Biostatistics III:
Survival analysis for epidemiologists in R

Alex Ploner
Department of Medical Epidemiology and Biostatistics
Karolinska Institutet
Stockholm, Sweden
http://www.biostat3.net/

6–15 November, 2023

http://kiwas.ki.se/katalog/katalog/kurs/9735, course 2992

Friendly faces

- Course director: Mark Clements (mark.clements@ki.se)
- Course administrator: Gunilla Nilsson Roos (gunilla.nilsson.roos@ki.se)
- Teachers: Alex Ploner, Anna Plym, Birzhan Akynkozhayev, Balram Rai, Nurgul Batyrbekova.

Overview of the course I

- Central concepts in survival analysis: censoring, truncation, survival function, hazard function.
- Estimating survival using the Kaplan-Meier method.
- Estimating rates and modelling them using Poisson regression.
- Cox proportional hazards model.
- The proportional hazards assumption.
- Modelling non-proportional hazards.
- Comparison of the Cox and Poisson regression models.
- Parametric survival models.
- Risk set sampling (e.g. nested case-control studies)
- Non-collapsibility of the hazard ratio

Teaching format

- Generally lecture Q&A and review in the morning followed by computing labs in the afternoon.
- We have constructed exercises and provided solutions to most exercises. We will suggest appropriate exercises for each afternoon, but you are welcome to diverge from those suggestions.
- Course participants have a wide range of backgrounds and diverse interests. It is hoped that the lab sessions will provide time for you to study or ask questions about topics of special interest.
- Zoom fatigue is a potential issue – we have tried to keep the late afternoons free. Moreover, most of the lectures have been recorded, so you can watch them at other times.

Textbook I

- We have not assigned any compulsory texts since experience has shown that course participants have varying preferences.
- We will provide extensive course notes. We suggest students interested in additional reading identify a textbook at a technical level suitable for them. Many books on medical statistics contain a chapter on survival analysis.
- Very few books are targeted at epidemiologists.
- The definitive text for epidemiologists is 'Statistical Methods in Cancer Research: Volume II - The Design and Analysis of Cohort Studies' by Breslow and Day [3] although it is rather advanced. [http://publications.iarc.fr/_publications/media/download/3494/fb469ed43c52f0c738915cca6a0f31544b9ed7b6.pdf]
- For R, KI has access to a chapter on survival from "The R Book" http://onlinelibrary.wiley.com.proxy.kib.ki.se/doi/10.1002/9781118448908.ch27/pdf.

Computing

- We will be using R in the lecture material and the labs (or see the Exercises link on the home page)
  `http://www.biostat3.net/download/index.php?dir=R/`.
- The biostat3 package, including the labs, is available from CRAN. The package can be installed by

  ```
  > install.packages("biostat3")
  ```

  This should only be needed once per course.

Key concepts for the course

- Special methods (i.e. survival analysis) are required when the outcome of interest has a time dimension.
- The outcome can be presented as a survival proportion or an event rate. The two measures are mathematically related.
- Epidemiological cohort studies can (and should) be analysed in the framework of survival analysis. 'Time' may be a confounder or an effect modifier.
- Cox regression and Poisson regression are very similar.
- Reinforce key concepts in statistical modelling of epidemiological data
  - Studying confounding and effect modification in a modelling framework
  - Reparameterising a statistical model to estimate interaction effects[1]

[1] In this course, we tend to use "effect modification" and "interaction" synonymously, although some authors defined them distinctly [7].

Learning outcomes I

- For the course plan, see `http://kiwas.ki.se/katalog/katalog/kurs/5412` for course 2212. The learning outcomes are listed in the course plan and reproduced below.

After successfully completing this course you should be able to:

1. Propose a suitable statistical model for assessing a specific research hypothesis using data from a cohort study, fit the model using standard statistical software, evaluate the fit of the model and interpret the results.
2. Explain the similarities and differences between Cox regression and Poisson regression.
3. Understand the concept of timescales in statistical models for time-to-event data, be able to control for different timescales using standard statistical software and argue for an appropriate timescale for a given research hypothesis.
4. Understand the concept of confounding in epidemiological studies and be able to control/adjust for it using statistical models.

Learning outcomes II

5. Apply and interpret appropriate statistical models for studying effect modification and be able to reparameterise a statistical model to estimate appropriate contrasts.
6. Critically evaluate the methodological aspects (design and analysis) of a scientific article reporting a cohort study.

Take-home examination I

The course grade is based solely on a take-home written examination. The content of the exam will be similar to the previous take-home exams (although you will *not* need to write code for this exam). The exam requires you to understand the concepts of survival analysis and interpret output from standard statistical software. Instructions:

- The examination is individual-based: **you are not allowed to cooperate with anyone**, although you are encouraged to consult the available literature. The teachers will use Urkund to check for plagiarism (`https://staff.ki.se/plagiarism-checks-in-doctoral-education`)
- The examination will be made available at 12:00 on Wednesday 15 November 2023 and the examination is due by 17:00 on Wednesday 22 November 2023.
- The examination will be graded and results will be returned to you by 1 December 2023.
- Students who do not obtain a passing grade in the first examination will be offered a second examination within 2 months of the final day of the course.
- Do not write answers by hand: please use Word, LaTeX or a similar format for your examination report.

Take-home examination II

- Motivate all answers and show all calculations in your examination report, but write as brief an answer as possible without loss of clarity. Define any notation that you use for equations. The examination report should be written in English.
- You are expected to interpret R computer code and output.
- Email the examination report containing the answers **as a pdf file** to `gunilla.nilsson.roos@ki.se`. **Write your name in the email, but do not write your name in the document containing the answers.**

Notation

Expression	Interpretation
$\exp(x)$	Natural exponential of x
$\log(x)$	Natural logarithm of x
$\sum_{i=1}^{n} \exp(\beta x_i)$	Sum of $\exp(\beta x_i)$ for i varying from 1 to n; $\exp(\beta x_1) + \exp(\beta x_2) + \ldots + \exp(\beta x_n)$
$\prod_{i=1}^{n}(1 - p_i)$	Product of $(1 - p_i)$ for i varying from 1 to n; $(1 - p_1) \times (1 - p_2) \times \ldots \times (1 - p_n)$
$\int_{t_0}^{t} f(u)du$	Area under the function f where the value u varies between t_0 and t
$\lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t)}{\Delta t}$	As Δt gets closer to 0, the probability of T being between t and $t + \Delta t$, divided by Δt

Properties of logs and exponentials

Expression	Interpretation
$\exp(a+b) = \exp(a)\exp(b)$	exponential of a sum equals the product of the exponentiated values
$\log(ab) = \log(a) + \log(b)$	log of a product equals the sum of the log values
$\log(a^b) = b\log(a)$	log of a^b equals b times log of a
$\log(\exp(x)) = x$	log is the inverse of exp; and
$\exp(\log(x)) = x$	exp is the inverse of log
$1 - x \approx \exp(-x)$	this approximation holds when x is close to zero

Definitions

Let T be a continuous random variable for the time to an event with time origin t_0 (e.g. $t_0 = 0$ if study entry is from time 0). Then:

Name	Symbol	Definition
Probability density function	$f(t)$	$\lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t)}{\Delta t}$
Cumulative distribution function	$F(t)$	$Pr(T \leq t)$
Survival function	$S(t)$	$Pr(T > t)$

Properties for $f(t)$:

- Non-negative: $0 \leq f(t) < \infty$ (where ∞ is the symbol for infinity)
- Area under f is one:[2] $\int_{t_0}^{\infty} f(t)dt = 1$
- Probability of an event between t and $t + \delta$ is approximately $f(t)\delta$.

[2] For a proper distribution with no cure.

Definitions

Properties for $F(t)$:

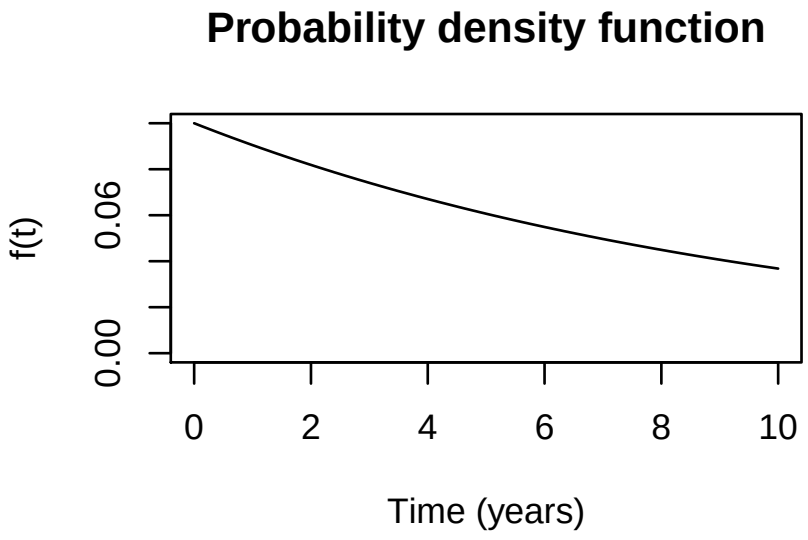
- Interpreted as the probability of having an event, or failure, by time t
- Bounded between zero and one: $0 \leq F(t) \leq 1$
- Zero at $t = t_0$: $F(t_0) = 0$
- One at $t = \infty$:[3] $F(\infty) = 1$
- Area under f between t_0 and t: $F(t) = \int_{t_0}^{t} f(u)du$

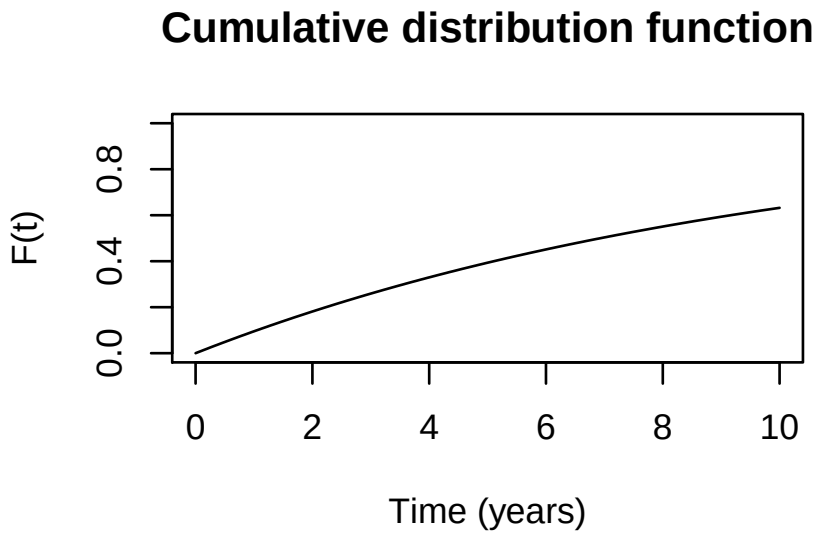
Properties for $S(t)$:

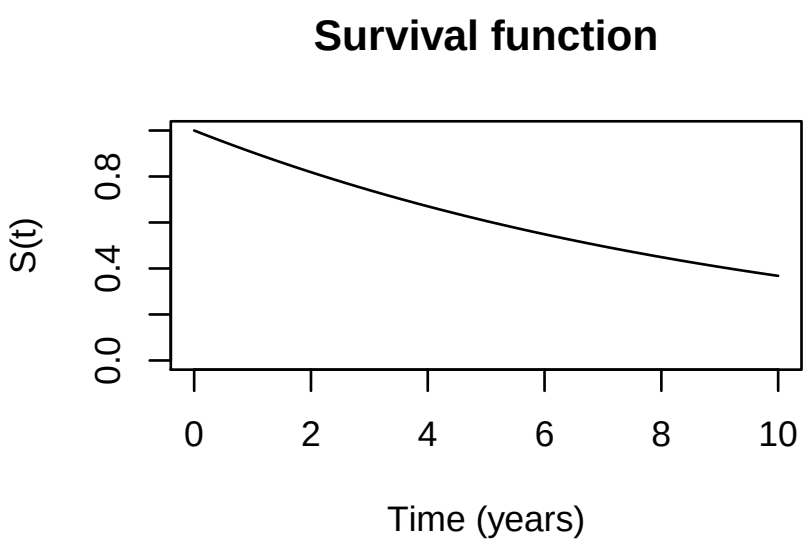
- $S(t) = 1 - F(t)$
- Interpreted as the probability of not having an event by time t
- Bounded between zero and one: $0 \leq S(t) \leq 1$
- One at $t = t_0$: $S(t_0) = 1$
- Zero at $t = \infty$:[3] $S(\infty) = 0$
- Area under $f(t)$ between t and ∞: $S(t) = \int_{t}^{\infty} f(u)du$

[3] For a proper distribution, otherwise some individuals may never have the event and $F(t) < 1$ and $S(t) > 0$.

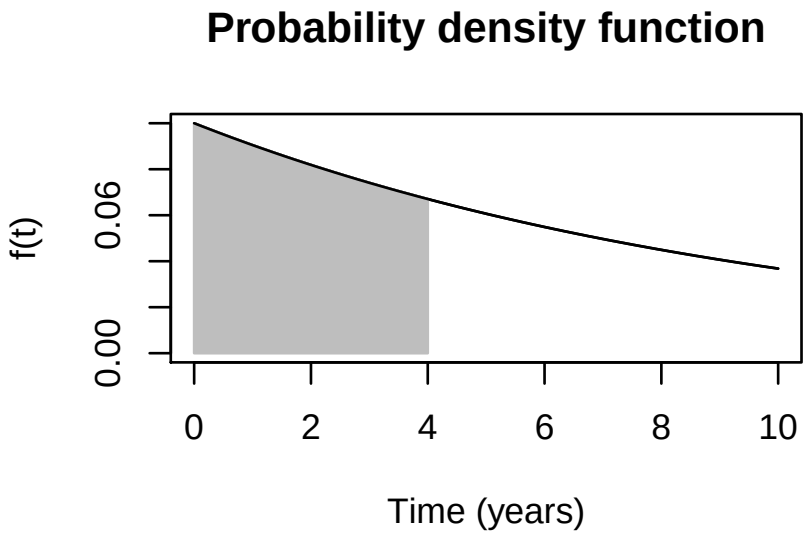
Three functions to represent an exponential distribution

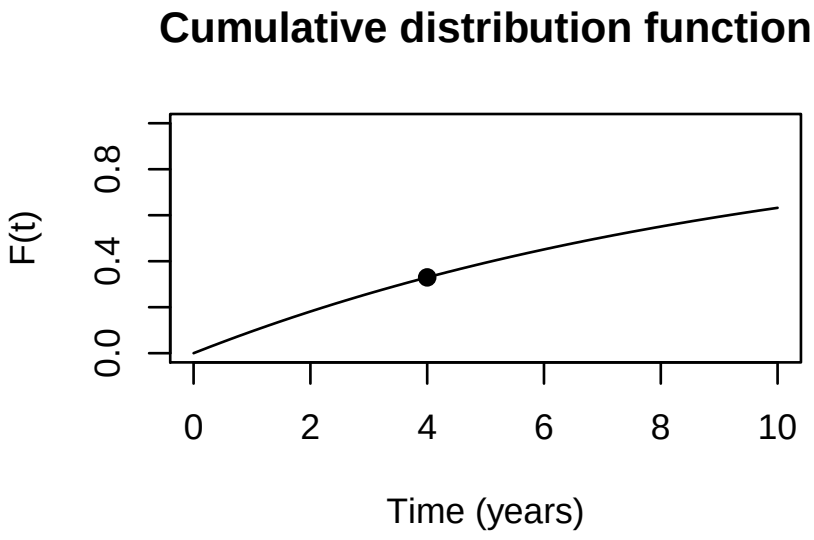


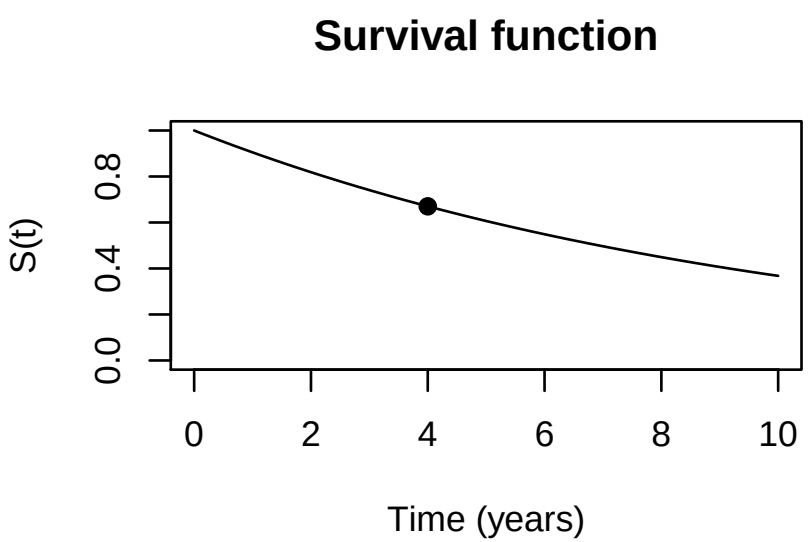




Three functions to represent an exponential distribution







Analysis of Time-to-Event Data (survival analysis)

- Survival analysis concerns analysing the time to the occurence of an event, e.g. time until a cancer patient dies.
- Time-to-event analysis is also known as failure time analysis, lifetime data analysis, event history analysis and survival analysis.
- Time-to-event analysis is used for cohort studies and randomised controlled trials (RCTs) for outcomes where study participants are followed from a well-defined entry time to an endpoint.

Formal requirements of time-to-event data

- Three basic requirements define time-to-event measurements
  1. precise definition of the start and end of follow-up time
  2. unambiguous origin for the measurement of 'time'; scale of time (e.g. time since diagnosis, attained age)
  3. precise definition of 'response,' or occurrence of the event of interest
- We will discuss the concept of timescales and how to choose an appropriate timescale later in the course.

What can we estimate from time-to-event data? I

- Survival probability, i.e. the proportion who have not experienced the event at a given time point during follow-up
- Median survival time, that is, the time when half of the individuals would have had the event
- Event rates (hazard rates, incidence rates), i.e. instantaneous risk that the event will occur at a given time point
- Hazard ratios, i.e. ratios of event rates between different groups (e.g. exposed vs. unexposed) while adjusting for confounders
- The focus of today's lecture is on how to estimate the survival probability. Later lectures will cover the other measures.

What can we estimate from time-to-event data? II

- In some studies, the event of interest (e.g. death) is bound to occur if we are able to follow-up each individual for a sufficient length of time.
- However, whether or not the event of interest is inevitable generally has no consequence for the design, analysis, or interpretation of the study[4].
- In some studies, the time-to-event (or survival probability) is of primary interest, whereas in many epidemiological cohort studies we are primarily interested in comparing the event rates between the exposed and unexposed.
- The basic statistical methodology is similar for randomised and observational studies, although the observational studies may have a stronger need to control for potential confounding.
- The characteristic that complicates the use of standard statistical methods is censoring — unobserved or interval values of the response measurement of interest. Censoring leads to differences in follow-up time between individuals.

[4] For completeness, an exception is when we are interesting in estimating the proportion "cured".

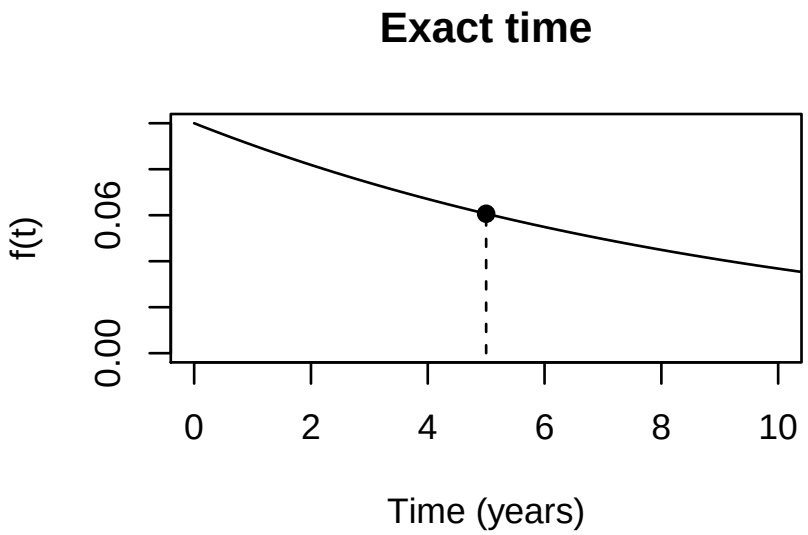
Why do we need survival analysis?

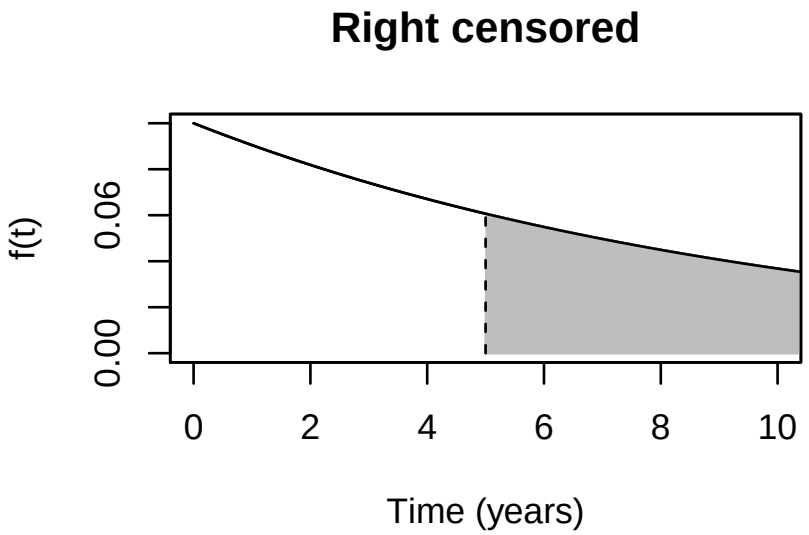
- A 2-by-2 table ignores time (exposure vs. outcome).
- If everyone has complete follow-up, then it is OK to ignore time.
- For example, if we follow 35 persons for 5 years and count how many who die during 5 years, then we can correctly estimate the risk of dying within 5 years.
- However, if not all persons will be followed for 5 full years, due to migration, death or other reasons, they are lost to follow-up. This means that we have incomplete follow-up for some persons.
- The deaths that we count during those 5 years will only be among those who are still being followed. Out of the 35 we started with, we will miss some deaths since they were unobserved (happened after we lost them from follow-up).
- In a situation with incomplete follow-up we must take time-at-risk (follow-up time) into account in the analysis, hence we use survival analysis.
- Note: In a situation with complete follow-up we can choose to either ignore time (by using a logistic regression type analysis) or by including it (by using a survival analysis).

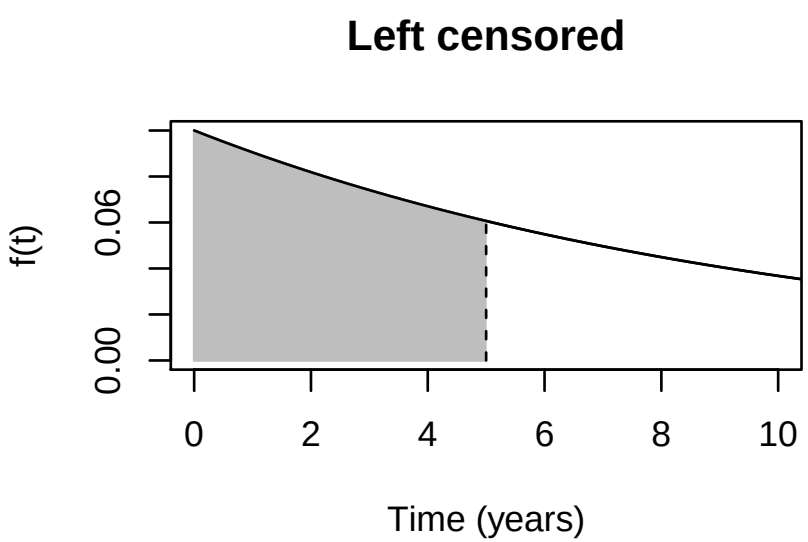
Right censoring and follow-up I

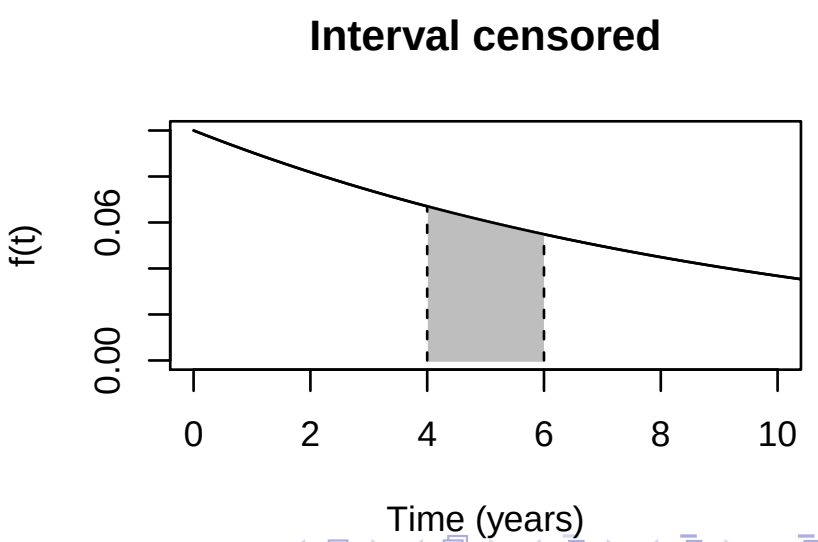
- Right censoring refers to the situation where the individual can no longer be followed up and the event of interest has not occurred.
- In studying the survival of cancer patients, for example, patients enter the study at the time of diagnosis (or the time of treatment in randomised trials) and are followed up until the event of interest is observed. Right censoring may occur in one of the following forms:
  - Termination of the study before the event occurs (administrative censoring);
  - Death due to a cause not considered to be the event of interest (in cause-specific survival analyses); and
  - Loss to follow-up, for example, if the patient emigrates.
- We say that the survival time is 'right censored'.
- With right censoring, we know that the event has not occurred during follow-up, but we are unable to follow-up the patient further. We know only that the true survival time of the patient is greater than a given value.
- If we do not account for these differences (by using survival analysis) then results may be biased.

Graphical representation of censoring









	sex	age	stage	mmdx	yydx	surv_mm	surv_yy	status	subsite
1	Male	72	Localised	2	1989	2	0.01	Dead: other	Descending and sigmoid
2	Female	82	Distant	12	1991	2	0.01	Dead: cancer	Descending and sigmoid
3	Male	73	Distant	11	1993	3	0.01	Dead: cancer	Descending and sigmoid
4	Male	63	Distant	6	1988	5	0.01	Dead: cancer	Transverse
5	Male	67	Localised	5	1989	7	0.01	Dead: cancer	Transverse
6	Male	74	Regional	7	1992	8	0.01	Dead: cancer	Coecum and ascending
7	Female	56	Distant	1	1986	9	0.01	Dead: cancer	Transverse
8	Female	52	Distant	5	1986	11	0.01	Dead: cancer	Coecum and ascending
9	Male	64	Localised	11	1994	13	1.00	Alive	Descending and sigmoid
10	Female	70	Localised	10	1994	14	1.00	Alive	Descending and sigmoid
11	Female	83	Localised	7	1990	19	1.00	Dead: other	Descending and sigmoid
12	Male	64	Distant	8	1989	22	1.00	Dead: cancer	Descending and sigmoid
13	Female	79	Localised	11	1993	25	2.00	Alive	Descending and sigmoid
14	Female	70	Distant	6	1988	27	2.00	Dead: cancer	Coecum and ascending
15	Male	70	Regional	9	1993	27	2.00	Alive	Coecum and ascending
16	Female	68	Distant	9	1991	28	2.00	Dead: cancer	Descending and sigmoid
17	Male	58	Localised	11	1990	32	2.00	Dead: cancer	Descending and sigmoid
18	Male	54	Distant	4	1990	32	2.00	Dead: cancer	Coecum and ascending
19	Female	86	Localised	4	1993	32	2.00	Alive	Descending and sigmoid
20	Male	31	Localised	1	1990	33	2.00	Dead: cancer	Coecum and ascending
21	Female	75	Localised	1	1993	35	2.00	Alive	Descending and sigmoid
22	Female	85	Localised	11	1992	37	3.00	Alive	Coecum and ascending
23	Female	68	Distant	7	1986	43	3.00	Dead: cancer	Descending and sigmoid
24	Male	54	Regional	6	1985	46	3.00	Dead: cancer	Transverse
25	Male	80	Localised	6	1991	54	4.00	Alive	Coecum and ascending
26	Female	52	Localised	7	1989	77	6.00	Alive	Transverse
27	Male	52	Localised	6	1989	78	6.00	Alive	Descending and sigmoid
28	Male	65	Localised	1	1989	83	6.00	Alive	Descending and sigmoid
29	Male	60	Localised	11	1988	85	7.00	Alive	Transverse
30	Female	71	Localised	11	1987	97	8.00	Alive	Descending and sigmoid
31	Male	58	Localised	8	1987	100	8.00	Alive	Descending and sigmoid

	yydx	status
1	1989	Dead: other
2	1991	Dead: cancer
3	1993	Dead: cancer
4	1988	Dead: cancer
5	1989	Dead: cancer
6	1992	Dead: cancer
7	1986	Dead: cancer
8	1986	Dead: cancer
9	1994	Alive
10	1994	Alive
11	1990	Dead: other
12	1989	Dead: cancer
13	1993	Alive
14	1988	Dead: cancer
15	1993	Alive
16	1991	Dead: cancer
17	1990	Dead: cancer
18	1990	Dead: cancer
19	1993	Alive
20	1990	Dead: cancer
21	1993	Alive
22	1992	Alive
23	1986	Dead: cancer
24	1985	Dead: cancer
25	1991	Alive
26	1989	Alive
27	1989	Alive
28	1989	Alive
29	1988	Alive
30	1987	Alive
31	1987	Alive

Choice of outcome measure I

- In Biostatistics I and II we covered statistical methods for comparing means and proportions (e.g. logistic regression). What happens if we apply these methods now?
- Let's assume a new treatment was introduced in late 1992 and we are interested in studying whether patient survival has improved for patients diagnosed 1993–94 compared to those diagnosed earlier.
- Let's compare the proportion of patients who die between two diagnosis periods.
- The patients were followed until end of 1995.
- This means that patients who were diagnosed 1993-1994 only had follow-up for at most 36 months (3 years) due to 'administrative' right censoring.
- Whereas, patients diagnosed 1985-1992 had follow-up for at most 11 years.
- Note that most patients do not have complete follow-up through to 11 years.

Choice of outcome measure II

R code and output

```
## Reminder: anything after a # is a comment
## In the following, code starts with ">"
> library(biostat3) # load biostat3 library to use colon_sample data-frame
> library(dplyr)    # load dplyr library to use the mutate function
> colon2 = mutate(colon_sample,
                  dx93=(yydx>=1993),        # binary variable
                  dead=(status != "Alive")) # binary variable
> m = stats::xtabs(~dx93+dead, data=colon2) # cross-tabulation
> m
> stats::chisq.test(m)

     dead
dx93    FALSE TRUE
  FALSE    10   18
  TRUE      6    1

Pearson's Chi-squared test with Yates' continuity correction

data:  m
X-squared = 3.8065, df = 1, p-value = 0.05105

Warning message:
In chisq.test(m) : Chi-squared approximation may be incorrect
```

Choice of outcome measure III

As the expected cell counts are less than five in the bottom row, we should use Fisher's exact test.

R code and output

```
> stats::fisher.test(m)

Fisher's Exact Test for Count Data

data:  m
p-value = 0.03182
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.001915923 0.990524462
sample estimates:
odds ratio
 0.0990587
```

Choice of outcome measure IV

- We see that only 1 of the 7 (14%) patients diagnosed in the recent period died compared to 18 of 28 (64%) in the early period and this difference is statistically significant.
- It is not surprising that the proportion of deaths was lower among patients diagnosed more recently since these patients had a shorter follow-up time: they did not have the same opportunity to die.
- Let's instead compare the average 'survival time' (the lengths of the lines) between the two groups while ignoring whether or not the patient died.

R code and output

```
> stats::t.test(surv_mm ~ dx93, data=colon2)

data:  surv_mm by dx93
t = 3.2604, df = 31.089, p-value = 0.002698
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 10.15265 44.06164
sample estimates:
mean in group FALSE  mean in group TRUE
           48.39286            21.28571
```

Choice of outcome measure V

- Patients diagnosed in 1985-92 'survived' on average for 48 months compared to 21 months for patients diagnosed 1993-94.
- Restricting this analysis to patients who died (i.e. mean survival time among those who died) is not appropriate either. By definition, the maximum survival time for patients diagnosed 1993-1994 is 36 months.

R code and output

```
> t.test(surv_mm ~ dx93, data=colon2, subset=dead)
Error in t.test.default(x = c(2L, 2L, 5L, 7L, 8L, 9L, 11L, 19L, 22L, 27L,  :
  not enough 'y' observations
> xtabs(~dx93, data=colon2, subset=dead)
dx93
FALSE  TRUE
   18     1
```

Choice of outcome measure VI

- What we would like is some measure of the risk of death adjusted for the fact that individuals were at risk for different lengths of time.
- Methods used for making inference about proportions (e.g. logistic regression) are only appropriate when all individuals have the same time at risk. This is typically not the case when we have survival data.
- There may, however, be situations where everyone has the same potential follow-up.
- That is, when we have a binary outcome and all individuals are at risk for the same length of time the proportion is an appropriate outcome measure.

$$\text{proportion who experience the event} = \frac{\text{number of events}}{\text{number of individuals}}$$

- Every individual contributes the same amount to the denominator.

Choice of outcome measure VII

- If, however, individuals are at risk for differing lengths of time we use 'person-time' as the denominator and estimate the event rate (a mortality rate in this example).

$$\text{event rate} = \frac{\text{number of events}}{\text{person-time at risk}}$$

We can calculate the rate in R using dplyr:

R code and output

```
> library(dplyr)
> group_by(colon2,dx93) %>%
    summarise(events=sum(dead), py=sum(surv_mm/12)) %>%
    mutate(rate = events/py)
# A tibble: 2 x 4
  dx93  events    py   rate
  <lgl>  <int> <dbl>  <dbl>
1 FALSE     18 113.  0.159
2 TRUE       1  12.4 0.0805
```

Choice of outcome measure VIII

- The main message is that, in survival analysis, the outcome has at least two dimensions – the event indicator and the time at risk[5].
- The event rate is not the only appropriate outcome measure; it is also possible to estimate the proportion surviving (or proportion dying) while controlling for the fact that individuals are at risk for different lengths of time. This, in fact, will be the focus for today's lectures.

[5] The time origin is a third dimension – which is particularly important if there is confounding by time.

Sample data sets

- The following data sets will be used during the course:

colon	Colon carcinoma diagnosed during 1975–1994 with follow-up to 31 December 1995.
melanoma	Skin melanoma diagnosed during 1975–1994 with follow-up to 31 December 1995.
colon_sample	A random sample of 35 patients from the colon data.
diet	Data from a pilot study evaluating the use of a weighed diet over 7 days in epidemiological studies. The primary hypothesis is the relationship between dietary energy intake and incidence of coronary heart disease (CHD).

- The diet data are analysed extensively by Clayton and Hills [4].

Brief review of R data types

Type	Name	Values	Comment
`num`	numeric	1.0, -2.0e-5	Double precision. Assumes 1 is numeric
`int`	integer	1L, 2L	
`char`	character	"A", "1234"	
`Factor`	factor	"A", "B"	Categories with ordered levels
`Date`	Date	as.Date("1969-08-01")	Difference of dates is a `difftime` – convert using `as.numeric`.

Variables in the colon carcinoma data set

R code and output

```
> str(colon)

'data.frame': 15564 obs. of  16 variables:
 $ sex     : Factor w/ 2 levels "Male","Female": 2 2 1 1 1 2 2 2 1 1 ...
 $ age     : int  77 78 78 76 80 75 81 77 77 78 ...
 $ stage   : Factor w/ 4 levels "Unknown","Localised",..: 4 2 4 4 2 2 4 3 2 1 ...
 $ yydx    : int  1977 1978 1978 1976 1980 1975 1981 1977 1977 1978 ...
 $ surv_mm : num  16.5 82.5 1.5 1.5 8.5 23.5 2.5 9.5 85.5 0.5 ...
 $ surv_yy : num  1.5 6.5 0.5 0.5 0.5 1.5 0.5 0.5 7.5 0.5 ...
 $ status  : Factor w/ 4 levels "Alive","Dead: cancer",..: 2 3 2 2 2 2 2 3 3 2 ...
 $ subsite : Factor w/ 4 levels "Coecum and ascending",..: 2 1 3 3 3 1 3 2 1 2 ...
 $ year8594: Factor w/ 2 levels "Diagnosed 75-84",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ agegrp  : Factor w/ 4 levels "0-44","45-59",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ dx      : Date, format: "1977-03-04" "1978-07-26" "1978-10-10" ...
 $ exit    : Date, format: "1978-07-20" "1985-06-11" "1978-11-25" ...
 $ id      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ ydx     : num  1977 1979 1979 1977 1981 ...
 $ yexit   : num  1979 1985 1979 1977 1982 ...
```

Variables in the skin melanoma data set

R code and output

```
> str(melanoma)

'data.frame': 7775 obs. of  16 variables:
 $ sex     : Factor w/ 2 levels "Male","Female": 2 2 2 2 2 2 2 2 2 2 ...
 $ age     : int  81 75 78 75 81 75 75 80 76 79 ...
 $ stage   : Factor w/ 4 levels "Unknown","Localised",..: 2 2 2 1 1 2 2 2 1 4 ...
 $ yydx    : int  1981 1975 1978 1975 1981 1975 1975 1980 1977 1980 ...
 $ surv_mm : num  26.5 55.5 177.5 29.5 57.5 ...
 $ surv_yy : num  2.5 4.5 14.5 2.5 4.5 1.5 5.5 4.5 18.5 1.5 ...
 $ status  : Factor w/ 4 levels "Alive","Dead: cancer",..: 3 3 3 2 3 2 3 3 1 2 ...
 $ subsite : Factor w/ 4 levels "Head and Neck",..: 1 1 3 4 1 2 1 1 3 4 ...
 $ year8594: Factor w/ 2 levels "Diagnosed 75-84",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ dx      : Date, format: "1981-02-02" "1975-09-21" "1978-02-21" ...
 $ exit    : Date, format: "1983-04-20" "1980-05-07" "1992-12-07" ...
 $ agegrp  : Factor w/ 4 levels "0-44","45-59",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ id      : int  1 2 3 4 5 6 7 8 9 10 ...
 $ ydx     : num  1981 1976 1978 1976 1982 ...
 $ yexit   : num  1983 1980 1993 1978 1986 ...
```

Variables in the diet data set

R code and output

```
> str(diet)

'data.frame': 337 obs. of  15 variables:
 $ id     : int  127 200 198 222 305 173 120 206 81 333 ...
 $ chd    : int  0 0 0 0 1 0 0 0 0 0 ...
 $ y      : num  16.79 19.96 19.96 15.39 1.49 ...
 $ hieng  : Factor w/ 2 levels "low","high": 1 1 1 1 1 1 1 1 1 1 ...
 $ energy : num  2023 2449 2281 2468 2363 ...
 $ job    : Factor w/ 3 levels "driver","conductor",..: 2 3 3 3 3 2 2 3 1 3 ...
 $ month  : int  2 12 12 2 1 12 7 1 12 6 ...
 $ height : num  174 178 NA 159 NA ...
 $ weight : num  61.5 73.5 NA 58.2 NA ...
 $ doe    : Date, format: "1960-02-16" "1956-12-16" "1956-12-16" ...
 $ dox    : Date, format: "1976-12-01" "1976-12-01" "1976-12-01" ...
 $ dob    : Date, format: "1910-09-27" "1909-06-18" "1910-06-30" ...
 $ yoe    : num  1960 1957 1957 1957 1960 ...
 $ yox    : num  1977 1977 1977 1973 1962 ...
 $ yob    : num  1911 1909 1910 1903 1913 ...
```

Colon carcinoma 1985–94

Table: Codes for vital status with corresponding frequency counts 1985–94

Code and description	Male	Female
0 Alive	1476	2081
1 Dead: due to colon carcinoma	1806	2618
2 Dead: other cause of death	519	586
4 Lost to follow-up	1	0
Total	3802	5285

Note that the sample data sets also include patients diagnosed 1975–1984.

Skin melanoma 1985–94 I

Table: Codes for vital status with corresponding frequency counts 1985–94

Code and description	Male	Female
0 Alive	1554	1786
1 Dead: melanoma was the cause	543	376
2 Dead: other cause of death	238	247
4 Lost to follow-up	0	0
Total	2335	2409

Note that the sample data sets also include patients diagnosed 1975–1984.

Examples of events and right censoring

Events	Right censoring
Death	Emigration
	End-of-study (e.g. 2006-12-31)
Cancer death	Death due to other causes than cancer
	Emigration
	End-of-study (e.g. 2006-12-31)
Breast cancer incidence	Death
	Emigration
	End-of-study (e.g. 2006-12-31)
	Mastectomy

The survival function I

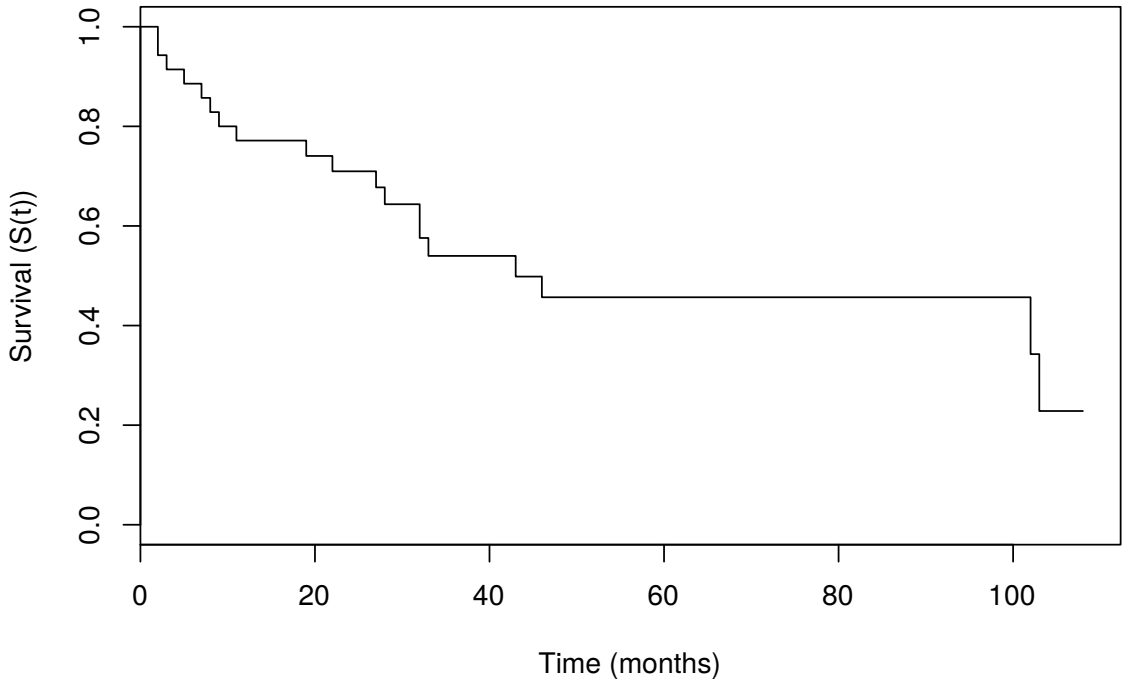


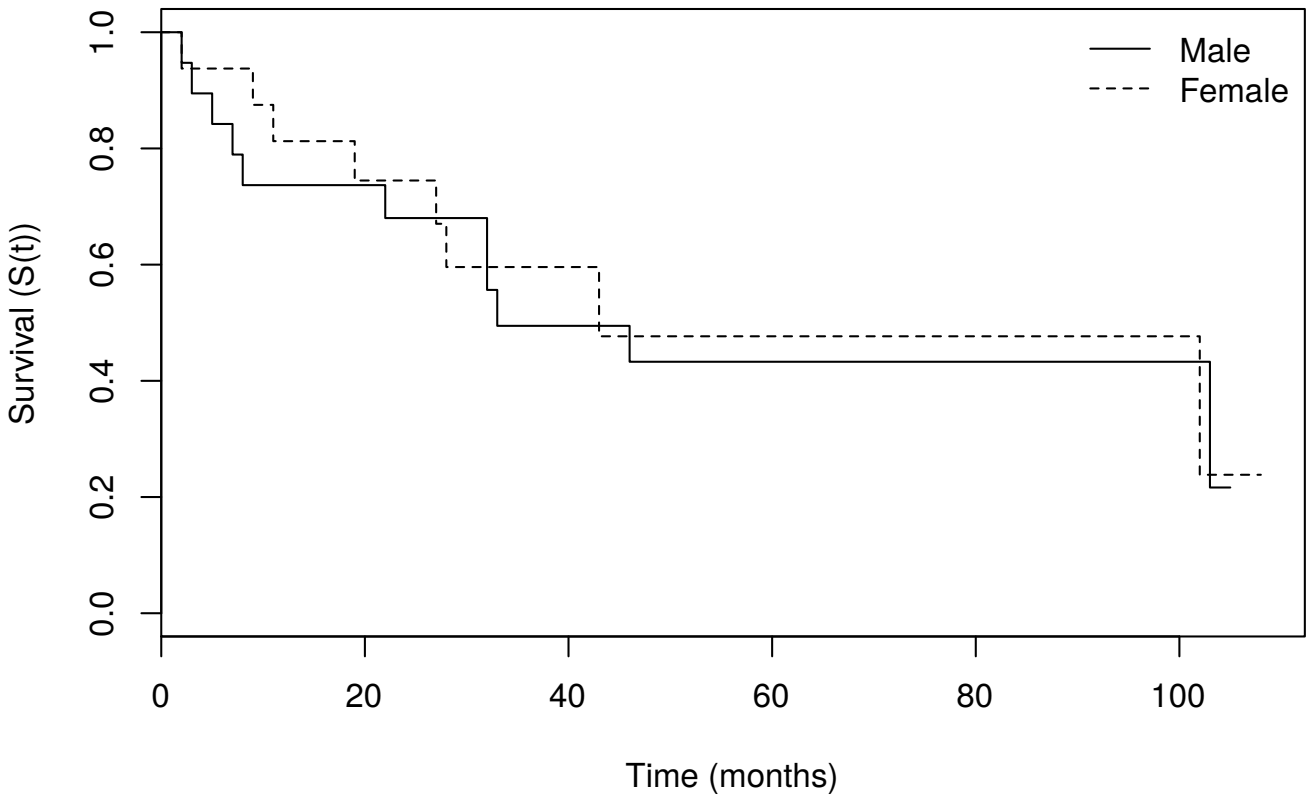
Figure: Estimates of $S(t)$ for 35 patients diagnosed with colon carcinoma. All deaths are considered events.

The survival function II

Review:

- $S(t)$ is a nonincreasing ('monotone decreasing') function with a value 1 at the time origin and a value 0 as t approaches infinity.
- Survival is a proportion.
- For example, the 5-year survival proportion for the data presented in Figure 1 is 45%.
- Nonparametric methods for estimating $S(t)$ (described later) generally involve estimating the survival proportion at discrete values of t and then interpolating these to obtain an estimate of $S(t)$.

Interpreting $S(t)$ and comparing estimates of $S(t)$ between groups I



Interpreting $S(t)$ and comparing estimates of $S(t)$ between groups II

- Individuals in group 1 experience slightly better survival compared to individuals in group 2.
- It is difficult to compare survival between groups simply by studying the plots.
- The rate of decline of the survival function is a measure of the risk of experiencing the event at time t (the instantaneous mortality rate at time t).
- In survival analysis, this is called the hazard function, $\lambda(t)$.
- Patients in group 1 have better survival for the interval up to 850 days following diagnosis but then have worse survival than group 2 after 850 days.
- This is an example of non-proportional hazards.
- The survival experience of a cohort can be expressed in terms of the survival proportion or the hazard rate.
- In epidemiological cohort studies where the incidence of a disease is the outcome (rather than death), we often present the failure proportion, given by $F(t) = 1 - S(t)$, rather than $S(t)$.

Interpreting $S(t)$ and comparing estimates of $S(t)$ between groups III

- We can model the hazard function (or incidence rate) and estimate the hazard ratio (or incidence rate ratio) for the exposed compared to the unexposed.
- The hazard ratio, rather than the survival function, may be of primary interest.

Definitions I

As defined earlier, we let T be a continuous random variable for the time to an event with time origin t_0 (e.g. $t_0 = 0$ if study entry is from time 0), with probability density function $f(t)$ and survival function $S(t)$. Then:

Name	Symbol	Definition
Hazard	$\lambda(t)$	$\lim_{\Delta t \to 0} \frac{Pr(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$
Cumulative hazard	$\Lambda(t)$	$\int_{t_0}^{t} \lambda(u) du$

Properties for $\lambda(t)$:

- Non-negative: $0 \leq \lambda(t) < \infty$
- Probability of an event between t and $t + \Delta t$ conditional on survival to time t is approximately $\lambda(t)\Delta t$.

Properties for $\Lambda(t)$:

- Non-negative: $0 \leq \Lambda(t) < \infty$
- $\Lambda(t_0) = 0$ and $\lim_{t \to \infty} \Lambda(t) = \infty$
- Interpreted as the area under λ between t_0 and t

Mathematical relationships between these functions

- Given the hazard or cumulative hazard, we can calculate survival:

$$\begin{aligned} S(t) &= \exp(-\Lambda(t)) \\ &= \exp\left(-\int_{t_0}^{t} \lambda(u)du\right) \end{aligned}$$

- Given survival or the cumulative hazard, we can calculate the hazard:

$$\begin{aligned} \lambda(t) &= -\frac{dS(t)}{dt}/S(t) && \text{(-slope/survival=rate of decline)} \\ \lambda(t) &= \frac{d}{dt}\Lambda(t) && \text{(derivative of cumulative hazard)} \\ &= -\frac{d}{dt}\log(S(t)) && \text{(derivative of cumulative hazard)} \\ \lambda(t) &= \frac{f(t)}{S(t)} && \text{(density/survival)} \end{aligned}$$

Estimating the survival function, $S(t)$ I

- Consider the 35 colon cancer patients introduced on slide 25.
- We may be interested in estimating $S(t)$ for death due to any cause.
- An estimate of $S(t)$ could be obtained by simply calculating the proportion of individuals still alive at selected values of t, such as completed years.
- We had 35 patients alive at start. Eight of the 35 patients died during the first year of follow-up so the estimate for $S(1)$ is $\hat{S}(1) = (35 - 8)/35 = 27/35 = 0.771$.
- We encounter problems when attempting to estimate $S(2)$. Ten patients died within two years of follow-up, but 2 patients (patients 9 and 10) could not be followed-up for a full 2 years.

Estimating the survival function, $S(t)$ II

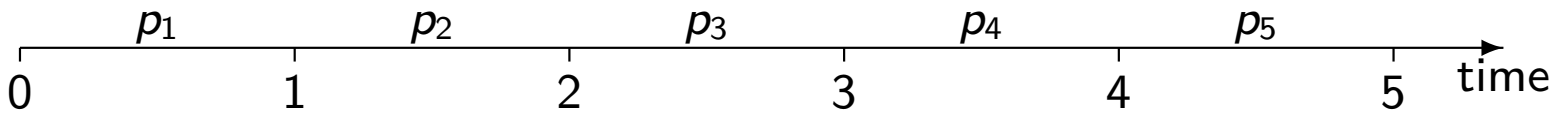
Consider the following table:

Interval	At risk (l)	Deaths (d)	Censored (w)	$S(t)$
$[0, 1)$	35	8	0	1-8/35
$[1, 2)$	27	2	2	?

- We could exclude these two patients from the analysis altogether and let $\hat{S}(2) = (33 - 10)/33$, but this will underestimate the true survival proportion since it ignores the fact that each of these two patients were at risk of death for between one and two years but did not die while under observation.
- If we instead use $\hat{S}(2) = (35 - 10)/35$ then we will overestimate the true survival proportion, since we are assuming that each of these two patients survived for a full two years.
- Two common (and similar) methods for estimating $S(t)$ in the presence of censoring are the the Kaplan-Meier (product-limit) method and the actuarial (life-table) method.

The general approach to nonparametric estimation of $S(t)$ I

- Assume we wish to estimate five year survival, $S(5)$.



- We start by estimating the following conditional survival probabilities:
  p_1, the probability of surviving at least 1 year from time 0
  p_2, the probability of surviving at least 2 years conditional on surviving 1 year
  p_3, the probability of surviving at least 3 years conditional on surviving 2 years
  p_4, the probability of surviving at least 4 years conditional on surviving 3 years
  p_5, the probability of surviving at least 5 years conditional on surviving 4 years
- The probability of surviving at least 5 years (from time zero) is then given by the product of these conditional survival probabilities.

$$S(5) = \prod_{i=1}^{5} p_i$$

The general approach to nonparametric estimation of $S(t)$ II

- That is, to survive five years, one must survive year 1, and year 2, . . . , and year 5.
- The advantage of this approach is that we can appropriately account for censoring when estimating the probability of surviving a small time interval (i.e. when estimating the conditional survival probabilities).
- This approach is employed by both the actuarial (life-table) method and the Kaplan-Meier (product-limit) method.
- We chose to estimate conditional probabilities for one year intervals (time-bands) but the intervals may be any width.
- The primary differences between the Kaplan-Meier and actuarial methods is the manner in which the default intervals are chosen (not really a difference in theory) and the method for dealing with ties.

Ties in survival data

- If two individuals have the same survival time (time to event or time to censoring), we say that the survival times are tied.
- Many of the standard methods for survival analysis, such as the Kaplan-Meier method and the Cox proportional hazards model, assume that survival time is measured on a continuous scale and that ties are therefore rare.
- In population-based survival analysis, however, ties are common.
- For example, among the 9087 people diagnosed with colon carcinoma during 1985–1994, 490 died during the first month of follow-up and 542 during the second month of follow-up (although there were no censorings during these months since every individual had a potential follow-up time of at least 12 months).

Summary: nonparametric estimation of $S(t)$ I

1. Split follow-up into intervals (timebands). If there are both deaths and censorings within an interval then (within the interval):

   K-M Assume the events precede those censored, that is, everyone is at risk when the events occur.

   Actuarial Assume half of the censored individuals are at risk when the events occur.

2. Estimate conditional probabilities of surviving each interval

$$p_i = 1 - d_i/n_i$$

   where d_i is the number of events and n_i number at risk for interval i.

3. $S(t)$ is the product of the conditional probabilities up to time t.

$$S(t_k) = \prod_{i=1}^{k} p_i$$

Summary: nonparametric estimation of $S(t)$ II

- The only difference between the Kaplan-Meier method and the actuarial method is the approach to dealing with ties (which affects the value of n_i in estimating the conditional probabilities).
- The Kaplan-Meier approach is slightly biased in the presence of ties so one should define time as accurately as possible (e.g. don't use time in months if you have time in days) to minimise the number of ties.
- If survival times are generated on a truly discrete scale (e.g. patients are contacted annually to ascertain vital status) and ties are common then the actuarial approach is preferable.
- The actuarial method can, however, also be used with many small intervals.

Continuing the example

Kaplan-Meier method

Interval	At risk (l)	Deaths (d)	Censored (w)	p	$S(t)$
$[0,1)$	35	8	0	1-8/35	1-8/35
$[1,2)$	27	2	2	1-2/27	(1-2/27)×(1-8/35)

Actuarial method

Interval	At risk (l)	Deaths (d)	Censored (w)	p	$S(t)$
$[0,1)$	35	8	0	1-8/35	1-8/35
$[1,2)$	27	2	2	1-2/(27-2/2)	(1-2/(27-2/2))×(1-8/35)

The Kaplan-Meier method for estimating $S(t)$ I

- Also known as the product-limit method but is more commonly known as the Kaplan-Meier method, after the two researchers who first published the method in English in 1958 [6].
- The method was published earlier (1912) in German [2].
- In essence, the Kaplan-Meier method is the life table method where the interval size is decreased towards zero so that the number of intervals tends to infinity. Each life table interval is of infinitesimal length, just enough for one event or time increment.
- In practice, survival time is measured on a discrete scale (e.g. minutes, hours, days, months, or years) so the interval length is limited by the accuracy by which survival time is measured.

The Kaplan-Meier method for estimating $S(t)$ II

- In practice, only those intervals containing an event contribute to the estimate, so we can ignore all other intervals.
- To estimate survival, the patient survival times are first ranked in increasing order.
- The times where events (deaths) occur are denoted by t_i, where $t_1 < t_2 < t_3 < \ldots$.
- The number of deaths occurring at t_i is denoted by d_i.
- If both censoring(s) and death(s) occur at the same time, then the censoring(s) are assumed to occur immediately after the death time.
- That is, individuals with survival times censored at t_i are assumed to be at risk at t_i.
- The Kaplan-Meier estimate of the cumulative survival function at time t is given by

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{t_i \leq t}(1 - \frac{d_i}{l_i}) & \text{if } t \geq t_1 \end{cases} \quad (1)$$

where l_i is the number of persons at risk.

The Kaplan-Meier method for estimating $S(t)$ III

- A plot of the Kaplan-Meier estimate of the survival function (slide 43) takes the form of a step function, in which the survival probabilities decrease at each death time and are constant between adjacent deaths times.
- Censorings contribute in Equation 1 by decreasing l_i at the next death time.
- If the largest observed survival time (which we will call t_{max}) is a censored survival time (or times), then $\hat{S}(t)$ is undefined for $t > t_{\text{max}}$, otherwise $\hat{S}(t) = 0$ for $t > t_{\text{max}}$.
- The standard error of the estimate can be obtained using Greenwood's method [5].

K-M estimates for the sample data (up to 19 months)

t	at risk	observed deaths	p_i	$S(t)$	SE
0	35	0	1.0000	1.0000	–
2	35	2	0.9429	0.9429	0.0392
3	33	1	0.9697	0.9143	0.0473
5	32	1	0.9688	0.8857	0.0538
7	31	1	0.9677	0.8571	0.0591
8	30	1	0.9667	0.8286	0.0637
9	29	1	0.9655	0.8000	0.0676
11	28	1	0.9643	0.7714	0.0710
13+	27	0			
14+	26	0			
19	25	1	0.9600	0.7406	0.0745

- At $t = 2$ months we observed 2 deaths among the 35 patients at risk, so $p_1 = 1 - 2/35 = 0.9428$.
- At $t = 3$ months we observed 1 death among the 33 patients at risk, so $p_2 = 1 - 1/33 = 0.9697$.
- Subsequently, $\hat{S}(t) = 0.9429 \times 0.9697 = 0.9143$ for $3 \leq t < 5$.

Survival analysis using R

- Some of the R survival analysis functions relevant to this course are given below. Further details can be found in the manuals or online help.

Function name	Description
`survival::survfit`	Kaplan-Meier survival curves
`survival::survSplit`	Split time-span records
`biostat3::survRate`	Tabulate failure rate
`biostat3::muhaz2`	Calculate smoothed hazards for single groups
`biostat3::lifetab2`	Calculate actuarial survival
`stats::glm`	Estimate Poisson regression
`survival::coxph`	Estimate Cox proportional hazards model
`survival::cox.zph`	Test of Cox proportional hazards assumption
`rstpm2::stpm2`	Estimate generalised survival models

Brief intro to the Surv function

- In the survival package in R (and many other R packages), survival is specified using the Surv() function. With two parameters, this represents right-censored data. For example:

```
> Surv(c(10,15), c(TRUE,FALSE))
[1] 10  15+
```

- This shows the survival time for two individuals, where the first individual had an event at 10 (assuming a time unit of years) and the second had no event after 15 years. The Surv function can also specify left- and interval-censored data and left-truncated data.

Example

- To calculate the Kaplan-Meier estimate of the cause-specific survival function by sex at 120 months and to fit a Cox proportional hazards model with sex and calendar period as covariates

R code

```
> fit = survfit(Surv(surv_mm,status=="Dead: cancer")~sex, data=colon)
> summary(fit, times=120)
> coxph(Surv(surv_mm,status=="Dead: cancer")~sex+year8594, data=colon)
```

Kaplan-Meier estimates in R

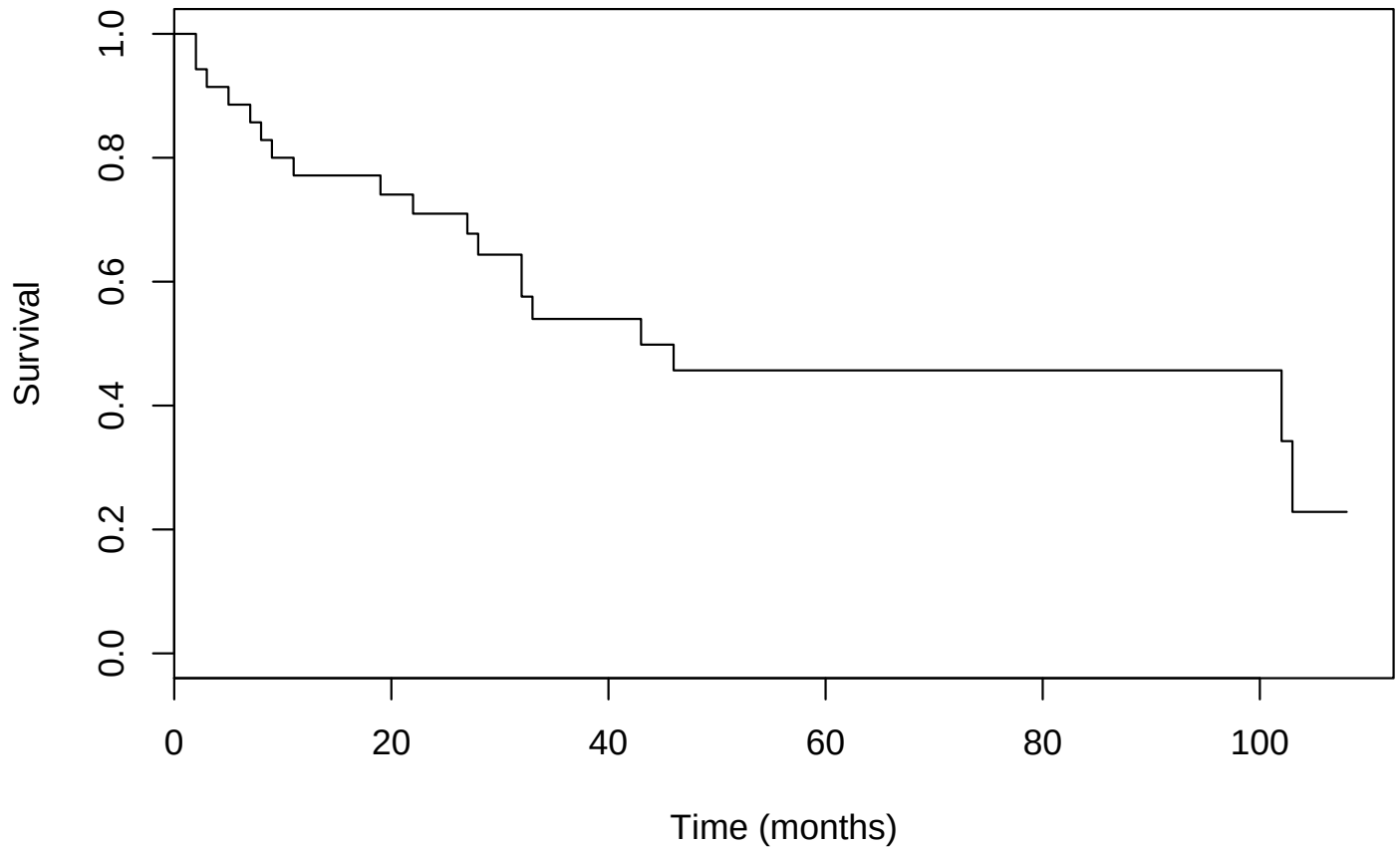
R code and output

```
> colon_sample = mutate(colon_sample,
                        dead=status != "Alive")
> fit = survfit(Surv(surv_mm, dead)~1, colon_sample)
> summary(fit)

Call: survfit(formula = Surv(surv_mm, dead) ~ 1, data = colon_sample)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    2     35       2    0.943  0.0392        0.8690        1.000
    3     33       1    0.914  0.0473        0.8261        1.000
    5     32       1    0.886  0.0538        0.7863        0.998
    7     31       1    0.857  0.0591        0.7487        0.981
    8     30       1    0.829  0.0637        0.7127        0.963
    9     29       1    0.800  0.0676        0.6779        0.944
   11     28       1    0.771  0.0710        0.6441        0.924
   19     25       1    0.741  0.0745        0.6080        0.902
   22     24       1    0.710  0.0776        0.5729        0.879
[snip]
```

Plotting Kaplan-Meier estimates of $S(t)$ using R I

R code

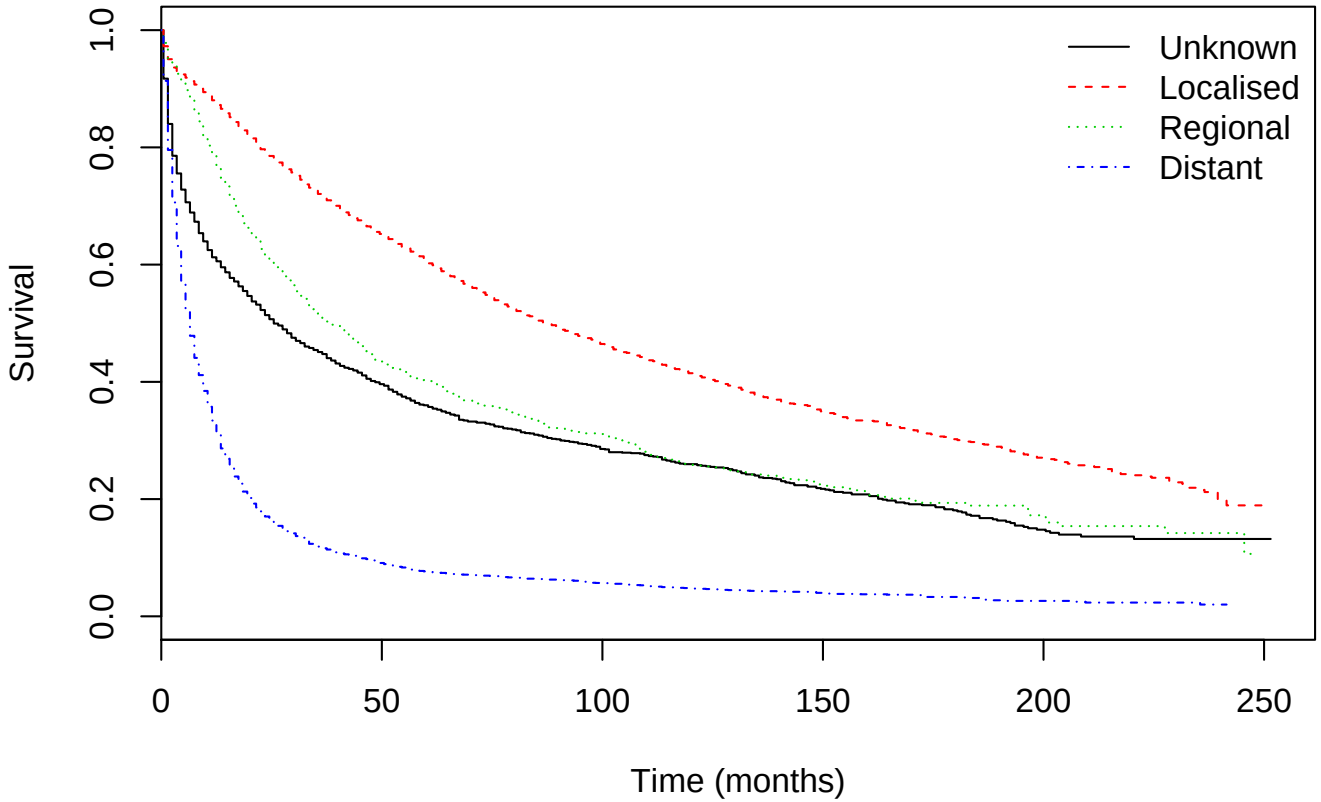
```
> plot(fit, xlab="Time (months)", ylab="Survival",
       conf.int=FALSE)
```



Plotting Kaplan-Meier estimates of $S(t)$ using R II

R code

```
> fit = survfit(Surv(surv_mm,dead)~stage, data=colon)
> plot(fit, xlab="Time (months)",
       ylab="Survival", lty=1:4, col=1:4)
> legend("topright", legend=levels(colon$stage), lty=1:4, col=1:4, bty="n")
```



Confidence intervals for estimated survival proportions I

- Confidence intervals can be calculated for any estimated survival proportion in order to provide a measure of uncertainty associated with the point estimate.
- A 95% confidence interval (CI) is an interval, i.e. a range of values, such that under repeated sampling, the true survival proportion will be contained in the interval 95% of the time (if the model is correct).
- The CI is often called an interval estimate for the true survival proportion, while the estimated survival proportion is called the point estimate.
- Estimated confidence intervals provide an indication of the level of statistical uncertainty in the estimated survival proportions. They do not represent the range of possible prognoses for an individual patient.
- A confidence interval for the true survival proportion can be obtained by assuming that the estimated survival proportion is normally distributed around the true value with estimated variance given by the square of the standard error.

Confidence intervals for estimated survival proportions II

- A two-sided $100(1-\alpha)\%$ confidence interval ranges from $p - z_{\alpha/2}\text{SE}(p)$ to $p + z_{\alpha/2}\text{SE}(p)$, where p is the estimated survival proportion (which can be an interval-specific or cumulative observed, cause-specific, or relative survival), $\text{SE}(p)$ the associated standard error, and $z_{\alpha/2}$ the upper $\alpha/2$ percentage point of the standard normal distribution.
- The standard error of the observed and cause-specific survival proportion can be obtained using Greenwood's method.
- As a rule of thumb, the normal approximation for a single interval i is usually appropriate when both $l_i' p_i$ and $l_i'(1-p_i)$ are greater than or equal to 5 [1].
- Confidence intervals obtained in this way are symmetric about the point estimate and can sometimes contain implausible values for the survival proportion, i.e. values less than zero or greater than one.
- The usual approach is to construct the confidence intervals on the log cumulative hazard scale and then back-transform.

Life table method for estimating $S(t)$ I

- Also known as the 'actuarial method'. The approach is to divide the period of observation into a series of time intervals and estimate the conditional (interval-specific) survival proportion for each interval.
- The cumulative survival function, $S(t)$, at the end of a specified interval is then given by the product of the interval-specific survival proportions for all intervals up to and including the specified interval.
- In the absence of censoring, the interval-specific survival proportion is $p = (I - d)/I$, where d is the number of events (deaths) observed during the interval and I is the number of patients alive at the start of the interval.
- In the presence of censoring, it is assumed that censoring occurs uniformly throughout the interval such that each individual with a censored survival time is at risk for, on average, half of the interval. This assumption is known as the actuarial assumption.
- The effective number of patients at risk during the interval is given by $I' = I - \frac{1}{2}w$ where I is the number of patients alive at the start of the interval and w is the number of censorings during the interval.

Life table method for estimating $S(t)$ II

- The estimated interval-specific survival proportion is then given by $p = (I' - d)/I'$.
- The actuarial estimate of the cumulative survival function at time t is given by

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{t_i \leq t}(1 - \frac{d_i}{I_i - w_i/2}) & \text{if } t \geq t_1 \end{cases} \quad (2)$$

- For the first interval, $I = I' = 35$ and $p = (35 - 8)/35 = 0.771$. The estimated 1-year survival proportion is therefore $\hat{S}(1) = 0.771$.
- For the second interval, $I' = 27 - \frac{1}{2} \times 2 = 26$ and $p = (26 - 2)/26 = 0.923$.
- The estimated 2-year survival proportion is then $\hat{S}(2) = 0.771 \times 0.923 = 0.71209$.
- The cumulative survival estimated is estimated as the product of conditional survival proportions, where the estimate of each conditional survival proportion is based upon only those individuals under follow-up.
- That is, the individuals who are censored are assumed to have the same prognosis as those individuals who could be followed up.

Life table method for estimating $S(t)$ III

- This requires the assumption that censoring is *non-informative*.
- That is, we make the assumption that, conditional on the values of any explanatory variables, censoring is unrelated to prognosis (the probable course and outcome of the disease).
- If censoring was informative, for example if censored were more likely to die, then we would be left with healthier patients in the study, showing a better survival than the true survival of the patients.
- In the first exercise you will construct (by hand) a life table on these same data but with death due to cancer as the outcome.

Life table method for estimating $S(t)$ IV

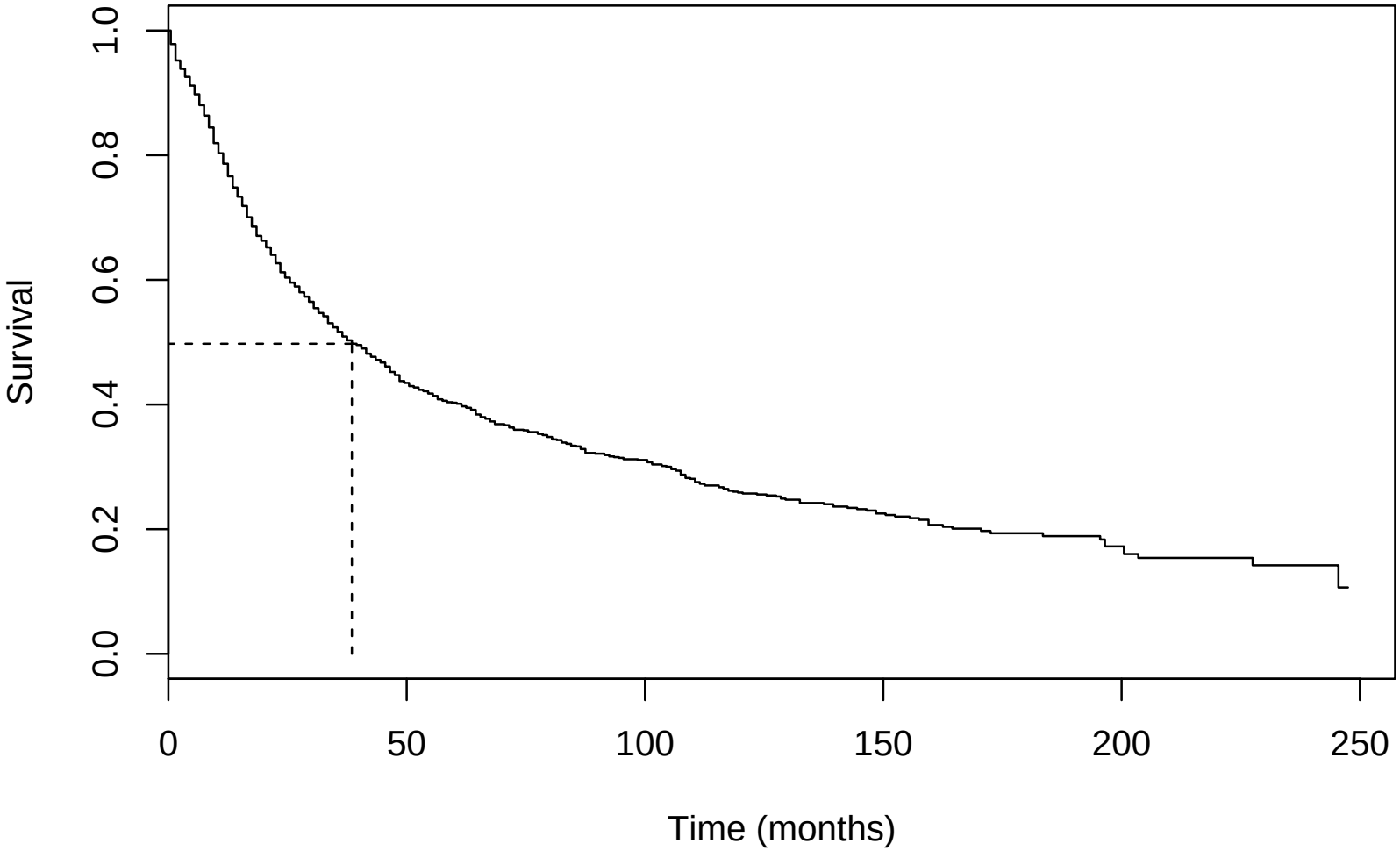
time	l	d	w	l'	p	$S(t)$
[0-1)	35	8	0	35.0	0.77143	0.77143
[1-2)	27	2	2	26.0	0.92308	0.71209
[2-3)	23	5	4	21.0	0.76190	0.54254
[3-4)	14	2	1	13.5	0.85185	0.46217
[4-5)	11	0	1	10.5	1.00000	0.46217
[5-6)	10	0	0	10.0	1.00000	0.46217
[6-7)	10	0	3	8.5	1.00000	0.46217
[7-8)	7	0	1	6.5	1.00000	0.46217
[8-9)	6	2	3	4.5	0.55556	0.25676
[9-10)	1	0	1	0.5	1.00000	0.25676

- l is the number alive at the start of the interval
- d is the number of events (deaths) during the interval
- w is the number of censorings (withdrawals) during the interval
- l' is the effective number at risk for the interval
- p is the interval-specific survival proportion
- $S(t)$ is the estimated cumulative survival function (proportion) at the end of the interval

Other measures of survival: Median survival time I

- The median survival time is another measure used to summarise the survival experience of the patients.
- The median survival time is the time at which $S(t) = 0.5$. That is, the time beyond which 50% of the individuals in the population are expected to survive.
- It is estimated by the time at which the estimate of $S(t)$ falls below 0.5.
- The median survival time for the example shown on the next slide is approximately 3.5 years.
- The median can be estimated by extrapolation if the survival function does not sink below 0.5 during the period the patients are under follow-up.

Other measures of survival: Median survival time II



Testing for differences in survival – Summary of key points

- Various tests are available for testing equality of survival curves, the most well-known being the log rank test.
- These tests are rarely used in observational epidemiology; we prefer to use modelling since it:
  1. provides estimates of the size of the effect (i.e. rate ratios); the log-rank test just gives a p-value;
  2. provides greater possibilities for confounder control and effect modification.
- The log-rank test assumes proportional hazards.
- Consider the situation where we have two groups; a Cox model with one explanatory variable gives us everything the log-rank test does (a p-value). It also gives us the estimated hazard ratio and CI but, more importantly, it is simple to extend the model to compare survival between the two groups while controlling for potential confounders.

Testing for differences in survival between groups I

- Comparing survival at a fixed time point (e.g. five years) wastes available information.
- It is invalid to compare the proportion surviving at a given time, based on the comparison of two binomial proportions, where the time point for comparison is chosen after viewing the estimated survival functions (e.g. testing for a difference at the point where the Kaplan-Meier curves show the largest difference).
- Various tests are available (parametric and non-parametric) for testing equality of survival curves. The most common is the log rank test, which is non-parametric.
- Start by tabulating the number at risk in each group and the total number of events (deaths) at every time point when one of more deaths occur.
- Under the null hypothesis that the two survival curves are the same, the expected number of deaths in each group will be proportional to the number at risk in each group.

Testing for differences in survival between groups II

- For example (see slide 80), at $t = 2$ months we observed 2 deaths (one male and one female). Conditional on 2 deaths being observed, we would expect $2 \times 19/35 = 1.086$ deaths among the 19 males at risk and $2 \times 16/35 = 0.914$ deaths among the 16 females at risk.
- Now calculate the totals of the observed and expected number of deaths for each group (1=males, 2=females), calling them O_1, O_2, E_1, and E_2, and calculate the following test statistic

$$\theta = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}. \tag{3}$$

Testing for differences in survival between groups III

- Under the null hypothesis, θ will approximately follow a χ^2 distribution with 1 degree of freedom. That is, if θ is greater than 3.84 then we reject the null hypothesis and conclude that there is a statistically significant difference between the two survival curves.

Log rank test for comparing survival of males and females I

event time	males			females		
	at risk	obs	exp	at risk	obs	exp
2	19	1	1.086	16	1	0.914
3	18	1	0.545	15	0	0.455
5	17	1	0.531	15	0	0.469
7	16	1	0.516	15	0	0.484
8	15	1	0.500	15	0	0.500
9	14	0	0.483	15	1	0.517
11	14	0	0.500	14	1	0.500
19	13	0	0.520	12	1	0.480
22	13	1	0.542	11	0	0.458
27	12	0	0.545	10	1	0.455
28	11	0	0.550	9	1	0.450
32	11	2	1.158	8	0	0.842
33	9	1	0.563	7	0	0.438
43	8	0	0.615	5	1	0.385
46	8	1	0.667	4	0	0.333
102	2	0	0.500	2	1	0.500
103	2	1	0.667	1	0	0.333
Totals: $O_1 = 11$, $E_1 = 10.488$, $O_2 = 8$, $E_2 = 8.512$						

Log rank test for comparing survival of males and females II

- The test statistic is $\theta = (O_1 - E_1)^2/E_1 + (O_2 - E_2)^2/E_2 = 0.056$, which is less than 3.84 implying no evidence of a difference in survival between males and females.
- For k groups, the log rank test statistic is

$$\theta = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{4}$$

  which has an approximate χ^2_{k-1} distribution under the null hypothesis.
- The log rank test is designed to be sensitive to departures from the null hypothesis in which the two hazards (instantaneous death rates) are proportional over time. It is very insensitive to situations in which the hazard functions cross.
- The log rank test puts equal weight on every failure (irrespective of the number at risk at the time of the failure).

Log rank test for comparing survival of males and females III

- An alternative test, the generalised Wilcoxon test, is constructed by weighting the contribution of each failure time by the total number of individuals at risk and is consequently more sensitive to differences early in the follow-up period (when the number at risk is larger).
- The Wilcoxon test is more powerful than the log rank test if the proportional hazards assumption does not hold.
- It is difficult to apply the log rank test while simultaneously controlling for potential confounding variables (a regression approach is preferable).
- In a randomised clinical trial, however, potential confounders are controlled for in the randomisation, so we can use the log rank test to compare survival curves for the different treatment groups.
- The log rank test provides nothing more than a test of statistical significance for the difference between the survival curves, it tells us nothing about the size of the difference. A regression approach allows us to both determine statistical significance and to estimate the size of the effect.

Log rank test in R

R code and output

```
> survdiff(Surv(surv_mm, dead)~sex, colon_sample)

Call:
survdiff(formula = Surv(surv_mm, dead) ~ sex, data = colon_sample)

             N Observed Expected (O-E)^2/E (O-E)^2/V
sex=Male    19       11    10.49    0.0250     0.057
sex=Female 16        8     8.51    0.0309     0.057

 Chisq= 0.1  on 1 degrees of freedom, p= 0.811
```

- The log rank test is non-significant (p-value 0.811) indicating no difference in survival between males and females (if the model is correct – such that there is no uncontrolled confounding).

The same test as a Cox model

R code and output

```
> coxph(Surv(surv_mm, dead)~sex, colon_sample)

Call:
coxph(formula = Surv(surv_mm, dead) ~ sex, data = colon_sample)

             coef exp(coef) se(coef)     z   p
sexFemale -0.116     0.891    0.467 -0.25 0.8

Likelihood ratio test=0.06  on 1 df, p=0.804
n= 35, number of events= 19
```

Summary of Day 1 I

- Time-to-event analysis (survival analysis) is necessary when
  - We are interested in studying the time to an event, e.g. time to diagnosis, time to death
  - Individuals in a study are followed for different lengths of time, and therefore are 'at risk' for different amounts of time, e.g. in cohort studies.
- The outcome in survival analysis consists of both an event indicator (0/1) and a time dimension (continuous).
- The outcome can be expressed as either a survival proportion or an event rate (hazard). Comparison between groups are primarily made using hazard ratios.
- The survival function (survival proportion) can be estimated using several alternative methods.
- The log rank test can be used to test for differences in survival, but is rarely used in observational epidemiology.

Summary of Day 1 II

- In observational epidemiology we prefer modelling since it:
  - enables us to compare survival between exposure categories while controlling for confounding (although we can also perform an adjusted log rank test).
  - places a focus on estimation rather than testing (i.e. we obtain estimated hazard ratios and CIs).
  - enables us to study effect modification.
  - is extendable in other useful ways.

Exercises for Monday afternoon

1a. Hand calculation: Kaplan-Meier estimates of cause-specific survival (35 patients)

1b. Kaplan-Meier estimates of cause-specific survival using R (35 patients)

2. Melanoma: Comparing survival proportions and mortality rates according to stage

3. Localised melanoma: Comparing estimates of cause-specific survival between periods; first graphically and then using the log rank test

4. Localised melanoma: Comparing various approaches to estimating the 10-year survival proportion

References I

D. G. Altman.

Practical Statistics for Medical Research.
London: Chapman and Hall, 1991.

P. E. Böhmer.

Theorie der unabhängigen Wahrscheinlichkeiten.
Rapports, Mémoires et Procès-verbaux de Septième Congrès International d'Actuaires, Amsterdam, 2:327–343, 1912.

N. E. Breslow and N. E. Day.

Statistical Methods in Cancer Research: Volume II - The Design and Analysis of Cohort Studies.
IARC Scientific Publications No. 82. Lyon: IARC, 1987.

D. Clayton and M. Hills.

Statistical Models in Epidemiology.
Oxford: Oxford University Press, 1993.

M. Greenwood.

The Errors of Sampling of the Survivorship Table, volume 33 of *Reports on Public Health and Medical Subjects*.
London: Her Majesty's Stationery Office, 1926.

E. L. Kaplan and P Meier.

Nonparametric estimation from incomplete observations.
Journal of the American Statistical Association, 53:457–481, 1958.

References II

Tyler J VanderWeele.
On the distinction between interaction and effect modification.
Epidemiology, 20(6):863–871, 2009.