# Survival data analysis

## Statistical modelling: Theory and practice

Gilles Guillot

gigu@imm.dtu.dk

November 24, 2014

# Survival data

- Historically: data occuring in the analysis of
  - time before death of a living organism
    (hence all the sinister vocabulary)


  - time before failure of an engineering system
    (e.g. electric component)
- In modern days data flood: any kind of time data.
  - medicine: life duration of a patient from initial infection by a virus
  - engineering: life duration of an electronic component
  - social science: time of unemployment
  - web data analysis: time spent browsing a certain web page

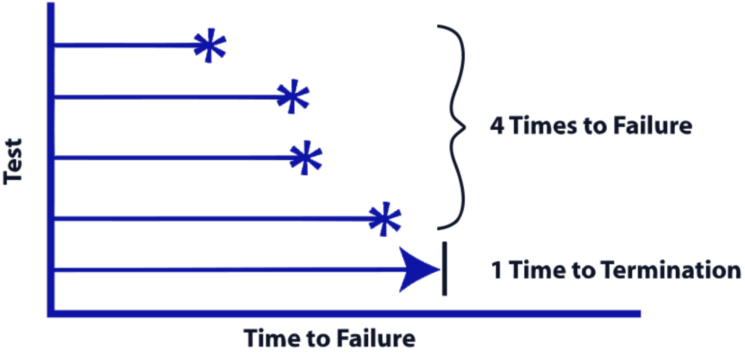
# Goals of survival data analysis I

- Describe variation
- Relate observed times to relevant covariates
    - medicine: Y ~ drug treatment + patient's genetic background + environmental factors
    - engineering: Y ~ manufacturing process + usage conditions
    - job market analysis: Y ~ education level + skills
    - web data analysis: Y ~ age + gender + living area + income
- Predict

# Need for new tools?

- Life durations are $>0$ and skewed (assymetric distrib).
  Whatever you say under a normality assumption is likely to be wrong (unrealistic p-values, inaccurate predictions, inappropriate model selection etc...)
- Life duration data are often censored
  - right censored, e.g. an electrical component lasted more than the duration of the experiment
  - left censored; e.g. a patient was already ill when he entered the study

# Censored data




(Reprinted from wikipedia)

# Probability concepts for survival data analysis

**Definition: Survivor (or survival) function**

$S(y)$ is defined as the probability of surviving beyond $y$

$S(y) = P(Y > y) = 1 - P(Y \leq y) = 1 - F(y)$

# Probability concepts for survival data analysis

## Definition: Hazard function

$h(y) =$ probability density of "death" in infinitesimal small time between $y$ and $y + \delta y$ given survival up to time $y$

Denote $E = \{y < Y \leq y + \delta y\}$ and $F = \{Y > y\}$, then

$$
\begin{aligned}
h(y) &\approx \frac{1}{\delta y} P(E|F) = \frac{1}{\delta y} \frac{P(E \cap F)}{P(F)} \\
&= \frac{1}{\delta y} \frac{P(y < Y \leq y + \delta y)}{S(y)} = \frac{1}{\delta y} \frac{F(y + \delta y) - F(y)}{S(y)}
\end{aligned}
$$

letting $\delta y \longrightarrow 0$ we get

$$
h(y) = \frac{f(y)}{S(y)}
$$

# Estimation of the survival function

## Estimation of the survival function for uncensored data

For a set of uncensored survival times $(y_1, ..., y_n)$, $S(y) = P(Y > y)$ can be estimated by

$$\widehat{S}(y) = \frac{\#\{y_i, y_i > y\}}{n}$$

- This function is non-increasing, piecewise constant and discontinuous.
- This estimator does not rely on any assumption regarding the distribution of survival times, it is known as a **non-paremetric estimator**.
- As such, it is general but converges slowly to the true survival function (when the data size $n$ increases)
- The estimatar can be improved if some knowledge is available on the distribution of survival times

## Estimator of the survival function for censored data

Let us assume that our dataset consists of 5 observations, one of them being censored with values $(1, 3, 3, 5^+, 7)$.
Here, $5^+$ means that the true value is at least as larger as $5$

# Estimator of the survival function for censored data

- For $y \leq 5$ we can estimate $S(y)$ as before by $\widehat{S}(y) = \dfrac{\#\{y_i, y_i > y\}}{n}$
- For $y > 5$, how do we handle the censored data points $5^+$?

  We can of course disregard this data point and apply the previous estimator to the restricted dataset $(1, 3, 3, 7)$.

  In math style: $\widehat{S}(y) = \dfrac{\#\{y_i, y_i > y, y_i \text{ uncensored}\}}{\#\{y_i, y_i \text{ uncensored}\}}$.

  This strategy would incur a loss information. It can be improved.

# Estimator of the survival function for censored data I

For $y > 5$, we can write

$$\begin{aligned} S(y) &= P(Y > y) \\ &= P(Y > 5 \cap Y > y) \\ &= P(Y > 5)P(Y > y|Y > 5) \end{aligned}$$

The fist term in the rhs can be estimated as

$$\widehat{P}(Y > 5) = \frac{\#\{y_i, y_i > 5\}}{n}$$

The second term in the rhs can be estimated as

$$\widehat{P}(Y > y|Y > 5) = \frac{\#\{y_i, y_i > y, y_i \text{ uncensored}\}}{\#\{y_i, y_i > 5, y_i \text{ uncensored}\}}$$

# Estimator of the survival function for censored data II

Combining the two terms above we get:

$$\widehat{S}(y) = \frac{\#\{y_i, y_i > 5\}}{n} \times \frac{\#\{y_i, y_i > y, y_i \text{ uncensored}\}}{\#\{y_i, y_i > 5, y_i \text{ uncensored}\}}$$

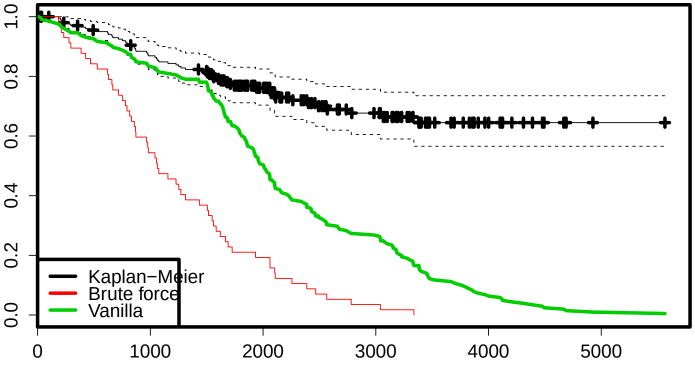
- Splitting the time interval $[0, y)$ into $[0, 5]$ and $(5, y)$ allows us to make use of all the information present in the data.
- This is known as the **Kaplan-Meier estimator** of the survival function.
- This idea can be extended for data with any number of censored observations.

# Estimator of the survival function with R

```
library("survival")
library("ISwR") ; data("melanom")  # for data


## creates an object of class survival
melanom.surv = Surv(time=melanom$days,
                    event= melanom$status==1, # TRUE for death
                    type='right')

## Kaplan-Meier
surv.func.km = survfit(melanom.surv~1)
plot(mfit)
```




Three estimations of the survival function for the `melanom` data

# Modelling survival data

# The exponential distribution

- Density: $f(y) = \frac{1}{\theta} e^{-y/\theta} I_{[0,+\infty[}(y)$ for some parameter $\theta > 0$
- Expectation: $E[Y] = \theta$
- CDF: $F(y) = 1 - e^{-y/\theta}$
- Survival function $S(y) = e^{-y/\theta}$
- Hazard function: $h(y) = \dfrac{f(y)}{S(y)} = 1/\theta$ for $y \geq 0$

The survival function is constant, a property known as "absence of memory". Under this distribution, life is "equally risky" at any age.

# A simple generalized linear model for survival data

## Exponential regression with log link function

- Data: survival times $y = (y_1, ..., y_n)$ and a covariate $x = (x_1, ..., x_n)$
- We assume that each $y_i$ arises from an exponential distribution, with mean $\theta_i$, for short: $Y_i \sim \mathcal{E}(\theta_i)$
- We assume further that $\theta_i$ is governed by $x_i$
  - $E[Y_i] = \theta_i = \beta_0 + \beta_1 x_i$ would be a poor choice
  - We assume instead: $E[Y_i] = \theta_i = e^{\beta_0 + \beta_1 x_i}$

The above is an instance of a proportional hazard model

# Exponential regression in R

Previous model not implemented in R function `glm`.
Work around: use a Gamma distribution and twist the results.

```
## fits a glm under the more general assumption
## of a Gamma distribution
glm.res <- glm(formula= (y~ x),
               family=(Gamma(link=log)))
## extract some results from model fit now under
## the assumption that the distribution was exponential
summary(glm.res,dispersion=1)
```

# Reading

- Regression with Linear Predictors P. K. Andersen L. T. Skovgaard, Springer Statistics for Biology and Health, 2010.
- Introductory statistics with R, Chapter Survival analysis, P. Dalgaard, Series Statistics and Computing, Springer, 2008.