# P8108

## Survival Analysis

Fall 2025

Xiao Wu, PhD
Department of Biostatistics, MSPH

Email: xw2892@cumc.columbia.edu
Twitter/X: @wu_xiao1993

# Some practical things

- Teaching Assistants:

| | |
|---|---|
| Elly Kipkogei | ek3235@cumc.columbia.edu |
| Zexi Cai | zc2626@cumc.columbia.edu |
| Yiming Li | yl4925@cumc.columbia.edu |

- Class Sessions (generally in Mailman 8th Floor Auditorium, with a few exceptions):

| | | |
|---|---|---|
| 11/21/2024 | 1:00-3:50pm | Hammer 312 |
| 11/28/2024 | no class | |

- **Office Hours:**

  Office hours will be announced soon (Zexi Cai) (Elly Kipkogei) (Yiming Li)

- **Course Website: Notes/readings available on class website :**
  https://courseworks2.columbia.edu/courses/228811

- **Tentative Schedule:**

| | Given out | Hand in | To earn |
|---|---|---|---|
| **HW1** | 09/12 | 09/26 | 10% |
| **HW2** | 10/03 | 10/17 | 10% |
| **Final Project Proposal** | 10/24 | 11/07 | 5% |
| **Midterm** | 10/24 | (in class) | 30% |
| **HW3** | 10/31 | 11/14 | 10% |
| **HW4** | 11/21 | 12/12 | 10% |
| **Final Project Presentation** | 12/05 | (in class) | 10% |
| **Final Project** | | 12/12 | 15% |

Notes:

- Homework assignments should be turned in electronically through course website, due in class (Friday 11:59 pm).
- Late Submission Penalties
  - 1-3 Days Late: 10% penalty per day late; 4-7 Days Late: 50% penalty.
  - More than 7 Days Late: receive a score of zero.
- You are encouraged to work together on homework assignments, but only in a spirit of learning and with Attribution. Plagiarism will be treated strictly.
- If computer work is required as part of an assignment, please avoid handing in pages of software output; choose relevant portions of your output to your answer sheets and explain the results using human language (in English).
- Software: we will focus on R. You may use SAS/Python, but you will be on your own.

**Project**

- Choose one category
  - Study design based on survival data
  - Analysis survival data from real-world applications
  - Literature review on specific survival analysis methods
- Presentation
  - PDF/PowerPoint
  - 10-15 minutes per group
- Final report
  - Minimum 10 pages including Figures and Tables, excluding References
    - Covers
    - Background
    - Methods
    - Results
    - Interpretation/Discussion
    - Conclusion
    - References
- Approximately 5 students per group
  - Groups will be randomly assigned via Course website on 10/24
  - Each group must select a team lead.
  - Divide and conquer – define each team member's role and contribution in your presentation and final report

Pre-requisites: P6103, P6104, or the Quant core module; at least one course in probability and statistical inference.

Basic expectation is that students should have:

- Strong quantitative skills
- Basic understanding of probability and statistical inference (common probability density functions like binomial, Poisson, normal, hypothesis testing and confidence intervals)
- Understanding of maximum likelihood estimation
- Familiarity with regression approaches for discrete data (eg., logistic regression) and/or continuous data (eg., linear regression).

Two packets of review notes have been posted on the course website (Modules/Background and Review Materials) for you:

- Discrete data and random variables
- Maximum Likelihood Estimation and Asymptotic Test Statistics

**Course Policy Regarding Use of Generative Intelligence (AI)**

**Permitted in this Course with Attribution.**

In this course, students are encouraged to use Generative AI Tools like ChatGPT to support their work. To maintain academic integrity, students must disclose any AI-generated material they use and properly attribute it, including in-text citations, quotations, and references.

A student should include the following statement in assignments to indicate use of a Generative AI Tool: "The author(s) would like to acknowledge the use of [*Generative AI Tool Name*], a language model developed by [*Generative AI Tool Provider*], in the preparation of [*Specific Sections and Tasks*] in this assignment. The [*Generative AI Tool Name*] was used in the following way(s) in this assignment [*e.g., brainstorming, grammatical correction, citation, which portion of the assignment*]."

- Where the use of AI is permitted, you should stick to the guidelines provided by the instructor. Under no circumstance can you submit any work generated by an AI program as your own. AI generated material should be cited like any other reference material using the APA or MLA guidelines.
- If unsure about policies on using generative AI, clarify with the course instructor or your TA.

# Chapter 1

# Survival Analysis - Introduction

## 1.1 Course focus

## 1.2 Some useful references

## 1.3 Basic definitions and notation

## 1.4 Types of censoring

## 1.5 Independent vs informative censoring

## 1.6 Some published studies

## 1.7 Some example datasets

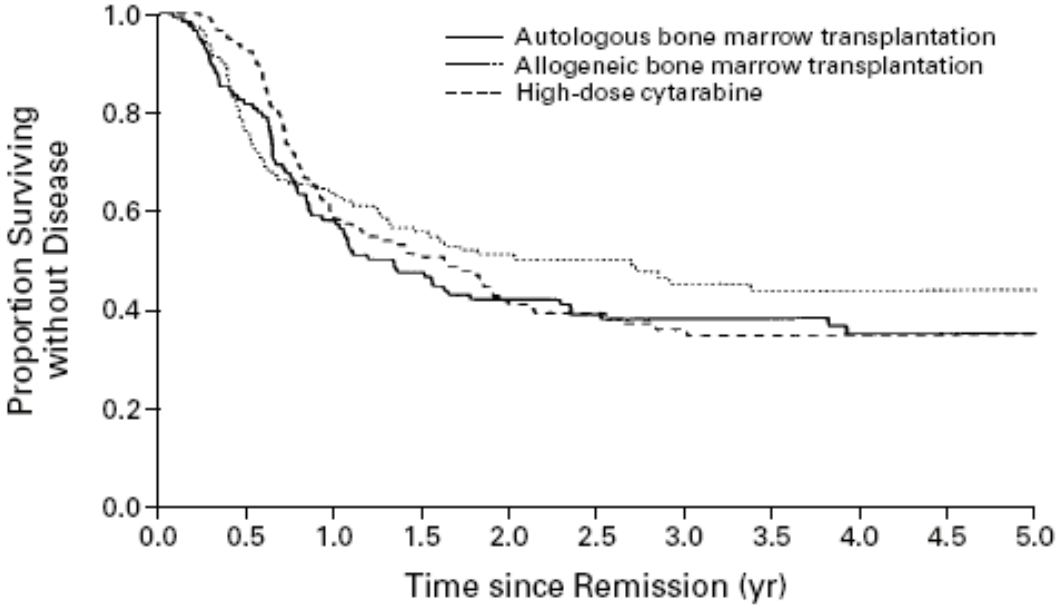
## 1.1 Course Focus

**"Survival Analysis" typically focuses on time to event data. In the most general sense, it consists of techniques for positive-valued random variables, such as:**

- **time to death**
- **time to onset (or relapse) of a disease, time to pregnancy**
- **length of stay in a hospital**
- **time to bankruptcy**
- **time to finishing a doctoral dissertation ...**
- **time from first submission of a paper to publication ...**

## An example:

### Cassileth et al, NEJM, 1998




| Group | No. of Events/No. at Risk | | | | |
|---|---|---|---|---|---|
| Autologous transplantation | 48/116 | 18/66 | 4/45 | 2/34 | 0/22 |
| Allogeneic transplantation | 41/113 | 14/71 | 5/55 | 1/32 | 0/22 |
| Cytarabine | 48/117 | 21/69 | 5/47 | 1/29 | 0/18 |

**Figure 1.** Probability of Disease-free Survival According to Postremission Therapy.

**Another Example:** **The table below shows results comparing the number of failures (deaths or progression) among cancer patients randomized to a standard treatment (called "CHOP") as compared to CHOP with an extra medication, called rituximab (*Haberman et al, JCO 2006*)**

| Maintenance Treatment ($X$) | Number of failures ($Y$) | | Total |
|---|---|---|---|
| | Yes | No | |
| CHOP+rituximab | 52 (30%) | 122 | 174 |
| CHOP (standard) | 74 (42%) | 104 | 178 |
| Total | 126 (36%) | 226 | 352 |

**The percentage of patients who failed was lower (30%) for those on the extra medication (CHOP + rituximab) as compared to the standard treatment (42%) (p=0.03 by Fisher's exact test)**

**However, not only was there a higher percentage of failures on the standard treatment - the failures also tended to occur earlier:**

**We can see this if we view the results in a different way:**

| Maintenance Treatment | Number of failures | | | | Total |
|---|---|---|---|---|---|
| | Year 1 | Year 2 | Year 3 | Years 4-6 | |
| CHOP+rituximab (N=174) | 26 | 12 | 12 | 2 | 52 |
| CHOP (standard) (N=178) | 53 | 13 | 6 | 2 | 74 |
| Total | 79 | 25 | 18 | 4 | 126 |

**Of the patients who failed:**

- **50% (26/52) on the CHOP+ arm failed in the first year**
- **72% (53/74) on the standard treatment failed in the first year**

**$\Rightarrow$ The timing of failures is different on the two arms**
**Therefore, combining the binary endpoint of failure and the time of the event into a survival endpoint is more informative**

**There can also be situations where the percent meeting endpoints is similar, but the timing is different. Survival analysis may be a more powerful approach for comparing treatment effects.**

## A hypothetical example:

**Say we have 10 subjects assigned to each of four treatments after cancer remission, and we follow them until death or end of study (at 36 months). The times below are their times to death (in months).**

**Treatment 1:** Deaths at: 2, 3, 7, 9, 15, 16 (4 alive at 36 months)

**Treatment 2:** Deaths at 1, 1, 2, 4, 4, 6, 7, 9, 11 (1 alive at 36 months)

**Treatment 3:** Deaths at 1, 1, 2, 4, 4, 5 (4 alive at 36 months)

**Treatment 4:** Deaths at: 2, 3, 7, 9, 15, 22, 27, 28, 29 (1 alive at 36 months)

## Questions:

- Looking just at Treatment 1, what is the average time to death?
- Comparing Trt 1 and 2, which treatment appears better?
- Comparing Trt 1 and 3, which treatment appears better?
- Comparing Trt 3 and 4, which appears better?
- Are any pairs of treatments "significantly" different?

## Descriptive Comparisons of "Average" death Time

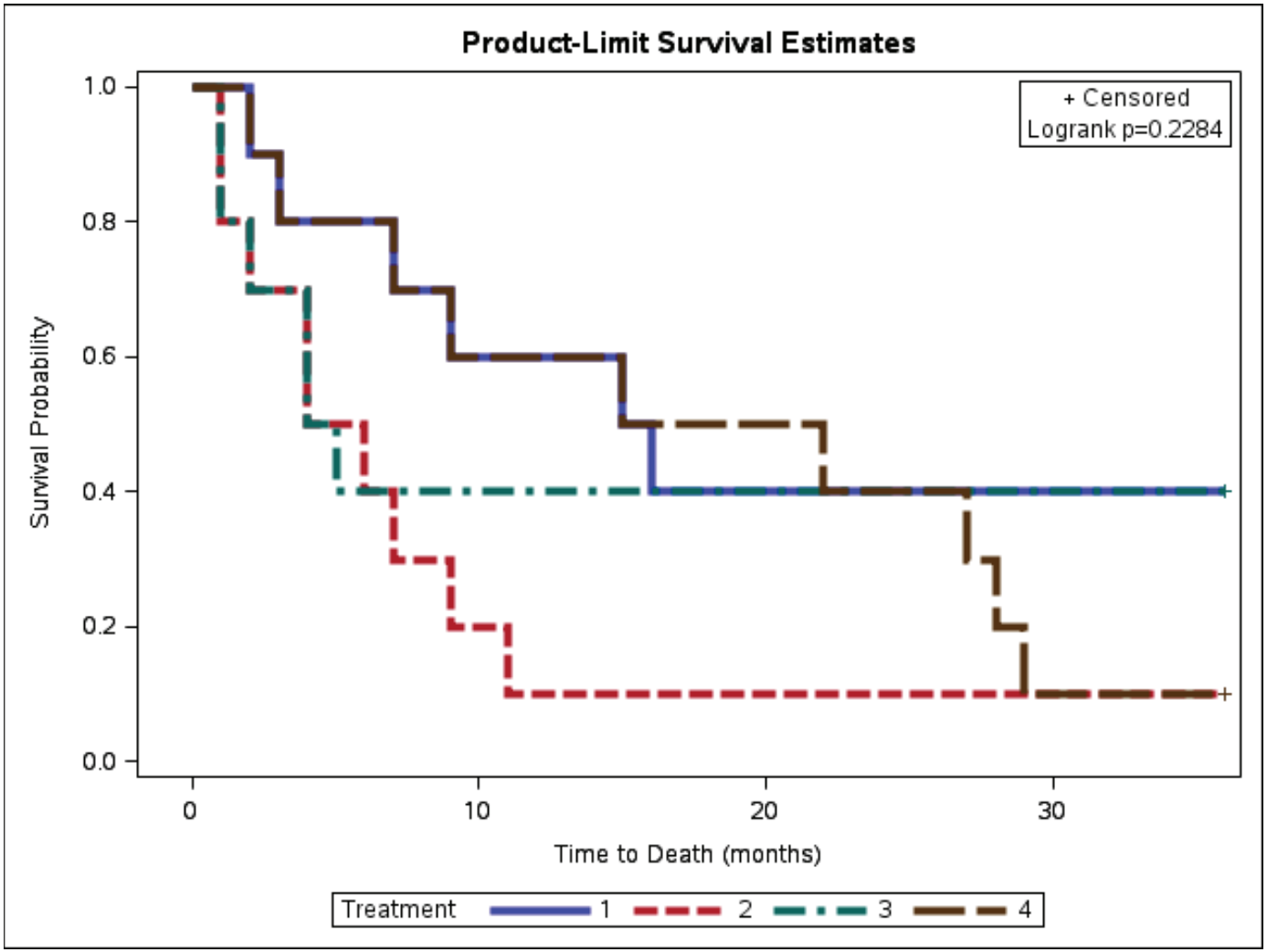
| Treatment Group | Among Deaths: Mean | Among Deaths: Median | Median adjusting for Censoring (KM) |
|---|---|---|---|
| 1 | 8.67 | 8.00 | 15.5 |
| 2 | 5.00 | 4.00 | 5.0 |
| 3 | 2.83 | 3.00 | 4.5 |
| 4 | 15.78 | 15.00 | 18.5 |

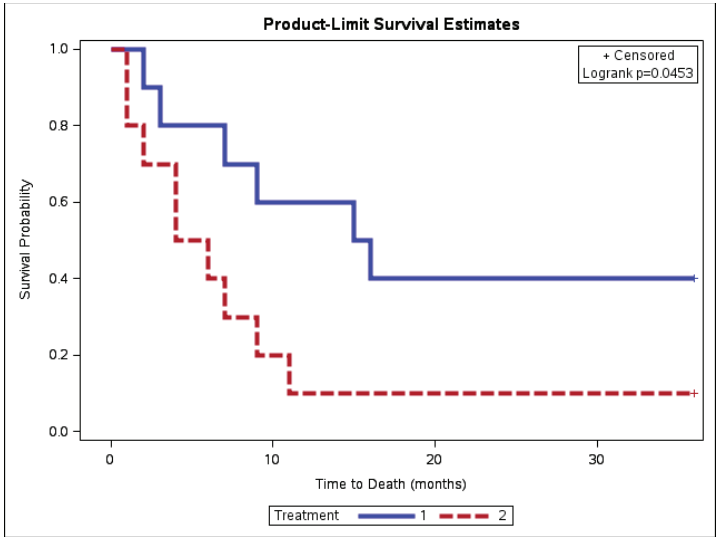
## Comparisons of Treatments (p-values):

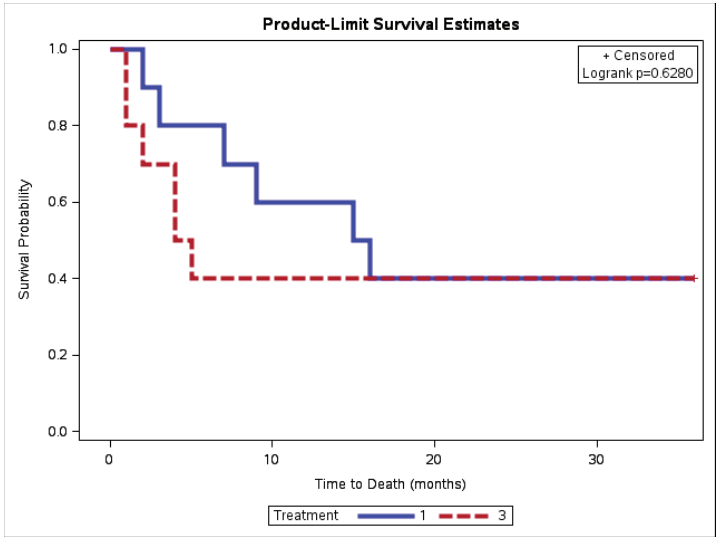
| Comparison | Logrank Test | Wilcoxon Test | Exponential model |
|---|---|---|---|
| 1 vs 2 | 0.045 | 0.048 | 0.014 |
| 3 vs 4 | 0.62 | 0.67 | 0.56 |
| 1 vs 3 | 0.63 | 0.37 | 0.73 |
| 2 vs 4 | 0.15 | 0.06 | 0.09 |
| 1 vs 4 | 0.37 | 0.68 | 0.34 |

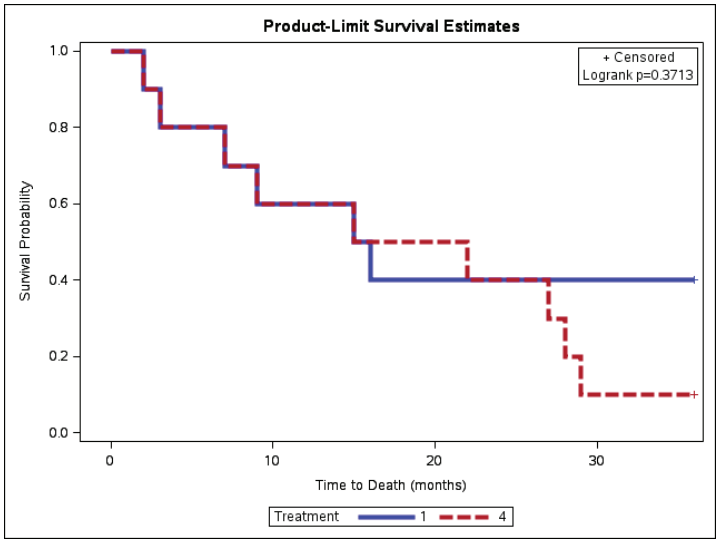
**Although this is a very small example, it illustrates some key principles of survival data:**

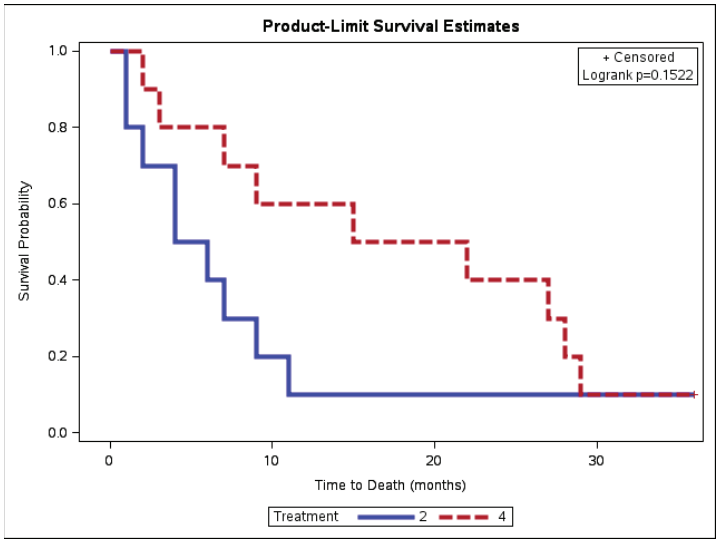
- **Survival times are often "skewed", so the median is usually a better measure of "average" than the mean**
- **Calculating summary statistics and comparing survival distributions needs to account for those without events**
- **Comparisons between survival distributions may give you different conclusions depending on what assumptions you make**


Product-Limit Survival Estimates
+ Censored
Logrank p=0.2284
Survival Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
Time to Death (months)
Treatment
1
2
3
4


Product-Limit Survival Estimates
+ Censored
Logrank p=0.0453
Survival Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
Time to Death (months)
Treatment 1 2


Product-Limit Survival Estimates
+ Censored
Logrank p=0.6280
Survival Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
Time to Death (months)
Treatment 1 3


Product-Limit Survival Estimates
+ Censored
Logrank p=0.3713
Survival Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
Time to Death (months)
Treatment 1 4


Product-Limit Survival Estimates
+ Censored
Logrank p=0.1522
Survival Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
10
20
30
Time to Death (months)
Treatment 2 4

## Types of survival studies

- **clinical trials**
- **prospective cohort studies**
- **retrospective cohort studies**

## Important feature requiring special analysis methods:

**Most often, survival data are not fully observed, but rather are censored .**
**We only get to observe the survival experience up to a certain time.**
**The fact that we are missing part of our variable of interest poses some fundamental problems:**

- **what information do we have?**
- **what untestable assumptions do we need to make - given the nature of the missing data - to learn more.**

**Even under standard (comforting) assumptions, censoring complicates analysis.**

## Next tasks (also see Syllabus):

- **further develop the assumption of non-informative censoring**
- **review probability distributions commonly used with survival outcomes**
- **describe survival data numerically and graphically**
- **compare survival of several groups**
- **explain and predict survival using baseline and time-varying covariates**
- **design studies with survival endpoints**
- **analyze multiple survival outcomes, including recurrence or competing risks**

## 1.2 Some useful references

- Collett: ***Modelling Survival Data in Medical Research, 3rd ed.***
- Allison: ***Survival Analysis Using the SAS System***
- Hosmer, Lemeshow & May: ***Applied Survival Analysis***
- Cox and Oakes: ***Analysis of Survival Data***
- Fleming and Harrington: ***Counting Processes and Survival Analysis***
- Kalbfleisch and Prentice: ***The Statistical Analysis of Failure Time Data, 2nd. ed.***
- Klein & Moeschberger: ***Survival Analysis: Techniques for censored and truncated data***
- Therneau and Grambsch: ***Modeling Survival Data: Extending the Cox Model (R)***

## 1.3 Basic Definitions and Notation

### General conventions

- **Random variables will be denoted by CAPITALS**
  **e.g.** $T$ **is the survival time (from today onwards) for a random person drawn from this class.**
- **Observed value of a random variable will be written as a small letter ,**
  **e.g.** $t = 80$ **(years or days)**
- **Sometimes we consider a set of independent random variables all having the same distribution as** $T$**,**
  **e.g.** $T_1, T_2, T_3, T_4$ **could be the results of repeating the experiment of a random draw from the class 4 times in an independent fashion ('with replacement'),**
- **Drawing inference about the study population is the most important activity in statistics: what can we claim with confidence about the population?**

## The Failure time random variable $T$

**is the focus of this course, to define it we need:**

**(1) an unambiguous time origin**
**(e.g. time of randomization, time of diagnosis, time of marketing intervention)**

**(2) a time scale**
**(e.g. real time (days, years), menstrual cycles)**

**(3) definition of the (occurrence of the) event**
**(e.g. death, recurrence, need a new implant)**
**To be specific, this often includes details of how it will be measured.**

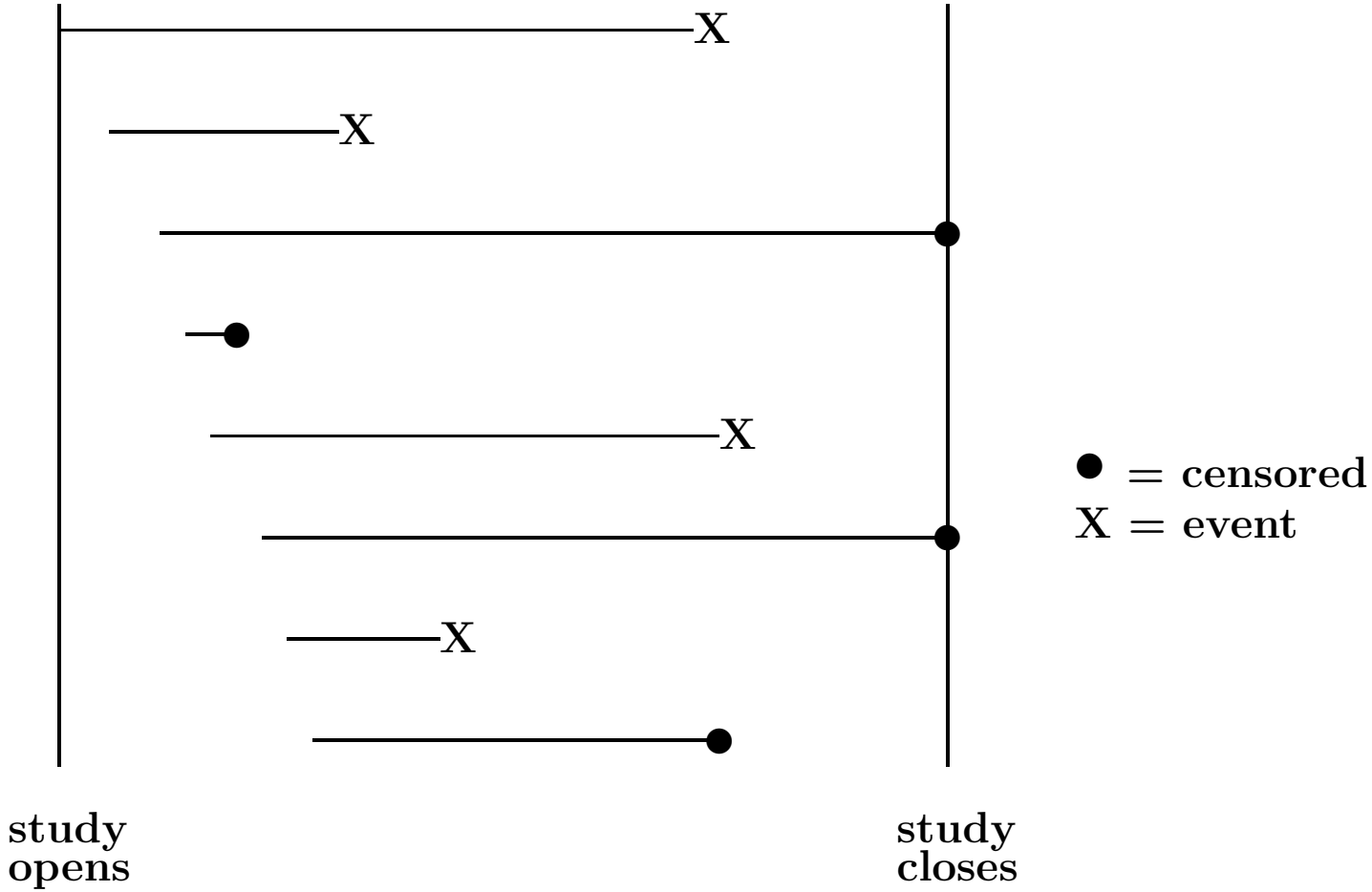
## Failure times, Censoring times and Observation times

- **Failure time random variables are always non-negative.**
  **That is, if we denote the failure time by $T$ , then $T \geq 0$.**
  - **$T$ can either be discrete (taking a finite set of values, e.g. $a_1, a_2, \ldots, a_J$)**
  - **or continuous (defined on $(0, \infty)$).**

- **Some process or mechanism, like end-of-study, may prevent us from continued observation. The potential time at which our observation of a time to failure would need to end is often called a censoring time, we denote it as $U$ (*Unavailable*)**

- **When a survival time of interest $T$ is possibly censored by time $U$,**
  **the observation time $X = \min(T, U)$ is called a censored failure time random variable**

**Figure. Illustration of survival data**



**The illustration of survival data in the previous Figure shows several features which are typically encountered in survival data:**

- **Subjects do not all enter the study at the same time**
  **This is referred to as "staggered entry"**
- **When the study ends, some individuals still haven't had the event yet ('administrative censoring')**
- **Other individuals drop out or get lost in the middle of the study, and all we know about them is the last time they were still "free" of the event ('lost to follow-up')**

**The last two features relate to "censoring" of the failure time events.**
**The censoring time will be the first of the following times (measured from the time origin):**

- **the study ends**
- **the subject drops out (not just from treatment: we stop observing the event)**

## 1.4 Types of censoring

- **Right-censoring**

  **Only observe $X = \min(T, U)$, due to:**

  – **loss to follow-up**

  – **drop-out**

  – **study termination (sometimes called 'administrative censoring')**

  **We call this right-censoring because the true unobserved event is to the right of our censoring time; i.e., all we know is that the event has not happened at the end of follow-up.**

  **In addition to observing $X$, we also get to see the failure indicator:**

$$\delta = \begin{cases} 1 \text{ if } T \leq U \\ 0 \text{ if } T > U \end{cases}$$

  **Right-censoring is the most common type of censoring we will deal with in survival analysis.**

- **Left-censoring**

  **Can only observe $Y = \max(T, U)$ and the "failure" (event) indicators:**

$$\delta = \begin{cases} 1 \text{ if } U \leq T \\ 0 \text{ if } U > T \end{cases}$$

  - **Example 1: (Miller) study of age at which African children learn a task. Some already knew (left-censored), some learned during study (exact), some had not yet learned by end of study (right-censored).**

  - **Example 2: when measuring the viral load burden in HIV positive subjects, i.e., the number of HIV virus particles in a milliliter of blood, one must account for the fact that levels below a certain threshold are not detectable. If we give those observations the value of the threshold, then they are left censored. Also called "limit of detection" censoring.**

- **Interval-censoring**

  **Observe** $(L, R)$ **where** $T \in (L, R]$

  – **Example 1: Time to HIV seroconversion occurs between a negative and a positive test result.**

  – **Example 2: Time to undetectable viral load in AIDS studies, based on measurements of viral load taken at each clinic visit**

  – **Example 3: Time to recurrence of colon cancer after surgery. Follow patients every 3 months after resection of primary tumor.**

  – **Example 4: Time to pubertal onset in HIV-infected youth Tanner Staging conducted at annual study visits**

**Suppose we have a sample of observations on $n$ people:**

$$(T_1, U_1), (T_2, U_2), ..., (T_n, U_n)$$

**There are three main types of right censoring times:**

- **Type I: All the $U$'s are the same (i.e., limit of detection)**
  **e.g. animal studies, all animals sacrificed after 2 years**

- **Type II: $U = T_{(r)}$, the time of the $r$th failure.**
  **e.g. study of light bulbs stops when 2/3 have failed**

- **Type III: 'Random' censoring the $U$'s are random variables, $\delta$'s are failure indicators:**

$$\delta = \begin{cases} 1 \text{ if } T \leq U \\ 0 \text{ if } T > U \end{cases}$$

## 1.5 Independent vs informative censoring

- We call censoring independent if $U$ is independent of $T$.
  - Example 1: If $U$ is the planned end of the study (say, 2 years after the study opens), then it is usually considered independent of the event times.

    Q: what if there is a trend over calendar time in the survival times?

  - Example 2: If $U$ is the time that a patient drops out of the study because the patient became much sicker and/or had to discontinue taking the study treatment, then $U$ and $T$ are probably not independent
- In principle:

  An individual censored at $U$ should be representative of all subjects who survive up to $U$

  Or: those who stay on are no different from those who disappear at a given time point in terms of future survival chances/distribution.

- **We call censoring non-informative for an analysis of survival time [conditional on (time-varying) covariates], if those who get actually censored at time $t$ are representative of all patients [with the same (time-varying) covariates] who survive up to that time $t$.**

  **This means that in general the probability of being censored at time $t$ *could* depend on prognostic characteristics measured up to that time, but that among all those who survived with the same measured characteristics up to time $t$, being censored at time $t$ does not further predict especially good or poor residual survival times.**

- **(Censoring is further considered informative if the distribution of $U$ contains information about the parameters characterizing the distribution of $T$.)**

## An example of informative censoring

**Consider a study conducted at two study centers A and B, which serve somewhat different patient populations:**

- **Center A has sicker patients with shorter survival times who are harder to recruit: these patients typically entered the study later**
- **Center B has healthier patients who tend to enroll earlier in the study and thus may stay on the study longer**
- **Assume that censoring times are equally distributed over individuals <u>within</u> both centers.**
- **If both centers stop the study at the same date, center A has shorter censoring times as well as lower survival chances. However within this center, the censoring time is independent of survival time.**

  **Suppose we get data on censoring times and survival times, and either**

  - **also data on the center or**
  - **no data on the center**

**Q: in which of these two situation(s) can the censoring be considered non-informative for estimating the survival distribution?**

- **With center information:**

  **within each center, the censoring does not discriminate between people and censoring is non-informative for calculating the survival distribution within each center.**

- **Without center information:**

  - **we are estimating the probability of surviving time point $t$, irrespective of the center to which a subject belongs (averaged over center).**
  - **the censoring distribution is informative about center, which in turn, is associated with survival**
  - **we have informative censoring if we do not account properly for center**

## 1.6 Some published studies: see /Modules/Readings

- **Example 1:**
  **The effect of changing the priority for HLA matching on the rates and outcomes of kidney transplantation in minority groups, NEJM 2004.**
- **Example 2.**
  **Cardiovascular events associated with the drug rofecoxib in a colorectal adenoma chemoprevention trial, NEJM 2005.**
- **Example 3:**
  **Chemotherapy compared with autologous or allogeneic bone marrow transplantation in the management of acute myeloid leukemia in first remission, NEJM 1998.**
- **Example 4:**
  **Breast cancer after prophylactic mastectomy in women with certain prognostic mutations (BRCA1 or BRCA2), NEJM 2001.**

## Example 1: Kidney transplantation and minority groups

(`Kidney_transplant_04.pdf`)

**This observational study addressed two different survival times:**

- $T_1$ : **Time from placement on waiting list to 'transplantation'**
- $T_2$ : **Time from 'first transplantation' to graft failure.**

**The analysis then considers the impact of the (HLA-matching) policy of priority on waiting times per race (we will not go into that for now).**

- **Consider in the abstract the definition of starting point and end point of the survival time $T_1$.**
- **Observe how these 'definitions' get qualified in the second column of page 546.**
- **Would you consider the censoring mechanism non-informative for the time until a transplant with at least one HLA mismatch becomes available?**

## Example 2: Cardiovascular risk with rofecoxib (`rofecoxib_05.pdf`)

**Recent studies have suggested that the anti-inflammatory drug and painkiller rofecoxib, although easier on the stomach, may cause an increased risk of cardiovascular events. The time to such events is studied in this article and compared between arms in a trial which randomized patients to either 25mg rofecoxib daily or placebo.**

- **The primary outcome is time since randomization until the occurrence of a first cardiovascular event.**
- **Consider the description of the outcome in the abstract and in the section on cardiovascular events, page 1094. Do you feel the outcome is well represented in the abstract? Interpret the choice of endpoint in light of the information in Figure 1.**
- **Assume again that risk of acquiring a cardiovascular event (and its timing) is exactly the same on both randomized arms. Can you think of a mechanism that would give us a different picture considering the way events have been censored here? Explain.**
- **What other choice of endpoint could one have made? What are the advantages and disadvantages of this?**
- **Look up the correspondance following this paper in `rofecoxib_06.pdf` ...**

## Example 3: Survival following bone marrow transplant as compared to chemotherapy

(`bone_marrow_98.pdf`)

- **Why are we not simply comparing survival times with time origin being the start of administration of treatment?**
- **Assume for a moment that the treatments 'high dose cytarabine' and 'Autologous bone marrow transplantation' work equally well. What would you then expect to happen with an analysis that uses as time origin for $T$ the time of start of randomized treatment?**
- **Do you see a similar problem for the analysis that has been proposed here instead?**

## Example 4: Breast cancer after prophylactic mastectomy

(`breast_cancer_01.pdf`)

- **Consider carefully how survival time has been defined in this observational study: for the treatment group and for the control group**
- **What do you see as possible problems with these definitions?**
- **Considering that no breast cancers occurred in the treatment group, would you doubt the conclusion drawn here?**

## 1.8 The population distribution of survival times

**There are several equivalent ways to characterize the (probability) distribution of a survival random variable. Some are special to survival analysis:**

- **The density *function* $f(t)$ (takes on a value for a range of possible t-values)**
- **The (cumulative) distribution function $F(t)$**
- **The survival function $S(t)$**
- **The hazard function $\lambda(t)$**
- **The cumulative hazard function $\Lambda(t)$**

**Remind yourselves of what these are for a normal outcome distribution... Before giving exact definitions of these parameters for the population distribution, we will look at some presentations of estimated distributions of survival times in the literature.**

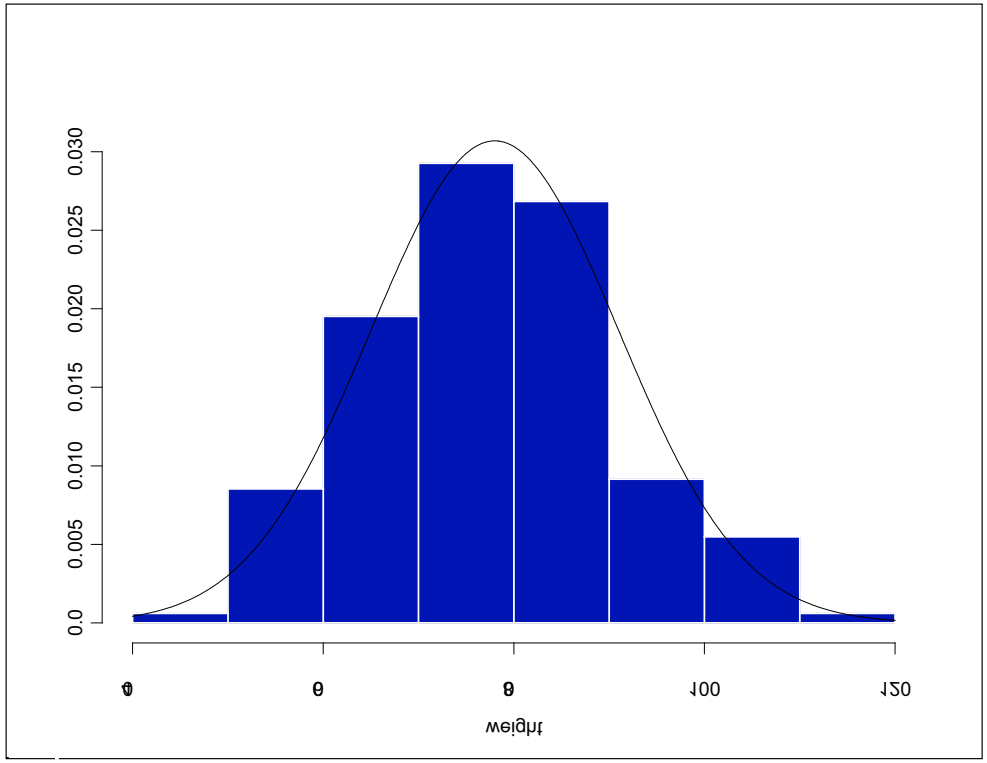



Figure 1.1: Weights in kg. of patients in a hypertension trial

**This picture shows a histogram with normal probability curve for the weights (in kg.) of individuals entering a hypertension trial.**

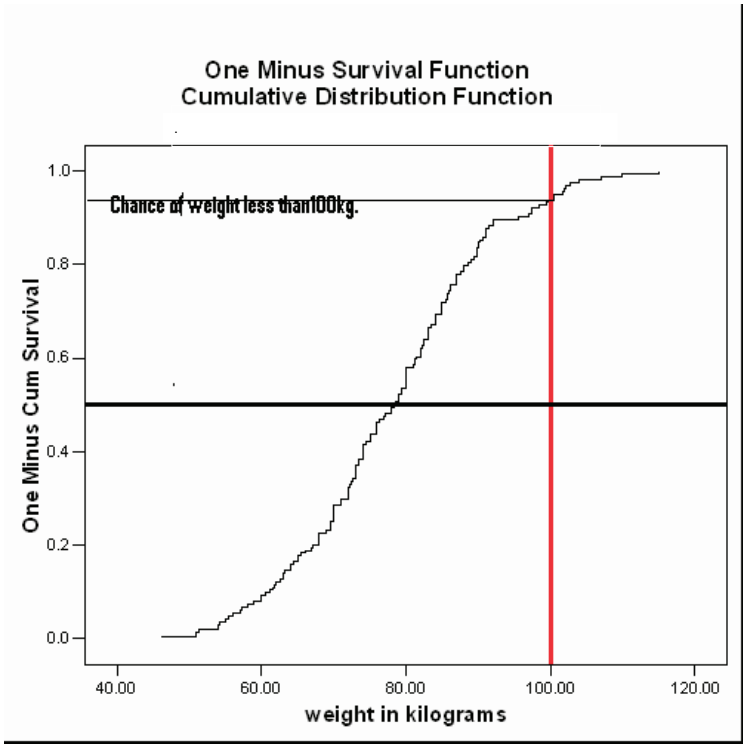



Figure 1.2: Percentage of weights below given weights in kg. - in a hypertension trial

**The same observed distribution is shown in this figure, but in the form of cumulative distribution: what observed percentage has an outcome less than the weight on the x-axis?**

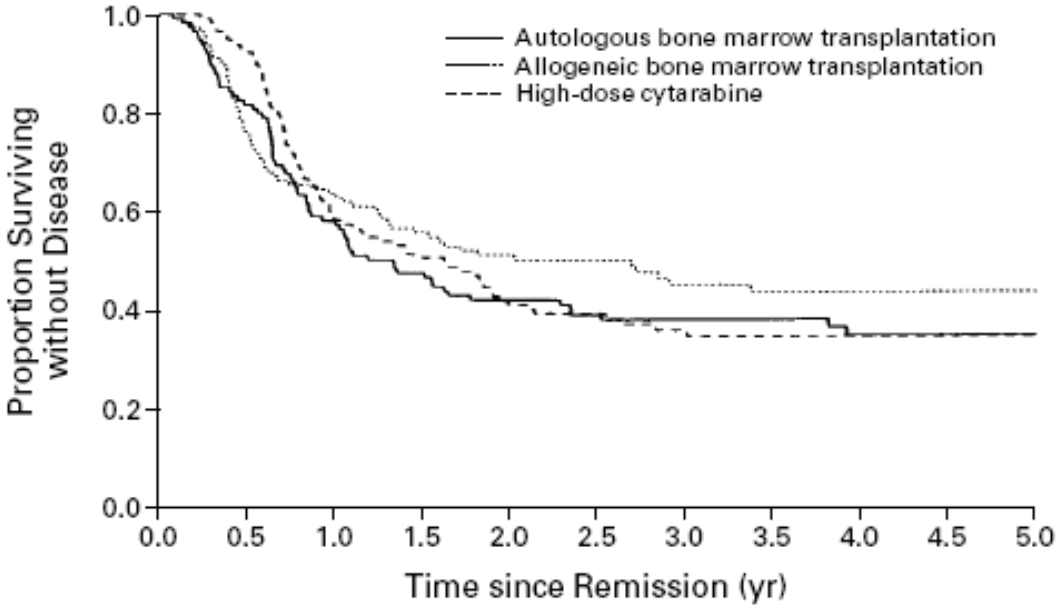
**Popular representations of the distributions of survival outcomes are:**

- **a survival function :**
  - **shows for different possible values of time $t$, the percentage of individuals who survive $t$,**
  - **this is the percentage that has an outcome for $T$ which is larger than $t$**
  - **this is one minus the CDF of the r.v. $T$: $S(t) = 1 - F(t)$**
- **a cumulative distribution function (CDF) (one minus the survival function)**
- **the hazard (rate) function sometimes called incidence (rate) : expected number of events per time in function of time $t$**

**Note how the word 'incidence' is sometimes used to refer to different parameters: the hazard rate, the cumulative hazard rate (see later) or a risk (probability) Be careful when you read the literature, to know what is meant. Also, the terminology for a rate and a risk are sometimes confused. We will give careful definitions later.**

**Each of those has been used in the example papers mentioned...**

## Cassileth et al, NEJM, 1998




| Group | No. of Events/No. at Risk | | | | |
|---|---|---|---|---|---|
| Autologous transplantation | 48/116 | 18/66 | 4/45 | 2/34 | 0/22 |
| Allogeneic transplantation | 41/113 | 14/71 | 5/55 | 1/32 | 0/22 |
| Cytarabine | 48/117 | 21/69 | 5/47 | 1/29 | 0/18 |

**Figure 1.** Probability of Disease-free Survival According to Postremission Therapy.

## Bresalier *et al.*, NEJM, 2005

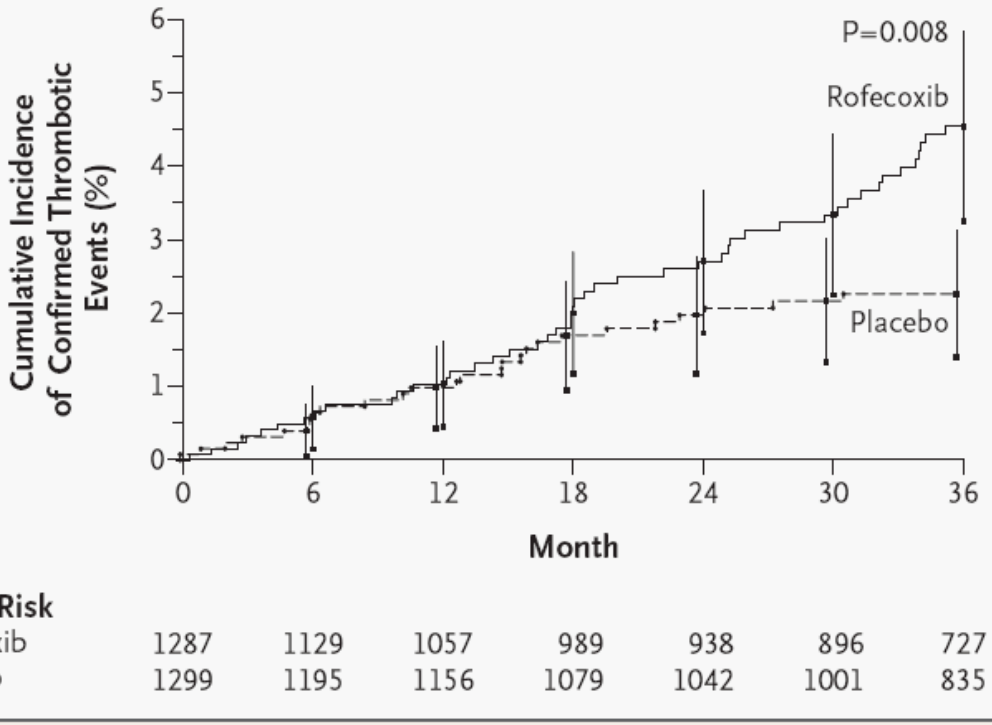



**Figure 2. Kaplan–Meier Estimates of the Cumulative Incidence of Confirmed Serious Thrombotic Events.**

Vertical lines indicate 95 percent confidence intervals.

## Bresalier *et al.*, NEJM, 2005

**Table 3. Summary of Rates and Relative Risks of Confirmed Serious Thrombotic Events and the APTC End Point.***

| Adverse Event | Rofecoxib Group | | | | Placebo Group | | | | Difference in Rate (95% CI) | Relative Risk (95% CI) |
|---|---|---|---|---|---|---|---|---|---|---|
| | No. at Risk | No. of Events | No. of Patient-yr at Risk | Rate/100 Patient-yr | No. at Risk | No. of Events | No. of Patient-yr at Risk | Rate/100 Patient-yr | | |
| **Confirmed event** | | | | | | | | | | |
| Overall | 1287 | 46 | 3059 | 1.50 | 1299 | 26 | 3327 | 0.78 | 0.72 (0.19 to 1.25) | 1.92 (1.19 to 3.11) |
| Month 0–18 | 1287 | 22 | 1656 | 1.33 | 1299 | 20 | 1765 | 1.13 | 0.20 (–0.55 to 0.94) | 1.18 (0.64 to 2.15) |
| Month 19–36 | 989 | 24 | 1403 | 1.71 | 1079 | 6 | 1561 | 0.38 | 1.33 (0.58 to 2.08) | 4.45 (1.77 to 13.32) |
| **APTC end point** | | | | | | | | | | |
| Overall | 1287 | 34 | 3070 | 1.11 | 1299 | 18 | 3334 | 0.54 | 0.57 (0.12 to 1.02) | 2.06 (1.16 to 3.64) |
| Month 0–18 | 1287 | 14 | 1658 | 0.84 | 1299 | 12 | 1769 | 0.68 | 0.17 (–0.42 to 0.75) | 1.25 (0.58 to 2.69) |
| Month 19–36 | 994 | 20 | 1412 | 1.42 | 1083 | 6 | 1565 | 0.38 | 1.03 (0.34 to 1.73) | 3.69 (1.43 to 11.24) |

* CI denotes confidence interval, and APTC Antiplatelet Trialists' Collaboration.

### 1.8.1 Density function (Probability Mass Function)

- **For discrete r.v.'s**

  **Suppose that $T$ can take values in $a_1, a_2, \ldots, a_J$.**

$$f(t) = P(T = t)$$

$$= \begin{cases} f_j & \textbf{if} \quad t = a_j, j = 1, 2, \ldots, J \\ 0 & \textbf{if} \quad t \neq a_j, j = 1, 2, \ldots, J \end{cases}$$

- **Density Function for continuous r.v.'s**
  **The continuous density function evaluated at $t$ : $f(t)$ approximates the probability that the (survival) outcome falls within a small unit of time $\Delta t$ around $t$:**

$$f(t)\Delta t \approx P(t \leq T < t + \Delta t)$$

  **Its formal definition is as follows:**

$$f(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t)$$

## Heuristics for the density

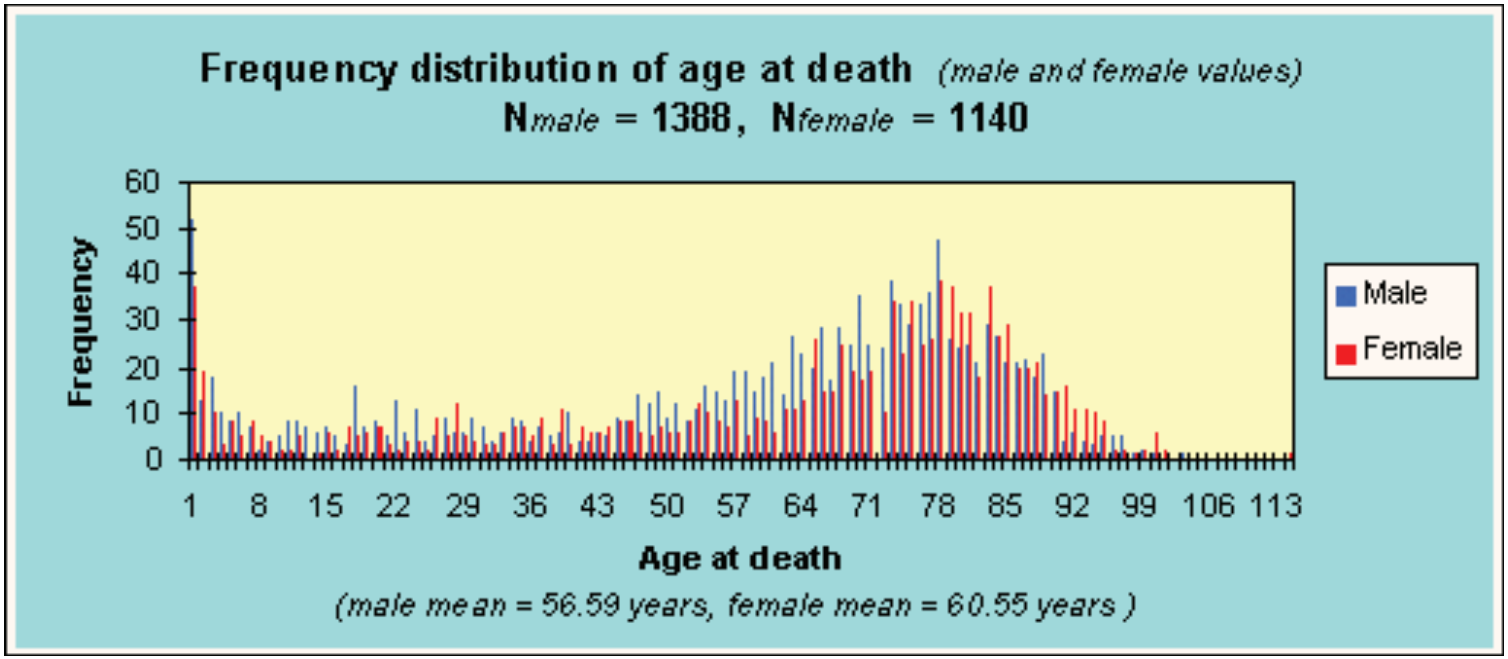
- **Think of a cohort of people born in 1900.**
- **Observe what percentage of them dies at each age; e.g. what percentage dies at age 51, what percentage at age 52, etc.**
- **The heights of the histogram representing these bars will approximate the density function well.**

**The picture below shows the distribution of age at death as it was found on cemeteries in New South Wales ...**

*This reflects the age-at-death distribution for a well defined birth population, if birth and death rates were stationary over time and there was essentially no migration.*

Cemeteries in and around Gunning Shire, New South Wales

http://www.gundaroo.info/genealogy/cemeteries/index.htm




Data starting in the 1800s…

### 1.8.2 Survivorship Function: $S(t) = P(T > t)$.

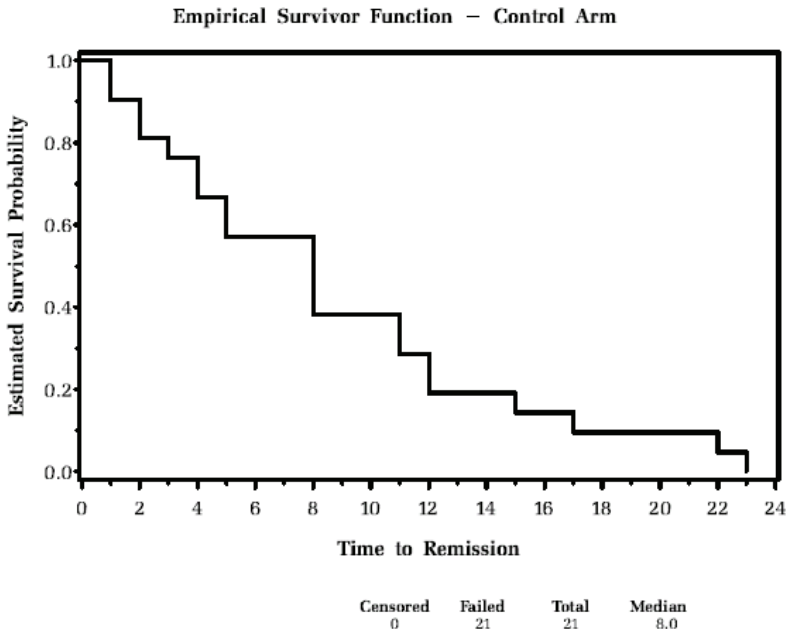
**In other settings, the cumulative distribution function, $F(t) = P(T \leq t) = 1 - S(t)$, is of interest. In survival analysis, our interest tends to focus on the survival function, $S(t)$.**
**For a discrete random variable:**

$$\begin{aligned} S(t) &= \sum_{u>t} f(u) \\ &= \sum_{a_j>t} f(a_j) \\ &= \sum_{a_j>t} f_j \end{aligned}$$

**For a continuous random variable:**

$$S(t) = \int_t^\infty f(u)du$$

**Figure 1.5: Example for leukemia data (control arm)**




**Consider the survival function for time to relapse of untreated leukemia patients who were in remission. Can you derive the density from that?**

$$f(t) = -S'(t)$$

### 1.8.3 Hazard Function $\lambda(t)$

**Sometimes called instantaneous failure rate , force of mortality , or incidence .**

- **Discrete random variables:**

$$\lambda(a_j) \equiv \lambda_j = P(T = a_j | T \geq a_j) = P(T = a_j | T > a_{j-1}) = \frac{P(T = a_j)}{P(T > a_{j-1})}$$

$$= \frac{f(a_j)}{S(a_{j-1})} = \frac{f(a_j)}{\sum_{k:a_k > a_{j-1}} f(a_k)} = \frac{f(a_j)}{\sum_{k:a_k \geq a_j} f(a_k)}$$

- **Continuous random variables:**

$$\lambda(t)\Delta t \approx P(t \leq T < t + \Delta t | T \geq t)$$

$$= \frac{P([t \leq T < t + \Delta t] \bigcap [T \geq t])}{P(T \geq t)}$$

$$= \frac{P(t \leq T < t + \Delta t)}{P(T \geq t)} = \frac{f(t)\Delta t}{S(t)}$$

**Formally:**

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t | T \geq t) = \frac{f(t)}{S(t)}$$

## Heuristics for the hazard

- **Think again of a cohort of people born in 1900.**
  **Consider the subgroup who has survived until a certain age $t$, 40 years, say.**
- **Observe what percentage of this subgroup dies within the year**
  **e.g. what percentage among the 40 year olds dies at age 41, and similarly for all possible ages at death**
- **The heights of bars representing these values will approximate the hazard function.**

## Some hazard shapes seen in applications

- **increasing**

  **e.g. aging after 65**

- **decreasing**

  **e.g. survival after surgery**

- **bathtub**

  **e.g. age-specific mortality**

- **constant**

  **e.g. survival of patients with advanced chronic disease**

### 1.8.4 Cumulative Hazard Function $\Lambda(t)$

- **Discrete random variables:**

$$\Lambda(t) = \sum_{k:a_k \leq t} \lambda_k$$

- **Continuous random variables:**

$$\Lambda(t) = \int_0^t \lambda(u)du$$

### 1.8.5 Relationship between $S(t)$ and $\lambda(t)$ :

**We've already shown that, for a continuous r.v.**

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

**For a right-continuous survivor function $S(t)$, we can show:**

$$f(t) = -\frac{d}{dt}S(t) \qquad \textbf{or} \qquad \frac{d}{dt}S(t) = -f(t)$$

**We can use this relationship to show that:**

$$-\frac{d}{dt}[\log S(t)] = -\left(\frac{1}{S(t)}\right) S'(t) = -\frac{-f(t)}{S(t)} = \frac{f(t)}{S(t)}$$

**So another way to write $\lambda(t)$ is as follows:**

$$\lambda(t) = -\frac{d}{dt}[\log S(t)]$$

### 1.8.6 Relationship between $S(t)$ and $\Lambda(t)$ :

- **Discrete case:**

  **Suppose that $a_j \leq t < a_{j+1}$. Then**

$$\begin{aligned} S(t) = P(T > t) &= P(T > a_1, T > a_2, \ldots, T > a_j) \\ &= P(T > a_1)P(T > a_2|T > a_1) \cdots P(T > a_j|T > a_{j-1}) \\ &= (1 - \lambda_1) \times \cdots \times (1 - \lambda_j) \\ &= \prod_{k:a_k \leq t} (1 - \lambda_k) \\ &= e^{\sum_{k:a_k \leq t} \log(1-\lambda_k)} \approx e^{\sum_{k:a_k \leq t} -\lambda_k} = e^{-\Lambda(t)} \end{aligned}$$

- **Continuous case:**

$$\begin{aligned} \Lambda(t) &= \int_0^t \lambda(u)du = \int_0^t \frac{f(u)}{S(u)}du \\ &= \int_0^t -\frac{d}{du} \log S(u)du \\ &= -\log S(t) + \log S(0) \\ &\Rightarrow S(t) = e^{-\Lambda(t)} \end{aligned}$$

## Relationships - an overview

$$f(t)\Delta t \approx P(t \leq T < t + \Delta t)$$

$$\lambda(t)\Delta t \approx P(t \leq T < t + \Delta t | T \geq t)$$

$$S(t) = P(T > t) = \int_t^\infty f(u)du$$

$$f(t) = -\frac{d}{dt}S(t)$$

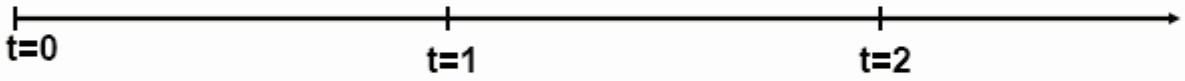
$$\lambda(t) = \frac{f(t)}{S(t)}$$

$$\lambda(t) = -\frac{d}{dt}[\log S(t)]$$

$$\begin{aligned} S(t) &= e^{-\Lambda(t)} \\ &\approx 1 - \Lambda(t) \textbf{ while the cumulative hazard is small} \end{aligned}$$

## Relationship between parameters in a simple example

| | | |
|---|---|---|
| S(0) =100% | S(1) =80% | S(2) =50% |
| f(0) =0% | f(1) =20% | f(2) =30% |
| λ(0) =0% | λ(1) =20% | λ(2) =30%/80%= 37.5% |




| | | |
|---|---|---|
| Λ(0) =0% | Λ(1) =20% | Λ(2) =57.5% |
| F(0) =0% | F(1) =20% | F(2) =50% |

Note: 80%*(100-37.5)%=50%

-log(50%) =69.3%

## Usefulness of the different parameters

- **Depending on the context, one or another parameter may be most relevant/revealing**

- **The survival probability is typically the easiest to interpret and most relevant**

- **The hazard is easiest to estimate directly from censored data, and it will lead to our basic regression models and estimators**

- **results of fitted hazard functions can subsequently be translated into survival functions to reveal how covariates matter in those terms**

## 1.9 Measuring Central Tendency in Survival Times

**Mean survival time - call this $\mu$**

f(t)=-dS(t)/dt

$$\mu = \sum_{j=1}^{J} a_j f_j \quad \textbf{for discrete T with } J \textbf{ possible outcomes.}$$

$$= \int_0^\infty u f(u) du = \int_0^\infty \int_0^u dv f(u) du = \int_0^\infty \int_v^\infty f(u) du dv$$

$$= \int_0^\infty S(v) dv \quad \textbf{for continuous } T$$

**The mean is sensitive to large outliers. With censored data, we are typically unable to estimate the entire survival curve. There is only information up to a certain time point $t$ (the end of our observation time)**
**Hence, we would be left with substantial uncertainty as to the actual value of the mean unless we are willing to make serious assumptions about how the survival curve will evolve (to zero) beyond our final observation time.**

**If our observation time reaches 'far enough', we should however be able to estimate the median survival time...**

**Median survival - call this $\tau$, is designed to satisfy**

$$S(\tau) = 0.5$$

**Similarly, any other percentile could be defined.**

**For discrete distributions, the median is more generally defined as 'the' value $m$ such that: at least ($\geq$) $50\%$ of the outcomes are larger than $m$ and at least ($\geq$) $50\%$ of the outcomes are smaller than or equal to $m$.**

**When more than 1 value satisfies this definition, one can either take the average, or the *smallest* time $\tau$ such that**

$$\hat{S}(\tau) \leq 0.5$$

$P(T > 3) = 50\%$

**Q: Apply this to outcomes $\{1, 2, 3, 4, 5\}$ and $\{1, 2, 3, 4, 5, 6\}$**

**Q: Given the survivor distribution $\hat{S}(t)$ for $t \in [0, 100]$**

- **When $\hat{S}(100) = 20\%$, can you derive the median(T) and/or the mean(T)?**
- **How about when $\hat{S}(100) = 80\%$?**

## Estimating the survival or hazard function

**We can estimate the survival (or hazard) function in two ways:**

- **by specifying a parametric model for $\lambda(t)$ or equivalently for the density function $f(t)$**
- **by developing an empirical non-parametric estimate of the survival function**

**If no censoring:**

**The empirical estimate of the survival function, $\tilde{S}(t)$, is the proportion of individuals with observed event times greater than $t$.**

**With censoring:**

**If there are censored observations, then $\tilde{S}(t)$ is not a good estimate of the true $S(t)$, so other non-parametric methods must be used to account for censoring (life-table methods, Kaplan-Meier estimator)**

## 1.10 Some Parametric Survival Distributions

- **The Exponential distribution This is the simplest distribution with just 1 unknown parameter. It plays a role similar to that of the normal distribution in linear regression.**

$$f(t) = \lambda e^{-\lambda t} \text{ for } t \geq 0$$

$$S(t) = \int_t^\infty f(u)du = e^{-\lambda t}$$

$$\lambda(t) = \frac{f(t)}{S(t)} = \lambda \qquad \text{constant hazard!}$$

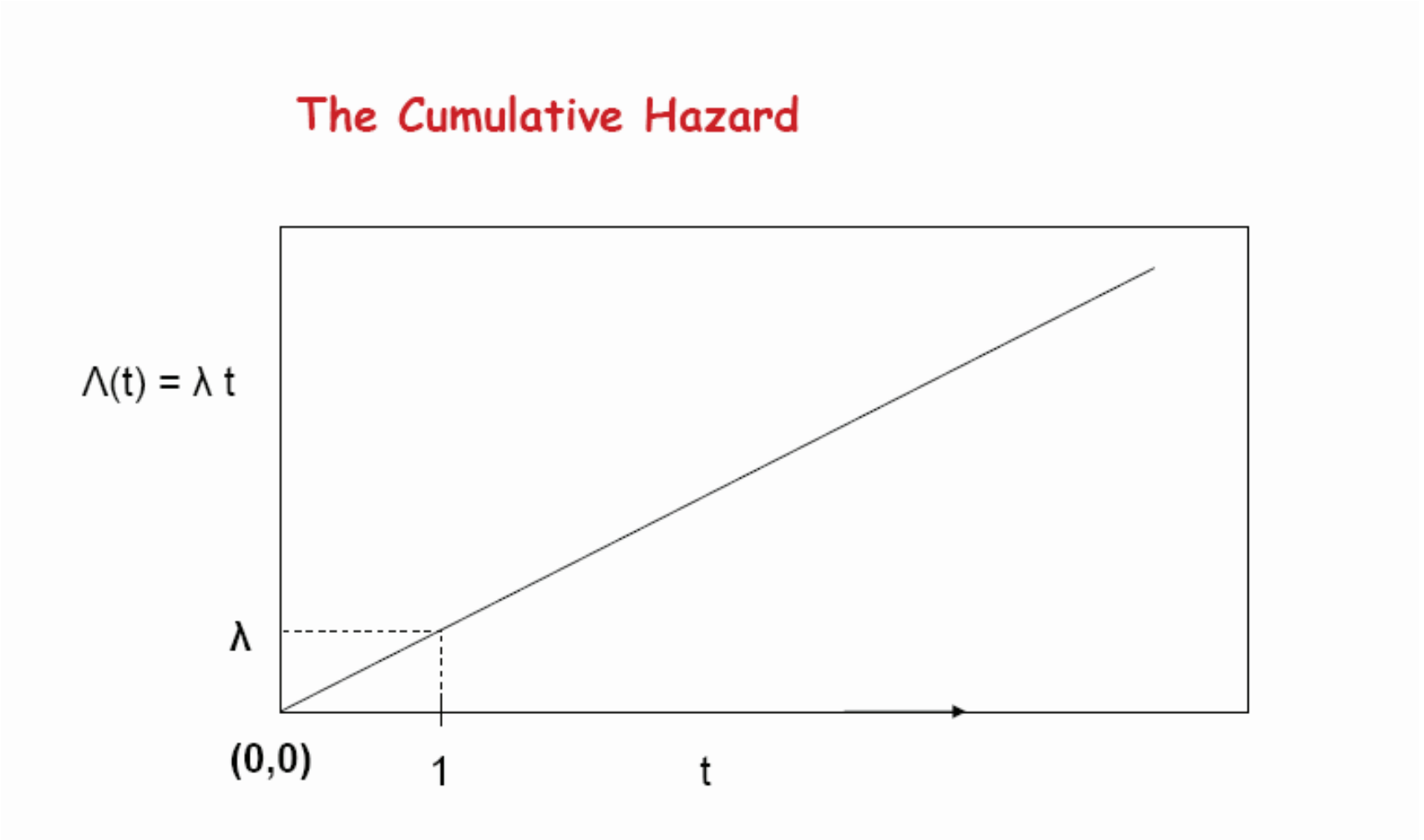
$$\Lambda(t) = \int_0^t \lambda(u)\,du = \int_0^t \lambda\,du = \lambda t$$

$$\frac{d \log L(\lambda)}{d\lambda}$$

$$F(T > t_n)$$

MLE $L(\lambda) = \prod_{r=1}^{R} f(t_r) \prod_{n=R+1}^{N} S(t_n)$

$= \lambda^R e^{-\lambda \Sigma t_r} \times e^{-\lambda \Sigma t_n}$


The Cumulative Hazard
Λ(t) = λ t
λ
(0,0)
1
t

## Exponential distribution

$$S(t) = \int_t^\infty f(u)du = e^{-\lambda t}$$

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

**Check: Does $S(t) = e^{-\Lambda(t)}$ ?**

**For a given dataset, the Maximum Likelihood Estimator $\hat{\lambda}$ turns out to be:**

$$\widehat{\lambda} = \frac{\text{The number of events}}{\text{Persontime: total number of time units observed on all individuals}}$$

**Can you find the mean and median of an Exp($\lambda$) random variable?**

$$\sum_{r=1}^{R} t_r + \sum_{n=r+1}^{N} t_n$$

## Median and mean of the Exponential distribution

- **median:** $0.5 = S(\tau) = e^{-\lambda\tau} : \Rightarrow \quad \tau = \frac{-\log(0.5)}{\lambda}$
- **mean:** $\int_0^\infty u\lambda e^{-\lambda u} du = \frac{1}{\lambda}$

| $\lambda$ | Median | Mean | Mean - Median |
|---|---|---|---|
| **10** | **0.069** | **0.1** | **0.03** |
| **2** | **0.35** | **0.5** | **0.15** |
| **1** | **0.69** | **1** | **0.31** |
| **0.5** | **1.39** | **2** | **0.61** |
| **0.1** | **6.93** | **10** | **3.07** |

**Exercise: Verify the *memoryless* property**

$$\forall a, b > 0 : P(T > a + b | T > a) = P(T > b)....$$

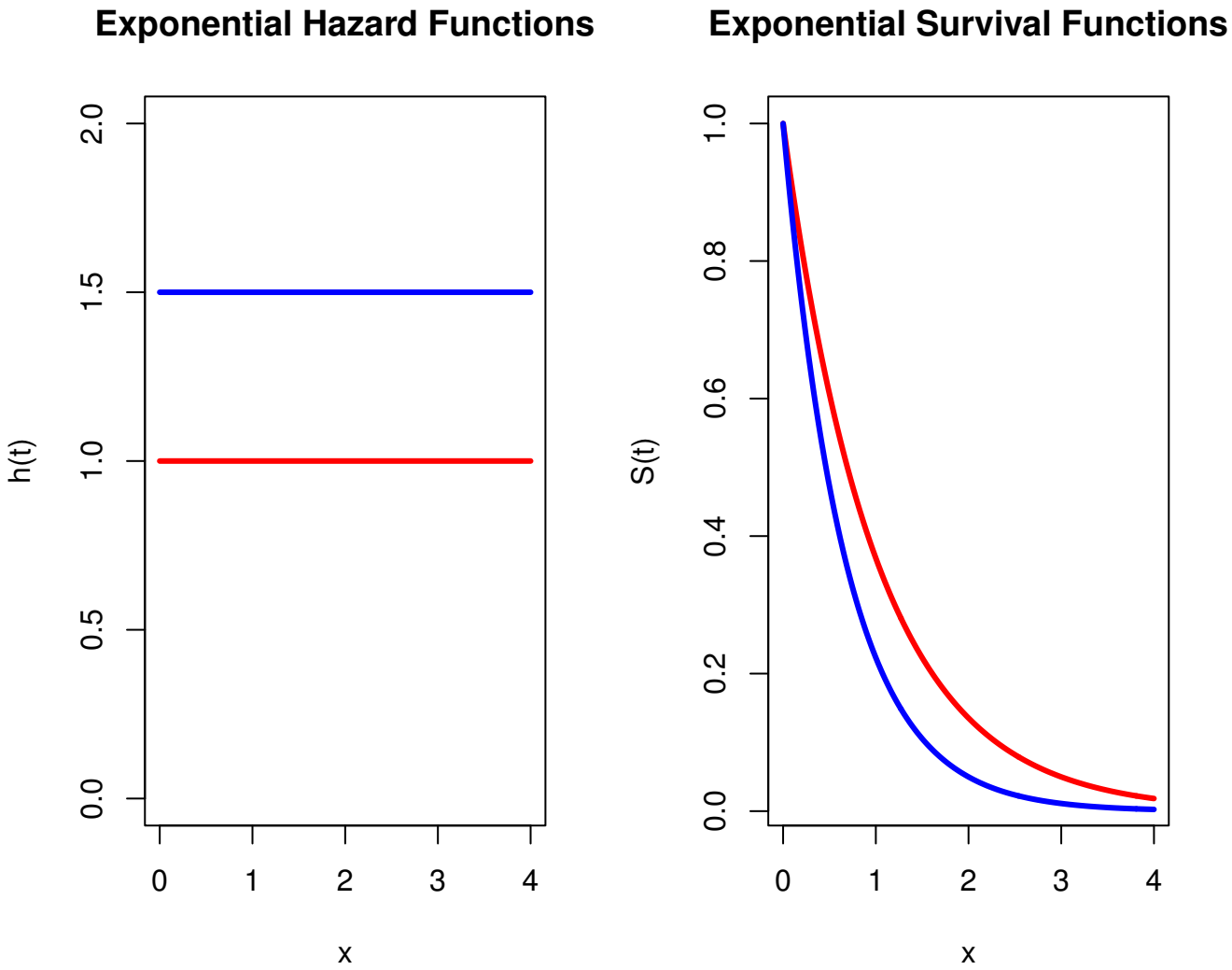
$$P(T>a+b \mid T>a) = \frac{P(T>a+b,\ T>a)}{P(T>a)} = \frac{P(T>a+b)}{P(T>a)}$$

$$\frac{S(a+b)}{S(a)}$$

$$\frac{\exp(-\lambda(a+b))}{\exp(-\lambda a)}$$

$$= \exp(-\lambda b)$$

$$= S(b)$$

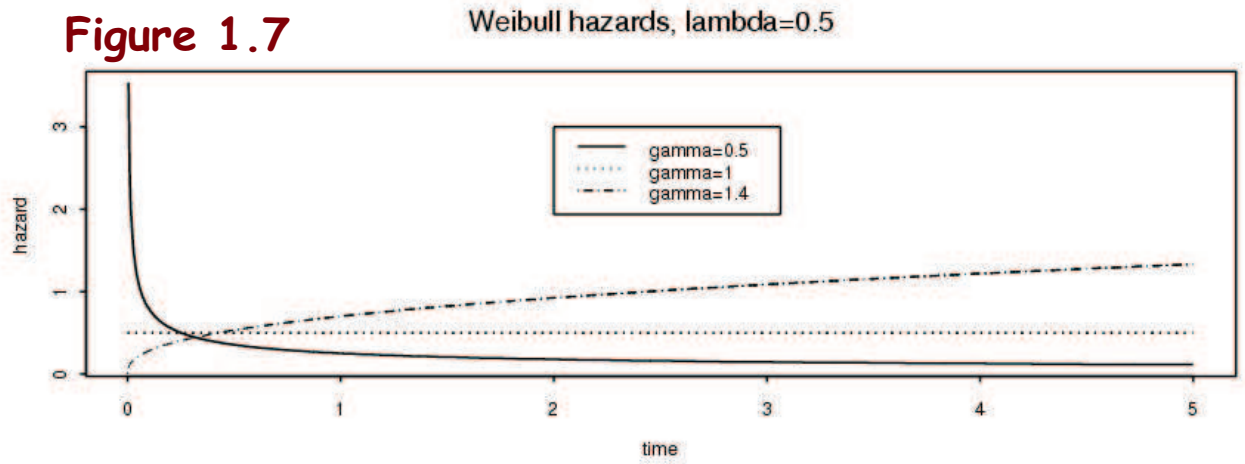

Exponential Hazard Functions
h(t)
0.0
0.5
1.0
1.5
2.0
0
1
2
3
4
x
Exponential Survival Functions
S(t)
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
x

- **The Weibull distribution (2 parameters)**
  **Generalizes exponential:**

$$
\begin{aligned}
S(t) &= e^{-\lambda t^{\gamma}} \\
f(t) &= \frac{-d}{dt} S(t) = \gamma \lambda t^{\gamma-1} e^{-\lambda t^{\gamma}} \\
\lambda(t) &= \gamma \lambda t^{\gamma-1} \\
\Lambda(t) &= \int_0^t \lambda(u) du = \lambda t^{\gamma}
\end{aligned}
$$

**$\lambda$ - the *scale* parameter**
**$\gamma$ - the *shape* parameter**
**$\gamma = 1 \rightarrow$ constant hazard**
**$0 < \gamma < 1 \rightarrow$ decreasing hazard**
**$\gamma > 1 \rightarrow$ increasing hazard**

**Figure 1.7**


Weibull hazards, lambda=0.5
gamma=0.5
gamma=1
gamma=1.4
hazard
time
0
1
2
3
4
5

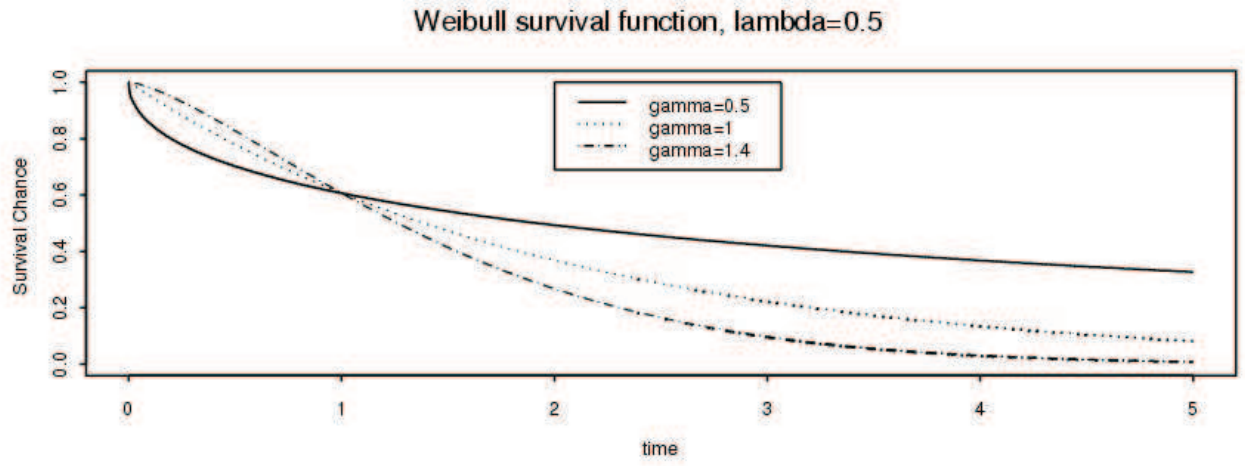
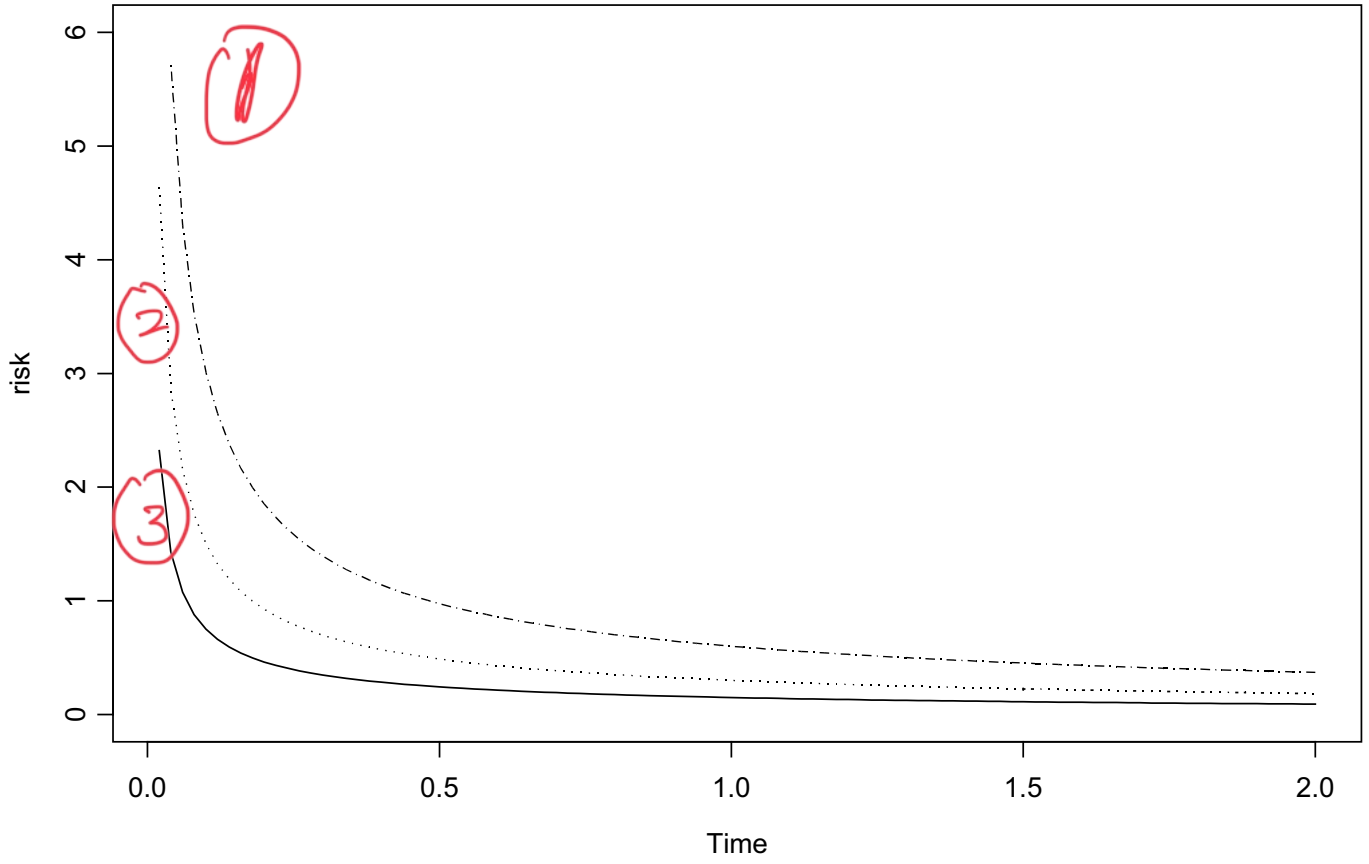

Weibull survival function, lambda=0.5
gamma=0.5
gamma=1
gamma=1.4
Survival Chance
time
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5

**Figure 1.8**

Weibull hazards with different scale parameter



- **Rayleigh distribution**

  **Another 2-parameter generalization of exponential:**

$$\lambda(t) = \lambda_0 + \lambda_1 t$$

- **Compound exponential distribution**

  **Per subject:** $T \sim \exp(\lambda)$,
  **Between subjects** $\lambda$ **varies according to some distribution:** $\lambda \sim g$
  **Yielding the (average) population density of failure time:**

$$f(t) = \int_0^\infty \lambda e^{-\lambda t} g(\lambda) d\lambda$$

  **Choosing as** $g$ **the density of the gamma density with mean** $\lambda_0$ **and index** $\kappa$ **:**

$$g(\lambda) = \frac{(\kappa/\lambda_0)(\kappa\lambda/\lambda_0)^{\kappa-1} e^{-\kappa\lambda/\lambda_0}}{\Gamma(\kappa)}$$

  **leads to the Pareto distribution**

- **Log-normal, Log-logistic:**

**Possible distributions for $T$ obtained by specifying for $\log T$, which is no longer constrained to be positive, any convenient family of distributions, e.g.**

$\log T \sim$ **normal (non-monotone hazard)**

$\log T \sim$ **logistic; log-logistic:** $\log \frac{S(t)}{1-S(t)} = -\theta - \kappa \log t$

**Figure 1.9**

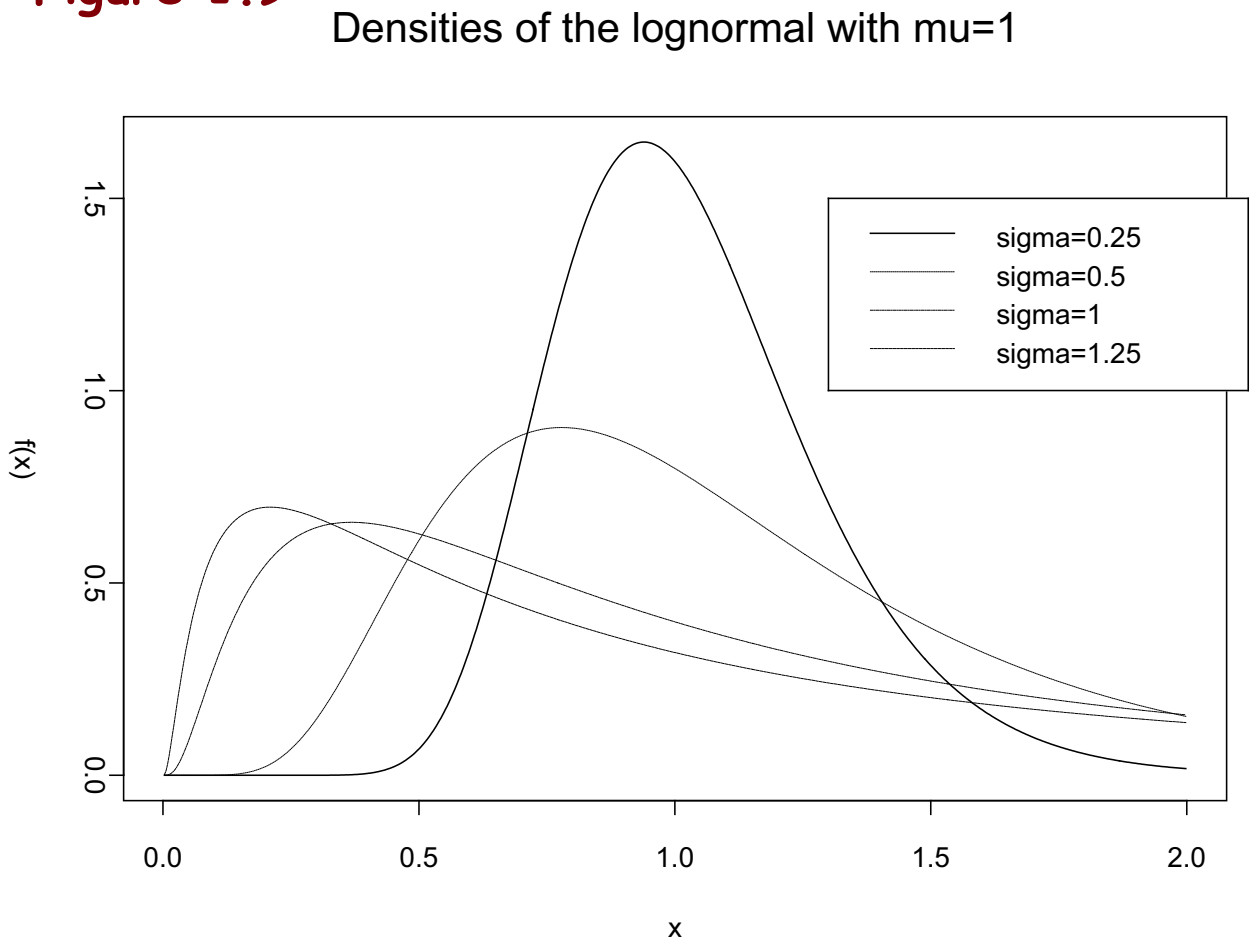
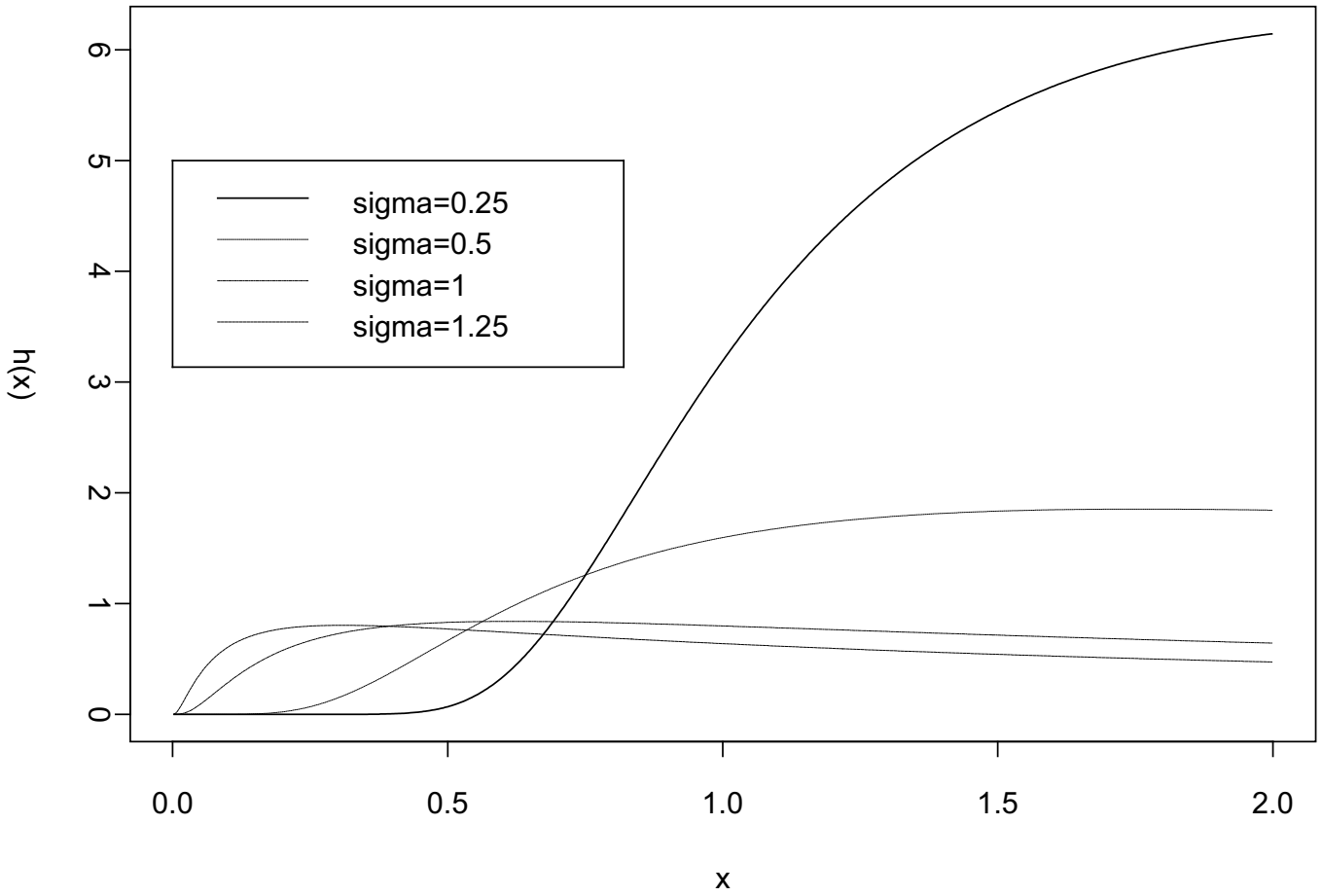

Densities of the lognormal with mu=1
sigma=0.25
sigma=0.5
sigma=1
sigma=1.25
f(x)
0.0
0.5
1.0
1.5
0.0
0.5
1.0
1.5
2.0
x

**Figure 1.10** Hazards of the lognormal with mu=1



## Piece-wise Exponential Distribution

**When the hazard is constant over different intervals of time.**
**For instance,**

| | | | |
|---|---|---|---|
| **From** | $t = 0$ **to** $t_1$ | **we have hazard** | $\lambda_1$ |
| **From** | $t_1$ **to** $t_2$ | **we have hazard** | $\lambda_2$ |
| **For** | $t > t_2$ | **we have hazard** | $\lambda_3$ |

## The Maximum Likelihood Estimator for $\lambda_i$ is then:

**(# events in $i-$th interval) / (person time follow-up in the $i-$th interval)**

**In "Clinical applications for change-point analysis of herpes zoster pain" (JPSM, 2002), the distribution of time to pain cessation is modeled, in terms of acute, subacute and chronic phases.**

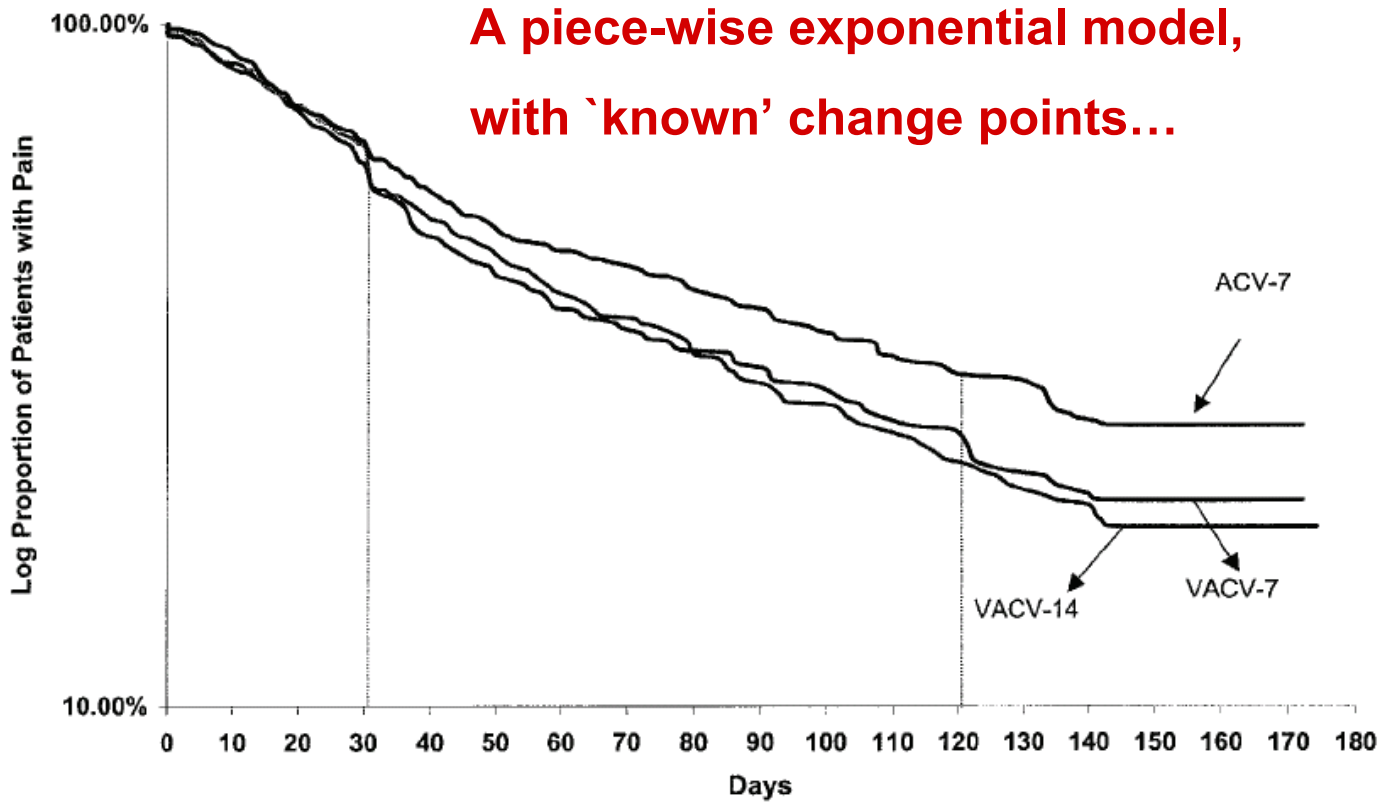



Fig. 2. Log of the Kaplan–Meier curves of time to cessation of pain for the VACV vs. ACV trial.

## Fitting a piece-wise exponential model...

*Table 4*
**Hazard Estimates for Each Phase in Study Populations**

| Study Population | $\lambda_1$ (SE) | $\lambda_2$ (SE) | $\lambda_3$ (SE) |
|---|---|---|---|
| VACV vs. ACV | 0.0136 (0.0008) | 0.0119 (0.0006) | 0.0131 (0.0014) |
| VACV ≥50 years | 0.0211 (0.0010) | 0.0171 (0.0008) | 0.0131 (0.0014) |
| VACV 18–49 years | 0.0419 (0.0019) | 0.0307 (0.0023) | 0.0020 (0.0048) |

## Why use one parametric model versus another?

- **technical convenience for estimation and inference**
- **explicit simple forms for $f(t), S(t)$, and $\lambda(t)$.**
- **qualitative shape of hazard function**
- **fit to the data (empirical survival distribution)**

**One can usually distinguish between a one-parameter model (like the exponential) and two-parameter (like Weibull or log-Normal) in terms of the adequacy of fit to a dataset.**

**Without a lot of data, it may be hard to distinguish between the fits of various 2-parameter models (i.e., Weibull vs log-normal)**

# Contents


1 Survival Analysis - Introduction 7

1.1 Course Focus . . . . . 8

1.2 Some useful references . . . . . 19

1.3 Basic Definitions and Notation . . . . . 20

1.4 Types of censoring . . . . . 25

1.5 Independent vs informative censoring . . . . . 29

1.6 Some published studies: see /Modules/Readings . . . . . 33

1.7 Some example datasets (will be used in notes and HW) . . 38

1.8 The population distribution of survival times . . . . . 45

1.8.1 **Density function (Probability Mass Function)** . . . . . 52

1.8.2 **Survivorship Function: $S(t) = P(T > t)$.** . . . . . 55

1.8.3 **Hazard Function $\lambda(t)$** . . . . . 57

1.8.4 **Cumulative Hazard Function $\Lambda(t)$** . . . . . 60

1.8.5 **Relationship between $S(t)$ and $\lambda(t)$ :** . . . . . 61

1.8.6 **Relationship between $S(t)$ and $\Lambda(t)$ :** . . . . . 62

1.9 **Measuring Central Tendency in Survival Times** . . . . . 66

1.10 **Some Parametric Survival Distributions** . . . . . 69