



“Survival”

...and Event History Analysis



Håkon K. Gjessing

Professor/Principal Investigator

⇒ Centre for Fertility and Health, Norwegian Institute of Public Health, Oslo
– a Norwegian Research Council Centre of Excellence

⇒ Department of Global Public Health and Primary Care, University of Bergen



Kathmandu
Wednesday, 7 October 2026



Event History Analysis: What and Why?

New England Journal of Medicine

Editorial, Jan. 6, 2000, p. 42-49

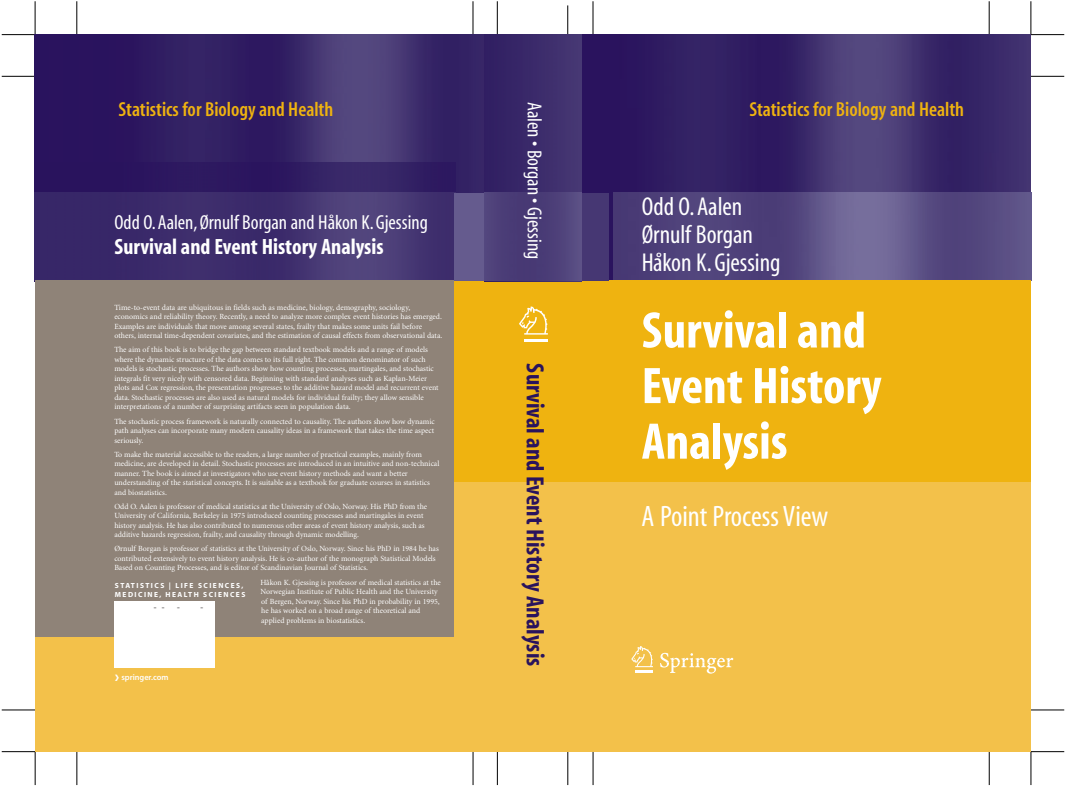
The eleven most important developments in medicine over the past millennium

- Elucidation of human anatomy and physiology
- Discovery of cells and their substructures
- Elucidation of the chemistry of life
- *Application of statistics to medicine*
- Development of anesthesia
- Discovery of the relation of microbes to disease
- Elucidation of inheritance and genetics
- Knowledge of the immune system
- Development of body imaging
- Discovery of antimicrobial agents
- Development of molecular pharmacotherapy

Event History Analysis: What and Why?

Application of statistics to medicine

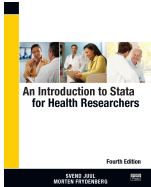
“Sir David Cox’s 1972 paper on proportional-hazards regression ignited the fields of survival analysis and semiparametric inference (using partial specification of the probability distribution of the outcomes under investigation). Rapid improvements in computer support were essential to the growing role of empirical investigation and statistical inference.”



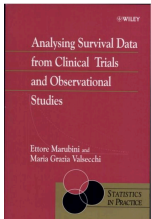
<https://link.springer.com/book/10.1007/978-0-387-68560-1>
<https://www.mn.uio.no/math/english/people/aca/borgan/data/> (Example data used in book)

A sample of books

An Introduction to Stata for Health Researchers, Fourth Edition
Svend Juul and Morten Frydenberg
Stata Press, 2014



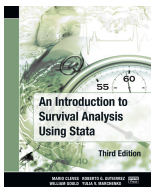
Analysing Survival Data from Clinical Trials and Observational Studies
Ettore Marubini, Maria Grazia Valsecchi
Wiley, 2004



Survival Analysis and Epidemiological Tables Reference Manual
Stata Press, 2013



An Introduction to Survival Analysis Using Stata, Third Edition
Mario Cleves, William Gould, Roberto G. Gutierrez, and Yulia V. Marchenko
Stata Press, 2010



Some examples

1. Time from cancer diagnosis to death
Censoring: Cancer patients get transferred to another hospital (loss-to-followup)
2. Time from started malaria treatment to cured
Censoring: Patients end the treatment when most severe symptoms end
3. Time from inserting a dental filling to when it fails
Censoring: Study ends after 5 years
4. Time from first birth to the second (for the same mother)
Censoring(?): Mother too high age, or decides not to have more children
5. Time from hip prosthesis operation to failure/re-operation
Censoring: The prosthesis lasts for the rest of the patients's life
6. Time from entering marriage to divorce
Censoring: The couple moves abroad (loss-to-followup), or never get divorced!

NOTE: Is censoring **independent**?

(not necessarily the case in every example above)

Independent Censoring

Intuitively:

- A censored individual is not “special” in any way, i.e.:
 - at any time before censoring, the risk of an event is the same as for everybody else.
 - after the time of censoring, the individual would have had the same risk profile as everybody else if it could have been observed.

Examples:

- Patients leave treatment/followup when most severe symptoms end.
Probably not independent
- End of study. Usually independent

Event history analysis: Choice of time scale

Three most common time scales:

1. Time from inclusion to event (study time)

Example: Time from cancer diagnosis to death

Zero: Date of inclusion (individual)

2. Calendar time

Example: Time from a fixed date (e.g. 01 Jan 2020) to infection with the SARS-CoV-2 virus

Zero: Start date (common)

3. Age

Example: Age at death

Zero: Date of birth (individual)

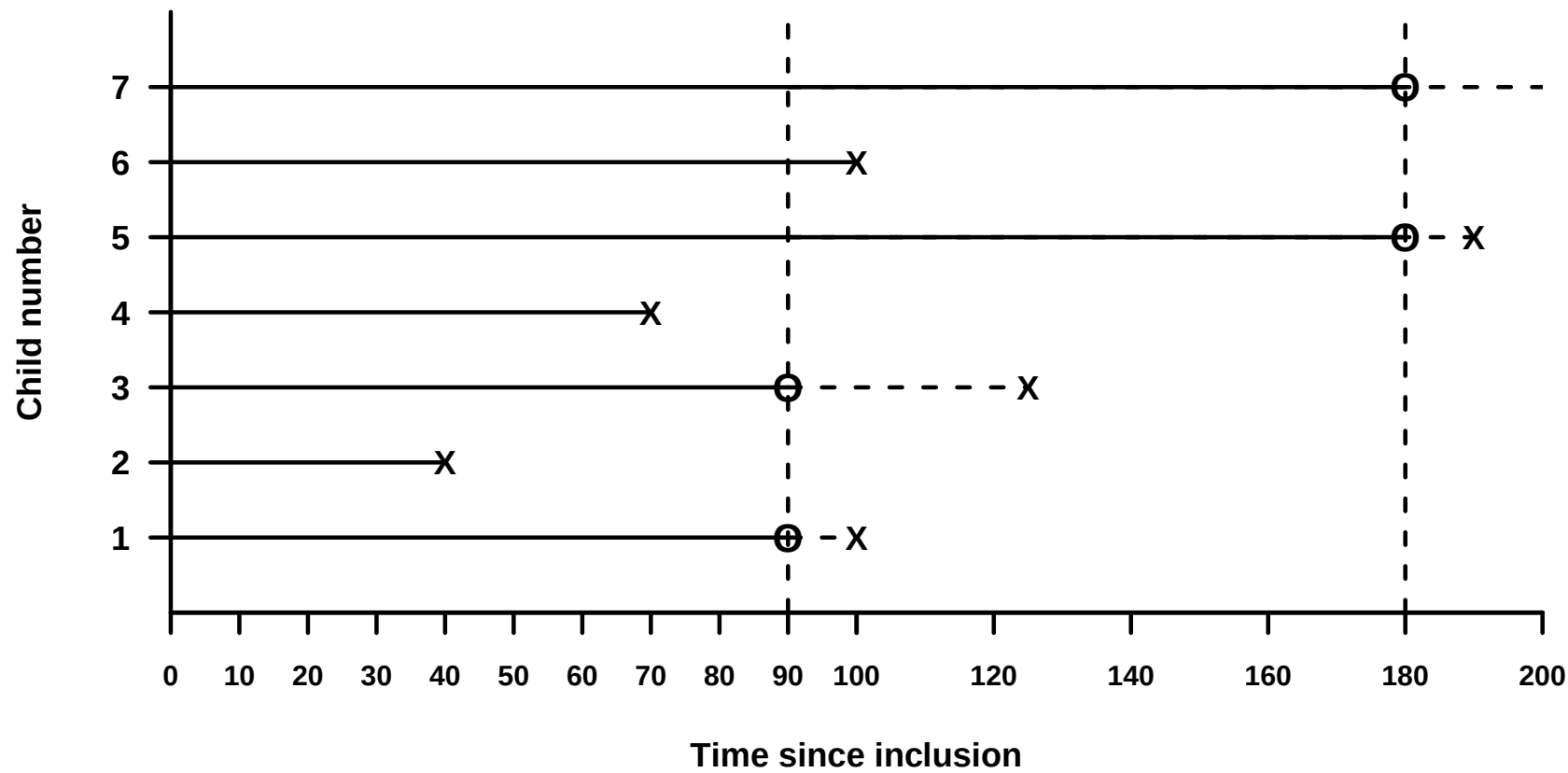
Nepal Study:

Time from first pneumonia admission to next

- **Outcome:**
Time from first admission with pneumonia until *next admission*
- Age range: 2 months to 3 years
- **Dates:** November 2003 to December 2007
- **Main exposure:** Zinc versus placebo
- A total of $719 + 350 = 1069$ children
- 719 children have two admissions
- 350 children have only the first admission (during follow-up period):
Censoring! To begin with, we will ignore these...

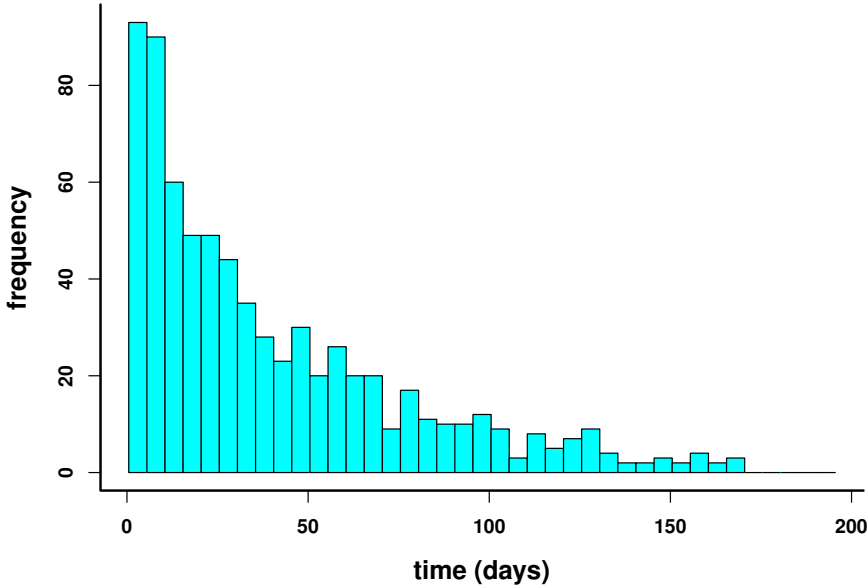
(Data from Tor Strand, Maria Mathisen, and others, Centre for International Health, UiB)

Timeline for events: Right censoring

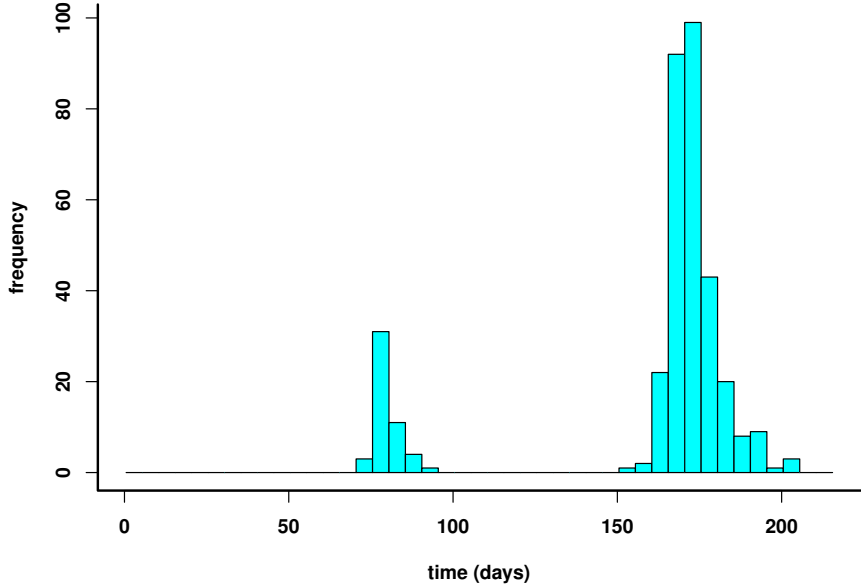


X = event O = censoring

Distribution of 719 new infections



Follow-up time of the 350 censored



Data, selected variables

	id	date	age	sex	treat.orig	time	event	time.14	treat
1	1	2004-01-25	10	1	1	69	1	56	1
2	2	2004-03-22	13	2	1	123	1	110	0
3	6	2003-11-30	7	1	1	190	0	177	1
4	8	2003-12-02	5	2	0	185	0	172	1
5	9	2003-12-03	4	1	0	93	0	80	1
6	13	2003-12-24	6	2	0	183	0	170	1
:		:			:				:
:		:			:				:

date: date of inclusion

age: age (at inclusion) in months

sex: boys = 1, girls = 2

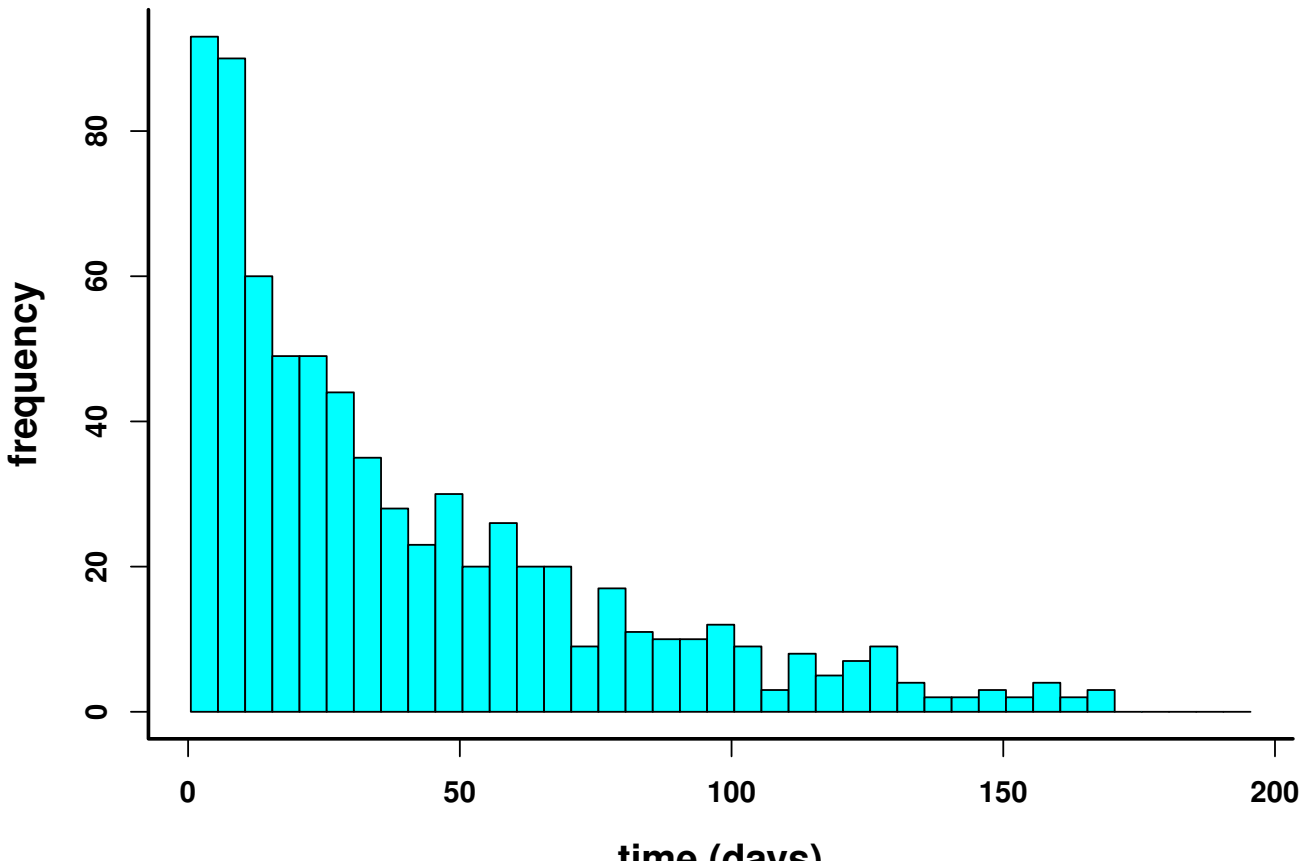
time: time since inclusion to event or censoring

event: new episode = 1, censored = 0

time.14 = time - 13: Starts counting after 14 days

treat: zinc = 1, placebo = 0

Days to next occurrence of pneumonia: histogram



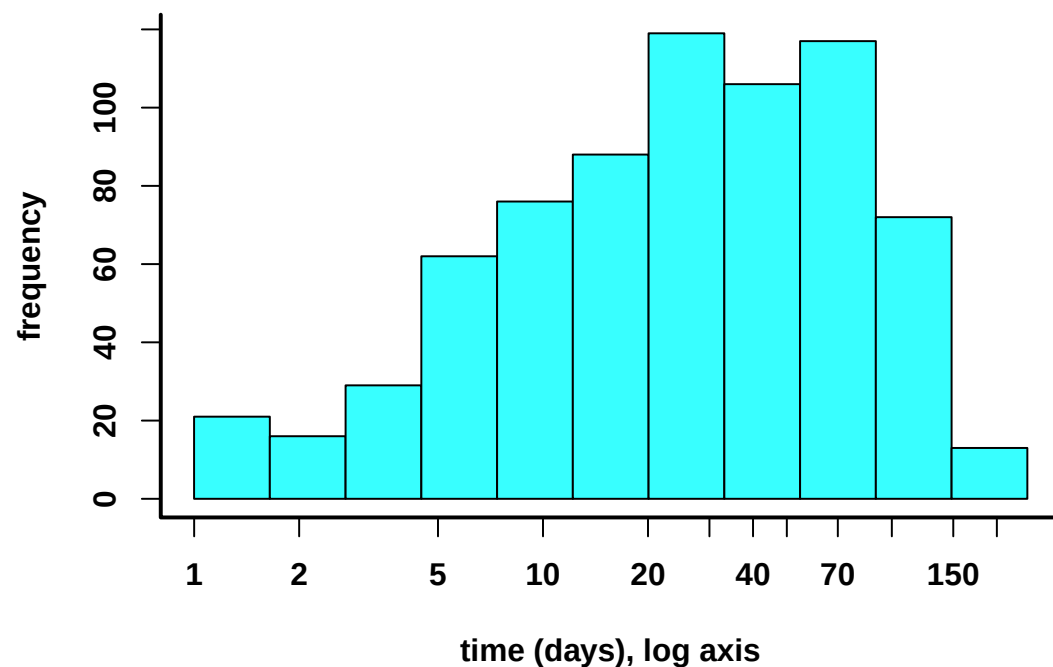
(Categories of 5 days, with 14 days “latency”)

Typical for a time-to-event study

- Always positive values
- NOT a normal distribution, i.e. no t-test nor ordinary regression
- Often skewed distribution, with a tail to the right
- What to do with the 350 that never got a new infection? (censoring)

Trick?? Try a log transform

Histogram with log scale:



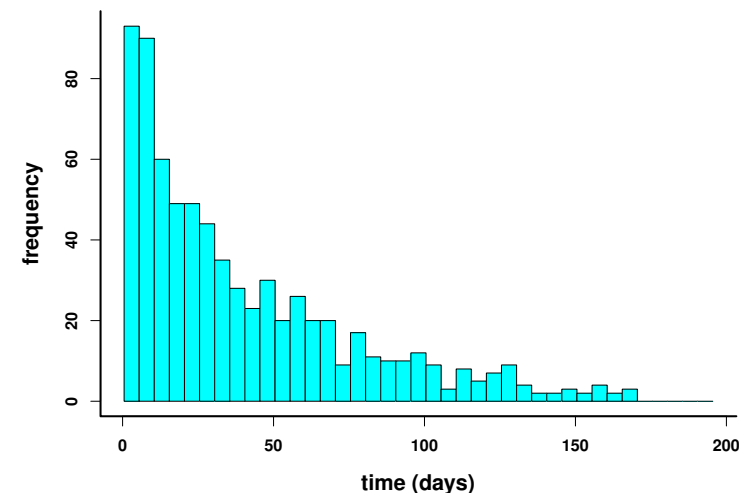
Did it help??

- Much closer to a normal distribution.
Can use ordinary t-test to compare zinc and placebo
- Problem...: we have still not dealt with the 350 censored
- Ordinary regression/t-test do NOT deal with censoring
- ... we need something better...

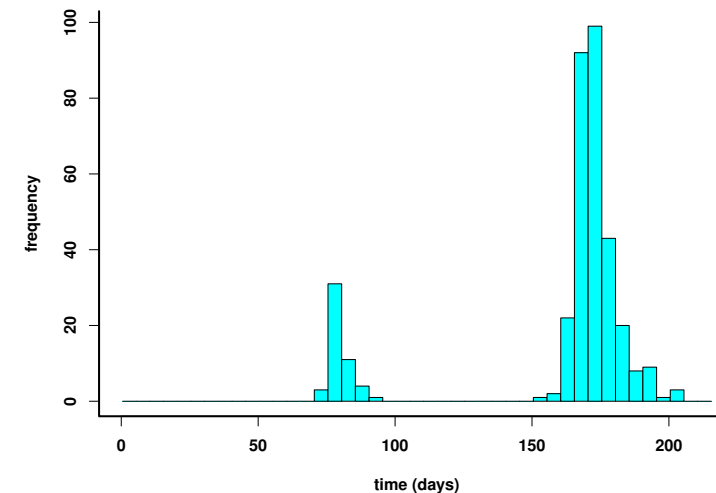
Trick?? Try dichotomization

- New infection before 50 days:
 - 1: New infection before 50 days
 - 0: Infection after 50 days, or censored
- This would actually work (using logistic regression)!
- ... but somewhat crude (looks only at a single time point)

Distribution of 719 new infections



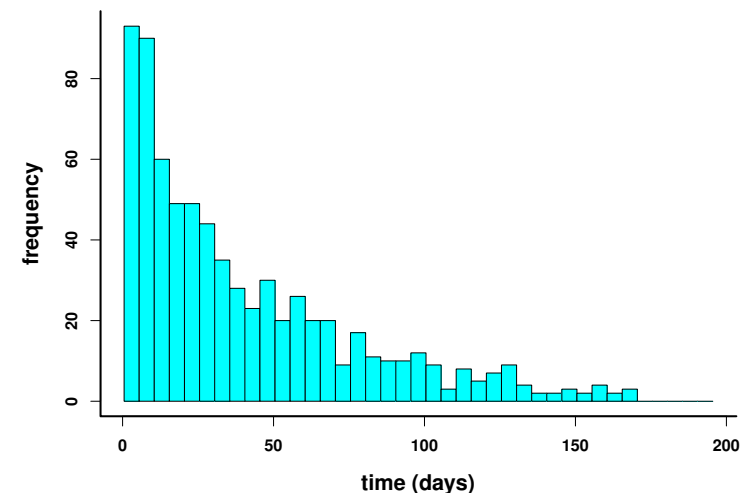
Follow-up time of the 350 censored



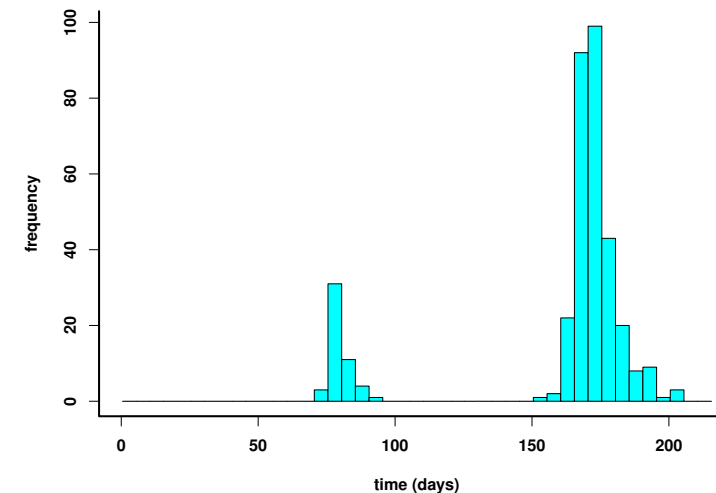
Trick?? Try dichotomization

- New infection before 100 days:
 - 1: New infection before 100 days
 - 0: Infection after 100 days, or censored??????
- This would NOT work!
- Censoring happens before cutoff value (100 days)!

Distribution of 719 new infections



Follow-up time of the 350 censored

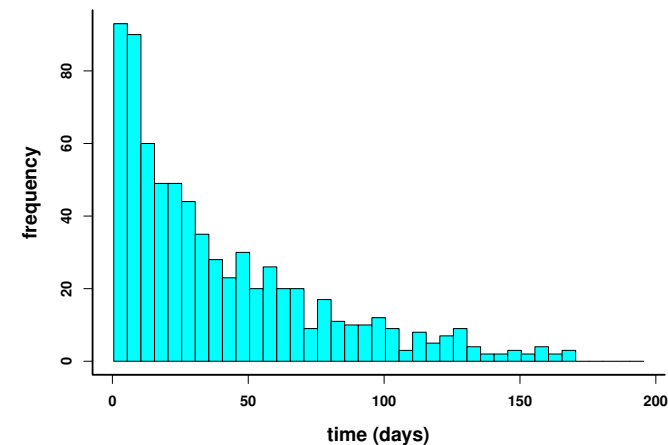


The “workhorse” of Survival Analysis

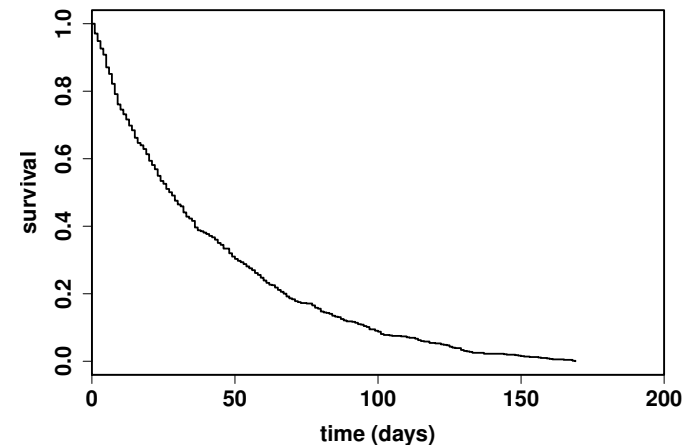
Survival curve

- T is time to event. Survival curve: $S(t) = P(T > t)$.
- I.e., the probability of “surviving” more than t days.

Distribution



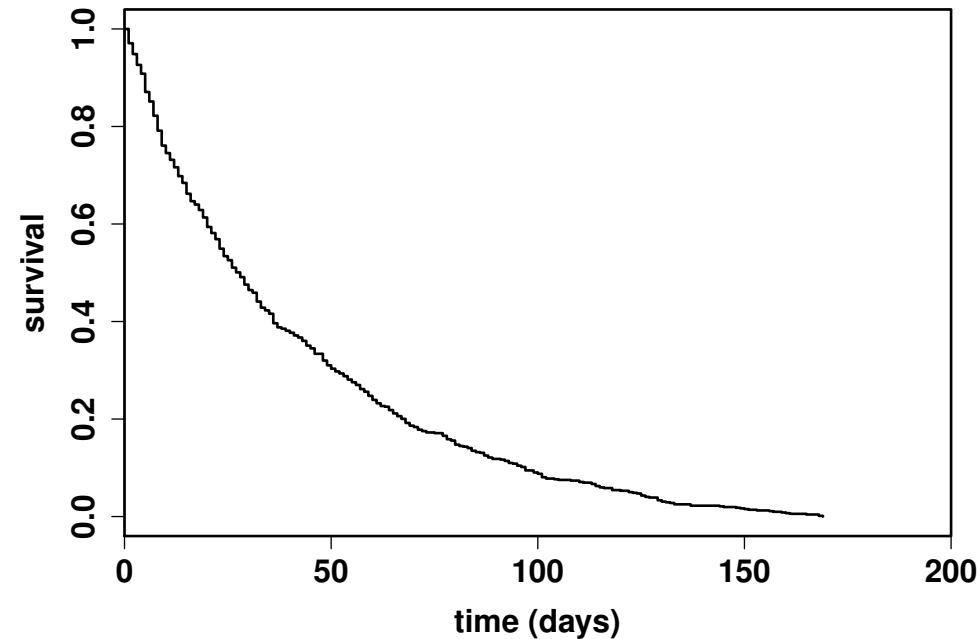
Survival curve



(Note: they look alike, but that’s a coincidence)

Kaplan-Meier estimate of the survival curve

Ignoring the censored, for the time being...



Shows the proportion that still have *not* had a new infection

E.g.: After 50 days there are about 30% that still haven't had a new infection

Kaplan-Meier: Computation without censoring

Compute $S(t)$ *without* censoring:

$$S(t) = \frac{\text{the number of children without new infection at time } t}{\text{total number included}}$$

Survival day-to-day (from software):

time	n.risk	n.event	survival
1	719	21	0.97079
2	698	16	0.94854
3	682	16	0.92629
4	666	13	0.90821
5	653	27	0.87065

Survival first day:

$$\frac{719 - 21}{719} = 0.97079$$

Survival first two days:

$$\text{survival first day} \times \text{survival second day} = 0.97079 \times \frac{698 - 16}{698} = 0.94854$$

etc., same result as the simple rule (above).

Kaplan-Meier: Computation with censoring

Compute $S(t)$ *with* censoring:

Just as an illustration: Assume 100 children were censored at day 3:

Survival day-to-day:

time	n.risk	n.event	n.censored	survival	
1	719	21	0	0.97079	
2	698	16	0	0.94854	
3	682	16	100	0.92629	new
4	566	13	0	0.90821	--> 0.90501
5	553	27	0	0.87065	--> 0.86083

Survival first 4 days:

$$\text{survival first 3 days} \times \frac{566 - 13}{566} = 0.92629 \times 0.9770318 = 0.90501$$

Survival first 5 days:

$$0.9050148 \times 0.9511754 = 0.86083$$

Effect of censoring accumulates!

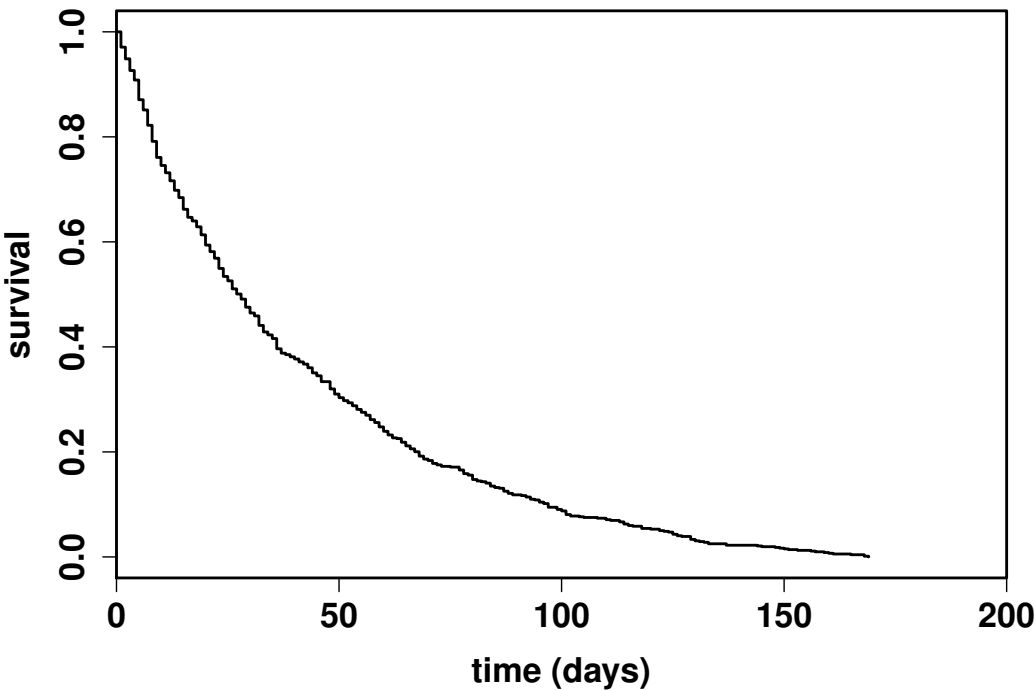
Kaplan-Meier: Numbers “at risk”

- The important issue is:

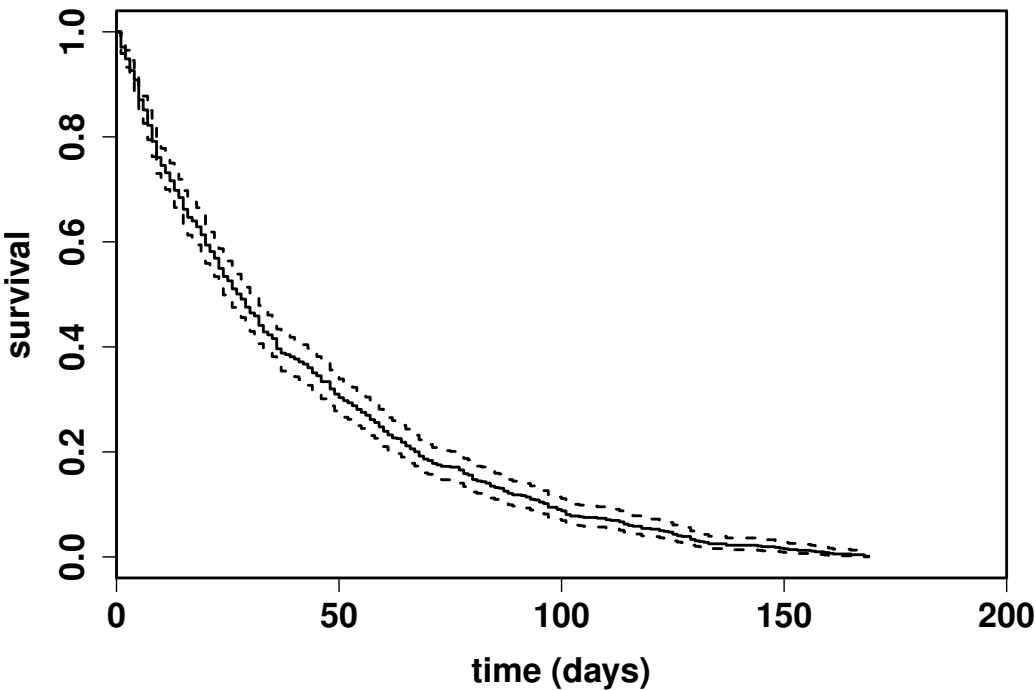
THE NUMBER OF CHILDREN
“AT RISK”
FOR A NEW INFECTION AT A GIVEN DAY

- That is, the number of children who – at a given day – are
WITHOUT A NEW INFECTION AND ALSO NOT CENSORED
- In other words, the number of children who still *can* experience an event
- Kaplan-Meier uses children *as long as they are at risk*,
then removes them from the computation

Kaplan-Meier estimate of the survival curve



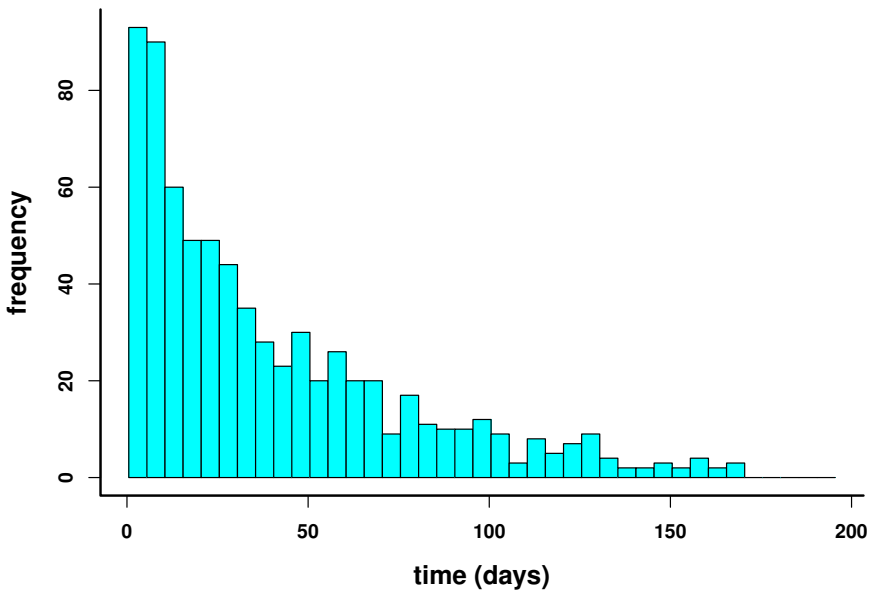
Kaplan-Meier estimate of the survival curve



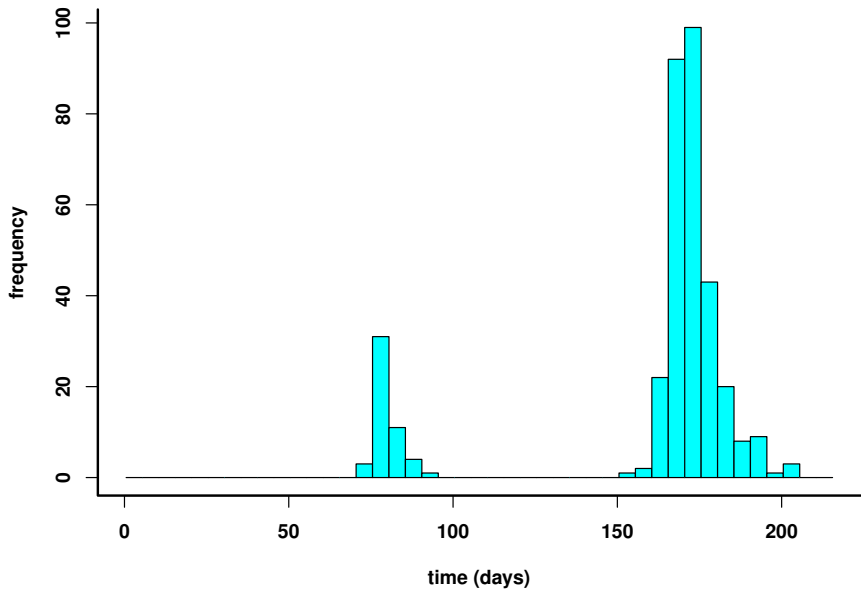
(With 95% “pointwise” confidence intervals)

Now, include the 350 CENSORED children

Distribution of new infections

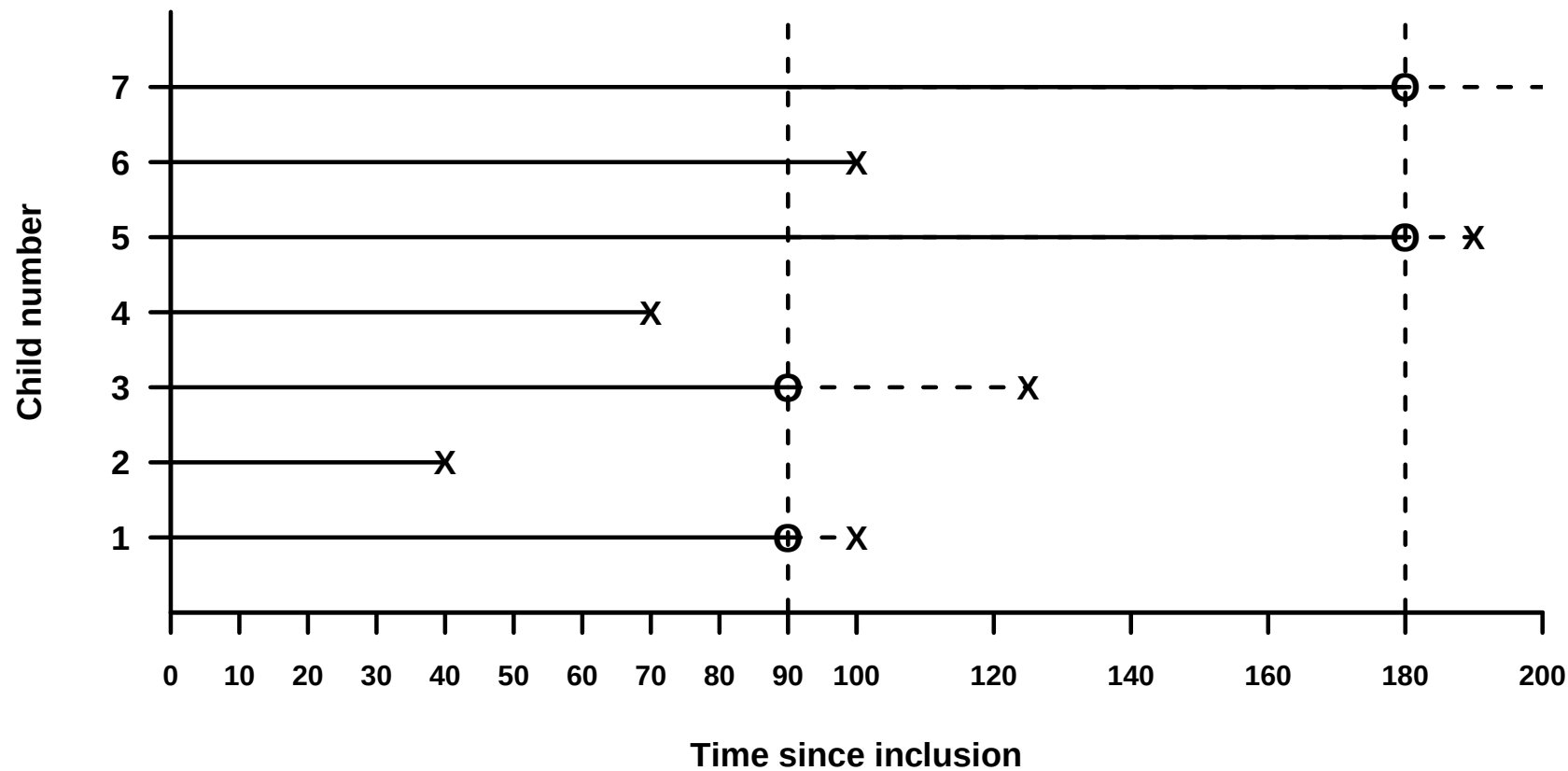


Follow-up time of the 350 censored



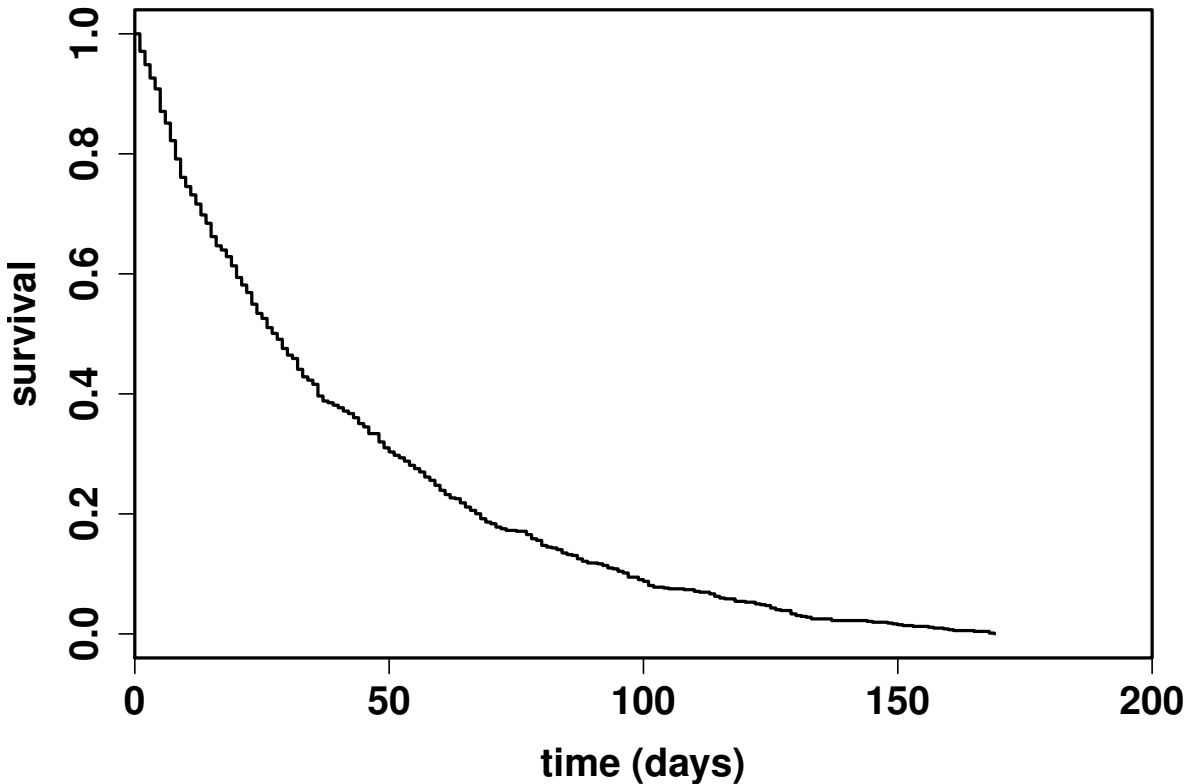
- Censoring has been ignored so far, but NOW we can deal with it....

Timeline for events: Right censoring



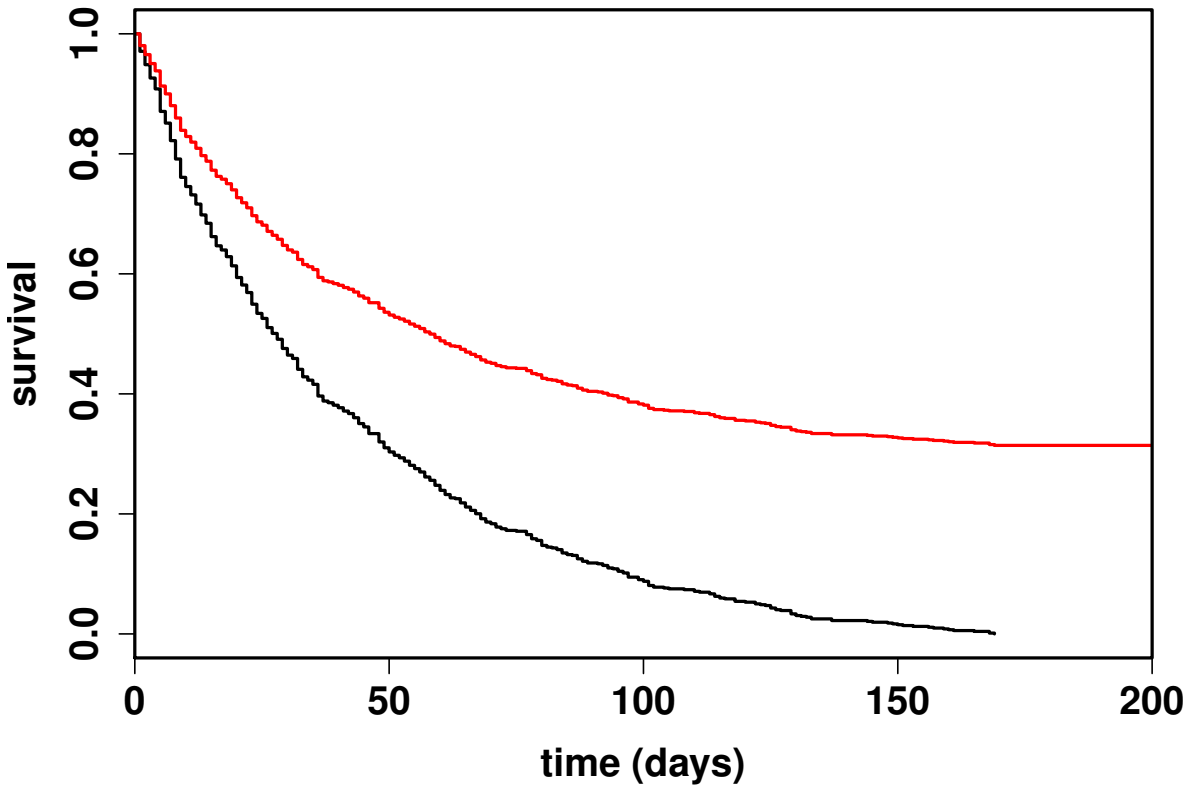
X = event O = censoring

Kaplan-Meier, *without* the censored



(Ignores censoring)

Kaplan-Meier, *with* the censored!

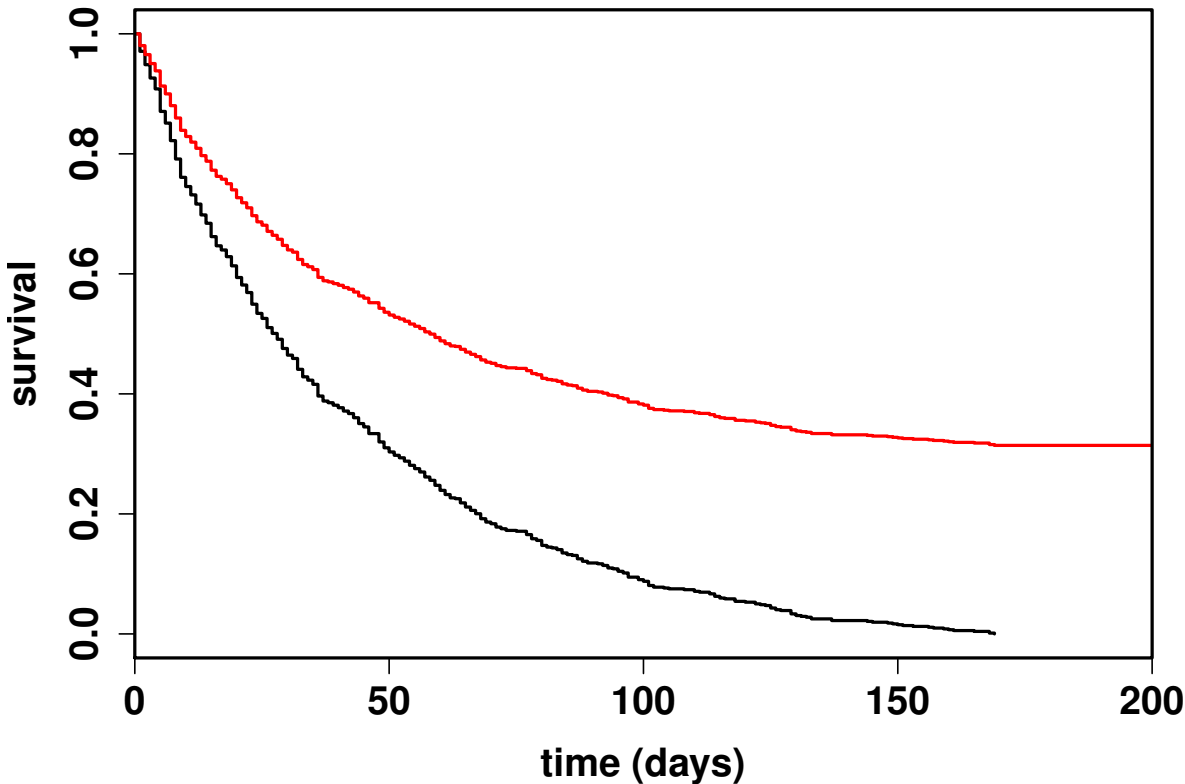


Red line: Kaplan-Meier with censoring

Two BAD approaches:

1. Remove all censored records before analysis.
Under-estimates the survival curve (as we have just seen)
2. Treat censored records as “real” endpoints, i.e. new admissions.
Under-estimates the survival curve

Kaplan-Meier, *with* the censored!



Red line: Kaplan-Meier with censoring

Mean versus Median: Who wins?

Compute:

- Mean time to next infection
- Median time to next infection

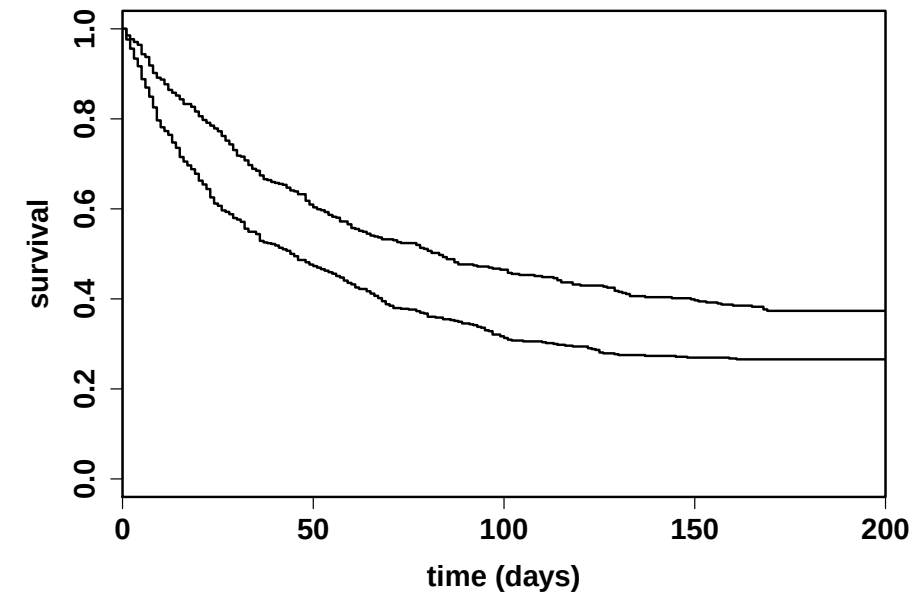
Results without censoring (**Wrong!**)

n	events	median	0.95LCL	0.95UCL
719	719	28	24	31
		mean		
		40.1	37.3	43.1

Results with censoring (**Correct!**)

n	events	median	0.95LCL	0.95UCL
1069	719	58	51	66
		mean		
		???	???	???

Kaplan-Meier: Compare groups



- Compare effect of zinc treatment with placebo
- Top curve is treatment, bottom curve is placebo

Data, selected variables

	id	date	age	sex	treat.orig	time	event	time.14	treat
1	1	2004-01-25	10	1	1	69	1	56	1
2	2	2004-03-22	13	2	1	123	1	110	0
3	6	2003-11-30	7	1	1	190	0	177	1
4	8	2003-12-02	5	2	0	185	0	172	1
5	9	2003-12-03	4	1	0	93	0	80	1
6	13	2003-12-24	6	2	0	183	0	170	1
:		:			:				:
:		:			:				:

date: date of inclusion

age: age (at inclusion) in months

sex: boys = 1, girls = 2

time: time since inclusion to event or censoring

event: new episode = 1, censored = 0

time.14 = time - 13: Starts counting after 14 days

treat: zinc = 1, placebo = 0

Testing the difference

(We don't really need to test.... difference is obvious here!)

(But still... a p-value might be useful)

Log-rank test

- Preferred when hazards are (roughly) proportional

```
> survdiff(Surv(time.14, event) ~ treat, data = .data0, rho = 0)
```

	N	Observed	Expected	$(O-E)^2/E$	$(O-E)^2/V$
treat=0	590	427	361	12.2	24.8
treat=1	479	292	358	12.2	24.8

Chisq= 24.8 on 1 degrees of freedom, p= 6.53e-07

```
time.14 = time - 13
event = 0 (censoring) or 1 (new episode)
treat = 0 (placebo) or 1 (zinc)
```

Testing the difference

Wilcoxon-type test

- Preferred when hazards are non-proportional

```
> survdiff(Surv(time.14, event) ~ treat, data = .data0, rho = 1)
```

	N	Observed	Expected	(O-E) ² /E	(O-E) ² /V
treat=0	590	296	245	10.7	30.6
treat=1	479	184	235	11.1	30.6

Chisq= 30.6 on 1 degrees of freedom, p= 3.11e-08

(NOTE: The observed number of events reported here are smaller than those actually observed because `survdiff` reports numbers weighted according to the numbers at risk)

Cox regression

The group difference should be measured, not only tested!

Cox regression:

- Can *test* difference (more or less like log-rank)
- Can *measure* difference (as “Hazard Ratio”, HR)
- Can produce confidence intervals for difference
- Can adjust for other variables/confounders (multiple regression)

... one of the most frequently used methods in medical statistics....

But remember that it assumes:

- “Independent” censoring
- Proportional hazards

Hazard rate

Hazard rate α

- “Instantaneous” probability of new event
- Same as incidence rate, but used in different settings:
 - **Incidence rate:**
 - Estimated directly from data
 - Over a time interval of some length
 - Often with an “open population”
 - **Hazard rate:**
 - A mathematical concept, estimated from the model
 - Instantaneous, i.e. over a “very short” time interval
 - Often with a “closed population” or at the individual level
- In our data:
 - Events first 5 days: $21 + 16 + 16 + 13 + 27 = 93$
 - Total (to begin with): 1069
 - Very roughly, $\alpha(0) = \frac{93}{1069 \times 5} = 0.017$

Thus, the hazard rate is about 1.7% new events per day to begin with

Four ways to describe survival

Survival time $T \geq 0$ (disregard censoring for the moment...)

Cumulative distribution function:

$$F(t) = P(T \leq t)$$

Survival function:

$$S(t) = P(T > t) = 1 - F(t)$$

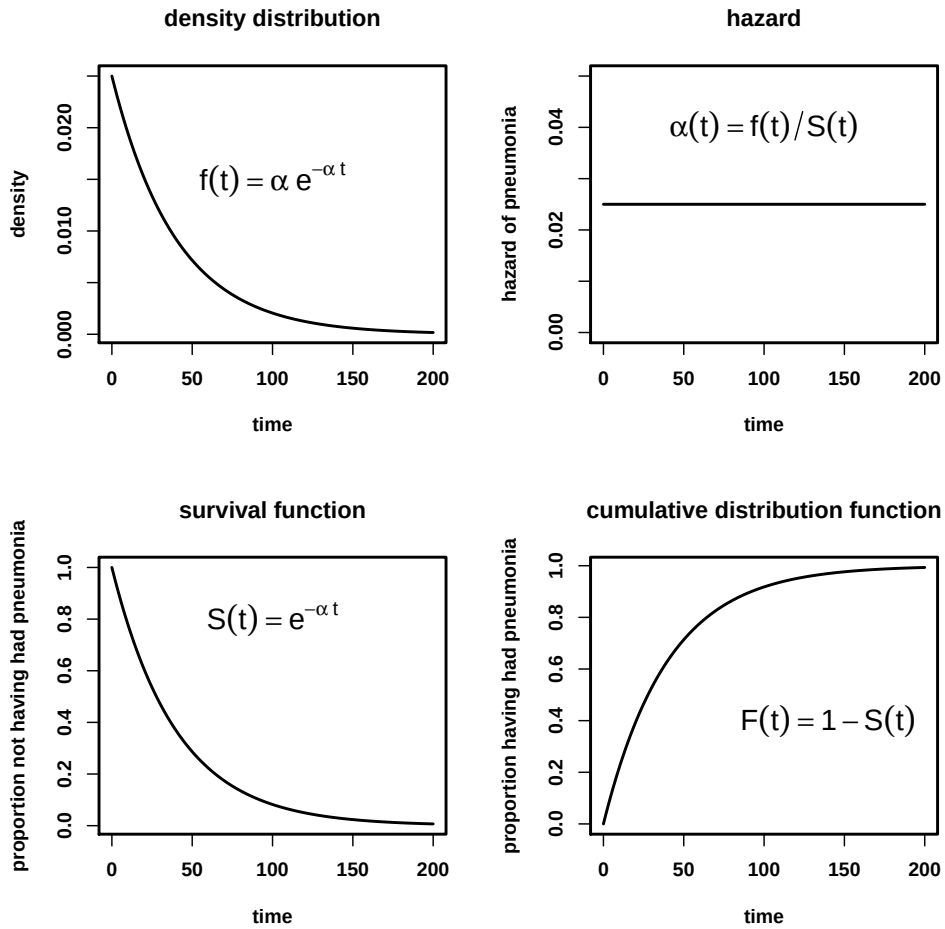
Distribution function (aka density):

$$f(t) = F'(t)$$

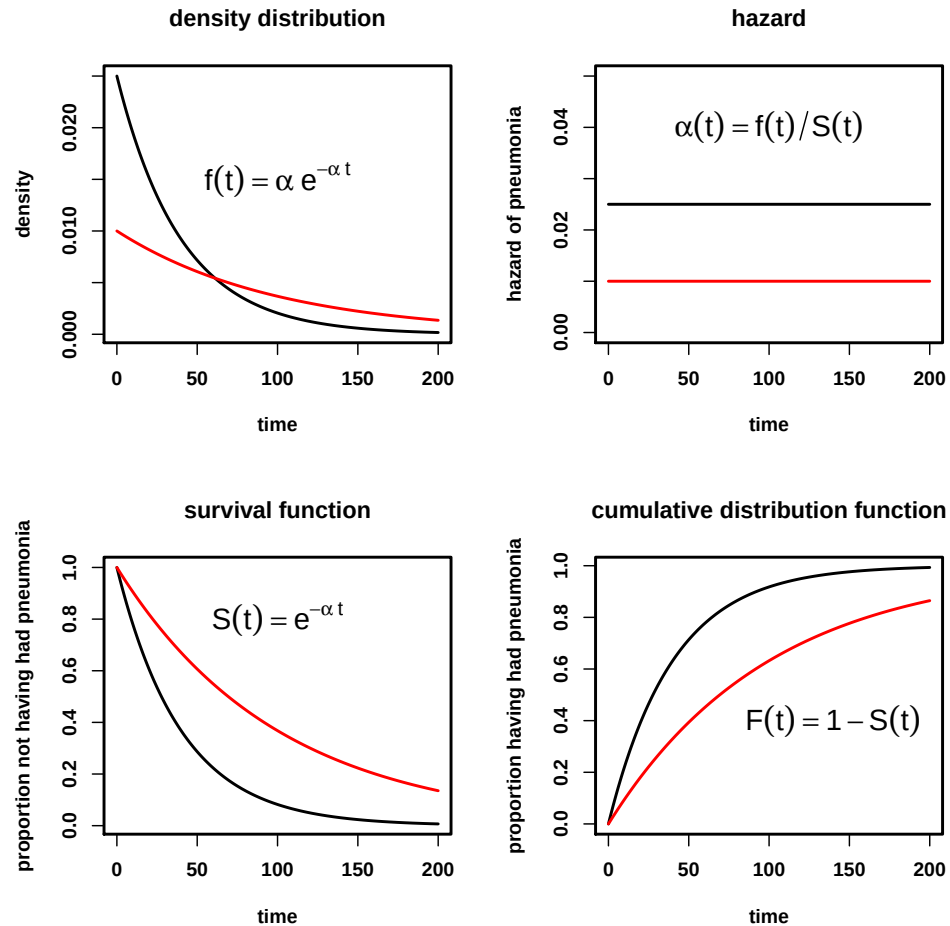
Hazard:

$$\alpha(t) = \lim_{\Delta t \rightarrow 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t \mid T \geq t)$$

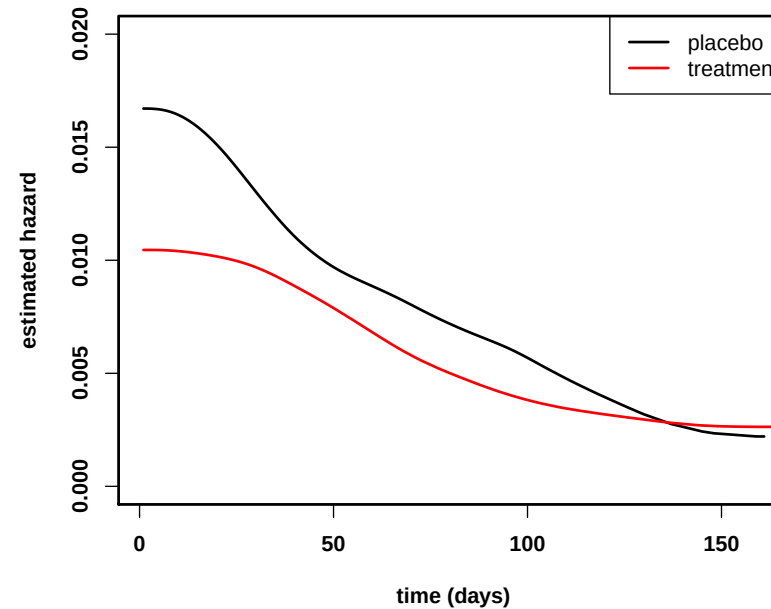
Four ways to describe survival



Four ways to describe survival



What do the hazards actually look like?



Note: Hazards are notoriously difficult to estimate!
... but almost never needed

Cox (proportional hazards) regression

Hazard:

$$\alpha(t) = \alpha_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \dots)$$

- $\alpha_0(t)$ is the *baseline hazard*
- $x_1, x_2, \dots$ are the covariates
- $\beta_1, \beta_2, \dots$ are the corresponding parameters

Covariates:

$x_1, x_2, \dots$ are covariates as in any other regression,
continuous or categorical (using dummy variables)

Baseline hazard:

β_0 not needed, α_0 takes its role

$\alpha_0(t)$ is thus the hazard (at time t) when all $x_1 = x_2 = \dots = 0$

Cox (proportional hazards) regression

For example:

$$\alpha(t) = \alpha_0(t) \exp(\beta_1 x_1)$$

$x_1 = 0$ (placebo) and $x_1 = 1$ (treatment)

$$\alpha_{\text{placebo}(t)} = \alpha_0(t)$$

$$\alpha_{\text{treatment}(t)} = \alpha_0(t) \exp(\beta_1)$$

Hazard (rate) ratio

$$HRR = \frac{\alpha_{\text{treatment}(t)}}{\alpha_{\text{placebo}(t)}} = \frac{\alpha_0(t) \exp(\beta_1)}{\alpha_0}(t) = \exp(\beta_1)$$

- $\alpha_0(t)$ is the hazard in the placebo group
- β_1 is the actual parameter estimate for treat (from software)

Cox regression (our data)

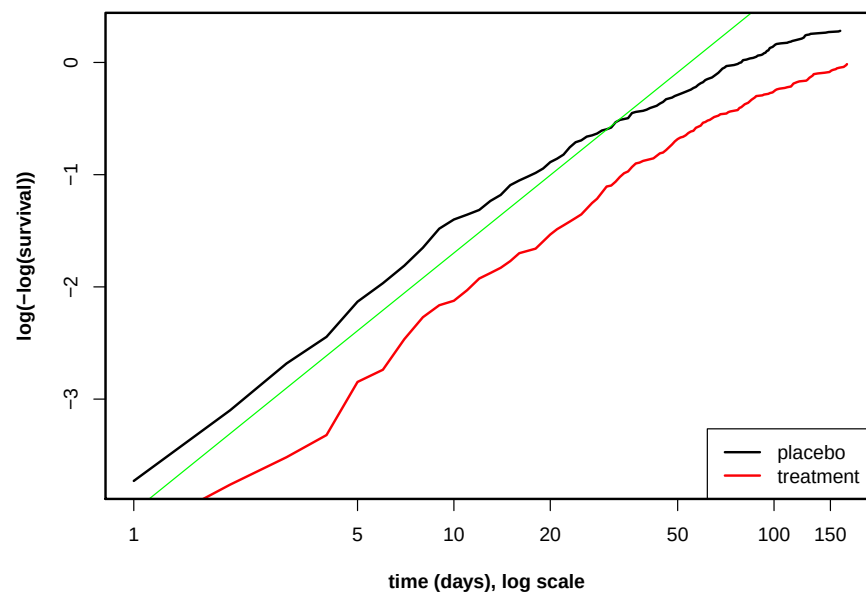
```
> coxph(Surv(time.14, event) ~ treat, data = .data0)
```

	coef	exp(coef)	se(coef)	z	p
treat	-0.377	0.686	0.076	-4.95	7.3e-07

	exp(coef)	exp(-coef)	lower .95	upper .95
treat	0.686	1.46	0.591	0.796

Hazard is reduced to about 69% relative to no treatment

Are hazards proportional? “Log-minus-log” plot



- **x-axis:** $\log(\text{time})$
- **y-axis:** $\log(-\log S(t))$

Rationale:

- Constant vertical distance $\rightarrow$ proportional hazards
- Lines with slope 1 $\rightarrow$ constant hazard (compare with green line)

-> Separate presentation, Cox

Cox regression, multivariable

```
> coxph(Surv(time.14, event) ~ treat + sex, data = .data0)
```

	coef	exp(coef)	se(coef)	z	p
treat	-0.375	0.687	0.0761	-4.93	8.1e-07
sex	-0.132	0.876	0.0753	-1.75	8.0e-02

	exp(coef)	exp(-coef)	lower .95	upper .95
treat	0.687	1.46	0.592	0.798
sex	0.876	1.14	0.756	1.016

- Adjustment for sex has little effect on treat
- sex is in itself not significant

(Note: In a randomized study you would not necessarily adjust for covariates)

Cox regression, stratified

```
> coxph(Surv(time.14, event) ~ treat + strata(sex), data = .data0)
```

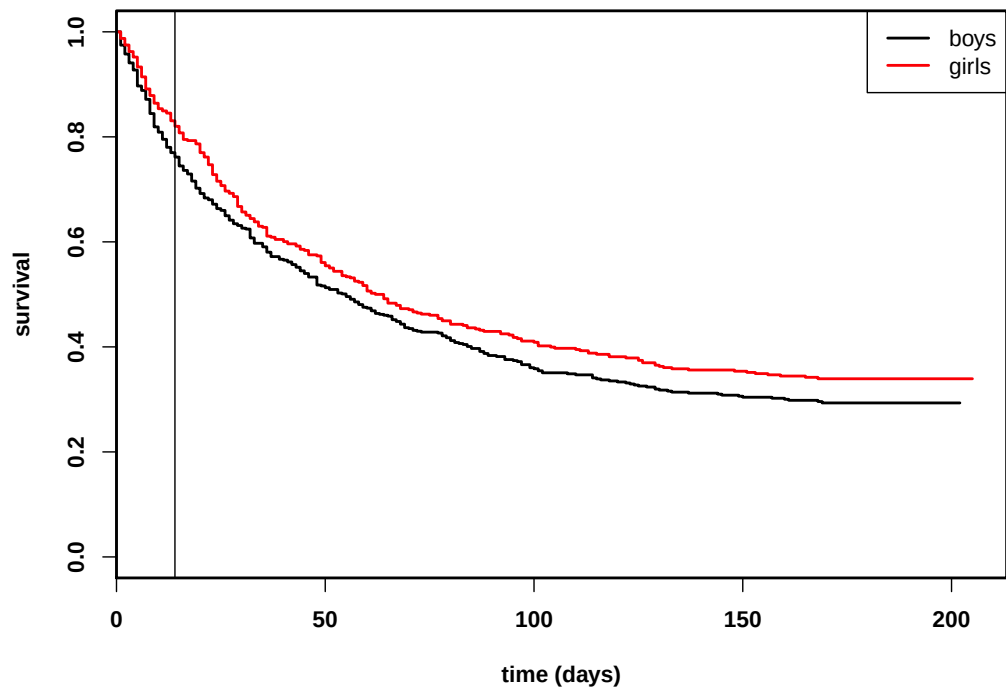
	coef	exp(coef)	se(coef)	z	p
treat	-0.374	0.688	0.0761	-4.92	8.7e-07

	exp(coef)	exp(-coef)	lower .95	upper .95
treat	0.688	1.45	0.592	0.798

- Treatment effect is assumed equal in both strata of sex
- But separate baselines are allowed for boys and girls

(Note: This is just as an illustration; not really necessary)

Survival for boys and girls



Somewhat peculiar. Does something happen the first two weeks?

Cox regression, interaction

Is the effect of treatment *different* for boys and girls?

Warning: When dealing with interactions, the exact coding of variables is very important.

- Dummy variables make interpretations much easier.
- treat is already a dummy.
- R can (could!) do this more simply than Stata, but the following works for both.

Create dummy:

$$\text{sex01} = \text{sex} - 1$$

that is, sex01 is a dummy variable for girl, with boys = 0, girls = 1

Create interaction term:

$$\text{treat.sex} = \text{treat} \times \text{sex01}$$

-> Separate presentation, Coding of interactions

Cox regression, standard interaction

```
> coxph(Surv(time.14, event) ~ treat + sex01 + treat.sex,  
data = .data0)
```

	coef	exp(coef)	se(coef)	z	p
treat	-0.507	0.603	0.103	-4.93	8.1e-07
sex01	-0.255	0.775	0.099	-2.57	1.0e-02
treat.sex	0.297	1.346	0.153	1.94	5.3e-02

	exp(coef)	exp(-coef)	lower .95	upper .95
treat	0.603	1.660	0.493	0.737
sex01	0.775	1.290	0.639	0.941
treat.sex	1.346	0.743	0.997	1.818

(Interaction is borderline significant, should probably just be dumped)

Interpretation:

Treatment HRR for girls = 1.346 × treatment HRR for boys

Cox regression, interaction, alternative coding

Create two interaction terms (and leave out `treat` from equation):

$$\text{treat.sex.0} = \text{treat} \times (1 - \text{sex01}) \quad (\text{Effect among boys})$$
$$\text{treat.sex.1} = \text{treat} \times \text{sex01} \quad (\text{Effect among girls})$$

```
> coxph(Surv(time.14, event) ~ sex01 + treat.sex.0 + treat.sex.1,  
        data = .data0)
```

	coef	exp(coef)	se(coef)	z	p
sex01	-0.255	0.775	0.099	-2.57	1.0e-02
treat.sex.0	-0.507	0.603	0.103	-4.93	8.1e-07
treat.sex.1	-0.209	0.811	0.114	-1.84	6.6e-02

	exp(coef)	exp(-coef)	lower .95	upper .95
sex01	0.775	1.29	0.639	0.941
treat.sex.0	0.603	1.66	0.493	0.737
treat.sex.1	0.811	1.23	0.649	1.014

- Risk of new pneumonia in boys is reduced to 60.3% with treatment
- Risk of new pneumonia in girls is reduced to 81.1% with treatment