La fonction de distribution ou répartition empirique

La function de survie empirique

Fonction de répartition empirique

- Observations $X_1, \ldots, X_n \; IID \sim F(x)$ ($F_{\theta_0}(x)$)
- La fonction de répartition empirique F_n(t) telle que $F_n(t) \xrightarrow{p}$F(t)
- F_n(t)=$\frac{nombre\ observations \leq t}{n}$
- F_n(t)=$\frac{1}{n}\sum_{i=1}^{n} I[x_i \leq t]$, t est fixé et connu.
- F_n(t) fait un saut de taille $\frac{1}{n}$ pour chaque observation rencontrée
- S_n(t)=$\frac{1}{n}\sum_{i=1}^{n} I[x_i > t]$, la fonction de survie empirique

Fonction de survie empirique

- $F_n(t)$ est une distribution discrete qui assigne des masses de taille $\frac{1}{n}$ a chaque point des observations:
- $--- - -x_{(1)}$--------$x_{(2)}$-------------------$x_{(n)}$------
- $\frac{1}{n}$ $\frac{1}{n}$ $\frac{1}{n}$
- $F_n(t)$: $\frac{1}{n}$ $\frac{2}{n}$ 1
- 1-$F_n(t)$: $1-\frac{1}{n}$ 1- $\frac{2}{n}$ 0
- $S_n(t)$
- La fonction de distribution théorique est continue
- $-x_{(1)} < x_{(2)} \ldots < x_{(n)}$

Notations:

- $E(X) = \int_0^\infty xf(x)dx = \int_0^\infty xdF(x)$.
- Estimateur empirique pour $E(X)$ est la moyenne empirique
- $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $\bar{X}$= $\int_0^\infty xdF_n(x)$
- Intégrer par parties
- $E(X)$=$\int_0^\infty S(x)dx = \int_0^\infty S_n(x)dx$
- Estimateur empirique pour $V(X)$ =E($(X-\mu)^2$)
- $\frac{1}{n}\sum_{i=1}^{n}(X_i$- $\bar{X})^2$

Percentile empirique

- F^{-1}(p) est bien définie, $F(.)$ est une fonction croissante
- F_n^{-1}(p) n'est pas bien définie, $F_n(.)$ est une fonction en escalier.
- Exemple:
- n=5

	$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$
$F_n(.)$	$\frac{1}{5}$	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{4}{5}$	$\frac{5}{5}$=1

- $F_n^{-1}(1/2)$ n'est pas définie
- $\left[x_{(2)}, x_{(3)}\right] \to F_n^{-1}(1/2)$

Moyenne et variance de F_n

- $E(F_n$ (t))=$F(t)$
- $V(F_n$ (t))=$V(S_n$ (t))=$\frac{1}{n}$ $(F(t))$(1- $F(t)$)

- F_n (t) est un estimateur sans biais pour $F(t)$
- F_n (t) est un estimateur convergent pour $F(t)$.
- L'approche conditionnelle pour retrouver F_n ou S_n .Cette approche sera réutilisée plus tard dans l'analyse de survie òu les données pourraient être censurées(incomplètes), chapitre 14 (Klugman et collègues).

- $0 - -t_1$----t_2 $---$ $-t_j$ $--$ $-t - -t_{j+1}$
- $S(t) = \frac{S(t_1)}{S(t=0)} \times \frac{S(t_2)}{S(t_1)} \times \cdots \times \frac{S(t)}{S(t_j)}$=$\frac{S(t)}{1}$
- $\frac{S(t_i)}{S(t_{i-1})} = P(T > t_i | \, T > t_{i-1})$=1$- P(T \leq t_i | \, T > t_{i-1})$

- $0 - -t_1$---- $---$ $-$ $t_2$$---$ $-t_j$ $--$ $-t - -t_{j+1}$
- 0 $x_{(1)}$]($x_{(1)}$ --- $x_{(2)}$] --- ($x_{(j)}$ t]

L'approche conditionnelle

- $1 - P(T \leq t_i | \, T > t_{i-1})$ peut être estimé par
- $1 - \frac{1}{n-i+1} = \frac{n-i}{n-i+1}$
- $\prod_{i \leq j} \left(\frac{n-i}{n-i+1}\right) = \frac{n-1}{n} \times \frac{n-2}{n-1} \times \cdots \times \frac{n-j}{n-j+1} = \frac{n-j}{n} = 1 - \frac{j}{n} = 1 - F_n(t)$.