## Incomplete observations

Censoring:
we don't see the exact value of a data point, only the information that it lies in some interval.

Right censoring:
for some individuals $i$, we observe only that the lifetime is in $(x_i, \infty)$. Very common, e.g. in a clinical trial

- trial has fixed length; event time for some patients is after the end of the trial;
- a patient leaves a trial early for some reason, "lost to follow-up";
- patient dies (where time to death is not the object under study).

## Left censoring.

Less common in practice. We observe that the lifetime is in $(0, x_i)$.
Examples:

- A childhood learning study: when do children acquire a particular skill? It may already be acquired before the time a particular child is enrolled in the study; the observation is left-censored.
- Malaria incubation time. A traveller develops malaria 25 days after arrival in a malarial region. Incubation time is known to be in (0, 25 days).

Maybe a combination:

- a study of carcinogenesis in animals. Existence of a tumour (but not its age) can be detected only at death. When the animal dies (or is killed) either a left-censored or a right-censored observation of the time to tumour development is made.

Or interval censoring:

- Traveller spends 7 days in a malaria zone and then returns to UK. 12 days after returning, develops malaria symptoms. This gives an *interval censored* observation: incubation time is in (12 days, 19 days).

## Truncation

A “potential data point” may **never be seen at all**, depending on its value. For example, only observed if it falls in $(Y_L, Y_R)$, otherwise invisible.

## Left truncation:

$Y_R = \infty$. Very common, and generally not problematic.

- mortality study in a retirement home. Age of arrival and age of death are recording. Only those reaching a sufficient age to enter the home are observed. Of course, this means we cannot estimate the survival probability to high ages, but we can, for example, estimate the conditional probability of survival to age 85 given survival to age 80. We have encountered this already in many contexts without difficulty.

## Right truncation:

$Y_L = 0$. Less common. For example, a June 1986 study of patients with transfusion-acquired AIDS.

Patient $i$:

transfusion date $a_i$

onset of symptoms $b_i$

incubation time $t_i = b_i - a_i$

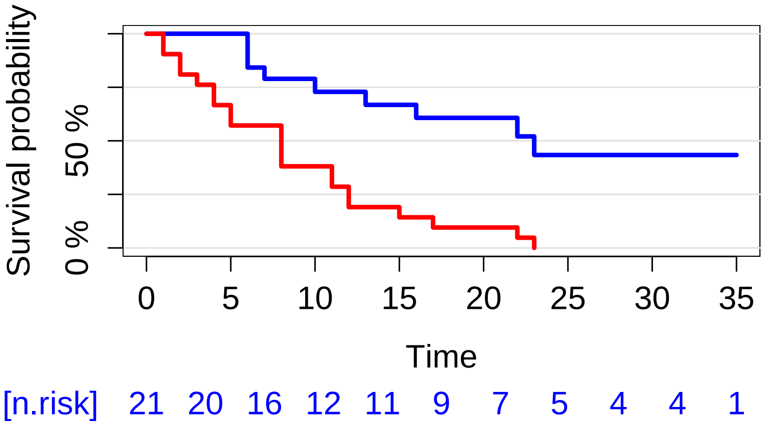
If $b_i >$ June 1986, i.e. if $t_i >$ June 1986 $-a_i$, then the patient does not appear in the study.
This is right truncation with $Y_R =$June 1986$-a_i$.

**N.B. different from censoring!** (where the patient is seen but has an imprecisely observed event time).

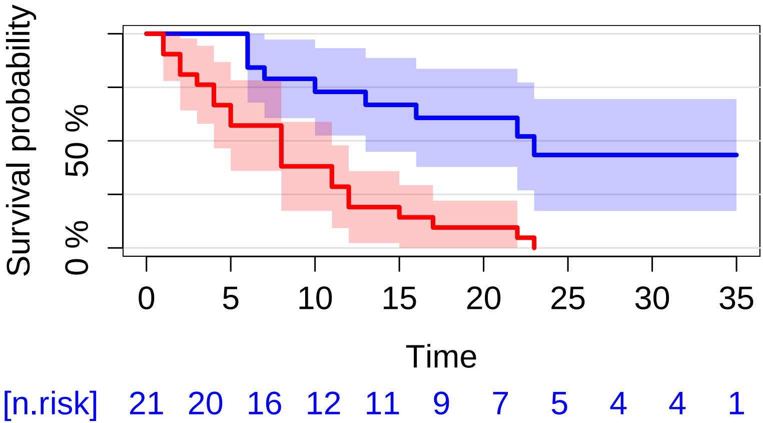
Length of remission periods (in weeks) of leukemia patients in a trial of 6-mercaptopurine.

| | |
|---|---|
| Treated group | 6, 6, 6, 6+, 7, 9+, 10, 10+, 11+, 13, 16, 17+, 19+, 20+, 22, 23, 25+, 32+, 32+, 34+, 35+ |
| Control group | 1,1,2,2,3,4,4,5,5,8,8,8,8,11,11,12,12,15,17,22,23 |

"+" indicates censoring ($\delta_i = 0$).



Kaplan-Meier estimate for the survival function for the control group (red) and treatment group (blue). Since there is no censoring in the control group, the red curve corresponds to the empirical distribution function.



Kaplan-Meier estimates showing pointwise confidence intervals. For the red curve, which corresponds to the empirical distribution function, these can be derived simply from the binomial likelihood. For the blue curve, we will see how to do this using Greenwood's formula.