2. Nonparametric analysis of survival data

September 1 - 25, 2015

Outline

- An introductory example (in discrete time)
  - Conditional survival probabilities
  - Estimation of (conditional) survival probabilities
- Life table methods
- Kaplan-Meier estimate of the survival function
- Nelson-Aalen estimate of the cumulative hazard
- Log-rank test

Why survival analysis?

In describing the distribution of *failure/life times*, special attention is needed because

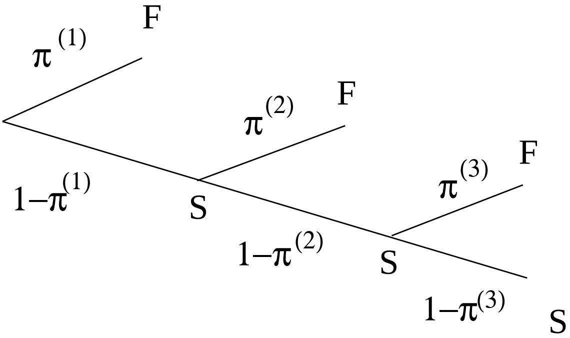
- failure times may be *censored*
  - the individual leaves the study cohort
    (lives until the end of follow-up, migrates, quits,...)
  - the individual may leave the study cohort for another event than what is being studied (competing risks)
- different studies are difficult to compare, if their follow-up times differ
- model specification and interpretation are often more convenient in terms of *conditional* failure rates

An introductory example

Break the total follow-up period into shorter time intervals (*bands*)

From a follow-up of one individual over three consecutive bands (next slide), there are four possible observations:

- failure (F) during the 1st band
- failure during the 2nd band
- failure during the 3rd band
- survival (S) until the end of follow-up



F
$\pi^{(1)}$
F
$\pi^{(2)}$
F
$\pi^{(3)}$
$1-\pi^{(1)}$
S
$1-\pi^{(2)}$
S
$1-\pi^{(3)}$
S

Conditional probabilities of failure

The three consecutive Bernoulli trials are described in terms of (three) conditional probabilities of failure:

- probability $\pi^{(1)}$ of failure during the 1st band
- probability $\pi^{(2)}$ of failure during the 2nd band, given survival until the end of the 1st band
- probability $\pi^{(3)}$ of failure during the 3rd band, given survival until the the end of the 2nd band

Unconditional probabilities

The probabilities of the three failure outcomes can be expressed in terms of conditional failure probabilities:

$$\pi^{(1)}$$
$$(1-\pi^{(1)})\pi^{(2)}$$
$$(1-\pi^{(1)})(1-\pi^{(2)})\pi^{(3)}$$

In general, P(failure during band i) $= \pi^{(i)} \prod_{j=1}^{i-1}(1-\pi^{(j)})$

In addition, the probability of surviving the entire follow-up can be calculated as

$$\prod_{j=1}^{i}(1-\pi^{(j)})$$

Cumulative survival probabilities

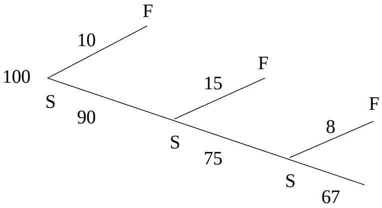
The probabilities to survive, i.e, to escape failure, up to the end of each time band:

$$
\begin{aligned}
&(1-\pi^{(1)})\\
&(1-\pi^{(1)})(1-\pi^{(2)})\\
&(1-\pi^{(1)})(1-\pi^{(2)})(1-\pi^{(3)})
\end{aligned}
$$

$$\mathrm{P}(\text{escape failure up to the end of band } i) = \prod_{j=1}^{i}(1-\pi^{(j)})$$

Estimation of conditional probabilities

Assume we have followed $N = 100$ individuals over three time bands:



The likelihood for conditional probabilities:

$$
\begin{aligned}
& \log L(\pi^{(1)}, \pi^{(2)}, \pi^{(3)}) \\
= \; & 10 \log(\pi^{(1)}) + 90 \log(1 - \pi^{(1)}) \\
& + 15 \log(\pi^{(2)}) + 75 \log(1 - \pi^{(2)}) \\
& + 8 \log(\pi^{(3)}) + 67 \log(1 - \pi^{(3)})
\end{aligned}
$$

Estimation of conditional probabilities cont.

- This is equivalent to the likelihood from three (conditionally) independent Bernoulli trials
- The maximum likelihood estimates are easily found to be:

$$\hat{\pi}^{(1)} = 10/100,\ \hat{\pi}^{(2)} = 15/90,\ \hat{\pi}^{(3)} = 8/75$$

Important lessions

- The *unit of observation* is *one individual's "experience" over one time band*
- A sufficient summary of data is the size of *risk set* Y_i and the number of failures D_i from each time band i
  - *The risk set* at a given time band includes all individuals still in the follow-up, that is, those that have until that
- The likelihood for conditional probabilities:
  $\log L = \sum_i \log L(\pi^{(i)})$, where
  $\log L(\pi^{(i)}) = D_i \log(\pi^{(i)}) + (Y_i - D_i) \log(1 - \pi^{(i)})$
- The maximum likelihood estimates are $\hat{\pi}^{(i)} = D_i / Y_i$

Survival function based on life tables

- When only grouped failure times are available, cencorings can be taken to occur sometime during the band.
- Assume that for band $t_{i-1} \leq t < t_i$ the observations are (Y_i, D_i, L_i), where

$$
\begin{aligned}
D_i &= \text{number of failures during time band} i \\
L_i &= \text{number of censorings during time band} i \\
Y_i &= \text{the size of the risk set at the beginning} \\
&\quad \text{of time band} i
\end{aligned}
$$

- To estimate cumulative survival, the size of the risk set at band i is taken to be

$$\begin{aligned} R_i &= Y_i - 0.5 * L_i \\ &= N - \sum_{j=0}^{i-1} D_j - \sum_{j=0}^{i-1} L_j - 0.5 L_i \end{aligned}$$

- We have thus assumed that a half of censorings took place at the beginning and another half at the end of the interval.
- The following table presents life times since diagnosis in two cancer treatment groups:

Example: two cancer treatments

Year t_i	Group I			Group II		
	Y_i	D_i	L_i	Y_i	D_i	L_i
1	110	5	5	234	24	3
2	100	7	7	207	27	11
3	86	7	7	169	31	9
4	72	3	8	129	17	7
5	61	0	7	105	6	13
6	54	2	10	85	6	6
7	42	3	6	73	5	6
8	33	0	5	62	3	10
9	28	0	4	49	2	13
10	24	1	8	34	4	6

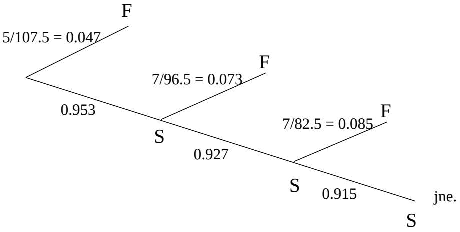
- For each time interval (year since diagnosis):
  - conditional survival probabilities

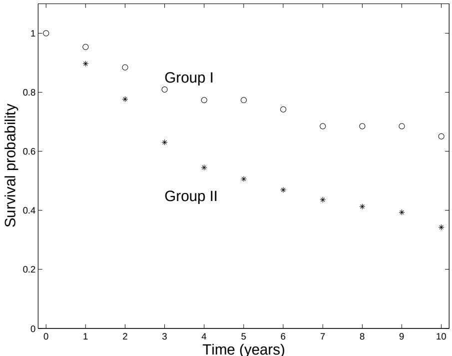
$$\mathrm{P}(T > t_i | T > t_{i-1}) = 1 - D_i/R_i$$

  - survival function

$$S(t_i) = \mathrm{P}(T > t_i) = \prod_{j=1}^{i}(1 - D_j/R_j)$$

Example cont.





The Kaplan-Meier estimate

- assume that exact failure times t_j are known
- break the follow-up period into bands so that each contains at most one time of failure*
- let the length of time bands go to zero $\Rightarrow$ survival function as function of time:

$$S(t) = \mathrm{P}(T > t) = \prod_{i;t_i \leq t} (1 - D_i/Y_i)$$

* There can be several failures and/or censorings at the same time. Censorings are assumed to take place after failures.

The risk set

- The size of the risk set at time t_i is

$$Y_i = N - \sum_{i=0}^{i-1} D_j - \sum_{j=0}^{i-1} L_j$$

$$\begin{aligned} D_j &= \text{number of failures at time} t_j \\ L_j &= \text{number of censorings at time} t_j \end{aligned}$$

- The conditional survival probabilities are now:

$$\mathrm{P}(T > t_i | T > t_{i-1}) = 1 - D_i / Y_i$$

- and the survival function:

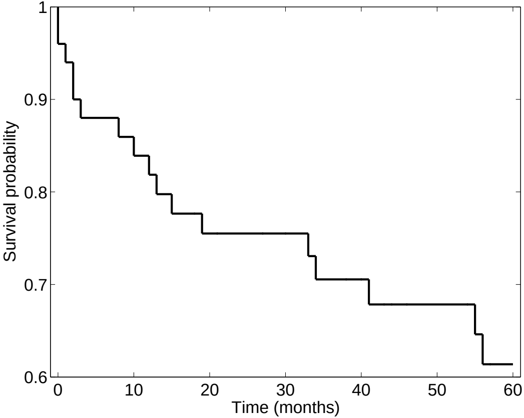
$$S(t_i) = \mathrm{P}(T > t_i) = \prod_{j=1}^{i} (1 - D_j / Y_j)$$

Example

t_i	Y_i	D_i	L_i	D_i/Y_i	$1 - D_i/Y_i$	$\mathrm{P}(T > t_i)$
0	50	2	0	0.0400	0.9600	0.9600
1	48	1	0	0.0208	0.9792	0.9400
2	47	2	0	0.0426	0.9574	0.9000
3	45	1	1	0.0222	0.9778	0.8800
8	43	1	0	0.0233	0.9767	0.8595
10	42	1	0	0.0238	0.9762	0.8391
...	...	...	...	...	...	...

Example cont.

A Kaplan-Meier estimate of the survival function



Properties of the KM estimate

- piecewise constant
- *non-parametric*
- jumps at the observation times only
- if no censoring at all, the size of the jump is d_j/N
- the precision of the estimate is poor towards the end of the follow-up
- confidence limits can be derived (next slide)

Confidence limits

Considering observations at each failure as binomial experiments ("drawing failures from the risk set"), one can derive the following standard deviation for $\Lambda(t)$ at the kth failure time:

$$\text{s.e.}(\Lambda_k) = \sqrt{\frac{D_1}{Y_1(Y_1 - D_1)} + \ldots + \frac{D_k}{Y_k(Y_k - D_k)}}$$

The so called Greenwood formula of standard error of the survival function at the kth failure time then is

$$\text{s.e.}(\Lambda_k) \times (S(t_k))^2$$

This can be easily calculated by the *survfit* function in R.

Hazard rate

- *The hazard* is the rate of change of the conditional failure probability:

$$\lambda(t) = \lim_{h \to 0} \frac{\mathsf{P}(T \in [t, t+h[|T \geq t)}{h}$$

- Assuming the h is short, the conditional failure probability over the time interval $[t, t+h[$, given survival until t, is

$$\pi_t = \mathsf{P}(T \in [t, t+h[|T \geq t) \simeq \lambda(t)h$$

The hazard function has many names and uses:

- *incidence rate* or *incidence density*
- *force of mortality*
- *force of morbidity*
- *force of infection*
- *...*

Going to the limit

Assume first that the hazard is constant in time. Based on the experience of one individual, when

- failure occurs at time t_i
- the individual's follow-up period $[0, t_i]$ is divided into M *"clicks"*
- the length of each click is h
- h goes to zero so that $Mh = t_i$ remains constant

$$\log L_i(\lambda)$$

$$= \log \left[\pi(1-\pi)^{M-1}\right] \simeq \log \left[(\lambda h)(1-\lambda h)^{M-1}\right]$$

$$= \log \lambda + \log h + (M-1)\log(1-\lambda h)$$

$$= \log \lambda + (M-1)\log(1-\lambda h) + \text{const.}$$

$$\to \log \lambda - M\lambda h + \text{const.} = \log \lambda - \lambda t_i + \text{const.}$$

as $h \to 0$.

The likelihood contribution from the observation on individual i failing at time t_i thus is

$$L_i(\lambda) = S(t_i) = f(t_i) = \overbrace{\lambda \exp(-\lambda t_i)}^{\text{probability density}} \quad .$$

Censored observations

Likewise, if the individual's failure is censored at time t_i,

$$\begin{aligned} \log L_i(\lambda) &= \log (1-\lambda h)^M \\ &= M \log (1-\lambda h) \to -\lambda M h = -\lambda t_i \end{aligned}$$

Thus, the likelihood contribution from the observation on individual i being censored at time t_i is

$$L_i(\lambda) = \exp(-\lambda t_i) \equiv S(t_i).$$

The survival likelihood

We can now construct the likelihood for a constant hazard λ, based on the follow-up of a study cohort with

- observed failure times $t_i, \; i = 1, ..., N$
- failure indicators $d_i, \; i = 1, ..., N$

The likelihood based on these observations is a product over individual contributions $L_i(\lambda)$:

$$\begin{aligned} L(\lambda) &= \prod_{i=1}^N L_i(\lambda) = \prod_{i=1}^N \lambda^{d_i} S(t_i) \\ &= \lambda^D \exp(-\lambda \sum_{i=1}^N t_i) = \lambda^D \exp(-\lambda Y) \end{aligned}$$

The maximum likelihood estimate:

$$\frac{dL}{d\lambda} = \left(D\lambda^{D-1} - \lambda^D Y\right)\exp(-\lambda Y) = 0$$

$$\Rightarrow \hat{\lambda} = D/Y = \frac{\text{number of failures}}{\text{person-time}}$$

In the example of the previous Figure, a sufficient data summary is: $D = 2$ (failures) ja $Y = 1.8$ (years of person-time). We obtain $\hat{\lambda} = 2/1.8 = 1.11$.

Interpretation of the hazard function

- Concerns one *individual* (cf. *risk*)
- $\lambda(u) \geq 0$, it is not bounded from above,
- So, it is not a probability but a rate
- Can be scaled apppropriately. For example, for a constant hazard the following expressions are equivalent:
  - 0.05/person/year
    = 0.0042/person/month
    = 5000/100000 person/year

N.B. Absolute incidence rates

In a large population, one can determine the (absolute) incidence:

- $N(t)\lambda$, where $N(t)$ is the risk set at time t
- if the population is *open* and stationary so that the size of the risk set stays constant, the incidence is $N\lambda$.
- the expected number of failures occurring from time 0 to time t is $N\lambda t$

Nelson–Aalen estimate

- In general, the cumulative hazard is defined as $\Lambda(t) = \int_0^t (u)du$
- An estimate can be calculated as follows
  - $\Lambda(t)$ jumps upwards at failure times t_j
  - the size of the jump is

$$\hat{\lambda}^{(j)}h = \left(\frac{D_j}{Y_j h}\right) h = D_j/Y_j$$

when Y_j is the size of the risk set and D_j the number of failures at t_j. We obtain (t) = $\sum_{j;t_j \leq t} \frac{D_j}{Y_j}$

There is a close relation between the Kaplan-Meier and Nelson–Aalen estimates: when $D_j/Y_j \simeq 0$,

$$\exp(-\overbrace{\sum_{j;t_j \le t} D_j/Y_j}^{Nelson-Aalen})$$

$$= \prod_{j;t_j \le t} \exp(-D_j/Y_j) \simeq \overbrace{\prod_{j;t_j \le t} (1 - D_j/Y_j)}^{Kaplan-Meier}$$

Comparison of two survival curves

For each failure time t_i, compare the expected and observed numbers of failures in the two groups. For time t_i, the data are

	Group 1	Group 2	Total
Failures	D_{1i}	D_{2i}	D_i
At risk	Y_{1i}	Y_{2i}	Y_i

Log-rank test

- Given that there were $D_i = 1$ failures, and assuming that survival is equal in the two groups, the expected number of failures in group j at time t_i is $\pi_{ji} = Y_{ji}/Y_i$.
- The expected total numbers are $E_j = \sum_i \pi_{ji}$, $j = 1, 2$. The log-rank test compares these to the observed numbers of failures $O_j = \sum_i D_{ji}$:

$$\frac{(E_1 - O_1)^2}{E_1} + \frac{(E_2 - O_2)^2}{E_2}$$

- This has a χ^2 distribution from which P values can be calculated