Analisi di sopravvivenza

L’analisi della sopravvivenza è un insieme di metodologie usate per descrivere e studiare il **tempo necessario affinché uno specifico evento si verifichi**.

I dati sono caratterizzati da tempi che intercorrono tra un istante 0, solitamente l’inizio dello studio, fino al verificarsi di un determinato evento che è detto end-point

L’analisi della sopravvivenza puo essere applicata a diversi campi, come medicina, sanita pubblica, scienze sociali, ingegneria.

Ad esempio in medicina il tempo di accadimento puo essere il tempo fino alla morte del paziente o il tempo fino al verificarsi di un’infezione. Nelle scienze sociali, l’interesse potrebbe essere l’analisi dei tempi di accadimento di un evento quale, per esempio, il cambiamento di lavoro, il matrimonio, la nascita di un bambino e così via

OBIETTIVI DELL'ANALISI DELLA SOPRAVVIVENZA

- **STIMARE** la funzione di sopravvivenza (ad esempio, probabilità cumulativa di sopravvivenza a 3 o 5 anni)
- **CONFRONTARE** le esperienze di vita di gruppi di pazienti sottoposti a trattamenti diversi
- **VALUTARE** la capacità prognostica di diverse variabili considerate separatamente e/o congiuntamente

Ad esempio, se volessimo studiare la sopravvivenza a due anni di un campione di persone (coorte dello studio) dovremmo seguire ciascuno di loro per 24 mesi.

Alla fine di questo periodo di osservazione ciascun paziente sarebbe descritto da una coppia di valori che indichiamo generalmente con (c, t).

“c” è la condizione (0=non morto, 1=morto)
“t” è la durata dell’osservazione

Per proceder con l’analisi servono tre concetti di base, ovvero: tempo di sopravvivenza, probabilità condizionata e rischio osservato

Il tempo di sopravvivenza

Il termine “tempo di sopravvivenza” va usato in senso estensivo perché si applica anche a eventi diversi dalla morte.

L’analisi di sopravvivenza riguarda infatti tutti quegli studi in cui si vuole analizzare l’incidenza di un determinato evento in un certo arco temporale (studi di coorte).

Perciò il tempo di sopravvivenza assume significati diversi in relazione al tipo di evento a cui è interessato il ricercatore.

Pertanto il tempo di sopravvivenza può essere: il tempo che intercorre tra l’inizio dello studio e la morte, l’incidenza di un evento cardiovascolare (come l’infarto o l’ictus), il tempo di insorgenza di una patologia (per esempio le peritoniti in una coorte di pazienti in CAPD), il tempo dopo il quale si osserva un aumento significativo dei livelli plasmatici di un certo marker biochimico (per esempio il raddoppio della creatinina o della proteinuria in una coorte di pazienti con insufficienza renale lieve) o il tempo in cui si
verifica un episodio di rigetto renale (in una coorte di pazienti portatori di trapianto di rene).

La probabilità condizionata

Il rapporto tra l'analisi di sopravvivenza e la probabilità condizionata è semplice ed intuitivo come del resto è chiara l'identità tra rischio e probabilità.

Infatti la probabilità che ha un paziente di sopravvivere dopo tre giorni dall'ingresso in uno studio (probabilità che indichiamo generalmente con P) è condizionata dal fatto che il paziente sia sopravvissuto nei due giorni precedenti.

Questa probabilità viene anche definita probabilità cumulativa o sopravvivenza cumulativa.

Esempio - Se indichiamo con:

p1 la probabilità che ha il paziente di sopravvivere il primo giorno
p2 la probabilità di sopravvivere il secondo giorno
p3 la probabilità di sopravvivere il terzo giorno

La probabilità condizionata

la sopravvivenza cumulativa è data dal prodotto di queste singole probabilità:

$$P = p_1 \cdot p_2 \cdot p_3$$

Per cui se

$$p_1 = probabilità\ di\ sopravvivere\ il\ primo\ giorno = 0.8$$
$$p_2 = probabilità\ di\ sopravvivere\ il\ secondo\ giorno = 0.68$$
$$p_3 = probabilità\ di\ sopravvivere\ il\ terzo\ giorno = 0.55$$

La probabilità cumulativa $P = 0.3$ (cioè il 30%).

Il rischio osservato

Il rischio osservato è il rischio misurato osservando la realtà.

Per comprendere meglio il concetto di rischio osservato supponiamo di avere a che fare con 11 pazienti ipertesi di cui 4 con severa ipertrofia ventricolare sinistra (IVS) e 7 senza IVS. Ipotizziamo di aver seguito questi pazienti per 1 anno. Il nostro obiettivo è quello di misurare il rischio relativo per un dato evento cardiovascolare (per esempio l'infarto del miocardio) associato alla presenza dell'IVS.

Costruiamo una tabella 2 x 2. Sulle colonne riportiamo il fattore di rischio (presente/assente) e sulle righe l'evento cardiovascolare (sì/no)

Fattore di rischio:
Ipertrofia ventricolare sinistra

		Presente	Assente
Evento	No	1	6
	Si	3	1
Totale		4	7

Fattore di rischio:
Ipertrofia ventricolare sinistra

		Presente	Assente
Evento	No	1	6
	Si	3	1
Totale		4	7

Alla fine del periodo 3 dei 4 pazienti con marcata IVS hanno avuto 1 evento cardiovascolare mentre solo 1 dei 7 pazienti senza IVS ha avuto un evento.

Per calcolare il rischio relativo di eventi cardiovascolari a cui sono esposti i pazienti del gruppo con IVS rispetto al gruppo senza IVS basta fare: (3/4)/(1/7) = 5.4

Ciò vuol dire che i pazienti con IVS hanno un rischio osservato di eventi cardiovascolari che è 5.4 volte maggiore rispetto ai pazienti senza IVS.

Lo strumento di analisi statistica che ci consente di costruire le curve di sopravvivenza (ovvero il grafico della relazione esistente tra la probabilità di sopravvivere e il tempo di osservazione) e di misurare il rischio osservato sono le curve di sopravvivenza di **KAPLAN-MEIER.**

L’ analisi di Kaplan-Meier

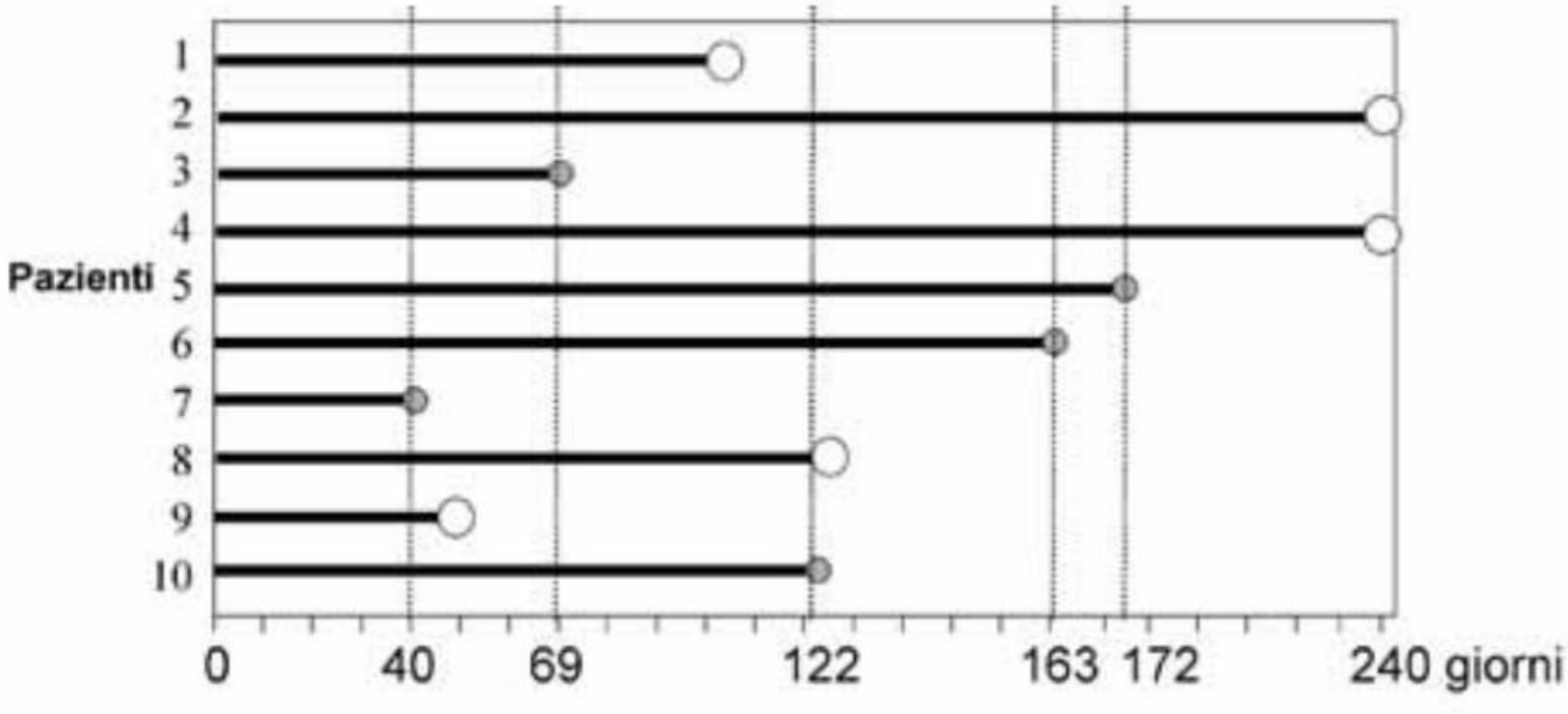
Come costruire una curva di sopravvivenza

Prendiamo in considerazione 10 pazienti con severa ipertrofia ventricolare sinistra IVS e supponiamo di seguirli per 240 giorni.

Il nostro obiettivo è quello di costruire la curva di sopravvivenza in relazione all’incidenza dell’infarto del miocardio fatale.

“censurati” = tutti i pazienti che non hanno l’infarto del miocardio durante il periodo di osservazione ovvero coloro che sopravvivono fino alla fine dell’osservazione oppure che escono dall’osservazione prima del termine dello studio per ragioni diverse dall’infarto del miocardio (per esempio pazienti persi al follow-up, trasferiti o morti per cause diverse da quella d’interesse).

I pazienti censurati rimangono nell’analisi fino al momento in cui sono disponibili dati certi sul loro stato di salute e la loro presenza



I cerchietti vuoti indicano i censurati mentre i cerchietti grigi i pazienti con l'evento.
Il paziente n. 1 viene censurato dopo 100 giorni perché perso al follow-up
il paziente n. 2 arriva vivo alla fine dell'osservazione
Il paziente n. 3 ha l'infarto dopo 59 giorni
il paziente n. 4 arriva vivo alla fine dell'osservazione
il paziente n. 5 ha l'infarto dopo 172 giorni, il paziente n. 6 ha l'infarto dopo 163 giorni, il paziente n. 7 dopo 40 giorni
il paziente n. 8 viene censurato dopo 125 giorni perché trasferito in un'altra città
il paziente n. 9 viene censurato dopo 50 giorni perché morto per neoplasia
il paziente n. 10 ha l'infarto dopo 122 giorni

Il passo successivo è quello di dividere il periodo di tempo (ovvero i 240 giorni) in intervalli.

Nell'analisi di Kaplan-Meier il numero degli intervalli è dettato dai tempi in cui ha luogo l'evento di interesse.

Nel nostro caso avremo complessivamente 6 intervalli, uno per ogni evento che si è verificato

Il primo intervallo è 0-40 giorni, il secondo è 41-69, il terzo è 70-122, il quarto è 123-163, il quinto è 164-172, il sesto è 173-240.

I dati sono riportati in una tabella:

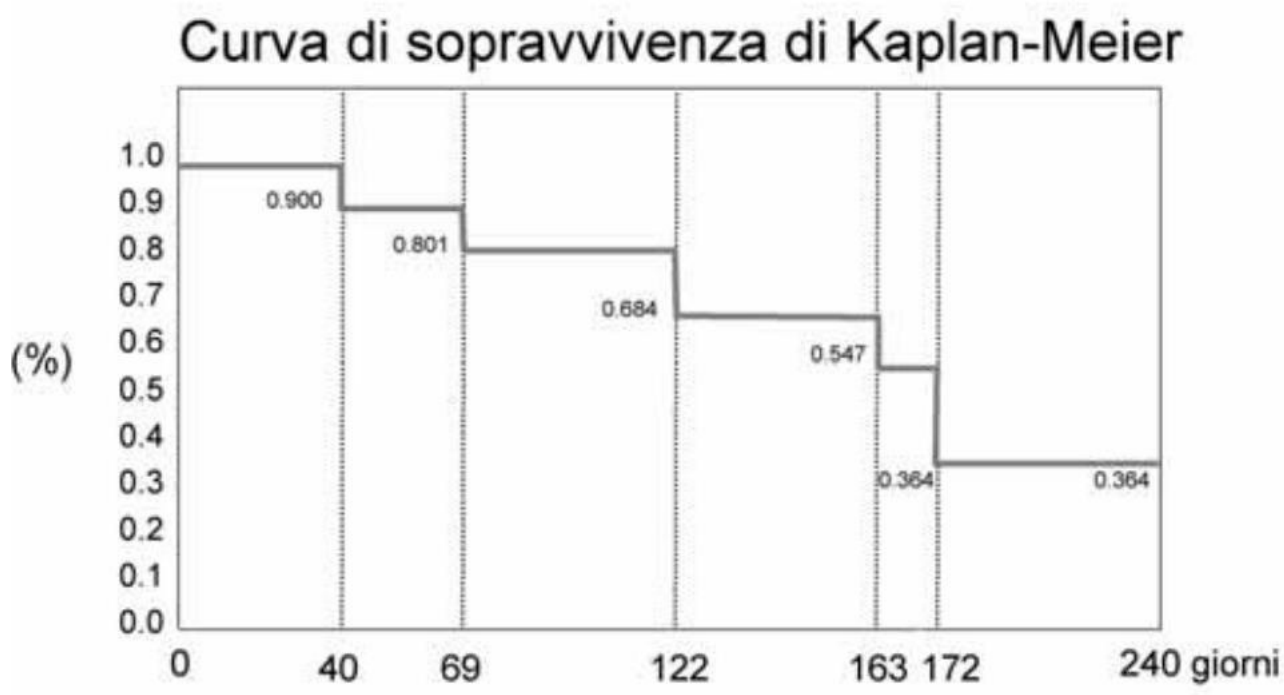
TABELLA I

Intervallo	Giorni	A rischio nel periodo	Eventi nel periodo	Censurati nel periodo	Sopravvivenza del periodo	Sopravvivenza cumulativa
1	0-40	10	1	0	0.900	0.900
2	41-69	9	1	1	0.890	0.801
3	70-122	7	1	1	0.857	0.684
4	123-163	5	1	1	0.800	0.547
5	164-172	3	1	0	0.666	0.364
6	173-240	2	0	2	1.000	0.364

Partiamo dal primo intervallo (0-40). In questo intervallo il numero dei pazienti a rischio è 10. I pazienti a rischio in un intervallo di tempo t1 sono i pazienti che possono potenzialmente avere l'evento nell'intervallo di tempo t1 ovvero quelli sopravvissuti e non censurati nell'intervallo di tempo precedente più i pazienti censurati nell'intervallo di tempo t1.

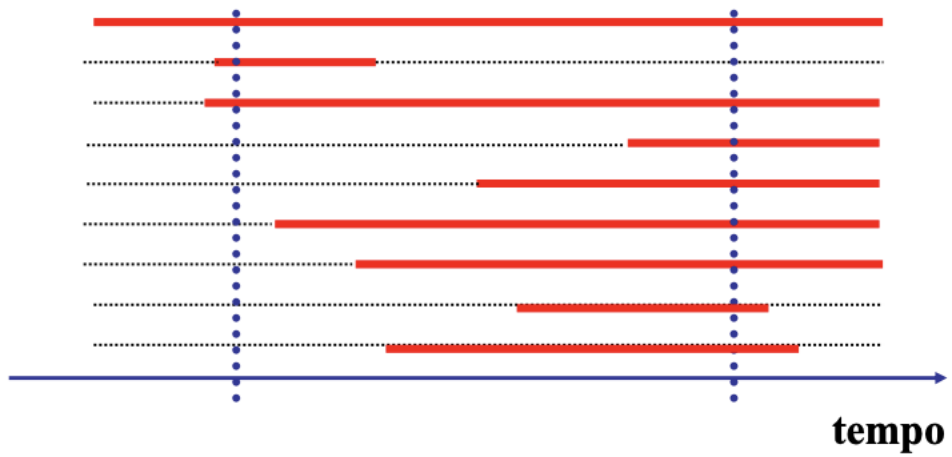
Ovviamente, all'inizio dell'osservazione i pazienti a rischio sono 10. Nel primo intervallo si verifica 1 evento e non vi sono censurati. Nel secondo intervallo (41-69 giorni) i pazienti a rischio sono 9

La sopravvivenza nel periodo è la percentuale di pazienti vivi in quel periodo. Perciò, per calcolare la sopravvivenza del periodo si utilizza la formula: $1-\frac{numero\ di\ eventi\ nell'intervallo}{numero\ di\ persone\ a\ rischio\ nell'intervallo}$


Curva di sopravvivenza di Kaplan-Meier
(%)
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0.900
0.801
0.684
0.547
0.364
0.364
0
40
69
122
163
172
240 giorni

IL MOMENTO INIZIALE (TEMPO 0) può essere lo stesso per tutti i pazienti, ovvero un momento del calendario.

Tuttavia, nella maggior parte dei casi il tempo zero varia da paziente a paziente perché coincide con il momento della diagnosi o con l'intervento chirurgico o con l'inizio della chemioterapia.



Si dice che il prodotto limite di Kaplan-Meier è dato da:

$$\hat{S}(t_j) = \prod \left(1 - \frac{d_i}{n_i}\right)$$

Dove ci sono n_i individui in vita immediatamente prima del tempo t_j mentre avvengono $d_i$decessi al tempo t_j.

Poiché i tempi di sopravvivenza non sono gaussiani, anzi sono molto asimmetrici, si usa il tempo di sopravvivenza mediano per riassumere in un solo valore i dati.

Una volta calcolata la curva, il tempo di sopravvivenza mediano è il più piccolo tempo di sopravvivenza osservato per il quale il valore della funzione di sopravvivenza è minore di 0.5

Come per tutte le statistiche basate su campioni, anche per la funzione di sopravvivenza c'è una distribuzione della statistica campionaria attorno al parametro di popolazione $S(t)$.

L'errore standard della funzione/curva di sopravvivenza può essere calcolato mediante la seguente equazione nota come formula di Greenwood:

$$SD_{\hat{S}(t_j)} = \hat{S}(t_j)\sqrt{\sum \frac{d_i}{n_i(n_i - d_i)}}$$

Dove la sommatoria è estesa a tutti i tempi nei quali avvengono gli eventi (es morti) fino al tempo t_j incluso.
Come per la curva di sopravvivenza, l'errore standard viene calcolato solo utilizzando i tempi in cui avvengono gli eventi (es morti).

Questo viene utilizzato per calcolare un intervallo di confidenza della curva di sopravvivenza.

Come confrontare due curve di sopravvivenza

Abbiamo finora visto come si costruisce una curva di sopravvivenza. Vediamo adesso come confrontare tra loro
due curve di sopravvivenza. Ciò è particolarmente importante quando, per esempio, vogliamo confrontare la sopravvivenza di due gruppi di pazienti: uno esposto ed uno non esposto ad un certo fattore di rischio (per esempio pazienti con e senza ipertrofia ventricolare sinistra, fumatori e non fumatori, ipertesi e normotesi e così via).

Il test da utilizzare è il **log-rank test**.

L’obiettivo finale di una grande parte della pratica medica è quello di prolungare la vita, e quindi sorge spontanea in molte ricerche cliniche la necessità di confrontare curve di sopravvivenza per gruppi di pazienti sottoposti a trattamenti diversi.

L’ipotesi nulla ➔ i trattamenti hanno gli stessi effetti sulla sopravvivenza

Problema: ci sono delle osservazioni censurate altrimenti potremmo considerare il Mann-Whitney.

Si usa il LONG-RANK

Alla base ci sono 3 ipotesi

1) I due campioni sono indipendenti
2) Le modalità di censura sono le stesse per entrambi i campioni
3) Le due curve di sopravvivenza presentano proporzionali così che essi sono in relazione secondo la formula $S_2(t) = [S_1(t)]^{\alpha}$ dove α è una costante che si chiama tasso di rischio

Le due curve di sopravvivenza sono identiche se $\alpha = 1$. Se $\alpha < 1$ i pazienti del gruppo 2 muoiono più lentamente di quelli del gruppo 1.

Intervallo	A rischio Gruppo 1	A rischio Gruppo 2	Totale a rischio (Gruppo 1+Gruppo II)	Censurati	Eventi Gruppo 1	Mortalità attesa Gruppo 1	Eventi Gruppo 2	Mortalità attesa Gruppo 2
1-23	6	6	12	0	1	6/12=0.50	0	6/12=0.50
24-25	5	6	11	0	0	5/11=0.454	1	6/11=0.546
26-26	5	5	10	0	0	5/10=0.50	1	5/10=0.50
27-36	5	4	9	0	0	5/9=0.556	1	4/9=0.444
37-43	5	3	8	1	1	5/8=0.625	0	3/8=0.375
44-61	3	3	6	1	1	3/6=0.50	0	3/6=0.50
62-71	2	2	4	1	0	2/4=0.50	1	2/4=0.50
72-78	1	1	2	0	0	1/2=0.50	1	1/2=0.50
79-83	1	0	1	0	1	1/1=1.00	0	0/1=0.00
					4 morti	5.13 morti	5 morti	3.87 morti

Per il Log-Rank Test, la mortalità attesa viene calcolata assumendo che il tasso di rischio sia costante e comune tra i due gruppi.

L'ipotesi nulla (H0) del Log-Rank Test è che non vi sia differenza tra i gruppi in termini di rischio di morte (o eventi).

Passaggi per calcolare la mortalità attesa:

1. Definizioni preliminari:

- d_j: Numero di **decessi osservati** al tempo t_j.
- n_j: Numero di soggetti **a rischio** immediatamente prima del tempo t_j.
- n_{jk}: Numero di soggetti **a rischio** nel gruppo k al tempo t_j.
- d_{jk}: Numero di decessi osservati nel gruppo k al tempo t_j.

2. Mortalità attesa al tempo t_j

L'evento atteso in ciascun gruppo è proporzionale al numero di soggetti a rischio in quel gruppo rispetto alla popolazione complessiva a rischio.

La formula per il numero atteso di decessi in ciascun gruppo k è:

$$E_{jk} = d_j \cdot \frac{n_{jk}}{n_j}$$

Dove:

- d_j = totale dei decessi osservati al tempo t_j (tutti i gruppi combinati),
- n_{jk} = soggetti a rischio nel gruppo k,
- n_j = soggetti a rischio totali (somma di tutti i gruppi).

3. Calcolo mortalità attesa totale per ciascun gruppo k

$$E_k = \sum_j E_{jk}$$

4. Usando i nostri dati:

n_j1	n_j2	n_j
Gruppo 1	Gruppo 2	Totale
6	6	12
5	6	11
5	5	10
5	4	9
5	3	8
3	3	6
2	2	4
1	1	2
1	0	1

decessi gruppo 1	decessi gruppo 2	decessi totali
1	0	1
0	1	1
0	1	1
0	1	1
1	0	1
1	0	1
0	1	1
0	1	1
1	0	1

E_j1 mortalità attesa gruppo 1	E_j2 mortalità attesa gruppo 2
0.50	0.50
0.45	0.55
0.50	0.50
0.56	0.44
0.63	0.38
0.50	0.50
0.50	0.50
0.50	0.50
1.00	0.00
5.14	3.86

La mortalità attesa è data dal calcolo della probabilità di evento per ogni intervallo temporale (numerosità gruppo / numerosità totale) moltiplicata per il numero di decessi attesi (osservati in totale) nel periodo

5.14 = somma di tutti gli elelmenti della colonna E_j1

3.86 = somma di tutti gli elementi della conìlonna E_j2

La sopravvivenza del gruppo 1 a fine dell'intervallo si osservazione è zero

$$\text{Log}\,rank\ test = \frac{(Mortalità\ attesa\ - Mortalità\ osservata)^2}{Somma\ dei\ prodotti\ delle\ mortalità\ attese\ nei\ due\ gruppi}$$

NUMERATORE

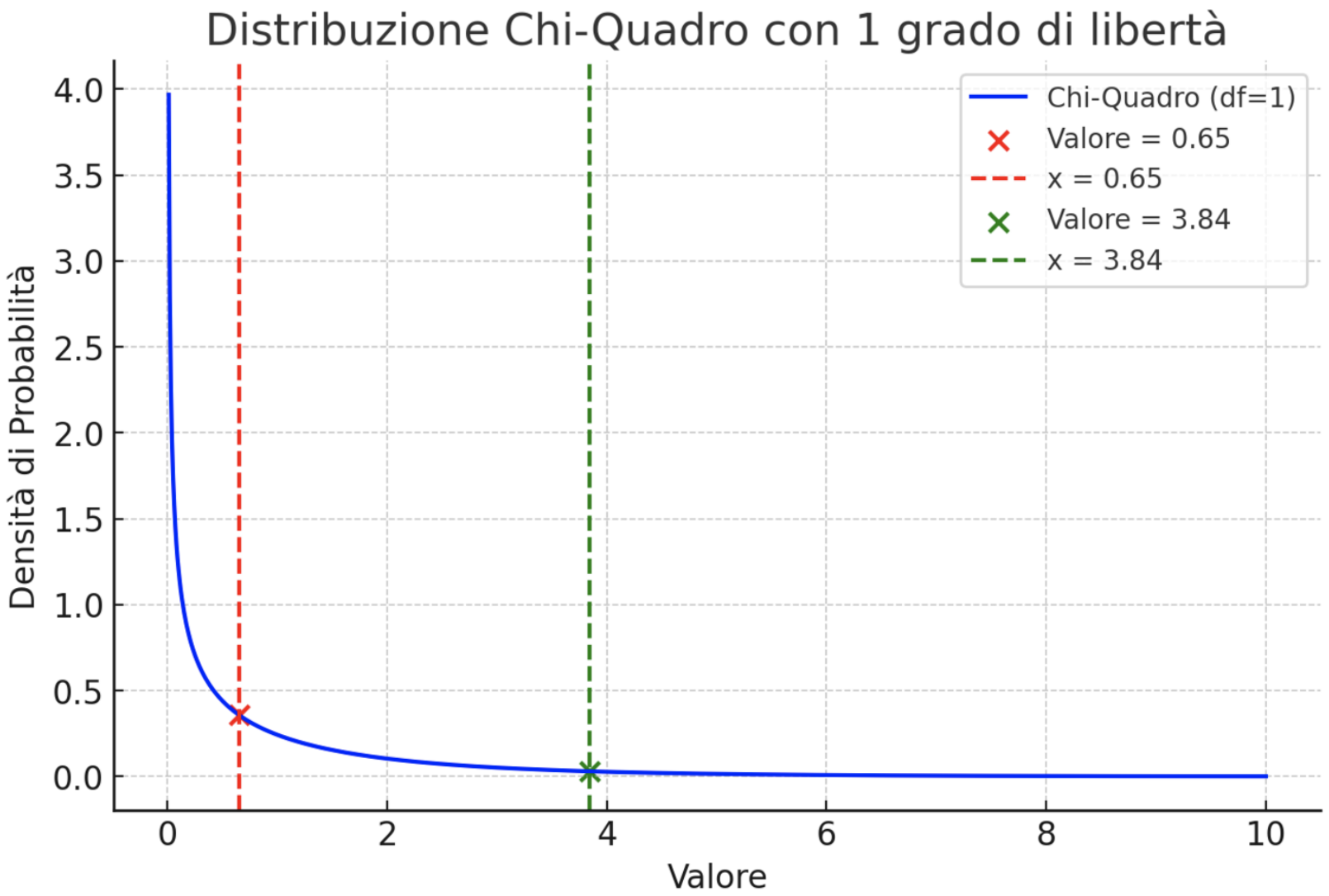
Mortalità attesa Gruppo 1 = 5.13
Mortalità osservata Gruppo 1 = 4
Differenza = $(5.13 - 4)^2 = 1.28$

Mortalità attesa Gruppo 2 = 3.87
Mortalità osservata Gruppo 1 = 5
Differenza = $(3.87 - 5)^2 = 1.28$

DENOMINATORE

E_j1 mortalità attesa gruppo 1	E_j2 mortalità attesa gruppo 2	E_j1*EJ_2
0.50	0.50	**0.25**
0.45	0.55	**0.25**
0.50	0.50	**0.25**
0.56	0.44	**0.25**
0.63	0.38	**0.23**
0.50	0.50	**0.25**
0.50	0.50	**0.25**
0.50	0.50	**0.25**
1.00	0.00	**0.00**
5.14	3.86	**1.98**

$$\text{Log}\,rank\ test = \frac{1.28}{1.98} = 0.65$$



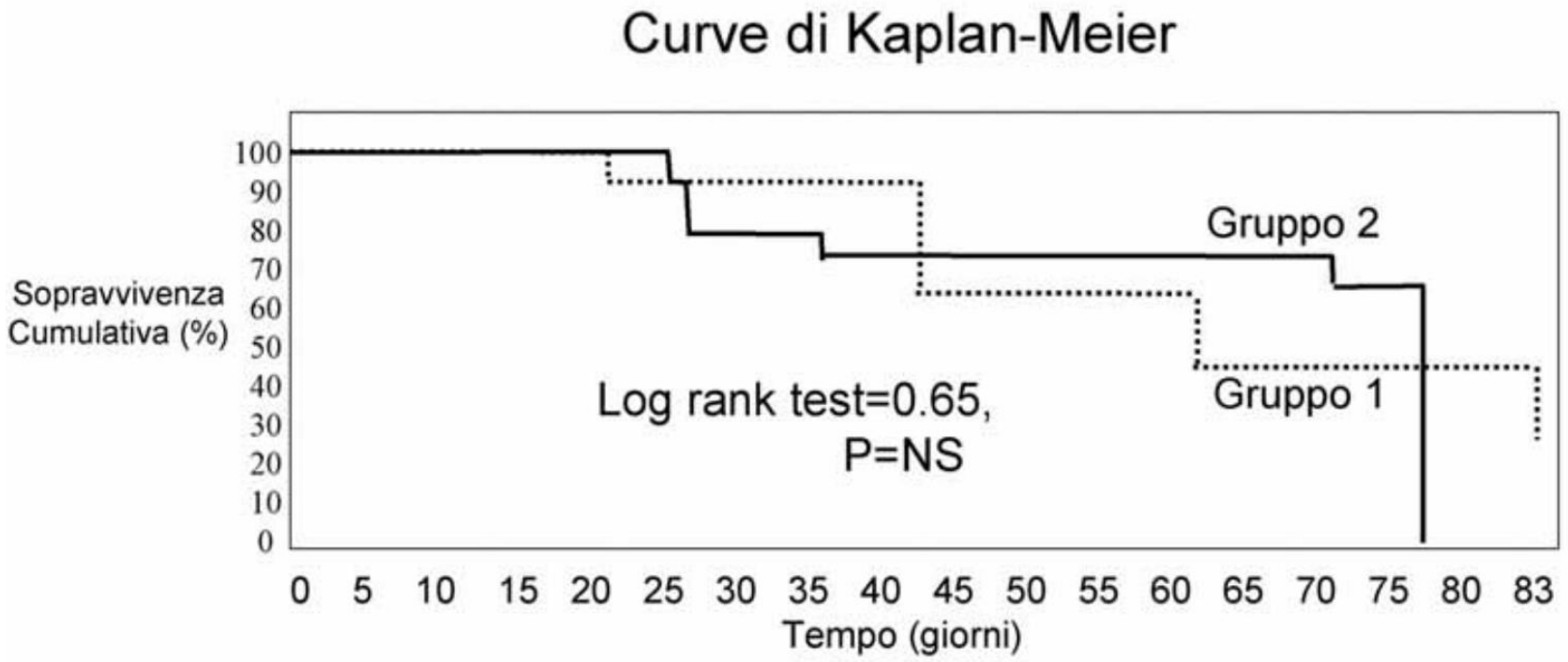
Il valore 3.84 rappresenta la soglia critica al livello di significatività 5% ($\alpha = 0.05$) per una distribuzione chi-quadro con 1 grado di libertà.

Il risultato ottenuto va quindi tradotto in una probabilità

Tale test, sotto l'ipotesi nulla, si distribuisce asintoticamente come un Chi-quadro con un grado di libertà (numero di gruppi -1).

Affinché un log rank test sia statisticamente significativo (p_value<0.05) è necessario che abbia un valore di almeno 3.84.

La differenza tra le due curve non è statisticamente significativa, ovvero la presenza del fattore di rischio non ha un impatto statisticamente significativo sulla sopravvivenza.



Funzione di sopravvivenza:

La funzione di sopravvivenza S(t) esprime la probabilità di sperimentare l'evento dopo un certo istante t.
Essa è monotona decrescente. In t = 0 nessun soggetto ha sperimentato l'evento, quindi la probabilità di sopravvivenza è pari a 1. All'aumentare del tempo la funzione di sopravvivenza decresce in modo più o meno ripido a seconda della frequenza dell'evento

Funzione rischio (O hazard):

Viene definita come:

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

h(t) è una funzione positiva.

Funzione rischio (O hazard):

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

Essa è utile per descrivere in che modo il rischio di sperimentare l'evento cambia nel tempo. La funzione di rischio, può assumere un andamento

- crescente (l'evento tende a verificarsi alla fine del periodo di osservazione: tipico di unità soggette a invecchiamento o usura)
- decrescente (l'evento tende a verificarsi all'inizio del periodo di osservazione: casi in cui vi è una selezione iniziale)
- costante
- non monotono

Se h=costante ➔ $S(t) = e^{-h \cdot t}$

Rischio cumulativo:

Una quantità connessa alla funzione di azzardo è la funzione di rischio cumulato H(t) definita da:

$$H(t) = -\ln S(t)$$

Può quindi essere ottenuta tramite la funzione di sopravvivenza per avere un'idea del possibile andamento della funzione di rischio h(t)

DEFINIZIONE:

Hazard (Rate)

L’Hazard di un evento (es. il decesso, evento cardiovascolare, progressione malattia …) rappresenta il tasso istantaneo dell’evento (misurato ad ogni istante) durante il periodo di osservazione

Hazard Ratio

Anche detto Relative Hazard; si ottiene dal rapporto tra gli Hazard di 2 gruppi (ad esempio di trattamento)

LIMITAZIONI DI Kaplan-Meier

Principalmente descrittivo
Non controlla le covariate
Non può trattare variabili dipendenti dal tempo

MODELLO DI COX

Nell'analisi di dati di sopravvivenza può essere d'interesse studiare la relazione fra il tempo all'evento e una serie di variabili esplicative/covariate. Questo non è possibile con il log rank test mentre lo è con il modello di Cox (1972).

E' un modello di regressione semiparametrico che esprime il rischio in funzione del tempo e delle covariate.

Dato un vettore di p covariate, $X^T = (X_1, X_2, \dots, X_p)$ il modello si presenta nella forma:

$$h(t \mid X) = h_0(t) \exp(\beta^T X) = h_0(t) \exp(\beta_1 X_1 + \dots + \beta_p X_p),$$

dove $\beta^T = (\beta_1, \dots, \beta_p)$ è il vettore dei parametri e $h_0(t)$ è la funzione rischio di base.

Quest'ultima è una componente non parametrica, positiva, ignota, uguale per tutti i soggetti, dipende solo dal tempo e rappresenta la distribuzione della funzione azzardo per il gruppo di base (X = 0). Ricordiamo che h deve essere positiva.

Se il modello h(t) = costante = μ allora

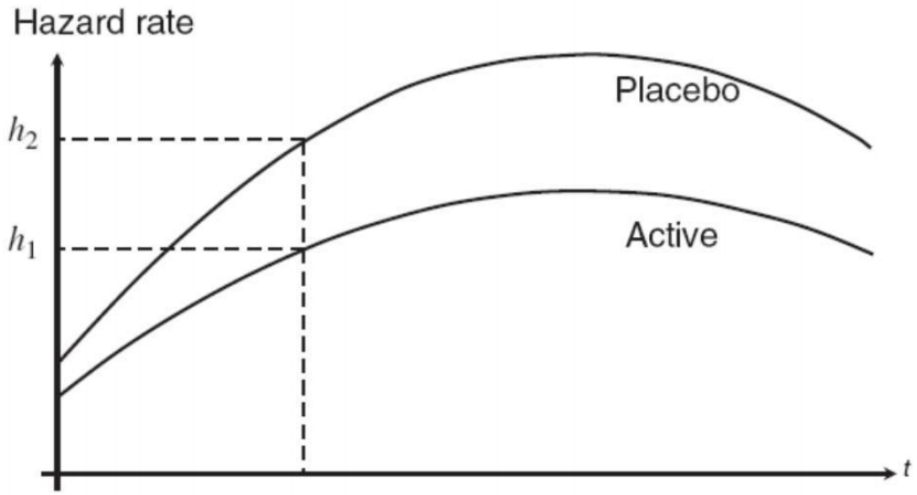
$$\ln h(t) = \mu + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

In generale comunque poiché la funzione di verosimiglianza dipende dalla funzione rischio di base, che non è specificata, è possibile derivare le stime dei parametri con il metodo della verosimiglianza parziale

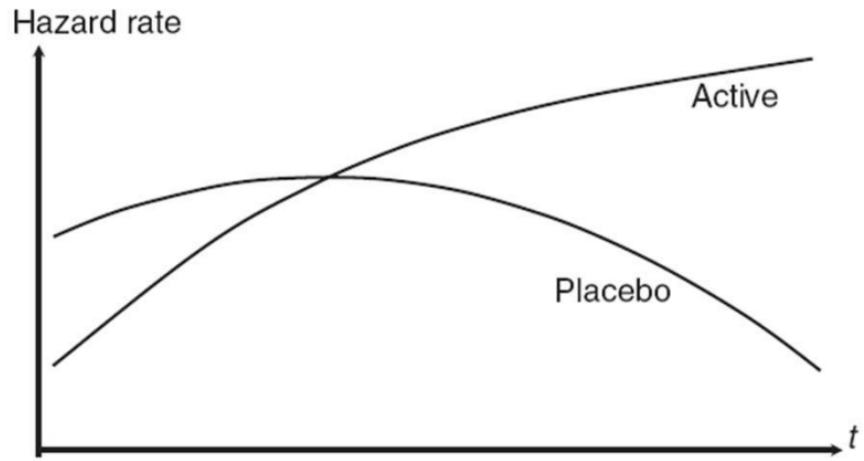
Nota: Il modello di Cox è valido per covariate fisse nel tempo e rispetta l'assunzione di proporzionalità dei rischi. Infatti, dati due individui con covariate X_A e X_B, il rapporto tra i rispettivi rischi HR (hazard ratio)

$$HR = \frac{h(t \mid X^A)}{h(t \mid X^B)} = \frac{\exp(\beta_1 X_1^A + \ldots + \beta_p X_p^A)}{\exp(\beta_1 X_1^B + \ldots + \beta_p X_p^B)}$$

è costante al variare del tempo. Tuttavia, è un modello molto flessibile e con opportuni accorgimenti può essere esteso a situazioni in cui le covariate dipendono dal tempo e l'assunzione di rischi proporzionali è violata.



Hazards variano nel tempo
Ma il loro rapporto è sostanzialmente stabile.
Si può usare Cox Model



Hazards variano nel tempo
Ma la costanza del rapporto e vistosamente violata.
Non si può usare Cox Model

Ottenuta la stima dei coefficienti del modello, è possibile risalire alla funzione di rischio cumulato e alla funzione di sopravvivenza per un soggetto con vettore di covariate X

$$H(t) = -\ln S(t)$$

$$\widehat{H}(t \mid X) = \widehat{H}_0(t)\exp(\widehat{\beta}_1 X_1 + \ldots + \widehat{\beta}_p X_p),$$

$$\widehat{S}(t \mid X) = \widehat{S}_0(t)^{\exp(\widehat{\beta}_1 X_1 + \ldots + \widehat{\beta}_p X_p)}.$$

$$h(t) = -\frac{d\big(\log\big(S(t)\big)\big)}{dt} \qquad h(t \mid X) = h_0(t)\exp(\beta^T X) = h_0(t)\exp(\beta_1 X_1 + \ldots + \beta_p X_p),$$

$$= -\frac{d}{dt}\left(e^{\boldsymbol{\beta}^T \boldsymbol{X}} \log\big(S_0(t)\big)\right)$$

$$= -\frac{d}{dt}\big(\log\big(S_0(t)\big)\big)e^{\boldsymbol{\beta}^T \boldsymbol{X}}$$

$$= h_0(t)e^{\boldsymbol{\beta}^T \boldsymbol{X}}$$

La **funzione di rischio** $h(t)$ è data da:

$$h(t|X) = h_0(t) \cdot e^{\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p}$$

- $h(t|X)$: rischio istantaneo all'istante t dato il vettore delle covariate X.
- $h_0(t)$: funzione di base del rischio (baseline hazard), che dipende solo dal tempo.
- e^{β_j}: **rischio relativo** associato alla covariata X_j.
- $X_1, X_2, \ldots, X_p$: covariate (es. età, sesso, trattamento).
- I coefficienti β_j vengono stimati tramite il metodo della **massima verosimiglianza parziale**.

IPOTESI:

- **Rischi proporzionali**: il rapporto dei rischi tra due individui è **costante nel tempo**.

$$\frac{h(t|X_1)}{h(t|X_2)} = e^{\beta(X_1 - X_2)} \quad \forall t$$

RISULTATI:

1. Se $\beta_j > 0$: la covariata X_j aumenta il rischio dell'evento.
2. Se $\beta_j < 0$: la covariata X_j riduce il rischio dell'evento.
3. Se $\beta_j = 0$: la covariata X_j non ha effetto sul rischio.

Rischio relativo (Hazard Ratio, HR):

$$HR = e^{\beta_j}$$

- **HR > 1**: maggiore rischio.
- **HR < 1**: minore rischio.

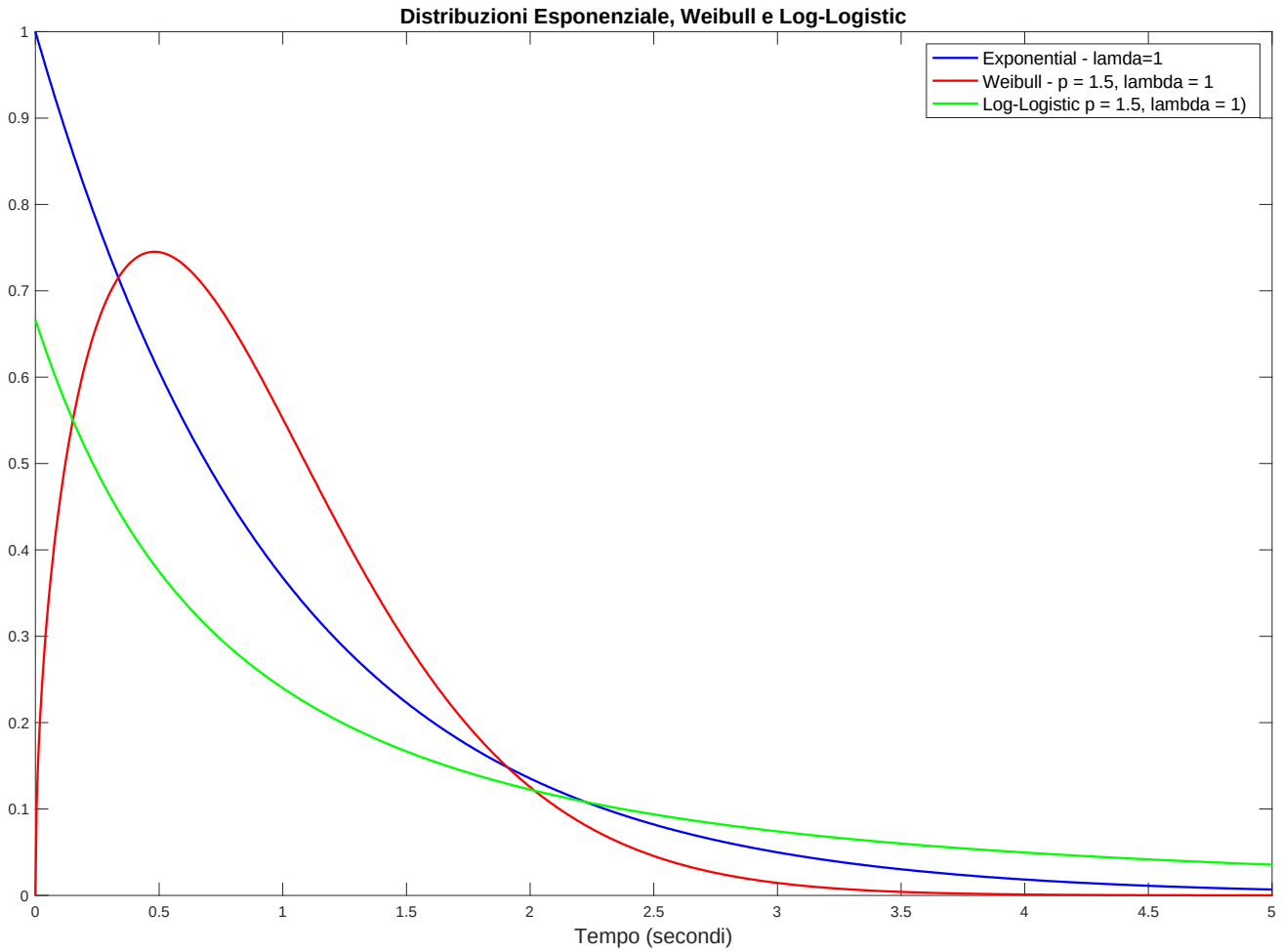
Il modello di Cox si dice non-parametrico perché non fa assunzioni su h(t).

Se assumiamo h(t) = costante allora stiamo utilizzando una informazione aggiuntiva e il metodo diventa parametrico.

I modelli parametrici di h(t) più frequentemente usati sono:

Modello esponezniale
Modello di Cox
Modello Log Logistic

f(t)


Distribuzioni Esponenziale, Weibull e Log-Logistic
Exponential - lamda=1
Weibull - p = 1.5, lambda = 1
Log-Logistic p = 1.5, lambda = 1)
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.5
1
1.5
2
2.5
3
3.5
4
4.5
5
Tempo (secondi)

MODELLO ESPONENZIALE

La funzione densità di riferimento è quella esponenziale che porta al fatto che il tasso di rischio, hazard rate, $\lambda > 0$ ` è costante; il che è un vantaggio matematico ma è uno svantaggio biomedico.
Il vantaggio matematico è che i conti diventano particolamente facili:

$$f(t;\lambda) = \lambda \exp(-\lambda t)$$

$$S(t;\lambda) = \exp(-\lambda t)$$

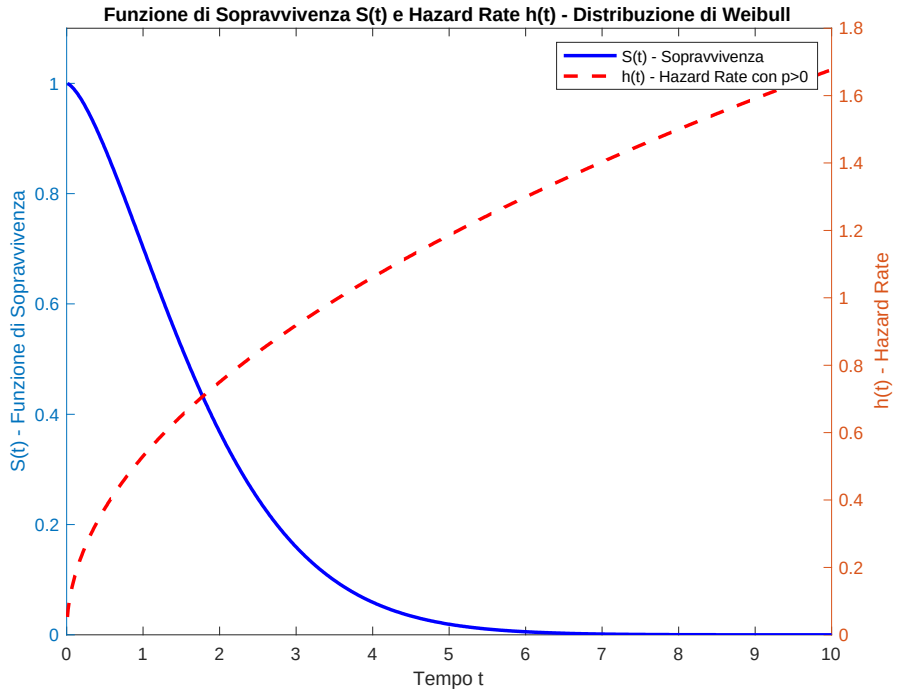
$$h(t;\lambda) = \lambda$$

Lo svantaggio biomedico sta nel fatto che, essendo il rate λ costante, un individuo giovane ed un individuo anziano hanno sempre la medesima probabilità di soccombere.

MODELLO WEIBULL

Si tratta di una distribuzione caratterizzata da due parametri (analogamente alla gaussiana, che ha un parametro di centralità, la media μ, e un parametro di forma, la deviazione standard σ):

- il parametro di forma (shape), p > 0
- il parametro di scala (scale), λ > 0



$$f(t) = k\,\lambda^k t^{k-1} e^{-(\lambda t)^k}$$

$$S(t) = e^{-(\lambda t)^k}$$

$$\mathrm{h}(t) = k\,\lambda^k t^{k-1} = \frac{k}{\lambda}\left(\frac{t}{\lambda}\right)^{k-1}$$

dove $\lambda > 0$ è un fattore di scala e $k > 0$ è detto parametron di forma

Lo svantaggio biomedico sta nel fatto che, essendo il rate λ costante, un individuo giovane ed un individuo anziano hanno sempre la medesima probabilità di soccombere.

Esempio di interpretazione dei risultati

Applicando il modello: $\ln h(t) = \mu + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$

Dove:

- $h(t)$è $il\ tasso\ di\ rischio\ dell'evento\ (stimato\ dal\ modello)\ al\ tempo\ t$
- μ *rappresenta il rischio di base (indipendente dalle covariate/fattore di rischio X)*
- β è $il\ coefficiente\ di\ regressione\ e\ indica\ di\ quanto\ aumenta\ in\ media$
  $il\ logaritmo\ naturale\ del\ tasso\ di\ incidenza\ dell'evento$

Esempio di interpretazione dei risultati

Supponiamo di voler analizzare il rapporto tra la presenza/assenza del diabete (una sola covariata) e il tasso di incidenza della mortalità nella coorte dei 2500 pazienti in dialisi inclusi nel registro dell'ERA-EDTA.

Applicando l'equazione di Cox avremo: $\ln h(t) = \mu + \beta_1 X_1$

- μ è il tasso di incidenza della mortalità nei non diabetici della coorte;
- β_1 indica di quanto aumenta in media il logaritmo naturale del tasso di rischio della mortalità nei diabetici rispetto ai non diabetici;
- X_1 rappresenta il fattore di rischio (nella fattispecie il diabete, che ammette solo due possibilità: 0=assente, 1=presente).

Nei diabetici l'equazione di Cox sarà: $\ln h(t) = \mu + \beta_1 1$ (1=diabete presente)

Nei non diabetici l'equazione di Cox sarà: $\ln h(t) = \mu + \beta_1 0$ (0=diabete assente)

Per sapere di quanto aumenta in media il logaritmo naturale del tasso di incidenza della mortalità nei diabetici rispetto ai non diabetici, è sufficiente calcolare la differenza tra le due equazioni:

$\ln(Ht_{diabetici}) - \ln(Ht_{non\ diabetici}) = \mu + \beta_1 1 - \mu$

Il termine μ ovviamente si annulla. L'equazione assume, perciò, la seguente forma:

$\ln(Ht_{diabetici}) - \ln(Ht_{non\ diabetici}) = \beta_1$

Per conoscere di quanto aumenta in media il tasso di incidenza dell'evento nei diabetici rispetto ai non diabetici è sufficiente calcolare l'esponenziale di entrambi i termini dell'equazione:

$$e^{\ln(h(t)_{diabetici}) - \ln(h(t)_{non\ diabetici})} = e^{\beta_1}$$

$$e^{\ln \frac{h(t)_{diabetici}}{h(t)_{non\ diabetici}}} = e^{\beta_1}$$

Ossia: $\frac{h(t)_{diabetici}}{h(t)_{non\ diabetici}} = e^{\beta_1}$ indica quante volte è più alto il tasso di incidenza della mortalità nei diabetici rispetto ai non diabetici.

Nella Tabella I sono riportati i risultati della regressione univariata di Cox per quanto attiene il rapporto tra un singolo fattore di rischio (il diabete) e il tasso di incidenza della mortalità nel campione dei 2500 pazienti in
dialisi appartenenti al registro dell'ERA-EDTA.

TABELLA I - ANALISI UNIVARIATA DI COX

Variabile	*b*	*Errore Standard*	*P*	*Hazard ratio*	*Intervallo di confidenza 95%*	
Diabete (sì/no)	0.54	0.07	<0.0001	1.71	1.5	1.9

Un coefficiente di regressione (b= β_1) pari a 0.54 indica che, nei diabetici, il logaritmo naturale del tasso di incidenza della mortalità è più alto di quello nei non diabetici di una quantità pari a 0.54. Calcolando l'esponenziale del coefficiente di regressione (2.71830.54=1.71) otteniamo l'hazard ratio.

In questo caso, il tasso di incidenza della mortalità è del 71% più alto nei diabetici rispetto ai non diabetici. Il programma fornisce in questo esempio, inoltre, l'errore standard del coefficiente di regressione, la significatività statistica (P) e l'intervallo di confidenza al 95% dell'hazard ratio.

METRICHE DI PERFORMANCE DEL MODELLO DI COX

Concordanza (C-Index)

Calibration (Grafico di calibrazione)

Residuati di Schoenfeld (validità dell'ipotesi di proporzionalità)

Grafici di Kaplan-Meier stratificati

Definizione

Il **C-Index** (Concordance Index) è una metrica utilizzata per valutare la capacità predittiva del Modello di Cox. Misura quanto bene il modello riesce a discriminare tra pazienti con tempi di sopravvivenza diversi in base alle covariate predette.

In altre parole, il C-Index è definito come la probabilità che il modello assegni risk score maggiori agli individui aventi tempi di sopravvivenza minori.

.

Definizione

$$C = P\left(Y_j > Y_i \middle| T_j < T_i\right)$$

Y_i, Y_j: variabili aleatorie rappresentanti gli score di rischio dei due individui $i\ e\ j$

T_i, T_j: variabili aleatorie rappresentanti I tempi di sopravvicenza dei due individui $i\ e\ j$

In pratica:

Pazienti confrontabili = Si considerano tutte le coppie di pazienti che possono essere confrontate:

- Entrambi hanno avuto l'evento (es: decesso).
- Un paziente ha avuto l'evento prima dell'altro.

Valutazione della concordanza:

- Se il paziente che ha l'evento prima ha un rischio predetto più alto, il modello è "concordante".
- Se non è così, il modello è "discordante".

STIMATORE DI HARRELL DEL C-INDEX

$$\hat{C} = \frac{n_{concordi} + 0.5 \cdot n_{pari}}{n_{confrontabili}}$$

- $n_{concordi}$ = numero di coppie concordi
- n_{pari} = numero di coppie pari
- $n_{confrontabili}$ = numero di coppie confrontabili

Interpretazione del C-Index

C =0.5: Il modello non discrimina meglio di una scelta casuale.

C =1: Concordanza perfetta, il modello predice perfettamente l'ordine degli eventi.

0.5<C<1: Buona capacità discriminatoria; valori più alti indicano un modello **migliore**.

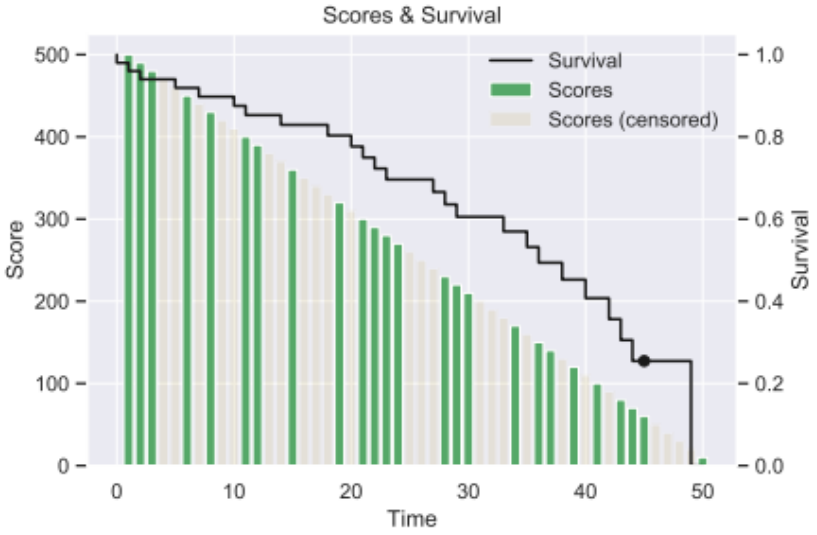
NOTA: Possiamo selezionare tra modelli di Cox diversi in base al C-Index

Esempio Pratico

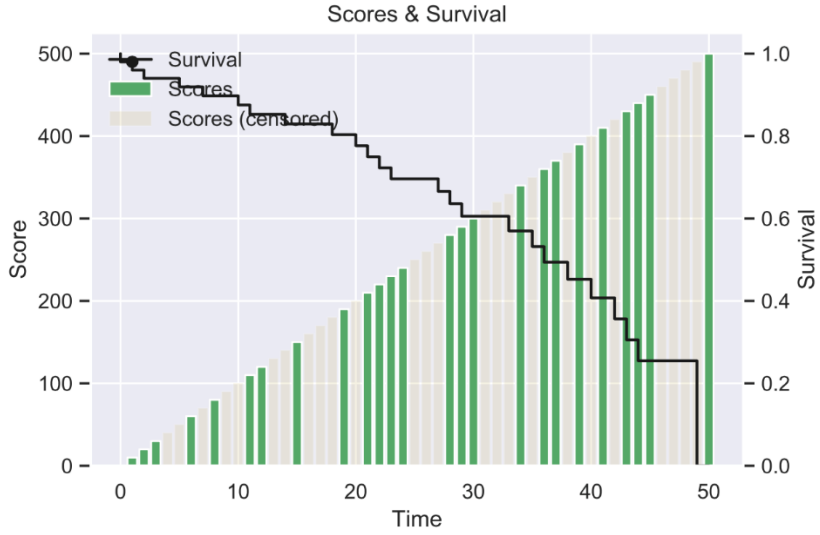
- **Scenario**:
  - Paziente A: tempo di sopravvivenza = 2 mesi
  - Paziente B: tempo di sopravvivenza = 5 mesi
  - Il modello assegna un rischio più alto al paziente A rispetto a B.
- **Risultato**:
  - Questa coppia è **concordante**, perché il paziente A ha avuto l'evento prima e il modello lo ha previsto correttamente.

Se si ripetono questi confronti per tutte le coppie, il **C-Index** rappresenterà la proporzione di coppie correttamente ordinate dal modello.

Concordanza perfetta tra risk score e tempi degli eventi→ C=1



Perfetta discordanza tra risk score e tempi degli eventi→ C=0



Modello che assegna i risk score in modo casuale→ C=0.5