Censored data and survival analysis

Filippo Biscarini

(Biostatistician, bioinformatician and quantitative geneticist) **CNR-IBBA**, Milan (Italy)

time-to-event data

Examples of types of events:

- relapse
- progression
- death
- in cows (livestock) also longevity

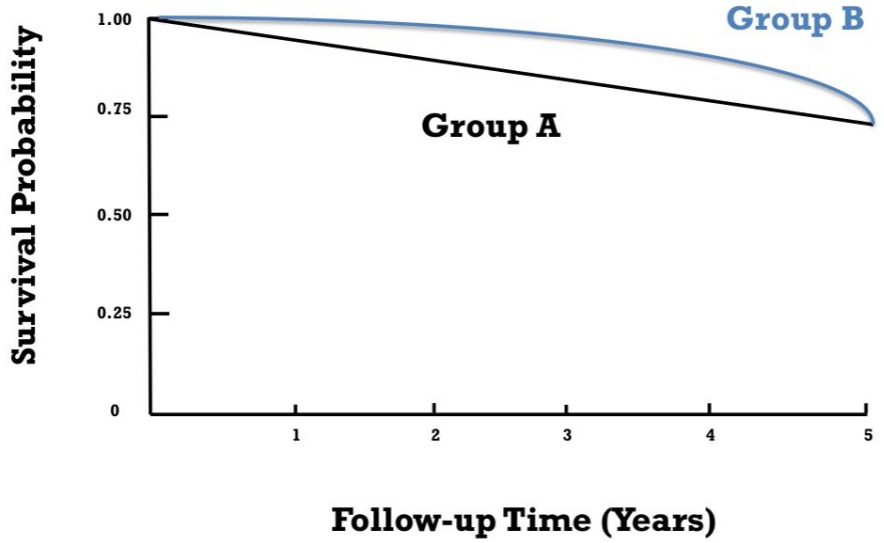
time-to-event data

Characteristics of time events:

- subjects enter at different times and have different duration of follow-up

- entire survival experience, not just the percentages who remain alive at the end of the study

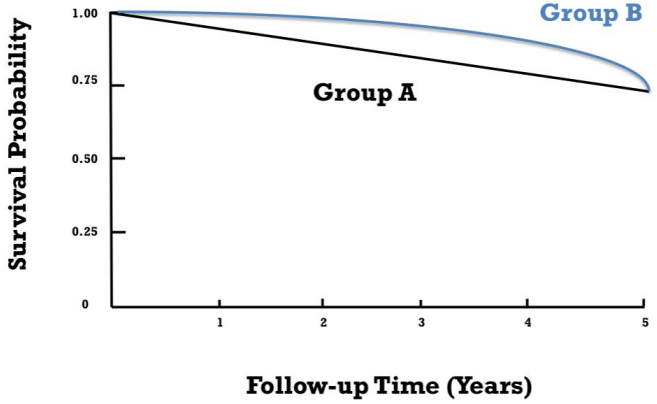
- the survival distributions may differ even though the five-year survival rates are similar



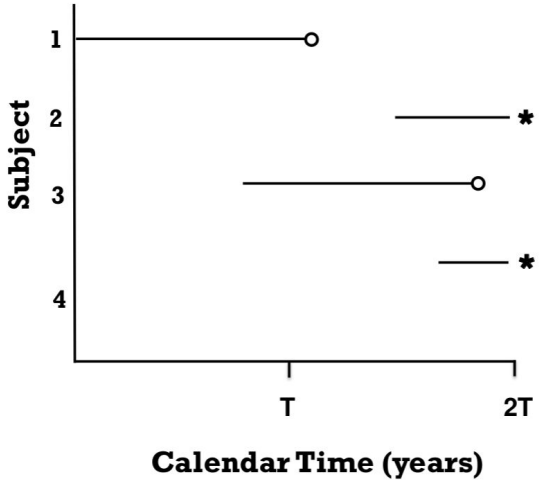
time-to-event data

Quantities of interest:

- **survival time**: time until the event occurs (death, failure, relapse)
- **survival probability** a.k.a. survival function ***S(t)***, is the probability that an individual survives from the time origin (e.g. diagnosis of cancer) to a specified **future time "t"**
- **hazard** (***h(t)***, or ***λ(t)***) is the probability that an individual who is under observation at a time "t" **has an event** at that time



survival data



⚠ we can't assume normality

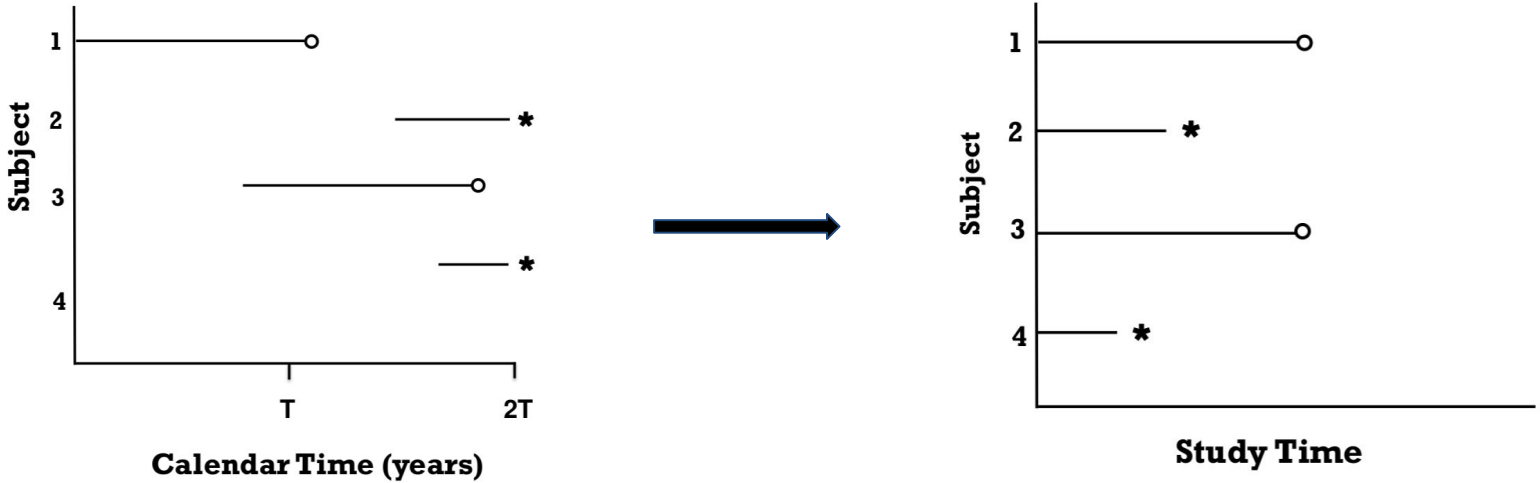
⚠ **censored data** (*right censoring*):

- study follow-up ends before a participant has experienced the event
- participants withdraw or are lost to follow-up, again prior to observing the event

Two patients have events (circles), two are censored (asterisks) because the study ended

survival data

Time to event data are **normalised** by representing each **time record relative to admission date/enrollment**



survival data

- ***T***: random variable representing **time to event** (e.g. death) for a subject
- ***F(t)***: the **probability** that the event (e.g. death) occurs before time ***t*** (end of study): **cumulative risk**, or **distribution function for time-to-event** (T)

$$F(t) = Pr(T < t)$$

- survival is the **complement of *F(t)***, defined as the probability that the subject has not had the event by time t

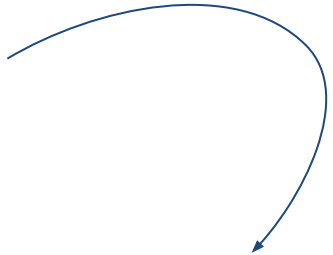
$$S(t) = 1 - F(t)$$

Kaplan-Meier estimate of S(t)

- **S(t)** would be easy to estimate if there were no censoring: however, we almost always have censored data → **Kaplan-Meier estimate of S(t)**
- K-M **updates S(t)** (**step function**) **when events occur** based on the proportion of study participants followed to that time point who have an event

Kaplan-Meier estimate of S(t)

group	year	deaths	survivors
1	1	20	80
2	1	25	75
1	2	20	60
2	2	NA	NA



group	deaths	survivors	year_1	year_2
1	20	80	0.8	NA
1	20	60	NA	0.75
2	25	75	0.75	NA
2	NA	NA	NA	NA

Kaplan-Meier estimate of S(t)

group	year	deaths	survivors
1	1	20	80
2	1	25	75
1	2	20	60
2	2	NA	NA

group	deaths	survivors	year_1	year_2
1	20	80	0.8	NA
1	20	60	NA	0.75
2	25	75	0.75	NA
2	NA	NA	NA	NA

155/200

year	Pr(S)
1	0.775
2	0.75

Kaplan-Meier estimate of S(t)

year	Pr(S)
1	0.775
2	0.75

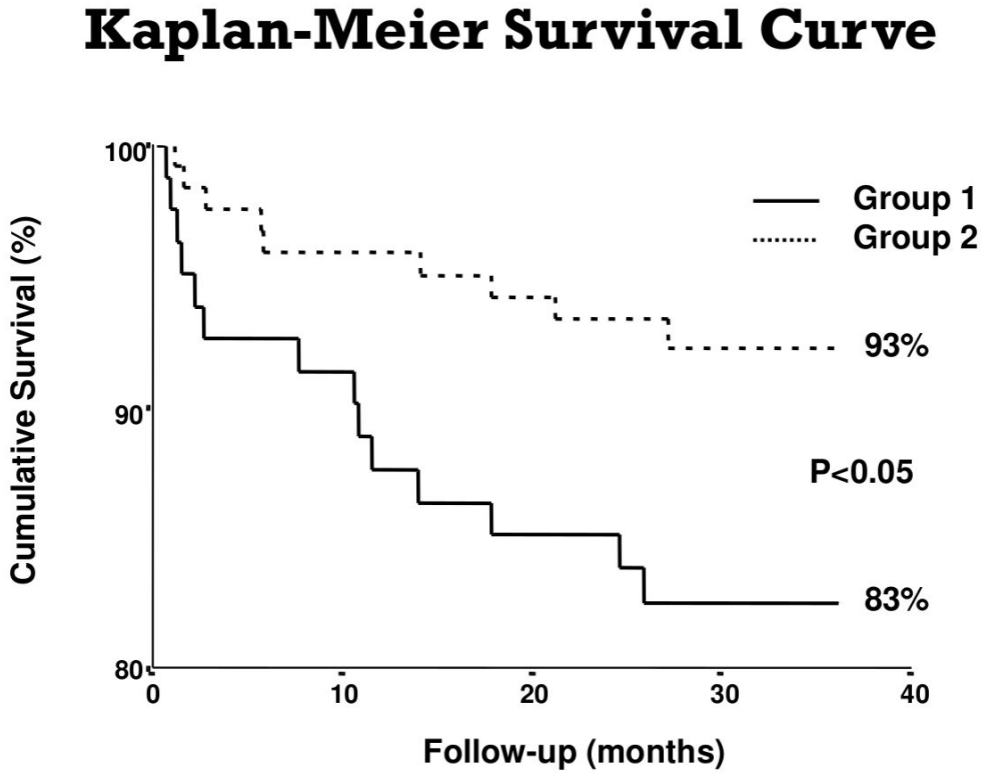
Conditional probabilities:

- **S(year1) = 0.775**
- **S(year2|year1) = 0.75** (alive at year 2 given they're alive at year 1)

Then, **S(year2) = S(year1)*S(year2|year1) = 0.775*0.75 = 0.58**

P(year1 AND year2)

Kaplan-Meier curves



comparing survival curves

- once we have constructed our survival curves, we usually want to know if they differ between groups (e.g. treatments, sexes, breeds etc.)

<u>Many ways to do this:</u>

1. Mantel-Haenszel test
2. Log-rank test

Both are based on multi-dimensional contingency (frequency) tables, comparing observed and expected frequencies accounting for stratification (e.g. treatment or sex or breed)

from Kaplan-Meier curves to Cox models

- Kaplan-Meier curves and log-rank tests are useful for univariate analysis, describing survival in terms of one factor under investigation, and typically work only with categorical predictors (e.g. sex, treatment A vs treatment B etc.)

- this is where **Cox proportional hazards regression analysis** comes in handy: it works for **both quantitative predictor** variables **and for categorical variables**. Furthermore, the Cox regression model extends survival analysis methods to **assess simultaneously the effect of several risk factors** on survival time

- Cox models examine how specific factors (covariates) influence the rate (hazard rate) of a particular event happening (e.g. infection, death) at a particular point in time

- Cox regression is based on the **proportional hazards assumption**: the hazard ratio between the two groups (e.g. treated/untreated) remains constant over time

from Kaplan-Meier curves to Cox models

The hazard function (**λ()** or **h()**) is defined as the **event rate at time *t* conditional on survival until time *t*** (or later, T ≥ t) → suppose a subject has survived for a time *t* and we want the probability that it will not survive for an additional time dt:

$$h(t) = \frac{P(t<T<(t+dt))}{P(T>t)dt} = h_0(t)exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p)$$

where $h_0(t)$ is baseline or reference hazard

from Kaplan-Meier curves to Cox models

we can then express h(t) relative to $h_0(t)$ and take the logarithm (rings a bell?):

$$\ln\left(\frac{h(t)}{h_0(t)}\right) = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

integrating over the elapsed time *t*, we obtain the cumulative risk (**F(t)**), which is related to the survival function (**S(t)**) [remember: **F(t) = P(T<t) → S(t) = 1- F(t)]**

Cox models: interpret the coefficients

HR = 1: no effect on the hazard of the event.
HR < 1: decreased hazard (lower risk) of the event.
HR > 1: increased hazard (higher risk) of the event.

$$\ln\left(\frac{h(t)}{h_0(t)}\right) = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

coefficients: change in hazard (or risk) associated with a one-unit change in the predictor variable, while holding other variables constant

- no effect: □ = 0 → $\mathrm{HR} = e^{\beta} = e^{0} = 1$
- increased risk: □ = 0.1 → $\mathrm{HR} = e^{0.1} = 1.105 - 1 = 0.105$ **+10.5%**
- decreased risk: □ = -0.15 → $\mathrm{HR} = e^{-0.15} = 0.861 - 1 = -0.139$ **-13.9%**