Notes for Biostatistics 140.641.01 'Survival Analysis' Department of Biostatistics

First Term, 2022

General Information

- Instructor: Yuxin Zhu <daisy@jhu.edu>
- Teaching Assistants:

  Chunnan Liu <cliu173@jhmi.edu>; Jiyang Wen <jwen22@jhu.edu>

  Chunnan Liu Laboratory: TBA Jiyang Wen Laboratory: TBA
- Lectures (Zhu): Tu/Thur 3:30-4:50pm, Room W2030

  Office hour by zoom: US EST Friday 9-10am

  https://jhjhm.zoom.us/j/91481312079?pwd=NFVIdDFnOFFxU1hXbC9oMzhzblkzUT09
- Evaluation based on 3 homework sets, one computing homework and one closed-book inclass exam. Grade is based on homework (50%) and one final exam (50%).
- Course material at intermediate level; proper math/stat background required
- Course material available at CoursePlus website http://courseplus.jhsph.edu

Course content

Chapter 1 - **Introduction and basics.** Survival and hazard functions; censoring; some basic concepts.

Chapter 2 - **Parametric modeling and estimation.**

Chapter 3 - **One sample estimation.** Concepts of nonparametric estimation; Kaplan-Meier estimate.

Chapter 4 - **Regression models.** With focus on PHM.

Chapter 5 - **PHM and beyond.**

Chapter 6 - **Hypothesis testing.**

Chapter 7 - **Competing risks models.**

Survival analysis

Survival analysis is a major branch of statistics/biostatistics which deals with modeling, estimation, prediction and testing for time-to-event outcome. In the collection of survival data, some of the sampling constraints (such as censoring) could arise due to specific features of time-to-event data. Survival analysis covers a very broad range of topics in epidemiology, medicine, finance, economics, imaging, engineering, genomics and genetics.

Chapter 1. Introduction and basics

Background: from binary outcome to survival time outcome

In public health or biomedical studies, the research interest frequently focuses on a binary outcome such as incidence of disease (Cancer, Alzheimer's Disease, HIV/AIDS, etc.). Suppose the binary outcome is the incidence of a specific disease. Let D denote a binary outcome variable with values $0, 1$ indicating the presence or absence of the disease. The prevalence rate is defined as $P(D = 1)$. With a $p \times 1$ vector of subject-level covariates X, the risk probability is defined as $P(D = 1 \mid X)$ and the risk-free probability is $P(D = 0|X)$. We understand that the logistic regression model is commonly adopted to find the relationship between X and D.

How to define the binary outcome D? The collection of binary variable D in a specified time interval (like 3 years follow-up time, or age interval 15∼60) is sometimes complicated by incomplete follow-up. Also, the variable D provides only limited information because it is a 0-1 variable. In contrast, a time-to-disease variable (survival time) conceptually provides much richer information than a binary outcome. Also, as will be seen in this course, the problem of incomplete follow-up for survival time can be properly handled by novel analytical approaches.

1.1 Survival time

Definition: A survival time, T, is a nonnegative-valued random variable. Note that we sometimes use survival time, failure time, lifetime interchangeably and they all represent 'time to event.' For most of the applications, the value of T is time from time-origin 0 to a failure event. For example,

a) in a clinical trial, T is time from start of treatment to diagnosis of disease.

b) to study an infectious disease such as HIV, T is time from onset of infection (HIV infection) to diagnosis of disease (AIDS). T is called the incubation time.

c) to study a genetic disease, T is set as the onset age of disease.

In some applications the survival time could be defined using alternative time unit. For example,

d) In women's pregnancy studies, T is the number of menstrual cycles since time origin.

e) When evaluating reliability of automobiles, the actual lifetime of a car is less meaningful than the mileage on the car. Thus, T could be defined as the mileage driven by the car until a failure event (car accident).

1.2 Definitions of survival and hazard functions

Definition. Survival function $S(t)$.

$$S(t) = \Pr(T \geq t) = 1 - \Pr(T < t)$$

Definition. Cumulative distribution function $F(t)$.

$$F(t) = \Pr(T \leq t)$$

Remarks:

1. If the random variable T is continuous, then point probabilty is 0; that is, $P(T = t) = 0$. In this case, $S(t) = 1 - F(t)$.

2. In some textbooks the survival function is defined as $S(t) = \Pr(T > t)$, which is slightly different from our definition here. We adopt $S(t) = \Pr(T \geq t)$ for better understanding of further/deeper methodology.

Characteristics of $S(t)$:

- $S(t) = 1$ if $t \leq 0$, $\quad S(\infty) = \lim_{t \to \infty} S(t) = 0$
- $S(t)$ is decreasing in t: $S(t_2) \leq S(t_1)$ if $t_2 \geq t_1$
- $S(t)$ is left-continuous
- In general, the survival function $S(t)$ provides useful summary information, such as the median or 60th-percentile survival time, t-year survival rate, etc.
- It is common to estimate the survival function $S(t)$ in exploratory data analysis. For example, in a randomized clinical trial such an exploratory analysis is useful for the comparison between treated and control groups.

Example: Bone marrow transplant data.

137 Leukemia patients underwent bone marrow transplant

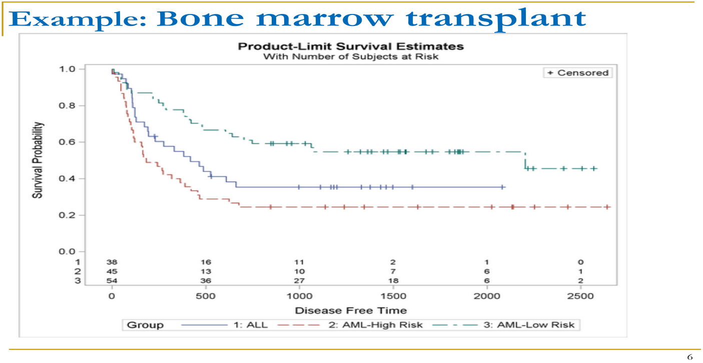
Time origin: date receiving transplant

Failure event: death

T: time from transplant to death

Research questions/issues:

- understanding distributional structure of time-to-death (so-called "disease free time" on the plot)
- 3 kinds of patients: ALL (acute lymphoblastic leukemia), AML (acute myeloid leukemia)-Low Risk, and AML-High Risk. Survival difference between groups?
- sampling constraint: you may not observe failure event for everyone (censoring)
- statistical/analytical methods?



Example: Bone marrow transplant

We will discuss how to estimate the survival probability, $S(t)$, in Chapter 3 (Kaplan-Meier curve).

Let $f(t)$ denote the probability density function of T.

Definition. Hazard function $\lambda(t)$.

a) If survival time T is discrete,

$$\lambda(t) = \mathrm{P}(T = t | T \geq t) = \frac{\mathrm{P}(T = t, T \geq t)}{\mathrm{P}(T \geq t)} = \frac{\mathrm{P}(T = t)}{\mathrm{P}(T \geq t)} = \frac{f(t)}{S(t)}.$$

Characteristics of $\lambda(t)$:

- $\lambda(t)$ is interpreted as the hazard probability.
- $0 \leq \lambda(t) \leq 1$
- $\lambda(t) = 0$ if t is not a discrete point of T
- If $t_1, ..., t_k$ are discrete points of T, then you only need $\lambda(t_1), ..., \lambda(t_k)$

b) If T is (absolutely) continuous,

$$\begin{aligned}\lambda(t) &= \lim_{\Delta\to 0^+} \frac{\Pr(t \le T < t+\Delta \mid T \ge t)}{\Delta} \\ &= \lim_{\Delta\to 0^+} \frac{P(T \in [t, t+\Delta))/S(t)}{\Delta} \\ &= \frac{1}{S(t)} \cdot \lim_{\Delta\to 0^+} \frac{P(T \in [t, t+\Delta))}{\Delta} \\ &= \frac{f(t)}{S(t)}\end{aligned}$$

Characteristics of $\lambda(t)$:

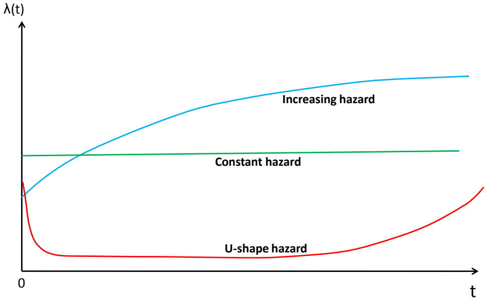
- $\lambda(t)$ is interpreted as instantaneous failure rate at t given survival up to t
- $0 \le \lambda(t) < \infty$
- $\lambda(t) \cdot \Delta \approx \Pr(t \le T < t+\Delta \mid T \ge t)$, the proportion of individuals experiencing failure in $[t, t+\Delta)$ to those surviving up to t

Example. Suppose T is a continuous survival time.

a. Constant hazard $\lambda(t) = \lambda_0$

b. Increasing hazard $\lambda(t_2) \geq \lambda(t_1)$ if $t_2 \geq t_1$

c. U-shape hazard (human mortality for age at death)



Some Insights

Probability theory. Suppose the survival time T has positive probabilities at n discrete points $x_1 < x_2 < \ldots < x_n$. Then, for $x_j < t \leq x_{j+1}$,

$$\begin{aligned}
\frac{S(t)}{S(x_1)} &= \frac{S(x_{j+1})}{S(x_1)} \\
&= \frac{P(T \geq x_2)}{P(T \geq x_1)} \times \frac{P(T \geq x_3)}{P(T \geq x_2)} \times \ldots \times \frac{P(T \geq x_{j+1})}{P(T \geq x_j)} \\
&= \frac{P(T \geq x_1) - P(T = x_1)}{P(T \geq x_1)} \times \ldots \times \frac{P(T \geq x_j) - P(T = x_j)}{P(T \geq x_j)} \\
&= (1 - \lambda(x_1)) \times (1 - \lambda(x_2)) \times \ldots \times (1 - \lambda(x_j))
\end{aligned}$$

Note that $S(x_1) = 1$ and $S(t) = \prod_{i=1}^{j} (1 - \lambda(x_i))$.

Statistical estimation. So, this motivates a nonparametric way to estimate $S(t)$ based on observed data (as a preparation for understanding the Kaplan-Meier estimator):
If we observe survival data $\{x_1, x_2, \ldots, x_n\}$ with large sample size n, where $x_1 < x_2 < \ldots < x_n$, then $S(x_1) \approx 1$ and $S(t) \approx \prod_{i=1}^{j} (1 - \lambda(x_i))$.

Definition. Cumulative hazard function $\Lambda(t)$.

a) If T is discrete, let x_i's be the discrete points,

$$\Lambda(t) = \sum_{x_i \le t} \lambda(x_i)$$

b) If T is (absolutely) continuous,

$$\begin{aligned} \Lambda(t) &= \int_0^t \lambda(u)du \\ \frac{d\Lambda(t)}{dt} &= \lambda(t) \end{aligned}$$

1.3 Relationship among Functions

a) If T is discrete,

$$\lambda(t) = \frac{f(t)}{S(t)}$$

b) If T is (absolutely) continuous,

$$\lambda(t) \quad = \quad \frac{f(t)}{S(t)}$$

A well known relationship among the density, hazard and survival functions is

$$\boxed{\lambda(t) = \frac{f(t)}{S(t)}} \; .$$

Facts:

- Probability distribution of T is determined by probability density function (pdf), or cumulative distribution function (cdf), or survival function (sf)
- Probability distribution of T is determined by hazard function
- Thus, modeling the hazard function is a valid approach for statistical modeling (as good as modelling pdf, or cdf, or sf!)

For continuous T,

$$\begin{aligned}\Lambda(t) &= \int_0^t \lambda(u)du = \int_0^t \frac{f(u)}{S(u)}du \\ &= \int_0^t \frac{\left(-\frac{dS(u)}{du}\right)}{S(u)}du = [-\log S(u)]\,\Big|_0^t \\ &= [-\log S(t)] - [-\log S(0)] = -\log S(t)\end{aligned}$$

Thus

$$\boxed{S(t) = e^{-\Lambda(t)} = e^{-\int_0^t \lambda(u)du}}$$

We now see that $\lambda(\cdot)$ is determined if and only if $f(\cdot)$ (or $S(\cdot)$) is determined.

When T is a **regular** continuous variable, we have

$$\boxed{\int_0^\infty \lambda(u)du = \infty}$$

This formula is implied by

$$0 = S(\infty) = e^{-\int_0^\infty \lambda(u)du}\ .$$

Note: We say a random variable X is ‘regular’ if $P(-\infty < X < \infty) = 1$. In a **cure model**, we are typically interested in an irregular survival time T where $P(0 \leq T < \infty) < 1$. That is,

$$\boxed{S(\infty) > 0}$$

In reality, it means

$$\boxed{S(\text{large no.}) > 0}$$

Example. $\lambda(t) = \lambda_0$, a positive constant, is hazard function for a regular survival time.

Example. $\lambda(t) = \lambda_0 + \lambda_1 t$, with $\lambda_0, \lambda_1 > 0$, is hazard function for a regular survival time.

Example. $\lambda(t) = e^{-\theta t}$, $\theta > 0$, is **not** a regular hazard function because

$$\int_0^\infty \lambda(u)du = \left[-\theta^{-1} \cdot e^{-\theta u}\right]_0^\infty = \theta^{-1} < \infty \ .$$

This can be a hazard function for a “cure model”.

Cure model. If a disease has 'cure'; that is, we assume $P(0 \leq T < \infty) = \gamma < 1$ and $P(T = \infty) = 1 - \gamma$, which implies that $\Lambda(\infty) < \infty$. This is allowed since T is not a 'regular random variable'. In this cure model, the pdf of T is

$$f_T(t) = I(0 \leq t < \infty)\gamma f_{T_0}(t) + I(t = \infty)(1 - \gamma)$$

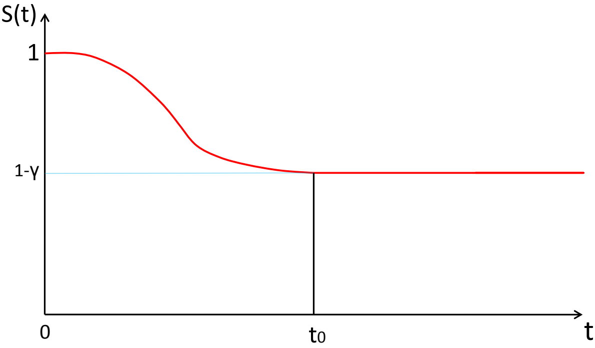
where $f_{T_0}(t)$ is the pdf for a regular survival time variable T_0. This is a mixture model.

Real life example: Consider ovarian cancer. Define T as the time from cancer diagonosis to death if $T \leq 10$ years and set $T = \infty$ if the patient survives longer than 10 years. Here $T = \infty$ represents a case of 'being cured.' Suppose $P(0 \leq T \leq 10) = \gamma < 1$. Then $P(T = \infty) = 1 - \gamma > 0$ and

$$f_T(t) = I(0 \leq t \leq 10)\gamma f_{T_0}(t) + I(t = \infty)(1 - \gamma)$$

where $f_{T_0}(t)$ is the pdf for a regular survival time variable T_0 with values ranging from 0 to 10 years. For exploratory analysis. we could be interested in knowing/estimating $S_{T_0}(t)$ and γ.

Survival curve for cure model



Remark: Nonparametric and semiparametric models could include cure models as special cases, such as the Kaplan-Meier estimator and the proportional hazards model that we'll see in later chapters. So we do not need to worry about if the underlying model is a regular or cure model.

Still, cure model itself is important as some of the parameters in the models may not be identifiable by standard nonparametric or semiparametric approaches.

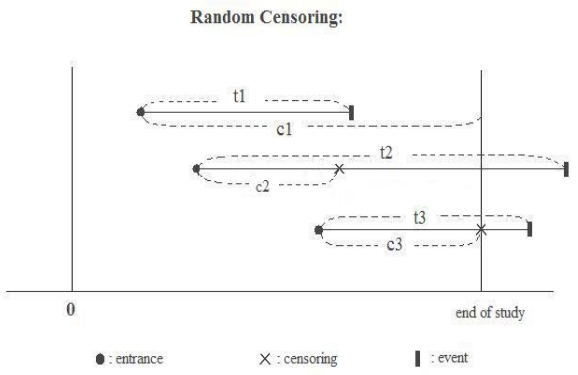
1.4 Censoring

Example. Suppose we have a 5-year longitudinal follow-up study, and that in the first 2 years, we recruit patients that are diagnosed with cancer. We then follow these patients and the event of interest is death. Say a patient was recruited at the end of the second year and remained alive throughout the study.

- We say that a subject is censored if we do not observe event of interest till the end of our follow-up period.
- The length of the (potential) observation window (3 years) is called the censoring time.
- More generally, censoring is a special missing data mechanism, under which the time-to-event is partially known.

Random censoring. This type of censoring will be the main censoring mechanism that we deal with in this course. It occurs when the censoring time varies from individual to individual and is unknown in advance. For example, in a follow-up study, the censoring occurs due to the end of the study, loss to follow-up, or early withdrawals.

Reasons for censoring – failures occur after the end of study (administrative censoring)
– patients drop out of study
– other reasons for loss to follow-up



We study right censoring in this course.

Theoretical setting. Suppose C is the censoring variable. Assume T and C are independent (the so-called 'independent censoring'). Define

$$Y = \min(T, C) = \begin{cases} T & \text{if} \quad T < C \\ C & \text{if} \quad T \geq C \end{cases}$$

and the censoring indicators

$$\Delta = I(T < C) = \begin{cases} 1 & \text{if data is uncensored,} \quad T < C \\ 0 & \text{if data is censored,} \quad T \geq C \end{cases}$$

Data. Assume $(Y_1, \Delta_1), (Y_2, \Delta_2), \ldots, (Y_n, \Delta_n)$ are iid copies of (Y, Δ). Under random censoring, what is the actually observed data? Ideally, we would like to observe the "complete data" $t_1, t_2, \ldots, t_n$. Due to censoring, we only observe "right-censored data" (y_1, δ_1), $(y_2, \delta_2), \ldots, (y_n, \delta_n)$ and possibly some covariate information.

Example. A set of observed survival data is

y_i	25	18	17	22	27
δ_i	1	0	1	0	1

The data can also be presented as

$$25 \quad 18^+ \quad 17 \quad 22^+ \quad 27$$

1.5 Probability Properties

Intuitively, the random variable Y tends to be 'shorter' than the survival time of interest, T. This is clear upon observing $Y = \min\{T, C\}$. Under the assumption that T and C are independent, the survival function of Y is

$$\begin{aligned} S_{_Y}(y) &= P(T \geq y, C \geq y) = P(T \geq y)P(C \geq y) \\ &= S_{_T}(y)S_{_C}(y) \leq S_{_T}(y) \ . \end{aligned}$$

Thus, as compared with $S_{_T}$, $S_{_Y}$ assigns more probability to smaller values.

Example. Suppose the censoring time is a fixed constant, $C = c_0$, $c_0 > 0$. Then the survival function of Y is $S_{_Y}(y) = S_{_T}(y)$ if $y \leq c_0$, and $S_{_Y}(y) = 0$ if $y > c_0$.

Example. Suppose $T \sim \mathrm{Exp}(\theta)$, $\theta > 0$, and $C \sim \mathrm{Unif}(0, \beta)$, $\beta > 0$. Then the survival function of Y is

$$S_{_Y}(y) = \begin{cases} 1 & \text{if} \quad y \leq 0 \\ e^{-\theta y}\left(\frac{\beta - y}{\beta}\right) & \text{if} \quad 0 < y \leq \beta \\ 0 & \text{if} \quad y > \beta \end{cases}$$

Property. Suppose T_0 and T_1 are continuous survival time variables. Denote by $\lambda_0(t)$ and $\lambda_1(t)$ the hazard function of T_0 and T_1, respectively. Then $\lambda_0(t) \geq \lambda_1(t)$ for all $t > 0$ implies $S_0(t) \leq S_1(t)$ for all $t > 0$.

Proof. Note that $\lambda_0(t) \geq \lambda_1(t)$ for all $t > 0$ implies

$$\Lambda_0(t) = \int_0^t \lambda_0(u)du \geq \int_0^t \lambda_1(u)du = \Lambda_1(t)$$

and

$$S_0(t) = e^{-\Lambda_0(t)} \leq e^{-\Lambda_1(t)} = S_1(t) \ .$$

Remark 1: $S_0(t) \leq S_1(t)$ for all $t > 0$ does NOT imply $\lambda_0(t) \geq \lambda_1(t)$ for all $t > 0$. So, it is possible that $S_0(t) \leq S_1(t)$ holds while there are cross-overs between hazards $\lambda_0(t)$ and $\lambda_1(t)$.

Remark 2: This property has been useful when statisticians developed statistical tests for hypothesis testing problems.

1.6 Other common sampling constraints

Interval censoring

The survival time t_i falls in an interval (ℓ_i, r_i) and observe only (ℓ_i, r_i). For example, let $T =$ time from treatment onset to disease onset. The onset of disease falls in the interval formed by two successive clinical visits. Let

$\ell_i =$ time of the visit when the ith patient is seen to be free of disease for the last time.

$r_i =$ time of the visit when the ith patient is seen to be diseased for the first time.

The best knowledge we have about the true survival time $T_i = t_i$ is $\ell_i < t_i \leq r_i$. If the disease is not present for all the visits, then $r_i = \infty$.

Data: $(\ell_1, r_1), \ldots, (\ell_n, r_n)$

Left truncation and right censoring

The presence of left truncation is usually due to the prevalent sampling scheme, that is, drawing samples from a disease prevalent population. Right censoring is encountered for the usual reasons (loss to follow-up etc.).

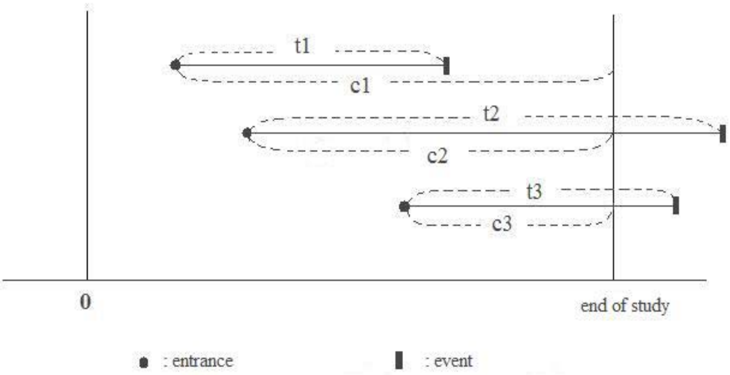
Example. Survival time $T =$ time from onset (or diagnosis) of ovarian cancer to death. A prevalent cohort includes a group of women who have developed ovarian cancer at the time of recruitment. Those with ovarian cancer who died before the recruiting time are excluded from the study. The study tends to recruit women with longer survival times.

*Appendix

More Types of Incomplete Data

Type-I Censoring. Type-I censoring occurs when a survival time t_i exceeds a pre-determined censoring time c_i. The censoring time c_i is considered as a constant in the study. For example, a clinical treatment study starts at the calendar time a and ends at b. Patients could enter the study at different calendar times. The survival time is the time between the start of treatment (entry) to a certain event. Assume no loss to follow-up. In this case, c_i is the time from entry to b. The actual survival time t_i cannot be observed if $t_i > c_i$.

Type-I Censoring:



Type-II Censoring. This type of censoring is frequently encountered in industrial applications. From n ordered survival times, only the first $r(r \leq n)$ times are observed, others are censored.

For example, put 100 transistors on test at the same time and stop the experiment when 50 transistors burn out. In this example, $n = 100$ and $r = 50$. Let $t_{(1)}, t_{(2)}, \ldots, t_{(50)}$ be the first 50 survival times. Note that $t_{(50)}$ is an estimate of the median survival time.

Left censoring

The survival time t_i could be too small to be observed. For example, consider a study in which interest centers on the time to recurrence of a particular cancer following surgical removal of the primary tumor. A few months after the operation, the patients are examined to determine if the cancer has recurred. Cancer recurrence is then monitored closely after the first examination. Let $T =$ time from operation to the recurrence of cancer. Some of the patients at this time may be found to have a recurrence and thus the actual time is less than the time from operation to the examination. These cases are said to be left censored. Note that in addition to left censoring, the observations could be subject to right censoring as well in the subsequent follow-up study.

Right truncation

The survival time t_i is too large to be included in data. A well known example is the reported AIDS incidences. The origin of HIV-AIDS disease started around late 70th or early 80th. In this example, $T =$ time from HIV infection to diagnosis of AIDS. An AIDS incidence is reported to a health institution only when AIDS develops. Those cases where AIDS occur after the closing date of data collection are excluded from the data set.

Current status data

In the above example, if there is no further follow-up of patients, then only the "status" of the cancer recurrence (i.e., yes or no) is known and the observations are

$$(w_1, \delta_1), (w_2, \delta_2), \ldots, (w_n, \delta_n)$$

where w_i is the time from operation to examination, and δ_i is the status of patient i.