Biostat 537: Survival Analaysis

TA Session 3

Ethan Ashby

January 22, 2024

A Review of Last time

1. Parametric survival models assume a particular shape of the distribution of survival times, which are governed by a finite set of parameters.
2. We showed parametric models are convenient for estimation and converting between hazards, survival functions, and single-number summaries of the survival experience.
3. Maximum likelihood enables estimation and inference on the parameters from data.

Some HW 1 Concepts to Review

1. Truncation
   1. Truncation versus censoring.
   2. Truncation $\subset$ Selection Bias.
   3. Redefining time origin ($t = 0$).
2. 'fitparametric' capabilities – "mean", "quantile", "survival", "condsurvival"
3. Significance testing
   1. Goodness of fit (different models fit to same data): likelihood ratio test
   2. Comparing survival distribution between parametric models: Wald test on derived model parameters
   3. Comparing survival distributiosn nonparametrically: Logrank test or variant.

Session Overview

1. Nonparametric Survival Curve Estimation
2. Nonparametric Estimation of Other Survival Quantities
3. Nonparametric comparison Survival Curves

Why go nonparametric?

The use of parametric models are often justified using

1. Convenience: ease of converting between survival quantities of interest, relatively simple estimation.
2. Efficient: *when correctly specified*, parametric models produce estimators w/ smallest possible variances.

Reasons why we may want to go nonparametric

1. Agnosticism around choice of model.
2. True survival experience unlikely to adhere to rigid parametric assumptions.
3. Conclusions that avoid making non-essential statistical assumptions.

Motivating Example for Kaplan-Meier Estimator

Consider the following survival data. Unique event times in red.

Patient	Survival Time	Status
1	7	0
2	6	1
3	6	0
4	5	0
5	2	1
6	4	1

Suppose we wish to estimate the survival function $S(t)$ without making parametric assumptions on the shape of the distribution of survival times.

Kaplan-Meier Curve Example

Suppose we construct the following table where t_i denotes a specific time of interest, n_i are the number of participants in the *risk set*, d_i are the number of events that occurred at t_i, q_i are the number censored at t_i.

t_i	n_i	d_i	q_i	$P(\text{Event at } t_i \mid \text{At Risk at } t_i) = \frac{d_i}{n_i}$
1	6	0	0	$\frac{0}{6}$
2	6	1	0	$\frac{1}{6}$
3	5	0	0	$\frac{0}{5}$
4	5	1	0	$\frac{1}{5}$
5	4	0	1	$\frac{0}{4}$
6	3	1	1	$\frac{1}{3}$
7	1	0	1	$\frac{0}{1}$

Kaplan-Meier Curve Example

Suppose we are interested in estimating the survivor curve

$$S(t) = P(T \geq t)$$

We can break time into a bunch of intervals of unit length.

$$\begin{aligned} S(t) &= P(T \geq t | T \geq t-1) \times P(T \geq t-1) \\ &= P(T \geq t | T \geq t-1) \times S(t-1) \\ &\ldots \text{Iterate} \\ &= \prod_{i=1}^{t} P(T \geq i | T \geq i-1) \\ &\equiv \prod_{i=1}^{t} (1 - P(T \leq i | T \geq i-1)) \\ &\equiv \prod_{i=1}^{t} (1 - P(\text{Event at time } i | \text{At Risk at Time } i)) \end{aligned}$$

Kaplan-Meier Curve Example

$$S(t) = \prod_{i=1}^{t} (1 - P(\text{Event at time } i | \text{At Risk at Time } i))$$

We can estimate pink term using the *observed number of events* at time *i* over the *observed number at risk* at time *i*.

$$\hat{P}(\text{Event at time } i | \text{At Risk at Time } i) = \frac{d_i}{n_i}$$

Yielding

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

Kaplan-Meier Curve Example

So far

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

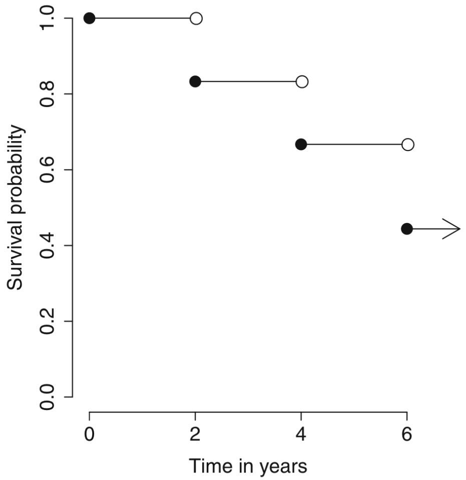
Recall that $d_i \neq 0$ if and only if an event ($d_i = 1$) occurred at time i. Hence, it suffices to consider the *product over the failure times*, $t_i \leq t$:

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

Kaplan-Meier Curve Example

t_i	n_i	d_i	q_i	$\hat{P}(\text{Event at } t_i \mid \text{At Risk at } t_i) = \frac{d_i}{n_i}$	$\hat{S}(t) = \prod_{t_i \le t} \left(1 - \frac{d_i}{n_i}\right)$
0	6	0	0	$\frac{0}{6}$	$\left(1 - \frac{0}{6}\right) = 1$
1	6	0	0	$\frac{0}{6}$	$\left(1 - \frac{0}{6}\right) = 1$
2	6	1	0	$\frac{1}{6}$	$1 \cdot \left(1 - \frac{1}{6}\right) = \frac{5}{6}$
3	5	0	0	$\frac{0}{5}$	$\frac{5}{6} \cdot \left(1 - \frac{0}{5}\right) = \frac{5}{6}$
4	5	1	0	$\frac{1}{5}$	$\frac{5}{6} \cdot \left(1 - \frac{1}{5}\right) = \frac{2}{3}$
5	4	0	1	$\frac{0}{4}$	$\frac{2}{3} \cdot \left(1 - \frac{0}{4}\right) = \frac{2}{3}$
6	3	1	1	$\frac{1}{3}$	$\frac{2}{3} \cdot \left(1 - \frac{1}{3}\right) = \frac{4}{9}$
7	1	0	1	$\frac{0}{1}$	$\frac{4}{9} \cdot \left(1 - \frac{0}{1}\right) = \frac{4}{9}$

Kaplan-Meier Curve Example



The Kaplan-Meier Estimator

The Kaplan-Meier Estimator is the product over the failure times of the conditional probabilities of surviving to the next failure time.

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

Where n_i is the number of individuals in the risk set at time t_i and d_i is the number of individuals who failed at time t_i.

The Kaplan-Meier Estimator

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

The Kaplan-Meier (KM) estimator

1. Makes no assumptions on the distribution of event times.
2. Accommodates censored data by letting censored observations contribute to the risk set n_i.
3. Assumes Non-informative Censoring: $\hat{P}(\text{Event at } t_i | \text{At Risk at } t_i) = \frac{d_i}{n_i}$ unbiased for $P(\text{Event at } t_i | \text{At Risk at } t_i)$.
4. Assumes that survival is constant between observed events.

NPMLE

The K-M estimator can be considered as the maximum likelihood estimator of the discrete hazard function. Let h_i be the hazard of experiencing an event at time t_i.

$$S(t) = \prod_{t_i \leq t} (1 - h_i)$$

Since failures are Bernoulli events, a binomial likelihood up to time t_i can be written as

$$L(h_j; j \leq i) = \prod_{j=1}^{i} h_j^{d_j} (1 - h_j)^{n_j - d_j} \binom{n_j}{d_j}$$

Hence, the maximum likelihood estimator for h_j is given by

$$\hat{h}_j = \frac{d_j}{n_j}$$

Plugging in $\hat{h}_j$ for h_j above gives the KM estimator.

R example

```
#Calculate KM estimator
library(survival)
tt<-c(7,6,6,5,2,4)
cens<-c(0,1,0,0,1,1)
Surv(tt,cens) #formatted as time to event
result.km<-survfit(Surv(tt,cens)~1,conf.type="log-log")
summary(result.km) #output table
plot(result.km) #plot KM curve w/ pointwise CIs
```

Nelson-Aalen Estimator

Suppose we wish to estimate the cumulative hazard.

$$H(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}$$

And recognizing that $S(t) = e^{-H(t)}$, we have similar estimator of the survival function. In R, we we fit it as follows.

```
result.km<-survfit(Surv(tt,cens)~1,conf.type="log-log",
    type="fh")
summary(result.km) #output table
```

Single-Number Summaries of the Survival Experience

We may be interested in the median survival time defined as

$$\hat{t}_{\text{med}} = \inf\{t : \hat{S}(t) \leq 0.5\}$$

By default, “survfit” prints out the estimate and 95% CI for the median.

Hazard Estimation Via Smoothing

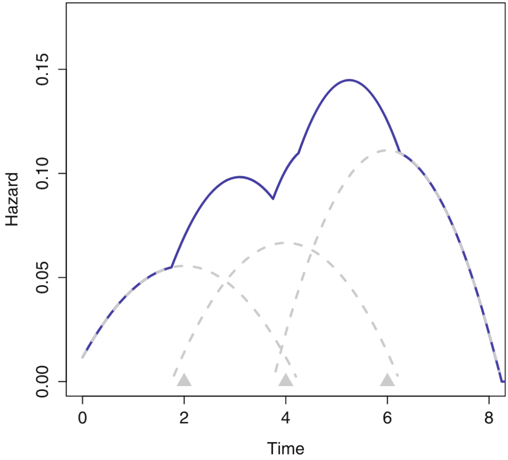
Suppose we are interested in estimating and examining the hazard function $h(t) := \lim_{\Delta t \to 0} \frac{P(t<T \leq t+\Delta t | T \geq t)}{\Delta t}$.

Nelson-Aalen estimates of the hazard function will be 0 with bumps of height d_i/n_i at each event time t_i. This is very unstable estimator with high error.

$$\hat{h}_{NA}(t) = \sum_{i=1}^{D} \mathbb{I}(t_{(i)} = t) \cdot \frac{d_i}{n_i}$$

Smoothing helps us reduce noise by borrowing local information to produce a more stable estimator.

Hazard Estimation Via Smoothing: Illustration



Not necessary but interesting

In mathematical terms, smoothed hazard estimation is accomplished using a *kernel estimator*

$$\hat{h}(t) = \frac{1}{b} \sum_{i=1}^{D} K\left(\frac{t - t_{(i)}}{b}\right) \frac{d_i}{n_i}$$

Where $t_{(1)}, \ldots, t_{(D)}$ are the unique failure times, and K is a non-negative function that assigns more weight to d_i/n_i if t is close to an observed failure time $t_{(i)}$.

The *bandwidth* b controls how much smoothing is performed.

Hazard Estimation Via Smoothing: In R

```
library(muhaz)
t.vec<-c(7,6,6,5,2,4)
cens.vec<-c(0,1,0,0,1,1)
result.smooth<-muhaz(t.vec,cens.vec,max.time=8, bw.grid
    =2.25, bw.method="global", b.cor="none") #smoothed
results.sparse<-pehaz(t.vec,cens.vec,width=1,max.time=8)
     #sparse
plot(result.smooth)
lines(results.sparse)
```

Set “bw.option”=“local” to automatically choose level of smoothing that adapts to the frequency of events in different regions.

Smoothed Estimation of Survival Function

Recall that $S(t) = e^{-\int_0^t h(u)du}$ by definition. Hence, we can replace $h(u)$ by its smoothed estimate $\hat{h}(u)$ to obtain a smoothed estimator of the survival function.

In R, we do this

```
haz <- result.smooth$haz.est
times <- result.smooth$est.grid
surv <- exp(-cumsum(haz[1:(length(haz)-1)]*diff(times)))
```

Roadmap

1. Nonparametric Survival Curve Estimation
2. Nonparametric Estimation of Other Survival Quantities
3. Nonparametric comparison Survival Curves

Motivation

Thus far, we have discussed nonparametric *estimation* of survival quantities of interest such as the survivor function, hazard function, etc.

In many practical situations, we may also wish to *test* whether the survivor curves are significantly different between two groups.

Testing equivalent survival between two groups

We propose a null hypothesis of $H_0 : S_0(t) = S_1(t)$.

We wish to develop a *test statistic* T which quantifies the discrepancy between $S_0(t)$ and $S_1(t)$ *based on the data* without relying on parametric assumptions.

A good starting point: Kaplan-Meier curves give us nonparametric estimates of $S_0(t)$, $S_1(t)$.

An idea: let T be the "distance" from group 1 K-M estimator, $\hat{S}_1(t)$, to pooled K-M estimator under H_0, $\hat{S}(t)$.

Logrank Test

Row	$t_{(i)}$	At risk n_{0i}	n_{1i}	Events d_{0i}	d_{1i}
1	2	10	10	1	0
2	5	9	10	1	0
3	7	8	10	1	0
4	8	7	10	1	1
5	11	6	9	1	0
⋮	⋮	⋮	⋮	⋮	

Under H_0, we assume $S_0(t) = S_1(t) = S(t)$. Hence, we can calculate an *expected* event count for each cell under H_0:

$$e_{ji} := \underbrace{\left(\frac{n_{ji}}{n_{0i} + n_{1i}}\right)}_{\text{Prop at Risk at } t_{(i)}} \times \underbrace{(d_{0i} + d_{1i})}_{\text{Total Failures at } t_{(i)}}$$

Logrank Test

Row	$t_{(i)}$	At risk n_{0i}	n_{1i}	Events d_{0i}	d_{1i}	Expected Events e_{0i}	e_{1i}
1	2	10	10	1	0	$\frac{10}{20} \times (1+0) = \frac{1}{2}$	$\frac{10}{20} \times (1+0) = \frac{1}{2}$
2	5	9	10	1	0	$\frac{9}{19} \times (1+0) = \frac{9}{19}$	$\frac{10}{19} \times (1+0) = \frac{10}{19}$
3	7	8	10	1	0	$\frac{8}{18} \times (1+0) = \frac{8}{18}$	$\frac{10}{18} \times (1+0) = \frac{10}{18}$
4	8	7	10	1	1	$\frac{7}{17} \times (1+1) = \frac{14}{17}$	$\frac{10}{17} \times (1+1) = \frac{20}{17}$
5	11	6	9	1	0	$\frac{6}{15} \times (1+0) = \frac{6}{15}$	$\frac{9}{16} \times (1+0) = \frac{9}{16}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

$$\text{Logrank Statistic} = \frac{\left(\sum_{t_{(i)}} (d_{1i} - e_{1i})\right)^2}{\mathrm{Var}(d_{1i} - e_{1i})}$$

Logrank Test: Some Extra Info

$$\text{Logrank Statistic} = \frac{\left(\sum_{t_{(i)}}(d_{1i} - e_{1i})\right)^2}{\text{Var}(d_{1i} - e_{1i})}$$

The variance formula is derived from a hypergeometric distribution which models the probability of d_{1i} group 1 failures in $d_{0i} + d_{1i}$ random draws when the size of group 1 is n_{1i} and the total at risk is $n_{1i} + n_{0i}$.

$$\text{Var}(d_{1i} - e_{1i}) = \sum_{t_{(i)}} \frac{n_{0i}n_{1i}(d_{0i} + d_{1i})(n_{0i} + n_{1i} - d_{0i} - d_{1i})}{(n_{0i} + n_{1i})^2(n_{0i} + n_{1i} - 1)}$$

Key Result!

When H_0 is true,

$$\text{Logrank Statistic} = \frac{\left(\sum_{t_{(i)}}(d_{1i} - e_{1i})\right)^2}{\text{Var}(d_{1i} - e_{1i})} \sim \chi^2_1$$

Hence, one can compare the logrank test statistic to the quantiles of a chi-square distribution with DOF=1.

If the statistic exceeds the $(1-\alpha)$ quantile, we can reject the null hypothesis H_0 at level α, and claim the survival curves are significantly different!

Logrank Test: in R

```
library(survival)
tt<-c(6,7,10,15,19,25)
delta<-c(1,0,1,1,0,1)
trt<-c(0,0,1,0,1,1)
survdiff(Surv(tt,delta)~trt)
```

Some Extensions

Look closely at the Logrank statistic

$$\text{Logrank Statistic} = \frac{\left(\sum_{t_{(i)}} (d_{1i} - e_{1i})\right)^2}{\text{Var}(d_{1i} - e_{1i})} \sim \chi_1^2$$

Each event time is weighted equally. We can generalize to include weights that treat failure times differently

$$\text{New Statistic} = \frac{\left(\sum_{t_{(i)}} w(i)(d_{1i} - e_{1i})\right)^2}{\text{Var}(w(i)(d_{1i} - e_{1i}))} \sim \chi_1^2$$

Some Extensions

Name	$w(i)$
Logrank	1
Wilcoxon-Breslow	n_i
Tarone-Ware	$\sqrt{n_i}$
Peto	$\tilde{s}(t_{(i)})$
Fleming-Harrison	$\hat{S}(t_{i-1})^p[1 - \hat{S}(t_{i-1})]^q$

Key Takeaways

1. Logrank test weights each failure time equally, Wilcoxon-Breslow/Traone-Ware/Peto weight earlier survival times more, Fleming-Harrison offers flexibility.
2. “Best test” is the one with the most power – where do you expect the survival curves to be most different?
3. Choice **MUST** be made *a priori* to avoid p-hacking.

Logrank Test Variants: in R

```
library(survival)
tt<-c(6,7,10,15,19,25)
delta<-c(1,0,1,1,0,1)
trt<-c(0,0,1,0,1,1)
survdiff(Surv(tt,delta)~trt, rho=0) #Logrank
survdiff(Surv(tt,delta)~trt, rho=1) # Peto-Prentice
```

Summary

1. Nonparametric survival methods are often preferred because they avoid making unnecessary assumptions which can invalidate inference.
2. The Kaplan-Meier estimator is the most common estimator of the survival curve.
3. Nonparametric estimators of the the hazard function require smoothing to account for noisy data.
4. The Logrank test and its variants are nonparametric tests of the equality of survival distributions.