

# **Survival Analysis - INTRODUCTION**

## **1.General Introduction to Survival Analysis**

### **1.1.What is Survival Analysis**

### **1.2.What is Survival Data**

### **1.3. Why use of Survival Analysis?**

### **1.4. Basic Approaches to the Survival Analysis**

## **2. Fundamentals of Survival Analysis**

## **3. Censoring & truncation**

## **4. Stochastic Process Approach**

## **5. Describing Survival Distributions – Parametric Models**

## **6. Non-parametric models**

### **6.1. Life Table Method**

### **6.2. Kaplan-Meier Method**

### **6.3. Nelson – Aalen Method and other estimators**

## **7. Semi-parametric models**

## **SUMMARY**

# 1. General Introduction to Survival Analysis

# 1.1.What is Survival Analysis?

***Survival analysis is extremely*** useful for studying many different kinds of events in both the social and natural sciences, including disease onset, equipment failures, earthquakes, automobile accidents, stock market crashes, revolutions, job terminations, births, marriages, divorces, promotions, retirements, and arrests.

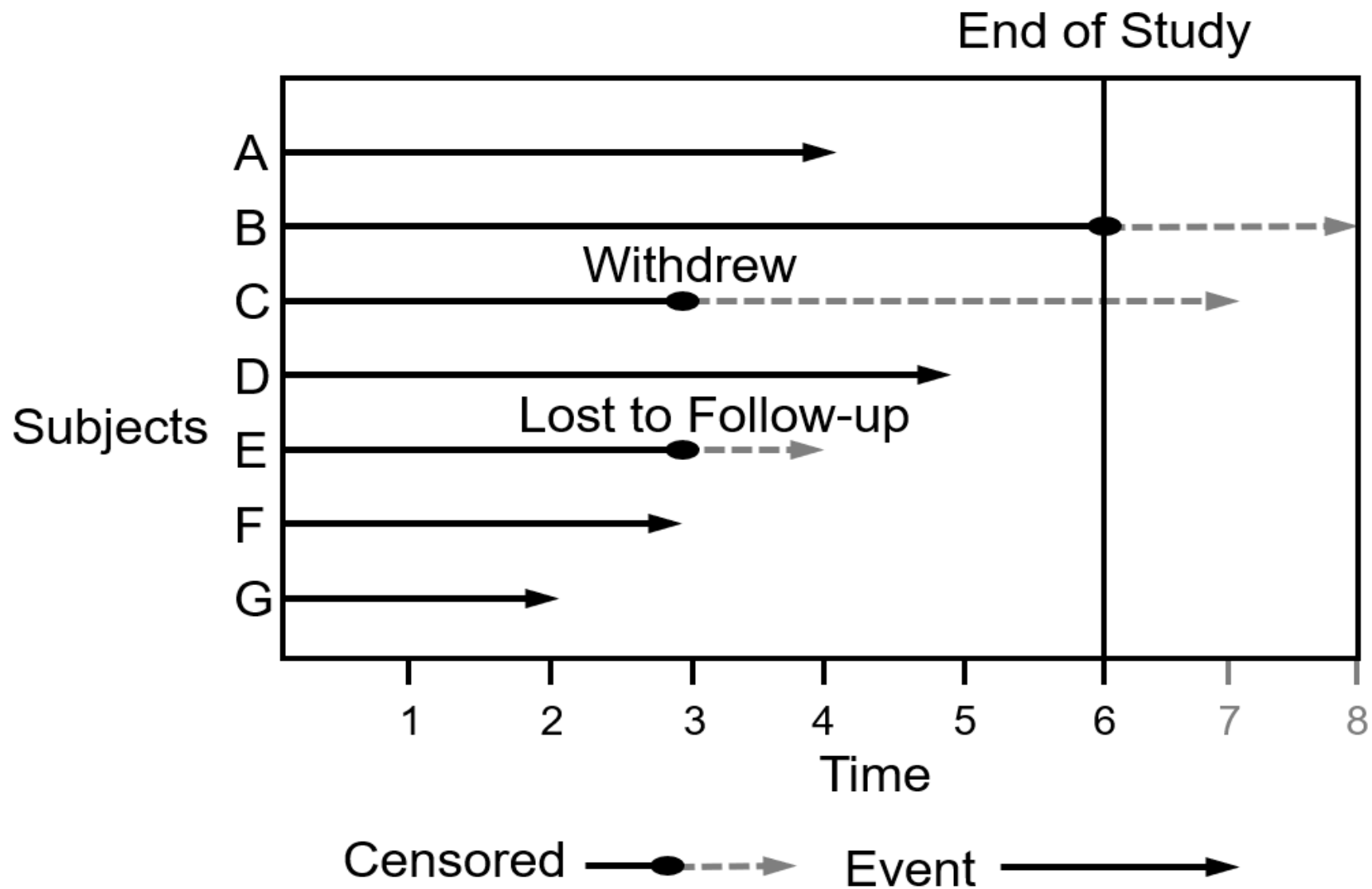
Because these methods have been adapted—and sometimes independently discovered—by researchers in several different fields, they also go by several different names:

**event history analysis** (sociology),  
**reliability analysis** (engineering),  
**failure time analysis** (engineering),  
**duration analysis** (economics), **and transition analysis** (economics).

These different names don't imply any real difference in techniques, although different disciplines may emphasize slightly different approaches. Because survival analysis is the name that is most widely used and recognized, it is the name I use here.

Although one have performed survival analysis with many different statistical packages.

# What Is Survival Analysis?



# Final answer :

## What Is Survival Analysis?

- *Survival analysis* is a class of statistical methods for which the outcome variable of interest is time until an event occurs.
- Time is measured from when an individual or organization first becomes a customer until the event occurs or until the end of the observation interval.
- In survival analysis, the basis of the analysis is tenure, or time at risk for the event, and not calendar time.

# 1.2.What is Survival Data?



Survival analysis was designed for longitudinal data on the occurrence of events.

But what is an event? Biostatisticians haven't written much about this question because they have been overwhelmingly concerned with deaths.

When you consider other kinds of events, however, it's important to clarify what is an event and what is not. I define an *event* as a qualitative change that can be situated in time.

By a *qualitative change*, I mean a transition from one discrete state to another. A marriage, for example, is a transition from the state of being unmarried to the state of being married.

**A promotion consists of the transition from a job at one level to a job at a higher level.**

But I have reservations about the application of survival methods when the threshold is arbitrarily set by the researcher. Ideally, statistical models should reflect the process generating the observations.

It's hard to see how such arbitrary thresholds can accurately represent the phenomenon under investigation.

For survival analysis, the best observation plan is prospective.

**It's not necessary that every individual experiences the event.**

**It's not necessary that every individual experiences the event.**

You can perform survival analysis when the data consist *only* of the times of events, but a common aim of survival analysis is to estimate causal or predictive models in which the risk of an event depends on covariates.

If this is the goal, the data set must obviously contain measurements of the covariates. Some of these covariates, like race and sex, may be constant over time. Others, like income, marital status, or blood pressure, may vary with time.

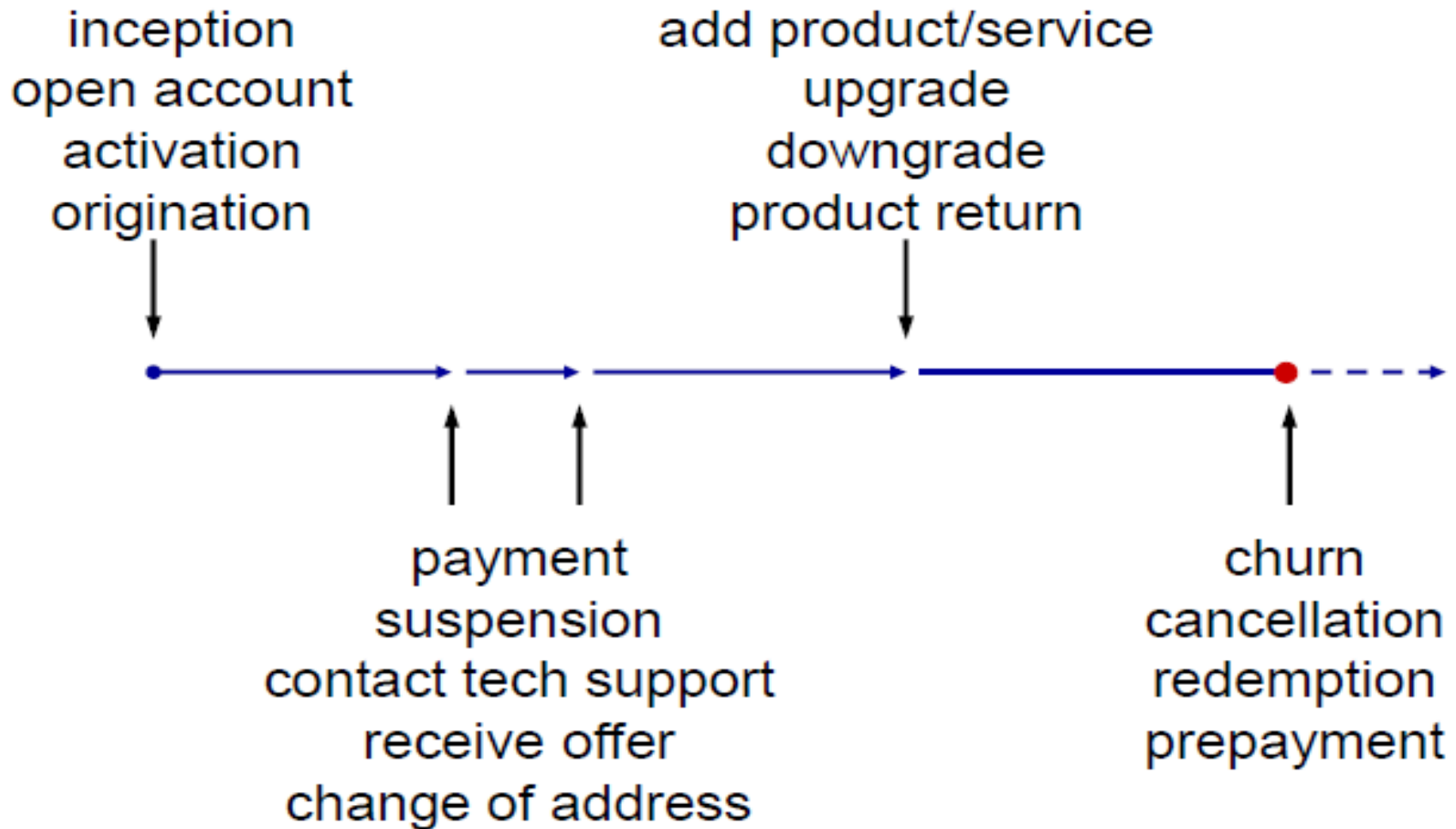
For time-varying covariates, the data set should include as much detail as possible on their temporal variation.

Survival analysis is frequently used with *retrospective* data in which people are asked to recall the dates of events like marriages, child births, and promotions.

There is nothing intrinsically wrong with this approach as long as you recognize the potential limitations.

For one thing, people may make substantial errors in recalling the times of events, and they may forget some events entirely. They may also have difficulty providing accurate information on time-dependent covariates.

# Customer History Data



# Time-Dependent Customer Outcomes

## Customer retention applications

- cancellation of all products and services
- severe downgrade or extreme inactivity
- unprofitable behavior

## Add-on selling applications

- acquisition of the target product or service
- more profitable behavior

## Credit risk management applications

- charge-off
- loan termination

# Extracting and Preparing the Relevant Data

## Customer Data

- identifier: account number, person, household
- billing details: address, amount, payment method
- application information: demographics, credit history

## Product Data

- features: description, category, rate plan
- status: start/stop date, cancellation reason, changes, suspensions, disconnections
- usage: activity, amount, duration, balance, payment

## Contact Data

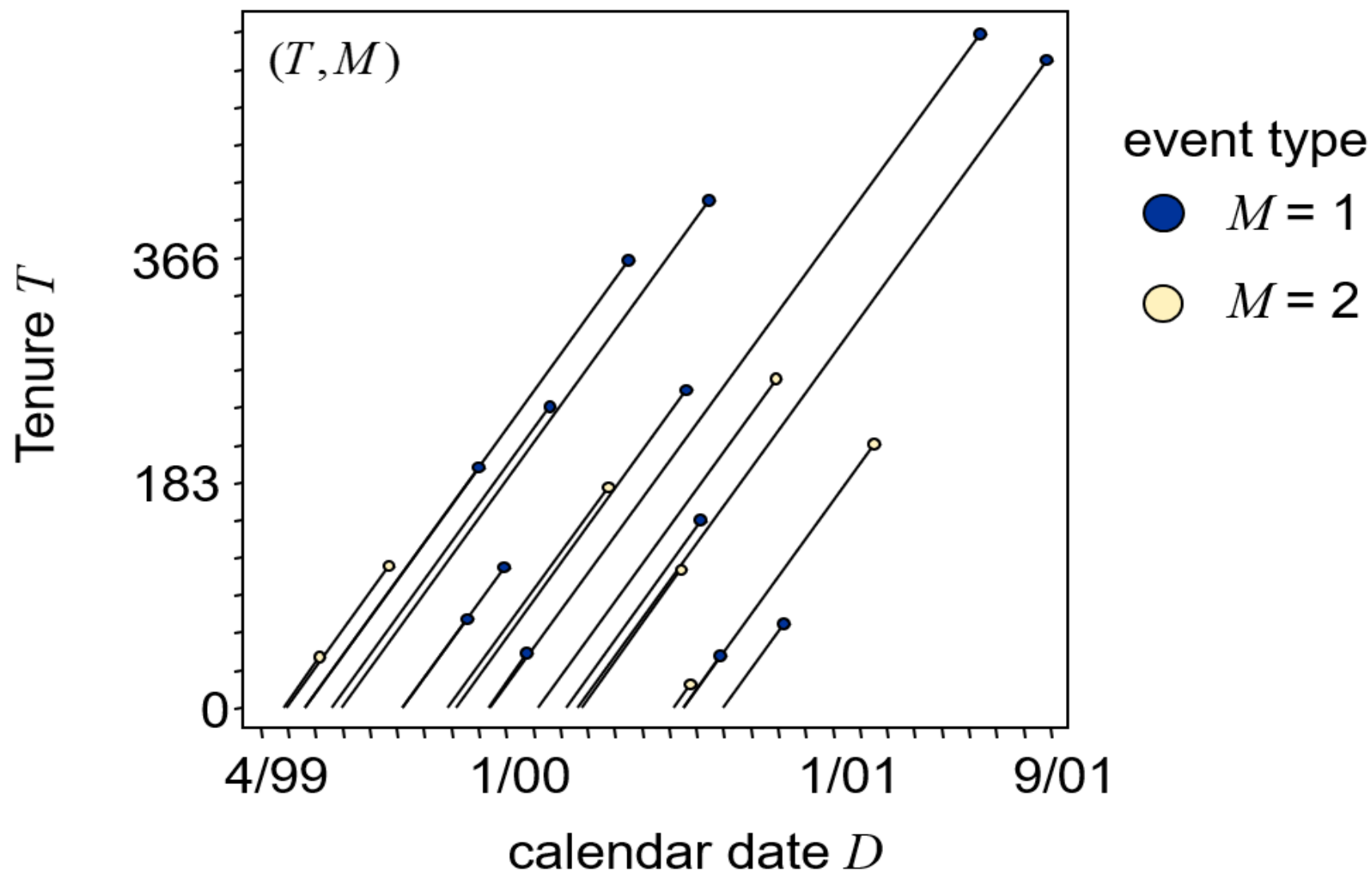
- marketing promotions: direct mail, call center
- customer service: technical support, billing inquiries

# Data Structure

Customer	Tenure	Status
A	4.0	1 (event)
B	6.0	0 (censored)
C	3.0	0
D	5.0	1
E	3.0	0
F	3.0	1
G	2.0	1



# Event Time and Type



# Event-Time Distribution

Duration between start and event  
for continuous time

$$T = D^{(\text{event})} - D^{(\text{start})}$$

Discrete random variable

- smallest meaningful unit
- days, months, billing cycles
- many ties

$$T = D^{(\text{event})} - D^{(\text{start})} = 0, 1, 2, \dots$$

Covariates

- time-independent covariates
- time-dependent covariates

$$\mathbf{x} = (x_1, \dots, x_p)$$

$$\mathbf{x}(t) = (x_1(t), \dots, x_p(t))$$

## 1.3. Why Use of Survival Analysis?

Survival data have two common features that are difficult to handle with conventional statistical methods: *censoring* and *time-dependent covariates* (sometimes called *time-varying explanatory variables*).

Some of these covariates (like race, age at release, and number of previous convictions) remained constant over the one-year interval.

Others (like marital and employment status) could change at any time during the follow-up period.

How do you analyze such data using conventional methods?

One possibility is to perform a logistic regression analysis with a dichotomous dependent variable ( 1 yes ; 0 no)

But this analysis ignores information on the timing of the process and event of the study.

By contrast, all methods of survival analysis allow for censoring, and many also allow for time-dependent covariates.

In the case of censoring, the trick is to devise a procedure that combines the information in the censored and uncensored cases in a way that produces consistent estimates of the parameters of interest.

You can easily accomplish this by the method of maximum likelihood or its close cousin, partial likelihood.

Time-dependent covariates can also be incorporated with these likelihood-based methods.

The next, very important research issues in event history analysis (especially in social and economic sciences ) are studying the mechanisms of causality and their modeling in the context of substantive process (processes).

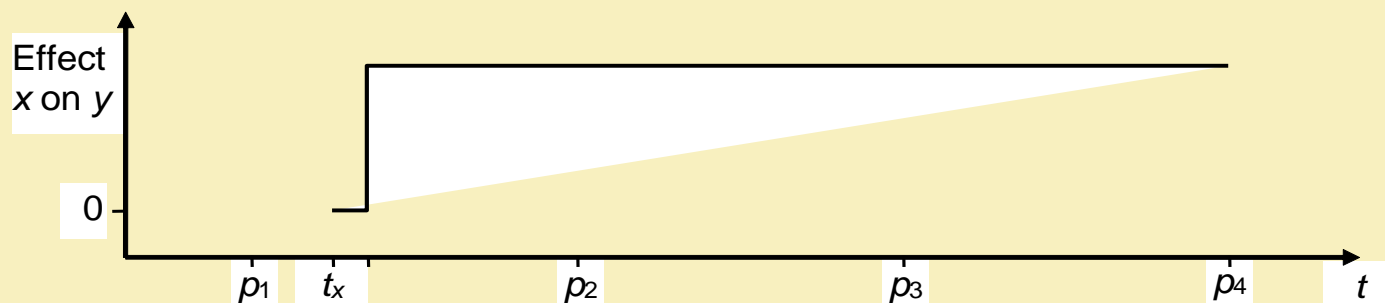
Substantive processes that generally speaking, the biological, psychological, economic and social changes that precede the occurrence of life events. As an example, processes: disease, decision making, social control.

One of the main challenges for event history analysis is the theory and modeling of causality of the events expressed in terms of (literally, in the circumstances) substantive process.

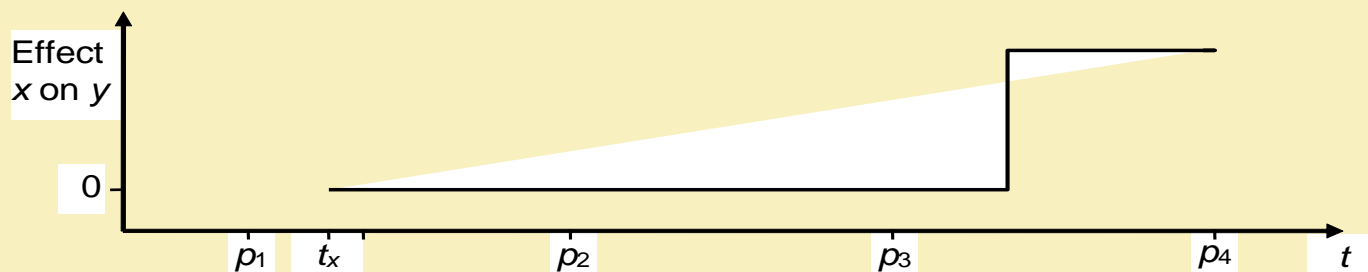


***Different shapes influence the outcome of time, showing how changes in the levels of the variable  $x$  occurring at time  $t$ , to affect changes in the variable  $y$***

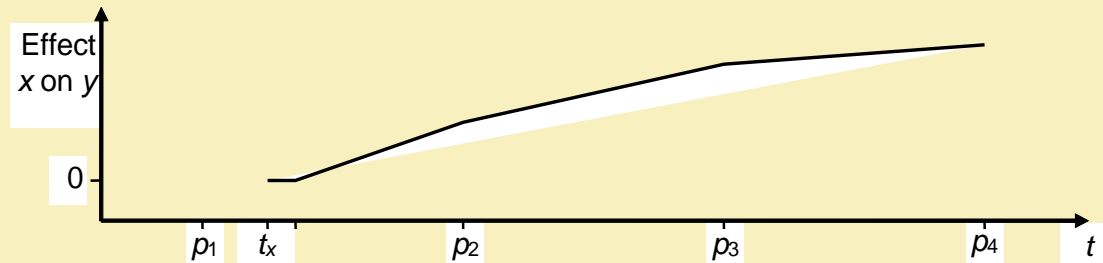
**2.2.A. Effect occurs almost immediately and is then time-constant**



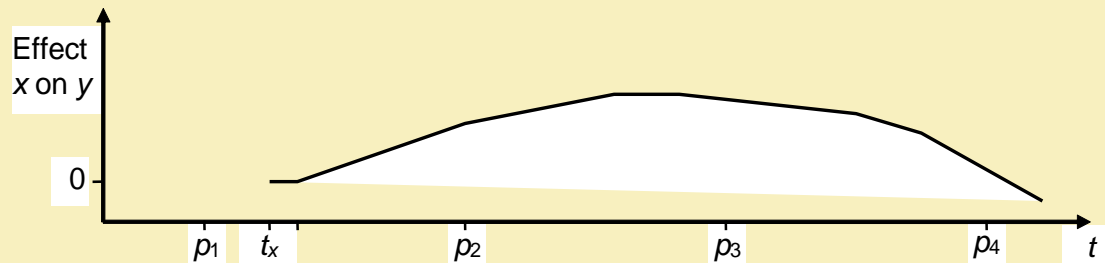
**2.2.B. Effect occurs with a certain time-lag and is then time-constant**



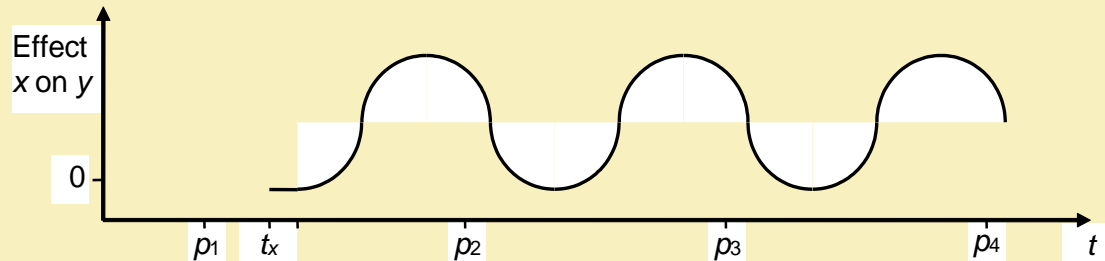
### 2.2.C. Effect occurs almost immediately and then increases continuously



### 2.2.D. Effect occurs almost immediately, rises monotonically at first, then declines and finally disappears



### 2.2.E. Effect occurs almost immediately and oscillates over time



## 1.4. Basic Approaches to the Survival Analysis

One of the confusing things about survival analysis is that there are so many different methods: life tables, Kaplan-Meier estimators, exponential regression, log-normal regression, proportional hazards regression, competing risks models, and discrete-time methods, to name only a few.

Sometimes these methods are complementary. Life tables have a very different purpose than regression models, for example, and discrete-time methods are designed for a different kind of data than continuous-time methods.

On the other hand, it frequently happens that two or more methods may seem attractive for a given application, and the researcher may be hard-pressed to find a good reason for choosing one over another.

# Survival Models

- Models in survival analysis are written in terms of the hazard function.
- They assess the relationship of predictor variables to survival time.
- They can be non-parametric, parametric or semi-parametric models.

# Type of the SURVIVAL MODELS

Type of model	Time * Failure = f(covariates)	
Non-Parametric	No Distribution assumed	No Distribution assumed
Semi-Parametric	No Distribution assumed	Some Distribution assumed
Parametric	Some Distribution assumed	Some Distribution assumed

Source: <https://www.analyticsvidhya.com/blog/2015/05/comprehensive-guide-parametric-survival-analysis/>

## 2. Fundamentals of Survival Analysis

**Basic notions: assumption single episode model (*one origin state and one destination state* ,  $T$  – *continuous variable*):**

**Basic measures of the T distribution are :**

*Cumulative Distribution Function*

*Probability Density Function*

*Survival Function*

*Hazard Function*

*Cumulative Hazard Function*

*Likelihood Function*



# Cumulative Distribution Function

One way that works for all random variables is the ***cumulative distribution function***, or *c.d.f.* The c.d.f. of a variable  $T$ , denoted by  $F(t)$ , is a function that tells us the probability that the variable will be less than or equal to any value  $t$  that we choose.

or,

**Lifetime Distribution Function (F) :**

This is the probability of failure happening before a time 'T'

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

# Probability Density Function

When variables are continuous, another common way of describing their probability distributions is the *probability density function*, or *p.d.f.* This function is defined as

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}$$

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{\Pr(t \leq T < t + \Delta t)}{\Delta t}$$

That is, the p.d.f. is just the derivative or slope of the c.d.f. Although this definition is considerably less intuitive than that for the c.d.f., it is the p.d.f. that most directly corresponds to our intuitive notions of distributional shape. For example, the familiar bell-shaped curve that is associated with the normal distribution is given by its p.d.f., not its c.d.f.

# Survival Function

Survival is the inverse of Cumulative Distribution Function (CDF). It is one minus CDF .

$$S(t) = \Pr(T > t) = \int_t^{\infty} f(u) du = 1 - F(t).$$

$$S(t) + F(t) = 1 ,$$

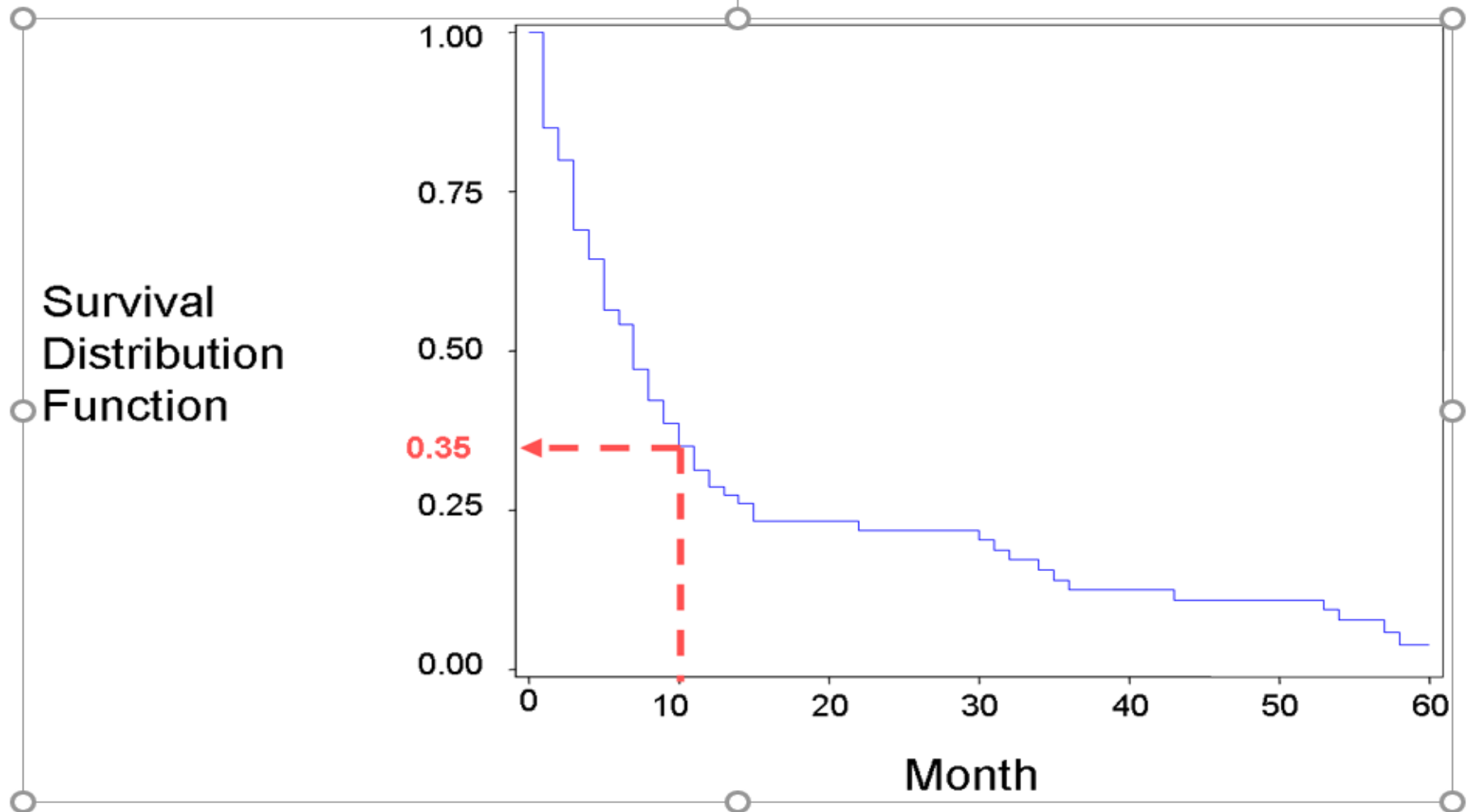
$$S(t) = 1 - F(t)$$

$$F(t) = 1 - S(t)$$

## Properties

- $S(0) = 1$
- $\lim_{t \rightarrow \infty} S(t) = 0$
- $S(t_a) \geq S(t_b) \iff t_a \leq t_b$
- $S(t) = 1 - F(t) = \int_t^{\infty} f(\tau) d\tau$

# Survival Function for Continuous Time



# Hazard Function

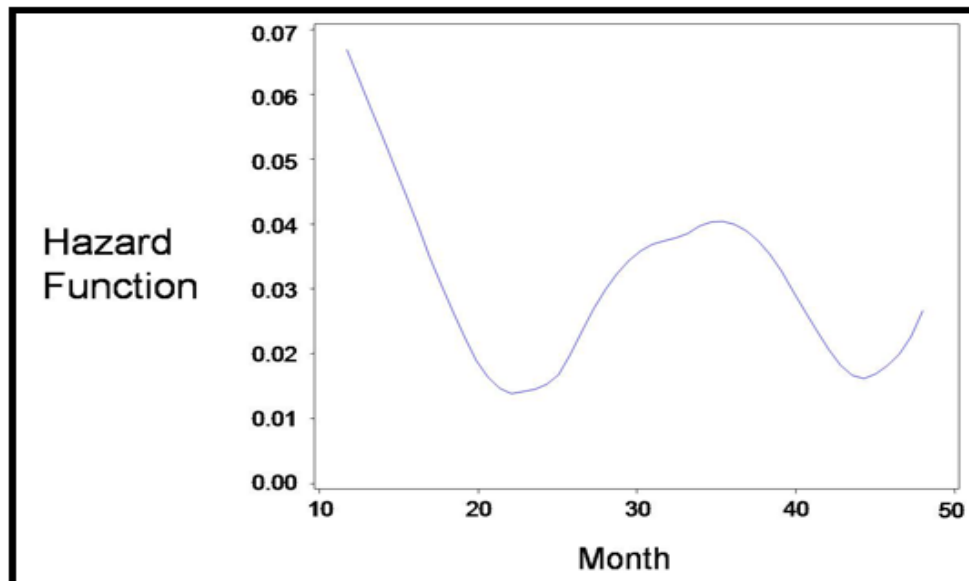
For continuous survival data, the *hazard function* is actually more popular than the p.d.f. as a way of describing distributions. The hazard function is defined as

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

Instead of  $h(t)$ , some authors denote the hazard by  $\lambda(t)$  or  $r(t)$ . Because the hazard function is so central to survival analysis, it is worth taking some time to explain this definition. The aim of the definition is to quantify the instantaneous risk that an event will occur at time  $t$ . Because time is continuous, the probability that an event will occur at exactly time  $t$  is necessarily 0. But we *can* talk about the probability that an event occurs in the small interval between  $t$  and  $t + \Delta t$ .

# Hazard Function for Continuous Time

- The hazard function is the instantaneous risk or potential that an event will occur at tenure  $t$ , given that the individual has survived up to tenure  $t$ .
- It takes the form of the expected number of events per interval of tenure.
- For tenure measured on a continuous scale, it is a rate, not a probability, that ranges from zero to infinity.



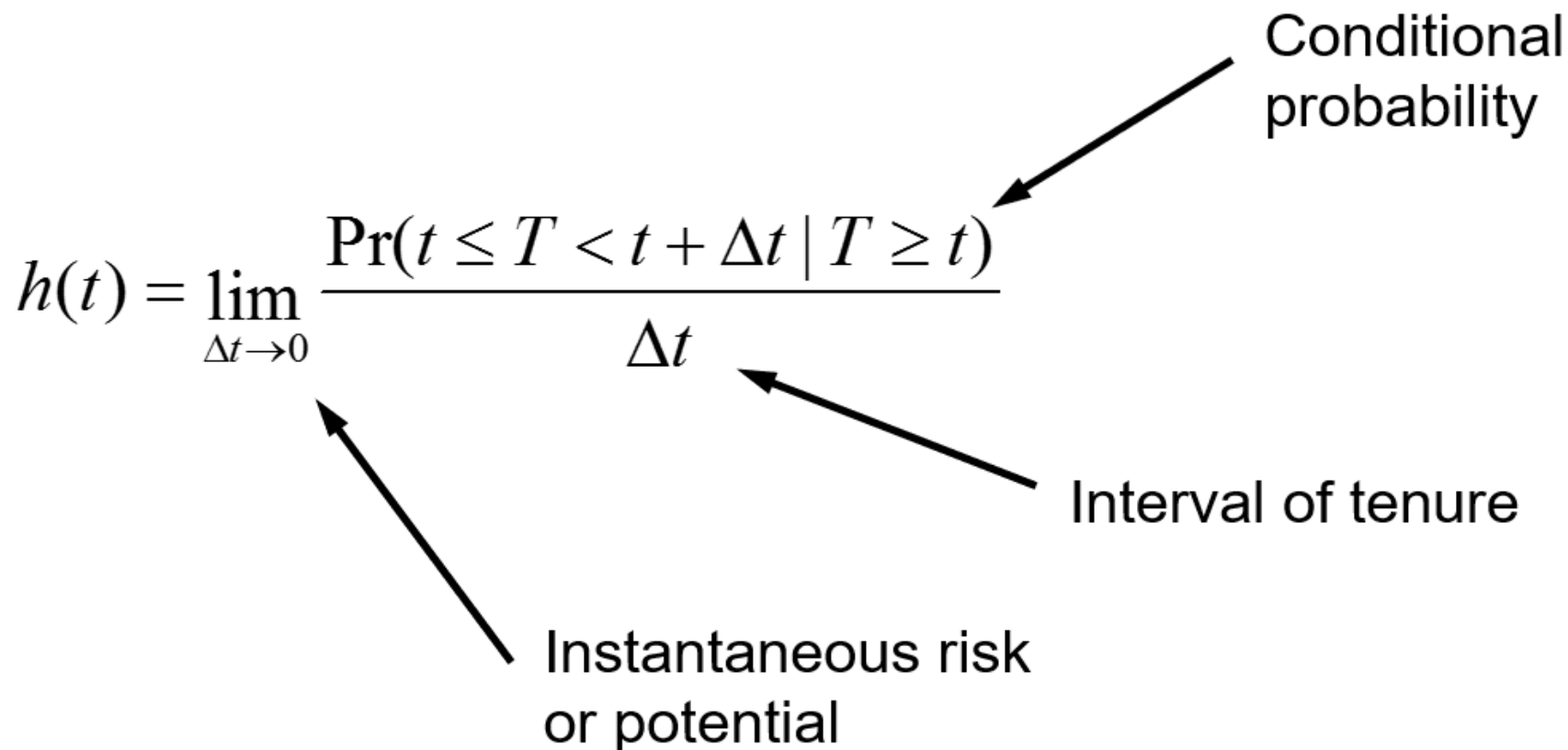
# Hazard Function for Continuous Time

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

Conditional probability

Interval of tenure

Instantaneous risk or potential



# Hazard Function for Discrete Time

$$h(t) = \Pr(T = t \mid T \geq t)$$

= probability of having the event at tenure  $t$  given no prior occurrence of the event

$$= 1 - \left( \frac{S(t)}{S(t-1)} \right)$$

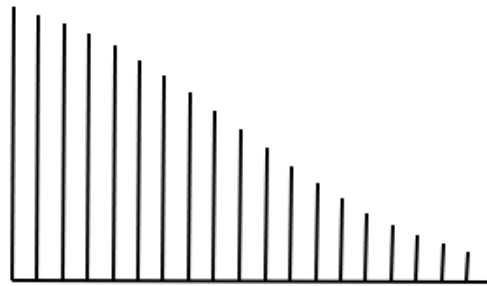


# Hazard Shapes

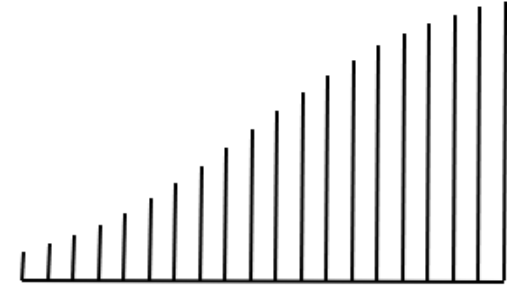
constant



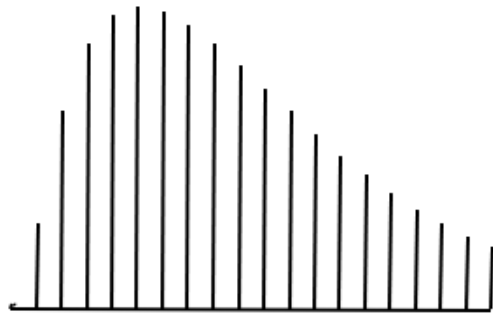
decreasing



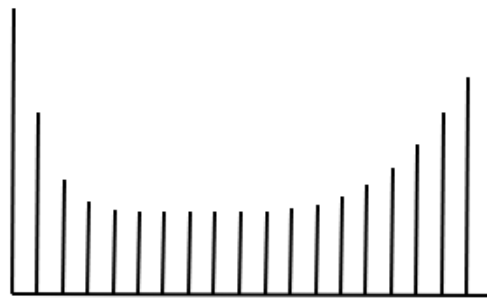
increasing



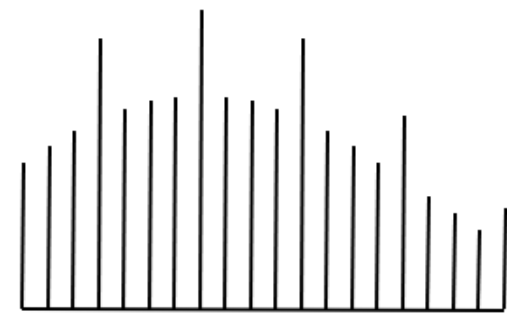
humped



bathtub



spiky



# Cumulative HAZARD Function

**Cumulative Hazard Function ( $\Lambda(t)$  or  $H(t)$  ):** This is simply the integral of the hazard function and is given as below.

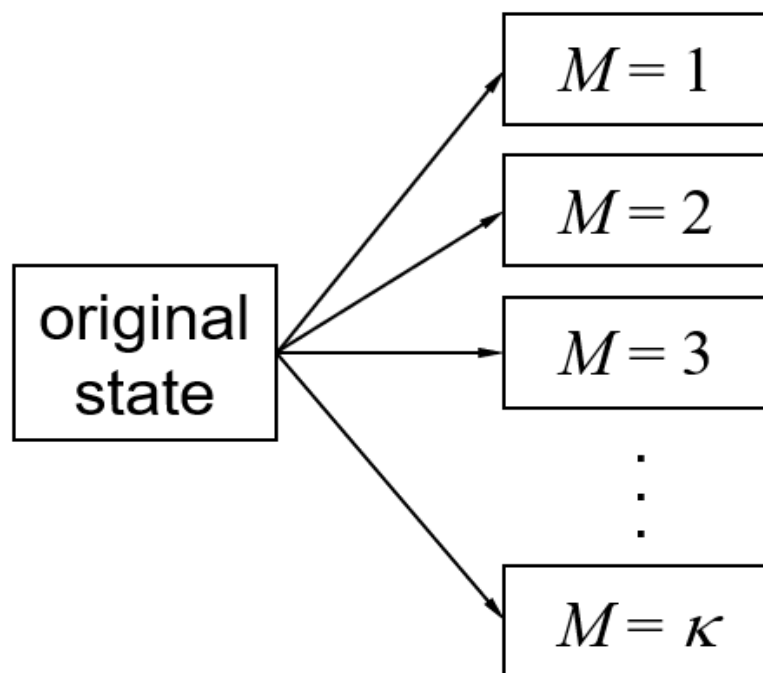
$$\Lambda(t) = \int_0^t \lambda(u) du$$

|

$$\Lambda(t) = -\log S(t)$$

# Competing Risks

$$M \in \{1, 2, \dots, \kappa\}$$



- multiple cancellation reasons (causes of death)
- voluntary or involuntary churn
- loan prepayment or default
- next product acquired
- outcome severity

# INTERPRETATIONS OF THE HAZARD FUNCTION

Before proceeding further, three clarifications need to be made:

- Although it may be helpful to think of the hazard as the instantaneous probability of an event at time  $t$ , it's not really a probability because the hazard can be greater than 1.0. This can happen because of the division by  $\Delta t$  in equation for hazard function. ***Although the hazard has no upper bound, it cannot be less than 0.***
- Because the hazard is defined in terms of a probability (which is never directly observed), it is itself an unobserved quantity. We may estimate the hazard with data, but that's only an estimate.
- It's most useful to think of the hazard as a characteristic of individuals, not of populations or samples (unless everyone in the population is exactly the same). Each individual may have a hazard function that is completely different from anyone else's.

The hazard function is much more than just a convenient way of describing a probability distribution. In fact, the hazard at any point  $t$  corresponds directly to intuitive notions of the risk of event occurrence at time  $t$ .

With regard to numerical magnitude, the hazard is a dimensional quantity that has the form *number of events per interval of time*, which is why the hazard is sometimes called a *rate*. To interpret the value of the hazard, then, you must know the units in which time is measured.

Suppose, for example, that I somehow know that my hazard for contracting flu at some particular point in time is .015, with time measured in months.

This means that if my hazard stays at that value over a period of one month, I would expect to contract influenza .015 times.

Remember, this is not a probability. If my hazard was 1.3 with time measured in years, then I would expect to contract flu 1.3 times over the course of a year (assuming that my hazard stays constant during that year).

To make this more concrete, consider a simple but effective way of estimating the hazard.

Suppose that we observe a sample of 10,000 people over a period of one month, and we find 75 cases of flu. If every person is observed for the full month, the total exposure time is 10,000 months.

Assuming that the hazard is constant over the month and across individuals, an optimal estimate of the hazard is  $75/10000 = .0075$ .

If some people died or withdrew from the study during the one-month interval, we have to subtract their *unobserved* time from the denominator.

## Basic relationships between functions:

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

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

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

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



# LIKELIHOOD FUNCTION

$$L = \prod_{i=1} f(t_i) \prod_j S(t_j) = \prod_k h(t_k)^{\delta_k} S(t_k)$$

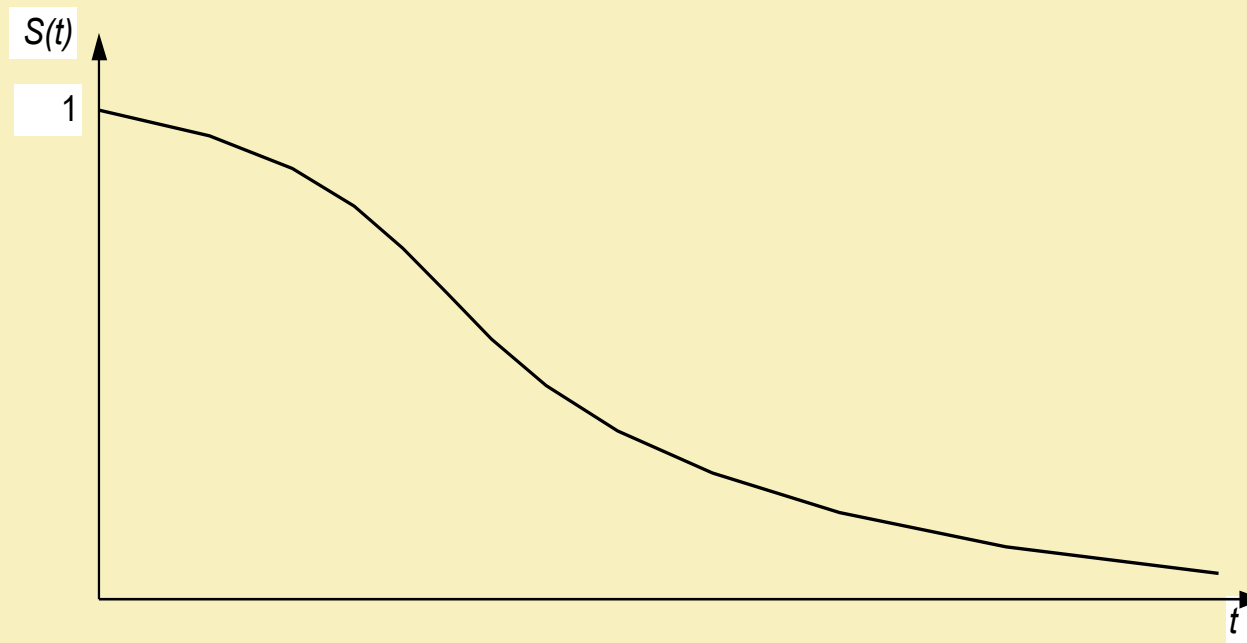
**Where :**

$\delta_i$  – censoring indicator with the values:

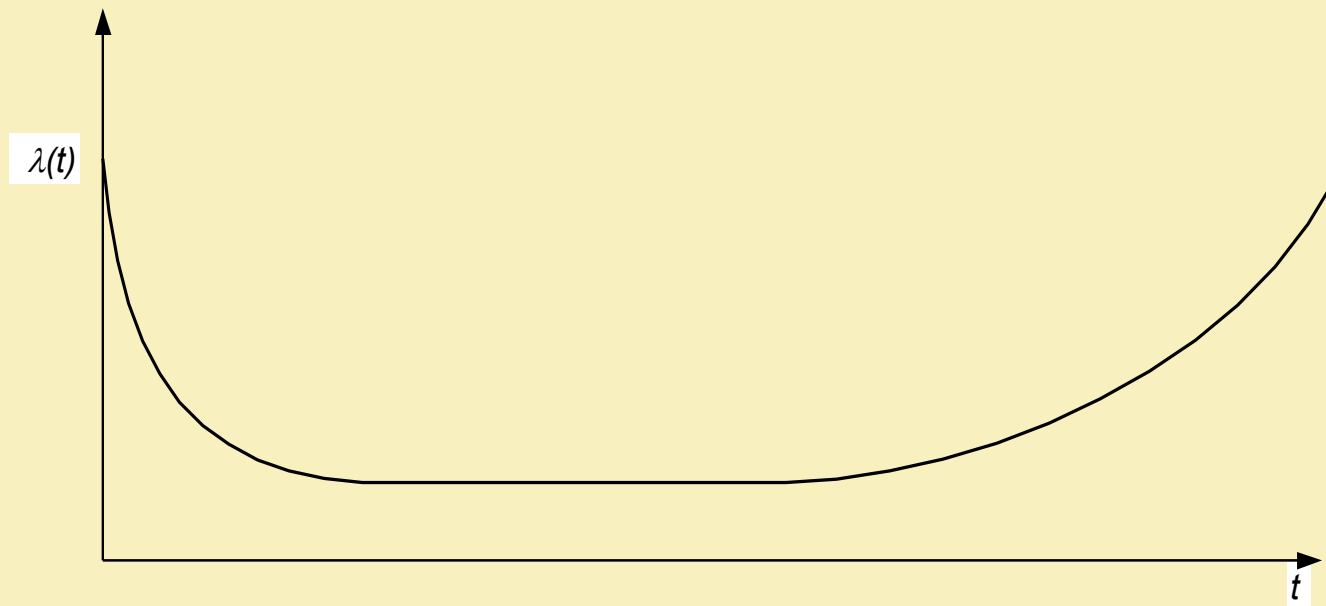
1,0 – in the case the event took place (at moment )  $t$ ,

0,0 – in the case , when the information was censored .

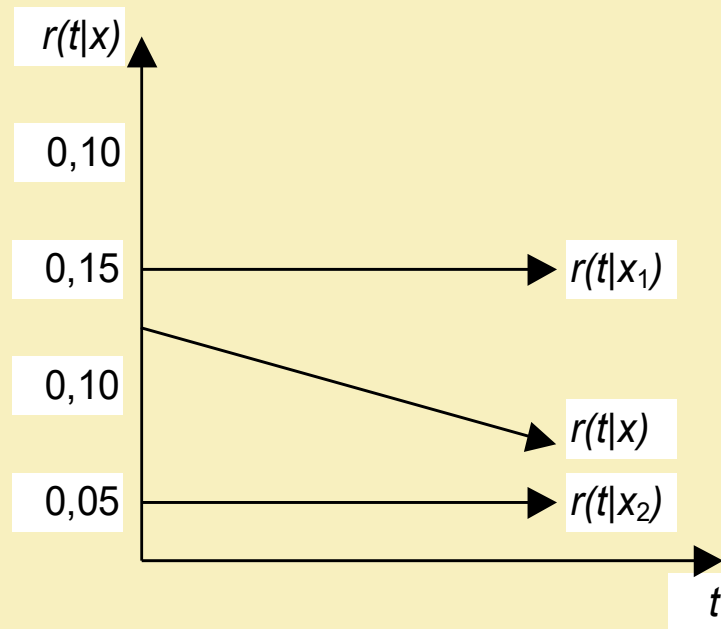
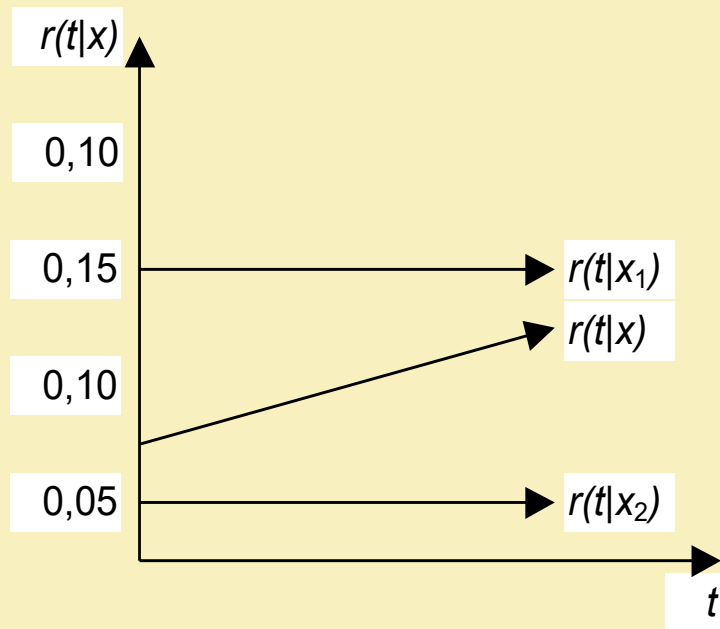
# Survival function for human population



# Hazard function for human population



## An example of hazard functions for: two sub-groups and population



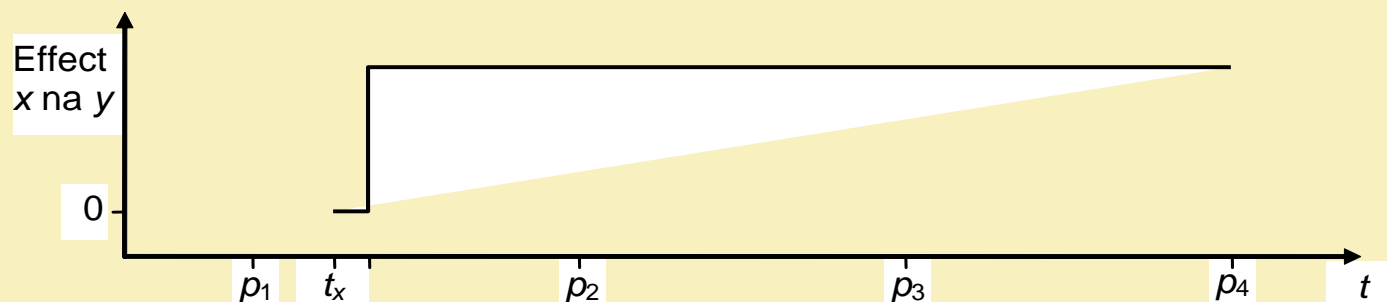
The next, very important research issues in event history analysis (especially in social and economic sciences ) are studying the mechanisms of causality and their modeling in the context of substantive process (processes).

Substantive processes that generally speaking, the biological, psychological, economic and social changes that precede the occurrence of life events. As an example, processes: disease, decision making, social control.

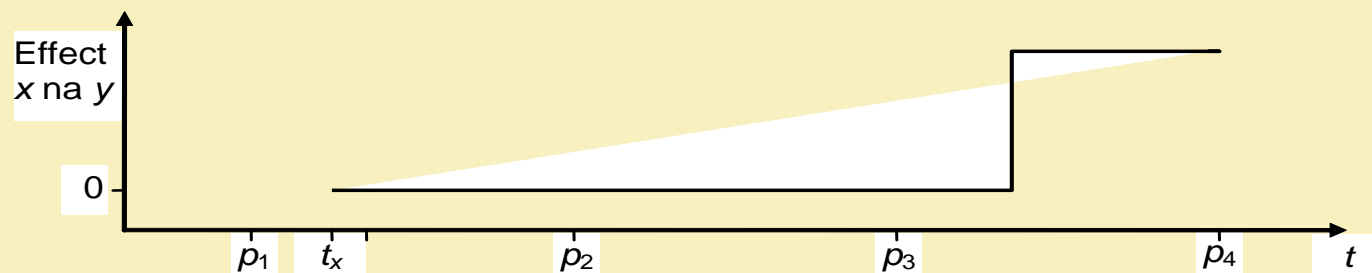
One of the main challenges for event history analysis is the theory and modeling of causality of the events expressed in terms of (literally, in the circumstances) substantive process.

***Different shapes influence the outcome of time, showing how changes in the levels of the variable  $x$  occurring at time  $t$ , to affect changes in the variable  $y$***

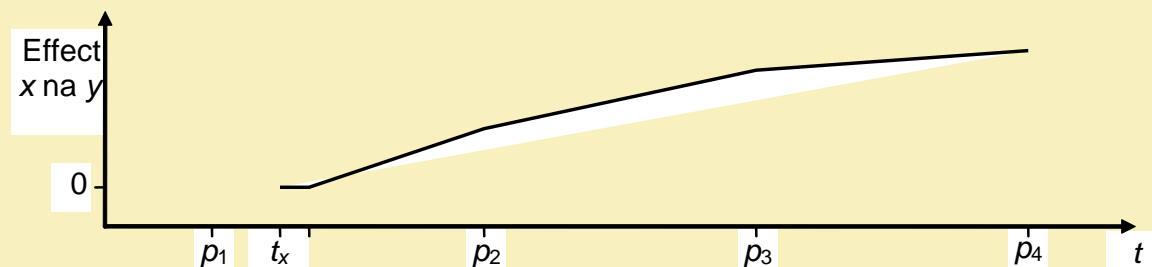
**2.2.A. Effect occurs almost immediately and is then time-constant**



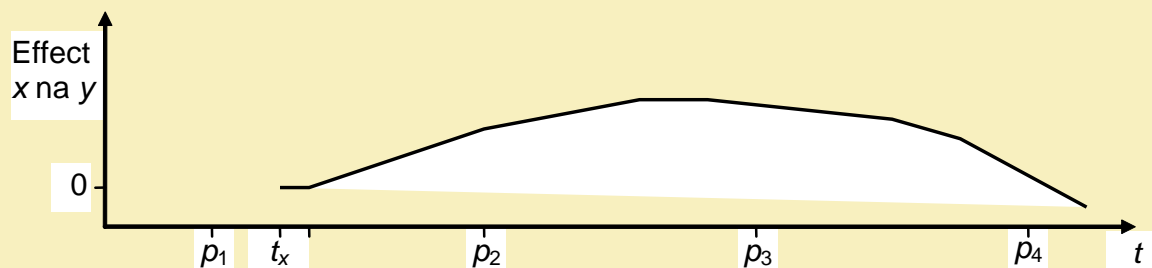
**2.2.B. Effect occurs with a certain time-lag and is then time-constant**



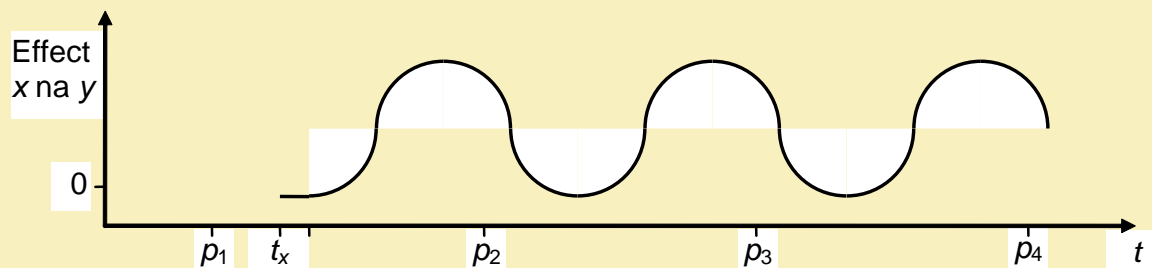
### 2.2.C. Effect occurs almost immediately and then increases continuously



### 2.2.D. Effect occurs almost immediately, rises monotonically at first, then declines and finally disappears



### 2.2.E. Effect occurs almost immediately and oscillates over time



# 3. Censoring & Truncation



# Censoring and Truncation

complete data  $\{(t_i, m_i, \mathbf{x}_i)\}_{i=1}^n$

- The observed data are not simply realizations of the random variables  $(T, M)$ .
- Survival data are incompletely observed.
- An observation is right-censored if the observation is terminated before the event occurs.
- An observation is left-truncated if the observation had the event before a certain time and the observation was omitted from the sample.
- Censoring is a property of the observation while truncation is a property of the sample.

Not all survival data contain censored observations, and censoring may occur in applications other than survival analysis. Nevertheless, because censored survival data are so common and because censoring requires special treatment, it is this topic more than anything else that unifies the many approaches to survival analysis.

RIGHT CENSORING

LEFT CENSORING

INTERVAL CENSORING

+ RANDOM

+ NOT RANDOM

INFORMATIVE & NOT INFORMATIVE CENSORING

Censoring comes in many forms and occurs for many different reasons. The most basic distinction is between ***left censoring*** and ***right censoring***. An observation on a variable  $T$  is right censored if all you know about  $T$  is that it is greater than some value  $c$ . In survival analysis,  $T$  is typically the time of occurrence for some event, and cases are right censored because observation is terminated before the event occurs. Thus, if  $T$  is a person's age at death (in years), you may know only that  $T > 50$ , in which case the person's death time is right censored at age 50. This notion of censoring is not restricted to event times. If you know only that a person's income is greater than \$75,000 per year, then income is right censored at \$75,000.

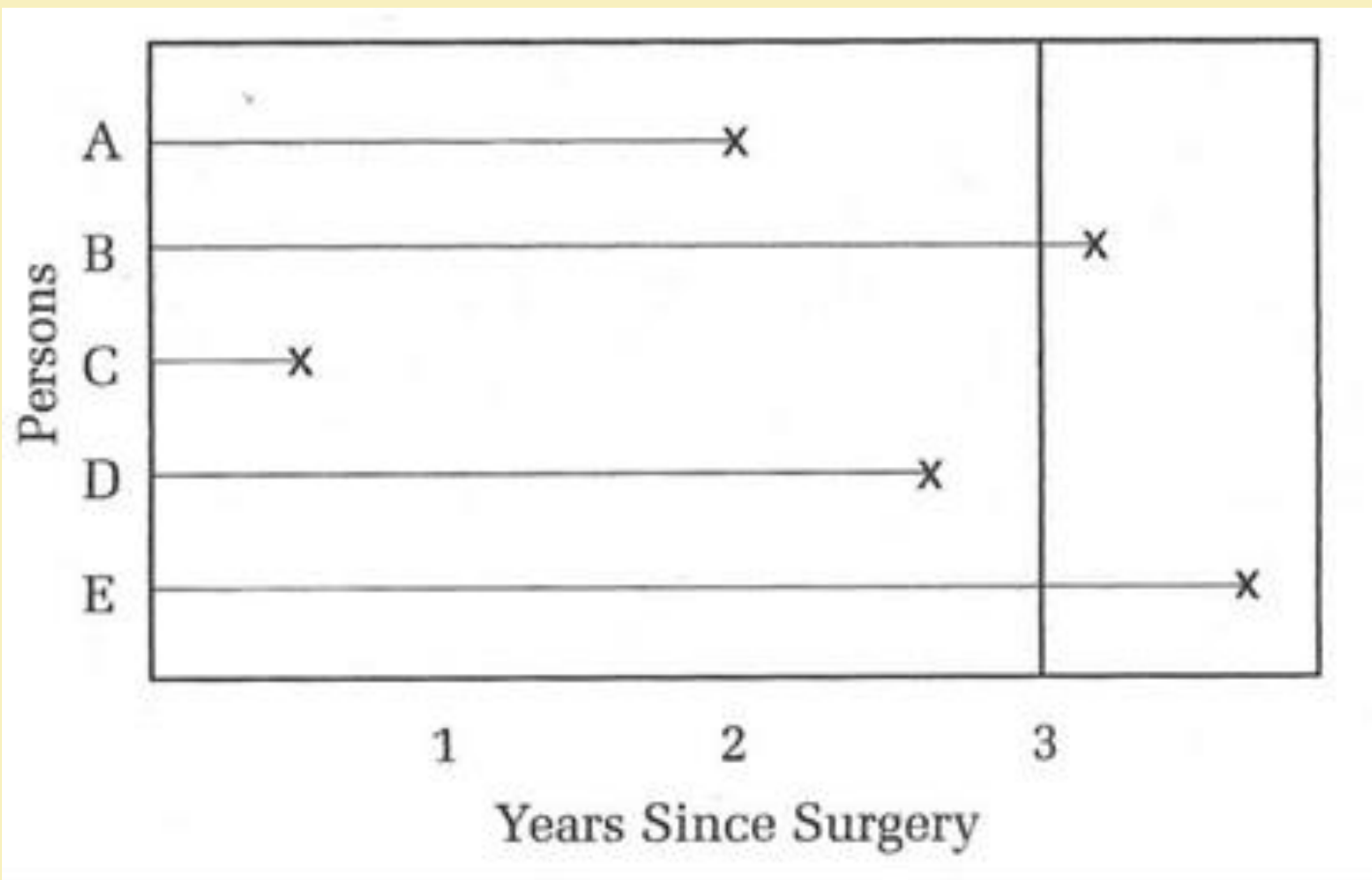
Symmetrically, **left censoring** occurs when all you know about an observation on a variable  $T$  is that it is *less* than some value.

Again, you can apply this notion to any sort of variable, not just an event time. In the context of survival data, left censoring is most likely to occur when you begin observing a sample at a time when some of the individuals may have already experienced the event. If you are studying menarche (the onset of menstruation), for example, and you begin following girls at age 12, you may find that some of them have already begun menstruating.

Unless you can obtain information on the starting date for those girls, the age of menarche is said to be left censored at age 12. (In the social sciences, *left censoring* often means something quite different.

Observations are said to be left censored if the *origin time*, not the event time, is known only to be less than some value. According to the definitions used here, such observations are actually right censored.)

# Figure *Singly Right-Censored Data*



**Type I** means that the censoring time is fixed (that is, under the control of the investigator), and *singly* refers to the fact that all the observations have the same censoring time. Even observations that are not censored are said to have a censoring time, in this case 3 years. It's just that their death times did not exceed their censoring time. Of course, censoring times can also vary across individuals. For example, you might want to combine data from two experiments, one with observation terminating after 3 years and another with observation terminating after 5 years. This is still Type I censoring, provided the censoring time is fixed by the design of the experiment.

**Type II** censoring occurs when observation is terminated after a prespecified number of events have occurred. Thus, a researcher running an experiment with 100 laboratory rats may decide that the experiment will stop when 50 of them have died. This sort of censoring is uncommon in the social sciences.

***Random censoring*** occurs when observations are terminated for reasons that are *not* under the control of the investigator. There are many possible reasons why this might happen. Suppose you are interested in divorces, so you follow a sample of couples for 10 years beginning with the marriage, and you record the timing of all divorces. Clearly, couples that are still married after 10 years are censored by a Type I mechanism. But for some couples, either the husband or the wife may die before the 10 years are up. Some couples may move out of state or to another country, and it may be impossible to contact them. Still other couples may refuse to participate in the study after, say, 5 years. These kinds of censoring are depicted in Figure 2.2, where the O for couples B and C indicates that observation is censored at that point in time. Regardless of the subject matter, nearly all prospective studies end up with some cases that didn't make it to the maximum observation time for one reason or another.



# Figure *Randomly Censored Data*



Random censoring can also be produced when there is a single termination time, but entry times vary *randomly* across individuals. Consider again the example in which people are followed from heart surgery until death. A more likely scenario is one in which people receive heart surgery at various points in time, but the study has to be terminated on a single date (say, December 31, 2010). All persons still alive on that date are considered censored, but their survival times from surgery will vary. This censoring is considered random because the entry times are typically not under the control of the investigator.

Standard methods of survival analysis do not distinguish among **Type I, Type II, and random censoring**. They are all treated as generic right-censored observations. Why make the distinctions, then? Well, if you have only **Type I or Type II censoring**, you're in good shape. The maximum likelihood and partial likelihood methods discussed in this book handle these types of censoring with no appreciable bias. Things are not so simple with random censoring, however. Standard methods require that random censoring be noninformative.

Unfortunately, there is no statistical test for **informative censoring versus non-informative censoring**. The best you can do is a kind of sensitivity analysis connected with the "Heterogeneity",

Repeated Events, and Other Topics." Aside from that, there are three important lessons here:

- 1.First, in designing and conducting studies, you should do everything possible to reduce the amount of random censoring. You can't rely on statistical methods to completely adjust for such censoring.
- 2.Second, you should make an effort to measure and include in the model any covariates that are likely to affect the rate of censoring.
- 3.Third, in studies with high levels of random censoring, you may want to place less confidence in your results than calculated confidence intervals indicate.

# 4. Stochastic Process Approach

# EVENT HISTORY ANALYSIS as a STOCHASTIC PROCESS

In event history analysis, stochastic process describing the following variables:

$T$  (time) - is a continuous variable and its realization (ie. the time of the event) are random,

$Y$  - determines the state space, which is finite.

Generally one can say that a stochastic process is described by the pair of variables  $\{(T, Y) = (T_k, Y_k) \text{ for } k = 1, 2, 3, \dots\}$ .

Additionally, you can take into account information:

$N$  - number of episodes,

$Y$  - vector of variables (the same for all the episodes or different)

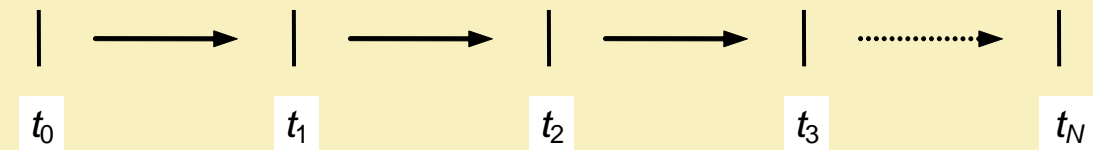
$d$  – censoring indicator (the value 0 or 1 and indicates whether the episode has been occurred or truncated).

**Staging (stochastic) process can be presented in the graphic form as:**

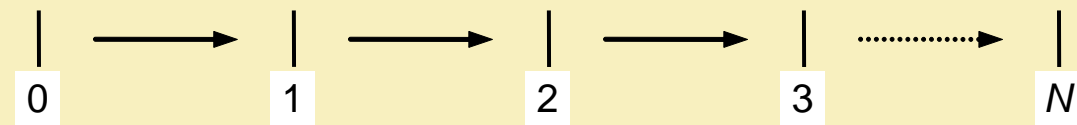
**Intensity**  
 $[r(t)]$



**Waiting time**  $[E(t)]$



**Number of event**  
 $[N]$



# Type of survival models :

**1. Non-parametric**

**2. Parametric**

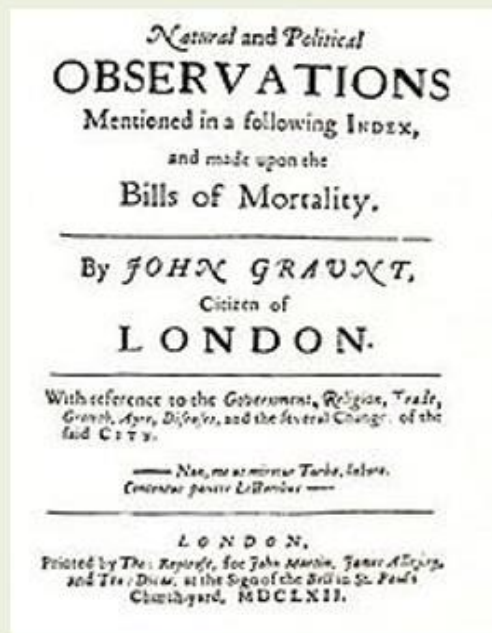
**3. Semi-parametric**



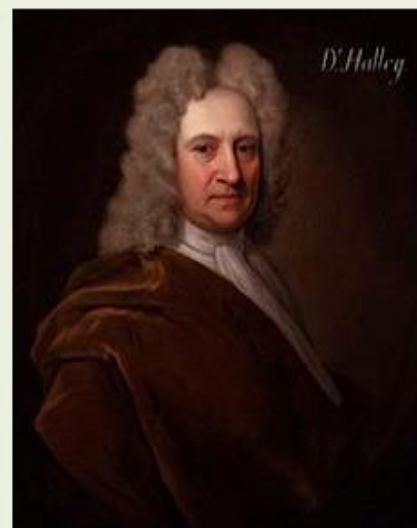
# ORIGIN STATE OF SURVIVAL ANALYSIS

## History of non-parametric models

In the mid-17th century, John Graunt, a merchant of mercy, started collecting and compiling in the form of a table the data contained in the Announcements on Mortality, which appeared at that time every week in London. These studies, probably created in collaboration with Sir William Petty, appeared in print in 1662 and contained a demographic summary of the causes of death in England and Wales



In 1693, the well-known astronomer Edmund Halley created the idea of survival tables in a form similar to the survival used in the analysis. His data was based on the register of births and deaths of the city of Breslau (currently Wrocław).



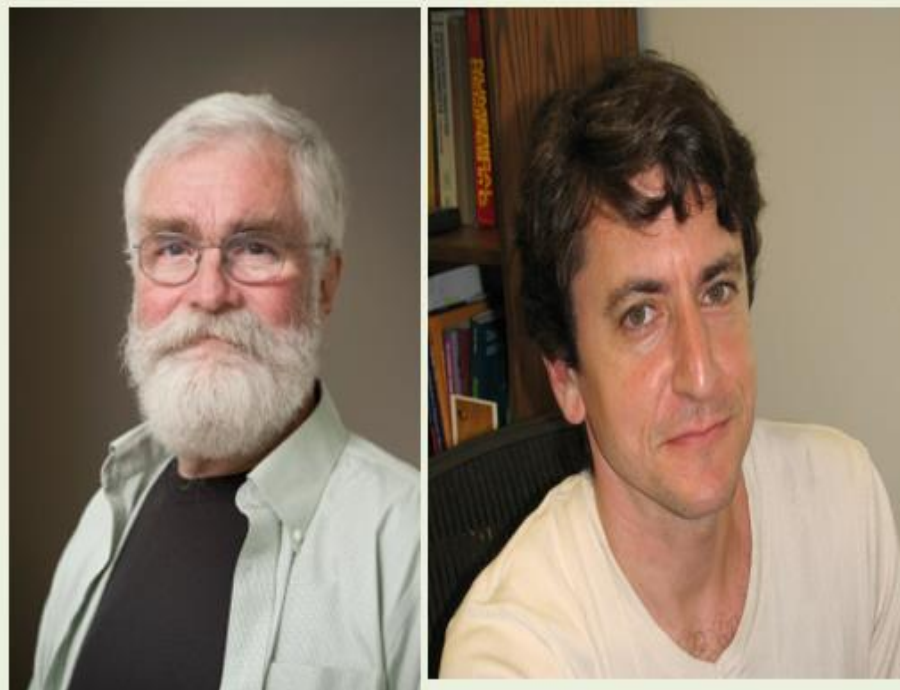
In 1760, Daniel Bernoulli used the Halley method to illustrate the effectiveness of smallpox vaccination. He calculated the increase in Halley's survival function if smallpox was eliminated as the cause of death. In this way, Bernoulli created a theory of competing risks. A summary of his work on this subject can be found in the work of David and Moeschberger (1978).





History has come full circle  
!!!!!!

What happened between  
1760 and 2018?  
We will briefly say !!!!



*From the 18th century, mathematics developed, statistics developed significantly, and prof. Ronald A. Fisher was commonly named FATHER OF MODERN STATISTICS. The great discoveries in the fields of biology and physics that took place in the nineteenth century led to a deterministic view of the world. However, with the continuous increase in precision of measurements, it became clear that there are also some unexplained, random facts affecting the events. In the twentieth century, partly due to the need to explain random variance, and partly due to the extraordinary development of computing power of computers, statistical theory has made significant progress. Based on Bernoulli's work, the theory of competing risks took root not only in actuarial sciences but also in medicine, social sciences, technical sciences, economics and many others.*

# 5. Describing Survival Distributions

## Parametric hazard functions

- Initially assume no explanatory variables: Model is for a single individual or a set of homogeneous individuals.
- Each type of hazard function defines a family of distributions for  $T$ , i.e., hazard functions are a way of characterizing a distribution.
- There are equivalences between hazard functions and other ways of describing distributions, thus:

$$h(t) = \frac{f(t)}{1 - F(t)} = \frac{f(t)}{S(t)}$$

$$f(t) = h(t) \exp\left(-\int_0^t h(u) du\right)$$

$$F(t) = 1 - \exp\left(-\int_0^t h(u) du\right)$$

Where  $f(t)$  is the probability density function and  $F(t)$  is the cumulative distribution function.

## **(a) constant hazard (exponential model)**

- Implies that  $T$  has an exponential distribution:  
Suppose  $h(t) = \lambda$  for all  $t$ , then density for  $T$  is  
$$f(t) = \lambda \exp(-\lambda t)$$
- Often useful as a baseline model, simplifies calculations

## EXPONENTIAL DISTRIBUTION – one more - the most popular and useful model in Survival Analysis

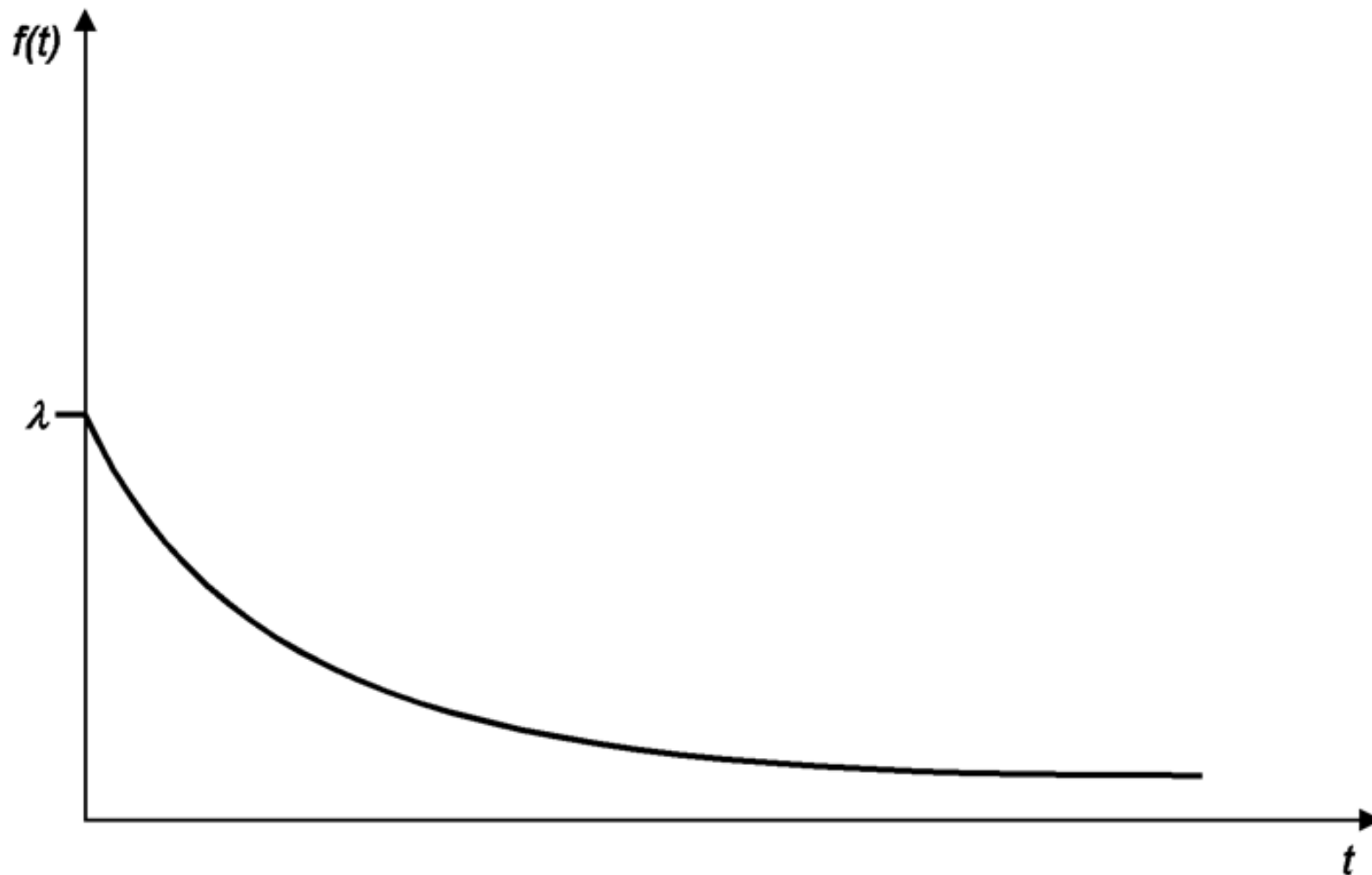
- **Basic functions :**  
density function  $f(t) = \alpha \exp(-\alpha t), \alpha > 0, t \geq 0$   
Survival function  $S(t) = G(t) = \exp(-\alpha t)$   
Hazard function  $\lambda(t) = r(t) = \alpha$
- **Basic parameters :**

**Expected value :**  $E(T) = \frac{1}{\alpha}$

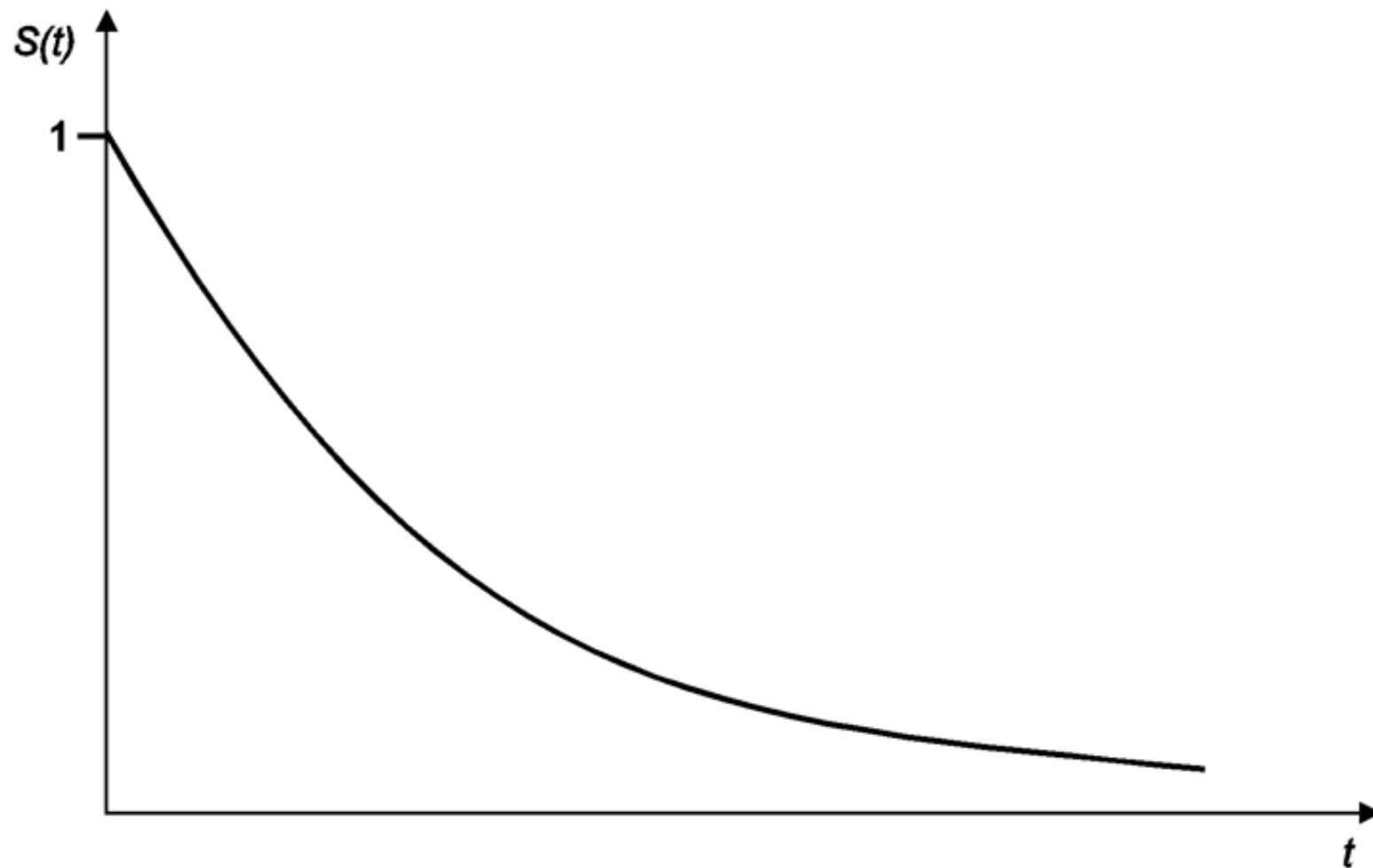
**Variance:**  $D^2(T) = \frac{1}{\alpha^2}$



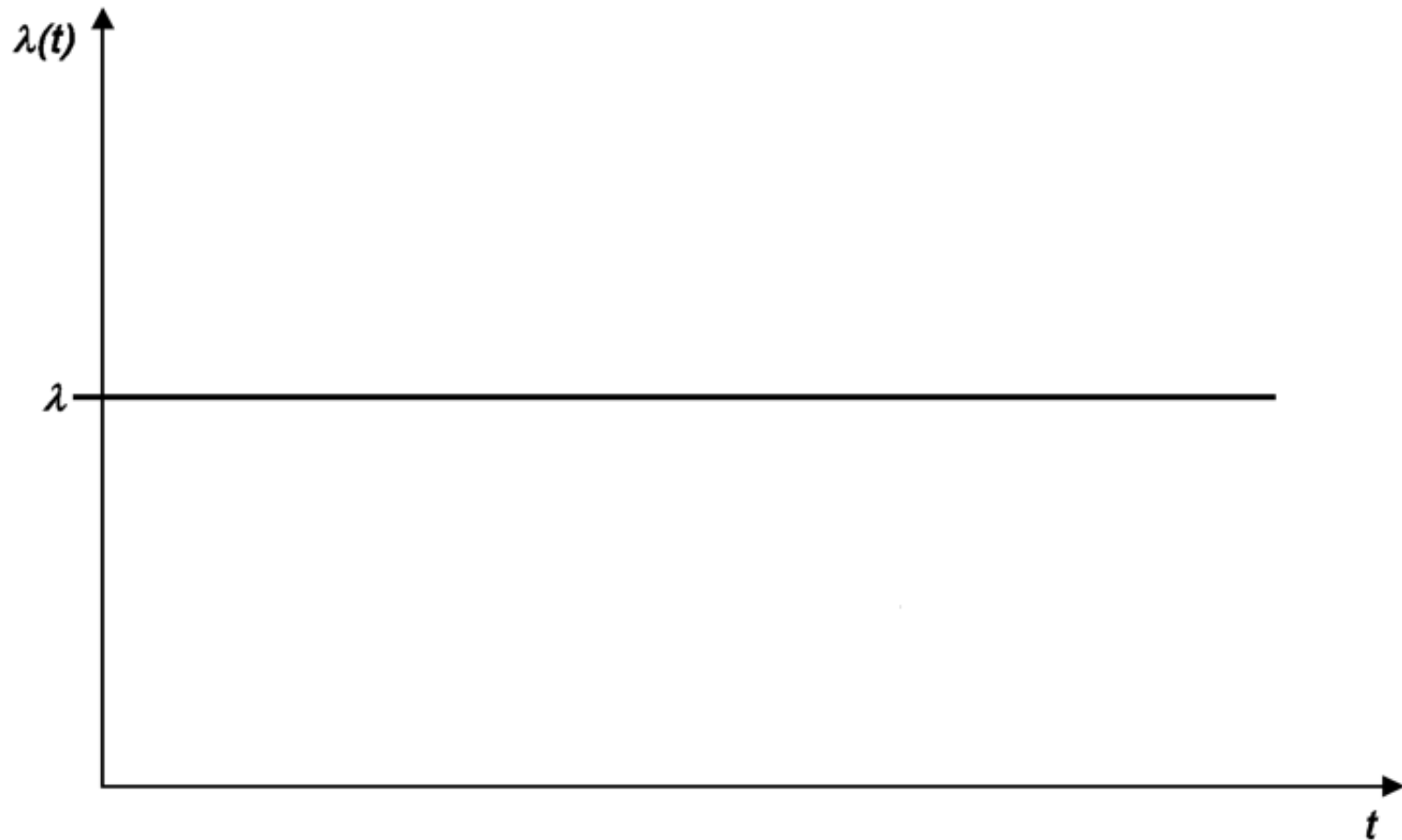
## Density function



## Survival function



## Hazard function

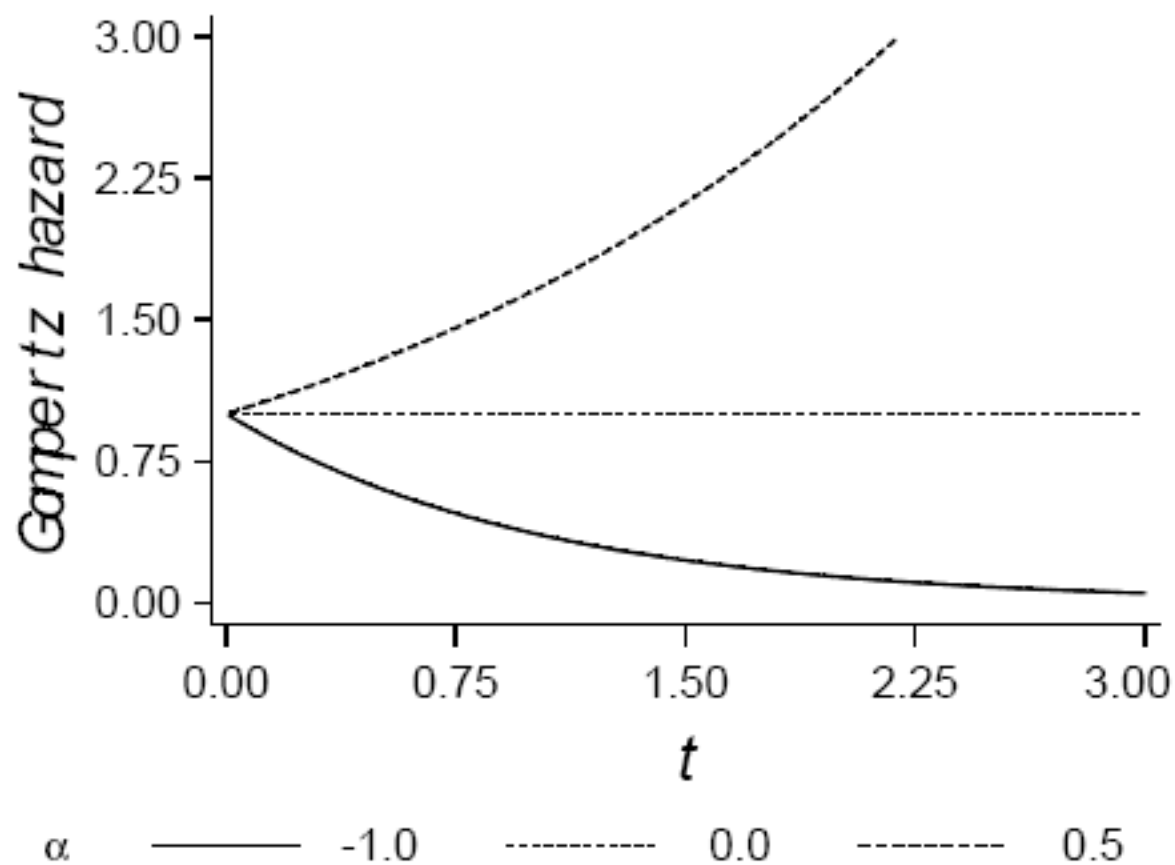


## (b) Gompertz model

$$\log h(t) = \mu + \alpha t$$

where  $\alpha$  can be positive or negative.

- Why log? Because  $h(t) \geq 0$ .
- Note: All logarithms to the base  $e = 2.71828....$
- Equivalently  
 $h(t) = \lambda_0 \exp\{\alpha t\}$  where  $\lambda_0 = e^\mu$ .
- Implies that  $T$  has a Gompertz distribution
- When  $t = 0$ ,  $h(t) = e^\mu$ .



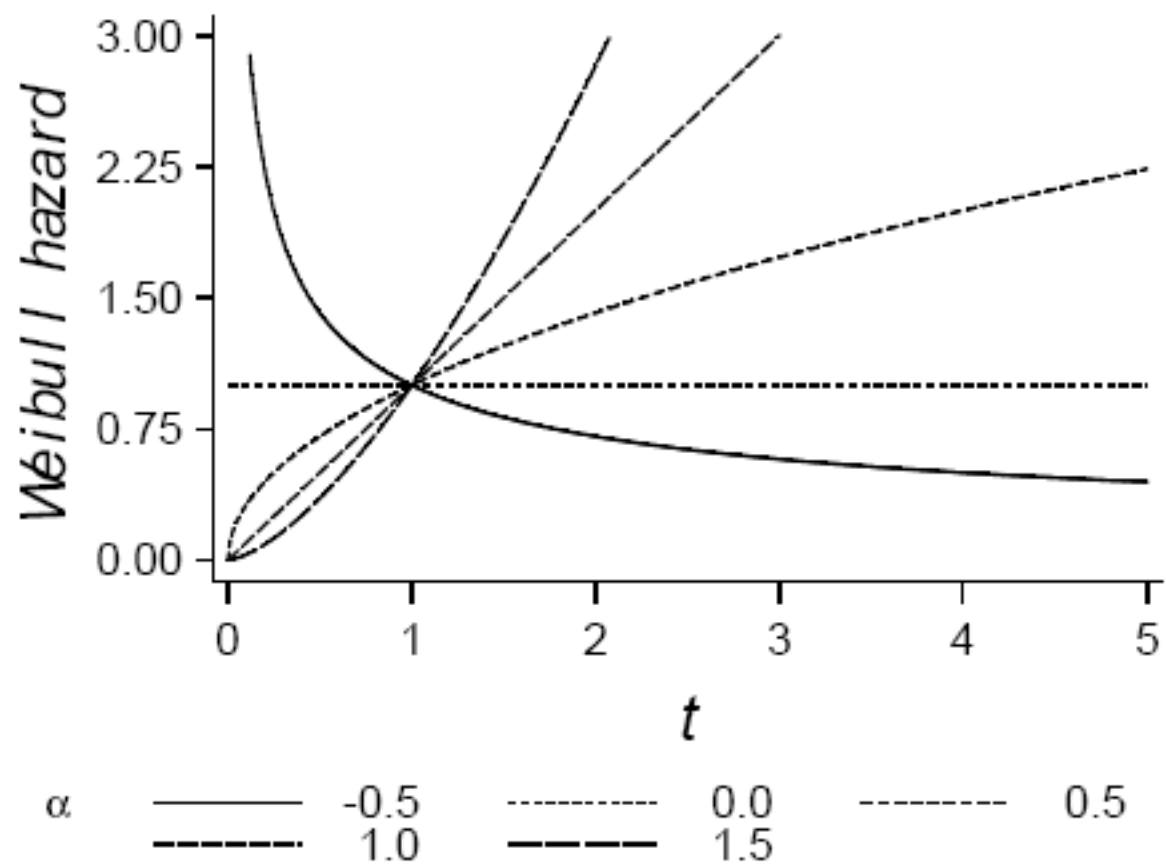
## (c) Weibull model

$\log h(t) = \mu + \alpha \log t$  where  $\alpha > -1$

Equivalently,

$$h(t) = \lambda_0 t^\alpha$$

- Implies that  $T$  has a Weibull distribution
- When  $t=0$ ,  $h(t) = 0$  or  $\infty$ .



## Inclusion of explanatory variables:

$\log h(t) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$  (exponential)

$\log h(t) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha t$  (Gompertz)

$\log h(t) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha \log t$  (Weibull)

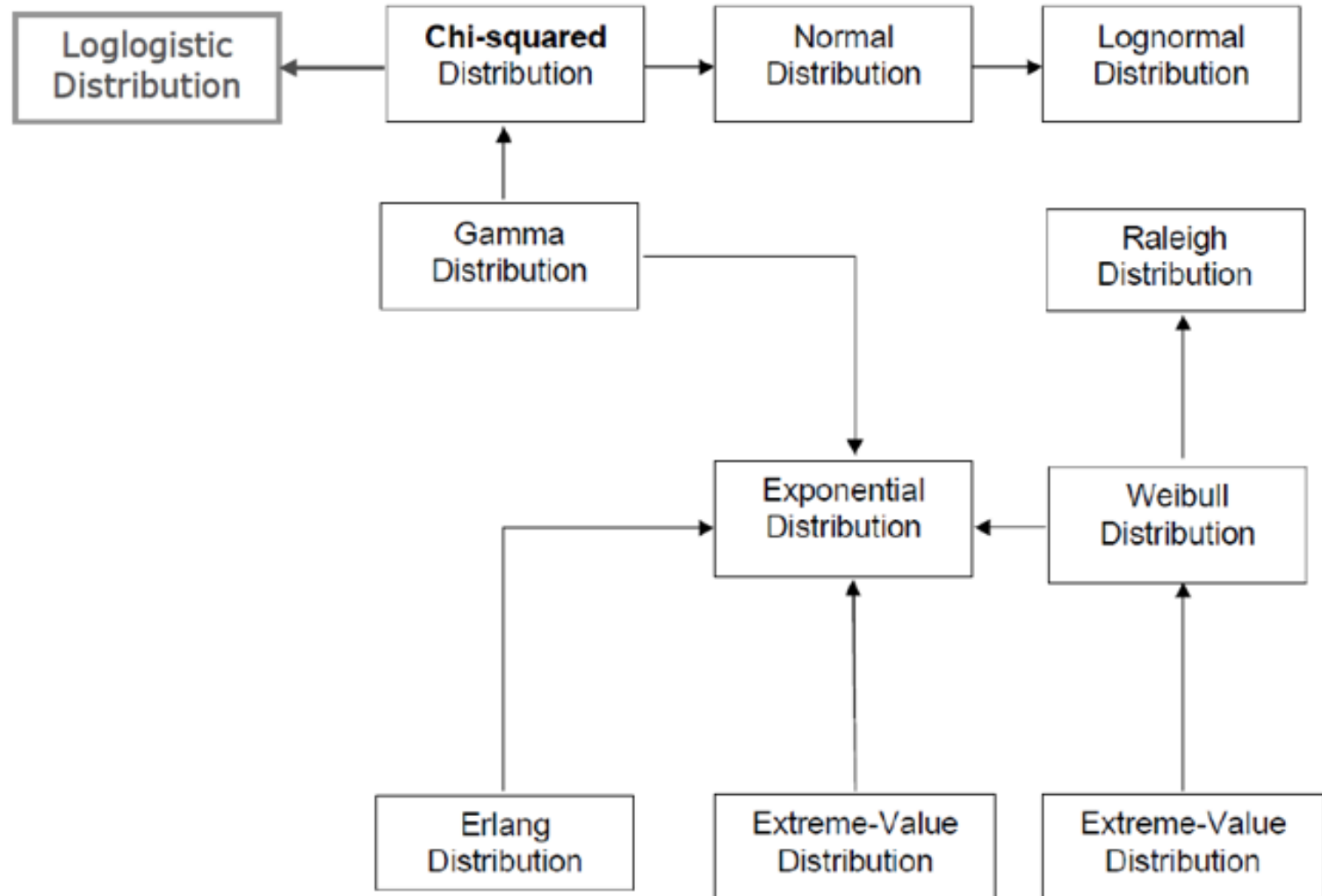
- All three models are members of a general class of models known as ***proportional hazards models***.
- Weibull (and exponential) is both a proportional hazards model and an ***accelerated failure-time model***.

**It is unique in this respect.**

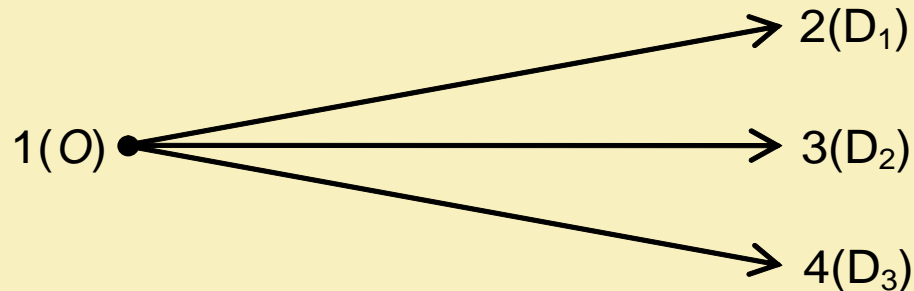
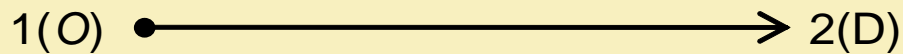
***Estimation: Maximum likelihood***



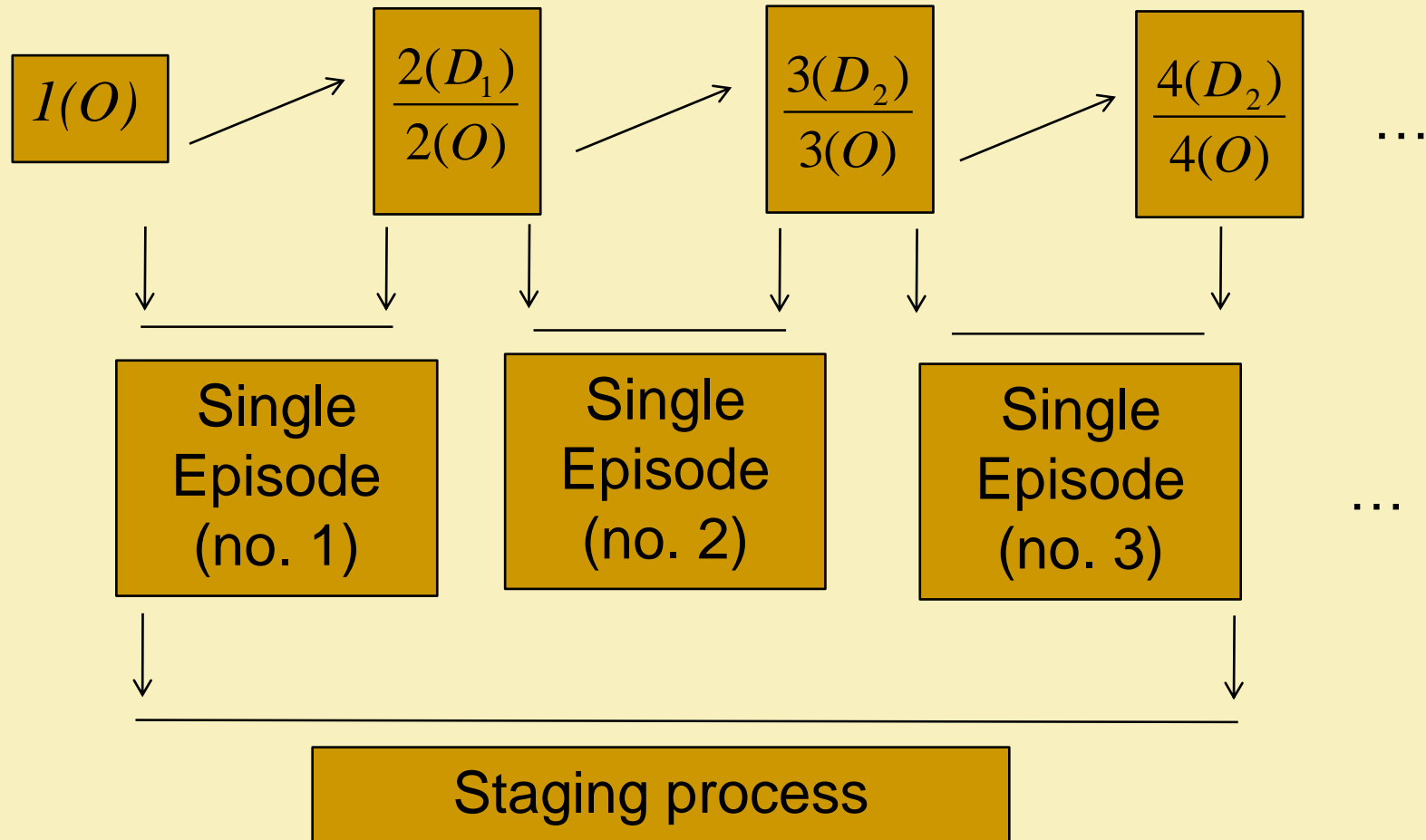
## Relations between different parametric models - arrows indicate the direction of transition from more general to special cases



**Single episode models: one origin state and one destination state or one origin state and many destinations:**



## *Multi episode models*



# Type of survival models :

1. Non-parametric

2. Parametric

3. Semi-parametric

# Nonparametric Estimation of Survivor Functions

Let  $S(t) = \Pr(T > t)$  where  $T$  is the time of the event, i.e., the probability that an individual "survives" to time  $t$ .

Recall that  $S(t) = 1 - F(t)$  where  $F(t)$  is the c.d.f.  $0 \leq S(t) \leq 1$ , a non-increasing function of  $t$ .

Every distribution for  $T$  has a corresponding survival curve.

When there is no censoring, it's easy to estimate the survivor function.

For any  $t$ , simply estimate  $S(t)$  by the proportion of cases that have survived past that point in time.

# Nonparametric Estimation of Survivor Functions

If all event times are less than all censored times, the survivor function can again be estimated by the proportion surviving, for all times up to the lowest censored time.

After that the estimate is undefined (clearly it must be between the last value and 0).

If some censored times are less than some event times, more complex methods are needed.

**The standard methods are the life table method and the Kaplan Meier method & .**

# 6.1. Life Table Method (ACTUARIAL METHOD)

This quite commonly used method has the following limitations:

- There is a need for grouping of observation time intervals of equal length (the definition of class intervals is arbitrary)
- Requires a relatively large number of episodes that obtained estimators for discrete time intervals were unloaded.
- If the number of episodes is sufficiently large, then the method gives good results and is easy to use.
- Otherwise, the results are subject to biased estimation.



Principle of the method is that the time axis (observation) is divided into interval of equal class in most (but not necessarily) range, namely:

$$0 \leq \tau_1 < \tau_2 < \tau_3 < \dots < \tau_q ; \quad \tau_{q+1} = \infty$$

In this way, q is total class intervals, and each is limited in left side:

$$I_L = \{t \mid \tau_l \leq t < \tau_{l+1}\} \quad l = 1, 2, 3, \dots, q$$

## **Estimation of parameters of life tables for a single episode model (of single transition).**

For each class interval  $I_L$ ,  $L = 1, 2, 3, \dots, q$  must be determined:

$E_l$  - the number of episodes ending events (number of events) in the range of  $I_L$

$Z_l$  - the number of episodes were cut off, ending in the range of  $I_L$

$R_l$  - a population exposed to the risk of events in the range of experiments  $I_L$  (a risk set).

Assuming that within the time frame may be truncated, it is a two-step procedure:

First - the number of episodes of  $N_l$ , the number of the inputs (included) to the  $l$ -th interval is defined as:

$$N_1 = N, N_l = N_{l-1} - E_{l-1} - Z_{l-1}$$

Second - it must be determined as the large number of episodes is cut off inside the interval, and to what extent they can be included in the set of risks (ie the population exposed to the risk experience of the event).

**A standard assumption is that, 1 / 2, although this assumption is sometimes arbitrary, particularly where the distribution of events is not uniform.**

**Next, one possible objective is to adopt a fixed  $w$ , where  $w$  is between 0 and 1, for functions of censored episodes that have come to the set of risks.**

The number of elements of the set of risk for that interval is then calculated as:

$$R_l = N_l - wZ_l$$

Assuming such signs is relatively easy to identify a set of indicators used in the simple life table, the conditional probability of an event in the  $l$ -th interval ( $q_l$ ), the function of survival ( $p_l$ ), etc.

$$q_l = \frac{E_l}{R_l}, \quad p_l = 1 - q_l$$

Survival function for the  $l$ -th interval is determined by the formula

$$S_1 = 1, \quad S_l = p_{l-1} S_{l-1}$$

After obtaining these results estimate the density function estymuje to the middle class interval by the formula

$$f_l = \frac{S_l - S_{l+1}}{\tau_{l+1} - \tau_l} \quad l = 1, 2, 3, \dots, q - 1$$

Of course, the last interval is the interval of an open upward, and it is not possible to determine precisely the function of survival.

Next estimation includes:

- Standard deviation of survival function

$$BS(S_l) = S_l \sqrt{\sum_{i=1}^{l-1} \frac{q_i}{p_i R_i}}$$

- Standard deviation of density function

$$BS(f_l) = \frac{q_l S_l}{\tau_{l+1} - \tau_l} \sqrt{\sum_{i=1}^{l-1} \frac{q_i}{p_i R_i} + \frac{p_l}{q_l R_l}}$$

- Standard deviation of hazard rate

$$BS(r_l) = \frac{r_l}{\sqrt{q_l R_l}} \sqrt{1 - \left[ \frac{r_l (\tau_{l+1} - \tau_l)}{2} \right]^2}$$

# An Example of Life Table Estimation

## Life Table (Actuarial) Method

Most useful with large number of observations, data grouped into intervals.

*Example: 65 myeloma patients*

DUR: time (in months) from diagnosis to death or censoring.

NIT: log of blood urea nitrogen at diagnosis

HEMO: hemoglobin at diagnosis

AGE: age at diagnosis

SEX: 0=male, 1=female

CALC: blood calcium at diagnosis

DEAD: 1=uncensored, 0 = censored.

17 cases are censored.



17 cases are censored.

OBS	DUR	NIT	HEMO	AGE	SEX	CALC	DEAD
1	1	2.218	9.4	67	0	10	1
2	1	1.940	12.0	38	0	18	1
3	2	1.519	9.8	81	0	15	1
4	2	1.748	11.3	75	0	12	1
5	2	1.301	5.1	57	0	9	1
6	3	1.544	6.7	46	1	10	1
7	5	2.236	10.1	50	1	9	1
8	5	1.681	6.5	74	0	9	1
9	6	1.362	9.0	77	0	8	1
10	6	2.114	10.2	70	1	8	1
11	6	1.114	9.7	60	0	10	1
12	6	1.415	10.4	67	1	8	1
13	7	1.978	9.5	48	0	10	1
14	7	1.041	5.1	61	1	10	1
61	28	1.230	7.3	82	1	9	0
62	41	1.756	12.8	72	0	9	0
63	53	1.114	12.0	66	0	11	0
64	57	1.255	12.5	66	0	11	0
65	77	1.079	14.0	60	0	12	0

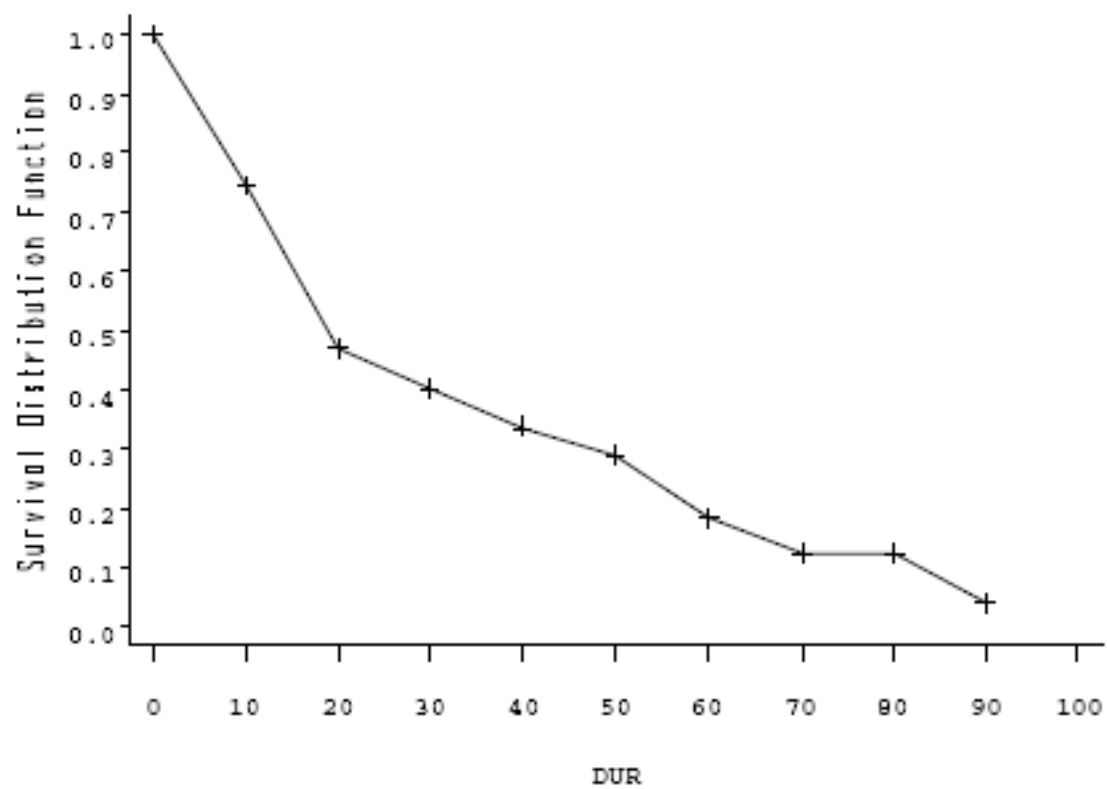
### Life Table Survival Estimates

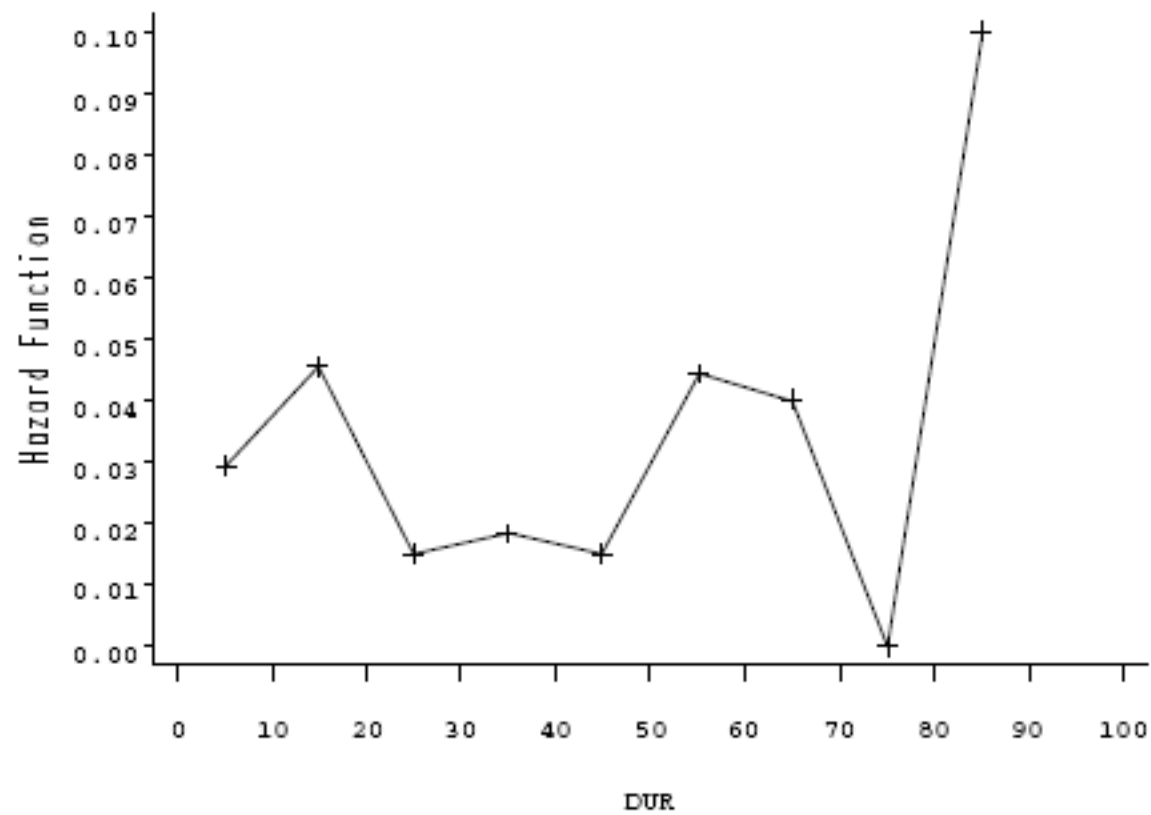
Interval [Lower, Upper)		Number Failed	Number Censored	Effective Sample Size	Conditional Probability of Failure
0	10	16	5	62.5	0.2560
10	20	15	7	40.5	0.3704
20	30	3	1	21.5	0.1395
30	40	3	0	18.0	0.1667
40	50	2	1	14.5	0.1379
50	60	4	2	11.0	0.3636
60	70	2	0	6.0	0.3333
70	80	0	1	3.5	0
80	90	2	0	3.0	0.6667
90	.	1	0	1.0	1.0000

Interval [Lower, Upper)		Conditional Probability Standard		Survival Standard	
		Error	Survival	Failure	Error
0	10	0.0552	1.0000	0	0
10	20	0.0759	0.7440	0.2560	0.0552
20	30	0.0747	0.4684	0.5316	0.0663
30	40	0.0878	0.4031	0.5969	0.0669
40	50	0.0906	0.3359	0.6641	0.0661
50	60	0.1450	0.2896	0.7104	0.0646
60	70	0.1925	0.1843	0.8157	0.0588
70	80	0	0.1228	0.8772	0.0528
80	90	0.2722	0.1228	0.8772	0.0528
90	.	0	0.0409	0.9591	0.0378

Interval [Lower, Upper)		Median Residual Lifetime	Median Standard Error
0	10	18.8548	2.2952
10	20	24.6264	8.7011
20	30	35.2562	4.7972
30	40	28.3600	4.5113
40	50	22.6571	7.1806
50	60	16.4286	7.1071
60	70	23.7500	4.5928
70	80	17.5000	4.0089
80	90	7.5000	4.3301
90	.	.	.

Evaluated at the Midpoint of the Interval					
Interval			PDF		Hazard
[Lower,Upper)		PDF	Standard Error	Hazard	Standard Error
0	10	0.0256	0.00552	0.029358	0.00726
10	20	0.0276	0.00600	0.045455	0.011429
20	30	0.00654	0.00362	0.015	0.008636
30	40	0.00672	0.00371	0.018182	0.010454
40	50	0.00463	0.00318	0.014815	0.010447
50	60	0.0105	0.00481	0.044444	0.021667
60	70	0.00614	0.00405	0.04	0.027713
70	80	0	.	0	.
80	90	0.00819	0.00486	0.1	0.061237
90	.	.	.	.	.





## 6.2. Kaplan - Meier Method



Estimator of survival functions such as "Product-Limit-Estimation (PLE) is calculated

$$\hat{S}(t) = \prod_{l: \tau_l < t} \left( 1 - \frac{E_l}{R_l} \right)$$

Standard error of estimator is defined as follows:

$$BS(\hat{S}(t)) = \hat{S}(t) \sqrt{\sum_{l: \tau_l < t} \frac{E_l}{R_l(R_l - E_l)}}$$

PLE method allows estimation of the cumulative hazard function

$$\hat{H}(t) = -\log(\hat{S}(t))$$

## Estimating confidence intervals for survival function

Estimating confidence intervals for survival function is based on the assumption that the statistic as

$$\frac{S(t) - \hat{S}(t)}{BS(\hat{S}(t))}$$

has asymptotic normal distribution. Under this assumption, the confidence interval for survival function is of the form:

$$P\left\{\hat{S}(t) - u_{\alpha} BS(\hat{S}(t)) < S(t) < \hat{S}(t) + u_{\alpha} BS(\hat{S}(t))\right\} = 1 - \alpha$$

# An example of PL Life Table Estimation

Product-Limit Survival Estimates

dur	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.0000	1.0000	0	0	0	65
1.0000	.	.	.	1	64
1.0000	0.9692	0.0308	0.0214	2	63
2.0000	.	.	.	3	62
2.0000	.	.	.	4	61
2.0000	0.9231	0.0769	0.0331	5	60
3.0000	0.9077	0.0923	0.0359	6	59
4.0000*	.	.	.	6	58
4.0000*	.	.	.	6	57
5.0000	.	.	.	7	56
5.0000	0.8758	0.1242	0.0411	8	55
6.0000	.	.	.	9	54
6.0000	.	.	.	10	53
6.0000	.	.	.	11	52
6.0000	0.8121	0.1879	0.0489	12	51

## Summary Statistics for Time Variable dur

### Quartile Estimates

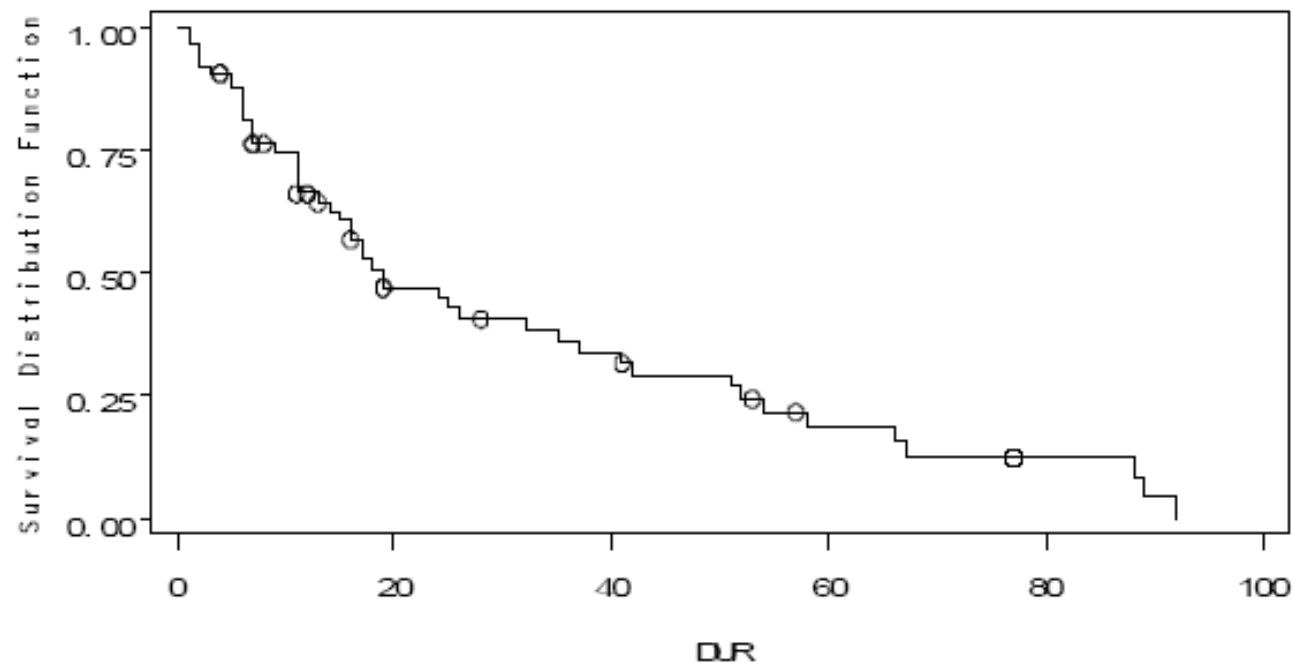
Percent	Point Estimate	95% Confidence Interval		
		Transform	[Lower	Upper)
75	52.0000	LOGLOG	35.0000	67.0000
50	19.0000	LOGLOG	14.0000	35.0000
25	9.0000	LOGLOG	6.0000	13.0000

Mean	Standard Error
------	----------------

32.1092	4.0270
---------	--------

Summary of the Number of Censored and Uncensored Values

Total	Failed	Censored	Percent Censored
65	48	17	26.15



Legend: — Product-Limit Estimate Curve  
 O O O Censored Observations

# Survival Models

- Models in survival analysis are written in terms of the hazard function.
- They assess the relationship of predictor variables to survival time.
- They can be non-parametric, parametric or semi-parametric models.

## 6.3. Nelson – Aalen Estimator

**Nelson-Aalen estimator:**

$$\hat{\Lambda}_{NA}(t) = \sum_{j: \tau_j \leq t} d_j / r_j.$$

Once we have  $\hat{\Lambda}_{NA}(t)$ , we can obtain the **Fleming-Harrington estimator** of  $S(t)$ :

$$\hat{S}_{FH}(t) = \exp(-\hat{\Lambda}_{NA}(t)).$$

## 6.3. Nelson – Aalen Estimator & Kaplan –Meier Estimator

$$\hat{S}_{FH}(t) = \exp(-\hat{\Lambda}_{NA}(t)).$$

The Kaplan-Meier estimator of the survivorship function (or survival probability)  $S(t) = P(T > t)$  is:

$$\begin{aligned}\hat{S}(t) &= \prod_{j:\tau_j \leq t} \frac{r_j - d_j}{r_j} \\ &= \prod_{j:\tau_j \leq t} \left(1 - \frac{d_j}{r_j}\right)\end{aligned}$$



## **7. Semiparametric model – COX MODEL**

**1972; 1975**

# Parametric versus Semi-Parametric Models

Parametric models require the following:

- that the distribution of survival time is known
- that the hazard function is completely specified except for the values of the unknown parameters

Examples include the Weibull model, the exponential model, and the log-normal model.

# Parametric versus Semi-Parametric Models


Properties of semi-parametric models are the following:

- the distribution of survival time is unknown
- the hazard function is unspecified


An example is the Cox proportional hazards model.

# Cox Proportional Hazards Model

$$h_i(t) = h_0(t) e^{\{\beta_1 X_{i1} + \dots + \beta_k X_{ik}\}}$$



Baseline Hazard function –  
involves time but not  
predictor variables



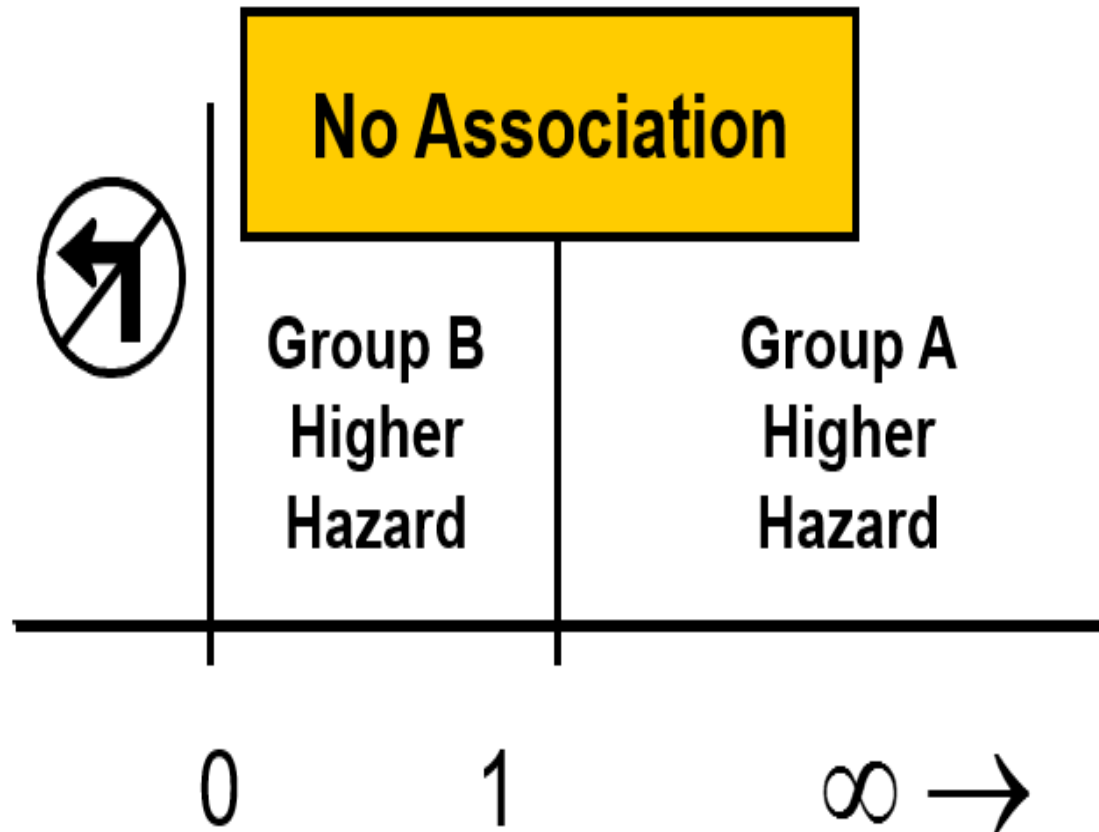
Linear function of a set  
of predictor variables –  
does **not** involve time

$$\ln h_i(t) = \ln h_0(t) + \beta_1 X_{i1} + \dots + \beta_k X_{ik}$$

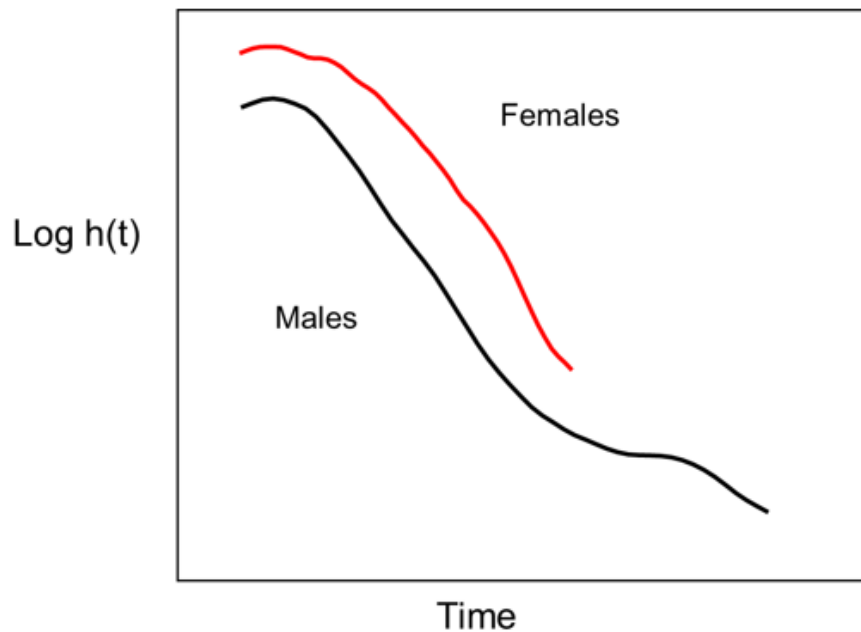
# Popularity of the Cox Model

- The Cox proportional hazards model provides the primary information desired from a survival analysis, hazard ratios, and adjusted survival curves, with a minimum number of assumptions.
- It is a robust model where the regression coefficients closely approximate the results from the correct parametric model.

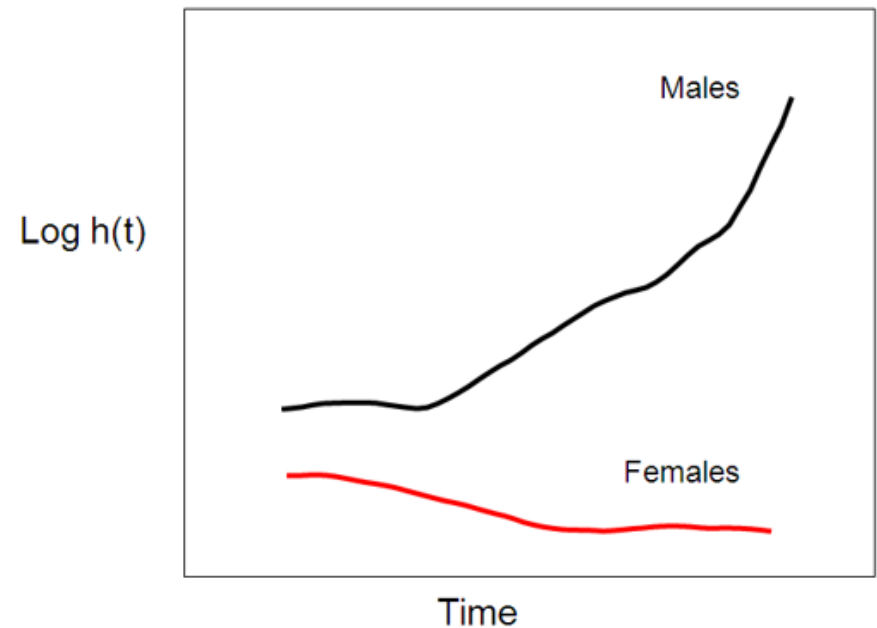
# Properties of the Hazard Ratio



# Proportional Hazards Assumption



Proportional Hazards



Non-Proportional Hazards

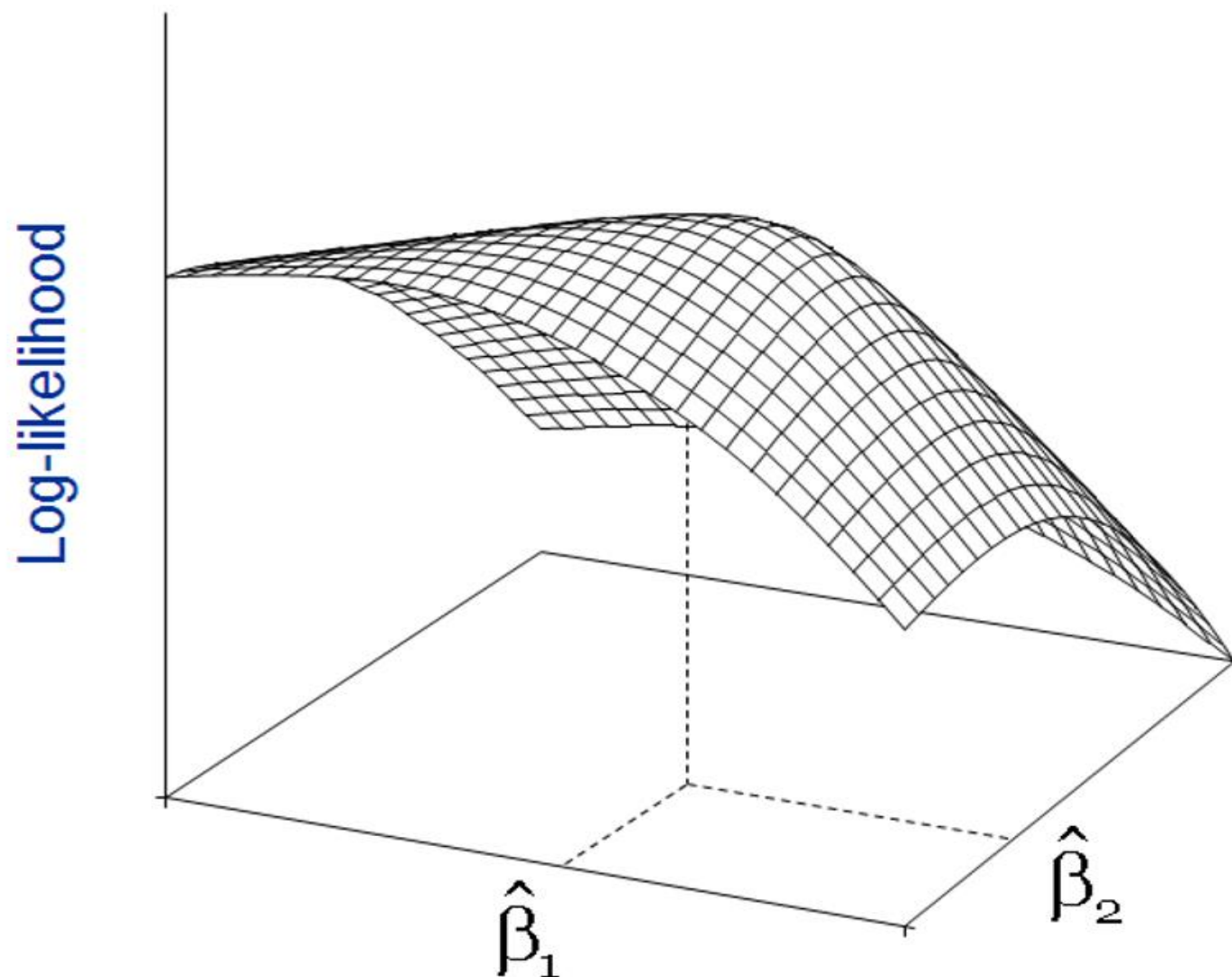
## **Types of COX – semiparametric models:**

1. Proportional hazard model
2. Non- proportional hazard model
3. Stratified model

## **METHOD OF ESTIMATION – PARTIAL LIKELIHOOD (PLM)**



# Maximum Partial Likelihood



# Partial Likelihood

Partial likelihood differs from maximum likelihood because of the following:

- it does not use the likelihoods for all subjects
- it only considers likelihoods for subjects that experience the event
- it considers subjects as part of the risk set until they are censored

## Regression Models and Life-Tables



D. R. Cox

*Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 34, No. 2  
(1972), 187-220.

## Partial Likelihood

D. R. Cox

*Biometrika*, Vol. 62, No. 2 (Aug., 1975), 269-276.

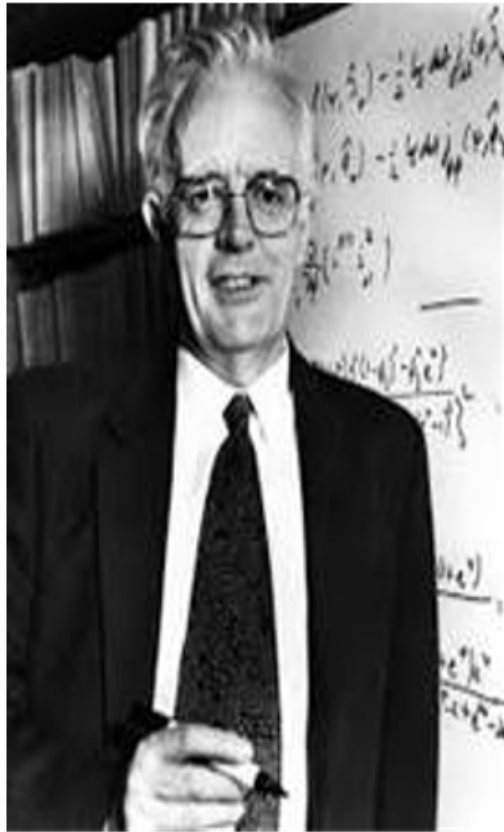
**October 20, 2016, -Special Edition- News from the World of Statistics**

**Inaugural International Prize in Statistics Awarded**

<https://www.youtube.com/watch?v=J9aiBi58uik&feature=youtu.be>



**International Prize in Statistics Awarded to Sir David Cox for Survival Analysis Model Applied in Medicine, Science, and Engineering**



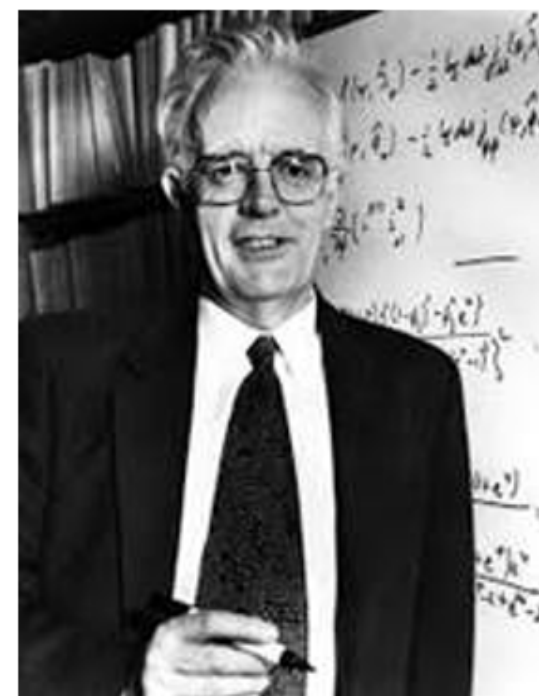
Prominent British statistician Sir David Cox has been named the inaugural recipient of the International Prize in Statistics. Susan Ellenberg, chair of the International Prize in Statistics Foundation, made the historic announcement [via video statement](https://www.youtube.com/watch?v=xwhEcXaWkh0).

<https://www.youtube.com/watch?v=xwhEcXaWkh0>

A giant in the field of statistics, Sir David Cox is being recognized by the International Prize in Statistics Foundation specifically for his pioneering 1972 paper in which he developed the proportional hazards model that today bears his name.

The Cox Model is widely used in the analysis of survival data and enables researchers to more easily identify the risks of specific factors for mortality or other survival outcomes among groups of patients with very different characteristics. From disease risk assessment and treatment evaluation to product liability to school dropout, re-incarceration, and AIDS surveillance systems, the Cox Model has been applied essentially in all fields of science, as well as in engineering, that involve discovering and understanding natural or human-induced risk factors on survival.

**His 50-year career included technical and research positions in the private and nonprofit sectors as well as numerous academic appointments as professor or department chair at Birkbeck College, Imperial College of London, Nuffield College and Oxford University. He obtained his PhD from the University of Leeds in 1949, and prior to that studied mathematics at St. Johns College. Though he retired in 1994, Sir David Cox remains active in the profession in Oxford, England.**





# Thank you for your attention !!

