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.

Rilevazioni longitudinali (*longitudinal studies*)

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.

Se per ogni individuo è osservato l'istante iniziale, e sono quindi osservate le durate dall'istante iniziale al verificarsi dell'evento, si dispone di **dati completi**. Si parla in tal caso di *cohort complete design*. Se si termina l'osservazione prima di avere osservato tutti gli eventi, si hanno dati incompleti (censurati a destra). In tal caso si perdono alcune informazioni sulla sopravvivenza della collettività oggetto di studio.

L'istante iniziale per i diversi individui osservati può coincidere con una stessa data di calendario, ma non necessariamente.

Rilevazioni trasversali (cross-sectional studies)

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

FUNZIONE DI SOPRAVVIVENZA EMPIRICA

Rilevazione longitudinale con dati completi

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

Si definisce **funzione di sopravvivenza empirica**

$$\hat{S}(t) = \begin{cases} 1 & t < t_1 \\ \dfrac{n-j}{n} & t_j \le t < t_{j+1}, \quad j = 1,2,\cdots n-1 \\ 0 & t \ge t_n \end{cases}$$

La funzione di sopravvivenza empirica è 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

$$E(N_t) = n\, S(t) \qquad Var(N_t) = n\, S(t)\,(1 - S(t))$$

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

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

Si ha inoltre

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

Quindi una stima di $Var\left(\tilde{S}(t)\right)$ è data da $\hat{Var}\left(\tilde{S}(t)\right) = \dfrac{\hat{S}(t)\,(1 - \hat{S}(t))}{n}$