Understanding Survival Analysis

The Statistical Study of Time, Risk, and Outcomes

Suchibrata

Introduction

Time isn't just ticking it's talking

Start/Follow-up Time -----> **Event** (Failure) [Death, Disease, Relapse, Recovery]

Outcome Variable (Survival Time)

[Years, months, weeks or even days from the beginning of the follow-up of an individual until an event occurs.]

Major Problem - Censoring !

We don't know the exact time when something happened. Only the part of the information is available to us.

Reason for Censoring

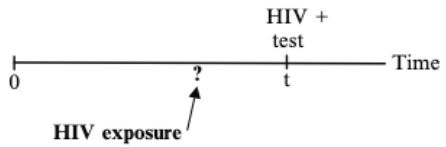
1. Study ends with no events.
   Example: A patient still in remission when the study wraps up.
2. Lost To Follow-up.
   (moved away, dropped out)
3. Withdrawn from the Study.

Right Censoring

We know the event didn't happen until a certain time, but we don't know when it exactly happened.

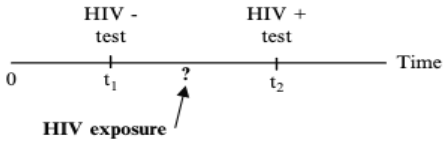
Left Censoring

We know the event already happened, but don't know exactly when.



Interval Censoring

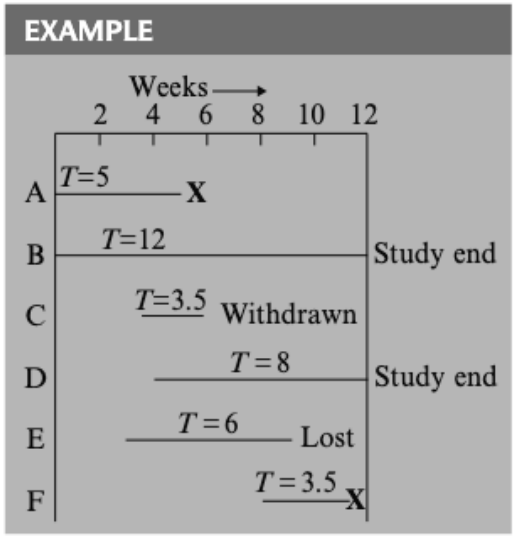
We know the event happened between two times, but not the exact moment.



Examples

1. Leukemia patients/time in remission.
(A study that follows leukemia patients in remission over several weeks to see how long they stay in remission)

2. Elderly Population (60+) /time until death. (13+ years follow-up of an elderly population (60+ years) to see how long remain alive.)



Notation:

T : Random Variable denoting a person's Survival Time.

t : Specific Value of interest of our Random Variable 'T'

d : Dichotomous variable denoting failure or Censoring. 1:Censored, 0:Failure.

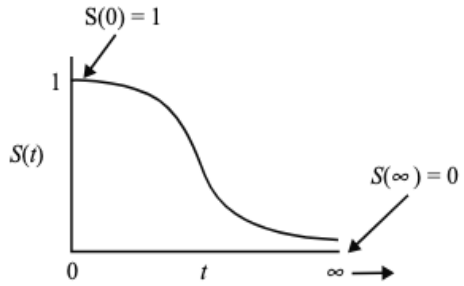
Terminology & Notation

Some Useful Metrics in Survival Analysis

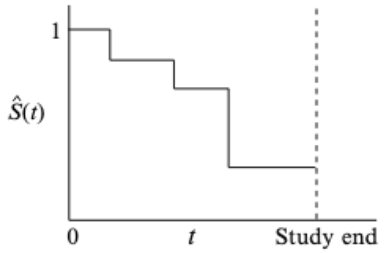
Survival Function

Intuitively, it is the probability that a person survives longer than some specified time 't'.
S(t) = P(T>t)

Theoretical $S(t)$:



$\hat{S}(t)$ in practice:



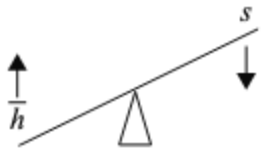
Properties

1. Survival functions are non-increasing. That is they head downwards as t increases.
2. At time t=0, S(t=0) = 1, i.e. at the beginning of the study, survival rate is the highest.
3. At time t=∞ , S(t=∞) = 0, i.e. if the study period increased without limit, eventually nobody would survive.

Relationship

$$S(t) = \exp\left[-\int_0^t h(u)du\right]$$

$$h(t) = -\left[\frac{d\,S(t)/dt}{S(t)}\right]$$



Hazard Function

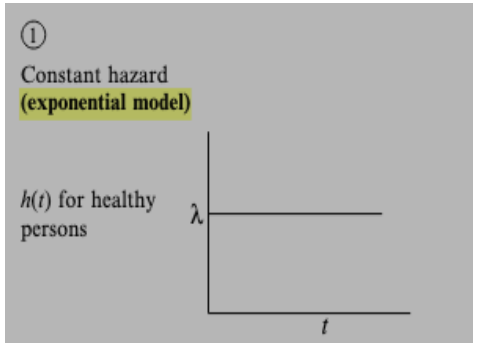
(Conditional Failure Rate)

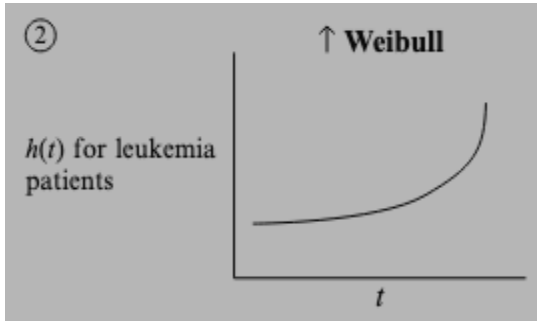
Gives the instantaneous potential per unit time for the event to occur, given that the individual

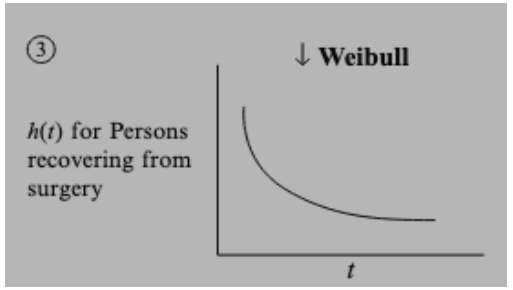
$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t}$$

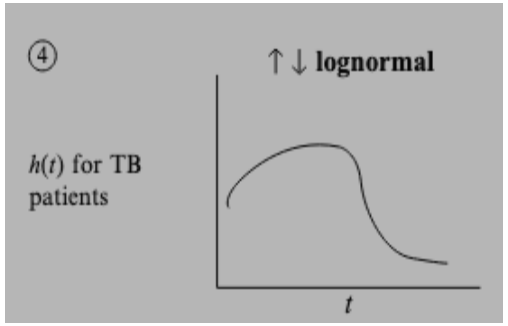
Properties

1. Hazard functions is always non negative.
2. It has No Upper bound.
3. People often think this as a probability, but this is nothing but a rate.









Des. Measures & their Glitches

Some useful Statistic to perform Analysis

Survival Function

leukemia trials conducted by the University of California, Berkeley [Berkson, Dr. David Cox]

Remission times (in weeks) for two groups of leukemia patients

Group 1 (Treatment) $n = 21$	Group 2 (Placebo) $n = 21$
6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23
$\bar{T}_1$ (ignoring + 's) = 17.1	$\bar{T}_2 = 8.6$
$\bar{h}_1 = \frac{9}{359} = .025$	$\bar{h}_2 = \frac{21}{182} = .115$

$$\text{Average hazard rate } (\bar{h}) = \frac{\#\text{failures}}{\sum_{i=1}^{n} t_i}$$

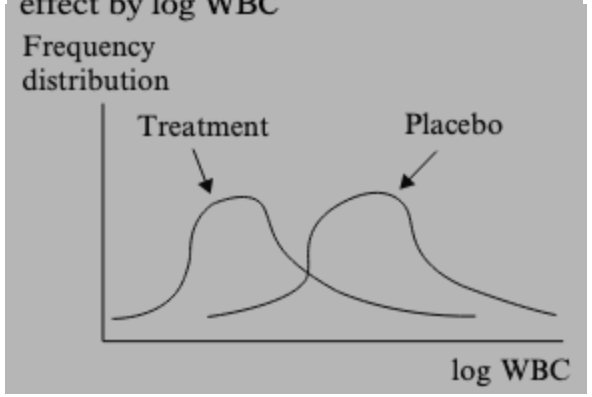
Using average hazard rates, we again see that the treatment group appears to be doing better overall than the placebo group; that is, the treatment group is less prone to fail than the placebo group.

Group 1		Group 2	
t (weeks)	log WBC	t (weeks)	log WBC
6	2.31	1	2.80
6	4.06	1	5.00
6	3.28	2	4.91
7	4.43	2	4.48
10	2.96	3	4.01
13	2.88	4	4.36
16	3.60	4	2.42
22	2.32	5	3.49
23	2.57	5	3.97
6+	3.20	8	3.52
9+	2.80	8	3.05
10+	2.70	8	2.32
11+	2.60	8	3.26
17+	2.16	11	3.49
19+	2.05	11	2.12
20+	2.01	12	1.50
25+	1.78	12	3.06
32+	2.20	15	2.30
32+	2.53	17	2.95
34+	1.47	22	2.73
35+	1.45	23	1.97

Confounding Effect

The table at the left gives the remission survival times for the two groups with additional information about white blood cell count for each person studied. In particular, each person's log white blood cell count is given next to that person's survival time.

Treatment group: $\overline{\log \text{WBC}} = 1.8$
Placebo group: $\overline{\log \text{WBC}} = 4.1$
Indicates **confounding** of treatment effect by log WBC



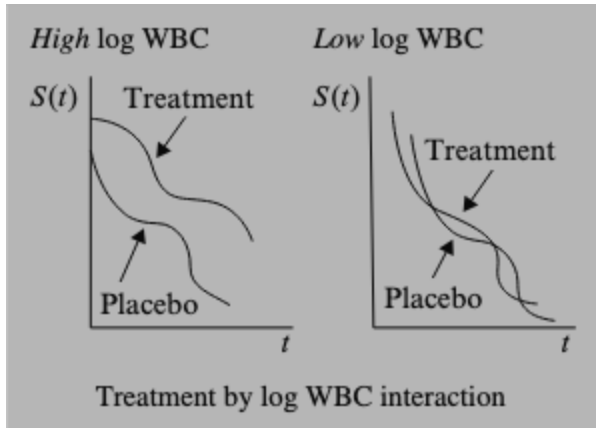
The treatment group may appear to survive longer due to low log WBC, not treatment efficacy — indicating the treatment effect is confounded by log WBC.

Des. Measures & their Glitches

Some useful Statistic to perform Analysis

Interaction Effect

Interaction means the treatment's effect varies by log WBC levels. For high log WBC, treatment improves survival over placebo; for low log WBC, there's no difference. This indicates a strong treatment-log WBC interaction - treatment efficacy depends on log WBC.



Alt. way of analysis

1. Stratify on log WBC.
2. Use math modelling.
   Ex - proportional hazard model

Remission times (in weeks) for two groups of leukemia patients

Group 1 (Treatment) $n = 21$	Group 2 (Placebo) $n = 21$
6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23
$\bar{T}_1$ (ignoring +'s) = 17.1	$\bar{T}_2 = 8.6$
$\bar{h}_1 = \frac{9}{359} = .025$	$\bar{h}_2 = \frac{21}{182} = .115$

Average hazard rate $(\bar{h}) = \frac{\#\text{ failures}}{\sum_{i=1}^{n} t_i}$

	Individual #	t (weeks)	d	X_1 (Group)	X_2 (log WBC)
Group 1	1	6	1	1	2.31
	2	6	1	1	4.06
	3	6	1	1	3.28
	4	7	1	1	4.43
	5	10	1	1	2.96
	6	13	1	1	2.88
	7	16	1	1	3.60
	8	22	1	1	2.32
	9	23	1	1	2.57
	10	6	0	1	3.20
	11	9	0	1	2.80
	12	10	0	1	2.70
	13	11	0	1	2.60
	14	17	0	1	2.16
	15	19	0	1	2.05
	16	20	0	1	2.01
	17	25	0	1	1.78
	18	32	0	1	2.20
	19	32	0	1	2.53
	20	34	0	1	1.47
	21	35	0	1	1.45

Problem

Compare two groups after adjusting for confounding and interaction.

we are now considering two explanatory variables in our extended example, whereas we previously considered a single variable, group status. The data layout for the computer needs to reflect the addition of the second variable, log WBC.

X1 – Primary interest
X2 – Extraneous variable used to get over the Confounding and/or Interaction Effect.

Censoring Assumptions

The assumptions about the censoring

Random Censoring

Censoring is **non-informative** and unrelated to the future failure time,
i.e. if T = failure time, C = censoring time,
then random censoring implies
T⫫C

Interpretation

At any time t, the subjects censored are representative of all those still at risk with respect to survival probability.

Example

In a heart disease study, if 5 out of 50 patients drop out at 6 months, we assume their risk of dying after 6 months is the same as the remaining 45.

Failure rate	
Censored	Not censored
$h_{Ce}(t) = h_{NCe}(t)$	

Independent Censoring

Censoring is independent of survival time within any subgroup defined by covariates Z.(T⫫C|Z)

Interpretation

Censoring may depend on group characteristics (e.g., age, gender), but within those groups, it's still non-informative.

Example

Among diabetic patients aged 60+, if some are censored at 1 year, we assume their risk of failure is the same as the uncensored patients in that same age - disease group.

	Failure rate	
Subgrp	Censored	Not censored
A	$h_{A,Ce}(t) = h_{A,NCe}(t)$	
B	$h_{B,Ce}(t) = h_{B,NCe}(t)$	

Note :

random ⇒ independent
Independent ⇏ random

Non Informtive Censoring

The **censoring mechanism** is **ignorable** in the likelihood function when estimating survival parameters.

Equivalence

In most classical settings, non-informative censoring implies independent censoring, and vice versa.

Example

If someone is censored after 3 years because the study ends, that doesn't tell you if they were likely to die at 3.1 years or 30 years. It's just a limit of observation.

Censoring Assumptions

The assumptions about the censoring

Independent Censoring

Formal Meaning : $T \perp C$

That is, the true event time T is independent of the censoring time C.

Implication

The reason someone is censored does not relate to how soon or late they would have experienced the event.

Example

A patient drops out of a medical trial because they move to another city - not because they are sicker or healthier.

Random

Formal Meaning : $T \perp C | Z$, where Z are covariates. So conditional on observed characteristics, the censoring is random.

Application:

You assume censoring behaves like a draw from a probability distribution, and not as a deterministic process.

Example: In a cohort study, people leave the study at random times due to unrelated life events - say, vacation, relocation, etc.

Independent Censoring

The **censoring mechanism** is **ignorable** in the likelihood function when estimating survival parameters.

Equivalence

In most classical settings, non-informative censoring implies independent censoring, and vice versa.

Example

If someone is censored after 3 years because the study ends, that doesn't tell you if they were likely to die at 3.1 years or 30 years. It's just a limit of observation.