Conference in Survival Analysis

2. Non-Parametric Inference for Survival Features and Kaplan-Meier Estimator

Jiajun Zhang

July 4, 2025

Content of this report

- A brief introduction to non-parametric inference: Empirical estimators and convergence;

Content of this report

- A brief introduction to non-parametric inference: Empirical estimators and convergence;
- Kaplan-Meier estimator for the Survival function and estimate for its variance: Greenwood's formula;

Content of this report

- A brief introduction to non-parametric inference: Empirical estimators and convergence;
- Kaplan-Meier estimator for the Survival function and estimate for its variance: Greenwood's formula;
- Nelson-Aalen estimator for the cumulative hazard function.

Content of this report

- A brief introduction to non-parametric inference: Empirical estimators and convergence;
- Kaplan-Meier estimator for the Survival function and estimate for its variance: Greenwood's formula;
- Nelson-Aalen estimator for the cumulative hazard function.
- A multi-state model and Aalen-Johansen estimator.

Empirical Estimators

In our last talk, we introduced likelihood inference for parametric families. But now we have a non-parametric family, where we do not know the form of the distribution function. In non-parametric inference, all we can do is to use our observed data and build our estimators based on those known information. An example is that if we have observed the data $x_1, \cdots, x_n$, then we may simply use $\bar{x}_n = \frac{x_1+\cdots+x_n}{n}$ to estimate its mean.

Empirical Estimators

Similarly, if we want to estimate the distribution function $F(x) = \mathbb{P}(X \leq x)$ at some point x, we may proceed our estimate by counting the proportion of individuals that the event time is less than or equal to x. We can put

$$\widehat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \leq x\} \tag{0.1}$$

as our estimate, and the observed sample is $x_1, \cdots, x_n$. The equation (0.1) is our empirical estimator.

Empirical Estimators

Now we may derive the expected value and the variance of the empirical estimator:

$$\mathbb{E}(\widehat{F}_n(x)) = \frac{1}{n} \cdot \sum_{i=1}^{n} \mathbb{E}[1(X_i \leq x)] \tag{0.2}$$

$$= \frac{1}{n} \cdot \sum_{i=1}^{n} [1 \cdot \mathbb{P}(1(X_i \leq x)) + 0 \cdot \mathbb{P}(1(X_i \leq x))] \tag{0.3}$$

$$= \frac{1}{n} \cdot \sum_{i=1}^{n} \mathbb{P}(X_i \leq x) \tag{0.4}$$

$$= \frac{1}{n} \cdot \sum_{i=1}^{n} F(x) \tag{0.5}$$

$$= F(x) \tag{0.6}$$

Empirical Estimators

The variance can be derived by

$$\mathbb{V}(\widehat{F}_n(x)) = \mathbb{E}(\widehat{F}_n^2(x)) - [\mathbb{E}(\widehat{F}_n(x))]^2 \tag{0.7}$$

$$= \mathbb{E}\left(\frac{1}{n}\sum_{j=1}^{n} 1(X_j \leq x)\right)^2 - F^2(x) \tag{0.8}$$

$$= \frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} 1(X_i \leq x)1(X_j \leq x)\right) - F^2(x) \tag{0.9}$$

$$= \frac{1}{n^2}\mathbb{E}\left(\sum_{i=j} 1^2(X_i \leq x) + \sum_{i\neq j} 1(X_i \leq x)1(X_j \leq x)\right) - F^2(x) \tag{0.10}$$

$$= \frac{1}{n^2}\left(nF(x) + n(n-1)F^2(x)\right) - F^2(x) \tag{0.11}$$

Empirical Estimators

Summary:

Theorem

The expected value and the variance of the empirical estimator are given by:

$$\mathbb{E}\widehat{F}_n(x) = F(x), \mathbb{V}(\widehat{F}_n(x)) = \frac{1}{n}F(x) - \frac{1}{n}F^2(x) \tag{0.12}$$

Empirical Estimators

How do we know the relation between $F(x)$ and its estimate $\widehat{F}_n(x)$?

Empirical Estimators

How do we know the relation between $F(x)$ and its estimate $\widehat{F}_n(x)$?

Theorem

(Glivenko-Cantelli Theorem) We have

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} |F(x) - \widehat{F}_n(x)| \overset{a.s}{\to} 0. \tag{0.13}$$

Towards Survival Analysis

Then, we need to consider potential censoring and truncation when constructing our non-parametric inference.

Kaplan-Meier Estimator

Now, suppose we have observed an ordered failure time $t_1, \cdots, t_m$ from our sample $X_1 . \cdots, X_n$ where $m \leq n$. Denote by d_j, the number of death at each observed time t_j, and Y_j, as the number of individuals at risk (i.e did not die yet) just prior to t_j.

Kaplan-Meier Estimator

Now, suppose we have observed an ordered failure time $t_1, \cdots, t_m$ from our sample $X_1 . \cdots, X_n$ where $m \leq n$. Denote by d_j, the number of death at each observed time t_j, and Y_j, as the number of individuals at risk (i.e did not die yet) just prior to t_j.

Kaplan-Meier estimator can be build from an argument of conditional probability.

Kaplan-Meier Estimator

Now, suppose we have observed an ordered failure time $t_1, \cdots, t_m$ from our sample $X_1 . \cdots, X_n$ where $m \leq n$. Denote by d_j, the number of death at each observed time t_j, and Y_j, as the number of individuals at risk (i.e did not die yet) just prior to t_j.

Kaplan-Meier estimator can be build from an argument of conditional probability.

We first investigate the time interval $[0, t_1)$, since we know the first few deaths were recorded at t_1, so it means from our observed sample everyone was alive in this interval, and the survival function in this interval is constant 1.

Kaplan-Meier Estimator

Now consider the time interval $[t_1, t_2)$, the survival function in this interval can be estimated by

$$\begin{aligned} S(t) &= \mathbb{P}(X > t_1) \\ &= \frac{\text{Number of individual survive beyond } t_1}{\text{Number of individual at risk prior to } t_1} \\ &= \frac{Y_1 - d_1}{Y_1} \end{aligned}$$

Kaplan-Meier Estimator

For the next interval, to compute the survival probability, we are conditioning on our sample: Those enter our study already survived the previous interval. So in the time interval $[t_2, t_3)$, we have

$$\begin{aligned} S(t) &= \mathbb{P}(X > t_2) \\ &= \mathbb{P}(X > t_2 | X > t_1)\mathbb{P}(X > t_1) \\ &= \frac{\text{Number of individual survive beyond } t_2}{\text{Number of individual at risk prior to } t_2} \times \frac{Y_1 - d_1}{Y_1} \\ &= \frac{Y_2 - d_2}{Y_2} \times \frac{Y_1 - d_1}{Y_1} \end{aligned}$$

Kaplan-Meier Estimator

So in general, we give the Kaplan-Meier estimator for the survival function:

Theorem

(Kaplan-Meier estimator) The estimate of the survival function is given by

$$\widehat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{Y_j}\right). \tag{0.14}$$

Kaplan-Meier Estimator

So in general, we give the Kaplan-Meier estimator for the survival function:

Theorem

(Kaplan-Meier estimator) The estimate of the survival function is given by

$$\widehat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{Y_j}\right). \tag{0.14}$$

Kaplan-Meier estimator is also called the product-limit estimator.

Kaplan-Meier Estimator

So in general, we give the Kaplan-Meier estimator for the survival function:

Theorem

(Kaplan-Meier estimator) The estimate of the survival function is given by

$$\widehat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{Y_j}\right). \tag{0.14}$$

Kaplan-Meier estimator is also called the product-limit estimator.

Fun Fact: Kaplan's original paper on this estimator was the most cited mathematics paper in 20th century.

Kaplan-Meier Estimator

The term $\frac{d_j}{Y_j}$ is in fact another estimate of the hazard rate. Note the hazard rate here refers to a discrete hazard and not the one in continuous case we defined. It can be viewed as the probability that an individual belongs to Y_j will belong to d_j instantly after time t_j, which is,

$$\lambda(t_j) = \mathbb{P}(X = t_j | X > t_j^-) \tag{0.15}$$

in continuous case we make $X = t_j + dt$ but here we must make it exactly to be t_j since we only observe death at t_j. Given this setting, we propose a lemma:

Lemma

When the hazard $\lambda(t)$ is small, we have

$$S(t) = \prod_{t_i \leq t} (1 - \lambda(t_i)) \tag{0.16}$$

Proof.

Recall that in discrete case, we simply have

$$S(t) = \exp\left(-\sum_{t_j \le t} \lambda(t_j)\right) = \prod_{t_j \le t} \exp(-\lambda(t_j)) \tag{0.17}$$

and the proof could be done via an obvious Taylor expansion

$$e^x = 1 + x + O(2). \tag{0.18}$$

□

Nelson-Aalen Estimator

As you might see, in Kaplan-Meier estimator, we simply replaced $\lambda(t)$ in our lemma by $\frac{d_j}{Y_j}$. This is possible because we can further show $\frac{d_j}{Y_j}$ is the MLE of $\lambda(t)$! And by using $\frac{d_j}{Y_j}$, another estimator arises: Nelson-Aalen estimator.

Theorem

(Nelson-Aalen estimator) The estimate of the cumulative hazard function is given by

$$\widehat{\Lambda}(t) = \sum_{t_j \leq t} \frac{d_j}{Y_j} \tag{0.19}$$

Greenwood's Formula

Now we are interested in the expected value and the variance of our estimator.

- Kaplan-Meier estimator is *almost unbiased*, what I mean is that, if we make our assumption that $Y(t) \neq 0$, then it is unbiased. To formally proof this we need counting processes and martingale theory, which will be introduced in later talks.

Greenwood's Formula

Now we are interested in the expected value and the variance of our estimator.

- Kaplan-Meier estimator is *almost unbiased*, what I mean is that, if we make our assumption that $Y(t) \neq 0$, then it is unbiased. To formally proof this we need counting processes and martingale theory, which will be introduced in later talks.
- The variance of the Kaplan-Meier estimator can be estimated by Greenwood's formula and its adjustment complimentary log-log transformation.

Greenwood's Formula

Theorem

(Greenwod's Formula) The estimate of the variance of $\widehat{S}(t)$ is given by

$$\mathbb{V}(\widehat{S}(t)) = \widehat{S}^2(t) \cdot \sum_{t_j \leq t} \frac{d_j}{Y_j(Y_j - d_j)}. \tag{0.20}$$

Proof:

To see this, first recall that for hazard function $\lambda(t)$, we may perform an empirical estimate:

$$\widehat{\lambda}(t_j) = \frac{1}{Y(t_j)} \sum_{k \in Y(t_j)} 1\{X_k \leq t_j\} \tag{0.21}$$

where its expected value and variance are given by

$$\mathbb{E}\widehat{\lambda}(t_j) = \lambda(t_j), \mathbb{V}(\widehat{\lambda}(t_j)) = \frac{\lambda(t_j)(1 - \lambda(t_j))}{Y(t_j)} \tag{0.22}$$

Greenwood's Formula

Then using the central limit theorem, we know that

$$\sqrt{n}\left(\widehat{\lambda}(t_j) - \lambda(t_j)\right) \xrightarrow{d} N\left(0, \frac{\lambda(t_j)(1-\lambda(t_j))}{Y(t_j)}\right). \tag{0.23}$$

Then we introduce a technical lemma:

Lemma

(Delta-Method) Let $\sqrt{n}(\bar{X}_n - \mu) \xrightarrow{d} V$, where V is some distribution, $g(x)$ be a real-valued function such that g' exists at $x = \mu$ and is non-zero. Then

$$\sqrt{n}(g(\bar{X}_n) - g(\mu)) \xrightarrow{d} g'(\mu) \cdot V \tag{0.24}$$

Greenwood's Formula

We choose $g = \log(1-x)$, now apply Delta-method and we have

$$\sqrt{n}\left(\log(1-\widehat{\lambda}(t_j)) - \log(1-\lambda(t_j))\right) \xrightarrow{d}$$

$$N\left(0, \frac{\lambda(t_j)(1-\lambda(t_j))}{Y(t_j)} \cdot \left[-\frac{1}{1-\lambda(t_j)}\right]^2\right). \tag{0.25}$$

So now we have

$$\mathbb{V}(\log(1-\widehat{\lambda}(t_j))) = \frac{\lambda(t_j)}{Y(t_j)(1-\lambda(t_j))}. \tag{0.26}$$

Recall the relation that $\widehat{S}(t) = \prod_{t_j \le t}(1-\widehat{\lambda}(t_j))$, by summing over all estimates, we have

$$\mathbb{V}(\log(\widehat{S}(t))) = \sum_{t_j \le t} \frac{\lambda(t_j)}{Y(t_j)(1-\lambda(t_j))}. \tag{0.27}$$

Greenwood's Formula

Finally we apply Delta method again, this time choose our function to be $g(x) = e^x$, so now $g(\log \widehat{S}(t)) = \widehat{S}(t)$ and $g'(x)|_{x=\log(\widehat{S}(t))} = \widehat{S}(t)$, and we replace $S(t)$ by $\widehat{S}(t)$ when we have large sample, so now we have

$$\mathbb{V}(\widehat{S}(t)) = \widehat{S}^2(t) \cdot \sum_{t_j \leq t} \frac{\lambda(t_j)}{Y(t_j)(1 - \lambda(t_j))} \tag{0.28}$$

and by replacing $\lambda(t_j)$ with its MLE $\frac{d_j}{Y_j}$ will give us Greenwood's formula. **Q.E.D**

Confidence Interval

After knowing the variance, it is straightforward to compute the confidence interval of $S(t)$.

Theorem

The $100(1-\alpha)\%$ *confidence interval of* $S(t)$ *is given by*

$$\left(\widehat{S}(t) - z_{\alpha/2}\sqrt{\mathbb{V}(\widehat{S}(t))}, \widehat{S}(t) + z_{\alpha/2}\sqrt{\mathbb{V}(\widehat{S}(t))}\right). \tag{0.29}$$

The proof is just applying central limit theorem.

Complimentary Log-log transformation

We may notice that, the confidence interval from the above theorem does not necessarily lie inside $[0, 1]$, however $S(t) \in [0, 1]$ by definition so we seek for a more accurate estimate.

Complimentary Log-log Transformation

To do so, define

$$Z(t) = \log(-\log \widehat{S}(t)). \tag{0.30}$$

Then by Delta-method again (here the function is automatically chosen to be $g(x) = \log(-x)$), we will have

$$\mathbb{V}[Z(t)] = \frac{1}{(\log \widehat{S}(t))^2} \cdot \sum_{t_j \leq t} \frac{d_i}{Y_i(Y_i - d_i)} \tag{0.31}$$

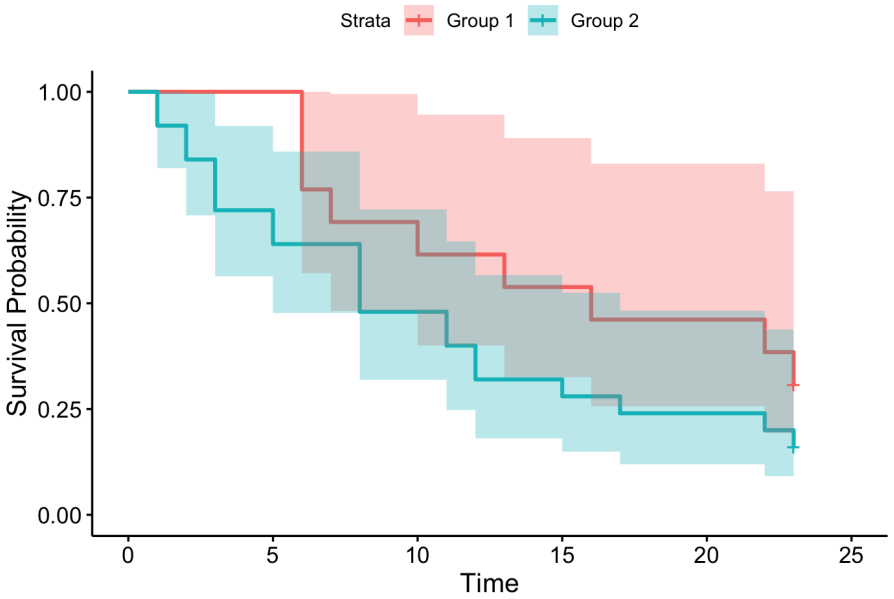
and this time if we compute the $100(1-\alpha)\%$ confidence interval it is guaranteed to fall in $[0, 1]$.

An Example

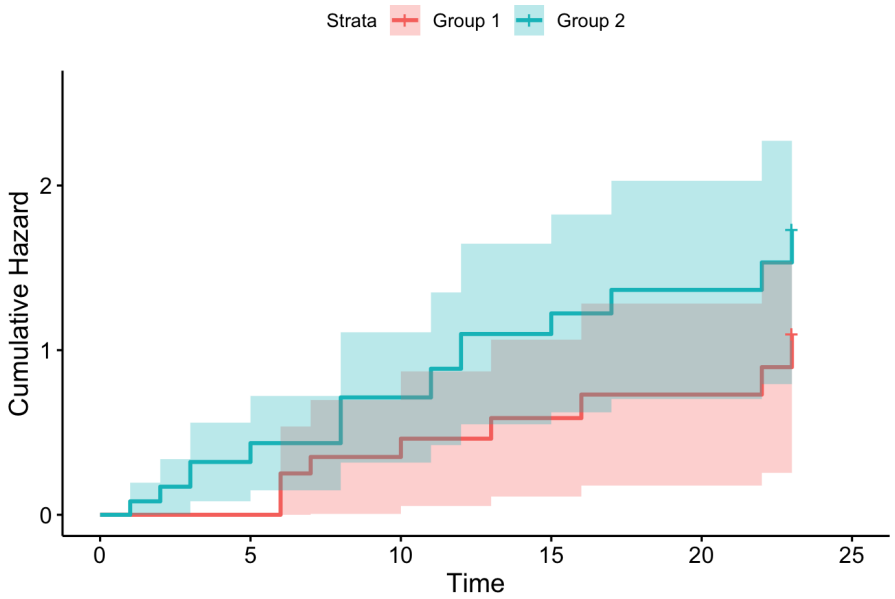
Let's go back to the Leukemia example I introduced in the first talk: You now should be more familiar with the table!

$t_{(f)}$	d_{1f}	d_{2f}	n_{1f}	n_{2f}
1	0	2	21	21
2	0	2	21	19
3	0	1	21	17
4	0	2	21	16
5	0	2	21	14
6	3	0	21	12
7	1	0	17	12
8	0	4	16	12
10	1	0	15	8
11	0	2	13	8
12	0	2	12	6
13	1	0	12	4
15	0	1	11	4
16	1	0	11	3
17	0	1	10	3
22	1	1	7	2
23	1	1	6	1

Survival Estimates



Kaplan–Meier curve with C.I



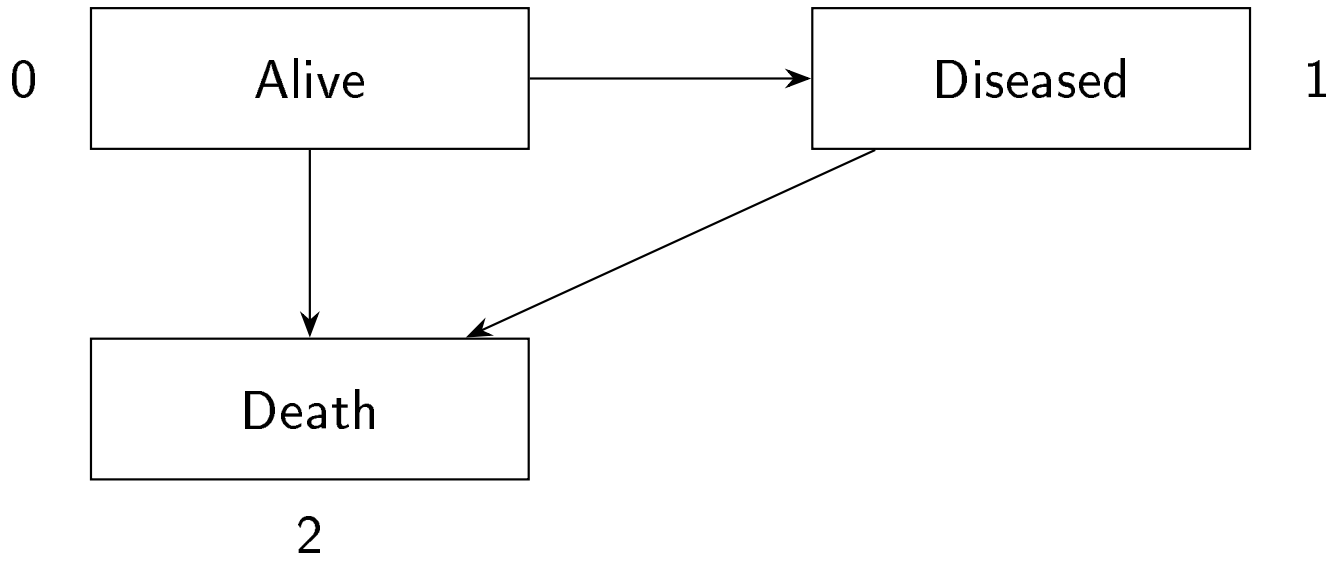
Nelson–Aalen curve with C.I

Aalen-Johansen Estimator

Previously, we only considered an easy model, that is, illness to death or illness to recovery. In real life, we may have a way more complicated model including different states, and Aalen-Johansen estimator is used in those cases. It is largely based on Kaplan-Meier estimator, and can be viewed as a multi-dimensional Kaplan-Meier estimator.

Alive-Illness-Death Model

Consider the following model with 3 different states:



An alive–illness–death model without recovery.

Alive-Illness-Death Model

Here is our mathematical set up:

- We denote $\alpha_{ij}(t)$ as the instantaneous transition hazard from state 1 to j at time $t + \Delta t$ $(i, j = 0, 1, 2)$, it can be viewed as a single case hazard function, with different transition states.

Alive-Illness-Death Model

Here is our mathematical set up:

- We denote $\alpha_{ij}(t)$ as the instantaneous transition hazard from state 1 to j at time $t + \Delta t$ $(i, j = 0, 1, 2)$, it can be viewed as a single case hazard function, with different transition states.
- We denote $\mathbb{P}_{ij}(s, t)$ as the probability that an individual in state i at time t will be in state j at time s.

Alive-Illness-Death Model

Here is our mathematical set up:

- We denote $\alpha_{ij}(t)$ as the instantaneous transition hazard from state 1 to j at time $t + \Delta t$ $(i, j = 0, 1, 2)$, it can be viewed as a single case hazard function, with different transition states.
- We denote $\mathbb{P}_{ij}(s, t)$ as the probability that an individual in state i at time t will be in state j at time s.
- We use $D_{ij}(t)$ as the observed number of individuals experiencing transition from state i to j at time t, $N_i(t)$ as the number of individual at risk in state i $(i = 1, 2)$ at t.

Alive-Illness-Death Model

By direct computation, we can obtain the followings:

$$\mathbb{P}_{00}(s,t) = \exp\left(-\int_s^t \alpha_{01}(u) + \alpha_{02}(u)du\right) \tag{0.32}$$

$$\mathbb{P}_{11}(s,t) = \exp\left(-\int_s^t \alpha_{12}(u)du\right) \tag{0.33}$$

$$\mathbb{P}_{12}(s,t) = \int_s^t \mathbb{P}_{11}(s,u)\alpha_{12}(u)du \tag{0.34}$$

$$\mathbb{P}_{01}(s,t) = \int_s^t \mathbb{P}_{00}(s,u)\alpha_{01}(u)\mathbb{P}_{11}(u,t)du \tag{0.35}$$

Alive-Illness-Death Model

Finally

$$\mathbb{P}_{02}(s,t) = \int_s^t \mathbb{P}_{00}(s,u) \cdot \alpha_{02}(u) du$$
$$+ \int_{v=u}^{v=t} \int_{u=s}^{u=t} \underbrace{\Big(\mathbb{P}_{00}(s,u) \cdot \alpha_{01}(u) \cdot \mathbb{P}_{11}(u,v)\Big)}_{\text{transition from 0 to 1}} \cdot \underbrace{\Big(\mathbb{P}_{11}(v,t) \cdot \alpha_{12}(v)\Big)}_{\text{transition from 1 to 2}} dudv$$

Alive-Illness-Death Model

What we got above are the parametric estimate of those features. Given the observed data defined earlier, we can also compute:

$$\widehat{\mathbb{P}}_{00}(s,t) = \prod_{s<t_j\leq t}\left(1-\frac{D_{0j}}{N_{0j}}\right); \widehat{\mathbb{P}}_{11}(s,t) = \prod_{s<t_j\leq t}\left(1-\frac{D_{12_j}}{N_{1j}}\right) \quad (0.36)$$

While $\mathbb{P}_{01}(s,t), \mathbb{P}_{12}(s,t)$ may be estimated by

$$\begin{aligned}\widehat{\mathbb{P}}_{01}(s,t) &= \sum_{s<t_j\leq t}\widehat{\mathbb{P}}_{00}(s,t_{j-1})\cdot\left(\frac{D_{01_j}}{N_{0j}}\right)\cdot\widehat{\mathbb{P}}_{11}(t_j,t); \widehat{\mathbb{P}}_{12}(s,t)\\ &= \sum_{s<t_j\leq t}\widehat{\mathbb{P}}_{11}(s,t_{j-1})\cdot\frac{D_{12_j}}{N_{1j}}.\end{aligned}$$

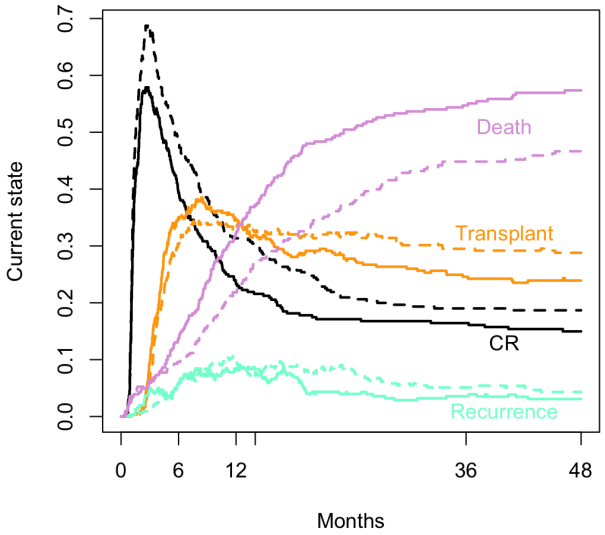
Finally, we have the following estimate of $\mathbb{P}_{02}(s,t)$:

$$\widehat{\mathbb{P}}_{02}(s,t) = \sum_{s<t_j\leq t} \widehat{\mathbb{P}}_{00}(s,t_{j-1}) \cdot \left(\frac{D_{02_j}}{N_{0j}}\right) \tag{0.37}$$

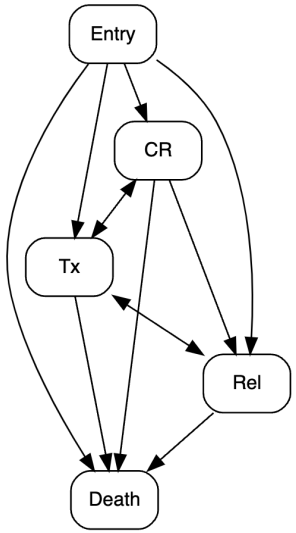
$$+ \sum_{s<t_j<t_k\leq t} \widehat{\mathbb{P}}_{00}(s,t_{j-1}) \cdot \left(\frac{D_{01_j}}{N_{0j}}\right) \cdot \widehat{\mathbb{P}}_{11}(t_j,t_k) \cdot \widehat{\mathbb{P}}_{11}(t_{k+1},t) \cdot \left(\frac{D_{12_j}}{N_{1j}}\right) \tag{0.38}$$

An Example

Below, scientists investigated a multi-state leukemia disease, and put the sample into treatment group and placebo group.



Transition Curves



Multi-State Graph

Aalen-Johansen Estimator

Now suppose we have observed the transition time $t_1 < t_2 < \cdots$ between any two states, also let $g, h \in \mathcal{I}, g \neq h$, we denote D_{ghj} as the number of individuals with transition from state g to state h at observed time t_j, and $D_{gj} = \sum_{h \neq g} D_{ghj}$ as the total number of transitions out of state g at observed time t, N_{gj} be the number of individuals at state g just prior to time t_j. Finally we define the $(k+1) \times (k+1)$ matrix $\widehat{\alpha}_j$ with entries (g, h) by

$$\widehat{\alpha}_j(g,h) = \begin{cases} \dfrac{D_{ghj}}{N_{gj}} & g \neq h \\ \\ -\dfrac{D_{gj}}{N_{gj}} & g = h \end{cases} \tag{0.39}$$

then the Aalen Johansen estimator takes the form

$$\widehat{\boldsymbol{P}}(s,t) = \prod_{s < t_j < t} (\boldsymbol{I} + \widehat{\alpha}_j) \tag{0.40}$$

Aalen-Johansen Estimator

Aalen-Johansen estimator can also be written as a product integral form, to see so, suppose we have a partition of the time interval (s, t) given by $s = \tau_0 < \tau_1, \cdots, \tau_K = t$, then we use Chapman-Kolmogorov equation from stochastic process, we have

$$\boldsymbol{P}(s, t) = \boldsymbol{P}(\tau_0, \tau_1) \times \boldsymbol{P}(\tau_1, \tau_2) \times \cdots \times \boldsymbol{P}(\tau_{k-1}, \tau_k). \tag{0.41}$$

Next Talk

In next talk, we shall formally prove some properties of the non-parametric estimators introduced in this talk, including "unbiasedness", asymptotic normality, etc. We will start by introducing counting processes and martingale theory and view the number of death from a process point of view. No pre-req in counting processes and martingale theory are needed.

Reference

[1] Survival Analysis: Analyzing Incident and Prevalent Cohort Survival Data, by *Jiajun Zhang*, 2025.

Thanks!