# Estimation of the survival function

Rasmus Waagepetersen
Department of Mathematics
Aalborg University
Denmark

September 12, 2024

# Estimation of the survival function - actuarial estimate

Suppose we are given data in terms of a lifetable for a population.

That is, for fixed times $0 = u_0 < u_1 < u_2 < \cdots$ we know for each $u_i$

- the number $r(u_i)$ of individuals *at risk* (not dead or censored) at time $u_i$ (i.e. both survival time and censoring time $\geq u_i$)
- the number of deaths $d_i$ in the interval $[u_{i-1}; u_i[$ and
- the number $c_i$ of censorings in $[u_{i-1}; u_i[$.

Note: $r(u_i) = r(u_{i-1}) - d_i - c_i$ and initial population size $n = r(u_0)$

We want to estimate $P(X \geq u_i)$

Usual estimate:

$$\hat{P}(X \geq u_i) = \frac{\#\text{alive up to time } u_i}{n}$$

If no censoring:

$$\hat{P}(X \geq u_i) = \frac{r(u_i)}{n}$$

Problem: due to censoring we often do not know numerator - typically larger than $r(u_i)$ ! (individuals censored prior to $u_i$ may well be alive)

# Factorization

$$P(X \geq u_l) = \prod_{k=1}^{l} P(X \geq u_k | X \geq u_{k-1}) =$$

$$\prod_{k=1}^{l}(1 - P(X < u_k | X \geq u_{k-1})) = \prod_{k=1}^{l}(1 - p_k)$$

Here $p_k$ is the probability of dying in the $k$th interval given alive at start of interval.

Suppose we obtain estimate $\hat{p}_k$. Then resulting estimate of $P(X \geq u_l)$ is

$$\hat{P}(X \geq u_l) = \prod_{k=1}^{l}(1 - \hat{p}_k)$$

# Estimation of $p_k$

Immediate idea:

$$\hat{p}_k = \frac{d_k}{r(u_{k-1})}$$

Requirement: individuals contributing to the denominator must be representative of those alive at time $u_{k-1}$. Thus the probability that a person dies in $[u_{k-1}; u_k[$ given that the person is at risk (not dead or censored) at time $u_{k-1}$ must coincide with $p_k$.

This is what we called independent censoring (or non-informative censoring in KM, unfortunately terminology is not consistent over text books)

OK if each $c_k$ represents a random sample of the $r(u_{k-1})$ persons at risk. Problematic if persons are censored because they appear very weak at time $u_{k-1}$.

In case of censoring we still have problem: if $c_k > 0$, numerator in

$$\hat{p}_k = \frac{d_k}{r(u_{k-1})}$$

may be too small. Adding $c_k$ to $d_k$ would not work - since not likely that all censored persons died in $[u_{k-1}, u_k[$. Instead we adjust the denominator.

Suppose all censoring takes place in the very beginning of the $k$th interval at time $u_{k-1}$. Then the effective number at risk in the $k$th interval is $r(u_{k-1}) - c_k$ and we let

$$\hat{p}_k = \frac{d_k}{r(u_{k-1}) - c_k}$$

If all censoring takes place at the very end of the interval then

$$\hat{p}_k = \frac{d_k}{r(u_{k-1})}$$

If the censoring times are uniformly dispersed on the interval then a censored individual is at risk on average half of the interval and we use

$$\hat{p}_k = \frac{d_k}{r(u_{k-1}) - c_k/2}$$

I.e. the so-called actuarial estimate - uses denominator given by average of previous denominators.

Note: $\hat{p}_k = 0$ if no deaths in the $k$th interval !

# Estimation using exact death times - reduced sample estimator

Suppose now we have observed the exact death or censoring times $(t_i, \delta_i)$ and we want to estimate $P(X > t)$ for an arbitrary $t$.

Suppose the censoring times $C_i$ are all observed and independent of the death times $X_i$ (e.g. type 1 censoring).

Unbiased reduced sample estimator:

$$\hat{S}_{\text{red}}(t) = \frac{\sum_{i=1}^{n} 1[x_i > t, c_i > t]}{\sum_{i=1}^{n} 1[c_i > t]}$$

Problem: inefficient use of observations. An observation censored at time $u$ does not contribute to $\hat{S}_{\text{red}}(t)$ for $t \geq u$.

Not applicable in case of competing risks when $C_i > t$ not observed if death happens prior to $t$.

Alternative idea: introduce discretization
$0 = u_1 < u_2 < \cdots < u_L = t$ and apply actuarial estimate.

Next consider limit $L \to \infty$ and $u_k - u_{k-1} \to 0$ (finer and finer discretization). Assume also that no censoring time coincides with a death time.

Let $D$ denote the set of distinct death times and let $d(t^*)$ denote the number of deaths at time $t^*$ for $t^* \in D$.

Then, for $L$ sufficiently large, there is a most one distinct death time in each interval and if there is a death time then there is no censoring.

Thus we have two possibilities $\hat{p}_k = 0$ (no death) or

$$\hat{p}_k = \frac{d(t^*)}{r(t^*)}$$

if $t^*$ is the unique death time falling in $[u_{k-1}; u_k[$.

Thus our estimate becomes

$$\hat{P}(X \geq t) = \prod_{\substack{t^* \in D: \\ t^* < t}} (1 - \frac{d(t^*)}{r(t^*)})$$

and

$$\hat{S}(t) = \hat{P}(X > t) = \prod_{\substack{t^* \in D: \\ t^* \leq t}} (1 - \frac{d(t^*)}{r(t^*)})$$

This is the Kaplan-Meier (product limit) estimate.

Estimate is right-continuous.

If last event, say $t_n$, is a death then $\hat{S}(t) = 0$ for $t \geq t_n$. If last event is a censoring then $\hat{S}(t) = \hat{S}(t_n) > 0$ for $t \geq t_n$.

# Nelson-Aalen estimator of cumulative hazard

$$H(t) = \int_0^t h(u)\mathrm{d}u \approx \sum_{k=1}^{L} h(u_{k-1})[u_k - u_{k-1}] \approx \sum_{k=1}^{L} p_k$$

Thus

$$\hat{H}(t) = \sum_{k=1}^{L} \hat{p}_k$$

In the limit (Nelson-Aalen estimator)

$$\hat{H}(t) = \sum_{\substack{t^* \in D: \\ t^* \leq t}} \frac{d(t^*)}{r(t^*)}$$

Recall $S(t) = \exp(-H(t))$. Estimates $\hat{H}(t)$ and $\hat{S}(t)$ related by $\log(1-x) \approx -x$ or $\exp(-x) \approx 1-x$ for $x$ close to 0.

# Asymptotic results

Consider the random censoring case where the $n$ survival and censoring times $X_i$ and $C_i$, $i = 1, \ldots, n$ have survival functions $S$ and $G$.

Consider any $0 < v < \infty$ with $S(v) > 0$, assume that $1 - S$ is absolute continuous with density $f$ and that $G$ is continuous. Then the random function

$$\sqrt{n}(\hat{S}(t) - S(t)), \quad 0 < t < v$$

converges in distribution to a zero mean Gaussian process $\{R(u)\}_{0<u<v}$ with covariance function

$$\mathbb{C}\mathrm{ov}(R(t_1), R(t_2)) = S(t_1)S(t_2) \int_0^{\min(t_1,t_2)} \frac{h(u)}{S(u)G(u)} \mathrm{d}u$$

(see e.g. Lawless, 1982).

# Implications of asymptotic result

For any $0 < t < v$:

$$\hat{S}(t) \approx N(S(t), \frac{\sigma_t^2}{n}) \text{ with } \sigma_t^2 = S(t)^2 \int_0^t \frac{h(u)}{S(u)G(u)} \mathrm{d}u$$

$\sqrt{n}$-consistency: for any fixed $c$,

$$P(\sqrt{n}|\hat{S}(t) - S(t)|/\sigma_t < c)$$

converges to $1 - 2\Phi(-c)$.

Loosely speaking, $\sqrt{n}(\hat{S}(t) - S(t))/\sigma_t$ is bounded with probability 1, thus $(\hat{S}(t) - S(t))$ converges to zero as $1/\sqrt{n}$.

95% Confidence interval (pointwise !):

$$\hat{S}(t) \pm 1.96\sigma_t/\sqrt{n}$$

# Estimation of asymptotic variance

In practice we need to estimate asymptotic variance $\sigma_t^2$:

$$\sigma_t^2 \approx S(t)^2 \sum_{k=1}^{L} \frac{h(u_{k-1})}{S(u_{k-1})G(u_{k-1})}(u_k - u_{k-1}) \approx S(t)^2 \sum_{k=1}^{L} \hat{p}_k \frac{n}{r(u_{k-1})}$$

Taking the limit $L \to \infty$ as before we obtain

$$\frac{\hat{\sigma}_t^2}{n} = \hat{S}(t)^2 \sum_{\substack{t^* \in D: \\ t^* \leq t}} \frac{d(t^*)}{r(t^*)} \frac{1}{r(t^*)}$$

Typically, the closely related Greenwoods formula is used:

$$\frac{\hat{\sigma}_t^2}{n} = \hat{S}(t)^2 \sum_{\substack{t^* \in D: \\ t^* \leq t}} \frac{d(t^*)}{r(t^*)} \frac{1}{r(t^*) - d(t^*)}$$

(recall: for $L$ sufficiently large $\hat{p}_k$ is either 0 or $d(t^*)/r(t^*)$ and in the latter case, $r(u_k) = r(t^*) - d(t^*)$)

Note: Greenwood's formula can be derived by heuristic arguments using

$$\hat{S}(t) = \prod_{k=1}^{L}(1 - \hat{p}_k) = g(\hat{p}_1, \dots, \hat{p}_L)$$

where $g(x_1, \dots, x_L) = \prod_{i=1}^{L}(1 - x_i)$ and the $\delta$-method.

We also assume $\hat{p}_k$ uncorrelated and estimate $\mathbb{V}\mathrm{ar}\hat{p}_k$ by

$$\hat{p}_k(1 - \hat{p}_k)/r(u_{k-1})$$

- see next slide.

# Some remarks on the $\hat{p}_k$

Consider for simplicity the case with no censoring. Let

$$N_k = \sum_{l=1}^{k} d_k$$

be counting process of deaths. $d_k = N_k - N_{k-1}$, $r(u_k) = n - N_k$.

Assume $d_k | N_1, \ldots, N_{k-1} \sim \text{bin}(r(u_{k-1}), p_k)$. Then

$$\mathbb{E}[\hat{p}_k - p_k | N_1, \ldots, N_{k-1}] = 0.$$

This implies $\mathbb{E}[\hat{p}_k] = E[E[\hat{p}_k | N_1, \ldots, N_{k-1}]] = p_k$ and for $k' > k$,

$$\mathbb{C}\text{ov}[\hat{p}_k, \hat{p}_{k'}] = \mathbb{E}[(\hat{p}_k - p_k)\mathbb{E}[\hat{p}_{k'} - p_{k'} | N_1, \ldots, N_{k'-1}]] = 0$$

Thus $\hat{p}_k$'s uncorrelated.

Moreover,

$$\mathbb{V}\mathrm{ar}\hat{p}_k = \mathbb{V}\mathrm{ar}[E[\hat{p}_k|N_1,\ldots,N_{k-1}] + \mathbb{E}\mathbb{V}\mathrm{ar}[\hat{p}_k|N_1,\ldots,N_{k-1}]] = 0 + \mathbb{E}[p_k(1-p_k)/r(u_{k-1})]$$

So we may estimate $\mathbb{V}\mathrm{ar}\hat{p}_k$ by

$$\hat{p}_k(1-\hat{p}_k)/r(u_{k-1})$$

Note $M_k = N_k - \sum_{l=1}^{k} p_l r(u_{l-1})$ is a martingale with respect to 'history' $N_1,\ldots,N_{k-1}$:

$$\mathbb{E}[M_k|N_1,\ldots,N_{k-1}] = M_{k-1} + \mathbb{E}[d_k - r(u_{k-1})p_k|r(u_{k-1})] = M_{k-1}$$

This implies uncorrelated increments $M_k - M_{k-1}$.

$M_k$ is centered/compensated version of $N_k$:

$$\mathbb{E}[M_k] = \mathbb{E}[M_{k-1}] = \ldots = \mathbb{E}[M_1] = 0$$

# Confidence intervals

Issues: $0 \leq S(t) \leq 1$. This is not respected by previously mentioned confidence intervals.

KM discusses various solutions including deriving confidence interval based on transformed $S(t)$ and transforming back.

KM section 4.4 also discusses simultaneous confidence bands.

# $\log(-\log(\cdot))$-transformation

$$L(t) = \log(H(t)) = \log(-\log(S(t))$$

is a function on $\mathbb{R}$ (unrestricted). Let $\hat{L}(t) = \log(-\log(\hat{S}(t))$ with standard error $\sigma_L$. Then approximate 95% confidence interval for $L(t)$ is

$$[\hat{L}(t) - 2\sigma_L; \hat{L}(t) + 2\sigma_L].$$

Transforming back we obtain approximate 95% interval for $S(t)$:

$$[(\hat{S}(t))^{\exp(-2\sigma_L)}; (\hat{S}(t))^{\exp(+2\sigma_L)}].$$

Finally, by $\delta$-method,

$$\sigma_L \approx \text{std.err}(\hat{S}(t))/(\log(\hat{S}(t))\hat{S}(t))$$

See KM (4.3.2).

# Log rank test

Non-parametric test for equality of survival distributions for two groups (e.g. different treatments) with hazard function $h_1$ and $h_2$.

I.e. null hypothesis is $H_0 : h_1(\cdot) = h_2(\cdot)$.

Use notation as for the Kaplan-Meier estimate:

- $D = D_1 \cup D_2$ where $D_1$ and $D_2$ are the sets of distinct death times for each group.
- $d_1(t^*)$ and $d_2(t^*)$ denote the deaths at time $t^* \in D$ in groups 1 and 2
- $r_1(t^*)$ and $r_2(t^*)$ denote the numbers at risk at time $t^* \in D$ in groups 1 and 2

# Heuristic derivation of log-rank test

For each $t^*$ we have $2 \times 2$ table:

$$
\begin{array}{c|cc}
r_1(t^*) & d_1(t^*) & r_1(t^*) - d_1(t^*) \\
r_2(t^*) & d_2(t^*) & r_2(t^*) - d_2(t^*) \\
\hline
r(t^*) & d(t^*) & r(t^*) - d(t^*)
\end{array}
$$

Conditional on $t^*$, $r_1(t^*)$ and $r_2(t^*)$ assume

$$d_i(t^*)|t^*, r_1(t^*), r_2(t^*) \sim \text{bin}(r_i(t^*), p_i(t^*))$$

and independent where $p_i(t^*) = h_i(t^*)\mathrm{d}t^*$, $i = 1, 2$

Under $H_0$, $d_1(t^*)|d(t^*), r_1(t^*), r_2(t^*)$ follows hypergeometric distribution (exercise) with mean and variance

$$e_1(t^*) = r_1(t^*)\frac{d(t^*)}{r(t^*)} \quad v_1(t^*) = \frac{r_1(t^*)r_2(t^*)(r(t^*) - d(t^*))d(t^*)}{r(t^*)^2(r(t^*) - 1)}$$

Note: this does not depend on the common unknown values of $h_1$ and $h_2$ !

Note: under the alternative $h_1(t^*) > h_2(t^*)$ we would expect $d_1(t^*) > e_1(t^*)$ - and vice versa

Log-rank test statistic

$$\frac{\sum_{t^* \in D}(d_1(t^*) - e_1(t^*))}{\sqrt{\sum_{t^* \in D} v_1(t^*)}}$$

Approximately $N(0, 1)$ under $H_0$.

- closely related to Fisher's exact test for contingency tables (conditioning on sufficient statistics under null hypothesis).
- same test statistic obtained with $d_2(t^*)$'s (symmetry).
- weak test if we do not have either $h_1(\cdot) > h_2(\cdot)$ or $h_1(\cdot) < h_2(\cdot)$.
- test is *non-parametric* since it does not involve any assumptions regarding individual shapes of $h_1$ and $h_2$.

Implemented in the `R` `survdiff()` procedure.

KM Section 7.3 gives further details.