Statistical Inference for Survival Data

Jiajun Zhang

Department of Mathematics and Statistics, McGill University

SUMM 2026
Montréal, QC

Outline

Motivation

Survival analysis studies the behavior of time-to-event data, which arise naturally in modern biostatistical and epidemiological studies:

- Study the life time of a certain disease;
- Study whether a treatment is effective;
- Study how different covariates effect certain disease.

What makes survival analysis unique?

- If we recall any course on statistical inference, we study different methodologies for the data: Maximum likelihood estimate (MLE), confidence interval, hypothesis testing, etc. .
- Those methodologies also apply for the survival data, but survival data itself can be problematic: They are characterized by incomplete observations. Those features are known by **censoring** and **truncation**, which we will study later.

What is censoring & truncation?

Briefly speaking,

- Censoring occurs when the observation is incompletely determined for some subjects;
- Truncation occurs when only a group of chosen individuals are observed and the rest are systematically excluded from the study.

Different types of censoring

There are three main types of censoring:

- **Right censoring:** when we only know that the event happens after a certain time threshold;
- **Left censoring:** when we only know that the event happens before a certain time threshold;
- **Interval censoring:** we only know that the event time happens within a certain time period.

Examples of censoring

Example 1. During 1982-1992, 863 patients underwent kidney transplantation at the Ohio state university transplant center, where researchers investigated the survival time of those patients after transplantation. Their maximum follow-up time was 9.47 years.

Example 2. In 1978, Turnbull and Weiss studied the first usage of marijuana among California high school boys, and some answered with "I used it before, but I cannot remember when".

Example 3. The National Longitudinal Survey of Youth investigated a random sample of females aged 14 to 21 yearly from 1979 to 1988. Females in the survey were asked about any pregnancies that have occurred since they were last interviewed.

Figure: Three examples where censoring is present. Examples are taken from Zhang, 2026, Klein and Moeschberger, 2003

Truncation

- Truncation, on the other hand, can be viewed as a design flaw where a group of individuals are systematically excluded.
- This feature becomes obvious in survival analysis, because due to cost, we do not have access to all the individuals and some individuals may have already died before entering our study.
- We mainly work with left truncation, that is, the individual already died before entering our study.

Why do we care about left truncation?

- Our conclusion may be biased!

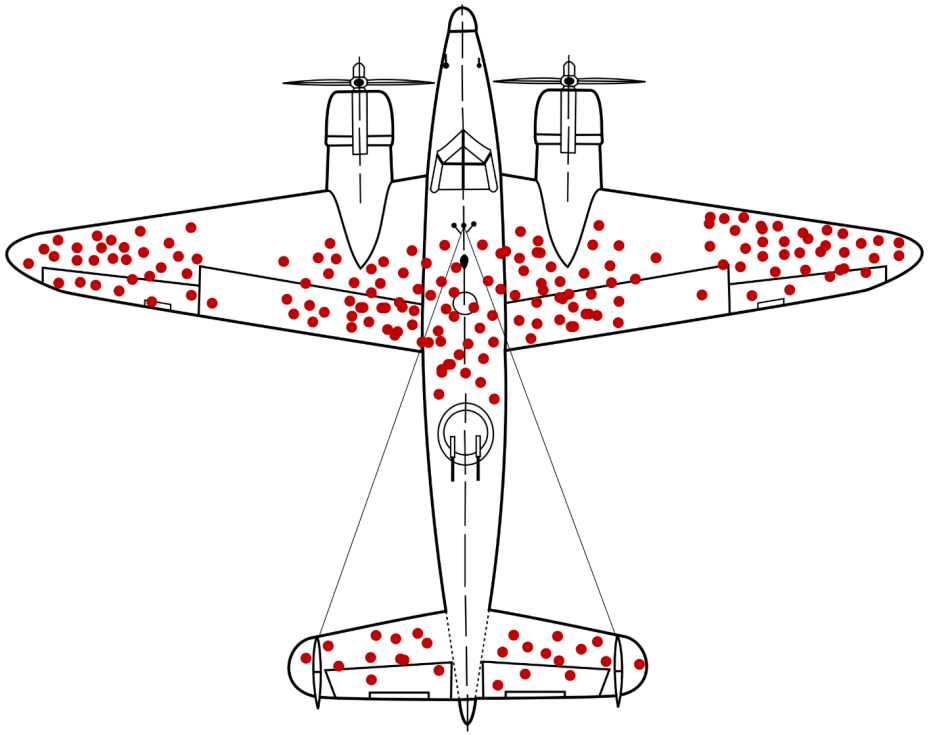

Figure: Survivorship bias: During WWII, researchers examined the damage done to aircrafts that had returned from mission.

Censoring and Truncation

- For simplicity and the purpose of this talk, we will assume only right censoring is present.

Survival Function

Definition

Let X be a non-negative random variable which denotes the life time, let $F(x)$ be the cumulative distribution function (cdf) of X, then the survival function is defined as

$$S(x) = \mathbb{P}(X \geq x) = 1 - F(x) = 1 - \int_0^x f(t)dt.$$

The survival function gives the probability that a subject will survive beyond the indicated time $t = x$.

Hazard Function

Definition

Let X be a non-negative random variable that denotes the life time, then the hazard function $h(x)$ is defined as

$$h(x) = \lim_{dx \to 0} \frac{\mathbb{P}(x \leq X \leq x + dx \big| X \geq x)}{dx}.$$

the hazard function is the instantaneous rate of death at time right after $t = x$ given that the individual is still alive at $t = x$.

Relation between survival function and hazard function

Proposition

Let X be a non-negative random variable, $S(x), h(x)$ denote the survival function and hazard function, then

$$h(x) = \frac{f(x)}{S(x)}.$$

Relation between survival function and hazard function

Proof.

$$
\begin{aligned}
h(x) &= \lim_{dx \to 0} \frac{\mathbb{P}((x \le X \le x + dx) \bigcap (X \ge x))}{\mathbb{P}(X \ge x) dx} \\
&= \lim_{dx \to 0} \frac{\mathbb{P}(x \le X \le x + dx)}{[1 - F(x)] dx} \\
&= \lim_{dx \to 0} \frac{F(x + dx) - F(x)}{[1 - F(x)] dx} \\
&= \frac{F'(x)}{1 - F(x)} = \frac{f(x)}{S(x)}.
\end{aligned}
$$

Relation between survival function and hazard function

In addition, if we apply the Chain rule, we have

$$h(x) = -\frac{d}{dx}\log(S(x)),$$

by integrating both sides, we have

$$H(x) = \int h(x)dx = -\log S(x),$$

where the integral term is defined as the **cumulative hazard function**.

Statistical inferences

In statistics, there are three main types of inferences on the sample we collected:

- **Parametric Inference:** The distribution of the sample is unknown up to some parameters. The methodologies are the easiest but less common in practice.
- **Non-parametric Inference:** We do not have any information on the distribution of the sample, and is very common in practice.
- **Semi-Parametric Inference:** This combines the parametric and non-parametric inference, where the distribution of the sample consists a parametric part and a non-parametric part.

Parametric Inference

Consider a random sample $X_1, \cdots, X_n \overset{i.i.d}{\sim} f(x, \theta)$.

- Suppose for each individual in the sample, X_i is the true life time, C_i is the censoring time independent of X_i.
- The true observation T_i is then defined by $T_i = \min\{X_i, C_i\}$.
- We introduce the following indicator δ_i:

$$\delta_i = \begin{cases} 1 & \text{if } X_i \leq C_i \\ 0 & \text{if } X_i > C_i \end{cases}. \tag{1}$$

- For uncensored individuals, their likelihood is simply given by $f(x_i, \theta)$. While for censored individuals, we know that $\mathbb{P}(X_i \geq C_i)$ hence their likelihood is proportional to $S(C_i, \theta)$.

Parametric Inference

Consider a random sample $X_1, \cdots, X_n \overset{i.i.d}{\sim} f(x, \theta)$.

- Thus the likelihood for the entire sample is given by

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} [f(x_i, \theta)]^{\delta_i} \cdot [S(C_i, \theta)]^{1-\delta_i},$$

  where δ_i is the censoring indicator defined in (1).
- After knowing the likelihood, we may use maximum likelihood estimate (MLE) method to find the best estimate of θ.

Early non-parametric methods

- In reality we often encounter non-parametric inferences where we do not know the distribution of the sample in advance: All we have is just a collection of observable data.
- One common approach is to estimate the **cumulative distribution function (cdf)** by its empirical form:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\}.$$

- Another estimation method is called **life table**.

Life table

	A	B	C	D
Age	2021 to 2023	2021 to 2023	2021 to 2023	2021 to 2023
60	91,805	552	0.00601	24.95
61	91,253	598	0.00655	24.10
62	90,655	648	0.00715	23.26
63	90,007	703	0.00781	22.42
64	89,304	763	0.00854	21.59
65	88,541	828	0.00935	20.77
66	87,713	898	0.01024	19.96
67	86,815	975	0.01123	19.17
68	85,840	1,058	0.01233	18.38
69	84,782	1,149	0.01355	17.60
70	83,634	1,246	0.01490	16.84
71	82,387	1,352	0.01641	16.08
72	81,035	1,466	0.01809	15.34
73	79,570	1,588	0.01996	14.62
74	77,982	1,719	0.02204	13.90
75	76,263	1,858	0.02437	13.21
76	74,405	2,007	0.02697	12.52
77	72,398	2,163	0.02988	11.86
78	70,235	2,327	0.03313	11.21
79	67,908	2,498	0.03678	10.57
80	65,410	2,674	0.04087	9.96

Table 1. The Life Table of Canadian Population Between Age 60 to 80 from year 2021 to 2023. The elements in the first rows represent: (A) The number of survivors at age x; (B) The number of deaths between age x and $x+1$; (C) Death probability between age x and $x+1$; (D) Life expectancy (in years) at age x. *Source: Statistics Canada. Table 13-10-0114-01 Life expectancy and other elements of the complete life table, three-year estimates, Canada, all provinces except Prince Edward Island.*

Life table

- We define l_x as the number of survivors at age x (in the table above, l_{60} would be $91,805$). So the theoretical deaths in year x would be $d_x = l_x - l_{x+1}$.
- Then the term $\frac{l_{x+1}}{l_x}$ is the conditional survival probability at year x.
- If we start our study from year 0, then the survival probability at age x can be estimated by

$$\hat{S}_x = \frac{l_x}{l_0} = \frac{l_1}{l_0} \times \frac{l_2}{l_1} \times \cdots \times \frac{l_x}{l_{x-1}} = \prod_{i=0}^{x} \left(1 - \frac{d_i}{l_i}\right).$$

Kaplan-Meier estimator

We introduce an estimator known as the **Kaplan-Meier (KM) Estimator**:

Theorem (Kaplan and Meier, 1958)

Suppose we have observed the ordered event time $t_1, \cdots, t_m$, where d_j deaths are observed at t_j, n_j individuals are at risk (event still not occurred) at t_j^-, then the Kaplan-Meier (KM) estimator of the survival function $S(t)$ is given by

$$\hat{S}(t) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{n_j}\right). \tag{2}$$

Kaplan-Meier estimator

We propose a plug in approach to derive the KM estimator given in equation (2). We consider each sub-interval $[t_i, t_{i+1})$:

- If $t \in [0, t_1)$, it means that no deaths had been observed in that interval, so the survival probability is simply 1.
- If $t \in [t_1, t_2)$, by definition we have

$$\begin{aligned} S(t) &= \mathbb{P}(X \geq t_1) \\ &= \mathbb{P}(\text{Survived after } t_1\ ) \\ &= \frac{\text{Number of Individuals survived after } t_1}{\text{Number of individuals at risk at } t_1^-} \\ &= \frac{n_1 - d_1}{n_1}. \end{aligned}$$

Kaplan-Meier estimator

- Now we consider $t \in [t_2, t_3)$, then we have the conditional probability given by

$$\begin{aligned} S(t) &= \mathbb{P}(X \geq t_2) \\ &= \mathbb{P}(\text{Survived after } t_2) \\ &= \mathbb{P}(\text{Survived after } t_2 \mid \text{Survived after } t_1) \cdot \mathbb{P}(\text{Survived after } t_1) \\ &= \frac{\text{Number of Individuals survived after } t_2}{\text{Number of individuals at risk at } t = t_2^-} \times \frac{n_1 - d_1}{n_1} \\ &= \frac{n_2 - d_2}{n_2} \times \frac{n_1 - d_1}{n_1}. \end{aligned}$$

Kaplan-Meier estimator

- Thus, we may use this plug in approach recursively, and we will get the KM estimator:

$$\hat{S}(t) = \prod_{t_j \leq t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{t_j \leq t} \left(1 - \frac{d_j}{n_j} \right).$$

- The KM curve $\hat{S}(t)$ with respect to time t is hence a non-increasing step function.

Bias and Variance of the KM estimator

Below, we briefly discuss the bias and variance of the KM estimator. The rigorous proof requires knowledge from martingale theory which can be found in Fleming and Harrington, 1991.

- The unbiasedness of KM estimator holds when assuming n_j is always non-zero.
- The variance of the KM estimator can be computed by **Greenwood's formula**:

$$\mathbf{Var}(\hat{S}(t)) = \hat{S}^2(t) \cdot \sum_{t_j \leq t} \frac{d_j}{n_j(n_j - d_j)}.$$

Nelson Aalen estimator

NA estimator estimates the cumulative hazard function $H(x)$. Recall that

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

by replacing $S(x)$ with KM estimator, we have

$$H(t) = -\sum_{t_j \le t} \log\left(1 - \frac{d_j}{n_j}\right).$$

Assume $d_j << n_j$, a Taylor expansion will yield

$$H(t) = \sum_{t_j \le t} \frac{d_j}{n_j} \quad \textbf{(Nelson-Aalen estimator)}.$$

Counting processes

Properties of KM, NA estimators can be rigorously studied using counting processes and martingale theory, which we only mention them briefly. More can be found in Li, n.d.

- The number of death $D(t)$ can be modeled by a **counting processes**, where $D(0) = 0$, $D(t) - D(s)$ represents the number of deaths occurred in $(s, t]$.
- On the other hand, suppose $h(t)$ is the true hazard rate, $N(t)$ is the number of individual at risk at t, then the **theoretical death** at time t is simply $N(t) \cdot h(t)$.
- Followed by **Doob-Meyer decomposition**, the following term is a mean zero martingale:

$$M(t) = D(t) - \int_0^t N(s)h(s)ds \quad \text{where } \mathbb{E}[M(t)] = 0. \quad (3)$$

Cox's Proportional Hazard (PH) model

- Cox's proportional Hazard Model (Cox, 1972) is a semi-parametric model which utilizes **covariates**, for example, the individual's gender, age, habits, etc.
- We usually denote $\boldsymbol{z} \in \mathbb{R}^p$ as a p-dimensional covariate vector, and the survival probability becomes the conditional probability given the covariates $\boldsymbol{z}$.
- Cox's PH model suggests that the hazard function given the covariates $\boldsymbol{z}$ takes the form

$$h(x|\boldsymbol{z}) = h_0(x) \cdot \exp\left(\boldsymbol{z}^\top \boldsymbol{\beta}\right), \tag{4}$$

where $h_0(x)$ is the **baseline hazard**, and $\boldsymbol{\beta}$ is the **coefficient vector**.

Partial likelihood

- Suppose at time t_0 we have m individuals with hazard function $h_1(t_0), \cdots, h_m(t_0)$ respectively, and we know there is exactly one death, then the probability that individual j has the event is given by

$$\mathbb{P}(\text{individual } j \text{ has the event} | t_0) = \frac{h_j(t_0)}{h_1(t_0) + \cdots + h_m(t_0)}.$$

- If we replace each $h_i(t_0)$ by equation (4), we have

$$\mathbb{P}(\text{individual } j \text{ has the event} | t_0) = \frac{\exp\left(\boldsymbol{z}_j^\top \boldsymbol{\beta}\right)}{\sum_{i=1}^m \exp\left(\boldsymbol{z}_i^\top \boldsymbol{\beta}\right)},$$

Partial likelihood

- We see the previous probability is independent of the event time, and only depends on the order in which events occur. Based on this we specify a model where we assume individual i fails at time t_i exactly. Then by multiplying the probability of each individual, we get the **partial likelihood** as follows:

$$\mathcal{L}(\boldsymbol{\beta}) = \prod_{i=1}^{m} \frac{\exp\left(\boldsymbol{z}_i^\top \boldsymbol{\beta}\right)}{\sum_{j \in R(t_i)} \exp\left(\boldsymbol{z}_j^\top \boldsymbol{\beta}\right)}, \tag{5}$$

  where $R(t_i)$ is the set of individuals at risk at time t_i.

- The equation (5) acts like the likelihood function, and we may use the method from maximum likelihood estimate to find the desired coefficients vector $\hat{\boldsymbol{\beta}}$.

Relative hazard

- If we choose two individuals with different covariates $\boldsymbol{z}_1, \boldsymbol{z}_2$, the ratio of the two hazard rates is then a constant given by

$$\frac{h(x|\boldsymbol{z}_1)}{h(x|\boldsymbol{z}_2)} = \exp((\boldsymbol{z}_1 - \boldsymbol{z}_2)^\top \boldsymbol{\beta}).$$

- Hence despite not knowing the baseline hazard $h_0(x)$, we may compare the hazard ratio among different covariates.

Example

- In a study of 863 kidney transplant patients, two different covariates were introduced: The patients' gender (male and female) and race (white and black).
- We will introduce 3 indicator covariates: $z_1 = 1$ if we have a black male and 0 otherwise; $z_2 = 1$ if we have a white male and 0 otherwise; $z_3 = 1$ if we have a black female and 0 otherwise.
- The white female is considered as the reference group and its covariate vector is set to be $\boldsymbol{z} = \mathbf{0}$, so $h_0(t)$ represents the hazard rate for white female.
- Our model is given by

$$h(t|\boldsymbol{z}) = h_0(t) \exp\Big(\beta_1 z_1 + \beta_2 z_2 + \beta_3 z_3\Big). \tag{6}$$

Results

- By solving equation (6) numerically using the dataset from Klein and Moeschberger, 2003, we have

$$\beta_1 = 0.1596, \beta_2 = 0.2484, \beta_3 = 0.6567.$$

- Now we can see the performances of non-reference group compared to the reference group:

$$\frac{h(t|z_1)}{h_0(t)} = 1.17, \frac{h(t|z_2)}{h_0(t)} = 1.28, \frac{h(t|z_3)}{h_0(t)} = 1.93.$$

- Thus the relative hazard risks for black male, white male and black female compared to white female are given by $1.17, 1.28, 1.93$ respectively.

References

- Cox, D. R. (1972). *Regression models and life tables.*. Journal of the Royal Statistical Society. Series B (Methodological).
- Fleming, T. R., & Harrington, D. P. (1991). *Counting processes and survival analysis.*. Wiley Series in Probability; Statistics.
- Kaplan, E. L., & Meier, P. (1958). *Non-parametric estimation from incomplete observations*. Journal of the American Statistical Association.
- Klein, J. P., & Moeschberger, M. L. (2003). *Survival analysis: Techniques for censored and truncated data*. Springer.
- Li, Y. (n.d.). *Lecture notes on survival analysis: A counting processes approach.*.
- Zhang, J. (2026). *Statistical inference for survival data*. MSURJ (Under Review).

Thanks!