ANALYSIS OF CENSORED DATABASE: THE WARRANTY CASE OF STUDY

- Censorship: Definition, causes and consquences
- Non parametric estimation for censored data
- Parametric estimation (graphical process)
- Maximum Likelihood for censored data
- Effect of the censorship on the goodness of fit

La censure: définition, causes et conséquences

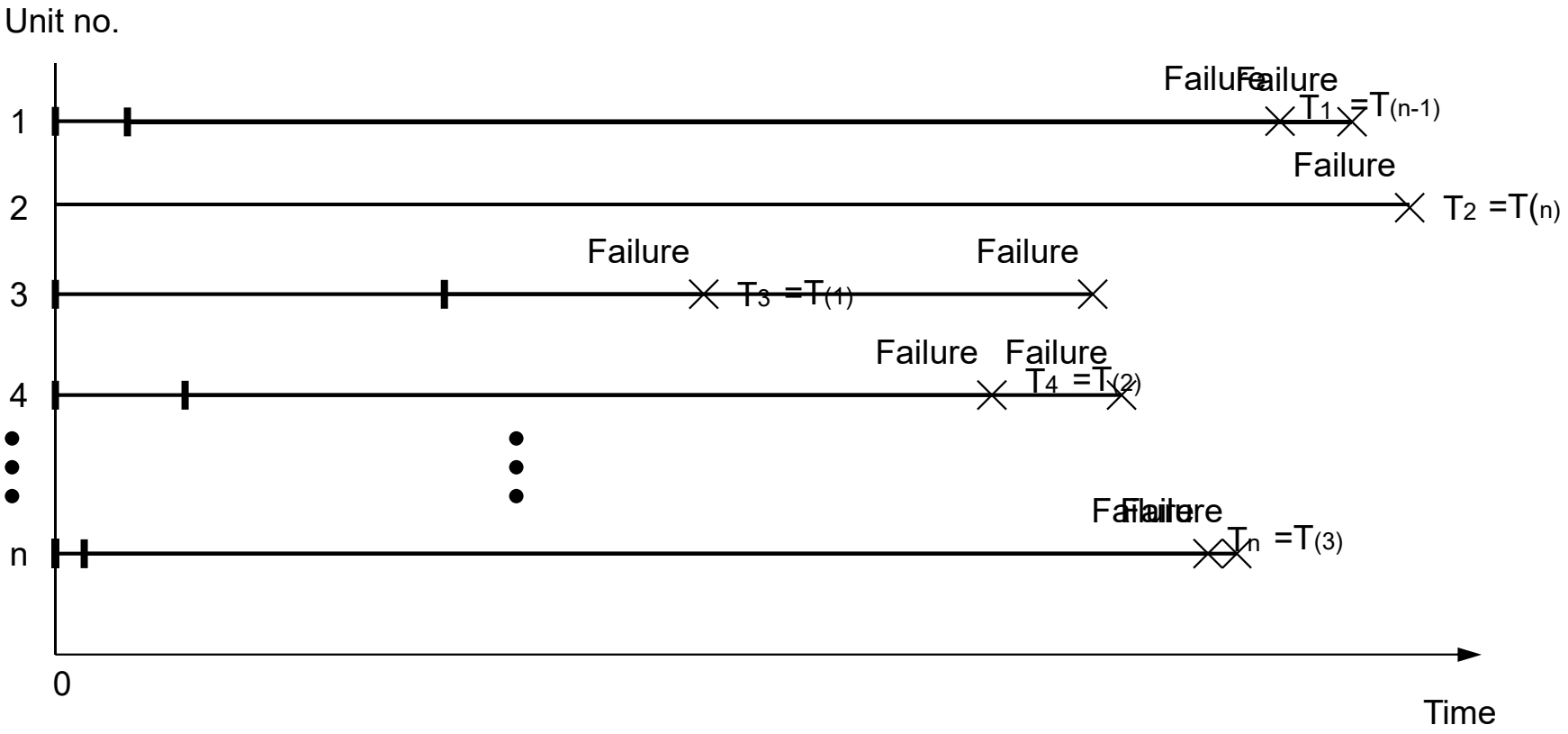
Table 1 Exemple de base de données de suivi exploitation pour 5 produits identiques jusqu'à aujourd'hui

Référence produit	Date de première mise en service	Date de retrait de service	Date de remise en service	Date de retrait de service (définitif)
1	23/05/2009	12/09/2011	28/09/2011	14/07/2015
2	12/06/2009	14/03/2010		
3	14/06/2009		01/02/2013	08/08/2014
4	14/06/2009	25/11/2012		26/05/2015
5		03/03/2011	09/05/2011	25/04/2011

A partir de ce tableau, définissez la nature de chaque donnée et identifiez le type de censure existant.

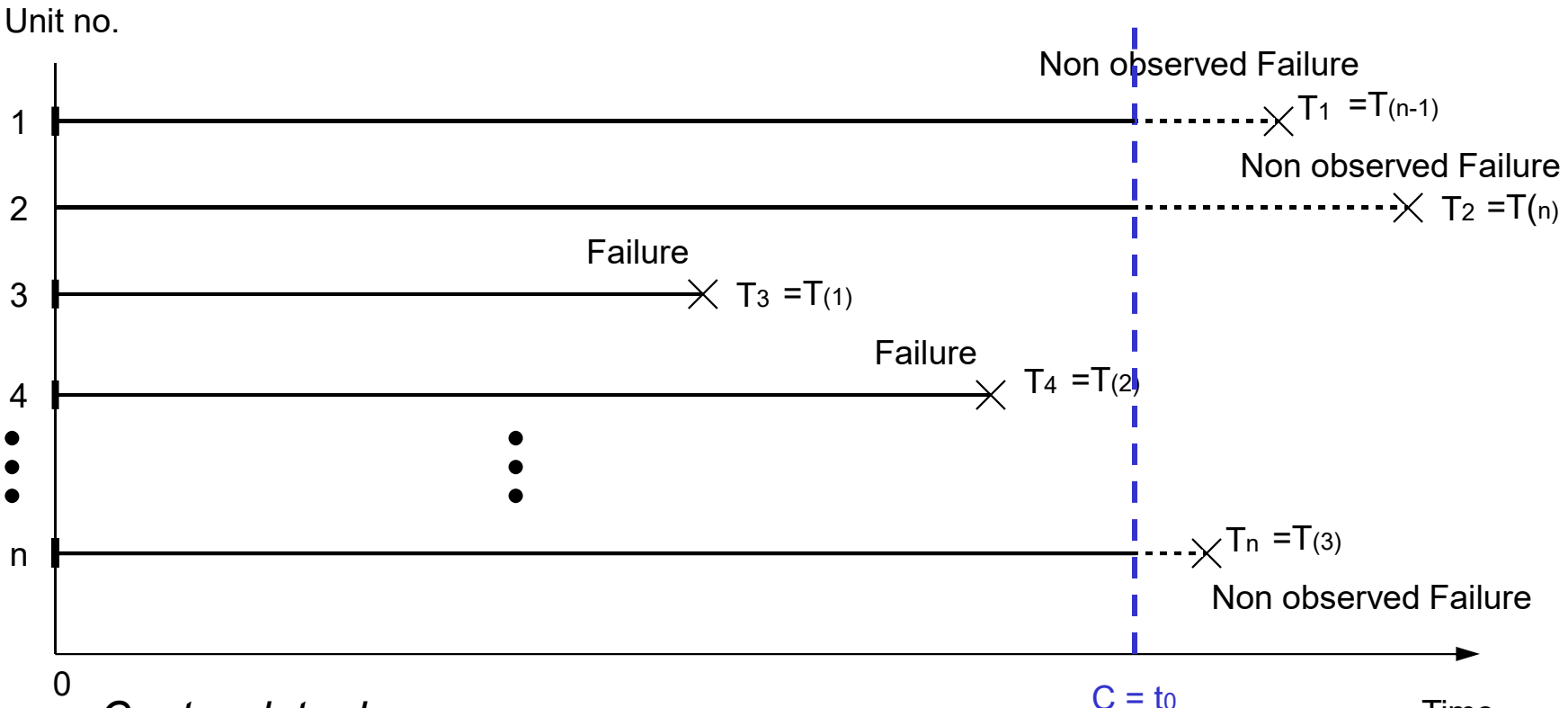
Les différents types de censure

1. L'échantillon complet



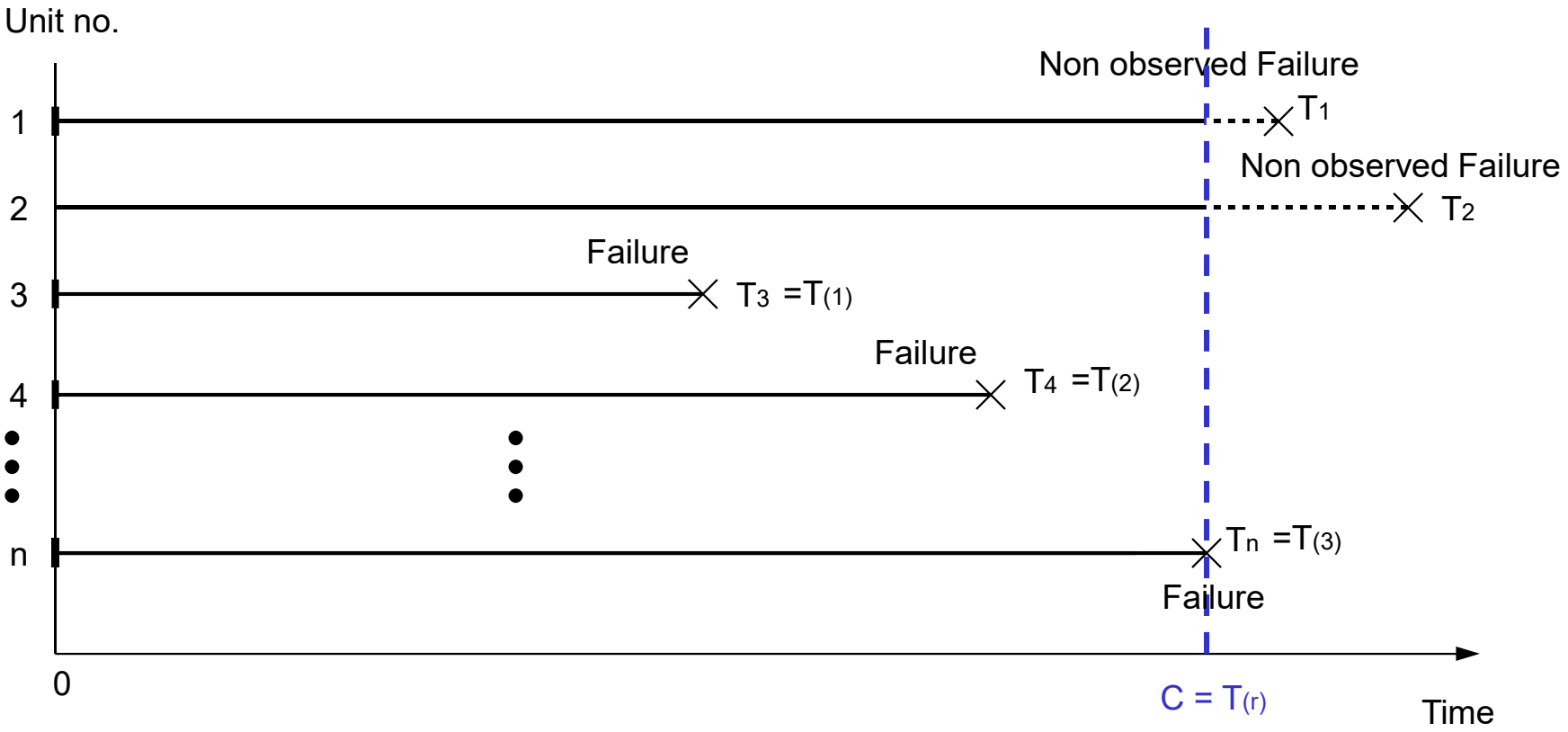
$T_{(i)}$ = i-ème ***statistique ordonnée*** *de l'échantillon - $T_{(i-1)} \leq T_{(i)}$*

Les différents types de censure



- *C = t0 : date de censure connue*
- *Censure de* **Type-1**
- *T(1) ≤ T(2) ≤ … ≤ T(S) < t0* = durées de vie observées mais une « certaine » information est contenue dans les *(n-S)* systèmes encore en service
- *C* peut être aléatoire

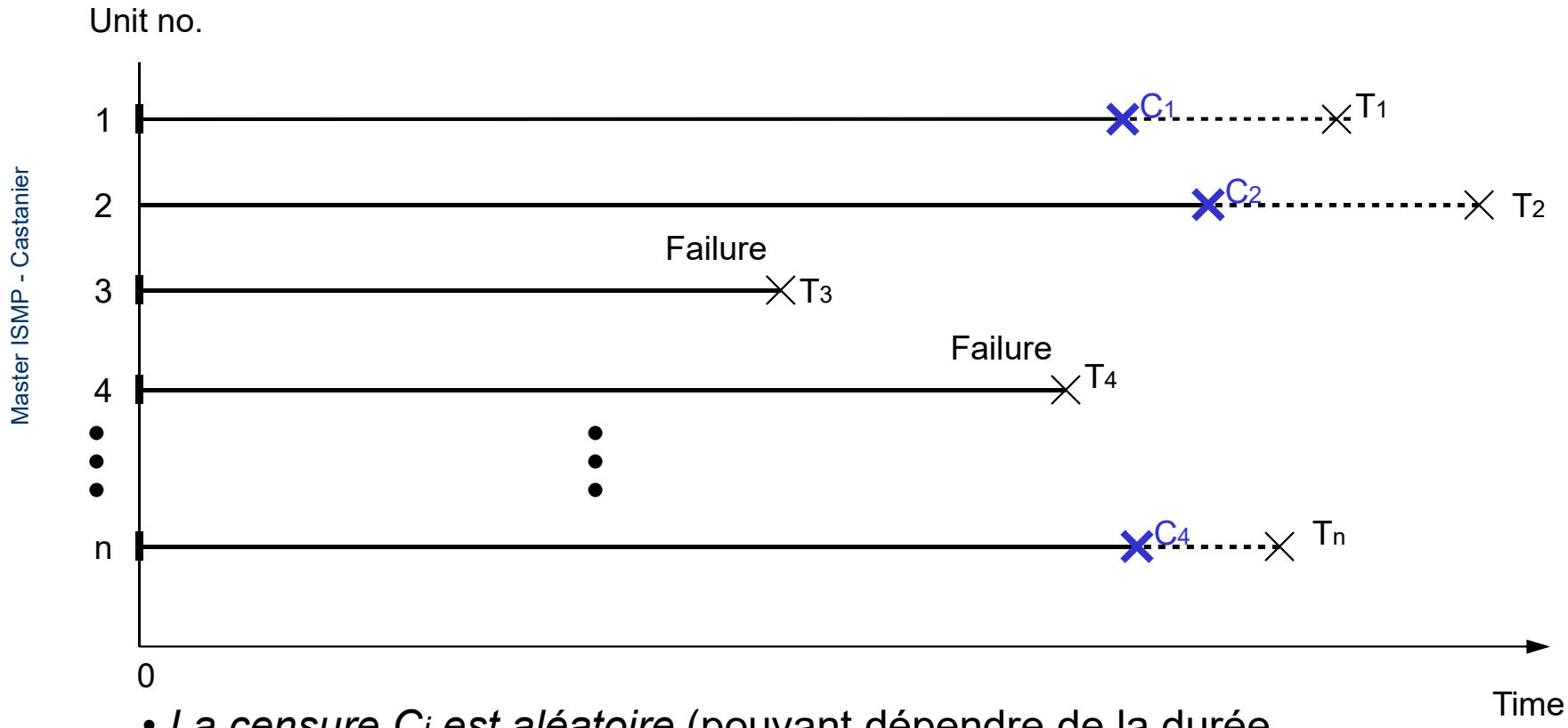
Les différents types de censure



- $C = T_{(r)}$: r-ème durée de vie où r est fixé avant l'expérimentation
- *Censure de* **Type-2**
- $T_{(r)}$ est aléatoire et peut être très grand

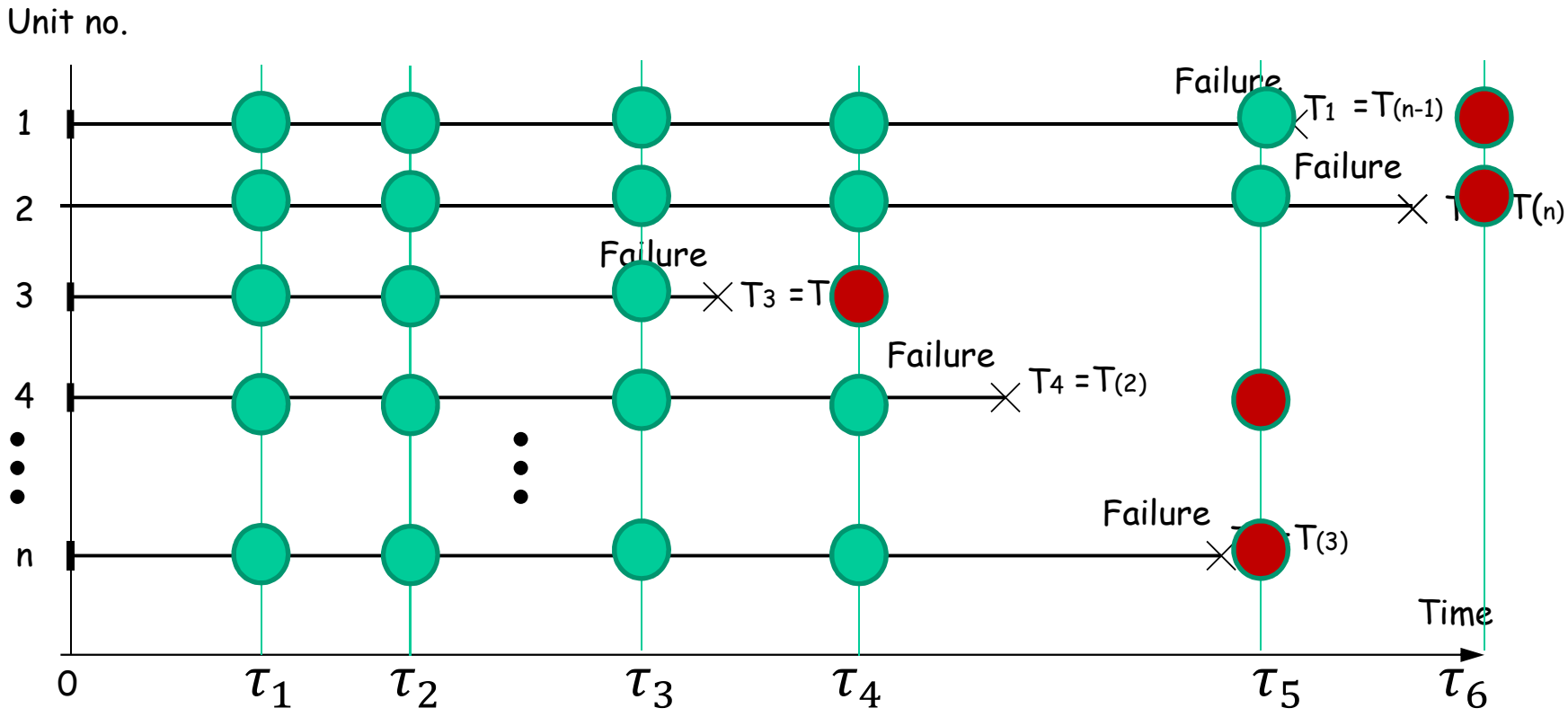
Les différents types de censure

- *C = min(T(r) , t0) est une variable aléatoire*
- *Censure de* **Type-3**



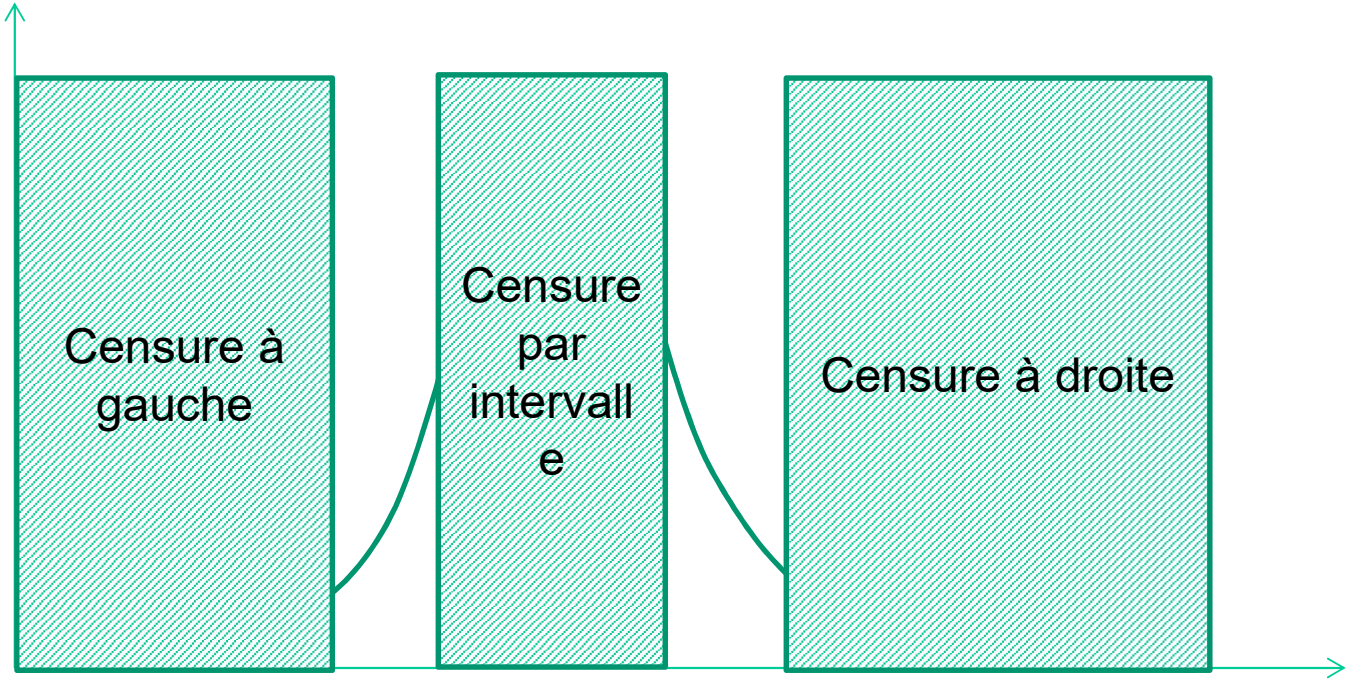
- *La censure Ci est aléatoire* (pouvant dépendre de la durée d'activation)
- Censure de **Type-4**

La censure par intervalle



Données = Dates d'inspection et Nombre de défaillances par intervalle

Effet d'une donnée censurée sur la densité estimée



Comment intégrer ces informations dans la procédure d'estimation ?

Démarche :

- Caractérisation de l'échantillon
  - Indépendance, homogénéité, ordonné
  - Définition du **type de censure** pour **chaque** donnée
- Recherche de la fonction empirique
  - Estimation non paramétrique de la loi de durée de vie
  - Test d'homogénéité de l'échantillon
- Choix d'un modèle paramétrique
- Estimation des paramètres du modèle
- Construction des Intervalles de Confiance

Estimation non paramétrique pour données censurées

- Etude de données censurées par intervalle
- Estimation pour des censures multiples avec des données exactes :
  - L'estimateur de Kaplan-Meïer (ou Product-Limit)
  - L'estimateur de Wayne-Nelson (ou Nelson-Aalen)
  - L'estimateur des rangs corrigés de Johnson (ou rangs médians généralisés)
  - L'estimateur de Kaplan-Meïer généralisé
  - La méthode actuarielle

Voici des données de SAV de tubes d'échangeur de chaleur fissurés

100 tubes	Année 1	Année 2	Année 3	Tubes non fissurés
Observations	1	2	2	95
Probabilité de défaillance	π_1	π_2	π_3	π_4

Que représente $\pi_i, i = 1, 2, 3$ et π_4 ?

Ecrire π_i en fonction d'une distribution de durée de vie $F(t)$?

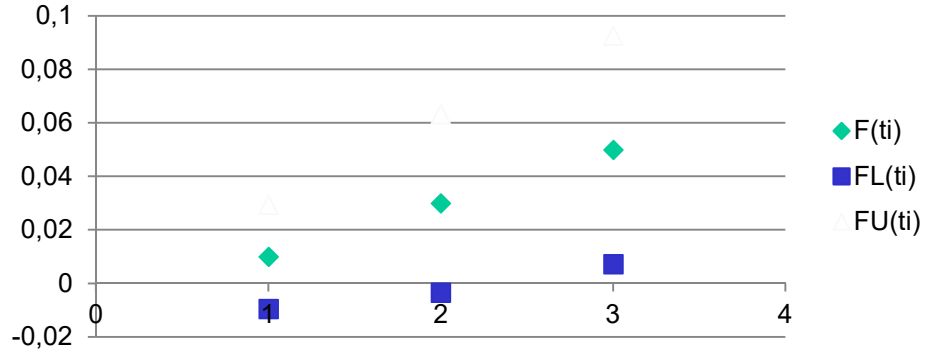
Donnez un estimateur de $\hat{F}(t_i)$

Intervalle de confiance sur $\hat{F}(t_i)$

$$IC_{1-\alpha} = [F_L(t_i), F_U(t_i)] = \left[\hat{F}(t_i) \pm z_{1-\frac{\alpha}{2}} \widehat{se}_{\hat{F}}(t_i)\right]$$

Avec $\widehat{se}_{\hat{F}}(t_i) = \sqrt{\frac{1}{n}\hat{F}(t_i)\left(1-\hat{F}(t_i)\right)}$

Année	ti	di	F(ti)	seF	FL(ti)	FU(ti)
(0-1]	1	1	0,01	0,00994987	-0,0095014	0,0295014
(1-2]	2	2	0,03	0,01705872	-0,0034345	0,06343448
(2-3]	3	2	0,05	0,02179449	0,00728358	0,09271642



Cas de censures multiples

	N	Année 1	Année 2	Année 3	Reste
Usine 1	100	1	2	2	
Usine 2	100	2	3		
Usine 3	100	1			
Toutes les usines	300				
Probabilité de défaillance conditionnelle	$\hat{p}_i$				

Soit $n_i = n - \sum_{j=1}^{i-1}(d_j + r_j)$ le nombre d'individus entrants dans $[t_{i-1}, t_i]$,
r_j nombre d'individus ayant survécu et censurés à droite

Que représente $\frac{d_i}{n_i}, i = 1, \dots, m$?

Déterminez l'expression de $\hat{F}(t_i)$

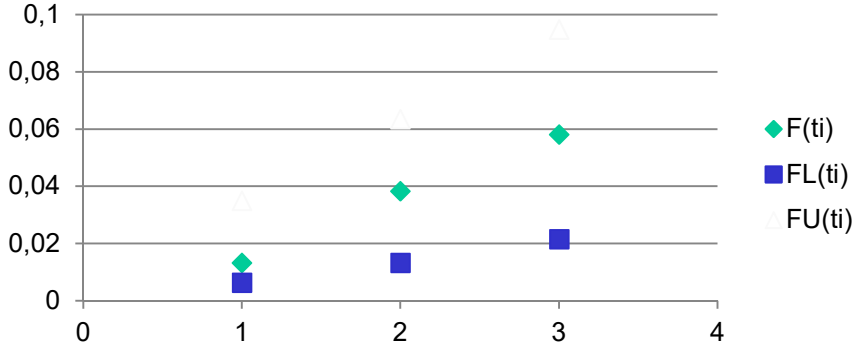
Etude de données censurées par intervalle

Intervalle de confiance sur $\hat{F}(t_i)$

$$IC_{1-\alpha} = [F_L(t_i), F_U(t_i)] = \left[\hat{F}(t_i) \pm z_{1-\frac{\alpha}{2}}\, \widehat{se}_{\hat{F}}(t_i)\right]$$

Avec $\widehat{se}_{\hat{F}}(t_i) = \hat{S}(t_i)\sqrt{\sum_{j=1}^{i} \frac{\hat{p}_j}{n_j(1-\hat{p}_j)}}$

		Défaillant	Censuré	Entrée				
Année	ti	di	ri	ni	F(ti)	seF	FL(ti)	FU(ti)
(0-1]	1	4	99	300	0,0133	0,0062	0,0064	0,1350
(1-2]	2	5	95	197	0,0384	0,0128	0,0133	0,0635
(2-3]	3	2	95	97	0,0582	0,0187	0,0216	0,0949



Estimateur de Kaplan-Meïer (ou Product-Limit)

$$\hat{S}(t) = \prod_{t_i<t} \frac{n_i - d_i}{n_i}$$

- n_i = nb de survivants juste avant t_i si non censuré
- n_i = nb de survivants – nb de censurés si censure

Intervalle de confiance

$$IC_{1-\alpha} = [F_L(t_i), F_U(t_i)] = \left[\hat{F}(t_i) \pm z_{1-\frac{\alpha}{2}} \widehat{se}_{\hat{F}}(t_i)\right]$$

Avec $\widehat{se}_{\hat{F}}(t_i) = \hat{S}(t_i)\sqrt{\sum_{t_i<t} \frac{d_i}{n_i(n_i-d_i)}}$

Exercice

Déterminer l'estimateur de Kaplan-Meïer pour les données observées suivantes (durées de vie d'amortisseurs).

kms	FM	kms	FM	kms	FM
6700	M1	9820	-	12870	M1
6950	-	11310	-	13150	M2
7820	-	11690	-	13330	-
8790	-	11850	-	13470	-
9120	M2	12140	M1	14040	-
9660	-	12200	-	14300	M1

- Dériver l'$IC_{95\%}$dans le cas des données complètes (seuls M1 et M2)
- Refaire l'étude avec l'ensemble de l'échantillon
- Comparer les IC

Estimateur de Wayne-Nelson (ou Nelson Aalen)

Condition d'emploi : Uniquement matériels non réparables

Principe : Taux de hasard cumulé $\Lambda(t) = \int_0^t \lambda(u)du\,, \forall t > 0$

Procédure : Echantillon = N_0 systèmes indépendants, k événements défaillances (observées), $(N_0 - k)$ censures à droite (suspension)

i. Classer les données et censures (à droite). On a alors :

ii. Détermination d'un Δt suffisamment petit pour contenir au plus une donnée

iii. Estimation des $\lambda(t_j), j = 1, \dots, m$

i. $$\hat{\lambda}(t_i) = \begin{cases} \frac{1}{n_i} & si\ t_i = défaillance \\ 0 & sinon \end{cases}$$

iv. Calculer $\widehat{\Lambda}(t) = \sum_{t_i<t} \hat{\lambda}(t_i)$

v. En déduire $\hat{R}_{W.N.}(t) = e^{-\sum_{t_i<t} \widehat{\lambda}(t_i)}$

Estimateur de Wayne-Nelson (ou Nelson Aalen)

Intervalle de confiance

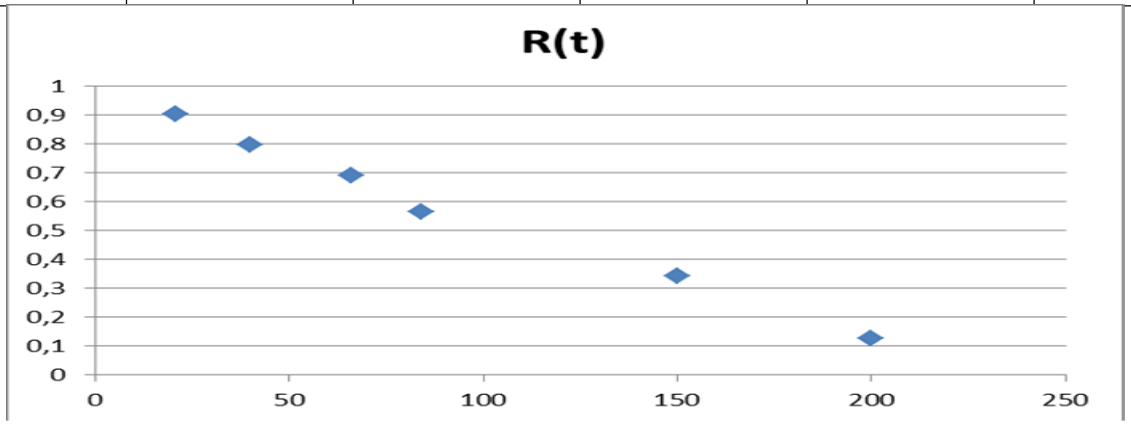
$$IC_{1-\alpha} = [F_L(t_i), F_U(t_i)] = \left[\hat{F}(t_i) \pm z_{1-\frac{\alpha}{2}} \widehat{se}_{\hat{F}}(t_i)\right]$$

Avec $\widehat{se}_{\hat{F}}(t_i) = \sqrt{\sum_{t_i<t} \frac{d_i}{n_i^2}}$

Propriété : L'estimateur de N.A. est conservatif par rapport à K.M

Exemple

Rang	ti	Nature	ni	1/ni	Lambda(t)	R(t)
1	21	D	10	0,1	0,1	0,90483742
2	33	S	9			
3	40	D	8	0,125	0,225	0,79851622
4	66	D	7	0,14285714	0,36785714	0,69221606
5	70	S	6			
6	84	D	5	0,2	0,56785714	0,56673858
7	100	S	4			
8	110	S	3			
9	150	D	2	0,5	1,06785714	0,34374432
10	200	D	1	1	2,06785714	0,12645647



Méthode des rangs corrigés de Johnson (ou des rangs médians généralisés)

$$F(t_i) = \frac{\theta_i - 0{,}3}{N_0 + 0{,}4}$$

Avec

- N_0: taille de l'échantillon
- $\theta_i = \begin{cases} i & si\ l'échantillon\ est\ complet \\ \theta_{i-1} + I_i & Si\ censuré \end{cases}$
- $I_i = \frac{N_0 + 1 - \theta_{i-1}}{1 + n_i}$, I_i est appelé l'incrément de la défaillance *i*

Méthode des rangs corrigés de Johnson (ou des rangs médians généralisés)

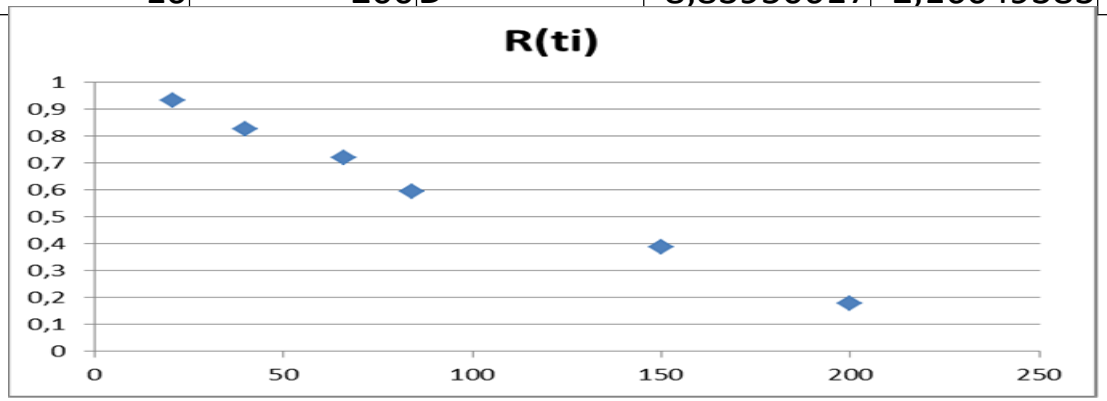
Initialisation : $i = 1\,; \theta_{i-1} = \theta_0 = 0\ et\ \theta_1 = I_1$

Analyse de cas particuliers :

- Pas de censure avant la t_k = *k*ème défaillance : $\theta_i = i, i = 1, \ldots k$
- Que des censures avant t_{k+1}, que vaut $\theta_i, i = 1, \ldots k$?
- Ecrire les relations entre I_{i-k}et I_i sachant aucune censure entre t_{i-k} et $\mathrm{t_i}$

Exemple

Rang	ti	Nature	theta_i	l_i	F(ti)	R(ti)
1	21	D	1	1	0,06730769	0,93269231
2	33	S	1			
3	40	D	2,11111111	1,11111111	0,1741453	0,8258547
4	66	D	3,22222222	1,11111111	0,28098291	0,71901709
5	70	S	3,22222222			
6	84	D	4,51851852	1,2962963	0,40562678	0,59437322
7	100	S	4,51851852			
8	110	S	4,51851852			
9	150	D	6,67901235	2,16049383	0,61336657	0,38663343
10	200	D	8,83950617	2,16049383	0,82110636	0,17889364



Estimation pour des censures multiples avec des données exactes

Méthode de Kaplan-Meïer généralisée

$$\hat{R}(t_i) = \frac{N_0 + 0{,}7}{N_0 + 0{,}4} \prod_{j=1}^{i} \left(1 - \frac{1}{n_j + 0{,}7}\right)$$

Avec

- N_0: taille de l'échantillon
- t_i : date de défaillance observée
- n_j: nombre de systèmes en test à t_{j-1}

Estimation pour des censures multiples avec des données exactes

Exemple

Rang	ti	Nature	ni	(1-1/(ni+0,7))	Π(1-1/(ni+0,7))	R(ti)	F(ti)
1	21	D	10	0,90654206	0,90654206	0,93269231	0,06730769
2	33	S	9		0,90654206		
3	40	D	8	0,88505747	0,80234182	0,8254863	0,1745137
4	66	D	7	0,87012987	0,69814158	0,71828028	0,28171972
5	70	S	6		0,69814158		
6	84	D	5	0,8245614	0,5756606	0,5922662	0,4077338
7	100	S	4		0,5756606		
8	110	S	3		0,5756606		
9	150	D	2	0,62962963	0,36245297	0,37290835	0,62709165
10	200	D	1	0,41176471	0,14924534	0,1535505	0,8464495

Estimation pour des censures multiples avec des données exactes

Méthode actuarielle

Discrétisation de l'axe temporel en intervalles $[t_{i-1}, t_i[$

$$\hat{R}(t_i) = \frac{N_0 + 0{,}7}{N_0 + 0{,}4} \prod_{j=1}^{i} \left(1 - \frac{d_j}{n_j + 0{,}7 - \frac{W_j}{2}}\right)$$

Avec

- N_0: taille de l'échantillon
- d_j: nombre de défaillances entre $[t_{j-1}, t_j[$
- n_j: nombre de systèmes en test à t_{j-1}
- W_j : nombre de censurés à droite sur $[t_{j-1}, t_j[$

Conclusion

De très nombreuses autres approches d'estimation non paramétriques sont proposées

- Transformation TTT (reposant sur le Temps Total on Test) qui offre une information sur la croissance de la fonction taux de défaillance
- Méthode d'estimation à « Noyau »

Ces approches offrent des conclusions plus ou moins équivalentes

Estimation paramétrique (méthode graphique)

Méthodologie

- Identique au cas « non censuré » en représentant les distributions empiriques des durées de vie **uniquement** aux points « **défaillance observée** »

Maximum de Vraisemblance pour un échantillon censuré

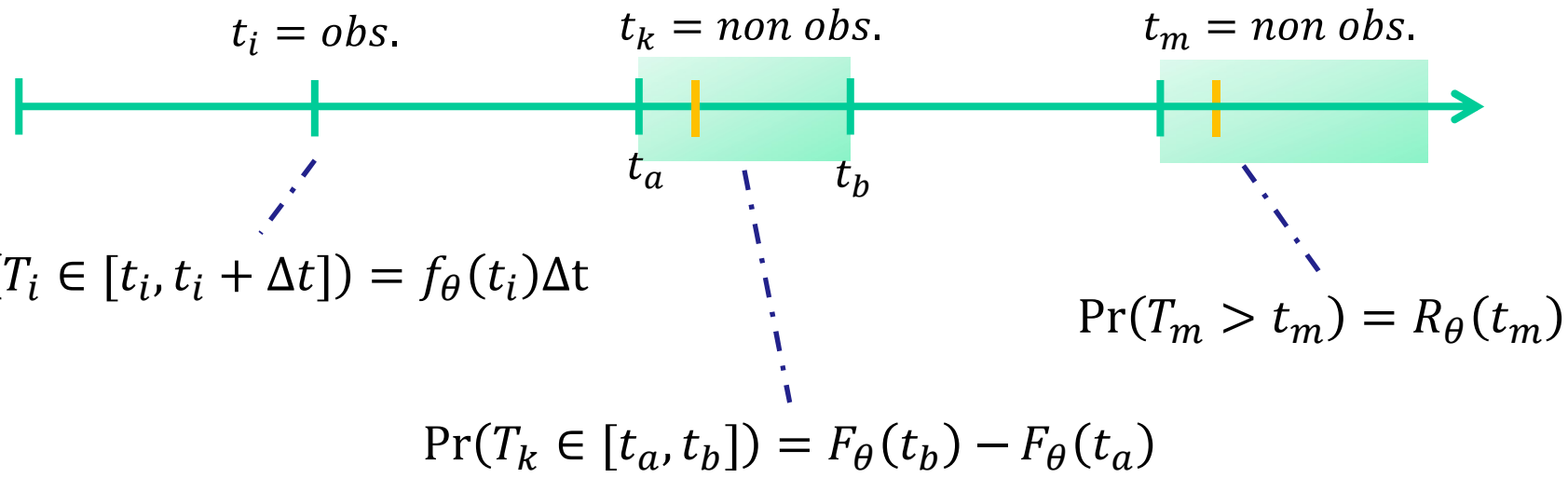
On rappelle que pour un échantillon complet

$$L_\theta(t_1, t_2, \dots, t_n) \overset{A.N.}{=} \prod_{i=1}^{n} \Pr(T_i = t_i | T_i \sim \mathcal{L}_\theta)$$

Dans le cas censuré,

- Si l'individu *i* est non censuré, on garde $\Pr(T_i = t_i | T_i \sim \mathcal{L}_\theta)$
- Si l'individu *i* est censuré à droite, on remplace $\Pr(T_i = t_i | T_i \sim \mathcal{L}_\theta)$ par $\Pr(T_i > t_i | T_i \sim \mathcal{L}_\theta)$
- Si l'individu *i* est censuré à droite, on remplace $\Pr(T_i = t_i | T_i \sim \mathcal{L}_\theta)$ par $\Pr(T_i \in Intervalle\ de\ censure | T_i \sim \mathcal{L}_\theta)$

Construction de la Vraisemblance pour un échantillon censuré



Ecrire la vraisemblance pour une loi quelconque de l'échantillon formé des observations sur n produits indépendants :

- $\{t_1, \ldots, t_g\}$ sont des dates d'inspection lors desquelles les systèmes correspondants sont défaillants
- $\{t_{g+1}, \ldots, t_{n-r}\}$ sont les dates observées des systèmes correspondants
- $\{t_{n-r+1}, t_n\}$ sont des dates pour lesquelles les systèmes correspondants ne sont pas défaillants

Vraisemblance d'un échantillon censuré

Ecrire la vraisemblance pour la loi exponentielle puis la loi de Weibull dans le cas où l'échantillon est formé de

- k dates de défaillance observées
- $n - k$ censures à droite

Déterminer les estimateurs du Maximum de Vraisemblance pour chacune des lois

Intervalle de confiance

Déterminer les IC de confiance des EMV en cas dans le même cas de données censurées

Intervalle de confiance (exponentielle)

Application à un IC sur λ dans le cas d'une censure de type I ou II

Intervalle bilatéral symétrique au niveau de confiance (1-α) :

- essai arrêté au bout d'un temps cumulé fixé T (essai tronqué) :

$$IC_\alpha = \left[\frac{z_{\alpha/2,2k}}{2T(t_0)}, \frac{z_{1-\alpha/2,2k+2}}{2T(t_0)}\right]$$

- essai arrêté à la k^{ième} défaillance (essai censuré)

$$IC_\alpha = \left[\frac{z_{\alpha/2,2k}}{2T(t_{(k)})}, \frac{z_{1-\alpha/2,2k}}{2T(t_{(k)})}\right]$$

Intervalle unilatéral à gauche au niveau de confiance (1-α) :

- essai arrêté au bout d'un temps cumulé fixé T (essai tronqué) :

$$\lambda \leq \frac{z_{1-\alpha,2k+2}}{2T(t_0)}$$

- essai arrêté à la k^{ième} défaillance (essai censuré)

$$\lambda \leq \frac{z_{1-\alpha,2k}}{2T(t_{(k)})}$$

Influence du niveau de censure

Procédure :

1. Construction d'un échantillon complet à partir de la vraie loi
2. Choix du niveau de censure r_c = pourcentage des données censurées dans l'échantillon
3. Détermination aléatoire des données censurées
   1. Identification des données censurées
   2. Tirage aléatoire d'une date de censure pour la donnée
4. Analyse statistique de l'échantillon
   1. Estimation des paramètres de la loi
   2. Construction des intervalles de confiance

Echantillon complet, *n* = 20 : 1.1 1.8 2.7 6. 6.1 7.3 9. 11.1 14. 14.9 18.7 20.8 21. 22.3 24.9 25.4 32.4 33.6 34.2 48.1

Influence du niveau de censure

Illustration de la procédure pour $r_c = 1/20 = 0{,}05$

3. Détermination aléatoire des données censurées
   1. Identification des données censurées

   On fait un tirage aléatoire suivant une loi binomiale sur les 20 données avec $p = 0{,}05$. On obtient par exemple 4. Donc $t_4$est censurée
   2. Tirage aléatoire d'une date de censure pour la donnée t_4

   On va utiliser la probabilité conditionnelle $\Pr(T < t | T \leq t_4)$

   Et on peut avoir par exemple, $t = 1{,}79$
4. Analyse statistique de l'échantillon
   1. Estimation des paramètres de la loi
   2. Construction des intervalles de confiance à 90%

Influence du niveau de censure

Taux de censure	$\hat{\lambda}(r_c)$	$\hat{\lambda}_L(r_c)$	$\hat{\lambda}_U(r_c)$
0	0,056	0,036	0,077
0,05	0,054	0,034	0,075
0,25	0,047	0,027	0,066
0,5	0,038	0,018	0,058
0,75	0,025	0,006	0,044
0,95	0,006	-0,004	0,016