Kaplan-Meier Method

Introduction

Introduction

- Survival function
  - The survival function $S(t)$ gives probability of surviving at least to time t.
- Our goal: to develop a *nonparametric estimator* for $S(t)$.
- Nonparametric estimation means no need to specify a distribution for survival times.

Introduction

- Kaplan-Meier(KM) estimator is the most widely used method for estimating survival functions.
- It is also known as the product-limit estimator.
- First proposed in

Kaplan, E. L.; Meier, P. (1958). "Nonparametric estimation from incomplete observations". J. Amer. Statist. Assn. 53 (282): 457-481.

- Meier and Kaplan had each submitted similar paper to *the Journal of the American Statistical Association*.
- The editor then convinced them to combine their work into one paper.

Background

- Edward Kaplan (1920-2006)
  - Got interested in lifetime of vacuum tubes while working at Bell laboratory
  - Was also a professor in Mathematics at Oregon State University (1961-1981)
- Paul Meier (1924-2011)
  - A professor at University of Chicago
  - Got interested in cancer patient survival
  - Made major contributions in randomized clinical trials

Background

- One of the most cited statistical papers.

Paper	Total Citation	Citation per year
Kaplan-Meier (1958)	46443	800
Cox model (1972)	40310	916
EM (Dempster 1977)	43524	1116
Bootstrap (Efron 1979)	12674	342
FDR (Benjamini & Hockberg 1995)	30345	1445

Kaplan-Meier estimator : a toy example

A toy example: suppose we have five patients

- Patient dropped out after 5 months.
- Patient B survived for the whole year long study
- Patient C died at 8 month
- Patient D survived for the whole year long study
- Patient E died at 4 month

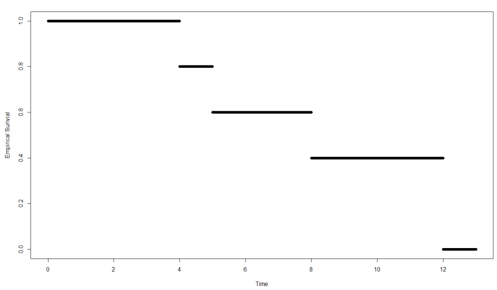
Data {5+, 12+, 8, 12+, 4 }

+: censoring

Uncensored data

- If no censoring, use empirical survival function
$$\hat{S}(t) = \frac{\#\{i:\ t_i > t\ \}}{n}$$
- If we have data {5, 12, 8, 12, 4 }, then
$\widehat{S}(5) = 3/5$, and $\widehat{S}(6.5) = 3/5$
- The resulting $\hat{s}(t)$ is a step-wise function, and decreases at event times.
- For censored observations, we don't know whether $t_i\ > t$ or not.

Empirical survival function



Right censored data

- Key idea of KM estimator
$$\hat{S}(t) = \hat{S}(t^{-})\hat{p}(T > t|T \geq t)$$
  - $S(t)$: probability surviving time t
  - $S(t^{-})$: probability surviving a time just before t
  - $p(T > t|T \geq t)$: probability surviving time t given still alive before t
- Conditional probability estimate:
  - $\hat{p}(T > t|T \geq t)$=1 if no failure at t
  - $\hat{p}(T > t|T \geq t)$=$1 - \frac{d(t)}{n(t)}$ if any failure at t
  - $d(t)$: # failures at t; $n(t)$: # still alive just before t.

Kaplan-Meier estimator

- KM estimator
  $$\hat{S}(t) = \hat{S}(t^-)\hat{p}(T > t|T \geq t)$$
  - $\hat{p}(T > t|T \geq t)$=1 if no failure at t
  - $\hat{p}(T > t|T \geq t)$=$1 - \frac{d(t)}{n(t)}$ if any failure at t
- KM estimator only decreases at event times
- KM estimator does not change
  - Between events
  - At censoring times

Definition

- Suppose the events (not censored) occur at D distinct times
  $t_1 < t_2 < \cdots < t_D$ (D≤n)
- At each event time t_i, define
  - d_i: number of events at t_i
  - n_i: number of individuals at risk just before t_i
  - $\frac{d_i}{n_i}$: conditional prob. of failure at t_i given alive before t_i.
  - $1 - \frac{d_i}{n_i}$: conditional prob. of surviving t_i given alive before t_i.
- At event time t_i, the K-M estimator is calculated recursively by
  $$\hat{S}(t_i) = \hat{S}(t_{i-1})(1 - \frac{d_i}{n_i}).$$

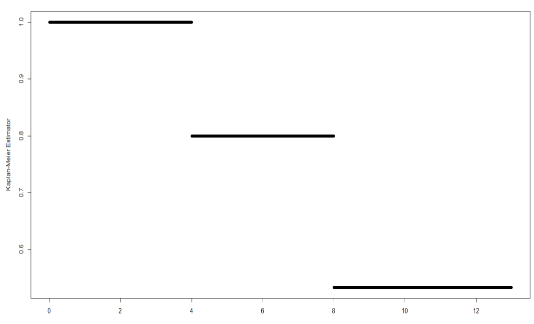
K-M estimator

Then the Kaplan-Meier (Product-limit) estimator of S(t) is

$$\hat{S}(t) = \begin{cases} 1 & if\ t < t_1 \\ \prod_{t_i \leq t} \left[1 - \frac{d_i}{n_i}\right] & if\ t \geq t_1 \end{cases}$$

An example

- Recall data {5+, 12+, 8, 12+, 4 }.
- $\hat{S}(4)$ at event times:
  - $\hat{S}(4)$=1-$\frac{d_1}{n_1} = 1 - \frac{1}{5} = \frac{4}{5}$
  - $\hat{S}(8) = \hat{S}(4) \times \left(1 - \frac{d_2}{n_2}\right)$=$\frac{4}{5} \times \left(1 - \frac{1}{3}\right)$=$\frac{8}{15}$



Summary

- K-M estimator: a nonparametric method for estimating survival function.
- The K-M estimator changes value only at event times.
- The size of change depends not only on the number of event observed at event time t_i, but also number of censored observations prior to t_i.
- If there is no censoring, K-M estimator is simply the empirical survival function.
- Another important property of the K-M estimator is that it is also the nonparametric MLE of S(t).