Partie III

Suite du cours

"Analyse de Survie"

§ Estimation paramétrique en survie

T v. a. de survie : f_T, S_T, F_T

C v. a. de censure : F_C, S_C, f_C

On suppose que la loi de T dépend d'un paramètre θ :

$$\mathbb{P}_T = \mathbb{P}_{\theta, T} \text{ dépend de } \theta$$

(loi paramétrique)

On observe T dans le cadre censuré :

on observe : $X_i = T_i \wedge C_i$, $i = 1, \dots, n$

$D = \mathbb{1}_{(T \leq C)}$ $T_i \perp\!\!\!\perp C_i$ censure à droite

Il s'agit d'estimer le paramètre θ à partir de l'échantillon : $x_1, x_2, \dots, x_n$.

Méthode de max de vraisemblance

Préliminaires de calculs :

Calcul de ① $\mathbb{P}(X \leq x, D = 1) = \int_0^x f_T(u) S_C(u) du$ (cf TD)

aussi on a :

② $\mathbb{P}(X \leq x, D = 0) = \int_0^x f_C(u) S_T(u) du$.

D'où les "densités" suivantes (en dérivant) :

① $\to$ si $D = 1$: $f_T(x) \cdot S_C(x) \cdot 1$ $- f(0) S(0)$

② $\to$ si $D = 0$: $f_C(x) \cdot S_T(x) \cdot 1$

Ainsi la densité de X_i s'écrit :

$$\left(f_T(x) S_C(x)\right)^D \cdot \left(f_C(x) S_T(x)\right)^{1-D}$$

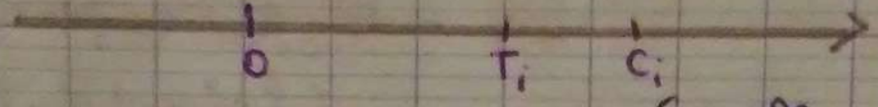
Or $\mathbb{P}(X \leq x) = \mathbb{P}(X \leq x, D = 1) + \mathbb{P}(X \leq x, D = 0)$.

D'où la vraisemblance de cas censuré :

$$L_n(x_1, \dots, x_n, \theta) = \prod_{i=1}^{n} \left(f_T(x_i, \theta) S_C(x_i) \right)^{D_i} \left(f_C(x_i) S_T(x_i \dots \right.$$

Remarque :

Dans le cas ~~de la~~ où il n'y a pas de censure : on observe $T_1 \dots T_n$



$C = +\infty$

$$S_C(t) = \mathbb{P}(C > t) = 1$$

$$F_C(t) = \mathbb{P}(C \leq t) = 0$$

$$L_n(T_1, \dots T_n, \theta) = \prod_{i=1}^{n} f_T(T_i, \theta)$$

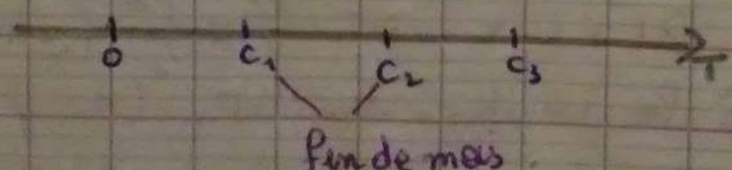
Ainsi on ~~Approcher~~ l'EMV :

$$\boxed{\hat{\theta}_n = \operatorname{Arg}\max_{\theta} L(x_1, \dots, x_n, \theta)}.$$

Exemple : $T \rightsquigarrow \mathcal{E}(\theta)$, C la v.a de censure

$C_i = c_i$ $i = 1, \dots, n$

→ constante donnée

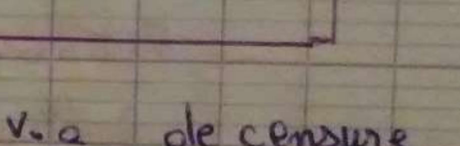


$$f_T(t, \theta) = \theta e^{-\theta t} \quad , t > 0 , \theta > 0$$

$$S_T(t, \theta) = e^{-\theta t} \quad , t > 0$$

$$f_{C_i}(t) = 1 \text{ si } t = c_i$$

$$= \mathbb{1}_{\{c_i\}}(t) = \delta_{c_i}(t)$$



$\mathbb{P}(C = c_i) = 1$

$S_{C_i}(t) = \mathbb{1}_{[0, c_i[}(t)$

D'où la vraisemblance :

$$L(x_1, x_2, \dots, x_n, \theta) = \prod_{i=1}^{n} \left(\theta e^{-\theta x_i} \mathbb{1}_{(x_i \geq 0)} \mathbb{1}_{[0, c_i[}(x_i) \right)^{D_i}$$

$$\cdot \left(\mathbb{1}_{\{c_i\}}(x_i) \cdot e^{-\theta x_i} \mathbb{1}_{(x_i \geq 0)} \right)^{1-D_i}$$

$$= \prod_{i=1}^{n} \left(\left(\theta e^{-\theta x_i} \mathbb{1}_{[0, c_i[}(x_i) \right)^{D_i} \left(e^{-\theta c_i} \right)^{1-D_i} \right)$$

$$\Rightarrow \ell_n(x_1, \dots, x_n, \theta) := \log L_n(x_1, \dots, x_n, \theta)$$

$$= \sum_{i=1}^{n} \left[D_i \log(\theta e^{-\theta x_i}) \mathbb{1}_{\{0 \leq x_i < c_i\}} + (1 - D_i) \log e^{-\theta c_i} \right]$$

$$\frac{\partial \ell_n}{\partial \theta} = \sum_{i=1}^{n} \left[D_i \left(\frac{1}{\theta} - x_i \right) \cdot \mathbb{1}_{(0 \leq x_i < c_i)} - (1 - D_i) \cdot c_i \right] = 0$$

$$0 = \frac{1}{\theta} \left(\sum_{i=1}^{n} D_i \right) - \left(\sum_{i=1}^{n} D_i x_i \right) - \left(\sum_{i=1}^{n} (1 - D_i) c_i \right)$$

si $0 \leq x_i < c_i$

$$\Rightarrow \boxed{\hat{\theta}_n = \frac{\sum_{i=1}^{n} D_i}{\sum_{i=1}^{n} D_i x_i + \sum_{i=1}^{n} (1 - D_i) c_i}}$$

si $0 \leq x_i < c_i$

Rem : $T \sim \mathcal{E}(\theta)$ $\quad T_1, \dots, T_n$ cas non censuré

l'EMV est :

$$L_n(T_1, \dots, T_n, \theta) = \prod_{i=1}^{n} f_T(T_i, \theta)$$

$$= \prod_{i=1}^{n} \left(\theta e^{-\theta T_i} \right) \mathbb{1}_{\{T_i \geq 0\}} = \theta^n e^{-\theta \sum_{i=1}^{n} T_i}$$

si $T_i \geq 0$

$L_n(T_1, \dots, T_n, \theta) = n \log \theta - \theta \sum_{i=1}^{n} T_i$

$$\frac{\partial \ell_n}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^{n} T_i = 0$$

$$\Rightarrow \boxed{\hat{\theta}_n^{a} = \frac{n}{\sum_{i=1}^{n} T_i}} \quad (D_i = 1)$$

§. Processus ponctuels :

. Processus de Poisson.

Soit (Y_n) v.a i.i.d de loi $B(p)$ $0<p<1$

Soit (B_n) une suite de v.a.r :

$$B_0 = 0 \quad , \quad B_n = \sum_{i=1}^{n} Y_i$$

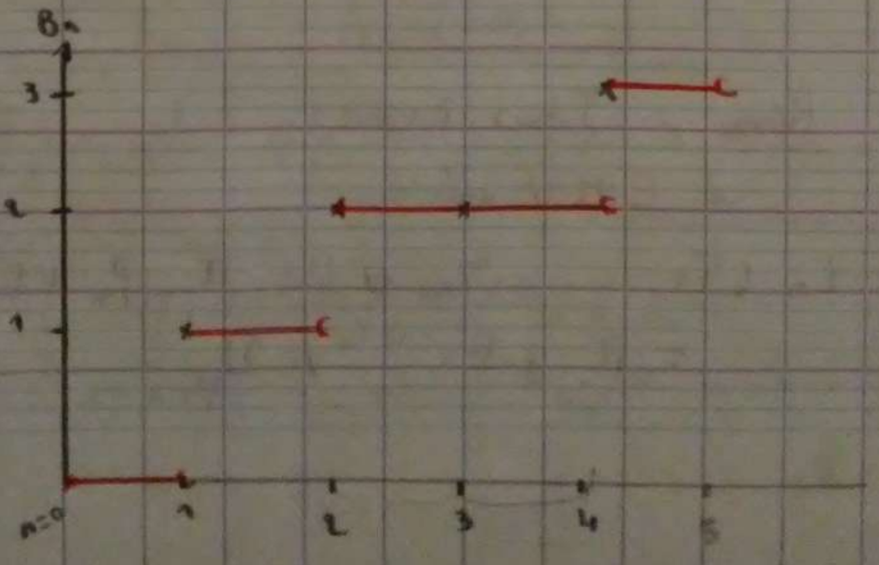
on a : $B_n = B_{n-1} + Y_n$

les trajectoires de (B_n) : (ω fixé (la suite $(B_n(\omega))$.

n varie : 1, 2, ...

$B_n(\omega) = \sum_{i=1}^{n} Y_i(\omega)$, $B_0(\omega) = 0$.

Ainsi pour voir $B_1(\omega), B_2(\omega), \dots, B_n(\omega)$



$Y_1(\omega) = 1 \qquad B_2(\omega) = \underset{1}{Y_1(\omega)} + \underset{1}{Y_2(\omega)}, \quad B_3(\omega) = Y_1 + Y_2 + Y_3$

les trajectoires de B_n sont ↗ (croissantes) au sens large.
$(B_n)_{n \geq 0}$ ~~est~~ appelé ~~le~~ processus de Bernoulli

Les instants de sauts de (B_n) : ω : 1, 2, 4 ...
on note les instants de sauts : $\underset{1^{er}\text{ saut}}{S_1}, S_2, \ldots, \underset{n^{ème}\text{ saut}}{S_n}$

$(S_n)_{n \geq 1}$

$S_1(\omega) = 1$

$S_2(\omega) = 2$

$S_3(\omega) = 4.$

$$S_{n+1} = \operatorname{Inf}\{k > S_n \mid B_k \neq B_{S_n}\}$$

on pose $E_n = S_n - S_{n-1}$, $E_{n+1} = S_{n+1} - S_n$

les écarts entre les sauts.

$E_n(\omega) \ldots$

on montre que : (E_n) sont i.i.d de loi $\mathcal{G}(p)$:

loi géométrique $\mathbb{P}(E_n = \ell) = p(1-p)^{\ell-1}$
$\ell \in \mathbb{N}^*$

$$\sum_{\ell=1}^{\infty} p(1-p)^{\ell-1} = p \sum_{\ell=1}^{\infty} (1-p)^{\ell-1} = p \frac{1}{1-(1-p)} = 1$$

$1 + (1-p) + (1-p)^2 + \ldots$

on a aussi $S_n = E_1 + E_2 + \ldots + E_n$.

car $S_n = S_{n-1} + E_n \stackrel{?}{=} S_{n-2} + E_{n-1} + E_n \ldots$

Propriétés sur la loi $g(p)$:

on a l'equivalence :

i) $\mathcal{L}(X) = g(p) \Leftrightarrow$ ii) $\mathcal{L}(X-n \mid X>n) = \mathcal{L}(X)$

caracterisation de la loi $g(p)$.

Dém : $X \sim g(p)$

$\Rightarrow$ on pose $G(n) = \mathbb{P}(X>n) = \sum_{\ell=n+1}^{\infty} P(X=\ell)$

$$= \sum_{\ell > n+1} p(1-p)^{\ell-1}$$

$$= \underbrace{\sum_{\ell=1}^{\infty} p(1-p)^{\ell-1}}_{1} - \sum_{\ell=1}^{n} p(1-p)^{\ell-1}$$

Mq : $\mathbb{P}(X-n=\ell \mid X>n) = p(1-p)^{\ell-1}$??

$$P(X=n+\ell \mid X>n) = \frac{P(X=n+\ell, X>n)}{P(X>n)}$$

$$= \frac{\mathbb{P}(X=n+\ell)}{\mathbb{P}(X>n)} = \frac{p(1-p)^{n+\ell-1}}{?}$$

exercice

Mq $G(n) \cdot G(m) = G(n+m)$? exercice

alors : $G(n) = G(1 + \overbrace{1+\dots+1}^{n-1}) = G(1) \cdot G(n-1)$

$$= (G(1))^2 G(n-2)$$

$$= (G(1))^n$$

$$= (P(X>1))^n$$

$$= (1 - \underbrace{P(X=1)}_{p})^n$$

Ainsi :

$$\mathbb{P}(X=\ell) = \mathbb{P}(X \geq \ell,\ X > \ell)$$

$$= \mathbb{P}(X \geq \ell \mid X > \ell)\, \mathbb{P}(X > \ell)$$

$P(X = 2\ell \mid X > \ell) = P(X = \ell)$ $\qquad = \sum_{j=\ell}^{\infty} P(X = j \mid X > \ell)\, P(X > \ell) = \dots = p(1-p)$

$G(n) = \mathbb{P}(X > n)$ **Autre methode :**

$$\mathbb{P}(X=\ell) = \mathbb{P}\big((X \geq \ell) \setminus (X > \ell)\big)$$

$$= \mathbb{P}(X \geq \ell) - \mathbb{P}(X > \ell) \quad P(X > \ell) \subset P(X \geq \ell)$$

$$= \mathbb{P}(X > \ell - 1) - P(X > \ell)$$

$= G(\ell-1) - G(\ell) = (G(1))^{\ell-1} - (G(1))^{\ell} =$ Reste à verifier que $G(1) = 1-p$??

$$\mathbb{P}(X=\ell) = (1-p)^{\ell-1} - (1-p)^{\ell} = (1-p)^{\ell-1}(1 - \underbrace{(1-p)}_{p})$$

$$= p(1-p)^{\ell-1}$$

Prop 1 : on a l'équivalence :

i) $\mathcal{L}(X) = \mathcal{E}(\lambda)$, $\lambda > 0$ $\iff$ ii) $\mathcal{L}(X - t \mid X > t) = \mathcal{L}(X)$

Prop 2 : Si $X \sim g(p_\ell)$, $0 < p_\ell < 1$ $\quad \forall \ell \geq 0 \quad t \in \mathbb{R}^+$

et $Z_\ell = \frac{X}{\ell}$, $\ell \geq 1$

Alors si $\ell p_\ell \xrightarrow[\ell \to \infty]{} \lambda > 0$ alors $Z_\ell \xRightarrow[\ell \to +\infty]{\text{en loi}} Z \sim \mathcal{E}(\lambda)$

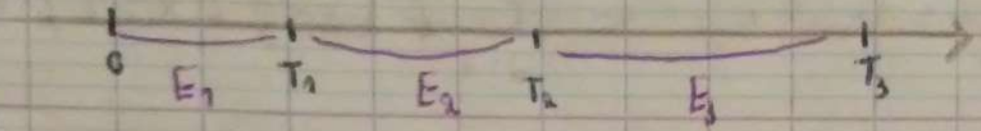
Rem : Si $Y \sim \mathcal{E}(\lambda)$ alors $[Y]$ et $Y - [Y]$ sont indép. et $1 + [Y] \sim g(e^{-\lambda})$ exercice

(X se fait avec les fonctions génératrices)

• **Processus de Poisson** :

Soit (E_n) v.a i.i.d $\sim \mathcal{E}(\lambda)$

$(T_n)_{n\geq 0}$ v.a.r : $T_0 = 0$, $T_n = \sum_{i=1}^{n} E_i$



$E_n = T_n - T_{n-1}$.

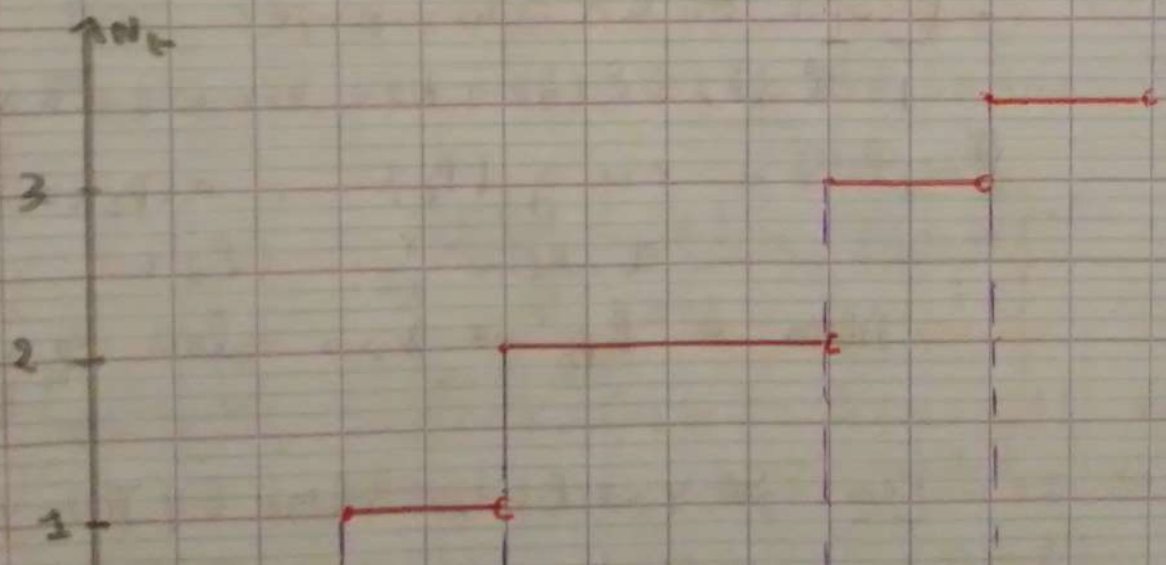
On définit N_t par

N_t = nbre de $T_i \in [0,t]$.

$$N_t(\omega) = \sum_{i\geq 1} \mathbb{1}_{[0,t]}(T_i(\omega)) = \sum_{i\geq 1} \mathbb{1}_{(T_i \leq t)}.$$

$N_t(\omega)$ = nbre de $T_i(\omega) \in [0,t]$

$(N_t)_{t\geq 0}$ est appelle un processus de Poisson (ou processus de comptage) d'intensité λ



Le trajectoir de processus de Poisson sont croissant au sens large et les T_i sont les de sauts de N_t.

Prop 3 on a :

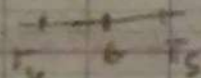
1) $N_{T_n} = n$

2) $N_t = \text{Inf}\{k \geq 0 \mid T_{k+1} > t\}$

$\forall t > 0 \quad N_t = \sum_{n \geq 0} n \, \mathbb{1}_{[T_n, T_{n+1}[}(t)$

Dans cette somme tout les termes sont nuls souf un terme où $t \in [T_n, T_{n+1}[$.

$\text{Inf}\{j - 1 \mid T_j > t\}$.

$5 - 1 = 4$ 

Thm (admis) : Soit $(N_t)_{t \geq 0}$ un pr. de Poisson d'intensité λ.

$(T_n)_{n \geq 1}$ les instants de sauts du processus de Poisson.

Alors :

1) $\mathcal{L}(T_1, T_2, \ldots, T_n)$ a pour densité :

$$f(t_1, \ldots, t_n) = \lambda^n e^{-\lambda t_n} \mathbb{1}_{(0 < t_1 < t_2 < \cdots < t_n)}(t_1, \ldots, t_n)$$

2) $\mathcal{L}(T_n)$ a pour densité :

$\forall n \geq 1$
$$f_{T_n}(t) = \frac{t^{n-1}}{(n-1)!} \lambda^n e^{-\lambda t}, \quad t > 0$$

3) $\mathcal{L}(T_1, T_2, \ldots, T_n / T_{n+1} = t)$ a pour densité :

$$f(t_1, \ldots, t_n) = n! \, t^n \mathbb{1}_{0 < t_1 < \cdots < t_n}$$

4) $\mathcal{L}(N_t) = \mathcal{P}(\lambda t)$

$$\mathbb{P}(N_t = \ell) = e^{-\lambda t} \frac{(\lambda t)^\ell}{\ell!}, \quad \mathbb{E}(N_t) = \lambda t$$

5) $\mathcal{L}(T_1, \dots, T_n \mid N_t = n) =$ loi donnée au 3).

Applications la dynamique des populations

§ Comparaison de 2 survies:

Dans certains applications médicales, on a 2 groupes de malades G_1 et G_2 à qui on donne 2 traitements ~~pour que~~ ~~[illegible]~~, On veut regarder lequel des traitements est efficace

Si $S_2(t)$ est plus grande que $S_1(t)$ cela signifie que la survie dans G_2 est meilleur que dans G_1 c.a.d le traitement 2 est plus efficace.

On estime S_1 et S_2 par $\hat{S}_{KM_1}$ et $\hat{S}_{KM_2}$ et il s'agit de comparer ces 2 estimateurs.

Ds certains graphes on arrive à conclure tout de suite par contre ds certains les courbes se chevauchent

1ère approche:

Il s'agit de calculer une distance entre les

2 survies :

Il s'agit d'un problème de test :

on veut tester $H_0 : S_1 = S_2$ (égalité des 2 survies ds les 2 groupes).

Les 2 traitements ont le m̂ effet.

contre $H_1 : S_1 \neq S_2$ (effet différent des 2 traitements ~~diff~~).

Pour cela on introduit la distance suivante :

$$T_n(t) = \frac{\left(\hat{S}_{n_1}(t) - \hat{S}_{n_2}(t)\right)^2}{V\left(\hat{S}_{n_1}(t)\right) + V\left(\hat{S}_{n_2}(t)\right)} \quad \text{et} \quad d^*(\hat{S}_{n_1}, \hat{S}_{n_2}) = \sup_{t \geq 0} T_n(t)$$

n nombre d'observations

$\hat{S}_{n_1}$, $\hat{S}_{n_2}$ estimateurs de S_1 et S_2

Règle : si cette distance est "voisine" de 0 (significativement nulle)

alors on accepte H_0.

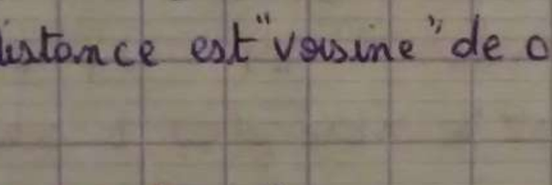
$$d^*(\hat{S}_{n_1}, \hat{S}_{n_2}) \approx 0 \Rightarrow \hat{S}_{n_1} \approx \hat{S}_{n_2}.$$

pour α donné on définit un région de rejet

$$D = \{T_n(t) > d_\alpha\}$$

Prop : Sous H_0 ! on a $T_n(t) \rightsquigarrow \chi^2_1$ qd n est assez grand.

$\alpha = \mathbb{P}(T_n(t) > d_\alpha) = \mathbb{P}(\chi^2_1 > d_\alpha)$ table → d_α



2ème approche : test du Log-Rank.

On a 2 groupes G_1 et G_2.

Il s'agit de tester $H_0 : S_1 = S_2$ (égalité de survie ds les 2 groupes)

On note : $t_1 < t_2 < \dots < t_L$ les instants de survenue de l'év^t E ou censures des 2 groupes.

Pour chaque t_k et $i = 1$ ou 2 :

$n_i(t_k)$ = nombre d'individus "à risque" à t_k ds le groupe i

$d_i(t_k)$ = nbre d'évènements E observés à t_k ds le groupe i

les 2 groupes {
on note $d(t_k) = d_1(t_k) + d_2(t_k)$
c'est le nbre d'év E observé à t_k ds les 2 groupes
$n(t_k) = n_1(t_k) + n_2(t_k)$
nbre d'individus à risque à t_k ds les 2 groupes.

On fait le tableau suivant (à l'instant t_k) :

	nbre d'ev E observé à t_k	Nbre d'individus à risque à t_k	nbre d'indiv en vie
G_1	$d_1(t_k)$	$n_1(t_k)$	$n_1(t_k)-d_1(t_k)$
G_2	$d_2(t_k)$	$n_2(t_k)$	$n_2(t_k)-d_2(t_k)$
Somme = Total (pour les 2 groupes)	$d(t_k)$	$n(t_k)$	$n(t_k)-d(t_k)$

On note : O_1 = nbre total d'evet E observés ds G_1

$$= \sum_{k \mid t_k \in G_1} d_1(t_k).$$

O_2 = nbre total d'event E observés ds G_2

$$= \sum_{\ell \mid t_\ell \in G_2} d_2(t_k).$$

O_1 et O_2 sont observés.

on ~~note~~ défint :

$$E_1 := \sum_k \overbrace{n_1(t_k)\,\frac{d(t_k)}{n(t_k)}}^{E_{1,k}} \quad \text{pour } G_1$$

$$E_2 := \sum_k n_2(t_k)\,\frac{d(t_k)}{n(t_k)} \quad \text{pour } G_2$$

Avec $E_{i,k} := n_i(t_k) \cdot \frac{d(t_k)}{n(t_k)}$, $i = 1,2$

~~Ex~~ : $E_{2,\ell} = n_2(t_\ell) \cdot \dfrac{d(t_\ell)}{n(t_\ell)} = (n(t_\ell) - n_1(t_\ell)) \dfrac{d(t_\ell)}{n(t_\ell)}$

$= d(t_\ell) - E_{1,\ell}$

$\Rightarrow E_{1,\ell} + E_{2,\ell} = d(t_\ell)$.

Sous H_0 : $S_1 = S_2$, on a : $O_1 + O_2 = E_1 + E_2$.

Car : $O_1 + O_2 = \sum_{\ell=1}^{L} d(t_\ell) = \sum_\ell (E_{1,\ell} + E_{2,\ell})$

$= \sum_\ell E_{1,\ell} + \sum_\ell E_{2,\ell} = E_1 + E_2$.

la statistique $O_1 - E_1$ (ou $O_2 - E_2$) est appelée statistique du LogRank.

E_1 (ou E_2) le nbre "espéré" d'event. E observés.

On introduit la statistique suivante :

$$\frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2}$$

(distance !)

Prop : Sous H_0 :

$$\frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} = (O_1 - E_1)^2 \left(\frac{1}{E_1} + \frac{1}{E_2}\right) \underset{n \to \infty}{\rightsquigarrow} \chi^2_1$$

D'où le test pour H_0 : $S_1 = S_2$

α donné $\to$ test de χ^2 : d_α

la région de rejet est : $D = \{T_n > d_\alpha\}$.

Rem: Sur E_1 et E_2 : nbre espéré d'év^t E observés

à t_ℓ : $\frac{d(t_\ell)}{n(t_\ell)} \simeq$ proba d'avoir "E" à t_ℓ.

$\simeq p_\ell$ ...

$$n_1 p_\ell = n_1 \frac{d(t_\ell)}{n(t_\ell)} = \underbrace{E(Z_\ell)}_{\text{d'avoir l'ev}} \sim \mathcal{B}(n_1, p_\ell) = Z_\ell$$

$$E_1 = \sum_k n_1 \frac{d(t_\ell)}{n(t_\ell)}.$$

On présente l'étude du logRank sous la forme du tableau suivant

t_ℓ	$n_1(t_\ell)$	$d_1(t_\ell)$	$n_2(t_\ell)$	$d_2(t_\ell)$	$n(t_\ell)$	$d(t_\ell)$	$n_1 \frac{d(t_\ell)}{n(t_\ell)}$	$n_2 \frac{d(t_\ell)}{n(t_\ell)}$
		$O_1 = \sum d_1(t_\ell)$		$O_2 = \sum d_2(t_\ell)$			$E_1 = \sum$	$E_2 = \sum$