Conference in Survival Analysis

1. Introduction to Censoring and Truncated Survival Data and Likelihood Inferences

Jiajun Zhang

June 29, 2025

Content of this report

- An Overview to Survival Analysis.

Content of this report

- An Overview to Survival Analysis.
- Properties of Survival Data: Censoring and Truncation.

Content of this report

- An Overview to Survival Analysis.
- Properties of Survival Data: Censoring and Truncation.
- Likelihood Inference on Survival Functions Based on Censoring and Truncation

An Overview of Survival Analysis

In survival analysis, we are interested in time-to-event data, for example, the lifetime of an electronic device. The general procedure to investigate our time-to-event data is:

An Overview of Survival Analysis

In survival analysis, we are interested in time-to-event data, for example, the lifetime of an electronic device. The general procedure to investigate our time-to-event data is:

- Identify our research interest;

An Overview of Survival Analysis

In survival analysis, we are interested in time-to-event data, for example, the lifetime of an electronic device. The general procedure to investigate our time-to-event data is:

- Identify our research interest;
- Conduct the experiment and collect our observed data based on our interest;

An Overview of Survival Analysis

In survival analysis, we are interested in time-to-event data, for example, the lifetime of an electronic device. The general procedure to investigate our time-to-event data is:

- Identify our research interest;
- Conduct the experiment and collect our observed data based on our interest;
- Estimate some key statistical features (eg. density function, confidence intervals, etc.) based on our observation;

An Overview of Survival Analysis

In survival analysis, we are interested in time-to-event data, for example, the lifetime of an electronic device. The general procedure to investigate our time-to-event data is:

- Identify our research interest;
- Conduct the experiment and collect our observed data based on our interest;
- Estimate some key statistical features (eg. density function, confidence intervals, etc.) based on our observation;
- Test the accuracy of our estimation (usually involves different types of hypothesis testing) and draw our conclusions.

An Overview of Survival Analysis

Survival analysis is widely used in bio-statistics and the time-to-event data is often the survival time of certain diseases. Here let's take a look of an example and see how we will proceed the 4-steps above:

An Overview of Survival Analysis

Survival analysis is widely used in bio-statistics and the time-to-event data is often the survival time of certain diseases. Here let's take a look of an example and see how we will proceed the 4-steps above:
In 1963, Freireich et al studied the effectiveness of a treatment (called 6-MP) to children with leukemia. They conducted this experiment on 42 children with leukemia, and divide them into 2 groups: 21 children received the drug 6-MP; while the other 21 received a placebo. The goal is to see if 6-MP acts differently than a placebo and would indeed lower the death rate of leukemia. The data we get is shown on the next slide:

$t_{(f)}$	d_{1f}	d_{2f}	n_{1f}	n_{2f}
1	0	2	21	21
2	0	2	21	19
3	0	1	21	17
3	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

$t_{(f)}$	d_{1f}	d_{2f}	n_{1f}	n_{2f}
1	0	2	21	21
2	0	2	21	19
3	0	1	21	17
3	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

The table on the left represents the leukemia data among the sample of 42 children. Each ordered time $t_{(f)}$ means deaths were recorded at that time, we use d_{if} to indicate the number of deaths at that time, separately by group i, followed by the numbers of subjects n_{if} at risk (i.e did not die at that time). Here 1 is the treatment group with 6-MP and 2 is the placebo group.

After we have obtained our data, we may use some standard techniques (as we will discuss in detail in later talks) to get the estimate of survival rate in 2 groups:

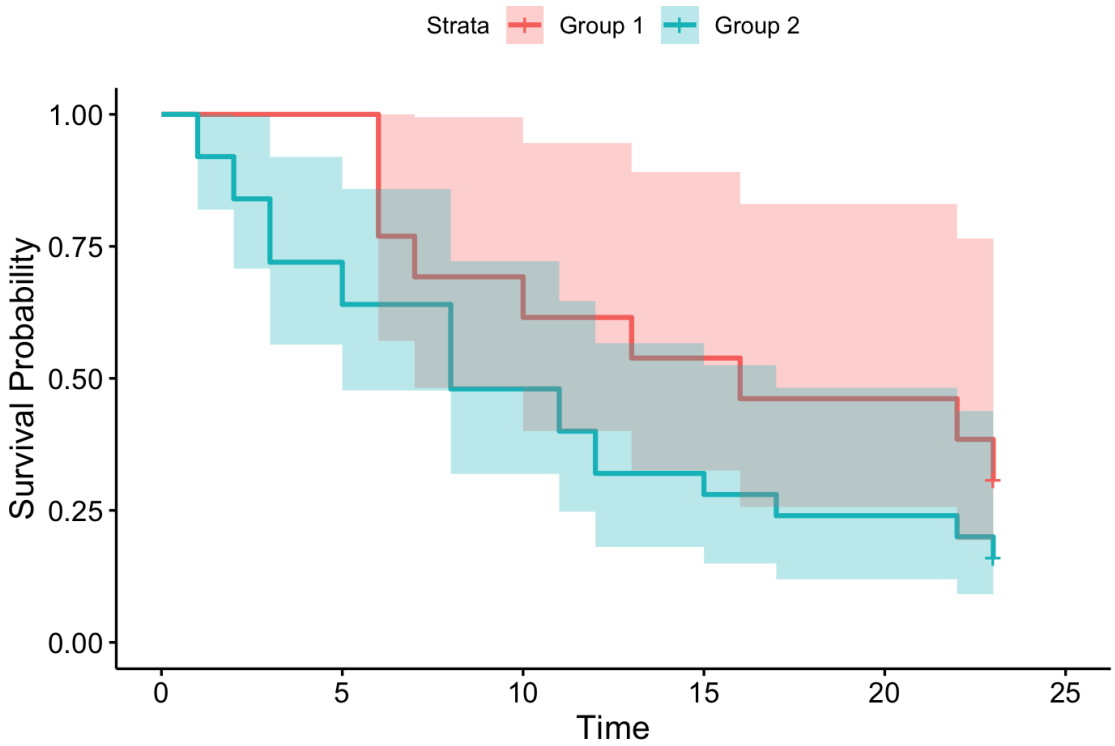

Figure: The estimated survival rate of two groups based on Kaplan-Meier estimator with 95% confidence interval (shaded region)

Finally, we would like to see if 6-MP plays an important role in lower the death rate of leukemia, and it may be carried out by a standard log-rank test. Denote $S_1(t), S_2(t)$ as the survival probability of children in group 1, 2 respectively, under the null hypothesis $\mathcal{H}_0 : S_1(t) = S_2(t)$, a theorem yields that

$$\frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \sim \chi^2_{(1)} \tag{0.1}$$

where $O_i = \sum_f d_{if}$ and $E_i = \sum_f \frac{n_{if}}{n_{1f} + n_{2f}} \times (d_{1f} + d_{2f})$.

Finally, we would like to see if 6-MP plays an important role in lower the death rate of leukemia, and it may be carried out by a standard log-rank test. Denote $S_1(t), S_2(t)$ as the survival probability of children in group 1, 2 respectively, under the null hypothesis $\mathcal{H}_0 : S_1(t) = S_2(t)$, a theorem yields that

$$\frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \sim \chi^2_{(1)} \tag{0.1}$$

where $O_i = \sum_f d_{if}$ and $E_i = \sum_f \frac{n_{if}}{n_{1f} + n_{2f}} \times (d_{1f} + d_{2f})$.

and by straightforward calculation, we obtain the p-value to be 4.17×10^{-5}, which suggests we shall reject the null hypothesis, meaning 6-MP plays a difference in the survival rate.

So What Makes Survival Analysis ”Hard”?

From the previous example, it just seems like regular statistics right? So what makes it "hard" and different? Well, we need to talk about **censoring** and **truncation**, a very common feature in survival analysis

Censoring and Truncation

Often times, the data we treated in regular statistics are too ”idealized”. In survival analysis, let's say when studying the survival time of a certain disease, due to cost and many other reasons, we do not have access to a complete set of data.

- In medical research, we often talk about the word "maximum follow-up time" as an indicator. That is, we do not and may not follow each patient until its death, because that will be inefficient and some follow-up times are way too long. Once we have designed this maximum follow-up time, all patients survive beyond this point are defined as ”survivors”, even though they may die one day after the follow-up time. In cancer research, the follow-up time is usually 5 years or depends, and the survival rate of the cancer actually means the survival rate within the follow-up time.

Censoring and Truncation

- Also, for those samples in our study, we may just simply lose track of them;

Censoring and Truncation

- Also, for those samples in our study, we may just simply lose track of them;
- Furthermore, since our research is in a specific time interval, there will be so many individuals before or after this time interval and we have (usually) absolutely no information on them.

So in general, the observed data is **incomplete** and not much accurate, which might lead to a huge bias if we try to estimate using standard techniques.

Censoring

Definition

We say an individual is **censored**, if it enters our study, but due to some reasons we are not able to record the exact event time (death time, etc).

Censoring

Definition

We say an individual is **censored**, if it enters our study, but due to some reasons we are not able to record the exact event time (death time, etc).

Based on the definition, there are may types of censoring, and we will mainly talk about: ① Right Censoring; ② Left Censoring; ③ Interval Censoring.

Censoring

- Right Censoring means the event time happens **after** some pre-determined time. For example, if we set our follow-up time of a certain disease to be 3 years, then the individual survive after 3 years is right-censored. (We only know that individual will die after that follow-up time, but we don't know when exactly).

Censoring

- Right Censoring means the event time happens **after** some pre-determined time. For example, if we set our follow-up time of a certain disease to be 3 years, then the individual survive after 3 years is right-censored. (We only know that individual will die after that follow-up time, but we don't know when exactly).
- Left Censoring means the event time happens **before** some pre-determined time. A straightforward example would be a survey on smoking. One may ask "When did you start smoking?", and if the answer is "I can't remember", then that individual is left-censored, since we only know the first time of smoking happened before the survey takes place, but we don't know when exactly. *In fact this example is also a special case of censoring called* ***Double Censoring****, where we only know if* $t < T$ *or* $t > T$*. If someone answers "I never smoke before", then no one will know when that person will smoke in the future.*

Censoring

- Interval Censoring means the event time happens **within** some pre-determined time intervals. If a person conducts yearly body check, and was discovered cancer in 2025 but not in 2024, then this individual is interval-censored, since we only know the cancer took place sometime within that year.

Some Examples

I have listed some examples, please identify what types of censoring we have in each:

Some Examples

I have listed some examples, please identify what types of censoring we have in each:

- During 1982 – 1992, 863 patients had their kidney transplant performed at the Ohio state university transplant center, and researchers investigated the survival time of those patients after the transplant. During the study, some patients were moved to other places where the researchers lost to follow up.
- The National Longitudinal Survey of Youth investigated a random sample of individuals 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.

More examples can be found in [1] Chapter 1.

Truncations

Definition

We say an individual is **truncated**, if it never enters our study and we have no information on them.

- Unlike censoring, we at least have some information, but for truncated data, we know nothing on them or they have been systematically excluded from our study.

Truncations

Definition

We say an individual is **truncated**, if it never enters our study and we have no information on them.

- Unlike censoring, we at least have some information, but for truncated data, we know nothing on them or they have been systematically excluded from our study.
- Truncation can also be viewed as a design flaw, but it's unpreventable. For example, if we start a research on a type of cancer in 2025, there might be individuals who already died because of this cancer many years ago, there also might be new patients getting this cancer after the end of our study, they are left-truncated and right-truncated respectively, follow the same definition as censoring.

Censoring & Truncation

Many would find left censoring and left truncation to be very similar, and here we give the main differences:

- Like I said, for censored observation, it enters our study, and we have partial information. But we know nothing for truncated observations, there might be or there might not be.

Censoring & Truncation

Many would find left censoring and left truncation to be very similar, and here we give the main differences:

- Like I said, for censored observation, it enters our study, and we have partial information. But we know nothing for truncated observations, there might be or there might not be.
- Plus, truncated data didn't even enter our study, so they are systematically excluded from our sample, say $X_1, \cdots, X_n$, and they do not have a huge impact on our final data since the sample has nothing to do with them (at least in this talk).

In survival analysis, we are mostly interested in **right censored, right truncated** *cohort* data, i.e we start our observation for all individuals in our sample at the same time.

Features of Survival Data

Now we would like to mathematically define some survival features. We will assume the sample that enters our study is $X_1, \cdots, X_n \overset{i.i.d}{\sim} f$ where f is a probability density function (or pmf) and $F(t) = \int_0^t f(s)ds$ indicates the probability of death at time t. Later we will see there are parametric, semi-parametric, non-parametric inferences but let's ignore them for now and assume we have parametric inference for a moment. Also let's ignore censoring and truncation for a moment.

Features of Survival Data

Definition

For our sample $X_1, \cdots, X_n \overset{i.i.d}{\sim} f$, the **survival function** is defined as

$$S(x) = 1 - F(x) = 1 - \int_0^t f(s)ds. \tag{0.2}$$

Features of Survival Data

Definition

For our sample $X_1, \cdots, X_n \overset{i.i.d}{\sim} f$, the **survival function** is defined as

$$S(x) = 1 - F(x) = 1 - \int_0^t f(s)ds. \tag{0.2}$$

Instead of modeling with probability of death, we often use the probability of survival to model our data, and this definition is equivalent as saying $S(x) = \mathbb{P}(X \geq x)$, which is the probability of survival at x.

Features of Survival Data

Another important feature is called the hazard rate, or the hazard function.

Definition

For our sample $X_1, \cdots, X_n \overset{i.i.d}{\sim} f$, the **hazard function** is defined as

$$h(x) = \lim_{dx \to 0} \frac{\mathbb{P}(x \leq X \leq x + dx | X \geq x)}{dx} \tag{0.3}$$

Features of Survival Data

Another important feature is called the hazard rate, or the hazard function.

Definition

For our sample $X_1, \cdots, X_n \overset{i.i.d}{\sim} f$, the **hazard function** is defined as

$$h(x) = \lim_{dx \to 0} \frac{\mathbb{P}(x \leq X \leq x + dx | X \geq x)}{dx} \tag{0.3}$$

Hazard function can be viewed as the "instantaneous" risk of death right after x given the individual survived beyond x.

Features of Survival Data

We may proceed the following on the hazard function $h(x)$:

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

So now we have an easier way to compute the hazard function.

Features of Survival Functions

We can also define cumulative hazard function, which is the cumulative hazard up to a given time x:

Definition

For our sample $X_1, \cdots, X_n \overset{i.i.d}{\sim} f$, the **cumulative hazard function** is defined as

$$H(x) = \int_0^x h(s)ds. \tag{0.4}$$

Features of Survival Functions

We also have the following relation between hazard function and survival function: If we apply the Chain rule, we have

$$h(x) = -\frac{d}{dx}\log(S(x)). \tag{0.5}$$

Taking the integral on both sides and interchanging the sign, we have

$$-H(x) = \log S(x) \tag{0.6}$$

then we take exponential on both sides, and we get

$$S(x) = \exp\{-H(x)\}. \tag{0.7}$$

This relation is very important when we derive Nelson-Aalen estimator from Kaplan-Meier estimator (in later talks).

Simple Parametric Likelihood Inference

Like I said before, in this talk we will assume we have a parametric inference, which means, all those functions we defined: $f(x), S(x), h(x)$, etc. are all known functions up to some unknown parameters, like $f \sim Exponential(\lambda)$, so it is easier to construct the likelihood inferences and use MLE methods to get our estimate of the function.

Definition

Given our sample $X_1, \cdots, X_n \overset{i.i.d}{\sim} f(x, \theta)$ where θ is the unknown parameter, the likelihood function of the sample, which is a function of θ, is defined as

$$L(\theta) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i, \theta) \tag{0.8}$$

Simple Parametric Likelihood Inference

In discrete case, the likelihood is equal to the probabiloty mass function (pmf) evaluated at x; In continuous case, it is okay to replace $\mathbb{P}(X_i = x, \theta)$ by $f(x_i, \theta)$ with some arguments: For small ϵ, we have

$$\mathbb{P}(x - \epsilon \leq X \leq x + \epsilon) = \int_{x-\epsilon}^{x+\epsilon} f(s)ds \approx 2\epsilon f(x) \tag{0.9}$$

so instead of measuring the support at a certain point like discrete case, we measure the support on a small neighborhood of x, and since ϵ is a constant with respect to θ, due to proportionality it is often times just use $f(x)$ to represent the likelihood.

Simple Parametric Likelihood Inference

In discrete case, the likelihood is equal to the probabiloty mass function (pmf) evaluated at x; In continuous case, it is okay to replace $\mathbb{P}(X_i = x, \theta)$ by $f(x_i, \theta)$ with some arguments: For small ϵ, we have

$$\mathbb{P}(x - \epsilon \leq X \leq x + \epsilon) = \int_{x-\epsilon}^{x+\epsilon} f(s)ds \approx 2\epsilon f(x) \tag{0.9}$$

so instead of measuring the support at a certain point like discrete case, we measure the support on a small neighborhood of x, and since ϵ is a constant with respect to θ, due to proportionality it is often times just use $f(x)$ to represent the likelihood.

so likelihood function is just simply change the variable from x to θ, and the MLE (maximum likelihood estimator) of θ is just the $\widehat{\theta}$ that maximizes $L(\theta)$, and it can be computed using standard techniques.

Simple Parametric Likelihood Inference

In discrete case, the likelihood is equal to the probabiloty mass function (pmf) evaluated at x; In continuous case, it is okay to replace $\mathbb{P}(X_i = x, \theta)$ by $f(x_i, \theta)$ with some arguments: For small ϵ, we have

$$\mathbb{P}(x - \epsilon \leq X \leq x + \epsilon) = \int_{x-\epsilon}^{x+\epsilon} f(s)ds \approx 2\epsilon f(x) \tag{0.9}$$

so instead of measuring the support at a certain point like discrete case, we measure the support on a small neighborhood of x, and since ϵ is a constant with respect to θ, due to proportionality it is often times just use $f(x)$ to represent the likelihood.

so likelihood function is just simply change the variable from x to θ, and the MLE (maximum likelihood estimator) of θ is just the $\widehat{\theta}$ that maximizes $L(\theta)$, and it can be computed using standard techniques.

Now, since our data contains censored observations, how can we construct the correct likelihood function?

Fixed Censoring Time

Let $X_1, \cdots, X_n \overset{i.i.d}{\sim} f(x, \theta)$, where X_i is a random variable indicating the death time of the individual. Suppose we have a fixed observation time C_i (often denoted as censoring time) for each individual X_i, then we know we will be able to observe the exact death time X_i if and only if $X_i \leq C_i$, otherwise we will have a right censored data.We often denote T_i as the true observed time and we have

$$T_i = \min\{X_i, C_i\} \quad \text{and} \quad \delta_i = 1\{X_i \leq C_i\} \tag{0.10}$$

Fixed Censoring Time

If at time x we have a uncensored observation, we will have $\delta_i = 1$, $T_i = X_i$ and now the likelihood will yield

$$L(\theta) = \mathbb{P}(T_i = X_i, \delta_i = 1) = f(X_i)$$

If at time x we have a censored observation, we will have $\delta_i = 0$, $T_i = C_i$, and now the likelihood will yield

$$\begin{aligned} L(\theta) &= \mathbb{P}(T_i = C_i, \delta_i = 0) \\ &= \mathbb{P}(T_i = C_i | X_i \geq C_i)\mathbb{P}(X_i \geq C_i) \\ &= \mathbb{P}(X_i \geq C_i) \\ &= S(C_i). \end{aligned}$$

Fixed Censoring Time

Denote $T_i = t_i$ as the observed time, so each individual can be written in a pair (T_i, δ_i) and combining all the observations, we have

$$L(\theta) = \prod_{i=1}^{n} [f(t_i)]^{\delta_i} \cdot [S(t_i)]^{1-\delta_i} \tag{0.11}$$

Use the relation between $S(t)$ and $H(t)$, we also have

$$L(\theta) = \prod_{i=1}^{n} h(t)^{\delta_i} \exp(-H(t_i)) \tag{0.12}$$

And now we can carry out the MLE method to get our estimate $\widehat{\theta}$, when **right censoring, and fixed censoring time** are present.

Random Censorship Model

Now we would like to introduce random censoring model, where censoring time C_i is no longer fixed and is in fact another random variable: $C_1, \cdots, C_n \overset{i.i.d}{\sim} f_{C_i}$. But an important assumption is that C_i, X_i are mutually independent. We shall use f_{X_i}, f_{C_i} as the probability density function (pmf) for X_i and C_i, and S_{X_i}, S_{C_i} as their survival functions respectively. We will again denote $T_i = \min\{X_i, C_i\}$ and $\delta_i = 1\{X_i \leq C_i\}$, and let t_i be the observed time.

Random Censorship Model

If we have a uncensored observation X_i, we have $T_i = X_i, \delta_i = 1$ and now the likelihood is

$$L(\theta) = \mathbb{P}(t_i = X_i, \delta_i = 1) = \mathbb{P}(t_i = X_i, X_i \leq C_i) \tag{0.13}$$

which then yields the following:

$$\begin{aligned}
\mathbb{P}(t_i = X_i, X_i \leq C_i) &= \frac{d}{dt_i} \int_0^{t_i} \int_{x_i}^{\infty} f_{X_i, C_i}(x_i, c_i) dc_i dx_i \\
&= \frac{d}{dt_i} \int_0^{t_i} f_{X_i}(x_i) dx \int_{x_i}^{\infty} f_{C_i}(c_i) dc_i \\
&= f_{X_i}(t_i) S_{C_i}(t_i).
\end{aligned}$$

Random Censorship Models

Similarly, if we have a censored observation X_i, we have $T_i = C_i, \delta_i = 0$ and now the likelihood is

$$L(\theta) = \mathbb{P}(t_i = C_i, \delta_i = 0) = \mathbb{P}(t_i = C_i, X_i \geq C_i) \tag{0.14}$$

which then yields the following:

$$\begin{aligned}
\mathbb{P}(t_i = C_i, X_i \geq C_i) &= \frac{d}{dt_i} \int_0^{t_i} \int_{c_i}^{\infty} f_{X_i, C_i}(x_i, c_i) dx_i dc_i \\
&= \frac{d}{dt_i} \int_0^{t_i} f_{C_i}(c_i) dc_i \int_{c_i}^{\infty} f_{X_i}(x_i) dx_i \\
&= f_{C_i}(t_i) S_{X_i}(t_i).
\end{aligned}$$

Random Censorship Models

So like we did before, combining all the above the likelihood when censoring are random is given by

$$L(\theta) = \prod_{i=1}^{n} [f_{X_i}(t_i) S_{C_i}(t_i)]^{\delta_i} \cdot [f_{C_i}(t_i) S_{X_i}(t_i)]^{1-\delta_i} \tag{0.15}$$

Note that, we assumed that X_i, C_i are mutually independent.

Random Censorship Models

So like we did before, combining all the above the likelihood when censoring are random is given by

$$L(\theta) = \prod_{i=1}^{n} [f_{X_i}(t_i) S_{C_i}(t_i)]^{\delta_i} \cdot [f_{C_i}(t_i) S_{X_i}(t_i)]^{1-\delta_i} \tag{0.15}$$

Note that, we assumed that X_i, C_i are mutually independent.

If the distribution of C_i does not depend on θ, then in the likelihood sense, f_{C_i}, S_{C_i} are all constant with respect to θ and we may simplify our likelihood as

$$L(\theta) = K \prod_{i=1}^{n} [f(t_i)]^{\delta_i} \cdot [S(t_i)]^{1-\delta_i} \tag{0.16}$$

which is proportional same as we obtained when censoring is not random. It will also yield the same MLE.

More Censoring Models

Further censoring models like left-censoring, interval censoring, type (ii) censoring, etc. can be found in [2], in Chapter 3, and we will mainly focus on right censoring throughout all the talks.

Estimating Using MLE

Since we have our parametric model, the best way to estimate θ is via the MLE method. Suppose we have our likelihood function given by

$$L(\theta) = K \prod_{i=1}^{n} [f(t_i)]^{\delta_i} \cdot [S(t_i)]^{1-\delta_i} \tag{0.17}$$

Suppose we have $f(x, \theta) \sim Exponential(\theta) = \theta e^{-\theta x}$, by direct computation we get $S(x) = e^{-\theta x}$, and we investigate the partial derivative of the log-likelihood:

$$\frac{\partial \log L(\theta)}{\partial \theta} = K \left(- \sum_{\text{censored}} x_i + \sum_{\text{uncensored}} \left(\frac{1}{\theta} - x_i \right) \right) \tag{0.18}$$

and the MLE is

$$\widehat{\theta}_{\text{MLE}} = \frac{\sum \delta_i}{\sum X_i}. \tag{0.19}$$

Confidence Intervals

A straightforward way to construct confidence interval is using the asymptotic normality of MLE

Confidence Intervals

A straightforward way to construct confidence interval is using the asymptotic normality of MLE

Theorem

Under regularity conditions, let $X \sim f(x, \theta_0)$, θ_0 *be the true parameter,* $\widehat{\theta}_n$ *be the MLE, then*

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_0\right) \xrightarrow{d} N\left(0, \mathcal{I}_1^{-1}(\theta_0)\right) \tag{0.20}$$

where $\mathcal{I}_1(\theta_0)$ *is the Fisher information defined by*

$$\mathcal{I}_1(\theta_0) = -\mathbb{E}\left\{ \frac{\partial^2}{\partial \theta^2} \log f(x, \theta) \Big|_{\theta=\theta_0} \right\} \tag{0.21}$$

Confidence Intervals

Also, we have a stronger convergence theorem:

Theorem

Under regularity conditions, let $\widehat{\theta}_0$ denote the MLE of θ_n, then

$$\widehat{\theta}_n \xrightarrow{a.s} \theta_0. \tag{0.22}$$

Hence we could replace θ_0 by $\widehat{\theta}_n$ in Fisher information, and by the previous theorems, it is straightforward to compute the following confidence interval:

$$\left(\widehat{\theta}_n - z_{\alpha/2} \cdot \sqrt{\frac{1}{n}[\mathcal{I}_1(\widehat{\theta}_n)]^{-1}}, \widehat{\theta}_n + z_{\alpha/2} \cdot \sqrt{\frac{1}{n}[\mathcal{I}_1(\widehat{\theta}_n)]^{-1}}\right) \tag{0.23}$$

where $100(1-\alpha)\%$ is the significance level, and $\mathbb{P}(X < z_{\alpha/2}) = \alpha/2$ for $X \sim N(0,1)$.

Hypothesis Testing

Like I said before, we are also interested in testing the parameter θ. We will introduce one type of hypothesis testing: **Likelihood Ratio Test**

Hypothesis Testing

Like I said before, we are also interested in testing the parameter θ. We will introduce one type of hypothesis testing: **Likelihood Ratio Test**

- In hypothesis testing, we **do not** estimate the parameter θ, instead, we will test the accuracy of a pre-made assumption $\theta = \theta_0$.

Hypothesis Testing

Like I said before, we are also interested in testing the parameter θ. We will introduce one type of hypothesis testing: **Likelihood Ratio Test**

- In hypothesis testing, we **do not** estimate the parameter θ, instead, we will test the accuracy of a pre-made assumption $\theta = \theta_0$.
- We will often involve a null hypothesis $\mathcal{H}_0 : \theta = \theta_0$ and an alternative hypothesis $\mathcal{H}_1$, where $\mathcal{H}_1$ could be $\theta = \theta_1$, $\theta \neq \theta_0$, $\theta > \theta_0$, $\theta < \theta_0$ or whatever, depending on our interest. We will then perform some tests and see if we tend to believe $\mathcal{H}_0$ or $\mathcal{H}_1$, or reject $\mathcal{H}_0$ or $\mathcal{H}_1$.

Hypothesis Testing

Like I said before, we are also interested in testing the parameter θ. We will introduce one type of hypothesis testing: **Likelihood Ratio Test**

- In hypothesis testing, we **do not** estimate the parameter θ, instead, we will test the accuracy of a pre-made assumption $\theta = \theta_0$.
- We will often involve a null hypothesis $\mathcal{H}_0 : \theta = \theta_0$ and an alternative hypothesis $\mathcal{H}_1$, where $\mathcal{H}_1$ could be $\theta = \theta_1$, $\theta \neq \theta_0$, $\theta > \theta_0$, $\theta < \theta_0$ or whatever, depending on our interest. We will then perform some tests and see if we tend to believe $\mathcal{H}_0$ or $\mathcal{H}_1$, or reject $\mathcal{H}_0$ or $\mathcal{H}_1$.
- If we have simple versus simple hypothesis, like $\mathcal{H}_0 : \theta = \theta_0$ and $\mathcal{H}_1 : \theta = \theta_1$, then we may use **Neyman-Pearson Lemma** to conduct our test.

Hypothesis Testing

Like I said before, we are also interested in testing the parameter θ. We will introduce one type of hypothesis testing: **Likelihood Ratio Test**

- In hypothesis testing, we **do not** estimate the parameter θ, instead, we will test the accuracy of a pre-made assumption $\theta = \theta_0$.
- We will often involve a null hypothesis $\mathcal{H}_0 : \theta = \theta_0$ and an alternative hypothesis $\mathcal{H}_1$, where $\mathcal{H}_1$ could be $\theta = \theta_1$, $\theta \neq \theta_0$, $\theta > \theta_0$, $\theta < \theta_0$ or whatever, depending on our interest. We will then perform some tests and see if we tend to believe $\mathcal{H}_0$ or $\mathcal{H}_1$, or reject $\mathcal{H}_0$ or $\mathcal{H}_1$.
- If we have simple versus simple hypothesis, like $\mathcal{H}_0 : \theta = \theta_0$ and $\mathcal{H}_1 : \theta = \theta_1$, then we may use **Neyman-Pearson Lemma** to conduct our test.
- Otherwise, we shall use **Likelihood Ratio Test**. But in all cases, the null hypothesis must be simple, i.e $\mathcal{H}_0 : \theta = \theta_0$ for some θ_0.

Likelihood Ratio Test

Like in confidence interval, we also need to specify a significance level $100(1-\alpha)\%$, due to **type 1** and **type 2** error, and it's like a bi-variable optimization problem if we do not specify α, so by fixing α will make our life easier.

Likelihood Ratio Test

The framework of likelihood ratio test is largely based on MLE theory. Suppose we have $X_1, \cdots, X_n \overset{i.i.d}{\sim} f(x, \theta)$ where θ is our parameter, and we have two tests: $\mathcal{H}_0 : \theta = \theta_0$ and $\mathcal{H}_1$, then we will compute two separate MLEs:

Likelihood Ratio Test

The framework of likelihood ratio test is largely based on MLE theory. Suppose we have $X_1, \cdots, X_n \overset{i.i.d}{\sim} f(x, \theta)$ where θ is our parameter, and we have two tests: $\mathcal{H}_0 : \theta = \theta_0$ and $\mathcal{H}_1$, then we will compute two separate MLEs:

- The first MLE is based on the null hypothesis, i.e the MLE of θ under the restriction of $\mathcal{H}_0$. Then it is just θ_0.
- The second MLE is based on the alternative hypothesis, i.e the MLE of θ under the restriction of $\mathcal{H}_1$. Then based on the type of $\mathcal{H}_1$, we will then have to solve an optimization problem to find the corresponding MLE. But usually we make $\mathcal{H}_1 : \theta \neq \theta_0$ so now the MLE is then simply the regular MLE over the entire sample space.

Likelihood Ratio Test

Theorem

Consider the random sample $X_1, \cdots, X_n \sim f(x, \theta)$, *and we have* $\mathcal{H}_0$ *(null hypothesis) and* $\mathcal{H}_1$ *(alternative hypothesis) as two tests, and we define the likelihood ratio (LR) statistic to be*

$$\lambda_n(\boldsymbol{X}) = \frac{L_n(\widehat{\theta}_{MLE,\mathcal{H}_0})}{L_n(\widehat{\theta}_{MLE})} \tag{0.24}$$

A test based on LR statistic has the following form

$$\phi(\boldsymbol{X}) = \begin{cases} 1 & \lambda_n(\boldsymbol{X}) < C \text{ (reject } \mathcal{H}_0) \\ 0 & \lambda_n(\boldsymbol{X}) > C \end{cases} \tag{0.25}$$

for $C \in [0, 1]$ *and the rejection region takes the form* $\mathcal{R} := \{\boldsymbol{x} \in \mathcal{X} : \lambda_n(\boldsymbol{X}) < C\}$.

Likelihood Ratio Test

Theorem

(Asymptotic Approximation of LR)
At significance level α, the rejection region $(\lambda_n(\boldsymbol{X}) < C)$ of the LR-based test under regularity conditions for large n is approximately

$$\mathcal{R} := \left\{ \boldsymbol{x} \in \mathcal{X} : -2\log[\lambda_n(\boldsymbol{X})] \geq \chi^2_{d,\alpha} \right\}, d = \dim\Theta - \dim\Theta_0 \tag{0.26}$$

where

$$-2\log[\lambda_n(\boldsymbol{X})] = 2\left\{ \sup_{\theta\in\Theta} \ell_n(\theta) - \sup_{\theta\in\Theta_0} \ell_n(\theta) \right\} \tag{0.27}$$

ℓ_n is the log-likelihood function, Θ is the entire parameter space, Θ_0 is the parameter space under null hypothesis.

Likelihood Ratio Test

So, the theorem tells us it is sufficient to compute $-2\log[\lambda_n(\boldsymbol{X})]$ and compare it with a χ^2 statistic from the table. There are other similar tests like **Wald Test**, **Rao Test**, but they are asymptotically equivalent to likelihood ratio test. More on those tests can be found in [2], Chapter 3. In most case if we are dealing with single parameter, we just have

$$-2\log[\lambda_n(\boldsymbol{X})] \xrightarrow{d} \chi^2_1. \tag{0.28}$$

Next Talk

In our next talk, we will see how we can build our inference with non-parametric inference. What we did this time was mainly a parametric inference approach but in reality we usually do not have this condition. We will build Kaplan-Meier estimator, which is a non-parametric estimator for the survival function. Before next talk, you can try to think how we can build this estimator if a set of data (with censoring) is given to you.

Reference

[1] Survival Analysis: Techniques for Censored and Truncated Data, by *John P. Klein and Melvin L. Moeschberger*, 2002.

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

Thanks!