RILEVAZIONE DEI DATI PER LA STIMA DI UN MODELLO DI SOPRAVVIVENZA

I dati per l'analisi della sopravvivenza si dicono **completi** se per ogni individuo osservato si osserva l'istante iniziale e l'evento; è quindi osservata la durata (time to event data).

Si possono avere due tipi di rilevazioni:

- le rilevazioni trasversali (cross-sectional studies)
- le rilevazioni longitudinali (longitudinal studies)

Rilevazioni trasversali

Si individua un gruppo di studio, cioè un gruppo di individui per i quali interessa studiare la sopravvivenza (per es. la popolazione di una certa zona o anche di un'intera nazione, gli assicurati di una compagnia di assicurazione, gli iscritti ad un fondo pensione, ...).

Si fissa un periodo di osservazione durante il quale viene osservato il gruppo di studio.

All'inizio dell'osservazione ci saranno individui già presenti, ai quali se ne aggiungeranno altri durante il periodo di osservazione; alcuni individui possono uscire per causa diversa dal decesso durante l'osservazione; ci saranno individui ancora in vita al termine dell'osservazione.

Tipicamente, nelle rilevazioni trasversali, i dati sono **incompleti**:

se non è osservato l'istante iniziale, l'osservazione è detta troncata a sinistra

se non è osservato l'evento, l'osservazione è detta censurata a destra

Rilevazioni longitudinali

Sono tipicamente usate negli studi clinici, quando le durate, dall'istante iniziale al verificarsi dell'evento, non sono molto lunghe.

Si seleziona un gruppo di studio, cioè un gruppo di individui per i quali interessa studiare la sopravvivenza, e si osserva il gruppo fino a quando per tutti gli individui è osservato l'evento e sono quindi osservate le durate dall'istante iniziale al verificarsi dell'evento.

Negli studi clinici, si possono spesso avere **dati completi**, se per ogni individuo è osservato l'istante iniziale, che può coincidere con una stessa data di calendario, ma non necessariamente. Si parla di *cohort complete design*.

Se si termina l'osservazione prima di avere osservato tutti gli eventi, si hanno dati incompleti (censurati a destra). Infatti, se per alcuni individui non si osserva l'evento, si perdono alcune informazioni sulla sopravvivenza della collettività oggetto di studio.

FUNZIONE DI SOPRAVVIVENZA EMPIRICA

Rilevazione longitudinale con dati completi

Osservazioni: n individui osservati; $t_1 < t_2 < \dots < t_n$ durate osservate

Si definisce **funzione di sopravvivenza empirica**

$$\hat{S}(t) = \begin{cases} 1 & t < t_1 \\ \frac{n-j}{n} & t_j \leq t < t_{j+1}, j = 1, 2, \dots, n-1 \\ 0 & t \geq t_n \end{cases}$$

La funzione di sopravvivenza empirica fornisce una stima della funzione di sopravvivenza $S(t)$

Sia

N_t il n.a. di individui in vita all'istante t a partire da $N_0 = n$ individui osservati all'istante 0

Lo stimatore $\tilde{S}(t)$ del quale la funzione di sopravvivenza empirica è la stima per $S(t)$ è

$$\tilde{S}(t) = \frac{N_t}{n}$$

Proprietà dello stimatore $\tilde{S}(t)$

Nell'ipotesi che le durate aleatorie di vita degli $N_0 = n$ individui osservati all'istante 0 siano ugualmente distribuite ed indipendenti con funzione di sopravvivenza $S(t)$, fissato $t > 0$ il n.a. N_t ha distribuzione Binomiale($n, S(t)$)

Si ha allora

$$E(N_t) = n S(t) \qquad \text{Var}(N_t) = n S(t) (1 - S(t))$$

Risulta quindi che $\tilde{S}(t)$ è uno stimatore non distorto, infatti per ogni $t > 0$ si ha

$$E(\tilde{S}(t)) = E\left(\frac{N_t}{n}\right) = S(t)$$

Si ha inoltre

$$\text{Var}(\tilde{S}(t)) = \text{Var}\left(\frac{N_t}{n}\right) = \frac{S(t) (1 - S(t))}{n}$$

Quindi una stima di $\text{Var}(\tilde{S}(t))$ è data da

$$\hat{\text{Var}}(\tilde{S}(t)) = \frac{\hat{S}(t) (1 - \hat{S}(t))}{n}$$

LA STIMA DI KAPLAN-MEIER PER LA FUNZIONE DI SOPRAVVIVENZA IN UN MODELLO DISCRETO

Rilevazione trasversale: per ogni individuo si osserva l'istante iniziale; alcuni individui danno luogo ad osservazioni censurate a destra. Quindi i dati sono incompleti.

Osservazioni: n individui osservati; m decessi osservati alle età $t_1 < t_2 < \dots < t_k$

$d_1, d_2, \dots, d_k$ numeri di decessi osservati alle età $t_1, t_2, \dots, t_k$

$$d_1 + d_2 + \dots + d_k = m$$

$n - m$ numero di osservazioni censurate

Sia t_0 l'età di ingresso in osservazione corrispondente all'istante iniziale

Si ripartisce l'intervallo $[t_0, +\infty[$ negli intervalli

$$[t_j, t_{j+1}[\quad j = 0, 1, \dots, k \quad \text{con} \quad t_{k+1} = +\infty$$

Sia c_j , $j = 0, 1, \dots, k$, il numero di osservazioni censurate nell'intervallo $[t_j, t_{j+1}[$

$$c_0 + c_1 + \dots + c_k = n - m$$

La stima di Kaplan-Meier per la funzione di sopravvivenza in un modello discreto

Con riferimento all'intervallo $[t_j, t_{j+1}[$, $j = 0, 1, \dots, k$, indichiamo con

$$t_{j1} \leq t_{j2} \leq \dots \leq t_{jc_j}$$

le età di uscita delle c_j osservazioni censurate

Sia

$$n_j, \quad j = 1, \dots, k,$$

il numero di individui in vita all'età t_j che danno luogo ai d_j decessi

Si ha

$$n_j = (d_j + d_{j+1} + \dots + d_k) + (c_j + c_{j+1} + \dots + c_k) \quad j = 1, \dots, k$$

Obiettivo: stimare la funzione di ripartizione $F(t)$
del n.a. T durata aleatoria di vita dall'età t_0

Disponendo delle osservazioni: decesso all'età t_j , $j = 1, \dots, k$
osservazione censurata all'età
 t_{ji} , $j = 0, \dots, k$, $i = 1, \dots, c_j$

La stima di Kaplan-Meier per la funzione di sopravvivenza in un modello discreto

Nell'ipotesi che i numeri aleatori osservati siano stocasticamente indipendenti ed ugualmente distribuiti con funzione di ripartizione $F(t)$ la verosimiglianza delle osservazioni è

$$L = \prod_{j=0}^k \left(\left[F(t_j) - F(t_j^-) \right]^{d_j} \cdot \prod_{i=1}^{c_j} [1 - F(t_{ji})] \right)$$

essendo

$$F(t_j^-) = \lim_{t \rightarrow t_j^-} F(t) \quad j = 1, \dots, k$$

$$\left[F(t_0) - F(t_0^-) \right]^{d_0} = 1$$

Obiettivo: determinare la stima $\hat{F}(t)$ di massima verosimiglianza della funzione di ripartizione $F(t)$

Osservazione: $L \neq 0 \Leftrightarrow F(t_j) - F(t_j^-) > 0 \quad \forall j = 1, \dots, k$, quindi
inoltre $1 - F(t_{ji})$ è massimo se $F(t_{ji}) = F(t_j)$

Quindi la verosimiglianza diventa

$$L = \prod_{j=1}^k \left[F(t_j) - F(t_j^-) \right]^{d_j} \cdot \prod_{j=0}^k [1 - F(t_j)]^{c_j}$$