Cours de M1 (Stat - Proba):

Analyse de Survie

$$\frac{1}{\sqrt{n}} \sum_{1}^{?} z_i \Rightarrow N(0, F(t) \cdot S(t))$$

or : $\frac{1}{\sqrt{n}} \left(\hat{\sum_{i=1}} (1)_{(T_i \le t)} - F(t) \right) \Rightarrow N(0, F(t) S(t))$

$\sqrt{n} (F_n(t) - F(t)) \Rightarrow N(0, F(t) S(t))$

Symét $\begin{cases} = -\sqrt{n} (S_n(t) - S(t)) \Rightarrow N(0, F(t) S(t)) \\ = \sqrt{n} (S_n(t) - S(t)) \Rightarrow N(0, F(t) S(t)) \end{cases}$

§ Estimateur de Kalplan - Maier des :

dans le paragraphe précedent, les observations $T_1, T_2, \dots T_n$ sont "observées" echantillon complet.

Il esciste des echantillon où les T_n ne sont pas tous "observés" echantillon incomplet ou données incomplète (observations censurées).

Soit $T_1, T_2, T_3^+, T_4, T_5^+, T_6, T_7^+, T_{20}^+$. (*)

où $T_i^+ = T_i \wedge C_i$ où C_i v.a de censure

$= X_i$.

Comment estimer S ?

$T_1, \dots, T_n$ $\quad S_n(t) = \frac{1}{n} \hat{\sum_{i=1}} 1_{(T_i > t)} \to S(t)$ $n \to \infty$

(*) $\to S_n(t) = S_{1n}(t) + S_{2n}(t)$ $\to$ sujets T_i non obs

contient les T_i observés — 4 termes — 16 termes

$S_n^+(t) = \frac{1}{20}(T_1 + T_2 + T_4 + T_6) + \frac{1}{20}(T_3^+ + T_5^+ + T_7^+ \dots$

$\downarrow$ $S_1(t)$ $+$ $S_2(t) \neq S(t)$

la présence de donnée incomplète modifie la limite $S(t)$. (voir exercice).

d'où la difficulté d'estimer $S(t)$ de la présence de donnée incomplète.

Kalplan-Marier ont proposé un estimateur de S.

On regarde d'abord 1 er cas :

1 er cas données complètes :

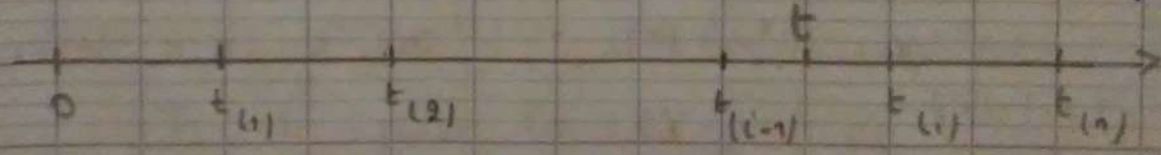
ech : $T_1, T_2, \dots T_n$ de T.

les réalisations : $T_1(\omega), T_2(\omega), \dots T_n(\omega)$.

$t_i = T_i(\omega)$.

$t_{(1)} < t_{(2)} < \dots < t_{(n)}$ instants de réalisation de "E" (événement d'intérêt).

$P(t_i = t_j) = 0$.



pb : estimer $S(t)$: Soit $t_{(i-1)} < t < t_{(i)}$; i fix

$S(t) = P(T > t) = P(T > t / T > t_{(i-1)}) \cdot P(T > t$

or : ① $P(T > t / T > t_{(i-1)})$ → proba que E ne se réalise pas ds $]t_{(i-1)}, t[$.

$P(T > t / T > t_{(i-1)}) = P(T > t \cap T > t_{(i-1)}$

Si je suis "en vie" à l'instant $t_{(i-1)}$ c'est la proba que je reste en vie à l'instant t.

On continue $\mathbb{P}(T > t_{(i-1)})$

$\mathbb{P}(T > t_{(i-1)}) = \mathbb{P}(T > t_{(i-1)} \mid T > t_{(i-2)}) \mathbb{P}(T > t_{(i-2)})$

donc :

$S(t) = \mathbb{P}(T > t \mid T > t_{(i-1)}) \cdot \mathbb{P}(T > t_{(i-1)} \mid T > t_{(i-2)})$

$\mathbb{P}(T > t_{(i-2)}) \dots$

on continue la règle

$$q_j = \mathbb{P}(T > t_{(j)} \mid T > t_{(j-1)}) = q_{t_{(j)}}$$

Ainsi :

$S(t) = \mathbb{P}(T > t \mid T > t_{(i-1)}) \, q_{i-1} \cdot q_{i-2} \dots q_2 \mathbb{P}(T > t_{(1)})$

$t_{(0)} = 0$

$q_1 = \mathbb{P}(T > t_{(1)} \mid T > 0)$.

$\mathbb{P}(T > t_1) = \mathbb{P}(T > t_{(1)} \mid T > t_{(0)} = 0) \cdot \mathbb{P}(T > t_{(0)})$

$= q_1 \cdot \mathbb{P}(T > 0) = q_1$

(1)

Donc

$$S(t) = \mathbb{P}(T > t \mid T > t_{(i-1)}) \cdot q_{i-1} \cdot q_{i-2} \dots q_1$$

on note $q_t := \mathbb{P}(T > t \mid T > t_{(i-1)})$ $\cdot$ $q_j = q_{t_{(j)}}$

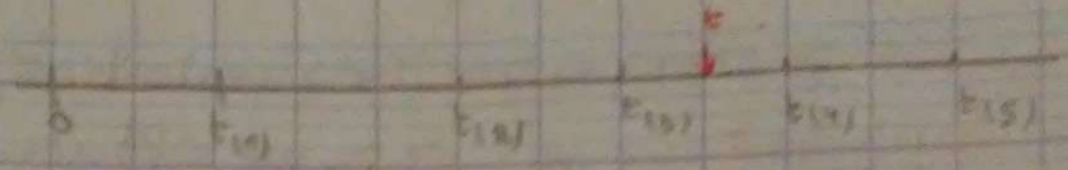
alors : $S(t) = q_t \cdot q_{i-1} \dots q_1$

$= q_t \cdot q_{t_{(i-1)}} \dots q_{t_{(1)}} = \prod_{u \le t} q_u$

u : instant

$= S(t)$

$\prod q_{t_{(j)}} = q_t \prod_{t_{(i-1)} \le t_{(j)} \le t_{(1)}} q_{t_{(j)}}$

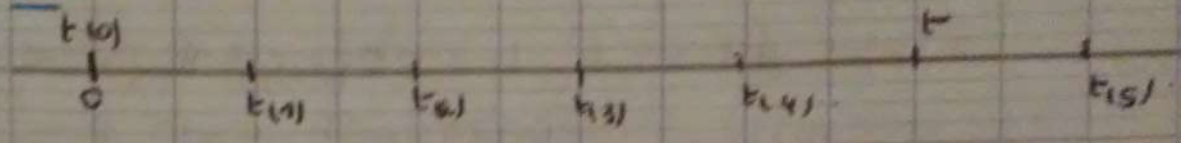


$t_{(3)} < t < t_{(4)}$

$S(t) = q_t \; q_{t_{(3)}} \; q_{t_{(2)}} \; q_{t_{(1)}}$

Rem :

g pas l'évem (E) à l'instant t



$S(t) = q_t \cdot q_4 \cdot q_3 \; q_2 \; q_1$

c'est la survie à l'instant t : c'est ne pas avoir l'evenement E à l'instant t, on a pas eu l'evenement dans $[t, t_{(4)}]$, $[t_{(4)}, t_{(3)}]$ ... $[t_{(1)}, 0]$.

problème : Comment estimer $q_{t_{(j)}}$? pour estimer $S(t)$.

Estimation de $q_{t_{(j)}} = q_j$

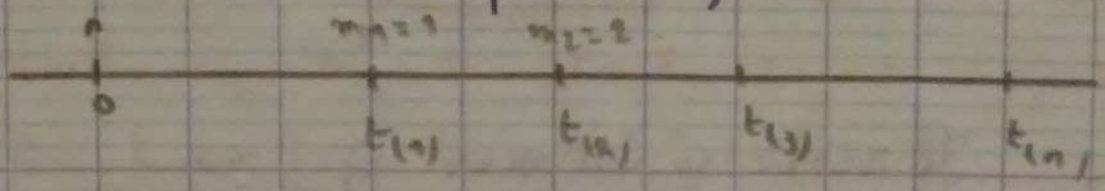
posons : $p_j := 1 - q_j$ = proba d'avoir E dans $]t_{(j-1)}, t_{(j)}]$

On va estimer p_j par les fréquences de l'evenement E $p_j = \frac{\text{cas fav}}{\text{cas poss}}$.

Soit : M_j = nombre "d'individus" envie à ensuite l'instant $t_{(j-1)}$

m_j : nombre d'evenement E à l'instant $t_{(j)}$ $(m_j \geq 1)$

Rem : n machines en fiabilité (on observe la panne E)



$t_{(0)} = 0$ $\quad P(x / x_i = x_j) = 0$

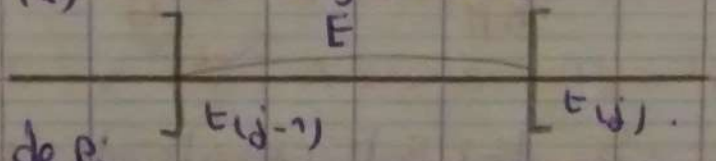
$n_1 \rightsquigarrow t_{(0)} = 0$ $\quad P(n_i) = 0$ n'est pas vide

$n_1 = n$

$n_2 \rightsquigarrow t_{(1)}$ $\quad n_2 = n - 1$

$n_3 \rightsquigarrow t_{(2)}$ $\quad n_3 = n - 3$.

Donc :



un estimateur de p_j

$$\hat{p}_j = \frac{\text{cas favorable dans }]t_{(j-1)}, t_{(j)}]}{\text{cas possible}} = \frac{m_j}{n_j}$$

nbre des individus en vie à t_{j-1}.

donc : $\boxed{\hat{q}_j = 1 - p_j = 1 - \frac{m_j}{n_j}}$ $\quad j = 1, \ldots, n.$

Ainsi l'estimateur de K. M.

on a : $S(t_{(i)}) = \prod_{t_{(j)} \leq t_{(i)}} q_j = q_i \, q_{i-1} \cdots q_1$

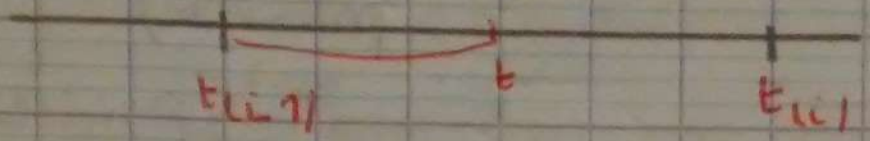
et un estimateur de $S(t)$ est :

$$\boxed{S_{KM}(t_{(i)}) = \prod_{i=1,\ldots,n \; t_{(j)} \leq t_{(i)}} \hat{q}_j = \prod_{1 \leq j \leq i} \left(1 - \frac{m_j}{n_j}\right)}$$

appelé estimateur de Kaplan Meier de $S(t_{(i)})$

$S_{Km}(t_{(i)})$ estimateur de $S(t_{(i)})$.

pour $S(t)$ où $t_{(i-1)} < t < t_{(i)}$.



Il reste à estimer q_t ⇝

$q_t = \mathbb{P}(T > t \mid T > t_{(i-1)})$

$p_t = 1 - q_t$ = proba d'avoir E ds $]t_{(i-1)}, t]$

$q_t = 1$.

$$S_{KM}(t) = \prod_{t_{(j)} \le t} \hat{q}_j = \prod_{t_{(j)} \le t} \left(1 - \frac{m_j}{n_j}\right).$$

$S_{KM}(t)$ est constant sur chaque intervalle $]t_{(i-1)}, t_{(i)}[$ $i = 1, \dots, n$ et décroissante

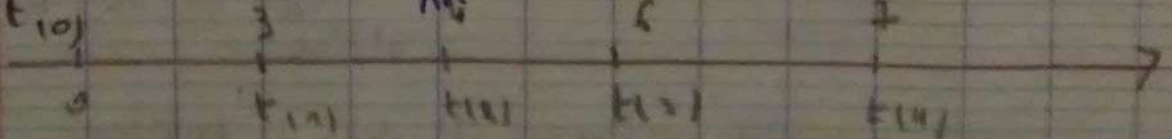
q_{10} q_9 $\dots$ q_1

q_{11} q_{10} $\dots$ q_1.

Ex: on observe 5 machines et leurs durées de fonctionnement jusqu'à la panne

3, 4.5, 6, 7 (mois)

Calculer $S_{KM}(t)$ $t > 0$



$$S_{KM}(t_{(i)}) = \prod_{j \le i} \left(1 - \frac{m_j}{n_j}\right)$$

$i = 1, \dots 4$

Faux

m_j : nbr d'even. "E" à l'instant $t_{(j)}$

n_j : nbr d'individus en vie à l'instant $t_{(j-1)}$

$m_1 = 1$, $m_2 = 1$, $m_3 = 2$, m_4 , $m_5 = 1$

$n_1 = 5$, $n_2 = 4$, $n_3 = 3$ $n_4 = n_5 = 1$, n_6 ?

Si $t \in]0,3[$: $S_{KM}(t) = 1$

$S_{KM}(3) = \left(1 - \frac{m_1}{n_1}\right) = 1 - \frac{1}{5} = \frac{4}{5} = q_1$

Si $t \in]3,4[$

$S_{KM}(t) = \frac{4}{5} = q_1$

$S_{KM}(4) = q_2 q_1 = \left(1 - \frac{m_2}{n_2}\right)\left(\frac{4}{5}\right) = \frac{3}{4} \cdot \frac{4}{5} = \frac{3}{5}$

$n_0 = 5$; $n_1 = 5$, $n_2 = 4$, $n_3 = 3$, $n_4 = 2$, $n_5 = 1$

$m_0 = 0$, $m_1 = 1$, $m_2 = 1$, $m_3 = 1$, $m_4 = 1$, $m_5 = 1$

$S_{KM}(t_{(0)}) = \left(1 - \frac{m_0}{n_0}\right) = 1 = S(0)$

$S_{KM}(t_{(1)}) = \left(1 - \frac{m_0}{n_0}\right)\left(1 - \frac{m_1}{n_1}\right) = \frac{4}{5} = S(3)$

$K_{KM}(t_{(2)}) = \frac{4}{5}\left(1 - \frac{m_2}{n_2}\right) = \frac{3}{5} = S(4)$

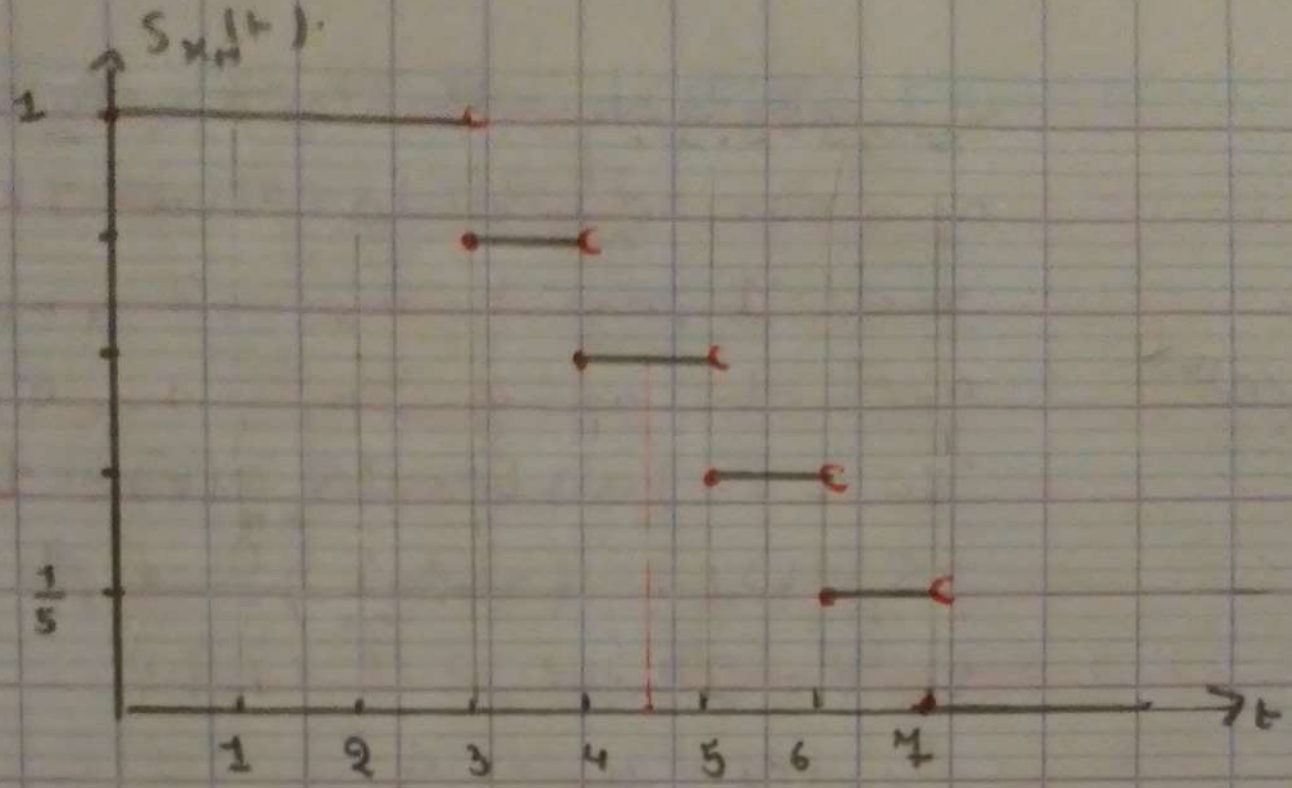
$S_{KM}(t_{(3)}) = \frac{3}{5}\left(1 - \frac{m_3}{n_3}\right) = \frac{2}{5} = S(5)$

$S_{KM}(t_{(4)}) = \frac{1}{5} = S(6)$

$S_{KM}(t_{(5)}) = \frac{1}{5}\left(1 - \frac{m_5}{n_5}\right) = \frac{1}{5}(1-1) = 0 = S(7)$

graphe : $S(4,5) = \frac{3}{5} = 0,6 = 60\%$

$t \in [0,3[$ $S(0) = 1$



$S(4,5) = \frac{3}{5} = 60\% = P(T > 4,5)$

$S(t) = P(T > t)$. , $S(5,5) = \frac{2}{5} = 40\%$.

Exemple 2) $\underbrace{3}_{t_{(1)}}, \underbrace{4}_{t_{(2)}}, \underbrace{6, 6}_{t_{(3)}}, \underbrace{7}_{t_{(4)}}$.

$n_0 = 5$, $n_1 = 5$, $n_2 = 4$, $n_3 = 3$, $n_4 = 1$

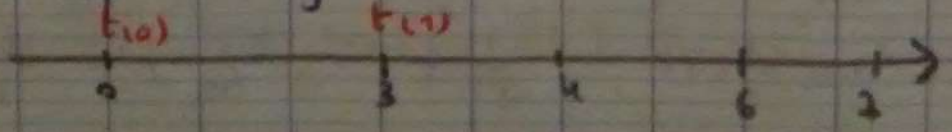
$m_0 = 0$, $m_1 = 1$, $m_2 = 1$, $m_3 = 2$, $m_4 = 1$

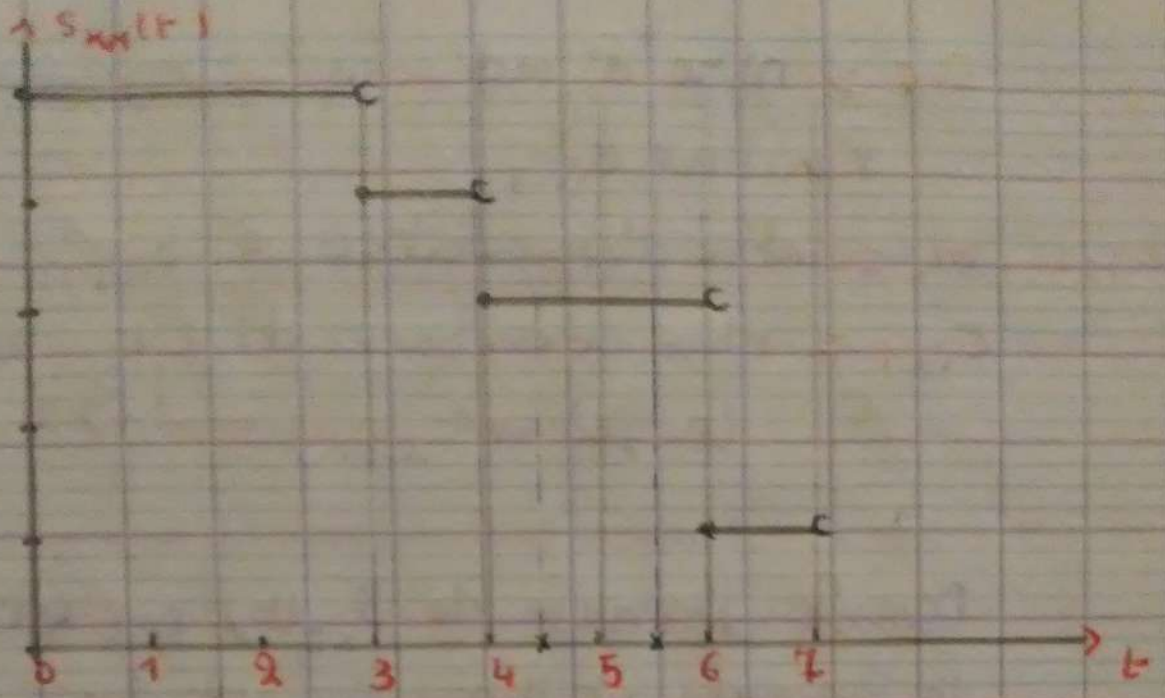
$S_{KM}(t_{(0)}) = \frac{4}{5} = S(3)$

$S_{KM}(t_{(2)}) = \frac{4}{5}\left(1 - \frac{1}{4}\right) = \frac{3}{5} = S(4)$

$S_{KM}(t_{(3)}) = \frac{3}{5}\left(1 - \frac{2}{3}\right) = \frac{1}{5} = S(6)$

$S_{KM}(t_{(4)}) = \frac{1}{5}(1 - 1) = 0 = S(7)$.





$S(4,5) = \frac{3}{5} = 60\%$

$S(5,5) = \frac{3}{5} = 60\%$

La survie est grande par rapport au 1er exemple car on a pas un ev^{t} en 5 par contre en 1er exple il y avait un evenement

Ex 1: $S_{KM}(8) = 0$, à 8: plus "personnes" en survie.

2ème cas: données incomplètes (avec des censures)

On suppose qu'on a un echantillon d'observation avec existence de donnée censure

on range par ordre croissant les instants $t_{(i)}$

($t_{(i)}$ = instant pour l'ev^{t} E ou une censure).

les données censurés sont signalés par $(+)$, 7^{+}

ex: (3, 5, 6, 6, 7^{+}, 8, 10, ...)

à 7 on a observé une censure.

on a observé E à ces instants.

n_i = nbre d'individus à risque (envie) à l'instant $t_{(i-1)}$.

m_i = nbre d'evenements E à $t_{(i)}$.

c_{i-1} = nbre de censures ds $[t_{(i-1)}, t_{(i)}[$. $i \geq 1$
(on fait-on $t_{(i-1)}$).

Dans la formule de $S_{KM}(t)$ qui estime $S(t) = P(T > t)$ la formule utilisée pour calculer $S(t)$ dans le 1er cas reste la même.

on arrive à $S_{KM}(t) = \prod_{t_{(j)} \leq t} \left(1 - \frac{m_j}{n_j}\right) = \prod_{t_{(j)} \leq t} \hat{q}_j$

la différence avec le 1er cas c'est l'estimation $\hat{q}_j$.

$S(t) = q_t \quad \prod q_j = \prod_{t_{(j)} \leq t} q_j$

$q_j = 1 - p_j \leftrightarrow \hat{q}_j = 1 - \hat{p}_j$

p_j = proba E ds $]t_{(j-1)}, t_{(j)}]$

$\rightarrow \hat{p}_j = \frac{m_j}{n_j}$

on a :

$t_{(0)} = 0$; $m_0 = 0$

c_0 = nbre de censure $]t_{(0)}, t_{(1)}[$ (convention)

1er cas :

on a : $\boxed{n_i = n_{i-1} - m_{i-1}}$.

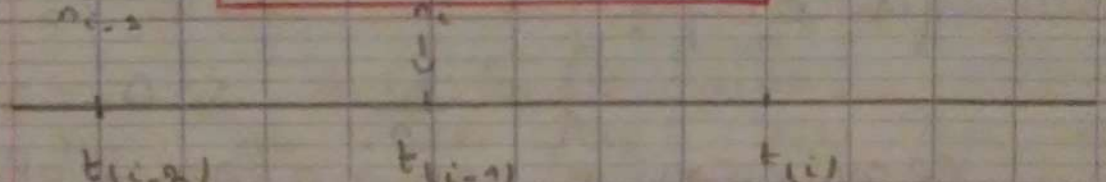
$n_2 = n_1 - m_1$

2ème cas

on a : $t_{(0)} = 0$, $m_0 = 0$

c_0 = nbre de censure $]\underbrace{t_{(0)}}_{0}, t_{(1)}[$

on a : $\boxed{n_i = n_{i-1} - m_{i-1} - c_{i-1}}$ $\quad c_0 = 0$



$n_i = n_{i-1} - m_{i-1} - c_{i-1}$

$n_1 = n_0 - \underbrace{m_0}_{0} - \underbrace{c_0}_{0}$ $\quad n_1 = n$

$n_2 = n_1 - m_1 - c_1$

On a :

$$\boxed{n_i = n - \sum_{j=1}^{i-1} m_j - \sum_{j=1}^{i-1} c_j}$$

(car : $n_i = n_{i-1} - m_{i-1} - c_{i-1}$

$= (n_{i-2} - m_{i-2} - c_{i-2}) - m_{i-1} - c_{i-1}$

$= n_{i-2} - (m_{i-1} + m_{i-2}) - (c_{i-1} + c_{i-2})$

$= \vdots$

$= n_0$

$= n$

L'estimateur de KM s'écrit :

$$\boxed{S_{KM}(t) = \prod_{t_{(i)} \leq t} \left(1 - \frac{m_i}{n_i}\right)^{D_i}} \quad (**)$$

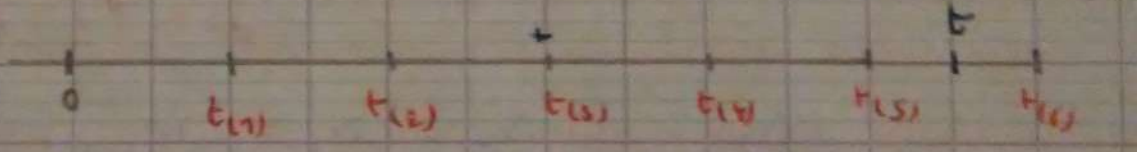
où $D_i = \begin{cases} 1 & \text{si l'evt E est observé en } t_{(i)} \\ 0 & \text{si il y a une censure en } t_{(i)} \end{cases}$

Rq : dans la formule de S_{KM} le facteur

$\left(1 - \frac{m_i}{n_i}\right)^{D_i} = \begin{cases} 1 & \text{si } D_i = 0 \\ 1 - \frac{m_i}{n_i} & \text{si } D_i = 1 \end{cases}$

ce qui signifie qu'il y a pas d'effet dans le calcul de produit en un point où il y a une ce mais il y a un changement dans le calcul de n_i.

$$S_{KM}(t) = \left(1 - \frac{m_1}{n_1}\right)^{D_1} \left(1 - \frac{m_2}{n_2}\right)^{D_2} \left(1 - \frac{m_3}{n_3}\right)^{D_3} \left(1 - \frac{m_5}{n_5}\right)$$



$$S_{KM}(t) = \left(1 - \frac{m_1}{n_1}\right)\left(1 - \frac{m_2}{n_2}\right) \cdot 1 \cdot \left(1 - \frac{m_4}{n_4}\right)\left(1 - \frac{m_5}{n_5}\right)$$

pour la formule (**) si il y'a pas de censure (on observe E de tous les $t_{(i)}$) donc on retrouve la formule du 1er cas.

si il y'a des censures : le facteur q_j de S_{KM} est égale à 1 (car $p_j = 1 - q_j = 0$)

p_j = proba d'avoir E de $]t_{(j-1)}, t_{(j)}[$

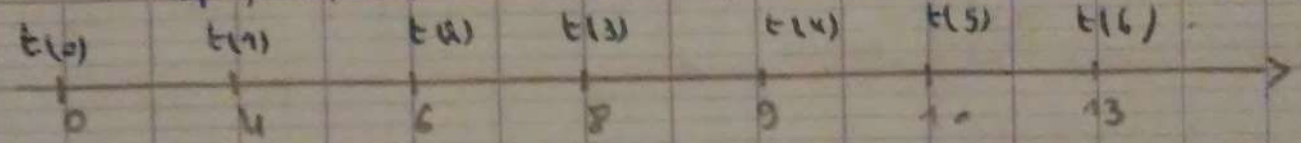
cette proba, elle vaut 0.

$p_3 \to]t_{(2)}, t_{(3)}]$ $\quad p_3 = 0 \quad q_3 = 1$.

exp: on observe le fonctionnement d'un nombre n de machine jusqu'à la panne.

$n = 7$ et on a les données suivantes.

4, 6, 6, 8, 9, 10, 13.



$n_0 = 7$	$m_0 = 0$	$c_0 = 0$
$n_1 = 7$	$m_1 = 1$	$c_1 = 0$
$n_2 = 6$	$m_2 = 2$	$c_2 = 0$
$n_3 = 4$	$m_3 = 0$	$c_3 = 1$
$n_4 = 3$	$m_4 = 1$	$c_4 = 0$
$n_5 = 2$	$m_5 = 1$	$c_5 = 0$
$n_6 = 1$	$m_6 = 1$	$c_6 = 0$

$S_{KM}(t) = 1 \quad , t \in [0, 4[$

$S_{KM}(t) = 1 \cdot \left(1 - \frac{m_1}{n_1}\right)^1 = 1 - \frac{1}{7} = \frac{6}{7} \quad t \in [4, 6[$

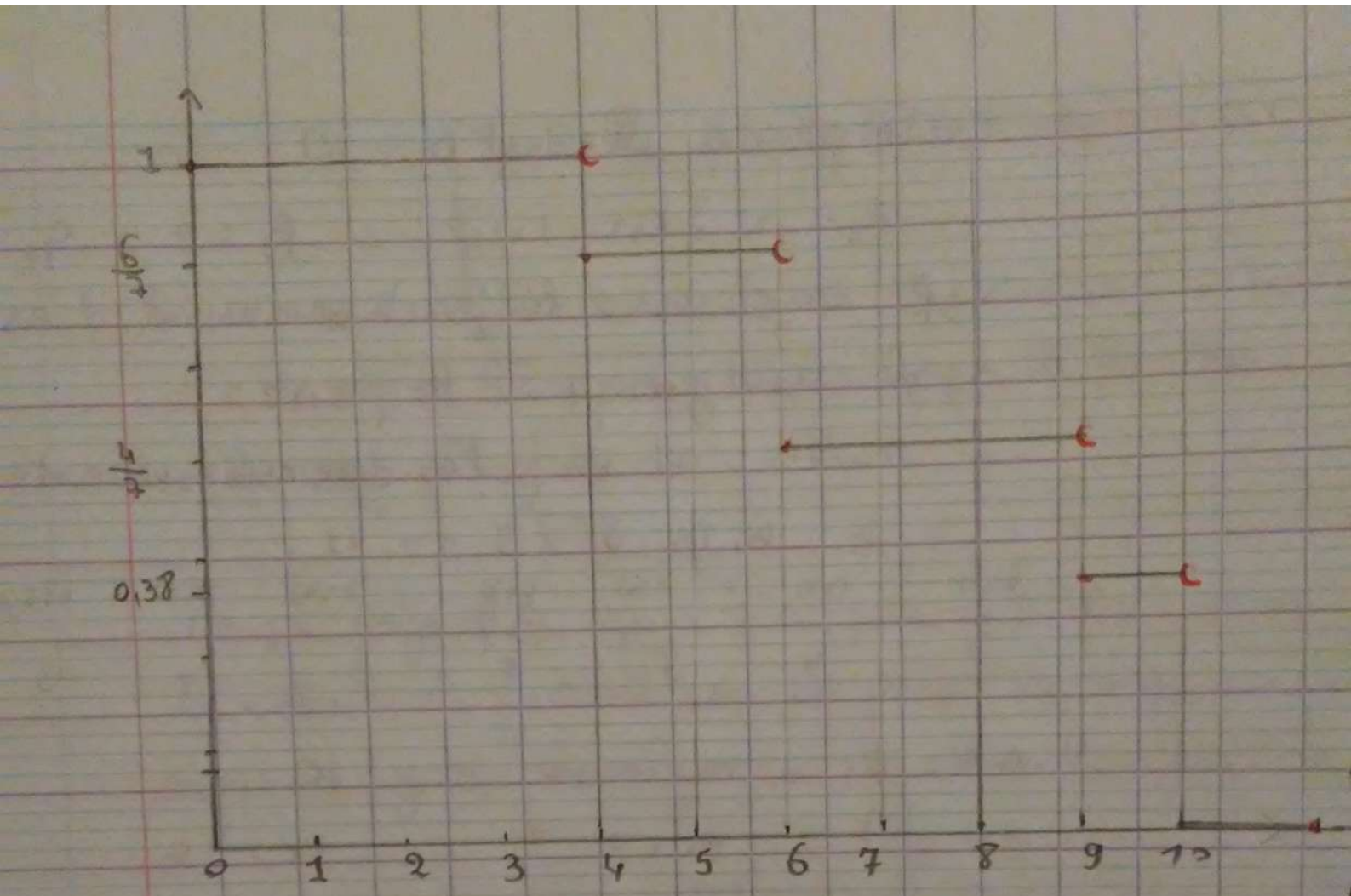
$S_{KM}(t) = \frac{6}{7}\left(1 - \frac{m_2}{n_2}\right)^1 = \frac{4}{7} \quad t \in [6, 8[.$

$S_{KM}(t) = \frac{4}{7}\left(1 - \frac{m_3}{n_3}\right)^0 = \frac{4}{7} \quad t \in [8, 9[.$

$S_{KM}(t) = \frac{4}{7}\left(1 - \frac{m_4}{n_4}\right)^1 = \frac{8}{21} \quad t \in [9, 10[.$

$S_{KM}(t) = \frac{8}{21}\left(1 - \frac{m_5}{n_5}\right) = \frac{4}{21} \quad t \in [10, 13[.$

$S_{KM}(t) = \frac{4}{21}\left(1 - \frac{m_6}{n_6}\right) = \frac{4}{21} \cdot 0 = 0 \quad t \geq 13$



$$S_n(t) = \frac{1}{n} \sum \mathbb{1}_{(T_i > t)} = \underbrace{S_{n,1}(t)}_{\text{obs}} + \underbrace{S_{n,2}(t)}_{\text{censure}} \xrightarrow{?} S(t)$$