

Survival Analysis

Autore: Mattia Pennati

Introduzione

- L'analisi di sopravvivenza è una branca della statistica che analizza l'**intervallo temporale** fra le occorrenze di un **evento** di interesse.
- Si concentra sui **tassi di sopravvivenza**, ovvero la probabilità di non avere occorrenze dell'evento nel **tempo di osservazione**.
(**oss**: si suppone che un soggetto possa avere una sola occorrenza dell'evento)
- È quindi centrata sulla **modellazione di eventi nel tempo**.
- È un'analisi di regressione con variabile risposta il tempo che intercorre fra l'inizio degli studi e l'occorrenza di un evento.

Terminologia

- **Evento** = trattandosi di analisi di sopravvivenza gli eventi in questione sono eventi che caratterizzano la fine della **lifetime** del soggetto, come :
 - Morte biologica
 - Occorrenza di fallimento di un sistema
 - Guasto meccanico
 - ...
- **Tempo di osservazione** = tempo che intercorre tra l'**inizio dell'osservazione** (t_0) e la sua **fine** che può essere :
 - Evento
 - Fine degli studi
 - Fine del tempo di osservazione a disposizione per quel soggetto
 - ...

Terminologia

- **Censored observation** = un'osservazione viene detta **censored** se il soggetto in questione non registra occorrenze dell'evento nel periodo di osservazione.
Dopo che un'osservazione viene considerata tale (tempo di censura) non si avrà più nessuna informazione sul soggetto.
- **Survival function $S(t)$** = funzione che modella la **probabilità** di un soggetto di sopravvivere oltre il tempo t .
- **Hazard** = situazione di **rischio**.
Un soggetto viene considerato a rischio in ogni tempo $t \geq t_0$ in cui ancora non ha registrato nessuna occorrenza dell'evento.

Dati di sopravvivenza

Considerando come evento la morte biologica i dati in questione sono dati **clinici** che, per ogni soggetto osservato, esprimono:

- Generalità (etnia, sesso, età, ...).
- Stato a fine osservazione :
 - ALIVE (o living/...) → osservazione censored (non ha registrato occorrenze dell'evento nel periodo di osservazione).
 - DEAD (o deceased/...) → ha registrato un'occorrenza dell'evento.
- Tempo di osservazione (espresso come un numero T che intende quanto tempo è passato dall'inizio dell'osservazione) :
 - Se lo stato è ALIVE T è assente e il tempo fornito corrisponde al **tempo di censura**.
 - Se lo stato è DEAD T corrisponde al tempo in cui è avvenuto l'evento.
- Eventuali informazioni relative a terapie o caratteristiche dei soggetti in analisi.

Dati di sopravvivenza

Esempio di dati di sopravvivenza ordinati per tempo di sopravvivenza. Le colonne sono:

- *Observation*: numero osservazione
- *Time*: tempo trascorso da inizio studi
- *Status*:
 - 1: occorrenza dell'evento, DEAD
 - 0: censored data, ALIVE
- *X*:
 - Maintained: sottoposto a chemioterapia
 - Nonmaintaned: non sottoposto a chemioterapia
 - È un esempio di un ulteriore dato fornito, non è sempre presente.

observation	time (weeks)	status	x
12	5	1	<u>Nonmaintained</u>
13	5	1	<u>Nonmaintained</u>
14	8	1	<u>Nonmaintained</u>
15	8	1	<u>Nonmaintained</u>
1	9	1	Maintained
16	12	1	<u>Nonmaintained</u>
2	13	1	Maintained
3	13	0	Maintained
17	16	0	<u>Nonmaintained</u>
4	18	1	Maintained
5	23	1	Maintained
18	23	1	<u>Nonmaintained</u>
19	27	1	<u>Nonmaintained</u>
6	28	0	Maintained
20	30	1	<u>Nonmaintained</u>
7	31	1	Maintained
21	33	1	<u>Nonmaintained</u>
8	34	1	Maintained
22	43	1	<u>Nonmaintained</u>
9	45	0	Maintained
23	45	1	<u>Nonmaintained</u>
10	48	1	Maintained
11	161	0	Maintained

Modelli

Dovendo formulare predizioni o stimare la sopravvivenza dei soggetti in analisi si necessita di un **modello** di riferimento per esprimere i dati osservati in funzione del tempo.

Sembrerebbe ragionevole utilizzare un modello di regressione lineare, ma si incontrerebbero due problemi principali:

- I dati in questione (tempi di sopravvivenza) sono numeri esclusivamente positivi
- Non gestirebbe efficacemente le osservazione censurate.

La scelta ricade quindi su differenti modelli in base ai dati a disposizione, alle scelte che si compiono, agli aspetti che si vuole evidenziare, alla precisione ricercata, ...

Modelli

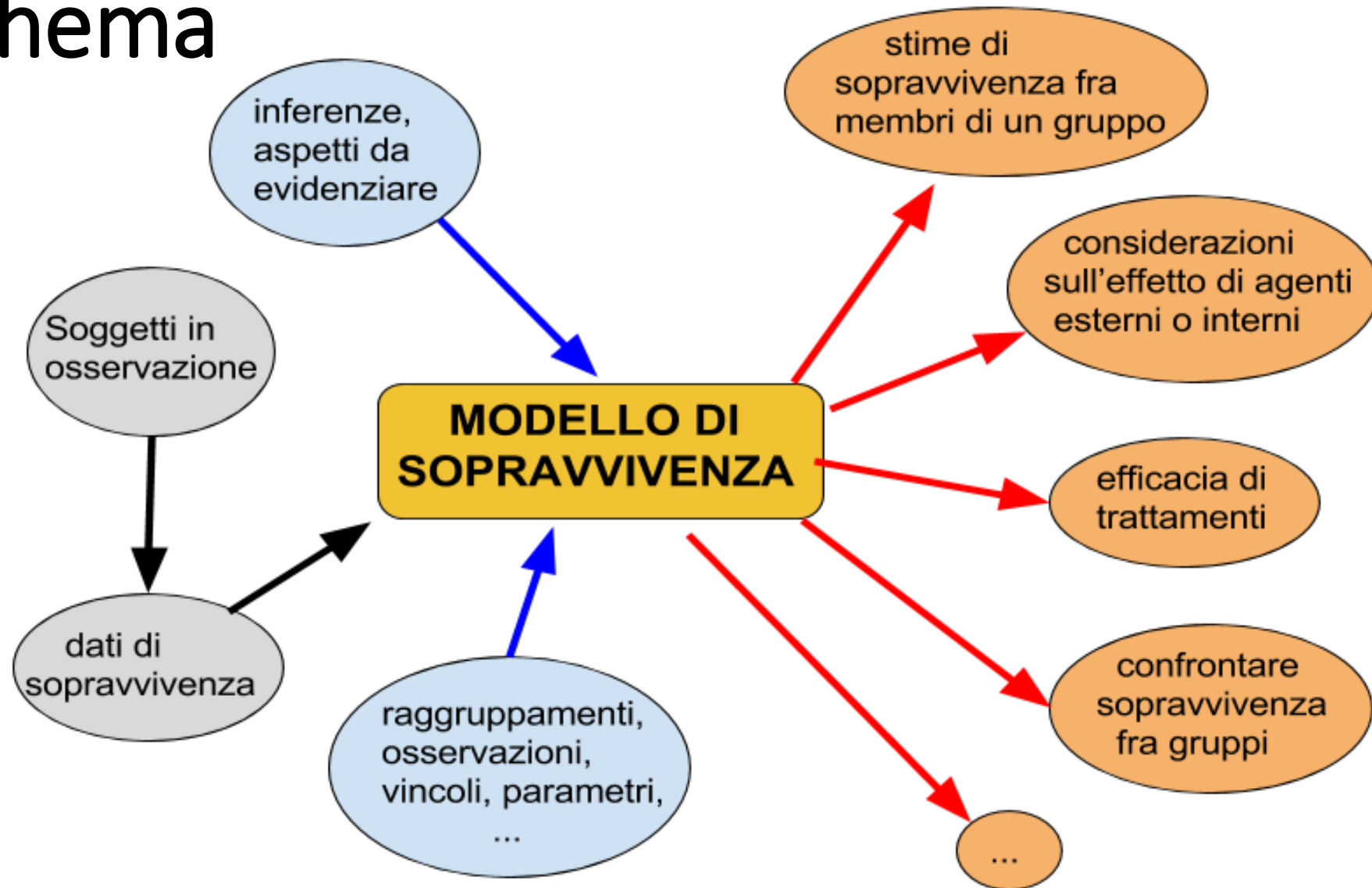
Si avranno quindi modelli diversi in base a :

- Stima della funzione di sopravvivenza
 - Parametrica → *parametric survival models*
 - Non parametrica → *non parametric survival models*
- Considerazioni del rapporto tra rischi (*hazards*) e il tempo
 - Proporzionali → *proportional hazards models* → *Cox model*
 - Non proporzionali → *accelerated failure time models*
- Eventuali differenziazioni su caratteristiche osservabili del campione (si ottengono stime di sopravvivenza più precise se trovate su raggruppamenti con caratteristiche simili che sul campione intero)
 - *Tree-structured survival models*
 - *Survival trees analysis*
 - *Survival random forest*

Applicazioni

- Stimare **tempo di sopravvivenza** dei membri di un gruppo
 - Survival Function
 - Hazard Function
 - curve di Kaplan-Meier
 - Life Tables
- Confrontare i **tempi di sopravvivenza di 2 gruppi**
 - Log-Rank Test
- Descrivere l'**effetto di variabili** quantitative o discrete relative ai soggetti (come l'età, la presenza di determinate mutazioni genetiche, il sesso, ...)
 - Parametric Survival Models
 - Cox Proportional Hazards Regression
 - Tree-structured Survival Models

Schema



Censoring

Solitamente è nota la **lifetime** di un soggetto, ovvero si conoscono:

- $t_0 \rightarrow$ tempo di inizio osservazione
- $T \rightarrow$ morte

Alcune volte però le osservazioni sono **censored** e T corrisponde al **tempo di censura**, ovvero il tempo di fine osservazione per quel soggetto.

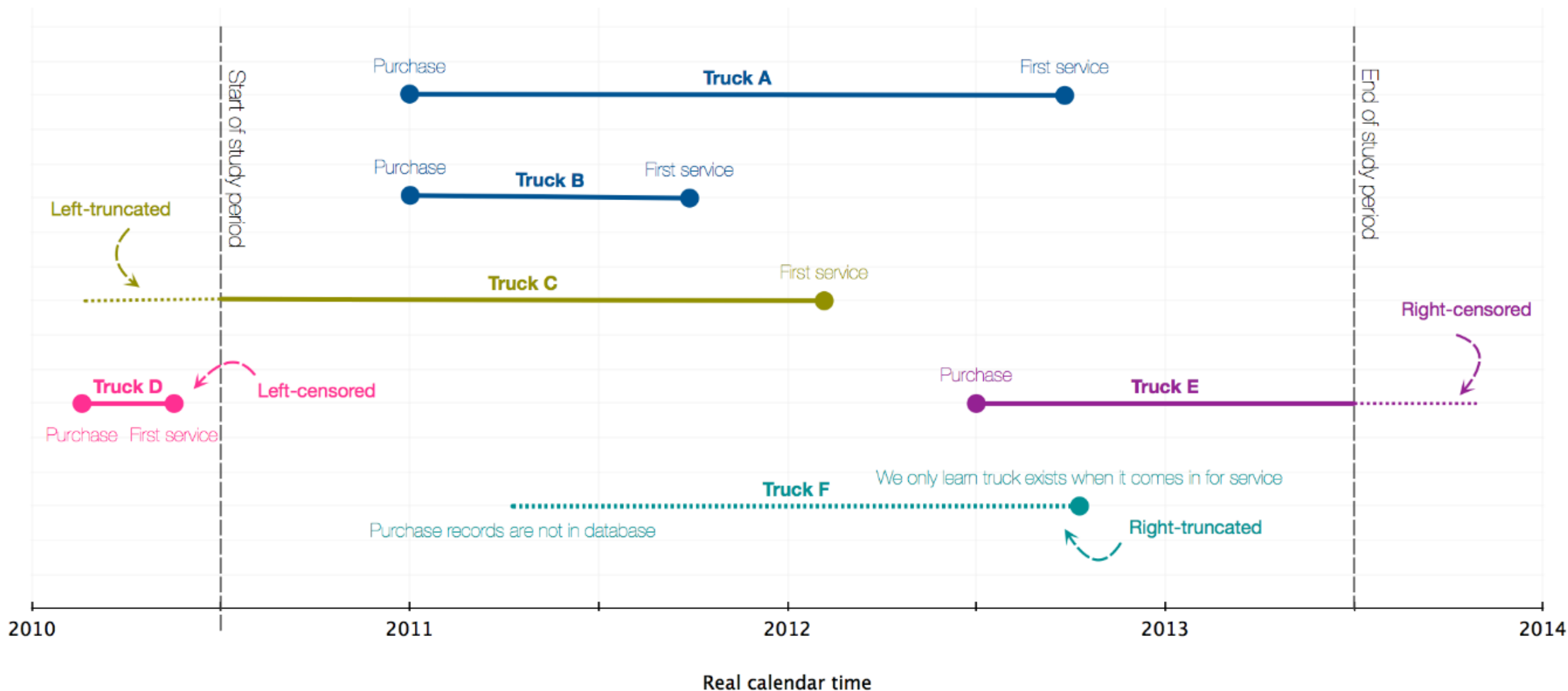
Durante il periodo di osservazione non è stata registrata nessuna occorrenza dell'evento e non si avranno informazioni riguardanti il soggetto in un tempo futuro ($t > T$).

Censoring

In base ai dati mancanti o impossibili da ottenere :

- **Right-Censoring** : si conosce t_0 , ma si perdono i contatti con il soggetto (o termina il periodo di osservazione) prima che si registri un'occorrenza dell'evento. Si sa quindi solo che T (inteso come evento morte) sarà successivo.
- **Left-Censoring** : l'evento era già accaduto a inizio studi. Si conosce solo che la lifetime di un soggetto è minore di un determinato tempo, ma non si hanno informazioni sull'occorrenza dell'evento.
 - **Truncation** (diverso dalla *left-censoring*) : il soggetto non è stato osservato o lo è stato solo parzialmente perché ha una lifetime minore di una certa soglia.
(es. In molti studi i bambini si osservano solo in età scolastica, che viene vista come la soglia)

Censoring



Survival Function & Hazard Function

Sono le due funzioni principali dell'analisi di sopravvivenza e sono semanticamente una l'opposto dell'altra.

- Survival function $S(t)$: probabilità che un soggetto sia vivo al tempo t .
- Hazard Function $\lambda(t)$: è l'**event rate** (tasso degli eventi) al tempo t condizionato dalla sopravvivenza fino o oltre t .

Entrambe sono basate sulla variabile aleatoria T_i , ovvero quella variabile corrispondente al tempo in cui il soggetto i registra l'occorrenza dell'evento.

Essa ha una propria distribuzione di probabilità che determinerà la sopravvivenza degli individui del campione sotto osservazione.

Variabile T_i

Come tutte le variabili aleatorie avrà una distribuzione con una densità di probabilità (*event density*) e una funzione di ripartizione (*lifetime distribution function*) :

- *Event density* $f(t) = F'(t) = \frac{dF(t)}{dt}$

- *Lifetime distribution function* $F(t) = P(T \leq t) = \int_0^t f(u)du$

OSS: considerando l'essere umano T **non** è distribuita normalmente, visto che la probabilità che il soggetto muoia sarà massima in età avanzata e comunque più alta a inizio vita rispetto all'età media.

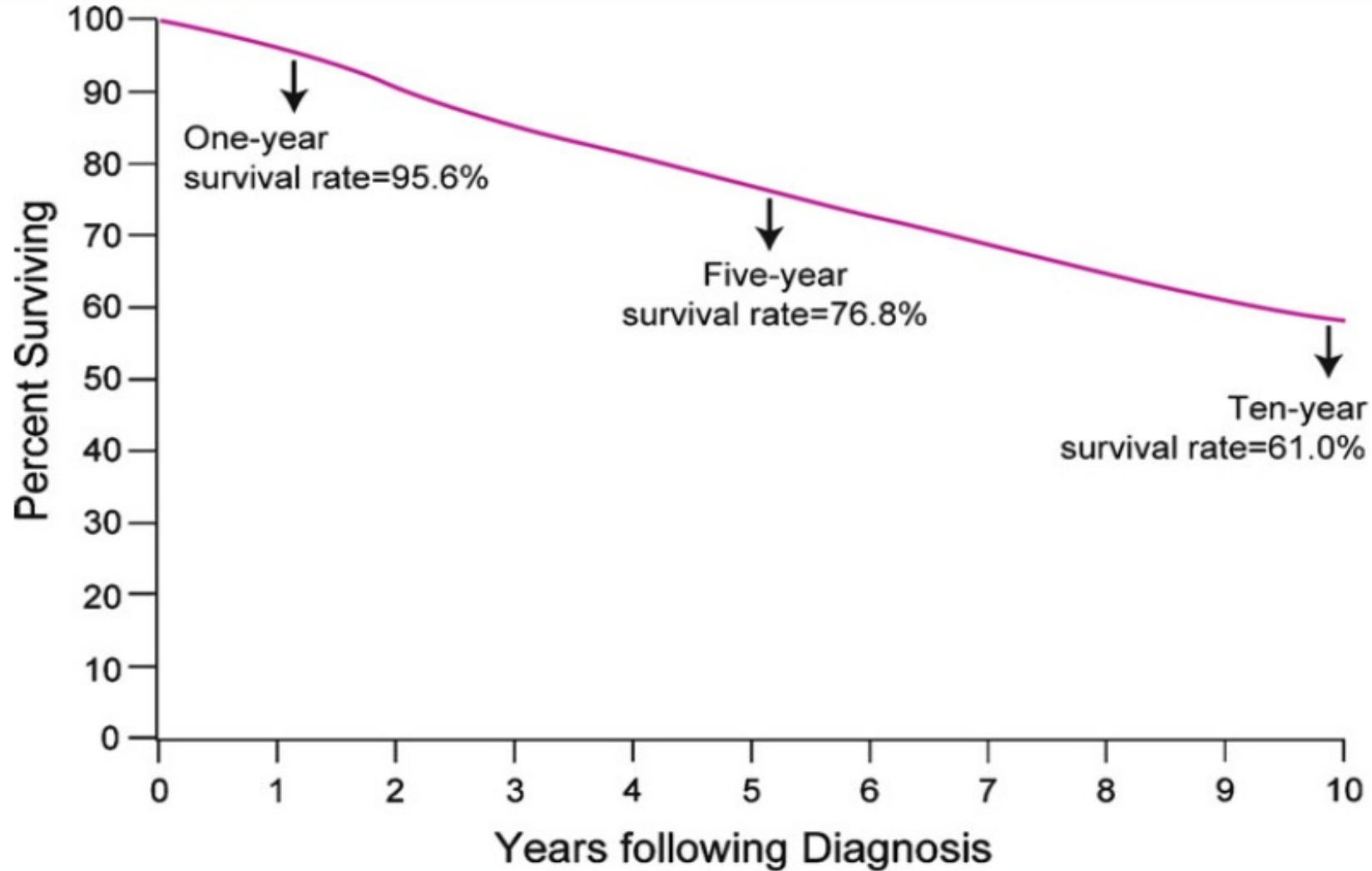
Survival Function $S(t)$

È la probabilità che un soggetto registri un'occorrenza dell'evento successivamente al tempo $t \rightarrow S(t) = P(T > t)$.

Essa non è nota a priori e si può ottenere in due modi :

- Approssimando/considerando $S(t)$ distribuita secondo una determinata *distribuzione parametrica di probabilità* \rightarrow *modelli parametrici*.
 - Implica l'utilizzo di parametri che, se non noti a priori, possono essere stimati.
- Stimandola tramite lo *stimatore di Kaplan-Meier* \rightarrow *modelli non parametrici*.
 - Non c'è la necessità di stimare o di conoscere nessun parametro.

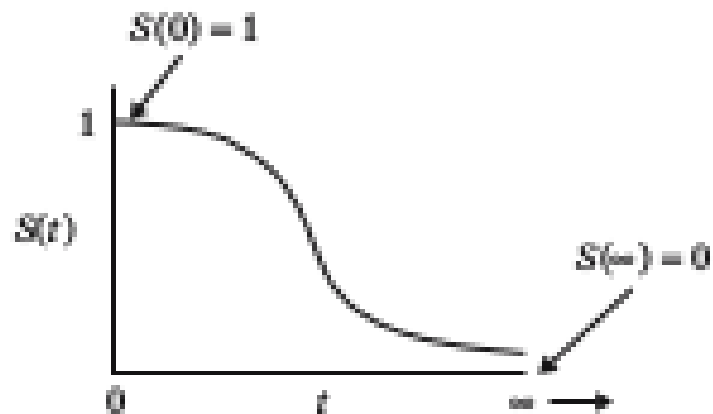
Survival Function $S(t)$



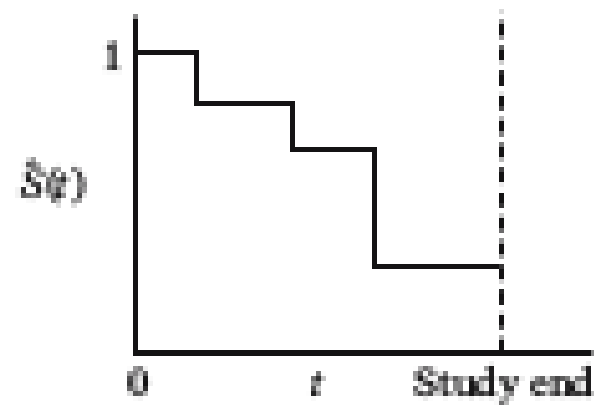
Survival Function: proprietà

- È una funzione **monotona NON crescente**, ovvero $S(u) \geq S(t) \quad \forall u > t$
 - $S(t) = 1$ se $t = 0$
 - $S(t) = 0$ se $t = \infty$ ($t \rightarrow \infty$)
- È una funzione *teoricamente* continua (perché il tempo è continuo), ma viene considerata *praticamente* come discreta perché si considera il tempo come discreto (giorni, settimane, mesi, ...)

Theoretical $S(t)$:



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



Survival Distribution Function

La funzione di sopravvivenza $S(t)$ è una funzione che $\forall t$ mi restituisce la probabilità $P(T > t)$.

Essa seguirà una determinata distribuzione e la sua funzione di ripartizione è il complemento della funzione di ripartizione $F(t)$ della variabile T_i , ovvero quella funzione che $\forall t$ mi restituisce la probabilità $P(T \leq t)$.

- $S(t) = P(T > t) = \int_t^{\infty} f(u)du = 1 - F(t)$
- $F(t) = P(T \leq t) = \int_0^t f(u)du = 1 - S(t)$

Survival Event Density

Dalla definizione di $S(t)$ si può ottenere anche la sua **densità di distribuzione $s(t)$** :

$$s(t) = S'(t) = \frac{dS(t)}{dt} = \frac{d}{dt} \int_t^{\infty} f(u) du = \frac{d}{dt} [1 - F(t)] = f(t)$$

Essa prende anche il nome di **first-passage time** distribution, dove con first-passage time si intende la prima volta che un processo stocastico raggiunge una soglia (che può essere uno stato, un limite, ...)

Future Lifetime

Dalla funzione di sopravvivenza $S(t)$ si possono derivare

- *Future Lifetime*
- *Expected Future Lifetime*

Future Lifetime: tempo mancante alla morte conoscendo la sopravvivenza ad un tempo $t_0 \rightarrow T - t_0$, ovvero la probabilità che $T \leq t_0 + t$

$$P(T \leq t_0 + t \mid T > t_0) = \frac{P(t_0 < T \leq t_0 + t)}{P(T > t_0)} = \frac{F(t_0 + t) - F(t_0)}{S(t_0)}$$

- Densità di probabilità: $\frac{d}{dt} \frac{F(t_0 + t) - F(t_0)}{S(t_0)} = \frac{f(t_0 + t)}{S(t_0)}$

Expected Future Lifetime

È il valore atteso della future lifetime

- $$\frac{1}{S(t_0)} \int_0^{\infty} t f(t_0 + t) dt = \int_{t_0}^{\infty} S(t) dt$$

- Se $t = t_0$ (nascita) → **expected lifetime**

OSS: presupponendo n individui con la stessa $S(t)$, se la sopravvivenza di ogni singolo individuo è indipendente allora il **numero atteso** di sopravvissuti al tempo t è $nS(t)$

Quindi il valore atteso della future lifetime è :

$$t * \frac{f(t_0+t)}{S(t_0)} = \frac{1}{S(t_0)} t f(t_0) = \frac{1}{S(t_0)} \int_0^{\infty} t f(t_0 + t) dt$$

Hazard Function $\lambda(t)$

- È la funzione che rappresenta il rischio di morire in un determinato momento t , ovvero indica la probabilità di registrare un evento condizionata dal fatto che si è sopravvissuti fino a tempo t .
- Ha come sinonimo **hazard rate/force of mortality** $\mu(x)$, ovvero il tasso di mortalità ad una certa età x .
- È l'**event rate** al tempo t (la sua istantanea), ossia la probabilità di registrare un evento a t condizionata dalla sopravvivenza fino a t :

$$\begin{aligned}\lambda(t) &= P(T \in (t, t + \Delta t) \mid T > t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t \mid T > t)}{\Delta t} \\ &= \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t S(t)} = \frac{f(t)}{S(t)} = \frac{S'(t)}{S(t)}\end{aligned}$$

Hazard Function: proprietà

Ha due uniche proprietà, qualsiasi funzione che le rispetti entrambe è una possibile hazard function.

Una funzione $h(x)$ è una hazard function sse :

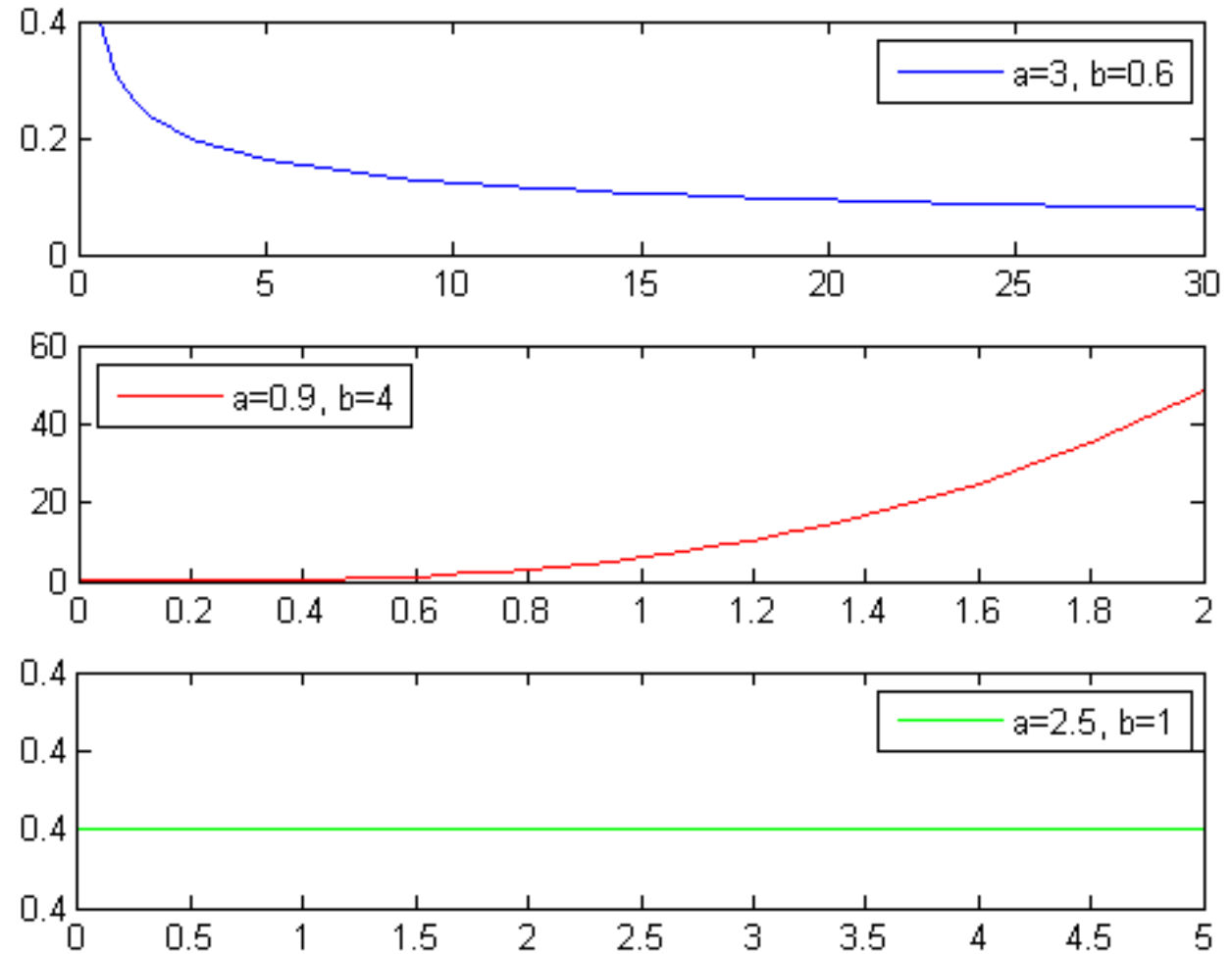
$$1) h(x) \geq 0 \quad \forall x \geq 0$$

$$2) \int_0^{\infty} h(x) dx = \infty$$

- Non ha quindi limitazioni come monotonicità, continuità, ...

Hazard Function

- Ci sono diversi tipi di hazard function :
 - Hazard **decrescenti**
 - Hazard **crescenti**
 - Hazard **costanti**
- Nel caso di modelli parametrici la forma della funzione sarà data da uno o più parametri di forma (*shape parameters*).

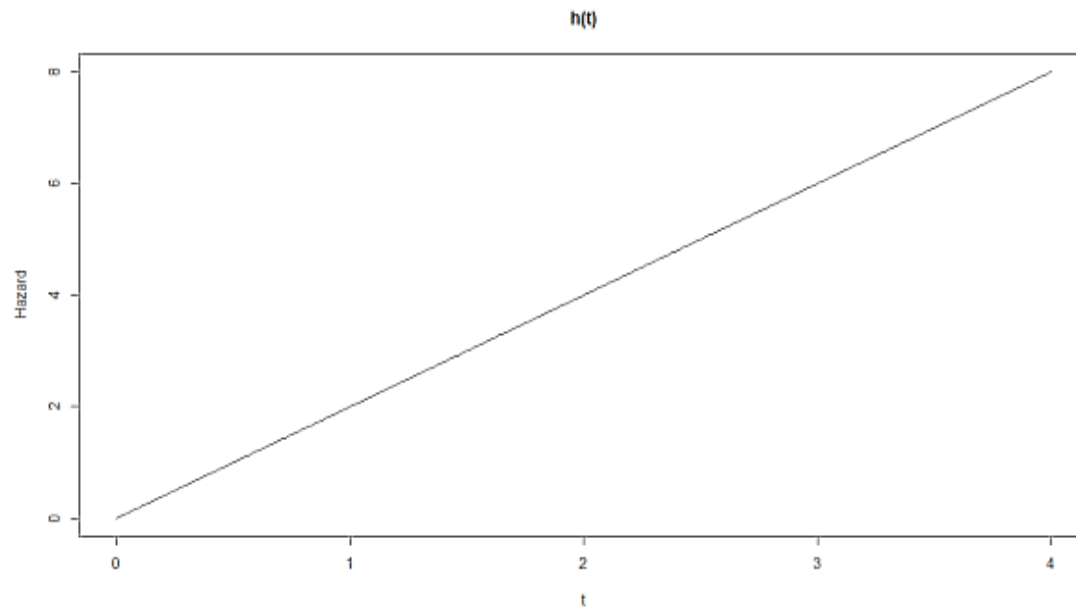


Cumulative Hazard Function $\Lambda(t)$

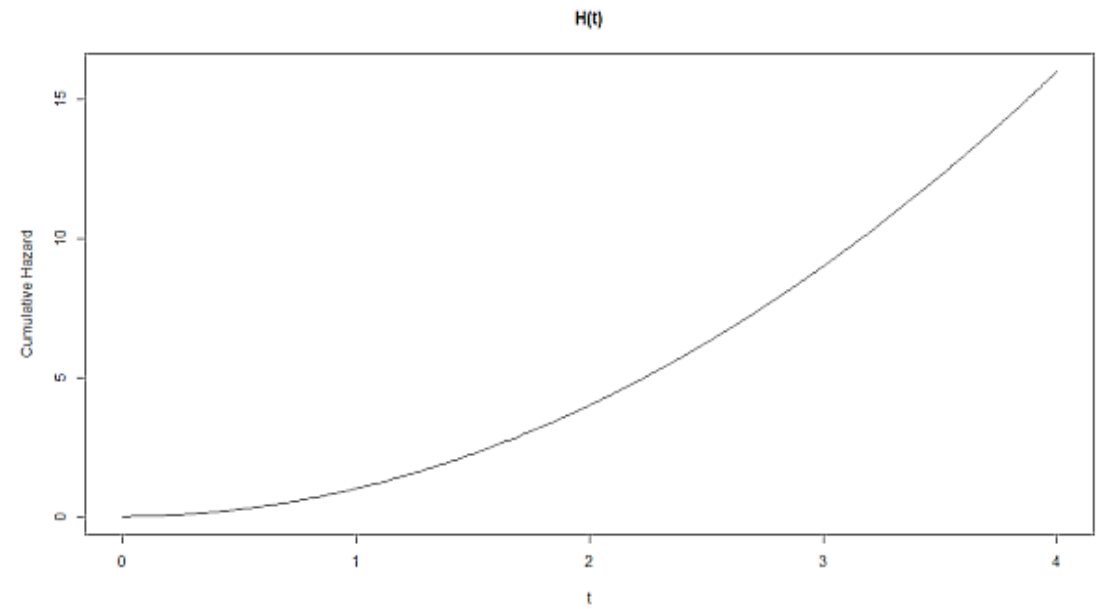
- Rappresenta l'accumulo dei rischi (*hazards*)
 - Definita come: $\Lambda(t) = -\log S(t)$
 - $S(t) = \exp(-\Lambda(t))$
 - $\frac{d}{dt}\Lambda(t) = -\frac{S'(t)}{S(t)} = \lambda(t)$
- Può essere vista come la **somma dei rischi** nel tempo che intercorre fra 0(nascita) e t .
 - $\Lambda(t) = \int_0^t \lambda(u)du$
- Per $t \rightarrow \infty$, supponendo $S(t) \rightarrow 0$, $\Lambda(t)$ cresce senza limiti.
 - Dato che $\Lambda(t)$ diverge $\lambda(t)$ non decresce rapidamente.

Hazard Function & Cumulative Hazard Function

- Hazard function



- Cumulative hazard function
ricavata dalla hazard function



Non-Parametric Survival Models

- Diventano di fondamentale importanza quando:
 - è impossibile ottenere un modello parametrico della situazione che si sta studiando.
 - si preferisce non sfruttare nessun parametro.
 - non è possibile conoscere la distribuzione di T .
 - Questo caso avviene spesso ed è quello che motiva il largo impiego di modelli non parametrici
- In questo caso si ricorre ad un modello non parametrico utilizzando degli **stimatori**. I più conosciuti sono:
 - *Kaplan-Meier* → utilizzato per stimare la **funzione di sopravvivenza**
 - *Nelson-Aalen* → utilizzato per stimare la **hazard function cumulativa**

Stimatore di Kaplan-Meier

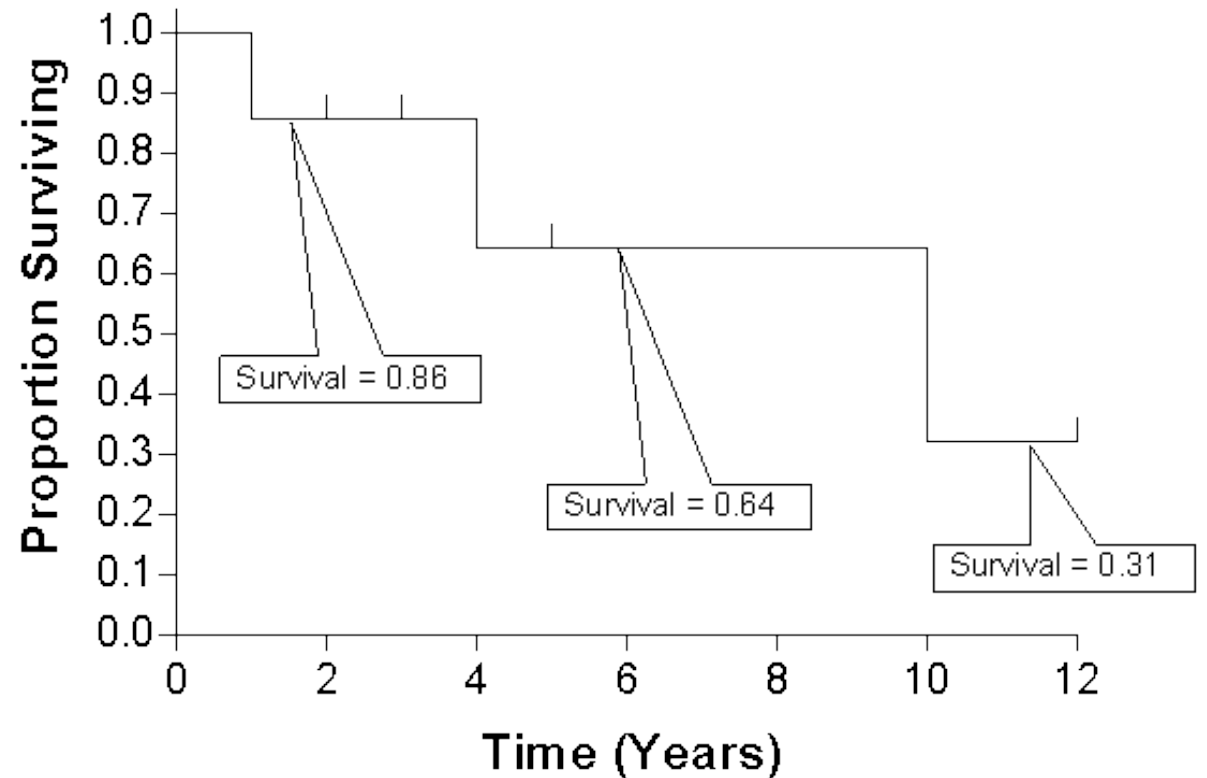
- È un **indicatore statistico** non-parametrico utilizzato per stimare la **survival function** dai **lifetime data** (dati di sopravvivenza) disponibili.
- È noto anche come **stimatore del prodotto limite**.
- È lo stimatore di $S(t)$ largamente più utilizzato.
- Considera dati non censurati o right-censored.
- Si utilizza per:
 - stimare quale porzione sopravviverà nel tempo dopo un determinato momento o dal momento in cui si hanno i dati
 - stimare l'efficacia di un trattamento o di una terapia
 - esaminare tassi di ricovero
 - ...

Kaplan Meier: formulazione

- $t_1, t_2, t_3, \dots, t_N \rightarrow$ tempi osservati fino alla morte degli N individui del campione di riferimento (o alla censura del dato). $\forall t_i$ corrispondono:
 - $n_i \rightarrow$ numero di individui a rischio prima di t_i
 - $d_i \rightarrow$ numero di morti a tempo t_i
- $\hat{S}(t) \rightarrow$ stima non-parametrica della massima verosimiglianza di $S(t)$
 - Definizione continua a sinistra:
$$\hat{S}(t) = \prod_{t_i < t} \frac{n_i - d_i}{n_i}$$
 - Definizione continua a destra equivalente:
(compatibile con una stima continua di $F(t)$)
$$\hat{S}(t) = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i}$$
- Rappresenta una funzione a gradini che ad ogni tempo t_i varia discretamente (con un gradino) di un fattore $\frac{n_i - d_i}{n_i}$

Kaplan Meier: diagramma

- È un diagramma composto da una serie di **gradini** di ampiezza decrescente.
 - Ascissa : tempo
 - Ordinata : **probabilità stimata** di sopravvivere
 - su campioni abbastanza ampi è una buona approssimazione di $S(t)$
 - La precisione aumenta più il campione è numeroso



Kaplan-Meier: varianza

- Kaplan-Meier è una statistica e la sua varianza può essere stimata tramite diversi metodi.
- Il metodo utilizzato principalmente per stimarla è la **formula di Greenwood**.

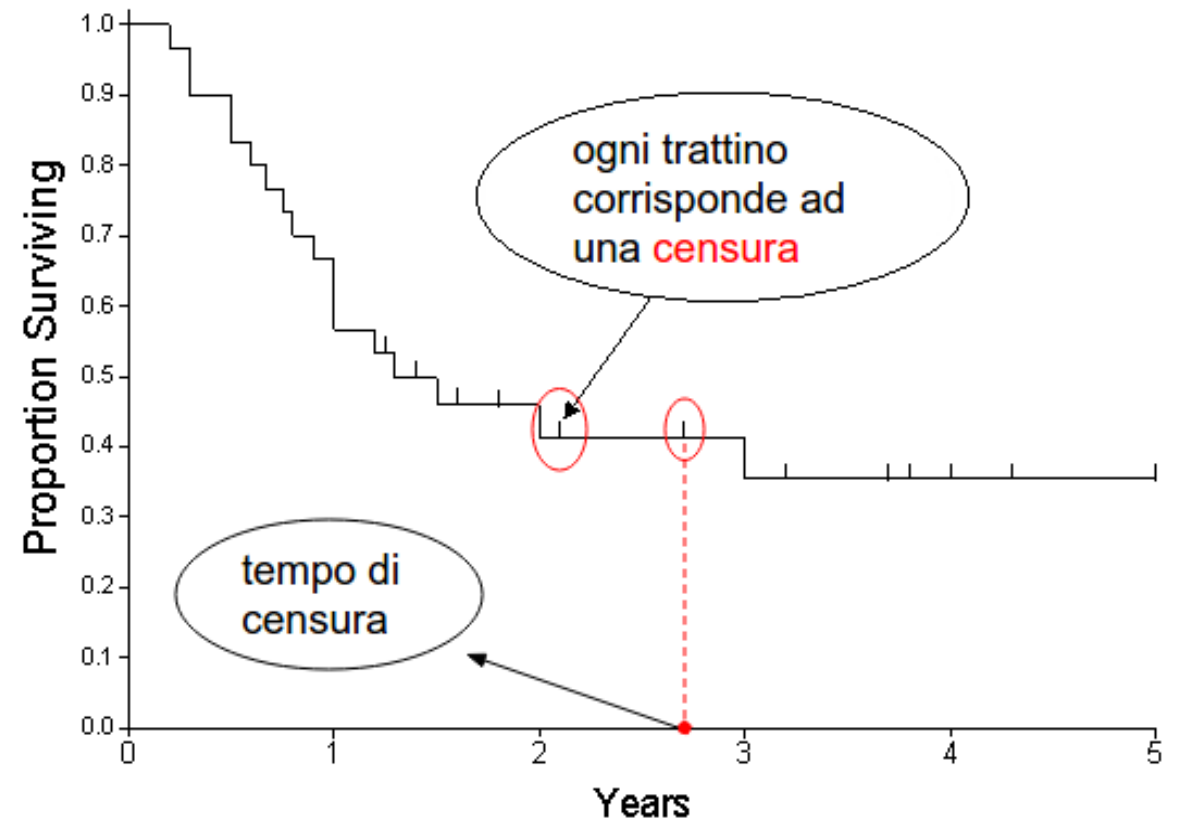
- Varianza:
$$\widehat{Var}(\hat{S}(t)) = \hat{S}(t)^2 \sum_{t_i < t} \frac{d_i}{n_i(n_i - d_i)}$$

- Assumendo la stima $\hat{S}(t)$ come distribuita normalmente, l'**intervallo di confidenza** per un test con coefficiente di confidenza $1-\alpha$ è :

- $$\hat{S}(t) \pm z_{1-\frac{\alpha}{2}} \sqrt{\widehat{Var}(\hat{S}(t))}$$

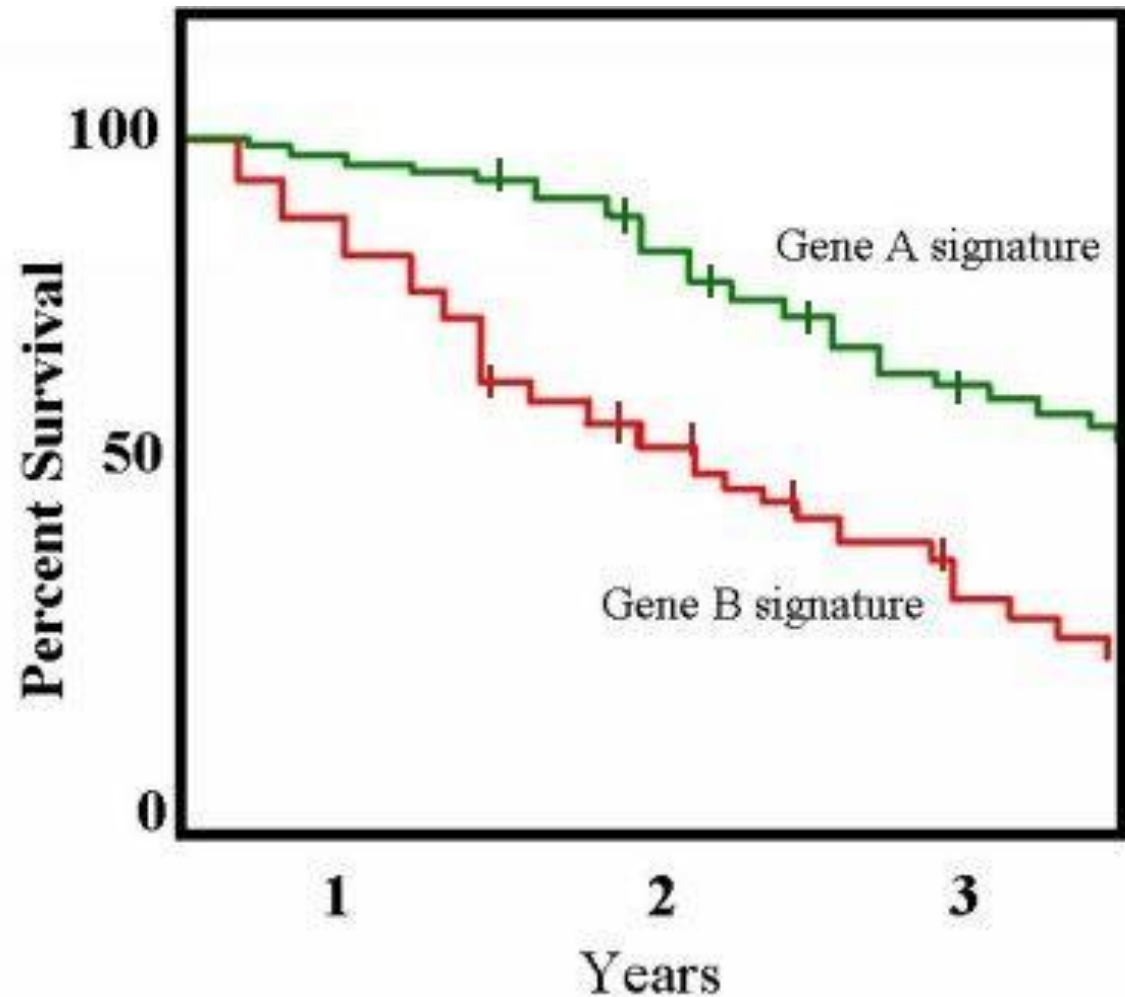
Kaplan-Meier: vantaggi

- Tiene conto senza ulteriori considerazioni anche dei dati **right-censored**
 - Si aggiunge un trattino verticale nel grafico per indicare la censura (al tempo in cui il dato viene censurato)
 - I dati censurati verranno quindi considerati fino a che si avranno informazioni, partecipando attivamente alla stima.
 - Vengono considerati come a **rischio** fino al tempo di censura.



Kaplan-Meier: vantaggi

- Permette di **raggruppare** i soggetti a disposizione secondo caratteristiche comuni.
- In questo modo si possono confrontare le diverse stime di sopravvivenza di diverse categorie dello stesso campione.



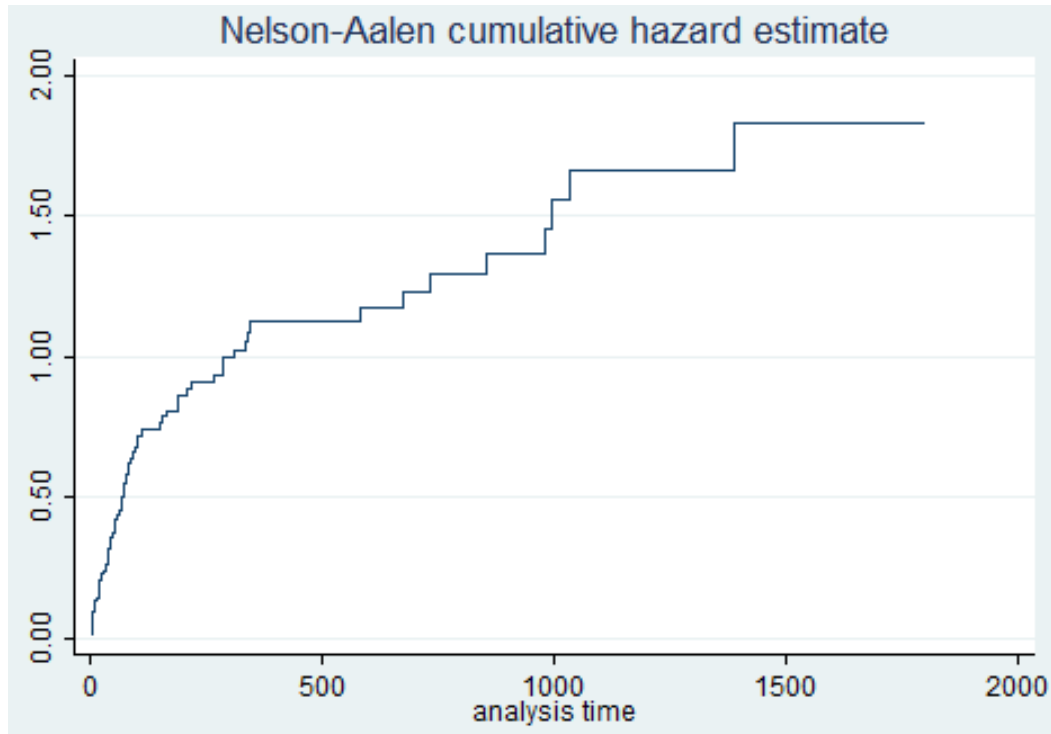
Nelson-Aalen

- È lo stimatore non parametrico della **cumulative hazard function**.
- Come Kaplan-Meier è utilizzabile anche con dati incompleti o right-censored.
- Utilizza gli stessi termini descritti per Kaplan-Meier.
- È così definito:

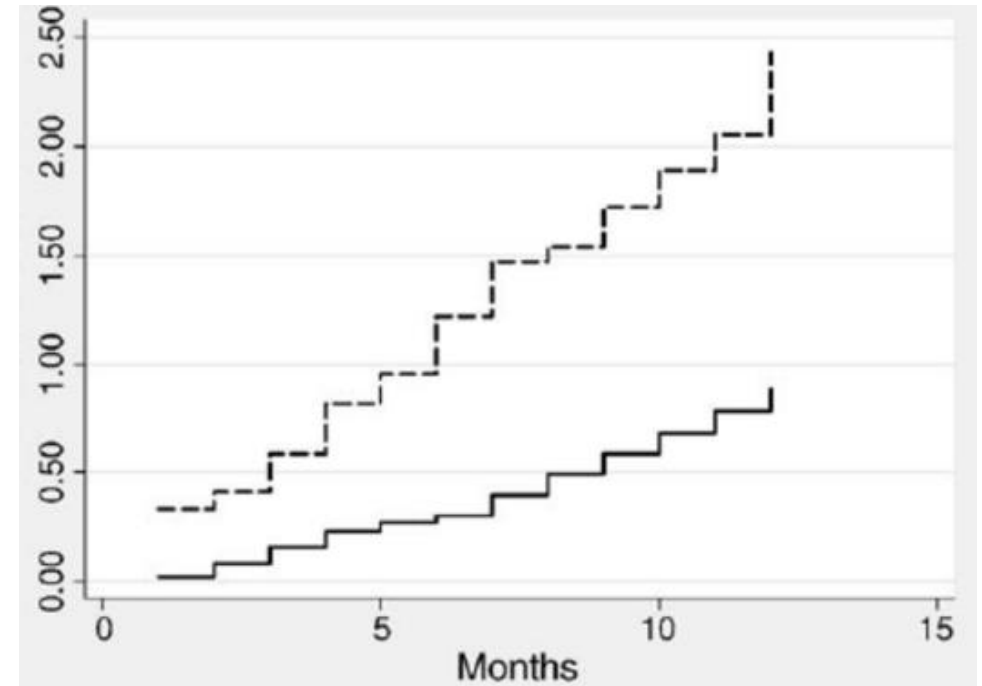
$$\Lambda(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}$$

Nelson-Aalen

- Esempio di stima di Nelson-Aalen per la cumulative hazard function.



- Confronto sul rischio cumulativo di schizofrenia fra consumatori (---) e non consumatori di Cannabis (—).



Log Rank Test

- È un **test d'ipotesi non parametrico** che confronta le distribuzioni di sopravvivenza di due campioni.
- Compara le stime delle hazard functions di due gruppi ad ogni event time osservato.
- Efficace anche su dati censurati.
- Utile per confrontare:
 - la stima di sopravvivenza di due gruppi
 - la reazione ad un trattamento di due gruppi con caratteristiche distinte
 - la reazione ad un determinato trattamento o alla somministrazione di droghe nello stesso gruppo (considerandolo sia prima che dopo)
 - ...

Log Rank Test & Kaplan Meier

- Il Log Rank Test è un **test chi-quadro** su campioni di grosse dimensioni.
- Confronta le stime di sopravvivenza di due campioni(e le conseguenti stime della hazard function) ottenute tramite Kaplan-Meier.
 - Utilizza una statistica che fornisce un resoconto del confronto delle due curve di Kaplan-Meier osservate.
- Si basa sul numero di **eventi osservati** e quello di **eventi stimati** ad ogni event time.
 - Si sommano per ottenere un resoconto totale di tutti i punti in cui c'è un evento.

Log Rank Test: Dati

- $j = 1, 2, \dots, J$: tempi in cui si sono osservati eventi (considerando entrambi i gruppi) $\rightarrow \forall j$:
 - N_{1j} e N_{2j} : numero di soggetti **a rischio** al tempo j (non hanno ancora riscontrato un evento fino a j) nei due gruppi
 - $N_j = N_{1j} + N_{2j}$ numero totale di soggetti a rischio
 - O_{1j} e O_{2j} : numero di eventi osservati a j nei due gruppi
 - $O_j = O_{1j} + O_{2j}$ numero totale di eventi osservati

Log Rank Test: Formulazione

- **Ipotesi nulla H_0** : *"i due gruppi hanno le stesse distribuzioni di sopravvivenza e di hazards"*
- O_{1j} distribuita secondo un **ipergeometrica** $O_{1j} \sim \mathcal{H}(N_j, N_{1j}, O_j)$
 - Valore atteso : $E_{1j} = \frac{O_j}{N_j} N_{1j}$
 - Varianza : $V_j = O_j \frac{N_{1j}}{N_j} (1 - \frac{N_{1j}}{N_j})(N_j - O_j)$
 - L'ipergeometrica è una distribuzione che descrive il comportamento di una variabile aleatoria che conta da un insieme A (N_j) quanti sono nel sottoinsieme B (N_{1j}) dati r elementi distinti (O_j) estratti casualmente.

Log Rank Test: Statistica Z

- Il test si basa su una **statistica Z** che confronta ogni O_{1j} con il valore atteso E_{1j} sotto l'ipotesi H_0 .

$$Z = \frac{\sum_{j=1}^j (O_{1j} - E_{1j})}{\sqrt{\sum_{j=1}^j V_j}}$$

- La statistica Z può essere considerata come distribuita secondo:
 - Normale → **asymptotic distribution**
 - Normale Bivariata → **joint distribution**

Z: asymptotic distribution

- Se i due gruppi hanno la stessa funzione di sopravvivenza $S(t)$ allora Z risulta distribuita secondo una normale.
- Un test con one-sided level α rifiuterà H_0 se $Z > Z_\alpha$ con α quantile superiore della normale standard.
- Con
 - λ = hazard ratio
 - n = numero di soggetti
 - d = probabilità che un soggetto abbia registrato un evento
 - nd = numero atteso di eventi
 - soggetti presi casualmente dai due gruppi con una proporzione del 50%

$$Z \sim N\left(\log\lambda\sqrt{\frac{nd}{4}}, \frac{4(z_\alpha + z_{1-\alpha})}{d\log^2\lambda}\right)$$

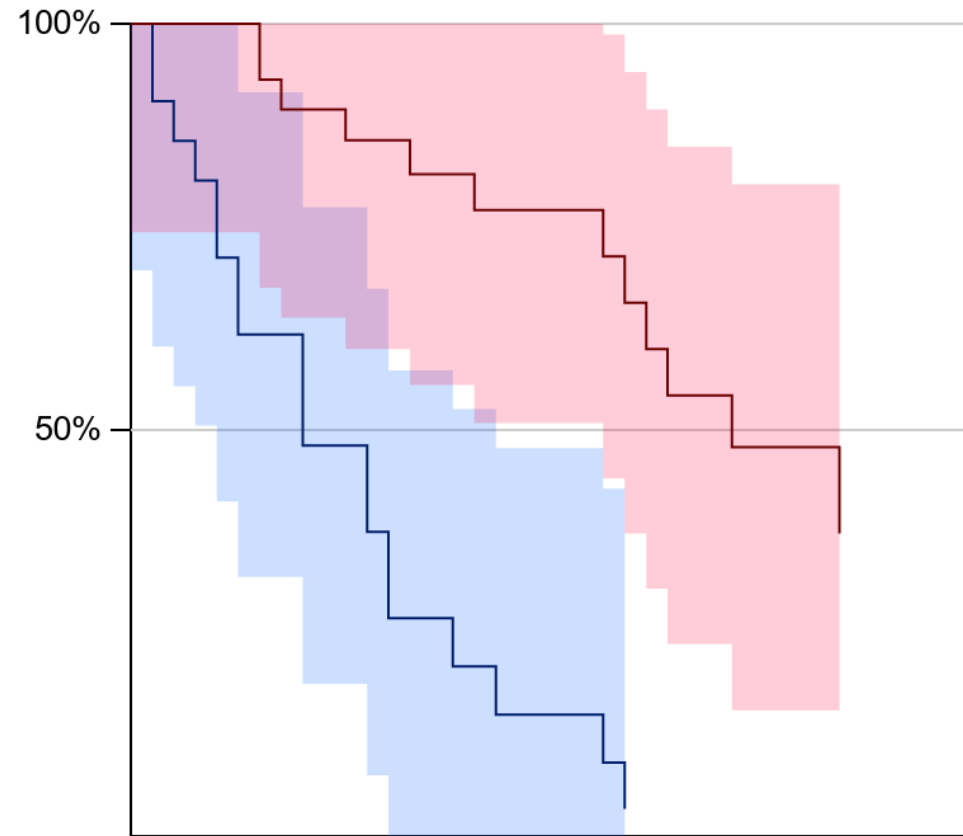
Z: joint distribution

- Si analizzano due tempi j diversi e se ne fornisce la distribuzione congiunta utilizzando una **normale multivariata** (in questo caso normale **bivariata**)
- Dati
 - Z_1, Z_2 statistiche Z a due differenti tempi j (con Z_1 che precede Z_2)
 - λ proporzionale nei due gruppi
 - d_1, d_2 probabilità di un evento nei due tempi analizzati
 - Z_1 e Z_2 risultano distribuite secondo una normale bivariata di parametri $\mu = [E[Z_1], E[Z_2]]$ e $\Sigma = [\text{cov}[E_1, E_2]]$

$$\bullet Z \sim N_2\left(\left[\log\lambda\sqrt{\frac{nd_1}{4}}, \log\lambda\sqrt{\frac{nd_2}{4}}\right], \sqrt{\frac{d_1}{d_2}}\right)$$

Log Rank Test: esempio

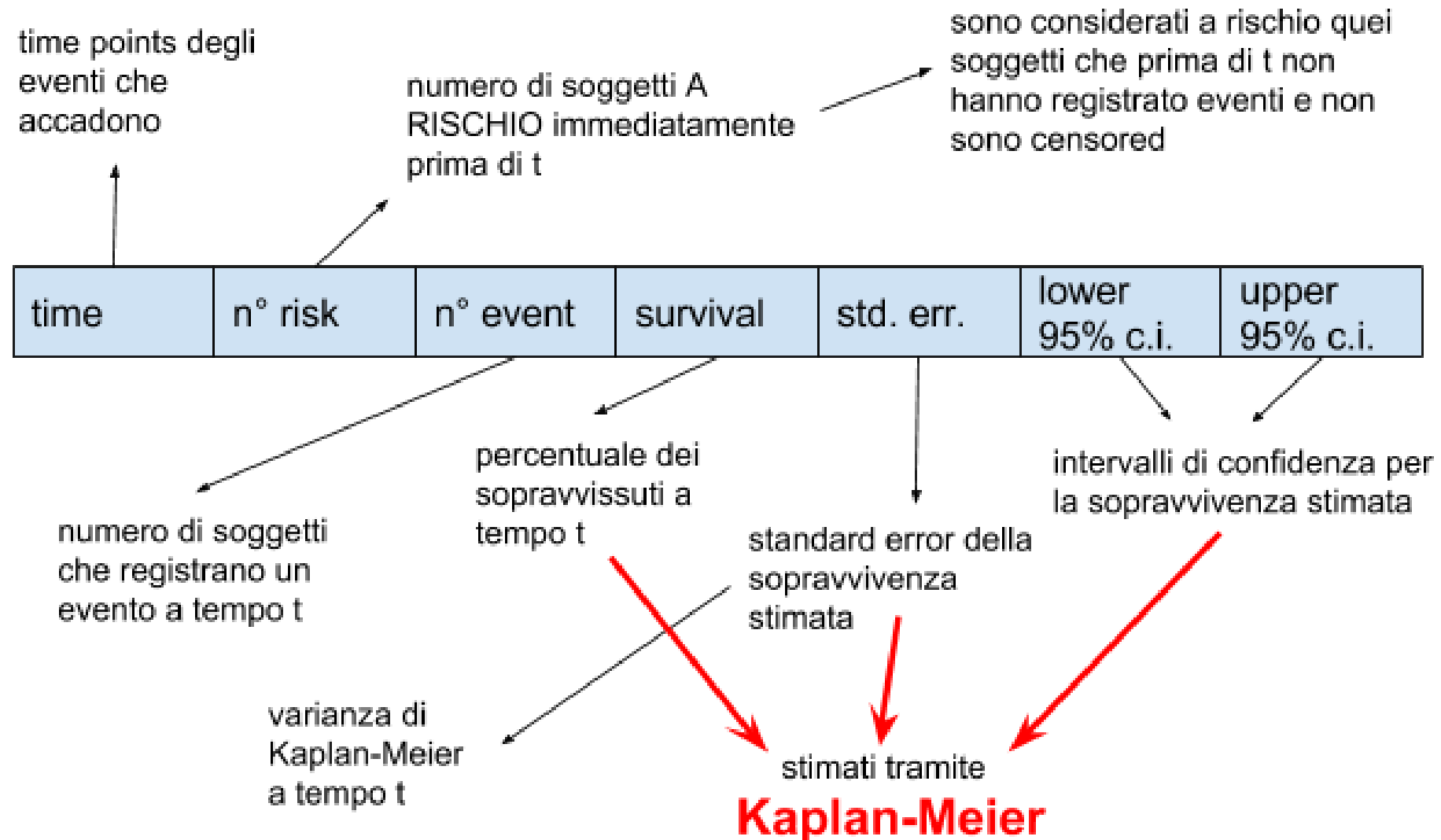
- Consideriamo due differenti curve ottenute da due differenti campioni.
 - **Blu** : ottenuta da un campione di 20 osservazioni di cui una censored.
 - **Rosso** : ottenuta da un campione di 28 osservazioni di cui 16 censored.
 - OSS: le bande colorate mostrano l'intervallo di confidenza (95%).
 - (Fonte: [link](#))
- Un **log rank test** sulle due curve produce esito negativo ($Z = 5.32$, $p < 0.001$) negando l'ipotesi nulla.
Le due curve non avranno quindi le stesse distribuzioni di sopravvivenza e di hazard.



Life Table & Kaplan Meier

- Una **life table**, detta anche **mortality table**, è una tabella che esprime i dati di sopravvivenza in termini di numero di eventi e porzione di sopravvissuti ad ogni unità temporale presente nei dati.
- Mostra per ogni tempo a disposizione la sopravvivenza degli individui di una certa popolazione.
- Utilizza le stime fornite da Kaplan Meier per ottenere la percentuale di sopravvissuti ad ogni tempo, standard error della sopravvivenza stimata e intervalli di confidenza.

Life Table



Life Table

- Esempio di life table ottenuta tramite R da dati relativi all'AML (*acute myeloid leukemia*) dove le ultime 4 colonne sono stimate tramite kaplan-meier.

time	<u>n.risk</u>	<u>n.event</u>	survival	<u>std.err</u>	lower 95% CI	upper 95% CI
5	23	2	0.913	0.0588	0.8049	1
8	21	2	0.8261	0.079	0.6848	0.996
9	19	1	0.7826	0.086	0.631	0.971
12	18	1	0.7391	0.0916	0.5798	0.942
13	17	1	0.6957	0.0959	0.5309	0.912
18	14	1	0.646	0.1011	0.4753	0.878
23	13	2	0.5466	0.1073	0.3721	0.803
27	11	1	0.4969	0.1084	0.324	0.762
30	9	1	0.4417	0.1095	0.2717	0.718
31	8	1	0.3865	0.1089	0.2225	0.671
33	7	1	0.3313	0.1064	0.1765	0.622
34	6	1	0.2761	0.102	0.1338	0.569
43	5	1	0.2208	0.0954	0.0947	0.515
45	4	1	0.1656	0.086	0.0598	0.458
48	2	1	0.0828	0.0727	0.0148	0.462

Modelli non-parametrici: limiti

- Il limite principale di questi modelli è dato dall'impossibilità di gestire situazioni in cui il rischio è influenzato dalle **covariate** (**variabili esplicative** osservabili dei soggetti)
 - **Covariate-adjusted regression**, situazioni in cui non sono osservabili né variabili di predizione né variabili di risposta, ma sono influenzate dagli effetti delle variabili esplicative osservabili che vengono viste come funzioni sconosciute che ne determinano il valore.
 - Un **vettore di covariate** è un vettore di **variabili di categoria**, ovvero un vettore di variabili binarie indicanti il possesso o meno di una determinata caratteristica per un soggetto. (es: genere, presenza di una mutazione, fumatore, ...)
- In questi casi sono preferibili :
 - *Parametric survival models*
 - *Proportional Hazards Model*
 - *Cox Model*

Survival Models: assunzioni sulle possibili variabili esplicative

- In un modello possono essere presenti delle **variabili esplicative** (**variabili indipendenti** in un modello di regressione) relative ai soggetti.
- Esse possono giocare un ruolo fondamentale perché **influenzano la lifetime** di un soggetto.
- L'influenza che possono avere dipende dalla **hazard function di base**.
- Sotto queste premesse un **modello di sopravvivenza** può essere fondamentalmente diviso in due parti :
 - **Hazard function** : descrive come varia il rischio per unità di tempo.
 - **Effect Parameters** : descrivono come varia il rischio in base alle variabili esplicative.
 - La loro gestione può differire in base all'assunzione del loro effetto sulla **hazard function** che si sceglie di usare.

Proportional Hazards vs Accelerated Failure Time

Un modello dovrà quindi assumere l'effetto delle variabili esplicative sul rischio come :

- Proporzionale alla hazard function di base.
 - **Proportional Hazards Model:**
 - In questi modelli l'effetto della variazione di una variabile esplicativa è moltiplicativo con la hazard function di base.
 - Possono essere sia parametrici che semiparametrici.
- Accelerante del tempo di realizzazione di un evento.
 - **Accelerated Failure Time:**
 - In questi modelli l'effetto della variazione di una variabile esplicativa aumenta o rallenta la velocità di realizzazione di un evento.

Parametric Survival Models

- Sono modelli basati su una distribuzione di probabilità parametrica, che consente di ottenere una stima di $S(t)$.
- Si assume quindi che il tempo di sopravvivenza (**survival time**) segua una distribuzione nota.
- Possono:
 - Non considerare la presenza di variabili esplicative.
 - Considerare la presenza di variabili esplicative e:
 - Assumere **PH** (proportional hazards).
 - Assumere **AFT** (accelerated failure time).

Parametric Survival Models: distribuzione parametrica di sopravvivenza

- Dalla definizione di $\Lambda(t)$ si può esprimere $S(t)$ come :
 - $S(t) = \exp(-\Lambda(t))$
 - Una **distribuzione parametrica di sopravvivenza** è una qualsiasi distribuzione per $t \in [0, \infty)$
- Si possono usare diverse distribuzioni come : Weibull, Exponential, Rayleigh, Log-Normal, Log-logistic, generalized Gamma, Gompertz, generalized F, inverse Gaussian, ...

Parametric Survival Distribution: terminologia dei parametri

Nelle distribuzioni utilizzate per modelli di sopravvivenza parametrici sono presenti tipologie di parametri ricorrenti:

- **λ : events rate/tasso degli eventi** =
$$\frac{\text{\textit{\#eventi}}}{\text{\textit{tempo totale di osservazione}}}$$
 - Tempo totale di osservazione = somma dei T_i (o del tempo di censura) di tutti i soggetti
 - **sopravvivenza media**: il reciproco dell'events rate.
- **Parametri di forma (*shape parameters*)**: parametri che modificano la forma delle curve ottenute, in particolare permettono di fare considerazioni sugli hazards.
 - **K** : parametro di scala che permette di modellare gli hazards come costanti, decrescenti o crescenti
- **Parametri di scala (*scale parameters*)**: modificano la 'scala' della distribuzione, ovvero cambiano quanto sia sparsa la densità di probabilità.
 - Ogni volta che è presente λ come parametro è una semplificazione per il significato che ha, in realtà si dovrebbe usare il corrispondente parametro di scala, ovvero il tempo medio di sopravvivenza **$1/\lambda$**

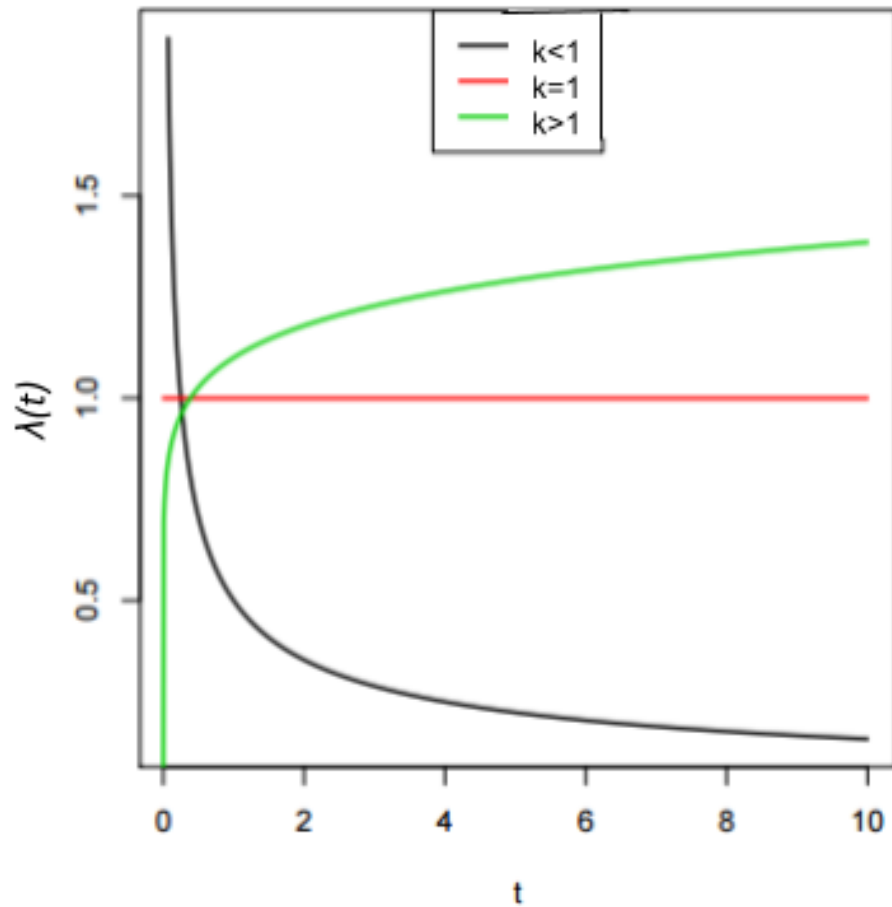
Weibull Model

- Utilizza la distribuzione di Weibull $T \sim W(\lambda, k)$
- I suoi parametri sono
 - λ : $1/\lambda$ *scale parameter*
 - k : *shape parameter*
 - $0 < k < 1 \rightarrow$ hazards decrescenti
 - $k = 1 \rightarrow$ hazard costanti \rightarrow **esponenziale**
 - $k > 1 \rightarrow$ hazard crescenti

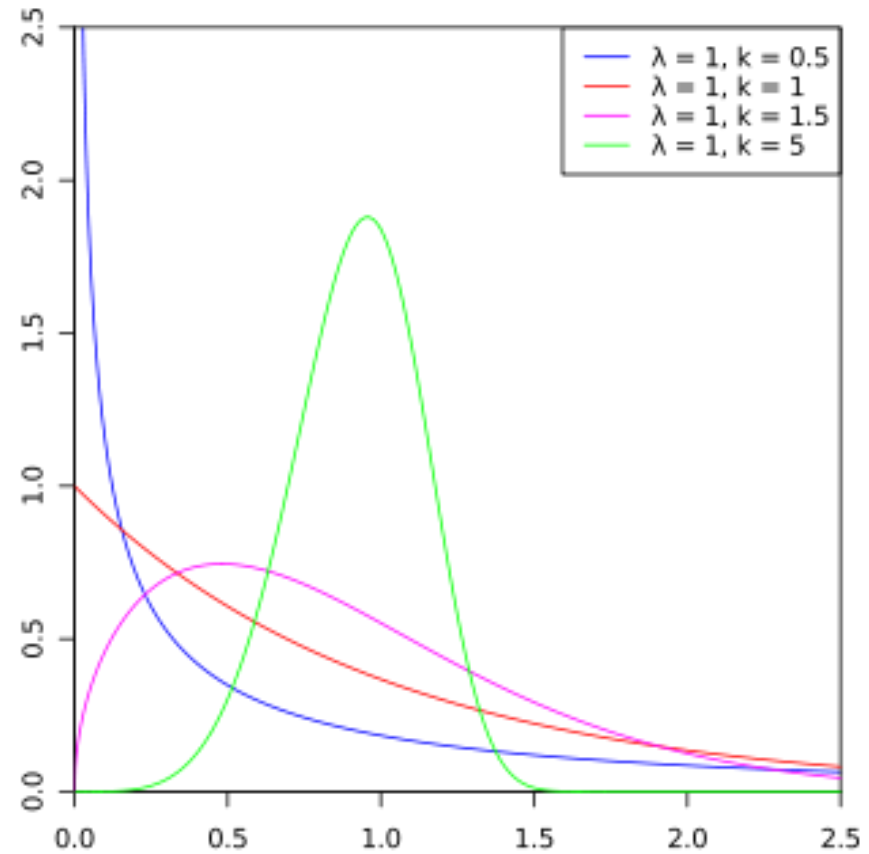
- $S(t) = e^{-\lambda t^k}$ $f(t) = \frac{-d}{dt} S(t) = k\lambda t^{k-1} e^{-\lambda t^k}$
 $\lambda(t) = k\lambda t^{k-1}$ $\Lambda(t) = \int_0^t \lambda(u) du = \lambda t^k$

Weibull Model

- $\lambda(t)$ in funzione di k

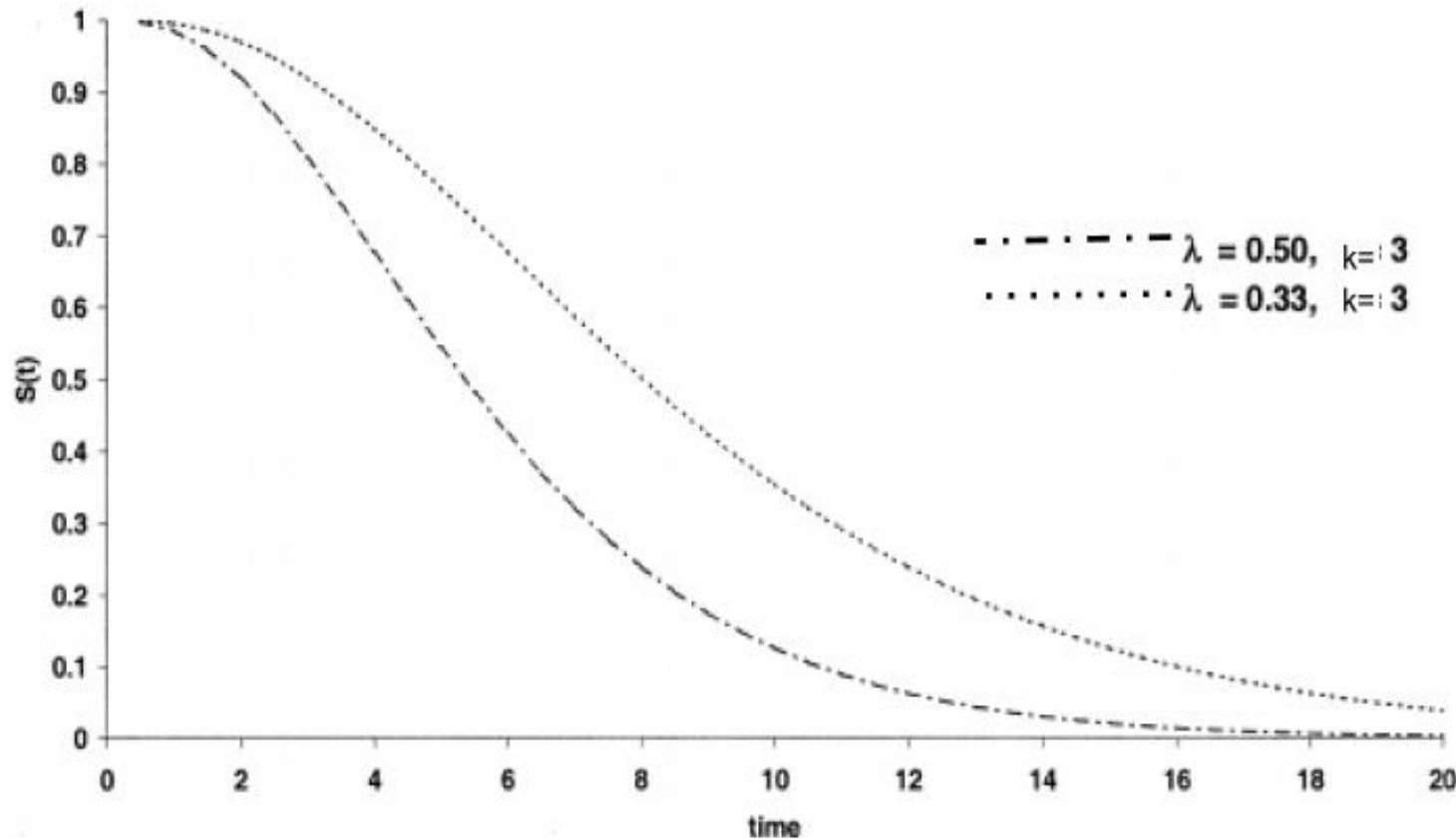


- $f(t)$ in funzione di k



Weibull Model

- $S(t)$ con tassi degli eventi diversi, più è grande λ più velocemente scende la curva di sopravvivenza



Exponential Model

- Utilizza la distribuzione esponenziale, ovvero un caso speciale della distribuzione di Weibull, in cui $k=1$.
 - $T \sim E(\lambda)$
- Ha un unico parametro λ indicante il 'rate' parameter, ovvero il **tasso degli eventi** che resta costante nel tempo (**assenza di memoria** della distribuzione esponenziale)

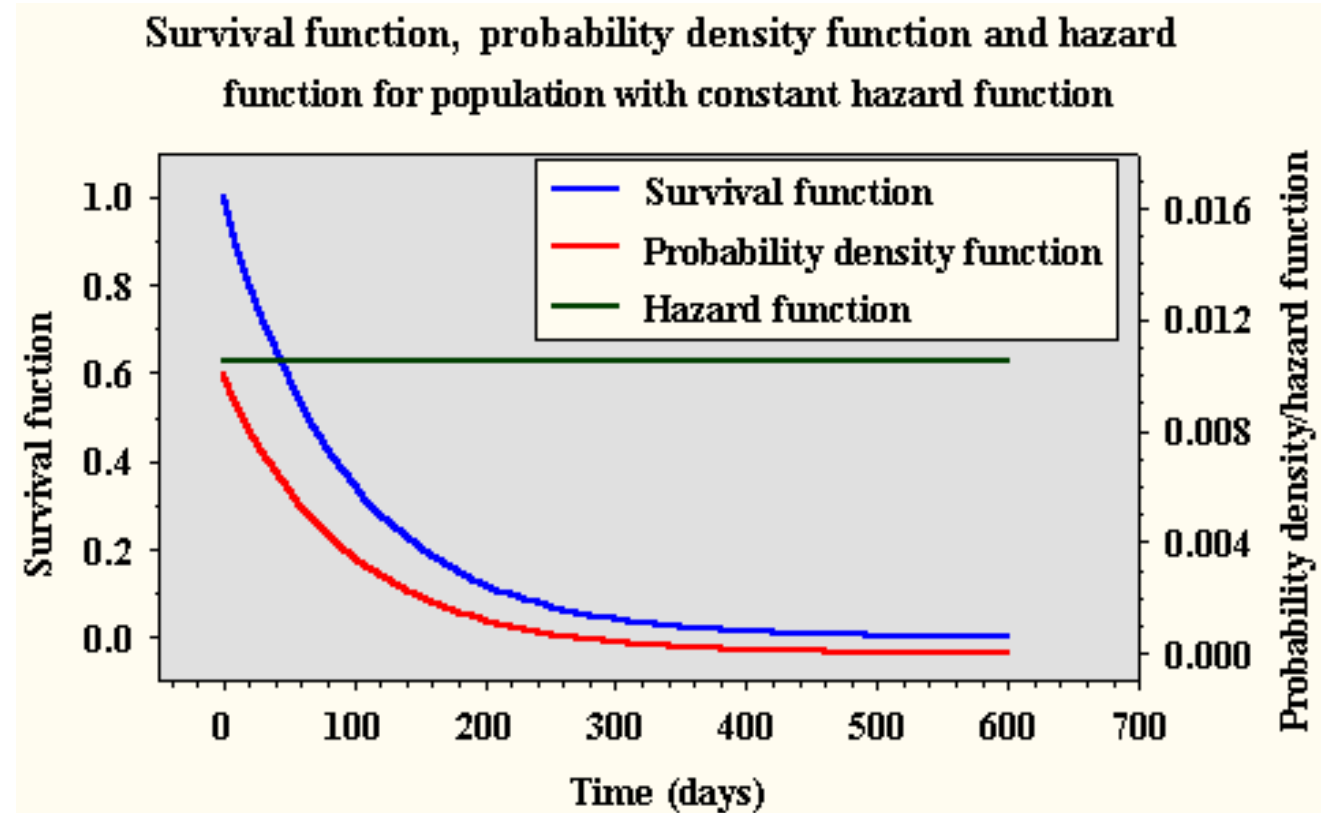
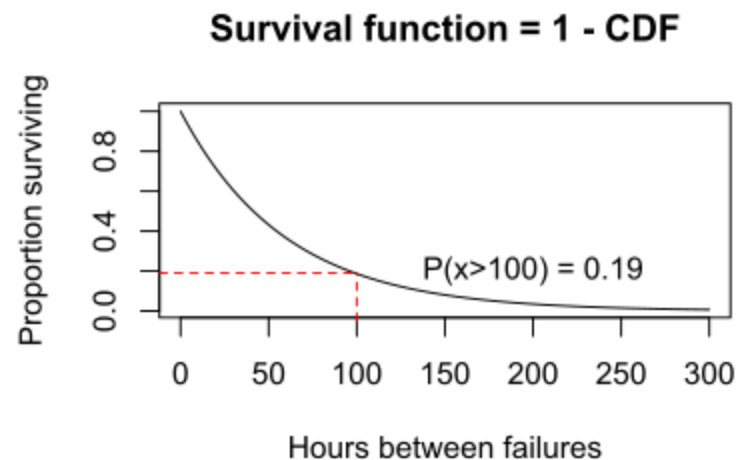
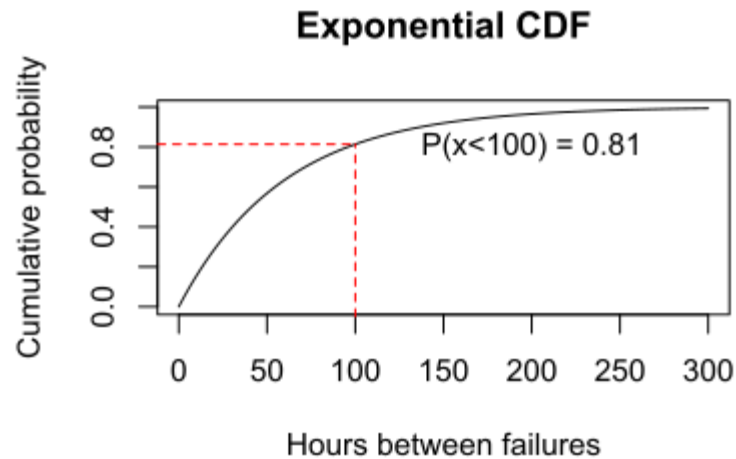
$$\bullet \quad f(t) = \lambda e^{-\lambda t}$$

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

$$\lambda(t) = \frac{f(t)}{S(t)} = \lambda \quad \text{hazard costanti !}$$

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

Exponential Model



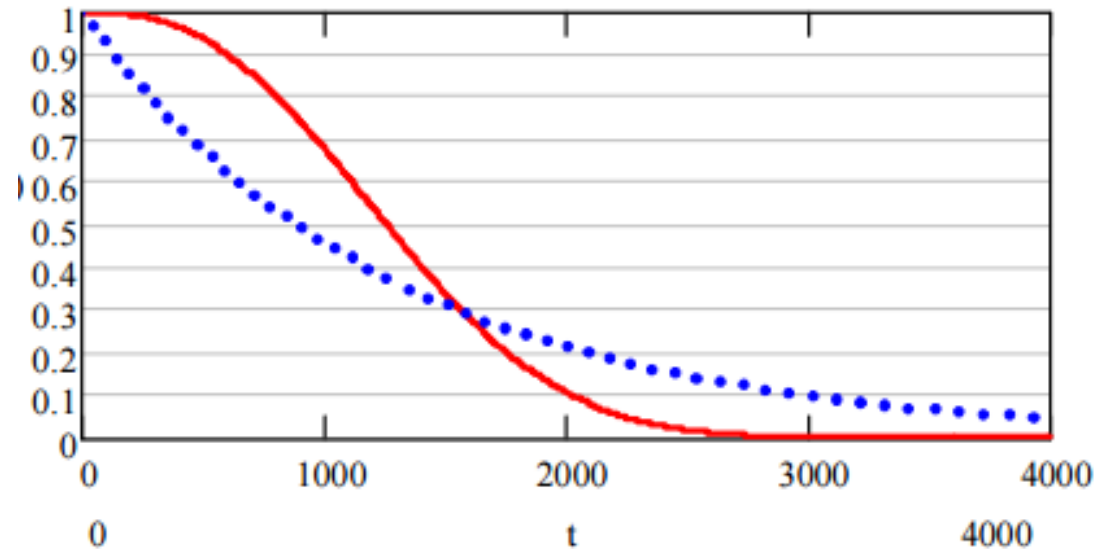
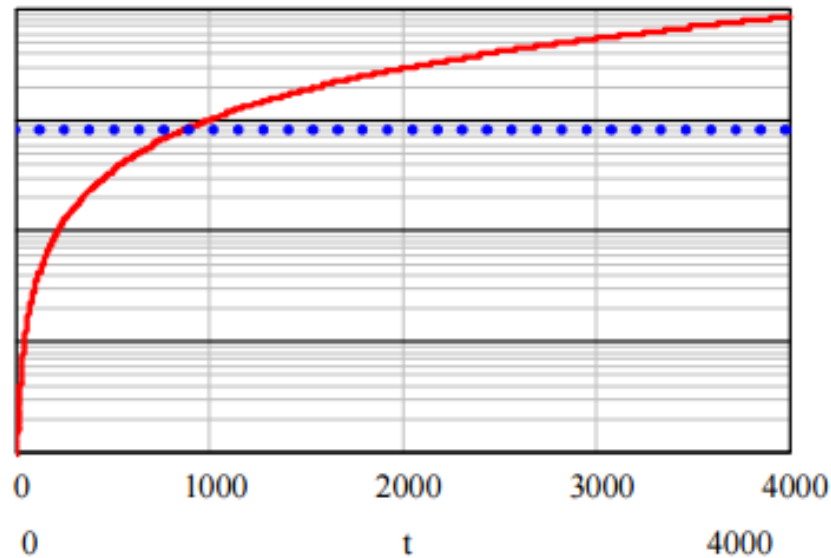
OSS: CDF = $F(t)$

Exponential vs Weibull

- La distribuzione **Esponenziale** presuppone il rischio di registrare un evento **costante** nel tempo, infatti utilizza il solo parametro λ , indicante il tasso degli eventi.
 - Spesso, soprattutto biologicamente, non è la considerazione più appropriata perché l'età influisce sul tasso di mortalità.
 - Il modello esponenziale non risulta quindi un buon modello su dati di sopravvivenza perché un soggetto avrebbe la stessa probabilità di morire a qualsiasi età.
- La distribuzione di **Weibull** invece considera un ulteriore parametro k , che permette di modellare il rischio come crescente o decrescente al passare del tempo.

Exponential vs Weibull

- Confronto tra le **Hazard Function** dello stesso campione.
(---exponential, —weibull)
- Confronto tra le **Survival Function** dello stesso campione.
(---exponential, —weibull)

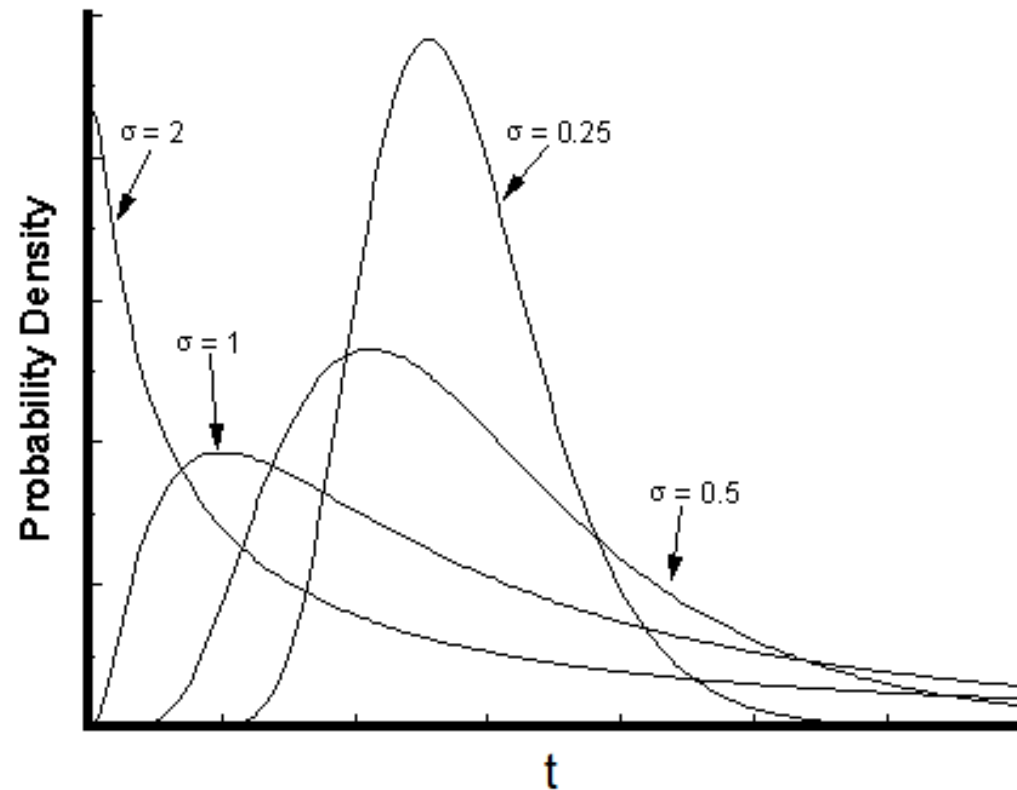


Log-Normal Model

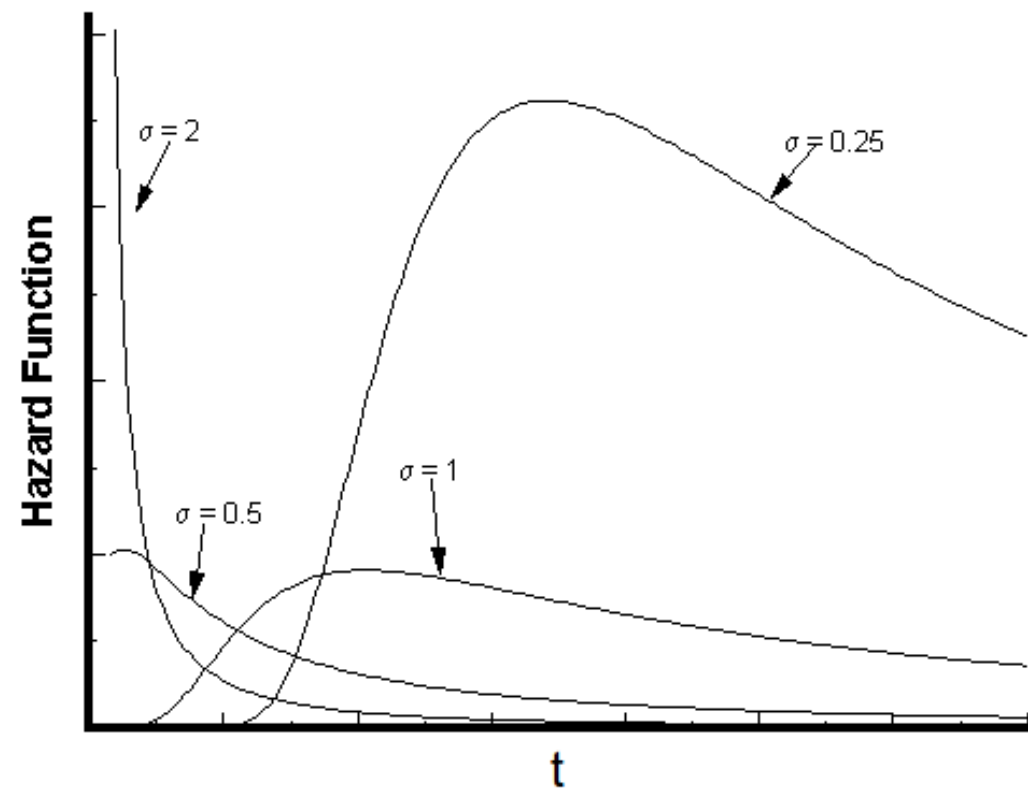
- È un modello che considera T distribuito secondo una log-normal, quindi $\log(T)$ sarà distribuito normalmente.
 - $T \sim LN(\mu, \sigma^2) \rightarrow \log T \sim N(\mu, \sigma^2)$
 - Ha come parametri i parametri della normale, dove però μ può essere un vettore modellato in funzione del vettore di variabili esplicative, così come σ^2 come matrice di correlazione in funzione del vettore di covariate. (stesso procedimento che si vedrà per riparametrizzare λ)
 - $S(t) = 1 - F(t) = 1 - \Phi\left(\frac{\ln(t) - \mu}{\sigma}\right)$
 - Oss: Φ è la funzione di ripartizione della normale
- È utilizzabile quando la variabile aleatoria che segue questa distribuzione è la somma di un largo numero di variabili indipendenti e identicamente distribuite, ovvero presupponendo che l'intero campione abbia la stessa $S(t)$.

Log-Normal Model

- $f(t)$ in funzione di σ

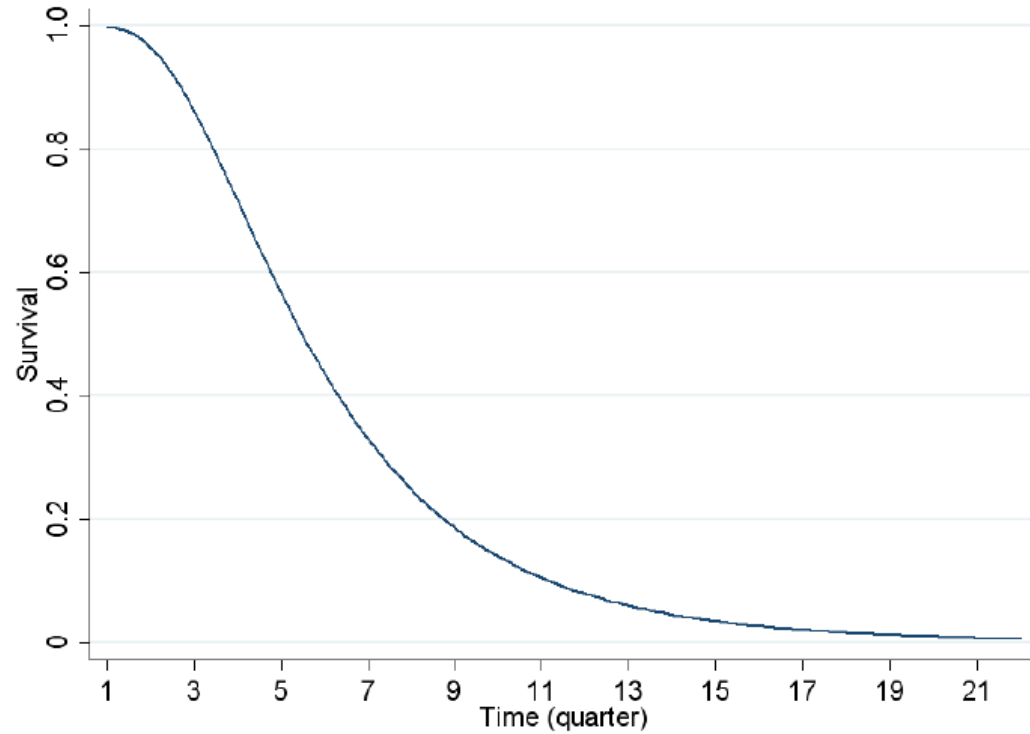


- $\lambda(t)$ in funzione di σ

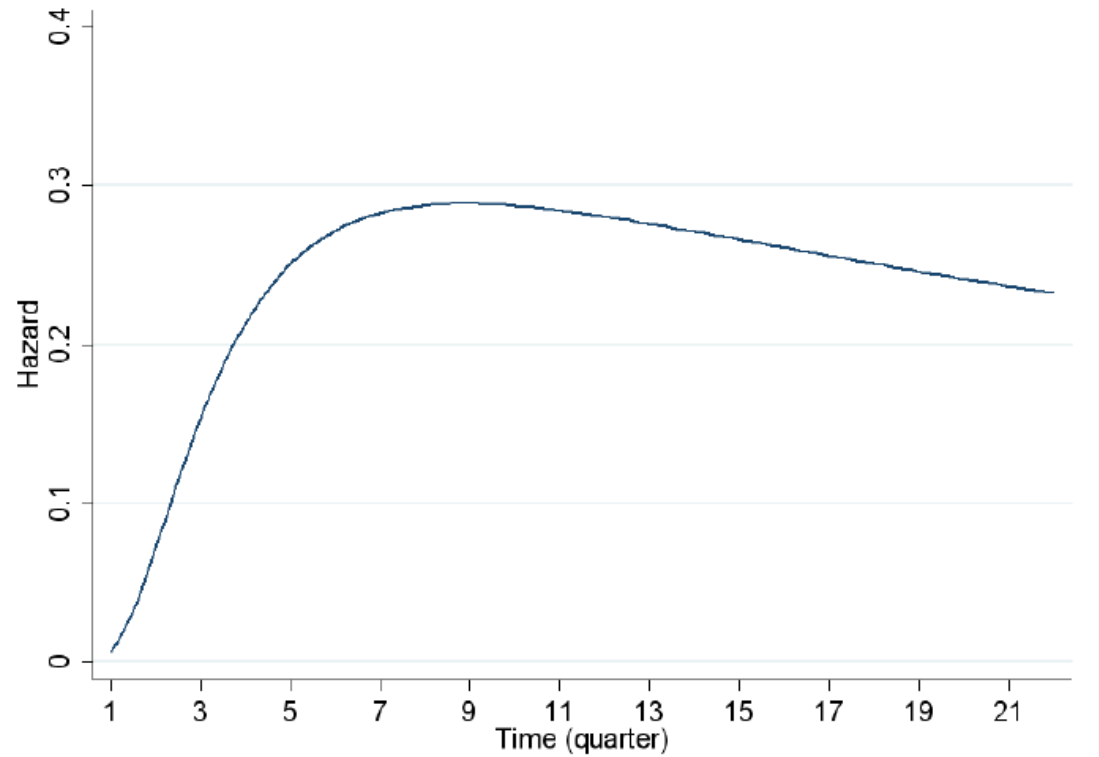


Log-Normal Model

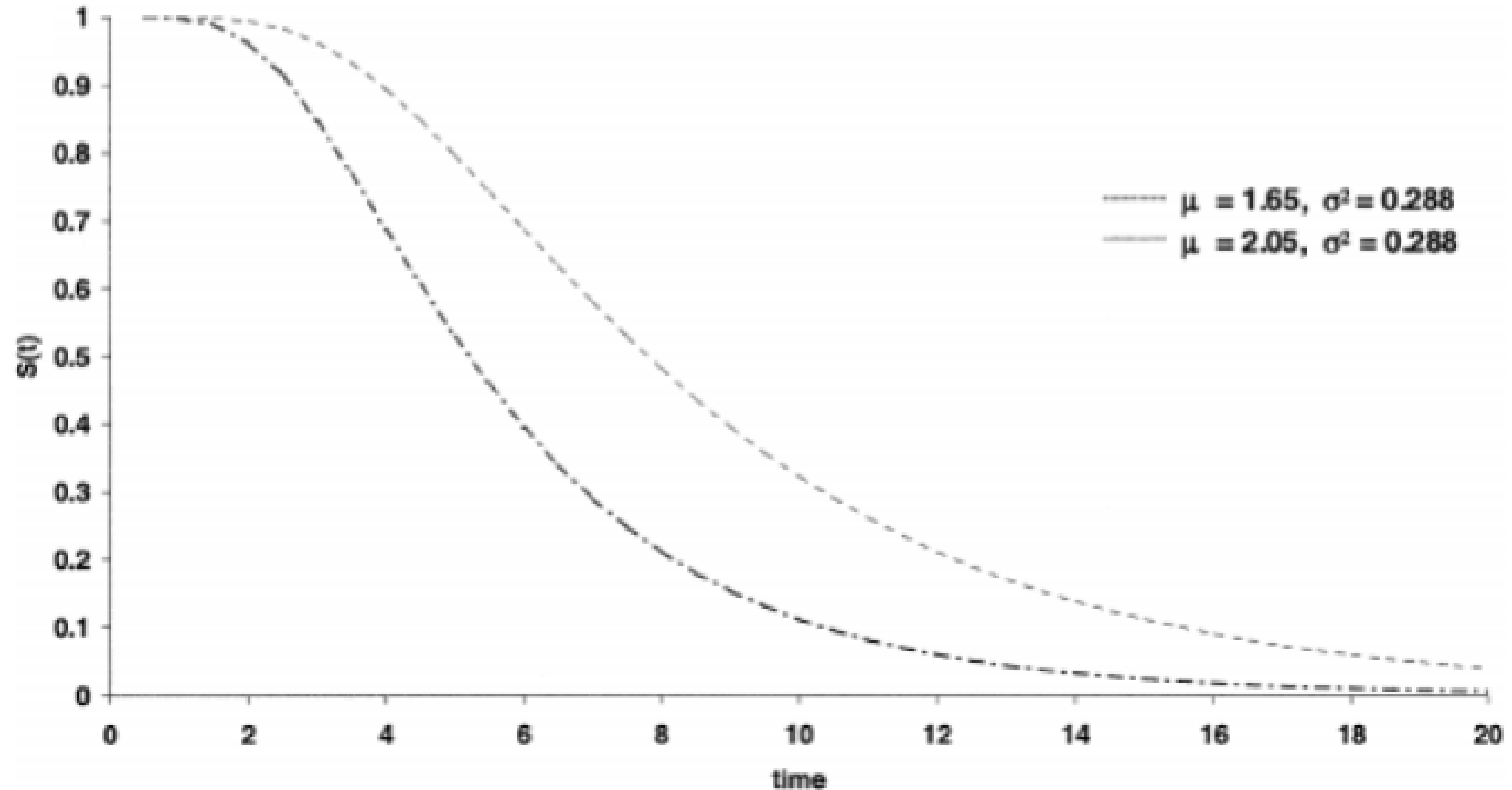
- $S(t)$



- $\lambda(t)$



Log-Normal Model



Log-Logistic Model

- Come il modello log-normal considera $\log(T)$ come distribuito logisticamente (*logistic distribution*) e quindi T distribuito secondo una Log-Logistic
 - $T \sim LL(\lambda, k)$
 - Ha come parametri λ (quindi $1/\lambda$ come *scale parameter*) e k come *shape parameter* che permette di considerare hazards come crescenti, costanti o decrescenti.

$$f(x) = \frac{k\lambda(\lambda t)^{k-1}}{(1 + (\lambda t)^k)^{-2}}$$

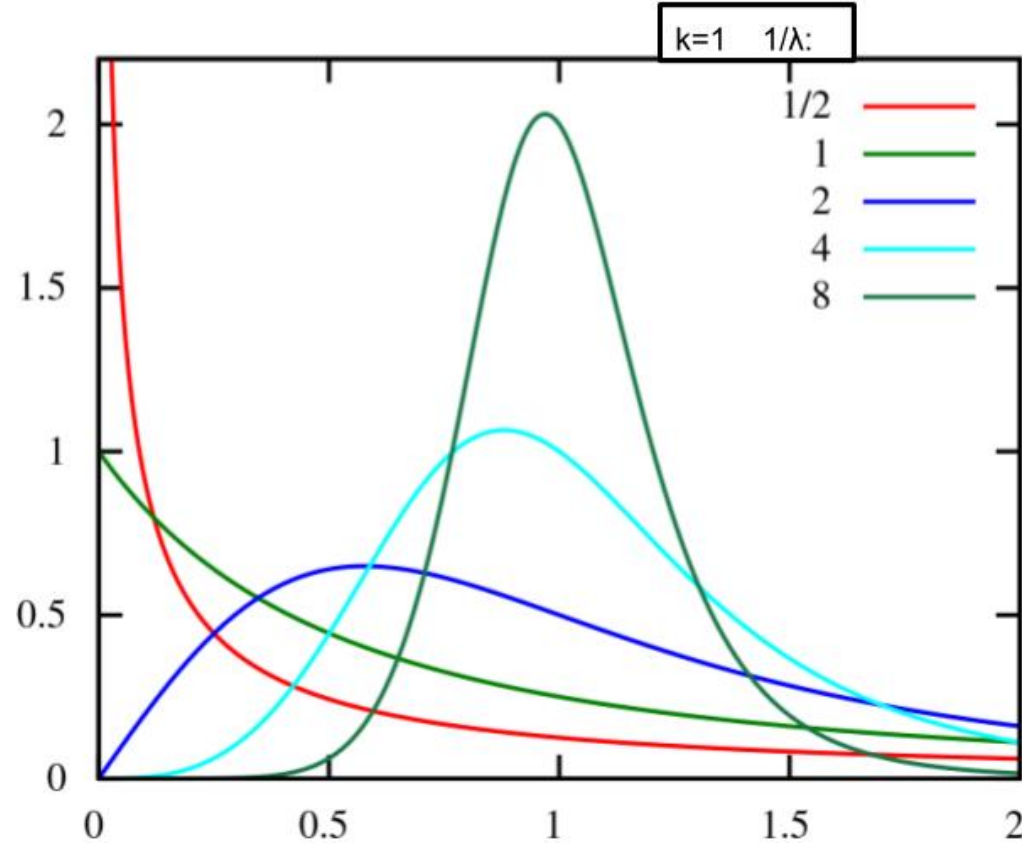
$$\lambda(x) = \frac{k\lambda(\lambda t)^{k-1}}{1 + (\lambda t)^k}$$

$$S(t) = 1 - F(t) = \frac{1}{1 + (\lambda t)^k}$$

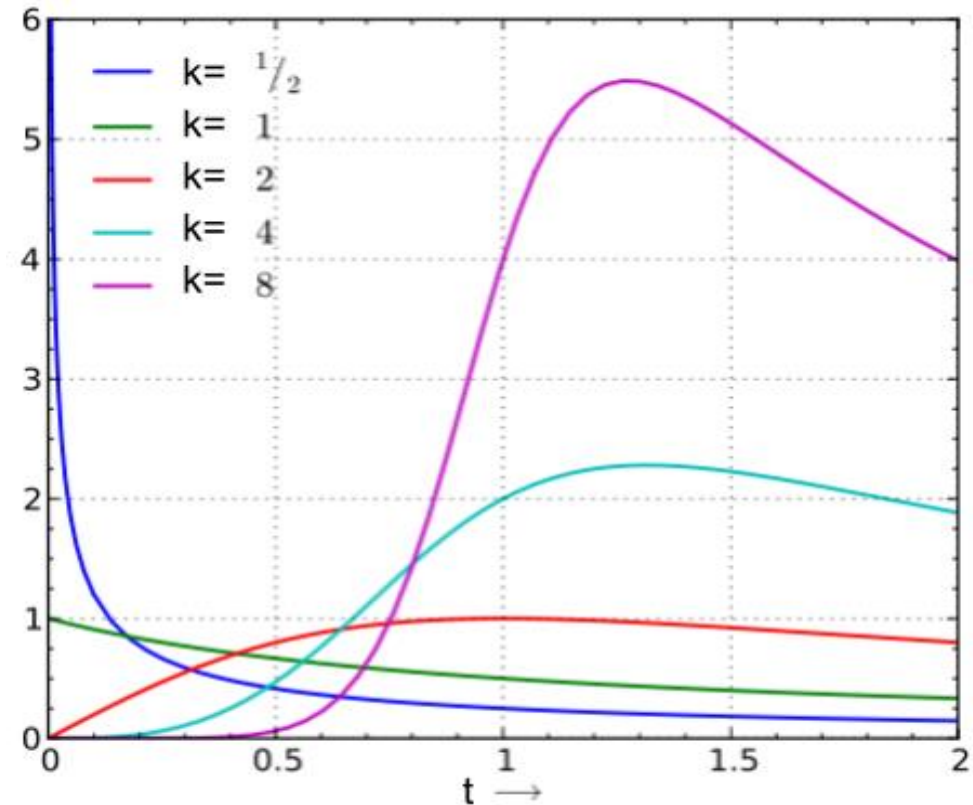
$$\Delta(t) = \ln(1 + (\lambda t)^k)$$

Log-Logistic Model

- $f(t)$ al variare di $1/\lambda$

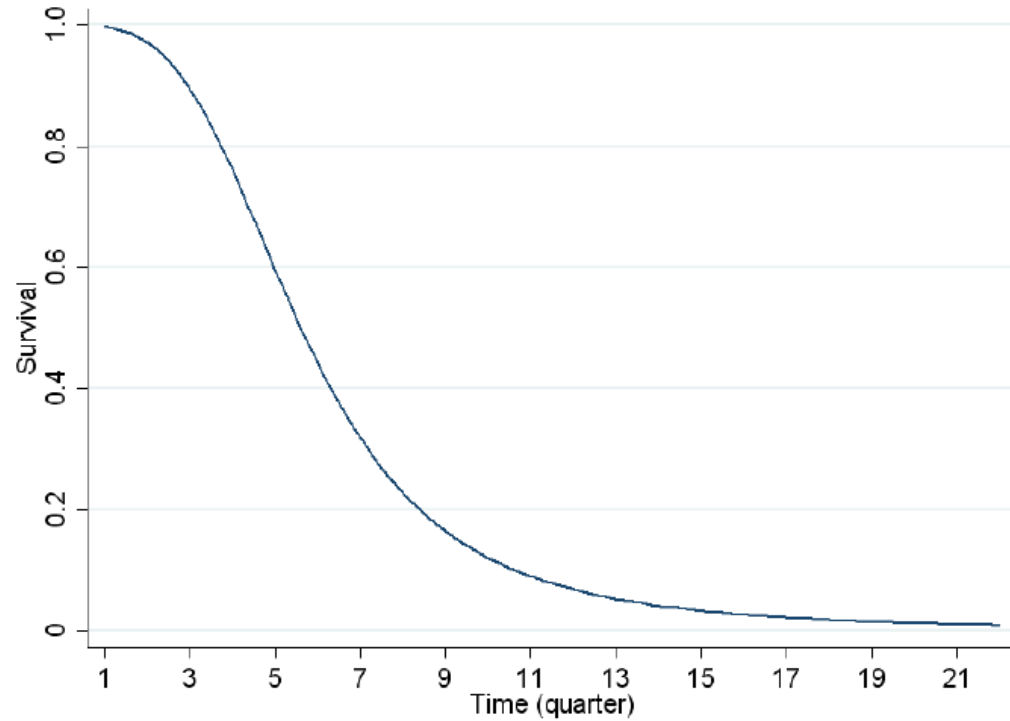


- $\lambda(t)$ al variare di k

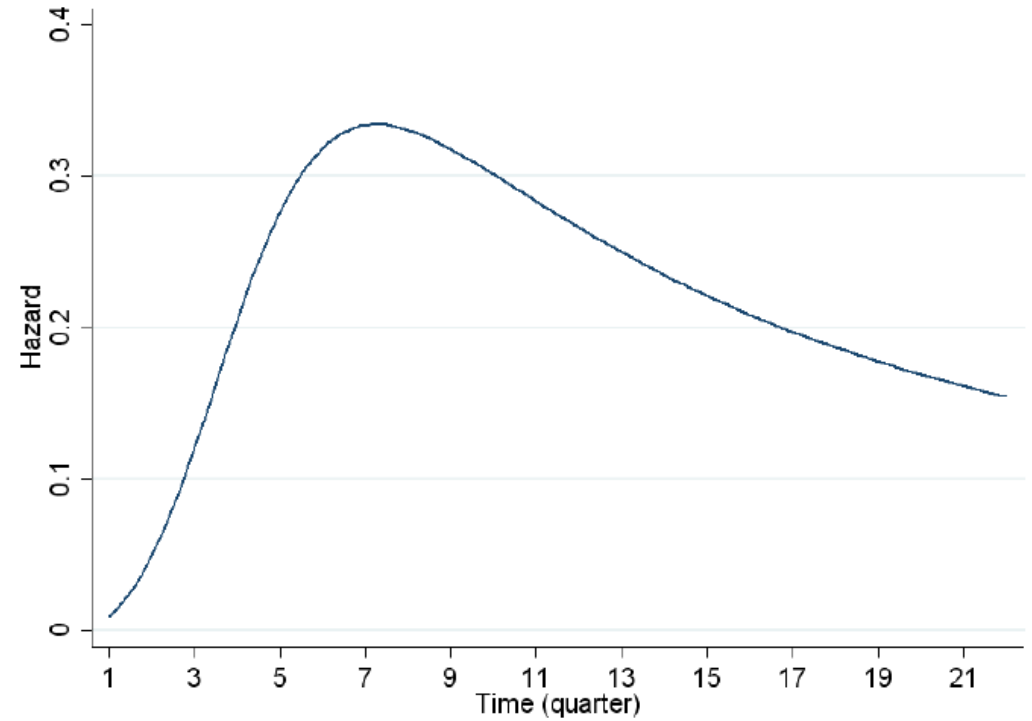


Log-Logistic Model

- $S(t)$



- $\lambda(t)$

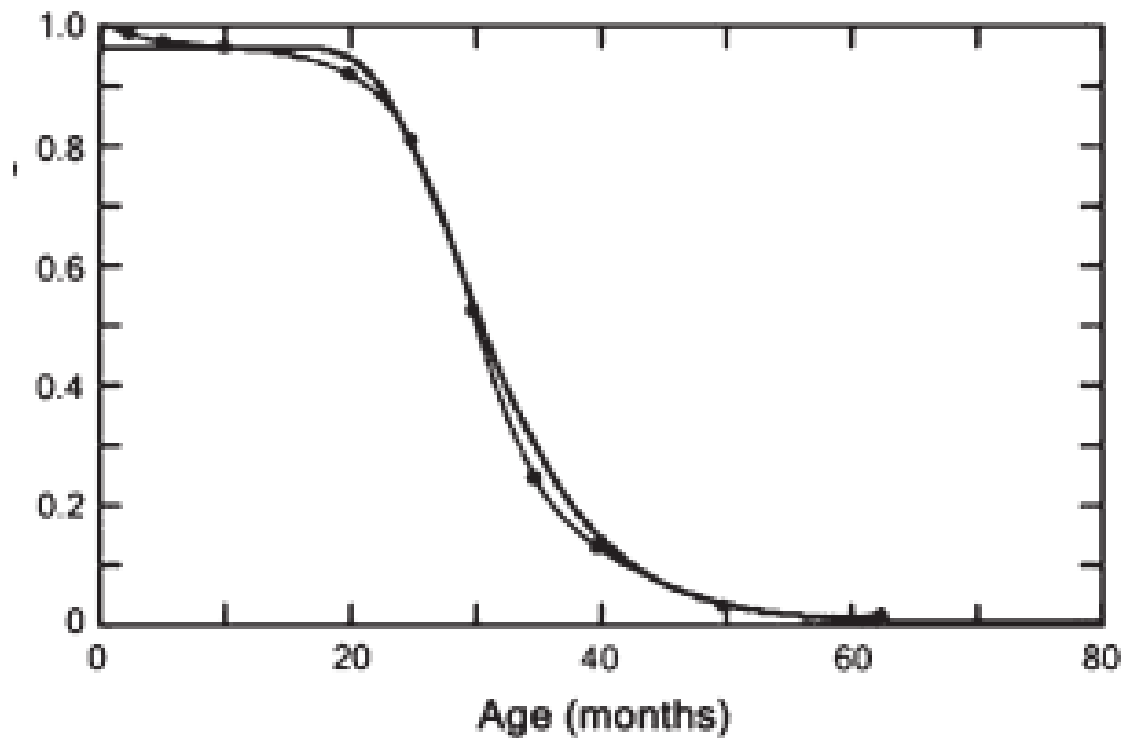


Gompertz Model

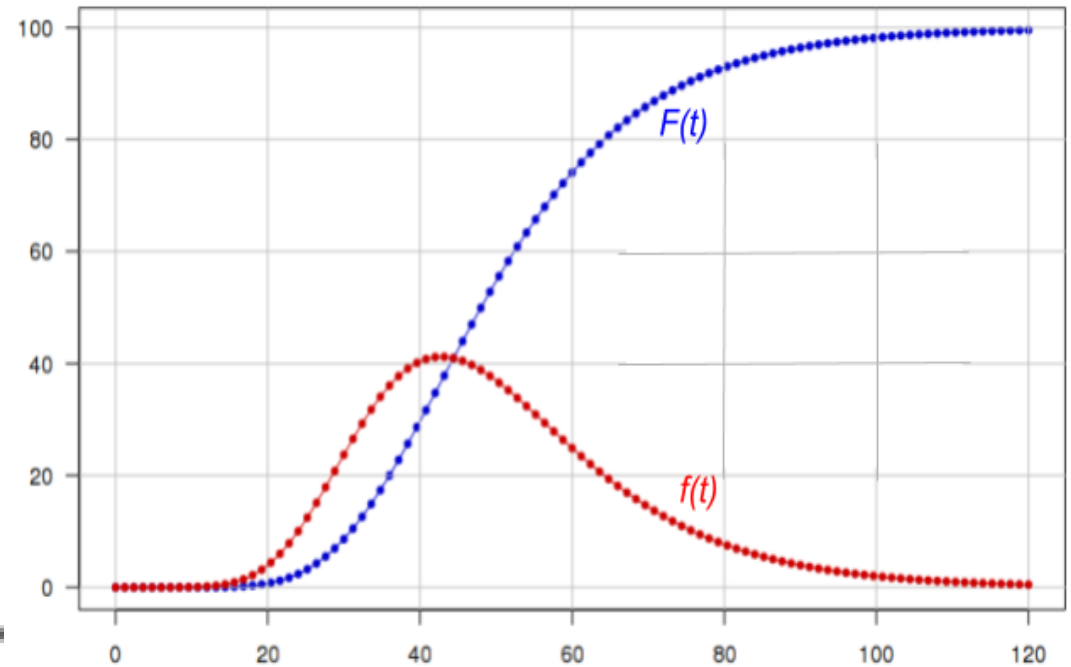
- Si basa su una relazione matematica che spiega perché esistono '*common age patterns of death*' dedotta da Gompertz osservando i fattori in comune in molte *life tables*.
- Utilizza la distribuzione di Gompertz : $T \sim \text{Gompertz}(\lambda, k)$
 - Ha parametri λ **tasso degli eventi** ($1/\lambda$ *scale parameter*) e k *shape parameter*
 - $f(t) = \lambda e^{kt} e^{-\frac{\lambda}{k}(e^{kt}-1)}$ $S(t) = 1 - F(t) = e^{-\frac{\lambda}{k}(e^{kt}-1)}$
 $\lambda(t) = \lambda e^{kt}$ $\Lambda(t) = \frac{\lambda}{k}(e^{kt} - 1)$
 - Il modello può essere generalizzato in **Gompertz-Makeham model** aggiungendo una costante c alla hazard function
 - $\lambda(t) = \lambda e^{kt} + c$

Gompertz Model

- $S(t)$ rispetto alle osservazioni



- $f(t)$ e $F(t)$



Generalized Gamma Model

- È un modello che utilizza la distribuzione gamma generalizzata, una distribuzione molto flessibile che contiene come casi speciali Weibull, Esponenziale, Log-Normal e Gamma (non generalizzata).
 - $T \sim GG(\alpha, \lambda, k)$
- Ha i parametri λ ($1/\lambda$ 'scale' parameter) e k come la distribuzione di Weibull e un ulteriore parametro di forma α .
 - Se $\alpha = 1 \rightarrow$ weibull distribution
 - Se $\alpha = k = 1 \rightarrow$ exponential distribution
 - Se $k = 1 \rightarrow$ gamma distribution

$$\begin{aligned}
 & \bullet \quad f(t) = \frac{k\lambda(\lambda t)^{\alpha-1} e^{-(\lambda t)^k}}{\Gamma(\frac{\alpha}{k})} \quad F(t) = \frac{\gamma(\frac{\alpha}{k}, (\lambda t)^k)}{\Gamma(\frac{\alpha}{k})} \quad \gamma(s, x) = \int_0^x t^{s-1} e^{-t} dt \quad \text{funzione gamma incompleta} \\
 & \quad \lambda(t) = \frac{k\lambda(\lambda t)^{\alpha-1} e^{-(\lambda t)^k}}{\gamma(\frac{\alpha}{k}, (\lambda t)^k)} \quad S(t) = 1 - F(t) \quad \Gamma(s) = \int_0^\infty t^{s-1} e^{-t} dt \quad \text{funzione gamma}
 \end{aligned}$$

Presenza di variabili esplicative

In presenza di un **vettore di covariate** bisogna assumere una delle due relazioni fra le variabili esplicative e il modello (PH o AFT).

Questo implica una **riparametrizzazione**, ovvero sostituire λ con una funzione di rischio in funzione della **baseline hazard function** e del vettore delle covariate.

Per questo introduciamo :

- λ_0 : **baseline hazard function**, ovvero effetto della hazard function senza considerare nessuna variabile esplicativa (spesso può essere il **tasso degli eventi**, così come può essere qualsiasi hazard function > 0)
- X : vettore delle covariate e quindi x_i i -esima variabile esplicativa
- β_i : parametro di influenza dell' i -esima variabile esplicativa
- α_i : parametro di accelerazione dell' i -esima variabile esplicativa

Baseline Hazard Function $\lambda_0(t)$

- Rappresenta il rischio che si incontra senza l'influenza di nessuna covariata.
- Segue una propria distribuzione, avendo quindi una forma propria
 - Ad esempio, utilizzando *weibull* considerando l'ipotesi di proporzionalità essa sarà la $\lambda(t)$ ottenuta tramite le formule del modello di weibull ed il modello ottenuto prenderà il nome di *weibull proportional hazards model*.
 - OSS: **Weibull** è l'unico modello che assumendo **AFT** mantiene comunque l'ipotesi di proporzionalità **PH**. (e viceversa)
- Alcune volte può essere vista come λ_0 tasso degli eventi, considerano però i rischi costanti nel tempo.

Assumere PH o AFT

OSS: $\exp(\arg) \equiv e^{\arg}$

- Assumendo PH

- $\lambda(t) = \exp(\lambda_0 + \sum_{i=1}^N (\beta_i X_i))$

- Esempio: Exponential PH model $S(t, X) = \exp\{-t \exp(\lambda_0 + \sum_{i=1}^N \beta_i X_i)\}$

- Assumendo AFT

- $\lambda(t) = \frac{1}{\exp(\lambda_0 + \sum_{i=1}^N (\alpha_i X_i))}$

- Esempio: Exponential AFT model $S(t, X) = \exp\{-t \exp(-\lambda_0 - \sum_{i=1}^N \alpha_i X_i)\}$

Adattamento parametri ai dati

Un modello di sopravvivenza può essere visto come un modello di regressione ordinaria con il tempo come variabile risposta. In presenza di dati censurati risulta molto utile la **funzione di verosimiglianza** (*likelihood*), ovvero la probabilità congiunta dei dati condizionata dai parametri del modello.

Essa ha un ruolo diverso dalla probabilità :

- Probabilità : funzione del risultato fissato un parametro (o un vettore di parametri)
- Verosimiglianza : funzione di un parametro dato un risultato.

La funzione di verosimiglianza si usa quindi per determinare quanto il valore di un parametro è verosimilmente corretto rispetto al suo valore osservato. Gioca un ruolo fondamentale nella **stima dei parametri** di un modello parametrico.

Adattamento parametri ai dati

Dati i parametri (ϑ), presupponendo che i dati siano fra loro **indipendenti**, la funzione di verosimiglianza $L(\vartheta)$ è il **prodotto** di quella di ogni singolo dato.

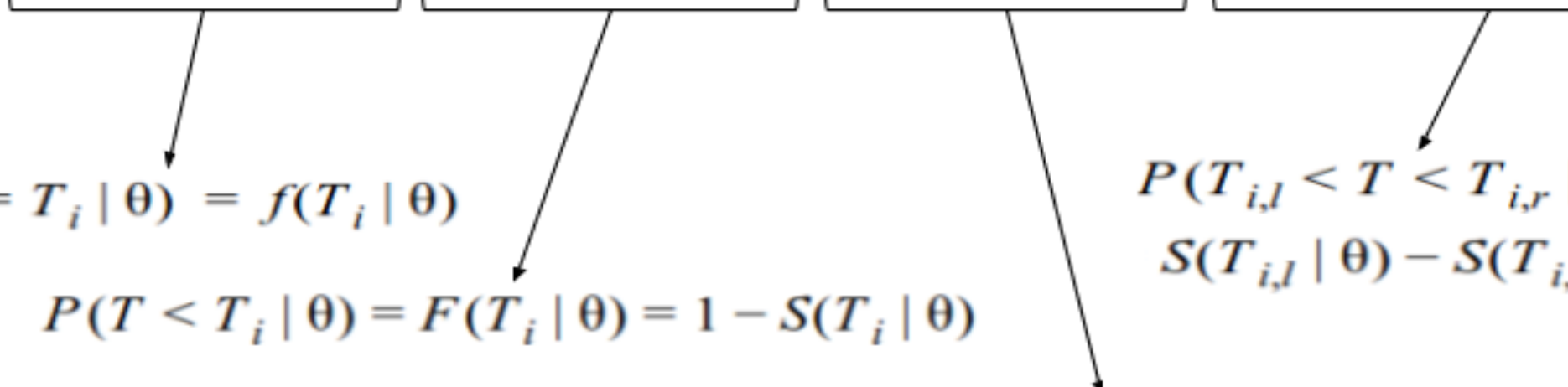
Si dividono quindi i dati in 4 categorie :

- *Uncensored* \rightarrow *UNC*
- *Left-censored* \rightarrow *L.C.*
- *Right-censored* \rightarrow *R.C.*
- *Interval-censored* \rightarrow *I.C.* (censura sia a destra che a sinistra)

Quindi $L(\vartheta)$ funzione di verosimiglianza dei parametri del modello \rightarrow prodotto della funzione di verosimiglianza dei 4 casi, dove ogni caso viene giustamente gestito diversamente.

Adattamento parametri ai dati

Componendo i 4 casi otteniamo la funzione di verosimiglianza utile per stimare i parametri del modello. (T_i = ultimo tempo registrato del soggetto i -esimo)

$$L(\theta) = \prod_{T_i \in UNC} P(T = T_i | \theta) \prod_{T_i \in L.C.} P(T < T_i | \theta) \prod_{T_i \in R.C.} P(T > T_i | \theta) \prod_{T_i \in UNC} P(T_{i,l} < T < T_{i,r} | \theta)$$


$P(T = T_i | \theta) = f(T_i | \theta)$

$P(T < T_i | \theta) = F(T_i | \theta) = 1 - S(T_i | \theta)$

$P(T > T_i | \theta) = 1 - F(T_i | \theta) = S(T_i | \theta)$

$P(T_{i,l} < T < T_{i,r} | \theta) = S(T_{i,l} | \theta) - S(T_{i,r} | \theta)$

Cox Proportional Hazards Model

- È un **proportional hazards model**
 - In contrasto con **accelerated failure time model**
 - L'effetto dell'incremento dell'influenza di una o più covariate è moltiplicativo con l'hazard rate (o la *baseline hazard function* che si è scelta), modificando la lifetime di un soggetto.
 - Basandosi sull'assunzione di proporzionalità stima l'effetto delle variabili esplicative senza considerare l'**hazard function**, che non verrà quindi specificata.
 - È uno dei motivi per cui il modello di Cox è uno dei più apprezzati.
- È un modello **semiparametrico**
 - Fornisce una stima di tutti i parametri, ma, a differenza dei modelli parametrici, non specifica una distribuzione per l'**event time T**.

Cox Model: formulazione

Terminologia:

- Y_i = tempo di osservazione per il soggetto i
- C_i = indicatore di corrispondenza tempo-eventi
 - $C_i = 1$ occorrenza dell'evento per il soggetto i
 - $C_i = 0$ il tempo per quel soggetto è un tempo di censoring
- $X_i = \{X_{i1}, X_{i2}, \dots, X_{ip}\}$ vettore delle covariate relative al soggetto i
 - Chiamati anche **parametri**

Formulazione:

$$\lambda(t | X_i) = \lambda_0 * \exp(\beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip}) = \lambda_0 * \exp\left(\sum_{j=1}^p \beta_j * X_{ij}\right)$$

The diagram illustrates the components of the Cox model equation. A green bracket under $\lambda(t | X_i)$ points to the label "hazard function" and "condizionata dal vettore delle covariate X_i ". A blue bracket under λ_0 points to the label "baseline hazard function". A red bracket under the exponential term $\exp(\beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_p X_{ip})$ points to the label "parameters effects : effetti del vettore delle covariate".

hazard function
condizionata dal
vettore delle covariate X_i

baseline hazard function

parameters effects : effetti del vettore
delle covariate

Cox Model: influenza di $\lambda_o(t)$

Essa è influente perchè nelle applicazioni di interesse si semplifica.

- **Hazard Rate:** calcolando l'hazard rate come quello di un individuo (o un gruppo di individui) fratto quello di un individuo (o un gruppo) differente:

$$HR = \frac{HR(t | X_i)}{HR(t | X_i^*)} = \frac{\lambda(t | X_i)}{\lambda(t | X_i^*)} = \frac{\lambda_0 * \exp(\sum_{j=1}^p \beta_j * X_{ij})}{\lambda_0 * \exp(\sum_{j=1}^p \beta_j * X_{ij}^*)} = \exp(\sum_{j=1}^p \beta_j (X_{ij} - X_{ij}^*))$$

- **Probabilità di un evento:** presupponendo un unico evento, ignorando possibili vincoli, considerando un soggetto i con $C_i = 1$ e $Y_i = t$ e definendo $\theta_i = \exp(\beta X_i)$:

$$\begin{aligned} P(\text{occorrenza evento per soggetto } i \mid \exists! \text{ evento}) &= \text{verosimiglianza } \beta = L_i(\beta) \\ &= \frac{P(\text{occorrenza evento} \cap \text{unico evento})}{P(\text{occorrenza dell'unico evento})} = L_i(\beta) = \frac{\lambda_0 \theta_i}{\sum_{j: Y_j \geq Y_i} \lambda_0 \theta_j} = \frac{\lambda_0 \theta_i}{\lambda_0 * \sum_{j: Y_j \geq Y_i} \theta_j} = \frac{\theta_i}{\sum_{j: Y_j \geq Y_i} \theta_j} \end{aligned}$$

Cox Model: probabilità e verosimiglianza

- Per stimare le probabilità di realizzazione di eventi si utilizza la verosimiglianza, dato che il suo valore è determinato dall'effetto dei **parametri**.
 - Viene quindi utilizzata la **funzione di verosimiglianza** per stimare i parametri del modello e rimuovere dall'equazione $\lambda_0(t)$.
- Essa prende il nome di **verosimiglianza parziale** perché considera nella probabilità solo i soggetti che hanno registrato un evento, non considerando i dati censored (che vengono però considerati nei rischi).
- Verosimiglianza parziale o *Partial Likelihood* : probabilità che si realizzino tutti gli eventi condizionata dall'esistenza di eventi in quei tempi presupponendo gli eventi dei soggetti come **indipendenti**

$$L(\beta) = \prod_{i: C_i=1} \frac{\theta_i}{\sum_{j: Y_j \geq Y_i} \theta_j}$$

Cox Model: log-likelihood

Alla verosimiglianza parziale corrisponde una sua versione logaritmica, ovvero la **log partial likelihood** $l(\beta)$, definita come segue :

$$l(\beta) = \sum_{i: C_i=1} (\beta * X_i - \log \sum_{j: Y_j \geq Y_i} \theta_j)$$

- Ad essa corrispondono :
 - **Score function parziale** $l'(\beta)$
 - Indica quanto una funzione di verosimiglianza $L(\beta, X)$ è **sensibile** ai suoi parametri β .
 - **Gradiente** della log-likelihood.
 - **Matrice hessiana** $l''(\beta)$
 - Il suo inverso valutato alla stima di β può essere usato come approssimazione della **matrice varianza-covarianza** per stimare **standard errors** per i **coefficienti della regressione**.

Cox Model: stima di parametri

- Massimizzando la funzione di verosimiglianza parziale logaritmica si ottengono le **stime di verosimiglianza massime dei parametri del modello**.
- Si può utilizzare il **metodo di Newton** (Newton-Raphson Algorithm) sfruttando $l'(\beta)$ come gradiente e $l''(\beta)$ come matrice hessiana.

$$\bullet \quad l'(\beta) = \sum_{i: C_i=1} \left(X_i - \frac{\sum_{j: Y_j \geq Y_i} \theta_j X_j}{\sum_{j: Y_j \geq Y_i} \theta_j} \right)$$

$$\bullet \quad l''(\beta) = - \sum_{i: C_i=1} \left(\frac{\sum_{j: Y_j \geq Y_i} \theta_j X_j X_j'}{\sum_{j: Y_j \geq Y_i} \theta_j} - \frac{\left[\sum_{j: Y_j \geq Y_i} \theta_j X_j \right] * \left[\sum_{j: Y_j \geq Y_i} \theta_j X_j' \right]}{\left[\sum_{j: Y_j \geq Y_i} \theta_j \right]^2} \right)$$

Cox Model: stimare $\lambda_0(t)$

- Il modello di Cox viene considerato semi-parametrico appunto perché nonostante fornisca un modello adeguato alle covariate non rappresenta nessun modello per la **baseline hazard function**.
- Allo stesso modo quando si utilizzava Kaplan-Meier non si specificava un modello per la funzione di sopravvivenza.
 - Adattando il suo metodo si può stimare $\lambda_0(t)$ come segue:

The diagram illustrates the estimation of the baseline hazard function $\hat{\lambda}_0(t_{(i)})$ in the Cox model. The equation is shown as
$$\hat{\lambda}_0(t_{(i)}) = \frac{d_i}{\sum_{j \in \mathcal{R}(t_{(i)})} \exp(\hat{\beta} X_j)}$$
 with several annotations and arrows: a red arrow points from $t_{(i)}$ to the text "tempo dell' i-esimo evento"; a green arrow points from d_i to "numero morti a $t_{(i)}$ "; a blue arrow points from $\hat{\lambda}_0(t_{(i)})$ to "stima della baseline hazard function"; and an orange arrow points from the risk set denominator to "risk set a tempo $t_{(i)}$, ovvero l'insieme di soggetti a rischio a quel tempo (quelli che ancora non hanno riscontrato un evento)".

tempo dell' i-esimo evento

numero morti a $t_{(i)}$

stima della baseline hazard function

risk set a tempo $t_{(i)}$, ovvero l'insieme di soggetti a rischio a quel tempo (quelli che ancora non hanno riscontrato un evento)

Cox Model: Survival Function

- Prende il nome di **adjusted survival curve** perché usa le variabili esplicative (il vettore di covariate) come predittori.
- Come tramite Kaplan-Meier se ne può ottenere solo una stima che sarà rappresentata da una funzione a gradini.

The diagram illustrates the derivation of the estimated survival function $\hat{S}_0(t_{(i)})$ from the Cox model equation. It features several mathematical expressions and descriptive text connected by arrows.

Top Left: The equation $S(t) = S_0(t) \exp\left(\sum_{i=1}^p \beta_i X_i\right)$ is shown. A blue bracket under $S_0(t)$ has a blue arrow pointing down to the text "baseline survival function: survival function non influenzata dalle covariate".

Top Right: The equation $S(t) = \exp(-\Lambda(t)) = \exp\left(-\int_0^t \lambda(t) dt\right)$ is shown. A black arrow points from the text "ricordando:" to this equation. A black bracket under the integral term has a black arrow pointing down to the text "approssimata dalla stima discreta precedente $\hat{\lambda}_0(t_{(i)})$ ".

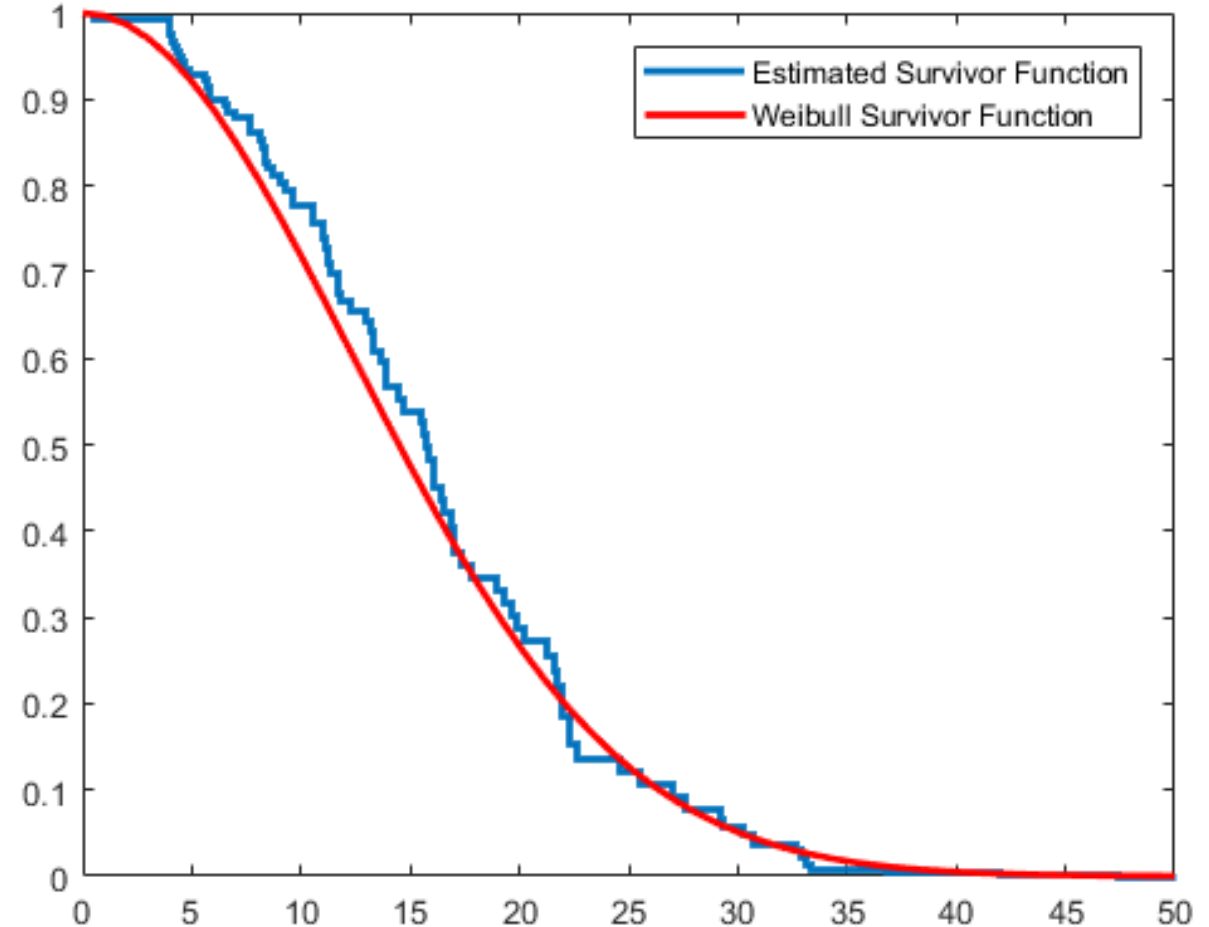
Center: The text "da stimare" is positioned between two black arrows. One arrow points from the $S_0(t)$ term in the top-left equation down to the estimated function. The other arrow points from the integral term in the top-right equation down to the same estimated function.

Bottom: The equation $\hat{S}_0(t_{(i)}) = \exp\left[-\sum_{j \leq i} \hat{\lambda}_0(t_{(j)})\right]$ is shown. A blue bracket under the entire expression has a blue arrow pointing up to the text "stima della baseline survival function".

Red Arrows: Two red arrows point from the text "da stimare" to the estimated function $\hat{S}_0(t_{(i)})$. One red arrow also points from the text "baseline survival function: survival function non influenzata dalle covariate" to the estimated function.

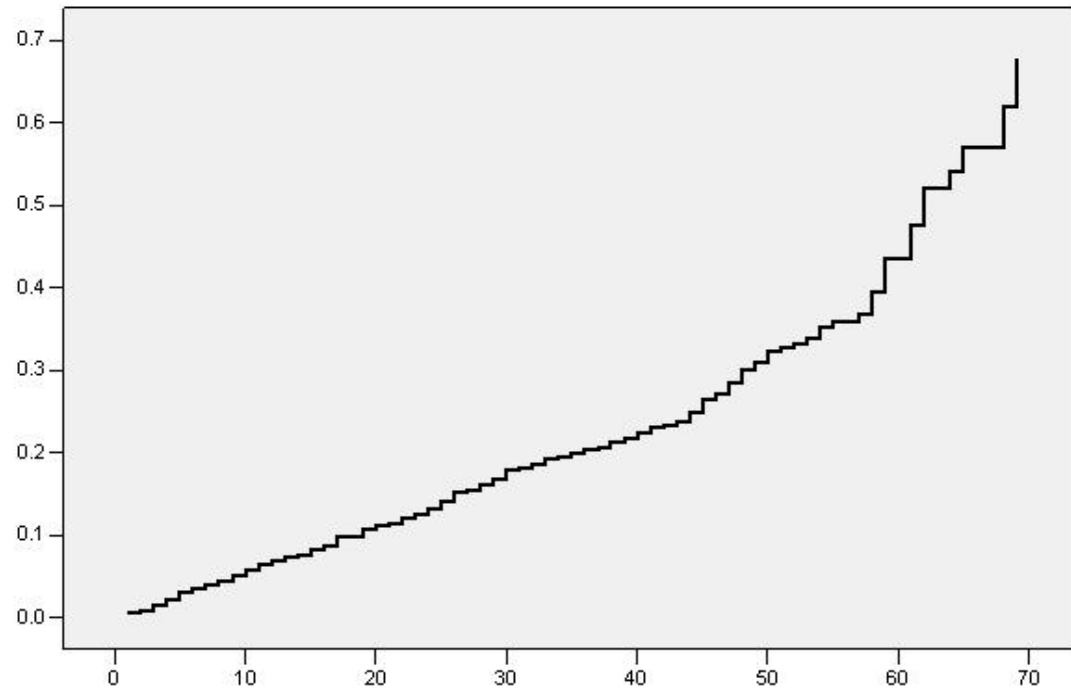
Cox Model: Survival Function

- Essendo ottenuta tramite una stima discreta sarà una funzione a gradini, non molto differente da quella ottenibile tramite kaplan-meier.
- Solitamente non si discosta troppo da quella che si otterrebbe con un qualsiasi modello parametrico adatto alla situazione.

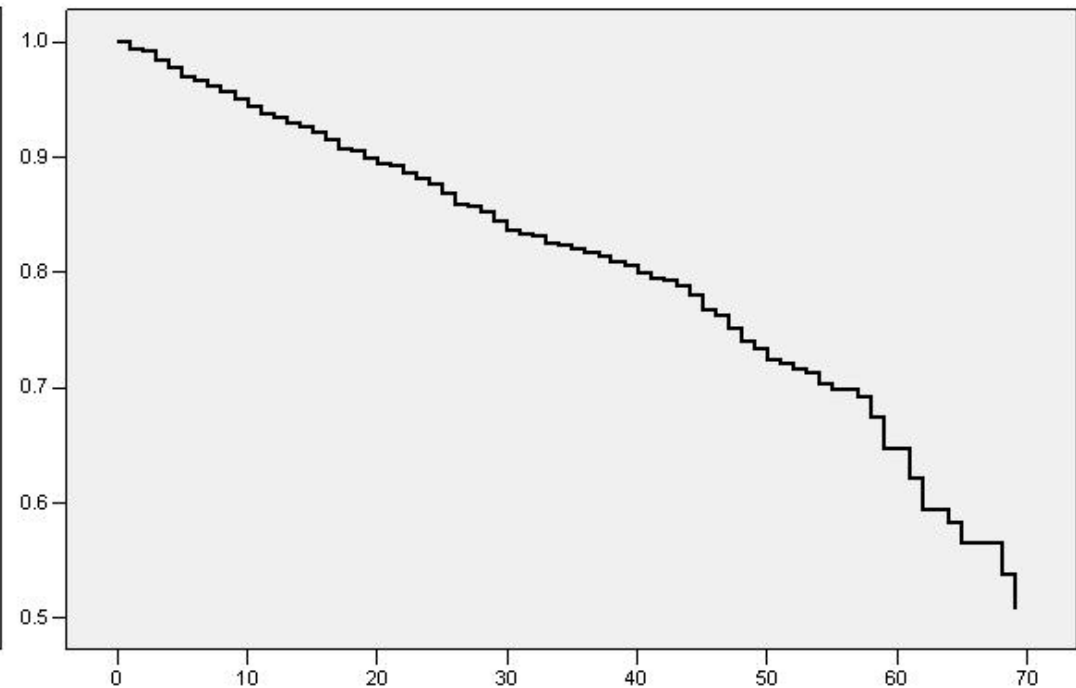


Cox Model: Survival function & Hazard function

- $\lambda(t)$



- $S(t)$



Cox Model: presenza di vincoli

- Nella definizione della funzione di verosimiglianza si è assunto il tempo come continuo, in assenza di vincoli temporali (*ties*).
- Nella pratica è molto comune incorrere in dati vincolati
- Ad esempio potrebbe capitare di avere dati come questi, dove i tempi (OSS: $x = t$) sono espressi in funzione degli altri, con i seguenti vincoli :
 - $x_1 = x_2 < x_3 < x_4 < x_5$
- In questi casi serve un'altra tecnica per costruire la funzione di verosimiglianza parziale per ottenere i parametri del modello.

Patient	x	δ	z
1	x_1	1	z_1
2	x_2	1	z_2
3	x_3	0	z_3
4	x_4	1	z_4
5	x_5	1	z_5

Cox Model: Efron's Method per dati vincolati

- È il metodo più utilizzato per dati di sopravvivenza fra loro vincolati.
- Se i soggetti con tempi vincolati hanno lo stesso hazard rate questo metodo risulta **esatto** (non solo un'approssimazione).
- Definisce:
 - t_j = tempo unico
 - H_j = set di indici t.c. $Y_i = t_j$ e $C_i = 1$ ($|H_j| = m_j$)
- Massimizza la funzione di verosimiglianza parziale

$$L(\beta) = \prod_{j=1}^{m-1} \frac{\prod_{i \in H_j} \theta_i}{\prod_{l=0}^j \left[\sum_{i: Y_i \geq t_j} \theta_i - \frac{l}{m} \sum_{i \in H_j} \theta_i \right]}$$

- A cui corrisponde la versione logaritmica: $l(\beta) = \sum_{j=1}^{m-1} \left(\sum_{i \in H_j} \beta X_i - \sum_{l=0}^{m-1} \log \left(\sum_{i: Y_i \geq t_j} \theta_i - \frac{l}{m} \sum_{i \in H_j} \theta_i \right) \right)$

Cox Model: Efron's Method per dati vincolati

- Per massimizzare $l(\beta)$ utilizziamo il metodo di Newton definendo:
 - **Gradiente/score function $l'(\beta)$**

$$l'(\beta) = \sum_j \left(\sum_{i \in H_j} X_i - \sum_{l=0}^{m-1} \frac{\sum_{i: Y_i \geq t_j} \theta_i X_i - \frac{l}{m} \sum_{i \in H_j} \theta_i X_i}{\sum_{i: Y_i \geq t_j} \theta_i - \frac{l}{m} \sum_{i \in H_j} \theta_i} \right)$$

- **Matrice Hessiana $l''(\beta)$**

$$l''(\beta) = - \sum_j \sum_{l=0}^{m-1} \left(\frac{\sum_{i: Y_i \geq t_j} \theta_i X_i X_i' - \frac{l}{m} \sum_{i \in H_j} \theta_i X_i X_i'}{\phi_{j,l,m}} - \frac{Z_{j,l,m} Z_{j,l,m}'}{\phi_{j,l,m}^2} \right)$$

$$\text{con } \phi_{j,l,m} = \sum_{i: Y_i \geq t_j} \theta_i - \frac{l}{m} \sum_{i \in H_j} \theta_i \text{ e } Z_{j,l,m} = \sum_{i: Y_i \geq t_j} \theta_i X_i - \frac{l}{m} \sum_{i \in H_j} \theta_i X_i$$

Cox Model: predittori e coefficienti che variano nel tempo

- In alcune situazioni l'effetto di alcuni fattori espressi dalle covariate, (come ad esempio un trattamento, una droga, una fase, un decadimento, ...) può variare nel tempo e perdere di efficacia.
- In questi casi servono modelli che prevedano questa assunzione.
- Si utilizza una differente formulazione per gli hazard, ideata da **Anderson e Gil** :

$$\lambda(t | X_i) = \lambda_0(t) + \beta_1 X_{i1} + \dots + \beta_p X_{ip} = \lambda_0(t) + \beta X_i$$

Cox Model: vettore di covariate di dimensioni elevate

- Se le dimensioni del vettore delle covariate sono maggiori delle dimensioni del campione ($p > n$) c'è bisogno di utilizzare un metodo differente per stimare i parametri del modello.
- Uno dei più usati è il **metodo di LASSO**
 - Si minimizza l'opposto della funzione di verosimiglianza parziale logaritmica vincolata dalla norma di tipo L1 (L1-Norm)

$$\bullet \quad l(\beta) = \sum_j \left(\sum_{i \in H_j} \beta X_i - \sum_{l=0}^{m-1} \log \left(\sum_{i: Y_i \geq t_j} \theta_i - \frac{l}{m} \sum_{i \in H_j} \theta_i \right) \right) + \lambda \|\beta\|_1$$

con $\|\beta\|_1 = \sum_{i=1}^p |\beta_i|$ L1-norm

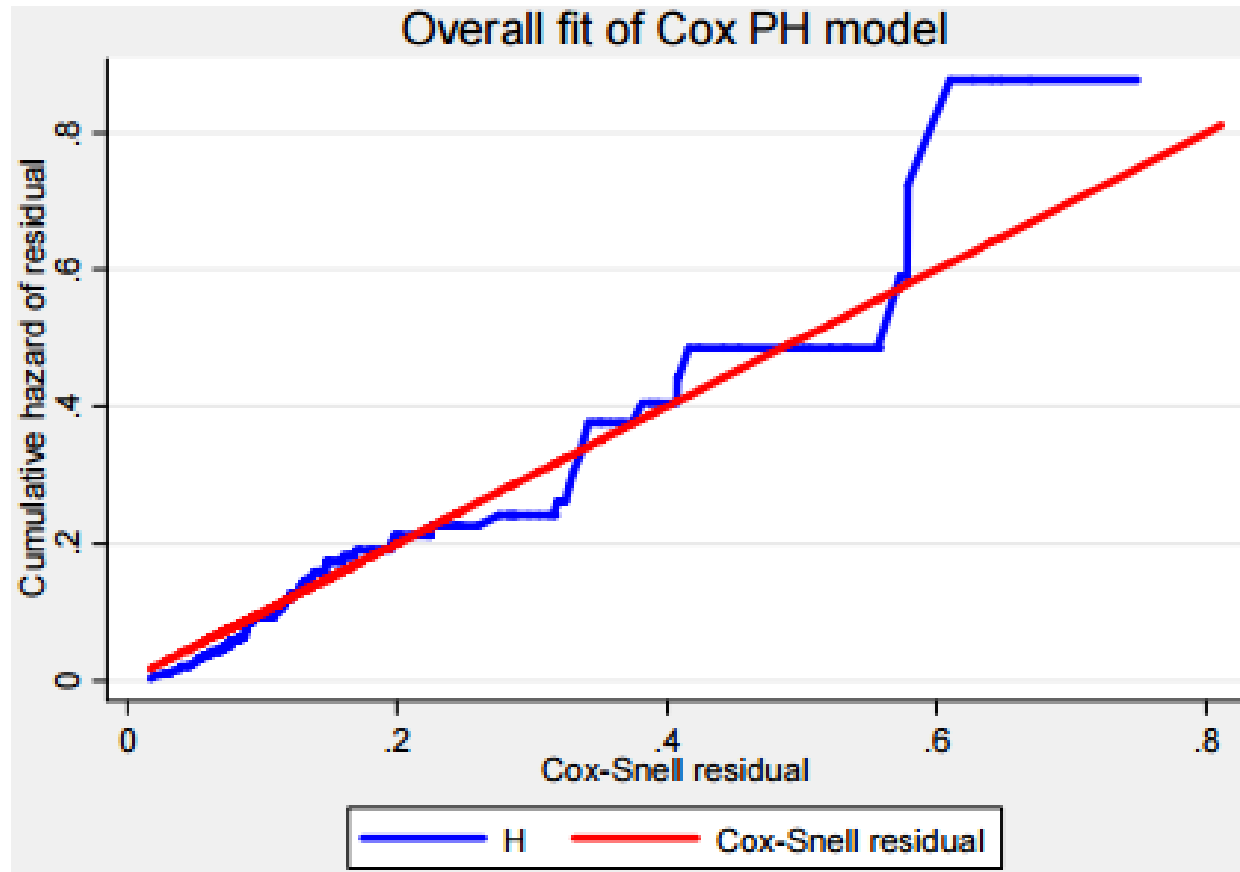
Cox Model: controllo del modello

- Una volta ottenuto il modello è necessario controllarne l'attinenza con i dati a disposizione.
- Per questo scopo si usano principalmente due metodi :
 - **Cox-Snell Residuals**
 - È un metodo basato sul controllo della disposizione dei residui.
 - Si usa per validare l'idoneità del modello di cox a hazard proporzionali.
 - **Controllo dell'assunzione di proporzionalità**
 - Si controlla che gli hazard siano effettivamente proporzionali tramite diversi metodi come, ad esempio:
 - Metodo grafico
 - Scaled Schoenfeld Residuals

Cox Model: Cox-Snell Residuals

- Utilizza il grafico dei residui per controllare l'idoneità del modello.
 - Residui per l' i -esimo individuo : $r_{ci} = \exp(\hat{\beta}x_i)\hat{\lambda}_0(t_{(i)})$
- Si usa lo stimatore di Nelson-Aalen per stimare $\hat{\Lambda}(r_{ci})$
- Se il modello è corretto e le stime di β sono vicine ai valori reali allora il grafico delle stime ottenute tramite Nelson-Aalen rispetto ai residui r_{ci} è una curva discreta che non si discosta troppo da una retta passante per l'origine e con pendenza 1.

Cox Model: Cox-Snell Residuals



- In questo esempio la stima ottenuta della cumulative hazard dei residui è rappresentata da **H** e si può notare come con il tempo si discosti troppo dalla **retta** passante per l'origine con pendenza 1.
- Il modello di Cox in questione non è troppo adeguato alla situazione da modellare. (qualche variabile esplicativa potrebbe non avere un effetto proporzionale)

Cox Model: controllo dell'assunzione di proporzionalità

L'assunzione di proporzionalità è verificabile dimostrando che l'**hazard ratio** è costante nel tempo.

- **Metodo grafico**

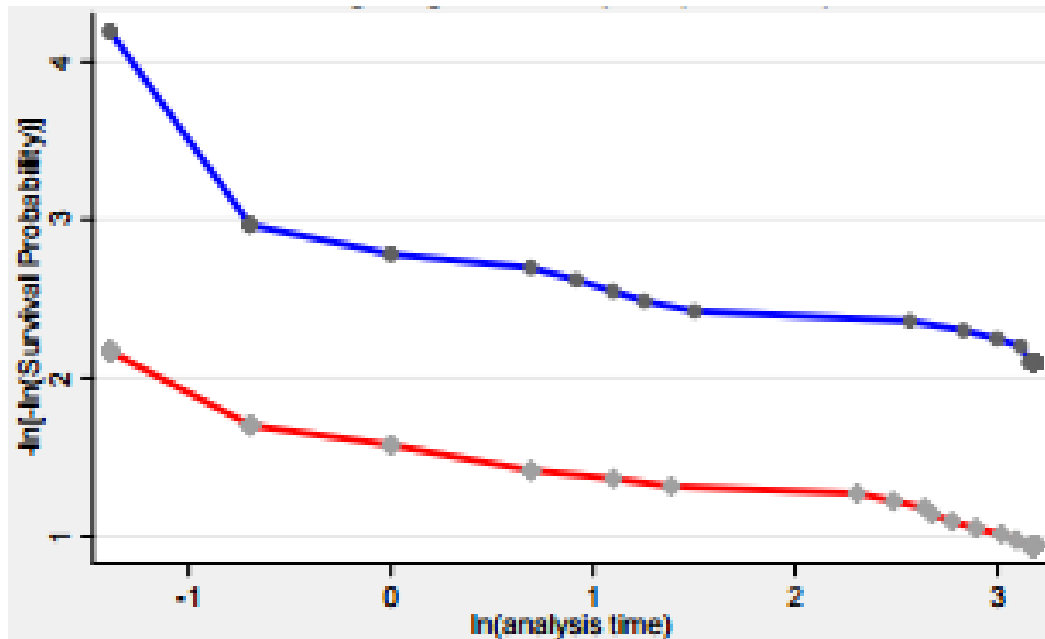
- Si considerano due individui (o gruppi di individui) e se ne confronta la funzione di sopravvivenza.
- Più nel dettaglio si confrontano le due relative curve $-\log(-\log(S(t)))$ ($S(t)$ a cui viene applicata due volte $-\log$).
- Se le due curve ottenute sono parallele, allora gli hazard sono proporzionali.

- **Scaled Schoenfeld Residuals**

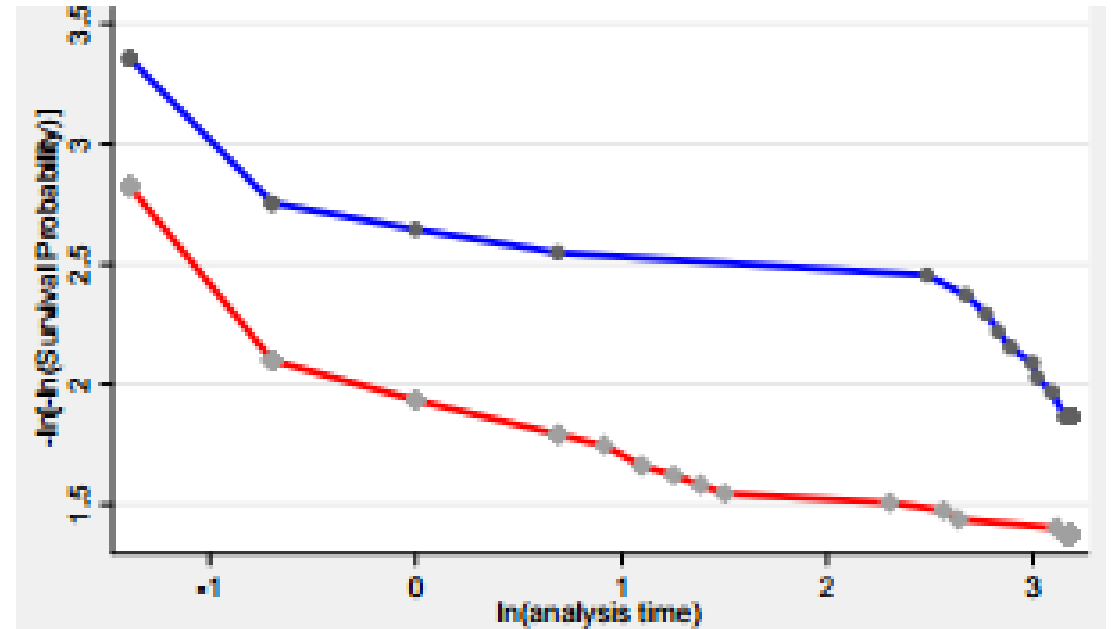
- Sono utilizzati per valutare tendenze temporali e mancanze di proporzionalità
- *Scaled Schoenfeld Residual k-esimo* = il prodotto tra l'inverso della matrice varianza-covarianza del k-esimo residuo di Schoenfeld e il *k-esimo residuo di Schoenfeld*
 - $r_{pij}^* = (V^{-1}) * r_{pij}$
- L'assunzione di proporzionalità risulta confermata se il grafico dei residui ottenuti è una retta con pendenza 0

Cox Model: controllo PH tramite metodo grafico

- Grafico relativo al valore di una variabile esplicativa binaria in cui la proporzionalità è mantenuta (le due curve sono assimilabili come parallele)

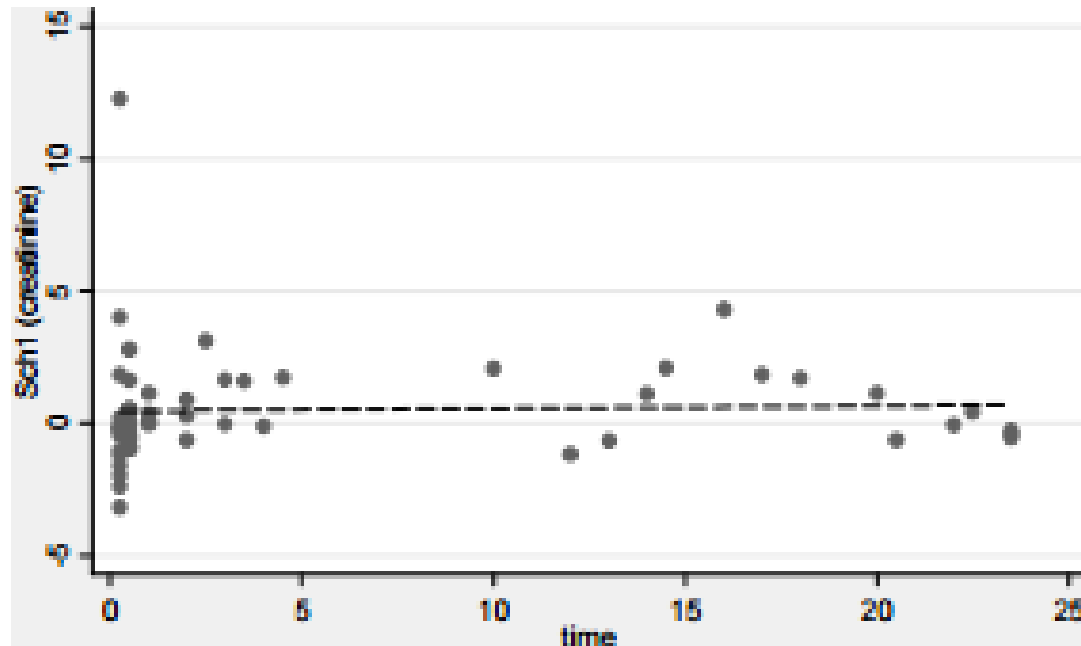


- Grafico relativo al valore di una variabile esplicativa binaria in cui la proporzionalità **non** è mantenuta (le due curve non sono parallele)

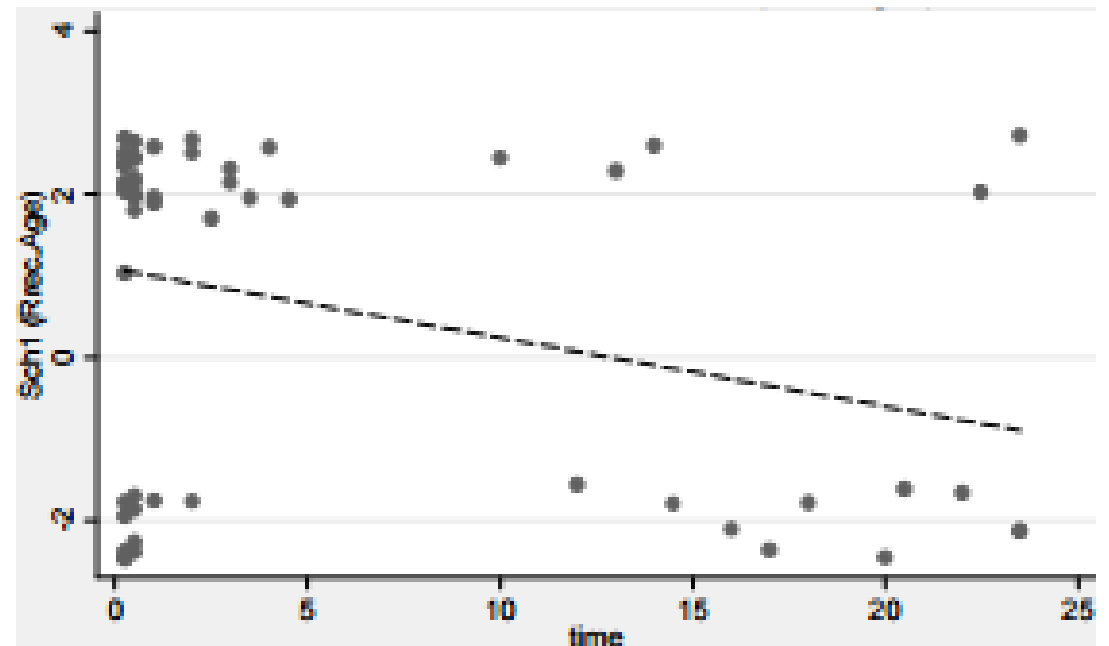


Cox Model: controllo PH tramite Scaled Schoenfeld Residuals

- Grafico dei residui relativo ad una covariata che rispetta la proporzionalità.



- Grafico dei residui relativo ad una covariata che **non** rispetta la proporzionalità.



Stratified Cox Model

- In alcuni casi qualche covariata potrebbe non rispettare l'assunzione di proporzionalità.
- In queste situazioni si deve adattare il modello di Cox per modellare quella determinata variabile con una differente **baseline hazard function** (negli altri casi diventa ininfluente appunto perché è comune fra tutti).
- Si presuppone quindi di avere un fattore di rischio a **k livelli**, dove ogni livello avrà una funzione di rischio differente.

$$\bullet \quad \forall k : \lambda_k(t | X(t)) = \lambda_{0k}(t) \exp\{\beta X(t)\}$$

Stratified Cox Model: likelihood

- Ad ogni livello k corrisponde una differente funzione di verosimiglianza parziale.
 - Ognuna rappresenta il contributo alla funzione di verosimiglianza totale degli individui di un determinato livello k
 - $\theta_i = \exp\{\beta X_{ki}\}$
 - likelihood k -esima : $L_k(\beta) = \prod_{i=1}^{n_k} \frac{\theta_i}{\sum_{j \in R_k} \theta_j}$, $R_k = n^\circ$ individui a rischio nel livello k
- La funzione di verosimiglianza totale sarà data dal prodotto di ogni singolo livello.
 - $L(\beta) = \prod_{k=1}^K L_k(\beta)$

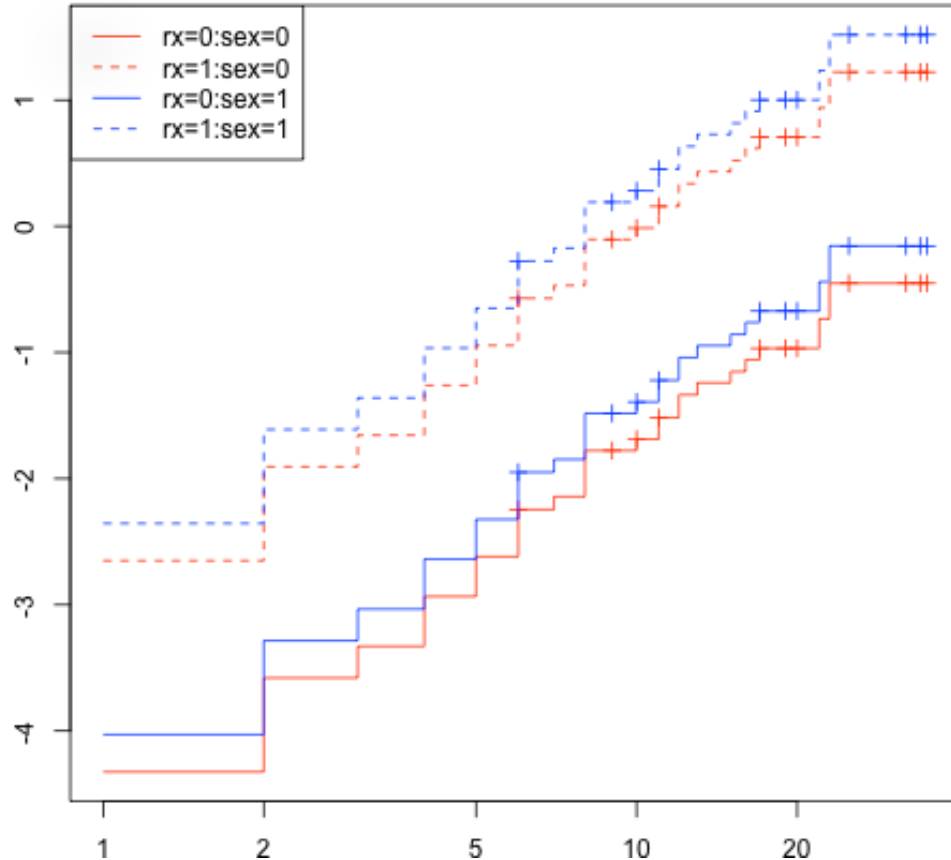
Stratified Cox Model: confronto fra $S(t)$

- Consideriamo una situazione in cui agiscono due variabili esplicative binarie, 'sex' e 'rx'.
- Analizziamo le differenze nei grafici delle curve di sopravvivenza logaritmiche in base alle possibili assunzioni in questi 4 casi:
 - Assunzione di proporzionalità per entrambe le variabili
 - Caso **1**: baseline hazard function e effect parameters uguali
 - Caso **2**: baseline hazard function uguale e effect parameters differenti
 - Stratificazione per 'sex'
 - Caso **3**: effect parameters uguali
 - Caso **4**: effect parameters differenti

Stratified Cox Model: confronto fra $S(t)$

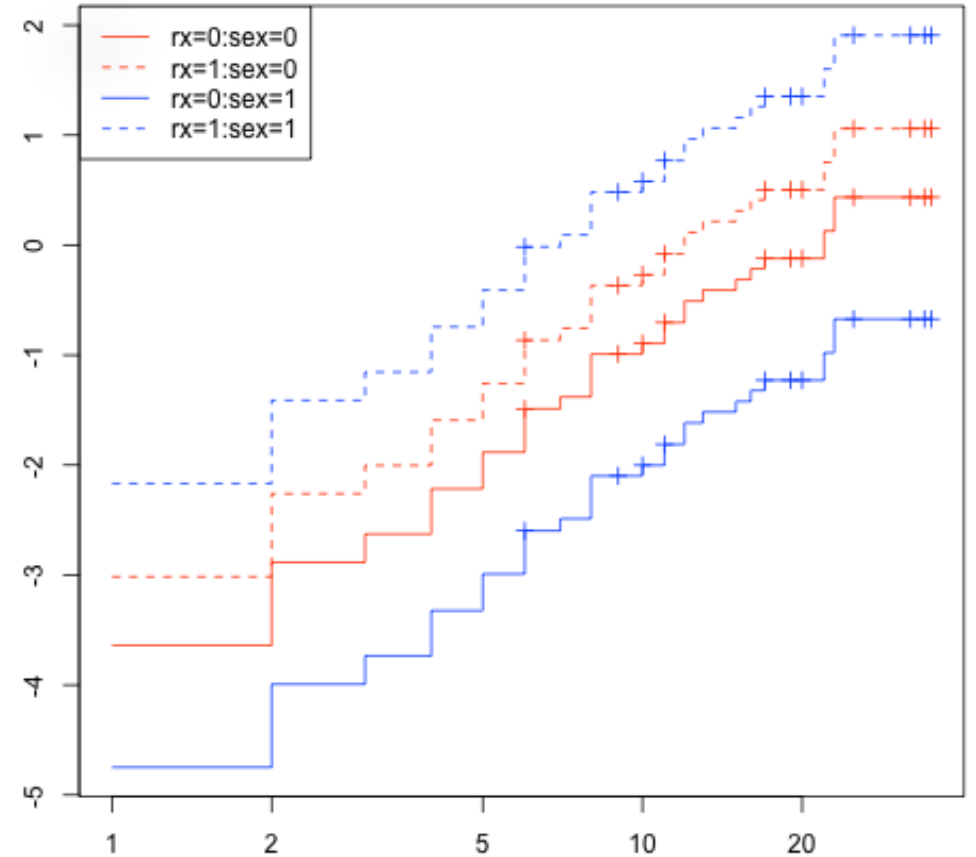
1

Same baselines, same effects
all parallel with same distance



2

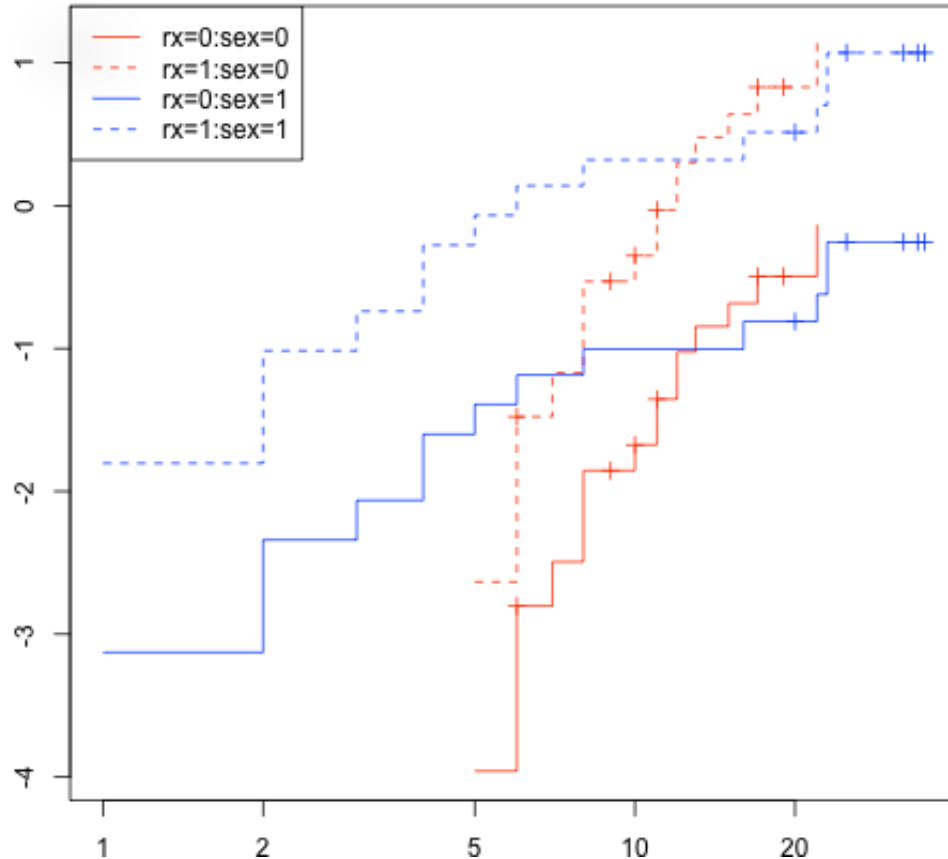
Same baselines, different effects
all parallel with different distances



Stratified Cox Model: confronto fra $S(t)$

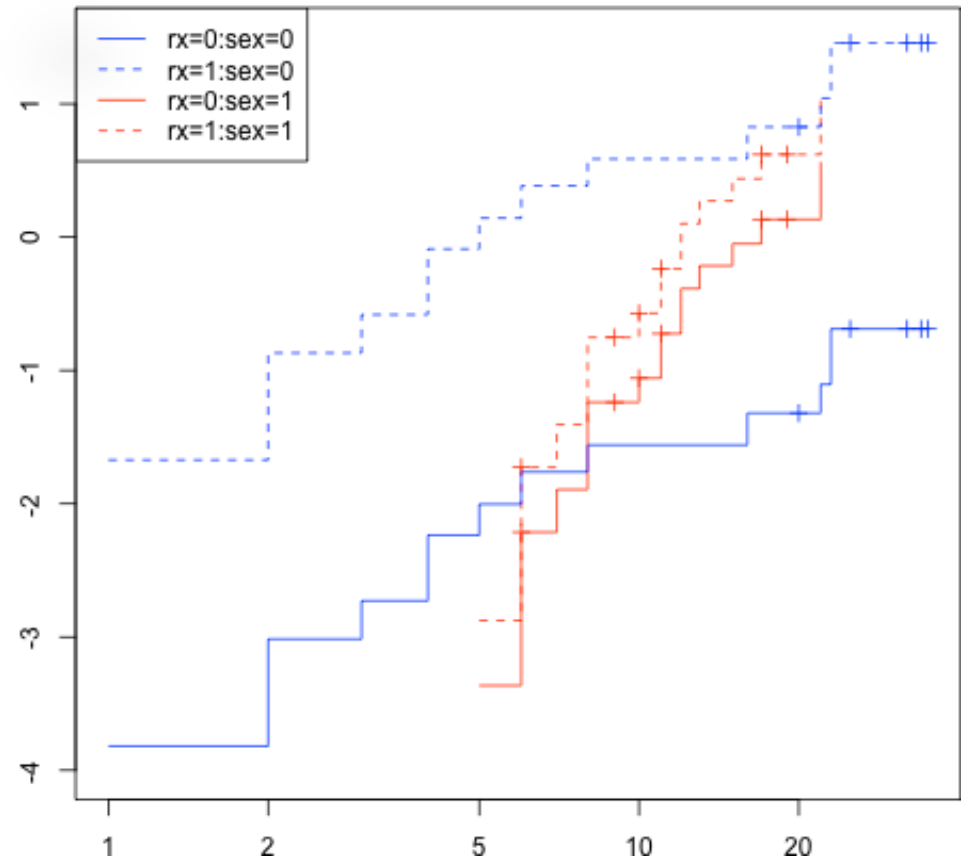
3

Different baselines, same effects
parallel within each stratum with equal distance



4

Different baselines, Different effects
parallel within each stratum with different distances



Tree Structured Survival Analysis

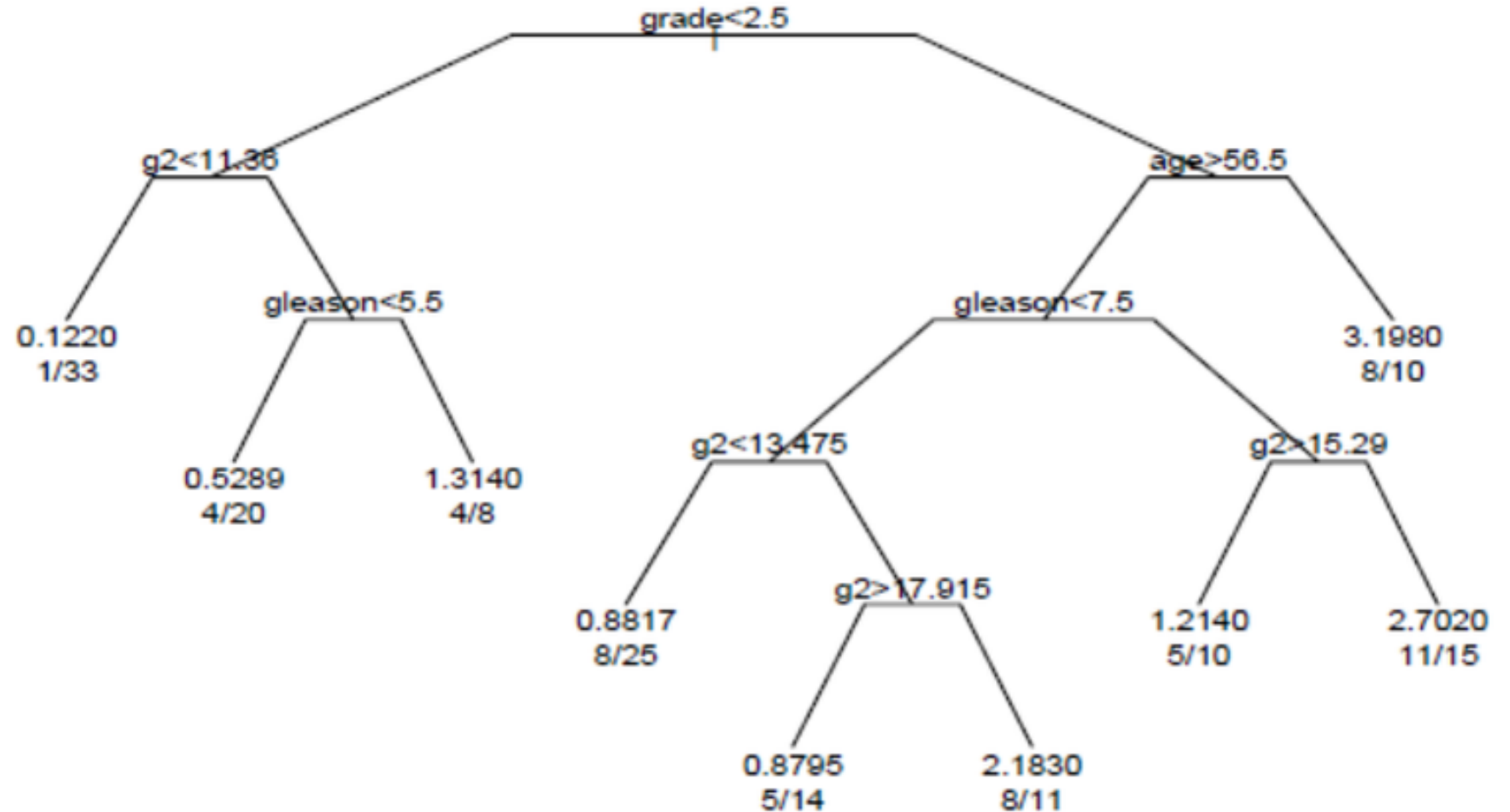
- Nei modelli precedenti si assumeva che una sola curva potesse stimare al meglio la sopravvivenza di un campione di individui.
- Spesso si potrebbero raggruppare i membri del campione per caratteristiche simili, ottenendo diversi sottogruppi.
 - Si riuscirebbe a fornire una stima più accurata per ogni sottogruppo
- L'analisi di sopravvivenza basata su modelli ad albero sfrutta questo concetto tramite l'utilizzo di due modelli principali:
 - **Survival Tree**
 - **Survival random Forest**

Survival Tree

- Sono modelli in cui si modellano le caratteristiche di interesse con delle variabili.
- Si utilizzerà un albero in cui ogni ramificazione indica uno **split** del **valore di una variabile**.
- In questo modo ogni foglia ha una particolare 'configurazione' di valori di variabili e per ognuna si potrà fornire una stima di sopravvivenza più accurata di quella generica ottenuta con i modelli precedenti.
 - La sopravvivenza dell'intero campione è comunque ottenibile facendo la media della sopravvivenza di ogni foglia.
- È un **decision tree** applicato all'analisi di sopravvivenza.

Survival Tree

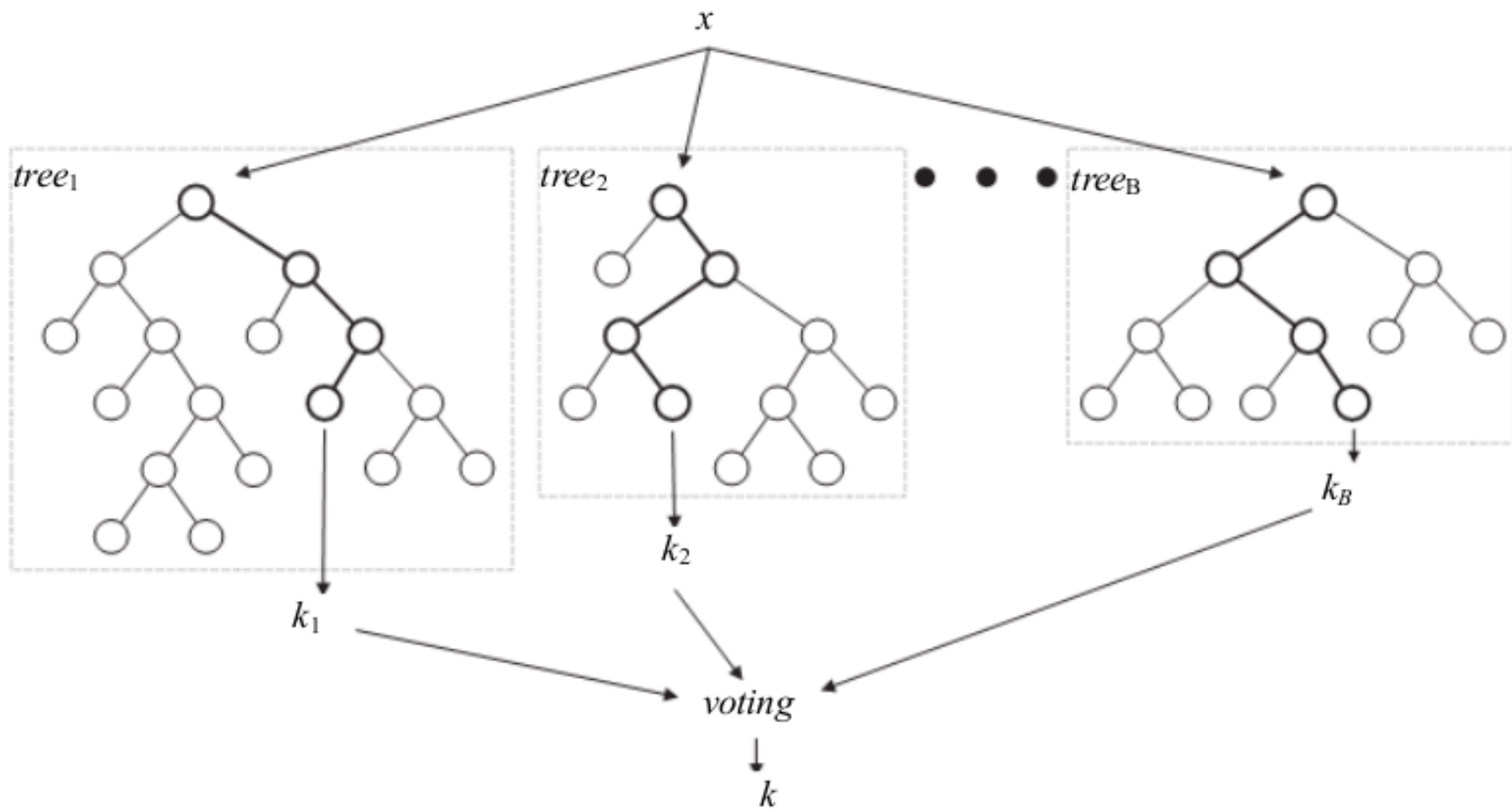
- Ogni ramificazione è uno split sul valore di una variabile
- Le foglie contengono l'**event rate**



Survival Random Forest

- Sono modelli basati sulla costruzione di più alberi, ognuno su un determinato campione della popolazione di interesse o, come accade maggiormente, sullo stesso campione, ma considerando covariate diverse nelle ramificazioni.
- Ogni albero funzionerà come un Survival Tree.
- La stima finale della sopravvivenza della popolazione sarà data dalla media di quella predetta per ogni albero.
- È una **random forest** (insieme di decision tree) applicata all'analisi di sopravvivenza, si differenzia perché sulle foglie contiene una stima degli hazard cumulativi.

Survival Random Forest



Esempio: non parametric model for Myelomatosis data

Dati di sopravvivenza per pazienti sottoposti a due diversi trattamenti (1 e 2), per i quali si desidera effettuare un'analisi di sopravvivenza sfruttando un **modello non parametrico**.

- Status:
 - 0 = censored
 - 1 = death
- Renal:
 - 0 = normal
 - 1 = impaired

Treat	Duration	Status	Renal	Treat	Duration	Status	Renal
1	8	1	1	2	180	1	0
1	852	0	0	2	632	1	0
1	52	1	1	2	2240	0	0
1	220	1	0	2	195	1	0
1	63	1	1	2	76	1	0
1	8	1	0	2	70	1	0
1	1976	0	0	2	13	1	1
1	1296	0	0	2	1990	0	0
1	1460	0	0	2	18	1	1
1	63	1	1	2	700	1	0
1	1328	0	0	2	210	1	0
1	365	0	0	2	1296	1	0
				2	23	1	1

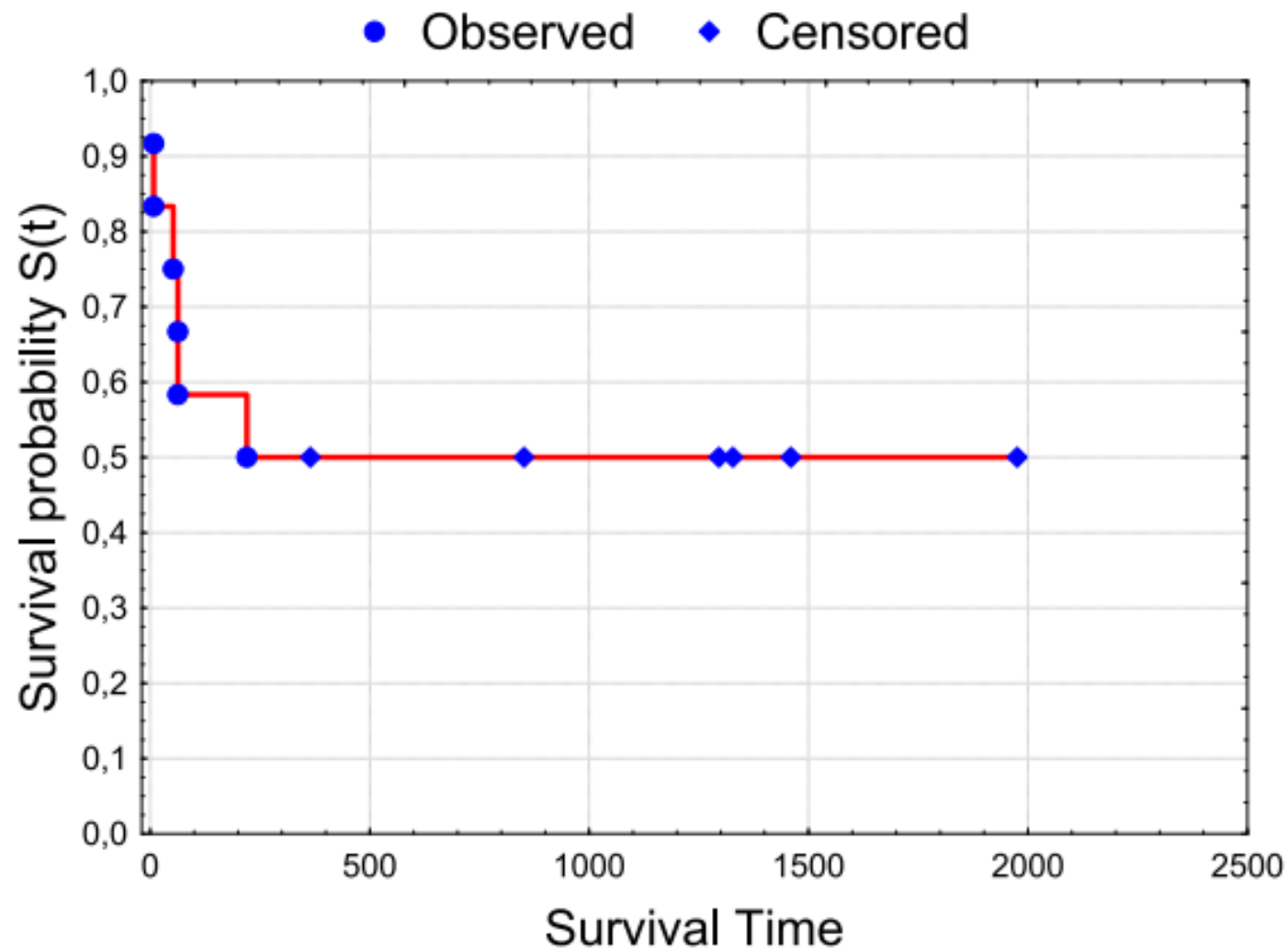
Esempio: non parametric model for Myelomatosis data

Stima di $S(t)$ ottenuta tramite lo stimatore di **Kaplan-Meier** per il gruppo di pazienti sottoposti al trattamento 1.

Time t	# at risk $N(t)$	#Failures	#Censored	Proportion dying	$\hat{S}(t)$
0	12	0	0	0	1.00
5	12	0	0	0/12	$1.00 \times (1 - 0/12) = 1.00$
8	12	2	0	2/12	$1.00 \times (1 - 2/12) = 0.83$
10	10	0	0	0/10	$0.83 \times (1 - 0/10) = 0.83$
52	10	1	0	1/10	$0.83 \times (1 - 1/10) = 0.75$
63	9	2	0	2/9	$0.75 \times (1 - 2/9) = 0.58$
220	7	1	0	1/7	$0.58 \times (1 - 1/7) = 0.50$
300	6	0	0	0/6	$0.50 \times (1 - 0/6) = 0.50$
365	6	0	1	0/6	$0.50 \times (1 - 0/6) = 0.50$
500	5	0	0	0/5	$0.50 \times (1 - 0/5) = 0.50$
852	5	0	1	0/5	$0.50 \times (1 - 0/5) = 0.50$
1296	4	0	1	0/4	$0.50 \times (1 - 0/4) = 0.50$
1460	3	0	1	0/3	$0.50 \times (1 - 0/3) = 0.50$
1976	2	0	1	0/2	$0.50 \times (1 - 0/2) = 0.50$
1328	1	0	1	0/1	$0.50 \times (1 - 0/1) = 0.50$

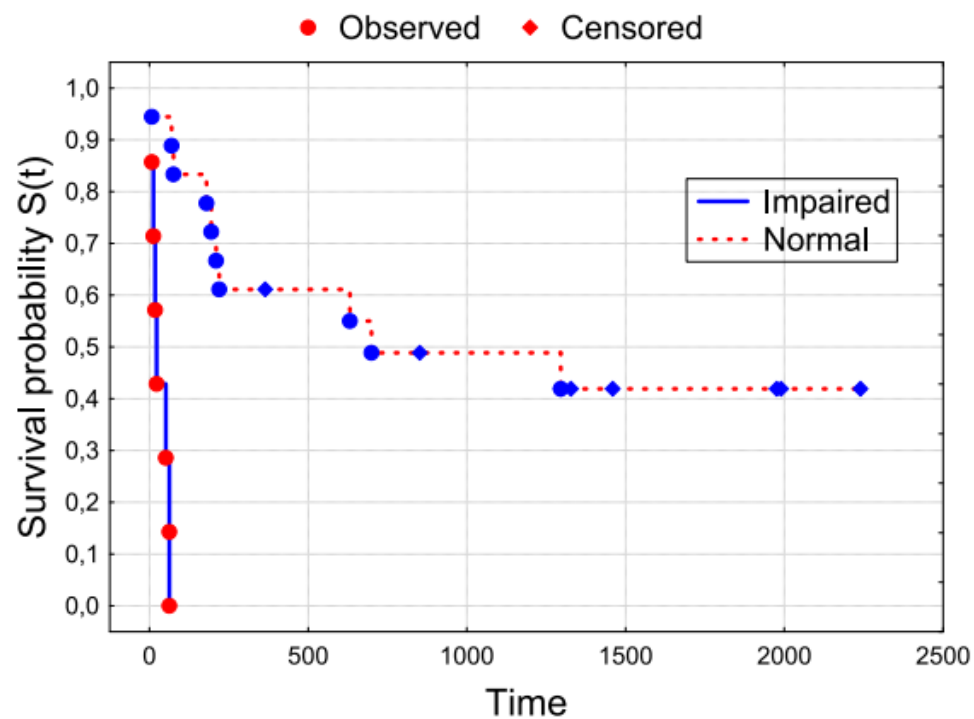
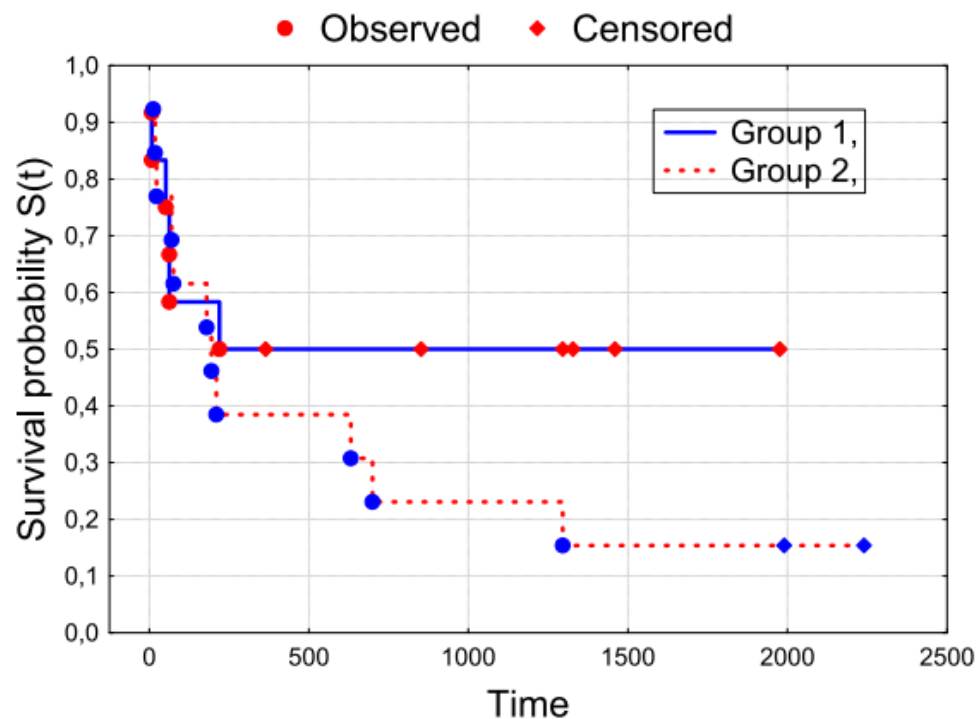
Esempio: non parametric model for Myelomatosis data

Grafico risultante dalla stima di **Kaplan-Meier** per i pazienti sottoposti al trattamento 1, dove i rombi corrispondono ai dati censurati.



Esempio: non parametric model for Myelomatosis data

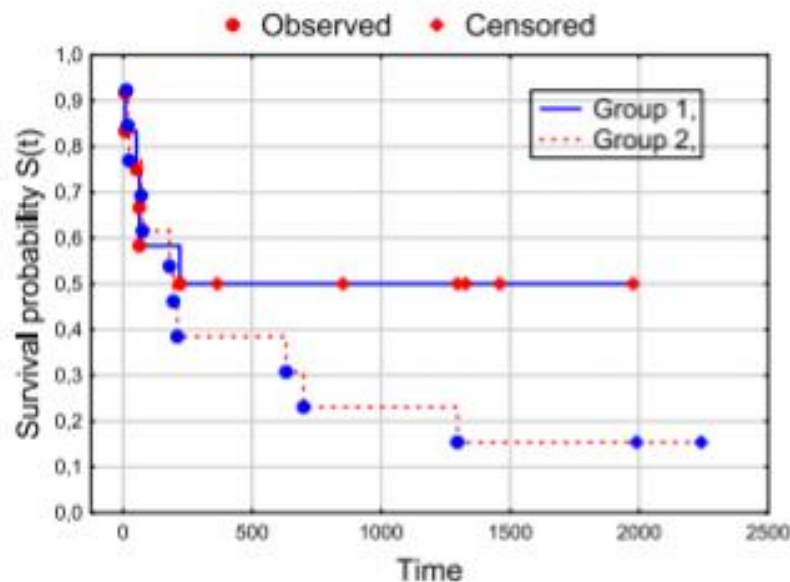
Lo stesso lavoro può essere fatto anche per il gruppo di pazienti sottoposti al trattamento 2, così come lo si potrebbe fare **raggruppando** i pazienti in base al tipo di funzionamento dei reni. In questo modo otterremo due diversi raggruppamenti ("Group 1 o Group 2" e "impaired o normal") con i seguenti risultati:



Esempio: non parametric model for Myelomatosis data

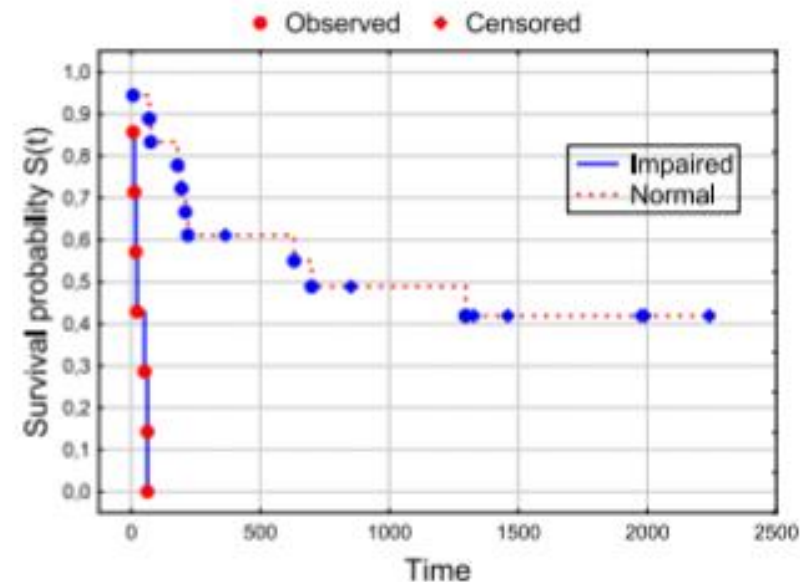
A questo punto si può usare il **Log Rank Test** per confrontare le curve di sopravvivenza fra i differenti gruppi, accettando o rifiutando l'ipotesi di stessa funzione di sopravvivenza in base al livello di confidenza desiderato.

Effect of treatment



Logrank: $p=0.2468$

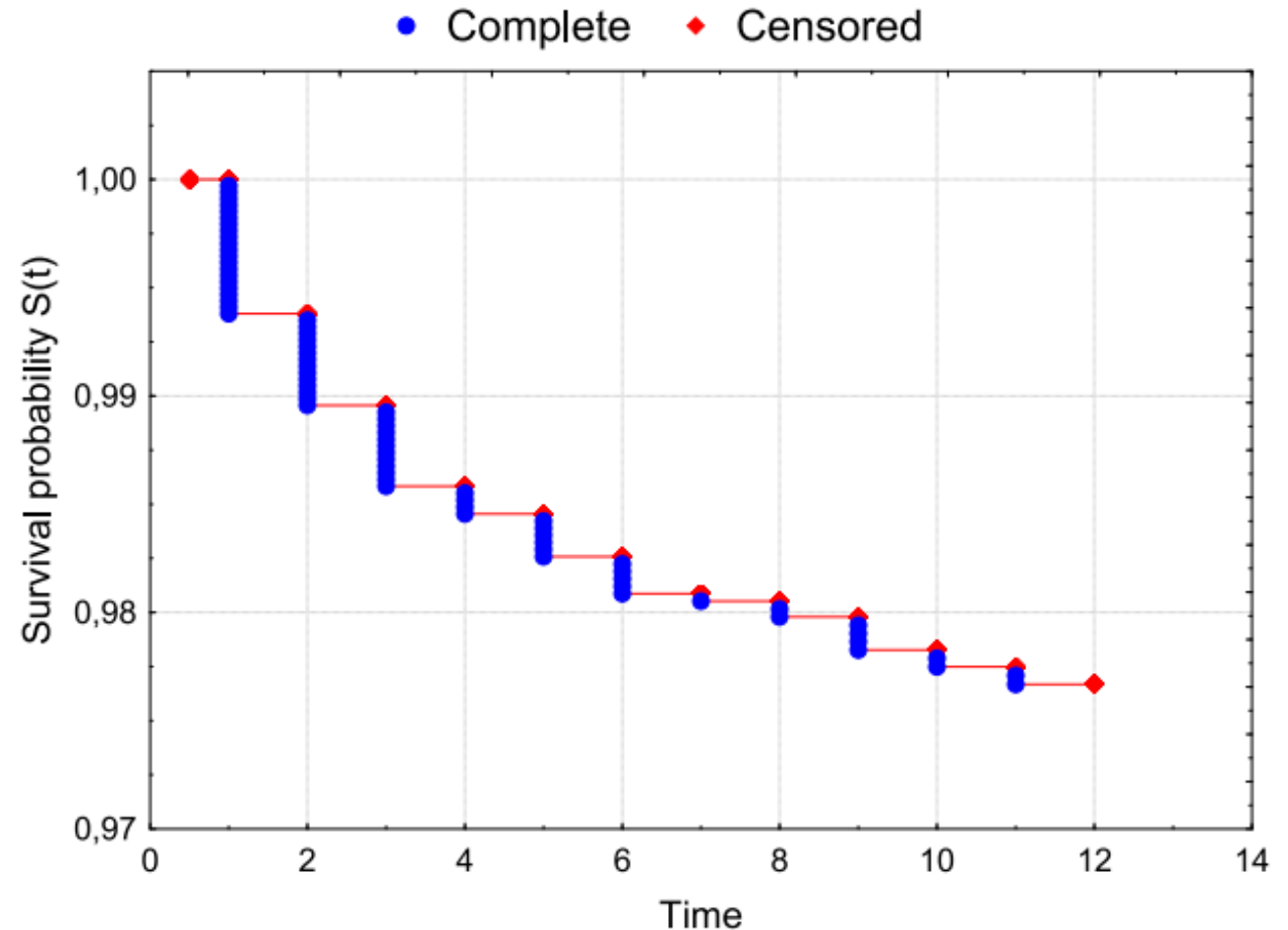
Effect of renal functioning



Logrank: $p=0.0029$

Esempio: Cox Model for Pneumonia data

In questo esempio si parte invece da un Data Set di 3470 bambini per studiare l'influenza di alcuni possibili **fattori di rischio** (considerabili come **covariate** che danno luogo a **effect parameters**) sulla probabilità di registrare un evento dovuto a Pneuomonia (Polmonite). Nel grafico si vede la stima della sopravvivenza per il campione in analisi.



Esempio: Cox Model for Pneumonia data

- Data la volontà di studiare l'effetto di alcuni fattori sulla sopravvivenza degli individui del campione si utilizza un modello di Cox.
- I fattori analizzati sono :
 - Age of mother (Years),
 - Presence of siblings (Yes: 48%, No: 52%),
 - Smoking status mother (Yes: 34%, No: 66%),
 - Urban environment (Yes: 76%, No: 24%),
 - Alcohol use mother (Yes: 36%, No: 64%),
 - Poverty status mother (Yes: 36%, No: 74%),
 - Normal birthweight child (≥ 5.5 pounds ≈ 2.5 kg. Yes: 92%, No: 8%).
- Ad ogni fattore corrisponderà un **parametro** β di influenza.
 - Se $\beta > 0 \rightarrow$ hazard crescenti sotto l'influenza di quella covariata (fattore).
 - Se $\beta < 0 \rightarrow$ hazard decrescenti sotto l'influenza di quella covariata.

Esempio: Cox Model for Pneumonia data

La tabella riporta l'effetto sulla sopravvivenza ottenuto considerando ogni covariata **singolarmente e valutandone il parametro di effetto** tramite la funzione verosimiglianza del modello di Cox.

Si nota, ad esempio, che:

- L'età della madre **influisce positivamente**, ovvero rende gli hazard decrescenti (una madre più anziana darà al mondo un figlio che avrà una probabilità minore di riscontrare polmonite rispetto ad uno con una madre più giovane)
- Il fatto che la madre fumi **influisce invece negativamente**, rendendo gli hazard crescenti (una madre fumatrice darà al mondo un figlio con una probabilità di riscontrare polmonite maggiore rispetto a quella del figlio di una non fumatrice).

Effect	$\hat{\beta}$	p-value
Age of mother	−0.0985	0.0275
Urban environment	−0.4523	0.0695
Alcohol use	−0.0535	0.8282
Normal birthweight child	−0.2412	0.5439
Smoking of mother	0.7958	0.0007
Poverty of mother	0.5963	0.0109
Presence of siblings	0.6436	0.0079

Esempio: Cox Model for Pneumonia data

Dopo aver considerato l'effetto delle covariate singolarmente, si prende in considerazione un modello in cui la **sopravvivenza è influenzata da tutte le covariate.**

Questo modello risulta decisamente più attendibile perché uniforma i dati.

Ad esempio:

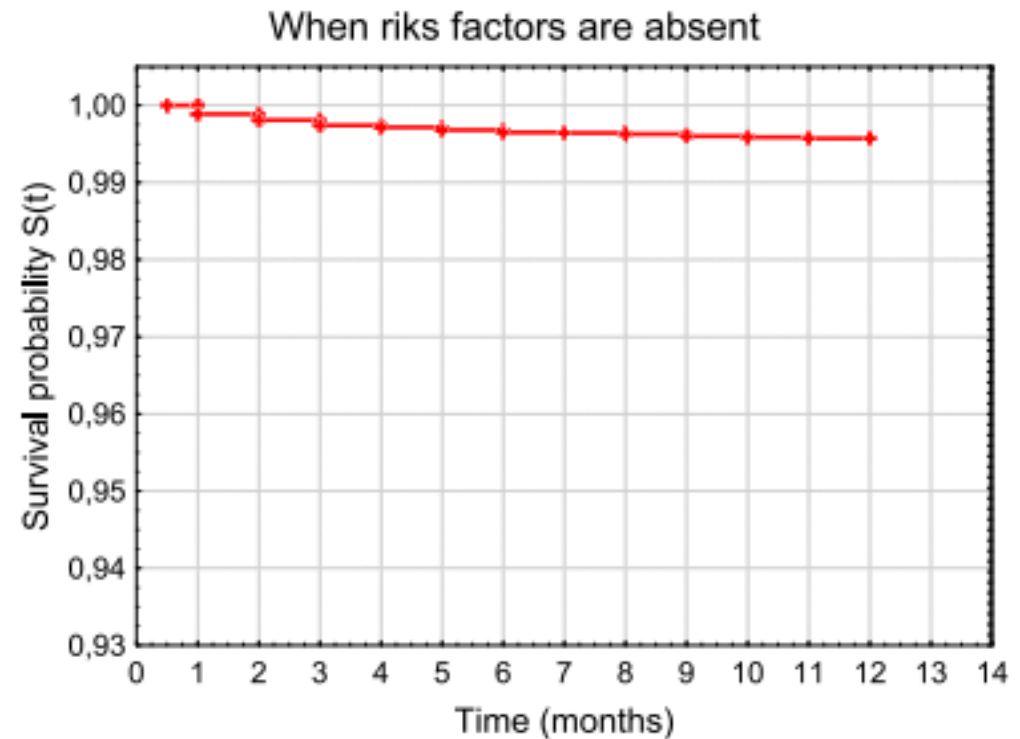
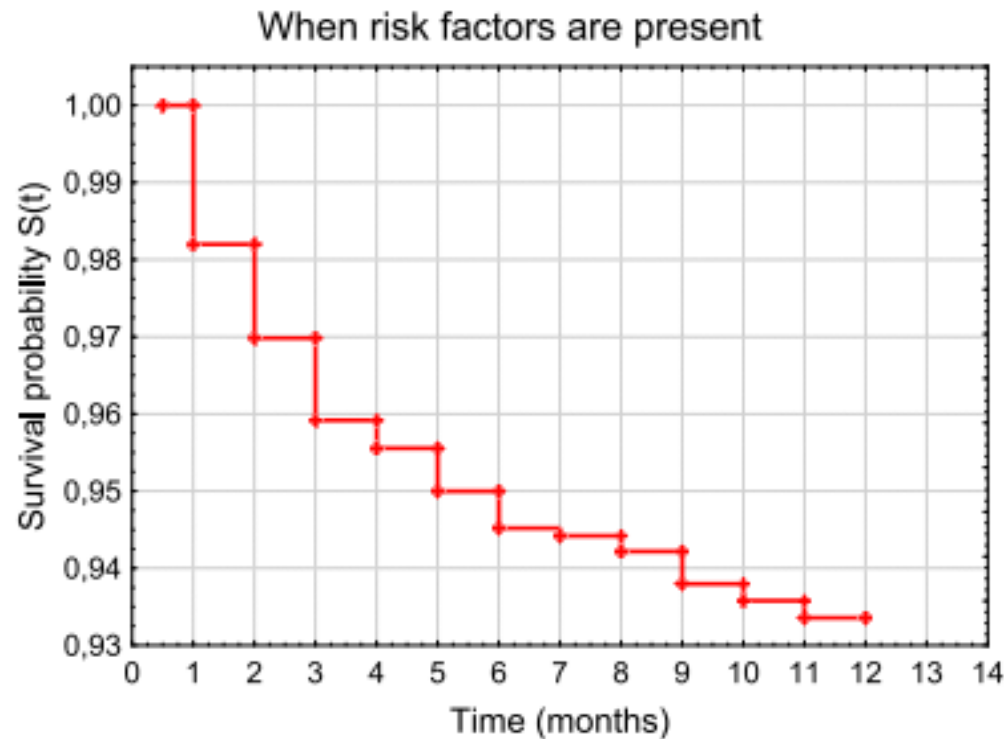
- Non si nota più un'influenza negativa così elevata data dalla povertà della madre.
 - Ci sono più madri fumatrici povere (40,30% contro il 30,69%), il che rende l'effetto congiunto e non dovuto semplicemente alla povertà.

Effect	Simple models		Multiple model	
	$\hat{\beta}$	p-value	$\hat{\beta}$	p-value
Age of mother	-0.0985	0.0275	-0.1287	0.0107
Urban environment	-0.4523	0.0695	-0.3509	0.1616
Alcohol use	-0.0535	0.8282	-0.1213	0.6374
Normal birthweight child	-0.2412	0.5439	-0.0152	0.9697
Smoking of mother	0.7958	0.0007	0.7289	0.0028
Poverty of mother	0.5963	0.0109	0.2778	0.2586
Presence of siblings	0.6436	0.0079	0.7557	0.0042

Esempio: Cox Model for Pneumonia data

Si può valutare l'effetto dei fattori di rischio confrontando le curve di sopravvivenza quando determinati rischi sono presenti o assenti e tenendo costanti gli altri. Nell'esempio :

- Presenti = madre di 20 anni, fumatrice e con altri figli
- Assenti = madre di 30 anni, non fumatrice e senza altri figli.



Applicazioni: package R

- Utilizzando R si possono sfruttare diversi package:
 - Librerie complete per l'**analisi di sopravvivenza**
 - survival
 - flexsurv
 - survsim
 - Olsurv (contiene ed estende survival e kmsurv)
 - Libreria per le **life table** e alcuni **data sets**
 - kmsurv
 - Librerie per i **survival tree**
 - rpart
 - Librerie per le **survival random forest**
 - RandomForestSRC
 - randomSurvivalForest

Referenze

Raccolta delle risorse utilizzate escludendo slide, siti internet e wikipedia:

- David G. Kleinbaum, Mitchel Klein, *Survival Analysis: a self-Learning Text*, Second Edition, Springer.
- David W. Hosmer, Stanley Lemeshow, *Applied Survival Analysis: Regression Modeling of Time to Event Data*.
- Frank E. Harrell, Jr., *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis*, Second Edition, Springer.
- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, *An Introduction to Statistical Learning with applications in R*, Springer.
- Germán Rodríguez, *Non-Parametric Estimation in Survival Models*, Spring, 2001; revised Spring 2005.

Referenze

- Germán Rodríguez, *Parametric Survival Models*, Spring, 2001; revised Spring 2005, Summer 2010.
- Kevin J. Carroll, *On the use and utility of the Weibull model in the analysis of survival data*, AstraZeneca Pharmaceuticals, Biostatistics Group, Global Clinical Science, Alderley Park, Macclesfield, UK.
- William R. Wessels, *Use of the Weibull versus Exponential to Model Part Reliability*, PhD, PE, CRE, CQE, University of Alabama in Huntsville.
- Matthew Witten, William Satzer, *Gompertz Survival Model Parameters: estimation and sensitivity*, *Appl. Math. Lett.* Vol. 5, No. 1, pp. 7-12, 1992.
- Medhat Mohamed Ahmed Abdelaal, Sally Hossam Eldin Ahmed Zakria, *Modeling Survival Data by Using Cox Regression Model*. American Journal of Theoretical and Applied Statistics. Vol. 4, No. 6, 2015, pp. 504-512.

Referenze per R

- Mai Zhou, *Use Software R to do Survival Analysis and Simulation. A tutorial*, Department of Statistics, University of Kentucky.
- Professor Mara Tableman, *Survival Analysis Using S/R*, Fariborz Maseeh Department of Mathematics & Statistics, Portland State University, Portland, Oregon, USA, August–September 2012.
- John Fox, *Cox Proportional-Hazards Regression for Survival Data: Appendix to An R and S-PLUS Companion to Applied Regression*, February 2002.
- Brian S. Everitt and Torsten Hothorn, *A Handbook of Statistical Analyses Using R*.
- Christopher H. Jackson, *flexsurv: A Platform for Parametric Survival Modeling in R*, MRC Biostatistics Unit.

Referenze per R

- David Moriña, CREAL, Albert Navarro, Universitat Autònoma de Barcelona, *The R Package survsim for the Simulation of Simple and Complex Survival Data*.
- David M Diez, *Survival Analysis in R*, June 2013.
- Terry M. Therneau, Elizabeth J. Atkinson, Mayo Foundation, *An Introduction to Recursive Partitioning Using the RPART Routines*, June 29, 2015.
- Maja Pohar, Janez Stare, *Relative survival analysis in R*, Department of Medical Informatics, University of Ljubljana, Slovenia