Statistical methods in public health
Analyzing time to event

Tommi Härkänen

National Institute for Health and Welfare (THL)
Department of Health (TERO)

September 8, 2015

Contents

Challenges in analyzing time as an outcome

Follow-up time: Subjects can have different time lengths until the outcome of interest occurs

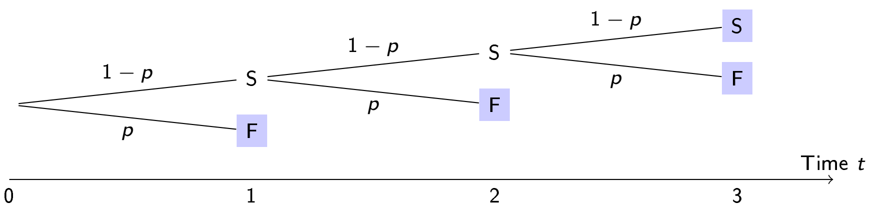
Right-censoring: Some subjects do not experience the outcome of interest before the end of the follow-up. Common reasons are
- follow-up ends at a specified (calendar) time
- follow-up ends due to another reason (e.g. death instead of cancer diagnosis)
- individual cannot be followed after some time point (e.g. emigration)

Time-dependent risk: Probability that "outcome occurs soon after time t after the baseline (assuming outcome did not occur before t)" depends on time t

Splitting the follow-up time

Consider a three-year follow-up data, and a split based on one-year intervals.
Possible outcomes can be
1. Failure during 1st year
2. Failure during 2nd year
3. Failure during 3rd year
4. Survival for the full three-year period

If the failure can occur only once, the failure during 2nd year (choice 2) is possible only if subject survived 1st year.



Conditional and unconditional probability of failure

In the example above, there was a (conditional) binary model for each follow-up year (until failure):

$$\mathbb{P}_p\{T = t \mid T \ge t\} = p, \ t = 1, 2, \ldots$$

where T is the year of failure. Let $Y_t \in \{S, F\}$ denote the binary random variable for failure in year t. If $T = t$, then $Y_1 = \cdots = Y_{t-1} = 0$ and $Y_t = 1$.
Unconditional failure probability is

$$\begin{aligned}
\mathbb{P}_p\{T=1\} &= \mathbb{P}_p\{Y_1=1\} = \mathbb{P}_p\{T=1 \mid T \ge 1\} = p \\
\mathbb{P}_p\{T=2\} &= \mathbb{P}_p\{Y_1=0,\ Y_2=1\} \\
&= \mathbb{P}_p\{Y_2=1 \mid Y_1=0\}\,\mathbb{P}_p\{Y_1=0\} = (1-p)p \\
&\vdots \\
\mathbb{P}_p\{T=t\} &= \mathbb{P}_p\{Y_1=0,\ Y_2=0,\ \ldots,\ Y_{t-1}=0,\ Y_t=1\} \\
&= \mathbb{P}_p\{Y_t=1 \mid Y_1=\cdots=Y_{t-1}=0\} \times \cdots \\
&\quad \times \mathbb{P}_p\{Y_2=0 \mid Y_1=0\}\,\mathbb{P}_p\{Y_1=0\} \\
&= (1-p)^{t-1}p
\end{aligned} \tag{1}$$

Terms in (1) are the likelihood terms $L(p;\, t) = \mathbb{P}_p\{T = t\}$ for subjects whose failure times t are observed during the follow-up.

Time-dependent failure probability

Failure probability can depend on the time band t:

$$\mathbb{P}_{p_t}\{T = t \mid T \ge t\} = p_t, \ t = 1, 2, \ldots$$

and the unconditional probability (1) becomes

$$\begin{aligned}
\mathbb{P}_p\{T=t\} &= \\
&= \mathbb{P}_{p_t}\{Y_t=1 \mid Y_1=\cdots=Y_{t-1}=0\} \times \cdots \\
&\quad \times \mathbb{P}_{p_2}\{Y_2=1 \mid Y_1=0\}\,\mathbb{P}_{p_1}\{Y_1=0\} \\
&= \prod_{s=1}^{t-1}(1-p_s)p_t.
\end{aligned} \tag{3}$$

The maximum likelihood estimate of failure probability in time band t is

$$\hat{p}_t = \frac{d_t}{m_t}, \tag{4}$$

where $d_t = \sum_i \mathbf{1}\{T_i = t\}$ is the number of failures and $m_t = \sum_i \mathbf{1}\{T_i \ge t\}$ is the *size of the risk set* in the beginning of time band t.

Survival function and right censoring

Survival function value at time t is the probability of survival longer than time t :

$$S_p(t) := \mathbb{P}_p\{T > t\} = 1 - \mathbb{P}_p\{T \le t\} = 1 - \sum_{s=1}^{t}\mathbb{P}_p\{T = t\} = (1-p)^t. \tag{2}$$

Subjects who were *right-censored*, that is, had not failed before end of follow-up time, say $u > 0$, have likelihood terms $L(p;\, u) = S_p(t)$.

Likelihood for subjects $i = 1, 2, \ldots, n$ with observations (t_i, δ_i) is

$$L(p;\, (t_i, \delta_i)_i) = \prod_{i=1}^{n} p^{\delta_i} \prod_{s=1}^{t_i - \delta_i}(1-p)$$

where $\delta_i \in \{0, 1\}$ is the *censoring indicator* with values

- 0 Subject i was right-censored at time t_i
- 1 Subject i failed at time t_i

The Kaplan-Meier estimator

Nonparametric estimate of survival function

What happens when the time bands become more and more narrow?

- Fewer failures (eventually only one) occur during a single time band
- More time bands contain no failures

Recall the definition of survival function (2), and the maximum likelihood estimate of p_t in (4):

$$S(t) := \prod_j 1 - \frac{d_{t_j}}{m_{t_j}}. \tag{5}$$

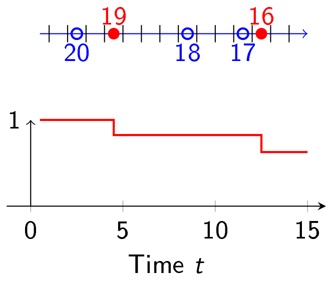
Terms in (5) equal 1 in bands with no failures.

The Kaplan-Meier estimator

Nonparametric estimate of survival function

The KM estimate drops at the failure times.
Step size at time t_j is $-d_{t_j}/m_{t_j}$

Right-censored failure times do not change the KM estimate.
(Right-censored failure time does influence the drop after the censoring as the size of risk set becomes smaller.)



The Kaplan-Meier estimator

Variance estimator

There are several variance estimators for KM. On of the most popular is based on *Greenwood's formula*:

$$\widehat{\mathrm{Var}}\left[\widehat{S(t)}\right] = \widehat{S(t)}^2 \sum_{i:\, t_i \le t} \frac{d_i}{m_i(m_i - d_i)}. \tag{6}$$

(6) is based on

1. log-transformation of $\widehat{S(t)}$
   traditional Greenwood formula was $f(t) = \log t$
   exponential Greenwood formula was $f(t) = \log(-\log t)$
2. *delta method* and
3. martingales (terms in (5) are not independent).

Hazard rate

Divide the follow-up time into short bands (as in the Kaplan-Meier estimator case) of length h (constant).

The shorter time band, the smaller probability of failure p. Assume that the probability of having two or more failures during one band is (very) small.

Reparameterize $p =: \lambda h$, where λ is called the *hazard rate* or *probability rate* or *instantaneous probability rate* or *force of mortality* or ....

Hazard rate

Poisson likelihood

The likelihood terms for binary model and probability rate λ:

$$L(\lambda;\, (t_i,\, \delta_i)_i) = \begin{cases} p\prod_{s=1}^{t_i/h-1}(1-p) & = \lambda h \prod_{s=1}^{t_i/h-1}(1-\lambda h), \quad \delta_i = 1 \\ \prod_{s=1}^{t_i/h}(1-p) & = \prod_{s=1}^{t_i/h}(1-\lambda h), \qquad \delta_i = 0 \end{cases} \tag{7}$$

Recall that for x close to zero $1 - x \approx \exp\{-x\}$.

It follows that $\prod_s(1-\lambda h) \approx \prod_s \exp\{-\lambda h\} = \exp\{-\sum_s \lambda h\}$, and we get the *Poisson likelihood*:

$$L(\lambda;\, t_i,\, \delta_i) = (\lambda h)^{\delta_i} \exp\{-t_j \lambda\}. \tag{8}$$

Hazard rate

Maximum likelihood estimate of λ and distribution of failure time

It is easy to see that

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i} = \frac{\text{Total number of failures}}{\text{Total observation time}}.$$

Survival function $S(\cdot)$ and density function $f_\lambda(\cdot)$ are

$$\begin{aligned} S(t) &= \exp\{-\lambda t\} \\ f_\lambda(t) &= \lambda \exp\{-\lambda t\}. \end{aligned} \tag{9}$$

Note that (9) correspond to *exponential distribution* with expectation $1/\lambda$ and variance $1/\lambda^2$.

Hazard rate

Time-dependent hazard rate

It may be unrealistic to assume that the hazard rate is constant over a (long) period of time.
A solution: Divide the follow-up time into time bands $(u_k, u_{k+1}]$ within which the hazard rate λ_k is constant.
E.g. follow-up time 15 years are divided into 5-year bands: