Osservazioni:

- $F(t)$ è la funzione di ripartizione di un n.a. con distribuzione discreta e determinazioni $t_1, t_2, \ldots, t_k$
- per stimare $F(t)$ si tratta quindi di stimare le probabilità

$$q_j = F(t_j) - F(t_j^-) \qquad\qquad j = 1, \ldots, k$$

- ovvero di stimare la funzione di rischio

$$\lambda_j = \lambda(t_j) = \frac{q_j}{S(t_{j-1})} = \frac{F(t_j) - F(t_j^-)}{1 - F(t_{j-1})} \qquad\qquad j = 1, \ldots, k$$

essendo

$$S(t_j) = 1 - F(t_j) = \prod_{t_i \le t_j} (1 - \lambda(t_i)) = \begin{cases} 1 & \text{se } j = 0 \\ (1-\lambda_1)(1-\lambda_2)\cdots(1-\lambda_j) & \text{se } j = 1, \ldots, k \end{cases}$$

Si ottiene allora la seguente espressione per la verosimiglianza

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

Data la verosimiglianza

$$L = \prod_{j=1}^{k} \lambda_j^{d_j} \left[1-\lambda_j\right]^{n_j-d_j}$$

Osservazione

$\lambda_j^{d_j} \left[1-\lambda_j\right]^{n_j-d_j}$ è proporzionale alla verosimiglianza di un n.a. con distribuzione Binomiale(n_j, λ_j)

Sia D_j il n.a. dei decessi all'età t_j a partire dagli n_j individui in vita all'età t_j prima che si verifichino i d_j decessi

Si ipotizza per D_j la distribuzione Binomiale(n_j, λ_j)

La stima di massima verosimiglianza di λ_j è

$$\hat{\lambda}_j = \frac{d_j}{n_j}$$

La stima di Kaplan-Meier per la funzione di sopravvivenza è allora

$$\hat{S}(t) = \prod_{t_j \le t} \left(1 - \hat{\lambda}_j\right) = \prod_{t_j \le t} \left(1 - \frac{d_j}{n_j}\right)$$

Osservazione

Poiché

$$S(t) = \prod_{t_j \le t} \left(1 - \lambda(t_j)\right) = \prod_{t_j \le t} P(T > t_j \,|\, T > t_{j-1})$$

si può allora interpretare la formula

$$\hat{S}(t) = \prod_{t_j \le t} \left(1 - \hat{\lambda}_j\right) = \prod_{t_j \le t} \left(1 - \frac{d_j}{n_j}\right)$$

come prodotto delle stime delle probabilità condizionate di sopravvivenza.

La stima di massima verosimiglianza, di Kaplan-Meier, della funzione di ripartizione $F(t)$ è allora

$$\hat{F}(t) = 1 - \hat{S}(t) = 1 - \prod_{t_j \le t} \left(1 - \frac{d_j}{n_j}\right)$$

Esempio

j	t_j	n_j	d_j	c_j	$1-d_j/n_j$	$\hat{S}(t_j)$
0	0	8	0	1	1	1
1	1,5	7	1	0	0,857143	0,857143
2	2	6	2	1	0,666667	0,571429
3	3,5	3	1	2	0,666667	0,380952

Osservazioni

- La stima di Kaplan-Meier $\hat{F}(t)$ è definita per $t \le t_k$
- Se l'età massima osservata è un dato censurato, cioè se $c_k > 0$, la $\hat{F}(t)$ non è definita per $t > t_k$ ed è $\hat{F}(t_k) < 1$; una possibilità è definire $\hat{F}(t) = \hat{F}(t_k)$ per $t > t_k$ ma ciò comporta la presenza di masse aderenti a $+\infty$
- La stima di Kaplan-Meier $\hat{F}(t)$ può essere definita anche in presenza di osservazioni troncate a sinistra, cioè di osservazioni per le quali non si è osservato l'istante iniziale essendo entrate in osservazione ad un'età maggiore di t_0. Tuttavia, in tal caso $\hat{S}(t)$ è la stima della funzione di sopravvivenza condizionata alla età minima di ingresso osservata $t_{\min}$, essendo $t_0 \le t_{\min} < t_1$; quindi è una stima di $P(T > t | T > t_{\min}) = S(t)/S(t_{\min})$

Proprietà dello stimatore di Kaplan-Meier

$$\hat{S}(t) = \prod_{t_j \le t} \left(1 - \frac{d_j}{n_j}\right)$$ è la stima di Kaplan-Meier di $S(t)$, $t \ge 0$

Sia

$$S(t) = \prod_{t_j \le t} p_j \qquad \text{con} \qquad p_j = 1 - \lambda(t_j) = P(T > t_j \mid T > t_{j-1})$$

si ha

$$\hat{S}(t) = \prod_{t_j \le t} \hat{p}_j \qquad \text{essendo} \qquad \hat{p}_j = 1 - \frac{d_j}{n_j}$$

dove $\hat{p}_j$ è una stima della probabilità $P(T > t_j \mid T > t_{j-1})$, $j = 1, \ldots, k$

Indichiamo con

$$\tilde{S}(t) = \prod_{t_j \le t} \tilde{p}_j \ , \qquad t \ge 0$$

lo stimatore del quale $\hat{S}(t)$ è la stima.

Per valutare speranza matematica e varianza dello stimatore $\tilde{S}(t)$ occorre formulare delle ipotesi sui n.a. $\tilde{p}_j$, $j = 1, \ldots, k$

Osservazione

Sia

D_j il n.a. dei decessi da una popolazione di n_j individui in vita all'età t_j

nell'ipotesi che D_j abbia distribuzione Binomiale(n_j, q_j) si ha

$$E(D_j) = n_j\ q_j \qquad Var(D_j) = n_j\ q_j\ (1-q_j)$$

si ha inoltre che il n.a.

$\frac{D_j}{n_j}$ ha distribuzione Binomiale scalata ed è

$$E\left(\frac{D_j}{n_j}\right) = q_j \qquad Var\left(\frac{D_j}{n_j}\right) = \frac{q_j\ (1-q_j)}{n_j}$$

Sia $\mathcal{I} = \{n_0, n_1, , \ldots, n_k\}$ lo stato di informazione sugli individui presenti alle età t_j, $j = 1, \ldots, k$ subito prima che si osservino i decessi d_j, $j = 1, \ldots, k$

Si formulano le seguenti ipotesi condizionate a $\mathcal{I} = \{n_0, n_1, , \ldots, n_k\}$ sui n.a.

$$\tilde{p}_j = 1 - \frac{D_j}{n_j}, \qquad j = 1, \ldots, k$$

Ipotesi

Condizionatamente a $\mathcal{I} = \{n_0, n_1, , \ldots, n_k\}$, i n.a.

$$\tilde{p}_j = 1 - \frac{D_j}{n_j}, \qquad j = 1, \ldots, k$$ siano stocasticamente indipendenti

e siano

$$E(\tilde{p}_j | \mathcal{I}) = p_j \qquad Var(\tilde{p}_j | \mathcal{I}) = \frac{p_j\,(1 - p_j)}{n_j} \qquad j = 1, \ldots, k$$

Risulta allora che $\tilde{S}(t) = \prod_{t_j \le t} \tilde{p}_j$ è uno stimatore non distorto, infatti

$$E\left(\tilde{S}(t)\middle|\mathcal{I}\right) = E\left(\prod_{t_j \le t} \tilde{p}_j \middle|\mathcal{I}\right) = \prod_{t_j \le t} p_j = S(t) \qquad t \ge 0$$

La varianza dello stimatore $\tilde{S}(t)$ è

$$Var\left(\tilde{S}(t)\middle|\mathcal{I}\right) = [S(t)]^2 \left[\prod_{t_j \le t}\left(1 + \frac{1-p_j}{p_j n_j}\right) - 1\right]$$

e può essere approssimata da

$$Var\left(\tilde{S}(t)\middle|\mathcal{I}\right) \cong [S(t)]^2 \sum_{t_j \le t} \frac{1-p_j}{p_j n_j}$$

dalla quale si ottiene la **formula di Greenwood**, che fornisce una stima della varianza dello stimatore di Kaplan-Meier

$$\hat{Var}\left(\tilde{S}(t)\middle|\mathcal{I}\right) = \left[\prod_{t_j \le t}\left(1 - \frac{d_j}{n_j}\right)\right]^2 \sum_{t_j \le t} \frac{d_j}{(n_j - d_j)n_j}$$

LA STIMA DI NELSON-AALEN PER LA FUNZIONE DI SOPRAVVIVENZA

Fornisce una formula approssimata per la stima della funzione di sopravvivenza di Kaplan-Meier.

Dalla stima di Kaplan-Meier della funzione di sopravvivenza $\hat{S}(t) = \prod_{t_j \le t} \left(1 - \frac{d_j}{n_j}\right) \quad t \ge 0$

si ottiene

$$\log\left(\hat{S}(t)\right) = \sum_{t_j \le t} \log\left(1 - \frac{d_j}{n_j}\right) \qquad t \ge 0$$

Considerando l'approssimazione lineare della funzione $f(x) = \log(1-x)$ si ottengono le seguenti approssimazioni: $\log\left(1 - \frac{d_j}{n_j}\right) \cong -\frac{d_j}{n_j}$

Si ottiene allora la seguente stima della funzione di sopravvivenza, detta **stima di Nelson-Aalen**

$$\hat{S}(t) = \exp\left(-\sum_{t_j \le t} \frac{d_j}{n_j}\right) \qquad t \ge 0$$