METODI STATISTICI PER LA BIOINGEGNERIA (B)

PARTE 14: ANALISI DI SOPRAVVIVENZA

A.A. 2024-2025

Prof. Martina Vettoretti

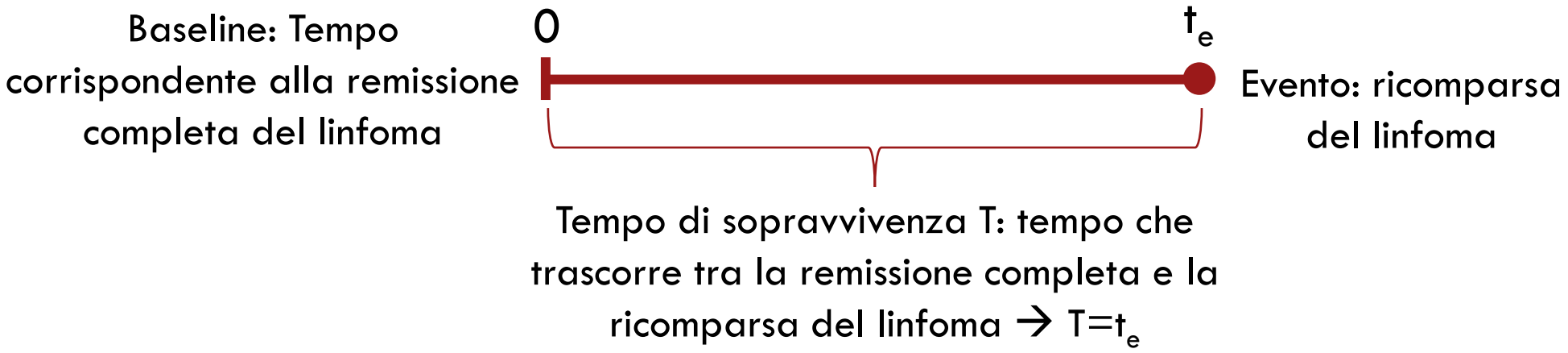
ANALISI DI SOPRAVVIVENZA

- L'**analisi di sopravvivenza**, o **survival analysis**, comprende una serie di metodi statistici per analizzare se sussiste una relazione tra una o più variabili indipendenti ed una outcome che rappresenta il **tempo ad un evento di interesse**, che si assume essere irreversibile.
  - Esempi:
    - Tempo alla rottura di un dispositivo
    - Tempo alla morte di un paziente
    - Tempo all'insorgenza di una patologia
    - Tempo alla necessità di sostituire un impianto protesico

	Regressione lineare	**Regressione logistica**	**Analisi di sopravvivenza**
Tipo di outcome	Variabile quantitativa	Variabile qualitativa binaria	Tempo ad un evento

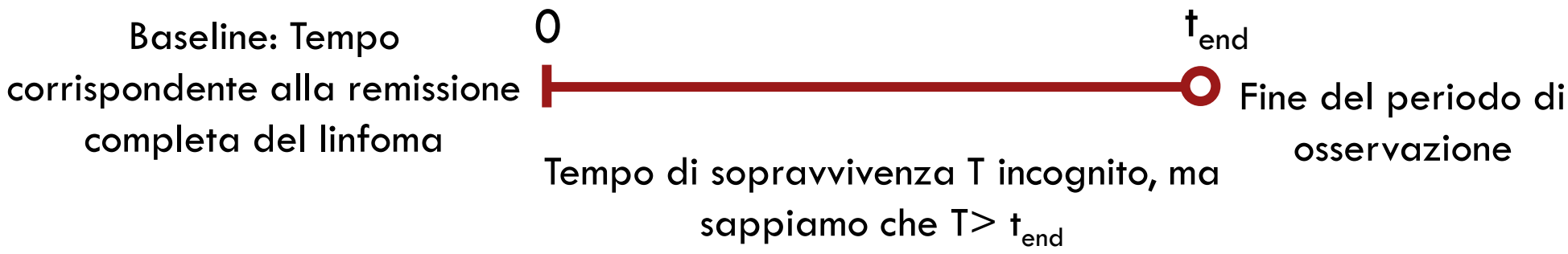
TEMPO DI SOPRAVVIVENZA

- Il tempo dell'evento di interesse, misurato rispetto ad un tempo 0 iniziale, viene chiamato **tempo di sopravvivenza** (*survival time*) oppure **tempo all'evento** (*time to event*).
- Il tempo 0 è normalmente l'inizio del periodo di osservazione o l'inizio dell'esperimento → chiamato anche *baseline*
- Esempio. Vogliamo studiare i fattori che influenzano la recidiva di un particolare tipo di linfoma. → Reclutiamo un insieme di soggetti affetti da linfoma per cui il trattamento ha portato ad una remissione completa della malattia e li monitoriamo nel tempo.



CENSORING

- L'evento di interesse potrebbe verificarsi durante il periodo di osservazione solo per alcuni elementi del campione analizzato.
  - Esempio. Molti soggetti guariti dal linfoma fortunatamente non avranno una recidiva. Alcuni pazienti potrebbero avere una recidiva dopo la fine del periodo di osservazione.
- Quando **l'evento non si verifica** durante il periodo di osservazione considerato, il tempo all'evento è incognito (non sappiamo se e quando l'evento si verificherà).
- Tuttavia sappiamo che fino ad un certo istante, tipicamente la fine del periodo di monitoraggio o dell'esperimento, l'evento non si è verificato.
- Dati di questo tipo vengono detti **censurati o *censored***.



PERCHE' NON UNA REGRESSIONE LINEARE?

- Il tempo ad un evento è di fatto una variabile quantitativa.
- Potremmo pensare di affrontare il problema mediante una regressione lineare avente come uscita il tempo all'evento:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m + \varepsilon$$

Tempo all'evento

Variabili esplicative

- Problema: la regressione lineare non può gestire i dati censurati!

PERCHE' NON UNA REGRESSIONE LOGISTICA?

- Potremmo pensare di affrontare il problema utilizzando una regressione logistica per predire se l'evento si verificherà o meno in un certo intervallo temporale.

$$\log\left(\frac{p}{1-p}\right) = \boldsymbol{\beta}^T \boldsymbol{X} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m$$

Probabilità che l'evento si verifichi

Variabili esplicative

- Problema: la regressione logistica non considera i tempi degli eventi! Potrà predire se l'evento si verifica o no, ma non quando si verifica.

OBIETTIVI DELL'ANALISI DI SOPRAVVIVENZA

- Se vogliamo studiare il tempo ad un evento di interesse pertanto abbiamo bisogno di altri metodi statistici → metodi dell'analisi di sopravvivenza
- Tre principali obiettivi dell'analisi di sopravvivenza:

1. Stimare il tempo ad un evento per un gruppo di individui
2. Confrontare il tempo ad un evento per due o più gruppi di individui
3. Studiare la relazione tra una o più variabili esplicative e il tempo all'evento

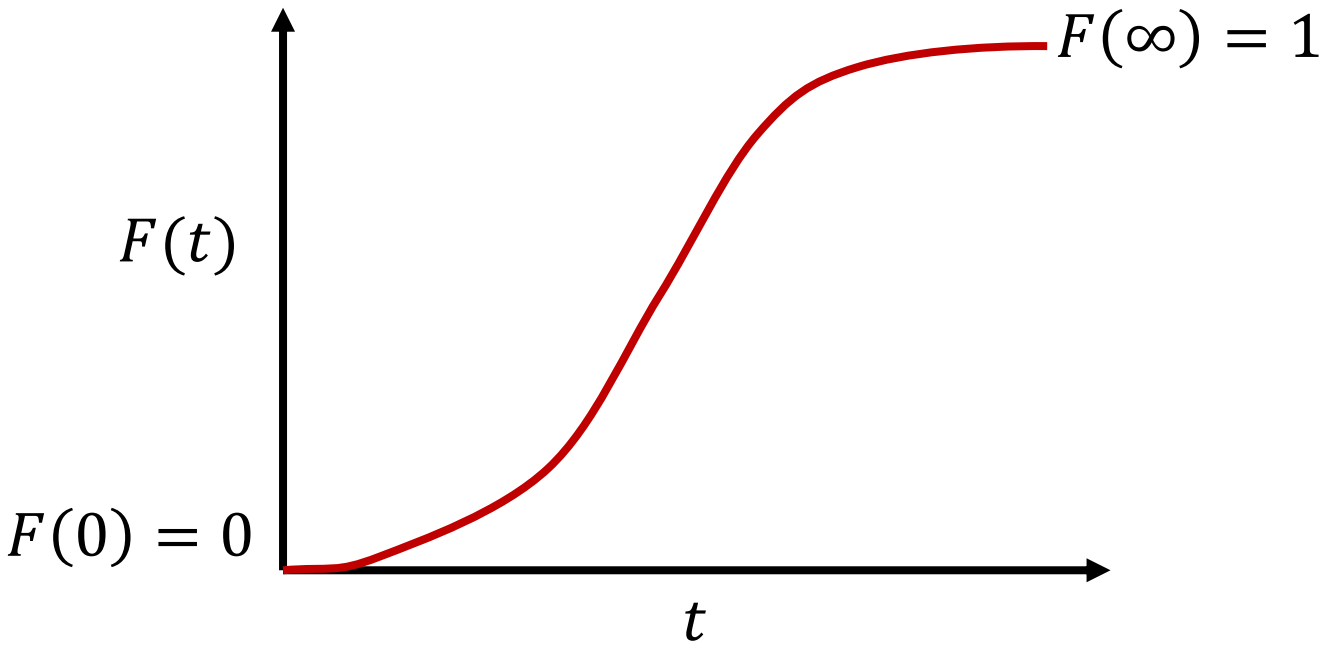
CARATTERIZZAZIONE DEL TEMPO DI SOPRAVVIVENZA

- T: variabile aleatoria che rappresenta il tempo di sopravvivenza di un certo individuo/elemento del campione statistico.
- t: valore osservato per T (una realizzazione).
- 5 funzioni per caratterizzare la distribuzione di T:
  - La funzione di ripartizione di T, o failure function, F(t)
  - La funzione di sopravvivenza, S(t)
  - La densità di probabilità di T, f(t)
  - La hazard function, o funzione di rischio, h(t)
  - La funzione cumulativa di rischio, H(t)

FAILURE FUNCTION

- La funzione di ripartizione di T è detta anche **failure function** (funzione di ripartizione del tempo di vita o *lifetime distribution function*):

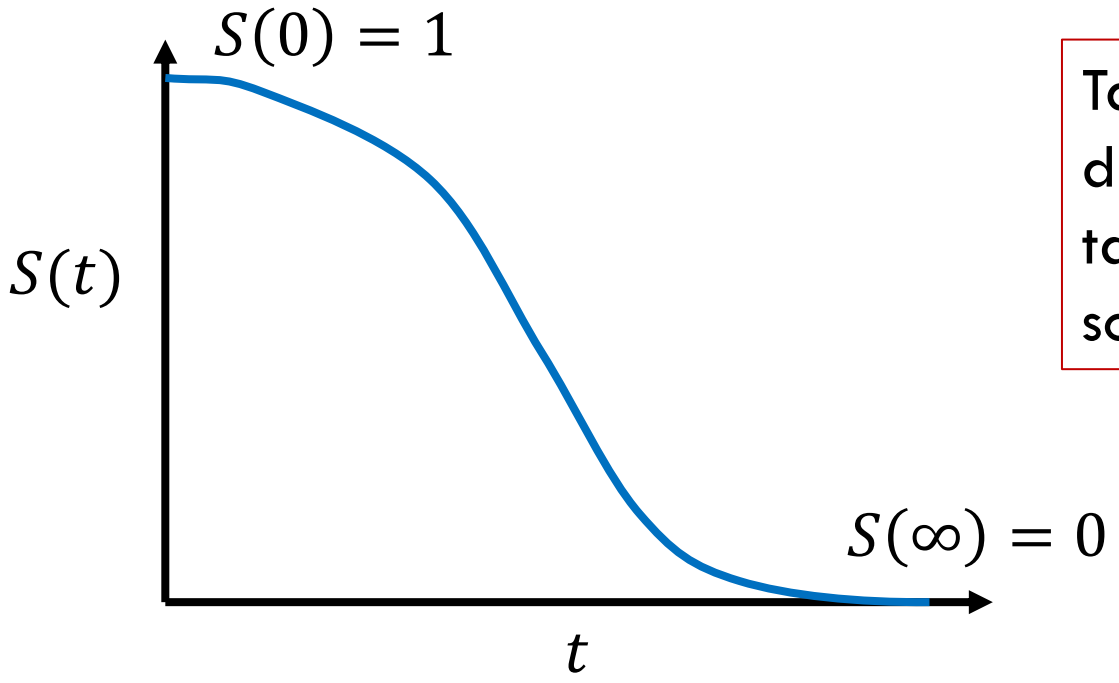
$$F(t) = P(T \leq t)$$



FUNZIONE DI SOPRAVVIVENZA

- **Funzione di sopravvivenza** (*survival function*, a volte *reliability function*):

$$S(t) = P(T > t) = 1 - F(t)$$



Tanto più rapida è la discesa della funzione di sopravvivenza, tanto più breve sarà il tempo di sopravvivenza.

LA DENSITA' DI PROBABILITA' DI T

- **Densità di probabilità** del tempo di sopravvivenza T (*lifetime density function*):

$$f(t) = \frac{dF(t)}{dt} = \frac{d(1 - S(t))}{dt} = -\frac{dS(t)}{dt}$$

assumendo che F(t) sia differenziabile.

HAZARD FUNCTION

- La **hazard function,** $h(t)$**,** detta anche **funzione di rischio,** è definita come:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t \,|T > t)}{\Delta t}$$

- Essa rappresenta la probabilità istantanea per unità di tempo che l'evento occorra al tempo t, sapendo che l'evento non si è verificato fino al tempo t.

LEGAME TRA h(t) E S(t)

- Ricordando la definizione di probabilità condizionata per cui $P(E|F)=P(E\cap F)/P(F)$:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t < T < t + \Delta t)}{\Delta t} \cdot \frac{1}{S(t)}$$

- Ricordando inoltre il significato di densità di probabilità per cui la probabilità che T sia in un intorno di ampiezza Δt intorno a t è circa $f(t) \cdot \Delta t$:

$$h(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)}$$

CUMULATIVE HAZARD FUNCTION

- La **cumulative hazard function,** $H(t)$, detta anche funzione cumulativa di rischio, è definita come l'integrale della hazard function, $h(t)$, nell'intervallo $[0\ t]$:

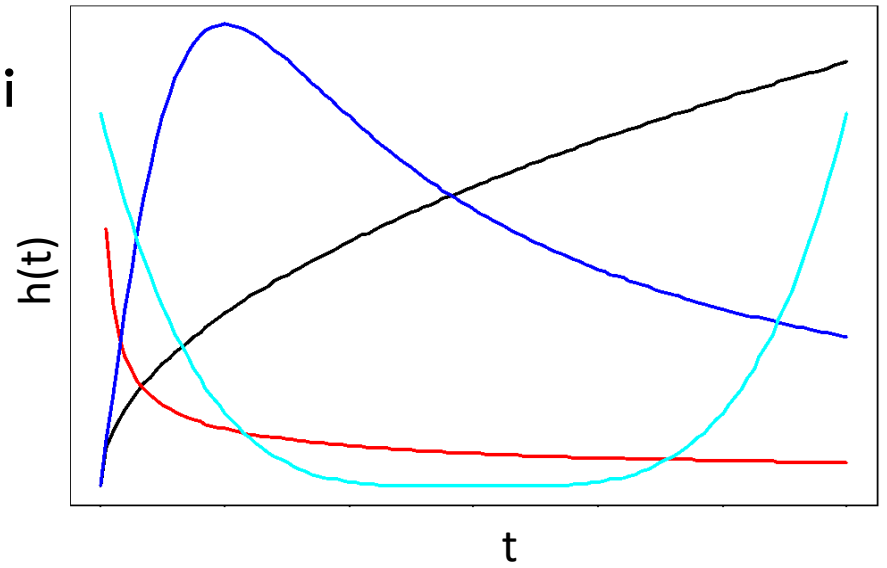
$$\mathrm{H}(t) = \int_0^t h(u)du = -\int_0^t \frac{S'(u)}{S(u)}du = -\log(S(t))$$

$$S(t) = e^{-H(t)}$$

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

NOTE SULLA HAZARD FUNCTION

- La hazard function non è una probabilità, ma una probabilità istantanea normalizzata per l'intervallo temporale Δt con $\Delta t \to 0$.
- Essa quindi assume valori positivi, senza limiti superiori: $h(t) \geq 0$.
- L'andamento nel tempo di $h(t)$ può essere di qualsiasi tipo (crescente, decrescente, crescente e poi decrescente…).
  - $h(t)$ crescente → situazione tipica quando l'evento di interesse è legato all'usura/invecchiamento per cui più passa il tempo maggiore è il rischio che si verifichi l'evento (es. rottura di un dispositivo).
  - $h(t)$ decrescente → situazione tipica quando più passa il tempo più il rischio che si verifichi l'evento diminuisce (es. l'evento è una complicanza post intervento chirurgico).



FORMULARIO DELL'ANALISI DI SOPRAVVIVENZA

- Funzione sopravvivenza: $S(t) = 1 - F(t) = \int_t^{+\infty} f(u)du$
- Densità del tempo di sopravvivenza: $f(t) = -\frac{dS(t)}{dt}$
- Hazard function: $h(t) = \frac{f(t)}{S(t)}$
- Cumulative hazard function: $\mathrm{H}(t) = \int_0^t h(u)du$
- $H(t) = -\log(S(t))$
- $h(t) = -\frac{d}{dt}(\log(S(t)))$
- $S(t) = e^{-H(t)}$
- $f(t) = h(t)\, e^{-H(t)}$

OBIETTIVI DELL'ANALISI DI SOPRAVVIVENZA

- Se vogliamo studiare il tempo ad un evento di interesse pertanto abbiamo bisogno di altri metodi statistici → metodi dell'analisi di sopravvivenza
- Tre principali obiettivi dell'analisi di sopravvivenza:

1. Stimare il tempo ad un evento per un gruppo di individui
2. Confrontare il tempo ad un evento per due o più gruppi di individui
3. Studiare la relazione tra una o più variabili esplicative e il tempo all'evento

CURVA DI SOPRAVVIVENZA

- Vogliamo analizzare il tempo ad un evento di interesse per una popolazione di individui.
  - Raccogliamo un set di dati su un campione della popolazione (tempi degli eventi + informazioni sul censoring).
  - Analizziamo i dati del campione mediante tecniche di analisi di sopravvivenza che consentono di stimare la funzione di sopravvivenza, $S(t)$, della popolazione di appartenenza.
  - Indichiamo con $\hat{S}(t)$ la stima di $S(t)$ ottenuta analizzando i dati del campione. Essa è anche chiamata **curva di sopravvivenza**.

DATI PER L'ANALISI DI SOPRAVVIVENZA (1 /2)

- Campione statistico raccolto su n individui appartenenti alla popolazione.
- Osserviamo gli n individui nel tempo. Ciascun individuo viene osservato fino a quando si verifica l'evento oppure il monitoraggio si interrompe senza che si sia verificato l'evento (censoring).
- Indichiamo con δ_i la funzione che indica se l'evento è avvenuto o meno per l'individuo i-esimo durante il periodo di monitoraggio:

$$\delta_i = \begin{cases} 1 & \text{se si è verificato l'evento} \\ 0 & \text{se non si è verificato l'evento} \end{cases}$$

- Indichiamo con t_i il tempo dell'evento o di censoring per l'individuo i-esimo.

DATI PER L'ANALISI DI SOPRAVVIVENZA (2/2)

n coppie (t_i, δ_i), una per ciascun individuo del campione

- Se $\delta_i = 1 \rightarrow t_i$ = tempo dell'evento per l'individuo i-esimo
- Se $\delta_i = 0 \rightarrow t_i$ = tempo di censoring per l'individuo i-esimo
  - L'individuo i-esimo è stato osservato fino al tempo t_i e l'evento non si è verificato. Non sappiamo per questo individuo se e quando si verificherà l'evento.

METODI PER LA STIMA DELLA CURVA DI SOPRAVVIVENZA

- Metodi non parametrici → si calcola una stima di S(t) senza fare assunzioni sulla densità di probabilità di T
  - Il metodo di Kaplan-Meier

- Modelli parametrici → si assume che T sia una variabile aleatoria avente una certa densità di probabilità (tipicamente esponenziale, di Weibull o lognormale) e si stima S(t) di conseguenza

IL METODO DI KAPLAN-MEIER (KM) (1/2)

- Supponiamo ci siano K tempi di sopravvivenza distinti.

$$t_1 < t_2 < \cdots < t_K$$

- Ad ogni tempo t_j, ci sono:
  - n_j individui a rischio → individui che non hanno avuto l'evento per tempi $t < t_j$ e che sono ancora all'interno del periodo di monitoraggio (non censurati per tempi $t < t_j$).
  - d_j → individui per cui l'evento si verifica al tempo t_j

- Probabilità dell'evento al tempo t_j dato che l'individuo è sopravvissuto fino a t_j:

$$P(T = t_j | T \geq t_j) = \frac{d_j}{n_j}$$

IL METODO DI KAPLAN-MEIER (KM) (2/2)

- Probabilità di sopravvivere al tempo t_j dato che l'individuo è sopravvissuto fino a t_j:

$$P(T > t_j | T \geq t_j) = 1 - \frac{d_j}{n_j}$$

- La stima KM del valore di S(t) al tempo t è il prodotto delle probabilità di essere sopravvissuti ai tempi $t_j \leq t$:

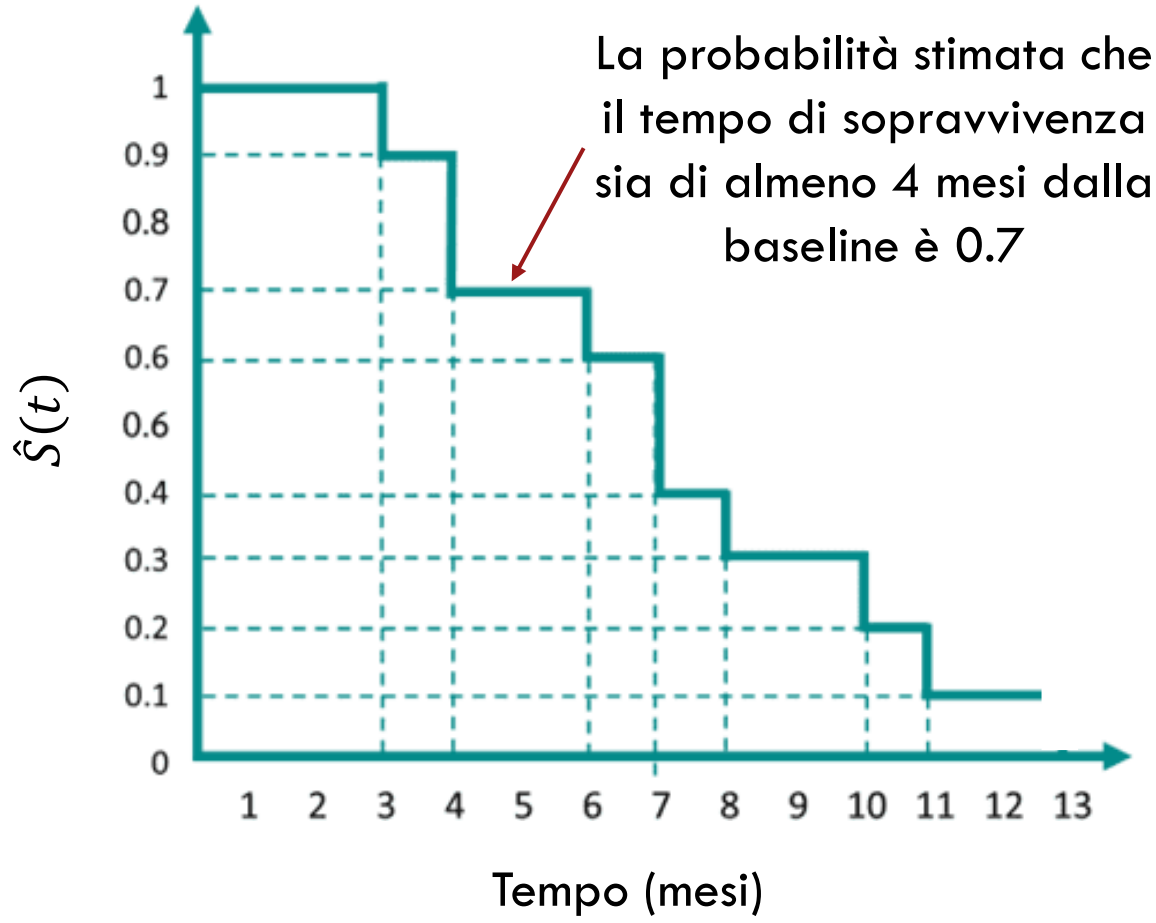
$$\hat{S}(t) = \mathrm{P}\Big(\bigcap_{j:\, t_j \leq t} \{T > t_j | T \geq t_j\}\Big) = \prod_{j:\, t_j \leq t} P(T > t_j | T \geq t_j) = \prod_{j:\, t_j \leq t} [1 - \frac{d_j}{n_j}]$$

CARATTERISTICHE DI $\hat{S}(t)$

$$\hat{S}(t) = \prod_{j:\, t_j \le t} [1 - \frac{d_j}{n_j}]$$

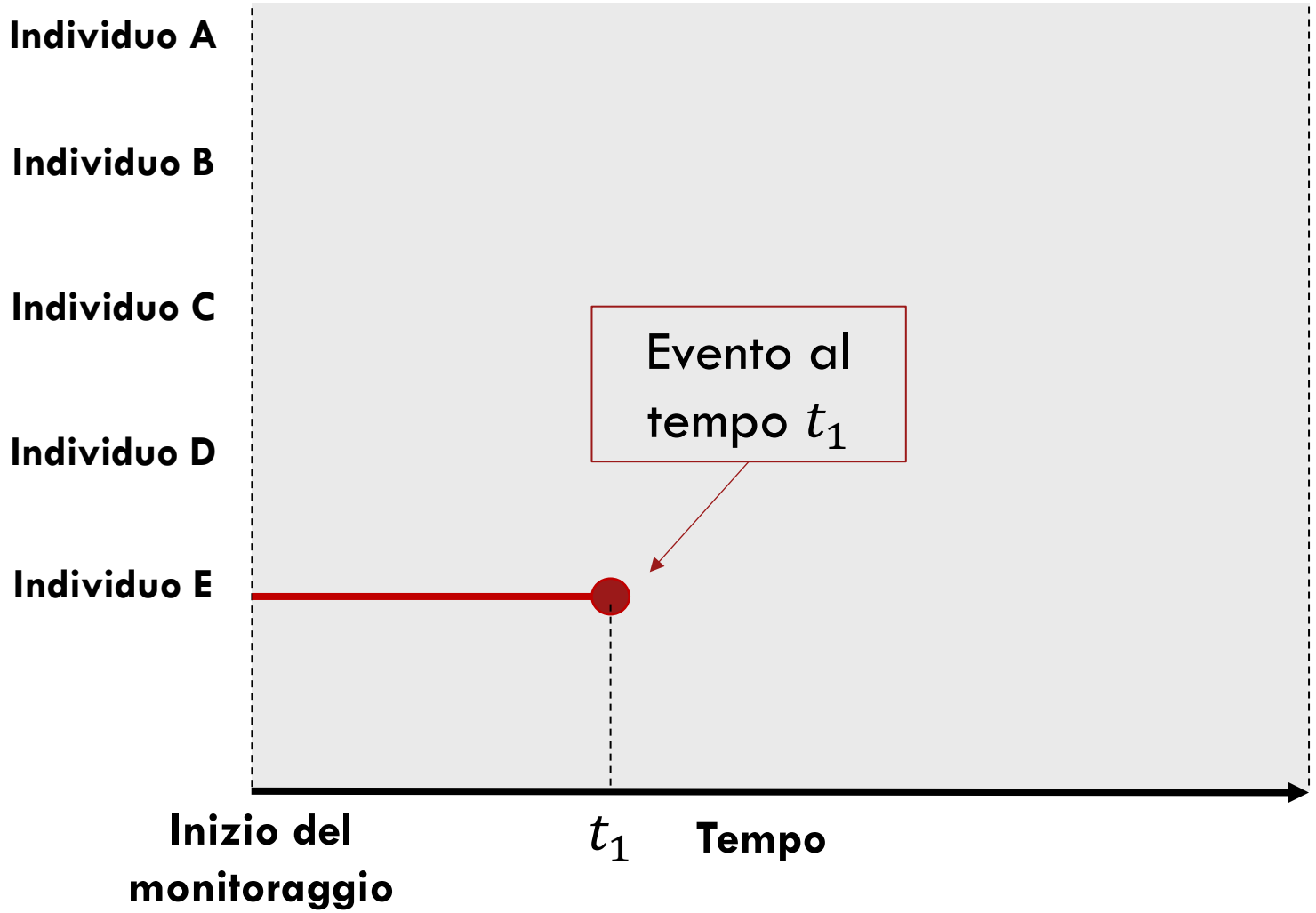
Si aggiunge un nuovo fattore alla produttoria in corrispondenza di ogni nuovo tempo in cui si verifica almeno un evento.

$\hat{S}(t)$ è una funzione decrescente a gradini che cambia valore in corrispondenza dei tempi distinti degli eventi $t_j, j = 1, \dots, K$.

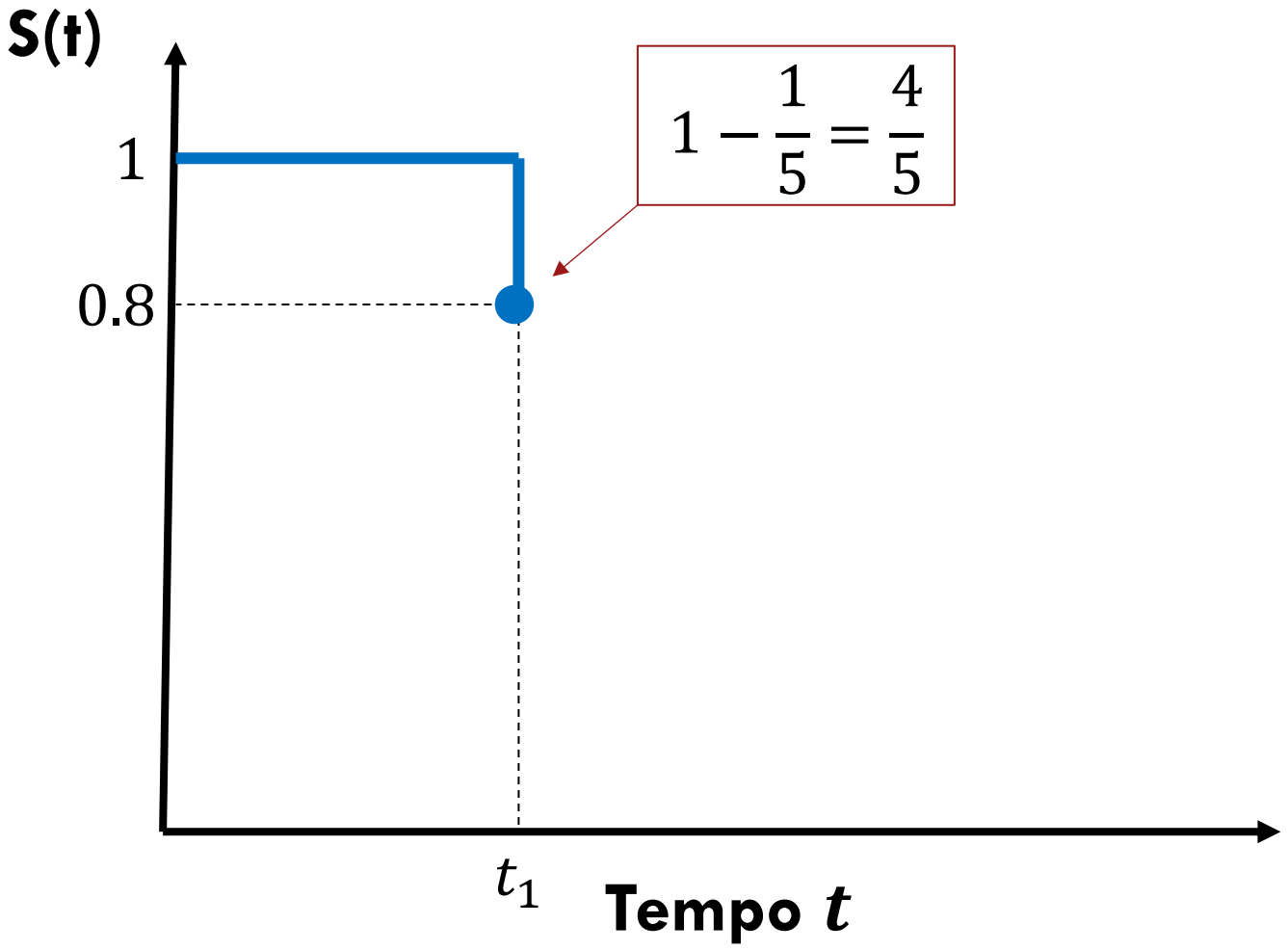


In questo esempio si sono verificati eventi dopo 3, 4, 6, 7, 8, 10, 11 mesi dalla baseline.

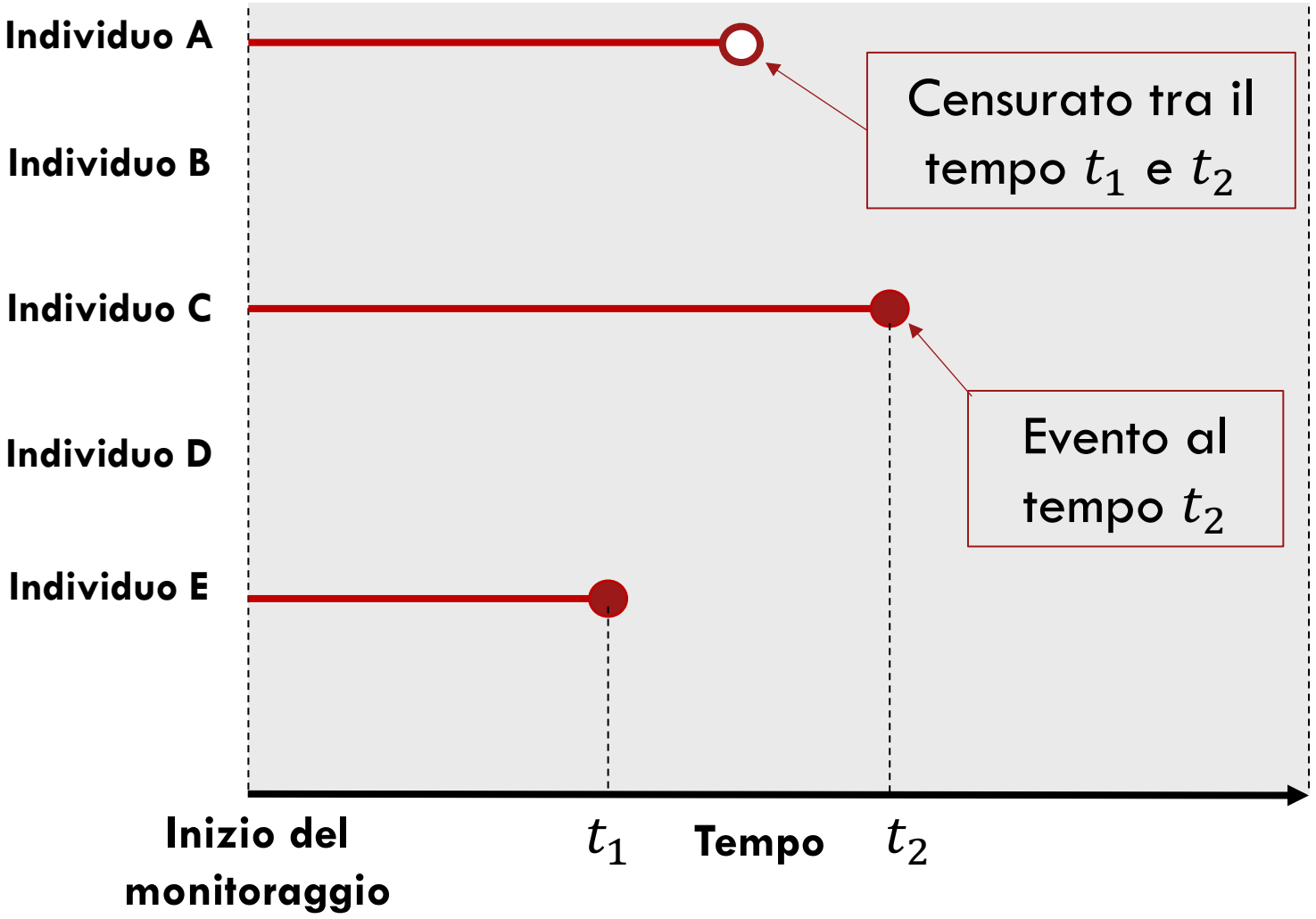
ESEMPIO


Individuo A
Individuo B
Individuo C
Individuo D
Individuo E
Evento al tempo t_1
Inizio del monitoraggio
t_1
Tempo

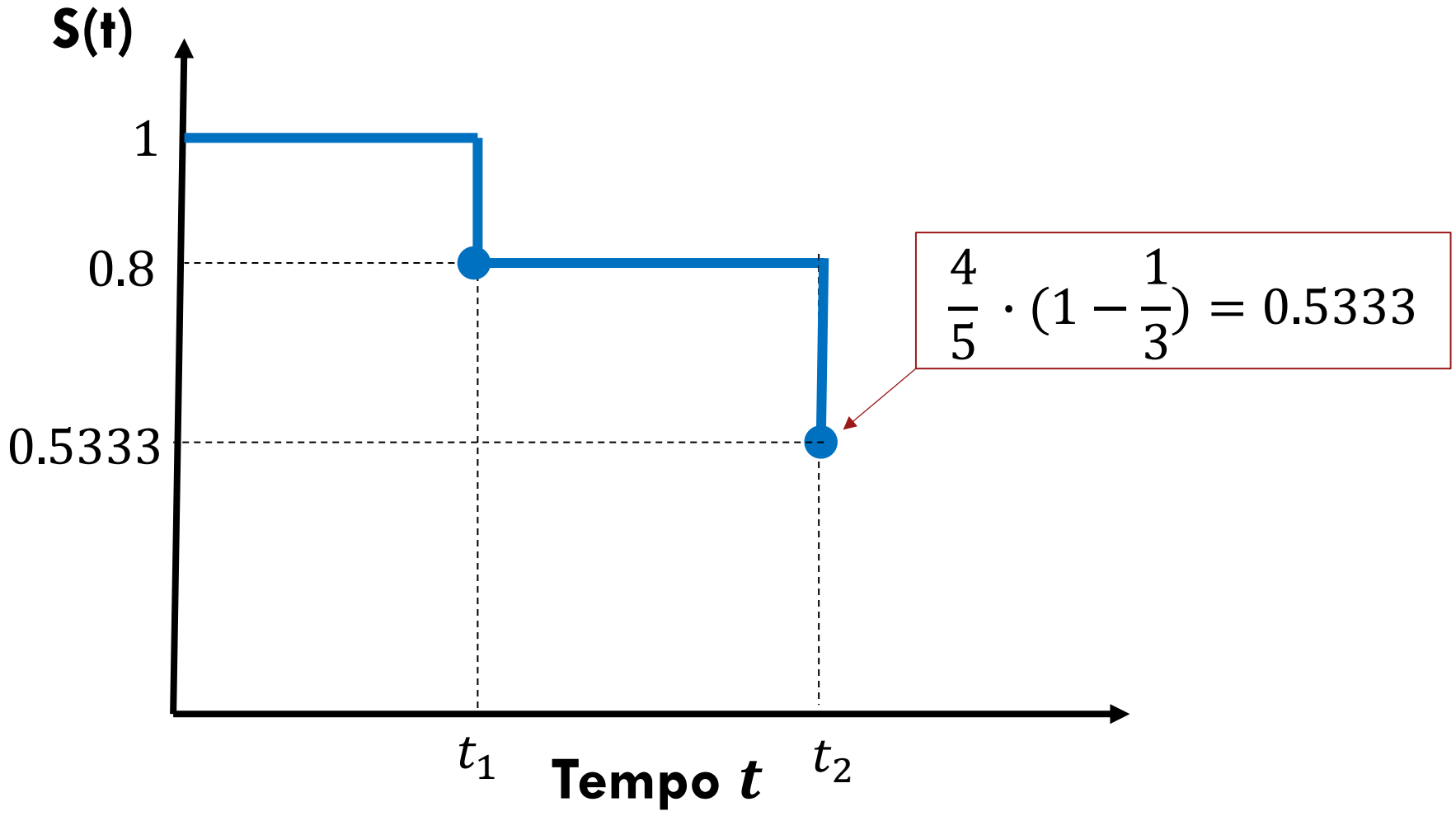
ESEMPIO



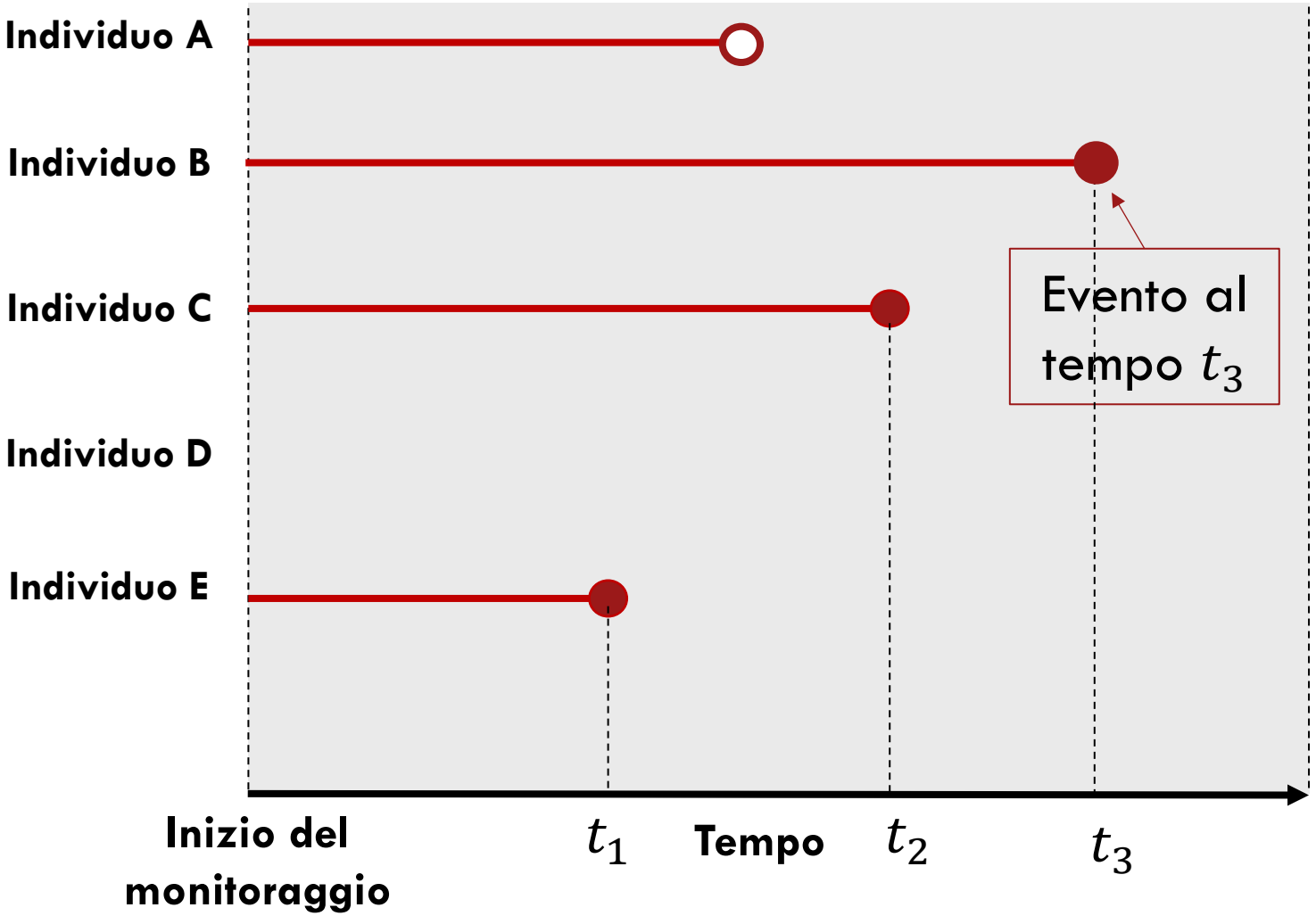
ESEMPIO


Individuo A
Individuo B
Individuo C
Individuo D
Individuo E
Censurato tra il tempo t_1 e t_2
Evento al tempo t_2
Inizio del monitoraggio
t_1
Tempo
t_2

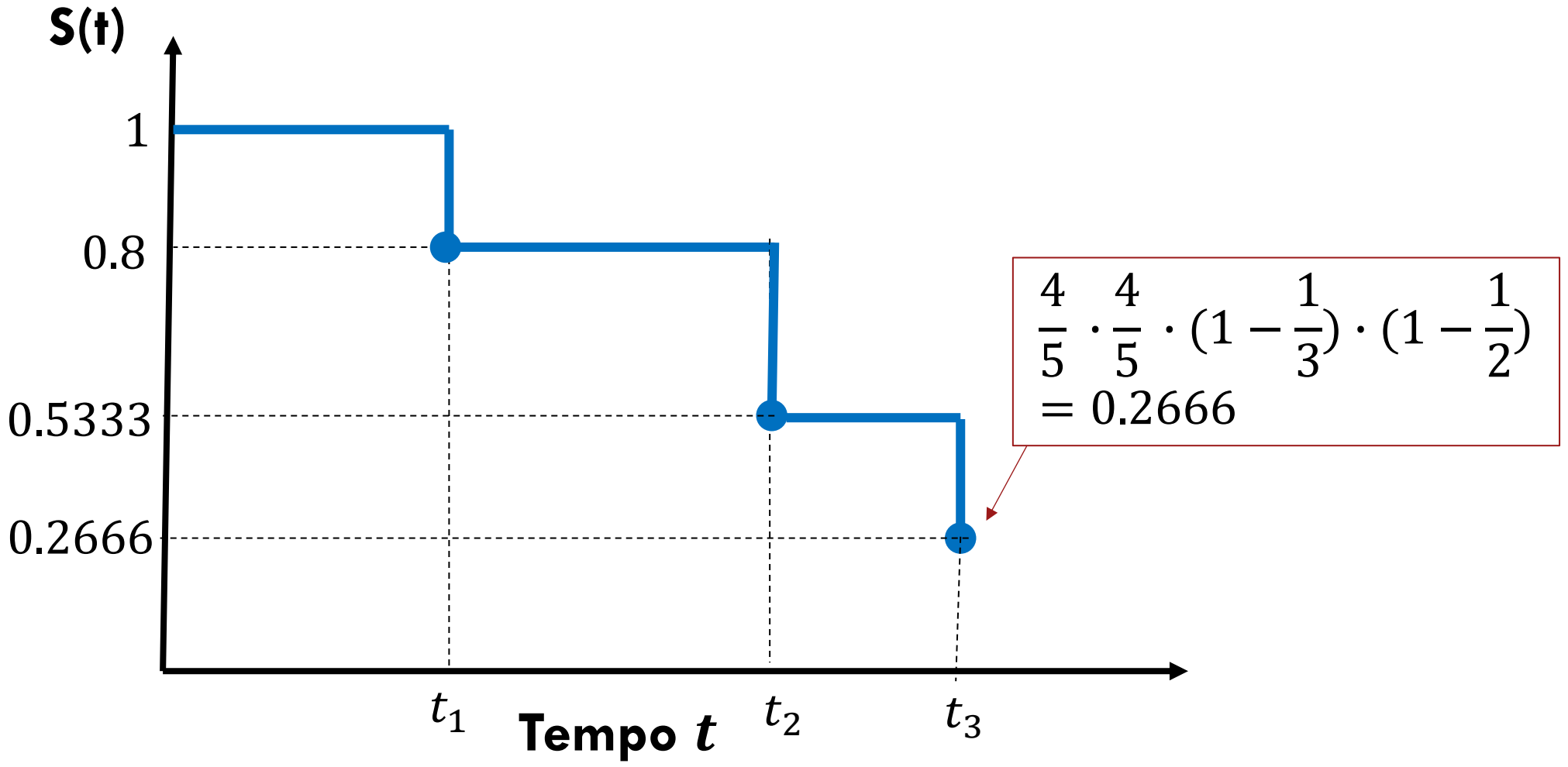
ESEMPIO



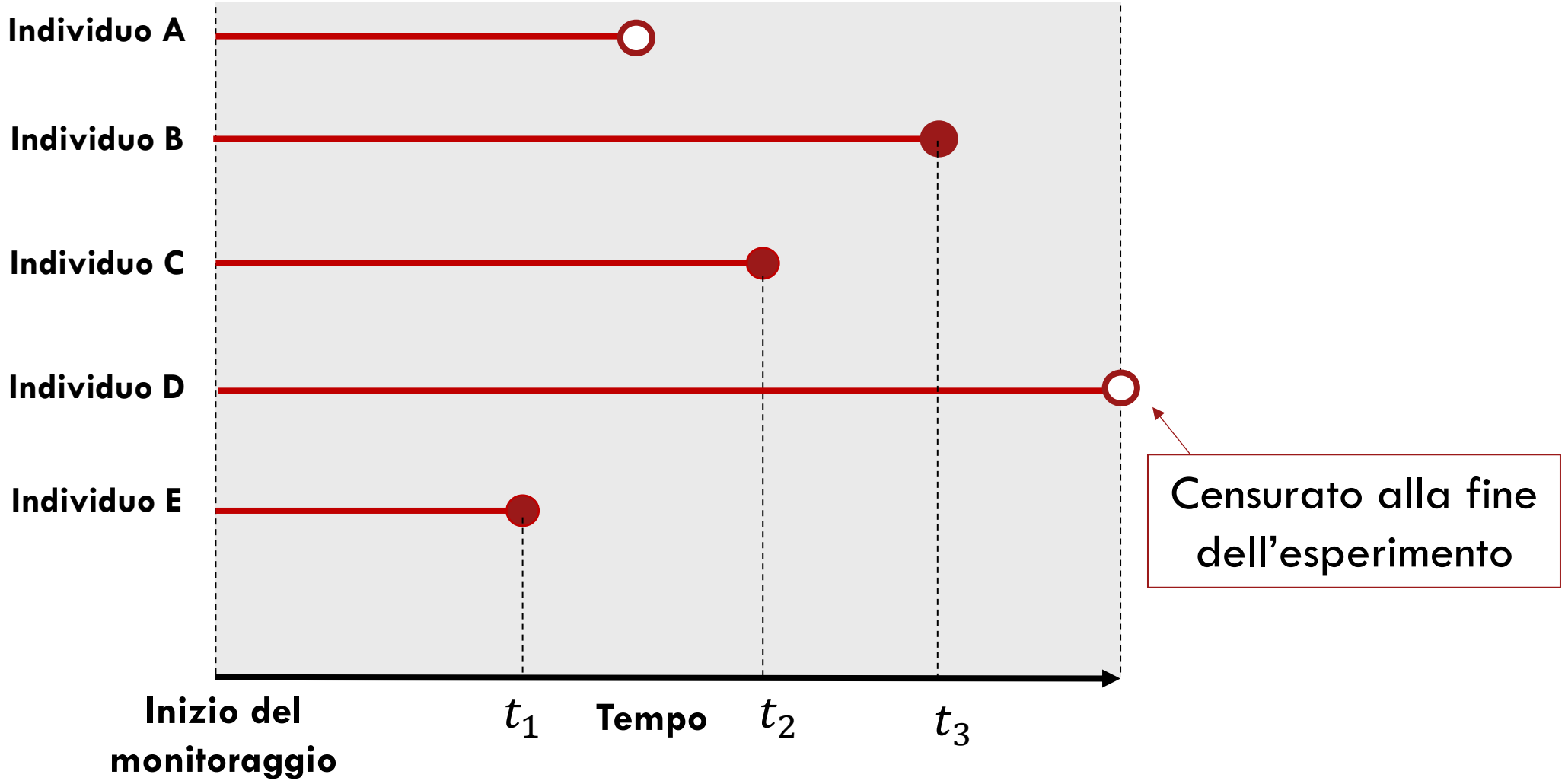
ESEMPIO


Individuo A
Individuo B
Individuo C
Individuo D
Individuo E
Evento al tempo t_3
Inizio del monitoraggio
t_1
Tempo
t_2
t_3

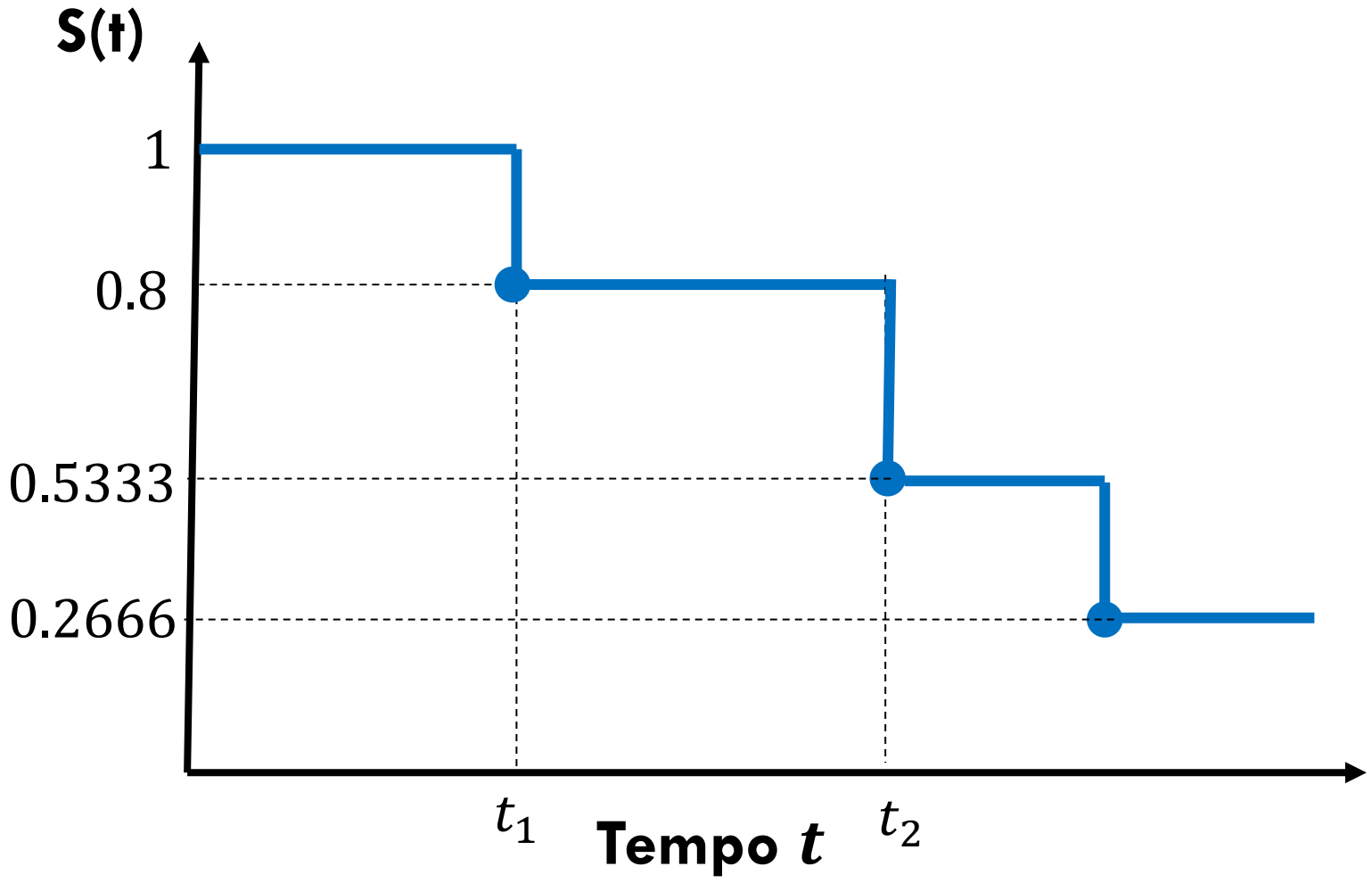
ESEMPIO


S(t)
1
0.8
0.5333
0.2666
t_1
Tempo t
t_2
t_3
$\frac{4}{5} \cdot \frac{4}{5} \cdot (1 - \frac{1}{3}) \cdot (1 - \frac{1}{2}) = 0.2666$

ESEMPIO


Individuo A
Individuo B
Individuo C
Individuo D
Individuo E
Censurato alla fine dell'esperimento
Inizio del monitoraggio
t_1
Tempo
t_2
t_3

ESEMPIO



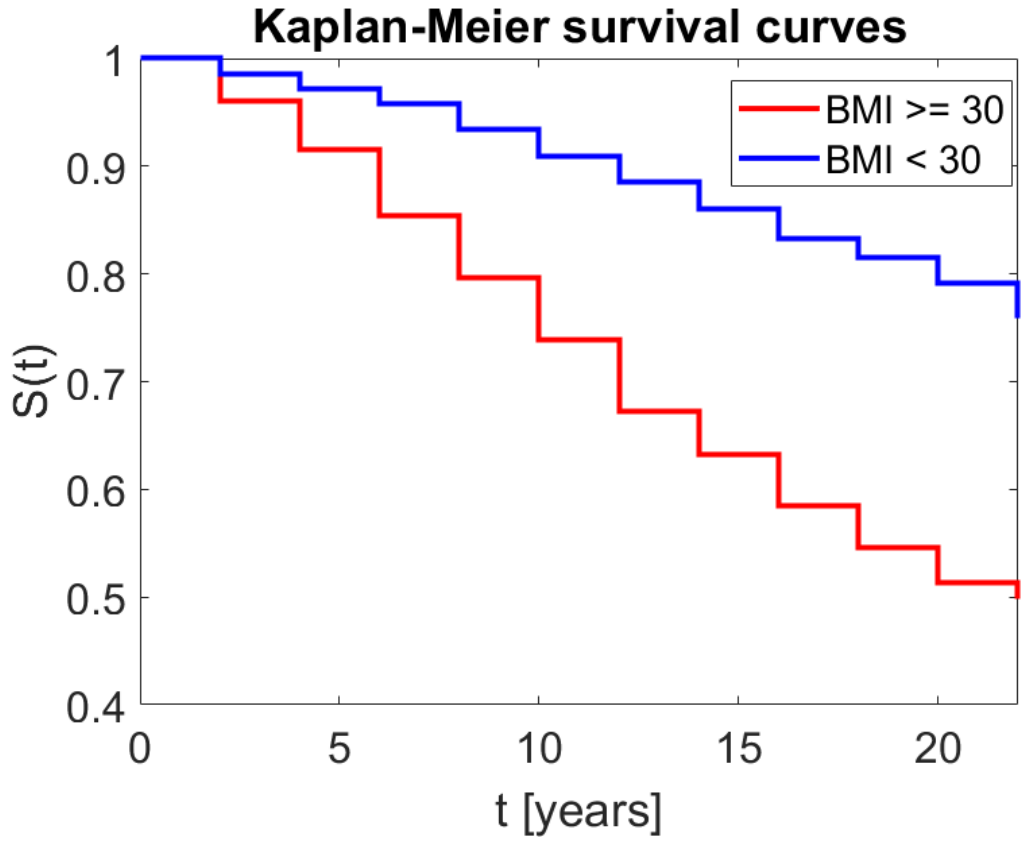
OBIETTIVI DELL'ANALISI DI SOPRAVVIVENZA

- Se vogliamo studiare il tempo ad un evento di interesse pertanto abbiamo bisogno di altri metodi statistici → metodi dell'analisi di sopravvivenza
- Tre principali obiettivi dell'analisi di sopravvivenza:

1. Stimare il tempo ad un evento per un gruppo di individui
2. Confrontare il tempo ad un evento per due o più gruppi di individui
3. Studiare la relazione tra una o più variabili esplicative e il tempo all'evento

CONFRONTO TRA CURVE DI SOPRAVVIVENZA (1/2)

- Evento di interesse: insorgenza di diabete di tipo 2 nella popolazione over 50.
- Confronto delle curve di sopravvivenza stimate per diversi sottogruppi di individui.

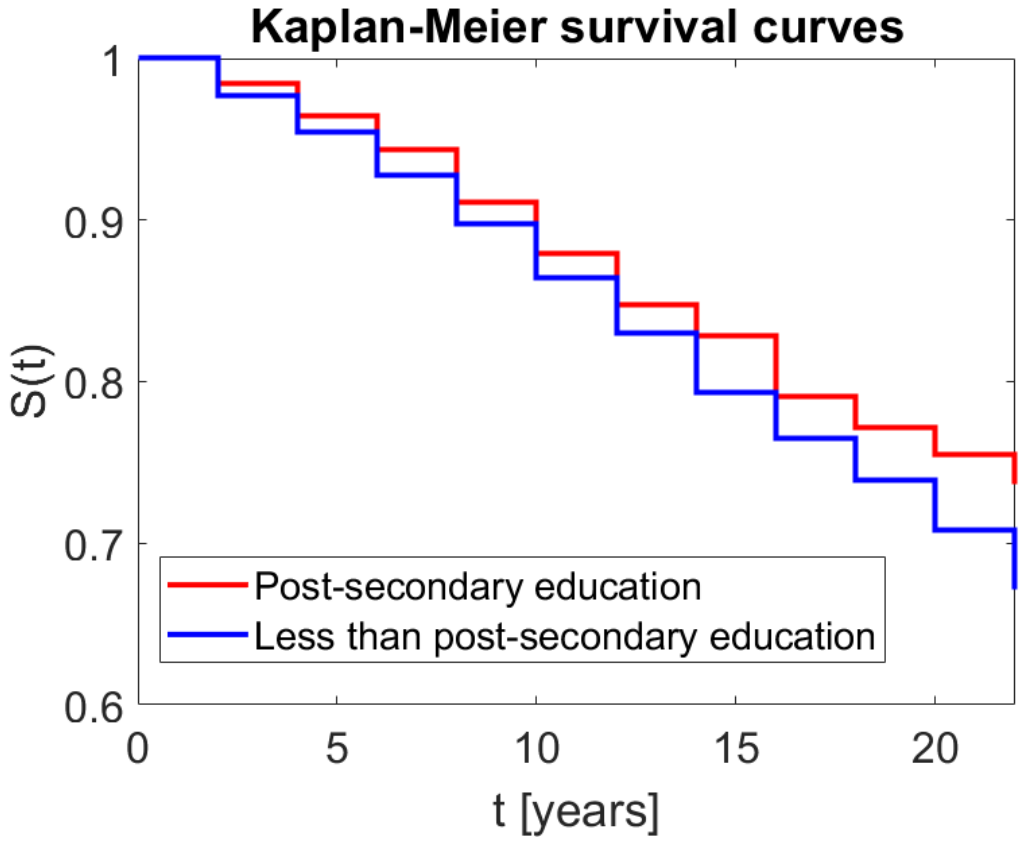


L'indice di massa corporea alla baseline (BMI) ha un impatto importante sulla probabilità di sopravvivenza.

Individui con BMI≥30 hanno un rischio di insorgenza di diabete di tipo 2 significativamente maggiore.

CONFRONTO TRA CURVE DI SOPRAVVIVENZA (2/2)

- Evento di interesse: insorgenza di diabete di tipo 2.
- Confronto delle curve di sopravvivenza stimate per diversi sottogruppi di individui.



Il livello di istruzione ha un impatto significativo sul tempo di sopravvivenza, ovvero sulla probabilità di insorgenza di diabete di tipo 2 nel tempo?

IL TEST DEI RANGHI LOGARITMICI

- **Test dei ranghi logaritmici** (***log-rank test***): test statistico non parametrico che ci consente di testare un'ipotesi statistica sulla differenza tra le funzioni di sopravvivenza di due popolazioni, analizzandone due campioni.
- Sistema di ipotesi:
  - H_0: le due popolazioni hanno la stessa funzione di sopravvivenza
  - H_1: la funzione di sopravvivenza delle due popolazioni è diversa
- La statistica del test confronta le stime delle funzioni di rischio dei due gruppi a confronto in ciascuno dei tempi distinti degli eventi.

LA STATISTICA DEL TEST (1/2)

BONUS

- Raccogliamo un campione di dati per due sottogruppi delle popolazioni a confronto. Per ciascun gruppo avremo un certo numero di individui per cui si verifica l'evento (di cui conosciamo i tempi).
- Supponiamo di avere K distinti tempi degli eventi (considerando sia gli eventi del gruppo 1, sia gli eventi del gruppo 2).

$$t_1 < t_2 < \cdots < t_K$$

- Per ogni tempo $t_j, j = 1, \dots, K$, calcoliamo:
  - n_{1j}: numero di individui a rischio per il gruppo 1 al tempo t_j
  - n_{2j}: numero di individui a rischio per il gruppo 2 al tempo t_j
  - $n_j = n_{1j} + n_{2j}$
  - d_{1j}: numero di eventi osservati per il gruppo 1 al tempo t_j
  - d_{2j}: numero di eventi osservati per il gruppo 2 al tempo t_j
  - $d_j = d_{1j} + d_{2j}$

LA STATISTICA DEL TEST (2/2)

BONUS

- Se vale H_0, d_{1j} ha distribuzione ipergeometrica di parametri n_j, n_{1j}, d_j e:

$$E(d_{1k}) = \frac{d_k}{n_k} n_{1k} \qquad Var(d_{1k}) = \frac{d_k (n_{1k}/n_k)(1 - n_{1k}/n_k)(n_k - d_k)}{n_k - 1}$$

- Statistica del test: quantifica la differenza tra i valori osservati per d_{1k} e quelli attesi sotto l'ipotesi nulla.

$$Z = \frac{\sum_{j=1}^{K} (d_{1k} - E(d_{1k}))}{\sqrt{\sum_{j=1}^{K} Var(d_{1k})}}$$

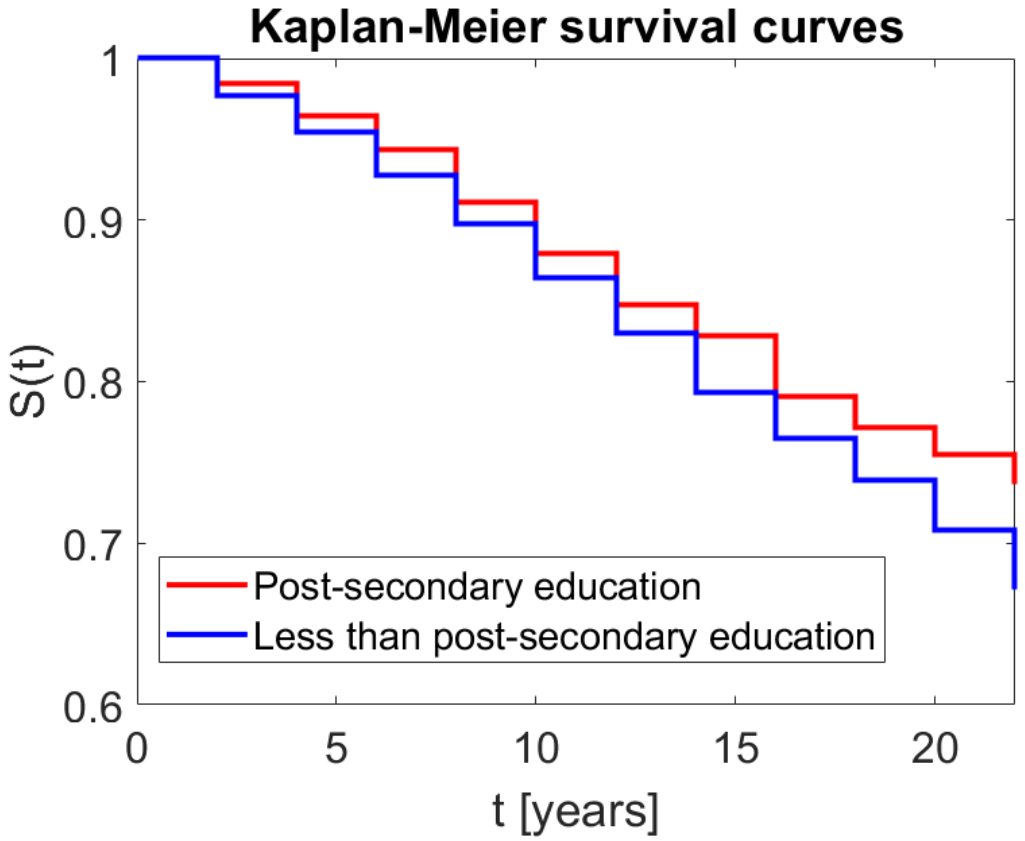
- Se vale H_0, Z ha distribuzione normale standard.

REGOLA DECISIONALE DEL TEST

- Calcoliamo il valore di Z a partire dai dati a disposizione.
- Livello di significatività α.
- Valore critico: $z_{\frac{\alpha}{2}}$ (per α=0.05, $z_{\frac{\alpha}{2}} = 1.96$)
- Regola decisionale:
  - Se $|Z| > z_{\frac{\alpha}{2}}$ → rifiuto H_0 → le due popolazioni hanno funzioni di sopravvivenza significativamente diverse
  - Se $|Z| \leq z_{\frac{\alpha}{2}}$ → non posso rifiutare H_0 → non possiamo dire nulla

ESEMPIO

- Evento di interesse: insorgenza di diabete di tipo 2.
- Confronto delle curve di sopravvivenza stimate per diversi sottogruppi di individui.



- α=0.05 → $z_{\frac{\alpha}{2}} = 1.96$
- z=-2.2329
- $|z| > z_{\frac{\alpha}{2}}$ → rifiutiamo H_0

Concludiamo che le due popolazioni hanno funzioni di sopravvivenza diverse → il livello di istruzione ha un impatto significativo sul rischio di insorgenza di diabete di tipo 2 nel tempo.