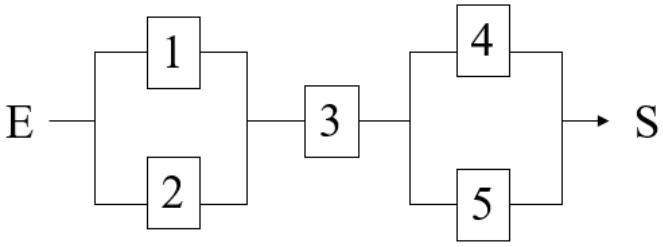
Master 2 de Statistique et Science des Données



Fiabilité des Systèmes

Notes de cours

Olivier Gaudoin

olivier.gaudoin@univ-grenoble-alpes.fr

Table des matières

Chapitre 1

Maîtrise des risques, sûreté de fonctionnement et fiabilité

1.1 Contexte

Ces dernières années, les problèmes liés à la **maîtrise des risques** ont vu leur importance et leur retentissement considérablement augmenter.

Exemples :

- risque sanitaire : covid 19, grippe aviaire, grippe A, chikungunya, virus Ebola, ...
- risque environnemental et climatique : canicules (2003, 2018-2020 en France), incendies (Californie 2018-2021, Amazonie 2019, Notre Dame de Paris 2019), inondations (Belgique et Allemagne 2021), tsunamis (2004, 2011), ouragans (Katrina 2005, Irma 2017, Dorian 2019, Ida 2021), tremblements de terre (Haïti 2010, Katmandou 2015, Italie 2016), dôme de chaleur (Canada 2021),...
- risque financier : risque de crédit, risque de marché, crise des subprimes (2008), affaires Kerviel (2008) et Madoff (2008), ...
- risque géopolitique : guerre, terrorisme, ...
- risque industriel : explosion de la plateforme pétrolière Deepwater Horizon (2010), du lanceur Falcon 9 de Space X (2016), accidents d'avion ou de train (Brétigny sur Orge, 2013), accidents nucléaires de Tchernobyl (1986) et Fukushima (2011), effondrement du pont de Gênes (2018), incendie de l'usine Lubrizol de Rouen (2019), explosion du port de Beyrouth (2020),...

Plus prosaïquement, les problèmes de fiabilité surgissent également dans la vie de tous les jours :

- panne électrique de 2 semaines à la gare Montparnasse (2018)
- panne de réseaux téléphoniques, de distributeurs de billets,...
- panne de voiture, retard de train,...
- panne de la machine à café ou du photocopieur de l'UFR,...

Un **risque** associe un danger potentiel et une perte quantifiable dus à un évènement indésirable et la probabilité d'occurrence de cet événement. La gravité et l'acceptabilité sont d'autres composantes du risque.

L'**analyse des risques** a pour objectif d'évaluer et d'anticiper les risques. Cela passe en particulier par :

— l'**évaluation** de la probabilité d'occurrence d'un évènement indésirable.
— la **prévision** de ces évènements et de leurs conséquences.
— l'évaluation des **facteurs de risque**.
— la mise en place de **mesures de prévention** et de réduction des risques (systèmes de surveillance et d'alerte par exemple), de façon à réaliser un **compromis** entre le **gain en sûreté** obtenu et le **coût** de ces mesures.
— l'étude de la **perception** des risques.
— etc.

L'opinion publique accepte de moins en moins la notion de risque et demande des niveaux de sûreté et de sécurité de plus en plus élevés. Pourtant, le risque zéro n'existe pas. Toutes les entreprises et les collectivités locales, nationales et internationales, sont concernées par la mesure, la gestion et la prévention des risques.

Ce cours porte sur la **fiabilité**, composante principale de la **sûreté de fonctionnement**, qui est l'élément clé de la gestion des risques industriels. La plupart de ses concepts peuvent se transposer à la gestion des autres types de risque.

1.2 Terminologie générale de la sûreté de fonctionnement

Un **système** est un ensemble de composants en interaction destiné à accomplir une tâche donnée. C'est le cas par exemple des systèmes de production, systèmes de transport, systèmes informatiques, etc...

La **sûreté de fonctionnement** (SdF, en anglais *dependability*) d'un système est la propriété qui permet à ses utilisateurs de placer une confiance justifiée dans le service qu'il leur délivre. On dit aussi que la SdF est la science des défaillances.

Un système subit une **défaillance** quand il ne peut plus délivrer le service attendu. La **panne** est l'état du système résultant d'une défaillance.

La sûreté de fonctionnement comprend 4 composantes : la fiabilité, la disponibilité, la maintenabilité et la sécurité.

- La **fiabilité** (*reliability*) est la caractéristique du système exprimée par la probabilité qu'il délivre le service attendu dans des conditions données et *pendant une durée déterminée*. La fiabilité exprime l'aptitude à la continuité du service (ex : envoyer un robot sur Mars).
- La **disponibilité** (*availability*) est exprimée par la probabilité que le système délivre le service attendu dans des conditions données et *à un instant donné*. La disponibilité caractérise donc l'aptitude du système à fonctionner quand on a besoin de lui (ex : avoir du réseau quand on veut téléphoner).
- La **maintenabilité** (*maintainability*) caractérise l'aptitude du système à être réparé quand il est défaillant, ou à évoluer.

- La **sécurité** (*safety*) caractérise l'aptitude du système à ne pas encourir de défaillances catastrophiques.

Un **système non réparable** est un système qui est mis au rebut dès qu'il tombe en panne. C'est le cas des petits systèmes (par exemple des ampoules) ou des systèmes qui coûtent plus cher à réparer qu'à remplacer.

Un **système réparable** est un système qui, après sa défaillance, peut être remis en état de marche par des actions de réparation ou maintenance. C'est le cas de quasiment tous les systèmes complexes (véhicules, usines, etc.). Dans la plupart des cas, un système réparable est constitué de composants non réparables. Quand un composant tombe en panne, on le remplace par un neuf, mais le système complet, lui, n'est pas remis à neuf.

Dans les systèmes réparables, on distingue en général les **systèmes matériels**, qui ont tendance à se dégrader à cause de l'usure, et les **systèmes logiciels**, qui ont tendance à s'améliorer du fait des corrections et mises à jour appliquées. De plus en plus de systèmes comportent à la fois une composante matérielle et une composante logiciele, dont les fiabilités peuvent évoluer en sens inverse.

La maintenance des systèmes est essentiellement de deux types :

- La **maintenance corrective (MC)** ou **réparation** est effectuée suite à une défaillance et a pour but de remettre le système en état de fonctionner.
- La **maintenance préventive (MP)** est effectuée alors que le système fonctionne et a pour but de ralentir le vieillissement pour retarder l'occurrence des défaillances futures.
  - La **MP planifiée** est effectuée à des instants prévus à l'avance.
  - La **MP conditionnelle** est effectuée suite à une surveillance du système si un état de dégradation avancé est détecté.

Par exemple, une voiture subit une maintenance corrective après un accident et une maintenance préventive planifiée lors des révisions.

Si les maintenances préventives sont prévues tardivement, on prend le risque que des défaillances surviennent avant. A l'inverse, faire des maintenances préventives très rapprochées permet de diminuer fortement le risque de défaillance mais cela est très coûteux. Il faut donc trouver un compromis coût-risque. L'**optimisation de la maintenance** est un enjeu industriel majeur. Par exemple, après le crash de l'Airbus d'Air France Rio-Paris en 2009, les sondes Pitot (qui mesurent la vitesse de l'avion par rapport à l'air) ont été incriminées. Les autorités de sûreté aérienne ont recommandé de remplacer 2 sondes sur 3 et de modifier la périodicité de leur maintenance.

La **maintenance prédictive** désigne l'ensemble de la démarche ayant pour objectif de déterminer les dates optimales pour effectuer les maintenances préventives. Si on arrive à prévoir avec précision quand aura lieu la défaillance, l'idéal serait de faire une maintenance préventive un peu avant. L'optimisation de la maintenance se base donc sur des études de fiabilité.

Dans les études de fiabilité, on distingue les approches boîte noire et boîte blanche :

- **Approche boîte blanche** ou **structurelle** : on considère qu'un système complexe est constitué de composants et que sa fiabilité dépend à la fois de la fiabilité de ses composants et de la façon dont le bon fonctionnement ou la panne de chaque composant influe sur le bon fonctionnement ou la panne du système tout entier.
- **Approche boîte noire** ou **globale** : on considère le système comme un tout, qu'on ne cherche pas à décomposer en composants. On s'intéresse alors à la suite des défaillances et réparations successives du système.

Il existe des méthodes pour construire des systèmes au fonctionnement sûr : tolérance aux fautes, redondance, prévention des défaillances, etc. Mais il faut aussi des méthodes pour évaluer la SdF des systèmes, prévoir les défaillances ultérieures, et vérifier que les objectifs de fiabilité ont bien été atteints.

1.3 Objectifs et plan du cours

Une étude de fiabilité classique comprend deux parties :

- La **modélisation probabiliste**. Construite à partir d'hypothèses sur le mécanisme du hasard engendrant les défaillances, elle permet de calculer toutes les caractéristiques de SdF des composants (qui vont être définies dans le chapitre suivant) : fiabilité, MTTF, taux de défaillance, etc. Elle permet également de calculer la fiabilité d'un système à partir de sa structure et de la fiabilité de ses composants.
- L'**analyse statistique**. Basée sur le retour d'expériences, c'est-à-dire l'observation des défaillances et réparations, elle permet de valider les hypothèses probabilistes retenues et d'estimer les valeurs numériques des caractéristiques de SdF, pour in fine prendre les décisions adéquates pour le système étudié (comme par exemple de changer une politique de maintenance).

Ce cours abordera les éléments essentiels de ces deux aspects. Dans une première partie probabiliste, nous commencerons par définir les principales mesures de fiabilité, pour les systèmes non réparables comme pour les systèmes réparables. Puis nous présenterons les lois de probabilités utilisées usuellement en fiabilité, principalement la loi exponentielle et la loi de Weibull. Enfin, nous expliquerons comment calculer la fiabilité d'un système en fonction de sa structure.

Dans une seconde partie statistique, nous étudierons comment évaluer la fiabilité d'un système à partir de l'observation de ses défaillances. Cela requiert des méthodes d'estimation et de tests statistiques, que l'on appliquera à des échantillons de lois exponentielle et de Weibull. Enfin, on abordera la problématique des données censurées, consistant à adapter les méthodes précédentes au cas où les données observées sont incomplètes.

Par exemple, le tableau 1.1 donne les durées de fonctionnement successives (en heures de vol) d'appareils d'air conditionné dans 13 Boeing 720. Un appareil défaillant est remplacé par un neuf. A l'aide de ces observations, on souhaite déterminer une loi de probabilité vraisemblable pour la durée de vie de ces appareils, ce qui permet-

tra d'estimer leur fiabilité. On pourra alors mettre en place un plan de maintenance approprié.

n°	Durées de fonctionnement
1	194, 15, 41, 29, 33, 181
2	413, 14, 58, 37, 100, 65, 9, 169, 447, 184, 36, 201, 118, 34, 31, 18, 18, 67, 57, 62, 7, 22, 34
3	90, 10, 60, 186, 61, 49, 14, 24, 56, 20, 79, 84, 44, 59, 29, 118, 25, 156, 310, 76, 26, 44, 23, 62, 130, 208, 70, 101, 208
4	74, 57, 48, 29, 502, 12, 70, 21, 29, 386, 59, 27, 153, 26, 326
5	55, 320, 56, 104, 220, 239, 47, 246, 176, 182, 33, 15, 104, 35
6	23, 261, 87, 7, 120, 14, 62, 47, 225, 71, 246, 21, 42, 20, 5, 12, 120, 11, 3, 14, 71, 11, 14, 11, 16, 90, 1, 16, 52, 95
7	97, 51, 11, 4, 141, 18, 142, 68, 77, 80, 1, 16, 106, 206, 82, 54, 31, 216, 46, 111, 39, 63, 18, 191, 18, 163, 24
8	50, 44, 102, 72, 22, 39, 3, 15, 197, 188, 79, 88, 46, 5, 5, 36, 22, 139, 210, 97, 30, 23, 13, 14
9	359, 9, 12, 270, 603, 3, 104, 2, 438
10	50, 254, 5, 283, 35, 12
11	130, 493
12	487, 18, 100, 7, 98, 5, 85, 91, 43, 230, 3, 130
13	102, 209, 14, 57, 54, 32, 67, 59, 134, 152, 27, 14, 230, 66, 61, 34

TABLE 1.1 – Durées de fonctionnement d'appareils d'air conditionné dans des Boeing 720

Chapitre 2

Les mesures de fiabilité

Les mesures de fiabilité sont différentes suivant que les systèmes concernés sont réparables ou non réparables.

2.1 Mesures pour les systèmes non réparables

Comme on l'a vu, un système non réparable est un système qui est mis au rebut dès qu'il tombe en panne. Les considérations sur les réparations ou corrections n'ont donc pas lieu d'être ici. Le seul point important est la **date de panne**, appelée aussi **instant de défaillance**, **durée de vie** ou **durée de bon fonctionnement** du système. Comme celle-ci n'est pas prévisible avec certitude à l'avance, on la modélise par une variable aléatoire, que l'on note X. Une durée étant un réel positif, cette variable aléatoire est à valeurs dans $\mathbb{R}^+$.

Si on s'intéressait à des systèmes pour lesquels le temps est exprimé par un nombre entier, comme un nombre de sollicitations, X serait à valeurs dans $\mathbb{N}$. On pourrait aussi prendre en compte la possibilité que le système soit en panne à l'instant initial, ce qui voudrait dire que $\mathbb{P}(X = 0) \neq 0$. Nous ne nous placerons ici dans aucun de ces deux cas. Par conséquent, X sera une variable aléatoire continue à valeurs dans $\mathbb{R}^+$. Sa loi de probabilité est définie par :

- sa **fonction de répartition** $F(x) = \mathbb{P}(X \leq x)$,
- sa **densité** $f(x) = F'(x)$.

Plus la durée de fonctionnement est grande, meilleure est la fiabilité du système. Donc on choisit de définir la fiabilité du système à l'instant x comme la probabilité que le système ne soit pas encore tombé en panne à l'instant x ou encore comme la probabilité que le système fonctionne sans défaillance entre 0 et x.

Définition 1 *La* **fiabilité** *d'un système non réparable est la fonction du temps R (R pour* **reliability***) définie par :*

$$\forall x \geq 0, \quad R(x) = \mathbb{P}(X > x) \tag{2.1}$$

On a évidemment $R(x) = 1 - F(x)$ et $R'(x) = -f(x)$. R est donc une fonction décroissante. Cela traduit le fait naturel que l'aptitude au bon fonctionnement d'un

système non réparable diminue avec le temps. Mais la monotonie de cette fonction fait que la fiabilité n'est pas suffisamment souple pour pouvoir clairement prendre en compte la diversité des types d'usure. Aussi la principale mesure de fiabilité n'est pas la fonction de fiabilité mais le taux de défaillance.

Définition 2 *Le* **taux de défaillance** *ou* **taux de panne** *ou* **taux de hasard** *d'un système non réparable est la fonction du temps h définie par :*

$$\forall x \geq 0, \quad h(x) = \lim_{\Delta x \to 0} \frac{1}{\Delta x} \mathbb{P}(x < X \leq x + \Delta x \mid X > x) \tag{2.2}$$

Dans cette expression, la probabilité considérée est la probabilité que le système tombe en panne entre x et $x + \Delta x$ sachant qu'il a bien fonctionné entre 0 et x. Notons que la fiabilité est une probabilité mais que le taux de défaillance n'en est pas une : $h(x)$ peut être supérieur à 1.

L'interprétation du taux de défaillance est liée à celle de la densité de la façon suivante. On sait que :

$$\begin{aligned} f(x) &= F'(x) = \lim_{\Delta x \to 0} \frac{F(x + \Delta x) - F(x)}{\Delta x} \\ &= \lim_{\Delta x \to 0} \frac{1}{\Delta x} \Big[\mathbb{P}(X \leq x + \Delta x) - \mathbb{P}(X \leq x) \Big] \\ &= \lim_{\Delta x \to 0} \frac{1}{\Delta x} \mathbb{P}(x < X \leq x + \Delta x) \end{aligned}$$

On a donc, pour Δx "petit" :

$$f(x)\, \Delta x \approx \mathbb{P}(x < X \leq x + \Delta x)$$

et :

$$h(x)\, \Delta x \approx \mathbb{P}(x < X \leq x + \Delta x \mid X > x)$$

La quantité $f(x)\, \Delta x$ peut donc être considérée comme la probabilité de défaillance juste après l'instant x alors que $h(x)\, \Delta x$ peut être considérée comme la probabilité de défaillance juste après l'instant x *sachant que* le système n'est pas tombé en panne avant x. Il y a donc une notion d'instantanéité dans $f(x)$ et une notion de durée dans $h(x)$ (comme dans $R(x)$).

On peut illustrer cette différence en comparant :

- la probabilité qu'un homme meure entre 100 et 101 ans ;
- la probabilité qu'un homme meure entre 100 et 101 ans sachant qu'il a vécu jusqu'à 100 ans.

La première (liée à la densité) est très faible : on a de très fortes chances de mourir avant 100 ans. La seconde (liée au taux de défaillance) est évidemment très forte, à cause du vieillissement.

On conçoit donc que le taux de défaillance est une mesure pratique de l'usure ou du vieillissement. Un taux de défaillance croissant correspond à un système qui se

dégrade, tandis qu'un taux de défaillance décroissant correspond à un système qui s'améliore avec le temps.

Il est facile d'établir les liens entre le taux de défaillance et la fiabilité :

$$\begin{aligned} h(x) &= \lim_{\Delta x \to 0} \frac{1}{\Delta x} \mathbb{P}(x < X \leq x + \Delta x \mid X > x) \\ &= \lim_{\Delta x \to 0} \frac{1}{\Delta x} \frac{\mathbb{P}(x < X \leq x + \Delta x \cap X > x)}{\mathbb{P}(X > x)} = \lim_{\Delta x \to 0} \frac{1}{\Delta x} \frac{\mathbb{P}(x < X \leq x + \Delta x)}{\mathbb{P}(X > x)} \\ &= \frac{1}{R(x)} \lim_{\Delta x \to 0} \frac{1}{\Delta x} \left[F(x + \Delta x) - F(x)\right] \\ &= \frac{f(x)}{R(x)} = \frac{f(x)}{1 - F(x)} = -\frac{R'(x)}{R(x)} = -\frac{d}{dx} \ln R(x) \end{aligned}$$

En intégrant et en prenant comme condition initiale $R(0) = 1$, car on a supposé que le système fonctionne à l'instant initial, on obtient la *formule d'exponentiation* :

$$R(x) = \exp\left(-\int_0^x h(u)\, du\right) \tag{2.3}$$

Puisque $f(x) = -R'(x)$, la densité de X s'exprime à l'aide du taux de défaillance sous la forme :

$$f(x) = h(x)\, \exp\left(-\int_0^x h(u)\, du\right) \tag{2.4}$$

Définition 3 *Le* **taux de défaillance cumulé** *ou* **taux de hasard cumulé** *d'un système non réparable est la fonction du temps H définie par :*

$$\forall x \geq 0, \quad H(x) = \int_0^x h(u)\, du = -\ln R(x) \tag{2.5}$$

La formule d'exponentiation s'écrit donc aussi $R(x) = \exp(-H(x))$.

Toutes les grandeurs caractéristiques de la loi de probabilité de X s'expriment à l'aide de la fonction h. Le taux de défaillance caractérise donc la loi d'une durée de vie. C'est pourquoi, en pratique, **construire un modèle de fiabilité de systèmes non réparables revient à se donner une forme particulière pour le taux de défaillance**.

Le choix de cette forme est basé sur des considérations de modélisation ou des constatations expérimentales. De nombreuses études pratiques ont montré que le graphe du taux de défaillance d'un système non réparable simple a très souvent une **forme de baignoire**, comme dans la figure 2.1. En effet, h se décompose dans ce cas en 3 parties :

- La **période de jeunesse** : quand un système est neuf, on observe souvent des défaillances précoces, dues à des défauts intrinsèques ou des fautes de conception. Le risque de défaillance est donc assez fort au tout début de la vie du système. Ensuite il diminue car, s'il y a des défauts initiaux, ils vont se manifester tôt. h est donc d'abord décroissant. C'est le **rodage** pour les matériels mécaniques et le **déverminage** pour les matériels électroniques. C'est aussi la **mortalité infantile** pour les êtres vivants.

- La **vie utile** : pendant cette période, le taux de défaillance est constant et les défaillances sont purement accidentelles.
- Le **vieillissement** : h se remet à croître car le risque de défaillance va finir par augmenter à cause de l'usure du système.

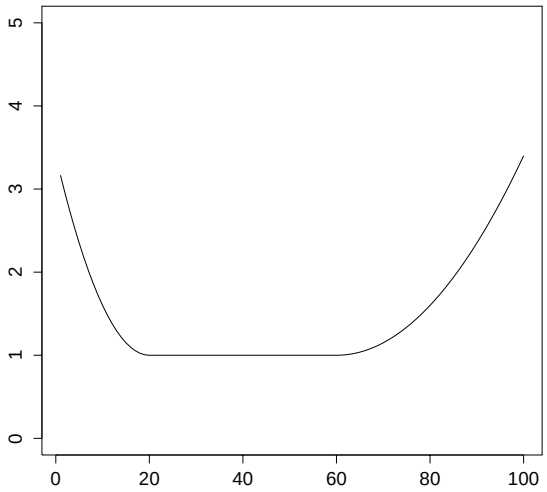


FIGURE 2.1 – Taux de défaillance en forme de baignoire

Du point de vue du consommateur cherchant à s'assurer contre les pannes du système, il est impératif d'avoir une garantie à court terme pour se prémunir contre les défauts de jeunesse. On peut souhaiter avoir une garantie à long terme contre le vieillissement, mais cela va coûter cher et les contrats ne garantissent en général pas les problèmes d'usure. En revanche, une garantie à moyen terme n'est pas forcément utile car, si le système a passé la période de jeunesse, il subira en général peu de défaillances en période de vie utile. Naturellement, pour pouvoir fixer de façon optimale les durées de garantie, il faut connaître ou estimer les dates de transition entre les différentes périodes, ce qui est généralement difficile.

La dernière mesure fondamentale de fiabilité est le MTTF.

Définition 4 *Le* **MTTF** *(***Mean Time To Failure***) d'un système non réparable est la durée de vie moyenne, ou durée moyenne de bon fonctionnement avant sa défaillance :*

$$\mathrm{MTTF} = \mathbb{E}[X] = \int_0^{+\infty} x\, f(x)\, dx \tag{2.6}$$

Une intégration par parties aboutit alors à :

$$\mathrm{MTTF} = \Big[-x\, R(x) \Big]_0^{+\infty} + \int_0^{+\infty} R(x)\, dx$$

En supposant que $R(x)$ tend vers 0 plus vite que $\frac{1}{x}$, ce qui sera toujours le cas, on obtient une formule plus usuelle pour le MTTF :

$$\mathrm{MTTF} = \int_0^{+\infty} R(x)\,dx \tag{2.7}$$

2.2 Mesures pour les systèmes réparables

Quand les systèmes sont réparables, deux cas de figure sont possibles, selon que l'on prend en compte ou pas les durées de réparation.

2.2.1 Durées de réparation comptabilisées

Le fonctionnement du système est une succession de **durées de bon fonctionnement** et de **durées de non fonctionnement** ou de **réparation**. On note traditionnellement $\{X_i\}_{i\geq 1}$ les durées de bon fonctionnement successives et $\{Y_i\}_{i\geq 1}$ les durées de réparation successives.

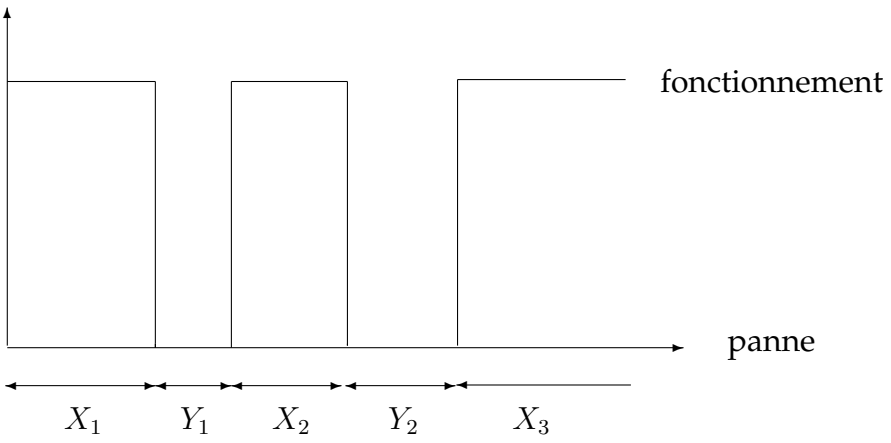


FIGURE 2.2 – Durées de bon fonctionnement et de réparation

La "durée de réparation" Y comprend en fait une durée de détection de la panne, une durée de réparation proprement dite et une durée de remise en service.

Pour une durée de réparation Y, on définit des quantités similaires à celles qui ont été définies pour une durée de bon fonctionnement X :

- La **maintenabilité** est la fonction de répartition de Y. La maintenabilité en y est la probabilité qu'un système en panne à l'instant 0 soit réparé avant l'instant y :

$$\forall y \geq 0, \quad M(y) = \mathbb{P}(Y \leq y)$$

- Le **taux de réparation** est défini par

$$\forall y \geq 0, \quad \mu(y) = \lim_{\Delta y \to 0} \frac{1}{\Delta y}\,\mathbb{P}(y < Y \leq y + \Delta y \mid Y > y)$$

- Le **MTTR** (**Mean Time To Repair**) est la durée moyenne de réparation :

$$\mathrm{MTTR} = \mathbb{E}[Y] = \int_0^{+\infty} [1 - M(y)]\, dy$$

Dans ce contexte, on peut toujours définir la fiabilité à l'instant x comme la probabilité que le système ne tombe pas en panne entre 0 et x. Cela signifie que l'instant de la première panne doit être supérieur à x, donc

$$\forall x \geq 0, \quad R(x) = \mathbb{P}(X_1 > x).$$

Mais on peut s'intéresser à une autre quantité particulièrement intéressante pour les systèmes réparables, la disponibilité.

Définition 5 *La* **disponibilité** *d'un système réparable est la fonction du temps A (A pour* **availability***) telle que :*

$$\forall t \geq 0, \quad A(t) = \textit{Probabilité que le système fonctionne à l'instant } t.$$

Donner une expression mathématique générale est beaucoup plus complexe pour la disponibilité que pour la fiabilité. Heureusement, il est souvent possible de donner des expressions simples de la **disponibilité asymptotique** :

$$A(\infty) = \lim_{t \to +\infty} A(t).$$

Remarque : La fiabilité implique une notion de durée (fonctionnement pendant une certaine durée) tandis que la disponibilité implique une notion d'instantanéité (fonctionnement à un instant donné).

Contrairement à ceux de $R(x)$ et $M(y)$, le sens de variation de $A(t)$ n'est pas déterminé. On a des systèmes à disponibilité croissante, d'autres à disponibilité décroissante et tous les sens de variation imaginables sont possibles.

Quand le système est remis à neuf après réparation, il est logique de supposer que X_2 a la même loi de probabilité que X_1. Plus généralement, on suppose couramment que les X_i sont indépendants et de même loi, et que les Y_i sont aussi indépendants et de même loi (mais pas la même que celle des X_i). Cela facilite grandement le calcul de la disponibilité. Mais dans la pratique, la réparation ne remet pas souvent le système à neuf, ce qui complexifie les calculs.

On parle parfois de MTBF = MTTF+MTTR (Mean Time Between Failures). Le MTBF est la durée moyenne entre deux défaillances successives, comprenant la durée moyenne de bon fonctionnement et la durée moyenne de réparation. On a alors souvent :

$$\lim_{t \to +\infty} A(t) = \frac{MTTF}{MTTF + MTTR},$$

ce qui se conçoit bien intuitivement.

2.2.2 Durées de réparation non comptabilisées

En pratique, il est fréquent que les durées de réparation soient négligeables par rapport aux durées de bon fonctionnement. Il est donc intéressant de modéliser la situation où les durées de réparation sont non comptabilisées. Dans ce cas, la notion de disponibilité n'a plus aucun sens.

Dans ces conditions, on considère que l'on observe le fonctionnement d'un système réparable à partir d'un instant $T_0 = 0$. Des défaillances se produisent à des instants que l'on note $T_1, T_2, \ldots$. Après chaque défaillance, le système est réparé ou corrigé puis relancé. Le **processus des défaillances** d'un tel système réparable est défini de manière équivalente par l'un des 3 processus aléatoires suivants.

- la suite des instants de défaillance $\{T_i\}_{i\geq 1}$, avec $T_0 = 0$.
- la suite des durées inter-défaillances $\{X_i\}_{i\geq 1}$ où $\forall i \geq 1, X_i = T_i - T_{i-1}$ est la durée entre la $(i-1)^{\text{ème}}$ et la $i^{\text{ème}}$ défaillance. $T_i = \sum_{j=1}^{i} X_j$.
- le processus de comptage des défaillances $\{N_t\}_{t\geq 0}$, où N_t est le nombre cumulé de défaillances survenues entre 0 et t.

La figure 2.3 illustre les quantités aléatoires ainsi définies en présentant une trajectoire quelconque du processus des défaillances.

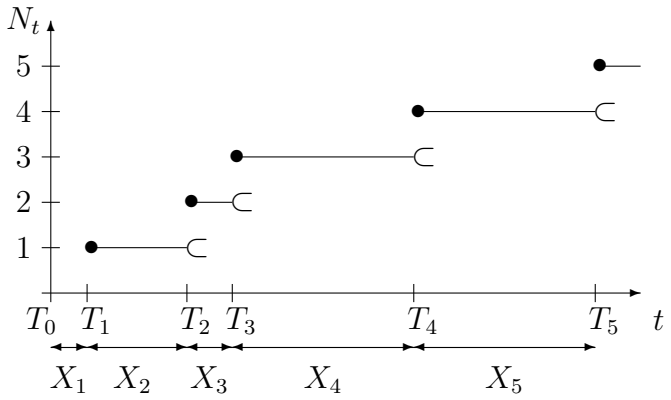


FIGURE 2.3 – Trajectoire quelconque du processus des défaillances

Si on définit la fiabilité comme précédemment, seul l'instant de la première défaillance est en jeu :

$$\forall x \geq 0, \quad R(x) = \mathbb{P}(X_1 > x) = \mathbb{P}(T_1 > x).$$

Or, pour un système réparable, on a évidemment envie de pouvoir calculer la probabilité que le système fonctionne correctement pendant une durée quelconque après n'importe quelle réparation. Aussi, on va modifier la définition de la fiabilité en considérant que la fiabilité du système à l'instant t exprime la probabilité qu'il fonctionne correctement pendant une certaine durée à partir de t. Pour être tout à fait complet,

on va considérer que cette probabilité peut éventuellement dépendre de tout ce qui s'est passé depuis la mise en service du système. Mathématiquement, cela signifie que c'est une probabilité conditionnelle au nombre et aux instants des défaillances ayant précédé l'instant présent t. D'où la définition suivante.

Définition 6 *La* **fiabilité** *d'un système réparable à l'instant t, ayant subi n défaillances avant t, est la fonction R_t définie par :*

$$\begin{aligned} \forall \tau \geq 0, \quad R_t(\tau; n, t_1, \ldots, t_n) &= \mathbb{P}(T_{n+1} > t + \tau | N_t = n, T_1 = t_1, \ldots, T_n = t_n) \\ &= \mathbb{P}(N_{t+\tau} - N_t = 0 | N_t = n, T_1 = t_1, \ldots, T_n = t_n) \end{aligned} \quad (2.8)$$

Autrement dit, $R_t(\tau; n, t_1, \ldots, t_n)$ est la probabilité que le système fonctionne sans défaillances pendant une durée au moins égale à τ après t, sachant qu'il y a eu exactement n défaillances entre 0 et t, aux instants $t_1, \ldots, t_n$. La première écriture exprime que la prochaine défaillance aura lieu après $t + \tau$ et la seconde exprime qu'il n'y aura aucune défaillance entre t et $t + \tau$. On conçoit bien la définition de la fiabilité à l'aide de la figure 2.4, pour laquelle n défaillances ont eu lieu entre 0 et t.

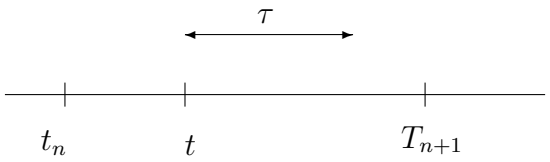


FIGURE 2.4 – Fiabilité pour une durée τ à partir de t, avec n défaillances observées

Quand on se place à l'instant t_n de la dernière défaillance, on est intéressé par la prévision de la durée X_{n+1} à attendre avant la prochaine défaillance. Sa loi de probabilité peut être influencée par le passé du processus de défaillance, donc on s'intéressera plutôt à la loi de X_{n+1} sachant $[T_1 = t_1, \ldots, T_n = t_n]$. Cette loi a pour taux de défaillance $h_{X_{n+1}|T_1=t_1,\ldots,T_n=t_n}(x)$.

Pour un système non réparable, la propension de défaillance à l'instant t est exprimée par le taux de défaillance $h(t)$. Pour un système réparable, il est logique d'exprimer la propension de défaillance à l'instant t par le taux de défaillance de la durée inter-défaillance courante à cet instant, autrement dit $h_{X_{n+1}|T_1=t_1,\ldots,T_n=t_n}(t - t_n)$. C'est ce qu'on appelle l'intensité de défaillance à l'instant t.

Définition 7 *L'***intensité de défaillance** *d'un système réparable à l'instant t est la fonction λ_t définie par :*

$$\lambda_t(n; t_1, \ldots, t_n) = h_{X_{n+1}|T_1=t_1,\ldots,T_n=t_n}(t - t_n) \quad (2.9)$$

Suivre l'intensité de défaillance au cours du temps revient donc à étudier les taux de défaillance conditionnels successifs des X_i sachant le passé. On parle alors de concaténation ou d'amalgame de taux de défaillance.

On montre que l'intensité de défaillance s'écrit aussi :

$$\begin{aligned}\lambda_t(n;t_1,\ldots,t_n) &= \lim_{\Delta t\to 0}\frac{1}{\Delta t}\,\mathbb{P}(t < T_{n+1} \leq t+\Delta t \mid N_t = n, T_1 = t_1,\ldots,T_n = t_n)\\ &= \lim_{\Delta t\to 0}\frac{1}{\Delta t}\,\mathbb{P}(N_{t+\Delta t} - N_t = 1 \mid N_t = n, T_1 = t_1,\ldots,T_n = t_n) \quad (2.10)\end{aligned}$$

La probabilité dans cette écriture est la probabilité que le système tombe en panne entre t et $t+\Delta t$ sachant tout le passé du processus des défaillances à l'instant t.

L'intensité de défaillance est aux systèmes réparables ce que le taux de défaillance est aux systèmes non réparables. Une intensité de défaillance croissante correspond à une fréquence de défaillance qui augmente avec le temps, donc à un système qui s'use malgré les réparations. Une intensité de défaillance décroissante correspond à un système qui s'améliore avec le temps. A priori, les matériels rentrent dans la première catégorie et les logiciels dans la seconde.

Définition 8 *Le* **MTTF** *d'un système réparable à l'instant t est la durée moyenne d'attente de la prochaine défaillance à l'instant t, sachant tout le passé du processus des défaillances à cet instant :*

$$\mathrm{MTTF}_t(n;t_1,\ldots,t_n) = \mathbb{E}\left[T_{n+1} - t \mid N_t = n, T_1 = t_1,\ldots,T_n = t_n\right] \quad (2.11)$$

Les résultats suivants sont les équivalents pour les systèmes réparables des formules (2.3) et (2.7).

$$R_t(\tau;n,t_1,\ldots,t_n) = \exp\left(-\int_t^{t+\tau}\lambda_u(n;t_1,\ldots,t_n)\,du\right) = \exp\left(-\int_0^{\tau}\lambda_{t+u}(n;t_1,\ldots,t_n)\,du\right) \quad (2.12)$$

$$\mathrm{MTTF}_t(n;t_1,\ldots,t_n) = \int_0^{+\infty} R_t(\tau;n,t_1,\ldots,t_n)\,d\tau \quad (2.13)$$

Une autre mesure importante de fiabilité des systèmes réparables est le nombre moyen de défaillances survenues à chaque instant. C'est ce qu'on appelle la fonction moyenne.

Définition 9 *La* **fonction moyenne** *(en anglais* mean value function*) du processus des défaillances est la fonction m définie par :*

$$\forall t \geq 0, \quad m(t) = \mathbb{E}[N_t] \quad (2.14)$$

Pour le cas où on ne prend pas en compte les durées de réparation, toutes les mesures de fiabilité des systèmes réparables s'expriment à l'aide de l'intensité de défaillance. Par conséquent, **construire un modèle de fiabilité des systèmes réparables revient à proposer une forme particulière pour l'intensité de défaillance**.

Les 3 classes de modèles les plus simples sont caractérisées par les formes d'intensité suivantes :

Processus de Poisson homogènes (HPP)	$\lambda_t(n; t_1, \ldots, t_n) = \lambda$
Modèles à durées inter-défaillances exponentielles (ETBF)	$\lambda_t(n; t_1, \ldots, t_n) = \lambda_{n+1}$
Processus de Poisson non homogènes (NHPP)	$\lambda_t(n; t_1, \ldots, t_n) = \lambda(t)$

2.3 Evaluation des mesures de fiabilité

On peut maintenant préciser le contenu des deux parties d'une étude classique de fiabilité, citées au chapitre 1. Pour évaluer la fiabilité, le taux de défaillance, la disponibilité ou l'intensité de défaillance d'un système, il faut passer par 2 étapes.

1. **Modélisation probabiliste**. A partir d'hypothèses sur le fonctionnement du système, l'effet des réparations et la nature du phénomène aléatoire ayant engendré les défaillances, il faut proposer des modèles réalistes pour les variables aléatoires impliquées. Par exemple, il faut proposer une loi de probabilité vraisemblable pour la durée de bon fonctionnement X d'un système non réparable ou pour la suite $\{X_i\}_{i\geq 1}$ des durées inter-défaillances d'un système réparable.

2. **Analyse statistique**. Les modèles proposés ont des paramètres inconnus qu'il va falloir estimer. Pour cela, il faut observer le fonctionnement des systèmes, relever les instants des défaillances et des réparations, et effectuer une analyse statistique de ces données (qu'on appelle le **retour d'expériences**). On va pouvoir ainsi estimer les caractéristiques de fiabilité des systèmes et utiliser ces estimations pour prendre des décisions qui peuvent être cruciales en termes de sûreté de fonctionnement. Par exemple, il faut décider si un produit est suffisamment fiable pour que l'on puisse le mettre en vente sans risque. Ou bien, on mettra en place un plan de maintenance ayant pour but de prolonger la durée de vie des systèmes, à un coût raisonnable. Pour effectuer les estimations, plusieurs méthodes sont possibles, mais on utilise la plupart du temps la méthode du **maximum de vraisemblance**.

Par exemple, dans le cas des Boeing, un appareil défaillant est remplacé par un neuf. On s'attend donc à ce que les durées de vie dans un Boeing donné soient des variables aléatoires indépendantes et de même loi. Il faut le vérifier, trouver une loi de probabilité adaptée (exponentielle, Weibull ou autre), et estimer ses paramètres. Il est intéressant également de déterminer si cette loi de probabilité est la même pour les 13 appareils. On doit estimer la fiabilité, le taux de défaillance et le MTTF de ces appareils. Au final, les estimations effectuées doivent permettre de prendre la décision de passer ou non à une nouvelle gamme d'appareils. On peut aussi décider de réparer les appareils au lieu de les remplacer par des neufs.

Chapitre 3

Les lois de probabilité usuelles en fiabilité

On a dit que la fiabilité d'un système non réparable est caractérisée par la loi de probabilité de sa durée de bon fonctionnement X. Quand on s'intéresse, dans une approche boîte blanche, à un système complexe constitué de composants interconnectés, on va chercher à déterminer la loi de X en fonction des lois des durées de bon fonctionnement des composants élémentaires. Dans une approche boîte noire, on s'intéressera directement à la loi de X, sans chercher à décomposer le système en composants.

Il est donc capital d'avoir des modèles de base pour les lois des durées de bon fonctionnement de systèmes non réparables simples. Dans ce chapitre, on présente les plus utilisées de ces lois, essentiellement la loi exponentielle et la loi de Weibull. Pour chaque loi, on donnera, quand c'est possible, l'expression de la fiabilité, du taux de défaillance et du MTTF. Dans tout ce chapitre, on supposera que les durées de bon fonctionnement sont à valeurs dans $\mathbb{R}^+$, donc x sera implicitement supposé être un réel positif.

3.1 La loi exponentielle $\exp(\lambda)$

Une variable aléatoire X est de loi exponentielle de paramètre $\lambda > 0$, notée $\exp(\lambda)$, si et seulement si sa fonction de répartition est :

$$F(x) = 1 - \exp(-\lambda x)$$

- La fiabilité est $R(x) = 1 - F(x)$ d'où :

$$R(x) = \exp(-\lambda x) \tag{3.1}$$

- La densité est $f(x) = F'(x) = \lambda \exp(-\lambda x)$.
- La durée de vie moyenne est :

$$\mathrm{MTTF} = \mathbb{E}[X] = \int_0^{+\infty} R(x)\,dx = \int_0^{+\infty} \exp(-\lambda x)\,dx = \frac{1}{\lambda} \tag{3.2}$$

- La variance de X est $Var[X] = 1/\lambda^2$.
- Le taux de défaillance est :

$$h(x) = \frac{f(x)}{R(x)} = \frac{\lambda \exp(-\lambda x)}{\exp(-\lambda x)} = \lambda \tag{3.3}$$

Le taux de défaillance est donc constant, ce qui signifie que la loi exponentielle modélise les durées de vie de systèmes qui ne s'usent pas et qui ne s'améliorent pas.

On dit aussi que la loi exponentielle est **sans mémoire**, ce qu'on exprime de la façon suivante : si le système n'est pas encore tombé en panne à l'instant t, c'est comme s'il était neuf à cet instant. Mathématiquement, cela s'écrit :

$$\forall t \geq 0, \quad \forall x \geq 0, \quad \mathbb{P}(X > t + x \mid X > t) = \mathbb{P}(X > x).$$

On a :

$$\mathbb{P}(X > t + x \mid X > t) = \frac{\mathbb{P}(X > t + x \cap X > t)}{\mathbb{P}(X > t)} = \frac{\mathbb{P}(X > t + x)}{\mathbb{P}(X > t)} = \frac{R(t + x)}{R(t)}$$

Si X est de loi $\exp(\lambda)$, alors :

$$\mathbb{P}(X > t + x \mid X > t) = \frac{\exp(-\lambda(t + x))}{\exp(-\lambda t)} = \exp(-\lambda x) = \mathbb{P}(X > x)$$

donc la loi exponentielle est sans mémoire.

Réciproquement, si X est une variable aléatoire sans mémoire, alors on a :

$$\mathbb{P}(X > t + x \mid X > t) = \frac{R(t + x)}{R(t)} = \mathbb{P}(X > x) = R(x)$$

Donc la propriété d'absence de mémoire implique que $R(t + x) = R(t)R(x)$ pour tout $x \geq 0$ et $t \geq 0$. On en déduit que pour tout $x \geq 0$:

$$\begin{aligned} R'(x) &= \lim_{\Delta x \to 0} \frac{R(x + \Delta x) - R(x)}{\Delta x} = \lim_{\Delta x \to 0} \frac{R(x)R(\Delta x) - R(x)}{\Delta x} \\ &= R(x) \lim_{\Delta x \to 0} \frac{R(\Delta x) - 1}{\Delta x} = R(x)\, R'(0) \end{aligned}$$

Comme R est décroissante, $R'(0)$ est une constante strictement négative que l'on note $-\lambda$, avec $\lambda > 0$. Ainsi, R est l'unique solution de l'équation différentielle $R'(x) = -\lambda R(x)$ avec comme condition initiale $R(0) = 1$. Autrement dit, on a $R(x) = \exp(-\lambda x)$. Ainsi la seule loi de probabilité à densité continue vérifiant la propriété d'absence de mémoire est la loi exponentielle. Dans le contexte de la fiabilité, cette absence de mémoire s'interprète comme une absence de vieillissement et une absence de rajeunissement.

Dans la pratique, on dit souvent que l'on peut modéliser par une loi exponentielle la durée de vie de systèmes qui sont dans leur période de vie utile, c'est-à-dire qui, dans

la courbe en baignoire, ont dépassé la période de jeunesse et ne sont pas encore entrés en période d'usure. Mais c'est une erreur méthodologique car la loi de probabilité de X doit pouvoir modéliser l'ensemble de la durée de vie du système.

Par conséquent, la loi exponentielle ne devrait être utilisée que pour des systèmes qui ne s'usent pas et ne s'améliorent pas. Or tous les systèmes matériels sont soumis à l'usure, donc leur durée de vie devrait avoir a priori un taux de défaillance croissant, au moins en fin de vie. Par contre, un logiciel ne s'use pas. Tant qu'il n'est pas modifié, sa propension à subir une défaillance reste constante. Aussi la loi exponentielle a-t-elle un rôle prépondérant en fiabilité des logiciels. Par ailleurs, l'expérience montre qu'un taux de défaillance constant est une hypothèse réaliste pour les composants électroniques après déverminage.

Remarque 1 : Le MTTF est exprimé par une unité de temps, par exemple l'heure. La relation $\mathrm{MTTF} = 1/\lambda$ implique donc qu'en pratique, on donne pour unité de λ l'inverse d'une unité de temps. Cela explique qu'un objectif de fiabilité soit souvent exprimé en terme de taux de panne "par heure".

Remarque 2 : Si l'on admet que la durée de vie d'un système est de loi exponentielle, toute maintenance préventive est inutile puisque le système est comme neuf à chaque instant tant qu'il n'est pas tombé en panne.

Si la loi exponentielle est de loin la loi de durée de vie la plus utilisée en raison de sa simplicité, elle ne permet de modéliser ni l'usure, ni le rajeunissement. Il est donc nécessaire de disposer de lois plus sophistiquées. En fiabilité, la loi de Weibull est la plus populaire d'entre elles.

3.2 La loi de Weibull $\mathcal{W}(\eta, \beta)$

Une variable aléatoire X est de loi de Weibull de paramètre d'échelle $\eta > 0$ et de paramètre de forme $\beta > 0$, notée $\mathcal{W}(\eta, \beta)$, si et seulement si sa fonction de répartition est :

$$F(x) = 1 - \exp\left(- \left(\frac{x}{\eta}\right)^{\beta} \right)$$

- La fiabilité est :

$$R(x) = \exp\left(- \left(\frac{x}{\eta}\right)^{\beta} \right) \tag{3.4}$$

- La densité est :

$$f(x) = F'(x) = \frac{\beta}{\eta} \left(\frac{x}{\eta}\right)^{\beta-1} \exp\left(- \left(\frac{x}{\eta}\right)^{\beta} \right)$$

- La durée de vie moyenne est $\mathrm{MTTF} = \int_0^{+\infty} \exp\left(-(x/\eta)^{\beta}\right)\, dx$. Un changement de variables $u = (x/\eta)^{\beta}$ permet d'obtenir :

$$\mathrm{MTTF} = \eta\, \Gamma\left(\frac{1}{\beta} + 1\right) \tag{3.5}$$

où Γ est la fonction gamma d'Euler définie par :

$$\Gamma(a) = \int_0^{+\infty} x^{a-1} \exp(-x)\, dx \tag{3.6}$$

En particulier, $\Gamma(n) = (n-1)!$ pour tout $n \in \mathbb{N}^*$.

- La variance de X est :

$$Var[X] = \eta^2 \left[\Gamma\left(\frac{2}{\beta} + 1\right) - \Gamma\left(\frac{1}{\beta} + 1\right)^2 \right]$$

- Le taux de défaillance est :

$$h(x) = \frac{f(x)}{R(x)} = \frac{\beta}{\eta} \left(\frac{x}{\eta}\right)^{\beta-1} \tag{3.7}$$

Le taux de défaillance de la loi de Weibull est donc une puissance du temps, ce qui permet de modéliser de nombreuses situations. En particulier :

- si $\beta < 1$, h est décroissant donc le système s'améliore ;
- si $\beta > 1$, h est croissant donc le système s'use ;
- si $\beta = 1$, h est constant et on retrouve la loi exponentielle comme cas particulier de la loi de Weibull.

La figure 3.1 donne les graphes des taux de défaillance de la loi de Weibull pour $\beta \in \{0.5, 1, 1.5, 3\}$.

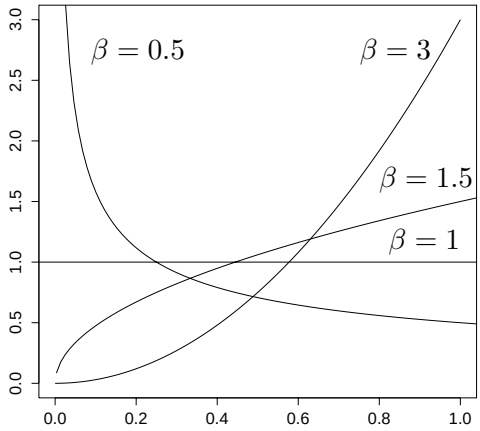


FIGURE 3.1 – Taux de défaillance de la loi de Weibull

Remarquons que pour $\beta \in]1, 2[$, h est concave, donc le système s'use, mais de moins en moins vite. L'interprétation de ce type d'usure est difficile et fait l'objet de controverses. Pour $\beta > 2$, h est convexe, ce qui correspond à une accélération de l'usure. Cela se conçoit plus facilement.

On dit parfois que la loi de Weibull permet de modéliser la période de jeunesse (pour $\beta < 1$), la vie utile (pour $\beta = 1$) et la période de vieillissement (pour $\beta > 1$). Là

encore, c'est une erreur méthodologique car on doit représenter l'ensemble de la durée de vie par une seule loi de probabilité.

Une propriété remarquable de la loi de Weibull est que c'est l'une des lois des valeurs extrêmes : pour n variables aléatoires $X_1, \ldots, X_n$ indépendantes et de même loi, la loi limite de variables aléatoires s'écrivant $a_n \min_{i=1}^{n} X_i + b_n$ quand on fait tendre n vers l'infini, est de trois types possibles. La seule dont le support soit $\mathbb{R}_+$ est la loi de Weibull. Concrètement, cela signifie que la loi de Weibull est un modèle naturel pour des systèmes constitués d'un très grand nombre de composants et dont la panne survient dès qu'un composant est défaillant (système série, voir chapitre 4).

3.3 Autres lois usuelles

3.3.1 La loi gamma $\mathcal{G}(\alpha, \lambda)$

X est de loi gamma de paramètre de forme $\alpha > 0$ et de paramètre d'échelle $\lambda > 0$, notée $\mathcal{G}(\alpha, \lambda)$, si et seulement si sa densité est :

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} \exp(-\lambda x)\ x^{\alpha-1}$$

où Γ est la fonction gamma définie en (3.6).

La fonction de répartition de la loi gamma n'a pas d'expression explicite, donc la fiabilité et le taux de défaillance non plus. En revanche, on dispose du MTTF et d'éléments qualitatifs sur le taux de défaillance :

- La durée de vie moyenne est : $\mathrm{MTTF} = \dfrac{\alpha}{\lambda}$.
- La variance de X est : $Var[X] = \dfrac{\alpha}{\lambda^2}$
- On peut montrer que :
  — si $\alpha < 1$, h est décroissant donc le système s'améliore ;
  — si $\alpha > 1$, h est croissant donc le système s'use ;
  — si $\alpha = 1$, h est constant et on retrouve la loi exponentielle.

  Ces 3 cas sont représentés dans la figure 3.2.

Pour n entier, $\mathcal{G}\left(\dfrac{n}{2}, \dfrac{1}{2}\right)$ est la **loi du chi-2** à n degrés de liberté, notée χ^2_n.

Proposition 1 *Si $X_1, \ldots, X_n$ sont indépendantes et de même loi $\exp(\lambda)$, alors $\sum_{i=1}^{n} X_i$ est de loi gamma $\mathcal{G}(n, \lambda)$.*

Proposition 2 *Si X est de loi $\mathcal{G}(\alpha, \lambda)$ et a est un réel strictement positif, alors aX est de loi $\mathcal{G}(\alpha, \lambda/a)$.*

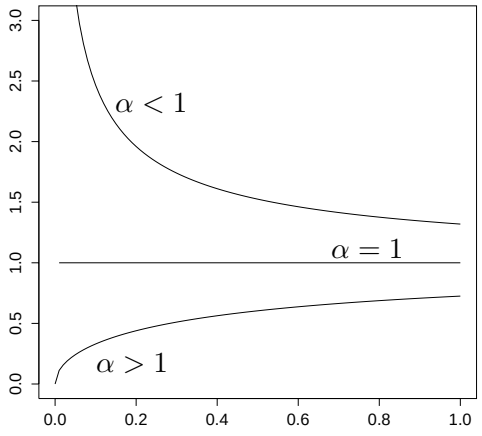


FIGURE 3.2 – Taux de défaillance de la loi gamma

3.3.2 La loi lognormale $\mathcal{LN}(m, \sigma^2)$

X est de loi lognormale de paramètres $m \in \mathbb{R}$ et $\sigma^2 > 0$, notée $\mathcal{LN}(m, \sigma^2)$, si et seulement si $\ln X$ est de loi normale $\mathcal{N}(m, \sigma^2)$.

- La densité est :
$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(\ln x - m)^2\right)$$

- Le MTTF vaut : $\mathrm{MTTF} = \exp\left(m + \frac{\sigma^2}{2}\right)$.
- La variance de X est : $Var[X] = \exp(2m)\ (\exp(2\sigma^2) - \exp(\sigma^2))$.

Là encore, la fonction de répartition, la fiabilité et le taux de défaillance de la loi lognormale n'ont pas d'expression explicite. En revanche, on peut vérifier que le taux de défaillance croît puis décroît en tendant vers 0 (voir la figure 3.3). Ceci peut modéliser des situations réelles : un système qui se détériore puis se met à s'améliorer au bout d'un moment. En fait l'expérience montre que la loi lognormale est plus à même de modéliser des durées de réparation que des durées de bon fonctionnement.

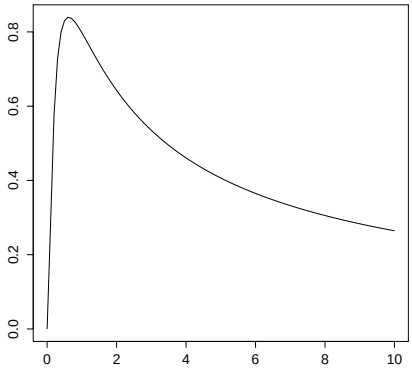


FIGURE 3.3 – Taux de défaillance de la loi lognormale

3.3.3 Lois avec taux de défaillance en baignoire

Il est étonnant de constater que, bien qu'il soit communément admis qu'en pratique le taux de défaillance d'un système non réparable a souvent une forme de baignoire, il existe peu de lois de probabilité de durées de vie possédant cette propriété. Par exemple, aucune des lois citées jusqu'à maintenant ne rentre dans ce cadre. La façon la plus simple de construire un taux de défaillance en baignoire est de "raccorder" trois taux de type Weibull respectivement décroissant, constant et croissant, en faisant en sorte que le taux résultant soit continu et à dérivée continue. Par exemple, la figure 2.1 a été obtenue à partir d'un taux de la forme :

$$h(x) = \begin{cases} \lambda + \dfrac{\beta_1}{\eta_1}\left(\dfrac{\tau_1 - x}{\eta_1}\right)^{\beta_1 - 1} & \text{si} \quad x \in [0, \tau_1[\\ \lambda & \text{si} \quad x \in [\tau_1, \tau_2] \\ \lambda + \dfrac{\beta_2}{\eta_2}\left(\dfrac{x - \tau_2}{\eta_2}\right)^{\beta_2 - 1} & \text{si} \quad x \in]\tau_2, +\infty[\end{cases}$$

Dans cette expression, la période de vie utile est l'intervalle $[\tau_1, \tau_2]$. D'autres lois de probabilité possèdent des taux de défaillance dont la forme se rapproche d'une baignoire, sans avoir pour autant une période de vie utile aussi bien délimitée. Notons par exemple :

$$\begin{aligned} h(x) &= \alpha\beta(\alpha x)^{\beta-1} + \frac{\alpha}{\beta}(\alpha x)^{1/\beta - 1} \\ h(x) &= \alpha(\beta + \lambda x)x^{\beta-1}\exp(\lambda x) \end{aligned}$$

La figure 3.4 donne les graphes des 3 taux de défaillance ci-dessus.

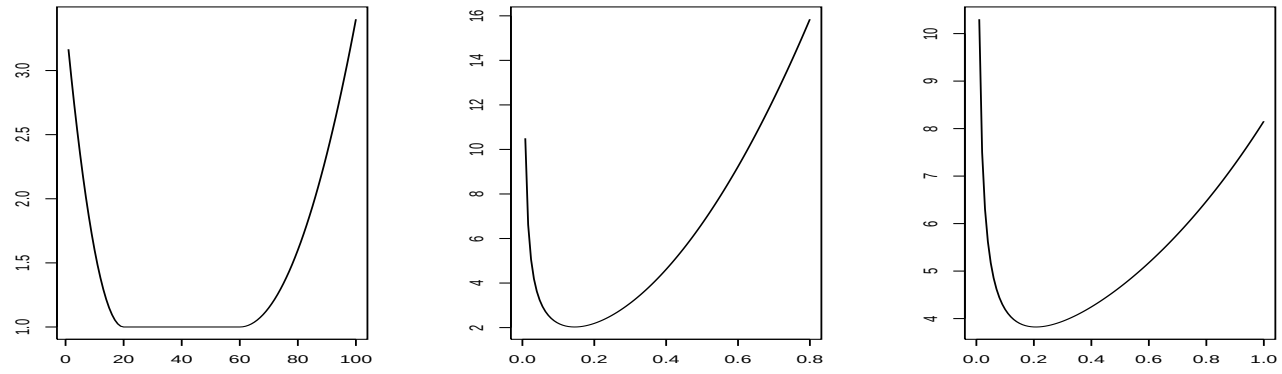


FIGURE 3.4 – Taux de défaillance en baignoire

Remarque : Quand on assemble des composants ayant chacun un taux de défaillance constant ou de type Weibull, il est possible que le système résultant ait un taux de défaillance en baignoire.

Chapitre 4

Calculs de fiabilité par structure

4.1 Principes

Le principe des calculs de fiabilité par structure (ou architecture) est de considérer qu'un système est constitué de composants élémentaires, et que sa fiabilité dépend à la fois de la fiabilité de ses composants et de la façon dont le bon fonctionnement ou la panne de chaque composant influe sur le bon fonctionnement ou la panne du système tout entier. Il est donc nécessaire de représenter la **logique de fonctionnement** du système.

Plusieurs types de représentations sont possibles : diagrammes de fiabilité, arbres de défaillance, graphes de Markov, réseaux de Petri, diagrammes de décision binaires, réseaux bayésiens, etc... On ne s'intéressera ici qu'à des systèmes non réparables et on représentera leur fonctionnement par un diagramme de fiabilité.

Le **diagramme de fiabilité** d'un système est un graphe sans circuit admettant une entrée E et une sortie S, dont :

- les sommets, appelés **blocs**, représentent les composants du système,
- les arcs traduisent les relations entre les différents composants, au sens où le système fonctionne si et seulement si il existe un chemin allant de E à S qui ne passe que par des composants en fonctionnement.

On peut faire l'analogie avec un réseau de distribution d'eau : l'eau n'est pas coupée tant qu'il existe un chemin dans le réseau qui lui permet d'aller de son point d'entrée à son point de sortie.

Remarque : le diagramme de fiabilité est une représentation *logique* du fonctionnement du système, qui n'a rien à voir avec une représentation *physique* des liaisons entre les différents composants. De même, il n'y a aucune contrainte de précédence dans ces diagrammes.

Exemple : une chaîne hi-fi comprend une platine CD (1), un tuner FM (2), un amplificateur (3) et deux enceintes (4 et 5). Le fonctionnement normal de la chaîne implique que tous ces éléments fonctionnent. Le diagramme de fiabilité est alors donné dans la figure 4.1. En effet, si un seul de ces éléments ne fonctionne pas, la chaîne ne fonctionne

pas correctement.

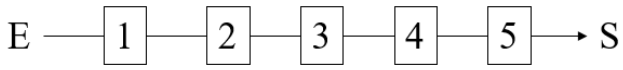


FIGURE 4.1 – Diagramme de fiabilité de la chaîne hi-fi en fonctionnement normal

Mais on peut admettre un fonctionnement dégradé dans lequel il est suffisant d'entendre au moins une des deux sources sonores sur au moins une des deux enceintes. Le diagramme de fiabilité est alors donné dans la figure 4.2.

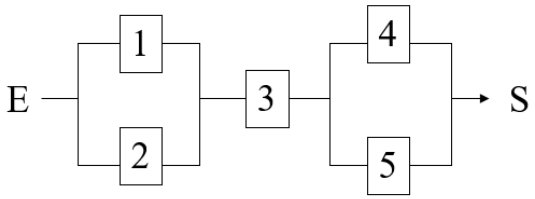


FIGURE 4.2 – Diagramme de fiabilité de la chaîne hi-fi en fonctionnement dégradé

Si on note $X_1, \ldots, X_5$ les durées de bon fonctionnement des 5 composants, il est facile de voir que, dans le premier cas, la durée de bon fonctionnement du système est $X = \min_{i=1}^{5} X_i$. Dans le deuxième cas, c'est moins évident mais on obtient que $X = \min\left(\max(X_1, X_2), X_3, \max(X_4, X_5)\right)$.

Quand le nombre de composants augmente, la structure du système peut se complexifier. Dans ce chapitre, nous allons étudier les structures de base les plus simples et donner une méthode permettant de calculer la fiabilité d'un système pour une structure complexe quelconque. Les critères de fiabilité sont un élément à prendre en compte dans le choix d'une architecture pour un système complexe.

Dans la suite, on considèrera des systèmes à n composants. Sauf mention contraire, les fonctionnements des n composants seront supposés indépendants. Pour le composant i, on note :

- X_i sa durée de bon fonctionnement,
- $r_i(x) = \mathbb{P}(X_i > x)$ sa fiabilité,
- $h_i(x)$ son taux de défaillance. $r_i(x) = \exp\left(-\int_0^x h_i(u)\,du\right)$.

Pour le système, on note X sa durée de bon fonctionnement, $R(x)$ sa fiabilité et $h(x)$ son taux de défaillance.

4.2 Systèmes série

Définition 10 *Un* **système série** *est un système qui ne fonctionne que si tous ses composants fonctionnent.*

C'est le cas de la chaine hi-fi en fonctionnement normal. Le diagramme de fiabilité est similaire à celui de la figure 4.1, avec n composants au lieu de 5. Il est donné dans la figure 4.3.

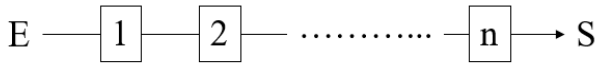


FIGURE 4.3 – Diagramme de fiabilité pour un système série

Un système série tombe en panne dès qu'un de ses composants tombe en panne. On a donc :

$$X = \min_{i=1}^{n} X_i$$

La fiabilité du système est alors :

$$R(x) = \mathbb{P}(X > x) = \mathbb{P}(\min_{i=1}^{n} X_i > x) = \mathbb{P}(\forall i, X_i > x) = \mathbb{P}\left(\bigcap_{i=1}^{n} [X_i > x]\right)$$

Comme on a supposé les composants indépendants, la probabilité ci-dessus est la probabilité d'une intersection d'évènements indépendants. Elle est donc égale au produit des probabilités de ces évènements :

$$R(x) = \prod_{i=1}^{n} \mathbb{P}(X_i > x) = \prod_{i=1}^{n} r_i(x)$$

On a donc :

$$R(x) = \prod_{i=1}^{n} \exp\left(-\int_0^x h_i(u)\,du\right) = \exp\left(-\sum_{i=1}^{n}\int_0^x h_i(u)\,du\right) = \exp\left(-\int_0^x \sum_{i=1}^{n} h_i(u)\,du\right)$$

Et comme $R(x) = \exp\left(-\int_0^x h(u)\,du\right)$, on en déduit que :

$$h(x) = \sum_{i=1}^{n} h_i(x)$$

Autrement dit, le taux de défaillance d'un système série à composants indépendants est égal à la somme des taux de défaillance de ses composants.

Il n'y a pas de résultat simple pour le MTTF :

$$MTTF = \int_0^{+\infty} R(x)\,dx = \int_0^{+\infty} \prod_{i=1}^{n} r_i(x)\,dx = \int_0^{+\infty} \exp\left(-\int_0^x \sum_{i=1}^{n} h_i(u)\,du\right)\,dx$$

Si tous les composants ont un taux de défaillance constant, $\forall i, \forall x, h_i(x) = \lambda_i$, donc X_i est de loi $\exp(\lambda_i)$ et $r_i(x) = \exp(-\lambda_i x)$. Alors $R(x) = \prod_{i=1}^{n} \exp(-\lambda_i x) = \exp\left(-[\sum_{i=1}^{n} \lambda_i]x\right)$ et $h(x) = \sum_{i=1}^{n} \lambda_i$ est encore constant.

On met donc là en évidence une propriété remarquable de la loi exponentielle : si $X_1, \ldots, X_n$ sont indépendantes et de lois respectives $\exp(\lambda_i)$, alors $X = \min_{i=1}^{n} X_i$ est de loi $\exp(\sum_{i=1}^{n} \lambda_i)$. Dans ce cas, on a un résultat simple pour le MTTF :

$$MTTF = \frac{1}{\sum_{i=1}^{n} \lambda_i}$$

De même, un système série constitué de composants indépendants et de durées de vie de lois de Weibull avec le même paramètre β a une durée de vie qui est encore de loi de Weibull. Enfin, on a aussi vu que la durée de vie d'un système série dont le nombre de composants tend vers l'infini a une loi qui tend vers une loi de Weibull.

4.3 Systèmes parallèles

4.3.1 Définition et propriétés

Définition 11 *Un* **système parallèle** *est un système tel qu'il suffit qu'un seul de ses composants fonctionne pour qu'il fonctionne.*

Autrement dit, la défaillance du système survient quand tous ses composants sont en panne.

Dans les systèmes parallèles, on distingue deux cas :

- La **redondance passive** ou **stand-by** : un seul composant fonctionne à la fois. Quand le composant qui fonctionne tombe en panne, il est instantanément remplacé par un des composants en attente. Dans ce cas, $X = \sum_{i=1}^{n} X_i$. La proposition 2 montre que si tous les composants sont indépendants et de même loi $\exp(\lambda)$, la durée de vie du système en redondance passive correspondant est de loi gamma $G(n, \lambda)$.
- La **redondance active** : les n composants fonctionnent en même temps.

On se place dans la suite de cette section dans le cas de la redondance active. Le diagramme de fiabilité est donné dans la figure 4.4.

On a évidemment :

$$X = \max_{i=1}^{n} X_i$$

La fiabilité du système est alors :

$$R(x) = \mathbb{P}(X > x) = \mathbb{P}(\max_{i=1}^{n} X_i > x) = 1 - \mathbb{P}(\max_{i=1}^{n} X_i \leq x) = 1 - \mathbb{P}(\forall i, X_i \leq x)$$

Avec des composants indépendants, on obtient :

$$R(x) = 1 - \prod_{i=1}^{n} \mathbb{P}(X_i \leq x) = 1 - \prod_{i=1}^{n} (1 - \mathbb{P}(X_i > x))$$

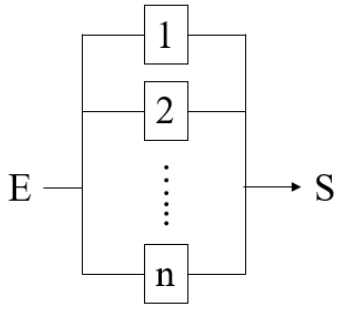


FIGURE 4.4 – Diagramme de fiabilité pour un système parallèle

d'où finalement :

$$R(x) = 1 - \prod_{i=1}^{n}(1 - r_i(x))$$

En écrivant $R(x) = 1 - \prod_{i=1}^{n}\left(1 - \exp\left(-\int_0^x h_i(u)\,du\right)\right)$ puis $h(x) = -\dfrac{R'(x)}{R(x)}$, on obtient que le taux de défaillance du système est :

$$h(x) = \frac{\sum_{i=1}^{n} h_i(x)\exp\left(-\int_0^x h_i(u)\,du\right)\prod_{j\neq i}\left(1 - \exp\left(-\int_0^x h_j(u)\,du\right)\right)}{1 - \prod_{i=1}^{n}\left(1 - \exp\left(-\int_0^x h_i(u)\,du\right)\right)}$$

Donc, contrairement au cas d'un système série, le taux de défaillance d'un système parallèle ne s'exprime pas facilement en fonction du taux de défaillance de ses composants.

Il n'y a pas non plus d'expression simple du MTTF.

4.3.2 Cas où tous les composants ont un taux de défaillance constant

On a :

- $\forall i, h_i(x) = \lambda_i$.
- $R(x) = 1 - \prod_{i=1}^{n}(1 - \exp(-\lambda_i x))$.
- $h(x) = \dfrac{\sum_{i=1}^{n}\lambda_i \exp(-\lambda_i x)\prod_{j\neq i}(1 - \exp(-\lambda_j x))}{1 - \prod_{i=1}^{n}(1 - \exp(-\lambda_i x))}$. Donc un système parallèle dont tous les composants ont un taux de défaillance constant, n'a pas un taux de défaillance constant !

En développant la fiabilité, on obtient :

$$\begin{aligned} R(x) &= \sum_{i=1}^{n} \exp(-\lambda_i x) - \sum_{i,j\ distincts} \exp(-(\lambda_i+\lambda_j)x) + \sum_{i,j,k\ distincts} \exp(-(\lambda_i+\lambda_j+\lambda_k)x) \\ &\quad - \ldots + (-1)^{n+1} \exp(-[\sum_{i=1}^{n} \lambda_i]x) \end{aligned}$$

d'où on déduit le MTTF :

$$MTTF = \sum_{i=1}^{n} \frac{1}{\lambda_i} - \sum_{i,j\ distincts} \frac{1}{\lambda_i+\lambda_j} + \sum_{i,j,k\ distincts} \frac{1}{\lambda_i+\lambda_j+\lambda_k} - \ldots + (-1)^{n+1} \frac{1}{\sum_{i=1}^{n} \lambda_i}$$

On montre que $\lim_{t\to+\infty} h(x) = \min_{i=1}^{n} \lambda_i$. C'est logique : c'est le composant le plus fiable qui a tendance à tomber en panne le dernier, donc à provoquer la panne du système.

4.3.3 Cas où tous les composants sont identiques

On a $\forall i, r_i(x) = r(x)$. Alors la fiabilité du système est :

$$R(x) = 1 - [1 - r(x)]^n$$

Comme $r(x) \in [0,1]$, on a $[1-r(x)]^n \geq [1-r(x)]^{n+1}$, donc $1-[1-r(x)]^n \leq 1-[1-r(x)]^{n+1}$. Par conséquent, quand on augmente le nombre de composants en redondance dans un système parallèle, on augmente la fiabilité du système.

Notons que c'est l'inverse pour les systèmes série puisque $[r(x)]^n \geq [r(x)]^{n+1}$.

4.4 Systèmes k/n

Définition 12 *Un* **système** k/n *est un système qui ne fonctionne que si au moins* k *composants parmi* n *fonctionnent.*

Par exemple, le système de contrôle-commande de la température d'un réacteur chimique ou nucléaire est conçu selon une architecture 2/3.

- $k = 1$ correspond à un système parallèle.
- $k = n$ correspond à un système série.

On ne peut pas représenter ce mode de fonctionnement par un diagramme de fiabilité usuel.

La fiabilité $R(x)$ est la probabilité que k composants au moins parmi n fonctionnent encore à l'instant x. Si on note N_x le nombre de composants qui fonctionnent à l'instant x, on a :

$$R(x) = \mathbb{P}(N_x \geq k)$$

Dans le cas général, on ne peut rien dire de plus. Mais si on suppose que tous les composants sont identiques et indépendants, de même fiabilité $r(x)$, alors la variable aléatoire N_x est de loi binomiale $\mathcal{B}(n, r(x))$, ce qui permet de calculer :

$$R(x) = \sum_{j=k}^{n} C_n^j \, r(x)^j \, [1 - r(x)]^{n-j}$$

- Pour $k = n$, on obtient $R(x) = r(x)^n$. C'est bien la fiabilité d'un système série.
- Pour $k = 1$, on obtient :

$$\begin{aligned} R(x) &= \sum_{j=1}^{n} C_n^j \, r(x)^j \, [1 - r(x)]^{n-j} = \sum_{j=0}^{n} C_n^j \, r(x)^j \, [1 - r(x)]^{n-j} - C_n^0 \, r(x)^0 \, [1 - r(x)]^{n-0} \\ &= [r(x) + 1 - r(x)]^n - [1 - r(x)]^n = 1 - [1 - r(x)]^n \end{aligned}$$

  C'est bien la fiabilité d'un système parallèle.

4.5 Systèmes mixtes

Les systèmes mixtes sont obtenus en combinant les systèmes série et les systèmes parallèles.

4.5.1 Systèmes série-parallèle

Définition 13 *Un* **système série-parallèle** *résulte de la mise en parallèle de sous-systèmes série.*

Le diagramme de fiabilité est donné dans la figure 4.5.

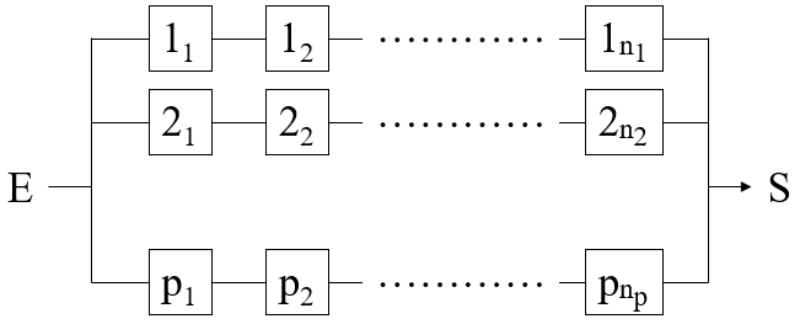


FIGURE 4.5 – Diagramme de fiabilité pour un système série-parallèle

Si on note $r_{ij}(x)$ la fiabilité du $j^{\text{ème}}$ composant de la $i^{\text{ème}}$ branche, les résultats précédents montrent que la fiabilité est :

$$R(x) = 1 - \prod_{i=1}^{p} \left[1 - \prod_{j=1}^{n_i} r_{ij}(x) \right]$$

4.5.2 Systèmes parallèle-série

Définition 14 *Un* **système parallèle-série** *résulte de la mise en série de sous-systèmes parallèles.*

Le diagramme de fiabilité est donné dans la figure 4.6.

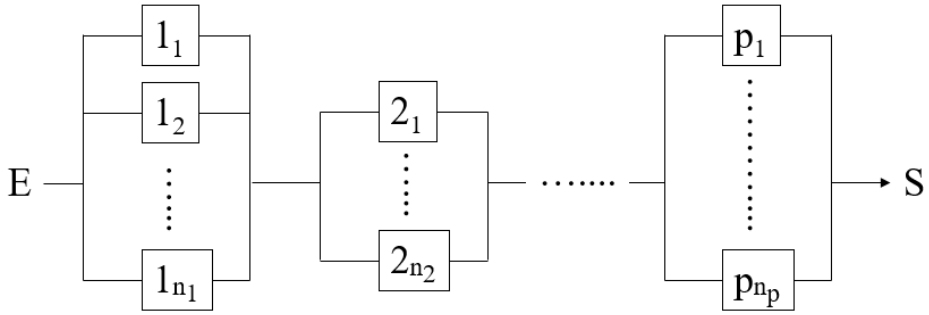


FIGURE 4.6 – Diagramme de fiabilité pour un système parallèle-série

Avec les mêmes notations que précédemment, on obtient que la fiabilité est :

$$R(x) = \prod_{i=1}^{p} \left[1 - \prod_{j=1}^{n_i} [1 - r_{ij}(x)] \right]$$

La chaîne hi-fi avec fonctionnement dégradé est un système parallèle-série. Sa fiabilité est :

$$R(x) = [1 - (1 - r_1(x))(1 - r_2(x))]\ r_3(x)\ [1 - (1 - r_4(x))(1 - r_5(x))]$$

4.6 La méthode de factorisation

De nombreux systèmes ne sont pas des systèmes série, parallèles, k/n ou mixtes. C'est le cas du système dit **en pont**, dont le diagramme de fiabilité est donné dans la figure 4.7.

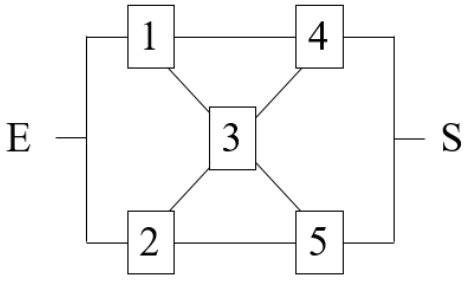


FIGURE 4.7 – Diagramme de fiabilité pour un système en pont

Pour calculer sa fiabilité, on va utiliser la **méthode de factorisation**. Celle-ci consiste à effectuer des conditionnements successifs qui vont permettre de se ramener à des systèmes mixtes. Ici, on va conditionner par rapport au fonctionnement du composant 3, qui a un rôle particulier dans le système en pont.

Le théorème des probabilités totales permet d'écrire :

$$\begin{aligned} R(x) &= \mathbb{P}(X > x) = \mathbb{P}(X > x | X_3 > x)\mathbb{P}(X_3 > x) + \mathbb{P}(X > x | X_3 \leq x)\mathbb{P}(X_3 \leq x) \\ &= R_A(x)\, r_3(x) + R_B(x)\,(1 - r_3(x)) \end{aligned}$$

où $R_A(x)$ est la fiabilité du système quand on sait que le composant 3 fonctionne, c'est-à-dire la fiabilité du système A donné par la figure 4.8, et $R_B(x)$ est la fiabilité du système quand on sait que le composant 3 ne fonctionne pas, c'est-à-dire la fiabilité du système B donné par la figure 4.9.

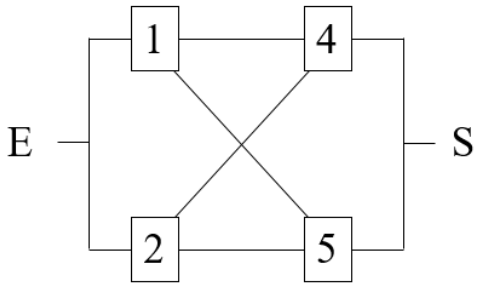


FIGURE 4.8 – Système en pont, 3 fonctionne

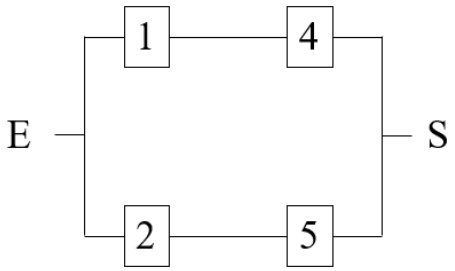


FIGURE 4.9 – Système en pont, 3 en panne

Il est clair que le système A est équivalent à un système parallèle-série, dont la fiabilité est :

$$R_A(x) = [1 - (1 - r_1(x))(1 - r_2(x))]\ [1 - (1 - r_4(x))(1 - r_5(x))]$$

De même, le système B est un système série-parallèle, dont la fiabilité est :

$$R_B(x) = 1 - [1 - r_1(x)r_4(x)]\ [1 - r_2(x)r_5(x)]$$

Finalement, la fiabilité du système en pont est :

$$\begin{aligned} R(x) &= r_3(x)\ [1 - (1 - r_1(x))(1 - r_2(x))]\ [1 - (1 - r_4(x))(1 - r_5(x))] \\ &\quad + (1 - r_3(x))\ [1 - [1 - r_1(x)r_4(x)]\ [1 - r_2(x)r_5(x)]] \end{aligned}$$

Si tous les composants sont identiques, on obtient :

$$\begin{aligned} R(x) &= r(x)\ \left[1 - (1 - r(x))^2\right]^2 + (1 - r(x))\ \left[1 - (1 - r^2(x))^2\right] \\ &= r^2(x)\ \left[2 + 2r(x) - 5r^2(x) + 2r^3(x)\right] \end{aligned}$$

On peut vérifier que $R(0) = 1$.

Si les composants ont un taux de défaillance constant λ, $r(x) = \exp(-\lambda x)$, d'où :

$$R(x) = 2\exp(-2\lambda x) + 2\exp(-3\lambda x) - 5\exp(-4\lambda x) + 2\exp(-5\lambda x)$$

On en déduit facilement la durée de vie moyenne du système :

$$MTTF = \int_0^{+\infty} R(x)\,dx = \frac{2}{2\lambda} + \frac{2}{3\lambda} - \frac{5}{4\lambda} + \frac{2}{5\lambda} = \frac{49}{60\lambda} = 0.82\,\frac{1}{\lambda}$$

Le MTTF du système vaut donc 82 % du MTTF de ses composants.

Le taux de défaillance est :

$$h(x) = -\frac{R'(x)}{R(x)} = \frac{4\lambda\exp(-2\lambda x) + 6\lambda\exp(-3\lambda x) - 20\lambda\exp(-4\lambda x) + 10\lambda\exp(-5\lambda x)}{2\exp(-2\lambda x) + 2\exp(-3\lambda x) - 5\exp(-4\lambda x) + 2\exp(-5\lambda x)}$$

On a $\lim\limits_{x\to+\infty} h(x) = 2\lambda$. Cela signifie qu'au bout d'un moment, le système se comporte comme deux composants en série, qui est la configuration minimale avant la panne définitive. La forme du taux de défaillance est donnée dans la figure 4.10.

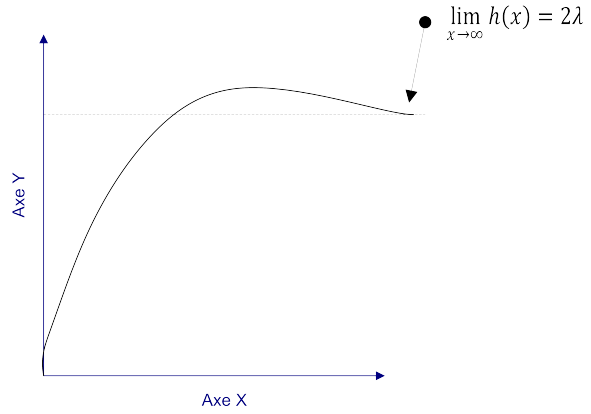


FIGURE 4.10 – Taux de défaillance du système en pont

Chapitre 5

Introduction à l'analyse statistique de données de fiabilité

5.1 Problèmes statistiques pour des données de fiabilité

Dans la première partie du cours, de nombreuses modèles ont été proposés pour calculer la fiabilité, le taux de défaillance, le MTTF ou toute autre quantité intéressante pour un système. Ces modèles reposent sur des hypothèses probabilistes. Par exemple, si on suppose que la durée de vie d'un système est de loi exponentielle de paramètre λ, la fiabilité est $R(x) = \exp(-\lambda x)$, le taux de défaillance est λ et le MTTF est $1/\lambda$.

Pour pouvoir appliquer ces résultats théoriques au calcul du taux de défaillance d'un système réel, il faut d'une part s'assurer que la durée de vie de ce système est bien de loi exponentielle, et d'autre part pouvoir calculer d'une manière ou d'une autre la valeur du paramètre λ.

Cela ne peut se faire qu'au moyen d'une analyse statistique de données expérimentales, issues du retour d'expérience, de résultats d'exploitation des systèmes en vie opérationnelle, ou d'essais de fiabilité en période de test des systèmes.

Les observations se présentent la plupart du temps sous la forme d'une suite de n durées $x_1, \ldots, x_n$, qu'on supposera être les réalisations de variables aléatoires $X_1, \ldots, X_n$. Si les x_i représentent les durées de vie de n systèmes similaires et indépendants placés dans des conditions de fonctionnement identiques, il est raisonnable de supposer que les variables aléatoires X_i sont indépendantes et de même loi (i.i.d., on dit alors que $X_1, \ldots, X_n$ est un échantillon)

Exemple des systèmes d'air conditionné dans les Boeing : comme un appareil défectueux est remplacé par un neuf, on s'attend à ce que les durées inter-défaillances dans un Boeing donné soient des variables aléatoires i.i.d. Il faut le vérifier, trouver une loi de probabilité adaptée, et estimer ses paramètres. Il est intéressant également de déterminer si cette loi de probabilité est la même pour les 13 appareils. On doit estimer la fiabilité, le taux de défaillance et le MTTF de ce système.

Plus généralement, passons rapidement en revue les principaux problèmes statistiques que nous allons étudier. Ce qui intéresse en premier lieu le concepteur et l'utilisateur d'un système, c'est l'**estimation** de sa fiabilité, de son taux de défaillance et de son MTTF. On peut effectuer ces estimations de manière **non paramétrique**, c'est-à-

dire sans faire aucune hypothèse sur la loi de probabilité des X_i. On utilise alors la **loi de probabilité empirique** des observations.

La plupart du temps, on fait des hypothèses **paramétriques**. Par exemple, on suppose que les X_i constituent un échantillon d'une loi exponentielle ou d'une loi de Weibull. Le problème majeur est alors l'estimation des paramètres de ces lois. On peut en donner une **estimation ponctuelle** (une seule valeur) ou ensembliste (un **intervalle de confiance**). La méthode d'estimation la plus usuelle est la **méthode du maximum de vraisemblance**, mais on peut également utiliser la **méthode des moments** ou une **estimation bayésienne**.

Mais pour pouvoir procéder à ces estimations, il faut s'assurer que les hypothèses probabilistes préalables sont bien vérifiées : par exemple, est-il raisonnable de penser que les X_i sont i.i.d. de loi exponentielle ? C'est un problème de **test d'adéquation** (ou d'ajustement). Il existe des tests d'adéquation graphiques, les **graphes de probabilité**, et des tests d'adéquation statistiques.

Si un test d'adéquation nous amène à adopter un modèle probabiliste donné, on peut s'intéresser à des problèmes précis sur ce modèle. Par exemple, on peut se préoccuper de savoir si le taux de défaillance d'un système est inférieur à une valeur critique donnée λ_0. Il faudra alors faire un **test d'hypothèses** de H_0 : "$\forall t \geq 0, \lambda(t) \geq \lambda_0$" contre H_1 : "$\forall t \geq 0, \lambda(t) < \lambda_0$" (ou plus simplement H_0 : "$\lambda \geq \lambda_0$" contre H_1 : "$\lambda < \lambda_0$" si on suppose le taux de défaillance constant).

Enfin, la qualité croissante des systèmes et les contraintes économiques fortes font que, bien souvent, on ne dispose que d'observations partielles. Par exemple, on peut faire fonctionner 10 systèmes identiques et indépendants pendant un mois dans les mêmes conditions, et n'observer que 4 défaillances dans cette période. Si on n'a pas la possibilité de poursuivre plus longtemps l'expérience, il va falloir tirer le maximum d'information de ces données dites **censurées**. Tous les problèmes statistiques précédemment évoqués peuvent se traiter avec des données censurées. Quand il n'y a pas de censure, on dit que les données sont **complètes**.

On peut également avoir recours à des **tests accélérés**, pour lesquels certaines conditions environnementales (température par exemple) seront modifiées pour accélérer la survenance de défaillances. Il s'agira ensuite, à partir des observations de défaillances en conditions accélérées, d'en déduire le comportement probable de l'appareil dans des conditions standard d'utilisation.

La deuxième partie de ce chapitre présente la méthode des graphes de probabilités, qui permet de manière très simple de proposer un modèle plausible pour des données. Les autres problèmes statistiques évoqués précédemment seront traités dans les chapitres suivants.

5.2 Les graphes de probabilités

On a relevé les durées de vie (en heures) de 10 ampoules identiques fonctionnant indépendamment les unes des autres :

91.6 35.7 251.3 24.3 5.4 67.3 170.9 9.5 118.4 57.1

On aura besoin par la suite de ces données ordonnées :

5.4 9.5 24.3 35.7 57.1 67.3 91.6 118.4 170.9 251.3

Si l'échantillon initial est noté $x_1, ..., x_n$, l'échantillon ordonné sera noté $x_1^*, ..., x_n^*$. On a donc, par exemple :

$x_1 = 91.6 =$ durée de vie de la première ampoule.

$x_1^* = \min(x_1, \ldots, x_n) = 5.4 =$ plus petite des durées de vie des 10 ampoules.

x_i^* est appelée la $i^{\text{ème}}$ statistique d'ordre.

On se demande si on peut faire l'hypothèse que ces données proviennent d'un modèle de fiabilité donné. Par exemple, la durée de vie de ces ampoules est-elle de loi exponentielle (auquel cas cela signifie qu'elles ont un taux de défaillance constant) ? Ou de loi de Weibull ? Une autre loi est-elle plus appropriée ?

Une première solution consiste à représenter un histogramme de ces données. Un histogramme est une estimation de la densité de probabilité sous-jacente. Si la forme de l'histogramme est proche de celle de la densité d'une loi connue, alors il est possible que les observations proviennent de cette loi.

Un inconvénient est qu'il n'existe pas de façon unique de définir un histogramme. Pour les données des ampoules, la figure 5.1 donne 3 histogrammes correspondant respectivement à l'histogramme brut de R, un histogramme à classes de même largeur selon la règle de Sturges, et un histogramme à classes de même effectif selon la même règle.

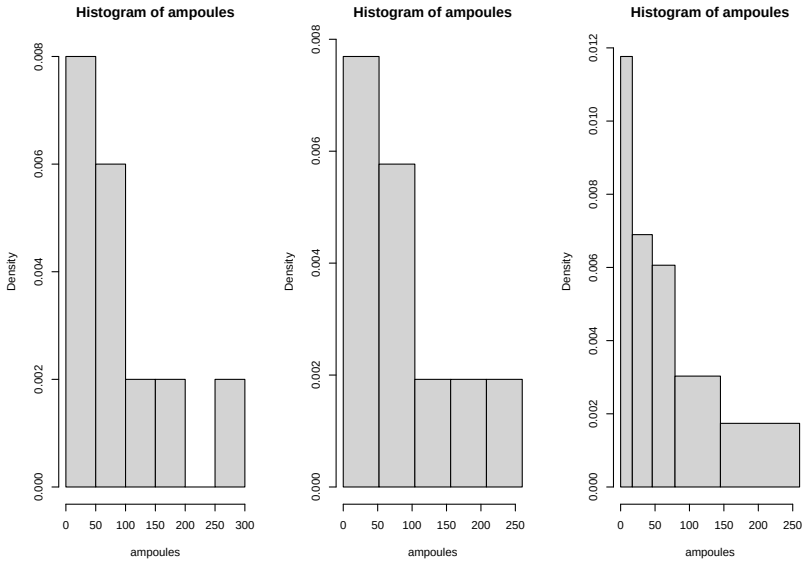


FIGURE 5.1 – Trois histogrammes pour les données des ampoules

La forme de ces histogrammes évoque vaguement la densité d'une loi exponentielle, surtout le troisième. De toutes façons, avec seulement 10 données, on ne peut pas espérer avoir une estimation satisfaisante de la densité de l'échantillon. Il est donc préférable d'avoir une méthode plus précise. C'est ce que fourniront les graphes de probabilité. Il faut d'abord définir le fonction de répartition empirique.

Définition 15 *La* **fonction de répartition empirique** *(FdRE)* F_n *associée à un échantillon*

$x_1, \ldots, x_n$ *est la fonction définie par :*

$$\forall x \in \mathbb{R}\ ,\ F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i \leq x\}} = \begin{cases} 0 & si\ \ x < x_1^* \\ \\ \dfrac{i}{n} & si\ \ x_i^* \leq x < x_{i+1}^* \\ \\ 1 & si\ \ x \geq x_n^* \end{cases}$$

La fonction de répartition de X, $F(x) = P(X \leq x)$, donne la probabilité qu'une observation soit inférieure à x, tandis que $F_n(x)$ est le pourcentage d'observations inférieures à x. Par conséquent, $F_n(x)$ est une estimation de $F(x)$. On peut montrer que cette estimation est d'excellente qualité.

$F_n(x)$ est une fonction en escalier qui fait des sauts de hauteur $1/n$ en chaque point de l'échantillon. La figure 5.2 représente la fonction de répartition empirique de l'échantillon des durées de vie d'ampoules.

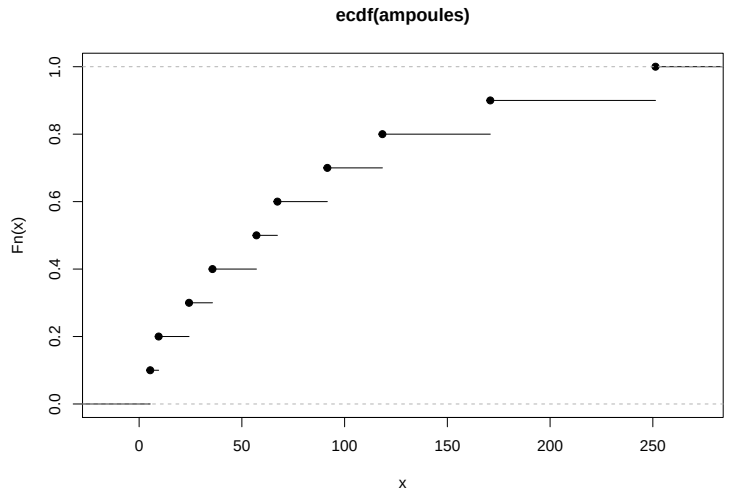


FIGURE 5.2 – Ampoules, fonction de répartition empirique

Si on veut déterminer un modèle probabiliste acceptable pour les observations, une première idée est de tracer le graphe de la fonction de répartition empirique et de déterminer si ce graphe ressemble à celui de la fonction de répartition d'une loi connue. En fait, il est très difficile de procéder ainsi car les fonctions de répartition de toutes les lois de probabilité se ressemblent : à vue d'oeil, il n'y a pas de grande différence entre les fonctions de répartition des lois normale et exponentielle, alors que leurs densités ne se ressemblent pas du tout.

Une seconde idée est alors d'appliquer une transformation à la fonction de répartition empirique qui permette de reconnaître visuellement une caractéristique d'une loi de probabilité. Un **graphe de probabilités** (en anglais **probability plot** ou **Q-Q plot**) est un nuage de points tracé à partir de la fonction de répartition empirique, tel que les points doivent être approximativement alignés si les observations proviennent d'une loi de probabilité bien précise.

Si on souhaite savoir si les observations sont issues de la loi de probabilité, dépendant d'un paramètre θ inconnu, dont la fonction de répartition est F, le principe est de chercher une relation affine du type $h[F(x)] = \alpha(\theta)g(x) + \beta(\theta)$, où h et g sont des fonctions qui ne dépendent pas de θ.

Ainsi, si la vraie fonction de répartition des observations est F, $h[F_n(x)]$ devrait être "proche" de $\alpha(\theta)g(x) + \beta(\theta)$, pour tout x. Pour $x = x_i^*$, $h[F_n(x_i^*)] = h(i/n)$. Donc, si la vraie fonction de répartition est F, les points $\Big(g(x_i^*), h(i/n)\Big)$ seront approximativement alignés. La pente et l'ordonnée à l'origine de cette droite fourniront des estimations de $\alpha(\theta)$ et $\beta(\theta)$, donc la plupart du temps de θ.

Définition 16 *: Soit F la fonction de répartition d'une loi de probabilité, dépendant d'un paramètre inconnu θ. S'il existe des fonctions h, g, α et β telles que,*

$$\forall x \in \mathbb{R},\ h[F(x)] = \alpha(\theta)g(x) + \beta(\theta)$$

alors le nuage des points

$$\Big(g(x_i^*), h(i/n)\Big),\ i \in \{1, \ldots, n\}$$

est appelé **graphe de probabilités** *pour la loi de fonction de répartition F. Si les points du nuage sont approximativement alignés, on admettra que F est une fonction de répartition plausible pour les observations.*

Exemple 1 : Graphe de probabilités pour la loi exponentielle

Si X est de loi $exp(\lambda)$, $F(x) = 1 - e^{-\lambda x}$, d'où $\ln(1 - F(x)) = -\lambda x$. Par conséquent, le graphe de probabilités pour la loi exponentielle est le nuage des points $\Big(x_i^*, \ln(1 - i/n)\Big),\ i \in \{1, \ldots, n-1\}$ (le point correspondant à $i = n$ doit être enlevé car $\ln(1-n/n) = -\infty$).

Si ces points sont approximativement alignés sur une droite de pente négative et passant par l'origine, on pourra considérer que la loi exponentielle est un modèle probabiliste vraisemblable pour ces observations. La pente de la droite fournit alors une estimation graphique de λ. Inversement, si ce n'est pas le cas, il est probable que les observations ne soient pas issues d'une loi exponentielle.

Sur l'exemple des ampoules, le graphe de probabilités pour la loi exponentielle est donné par la figure 5.3. A vue d'oeil, les points peuvent être considérés comme approximativement alignés sur une droite de pente négative et passant par l'origine. Donc on peut considérer qu'il est vraisemblable que la durée de vie d'une ampoule soit une variable aléatoire de loi exponentielle. Cette conclusion est cohérente avec celle des histogrammes.

La droite en question a pour équation $y = -\lambda x$. Sa pente fournit donc une estimation du paramètre λ. Pour déterminer cette pente, il suffit de tracer la droite des moindres carrés pour ce nuage de points (figure 5.4). On obtient ici une estimation de l'ordre de 0.013.

Exemple 2 : Graphe de probabilités pour la loi de Weibull

A faire en exercice : construire le graphe de probabilités pour la loi de Weibull et donner des estimations graphiques de η et β.

Au final, la méthode des graphes de probabilités permet de déterminer très simplement si un modèle probabiliste est adapté ou non à un jeu de données. L'inconvénient de cette méthode est qu'on ne dispose pas de critère objectif pour décider à partir de quand les points peuvent être considérés comme raisonnablement alignés. Il est donc nécessaire de compléter cette méthode empirique par des méthodes statistiques plus précises, les tests d'adéquation (voir chapitre suivant). Un modèle ayant été retenu, on peut passer à l'étape d'estimation paramétrique.

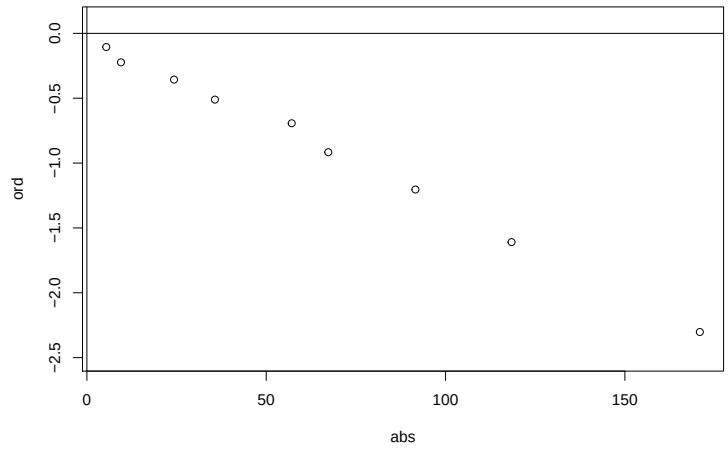


FIGURE 5.3 – Ampoules, graphe de probabilités pour la loi exponentielle

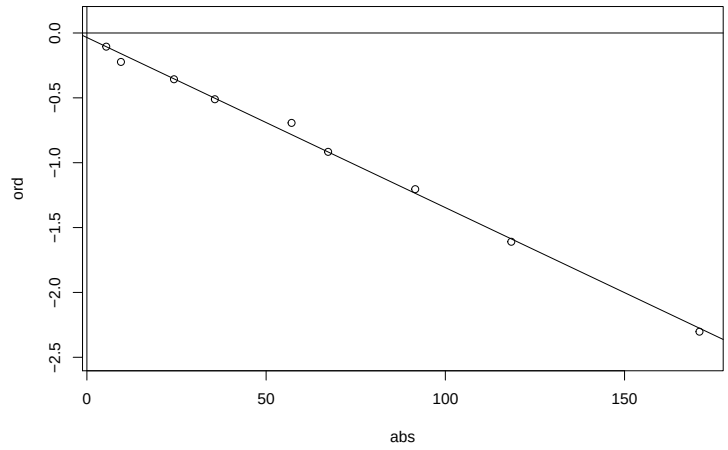


FIGURE 5.4 – Ampoules, graphe de probabilités pour la loi exponentielle, avec la droite de régression

Chapitre 6

Méthodes paramétriques d'analyse d'échantillons complets

On suppose ici que les observations $x_1, \ldots, x_n$ sont les réalisations de variables aléatoires $X_1, \ldots, X_n$ i.i.d. de loi dépendant d'un paramètre inconnu θ. La fonction de répartition et la densité de cette loi seront maintenant notées respectivement $F(x;\theta)$ et $f(x;\theta)$.

Le but du chapitre est de présenter rapidement les méthodes basiques d'analyse statistique paramétrique pour des échantillons de ce type : estimation paramétrique, intervalles de confiance et tests d'hypothèses.

6.1 Estimation paramétrique

6.1.1 Principes de l'estimation

Estimer une grandeur inconnue θ, c'est en proposer une approximation que l'on espère la plus proche possible de la vraie valeur de cette grandeur. Cette approximation est une fonction des observations $x_1, \ldots, x_n$, c'est-à-dire une **statistique** $t_n = t(x_1, \ldots, x_n)$. La variable aléatoire correspondante $T_n = t(X_1, \ldots, X_n)$ est appelée un **estimateur** de θ et sa réalisation t_n une **estimation** de θ. Pour être performant, un estimateur doit être :

- **sans biais** : $E[T_n] = \theta$.
- **convergent** : T_n tend vers θ quand n tend vers l'infini (plus il y a d'observations, plus l'estimateur est proche de la valeur à estimer).

On retient en général la convergence en moyenne quadratique : $\lim_{n\to\infty} E[(T_n-\theta)^2] = 0$, qui devient, si T_n est sans biais, $\lim_{n\to\infty} Var[T_n] = 0$.

Un estimateur optimal est un **estimateur sans biais et de variance minimale (ESBVM)**.

6.1.2 Méthode des moments

L'idée de la méthode des moments est d'estimer une espérance mathématique par une moyenne empirique, une variance par une variance empirique, etc...

Si le paramètre à estimer est l'espérance de la loi des X_i, alors on peut l'estimer par la moyenne empirique de l'échantillon. Autrement dit, si $\theta = E[X]$, alors l'**estimateur** de θ **par la méthode des moments (EMM)** est $\tilde{\theta}_n = \bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Plus généralement, pour $\theta \in \mathbb{R}$, si $E[X] = \varphi(\theta)$, où φ est une fonction inversible, alors l'estimateur de θ par la méthode des moments est $\tilde{\theta}_n = \varphi^{-1}\left(\bar{X}_n\right)$.

De la même manière, on estime la variance de la loi des X_i par la variance empirique de l'échantillon $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \bar{X}_n^2$.

Plus généralement, si la loi des X_i a deux paramètres θ_1 et θ_2 tels que $(E[X], Var[X])$ = $\varphi(\theta_1, \theta_2)$, où φ est une fonction inversible, alors les estimateurs de θ_1 et θ_2 par la méthode des moments sont $\left(\tilde{\theta}_{1n}, \tilde{\theta}_{2n}\right) = \varphi^{-1}\left(\bar{X}_n, S_n^2\right)$.

Ce principe peut naturellement se généraliser aux moments de tous ordres, centrés ou non centrés : $E\left[(X - E[X])^k\right]$ et $E\left[X^k\right]$, $k \geq 1$.

L'avantage de cette méthode est qu'elle est naturelle et qu'elle permet d'obtenir des estimateurs ayant une forme explicite, alors que les estimateurs de maximum de vraisemblance sont souvent obtenus comme solutions non explicites d'un système d'équations.

6.1.3 Méthode du maximum de vraisemblance

Définition 17 *Pour des observations $x_1, \ldots, x_n$ toutes discrètes ou toutes continues, la* **fonction de vraisemblance** *(ou plus simplement* **vraisemblance***) est la fonction du paramètre θ :*

$$\mathcal{L}(\theta; x_1, \ldots, x_n) = \begin{cases} P(X_1 = x_1, \ldots, X_n = x_n; \theta) & \text{si les } X_i \text{ sont discrètes} \\ \\ f_{X_1,\ldots,X_n}(x_1, \ldots, x_n; \theta) & \text{si les } X_i \text{ sont continues} \end{cases}$$

Pour les modèles de fiabilité étudiés dans ce cours, les X_i sont indépendantes et de même loi continue de densité f. Dans ce cas, la vraisemblance s'écrit :

$$\mathcal{L}(\theta; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta)$$

Remarque : La densité f est une fonction des observations $x_1, \ldots, x_n$, dépendant du paramètre θ. A l'inverse, la fonction de vraisemblance est considérée comme une fonction de θ dépendant des observations $x_1, \ldots, x_n$, ce qui permet, par exemple, de dériver cette fonction par rapport à θ.

Le principe de la méthode est de considérer que la valeur la plus vraisemblable de θ est la valeur pour laquelle la probabilité d'observer $x_1, \ldots, x_n$ est la plus forte possible (cas d'observations discrètes). Cela revient à faire comme si c'était l'éventualité la plus probable qui s'était produite au cours de l'expérience.

Définition 18 *L'***estimation de maximum de vraisemblance** *de θ est la valeur $\hat{\theta}_n$ de θ qui rend maximale la fonction de vraisemblance $\mathcal{L}(\theta; x_1, \ldots, x_n)$. L'***estimateur de maximum de vraisemblance (EMV)** *de θ est la variable aléatoire correspondante.*

La valeur qui maximise une fonction maximise aussi son logarithme. Comme la fonction de vraisemblance s'exprime comme un produit, il est plus pratique de maximiser la log-vraisemblance $\ln \mathcal{L}(\theta; x_1, \ldots, x_n)$.

La méthode la plus usuelle pour maximiser une fonction est d'annuler sa dérivée. Quand $\theta = (\theta_1, \ldots, \theta_d) \in \mathbb{R}^d$ et que toutes les dérivées partielles ci-dessous existent, $\hat{\theta}_n$ est solution du système d'équations appelées **équations de vraisemblance** :

$$\forall j \in \{1, \ldots, d\}, \quad \frac{\partial}{\partial \theta_j} \ln \mathcal{L}(\theta; x_1, \ldots, x_n) = 0$$

Il est fréquent que le système des équations de vraisemblance n'ait pas de solution explicite. Dans ce cas, on le résoud par des méthodes numériques, comme la méthode de Newton-Raphson. En `R`, la maximisation numérique peut se faire à l'aide des commandes `uniroot` ou `optim`.

Un estimateur de maximum de vraisemblance n'est pas forcément unique (la vraisemblance peut avoir plusieurs maxima), ni sans biais, ni de variance minimale. Mais il possède d'excellentes propriétés asymptotiques, pour peu que la loi des observations vérifie certaines conditions de régularité (ce qui sera le cas des modèles usuels de fiabilité).

Pour comprendre ces propriétés, il faut d'abord définir la notion de quantité d'information.

Définition 19 *Pour $\theta \in \mathbb{R}$, si la loi des observations vérifie des conditions de régularité, on appelle* **quantité d'information** *(de Fisher) sur θ apportée par l'échantillon $x_1, \ldots, x_n$, la quantité :*

$$\mathcal{I}_n(\theta) = Var\left[\frac{\partial}{\partial \theta} \ln \mathcal{L}(\theta; X_1, \ldots, X_n)\right] = -E\left[\frac{\partial^2}{\partial \theta^2} \ln \mathcal{L}(\theta; X_1, \ldots, X_n)\right]$$

L'intérêt de la quantité d'information de Fisher est qu'elle fournit une borne inférieure pour la variance de n'importe quel estimateur sans biais de θ. Ce résultat s'exprime sous la forme de la propriété suivante :

Proposition 3 Inégalité de Fréchet-Darmois-Cramer-Rao (FDCR) *: Si la loi des observations vérifie les conditions de régularité, alors pour tout estimateur T_n de θ, on a :*

$$Var(T_n) \geq \frac{\left[\frac{\partial}{\partial \theta} E[T_n]\right]^2}{\mathcal{I}_n(\theta)}$$

Ce résultat est particulièrement intéressant pour les estimateurs sans biais. En effet, si T_n est un estimateur sans biais de θ, alors $E[T_n] = \theta$, donc $Var[T_n] \geq \frac{1}{\mathcal{I}_n(\theta)}$.

La quantité $\frac{1}{\mathcal{I}_n(\theta)}$ est appelée la **borne de Cramer-Rao**. L'inégalité FDCR dit donc que la variance d'un estimateur sans biais quelconque de θ est forcément supérieure à cette borne.

Si un estimateur T_n est sans biais et tel que $Var[T_n] = \frac{1}{\mathcal{I}_n(\theta)}$, alors sa variance est minimale, donc c'est un ESBVM de θ.

Revenons maintenant aux propriétés asymptotiques de l'estimateur de maximum de vraisemblance.

Proposition 4 *Si les X_i sont indépendants et de même loi dépendant d'un paramètre réel θ, cette loi vérifiant des conditions de régularité, on a :*

- *$\hat{\theta}_n$ converge presque sûrement vers θ.*
- *$\sqrt{\mathcal{I}_n(\theta)}(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$, ce qui signifie que, quand n est grand, $\hat{\theta}_n$ est approximativement de loi $\mathcal{N}\left(\theta, \frac{1}{\mathcal{I}_n(\theta)}\right)$. On en déduit que $\hat{\theta}_n$ est asymptotiquement gaussien, sans biais (son espérance tend vers θ) et de variance minimale (sa variance tend vers la borne de Cramer-Rao $1/\mathcal{I}_n(\theta)$). Cette propriété peut aussi s'écrire :*

$$\sqrt{n}\,(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{1}{\mathcal{I}_1(\theta)}\right)$$

- *Si $\hat{\theta}_n$ est l'EMV de θ, alors $\varphi(\hat{\theta}_n)$ est l'EMV de $\varphi(\theta)$. De plus, si φ est dérivable, on a :*

$$\sqrt{n}\left[\varphi(\hat{\theta}_n) - \varphi(\theta)\right] \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{\varphi'(\theta)^2}{\mathcal{I}_1(\theta)}\right)$$

  Ce résultat est connu sous le nom de **méthode delta**. *Quand n est grand, $\varphi(\hat{\theta}_n)$ est donc approximativement de loi $\mathcal{N}\left(\varphi(\theta), \frac{\varphi'(\theta)^2}{\mathcal{I}_n(\theta)}\right)$.*

- *En général, l'EMV est meilleur que l'EMM au sens où $Var[\hat{\theta}_n] \leq Var[\tilde{\theta}_n]$. C'est au moins vrai asymptotiquement.*

Quand θ est un paramètre de dimension $d > 1$, on a des résultats équivalents basés sur la propriété suivante :

$$\sqrt{n}\,(\hat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}_d\left(0, \mathcal{I}_1^{-1}(\theta)\right)$$

où $\mathcal{N}_d$ est la loi normale dans $\mathbb{R}^d$ et $\mathcal{I}_n(\theta)$ est la matrice d'information de Fisher, matrice de taille $d \times d$ dont le terme d'ordre (j,k) est :

$$\mathcal{I}_{jk}(\theta) = -E\left[\frac{\partial^2}{\partial\theta_j \partial\theta_k} \ln \mathcal{L}(\theta; X_1, \ldots, X_n)\right]$$

Le fait que l'EMV soit asymptotiquement sans biais et de variance minimale fait que, si on a beaucoup de données, on est pratiquement certains que la méthode du maximum de vraisemblance est la meilleure méthode d'estimation possible. C'est pourquoi cette méthode est considérée comme globalement la meilleure et est utilisée de préférence à toute autre méthode, y compris celle des moments.

6.2 Intervalles de confiance

Un intervalle de confiance est un ensemble de valeurs vraisemblables pour θ, qu'il est logique de prendre proches de $\hat{\theta}_n$. Dire que toutes les valeurs de cet intervalle sont vraisemblables pour θ, c'est dire qu'il y a une forte probabilité que θ appartienne à cet intervalle.

Définition 20 *Un* **intervalle de confiance** *de* **seuil** *(ou* **niveau de signification***)* $\alpha \in [0,1]$ *pour un paramètre* θ*, est un intervalle aléatoire* I *tel que*

$$P(\theta \in I) = 1 - \alpha.$$

α est la probabilité que le paramètre θ n'appartienne pas à l'intervalle I, c'est à dire la probabilité que l'on se trompe en affirmant que $\theta \in I$. C'est donc une probabilité d'erreur, qui doit être assez petite. Les valeurs usuelles de α sont 10%, 5%, 1%, etc. Si I est un intervalle de confiance de seuil 5% pour θ, cela signifie qu'on a une confiance de 95% dans le fait que θ est dans I.

Pour trouver un intervalle de confiance, il existe plusieurs méthodes. La plus efficace consiste à chercher une **fonction pivotale**, c'est à dire une variable aléatoire fonction à la fois du paramètre θ et des observations $X_1, \ldots, X_n$, dont la loi de probabilité ne dépend pas de θ.

Quand on n'a pas d'intervalle de confiance exact, on peut en général déterminer un **intervalle de confiance asymptotique** de seuil α, c'est-à-dire un intervalle I_n tel que

$$\lim_{n \to +\infty} P(\theta \in I_n) = 1 - \alpha.$$

Un intervalle de confiance asymptotique peut être obtenu grâce aux propriétés asymptotiques de l'estimateur de maximum de vraisemblance.

Proposition 5 *Un intervalle de confiance asymptotique de seuil* α *pour* θ *est :*

$$\left[\hat{\theta}_n - \frac{u_\alpha}{\sqrt{\mathcal{I}_n(\hat{\theta}_n)}}, \hat{\theta}_n + \frac{u_\alpha}{\sqrt{\mathcal{I}_n(\hat{\theta}_n)}} \right].$$

où u_α *est le quantile d'ordre* $1 - \alpha/2$ *de la loi* $\mathcal{N}(0,1)$ *: si* U *est de loi* $\mathcal{N}(0,1)$*,* $P(U \leq u_\alpha) = 1 - \alpha/2$ *(commande* `R : qnorm(1-alpha/2)`*).*

6.3 Tests d'hypothèses

6.3.1 Principe des tests d'hypothèses

Un **test d'hypothèses** est une méthode qui permet de trancher entre une hypothèse nulle H_0 et une hypothèse alternative H_1 au vu des observations $x_1, \ldots, x_n$.

L'erreur de 1ère espèce consiste à rejeter à tort H_0, c'est-à-dire décider H_1 alors que H_0 est vraie. L'erreur de 2ème espèce consiste à ne pas rejeter H_0, à tort (il est préférable de parler de "ne pas rejeter H_0" plutôt que "accepter H_0"), c'est-à-dire à décider H_0 alors que H_1 est vraie.

Étant donné qu'on ne peut pas limiter les deux risques d'erreur en même temps, on choisit de privilégier l'erreur de 1ère espèce. En pratique, cela signifie qu'on se fixe la probabilité d'erreur de 1ère espèce, notée α et appelée **seuil** ou **niveau de signification** du test.

Construire un test, c'est trouver sa région critique W, c'est-à-dire l'ensemble des valeurs des observations pour lesquelles on rejettera H_0. La probabilité maximale de rejeter à tort H_0 est alors

$$\alpha = \underset{H_0}{Sup}\, P\left((X_1, \ldots, X_n) \in W\right).$$

W est donc entièrement déterminée par la valeur de α et la loi de probabilité de $(X_1, \ldots, X_n)$ sous H_0. Pour déterminer W, on peut utiliser des méthodes spécifiques, comme le lemme de Neyman-Pearson. Mais le plus souvent, on construit un test à l'aide du simple bon sens.

En général, la région critique d'un test est de la forme

$$W = \{(x_1, \ldots, x_n); g(x_1, \ldots, x_n) > k_\alpha\}.$$

On a alors $\alpha = \underset{H_0}{Sup}\, P\left(g(X_1, \ldots, X_n) > k_\alpha\right) \in W$. Pour que cette probabilité puisse prendre une valeur connue, il faut que la loi de $g(X_1, \ldots, X_n)$ sous H_0 ne dépende pas du paramètre inconnu θ. On retrouve ici la notion de fonction pivotale déjà utilisée pour les intervalles de confiance. En fait, intervalles de confiance et tests d'hypothèses sont des notions duales.

6.3.2 Tests d'adéquation

La méthode des graphes de probabilité est très utile pour permettre de choisir un modèle adapté pour un jeu de données. Mais on a vu qu'elle était subjective (il faut juger si des points sont suffisamment alignés ou pas). Les tests d'adéquation permettent de résoudre le problème de manière objective.

Tester l'adéquation d'un échantillon $(x_1, \ldots, x_n)$ à une loi de probabilité donnée, c'est déterminer s'il est vraisemblable que $x_1, \ldots, x_n$ soient les réalisations de variables aléatoires $X_1, \ldots, X_n$ indépendantes et de cette loi.

On note F la fonction de répartition inconnue de l'échantillon, supposée ici continue. On distingue deux cas, suivant que l'on veut tester l'adéquation de l'échantillon à une loi de probabilité entièrement spécifiée ou à une famille de lois de probabilité.

- *Cas 1* : Test d'adéquation à une loi entièrement spécifiée.

  $$\text{Test de } H_0 : \text{“}F = F_0\text{”} \text{ contre } H_1 : \text{“}F \neq F_0\text{”}.$$

  Par exemple, on se demande si les observations sont issues d'une loi exponentielle de paramètre $\lambda = 2$..

- *Cas 2* : Test d'adéquation à une famille de lois de probabilité.

  $$\text{Test de } H_0 : \text{“}F \in \mathcal{F}\text{”} \text{ contre } H_1 : \text{“}F \notin \mathcal{F}\text{”}.$$

  Le plus souvent, la famille $\mathcal{F}$ est une famille paramétrée : $\mathcal{F} = \{F(.;\theta); \theta \in \Theta\}$. C'est le cas quand on se demande simplement si les observations sont issues d'une loi exponentielle, sans donner de valeur a priori au paramètre λ. Si le modèle de loi exponentielle est adopté, on pourra toujours estimer λ ultérieurement.

En théorie, on devrait toujours appliquer un test d'adéquation avant d'utiliser n'importe quel modèle probabiliste sur des données. En pratique, on ne le fait pas toujours, ce qui entraine parfois l'utilisation de modèles complètement erronés.

6.3.2.1. Cas d'une loi entièrement spécifiée

Rappelons que la fonction de répartition F est estimée par la fonction de répartition empirique F_n. La figure 6.1 donne un exemple de représentation simultanée de F et F_n. Les 2 courbes sont bien proches.

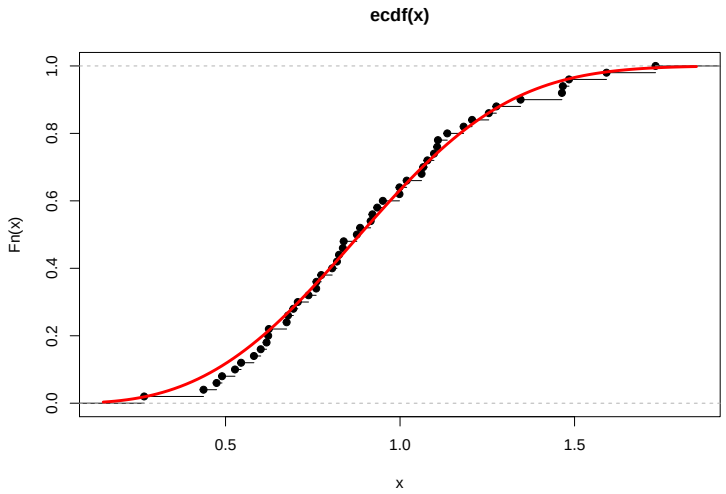


FIGURE 6.1 – Fonctions de répartition empirique et théorique

Par conséquent, quand on doit tester si “$F = F_0$”, il est logique de ne pas rejeter cette hypothèse si F_n et F_0 sont significativement proches.

Il s'agit donc de définir une distance, ou plutôt un écart, entre F_n et F_0, et de rejeter H_0 : “$F = F_0$” si cet écart est “trop grand”. Les mesures d'écart les plus usuelles sont :

- La **statistique de Kolmogorov-Smirnov** (KS) :

$$K_n = \sqrt{n} \sup_{x \in \mathbb{R}} |F_n(x) - F_0(x)| .$$

- La **statistique de Cramer-von Mises** (CM) :

$$W_n^2 = n \int_{-\infty}^{+\infty} [F_n(x) - F_0(x)]^2 \, dF_0(x).$$

- La **statistique d'Anderson-Darling** (AD) :

$$A_n^2 = n \int_{-\infty}^{+\infty} \frac{[F_n(x) - F_0(x)]^2}{F_0(x)\,(1 - F_0(x))} \, dF_0(x).$$

Un test de seuil α de H_0 : "$F = F_0$" contre H_1 : "$F \neq F_0$" aura donc une région critique de la forme $W = \{K_n > k_\alpha\}$, avec $\alpha = P_{H_0}(K_n > k_\alpha)$. Il faut donc connaitre la loi des variables aléatoires K_n, W_n^2 et A_n^2 sous H_0. Ces lois ne sont pas facilement accessibles pour n fini. En revanche, on a un résultat asymptotique.

Proposition 6 . *Sous H_0, K_n converge en loi vers la loi de Kolmogorov-Smirnov, de fonction de répartition :* $\forall z \in \mathbb{R}^+,\ F_{KS}(z) = 1 - 2\sum_{k=1}^{+\infty} (-1)^{k+1} e^{-2k^2 z^2}$.

Ce qui est remarquable dans ce résultat, c'est que la loi limite de K_n est la même, quelle que soit la loi de l'échantillon. C'est pour cela que l'on peut construire un test.

Ainsi, dans la région critique définie plus haut, k_α est le quantile d'ordre $1 - \alpha$ de la loi de Kolmogorov-Smirnov. Ce test est disponible en `R` grâce à la commande `ks.test`.

Pour les tests de Cramer-von Mises et Anderson-Darling, on montre que, sous H_0, W_n^2 et A_n^2 convergent aussi en loi vers des lois qui ne dépendent pas de F. En `R`, ces tests sont disponibles dans le package `goftest`.

Des études intensives prenant en compte un grand nombre d'alternatives possibles ont montré que, de manière générale, le test d'Anderson-Darling était le meilleur des trois et le test de Kolmogorov-Smirnov le moins bon.

6.3.2.2. Cas d'une famille de lois

On teste H_0 : "$F \in \mathcal{F} = \{F(.;\theta); \theta \in \Theta\}$" contre H_1 : "$F \notin \mathcal{F}$".

Puisque θ est un paramètre inconnu, une démarche naturelle est d'en déterminer un estimateur $\hat{\theta}(X_1, \ldots, X_n)$ et de calculer les statistiques K_n, W_n^2 et A_n^2 en remplaçant $F_0(x)$ par $F(x; \hat{\theta}(X_1, \ldots, X_n))$. On notera $\hat{K}_n$, $\hat{W}_n^2$ et $\hat{A}_n^2$ les statistiques correspondantes.

Malheureusement, le fait d'estimer θ entraine que les lois limites sous H_0 de $\hat{K}_n$, $\hat{W}_n^2$ et $\hat{A}_n^2$ ne sont pas les mêmes que celles de K_n, W_n^2 et A_n^2. Dans le cas général, les lois limites des statistiques de test dépendent de la loi testée F, de la procédure d'estimation utilisée (maximum de vraisemblance, moments, moindres carrés, ...), et de la vraie

valeur de θ. Contrairement au cas d'une loi entièrement spécifiée, on ne peut donc pas obtenir de test d'adéquation applicable dans tous les cas de figure.

Des résultats mathématiques permettent de déterminer dans quels cas on peut appliquer ces tests. On montre que c'est le cas de la loi exponentielle mais malheureusement pas de la loi de Weibull. On présentera plus tard comment construire quand même un test d'adéquation à la loi de Weibull.

En `R`, les packages `goftest` et `EWGoF` fournissent quelques uns des tests disponibles.

Chapitre 7

Analyse statistique d'échantillons complets de lois exponentielle et de Weibull

7.1 Loi exponentielle

On suppose ici que les observations $x_1, \ldots, x_n$ sont les réalisations de variables aléatoires $X_1, \ldots, X_n$ i.i.d. de loi exponentielle $\exp(\lambda)$. On va appliquer les résultats du chapitre précédent avec $\theta = \lambda$, $F(x;\lambda) = 1 - \exp(-\lambda x)$ et $f(x;\lambda) = \lambda \exp(-\lambda x)$.

7.1.1 Estimation de λ

On sait que l'espérance de la loi $\exp(\lambda)$ est $MTTF = E[X] = 1/\lambda$. Donc $\lambda = 1/E[X]$ et l'estimateur de λ par la méthode des moments (EMM) est $\tilde{\lambda}_n = 1/\bar{X}_n$.

La fonction de vraisemblance est :

$$\mathcal{L}(\lambda; x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i;\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

D'où $\ln \mathcal{L}(\lambda; x_1, \ldots, x_n) = n \ln \lambda - \lambda \sum_{i=1}^{n} x_i$.

Alors $\dfrac{\partial}{\partial \lambda} \ln \mathcal{L}(\lambda; x_1, \ldots, x_n) = \dfrac{n}{\lambda} - \sum_{i=1}^{n} x_i$, qui s'annule pour $\lambda = \dfrac{n}{\sum_{i=1}^{n} x_i} = \dfrac{1}{\bar{x}_n}$.

Par conséquent, l'estimateur de maximum de vraisemblance (EMV) de λ est

$$\hat{\lambda}_n = \frac{1}{\bar{X}_n}.$$

On constate donc que, pour la loi exponentielle, les deux méthodes donnent le même résultat. Ce ne sera pas le cas pour les autres lois usuelles en fiabilité.

Déterminons les propriétés de cet estimateur. Pour savoir si $\hat{\lambda}_n$ est sans biais, il faut

calculer $E[\hat{\lambda}_n] = E\left[\frac{1}{\bar{X}_n}\right] = E\left[\frac{n}{\sum_{i=1}^{n} X_i}\right]$.

D'après la proposition 1, on sait que $\sum_{i=1}^{n} X_i$ est de loi $G(n,\lambda)$, donc $E[\hat{\lambda}_n] = E\left[\frac{n}{Y}\right]$, où Y est de loi $G(n,\lambda)$. On obtient :

$$\begin{aligned} E[\hat{\lambda}_n] &= E\left[\frac{n}{Y}\right] = \int_0^{+\infty} \frac{n}{y}\,\frac{\lambda^n}{(n-1)!}\exp(-\lambda y)\,y^{n-1}\,dy \\ &= \frac{n\lambda^n}{(n-1)!}\int_0^{+\infty}\exp(-\lambda y)\,y^{n-2}\,dy = \frac{n\lambda}{n-1}\int_0^{+\infty}\frac{\lambda^{n-1}}{(n-2)!}\exp(-\lambda y)\,y^{n-2}\,dy \\ &= \frac{n\lambda}{n-1}\int_0^{+\infty} f_{G(n-1,\lambda)}(y)\,dy = \frac{n\lambda}{n-1} \end{aligned}$$

$E[\hat{\lambda}_n] \neq \lambda$, donc $\hat{\lambda}_n$ est un estimateur biaisé de λ. En revanche, il est clair que $\hat{\lambda}'_n = \frac{n-1}{n}\hat{\lambda}_n = \frac{n-1}{\sum_{i=1}^{n} X_i}$ est un estimateur sans biais de λ.

Un calcul analogue au précédent montre que $Var[\hat{\lambda}'_n] = \frac{\lambda^2}{n-2}$. Cette variance tend vers 0 quand n tend vers l'infini, donc $\hat{\lambda}'_n$ est un estimateur convergent de λ.

Par ailleurs, la quantité d'information est

$$\mathcal{I}_n(\lambda) = Var\left[\frac{\partial}{\partial\lambda}\ln\mathcal{L}(\lambda;X_1,\ldots,X_n)\right] = Var\left[\frac{n}{\lambda} - \sum_{i=1}^{n} X_i\right] = Var[\sum_{i=1}^{n} X_i] = nVar[X] = \frac{n}{\lambda^2}$$

On constate que $Var[\hat{\lambda}'_n] > 1/\mathcal{I}_n(\theta)$, ce qui ne permet pas de conclure directement sur l'optimalité de $\hat{\lambda}'_n$. Mais en fait, on peut montrer qu'il n'existe aucun estimateur sans biais dont la variance est égale à la borne de Cramer-Rao, et que c'est $\hat{\lambda}'_n$ qui a la variance minimale. D'où le résultat :

Proposition 7 $\hat{\lambda}'_n = \frac{n-1}{\sum_{i=1}^{n} X_i}$ *est l'estimateur sans biais et de variance minimale (ESBVM) de* λ.

7.1.2 Intervalles de confiance pour λ

On cherche un intervalle de confiance de seuil α pour λ, c'est-à-dire un intervalle I fonction des X_i tel que $P(\lambda \in I) = 1-\alpha$. On a dit dans le chapitre précédent que le meilleur moyen de le trouver est de chercher une fonction pivotale pour λ, c'est-à-dire une fonction de λ et des observations $X_1,\ldots,X_n$ dont la loi de probabilité ne dépend pas de λ.

En utilisant les propriétés des lois exponentielle, gamma et chi-deux vues au chapitre 3, on obtient que $\sum_{i=1}^{n} X_i$ est de loi gamma $G(n,\lambda)$, puis que $2\lambda\sum_{i=1}^{n} X_i$ est de loi $G(n,1/2) = \chi^2_{2n}$. Donc $2\lambda\sum_{i=1}^{n} X_i$ est une fonction pivotale pour λ.

Pour $0 < a < b$, on a

$$P(a < 2\lambda \sum_{i=1}^{n} X_i < b) = F_{\chi^2_{2n}}(b) - F_{\chi^2_{2n}}(a) = P\left(\frac{a}{2\sum_{i=1}^{n} X_i} < \lambda < \frac{b}{2\sum_{i=1}^{n} X_i} \right)$$

Il y a une infinité de façons possibles de choisir a et b de sorte que cette probabilité soit égale à $1 - \alpha$. La façon la plus usuelle de procéder est d'"équilibrer les risques", c'est-à-dire de prendre $a = F^{-1}_{\chi^2_{2n}}(\alpha/2) = z_{2n,1-\alpha/2}$ et $b = F^{-1}_{\chi^2_{2n}}(1 - \alpha/2) = z_{2n,\alpha/2}$.

On note usuellement $z_{n,\alpha} = F^{-1}_{\chi^2_n}(1 - \alpha)$ (commande `R : qchisq(1-alpha,n)`). D'où le résultat :

Proposition 8 *Un intervalle de confiance bilatéral de seuil α pour λ est*

$$\left[\frac{z_{2n,1-\frac{\alpha}{2}}}{2\sum_{i=1}^{n} X_i} ; \frac{z_{2n,\frac{\alpha}{2}}}{2\sum_{i=1}^{n} X_i} \right]$$

Remarques : La largeur de l'intervalle de confiance est $\dfrac{z_{2n,\frac{\alpha}{2}} - z_{2n,1-\frac{\alpha}{2}}}{2n\bar{X}_n}$.

1. Quand n grandit, on peut supposer que $\bar{X}_n$ varie peu puisque c'est une (bonne) estimation de $E[X]$. On peut voir dans les tables du χ^2 ou avec `R` que n croit plus vite que $z_{2n,\frac{\alpha}{2}} - z_{2n,1-\frac{\alpha}{2}}$. Donc plus n est important, plus l'intervalle de confiance est étroit. C'est logique, car plus on dispose d'observations, plus on a d'information sur le phénomène, donc meilleure est la précision des estimateurs.
2. Par contre, plus α diminue, plus l'intervalle de confiance est large. A la limite, l'intervalle de confiance est $\mathbb{R}^+$ tout entier pour $\alpha = 0$. C'est encore logique : moins on accepte de prendre de risque sur l'estimation de λ, plus on donnera de valeurs vraisemblables pour lui. Si on dit que toutes les valeurs sont vraisemblables, on est ainsi sûr de ne pas se tromper, mais cela n'a aucun intérêt.

Comme vu dans le chapitre précédent, un intervalle de confiance asymptotique de seuil α pour λ est :

$$\left[\hat{\lambda}_n - \frac{u_\alpha}{\sqrt{\mathcal{I}_n(\hat{\lambda}_n)}}, \hat{\lambda}_n + \frac{u_\alpha}{\sqrt{\mathcal{I}_n(\hat{\lambda}_n)}} \right].$$

Pour la loi exponentielle, $\mathcal{I}_n(\lambda) = n/\lambda^2$. On obtient donc qu'un intervalle de confiance asymptotique de seuil α pour λ est :

$$\left[\hat{\lambda}_n \left(1 - \frac{u_\alpha}{\sqrt{n}} \right), \hat{\lambda}_n \left(1 + \frac{u_\alpha}{\sqrt{n}} \right) \right].$$

On peut facilement vérifier que quand n est très grand, les intervalles de confiance exact et asymptotique sont équivalents.

7.1.3 Tests d'hypothèses sur λ

En fiabilité, il est normal de vouloir vérifier si un système est "suffisamment fiable" c'est-à-dire si son taux de défaillance λ est inférieur à une certaine valeur fixée λ_0, ou si son MTTF est supérieur à une certaine valeur fixée MTTF_0. Par exemple, certaines normes exigent que le taux de défaillance d'un système soit inférieur à $10^{-7}h^{-1}$. Dans le cas de la loi exponentielle, MTTF = $1/\lambda$, donc les deux questions sont identiques avec $\text{MTTF}_0 = 1/\lambda_0$.

Comme l'hypothèse sur laquelle on peut se prononcer est H_1, il s'agit de tester

$$H_0 : \text{“}\lambda \geq \lambda_0\text{”} \text{ contre } H_1 : \text{“}\lambda < \lambda_0\text{”}.$$

Étant donné que l'ESBVM de λ est $\hat{\lambda}'_n = (n-1)/\sum_{i=1}^{n} X_i$, il est logique de rejeter H_0 si et seulement si $\hat{\lambda}'_n$ est "suffisamment petit", c'est-à-dire si $\sum_{i=1}^{n} x_i$ ou $\bar{x}_n$ est "suffisamment grand". On proposera donc une région critique de la forme $W = \{\bar{x}_n > l_\alpha\}$.

Pour déterminer l_α, on va utiliser la fonction pivotale déjà vue dans la section précédente : $2\lambda \sum_{i=1}^{n} X_i = 2n\lambda\bar{X}_n$ est de loi χ^2_{2n}. On a alors :

$$\begin{aligned}
\alpha &= \sup_{H_0} P\left((X_1,\ldots,X_n) \in W\right) = \sup_{\lambda\geq\lambda_0} P\left(\bar{X}_n > l_\alpha\right) \\
&= \sup_{\lambda\geq\lambda_0} P\left(2n\lambda\bar{X}_n > 2n\lambda l_\alpha\right) = \sup_{\lambda\geq\lambda_0}\left[1 - F_{\chi^2_{2n}}(2n\lambda l_\alpha)\right] \\
&= 1 - F_{\chi^2_{2n}}(2n\lambda_0 l_\alpha)
\end{aligned}$$

car $1 - F_{\chi^2_{2n}}(2n\lambda l_\alpha)$ est une fonction décroissante de λ.

On obtient donc que $2n\lambda_0 l_\alpha = F^{-1}_{\chi^2_{2n}}(1-\alpha) = z_{2n,\alpha}$. Finalement, la région critique du test est :

$$W = \left\{\bar{x}_n > \frac{z_{2n,\alpha}}{2n\lambda_0}\right\} = \left\{2n\lambda_0\bar{x}_n > z_{2n,\alpha}\right\}.$$

De la même manière, on peut construire les tests de

$$H_0 : \text{“}\lambda \leq \lambda_0\text{”} \text{ contre } H_1 : \text{“}\lambda > \lambda_0\text{”}.$$

$$H_0 : \text{“}\lambda = \lambda_0\text{”} \text{ contre } H_1 : \text{“}\lambda \neq \lambda_0\text{”}.$$

7.1.4 Tests d'adéquation à la loi exponentielle

7.1.4.1. Tests d'adéquation à une loi entièrement spécifiée

En `R`, la commande `ks.test` permet de faire un test d'adéquation à une loi entièrement spécifiée. Par exemple, `ks.test(x,"pexp",0.1)` teste l'adéquation de l'échantillon x à la loi $\exp(0.1)$.

Exemple 1.

```
> x<-rexp(100,0.1)
```

```
> ks.test(x,"pexp",0.1)

One-sample Kolmogorov-Smirnov test

data:  x
D = 0.052849, p-value = 0.9428
alternative hypothesis: two-sided
```

Quand l'échantillon est vraiment de loi $\exp(0.1)$, la p-valeur est très élevée, donc on ne rejette pas l'hypothèse que les données sont de loi $\exp(0.1)$.

Exemple 2.

```
> x<-rexp(100,0.5)
> ks.test(x,"pexp",0.1)

One-sample Kolmogorov-Smirnov test

data:  x
D = 0.58578, p-value < 2.2e-16
alternative hypothesis: two-sided

> x<-rweibull(100,shape=11,scale=3)
> ks.test(x,"pexp",0.1)

One-sample Kolmogorov-Smirnov test

data:  x
D = 0.3395, p-value = 1.949e-10
alternative hypothesis: two-sided
```

Quand l'échantillon n'est pas de loi $\exp(0.1)$, la p-valeur est très faible, donc on rejette l'hypothèse que les données sont de loi $\exp(0.1)$.

7.1.4.2. Tests d'adéquation à une famille de lois

La loi exponentielle vérifie les propriétés permettant d'appliquer les tests de Kolmogorov-Smirnov, Cramer-von Mises et Anderson-Darling d'adéquation à la loi exponentielle.

En `R`, il faut utiliser le package `EWGoF` (développé au LJK) et la commande `EDF_NS.test` avec comme paramétrage `type="KS"`, `"CM"` ou `"AD"`.

Exemple 1.

```
> library(EWGoF)
> x<-rexp(100,0.1)
> EDF_NS.test(x,type="KS")
```

```
Test of Kolmogorov-Smirnov for the Exponential distribution

data:  x
S = 0.40578, p-value = 0.985
sample estimates:
[1] 0.1313343
```

Quand l'échantillon est vraiment de loi exponentielle, la p-valeur est très élevée, donc on ne rejette pas l'hypothèse que les données sont de loi exponentielle.

Exemple 2.

```
> x<-rweibull(100,scale=11,shape=3)
> EDF_NS.test(x,type="AD")

Test of Anderson Darling for the Exponential distribution

data:  x
S = 21.038, p-value < 2.2e-16
sample estimates:
[1] 0.09607801[1]
```

Quand l'échantillon n'est pas de loi exponentielle, la p-valeur est très faible, donc on rejette l'hypothèse que les données sont de loi exponentielle.

On montre que, globalement, les meilleurs tests d'adéquation à la loi exponentielle sont les tests d'Anderson-Darling et de Cox-Oakes (CO), disponibles dans le package EWGoF.

7.2 Loi de Weibull

On suppose ici que les observations $x_1, \ldots, x_n$ sont les réalisations de variables aléatoires $X_1, \ldots, X_n$ i.i.d. de loi de Weibull $\mathcal{W}(\eta, \beta)$. On va appliquer les résultats du chapitre précédent avec $\theta = (\eta, \beta)$, $F(x; \eta, \beta) = 1 - \exp\Big(-\big(\frac{x}{\eta}\big)^{\beta}\Big)$ et $f(x; \eta, \beta) = \frac{\beta}{\eta}\left(\frac{x}{\eta}\right)^{\beta-1}$ $\exp\Big(-\big(\frac{x}{\eta}\big)^{\beta}\Big)$.

7.2.1 Estimation de η et β

7.2.1.1. Méthode des moments

Rappelons que $E[X] = \eta\,\Gamma\left(\frac{1}{\beta}+1\right)$ et $Var[X] = \eta^2\left[\Gamma\left(\frac{2}{\beta}+1\right) - \Gamma\left(\frac{1}{\beta}+1\right)^2\right]$. On a donc :

$$\frac{Var[X]}{E[X]^2} = \frac{\Gamma\left(\frac{2}{\beta}+1\right)}{\Gamma\left(\frac{1}{\beta}+1\right)^2} - 1$$

Par conséquent, l'EMM $\tilde{\beta}_n$ de β est solution de l'équation implicite

$$\frac{S_n^2}{\bar{X}_n^2} = \frac{\Gamma\left(\frac{2}{\tilde{\beta}_n}+1\right)}{\Gamma\left(\frac{1}{\tilde{\beta}_n}+1\right)^2} - 1$$

et l'EMM de η est

$$\tilde{\eta}_n = \frac{\bar{X}_n}{\Gamma\left(\frac{1}{\tilde{\beta}_n}+1\right)}$$

7.2.1.1. Méthode du maximum de vraisemblance

La fonction de vraisemblance est :

$$\mathcal{L}(\eta,\beta;x_1,\ldots,x_n) = \prod_{i=1}^n \frac{\beta}{\eta}\left(\frac{x_i}{\eta}\right)^{\beta-1} \exp\Big(-\Big(\frac{x_i}{\eta}\Big)^\beta\Big) = \frac{\beta^n}{\eta^{n\beta}}\prod_{i=1}^n x_i^{\beta-1}\,\exp\Big(-\frac{1}{\eta^\beta}\sum_{i=1}^n x_i^\beta\Big)$$

La log-vraisemblance est :

$$\ln\mathcal{L}(\eta,\beta;x_1,\ldots,x_n) = n\ln\beta - n\beta\ln\eta + (\beta-1)\sum_{i=1}^n \ln x_i - \frac{1}{\eta^\beta}\sum_{i=1}^n x_i^\beta$$

On a donc :

$$\frac{\partial}{\partial\eta}\ln\mathcal{L}(\eta,\beta;x_1,\ldots,x_n) = -\frac{n\beta}{\eta} + \frac{\beta}{\eta^{\beta+1}}\sum_{i=1}^n x_i^\beta$$

Cette quantité s'annule pour $\eta = \left[\frac{1}{n}\sum_{i=1}^n x_i^\beta\right]^{1/\beta}$.

Par ailleurs,

$$\frac{\partial}{\partial\beta}\ln\mathcal{L}(\eta,\beta;x_1,\ldots,x_n) = \frac{n}{\beta} - n\ln\eta + \sum_{i=1}^n \ln x_i - \frac{1}{\eta^{2\beta}}\left[\eta^\beta\sum_{i=1}^n x_i^\beta\ln x_i - (\ln\eta)\eta^\beta\sum_{i=1}^n x_i^\beta\right].$$

Quand on annule cette dérivée en prenant en compte l'expression de η en fonction de β, on obtient que l'EMV $\hat{\beta}_n$ de β est solution de l'équation implicite :

$$\frac{n}{\hat{\beta}_n} + \sum_{i=1}^{n} \ln X_i - n \frac{\sum_{i=1}^{n} X_i^{\hat{\beta}_n} \ln X_i}{\sum_{i=1}^{n} X_i^{\hat{\beta}_n}} = 0$$

et que l'EMV $\hat{\eta}_n$ de η est :

$$\hat{\eta}_n = \left[\frac{1}{n} \sum_{i=1}^{n} X_i^{\hat{\beta}_n} \right]^{1/\hat{\beta}_n}$$

Par conséquent, pour la loi de Weibull, les EMM et les EMV sont différents. L'EMM et l'EMV de β n'ont pas d'expressions explicites, il faut donc les déterminer par des méthodes d'optimisation numérique. En `R`, on utilisera les commandes `uniroot` ou `optim`.

Le fait que ces estimateurs n'aient pas d'expressions explicites rend difficile la détermination de leurs propriétés. En particulier, on ne peut pas calculer exactement leur biais, on ne peut que l'évaluer numériquement. Les EMM et EMV sont biaisés et on ne peut pas facilement les débiaiser.

Il existe de nombreuses autres méthodes pour estimer les paramètres de la loi de Weibull. Elles nécessitent en général le recours à des tables de constantes dépendant de la taille de l'échantillon. Au final, on considère en général que la meilleure méthode est le maximum de vraisemblance. On sait que cette méthode est asymptotiquement optimale.

7.2.2 Intervalles de confiance asymptotiques

Comme précédemment, le fait que les estimateurs des paramètres n'aient pas d'expression explicite ne permet pas de déterminer facilement des intervalles de confiance exacts. Il existe des méthodes d'approximation d'intervalles de confiance exacts qui, là encore, nécessitent le recours à des tables de constantes. Aussi, on préfère en général donner des intervalles de confiance asymptotiques qui, eux, ont des expressions explicites.

Pour cela, on utilise les propriétés asymptotiques de l'estimateur de maximum de vraisemblance de $\theta = (\eta, \beta)$, vues dans le chapitre 6. Il faut calculer la matrice d'information de Fisher et l'inverser. On obtient :

$$\sqrt{n} \begin{pmatrix} \hat{\eta}_n - \eta \\ \hat{\beta}_n - \beta \end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}_2\left(0, \mathcal{I}_1^{-1}(\eta, \beta)\right)$$

où $\mathcal{I}_1^{-1}(\eta, \beta) = \begin{pmatrix} \frac{\eta^2}{\beta^2}\left(1 + \frac{6(1-\gamma_E)^2}{\pi^2}\right) & \eta\frac{6(1-\gamma_E)}{\pi} \\ \eta\frac{6(1-\gamma_E)}{\pi} & \beta^2\frac{6}{\pi^2} \end{pmatrix}$ et $\gamma_E = 0.5772156$ est la constante d'Euler.

Cela permet de déduire des intervalles de confiance asymptotiques de seuil α pour respectivement η et β :

$$\left[\hat{\eta}_n - \frac{u_\alpha}{\sqrt{n}}\frac{\hat{\eta}_n}{\hat{\beta}_n}\sqrt{1+\frac{6(1-\gamma_E)^2}{\pi^2}}, \hat{\eta}_n + \frac{u_\alpha}{\sqrt{n}}\frac{\hat{\eta}_n}{\hat{\beta}_n}\sqrt{1+\frac{6(1-\gamma_E)^2}{\pi^2}}\right]$$

$$\left[\hat{\beta}_n - \frac{u_\alpha}{\sqrt{n}}\frac{\hat{\beta}_n\sqrt{6}}{\pi}, \hat{\beta}_n + \frac{u_\alpha}{\sqrt{n}}\frac{\hat{\beta}_n\sqrt{6}}{\pi}\right]$$

7.2.3 Tests d'adéquation

Comme dit dans le chapitre 6, il n'est pas possible d'appliquer directement les tests de Kolmogorov-Smirnov, Cramer-von Mises et Anderson-Darling d'adéquation à la famille des lois de Weibull, car la loi de Weibull ne vérifie pas les conditions d'application.

Heureusement, il existe une solution grâce à une transformation logarithmique : si X est de loi de Weibull $\mathcal{W}(\eta, \beta)$, $\ln X$ est de loi des valeurs extrêmes $\mathcal{EV}_1(\ln\eta, 1/\beta)$. Or la loi des valeurs extrêmes vérifie la condition d'application des tests. Par conséquent, on peut tester l'adéquation des x_i à la loi de Weibull en testant l'adéquation des $\ln x_i$ à la loi des valeurs extrêmes, par exemple avec un test d'Anderson-Darling.

Mais ces tests ne sont pas les meilleurs. On montre que, globalement, les meilleurs tests d'adéquation à la loi de Weibull sont les tests de Tiku-Singh (TS) et d'Öztürk-Korukogu (OK), disponibles dans le package EWGoF.

Exemple 1.

```
> x<-rweibull(100,shape=11,scale=3)
> WNS.test(x, type = "TS")

Test of Tiku and Singh for the Weibull distribution

data:  x
S = 1.0646, p-value = 0.21
sample estimates:
      eta      beta
11.312596  2.992414
```

Pour des données de loi de Weibull, la p-valeur est élevée, donc on ne rejette pas la loi de Weibull.

Exemple 2.

```
> x<-rgamma(100,scale=2,shape=10)
> WPP.test(x, type = "OK")

Test of Ozturk and Koruglu for the Weibull distribution
```

```
data:  x
S = 2.3633, p-value = 0.02
sample estimates:
      eta      beta
21.450802  3.232768
```

Pour des données de loi gamma, la p-valeur est faible, donc on rejette la loi de Weibull. Notons ici qu'une loi gamma est assez proche d'une loi de Weibull et que le test OK a quand même réussi à différencier ces deux lois.

Chapitre 8

Analyse de données censurées

8.1 Introduction

Les échantillons que l'on traite en fiabilité sont généralement issus d'**essais de fiabilité**. Les conditions expérimentales de l'essai sont déterminées par un **plan d'essai**. Un plan d'essai va fixer à l'avance le nombre d'appareils à tester, le critère d'arrêt du test, déterminer si les appareils défaillants sont remplacés au fur et à mesure, etc.

Lorsque le plan d'essai nous amène à stopper l'expérience avant que tous les systèmes testés soient tombés en panne, on dit que les données obtenues sont **censurées**. Si les critères d'arrêt sont les mêmes pour tous les systèmes testés, on dit que l'on a affaire à une **censure simple**. Dans le cas contraire, on parle de **censure multiple** ou **multicensure**.

Par exemple, lorsqu'on traite de données provenant de retours d'expériences (REX), la censure provient du fait qu'on ne sait pas forcément quand les appareils ont commencé à fonctionner, et beaucoup restent encore en fonctionnement lorsqu'on fait l'analyse.

Si le fonctionnement des systèmes dépend de facteurs externes (température, vibrations, etc.), les conditions de fonctionnement doivent être identiques pour tous les systèmes en essai. En recommençant l'essai sous diverses contraintes, on pourra alors déterminer l'influence de l'environnement.

Pour des systèmes très fiables, on risque d'attendre longtemps les défaillances, ce qui coûte très cher. On utilise donc parfois des **plans d'essais accélérés**. Ceux-ci consistent à "durcir" les conditions d'environnement (ce qu'on appelle aussi le **stress**), de façon à "user" le système plus vite qu'en conditions normales, et faire en sorte que les pannes se produisent plus rapidement. Il reste alors à extrapoler les résultats au cas d'un niveau de stress normal ; cela nécessite de bien connaître la manière dont l'environnement influe sur la fiabilité d'un système, via des lois d'accélération.

Dans ce qui suit, on suppose que l'on teste n systèmes de durées de vie respectives $X_1, \ldots, X_n$ indépendantes et de même loi, de fonction de répartition $F(x;\theta)$ et de densité $f(x;\theta)$. Le fait que les données soient incomplètes a un impact sur l'estimation du paramètre θ.

On définira les cas les plus fréquents de censure et on déterminera la fonction de vraisemblance de l'échantillon dans chaque cas, ce qui permettra d'estimer θ. Il existe de la censure à gauche (on ne sait pas quand un système a commencé à fonctionner),

mais nous ne parlerons que de censure à droite (on ne sait pas quand un système tombera en panne). Il existe des plans d'essais où l'on remplace les systèmes tombés en panne, mais on ne considèrera que le cas où les systèmes en panne ne sont pas remplacés. On peut prendre en compte des systèmes mis en marche à des dates différentes, mais on ne s'intéressera qu'au cas où les systèmes sont tous mis en marche en même temps à la date 0 (ou on translatera les données pour se ramener à ce cas).

8.2 Les différents types de censure

8.2.1 Plan censuré de type 1

Le plan d'essais est censuré de type 1 lorsqu'on décide de stopper les essais au bout d'une durée c fixée à l'avance. Le nombre de systèmes tombés en panne à l'instant c est une variable aléatoire R. Comme les durées de vie des n systèmes sont indépendantes et que chaque système a une probabilité $P(X < c) = F(c;\theta)$ de tomber en panne entre 0 et c, R est de loi binomiale $\mathcal{B}(n, F(c;\theta))$.

Les observations sont la valeur observée r de R et les dates de panne des r systèmes tombés en panne avant c. Ces dates ne sont autres que les r premières statistiques d'ordre de l'échantillon $x_1, \ldots, x_n$, que l'on note usuellement $x_1^* \leq \cdots \leq x_r^*$.

On sait donc qu'il y eu r pannes entre 0 et c, que les dates de ces r pannes sont $x_1^* \leq \cdots \leq x_r^*$, et que $n - r$ systèmes sont toujours en vie à la date c. Cela permet d'établir que la fonction de vraisemblance pour un plan censuré de type 1 est :

$$\mathcal{L}(\theta; r, x_1^*, \ldots, x_r^*) = \frac{n!}{(n-r)!} \left[\prod_{i=1}^{r} f(x_i^*;\theta)\right] [1 - F(c;\theta)]^{n-r}$$

8.2.2 Plan censuré de type 2

Le plan d'essais est censuré de type 2 lorsqu'on décide que l'on stoppera les essais au moment où surviendra la $r^{\text{ème}}$ défaillance. Cette fois-ci, le nombre de systèmes en panne au moment de l'arrêt des tests est une constante r connue et la durée des essais X_r^* est aléatoire.

Les r premières défaillances ont lieu aux instants $x_1^* \leq \cdots \leq x_r^*$, et les $n - r$ systèmes restant ont une durée de vie supérieure à x_r^*. Cela permet de montrer que la fonction de vraisemblance pour un plan censuré de type 2 est :

$$\mathcal{L}(\theta; x_1^*, \ldots, x_r^*) = \frac{n!}{(n-r)!} \left[\prod_{i=1}^{r} f(x_i^*;\theta)\right] [1 - F(x_r^*;\theta)]^{n-r}$$

8.2.3 Plan multicensuré

Les deux cas précédents sont des cas de censure simple, c'est-à-dire où le critère de censure est le même pour tous les systèmes testés. Le plan d'essais est multicensuré lorsque les critères de censure sont différents suivant les systèmes. On considèrera ici que la date de censure du $i^{\text{ème}}$ système est c_i.

Lors d'essais de fiabilité, on n'a pas de raison d'imposer des dates de censure différentes pour chaque système. En revanche, la multicensure correspond plus à l'observation de systèmes en exploitation, pour lesquels les c_i sont les dates où l'on procède aux mesures, contrôles ou inspections.

Parmi les n systèmes testés, un nombre aléatoire R d'entre eux vont tomber en panne avant leurs instants de censure, aux instants $t_1 \leq t_2 \leq \ldots \leq t_r$. Notons que la multicensure entraîne que t_i n'est pas forcément égal à x_i^*. Soient $i_1, \ldots, i_r$ les indices des r systèmes tombés en panne. Les $n-r$ autres, d'indices $i_{r+1}, \ldots, i_n$, vont tomber en panne après leurs instants de censure : le système i_j tombe en panne après c_{i_j} pour $j \geq r+1$. Cela permet de montrer que la fonction de vraisemblance pour un plan multicensuré est :

$$\mathcal{L}(\theta; r, t_1, \ldots, t_r) = r! \left[\prod_{i=1}^{r} f(t_i;\theta)\right] \left[\prod_{j=r+1}^{n} \Big(1 - F(c_{i_j};\theta)\Big)\right]$$

8.2.4 Présentation unifiée

Dans les 3 plans d'essai précédents, on considère que, quand une défaillance est observée, on ne sait pas quel est le système qui est tombé en panne parmi les n systèmes testés. Si on fait l'hypothèse que l'on sait identifier quel est le système qui est tombé en panne (ou qui est censuré), cela change le point de vue et permet de faire une présentation unifiée des 3 cas précédents.

En effet, cela signifie que le $i^{\text{ème}}$ système a une durée de vie x_i et une date de censure c_i, et que l'on observe n couples (y_i, δ_i) avec :

- $\delta_i = 1$ si la durée de vie du $i^{\text{ème}}$ système est observée, et 0 si la durée de vie est censurée. Autrement dit, $\delta_i = \mathbb{1}_{\{x_i \leq c_i\}}$.
- $y_i = \min(x_i, c_i)$ est la durée observée pour le $i^{\text{ème}}$ système, qui est soit sa durée de vie soit sa date de censure.

Pour un plan censuré de type 1, $\forall i, c_i = c$. Pour un plan censuré de type 2, $\forall i, c_i = x_r^*$. Dans le premier cas, c_i est déterministe, dans le second cas, c'est la réalisation de la variable aléatoire X_r^*. Plus généralement, cette formulation permet de prendre en compte des dates de censures aléatoires quelconques C_i.

Le nombre total de défaillances est $r = \sum_{i=1}^{n} \delta_i$.

Alors la fonction de vraisemblance est

$$\mathcal{L}(\theta; y_1, \delta_1, \ldots, y_n, \delta_n) = \prod_{i=1}^{n} \left[f(y_i;\theta)\right]^{\delta_i} \left[1 - F(y_i;\theta)\right]^{1-\delta_i}$$

On voit que l'on retrouve bien les 3 vraisemblances des 3 plans d'essais, hormis les termes en factorielles. Ceux-ci disparaissent car toutes les observations y_i sont associées à un système bien identifié. Comme ces termes sont constants, ils n'interviennent pas dans la maximisation de la vraisemblance, donc cela ne change rien sur l'estimation de θ.

8.3 Analyse d'échantillons censurés de loi exponentielle

Pour la loi exponentielle $\exp(\lambda)$, $\theta = \lambda$, $F(x;\lambda) = 1-\exp(-\lambda x)$ et $f(x;\lambda) = \lambda\exp(-\lambda x)$. On suppose que l'on a un échantillon censuré de taille n de loi $\exp(\lambda)$ et on va estimer λ.

Pour simplifier les calculs, on adopte la présentation unifiée : le $i^{\text{ème}}$ système a une durée de vie X_i de loi $\exp(\lambda)$, et une date de censure C_i, qui peut être aléatoire ou déterministe. On observe n couples (Y_i, Δ_i) avec $Y_i = \min(X_i, C_i)$ et $\Delta_i = \mathbb{1}_{\{X_i \leq C_i\}}$.

La vraisemblance est :

$$\begin{aligned}
\mathcal{L}(\lambda; y_1, \ldots, \delta_n) &= \prod_{i=1}^{n} \left[\lambda\exp(-\lambda y_i)\right]^{\delta_i} \left[\exp(-\lambda y_i)\right]^{1-\delta_i} \\
&= \lambda^{\sum_{i=1}^{n}\delta_i} \exp\left(-\lambda\sum_{i=1}^{n} y_i\delta_i - \lambda\sum_{i=1}^{n} y_i(1-\delta_i)\right) \\
&= \lambda^r \exp(-\lambda\sum_{i=1}^{n} y_i)
\end{aligned}$$

D'où $\ln\mathcal{L}(\lambda; y_1, \ldots, \delta_n) = r\ln\lambda - \lambda\sum_{i=1}^{n} y_i$.

Alors $\dfrac{\partial}{\partial\lambda}\ln\mathcal{L}(\lambda; y_1, \ldots, \delta_n) = \dfrac{r}{\lambda} - \sum_{i=1}^{n} y_i$, qui s'annule pour $\lambda = \dfrac{r}{\sum_{i=1}^{n} y_i}$.

Par conséquent, l'estimateur de maximum de vraisemblance de λ est

$$\hat{\lambda}_n = \frac{R}{\sum_{i=1}^{n} Y_i} = \frac{\text{nombre de défaillances}}{\text{durée de fonctionnement cumulée}}$$

Regardons maintenant quelques cas particuliers.

8.3.1 Données complètes

Quand les données sont complètes, il n'y a pas de censure. Donc $R = n$ et $\forall i, Y_i = X_i$. On retrouve le résultat déjà vu que l'EMV de λ est $\hat{\lambda}_n = n/\sum_{i=1}^{n} X_i$.

8.3.2 Plan censuré de type 1

Dans un plan censuré de type 1, $\forall i, C_i = c$. Le nombre de défaillances R est aléatoire et de loi binomiale $\mathcal{B}(n, 1-\exp(-\lambda c))$. Les dates des défaillances sont $X_1^* \leq \cdots \leq X_R^*$, et il y a $n-R$ censures à la date c. On en déduit que l'EMV de λ est :

$$\hat{\lambda}_n = \frac{R}{\sum_{i=1}^{R} X_i^* + (n-R)c}$$

Le fait que R soit aléatoire rend difficile la détermination de la loi de $\hat{\lambda}_n$. On ne peut pas calculer le biais de l'estimateur, ni s'en servir pour déterminer un intervalle de confiance pour λ.

Mais comme R est de loi $\mathcal{B}(n, 1-\exp(-\lambda c))$, on a $E[R] = n(1-\exp(-\lambda c))$, d'où $\lambda = -\frac{1}{c}\ln\left(1-\frac{E[R]}{n}\right)$. Le principe de la méthode des moments permet d'en déduire qu'un nouvel estimateur de λ est

$$\tilde{\lambda}_n = -\frac{1}{c}\ln\left(1-\frac{R}{n}\right)$$

Aucun des deux estimateurs $\hat{\lambda}_n$ et $\tilde{\lambda}_n$ n'est sans biais et on ne peut pas les débiaiser facilement. Mais les propriétés de la loi binomiale permettent de montrer qu'un intervalle de confiance asymptotique de seuil α pour λ est :

$$\left[-\frac{1}{c}\ln\left(1-\frac{R}{n}+u_\alpha\sqrt{\frac{R(n-R)}{n^3}}\right), -\frac{1}{c}\ln\left(1-\frac{R}{n}-u_\alpha\sqrt{\frac{R(n-R)}{n^3}}\right)\right]$$

8.3.3 Plan censuré de type 2

Dans un plan censuré de type 2, le nombre de défaillances r est fixé et $\forall i, C_i = X_r^*$. Les dates des défaillances sont $X_1^* \leq \cdots \leq X_r^*$, et il y a $n-r$ censures à la date X_r^*. On en déduit que l'EMV de λ est :

$$\hat{\lambda}_n = \frac{r}{\sum_{i=1}^{r} X_i^* + (n-r)X_r^*}$$

Comme r n'est pas aléatoire, les propriétés de cet estimateur sont bien plus faciles à déterminer que dans le cas du plan censuré de type 1. On montre que :

- $\hat{\lambda}'_n = \dfrac{r-1}{\sum_{i=1}^{r} X_i^* + (n-r)X_r^*}$ est l'ESBVM de λ.
- $2\lambda\left(\sum_{i=1}^{r} X_i^* + (n-r)X_r^*\right)$ est de loi χ^2_{2r}.
- Un intervalle de confiance bilatéral de seuil α pour λ est

$$\left[\frac{z_{2r,1-\frac{\alpha}{2}}}{2\sum_{i=1}^{r} X_i^* + 2(n-r)X_r^*}; \frac{z_{2r,\frac{\alpha}{2}}}{2\sum_{i=1}^{r} X_i^* + 2(n-r)X_r^*}\right]$$

- Pour tout $x \geq 0$, l'ESBVM de $R(x)$ est

$$\hat{R}_x = \left(1-\frac{x}{\sum_{i=1}^{r} X_i^* + (n-r)X_r^*}\right)^{r-1} \mathbb{1}_{\{x \leq \sum_{i=1}^{r} X_i^* + (n-r)X_r^*\}}$$

On retrouve les propriétés obtenues pour des données complètes en prenant $r = n$. On voit donc que le cas des données censurées de type 2 est bien plus pratique d'un point de vue statistique que celui des données censurées de type 1.

8.4 Estimateur de Kaplan-Meier de la fiabilité

Quand les données sont censurées, il est plus complexe de choisir un modèle approprié pour un échantillon que quand les données sont complètes. Il existe des adaptations des graphes de probabilité et des tests d'adéquation au cas des données censurées, mais ces méthodes sont plus complexes et moins précises. Par conséquent, il est intéressant d'avoir des méthodes permettant de faire des analyses statistiques sans faire l'hypothèse que les observations proviennent d'une loi donnée. C'est ce qu'on appelle de la statistique non paramétrique.

On s'intéressera ici à l'estimation non paramétrique de la fiabilité. Quand les données sont complètes, c'est très simple. En effet, $R(x) = 1 - F(x)$ et un estimateur non paramétrique de la fonction de répartition est la fonction de répartition empirique $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}$. Par conséquent, un estimateur non paramétrique de la fiabilité est la **fiabilité empirique**, définie par :

$$R_n(x) = 1 - F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i > x\}} = \begin{cases} 1 & \text{si } x < X_1^* \\ \frac{n-i}{n} & \text{si } X_i^* \leq x < X_{i+1}^* \\ 0 & \text{si } x \geq X_n^* \end{cases}$$

Quand les données sont censurées, on ne peut pas calculer cet estimateur car toutes les statistiques d'ordre ne sont pas observées. L'estimateur de Kaplan-Meier a pour objectif d'estimer la fiabilité dans ce cas.

On cherche à estimer $R(x) = P(X > x)$. Considérons j dates entre 0 et x : $0 < t_1 < \ldots < t_j < x$. On a :

$$\begin{aligned} R(x) &= P(X > x) = P(X > x \cap X > t_j) = P(X > x | X > t_j) P(X > t_j) \\ &= P(X > x | X > t_j) P(X > t_j | X > t_{j-1}) P(X > t_{j-1}) \\ &= \ldots \\ &= P(X > x | X > t_j) P(X > t_j | X > t_{j-1}) \ldots P(X > t_1 | X > 0) P(X > 0) \\ &= P(X > x | X > t_j) \prod_{i=1}^{j} P(X > t_i | X > t_{i-1}) \text{ en posant } t_0 = 0 \end{aligned}$$

On considère des données censurées selon la présentation unifiée : le $i^{\text{ème}}$ système a une durée de vie X_i et une date de censure C_i. On observe n couples (Y_i, Δ_i) avec $Y_i = \min(X_i, C_i)$ et $\Delta_i = \mathbb{1}_{\{X_i \leq C_i\}}$. Les statistiques d'ordre des Y_i sont notées Y_i^*.

Pour un x donné, il existe un j tel que $x \in [Y_j^*, Y_{j+1}^*[$. On doit donc estimer $R(x) = P(X > x | X > Y_j^*) \prod_{i=1}^{j} P(X > Y_i^* | X > Y_{i-1}^*)$.

Dans les données, il n'y a pas de défaillance observée entre Y_j^* et Y_{j+1}^*. Donc on peut estimer $P(X > x|X > Y_j^*)$ par 1. De même, il est logique d'estimer $P(X > Y_i^*|X > Y_{i-1}^*)$ par le pourcentage de durées de vie supérieures à Y_i^* parmi celles supérieures à Y_{i-1}^*. Jusqu'à Y_{i-1}^*, il y a eu $i-1$ observations, qui sont des dates de défaillances ou des censures. On sait donc qu'il y a au moins $n-i+1$ systèmes qui tomberont en panne après Y_{i-1}^*. Parmi ceux-là, on n'a aucune défaillance observée entre Y_{i-1}^* et Y_i^*. Il y a une défaillance en Y_i^* si $\Delta_i = 1$ et une censure en Y_i^* si $\Delta_i = 0$. Donc on a soit une défaillance sur $n-i+1$ possibles entre Y_{i-1}^* et Y_i^*, soit 0. On peut donc estimer $P(X > Y_i^*|X > Y_{i-1}^*)$ par $1 - \dfrac{\Delta_i}{n-i+1} = \left(\dfrac{n-i}{n-i+1}\right)^{\Delta_i}$. D'où le résultat :

Définition 21 *L'***estimateur de Kaplan-Meier** *de la fiabilité est :*

$$R_{KM}(x) = \prod_{i;Y_i^* \leq x} \left(1 - \frac{\Delta_i}{n-i+1}\right) = \prod_{i;Y_i^* \leq x} \left(\frac{n-i}{n-i+1}\right)^{\Delta_i}$$

Pour les données complètes, $\forall i, \Delta_i = 1$ et $Y_j = X_j$. Donc $R_{KM}(x) = \prod_{i;X_i^* \leq x} \left(\dfrac{n-i}{n-i+1}\right) = \dfrac{n-j}{n}$, où j est le plus grand indice tel que $X_j^* \leq x$. On voit que dans ce cas, l'estimateur de Kaplan-Meier n'est autre que la fiabilité empirique.

En `R`, l'estimateur de Kaplan-Meier s'obtient grâce à la fonction `survfit` du package `survival`.

Exemple : On considère un jeu de $n = 15$ données, comprenant 11 durées de vie et 4 censures. Les censures sont indiquées par une étoile :

$$14.1^* \quad 11.8 \quad 3.5 \quad 8.3^* \quad 21.7 \quad 19.7 \quad 9.1 \quad 2.7 \quad 4.1^* \quad 2.6 \quad 3.7 \quad 4.6 \quad 5.0^* \quad 6.4 \quad 13.6$$

On commence par créer un dataframe pour ce jeu de données :

```
> y<-c(14.1, 11.8, 3.5, 8.3, 21.7, 19.7, 9.1, 2.7, 4.1, 2.6, 3.7,
4.6, 5.0, 6.4, 13.6)
> delta<-c(0,1,1,0,1,1,1,1,0,1,1,1,0,1,1)
> donnees<-data.frame(y,delta)
```

Puis on représente l'estimateur de Kaplan-Meier (avec des bandes de confiance) :

```
> library(survival)
> RKM<-survfit(Surv(y,delta) ~ 1, data=donnees)
> plot(RKM)
```

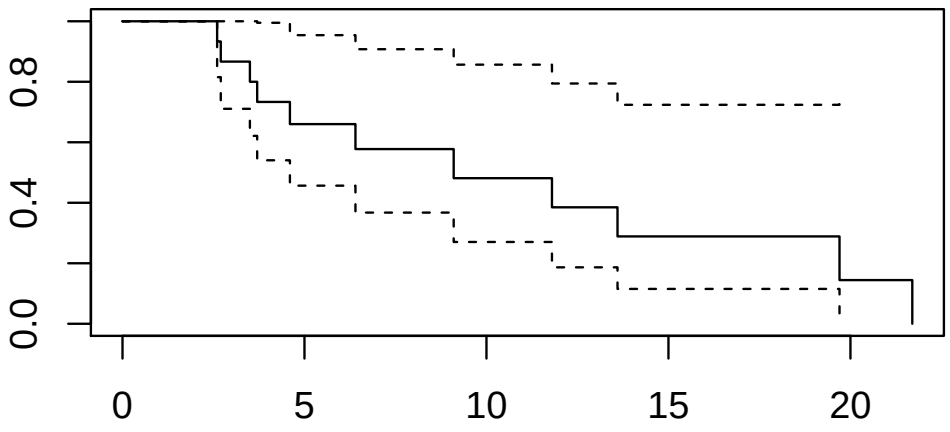


FIGURE 8.1 – Estimateur de Kaplan-Meier