Applied Statistical Modeling 2

Lorenzoni Valentina
valentina.lorenzoni@santannapisa.it

Institute of Management - Scuola Superiore Sant' Anna, Pisa

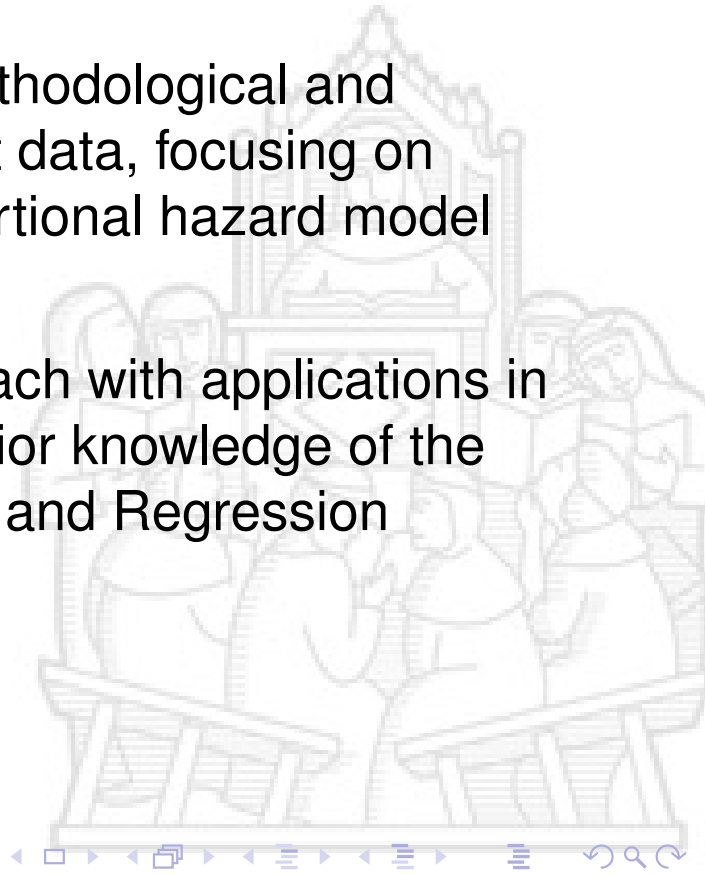
16th April 2025

Course description and practical information

Overview

The course aims to provide students with a methodological and applied background of models for time-to-event data, focusing on survival analysis and specifically on Cox proportional hazard model and on models for competing risk

The course provides a practice-oriented approach with applications in the context of social sciences, and assumes prior knowledge of the foundations of Probability, Inferential Statistics, and Regression models

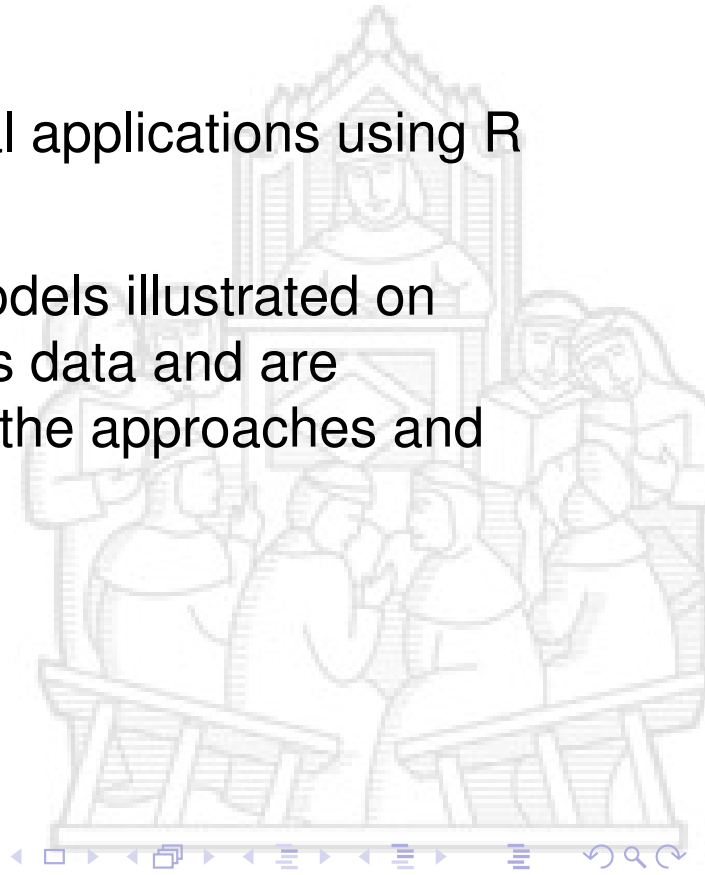


Course description and practical information

Classes

The course foresees both lectures and practical applications using R software

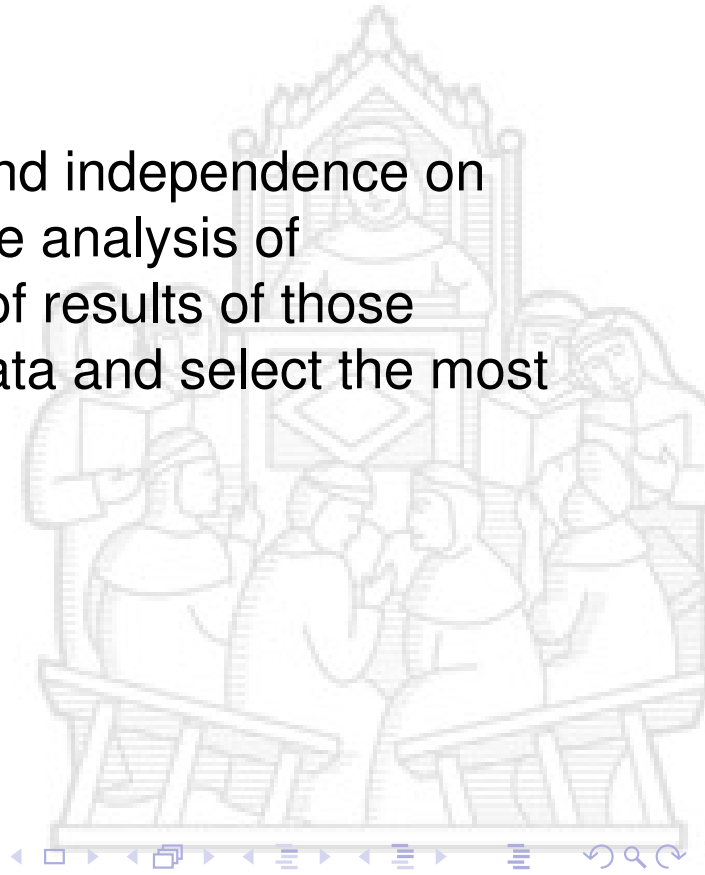
Practical applications deal with examples of models illustrated on some open source database on social sciences data and are intended to let students acquire familiarity with the approaches and the use of R software to perform analysis



Course description and practical information

Goals

Students are expected to acquire knowledge and independence on the use of the most common approaches for the analysis of time-to-event data and on the comprehension of results of those analyses being able to critically comment on data and select the most appropriate approaches for the analysis



Course description and practical information

Final evaluation

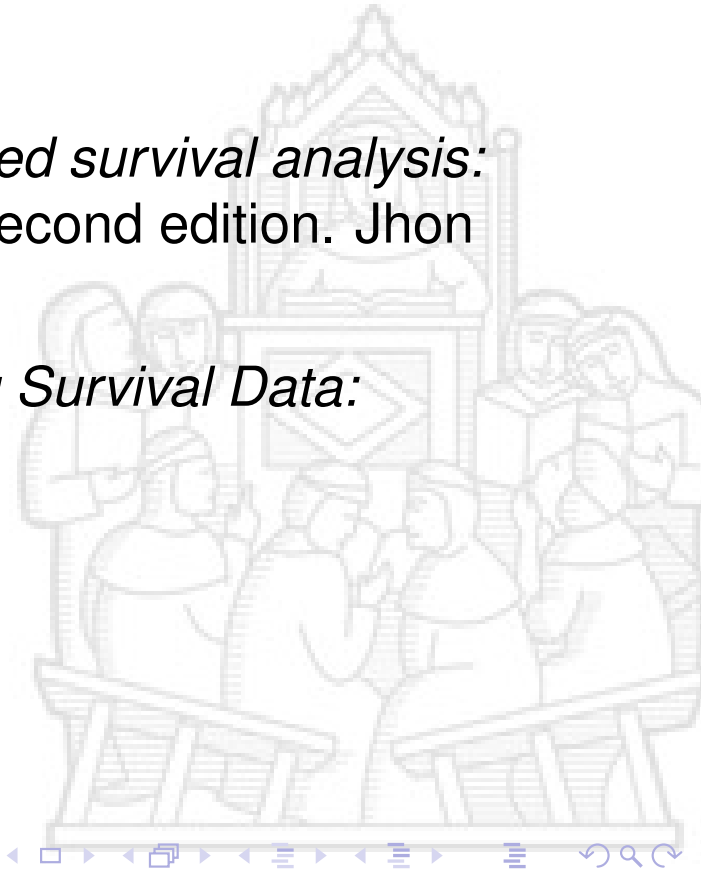
The final evaluation consists of a presentation discussing an assignment related to the analysis of an example dataset

Course description and practical information

Textbooks

Hosmer DW, Lemeshow S, May S (2008).*Applied survival analysis: Regression Modeling of Time-to-Event Data.* Second edition. Jhon Wiley and Sons, Inc., Hoboken, New Jersey

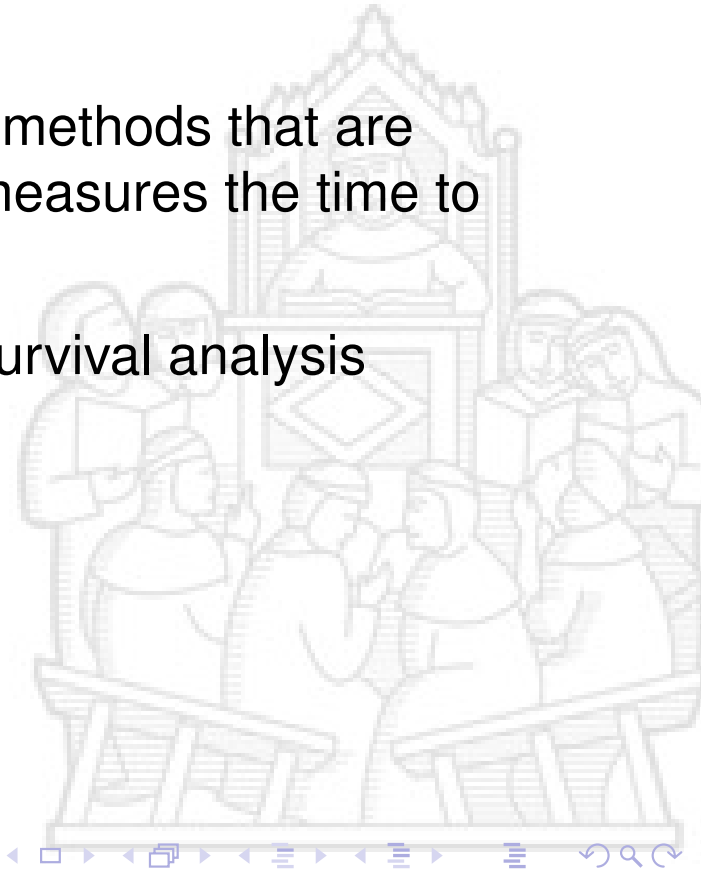
Therneau TM, Grambsch PM (2000). *Modeling Survival Data: Extending the Cox Model.* Springer



Introduction

Survival analysis deals with all those statistical methods that are suited for modeling a dependent variable that measures the time to the occurrence of an event of interest

Time to an event is the outcome of interest in survival analysis



What are time-to-event data?

Consider some possible events of interest:

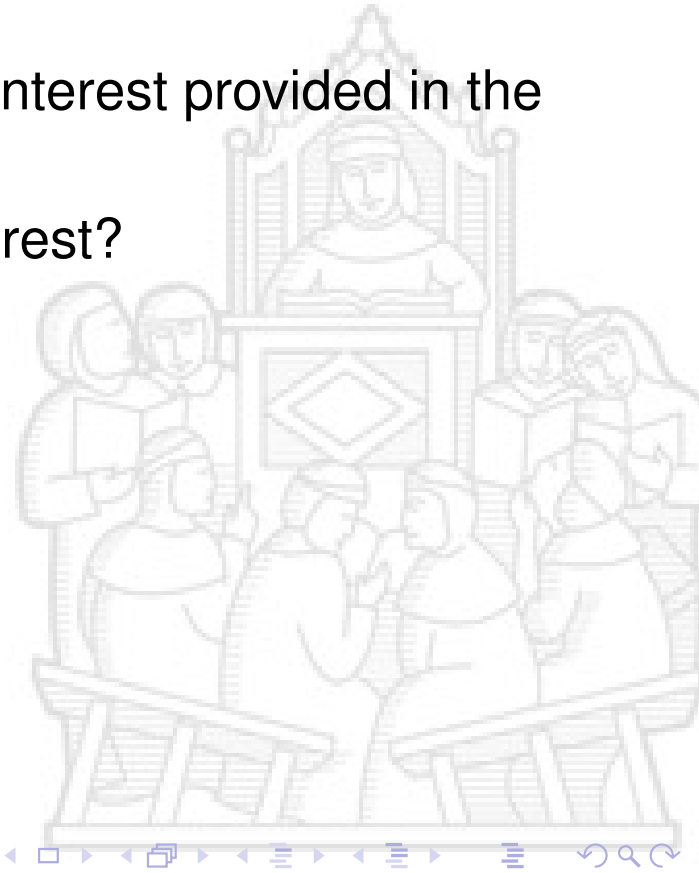
- First occupation
- First child
- Business failure

that could be evaluated through a longitudinal study observing a population of interest over time

What are time-to-event data?

Considering the example of possible events of interest provided in the previous slide

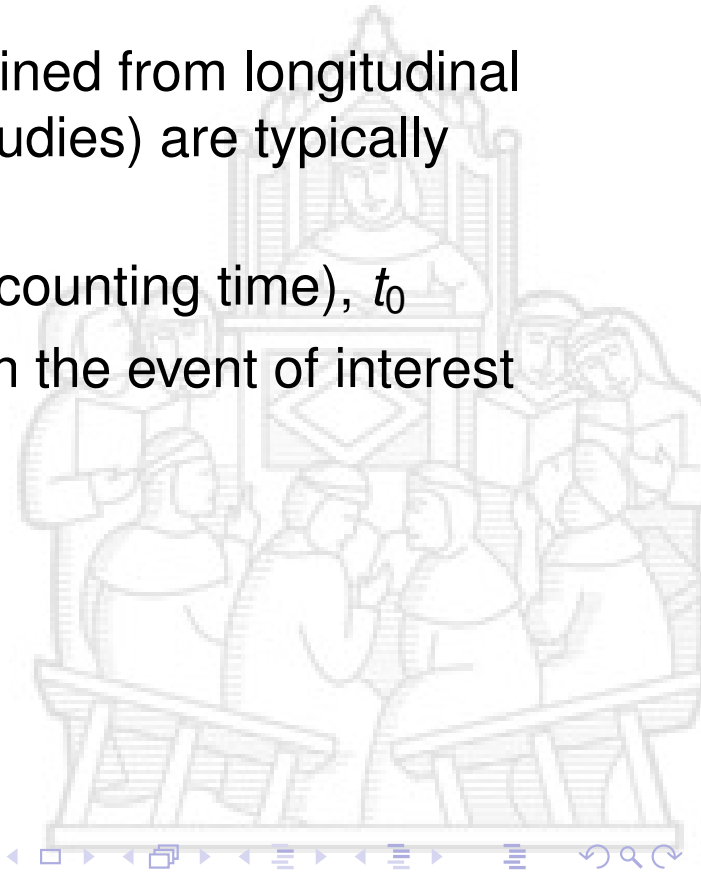
- How do we define the specific event of interest?
- How do we measure time to the event?
- ...



What are time-to-event data?

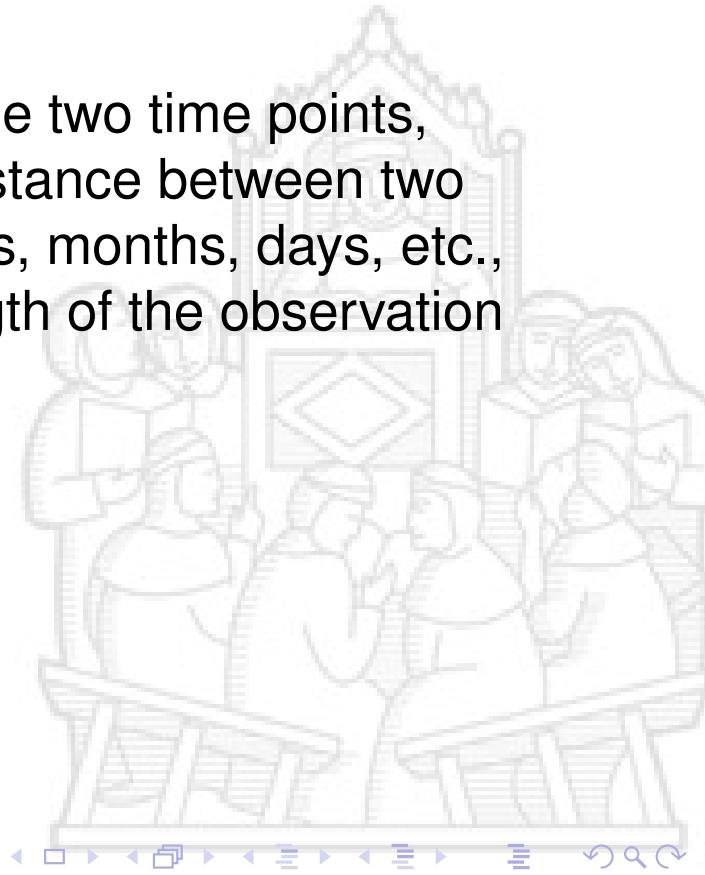
Time-to-event data (survival data) that are obtained from longitudinal studies (either observational or experimental studies) are typically characterized by

- a beginning (time) point (the time we start counting time), t_0
- an ending (time) point that is reached when the event of interest occurs



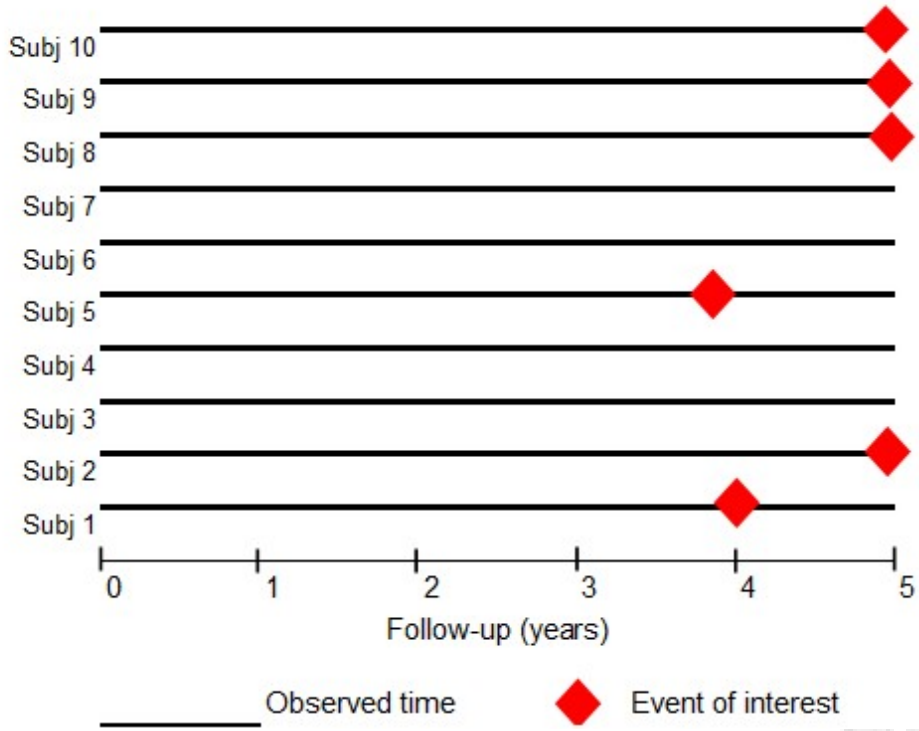
What are time-to-event data?

Survival time is thus the distance between those two time points, which in practice is typically obtained as the distance between two calendar dates and could be measured in years, months, days, etc., depending on the type of study and on the length of the observation (i.e., follow-up time)



Characteristics of time-to-event data

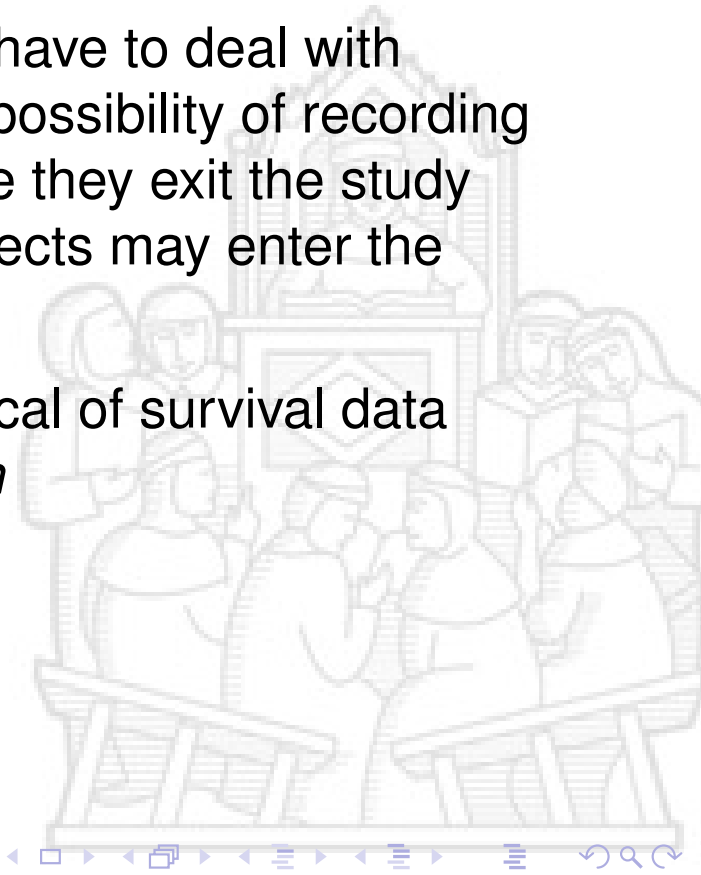
"Ideal" survival data,



Characteristics of time-to-event data

When doing a longitudinal study we frequently have to deal with incomplete data, which could arise from the impossibility of recording the event of interest on some subjects, because they exit the study before the event occurs, or because some subjects may enter the study later to the pre-defined starting date set

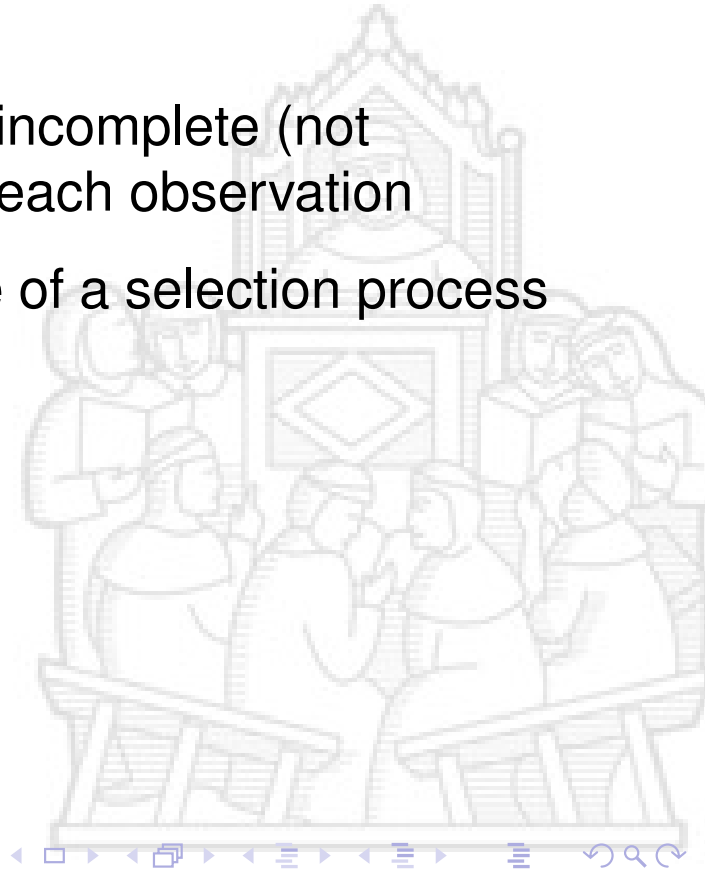
Those incomplete observations of time are typical of survival data and are referred to as *censoring* and *truncation*



Characteristics of time-to-event data

A *censored* observation is one whose value is incomplete (not known) because of factors that are random for each observation

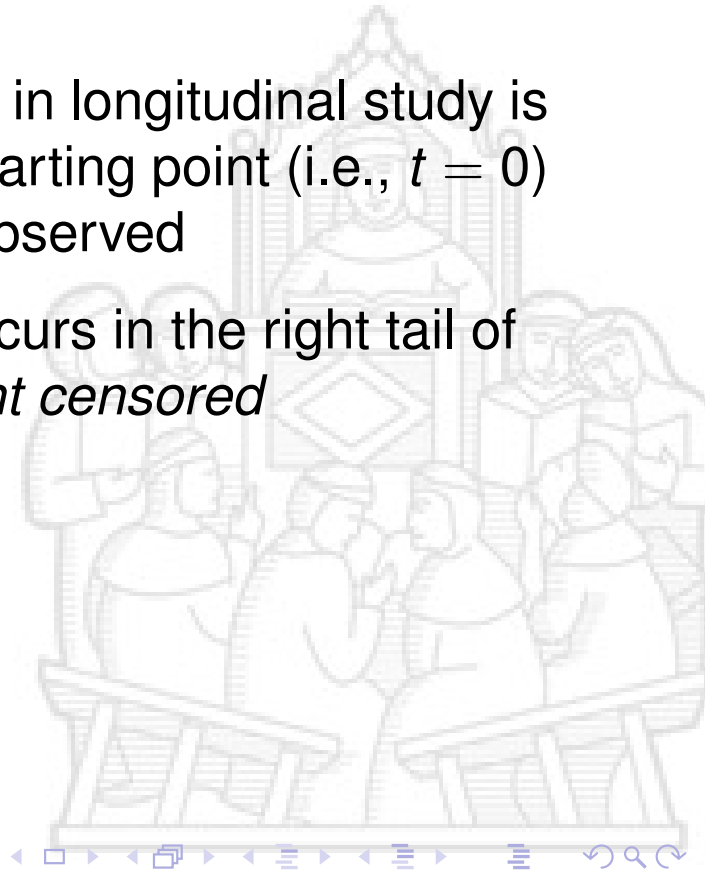
A *truncated* observation is incomplete because of a selection process that is inherent in the study design



Characteristics of time-to-event data?

The most common form of *censoring* occurring in longitudinal study is one where observation begins at the defined starting point (i.e., $t = 0$) and terminates before the event of interest is observed

As the incomplete nature of the observation occurs in the right tail of the time axis, that observation is said to be *right censored*

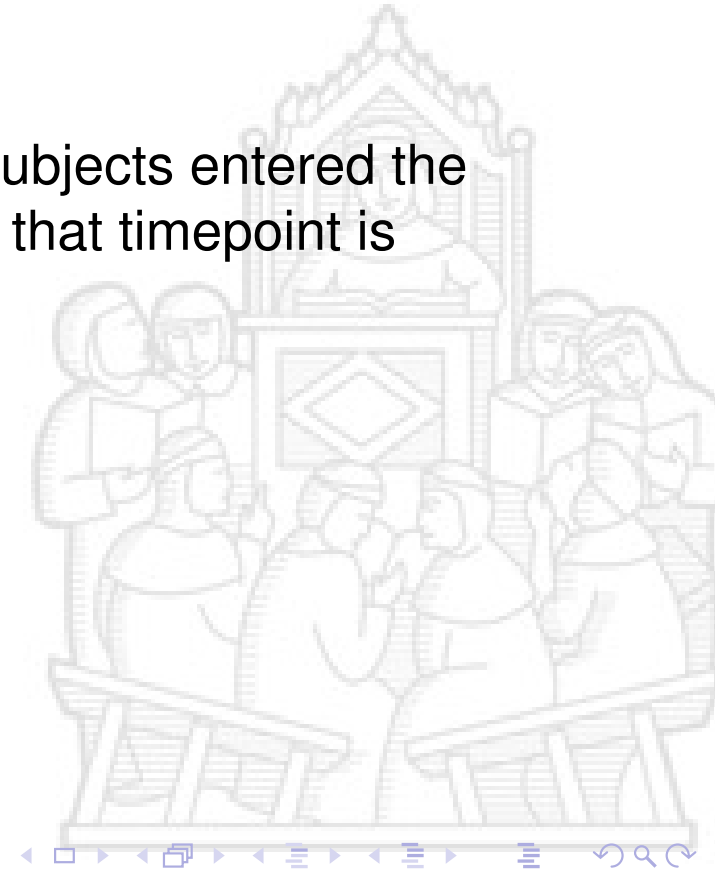


Characteristics of time-to-event data

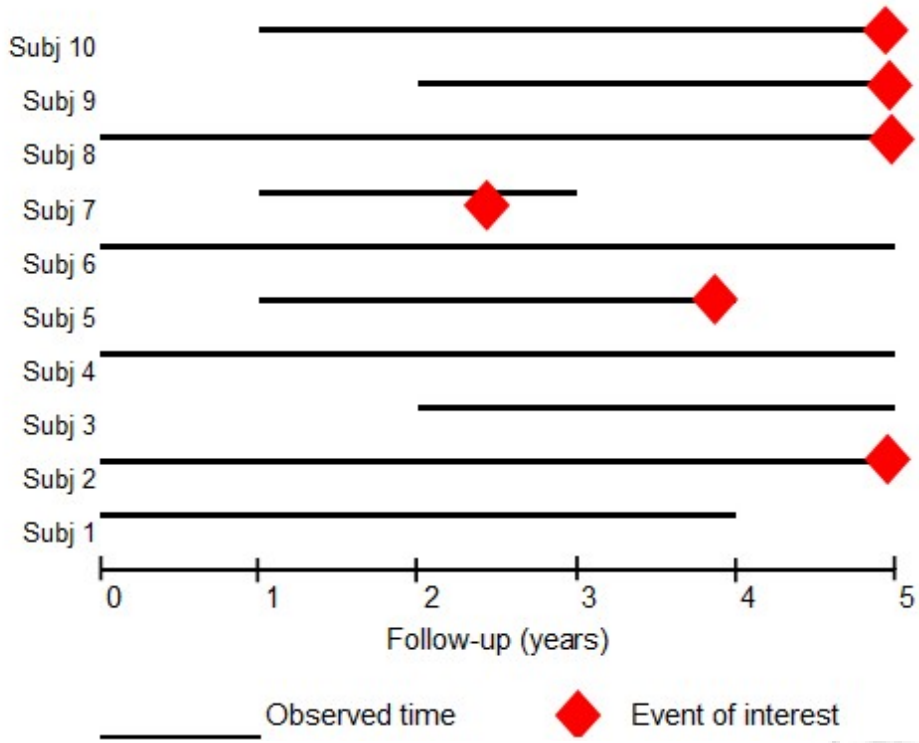
Another form of *censoring* that sometimes happens when conducting longitudinal studies is *left censored* which occurs when the event of interest has already occurred when the observation begins

Characteristics of time-to-event data

Left truncation or *delayed entry* occurs when subjects entered the study not at the beginning time points but after that timepoint is already passed

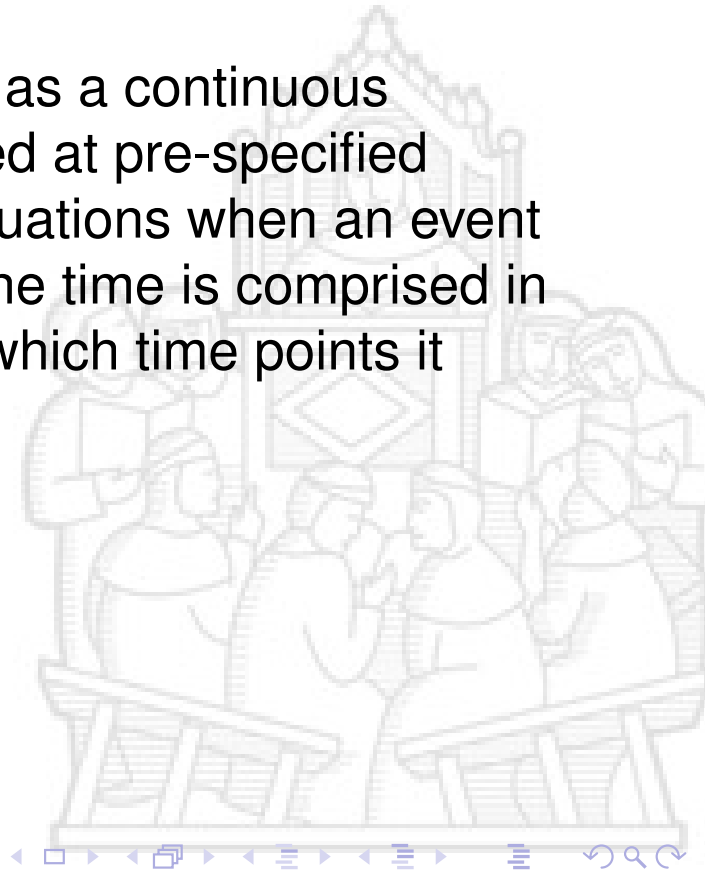


Characteristics of time-to-event data



Characteristics of time-to-event data

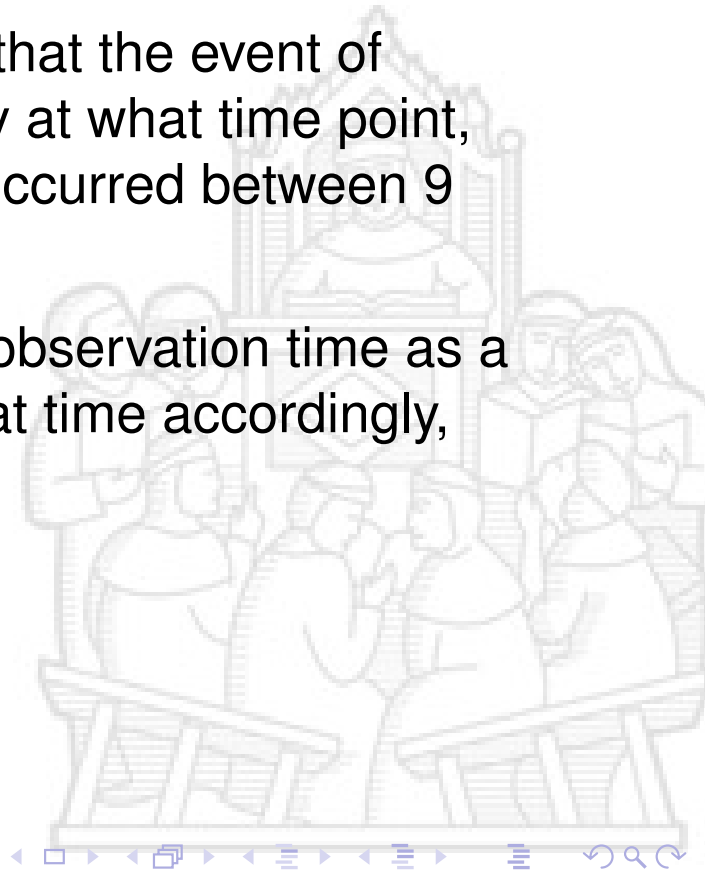
In most studies, time is collected and recorded as a continuous variable while sometimes subjects are contacted at pre-specified intervals (i.e., every 3 months), and in those situations when an event occurs at a certain interval, we just know that the time is comprised in a certain interval but we don't know exactly at which time points it happens



Characteristics of time-to-event data

For example, if a subject reports at 12 months that the event of interest has occurred but we don't know exactly at what time point, the only information we have is that the event occurred between 9 and 12 months

In those cases, it is not recommended to treat observation time as a continuous variable and to use models that treat time accordingly, those data are said to be *interval censored*



Describing time-to-event data

Because of the characteristics of time-to-event data, the description of those data must begin with the estimation of the cumulative distribution function

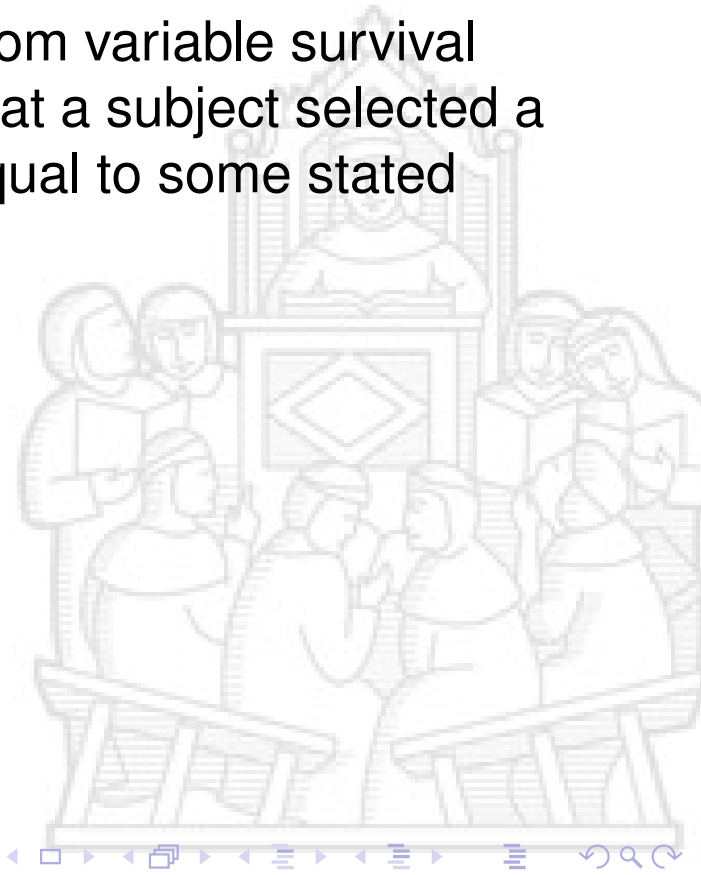
Remember that, given a random variable X, the cumulative distribution function is defined as

$$F(x) = P(X \leq x^*)$$

Describing time-to-event data

The cumulative distribution function of the random variable survival time, denoted T, is defined as the probability that a subject selected a random will have a survival time less than or equal to some stated value, t

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



Describing time-to-event data

The survival is thus the probability of observing a survival time greater than some stated value, t

$$S(t) = P(T > t)$$

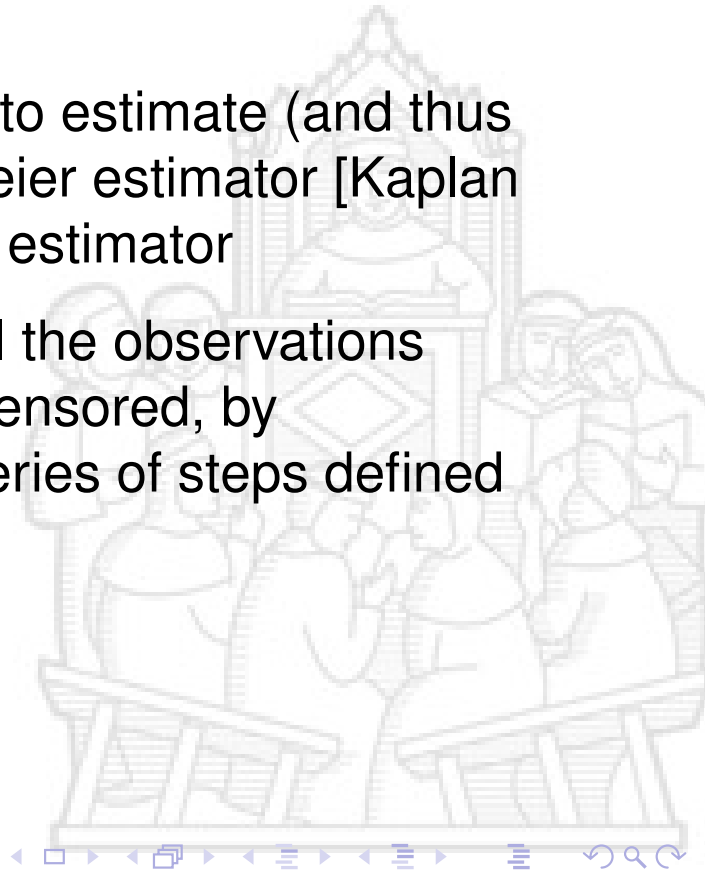
and by definition

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

Describing time-to-event data (Kaplan-Meier estimator)

One of the most widespread approaches used to estimate (and thus describe) the survival function is the Kaplan-Meier estimator [Kaplan and Meier (1958)], also called the *product limit* estimator

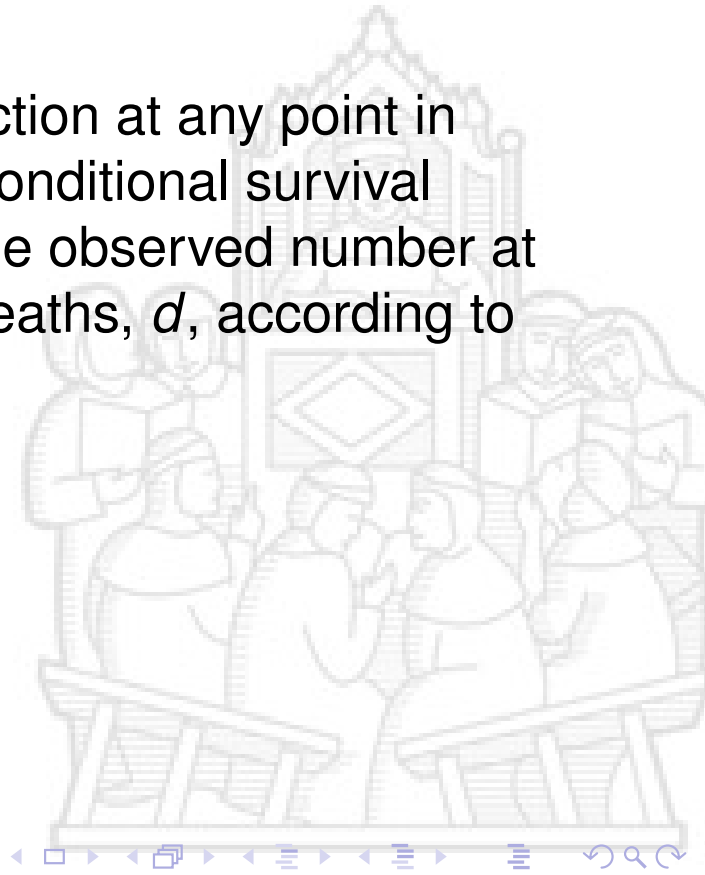
This estimator incorporates information from all the observations available, both uncensored (event times) and censored, by considering survival to any point in time as a series of steps defined at the observed survival and censored times



Describing time-to-event data (Kaplan-Meier estimator)

The Kaplan-Meier estimator of the survival function at any point in time is obtained by multiplying a sequence of conditional survival probability estimators which is obtained from the observed number at risk of death, n, and the observed number of deaths, d, according to the formula

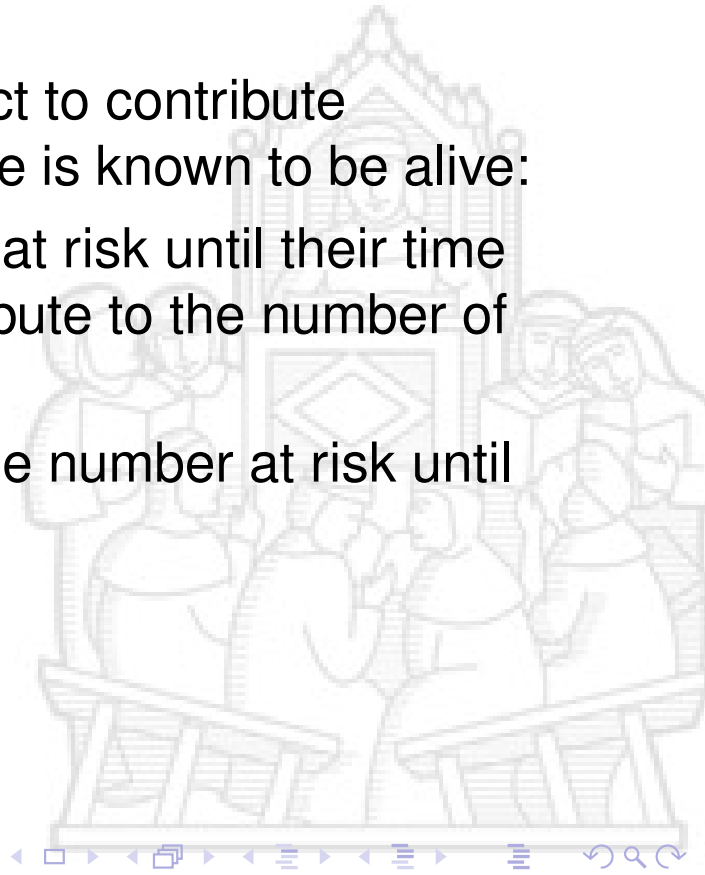
$$\frac{(n-d)}{n}$$



Describing time-to-event data (Kaplan-Meier estimator)

The Kaplan-Meier estimator allows each subject to contribute information to the calculations as long as he/she is known to be alive:

- subjects who die contribute to the number at risk until their time of death, and at this point, they also contribute to the number of deaths
- subjects who are censored contribute to the number at risk until they are lost to follow-up

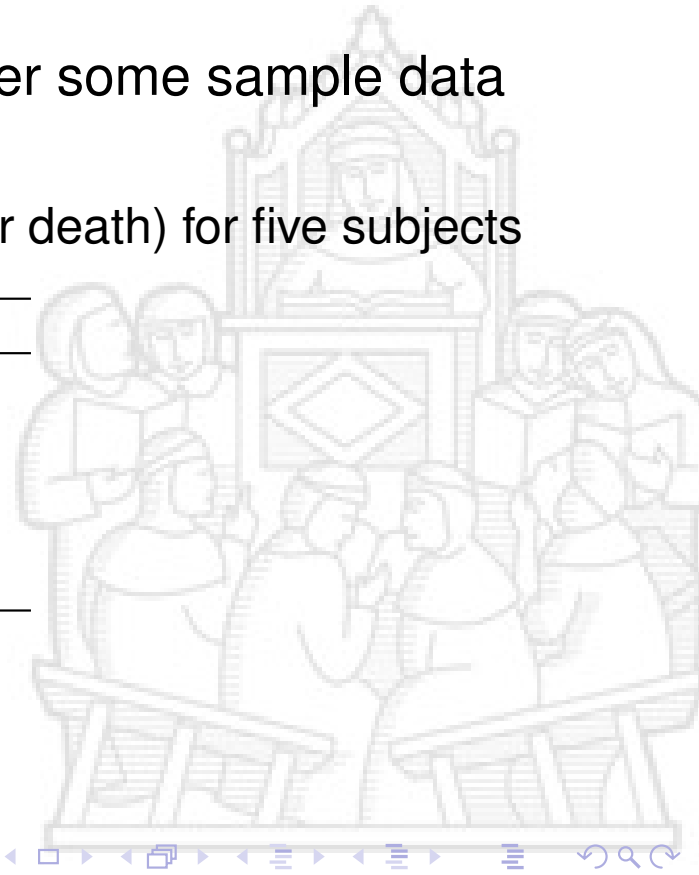


Describing time-to-event data (Kaplan-Meier estimator)

To illustrate the Kaplan-Meier estimator, consider some sample data

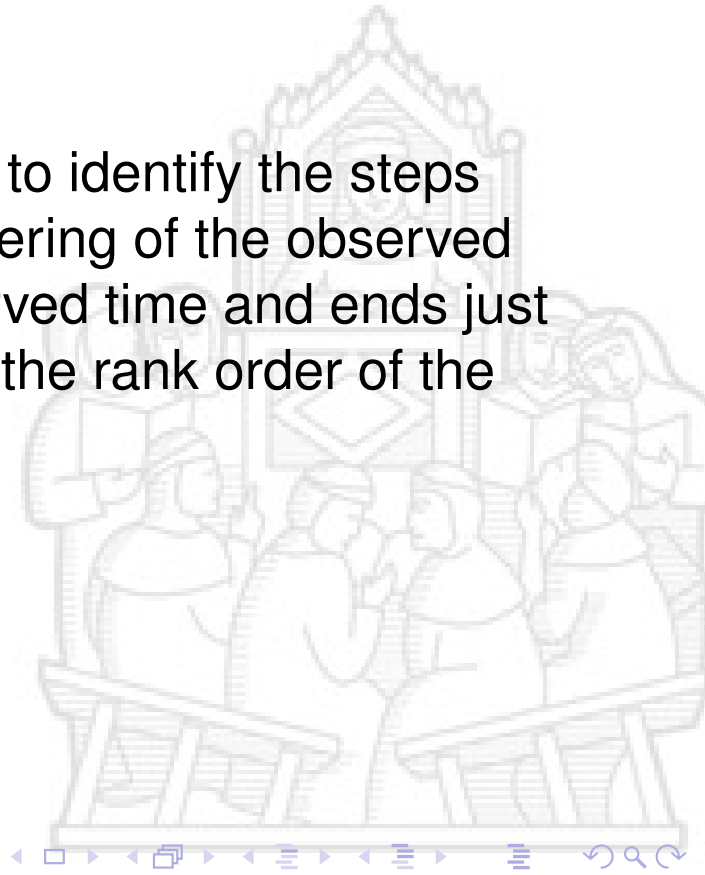
Table: Survival times and vital status (Censor==1 for death) for five subjects

Subject	Time	Censor
1	6	1
2	44	1
3	21	0
4	14	1
5	62	1



Describing time-to-event data (Kaplan-Meier estimator)

To derive the Kaplan-Meier estimator, we need to identify the steps that consist of intervals defined by the rank ordering of the observed time, in details each interval begins at an observed time and ends just before the next ordered time and is indexed by the rank order of the time points defining its beginning



Describing time-to-event data (Kaplan-Meier estimator)

Using the example data shown before

Table: Survival times and vital status (Censor==1 for death) for five subjects

Subject	Time	Censor
1	6	1
2	44	1
3	21	0
4	14	1
5	62	1

the first rank-ordered of observed time is 6 days for the first subject, followed by subject 4 with 14 days of survival time, then subject 3 with 21 days of survival time, and then subjects 2 and 5 with respectively a survival time of 44 and 62 days

Describing time-to-event data (Kaplan-Meier estimator)

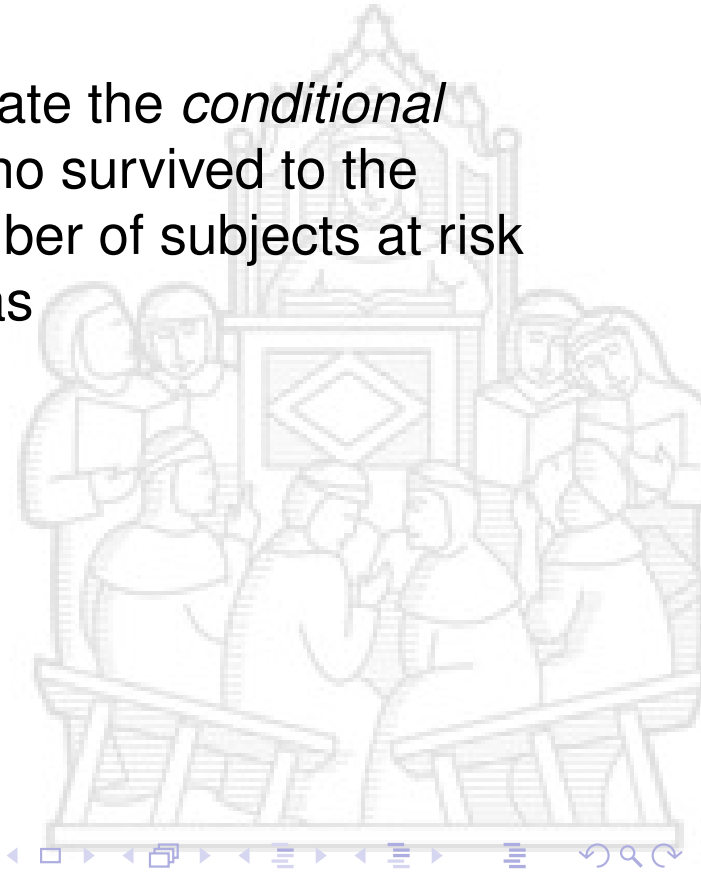
The intervals of rank-ordered time that could be identified are thus

$$I_0 = t : 0 \leq t < 6 = [0, 6)$$
$$I_1 = t : 6 \leq t < 14 = [6, 14)$$
$$I_2 = t : 14 \leq t < 21 = [14, 21)$$
$$I_3 = t : 21 \leq t < 44 = [21, 44)$$
$$I_4 = t : 44 \leq t < 62 = [44, 62)$$
$$I_4 = t : t \geq 62 = [62, \infty)$$

Describing time-to-event data (Kaplan-Meier estimator)

For each interval defined, it is possible to estimate the *conditional probability* of surviving (conditional on those who survived to the interval under consideration) based on the number of subjects at risk n_i and the number of death d_i in each interval as

$$\frac{(n_i - d_i)}{n_i}$$



Describing time-to-event data (Kaplan-Meier estimator)

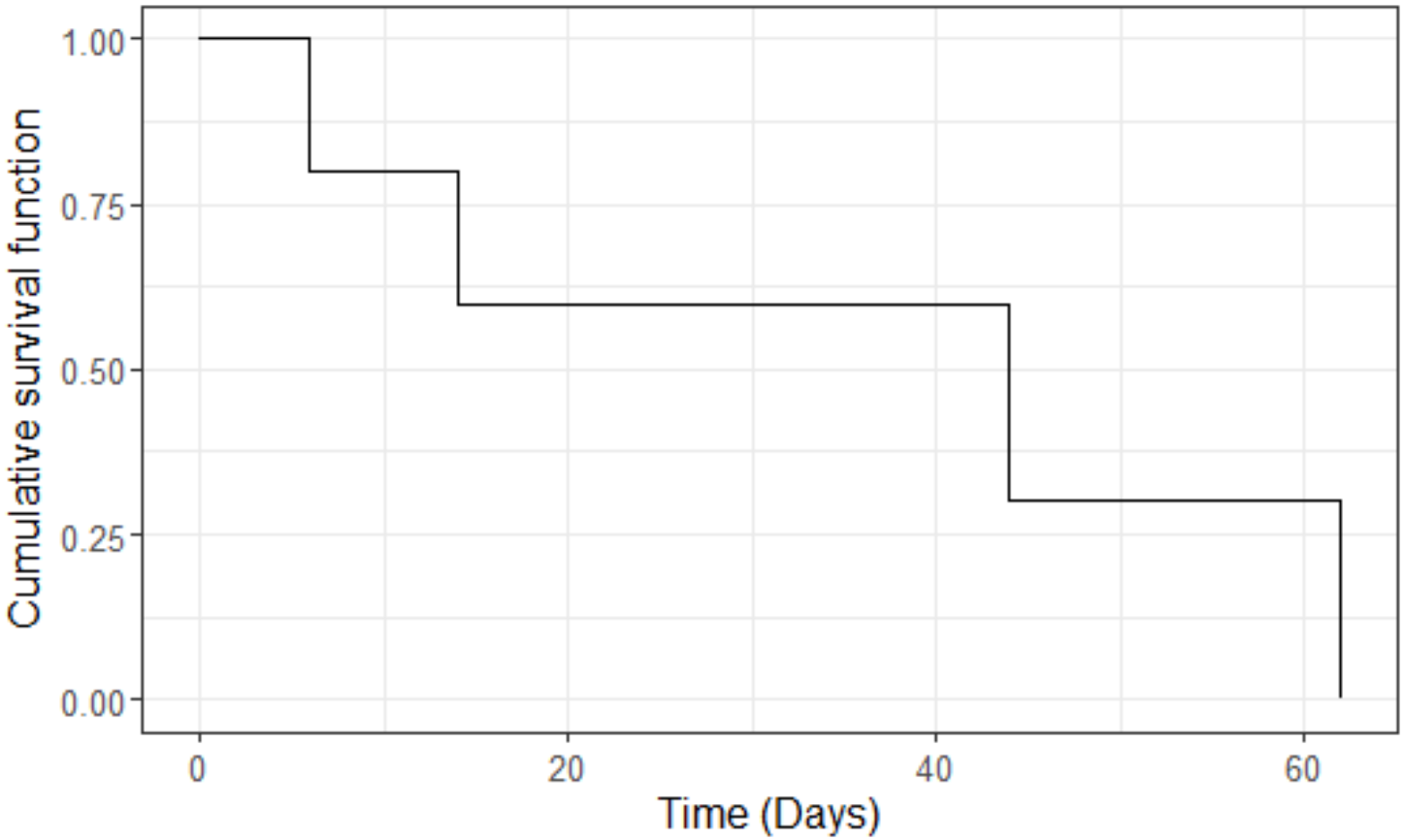
The Kaplan-Meier estimator of the survival function is defined the cumulative product of the conditional probability over all steps

$$\hat{S} = \prod_i \frac{(n_i - d_i)}{n_i}$$

Table: Estimated survival function estimated from Survival times and vital status (Censor==1 for death) for five subjects

Interval	Number at risk	Death	$\frac{(n-d)}{n}$	$\hat{S}$
$[0, 6)$	6	0	$(6-0)/6 = 1$	1.0
$[6, 14)$	5	1	$(5-1)/5 = 0.8$	0.8
$[14, 21)$	4	1	$(4-1)/4 = 0.6$	0.6
$[21, 44)$	3	0	$(3-0)/3 = 0.6$	0.6
$[44, 62)$	2	1	$(2-1)/2 = 0.3$	0.3
$[62, \infty)$	1	1	$(1-1)/0 = 0$	0

Describing time-to-event data (Kaplan-Meier estimator)



Describing time-to-event data (Kaplan-Meier estimator)

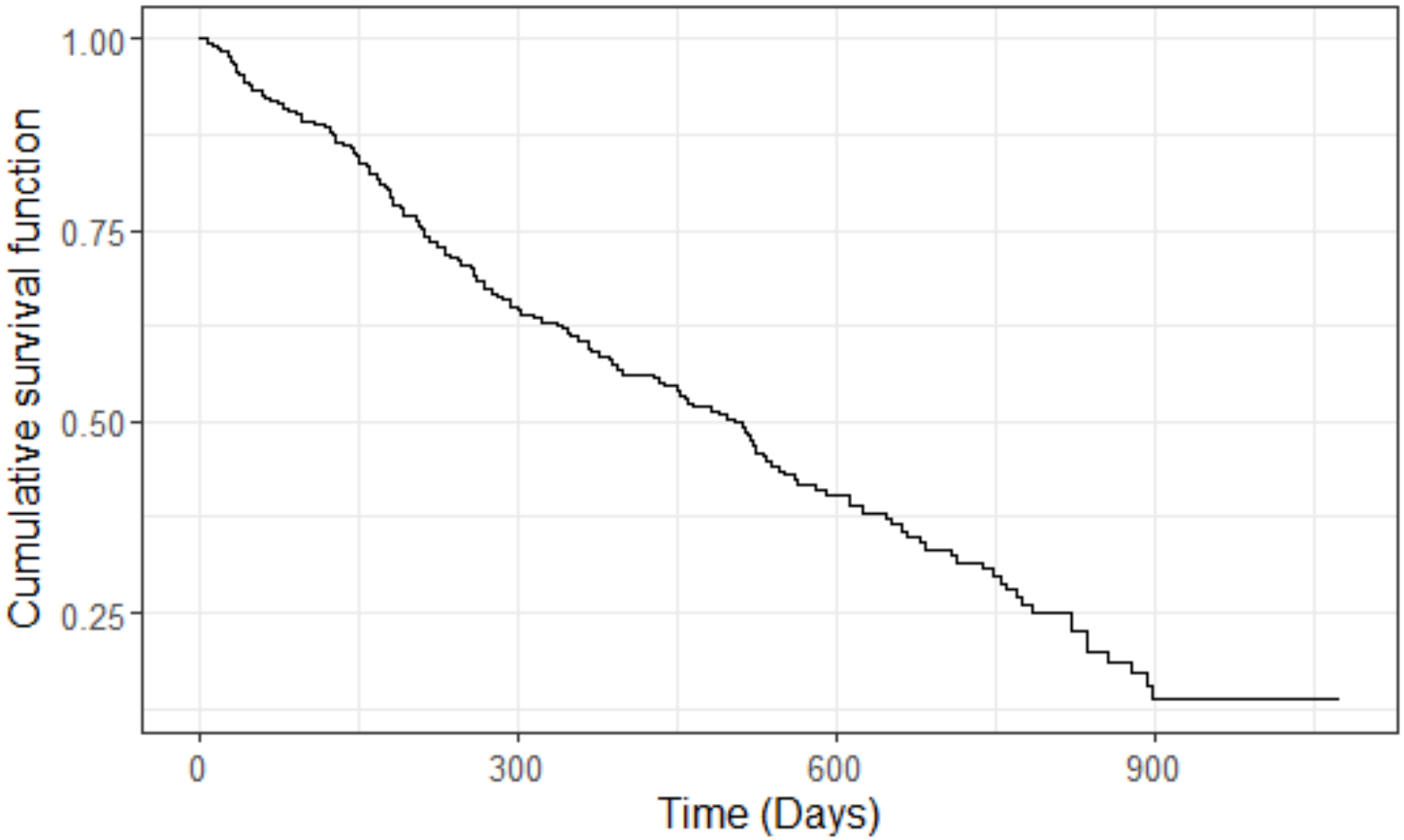
In general, assume to have a sample of n independent observation $(t_i, c_i), i = 1, 2, .., n$ of the underlying survival variable

The Kaplan-Meier estimator of the survival time variable T and the censoring indicator variable C, having observed $m \leq n$ recorded time of failure and $n - m$ censored values, the Kaplan-Meier estimator is given by

$$\hat{S(t)} = \prod_{t_{(i)} \leq t} \frac{(n_i - d_i)}{n_i}$$

where n_i and d_i represent respectively the number at risk of dying and the number of deaths over the rank-ordered survival times t_i with $t_1 < t_2 < ... < t_m$

Describing time-to-event data (Kaplan-Meier estimator)

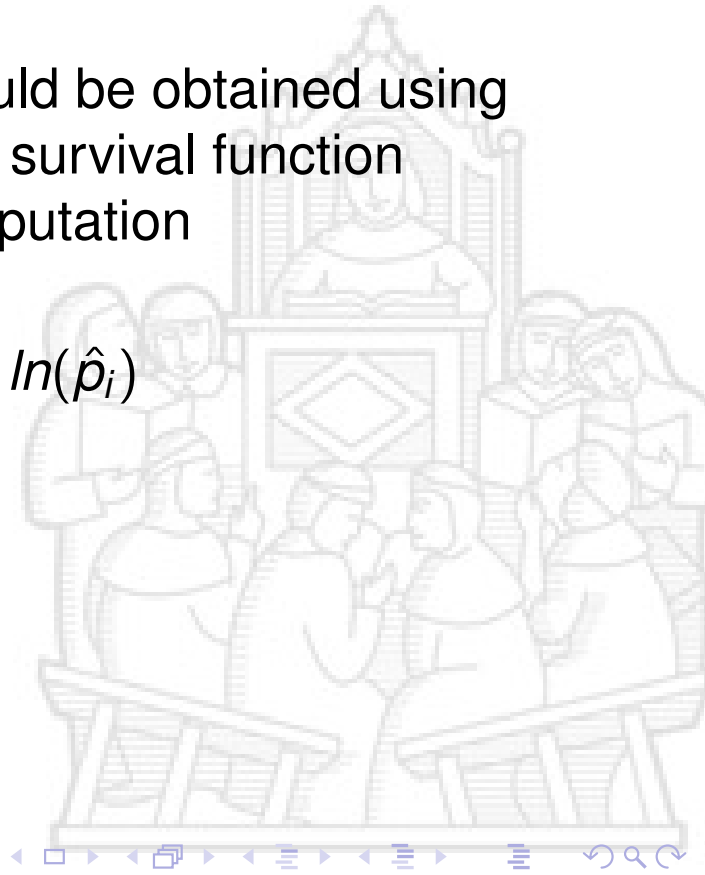


Describing time-to-event data (Kaplan-Meier estimator)

The variance of the Kaplan-Meier estimator could be obtained using the *delta method* and referring to the log of the survival function instead of the function itself to simplify the computation

$$ln(\hat{S(t)}) = \sum_{t_{(i)} \leq t} ln(\frac{(n_i - d_i)}{n_i}) = ln(\hat{p}_i)$$

where $\hat{p}_i = (n_i - d_i)/n_i$

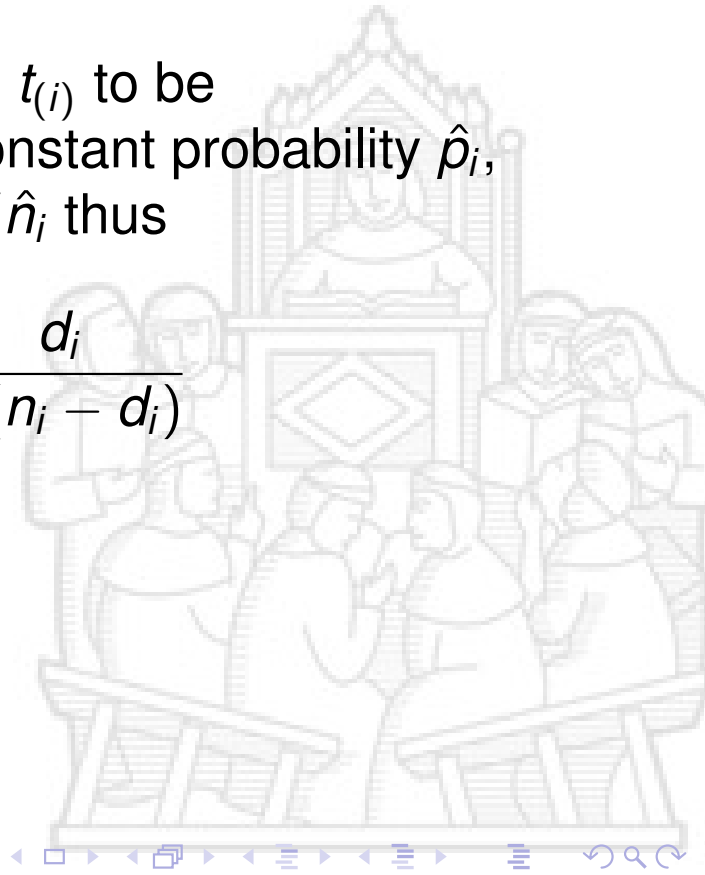


Describing time-to-event data (Kaplan-Meier estimator)

Considering observations in the risk set at time $t_{(i)}$ to be independently distributed as a Bernuolli with constant probability $\hat{p}_i$, its variance could be estimated as $(\hat{p}_i(1-\hat{p}_i))/\hat{n}_i$ thus

$$\hat{Var}[ln(\hat{p}_i)] \approx \frac{1}{\hat{p}_i^2}\frac{\hat{p}_i(1-\hat{p}_i)}{n_i} \approx \frac{d_i}{n_i(n_i-d_i)}$$

as $Var[ln(X)] \approx \frac{1}{\mu_{i2}}\sigma_x^2$



Describing time-to-event data (Kaplan-Meier estimator)

Assuming independent observation at each time the estimator of the variance of the log of the survival function is thus

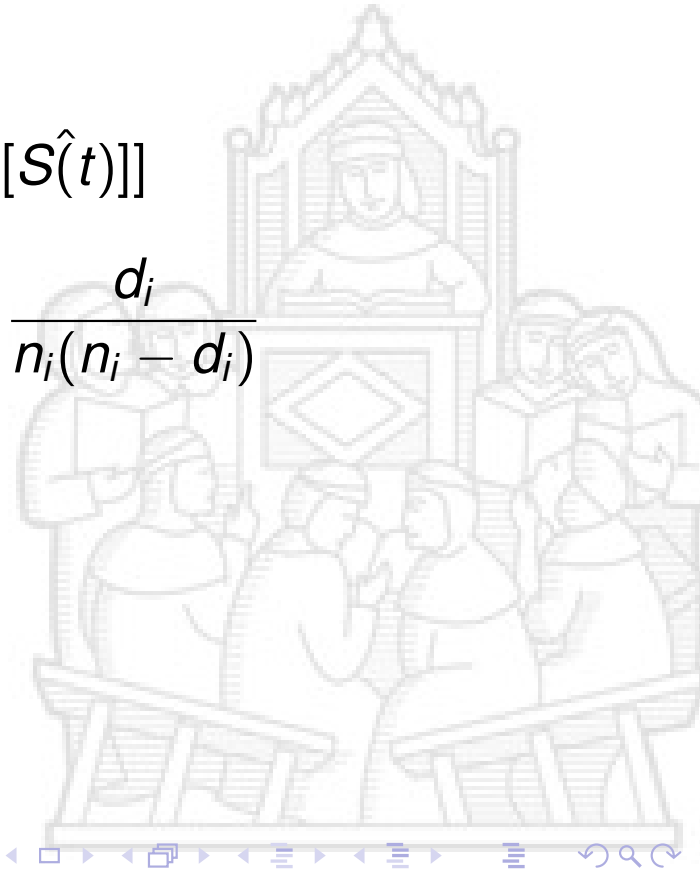
$$\hat{Var}[ln(\hat{S(t)})] = \sum_{t_{(i)} \leq t} \hat{Var}[ln(\hat{p}_i)] = \sum_{t_{(i)} \leq t} \frac{d_i}{n_i(n_i - d_i)}$$

$$\Longrightarrow \hat{Var}[\hat{S(t)}] = \hat{S(t)}^2 \sum_{t_{(i)} \leq t} \frac{d_i}{n_i(n_i - d_i)}$$

Describing time-to-event data (Kaplan-Meier estimator)

Similarly for the *log-log survival function*, $ln[-ln[\hat{S(t)}]]$

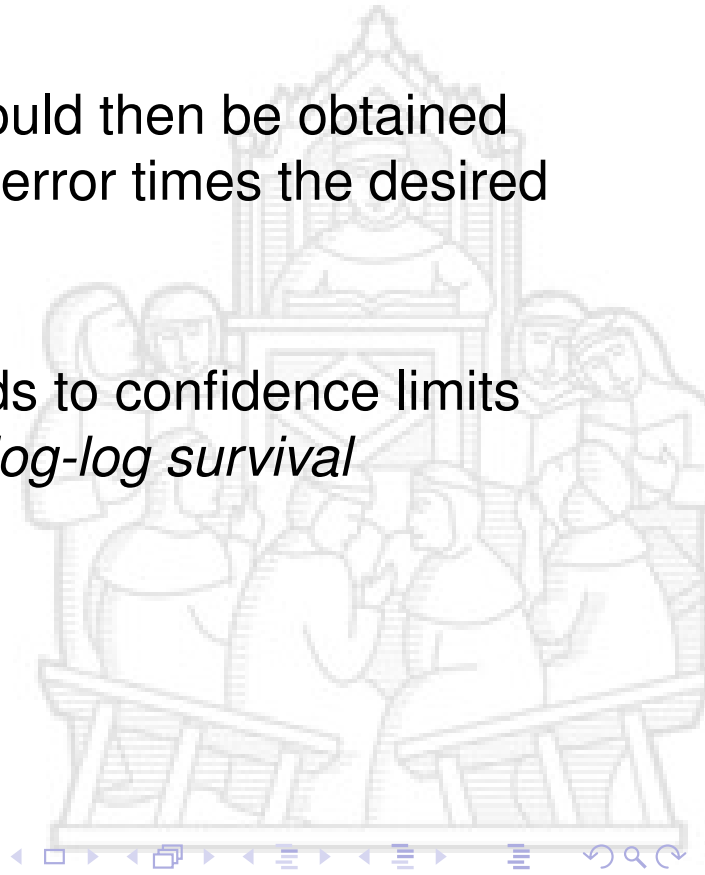
$$\hat{Var}[ln[-ln(\hat{S(t)})]] = \frac{1}{[ln(\hat{S(t)})]^2} \sum_{t_{(i)} \leq t} \frac{d_i}{n_i(n_i - d_i)}$$



Describing time-to-event data (Kaplan-Meier estimator)

Confidence intervals for the survival function could then be obtained by using the product of the estimated standard error times the desired quantile of the standard normal distribution

As the use of the *delta method* sometimes leads to confidence limits below zero or greater than one, the use of the *log-log survival function* is preferred to address this problem



Describing time-to-event data (Kaplan-Meier estimator)

Point and interval estimates of the survival function could also be used to compare survival experience between groups

The general idea behind the calculation of each test is based on a contingency table of groups by status at each observed survival time

Table: Table used for test of equality of the survival function in two groups at observed survival time t_i

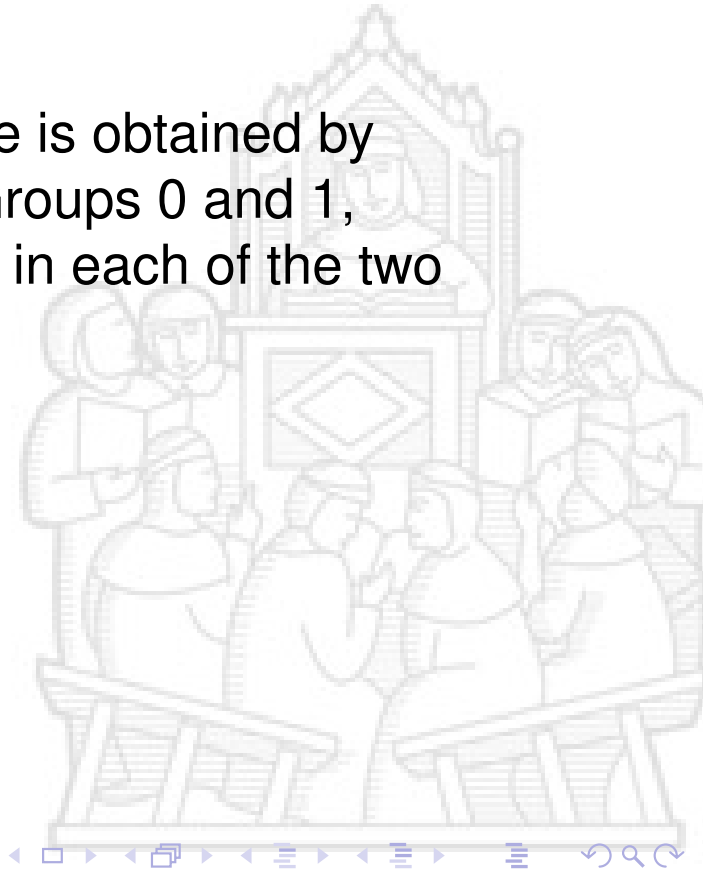
Event	Group 1	Group 0	Total
Die	d_{1i}	d_{0i}	d_i
Not die	$n_{1i} - d_{1i}$	$n_{0i} - d_{0i}$	$n_i - d_i$
At risk	n_{1i}	n_{0i}	n_i

where n_{ji} and d_{ji}, $j = 0, 1$ are respectively the number at risk and observed deaths in Group 0 and 1 at time t_i

Describing time-to-event data (Kaplan-Meier estimator)

The contribution to the test statistic at each time is obtained by calculating the expected number of deaths in Groups 0 and 1, assuming that the survival function is the same in each of the two groups, that is given by

$$\hat{e}_{ji} = \frac{n_{ji} d_i}{n_i}$$



Describing time-to-event data (Kaplan-Meier estimator)

The tests for comparison between groups may be defined in general as

$$Q = \frac{[\sum_{i=1} w_i(d_{ji} - \hat{e}_{ji}]}{\sum_{i=1} w_i^2 \hat{v}_{ji}}$$

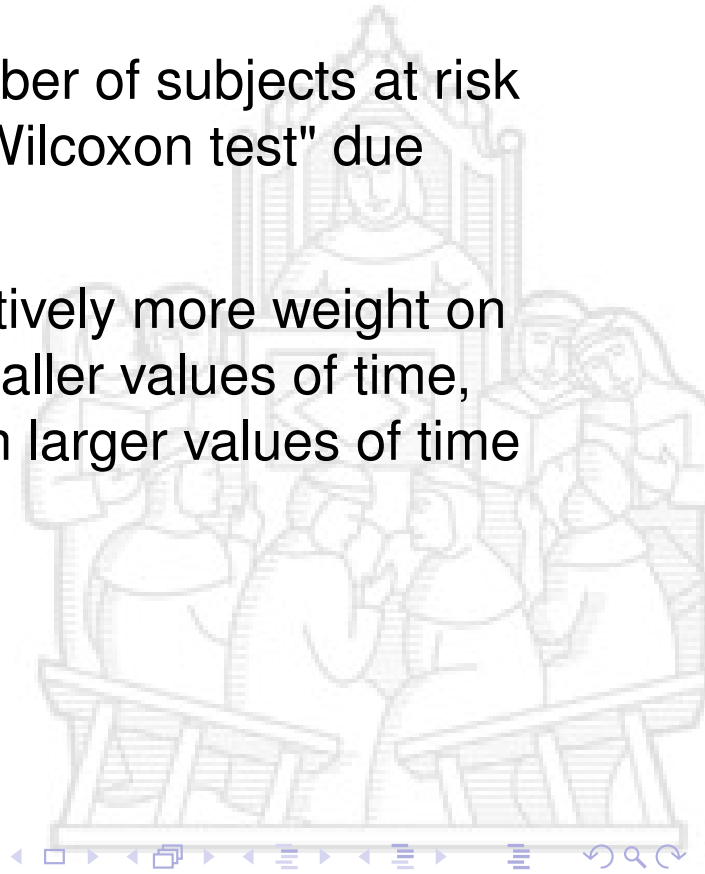
which follows a Chi-square distribution with one degree of freedom and w_i are weights whose values depend on the specific test and $\hat{v}_{ji}$ is the estimator of the variance of d_{ji}

Under the null hypothesis, the two survival functions are the same in the two groups, the censoring experience is assumed to be independent of the group, and the total number of observed events and the sum of the expected number of events is large

The well-known Log-rank test, due to Peto and Peto (1972) consider $w_i = 1$

When $w_i = n_i$, so weights are equal to the number of subjects at risk at each survival time we have the "generalizd Wilcoxon test" due Gehan (1965) and Breslow (1970)

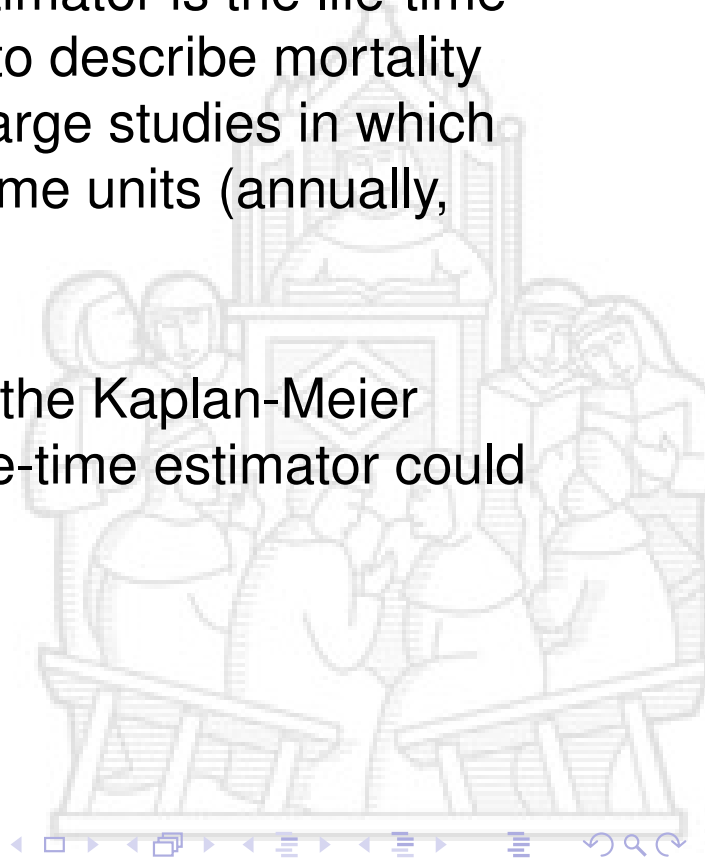
While the "generalized Wilcoxon test" puts relatively more weight on differences between the survival function at smaller values of time, the Log-rank test puts relatively more weight on larger values of time



Describing time-to-event data (Life-time estimator

A grouped-data analog of the Kaplan-Meier estimator is the life-time estimator that has been used over a long time to describe mortality data and well-fit situations when we deal with large studies in which mortality experience is presented in calendar time units (annually, semi-annually, etc.)

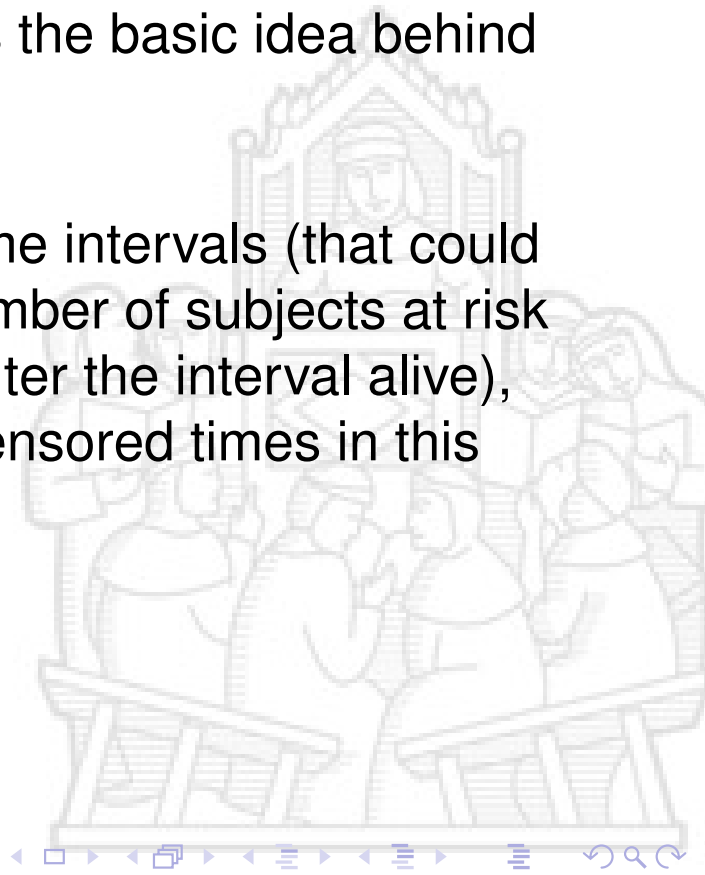
In these situations, the tabulation and graph of the Kaplan-Meier estimator can be quite cumbersome and the life-time estimator could be preferable



Describing time-to-event data (Life-time estimator

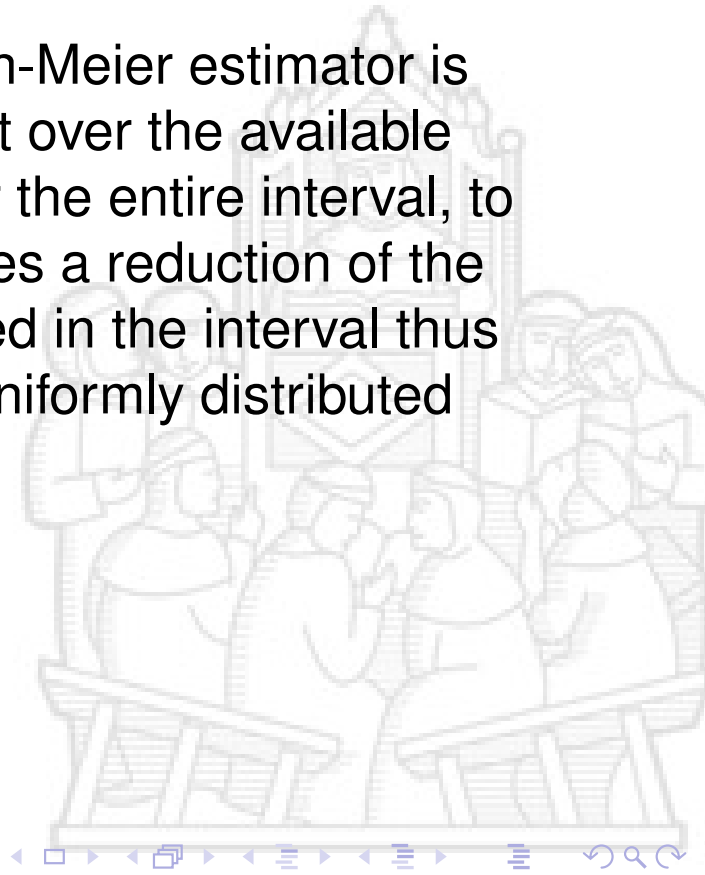
The derivation of the life-table estimator follows the basic idea behind the Kaplan-Meier estimator

Assume to having observed data over some time intervals (that could be year or semester) and let again n be the number of subjects at risk of dying at time t (also said the number who enter the interval alive), and d and c be subjects having survival and censored times in this interval



Describing time-to-event data (Life-time estimator)

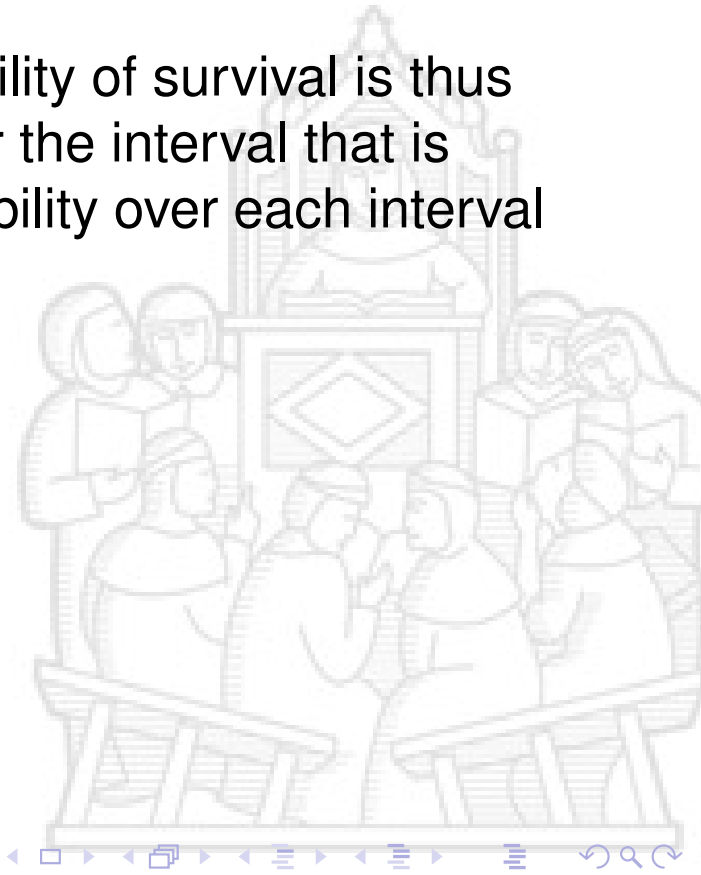
What differentiates the life-time from the Kaplan-Meier estimator is that the first one takes into account the fact that over the available intervals, not all subjects are at risk of dying for the entire interval, to account for that the life-time estimator introduces a reduction of the size of the risk set by one-half of those censored in the interval thus assuming that the censored observation was uniformly distributed over the interval



Describing time-to-event data (Life-time estimator)

In the life-time estimator the conditional probability of survival is thus estimated considering the average risk set over the interval that is $n - (c - 2)$, it follows that the conditional probability over each interval i is derived as

$$\frac{(n_i - (c_i/2) - d_i)}{n_i - (c_i/2)}$$



Describing time-to-event data (Life-time estimator)

Using the example data shown before

Table: Survival times and vital status (Censor==1 for death) for five subjects

Interval	Enter	Die	Censored	S
$[0,1)$	100	20	0	$\frac{(100-(0/2)-20)}{(100-(0/2))} = 0.80$
$[1,2)$	80	5	0	0.75
$[2,3)$	75	7	0	0.68
$[3,4)$	68	4	0	0.64
$[4,5)$	64	6	0	0.58
$[5,6)$	58	58	38	0.51
$[6,7)$	15	2	1	0.44
$[7,8)$	12	2	10	0.31

Describing time-to-event data (Life-time estimator)

The life-time estimator could be represented graphically using a step function or a polygon representation

Being derived from grouped data, the life-time estimator is not as precise as the Kaplan-Meier estimator which is based on individual data