LIFE TIME DATA ANALYSIS: THE COMPLETE DATABASE CASE

1. Operational reliability: Objective and analysis process
2. Methods for the life time distribution estimation
3. Estimation with the Maximum Likelihood method
4. Data analysis in case of multiple failure modes

1. Objectif et processus d'analyse

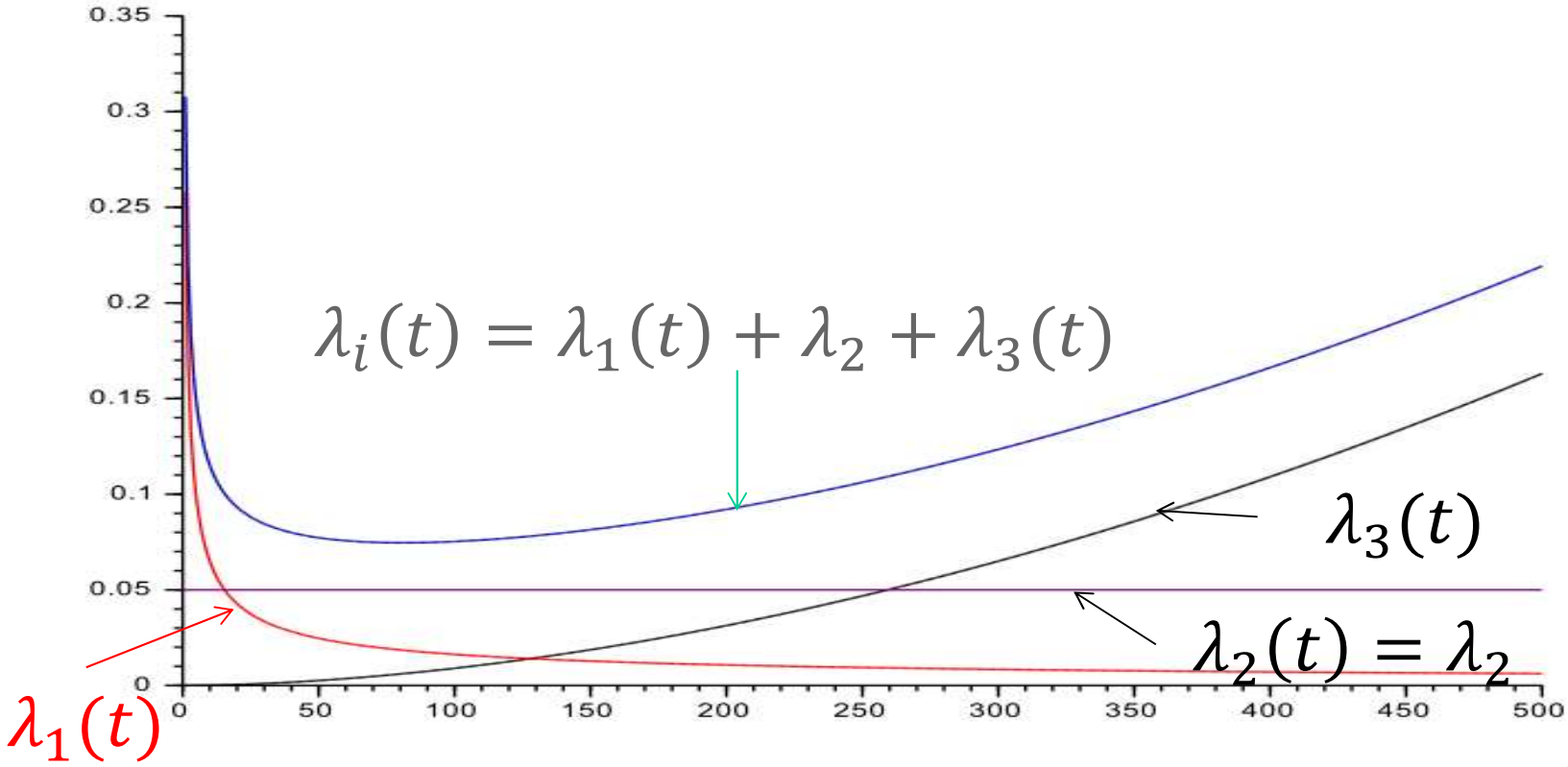
Objectif

Mettre en place une analyse de retour d'expérience sur des données de défaillance « brutes » et complètes pour la caractérisation du fonctionnement en phase opérationnelle

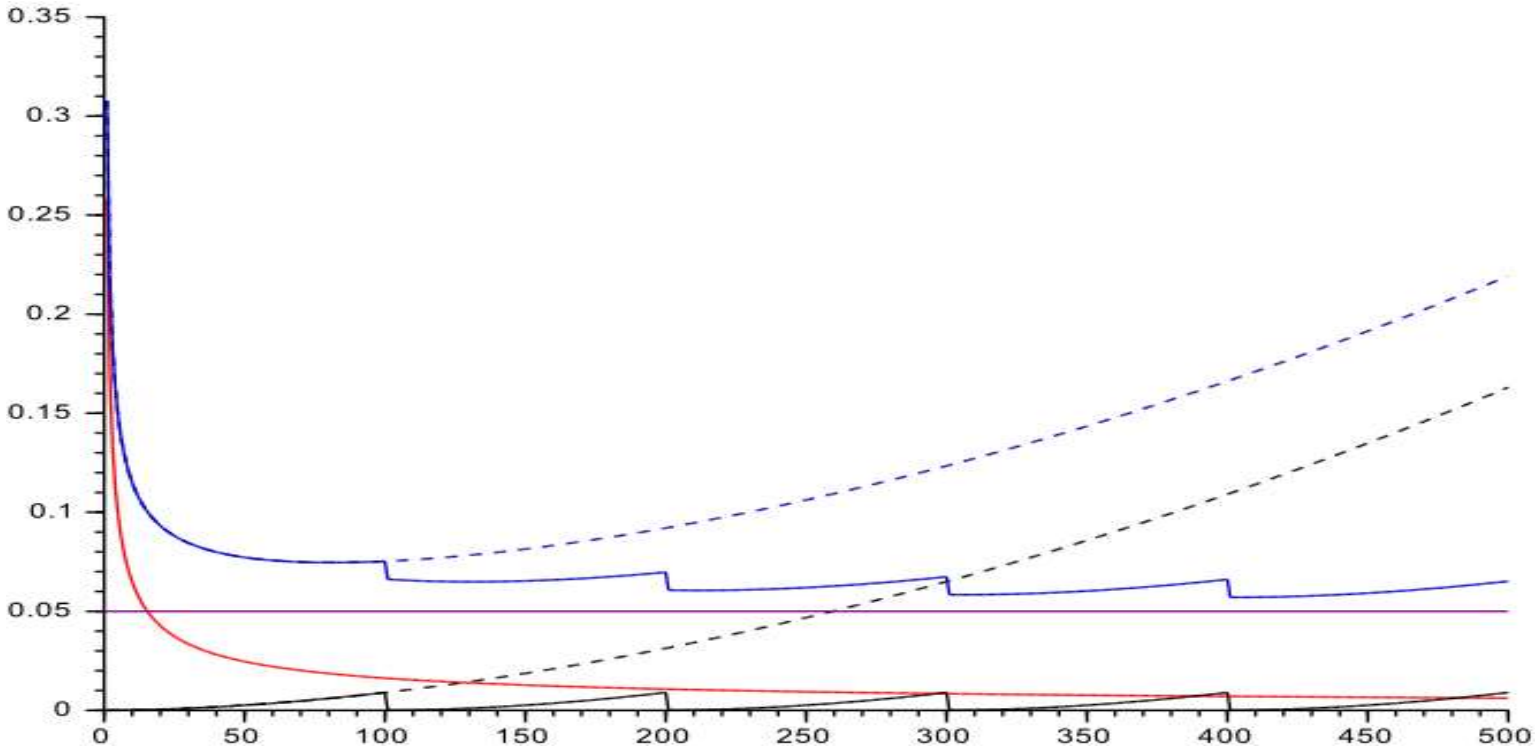
1. Objectif et processus d'analyse

Discussion sur le taux de fiabilité en phase opérationnelle (1)

Les différentes phases de maturité d'un produit
(défaillances intrinsèques)



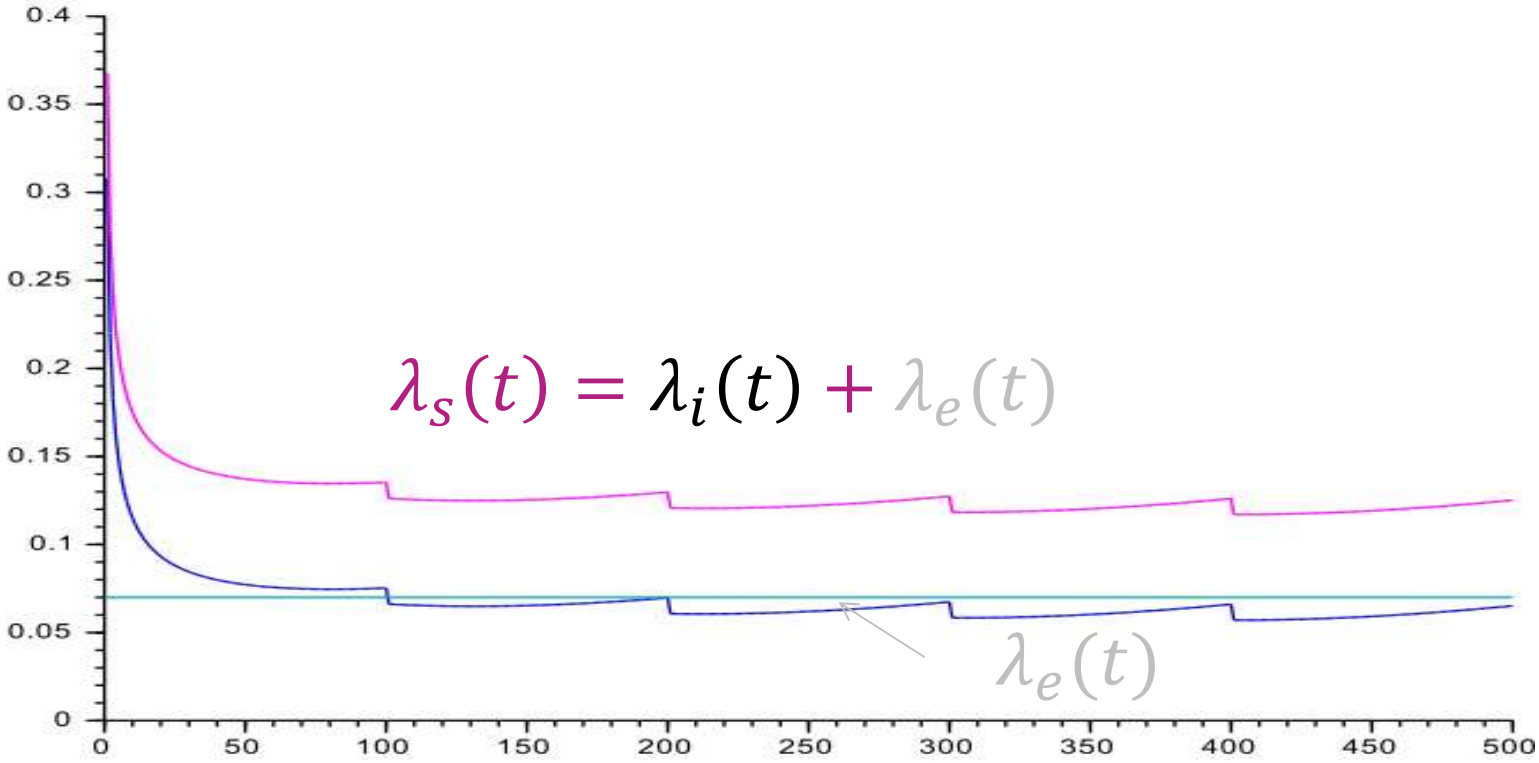
Effet d'une maintenance périodique préventive sur les différentes phases de maturité d'un produit



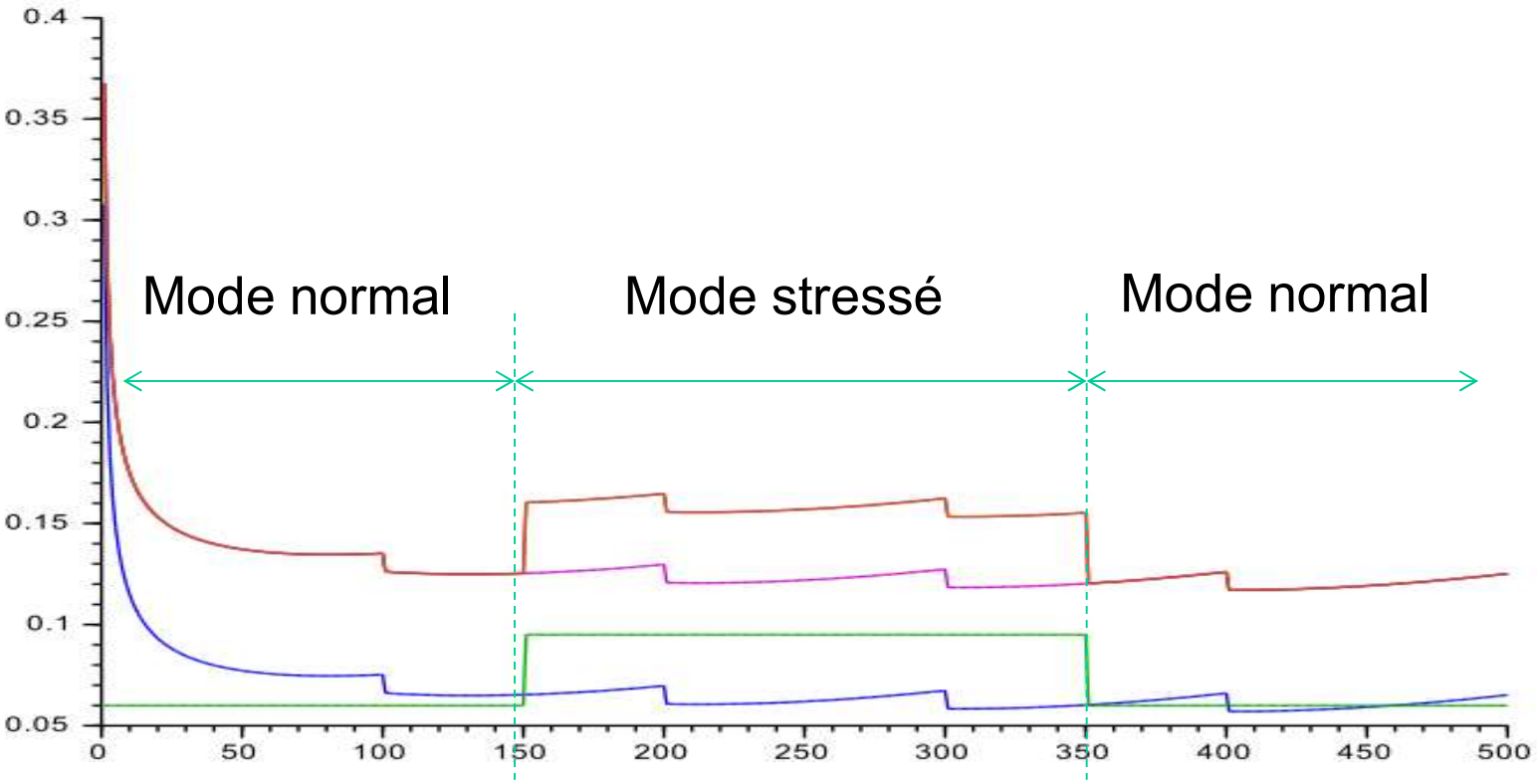
Pratiques industrielles : Approximation par un taux constant

$$\lambda_c = \frac{1}{\tau}\int_0^{\tau} \lambda_s(t)dt$$

Prise en compte des défaillances « extrinsèques » - Hyp. : taux de défaillance constant

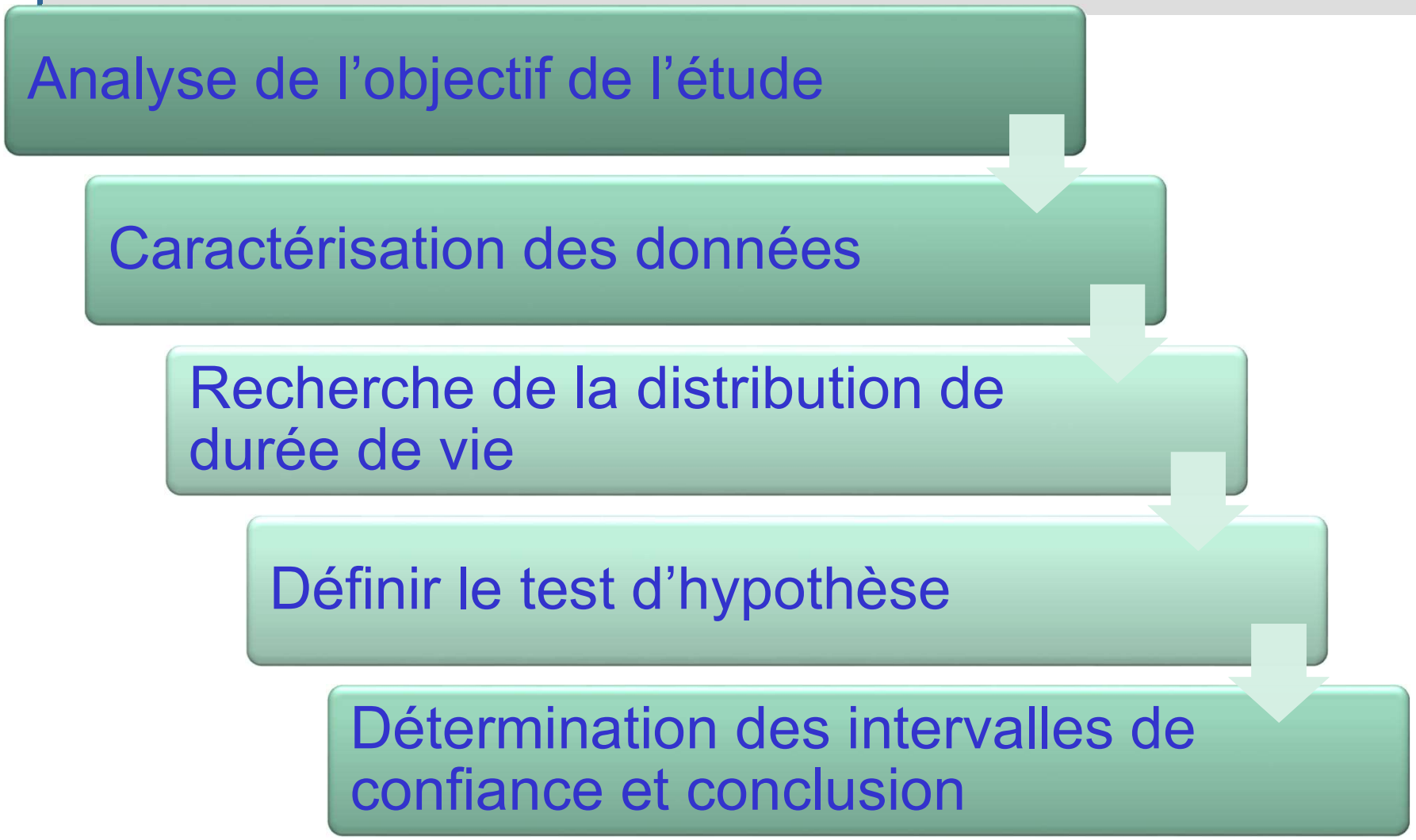


Prise en compte des défaillances « extrinsèques » - changement de modes de fonctionnement



1. Objectif et processus d'analyse

Processus d'Analyse d'une base de données



Vérification d'une exigence (de fiabilité)

- Analyse de la fonction du taux de défaillance

Exemple : Spécification de sécurité définie par un taux de défaillance maximal pour une défaillance critique

- L'exigence est-elle respectée ?
- Causes éventuelles de non respect :
  1. Estimation erronée en phase prévisionnelle
  2. Evolution du taux de défaillance dans le temps

Hypothèse : $\lambda_0 = 1,2 \cdot 10^{-6} h^{-1}$

1. Objectif et processus d'analyse

Caractérisation des données

- Type de données :
  - Durée de vie d'un composant ou d'un système
  - Données de dégradation
- Qualité de l'information :
  - Donnée observée
  - Donnée erronée
  - Donnée manquante
  - Donnée influente
  - Donnée censurée

Discussion

- Supposons un échantillon complet $\{t_1, t_2\}$ avec $t_1 = 86294{,}66\ h$ et $t_2 = 150205{,}02\ h$ (données génériques simulées à partir d'une loi exponentielle de paramètre λ_0)

$$\hat{\lambda}_2 = \frac{2}{t_1 + t_2} = 8{,}5 \cdot 10^{-6}\ h^{-1}$$

Conclusion : la loi opérationnelle n'est pas égale à la loi prévisionnelle ?

Questions :

1. Comment améliorer la qualité de la réponse ?
2. Comment fournir une garantie pour la réponse (qualification) ?

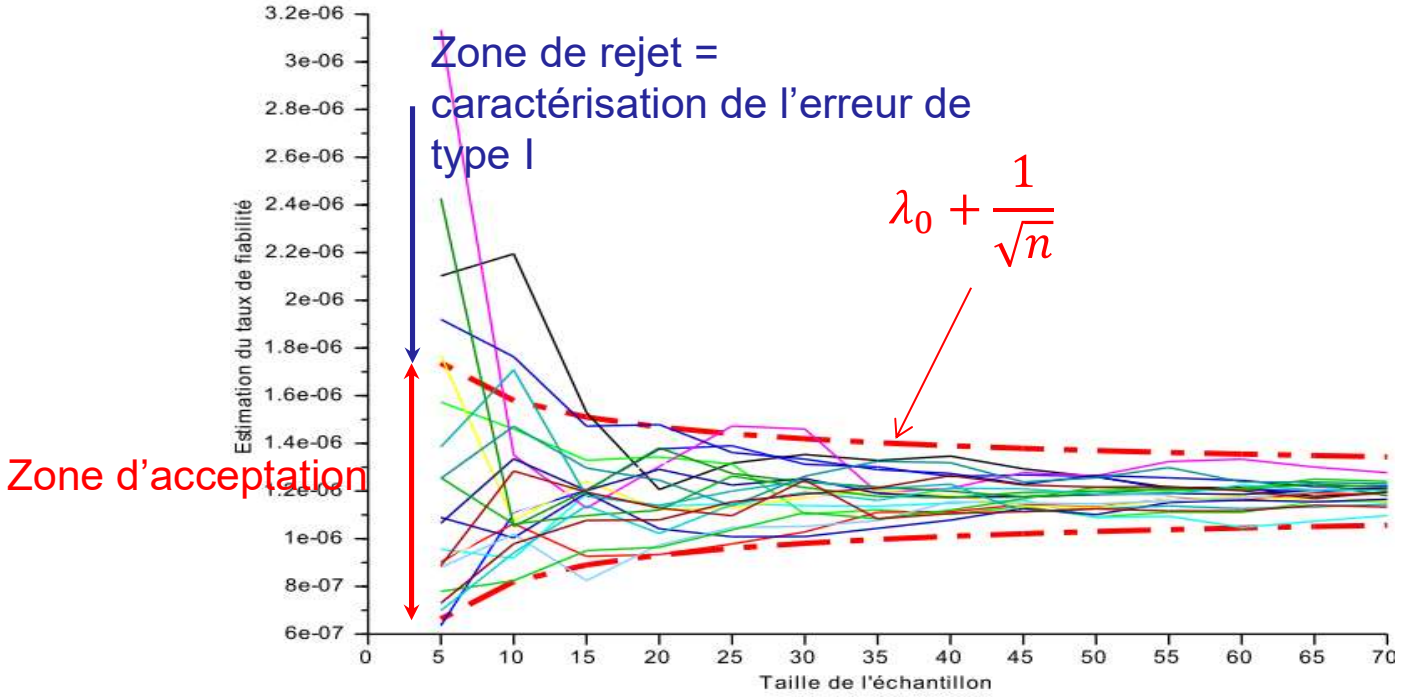
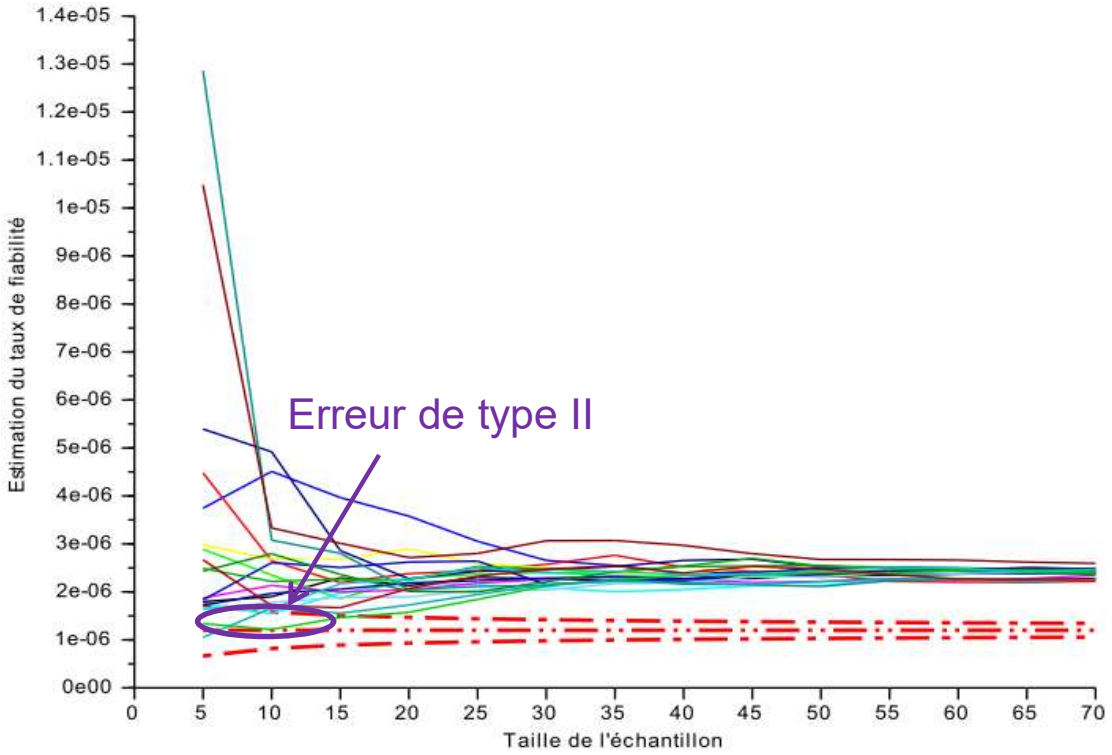


Illustration de la convergence d'un estimateur en fonction de la taille d'un échantillon

Définition : *l'intervalle de confiance pour* Θ *à un niveau de confiance* α *donné est déterminé par les deux fonctions* $\theta_L(T_1, \dots, T_n, \alpha)$ *et* $\theta_U(T_1, \dots, T_n, \alpha)$ *vérifiant*

$$\Pr(\Theta \in [\theta_L(T_1, \dots, T_n, \alpha);\ \theta_L(T_1, \dots, T_n, \alpha)]) = 1 - \alpha$$



Loi opérationnelle ≠ Loi prévisionnelle

Objectif : *Définir la région d'acceptation pour offrir le meilleur compromis entre les risques I et II*
⇒ *Construction d'un* ***test statistique d'hypothèses*** *(ou* ***test d'hypothèses paramétrique****)*

Un test est composé de 2 hypothèses :

$$H_0 : \theta \in \Theta_0 \qquad vs \qquad H_1 : \theta \in \Theta_1$$

Θ_0 et Θ_1 : ensembles respectifs et disjoints des propriétés à tester.

Généralement,

$$H_0 : \theta = \theta_0 \text{ vs } H_1 : \theta \neq \theta_0$$

Définition : $\alpha = \Pr(H_1 | H_0) \quad et \quad \beta = \Pr(H_0 | H_1)$

Objectif d'un test paramétrique (Newman-Pearson) :
Minimiser la région critique *W* ou encore minimiser
$\beta(W) = \Pr(H_0|H_1) = \Pr\big((T_1, T_2, \dots, T_n) \in \bar{W} | H_1\big) = 1 - \Pr\big((T_1, T_2, \dots, T_n) \in W | H_1\big)$
pour $\alpha(W) = \alpha$ fixé.

- **Approche :** Multiplicateur de Lagrange

Détermination de $W_\delta = \left\{ \frac{f(T_1,T_2,\dots,T_n ;\theta_1)}{f(T_1,T_2,\dots,T_n ;\theta_0)} = \frac{\prod_{i=1}^n f(T_i;\theta_1)}{\prod_{i=1}^n f(T_i;\theta_0)} \geq \delta \right\}$

avec $\delta > 0$ le multiplicateur de Lagrange.

Cas de la loi exponentielle

- Retour à l'exemple de sécurité :

$$H_0 : \lambda = \lambda_0 \qquad vs \qquad H_1 : \lambda = \lambda_1 \quad \text{avec} \quad \lambda_1 > \lambda_0$$

1. Ecrire l'inégalité des rapports des densités jointes
2. Simplifier l'expression
3. On cherche la loi de la statistique de test, ici $\sum_{i=1}^{n} T_i$

   Rappel : si $T_i \sim Exp(\lambda)$, $2\lambda \sum_{i=1}^{n} T_i$ suit une loi du Chi-Deux à $2n$ degrés de liberté
4. Détermination de la région critique *W*
5. Trouver la p-value du test

Exemple :

Le retour d'expérience de 10 pompes de sécurité identiques en fonction est formé des heures de fonctionnement avant défaillance suivantes :
16468, 26928, 27395, 244916, 262322, 294597, 418182, 444294, 506014, 561137.

Il est alors demandé de vérifier la conformité avec l'exigence d'un taux de défaillance $\lambda_0 = 1{,}2 \cdot 10^{-6}\ h^{-1}$ pour un niveau de confiance de $\alpha = 5\%$

Le tableau donne x tel que **P(K > x) = p**

p	0,999	0,995	0,99	0,98	0,95	0,9	0,8	0,2	0,1	0,05	0,02	0,01	0,005	0,001
ddl														
1	0,0000	0,0000	0,0002	0,0006	0,0039	0,0158	0,0642	1,6424	2,7055	3,8415	5,4119	6,6349	7,8794	10,8276
2	0,0020	0,0100	0,0201	0,0404	0,1026	0,2107	0,4463	3,2189	4,6052	5,9915	7,8240	9,2103	10,5966	13,8155
19	5,4068	6,8440	7,6327	8,5670	10,1170	11,6509	13,7158	23,9004	27,2036	30,1435	33,6874	36,1909	38,5823	43,8202
20	5,9210	7,4338	8,2604	9,2367	10,8508	12,4426	14,5784	25,0375	28,4120	31,4104	35,0196	37,5662	39,9968	45,3147
21	6,4467	8,0337	8,8972	9,9146	11,5913	13,2396	15,4446	26,1711	29,6151	32,6706	36,3434	38,9322	41,4011	46,7970

Comment choisir sa loi ?

Première approche :

- L'expertise en fonction des contextes d'utilisation et en analysant les tendances des taux de défaillance

Deuxième approche :

- Test d'adéquation graphique
- Test d'adéquation de type Chi-Deux
- Critère de type R^2, R^2 ajusté, Cp de Marlow, AIC, BIC, …

2. Analyse d'une base de données complète pour un unique mode de défaillance

Etude de cas

Objectif de l'étude :

On veut s'assurer de moins de 10 interventions sur une période d'une année d'exploitation supposée en continue sur une population de 20 calculateurs de navigation aéronautiques identiques, avec une probabilité supérieure à 80%.

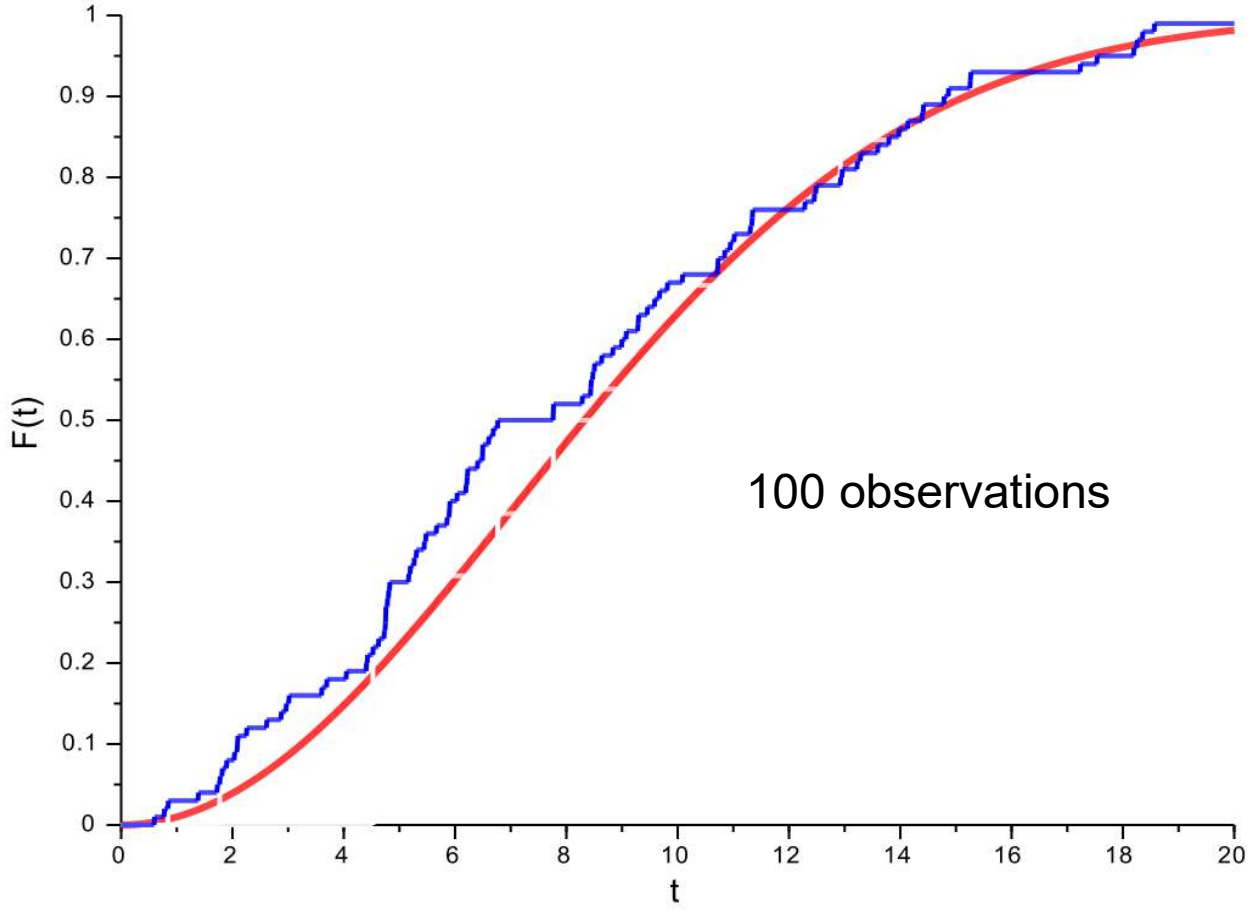
Pour cela, on utilisera les données d'intervention (Data 1) : 3723; 25406 ; 9394 ; 13510 ; 1753 ; 1926 ; 11174 ; 9501 ; 16350 ; 12360 ; 13610 ; 777 ; 422 ; 3351 ; 5189 ; 226 ; 19611 ; 10419 ; 644 ; 2306

Remarque : les données sont ici simulées

1. Formalisation du problème

2. Caractérisation de l’échantillon

3. Recherche de la loi de durée de vie



3.1. Construction de la loi empirique

Taille de l'échantillon	Fréquence cumulée
$n \leq 20$	$F_n(t_i) = \frac{i - 0{,}3}{n + 0{,}4}$
$20 < n \leq 50$	$F_n(t_i) = \frac{i}{n + 1}$
$n > 50$	$F_n(t_i) = \frac{i}{n}$

Exercice : Représentez graphiquement la distribution empirique de notre échantillon Data1

3.2. Tester l'homogénéité

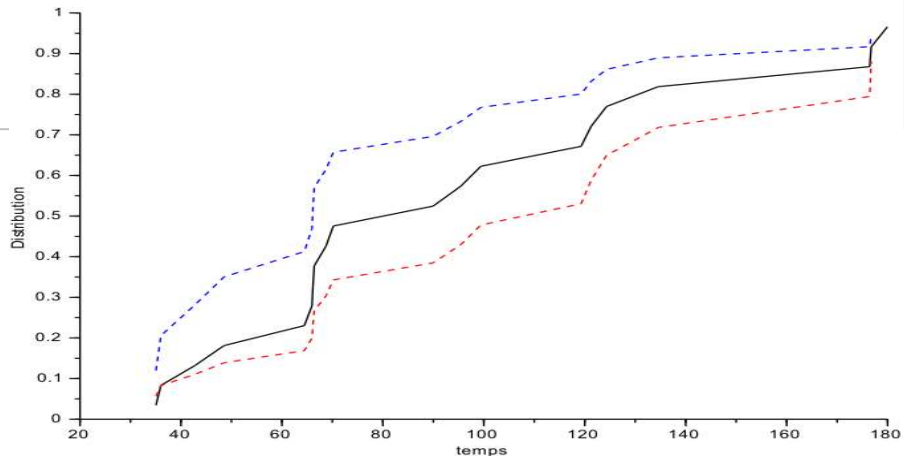
Démarche :

1. Pour chaque $F_n(t_i), i \in \{1, \dots, n\}$, déterminez l'intervalle de confiance bilatéral symétrique associé au niveau de confiance $(1-\alpha)$:

$$F_i(t_i) = \frac{i}{i + (n-i+1)F^{-1}_{student}\left(1-\frac{\alpha}{2}, 2(n-i+1), 2i\right)}$$

$$F_u(t_i) = \frac{(i+1)F^{-1}_{student}\left(1-\frac{\alpha}{2}, 2(i+1), 2(n-i)\right)}{n-i+(i+1)F^{-1}_{student}\left(1-\frac{\alpha}{2}, 2(i+1), 2(n-i)\right)}$$

2. Tracer les trois courbes $F_i(t_i), F_u(t_i)$ et $F_n(t_i)$
3. S'assurer que tous $[F_i(t_i), F_u(t_i)]_{i\in\{1,\dots,n\}}$.c



3.3. Choix d'un modèle paramétrique

Objectif :

$\Rightarrow$ trouver une loi paramétrique $F_\Theta(t)$ qui approchera au mieux le nuage de points formé par la distribution empirique.

Démarche :

1. Estimation de la fonction empirique
2. Choix d'une famille de loi paramétrique connue (exponentielle, Weibull, normale, lognormale)
3. Construction d'un test
    1. Graphique (visuel) par linéarité de la transformée de $g\left(F_{\mathrm{n}}(t_i)\right), i = 1, \ldots n$ en fonction du choix de la famille.
    2. Des tests d'hypothèses

Exemple : Soit l'échantillon suivant formé de $n = 20$ dates de défaillance observées : 91 ; 140 ; 159 ; 170 ; 171 ; 174 ; 186 ; 196 ; 198 ; 205 ; 221 ; 225 ; 226 ; 237 ; 250 ; 275 ; 282 ; 313 ; 366 ; 378.

Etape 1 : Construction de la fonction empirique

$$n = 20 \Rightarrow F_n(t_i) = \frac{i - 0{,}3}{n + 0{,}4}$$

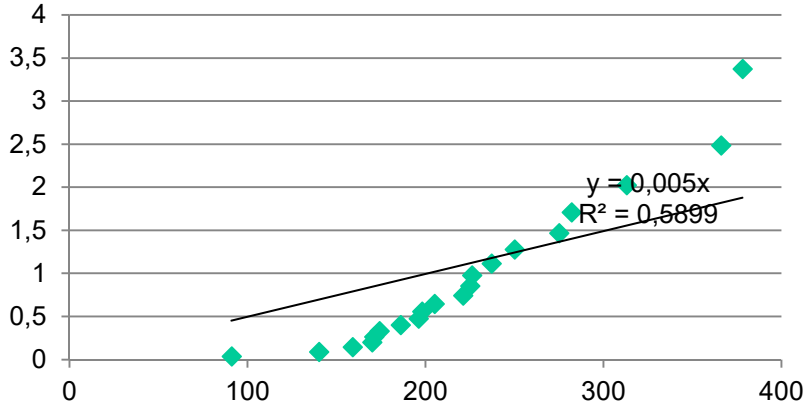
t_i	$F_n(t_i)$	t_i	$F_n(t_i)$	t_i	$F_n(t_i)$	t_i	$F_n(t_i)$
91	0,034	174	0,28	221	0,52	275	0,77
140	0,083	186	0,33	225	0,57	282	0,82
159	0,13	196	0,38	226	0,62	313	0,87
170	0,18	198	0,43	237	0,67	366	0,92
171	0,23	205	0,48	250	0,72	378	0,97

Exemple : Soit l'échantillon suivant formé de $n = 20$ dates de défaillance observées : 91 ; 140 ; 159 ; 170 ; 171 ; 174 ; 186 ; 196 ; 198 ; 205 ; 221 ; 225 ; 226 ; 237 ; 250 ; 275 ; 282 ; 313 ; 366 ; 378.

Etape 1 : Construction de la fonction empirique
Etape 2.a. : Test de la linéarité pour la famille exponentielle

t_i	$-\ln(1-F_n(t_i))$	t_i	$-\ln(1-F_n(t_i))$	t_i	$-\ln(1-F_n(t_i))$	t_i	$-\ln(1-F_n(t_i))$
91	0,034	174	0,28	221	0,52	275	0,77
140	0,083	186	0,33	225	0,57	282	0,82
159	0,13	196	0,38	226	0,62	313	0,87
170	0,18	198	0,43	237	0,67	366	0,92
171	0,23	205	0,48	250	0,72	378	0,97

On trace la courbe $\{(t_i, -\ln(1-F_n(t_i)))\}_{i\in\{1,\dots,n\}}$:

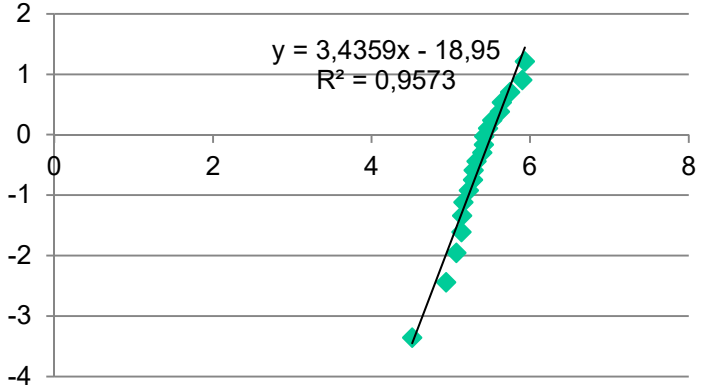


Exemple : Soit l'échantillon suivant formé de $n = 20$ dates de défaillance observées : 91 ; 140 ; 159 ; 170 ; 171 ; 174 ; 186 ; 196 ; 198 ; 205 ; 221 ; 225 ; 226 ; 237 ; 250 ; 275 ; 282 ; 313 ; 366 ; 378.

Etape 1 : Construction de la fonction empirique
Etape 2.a. : Test de la linéarité pour la famille exponentielle
Etape 2.b. : Test de la linéarité pour la famille Weibull

$\ln(t_i)$	$\ln(-ln(1-F_n(t_i)))$	$\ln(t_i)$	$\ln(-ln(1-F_n(t_i)))$	$\ln(t_i)$	$\ln(-ln(1-F_n(t_i)))$	$\ln(t_i)$	$\ln(-ln(1-F_n(t_i)))$
4,51	-3,38	5,16	-1,27	5,4	-0,65	5,62	-0,26
4,94	-2,49	5,23	-1,11	5,42	-0,56	5,64	-0,2
5,07	-2,04	5,28	-0,97	5,42	-0,48	5,75	-0,14
5,14	-1,71	5,29	-0,84	5,47	-0,4	5,9	-0,08
5,14	-1,47	5,32	-0,73	5,52	-0,33	5,93	-0,03

On trace la courbe $\{(\ln(t_i), \ln(-\ln(1-F_n(t_i))))\}_{i\in\{1,\dots,n\}}$:



3.3. Application sur notre échantillon

Testez graphiquement l'adéquation de notre échantillon à la loi exponentielle

3.4. Estimation ponctuelle des paramètres

Objectif :

⇒ Estimer les valeurs des paramètres $\widehat{\Theta}(t_1, \dots, t_n)$ pour la loi paramétrique choisie $F_\Theta(t)$.

Démarche :

1. Si $F_\Theta(t) \in \mathcal{F}(exponentielle)$, $\hat{\lambda} = g\left(F_\Theta(1)\right)$, coefficient directeur de la droite de régression
2. Si $F_\Theta(t) \in \mathcal{F}(Weibull)$,
   - $\hat{\beta}$ =coefficient directeur de la droite de régression
   - $\hat{\eta} = e^{-\frac{b}{\hat{\beta}}}$ où b = ordonnée à l'origine de la droite de régression
3. Tests d'adéquation

3.4.1 Test d'adéquation

Test du Chi-Deux

Objectif :

$\Rightarrow$ mesurer, en termes de probabilité, l'écart des données observées à des données théoriques obtenues à partir de la loi $F_{\Theta}(t)$

Démarche :

1. Construire la mesure des écarts :

$$E = \sum_{i=1}^{r} \frac{(n_i - np_i)^2}{np_i}$$

   - r le nombre de classes (> 4 individus par classe)
   - n_i le nombre d'individus dans la classe i
   - $p_i = F_{\widehat{\Theta}}(t_{i+1}) - F_{\widehat{\Theta}}(t_i)$

2. $E \sim$ Chi-deux à $\nu = r - k - 1$ ddl où k nombre de paramètres ($k = 1$ pour exponentielle et $k = 2$ pour Weibull).
3. Si $E \leq F_{\chi^2}^{-1}(1 - \alpha, \nu)$, alors l'hypothèse est acceptée.

Test du Chi-Deux

Objectif :

$\Rightarrow$ mesurer, en termes de probabilité, l'écart des données observées à des données théoriques obtenues à partir de la loi $F_\Theta(t)$

Exemple : $n = 50,\ \alpha = 0{,}05$

Classe (h)	0-50	50-100	100-150	150-200	200-250	250-300	300-350	350-400
n_i	0	0	5	7	5	16	9	3

1. On se restreint aux classes strictement positives donc $r = 6$,
2. Test Exponentiel :
   1. $F_{\hat{\lambda}=5e-3}(t) = 1 - e^{-\hat{\lambda}\cdot t}$, *d'où* $E = 207.$
   2. $k = 1$, d'où $\nu = 6.$
   3. $F^{-1}_{\chi^2}(1-\alpha, \nu) = 12{,}59$
   4. On rejette le test de la loi exponentielle $(E \gg 12{,}59)$
3. Test Weibull :
   1. $F_{(\hat{\beta}=3{,}43,\hat{\eta}=246{,}78)}(t) = 1 - \exp\left(-(t/\hat{\eta})^{\hat{\beta}}\right)$, *d'où* $E = 13{,}6$
   2. $k = 2$ d'où $\nu = 6.$
   3. $F^{-1}_{\chi^2}(1-\alpha, \nu) = 14{,}07$
   4. L'hypothèse est donc acceptée.

Test de Kolmogorov-Smirnov

Objectif :

$\Rightarrow$ mesurer, en termes de probabilité, l'écart des données observées à des données théoriques obtenues à partir de la loi $F_\Theta(t)$

Démarche :

1. Statistique de test : $D = \max_{i\in\{1,\dots n\}} |F_n(t_i) - F_\Theta(t_i)|$
2. $D \sim$ loi *D* ou loi de Kolmogorov-Smirnov
3. Comparez *D* avec $1{,}36/\sqrt{n}$ pour $n \geq 35$

Exemple : $n = 50$, $\alpha = 0{,}05$ pour la loi de Weibull

1. $D = \max_{i=1,\dots,50} \left| \frac{i}{n} - \left(1 - e^{-\left(\frac{t_i}{\eta_1}\right)^{\beta_1}}\right) \right| = 0{,}10$
2. $D_{0.05,50} = \frac{1{,}36}{\sqrt{50}} = 0{,}19$
3. Acceptation de l'hypothèse H_0.

Test de Anderson-Darling

Objectif :

⇒ mesurer, en termes de probabilité, l'écart des données observées à des données théoriques obtenues à partir de la loi $F_\Theta(t)$

Démarche :

1. Statistique de test : $AD_n = n\int_{-\infty}^{+\infty} \frac{[F_n(t)-\hat{F}_0(t)]^2}{\hat{F}_0(t)(1-\hat{F}_0(t))}\hat{f}_0(t)dt = -n + 1/n\sum_{i=1}^{n}[(2i-1-2n)\ln(1-U_i) - (2i-i)\ln U_i]$

   Avec $U_i = \hat{F}_0(t_i) = 1 - e^{-\frac{t_i}{\bar{T}_n}}$ (test exponentiel)

2. $AD_n \sim$ loi d'Anderson-Darling tabulée (ici on calcule la modifiée AD_n^*)

Table No.	Modified A^*	Upper tail percentage level α								
		.25	.20	.15	.10	.05	.025	.01	.005	.0025
1.	For all $n \geq 5$			1.610	1.933	2.492	3.070	3.853		
2. Case 1	See below	.644		.782	.894	1.087	1.285	1.551	1.756	1.964
Case 2	See below	1.072		1.430	1.743	2.308	2.898	3.702	4.324	4.954
Case 3	$A^* = A^2(1.0+0.75/n+2.25/n^2)$	.472	.509	.561	.631	.752	.873	1.035	1.159	1.283
3.	$A^* = A^2(1.0+0.3/n)$	.736	.816	.916	1.062	1.321	1.591	1.959	2.244	2.534

4. Estimation par intervalle

Objectif : *Construire un intervalle de confiance*

$\Rightarrow$ trouver les bornes de l'intervalle de confiance $\theta_L(T_1, \ldots, T_n, \alpha)$ et $\theta_U(T_1, \ldots, T_n, \alpha)$ fonction des valeurs des estimateurs. On parle aussi d'*estimation par intervalle*.

Cas de la loi exponentielle :

$$\Pr\left(F_{\chi^2}^{-1}\left(\frac{\alpha}{2}, 2n\right) \leq 2\lambda \sum_{i=1}^{n} T_i \leq F_{\chi^2}^{-1}\left(1-\frac{\alpha}{2}, 2n\right) \right) = 1-\alpha$$

$$\Rightarrow \boxed{IC_{1-\alpha}(\lambda) = \left[\frac{F_{\chi^2}^{-1}\left(\frac{\alpha}{2}, 2n\right)}{2\sum_{i=1}^{n} T_i} ; \frac{F_{\chi^2}^{-1}\left(1-\frac{\alpha}{2}, 2n\right)}{2\sum_{i=1}^{n} T_i} \right]}$$

Objectif : *Construire un intervalle de confiance*

$\Rightarrow$ trouver les bornes de l'intervalle de confiance $\theta_L(T_1, \ldots, T_n, \alpha)$ et $\theta_U(T_1, \ldots, T_n, \alpha)$ fonction des valeurs des estimateurs. On parle aussi d'*estimation par intervalle*.

Cas de la loi de Weibull :

Pas possible de construire une statistique pour la construction des deux intervalles de confiance.

Alternative : Utilisation de résultats de probabilité (borne de Cramer-Rao) et la *matrice d'information de Fisher* construite à partir des *Estimateurs du Maximum de Vraisemblance*.

5. Autres approches classiques d'estimation

La méthode des « moments »

Le Maximum de Vraisemblance

5.1 La méthode des « moments »

Approche :

On considère une loi de durée de vie de paramètre θ.

On a les moments successifs $E(T^s) = m_s(\theta)$ dont les estimations respectives sont

$$\hat{m}_s = \frac{1}{n}\sum_{i=1}^{n} t_i^s$$

L'estimateur des moments $\hat{\theta}$ est alors solutions du système

$$\left\{\hat{\theta} = m_s^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} t_i^s\right) \text{ où } s \text{ est la dimension de } \theta\right.$$

Exemple de la loi exponentielle :

$$E(T) = \frac{1}{\lambda}; \hat{m}_1 = \frac{1}{n}\sum_{i=1}^{n} t_i \Rightarrow \hat{\lambda} = \frac{n}{\sum_{i=1}^{n} t_i} = \frac{1}{\widehat{MTTF}}$$

5.2 Méthode du Maximum de Vraisemblance

Définition : *On appelle la vraisemblance d'un échantillon* $\{t_1, \dots, t_n\}$ *la probabilité que l'échantillon soit la réalisation de variables aléatoires* $(T_1, \dots, T_n)$ *qui suivent toute la même loi supposée connue et de paramètre* θ.

Remarque 1 :
Si on note $f_\theta(t_1, \dots, t_n)$ la densité de probabilité conjointe de l'échantillon, la *vraisemblance* est la variable aléatoire $L_\theta(T_1, \dots, T_n)$ définie par :

$$L_\theta(T_1, \dots, T_n) = f_\theta(T_1, \dots, T_n)$$

Remarque 2 :
Ainsi, si on note $E(t_1, t_2, \dots, t_n)$ l'événement $\{T_1 = t_1, T_2 = t_2, \dots, T_n = t_n\}$, la probabilité de réalisation de cet événement s'écrit :

$$L_\theta(t_1, t_2, \dots, t_n) = P_\theta(T_1 = t_1, T_2 = t_2, \dots, T_n = t_n) = f_\theta(t_1, t_2, \dots, t_n)$$

Remarque 3 :
Si on suppose les variables $T_i, i \in \{1, \dots, n\}$ indépendantes, on a :

$$L_\theta(t_1, t_2, \dots, t_n) = \prod_{i=1}^{n} f_\theta(t_i)$$

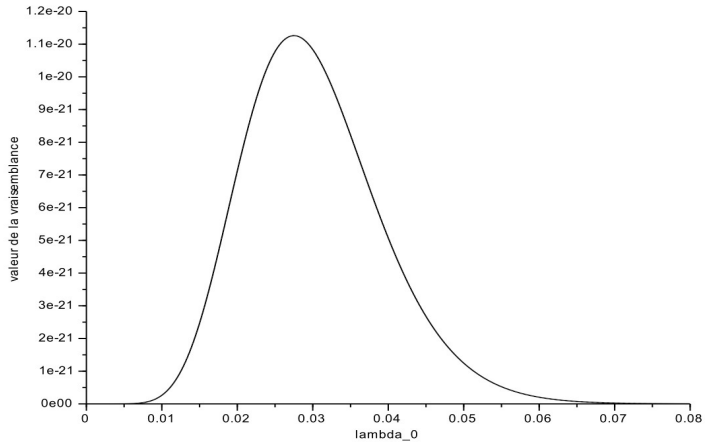
Définition de la vraisemblance d'un échantillon

Exemple 1 : Supposons

1. l'échantillon $\{3.24, 11.30, 15.72, 19.23, 29.15, 36.09, 38.58, 42.27, 77.81, 90.14\}$ (Data2)
2. la loi de l'échantillon est une loi exponentielle de paramètre $\lambda_0 = 0{,}03$

$$\Rightarrow \boxed{L_{\theta=\lambda_0}(t) = \prod_{i=1}^{n} \lambda_0 e^{-\lambda_0 t_i} = 1{,}083\ e-20}$$

Exemple 2 : Pour le même échantillon, si on fait varier λ_0 **:**



Méthode du Maximum de Vraisemblance - EMV

Définition : *On appelle l'Estimateur du Maximum de Vraisemblance, $\hat{\theta}_{EMV}$, la valeur qui maximise la fonction de vraisemblance pour une fonction donnée :*

$$\theta_{EMV} = arg \max_{\theta \in \mathbb{R}^k} L_\theta(t_1, t_2, \dots, t_n)$$

Où k est le nombre de paramètres de la loi paramétrique choisie.

Remarque 1 : Trouver l'EMV revient à rechercher le maximum de fonction $L_\theta(t_1, t_2, \dots, t_n)$ en θ.

Remarque 2 :

$$\max_\theta L_\theta(t_1, t_2, \dots, t_n) = \max_\theta \prod_{i=1}^n f_\theta(t_i) = \max_\theta \ln \prod_{i=1}^n f_\theta(t_i) = \max_\theta \sum_{i=1}^n \ln f_\theta(t_i)$$

Remarque 3 : Ceci est un problème classique d'optimisation continu. On pourra le résoudre par la méthode du gradient et l'EMV sera alors la valeur qui annule le système des dérivées partielles

Méthode du Maximum de Vraisemblance - EMV

Procédure :

1. Choisir une loi paramétrique $f_\theta(t)$. Ce choix pourra être réalisé par le biais d'une approche comme vue précédemment.
2. Ecrire la fonction de vraisemblance
3. Ecrire la fonction Log-vraisemblance
4. Dériver la fonction Log-vraisemblance en fonction de chacun des paramètres de la fonction $f_\theta(t)$
5. Résoudre le système formé des équations aux dérivées partielles.

Propriétés de l'EMV :

1. L'EMV est convergent
2. Il est asymptotiquement efficace, il converge vers la borne de Cramer-Rao
3. Il est asymptotiquement distribué selon une loi normale
4. Il peut être biaisé pour un échantillon fini (voir pour la loi exponentielle)

Méthode du Maximum de Vraisemblance - IC

Préambule : La procédure de construction d'un intervalle de confiance n'est pas unique. Par exemple :

1. IC pour l'exponentielle reposant sur la statistique $\sum_{i=1}^{n} T_i$
2. Test de Wald ou IC par approximation asymptotique (ou normale)
3. Test du rapport de vraisemblance

Notations : On note

- $\hat{\theta} = \theta_{EMV}$
- $L_{\hat{\theta}} = L_{\theta_{EMV}}(t_1, t_2, \dots, t_n)$

Objectif : On veut construire des intervalles de confiance sur la base de tests d'hypothèse :

$$H_0 : \theta = \theta_0 \qquad vs \qquad \theta \neq \theta_0$$

Avec θ_0 le vecteur des paramètres, $\theta_0 \in \mathbb{R}^k$

Méthode du Maximum de Vraisemblance - IC

Test de Wald : (application au cas $\theta_0 \in \mathbb{R}$)

- Sous H_0, $W_{\theta_0} = \frac{\hat{\theta} - \theta_0}{\widehat{Se}_{\hat{\theta}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$

  Avec $\widehat{Se}_{\hat{\theta}}$ l'estimation de l'écart-type de l'estimateur de $\hat{\theta}$
- Donc $\boldsymbol{IC_{1-\alpha} = [\theta_L, \theta_U] = \hat{\theta} \pm z_{1-\alpha/2} \cdot \widehat{Se}_{\hat{\theta}}}$

  Avec $z_{1-\alpha/2} = \Phi^{-1}(1 - \alpha/2)$, le $(1 - \alpha/2)$-quantile de la loi normale et $\Phi^{-1}(\cdot)$ l'inverse de la distribution de la loi normale $\mathcal{N}(0,1)$.

Inconnues ?

-

Calcul de $\widehat{Se}_{\hat{\theta}}$ l'estimation de l'écart-type de l'estimateur de $\hat{\theta}$

Théorème : Un estimateur de la matrice de variance-covariance peut être évalué à partir de l'information de l'échantillon ou information observée par la formule suivante :

$$\hat{V}(\hat{\theta}) = [\mathcal{F}]^{-1}$$

Avec $[\mathcal{F}]$ *la matrice d'information de Fisher définie par :*

$$[\mathcal{F}] = \left[-\frac{\partial^2 \ln L_{\hat{\theta}}}{\partial\theta\partial\theta'} \right]$$

$$\widehat{Se}_{\hat{\theta}} = \hat{V}(\hat{\theta})^{\frac{1}{2}}$$

Remarques :

- Si $\theta \in \mathbb{R}, [\mathcal{F}] = -\frac{\partial^2 \ln L_{\hat{\theta}}}{\partial\theta^2}$
- Si $\theta = (\theta_1, \theta_2) \in \mathbb{R}^2, [\mathcal{F}] = \begin{bmatrix} -\frac{\partial^2 \ln L_{\hat{\theta}}}{\partial\theta_1^2} & -\frac{\partial^2 \ln L_{\hat{\theta}}}{\partial\theta_1\partial\theta_2} \\ -\frac{\partial^2 \ln \hat{\theta}}{\partial\theta_1\partial\theta_2} & -\frac{\partial^2 \ln \hat{\theta}}{\partial\theta_2^2} \end{bmatrix}$

Application à la loi exponentielle

Recherchez l'EMV pour la loi exponentielle λ_0

Déterminez l'$IC_{1-\alpha}$ par le biais de la statistique de Wald

Intervalle de confiance pour les grandeurs de fiabilité

Contexte : Les spécifications de fiabilité sont généralement données en fonction de quantités autres que les valeurs directes des lois mais plutôt par des seuils de fonctions de ces paramètres, $g(\theta)$

Corollaire : L'estimateur ponctuel de la grandeur $y = g(\theta)$ est donné par

$$\hat{y} = g(\hat{\theta})$$

Exemples :

- La L_{10} pour une loi exponentielle est donnée par $L_{10} = \frac{-\ln 0{,}9}{\lambda}$, d'où

$$\hat{L}_{10} = \frac{-\ln 0{,}9}{\hat{\lambda}}$$

Intervalle de confiance pour les grandeurs de fiabilité

Pour la construction de $IC_{1-\alpha}$, la démarche reste identique. On aura

$$[y_L, y_U] = \left[\hat{y} \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\hat{V}(\hat{y})}\right] \quad si \quad y \in (-\infty, +\infty)$$

$$[y_L, y_U] = \left[\hat{y} e^{\pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{\hat{V}(\hat{y})}}{\hat{y}}}\right] \quad si \quad y \in (0, +\infty)$$

$$[y_L, y_U] = \left[\frac{\hat{y}}{\hat{y} + (1-\hat{y}) e^{\pm z_{1-\frac{\alpha}{2}} \cdot \frac{\sqrt{\hat{V}(\hat{y})}}{\hat{y}}}}\right] \quad si \quad y \in (0,1)$$

Rappel : Si $\hat{y} = g(\hat{\theta})$,

$$\hat{V}(\hat{y}) = \hat{V}\left(g(\hat{\theta})\right) = \nabla g(\hat{\theta})\, \hat{V}(\hat{\theta}) \nabla g(\hat{\theta})'$$

Où $\nabla g(\hat{\theta})$ est le gradient de la fonction $g(\cdot)$ au point $\hat{\theta}$ et $\nabla g(\hat{\theta})'$ sa transposée.

Analyse d'une base de données pour un système à plusieurs modes de défaillance

Cas de mélange de lois

Contexte : L'analyse prévisionnelle d'un système nous assure l'existence de 2 modes de défaillance indépendants entrainant chacun l'arrêt total du système.

Problématique : Comment estimer les lois de durée de vie pour chacun des modes si

1. Les modes sont identifiés dans la base de données
2. Les modes ne sont pas identifiés dans la base de données

Cas de mélange de lois

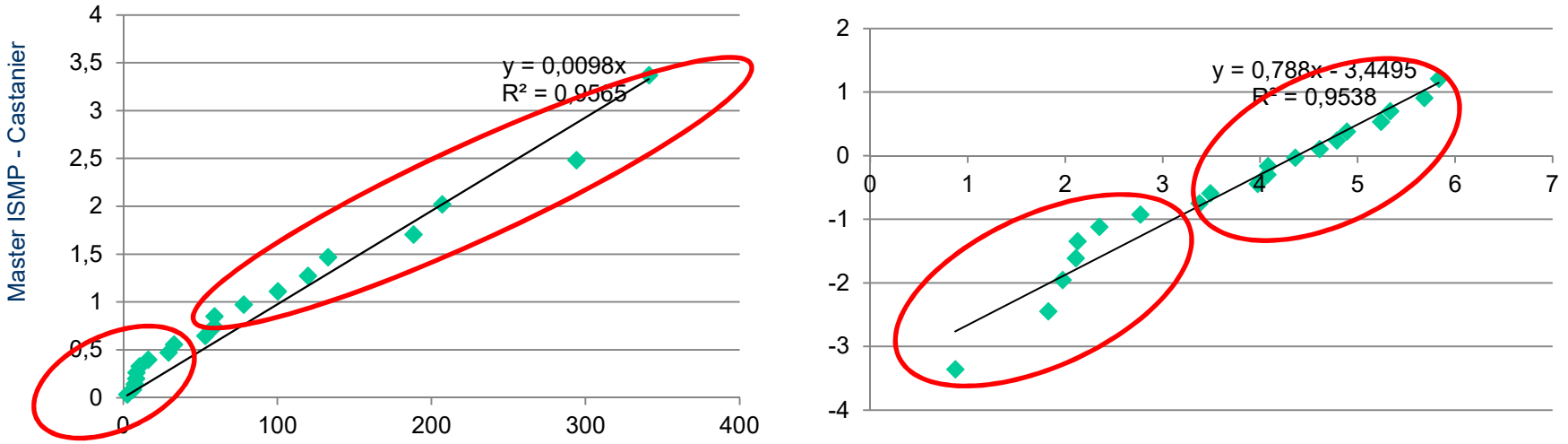
Cas 1 : Modes de défaillance connus (Data3)

Mode 1		Mode 2	
n_1	Observation	n_2	
1	2,39	1	6,20
2	29,26	2	7,19
3	32,72	3	8,23
4	53,10	4	8,37
5	58,79	5	10,46
6	59,04	6	15,94
7	78,04	7	119,76
8	100,20		
9	132,81		
10	188,13		
11	206,80		
12	293,87		
13	341,15		
Moyenne	121,25	Moyenne	25,17

1. Estimation de la proportion π_1 de défaillances de mode 1 : $\hat{\pi}_1 = n_1/n$
2. Conduire les estimations pour chacun des modes de manière indépendantes
3. Construire la loi générale : $\hat{R}(t) = \hat{\pi}_1 \hat{R}_1(t) + (1 - \hat{\pi}_1)\hat{R}_2(t)$

Cas de mélange de lois

Cas 2 : Modes de défaillance non connus
Éch = {2,39 ; 6,20 ; 7,18 ; 8,23 ; 8,37 ; 10,46 ; 15,94 ; 29,26 ; 32,72 ; 53,10 ; 58,79 ; 59,04 ; 78,04 ; 100,20 ; 119, 76 ; 132,81 ; 188,13 ; 206,80 ; 293,87 ; 341,15}



Hypothèse (fins pédagogiques) : Mélange de 2 lois exponentielles
Question : Comment affecter les points à chacune des lois ?
Réponse :

1. Construction d'une approche empirique
2. Construction d'un modèle d'optimisation

Approche empirique :

1. Sélection « intuitive » des données pour chacun des modes. Par ex :
   1. $\{t_7, \dots, t_{20}\} \rightarrow f_{\lambda_1}(t)$
   2. $\{t_1, \dots, t_6\} \rightarrow f_{\lambda_2}(t)$
2. Estimation de $\pi_1$et des λ_i

Réponse : $\hat{\pi}_1 = \frac{14}{20}$, $\hat{\lambda}_1 = \frac{14}{\sum_{i=7}^{20} t_i} = 0{,}008h^{-1}$ et $\hat{\lambda}_2 = \frac{6}{\sum_{i=1}^{6} t_i} = 0{,}14h^{-1}$

Identifier les risques de la procédure

Cas de mélange de lois

Approche par optimisation (ou du Maximum de Vraisemblance) :

1. La densité de probabilité s'écrit :

$$f_{(\lambda_1,\lambda_2,\pi)}(t) = \pi_1\lambda_1 e^{-\lambda_1 t} + \pi_2\lambda_2 e^{-\lambda_2 t}$$

2. On écrit la Log-Vraisemblance

$$\ln L_{(\lambda_1,\lambda_2,\pi)}(t_1,\dots,t_n) = \sum_{i=1}^{20} \ln\left(\pi_1\lambda_1 e^{-\lambda_1 t_i} + \pi_2\lambda_2 e^{-\lambda_2 t_i}\right)$$

3. Problème d'optimisation sous contrainte $\sum_{k=1}^{2}\pi_k = 1 \Rightarrow$ Méthode du multiplicateur de Lagrange

$$g(\lambda_1,\lambda_2,\pi,b) = \sum_{i=1}^{20} \ln\left(\pi_1\lambda_1 e^{-\lambda_1 t_i} + \pi_2\lambda_2 e^{-\lambda_2 t_i}\right) + b\left(\sum_{k=1}^{2}\pi_k - 1\right)$$

Approche par optimisation (ou du Maximum de Vraisemblance) : (suite)

4. Systèmes des dérivées partielles

$$\begin{cases} \dfrac{\partial g(\lambda_1,\lambda_2,\pi,b)}{\partial \lambda_k} = \displaystyle\sum_{i=1}^{n} \frac{\pi_k e^{-\lambda_k t_i} - \lambda_k t_i e^{-\lambda_k t_i}}{\pi_1 \lambda_1 e^{-\lambda_1 t_i} + (1-\pi_1)\lambda_2 e^{-\lambda_2 t_i}} = 0 \quad (1) \\ \dfrac{\partial g(\lambda_1,\lambda_2,\pi,b)}{\partial \pi_k} = \displaystyle\sum_{i=1}^{n} \frac{\lambda_k e^{-\lambda_k t_i}}{\pi_1 \lambda_1 e^{-\lambda_1 t_i} + (1-\pi_1)\lambda_2 e^{-\lambda_2 t_i}} + b = 0 \quad (2) \\ \dfrac{\partial g(\lambda_1,\lambda_2,\pi,b)}{\partial b} = -\left(\displaystyle\sum_{k=1}^{2} \pi_k - 1\right) = 0 \quad (3) \end{cases}$$

5. La quantité $\gamma(z_{i,k}) = \frac{\pi_k \lambda_k e^{-\lambda_k t_i}}{\pi_1 \lambda_1 e^{-\lambda_1 t_i} + (1-\pi_1)\lambda_2 e^{-\lambda_2 t_i}}$ représente la probabilité que l'individu T_i appartienne à la classe k.

EMV =

$$\hat{\lambda}_k = \frac{n_k}{\sum_{i=1}^{n} \gamma(z_{i,k}) t_i} \quad pour\ k = 1\ et\ 2$$

$$\hat{\pi}_1 = 1 - \hat{\pi}_2 = \frac{\sum_{i=1}^{n} \gamma(z_{i,1})}{n} = \frac{n_1}{n}$$

Approche par optimisation (ou du Maximum de Vraisemblance) : (suite)

4. Procédure de résolution : Algorithme EM

   a. Initialisation : $\pi^{(0)} = (0{,}5\,;0{,}5)$; $\lambda_1^{(0)} = 0{,}005$; $\lambda_2^{(0)} = \frac{\left(n/\sum_{i=1}^{n} t_i - \pi_1^{(0)}\lambda_1^{(0)}\right)}{\pi_2^{(0)}} = 0{,}018.$

   b. Pour $\nu = 2{:}\,50$

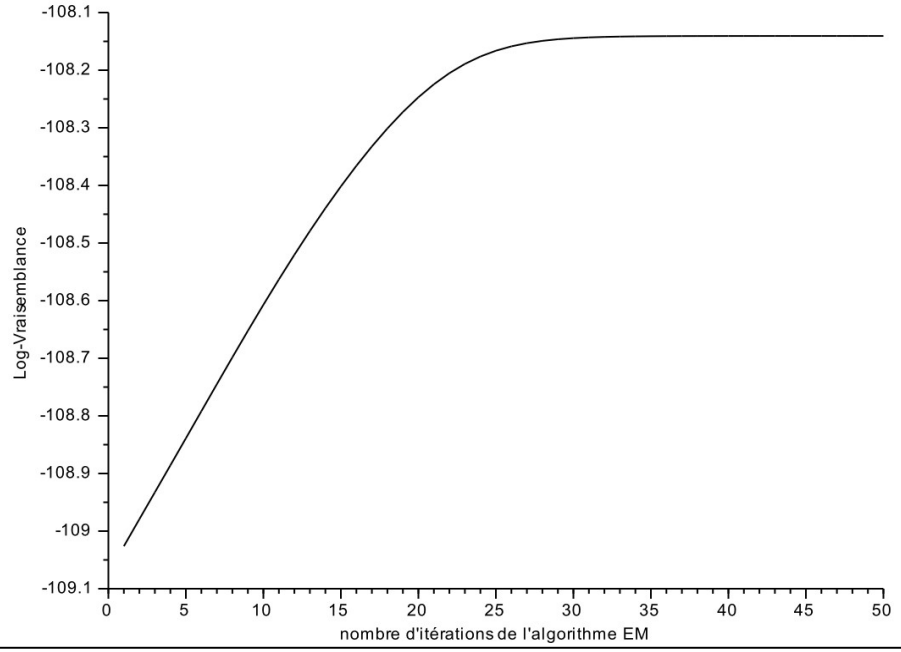
   - Etape E : évaluer $\gamma^{(\nu)}(z_{i,k}) = \frac{\pi_k^{(\nu-1)}\lambda_k^{(\nu-1)}e^{-\lambda_k^{(\nu-1)}t_i}}{\sum_{k=1}^{2}\pi_k^{(\nu-1)}\lambda_k^{(\nu-1)}e^{-\lambda_k^{(\nu-1)}t_i}}, \forall(i,k)$
   - Etape M : évaluer $\begin{cases} \hat{\pi}_k^{(\nu)} = \frac{\sum_{i=1}^{n}\gamma^{(\nu)}(z_{i,k})}{n} = \frac{n_k}{n} \\ \hat{\lambda}_k^{(\nu)} = \frac{n_k}{\sum_{i=1}^{n}\gamma^{(\nu)}(z_{i,k})t_i} \end{cases} pour\ k = 1\ et\ 2$
   - Evaluation de la fonction-objectif : le Log-Vraisemblance

$$\ln L_{(\lambda_1,\lambda_2,\pi)}(t_1,\ldots,t_n) = \sum_{i=1}^{20}\ln\left(\sum_{k=1}^{2}\pi_k^{(\nu)}\lambda_k^{(\nu)}e^{-\lambda_k^{(\nu)}t_i}\right)$$

End

Approche par optimisation (ou du Maximum de Vraisemblance) : (suite)

4. Résultat :



	Exponentielle 1		Exponentielle 2	
	λ_1	π_1	λ_2	π_2
Modes connus	0,008	0,65	0,04	0,35
Approche intuitive	0,008	0,7	0,14	0,3
EM	0,009	0,76	0,09	0,24