Time to Event Analysis (Part 1)

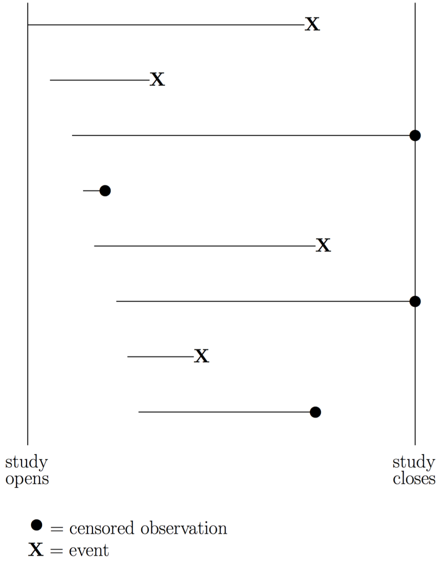
Christof Seiler

Stanford University, Spring 2016, STATS 205

Overview

- Survival data
- Survival function
- Hazard function
- Kaplan-Meier estimator

Survival Data



Source: Ibrahim (2005)

Survival Data

- We are interested in time to an event of interest as the outcome variable
- In medicine:
  - Often in a clinical trial the goal is to evaluate the **effectiveness of a new treatment at prolonging survival**
  - For example to extend the time to the **event of death**
  - It is usually the case that at the end of followup a **portion of the subjects** in the trial have **not experienced the event**
  - For these subjects the outcome variable is **censored**

Survival Data

- In engineering:
  - Similarly, in engineering studies, often the lifetimes of **mechanical or electrical parts** are of interest
  - In a typical experimental design, lifetimes of these parts are recorded along with covariates (including design variables)
  - Often the lifetimes are called **failure times**, i.e., times until failure
  - As in a clinical study, at the end of the experiment, there may be **parts** which are **still functioning** (censored observations)

Survival Data

- One object of interesting is the **survival function**
- An popular estimator of this function is the **Kaplan-Meier estimator**
- It is common to estimate **two functions** (e.g. in a case/control experiment) and to test the hypothesis that they are the same
- A popular test is the **Log rank test**
- An alternative test is the **Gehan's test**
- Another object of interst is the **hazard function**
- Which can be interpreted as the instantaneous chance of the event (death)
- In this context, we'll talk about the **Cox proportional hazards models**

Survival Data

- Failure time random variables are always **non-negative**
- Denote the failure time by T, then $T \geq 0$
- T can either be discrete (taking a finite set of values $a_1, a_2, \ldots, a_n$) or continuous (defined on $(0, \infty)$)
- A random variable X is called a censored failure time random variable if $X = \min(T, U)$, where U is a non-negative censoring variable
- In order to define a failure time random variable, we need:
    1. an unambiguous **time origin** (e.g. randomization to clinical trial, purchase of car)
    2. a **time scale** (e.g. real time (days, years), mileage of a car)
    3. definition of the **event** (e.g. death, need a new car transmission)

Survival Data

- Several features which are typically encountered in analysis of survival data:
  - individuals do not all enter the study at the same time
  - when the study ends, some individuals still haven't had the event yet
  - other individuals drop out or get lost in the middle of the study, and all we know about them is the last time they were still "free" of the event
- The first feature is referred to as **"staggered entry"**
- The last two features relate to **"censoring"** of the failure time events

Survival Data

- Right-censoring: only the $X_i = \min(T_i, U_i)$ is observed due to
  - loss to follow-up
  - drop-out
  - study termination
- We call this right-censoring because the true unobserved event is to the right of our censoring time
- All we know is that the event has not happened at the end of the study

Survival Data

- Suppose we have a sample of observations on n people:

$$(T_1, U_1), (T_2, U_2), \ldots, (T_n, U_n)$$

- There are three main types of (right) censoring times:
  - Type I: All the U_i's are the same (e.g. animal studies, all animals sacrificed after 2 years)
  - Type II: $U_i = T(r)$, the time of the rth failure (e.g. animal studies, stop when $4/6$ have tumors)
  - Type III: the U_i's are random variables
- Type I and Type II are called singly censored data, Type III is called randomly censored (or sometimes progressively censored)

Survival Function

- Let T denote the time to an event
- Assume T is a continuous random variable with cdf $F(t)$
- The survival function is defined as the probability that a subject survives until at least time t

$$S(t) = P(T > t) = 1 - F(t)$$

- When all subjects in the trial experience the event during the course of the study
- So that there are no censored observations, an estimate of $S(t)$ may be based on the empirical cdf

Survival Function (Example: No Censored Observations)

- Treatment of pulmonary metastasis (cancer spreads to lung), survival time (in months) was collected

```
survTimes = c(11,13,13,13,13,13,14,14,15,15,17)
```

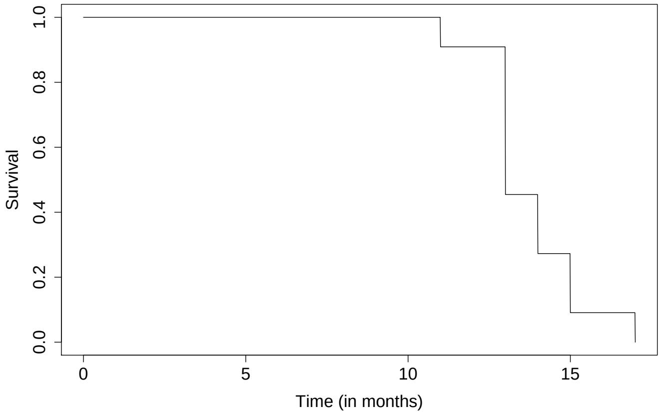
- No censored observation, we can estimate survival function at time t with empirical cdf

$$\widehat{S}(t) = \frac{\#\{t_i > t\}}{n}$$

- Estimate by counting

$$\widehat{S}(t) = \begin{cases} 1 & 0 \leq t < 11 \\ \frac{10}{11} & 11 \leq t < 13 \\ \frac{5}{11} & 13 \leq t < 14 \\ \frac{3}{11} & 14 \leq t < 15 \\ \frac{1}{11} & 15 \leq t < 17 \\ 0 & t \geq 17 \end{cases}$$

Survival Function (Example: No Censored Observations)



Kaplan-Meier Estimator

- In most clinical studies, at the end of the study there are subjects who have yet to experience the event being studied
  - Either hasn't happened
  - Or the subject died before
- In such cases the Kaplan-Meier product limit estimate can be used
- Let $t(1) < \cdots < t(k)$ denote the ordered distinct event times
- If there are censored responses, then $k < n$
- Let $n_i = \#$subjects at risk at the beginning of time $t(i)$
- Let $d_i = \#$events occurring at time $t(i)$
- The **Kaplan-Meier estimate** of the survival function is defined as

$$\widehat{S}(t) = \prod_{t(i) \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

Kaplan-Meier Estimator (Example)

- Event: time to relapse
- Data:
  - Relapse: $3, 6.5, 6.5, 10, 12, 15$
  - Lost to followup: 8.4
  - Alive and in remission at at end of study: $4, 5.7, 10$
- Step-by-step Kaplan-Meier estimate:

t	n	d	$1 - d/n$	$S(t)$
3	10	1	$9/10 = 0.9$	0.9
6.5	7	2	$5/7 = 0.71$	$0.9 \times 0.71 = 0.64$
10	4	1	$3/4 = 0.75$	$0.64 \times 0.75 = 0.48$
12	2	1	$1/2 = 0.5$	$0.48 \times 0.5 = 0.24$
15	1	1	$0/1 = 0$	0

Kaplan-Meier Estimator

- Intuition behind the Kaplan-Meier Estimator
- Think of dividing the observed timespan of the study into a series of fine intervals so that there is a separate interval for each time of death or censoring
- Using the law of conditional probability,

$$P(T \geq t) = \prod_i P(\text{ survive } i\text{th interval } I_i | \text{ survived to start of } I_i)$$

where the product is taken over all the intervals including or preceding time t

Kaplan-Meier Estimator

- 4 possibilities for each interval:
  1. **No events (death or censoring)**: conditional probability of surviving the interval is 1
  2. **Censoring**: assume they survive to the end of the interval, so that the conditional probability of surviving the interval is 1
  3. **Death, but no censoring**: conditional probability of not surviving the interval is # deaths (d) divided by # at risk (n) at the beginning of the interval. So the conditional probability of surviving the interval is $1 - (d/n)$
  4. **Tied deaths and censoring**: assume censorings last to the end of the interval, so that conditional probability of surviving the interval is still $1 - (d/n)$
- The general formula for the conditional probability of surviving the ith interval that holds for all 4 cases:

$$1 - \frac{d_i}{n_i}$$

Kaplan-Meier Estimator

- We could use the same approach by **grouping the event times into intervals** (say, one interval for each month), and then **counting** up the number of deaths (events) in each to estimate the probability of surviving the interval (this is called the lifetable estimate)
- However, the assumption that those **censored last until the end of the interval wouldn't be quite accurate**, so we would end up with a cruder approximation
- As the **intervals get finer and finer**, the approximations made in estimating the probabilities of getting through each interval become smaller and smaller, so that the estimator **converges** to the true $S(t)$
- This intuition clarifies why an alternative name for the KM is the product limit estimator

Kaplan-Meier Estimator

- Suppose $a_k < t < a_{k+1}$, then

$$S(t) = P(T \geq a_{k+1}) = P(T \geq a_1, T \geq a_2, \ldots, T \geq a_{k+1})$$

- Rewrite in terms of conditional probabilities

$$P(T \geq a_1) \times P(T \geq a_2 | T \geq a_1) \times \cdots \times P(T \geq a_k + 1 | T \geq a_k)$$

$$= P(T \geq a_1) \times \prod_{i=1}^{k} P(T \geq a_{i+1} | T \geq a_i)$$

- Now use
$1 - P(T = a_i | T \geq a_i) = P(T \geq a_i) P(T \geq a_{i+1} | T \geq a_i)$

$$= \prod_{i=1}^{k} (1 - P(T = a_i | T \geq a_i)) = \prod_{i=1}^{k} (1 - \lambda_i)$$

- λ_i is called the discrete Hazard function

Kaplan-Meier Estimator

- The Hazrd function $\lambda(t)$ is sometimes called an instantaneous failure rate, the force of mortality, or the age-specific failure rate
- For continuous random variables:

$$\begin{aligned}
\lambda(t) &= \lim_{\Delta t \to 0} \frac{1}{\Delta} P(t \leq T < t + \Delta | T \geq t) \\
&= \lim_{\Delta t \to 0} \frac{1}{\Delta} \frac{P(t \leq T < t + \Delta, T \geq t)}{P(T \geq t)} \\
&= \lim_{\Delta t \to 0} \frac{1}{\Delta} \frac{P(t \leq T < t + \Delta)}{P(T \geq t)} \\
&= \frac{f(t)}{S(t)}
\end{aligned}$$

Kaplan-Meier Estimator

- For discrete random variables ($\lambda(a_i) := \lambda_i$):

$$\lambda_i = P(T = a_j | T \geq a_i) = \frac{P(T = a_i)}{P(T \geq a_i)}$$

$$= \frac{f(a_i)}{S(a_i)} = \frac{f(t)}{\sum_{k:a_k \geq a_i} f(a_k)}$$

- Using an estimate d_i/n_i of the the Hazard function
  - d_i is the number of deaths at a_i and
  - n_i is the number at risk at a_i

$$\widehat{S}(t) = \prod_{i=1}^{k} \left(1 - \frac{d_i}{n_i}\right)$$

Efron's Redistribute-to-the-Right Algorithm

- Example data: `4, 5, 5+, 6+, 7, 8+, 9, 11`
- If all were non-cencered then empirical survival function would assign mass $\frac{1}{8}$ to each of the values
- At first censored time `5+`, a death has not occured but will occur somewhere to the right of `5`
- Efron's algorithm takes the mass of $\frac{1}{8}$ of `5+` and redistributes it equally among the five times `6+, 7, 8+, 9, 11+` to the right of `5+`, adding $\frac{1}{5}(\frac{1}{8})$ to the mass at `6+, 7, 8+, 9, 11+`
- Now go to the next censored time `6+` and redistribute the new mass $\frac{1}{5}(\frac{1}{8}) + \frac{1}{8}$ equally among the observations to the right of `6+`
- Continue this process until you reach the last observation
- Efron (1967) showed that this yields the Kaplan-Meier estimator

Properties of Kaplan-Meier Estimator

- In case of no censoring (number of subjects n):

$$\widehat{S}(t) = \frac{\#\{t_i > t\}}{n}$$

- This is like the binomial proportion problem (where we estimated success probability):

$$\widehat{S}(t) \xrightarrow{d} N(S(t), S(t)(1 - S(t))/n)$$

- Much harder in the censored case
- $\widehat{S}(t)$ still is approximately normal
- The mean $\widehat{S}(t)$ converges to the true $S(t)$
- The variance is more complicated (since the denominator n includes some censored observations)

Properties of Kaplan-Meier Estimator

- We can calculate the variance using Greenwood's formula

$$\mathrm{Var}(\widehat{S}(t)) = \widehat{S}(t)^2 \sum_{t(i)<t} \frac{d_i}{(n_i - d_i)n_i}$$

- Think of the KM estimator as

$$\widehat{S}(t) = \prod_{t(i)<t} (1 - \widehat{\lambda}_i)$$

- With $\widehat{\lambda}_i = d_i/n_i$, and since $\widehat{\lambda}_i$ are binomial proportions, we have

$$\widehat{S}(t) \xrightarrow{d} N(\lambda_i, \widehat{\lambda}_i(1 - \widehat{\lambda}_i)/n_i)$$

- Also assume that $\widehat{\lambda}_i$ are independent
- Since $\widehat{S}(t)$ is a function of the $\widehat{\lambda}_i$'s, we can estimate its variance using the delta method

Properties of Kaplan-Meier Estimator

- Delta method: If Y is normal with mean μ and variance σ^2, then $g(Y)$ is approximately normally distributed with mean $g(\mu)$ and variance $\left(g'(\mu)\right)^2 \sigma^2$
- Take the log so that we can work with sums

$$\log(\widehat{S}(t)) = \sum_{t(i)<t} \log(1 - \widehat{\lambda}_i)$$

- And using the independence assumption

$$\text{Var}\left(\log(\widehat{S}(t))\right) = \sum_{t(i)<t} \text{Var}\left(\log(1 - \widehat{\lambda}_i)\right)$$

- If $Z = \log(Y)$, then $Z \sim N(\log(\mu), \left(\frac{1}{\mu}\right)^2 \sigma^2)$

$$\text{Var}\left(\log(\widehat{S}(t))\right) = \sum_{t(i)<t} \left(\frac{1}{1 - \widehat{\lambda}_i}\right)^2 \text{Var}\left(\widehat{\lambda}_i\right)$$

Properties of Kaplan-Meier Estimator

- Then replace $\mathrm{Var}(\widehat{\lambda}_i)$ with $\widehat{\lambda}_i(1-\widehat{\lambda}_i)/n_i$ and set $\widehat{\lambda}_i = d_i/n_i$

$$\mathrm{Var}\left(\log(\widehat{S}(t))\right) = \sum_{t(i)<t} \frac{d_i}{(n_i - d_i)n_i}$$

- Take exp to transform it back $\widehat{S}(t) = \exp(\log(\widehat{S}(t)))$
- If $Z = \exp(Y)$, then $Z \sim N\left(e^{\mu}, (e^{\mu})^2\sigma^2\right)$

$$\mathrm{Var}(\widehat{S}(t)) = \widehat{S}(t)^2\,\mathrm{Var}(\log(\widehat{S}(t)))$$

- And we end up with

$$\mathrm{Var}(\widehat{S}(t)) = \widehat{S}(t)^2 \sum_{t(i)<t} \frac{d_i}{(n_i - d_i)n_i}$$

References

- The Statistical Analysis of Failure Time Data (2002). Kalbfleisch and Prentice
- Analysis of Survival Data (1984). Cox and Oakes
- Lecture Notes (2005). Ibrahim
- Hollander and Wolfe, and Chicken (2013). Nonparametric Statistical Methods