

جلسه اول

Introduction to Survival Analysis

Types of Survival Analysis and Challenges in Learning Them

- The Kaplan-Meier
- Nelson Aalen curve
- Lifetable
- The log-rank test & Other Test to Comparison
- Cox proportional hazards regression
- Parametric models
- Accelerated failure time model
- Frailty models
- Recurrent Event Time Model
- Competing risk models
- Cure Models

Texts and Software

- **Required:**
- Survival Analysis, A Self-Learning Text: Third Edition
- David G. Kleinbaum & Mitchel Klein
- **SoftWare:** An Introduction Survival Analysis Using STATA.
- MARIO CLEVES, WILLIAM GOULD, ROBERTO G. GUTIERREZ, YULIA V. MARCHENKO
- Example analyses will be in STATA.
- You may use any software capable of the analysis,

Time to Event Data

- Survival Analysis is a term for analyzing time-to-event data.
- This is used in clinical and epidemiological studies, where the event is often death or incidence or recurrence of disease.
- It is used in engineering **reliability analysis**, where the event is often failure of a device or system.
- It is used in insurance, particularly life insurance, where the event is death, disability, or damage from an accident.

Time to Event Data

- The distribution of 'failure' times is usually asymmetric and can be long-tailed.
- The base distribution is not normal, but exponential. This is the simplest distribution to model failure time data.
- There are often **censored & truncated** observations, which are ones in which the failure time is not observed.
- Typically, this is because the failure has not yet happened, though there are other patterns.

Time to Event Data

- Often, these are **right-censored**, meaning that we know that the event occurred after some known time t , but we don't know the actual event time, as when a patient is still alive at the end of the study.
- Observations can also be **left-censored**, meaning we know the event has already happened at time t , **or interval-censored**, meaning that we only know that the event happened between times t_1 and t_2 .
- Analysis is difficult if censoring is associated with treatment or other predictors of the event in question.

Pre Assumptions in Survival Analysis

- An implicit concept in the definition of censoring is that, if the study had been prolonged (or if subjects had not dropped out), eventually the outcome of interest would have been observed to occur for all subjects.
- Conventional statistical methods for the analysis of survival data make the important assumption of **independent** or **non informative censoring**.
- This means that, at a given point in time, subjects who remain under follow up have the same future risk for the occurrence of the event as those subjects no longer being followed (either because of censoring or study dropout), as if losses to follow-up were **random** and thus **non informative**.

Objectives of survival analysis

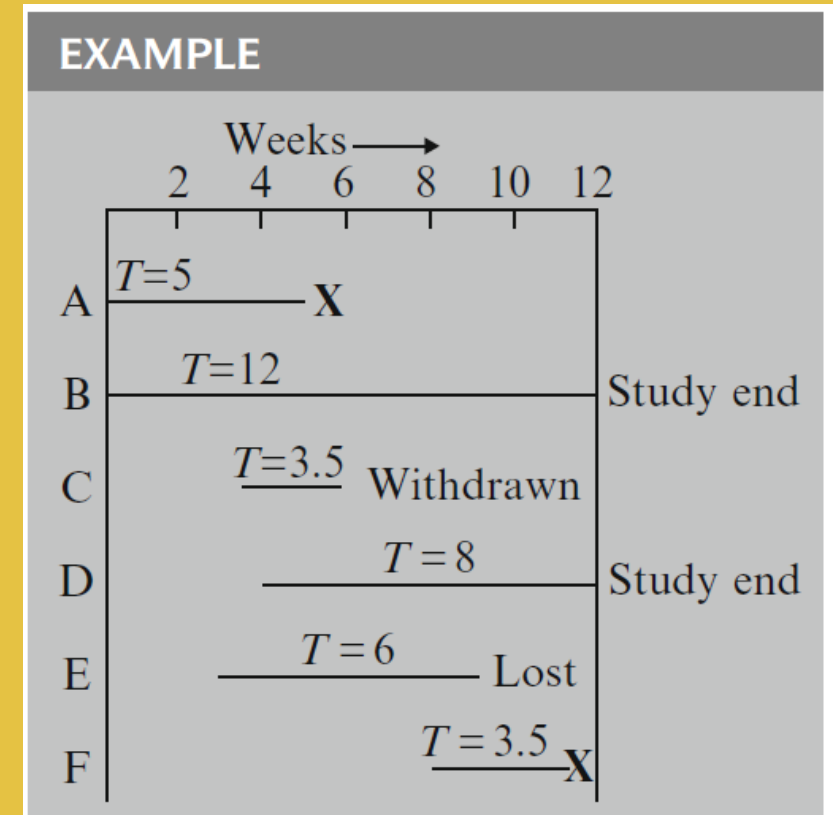
- **Estimate time-to-event for a group of individuals**, such as time until second heart-attack for a group of MI patients.
- **To compare time-to-event between two or more groups**, such as treated vs. placebo MI patients in a randomized controlled trial.
- **To assess the relationship of co-variables to time-to-event**, such as: BMI, insulin resistance, or cholesterol influence survival time of MI patients?

Note: expected time-to-event = $1/\text{incidence rate}$

This equation applies when the failure rate is “constant” over time. However, the failure rate may be an "increasing function", a "decreasing function" or a "bathtub curve" over time.

Survival Analysis: Terms

- **Time-to-event**: The time from entry into a study until a subject has a particular outcome.
- **Censoring**: Subjects are said to be censored if they are **lost to follow up** or **withdraws from the study**, or if the **study ends before they die** or have an **outcome of interest**. They are counted as **alive** or **disease-free** for the time they were enrolled in the study.



Why use Survival Analysis?

1. Why not compare mean time-to-event between your groups using a t-test (ANOVA) or linear regression?
 - ignores censoring
2. Why not compare proportion of events in your groups using risk/odds ratios or logistic regression?
 - ignores time

Why use Survival Analysis?

- An important distinction among modeling methods is the type of outcome variable being used. In survival analysis, the outcome variable is “time to an event,” and there may be censored data.
 - In linear regression modeling, the outcome variable is generally a continuous variable, like blood pressure.
 - In logistic regression, the outcome variable is a dichotomous variable, like CHD status, yes or no.
 - In Poisson regression, The outcome is a numerical variable, like the number of decay teeth.
- Note:** With linear, logistic or Poisson modeling, we usually do not have information on follow-up time available.

Measure of effect in linear, logistic, Poisson and survival analysis

- In Linear regression modeling, the measure of effect is usually some regression coefficient b .
- In Logistic modeling, the measure of effect is an odds ratio expressed in terms of an exponential of one or more regression coefficients in the model.
- In Poisson modeling, the measure of effect is usually on the log of the mean.
- In survival analysis, the measure of effect typically obtained is called a hazard ratio; as with the logistic model, this hazard ratio is expressed in terms of an exponential of one or more regression coefficients in the model.
- **Note:** The hazard ratio, although a different measure from an odds ratio, nevertheless has a similar interpretation of the strength of the effect. A hazard ratio of 1, like an odds ratio of 1, means that there is no effect; that is, 1 is the null value for the exposure–outcome relationship.
- A hazard ratio of 10, on the other hand, is interpreted like an odds ratio of 10; that is, the exposed group has ten times the hazard of the unexposed group. Similarly, a hazard ratio of $1/10$ implies that the exposed group has one-tenth the hazard of the unexposed group.

Review Question 1

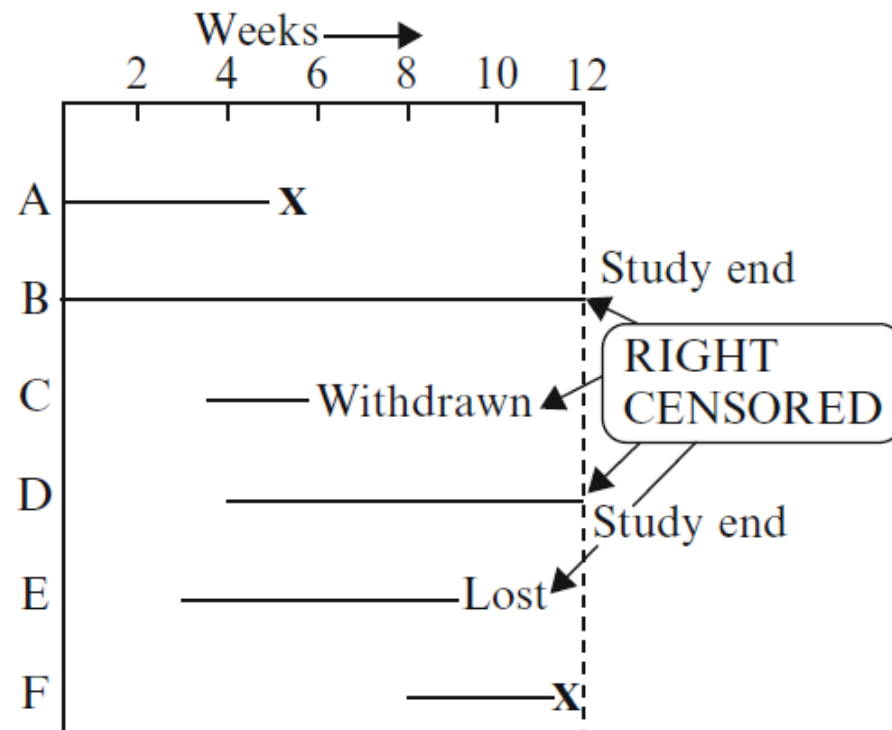
- Which of the following data sets is likely to lend itself to survival analysis?
 - a. A case-control study of caffeine intake and breast cancer.
 - b. A randomized controlled trial where the outcome was whether or not women developed breast cancer in the study period.
 - c. A cohort study where the outcome was the time it took women to develop breast cancer.
 - d. A cross-sectional study which identified both whether or not women have ever had breast cancer and their date of diagnosis.

Right Censoring

- Patients are in a clinical trial for cancer, some on a new treatment and some on standard of care.
- Some patients in each group have died by the end of the study.
- We know the survival time (measured for example, from time of diagnosis—each person on their own clock).
- Patients still alive at the end of the study are right censored.
- Patients who are lost to follow-up or withdraw from the study may be right-censored.

Right Censoring

Right-censored: true survival time is equal to or greater than observed survival time



Left & Interval Censoring

- when a person's true survival time is less than or equal to that person's observed survival time.
- An individual tests positive for HIV.
- If the event is infection with HIV, then we only know that it has occurred before the testing time t , so this is **left censored**.
- If an individual has a negative HIV test at time t_1 and a positive HIV test at time t_2 , then the infection event is **interval censored**.
- Interval-censoring actually incorporates both right-censoring and left-censoring as special cases.

Truncation

- We note that there are two types of left truncation at t_0 . The first type of left truncation occurs if the subject has the event before t_0 and thus is not included in the study.
- If, for example, the exposure (E) under study causes individuals to die before they could enter the study, this could lead to a (selective) survival bias that would underestimate the effect of exposure.
- The second type of left truncation occurs if the subject survives beyond time t_0 (i.e., $t > t_0$). This is required in order for the subject to have his/her survival time observed. Thus, a condition of the subject's entry into the study is that they survive until time t_0 .
- If they do not meet that condition, then their left truncation is of the first type and thus not included in the study.
- If they do survive past time t_0 , then their left truncation is of the second type.

Truncation vs Censoring

- Left truncation (of both types) at time t is commonly confused with left censorship at time t .
- If a subject is **left censored** at time t , then that subject is
 - (i) included in the study,
 - (ii) known to be event free at time 0,
 - (iii) known to be at risk for the event after time 0, and
 - (iv) known to have had the event before time t but with the exact time of event being unknown.

Censoring vs Truncation

- Suppose that all individuals in the population or sample of interest are observed, but if the event of interest happens before time t the exact time is not known, only that it occurred before time t , then the data are left censored at time t .
- If individuals in the population are **not observed** if the event occurs before time t , then the data are said to be left truncated at time t .
- This means that the potentially observed population is not the same as the actual population.

Censoring vs Truncation

- This is a form of selection bias and like other sources of selection bias one can account for it in the analysis to the extent possible, or redefine the population of interest. This requires careful thought beforehand.
- If UC Davis runs a clinical trial of patients diagnosed with stage 3 colon cancer at a regularly scheduled colonoscopy, then strictly speaking the population from which these patients come is not the set of people with stage 3 colon cancer.

Censoring vs Truncation

- The population here is that of patients who get regularly scheduled colonoscopies and who have stage 3 colon cancer at the time of the examination.
- This group may differ from the general population of the same age in income, occupation, race/ethnicity, and other factors which might influence the course of the disease.

Note: Right-truncated data typically occurs in registries. For example, a cancer registry includes only subjects who developed cancer, and thus survival data obtained from this registry will be right-truncated.

Censoring vs Truncation

- The internal validity of comparisons within the clinical trial between a new treatment and standard of care is not compromised.
- But the external validity of the study, the population that it can be generalized to, is compromised since all the patients are those who routinely receive colonoscopies.
- When patients in a study are selected by the time of the event, where patients with events occurring at an early age are implicitly excluded, both problems can occur.

Censoring vs Truncation

- Suppose that patients are selected for a clinical trial of a new cancer therapy intended to prevent recurrence from those who have been on standard of care for one year after diagnosis and have not had a recurrence.
- Then the population of interest perforce excludes those who had died or recurred before one year, and thus may exclude the sickest patients.

Censoring vs Truncation

- **Censoring period:**
 - during this period the subject is no longer under observation, but **she may experience** the event of interest
- **Truncation period**
 - during this period the subject is no longer under observation, but **she cannot experience** the event of interest

Truncation

- Similarly to censoring, there are 2 types of truncation
- **Left truncation or Delay entry time:** a subject **enters** the population at risk at some stage after the start of the study, and we know that there is no way that the event of interest could have occurred before this date.
- Example: We are interested in the years spent in retirement. However, some died before getting retired
- **Right truncation:** a subject **leaves** the population at risk at some stage after the start of the study, and we know that there is no way that the event of interest could have occurred after this date
- Example: we are interested in the years spent in marriage, married people who died is right truncation.

Mechanisms of Censoring

- An important assumption of most survival analysis methods is that censoring is not statistically associated with the treatment, covariates, or with the risk of the event. This assumption is crucial but can be hard to check.
- Description of censoring mechanisms requires a clock. Time 0 may be birth, or the beginning of the study, or the entry of an individual subject into the study.
- As we will see later, the progression of time may not be in physical time units, but in some related measure.

Types of Right Censoring

- **Type I Censoring** is when the censoring time of all subjects is fixed and known in advance, such as the known end of an animal study when the study start time is the same for all animals.
- We might study mice with explanted human tumor tissue. Six mice are controls and six receive a form of chemotherapy. At 12 weeks, mice that have not died are sacrificed so the time at which they would have died is unknown.
- All censored observations are censored at 12 weeks.

Types of Right Censoring

- If the censoring time is fixed in advance for each animal, but may differ between animals, then this is called **Progressive Type I Censoring**.
- We might have the same design as the previous slide, but at 6 weeks a pre-specified half of each group of mice is sacrificed to obtain pathology tissues at that time.
- Some of the censored observations are censored at 6 weeks and some at 12 weeks, and the possible censoring time is fixed for each animal at the start of the study.

Types of Right Censoring

- **Generalized Type I Censoring** is when the censoring time is the end of the study and is known in advance, but subjects enter the study at different times.
- Usually, we define time 0 as the time of entry for each subject, so that t is the time on study which may be different at the same physical time for different subjects.
- This might happen in human clinical trials with a known end date.

Types of Right Censoring

- **Type II Censoring** is when the end of the study is determined by the accumulation of a fixed number of events. For example, this is the case if a study using 100 mice ends at the death of the 25th mouse.
- This is often used in engineering reliability analysis in which one may not want to wait until the last unit on test has failed, but it is important for analysis that at least some failures have occurred.

Types of Right Censoring

- Censoring can also be the result of a random process which generally should be independent of the event process. For example, patients may drop out of a clinical trial or be otherwise lost to follow up.
- The chance of censoring from this random process should be statistically independent from the process leading to the main event (e.g., death), and this implies that it should be independent of the treatment or other variables used to predict the event.
- Complications can ensue if this is not true. We call this **noninformative censoring**

Types of Right Censoring

- A more formal way to represent random censoring is that, for each subject i , there are two associated random
- variables, T_i , the event time for subject i and C_i , the random censoring time. We observe only the smaller of the two times since each subject either has an event or is censored.
- If Z_i is a collection of potential predictors of the event time, then the censoring is non-informative if C_i is independent of T_i conditional on Z_i .

Example 1: Truncation vs Censoring

- Leukemia patients are given a drug or placebo. Survival time is the duration from remission to relapse.

The study ends at 52 weeks with some patients yet to relapse.

- A left censoring
- B right censoring
- C left truncation
- D right truncation

Example 2: Truncation vs Censoring

- College students are asked the age at which they first tried smoking.
Some answer: using it but forget when
- A left censoring
- B right censoring
- C left truncation
- D right truncation

Example 3: Truncation vs Censoring

- For patients who have been hospitalized for a heart attack, we are interested in testing whether a new treatment that they take after they have been discharged prolongs survival. A patient died in the hospital is
- A left censoring
- B **left truncation**
- C true event
- D interval censoring

Example 4: Truncation vs Censoring

- For patients who have been infected with the AIDS virus and developed AIDS by June 30, 1986 ,If we considered patients who have been developed AIDS prior to the end of the study period. Infected individuals who have yet to develop AIDS are not included in the sample, so this patients will be considered
- A left censoring
- B right censoring
- C left truncation
- D right truncation
- **Note:** In the presence of right-truncation only, Lagakos, Barraj, and De Gruttola (1988) describe an estimation procedure based on a reverse time scale that transforms right truncation into left-truncation. In the presence of both right-truncation and left-truncation or censoring, Alioum and Commenges (1996) present a more general approach using nonparametric maximum likelihood estimation (outside the scope of this course)

Censoring Assumptions

There are three assumptions about censoring often considered for survival data:

- **Independent censoring**
- **Random censoring**
- **Non-informative censoring**

Independent (VS. Non-Independent) Censoring

- The assumption of independent censoring is the **most useful** of the three types for drawing correct inferences that compare the survival experience of two or more groups.
- In particular, the presence of non-independent censoring typically **affects the validity** of one's estimated effect.
- Independent censoring essentially means that within any subgroup of interest, the subjects who are censored at time t should be **representative** of all the subjects in that subgroup who remained at risk at time t with respect to their survival experience.
- In other words, censoring is independent provided that it is **random** within any subgroup of interest.

Random (VS. Non-Random) Censoring

- Random censoring is a stronger assumption and more restrictive than independent censoring.
- To be more specific, random censoring essentially means that subjects who are censored at time t should be representative of all the study subjects who remained at risk at time t with respect to their survival experience.
- In other words, the failure rate for subjects who are censored is assumed to be equal to the failure rate for subjects who remained in the risk set who are not censored.
- more restrictive than independent,
- i.e., random  indep,
- whereas indep  random.

Non-Informative Censoring

- Non-informative censoring occurs if the distribution of survival times (T) provides no information about the distribution of censorship times (C), and vice versa.
- Otherwise, the censoring is informative. Note, however, that the data must still identify which subjects are or are not censored.
- The assumption of non-informative censoring is often justifiable when censoring is independent and/or random; nevertheless, these three assumptions are not equivalent.

T distribution $\overset{\text{no information}}{\nRightarrow}$ *C distribution*

Note: must still need to know which subjects are censored or not censored.

بخش دوم

Basic Quantities and Models

Basic Quantities and Models

- The probability density function $f(x)$ is defined as with any continuous distribution. For any short interval of time, it can be thought of as the relative chance that the event will occur in that short interval. The cumulative distribution function is

$$F(x) = \Pr(X \leq x) = \int_0^x f(x)dx$$

- For survival data, a more relevant quantity is the survival function

$$S(x) = 1 - F(x) = \Pr(X > x) = \int_x^{\infty} f(x)dx$$

- The survival function $S(x)$ is the probability that the event time is later than x . If the event in a clinical trial is death, then this is the fraction of the original population at time 0 that is still alive at time x ; that is, the fraction surviving to time x

Properties of the survival function

- it is constrained between 0 and 1
- it is a decreasing function of time, i.e.,
- * at time $t = 0$ all patients are alive
- * at time $t = \infty$ all patients have died
- **Note:** in some settings patients can be cured and thus, it may not be reasonable to assume that all patients would die from the disease under study
- A class of statistical models (aka Cure rate models) has been developed to deal with such phenomena.

The Hazard Function

- Another important function is the hazard function, which is the probability that the event will occur in the next very short interval, given that it has not occurred yet.

$$h(x) = \lim_{\Delta x \rightarrow 0} \frac{\Pr[x \leq X < x + \Delta x | X \geq x]}{\Delta x}$$

- The expression in the numerator is the probability of survival until at least time $x + \Delta x$ conditional on surviving until time x . This might be the chance of someone who has just turned 30 still being alive one day later.
- This might be the chance of someone who has just turned 30 still being alive one day later. You can see that this is different than the probability at birth of surviving until age 30 plus one day. The latter is clearly smaller.
- **Note:** the hazard is not a probability $\Rightarrow$ can be interpretable as the expected number of events per individual per unit of time
- it is always nonnegative, that is, equal to or greater than zero;
- but it can be (much) greater than 1, also it has no upper bound.

The Hazard Function

- The hazard function $h(t)$ gives the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t .
- Note that, in contrast to the survivor function, which focuses on not failing, the hazard function focuses on failing, that is, on the event occurring. Thus, in some sense, the hazard function can be considered as giving the opposite side of the information given by the survivor function.
- In the hazard formula, the conditional probability gives the probability that a person's survival time, T , will lie in the time interval between t and $t + Dt$, given that the survival time is greater than or equal to t . Because of the given sign here, the hazard function is sometimes called a **conditional failure rate**.

The Hazard Function

- We now explain why the hazard is a rate rather than a probability. Note that in the hazard function formula, the expression to the right of the limit sign gives the ratio of two quantities.
- The numerator is the conditional probability and the denominator is Δt , which denotes a small time interval. By this division, we obtain a probability per unit time, which is no longer a probability but a rate.
- In particular, the scale for this ratio is not 0 to 1, as for a probability, but rather ranges between 0 and infinity, and depends on whether time is measured in days, weeks, months, or years, etc.

Hazard function $\equiv$ conditional failure **rate**

$$\lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$

Probability per unit time

Rate: 0 to ∞

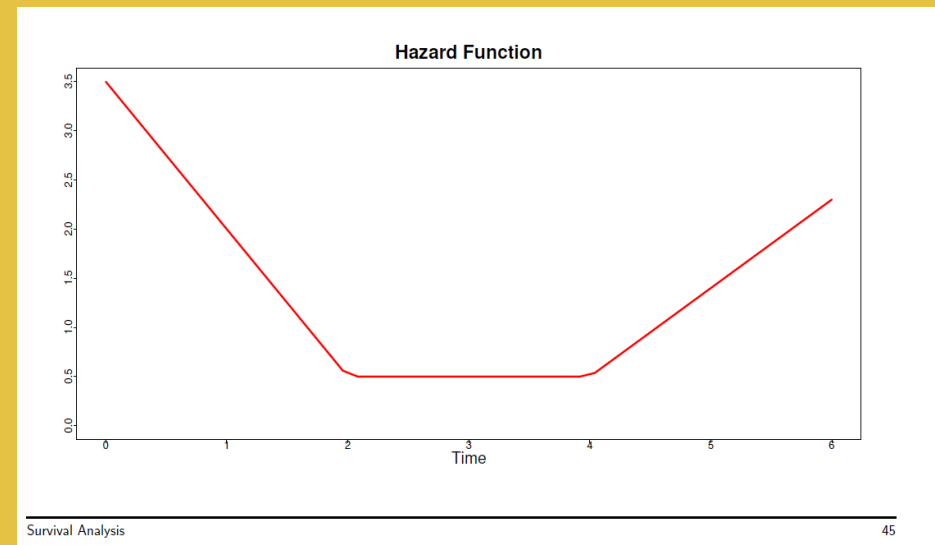
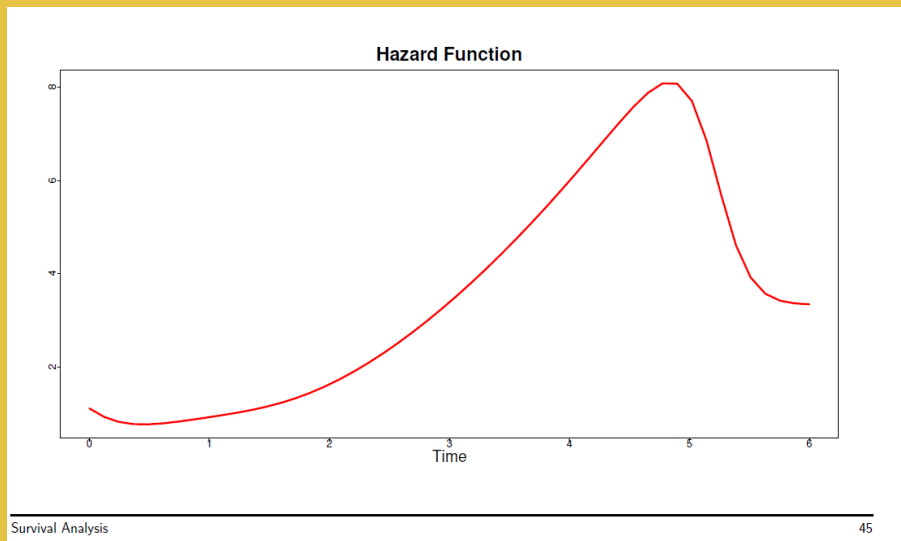
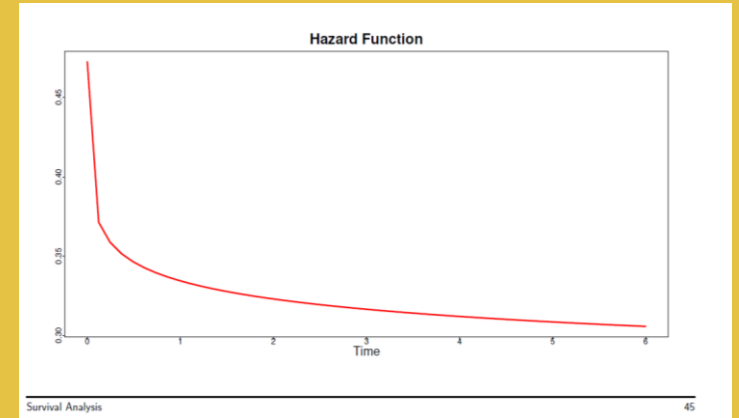
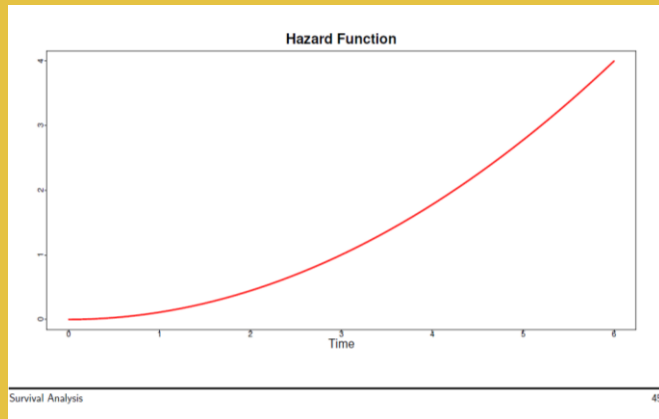
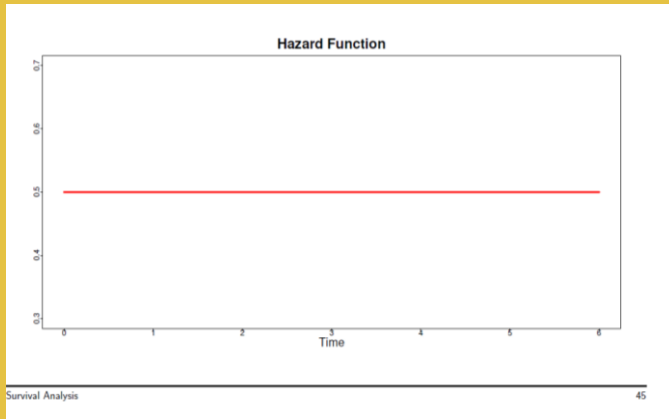
Relationship between Survival and Hazard Functions

- The limit takes the difference quotient into a derivative (by definition of the derivative) and the result is because the density $f(x)$ is the derivative of the CDF $F(x)$.

$$\begin{aligned}h(x) &= \lim_{\Delta x \rightarrow 0} \frac{\Pr[x \leq X < x + \Delta x | X \geq x]}{\Delta x} \\&= S^{-1}(x) \lim_{\Delta x \rightarrow 0} \frac{\Pr[x \leq X < x + \Delta x]}{\Delta x} \\&= f(x)/S(x)\end{aligned}$$

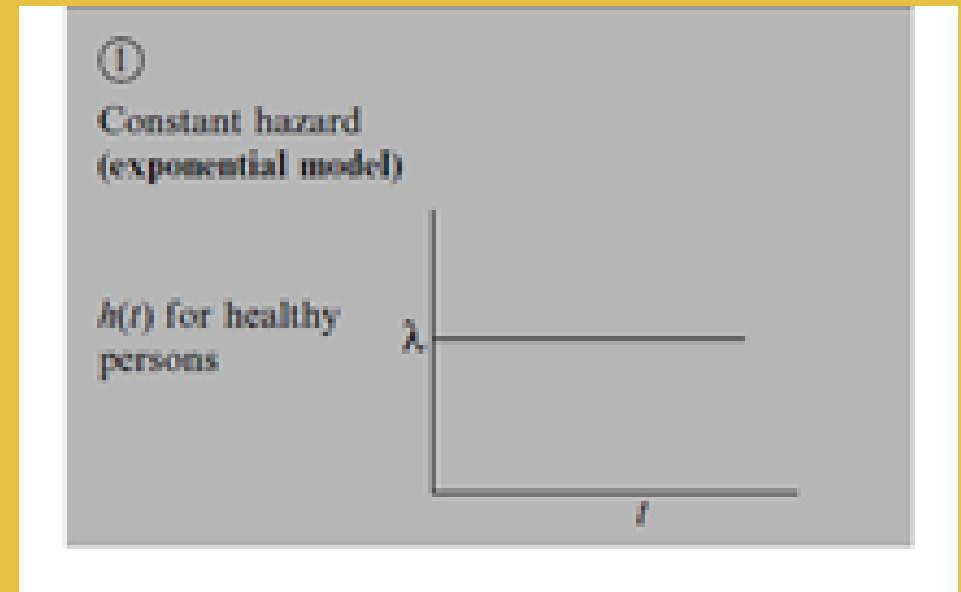
$$\begin{aligned}f(x) &= -\frac{dS(x)}{dx} \quad \text{Because } F' = f \\h(x) &= -\frac{d \ln(S(x))}{dx} = -S^{-1}(x) \frac{dS(x)}{dx}\end{aligned}$$

The Hazard Function



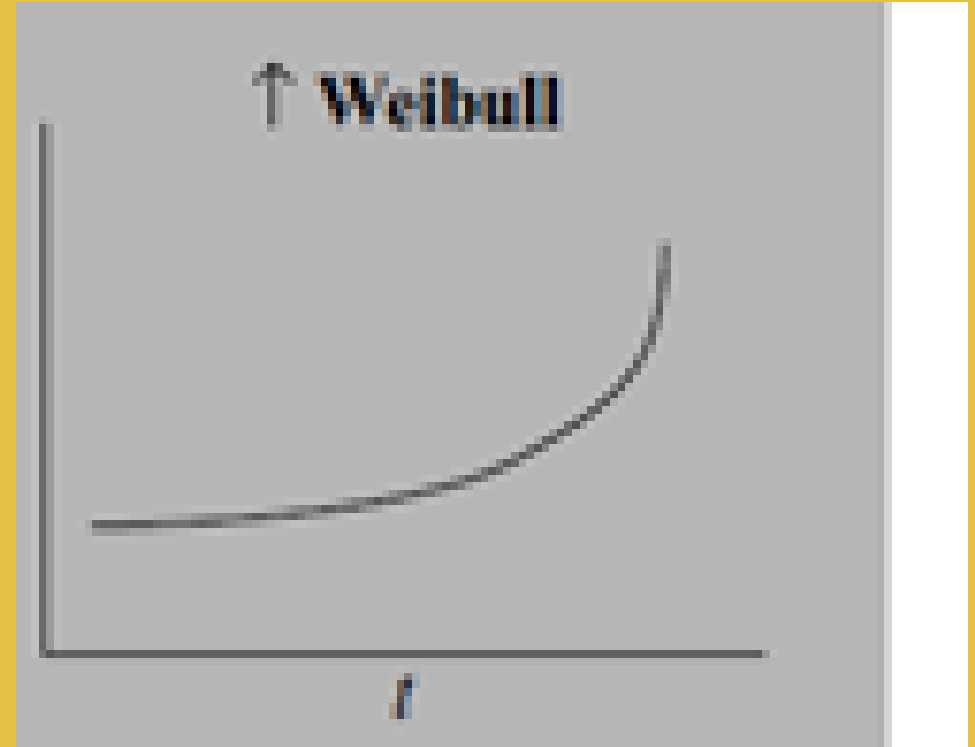
The Hazard Function

- Constant Hazard
- In this graph, no matter what value of t is specified, $h(t)$ equals the same value.
- Note that for a person who continues to be healthy throughout the study period, his/her instantaneous potential for becoming ill at any time during the period remains constant throughout the follow-up period.
- When the hazard function is constant, we say that the survival model is exponential. This term follows from the relationship between the survivor function and the hazard function.
- We will return to this relationship later.



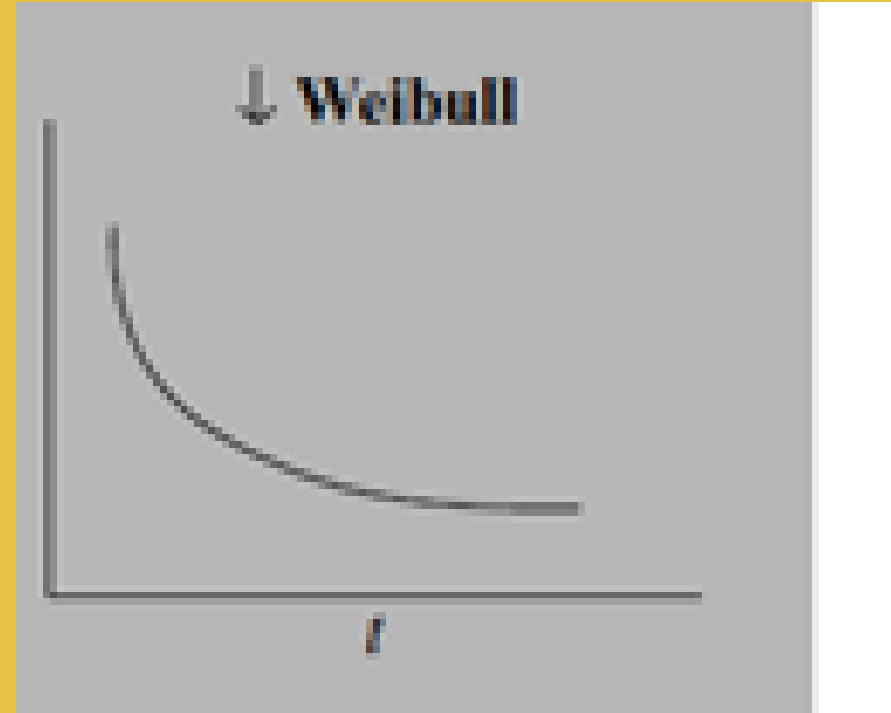
The Hazard Function

- The hazard function that is increasing over time.
- An example of this kind of graph is called an increasing Weibull model. Such a graph might be expected for cancer patients not responding to treatment, where the event of interest is death.
- As survival time increases for such a patient, and as the prognosis accordingly worsens, the patient's potential for dying of the disease also increases.



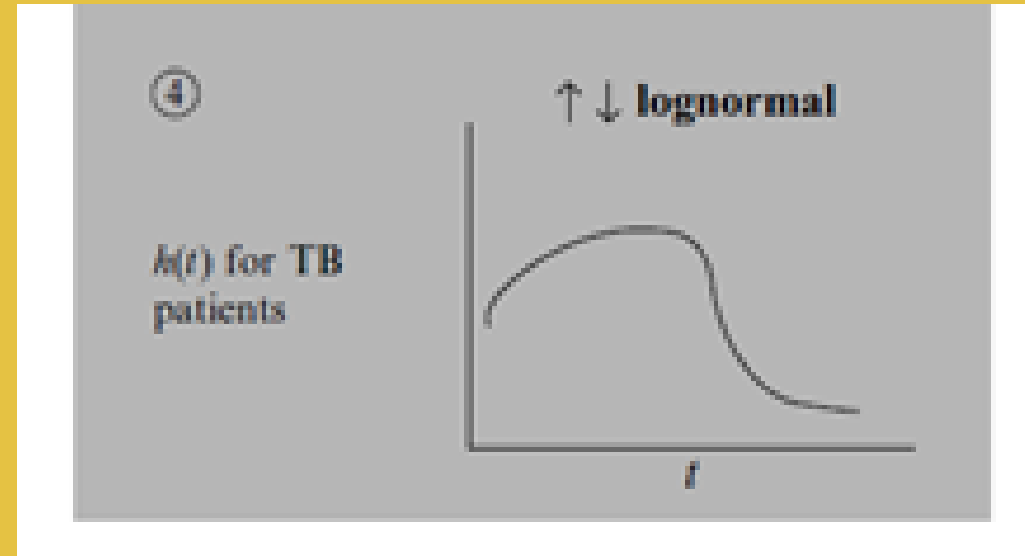
The Hazard Function

- The hazard function is decreasing over time. An example of this kind of graph is called a decreasing Weibull.
- Such a graph might be expected when the event is death in persons who are recovering from surgery, because the potential for dying after surgery usually decreases as the time after surgery increases.



The Hazard Function

- The Hazard function that is first increasing and then decreasing.
- An example of this type of graph is the lognormal survival model.
- We can expect such a graph for tuberculosis patients, since their potential for dying increases early in the disease and decreases later.



The Hazard Function

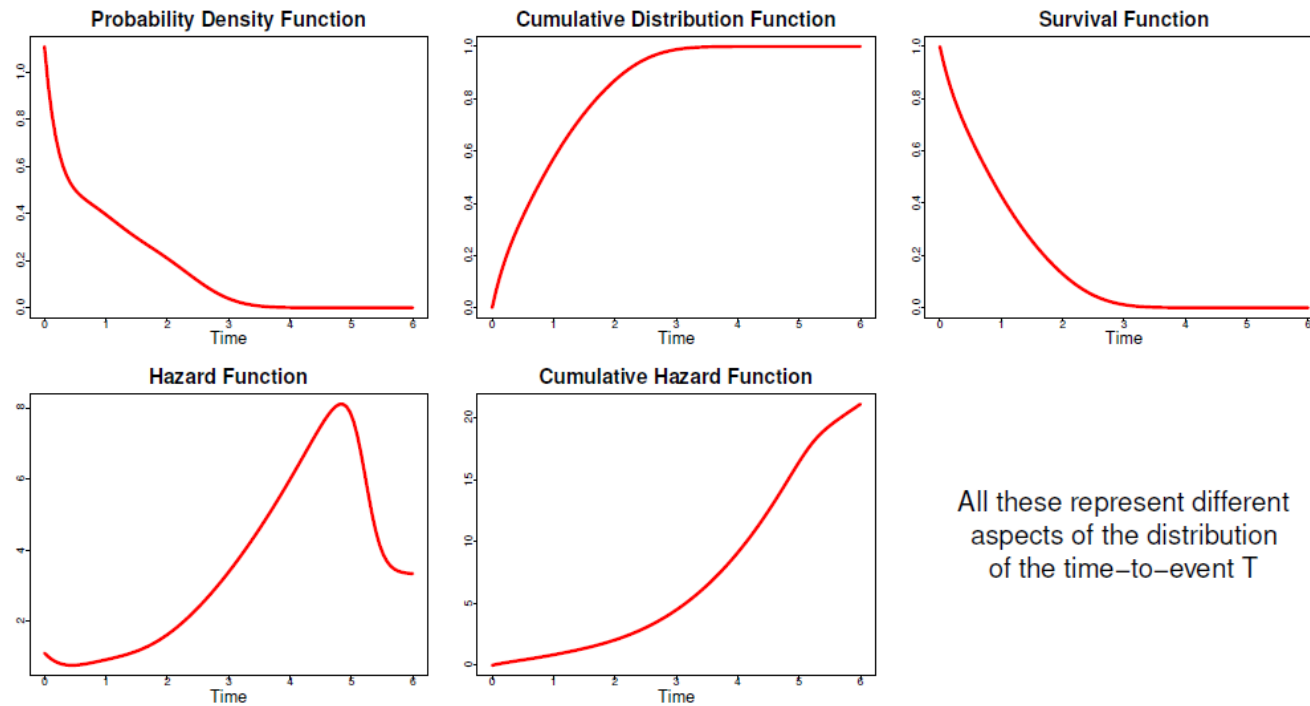
The hazard function $h(t)$ has the following characteristics:

- it is always nonnegative, that is, equal to or greater than zero;
- it has no upper bound.
- These two features follow from the ratio expression in the formula for $h(t)$, because both the probability in the numerator and the Dt in the denominator are nonnegative, and since Dt can range between 0 and 1.
- In contrast to a survivor function, the graph of $h(t)$ does not have to start at 1 and go down to zero, but rather can start anywhere and go up and down in any direction over time.

Cumulative Hazard

- The cumulative hazard function is the expected number of events that have occurred by time t $H(x) = \int_0^x h(t)dt = -\ln(S(x))$
- This function is easier to estimate than the hazard function, and we can then approximate the hazard function by the approximate derivative of the cumulative hazard.
- Note: this is not a (cumulative) probability has to be positive
- increasing function of t (as the time progresses we expect more events to have occurred)

Different Aspects of Data



All these represent different aspects of the distribution of the time-to-event T

Survival and Hazard Function

- The survivor function is more naturally appealing for analysis of survival data, simply because $S(t)$ directly describes the survival experience of a study cohort.
- However, the hazard function is also of interest for the following reasons:
- It is a measure of instantaneous potential whereas a survival curve is a cumulative measure over time;
- It may be used to identify a specific model form, such as an exponential, a Weibull, or a lognormal curve that fits one's data
- It is the tool by which mathematical modeling of survival data is carried out; that is, the survival model is usually written in terms of the hazard function.

Survival and Hazard Function

- Regardless of which function $S(t)$ or $h(t)$ one prefers, there is a clearly defined relationship between the two.
- In any actual data analysis, a computer program can make the numerical transformation from $S(t)$ to $h(t)$, or vice versa, without the user ever having to use either formula.
- The point here is simply that if you know either $S(t)$ or $h(t)$, you can get the other directly.

$$S(t) = \exp\left[-\int_0^t h(u)du\right]$$
$$h(t) = -\left[\frac{dS(t)/dt}{S(t)}\right]$$

Summary Notation

- The key notation is T for the survival time variable, t for a specified value of T , and d for the dichotomous variable indicating event occurrence or censorship.
- The key terms are the survivor function $S(t)$ and the hazard function $h(t)$, which are in essence opposed concepts, in that the survivor function focuses on surviving whereas the hazard function focuses on failing, given survival up to a certain time point.

SUMMARY

T = survival time random variable

t = specific value of T

d = (0,1) variable for failure/censorship

$S(t)$ = survivor function

$h(t)$ = hazard function

Homework

- Practice: Read Chapter 1 of the Survival Analysis, A Self-Learning Text Book; 2-15 pages