

Survival analysis

Manuela Zucknick

Oslo Center for Biostatistics and Epidemiology, UiO
manuela.zucknick@medisin.uio.no

MF9130E – Introductory Statistics

May 11, 2023

Overview

Aalen chapter 13, Kirkwood and Sterne chapter 26

- Life tables and **survival data**
- **Univariable survival analysis**
 - ▶ Survival curves
 - ▶ Kaplan-Meier
- **Multivariable survival analysis**
 - ▶ Cox regression and other alternatives

① Introduction: What makes survival data special?

Survival analysis

- Want to analyse data where time until an event is of interest - often called **failure time**, **survival time** or **event time**
- **One of the most applied statistical methodologies in medicine**
- Reinvented many times and **also used extensively in other fields**, such as reliability engineering, sociology, demography and actuarial science

Statistical methods in the NEJM



The NEW ENGLAND
JOURNAL of MEDICINE

CORRESPONDENCE

Statistical Methods in the *Journal*

N Engl J Med 2005; 353:1977-1979

- data from 311 articles published in volumes 350 through 352 (January 2004 through June 2005)
- Use of t-tests decreased from 44% in 1978-79 to 26% in 2004-05
- Only 21% of the articles are accessible to a reader with only an introductory course in statistics
- >50% of the papers use more advanced statistical methods, e.g. multiple regression and survival analysis

Table 1. Statistical Content of Original Articles in the *New England Journal of Medicine* over Time.^a

Statistical Procedure	Original Articles Containing Methods			Accumulation by Article‡
	1978–1979	1989	2004–2005	2004–2005
	<i>number (percent)</i>			
No statistical methods or descriptive statistics only	91 (27)	14 (12)	39 (13)	39 (13)
t-Tests	147 (44)	45 (39)	80 (26)	42 (14)
Contingency tables	91 (27)	41 (36)	166 (53)	47 (15)
Nonparametric tests	38 (11)	24 (21)	85 (27)	53 (17)
Epidemiologic statistics	33 (10)	25 (22)	110 (35)	55 (18)
Pearson's correlation	40 (12)	22 (19)	10 (3)	56 (18)
Simple linear regression	28 (8)	10 (9)	19 (6)	56 (18)
Analysis of variance	25 (8)	23 (20)	50 (16)	61 (20)
Transformation	23 (7)	8 (7)	31 (10)	62 (20)
Nonparametric correlation	13 (4)	1 (1)	14 (5)	65 (21)
Survival methods	36 (11)	37 (32)	190 (61)	74 (24)
Multiple regression	15 (5)	16 (14)	160 (51)	122 (39)
Multiple comparisons	11 (3)	10 (9)	70 (23)	127 (41)
Adjustment and standardization	9 (3)	10 (9)	3 (1)	128 (41)
Multitway tables	12 (4)	11 (10)	39 (13)	136 (44)
Power analyses	10 (3)	4 (3)	121 (39)	211 (68)
Cost-benefit analysis	3 (1)	0	1 (<1)	212 (68)
Sensitivity analysis‡	0	0	18 (6)	223 (72)
Repeated-measures analysis	—	—	37 (12)	249 (80)
Missing-data methods	—	—	26 (8)	272 (87)
Noninferiority trials	—	—	11 (4)	283 (91)
Receiver-operating characteristic	—	—	7 (2)	288 (93)
Resampling	—	—	5 (2)	293 (94)
Principal component analysis and cluster analysis	—	—	5 (2)	298 (96)
Other methods§	12 (4)	10 (9)	13 (4)	311 (100)
Total				
Articles	332	115	311	311
Article-method uses	637	311	1310	
Average uses of methods per article	1.9	2.7	4.2	

History of survival analysis

- Roots back to **John Graunt**, who published *Natural and Political Observations Made upon the Bills of Mortality* in 1662
- Graunt was interested in **mortality during the last great plague** in Europe. He produced tables with commentaries, did basic calculations, and compared the number of male and female births and deaths
- Until well after the Second World War the field was dominated by the classical approaches developed by the **early actuaries**
- Modern survival analysis started with Kaplan-Meier (1958), Cox (1972) and **Aalen** (1975)

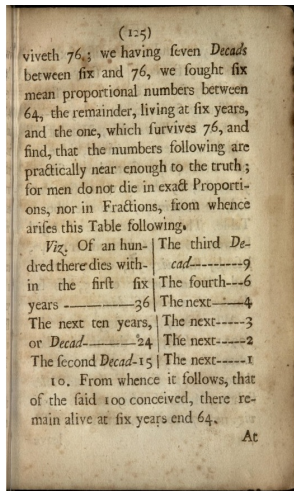
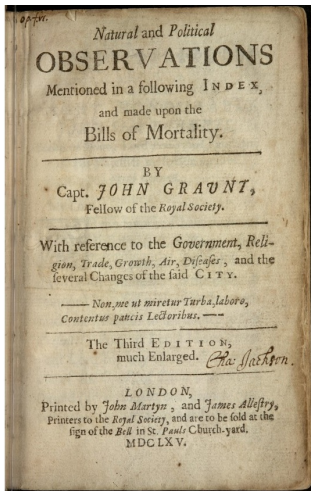


Figure: *Observations* and the first known primitive life table, which became one of the main tools of demography and insurance mathematics. Grant is considered one of the first demographers and epidemiologists.

Life tables

- Show the **probability of surviving any particular year** of age
- Can be used to calculate remaining **life expectancy** for people at different ages
- Graunt's data on deaths from *Observations* 1662 in a life table:

Age	Deaths	Survivors
-	-	100
0-6	36	64
6-16	24	40
16-26	15	25
26-36	9	16
36-46	6	10
46-56	4	6
56-66	3	3
66-76	2	1
76-	1	0

Survival curves

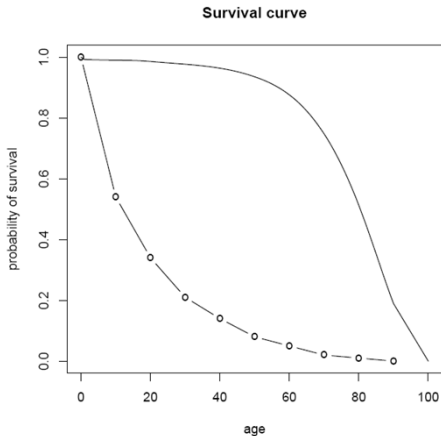


Figure: Graunt's **survival curve** compared to US 2000 mortality.

- Expected years of life in Graunt's data: 18 (median survival)

Response variables in survival analysis

- In regular survival analysis we study **time until a dichotomous (binary) outcome**, e.g.:
 - ▶ Time until death
 - ▶ Time until tumor recurrence
 - ▶ Time until AIDS for HIV patients
 - ▶ Time until machine part fails
 - ▶ Age at breast cancer diagnosis
- Durations are **important clinical and epidemiological outcome parameters**
 - ▶ What is the expected survival time for specific patient?
 - ▶ Do patients live longer?
 - ▶ Does the remission period increase?
 - ▶ Can we postpone disease?

What makes survival data special?

- **Right skewed data**

- ▶ Survival times are non-negative and therefore typically skew to the right
- ▶ Naive analysis of un-transformed survival times unpromising

- **Censoring**

- ▶ Incompletely observed times
- ▶ Typically due to either 1) dropout or 2) end of study
- ▶ Not taking censoring into account can cause seriously biased results

Censoring and survival analysis

- Censoring **rules out ordinary statistical methods** for survival time data
- In reality we keep track of time until two different types of events: **the event of interest *and* censoring**, where the latter includes any other events terminating observation

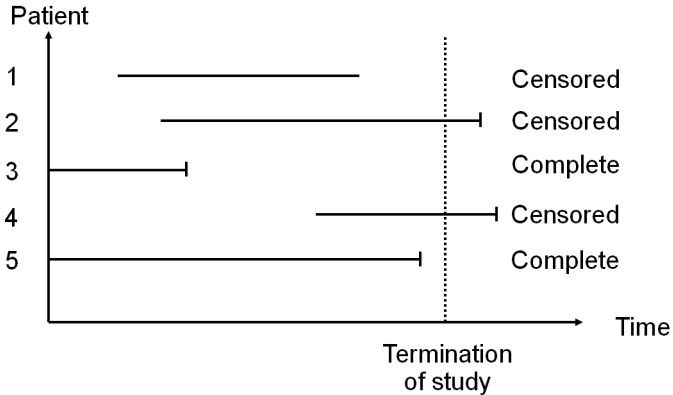


Figure: Illustration of typical survival data on **calendar scale**.

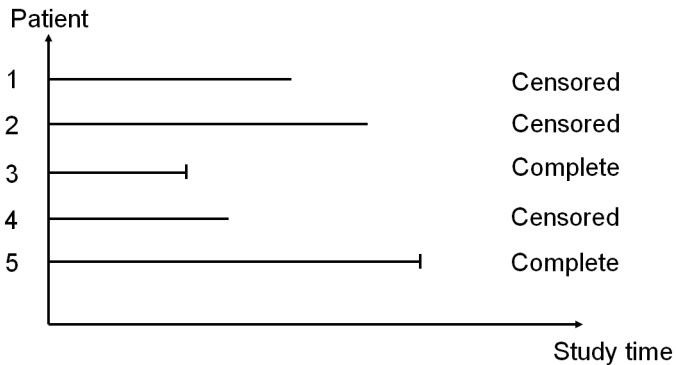


Figure: Illustration of typical survival data on **study time scale**.

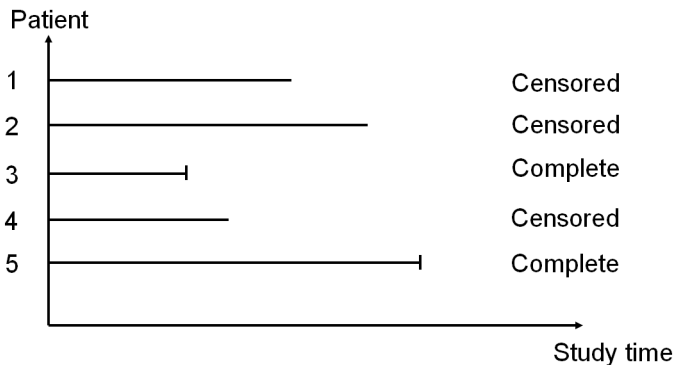


Figure: Illustration of typical survival data on **study time scale**.

Key concept

The risk set at time t - the individuals under observation at time t

Censoring (cont)

- Traditional censoring is often called **right censoring**
- A related term is **left truncation**: e.g. if patients are not included in the study from baseline, but come in later
Left censoring and right truncation also exists, but are not common
- The key assumption of all basic survival methods is **independent censoring** – The individuals which get censored at any give time shall not differ (on average) from those observed

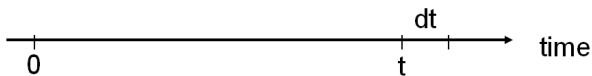
A small data example

- **Dataset:** 26, 17, 7*, 41, 34*, 9, 13, 25*, 37, 18
* denotes censoring time
- The same **data ordered:**
7*, 9, 13, 17, 18, 25*, 26, 34*, 37, 41
- The typical set-up **for most software:**

Time	Event
26	1
17	1
7	0
41	1
⋮	⋮

The survival and hazard

What makes survival data special?



- **Survival function:**

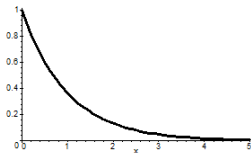
$$S(t) = P(\text{event does not occur before time } t)$$

- **Hazard function** or hazard rate:

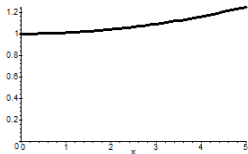
$$h(t) = \frac{1}{dt} P(\text{event occur in } (t, t + dt), \text{ given no event before } t)$$

Illustrations

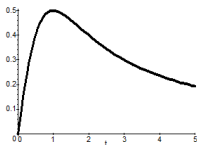
- **Survival curve:** describe the proportion that survives up to some time



- **Hazard curve:** describe the risk of the event (death, relapse etc) as function of time



Hazard rates:



Survival curves:

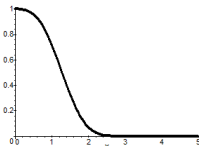
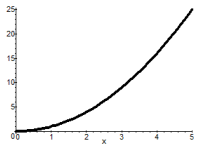
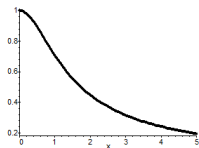
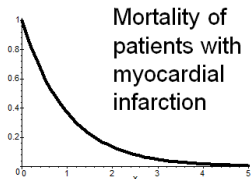
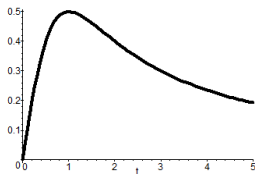


Figure: Connecting hazard (left) and survival curves (right).

Divorce rates
Mortality of cancer
patients
Incidence of
childhood leukemia



General
mortality
Incidence
of most
cancers

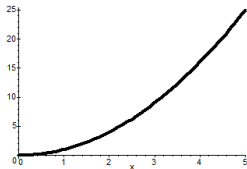


Figure: The shapes of hazard curves - three examples.

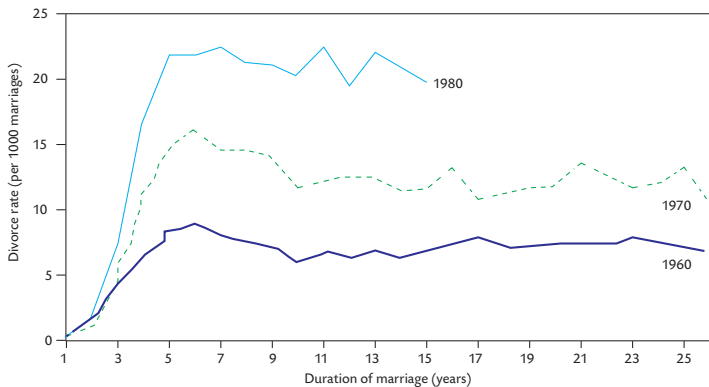


Figure 5.2 Hazard rates of divorce for Norwegian couples married in 1960, 1970, and 1980. (Based on data from Statistics Norway.)

Figure: Rates of divorces for couples married in Norway in 1960, 1970 and 1980 (hazard/incidence rate).

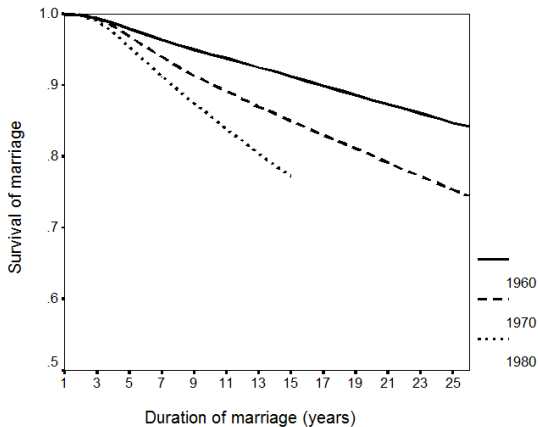
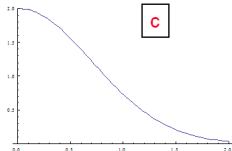
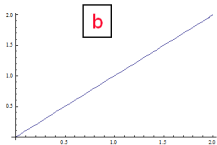
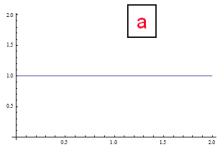


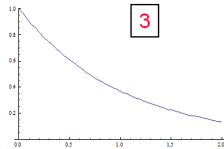
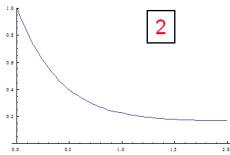
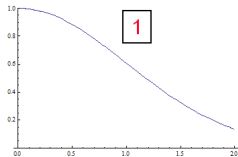
Figure: “Survival” of marriages among the same couples.

Exercise: Connect the hazard and survival functions!

Hazard rates:



Survival functions:



Formal notation

- T denotes the response variable, $T \geq 0$
- $S(t) = P(T > t)$
- $h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t < T \leq t + \Delta t | T > t)}{\Delta t}$

Formal connection of hazards and survival

- Cumulative hazard rate: $H(t) = \int_0^t h(s) ds$
- Survival function: $S(t) = \exp(-H(t)) = \exp(-\int_0^t h(s) ds)$
- Hazard rate: $h(t) = -\frac{d}{dt} \ln(S(t)) = -\frac{S'(t)}{S(t)}$

② Univariable survival analysis: Kaplan-Meier & Logrank

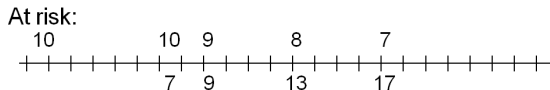
Estimating the survival function: The Kaplan-Meier estimator

- Let all event times be ordered and t_j be the j 'th event
Let r_j be the number *at risk* at time t_j
The Kaplan-Meier **estimator for the probability of surviving until time t** is given by

$$\hat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{1}{r_j}\right)$$

Example: Kaplan-Meier

- **Multiply survival probabilities** for small intervals



$$\text{Survival: } \left(1 - \frac{1}{9}\right) \times \left(1 - \frac{1}{8}\right) \times \left(1 - \frac{1}{7}\right) \times \dots$$

Example: Kaplan-Meier

- Small **example from earlier**:

Patient no.	Survival time	Patients at risk	Survival factors	K-M estimator
1	7*	10	1	1.00
2	9	9	1 - 1/9	0.89
3	13	8	1 - 1/8	0.78
4	17	7	1 - 1/7	0.67
5	18	6	1 - 1/6	0.56
6	25*	5	1	0.56
7	26	4	1 - 1/4	0.42
8	34*	3	1	0.42
9	37	2	1 - 1/2	0.21
10	41	1	1 - 1/1	0

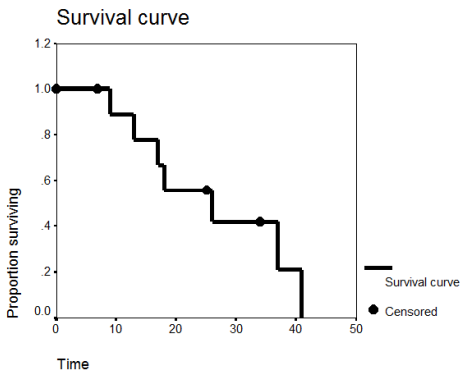


Figure: Kaplan-Meier plot.

- Median survival: 26 days

Exercise: Kaplan-Meier

- **Compute** the Kaplan-Meier survival probabilities for the following survival data :

5, 12*, 14, 16*, 20

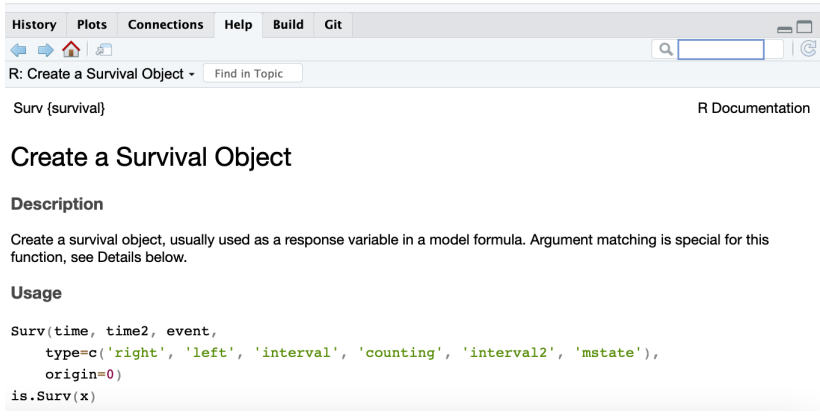
(e.g. by filling in a table as earlier)

- Make a **sketch** of the survival curve

Basic survival analysis in R: Getting started

- We need the R package `survival`

```
> # install.packages("survival")
> library(survival)
>
> ?Surv
> |
```



The screenshot shows the RStudio interface with the documentation for the `Surv` function. The top menu bar includes History, Plots, Connections, Help, Build, and Git. Below the menu bar is a search bar and a "Find in Topic" input field. The main content area displays the title "Create a Survival Object" and the "Description" section, which states: "Create a survival object, usually used as a response variable in a model formula. Argument matching is special for this function, see Details below." The "Usage" section shows the function signature: `Surv(time, time2, event, type=c('right', 'left', 'interval', 'counting', 'interval2', 'mstate'), origin=0)` and `is.Surv(x)`. The text "Surv {survival}" and "R Documentation" are visible in the top right of the content area.

Surv {survival} R Documentation

Create a Survival Object

Description

Create a survival object, usually used as a response variable in a model formula. Argument matching is special for this function, see Details below.

Usage

```
Surv(time, time2, event,
      type=c('right', 'left', 'interval', 'counting', 'interval2', 'mstate'),
      origin=0)
is.Surv(x)
```

Set up a survival variable

See Also

[coxph](#), [survfit](#), [survreg](#), [lung](#).

Examples

[Run examples](#)

```
with(aml, Surv(time, status))  
survfit(Surv(time, status) ~ ph.ecog, data=lung)  
Surv(heart$start, heart$stop, heart$event)
```

- We use the function `Surv(time, status)` to set up a survival variable.
- The **data need to be on the following form:**

id	time	status
1	26	1
2	17	1
3	7	0
4	41	1
⋮	⋮	⋮

R help page for **Surv**

History Plots Connections Help Build Git

R: Examples for 'survival::Surv' - Find in Topic

Examples for 'survival::Surv'

Create a Survival Object

Aliases: [Surv](#) [is.Surv](#) [.Surv](#)

Keywords: [survival](#)

```
### ** Examples
```

```
with(aml, Surv(time, status))
```

[1]	9	13	13+	18	23	28+	31	34	45+	48	161+	5	5	8	8
[16]	12	16+	23	27	30	33	43	45							

Example data set `aml` in the `survival` package

```
> head(aml)
  time status      x
1    9      1 Maintained
2   13      1 Maintained
3   13      0 Maintained
4   18      1 Maintained
5   23      1 Maintained
6   28      0 Maintained

> tail(aml)
  time status      x
18   23      1 Nonmaintained
19   27      1 Nonmaintained
20   30      1 Nonmaintained
21   33      1 Nonmaintained
22   43      1 Nonmaintained
23   45      1 Nonmaintained

> Surv(aml$time, aml$status)
 [1]  9  13  13+ 18  23  28+ 31  34  45+ 48 161+  5  5  8  8 12
[17] 16+ 23  27  30  33  43  45
```

History Plots Connections Help Build Git

R: Acute Myelogenous Leukemia survival data - Find in Topic

aml {survival} R Documentation

Acute Myelogenous Leukemia survival data

Description

Survival in patients with Acute Myelogenous Leukemia. The question at the time was whether the standard course of chemotherapy should be extended ('maintenance') for additional cycles.

Usage

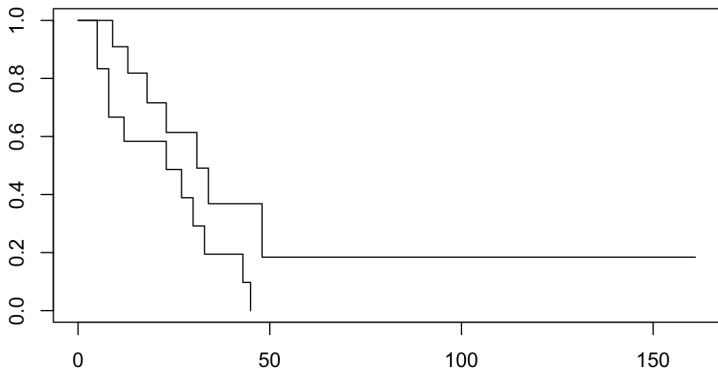
```
aml  
leukemia  
data(cancer, package="survival")
```

Format

time: survival or censoring time
status: censoring status
x: maintenance chemotherapy given? (factor)

Plotting **Kaplan-Meier** curves

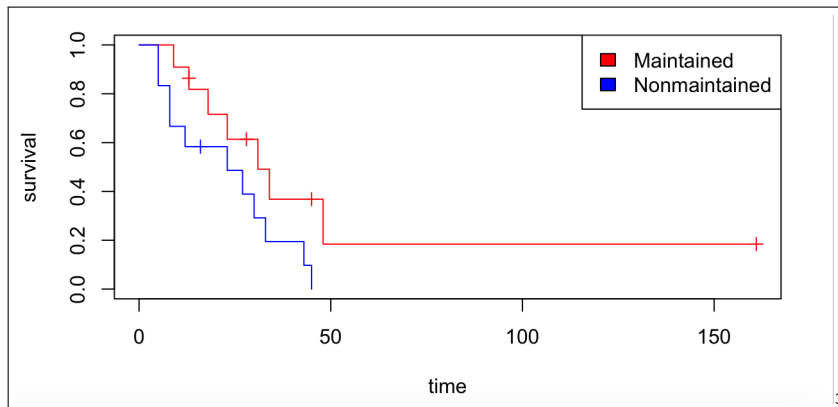
```
>  
> plot(survfit(Surv(time, status) ~ x, data=aml))  
>
```



Annotate Kaplan-Meier curves

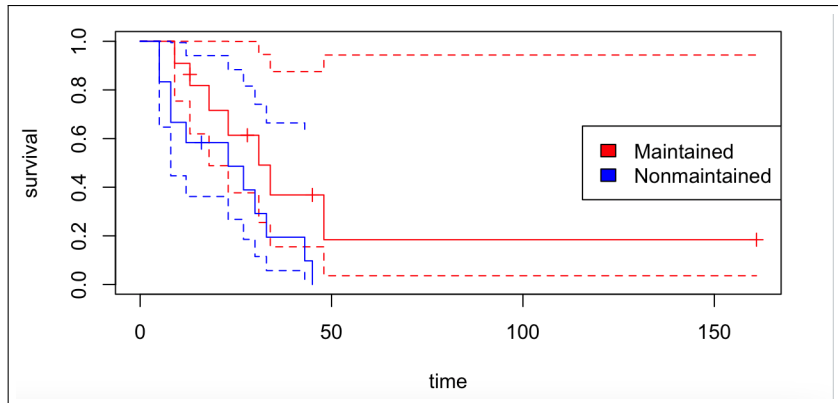
- with censoring marks, axis labels and labelled curves:

```
> plot(fit, col=c("red","blue"), mark.time=TRUE,  
+       xlab="time", ylab="survival")  
> legend("topright", fill=c("red","blue"),  
+       legend=levels(aml$x))
```



Plotting Kaplan-Meier curves with CIs

```
> plot(fit, col=c("red","blue"), mark.time=TRUE,  
+       xlab="time", ylab="survival", conf.int=TRUE)  
> legend("right", fill=c("red","blue"),  
+       legend=levels(aml$x))
```



Comparing groups

Option 1: Compare survival rates at a specified time point, e.g 5-year survival after melanoma surgery

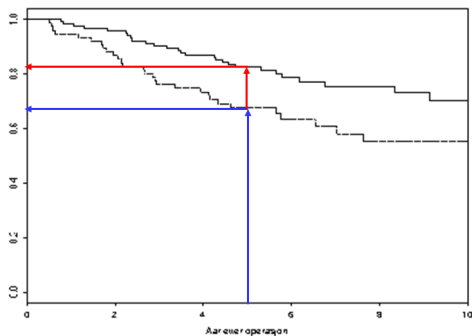


Figure: Kaplan-Meier plot for woman (red) and men (blue).

- Five year survival for women: 82%; and for men: 68%

Comparing groups

Option 2: Report median survival probabilities

```
> survfit(Surv(time, status) ~ ph.ecog, data=lung)
Call: survfit(formula = Surv(time, status) ~ ph.ecog, data = lung)
```

```
1 observation deleted due to missingness
```

	n	events	median	0.95LCL	0.95UCL
ph.ecog=0	63	37	394	348	574
ph.ecog=1	113	82	306	268	429
ph.ecog=2	50	44	199	156	288
ph.ecog=3	1	1	118	NA	NA

- Time when the survival probability is 50% in both groups with 95% confidence intervals

Comparing groups

Option 2: Report median survival probabilities

```
> survfit(Surv(time, status) ~ x, data=aml)
Call: survfit(formula = Surv(time, status) ~ x, data = aml)
```

	n	events	median	0.95LCL	0.95UCL
x=Maintained	11	7	31	18	NA
x=Nonmaintained	12	11	23	8	NA

- If the median survival probability is not reached in the observed time frame, then the estimate is reported as NA (missing); equivalently for the confidence limits

The log-rank test

- **The most common test** for the difference between two survival curves; also called the Mantel-Cox test
- Test two general hazard functions, $h_1(t)$ and $h_2(t)$ which we assume to have a proportional relationship.

Test hypotheses:

- ▶ $H_0 : h_1(t) = h_2(t)$
 $H_a : h_1(t) \neq h_2(t)$

- **P-value less than 0.05** \Rightarrow **the hazards are different** between groups

Proportional hazards and log-rank

- **The log-rank test is optimal for proportional hazards** type of comparisons (in term of power)
- The **log-rank test can fail** if the hazards are crossing (do not confuse crossing hazards with crossing survival curves)
- **Other tests** are more suitable for crossing hazards

Log-rank test for difference between groups in R

```
> survdiff(Surv(time, status) ~ x, data=aml)
Call:
survdiff(formula = Surv(time, status) ~ x, data = aml)

              N Observed Expected (O-E)^2/E (O-E)^2/V
x=Maintained  11         7    10.69     1.27     3.4
x=Nonmaintained 12        11     7.31     1.86     3.4

Chisq= 3.4  on 1 degrees of freedom, p= 0.07
```

- Example for how to report this result in a paper:

There was no statistically significant difference between the survival curves for the groups with versus without maintenance chemotherapy (log-rank test, chisq=3.4, p=0.07).

③ Multivariable survival analysis: Cox regression and Co.

Cox regression is the standard approach to add covariates to the analyses

- Model

$$\begin{aligned}h_i(t) &= h_0(t) \cdot \exp(\beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_k x_{ik}) \\ &= h_0(t) \cdot \exp(\beta_1 x_{i1}) \cdot \exp(\beta_2 x_{i2}) \cdot \dots \cdot \exp(\beta_k x_{ik})\end{aligned}$$

- For example: Say that x_1 is smoking (0/1).

$$HR_{smokers/non-smokers} = \frac{h_0(t) \exp(\beta_1 \cdot 1)}{h_0(t) \exp(\beta_1 \cdot 0)} = \exp(\beta_1)$$

- Assumption:
 - ▶ **Proportional hazards**, hence; ***The Cox PH model***
 - ▶ **Multiplicative risk**

Cox vs logistic regression

- Cox model **the hazard rate** (rate per unit time), while logistic regression model **the proportion** in a given time period
- Logistic regression aim to estimate **the odds ratio**, while Cox estimate **the hazard ratio**
- Main practical difference: **survival models handle censoring**
- Note: there are other regression models for survival than Cox, e.g. Aalen's additive model

Other models for multivariable survival analysis

- Alternatives exist, for example **Aalen's additive model**
 - ▶ Does not assume proportional hazards or multiplicative risk
- **Combinations** of Cox and additive models
- **Accelerated failure-time models**

Cox regression is however the dominating model, and the focus in this course.

④ Summary

Key words

- Survival times and censored data
- The survival function and the Kaplan-Meier estimator
- Compare groups: Median survival probabilities or survival rates at a fixed time point (e.g. 5-year survival rates)
- Log-rank test

- (Cox proportional hazards regression)

Notation

- $S(t)$, $H(t)$ and $h(t)$