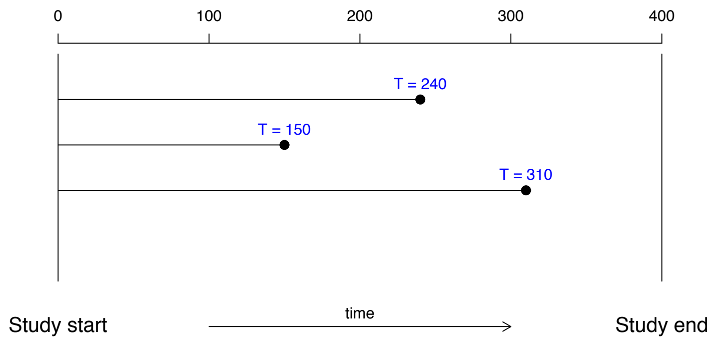
Survival Analysis

Survival analysis

Survival analysis: Study of durations between events

⟶ Outcome:
Time until an event occurs, i.e. *survival time* or *failure time*.



Examples: Age at death, age at first disease diagnosis, waiting time to pregnancy, duration between treatment and death, . . .

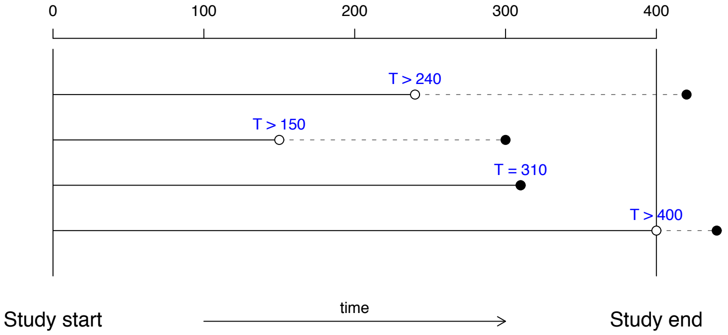
The censoring problem in survival analysis

⟶ Censoring:

Incomplete observations of the survival time.

⟶ Right censoring:

Some individuals may not be observed for the full time to failure, because of loss to follow-up, drop out, termination of the study, . . .



Basic goals of survival analysis

1. To estimate and interpret survival characteristics

   ⟶ Kaplan-Meier plots

2. To compare survival in different groups

   ⟶ Log-rank test

3. To assess the relationship of explanatory variables to survival

   ⟶ Cox regression model

Survival function

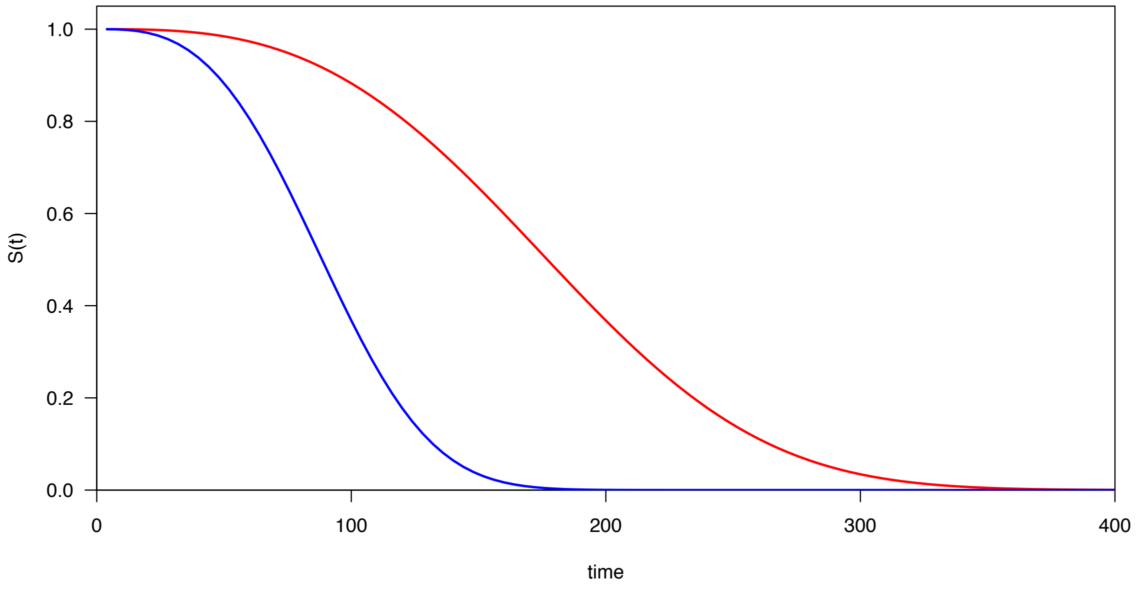
Survival function: S(t) = P(T $>$ t)

$\longrightarrow$ S(t) describes the probability of surviving to time t, or what fraction of subjects survive (on average) to time t.

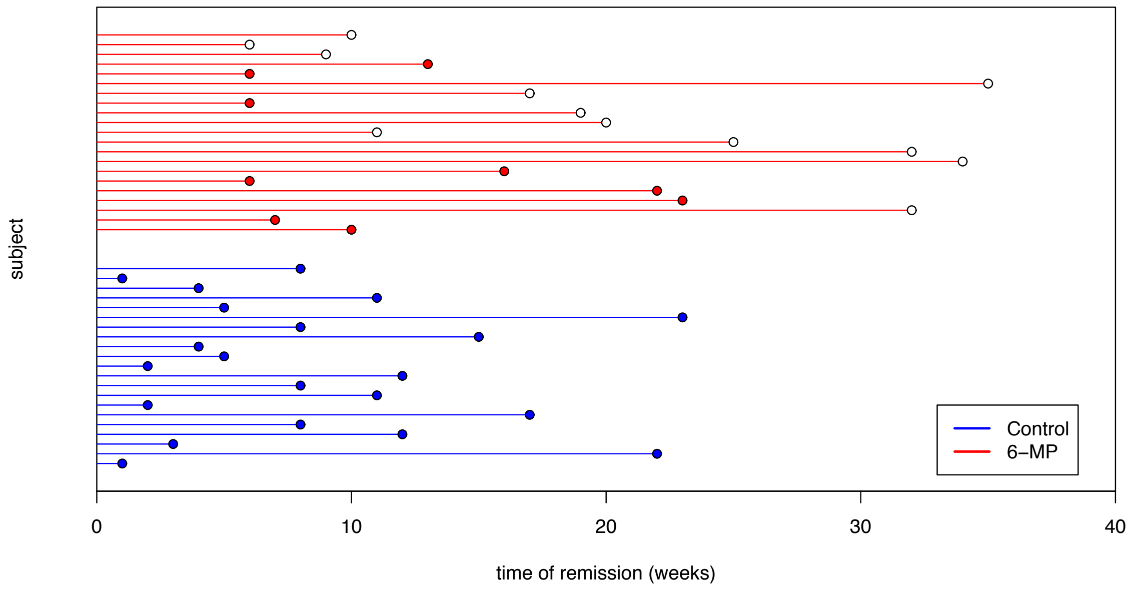
Properties:

- ○ S(t) is a smooth function in t.
- ○ S(0) = 1 and S(∞) = 0.
- ○ S(t) is a decreasing function in t.
- ○ Describes *cumulative* survival characteristics.

Survival functions



Example



Example

```
> library(survival)
> library(MASS)
> attach(gehan)

> str(gehan)
‘data.frame’:   42 obs. of  4 variables:
 $ pair : int  1 1 2 2 3 3 4 4 5 5 ...
 $ time : int  1 10 22 7 3 32 12 23 8 22 ...
 $ cens : int  1 1 1 1 1 0 1 1 1 1 ...
 $ treat: Factor w/ 2 levels "6-MP","control": 2 1 2 1 2 1 2 1

> Surv(time,cens)
 [1]  1  10  22   7   3  32+ 12  23   8  22  17
[12]  6   2  16  11  34+  8  32+ 12  25+  2  11+
[23]  5  20+  4  19+ 15   6   8  17+ 23  35+  5
[34]  6  11  13   4   9+  1   6+  8  10+
```

Kaplan-Meier estimate

The Kaplan-Meier or product-limit estimate $\hat{S}(t)$ is an estimate of S(t) from a finite sample.

Suppose that there are observations on n individuals and assume that there are k ($k \leq n$) distinct times $t_1, \ldots, t_k$ at which deaths occur. Let d_j be the number of deaths at time t_j. Define

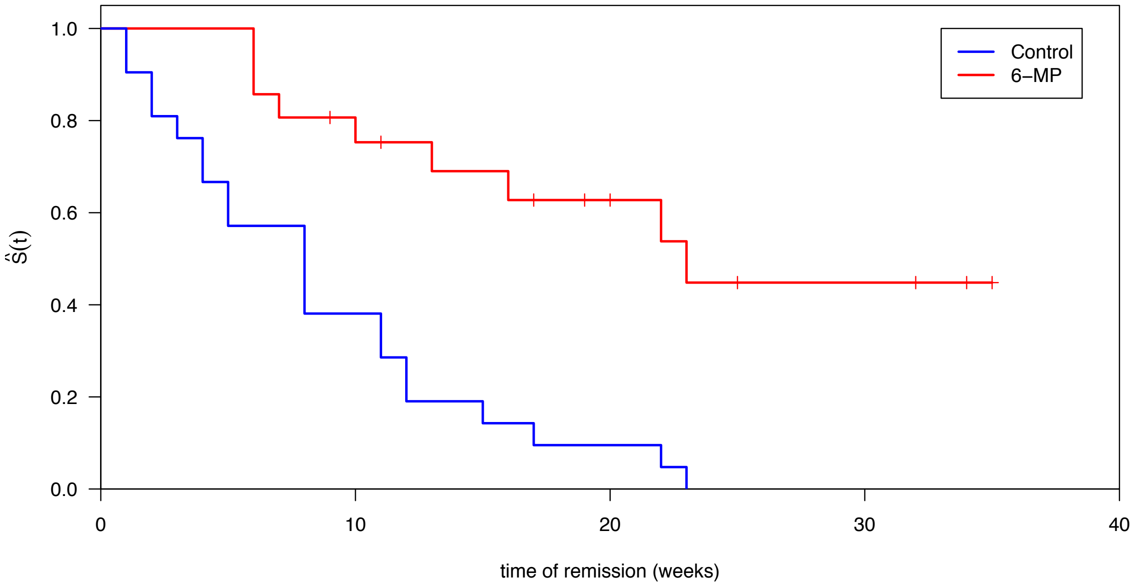
$$\hat{S}(t) = \prod_{j:\ t_j < t} \frac{n_j - d_j}{n_j},$$

where n_j is the number of individuals at risk (e.g., the individuals alive and uncensored) at time t_j.

$\longrightarrow$ If there are no censored observations, this reduces to

$$\hat{S}(t) = (\text{number of observations} \geq t) / n.$$

Example



```
> gehan.surv = survfit(Surv(time, cens) ~ treat, data = gehan)
> plot(gehan.surv)
```

Some facts about the Kaplan-Meier estimate

$\longrightarrow$ The Kaplan-Meier method is *non-parametric*. The survival curve is step-wise, not smooth. Any jumping point is a failure time point. The jump size is proportional to the number of deaths at a failure time point. Note that having a small sample means having big steps!

$\longrightarrow$ If the largest observed study time t_k corresponds to a death time, then the estimated Kaplan-Meier survival curve is 0 beyond t_k. If the largest observed study time is censored, then the survival curve is not 0 beyond t_k.

$\longrightarrow$ $\hat{S}(t)$ is a decreasing function in t with $\hat{S}(0) = 1$. Further $\hat{S}(t)$ converges to $S(t)$ as $n \to \infty$.

Comparison of two survival distributions

We test H_0: $S_1(t) = S_2(t)$ versus H_a: $S_1(t) \neq S_2(t)$

$\longrightarrow$ The main idea behind the two-sample log-rank test: if survival is unrelated to group effect, then at each time point, roughly the same proportion in each group will fail.

The test is based on χ^2-types of statistics:

$$Q = \sum_{i=1}^{D} (O_{1i} - E_{1i})$$

where the summation is over the pooled failure time points among the 2 groups. O_{1i} and E_{1i} are the observed number of death for group 1 at the i^{th} pooled failure time. The log-rank test statistic under H_0 is

$$\log RT = \frac{Q^2}{Var(Q)} \sim \chi_1^2$$

Example

```
> survdiff(Surv(time,cens)˜treat,data=gehan)

Call:
survdiff(formula = Surv(time, cens) ˜ treat, data = gehan)

               N Observed Expected (O-E)^2/E (O-E)^2/V
treat=6-MP    21        9     19.3      5.46      16.8
treat=control 21       21     10.7      9.77      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4.17e-05
```

Comparison of survival distributions

The log-rank test can be extended to $k > 2$ groups. Under H_0 the null distribution of the test statistic is

$$\text{logRT} \sim \chi^2_{k-1}$$

However, these test also have some shortcomings:

- ○ The tests have a bad performance when the two survival functions are overcrossing.
- ○ The test can only be used for comparing groups defined by single categorical covariates.
- ○ They are not very useful to quantify the differences.

Hazard function

The hazard function is defined as

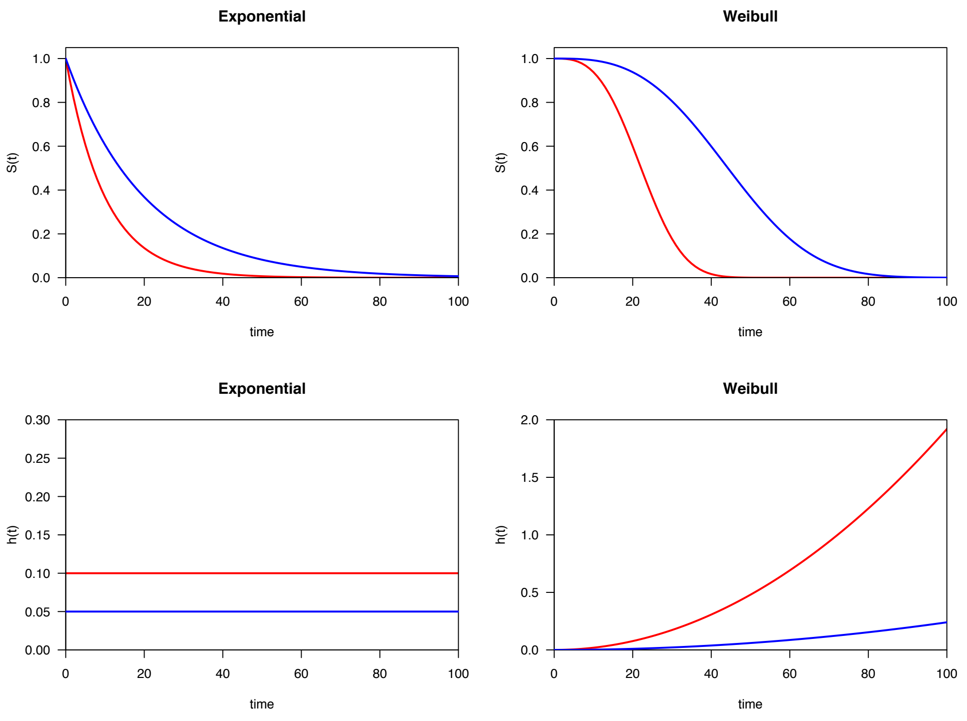
$$h(t) = -\frac{d}{dt}\log(S(t))$$

In other words, it is the slope of $-\log(S(t))$. You can think of it as the propensity for failure for an individual at each time point, e.g. the instantaneous risk of failure.

Properties:

- ○ Closely related to the incidence rate.
- ○ Not a probability!
- ○ May increase or decrease or both.
- ○ Describes *instantaneous* survival characteristics.

Hazard functions



Cox regression model

- Goal:

  To assess the relationship of explanatory variables (e.g. sex, age, treatment type, etc) to survival time.

- One idea (Sir David Cox):

  Use a proportional hazards regression model, defined as

  $$h(t|x) = h_0(t)e^{\beta x}$$

  Here, $h_0(t)$ is a baseline hazard function, and β is a regression coefficient.

Cox regression model

What does $h(t|x) = h_0(t)e^{\beta x}$ mean?

For example, assume we a treatment group ($x = 1$) and a control group ($x = 0$).

- In the control group, the hazard function is

  $$h(t|x = 0) \;=\; h_0(t)e^{\beta \times 0} \;=\; h_0(t)$$

- In the treatment group, the hazard function is

  $$h(t|x = 1) \;=\; h_0(t)e^{\beta \times 1} \;=\; h_0(t)e^{\beta}$$

- The relative risk for treatment versus control group is

  $$\text{RR} \;=\; \frac{h(t|x = 1)}{h(t|x = 0)} \;=\; e^{\beta}$$

Cox regression model

⟶ Interpretation of the parameters:

$\beta > 0$ RR > 1 and $h(t|x = 1) > h(t|x = 0)$

$\beta = 0$ RR $= 1$ and $h(t|x = 1) = h(t|x = 0)$

$\beta < 0$ RR < 1 and $h(t|x = 1) < h(t|x = 0)$

⟶ Hypothesis of interest:

$\mathrm{H}_0 : \beta = 0$ (no treatment effect)

$\mathrm{H}_a : \beta \neq 0$ (treatment influences survival)

Example

```
> gehan.cox = coxph(Surv(time, cens) ~ treat, gehan)
> summary(gehan.cox)

Call:
coxph(formula = Surv(time, cens) ~ treat, data = gehan)

  n= 42
             coef exp(coef) se(coef)    z       p
treatcontrol 1.57      4.82    0.412 3.81 0.00014

             exp(coef) exp(-coef) lower .95 upper .95
treatcontrol      4.82      0.208      2.15      10.8
```

Another example

```
> leuk.cox = coxph(Surv(time)~ ag + log(wbc), data = leuk)
> summary(leuk.cox)

Call:
coxph(formula = Surv(time) ~ ag + log(wbc), data = leuk)

  n= 33
            coef exp(coef) se(coef)     z      p
agpresent -1.069     0.343    0.429 -2.49 0.0130
log(wbc)   0.368     1.444    0.136  2.70 0.0069

          exp(coef) exp(-coef) lower .95 upper .95
agpresent     0.343      2.913     0.148     0.796
log(wbc)      1.444      0.692     1.106     1.886
```