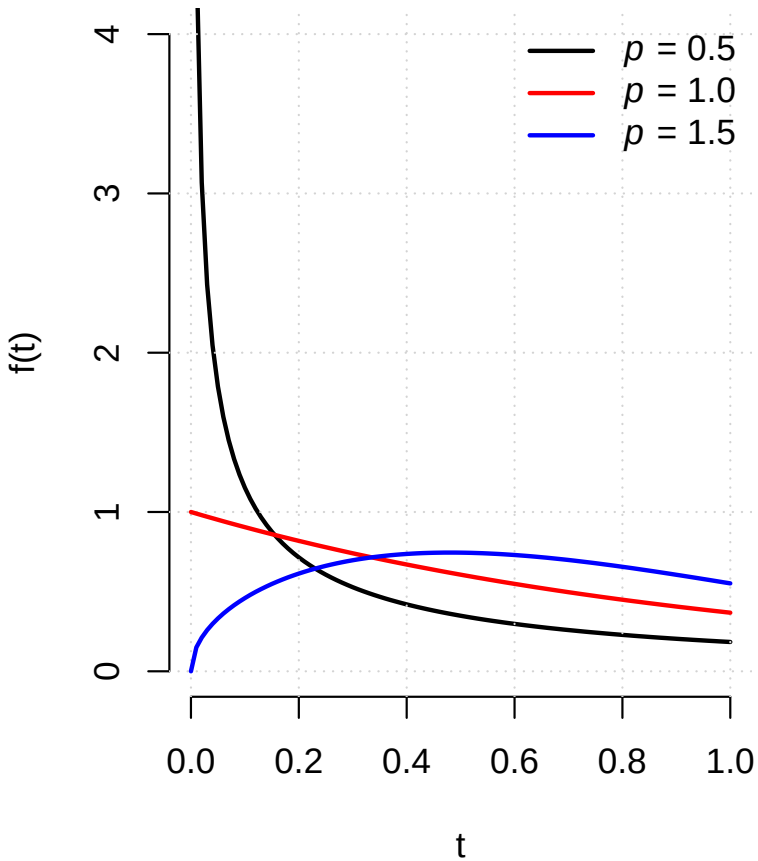
Properties

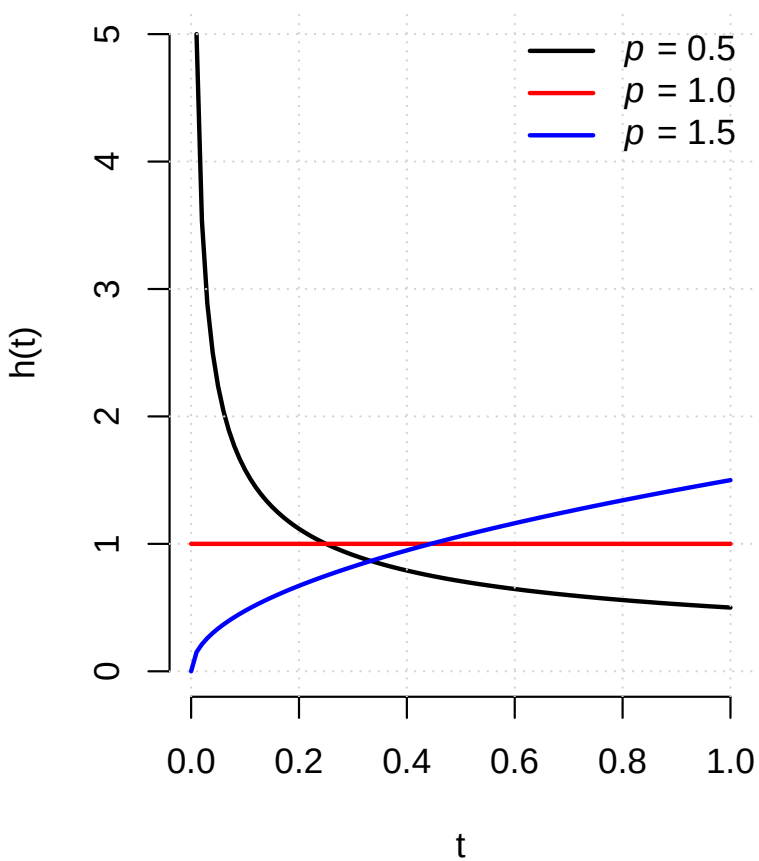
- Lifetime distribution: $f(t) = \lambda p t^{p-1} \exp(-\lambda t^p)$, $\lambda > 0$, $p > 0$
- Hazard function: $h(t) = \lambda p t^{p-1}$
- Survival function: $S(t) = \exp(-\lambda t^p)$
- Cumulative hazard: $H(t) = \lambda t^p$

Key features

- Parameterised by rate, λ, and shape, p
- Flexible: $p = 1 \rightarrow$ exponential; $p < 1 \rightarrow$ monotonically decreasing hazard with time; $p > 1 \rightarrow$ monotonically increasing hazard with time

There are many other distributions, e.g. log-logistic, Gompertz, etc., but exponential and Weibull common choices as they are sufficiently flexible for many applications





Which one to use?

Simple test

1. Plot $\log\left[-\log\left(\hat{S}(t)\right)\right]$ against $\log(t)$
2. If a straight-line, then test slope: If $= 1 \Rightarrow$ exponential; otherwise $\Rightarrow$ Weibull
3. If not straight-line, need a different model

How to estimate parameters

To estimate the model parameters θ, e.g. $\theta = (p, \lambda)$ for Weibull model:

- For non-censored subjects, the contribution to the likelihood function is $f(t_i \mid \theta)$
- For right-censored subjects, the contribution to the likelihood function is $S(t_i \mid \theta)$
- Assuming censoring independent of t_i, the likelihood function is:

$$\left(\prod_{\substack{i:\ \text{failure time} \\ \text{observed}}} f(t_i \mid \theta) \right) \times \left(\prod_{\substack{i:\ \text{failure time} \\ \text{right-censored}}} S(t_i \mid \theta) \right)$$

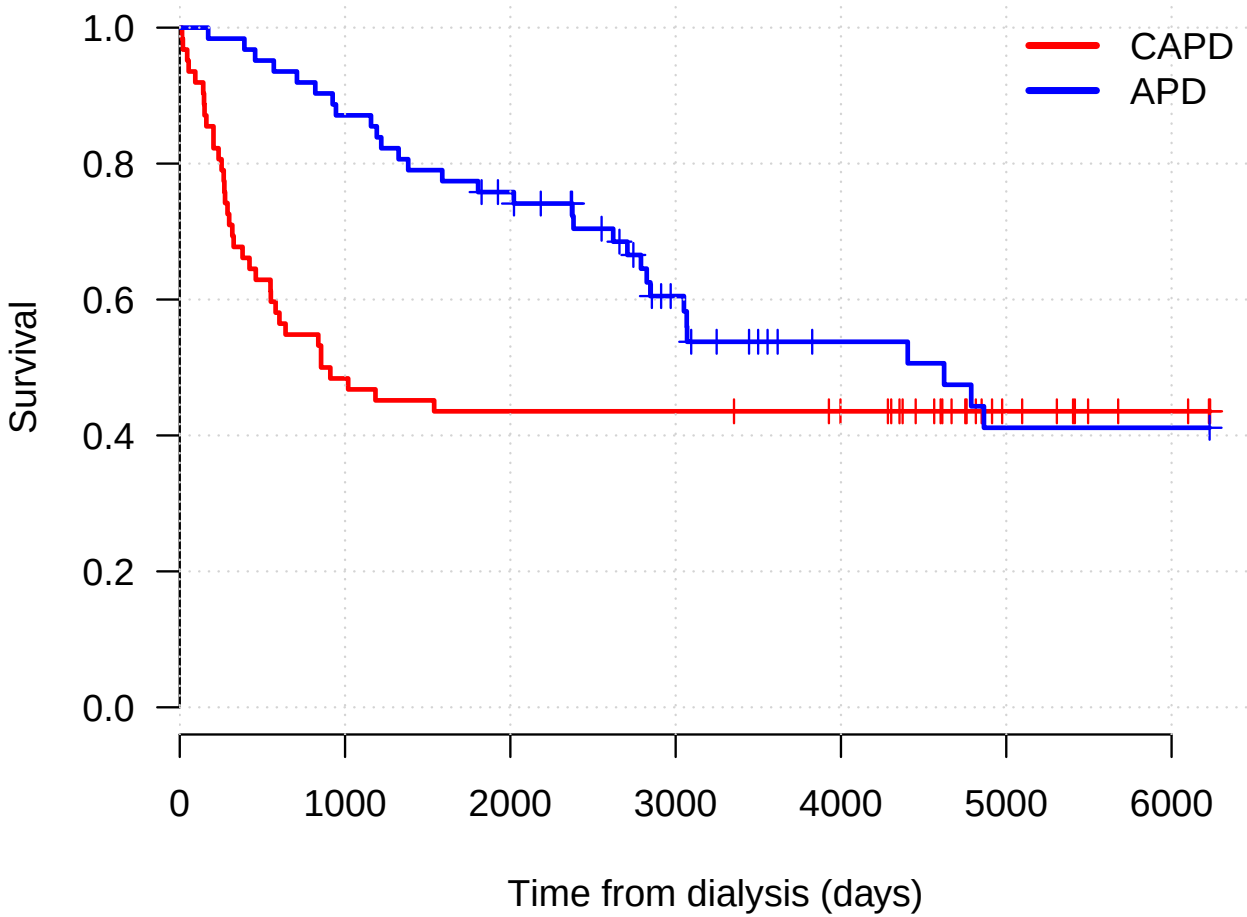
How can we compare survival between groups?

Examples

- Do patients survive longer after treatment that those without?
- Do males take longer to quit smoking compared to females?

If no censoring present, then can use Mann-Whitney U-test (or Kruskal-Wallis one-way ANOVA test)

Recall the peritoneal dialysis data had two treatment arms: APD ($n = 62$) and CAPD ($n = 62$) — do they have different survival distributions?



The log-rank test

Null hypothesis

Survival times are from the same distribution, i.e. there is no difference in group survivorship

Test

1. At each failure time t_i, there are $n_i = n_{i1} + n_{i2}$ subjects at risk just before, and there are $d_i = d_{i1} + d_{i2}$ failures
2. The probability of any subject experiencing the event under the null hypothesis is $p_i = d_i/n_i$
3. The expected number of failures in group 1 and 2 at t_i is $n_{i1}p_i$ and $n_{i2}p_i$ respectively
4. Repeat for every failure time and aggregate to calculate a Cochran-Mantel-Haenszel statistic, which has an (approximate) χ^2 distribution on 1 *df*

The log-rank test

Null hypothesis

Survival times are from the same distribution, i.e. there is no difference in group survivorship

Test

1. At each failure time t_i, there are $n_i = n_{i1} + n_{i2}$ subjects at risk just before, and there are $d_i = d_{i1} + d_{i2}$ failures
2. The probability of any subject experiencing the event under the null hypothesis is $p_i = d_i/n_i$
3. The expected number of failures in group 1 and 2 at t_i is $n_{i1}p_i$ and $n_{i2}p_i$ respectively
4. Repeat for every failure time and aggregate to calculate a Cochran-Mantel-Haenszel statistic, which has an (approximate) χ^2 distribution on 1 *df*

The log-rank test

1. Consider the 10 dialysis patient (3 CAPD and 7 APD) failure times[3] shown earlier, the first failure time was 147 days
2. The probability of a death is $\frac{1}{10}$ just before this time, so under the null hypothesis we would have expected $3 \times \frac{1}{10} = 0.3$ deaths in the CAPD group and $7 \times 110 = 0.7$ deaths in the APD group
3. We observed 1 failure in the CAPD group, and 0 in the APD group
4. This gives us the data for the first contribution to the χ^2 test statistic
5. And so on...

[3] In the interests of brevity, we pretend there are only 10 patients, although the actual dataset has 124 patients

Situations that require more complex methods

- **Interval censoring**
  Example: survival times subject to gross round-off error
- **Competing risks**
  Example: multiple causes of death
- **Informative censoring**
  Example: subjects censored because their condition is deteriorating

Suggested reading

- Diggle PJ, Chetwynd AG (2011). Statistics and Scientific Method: An Introduction for Students and Researchers. Oxford: *Oxford University Press*.
  - ✎ Chapter 8 covers much of the material presented in this course
- Collett D. Modelling Survival Data in Medical Research (1994). Boca Raton: *Chapman & Hall/CRC*.
  - ✎ Comprehensive text on survival data
- Bland JM, Altman DG (2004). The logrank test. *BMJ*, 328:1073.
  - ✎ 1-page round-up of the log-rank test
- Guo Z, et al. (2009). Modeling repeated time-to-event health conditions with discontinuous risk intervals: an example of a longitudinal study of functional disability among older persons. *Methods of Information in Medicine*, 47(2), 107-116.
  - ✎ Extensions to repeated/multiple events
- Putter H, et al. (2007). Tutorial in biostatistics: Competing risks and multi-state models. *Statistics in Medicine*, 26, 2389-2430.
  - ✎ Introduction to competing risk models