**École doctorale d'astronomie et d'astrophysique d'Île de France.**
**I.A.P., février 2013**

**Post-master.**
**Approche statistique bayésienne par l'exemple**

# Analyse de survie

**Michel Fioc**

**(Michel.Fioc@iap.fr,**
**www2.iap.fr/users/fioc/enseignement/analyse_de_survie/)**

L'*analyse de survie* désigne un ensemble de techniques statistiques permettant de traiter des données soumises à une *censure*, c'est-à-dire de données dont l'on ne connaît, pour certaines d'entre elles, qu'une borne inférieure ou supérieure et non une valeur précise.

Elle est appelée *survival analysis* ou *lifetime data analysis* dans la littérature anglo-saxonne ; voir aussi *censoring*, *censored data*.

## I. Domaines d'application

En *médecine*, l'analyse de survie permet d'évaluer l'efficacité d'un traitement. On souhaite par exemple estimer la durée de survie probable d'un patient. On utilise pour cela un échantillon de malades dont on connaît, pour chacun d'eux,

- soit la vraie durée de survie (*donnée non censurée* ou *détection*),
- soit une borne inférieure de cette durée (*donnée censurée*).

Le deuxième cas se produit lorsqu'on perd la trace d'un patient, par exemple en raison d'un déménagement, ou lorsqu'il décède pour une cause indépendante.

En *démographie*, l'analyse de survie sert à construire des tables de mortalité. Celles-ci sont utilisées par les actuaires pour déterminer le montant des assurances-vie et rentes viagères, entre autres ; on parle de *tables actuarielles* quand les données sont regroupées dans des intervalles.

En *ingénierie*, l'analyse de survie permet d'estimer la fiabilité de machines, de composants électroniques. . .

L'analyse de survie est aussi utile en *astrophysique*. Imaginons que l'on ait détecté des sources à une longueur d'onde $\lambda$ et qu'on observe aux mêmes positions à une autre longueur d'onde, $\lambda'$. Certaines sources ne sont pas détectées à $\lambda'$ car le rapport signal/bruit, $\ell'/b$, est trop faible. Comment calculer alors la distribution des couleurs $m - m' = 2{,}5 \log_{10}(\ell'/\ell) + \mathrm{c^{te}}$ ?

## II. Notations et définitions

Reprenons l'exemple d'un traitement médical.

On considère une population de patients dont la durée de survie est décrite par une variable aléatoire $T$. Un échantillon de $n$ individus est tiré aléatoirement de cette population. Les variables aléatoires $T_i$ (avec $i \in [\![1, n]\!]$) décrivant la durée de survie des $n$ patients sont supposées indépendantes et identiquement distribuées (selon la loi de $T$).

La durée de vie effective du patient n$^{\mathrm{o}}$ $i$ est notée $t_i$. Elle est censurée si $t_i > s_i$. Posons

$$\tau_i = \min\{t_i, s_i\}$$

et

$$\delta_i = \mathbb{1}(t_i \leqslant s_i) = \begin{cases} 1 & \text{si } t_i \leqslant s_i \text{ (détection),} \\ 0 & \text{si } t_i > s_i \text{ (donnée censurée).} \end{cases}$$

Pour chaque patient, on connaît $\tau_i$ et $\delta_i$. On connaît aussi le seuil $s_i$ pour les données non censurées, mais cette information ne sera pas utilisée ici (elle pourrait l'être en revanche pour des simulations).

La probabilité de survie sera décrite par l'une des fonctions suivantes :

- *Fonction de répartition* (ou *probabilité cumulée*) $F(t)$ :

$$F(t) = \mathbb{P}(T \leqslant t).$$

- *Densité de probabilité* $f(t)$ :

$$f(t)\,\mathrm{d}t = \mathbb{P}(t \leqslant T < t + \mathrm{d}t), \quad \text{soit} \quad f(t) = \frac{\mathrm{d}F}{\mathrm{d}t}.$$

- *Fonction de survie* $S(t)$ :

$$S(t) = 1 - F(t) = \mathbb{P}(T > t).$$

On rencontre parfois aussi la *fonction de hasard*, $h(t) = f(t)/S(t)$.

On va chercher des estimateurs de ces quantités à partir d'un échantillon de taille $n$. On adoptera les notations suivantes :

- $\tilde{X}$ ou $\tilde{X}_n$ : estimateur de la variable aléatoire $X$ en l'absence de censure.
- $\hat{X}$ ou $\hat{X}_n$ : estimateur de $X$ dérivé de l'estimateur de Kaplan & Meier.
- $\check{X}$ ou $\check{X}_n$ : autre estimateur de $X$ en présence de censure.

Si forme de la loi est supposée connue (loi normale par exemple) mais que ses paramètres (moyenne, dispersion) sont inconnus, on procède à une *estimation paramétrique*. Si la forme de la loi elle-même est à déterminer, il s'agit d'une *estimation non paramétrique*.

L'estimation paramétrique est plus facile à réaliser (par exemple en maximisant la vraisemblance ; cf. § V) ; elle permet d'extrapoler hors du domaine des valeurs observées et de prendre en compte les incertitudes observationnelles ; elle est plus efficace si la forme de la loi de probabilité adoptée est correcte (cf. Miller (1983)). Sinon, l'estimation non paramétrique est préférable, mais les estimateurs non paramétriques sont plus difficiles à construire.

# III. Types de censure

## III.1. Nature de la censure

**Censure à droite** : on ne connaît pour certains $i$ qu'une borne inférieure de $t_i$ (cf. exemple médical).

**Censure à gauche** : on ne connaît pour certains $i$ qu'une borne supérieure de $t_i$ (cf. exemple astrophysique).

Il suffit de remplacer $\tau_i$ par $-\tau_i$ pour obtenir une censure à droite, ou mieux, par

$$\max_{j \in [\![1,\, n]\!]} \{\tau_j\} - \tau_i$$

pour que les $\tau_i$ soient tous positifs, ce qu'on supposera par la suite pour l'estimation non paramétrique.

**Censure par intervalle** : on ne connaît pour certains $i$ qu'une borne supérieure et une borne inférieure de $t_i$ (non traité ici ; cf. littérature). Ceci se produit notamment si un patient se rend à l'hôpital à des dates régulières : s'il ne se présente pas à un rendez-vous, on sait seulement que son décès s'est produit dans l'intervalle entre la dernière visite et le rendez-vous.

## III.2. Processus de censure

La terminologie est passablement fluctuante et confuse.

**Censure de type I** : les seuils de détection sont fixés a priori. Si le seuil est commun à toutes les valeurs, on parle de *censure simple* ; sinon, de *censure progressive ou généralisée*.

Une étude limitée à une durée prédéfinie de l'efficacité d'un traitement médical appliqué, dès $t = 0$, à tous les patients d'un échantillon constitue un exemple de censure simple de type I.

Si des patients rentrent dans l'échantillon en cours d'étude, il s'agit d'une censure progressive de type I.

Dans l'exemple astrophysique, on a une censure simple de type I pour les luminosités $\ell'_i$ lorsque les sources détectées à la longueur d'onde $\lambda$ sont considérées comme non détectées à $\lambda'$ si $\ell'_i$ est inférieure à un seuil uniforme (par exemple, 3 $b$, où $b$ est le bruit dans le champ). À cause de l'éparpillement des magnitudes $m_i$, les couleurs $m_i - m'_i$ subissent une censure généralisée de type I.

**Censure aléatoire** (parfois qualifiée de censure de type III) : pour chaque élément $i$ de l'échantillon, le seuil est aléatoire et indépendant de $t_i$.

En pratique, la censure aléatoire est proche d'une censure généralisée de type I.

La censure simple de type I est un cas particulier de censure aléatoire dans le cas où les $s_i$ sont tous distribués de la même manière avec une dispersion nulle.

**Censure de type II** : on fixe a priori le nombre $k$ de détections. Toutes les valeurs non observées une fois ce nombre atteint sont censurées. Le seuil est donc fixé a posteriori.

Dans l'exemple médical, on arrête l'étude dès que le nombre de décès dépasse $k$.

Dans l'exemple astrophysique, on observe, à une longueur d'onde $\lambda'$, un champ où $n$ sources son déjà connues à $\lambda$. L'observation dure jusqu'à ce qu'un nombre $k$ de ces $n$ sources soient détectées à $\lambda'$ (ce qui finit par se produire, le rapport signal/bruit augmentant normalement avec le temps...).

Inconvénient : la durée totale d'observation n'est pas maîtrisée.

**Troncature** : l'existence même des individus non détectés est inconnue.

Dans l'exemple astrophysique, ceci se produit pour les sources qui ne sont détectées ni à la longueur d'onde $\lambda$, ni à $\lambda'$.

### III.3. Hypothèse de la censure non informative

On supposera la censure *non informative*, c'est-à-dire que

$$\mathbb{P}(t \leqslant T_i < t + \mathrm{d}t \mid T_i > \tau_i \text{ et } \delta_i = 0) = \mathbb{P}(t \leqslant T < t + \mathrm{d}t \mid T > s_i).$$

En clair, le fait que l'observation n° $i$ soit censurée ne dit rien d'autre que $t_i > \tau_i$ : on ne peut notamment pas en conclure que $t_i$ est « loin » ou « près » de $\tau_i$.

Exemple de censure informative : patients abandonnant un traitement médical car ils ont le sentiment, peut-être justifié, que leur état ne s'améliore pas.

## IV. Estimation non paramétrique

### IV.1. En l'absence de censure

#### IV.1.*a*. Méthode directe

Un estimateur empirique $\tilde{S}_n(t)$ de $S(t)$ est la fraction de valeurs $t_i$ strictement supérieures à $t$. Si les $t_i$ sont rangés par ordre croissant ($t_1 \leqslant t_2 \leqslant \ldots \leqslant t_n$), on a

$$\tilde{S}_n(t) = \begin{cases} 1 & \text{si } t < t_1, \\ (n-i)/n & \text{si } t \in [t_i, t_{i+1}[, \\ 0 & \text{si } t \geqslant t_n. \end{cases}$$

On peut aussi écrire

$$\tilde{S}_n(t) = \sum_{i=1}^{n} \frac{\mathbb{1}(t > t_i)}{n},$$

$$\tilde{F}_n(t) = \sum_{i=1}^{n} \frac{\mathbb{1}(t \leqslant t_i)}{n}$$

et

$$\tilde{f_n}(t) = \frac{\mathrm{d}\tilde{F}_n}{\mathrm{d}t} = \sum_{i=1}^{n} \frac{\delta(t - t_i)}{n},$$

où $\delta$ est la distribution de Dirac. Le poids de chaque donnée est donc $1/n$.

### IV.1.*b*. Produit des probabilités conditionnelles

On suppose ici les $t_j$ distincts.
Posons $t_0 = 0$, $p_0 = \mathbb{P}(T > t_0)$ $(= 1)$ et

$$\forall\ j \in [\![1, n]\!], \quad p_j \coloneqq \mathbb{P}(T > t_j \mid T > t_{j-1}) = \frac{\mathbb{P}([T > t_j] \text{ et } [T > t_{j-1}])}{\mathbb{P}(T > t_{j-1})} = \frac{\mathbb{P}(T > t_j)}{\mathbb{P}(T > t_{j-1})}.$$

Un estimateur de $p_j$ est

$$\tilde{p}_j = \frac{n - j}{n - j + 1}.$$

De même, si $t \in [t_i, t_{i+1}[$, un estimateur de $p_{i,\,t} \coloneqq \mathbb{P}(T > t \mid T > t_i)$ est

$$\tilde{p}_{i,\,t} = \frac{n - i}{n - i} = 1.$$

Or

$$S(t) = \mathbb{P}(T > t \mid T > t_i) \times \mathbb{P}(T > t_i \mid T > t_{i-1}) \times \ldots \times \mathbb{P}(T > 0).$$

Pour $t \in [t_i, t_{i+1}[$, un estimateur de $S(t)$ est donc

$$\tilde{S}_n(t) = \tilde{p}_{i,\,t} \prod_{j=1}^{i} \tilde{p}_j = \prod_{j=1}^{i} \frac{n - j}{n - j + 1} = \frac{n - i}{n}.$$

On retrouve bien le résultat obtenu par la méthode directe.

## IV.2. En présence de censure. Estimateur de Kaplan & Meier

On suppose ici que certaines données sont soumises à une censure non informative à droite (peu importe le processus).

### IV.2.*a*. Cas où les valeurs sont toutes distinctes

On a toujours

$$\hat{S}_n(t) = \sum_{i=1}^{n} \frac{\mathbb{1}(t > t_i)}{n},$$

mais les valeurs vraies $t_i$ sont désormais inconnues pour les données censurées. L'estimation directe est donc impossible [※1] (cf. § IV.2.*e*.ii cependant).

En revanche, si la méthode du produit des probabilités conditionnelles n'apporte rien en l'absence de censure, elle présente l'avantage de se généraliser au cas où des données sont censurées : on construit ainsi l'estimateur de Kaplan & Meier.

Supposons les $\tau_i$ distincts et rangés par ordre croissant. Comme précédemment,

$$\hat{S}_n(t) = \hat{p}_{i,\,t} \prod_{j=1}^{i} \hat{p}_j,$$

1. Notons qu'éliminer toutes les données censurées ou les considérer comme des détections revient à surestimer $F(t)$ ; au contraire, remplacer les valeurs censurées par la plus grande valeur, $\tau_n$, qu'elle soit détectée ou non, conduit à sous-estimer $F(t)$.

mais le calcul des $\hat{p}_j$ et de $\hat{p}_{i,t}$ doit être modifié pour prendre en compte les données censurées.

Si $j$ correspond à une détection, il y a, parmi les $t_{k,k\geqslant j}$ (les valeurs « à risque »), $n-j$ valeurs vraies $t_k$ dans $]\tau_j, \infty[$ (tous les $\tau_{k,k>j}$, censurés ou non) et $n-j+1$ valeurs vraies dans $]\tau_{j-1}, \infty[$ (les mêmes plus $\tau_j$). On a donc

$$\hat{p}_j = \frac{n-j}{n-j+1}.$$

Si la donnée n° $j$ est censurée, il y a $n-j+1$ valeurs vraies dans $]\tau_j, \infty[$ (car $t_j > \tau_j$) parmi les $k \geqslant j$, et autant dans $]\tau_{j-1}, \infty[$, donc

$$\hat{p}_j = 1.$$

**Remarque** : il est inutile de tenir compte des valeurs censurées $\tau_{k,k\leqslant j-1}$ : les $t_{k,k\leqslant j-1}$ peuvent certes tomber dans les intervalles $]\tau_{j-1}, +\infty[$ et $]\tau_j, +\infty[$, mais, si on en tenait compte, en raison de l'hypothèse de censure non informative, on devrait les affecter dans ces deux intervalles proportionnellement à $n-j$ et $n-j+1$ respectivement si $j$ n'est pas censuré, et à $n-j+1$ et $n-j+1$ sinon.

Que $j$ soit censuré ou non, on peut écrire

$$\hat{p}_j = \left(\frac{n-j}{n-j+1}\right)^{\delta_j}.$$

Enfin,

$$\hat{p}_{i,t} = 1 \quad \text{si } t \leqslant \tau_n.$$

$\hat{p}_{i,t}$ n'est en revanche pas défini si $t > \tau_n$ (« 0/0 »).

Au final,

$$\hat{S}_n(t) = \begin{cases} 1 & \text{si } t < \tau_1, \\ \displaystyle\prod_{j=1}^{i} \left(\frac{n-j}{n-j+1}\right)^{\delta_j} & \text{si } i \in [\![1, n-1]\!] \text{ et } t \in [\tau_i, \tau_{i+1}[, \\ \displaystyle\prod_{j=1}^{n-1} \left(\frac{n-j}{n-j+1}\right)^{\delta_j} & \text{si } t = \tau_n \text{ et } \delta_n = 0, \\ 0 & \text{si } t = \tau_n \text{ et } \delta_n = 1. \end{cases}$$

Si $t > \tau_n$, on sait que $0 \leqslant \hat{S}_n(t) \leqslant \hat{S}_n(\tau_n)$ car $S(t)$ est une fonction décroissante positive. Si $\delta_n = 1$, on a donc $\hat{S}_n(t) = 0$ pour tout $t > \tau_n$ ; si $\delta_n = 0$, $\hat{S}_n(t)$ n'est pas définie précisément.

**Remarque** : les $\hat{p}_j$ sont évidemment très incertains : pour obtenir des valeurs plus fiables, il faut regrouper les observations dans des intervalles assez larges ; on obtient alors les « estimateurs actuariels ». Nous allons voir en revanche que $\hat{S}_n(t)$ est un bon estimateur non paramétrique de $S(t)$ ; Kaplan & Meier (1958) l'ont appelé *product-limit estimator* car c'est la limite des estimateurs actuariels quand le nombre de valeurs par intervalle tend vers 1.

En l'absence de censure, on retrouve la fonction de survie empirique.

### IV.2.*b*. Cas général

Dans le cas où les $\tau_i$ ne sont pas distincts, on applique la procédure suivante.

Notons $n'$ le nombre de valeurs *distinctes* et $\tau'_{i,\, i\in[\![1,n']\!]}$ ces valeurs, rangées par ordre croissant. Soit $n_i$ le nombre de valeurs $\tau_j$ supérieures ou égales à $\tau'_i$, $d_i$ le nombre de $\tau_j$ détectés égaux à $\tau'_i$ et $c_i$ le nombre de $\tau_j$ censurés égaux à $\tau'_i$.

On a $n_1 = n$ et $n_j = n_{j-1} - d_{j-1} - c_{j-1}$ pour tout $j \geqslant 2$.

La procédure appliquée dans la section précédente reste valable en convenant que les valeurs censurées égales à $\tau'_i$ sont infinitésimalement supérieures aux valeurs détectées.

On obtient alors

$$\hat{S}_n(t) = \begin{cases} 1 & \text{si } t < \tau'_1, \\ \displaystyle\prod_{j=1}^{i} \frac{n_j - d_j}{n_j} & \text{si } i \in [\![1, n']\!] \text{ et } t \in [\tau'_i, \tau'_{i+1}[. \end{cases}$$

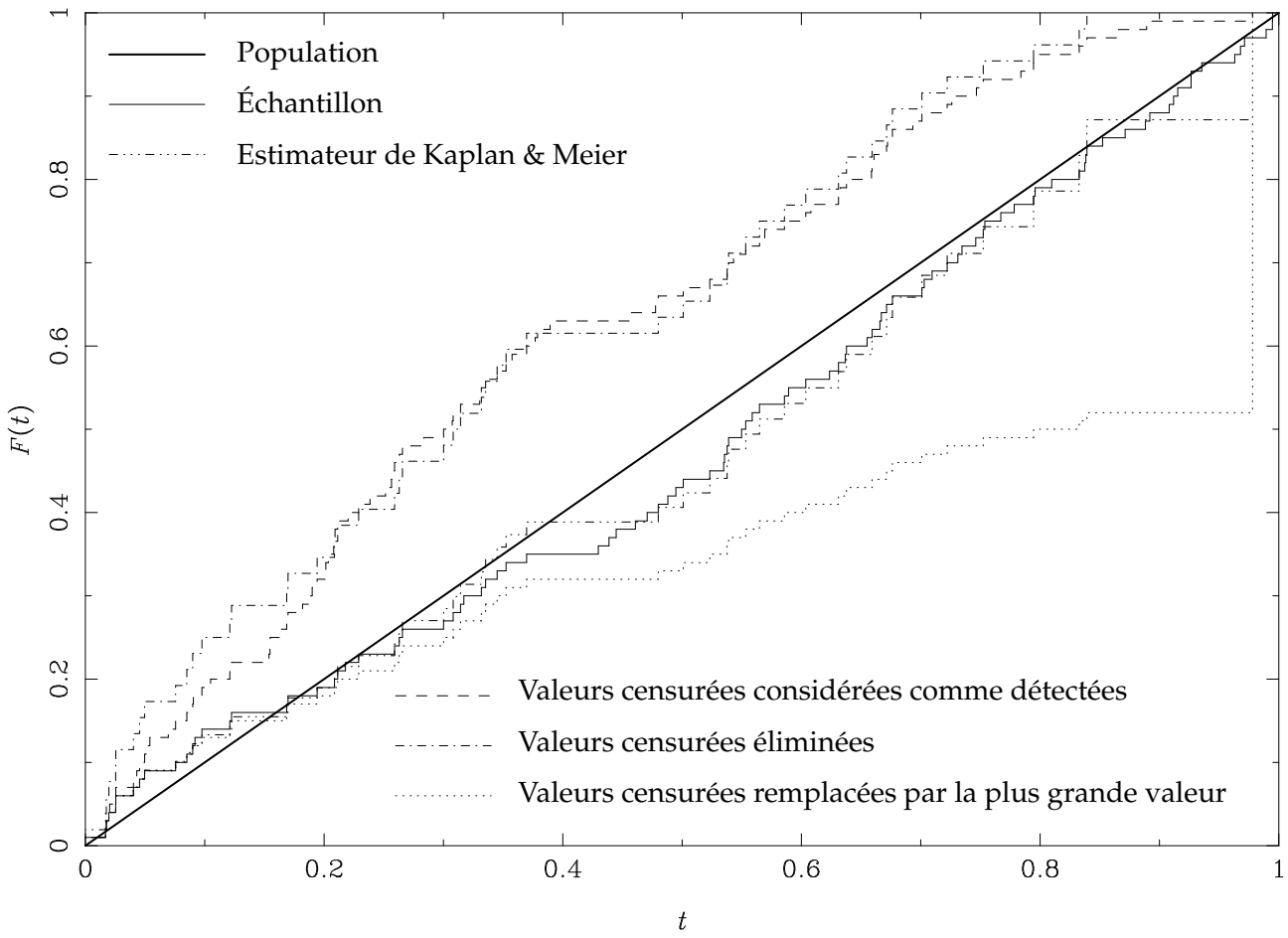



**Figure 1.** Fonction de répartition $F(t)$ pour $n = 100$ couples $(t_i, s_i)$ tirés aléatoirement et uniformément dans $[0, 1]$.

L'estimateur de Kaplan & Meier est une fonction en escalier, continue à droite et discontinue à gauche aux valeurs détectées seulement. On a

$$\hat{f}_n(t) = \sum_{i=1}^{n'} w_i \, \delta(t - \tau'_i),$$

avec $w_i = \hat{S}_n(\tau'^{-}_i) - \hat{S}_n(\tau'_i) = \hat{S}_n(\tau'_{i-1}) - \hat{S}_n(\tau'_i)$. Le poids $w_i$ est nul pour les valeurs censurées, et différent de $1/n$ et variable pour les valeurs détectées.

### IV.2.*c*. Application

On veut estimer la durée de vie typique d'une ampoule neuve. On dispose d'un échantillon de $n = 11$ ampoules qu'on allume à l'instant $t = 0$ et qu'on étudie pendant 10 mois. Par mesure d'économie, on éteint certaines ampoules en cours d'étude à des instants définis a priori ; d'autres claquent avant qu'on ne les éteigne.

| N° de l'ampoule | Date d'extinction (en mois) | Date de claquage (en mois) |
|---|---|---|
| 1 | — | 2 |
| 2 | — | 1 |
| 3 | — | — |
| 4 | 1 | — |
| 5 | 8 | — |
| 6 | — | 6 |
| 7 | — | 4 |
| 8 | — | 3 |
| 9 | 2 | — |
| 10 | — | 3 |
| 11 | 6 | — |

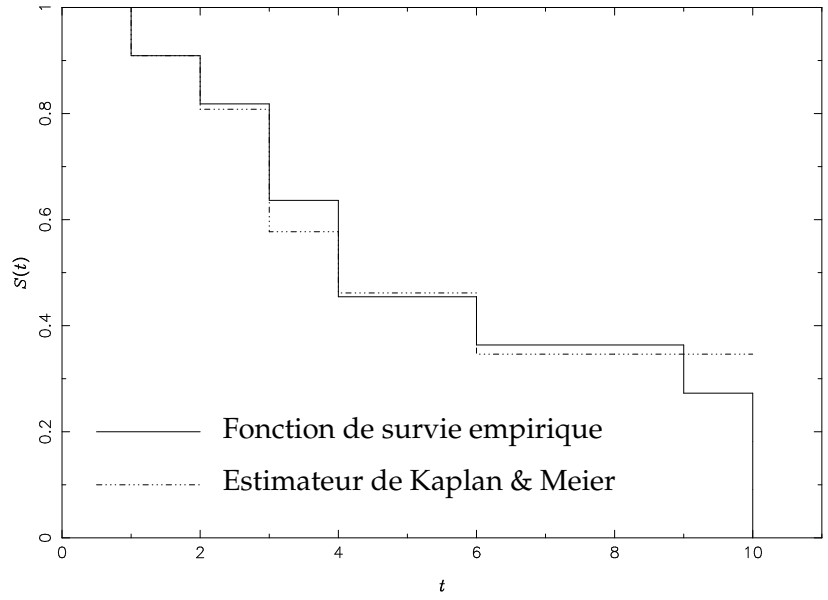
Construisons l'estimateur de Kaplan & Meier de la fonction de survie d'une ampoule. Les $\tau_j$ prennent les valeurs suivantes :

| Ampoule n° $j$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\tau_j$ | $= 2$ | $= 1$ | $> 10$ | $> 1$ | $> 8$ | $= 6$ | $= 4$ | $= 3$ | $> 2$ | $= 3$ | $> 6$ |

Il faut d'abord ordonner et regrouper les $\tau_j$. Il y a $n' = 7$ valeurs distinctes $\tau'_i$.

| $i$ | $\tau'_i$ | $n_i$ | $d_i$ | $c_i$ | $(n_i - d_i)/n_i$ | $\hat{S}_n(\tau'_i)$ |
|---|---|---|---|---|---|---|
| 1 | 1 | 11 | 1 | 1 | 10/11 | $10/11 = 0{,}909$ |
| 2 | 2 | 9 | 1 | 1 | 8/9 | $8/9 \times 0{,}909 = 0{,}808$ |
| 3 | 3 | 7 | 2 | 0 | 5/7 | $5/7 \times 0{,}808 = 0{,}577$ |
| 4 | 4 | 5 | 1 | 0 | 4/5 | $4/5 \times 0{,}577 = 0{,}462$ |
| 5 | 6 | 4 | 1 | 1 | 3/4 | $3/4 \times 0{,}462 = 0{,}346$ |
| 6 | 8 | 2 | 0 | 1 | 1 | $1 \times 0{,}346 = 0{,}346$ |
| 7 | 10 | 1 | 0 | 1 | 1 | $1 \times 0{,}346 = 0{,}346$ |

$n_i$ : nombre de $\tau_j \geqslant \tau'_i$.
$d_i$ : nombre de $\tau_j$ détectés et égaux à $\tau'_i$.
$c_i$ : nombre de $\tau_j$ censurés et égaux à $\tau'_i$.
On a $n_1 = n$ et $n_{i+1} = n_i - d_i - c_i$.
$\hat{S}_n(\tau'_0) = 1$ et $\hat{S}_n(\tau'_i) = \dfrac{n_i - d_i}{n_i}\, \hat{S}_n(\tau'_{i-1})$.




$$\hat{S}_n(t) = \begin{cases} 1 & \text{si } t < \tau'_1, \\ \hat{S}_n(\tau'_i) & \text{si } t \in [\tau'_i, \tau'_{i+1}[ \\ & \text{et } i \in [\![1, n'-1]\!], \\ \leqslant \hat{S}_n(\tau'_n) & \text{si } t \geqslant \tau'_{n'}. \end{cases}$$

**Figure 2.** Fonction de survie d'une ampoule.

### IV.2.*d*. Redistribution des données censurées à droite

L'estimateur de Kaplan & Meier peut être construit selon la procédure itérative suivante, plus intuitive mais moins efficace numériquement :

1. Ordonner les $\tau_i$ de manière croissante. Si des valeurs censurées sont égales aux valeurs détectées, placer celles-ci avant celles-là ;
2. Considérer la valeur n$^o$ $n$ comme détectée ;
3. Attribuer initialement le poids $w_i = 1/n$ à toutes les valeurs, détectées ou non ;
4. En allant de $i = n - 1$ jusqu'à $i = 1$, si la valeur n$^o$ $i$ est censurée, attribuer son poids $w_i$ à tous les $j > i$ qui sont détectés en proportion de leur poids, c'est-à-dire calculer

$$W_i = \sum_{j=i+1}^{n} \delta_j \, w_j$$

et faire les affectations suivantes :
   - $\forall\ j \in [\![i+1, n]\!], \quad w_j \leftarrow w_j + w_i \, \delta_j \, \dfrac{w_j}{W_i}$ ;
   - $w_i \leftarrow 0$ ;
5. Calculer $\hat{F}_n(t) = \sum_{i=1}^{n} w_i \, \mathbb{1}(t \leqslant \tau_i)$.

### IV.2.*e*. Propriétés de l'estimateur de Kaplan & Meier

#### IV.2.*e*.i. Biais et convergence

L'estimateur de Kaplan & Meier est légèrement biaisé : en général,

$$\mathbb{E}(\hat{F}_n[t]) < F(t),$$

où $\mathbb{E}$ désigne l'espérance.

Il est en revanche convergent (*consistent estimator*) :

$$\forall\ \epsilon > 0, \quad \lim_{n\to\infty} \mathbb{P}\Big(\big|\hat{F}_n[t] - F[t]\big| \geqslant \epsilon\Big) = 0.$$

Il est donc asymptotiquement non biaisé :

$$\lim_{n\to\infty} \mathbb{E}(\hat{F}_n[t]) = F(t).$$

#### IV.2.*e*.ii. Auto-cohérence

En l'absence de censure, un estimateur de $S(t)$ est

$$\tilde{S}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(t_i > t).$$

En présence de censure, on peut encore écrire

$$\tilde{S}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \left(\delta_i \, \mathbb{1}[t_i > t] + [1 - \delta_i] \, \mathbb{1}[t_i > t]\right),$$

mais la valeur de $\mathbb{1}(t_i > t)$ n'est pas connue pour les données censurées. Si un estimateur $\check{S}_n(t)$ de $S(t)$ est connu, on peut estimer l'espérance de $\mathbb{1}(t_i > t)$ sachant que $\delta_i = 0$ et $t_i > s_i$ : on a

$$\mathbb{E}(\mathbb{1}[t_i > t] \mid \delta_i = 0 \text{ et } t_i > s_i) = \frac{\mathbb{P}(T_i > t)}{\mathbb{P}(T_i > s_i)},$$

donc

$$\check{\mathbb{E}}(\mathbb{1}[t_i > t] \mid \delta_i = 0 \text{ et } t_i > s_i) = \frac{\check{S}_n(t)}{\check{S}_n(s_i)}.$$

L'estimateur de Kaplan & Meier présente la propriété d'être *auto-cohérent* (*self-consistent* ; Efron, 1967), c.-à-d. que

$$\hat{S}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \left( \delta_i \, \mathbb{1}[t_i > t] + [1 - \delta_i] \, \frac{\hat{S}_n(t)}{\hat{S}_n(s_i)} \right).$$

Si l'on part d'une fonction de survie arbitraire (mais compatible avec les contraintes sur une fonction de survie) $\check{S}_n^{(0)}$, on peut calculer itérativement une estimation $\check{S}_n^{(k)}$ par

$$\check{S}_n^{(k)}(t) = \frac{1}{n} \sum_{i=1}^{n} \left( \delta_i \, \mathbb{1}[t_i > t] + [1 - \delta_i] \, \frac{\check{S}_n^{(k-1)}(t)}{\check{S}_n^{(k-1)}(s_i)} \right).$$

et l'on a $\lim_{k\to\infty} \check{S}_n^{(k)} = \hat{S}_n$.

### IV.2.*f*. Limitations de l'estimateur de Kaplan & Meier

L'estimateur de Kaplan & Meier est discontinu. Pour certaines applications, il est nécessaire de le lisser en le convoluant avec un noyau.

Il ne prend pas en compte les incertitudes sur les valeurs $t_i$ et les censures $s_i$. Elles sont négligeables en statistiques médicales (la date de décès ou de sortie de l'échantillon d'un patient est connue précisément), mais souvent cruciales en astrophysique.

Ces incertitudes s'ajoutant à la dispersion intrinsèque de la loi de distribution (qu'on peut caractériser par l'écart-type autour de la moyenne), la dispersion apparente estimée à partir de l'estimateur de Kaplan & Meier surestime la dispersion intrinsèque.

Par ailleurs, le seuil de censure est souvent arbitraire en astrophysique : on peut ainsi considérer qu'une source n'est pas détectée si son flux est inférieur à 2, 3 ou 5 fois le bruit (cf. Kashyap *et al.*).

Des simulations, de type bootstrap par exemple, peuvent permettre de modéliser ces phénomènes et de corriger leurs effets.

### IV.2.*g*. Estimation de la médiane

La médiane $m$ est définie par $F(m)$ $(= 1 - S[m]) = 1/2$. Un estimateur de la médiane est donc obtenu en cherchant la valeur $\hat{m}_n$ telle que

$$\hat{S}_n(\hat{m}_n) = \frac{1}{2}.$$

$\hat{m}_n$ tombe forcément sur une des valeurs détectées, sauf dans le cas où l'une des marches de $\hat{S}_n$ vaut précisément $1/2$ : $\hat{m}_n$ n'est alors pas précisément défini (le même cas se produit en l'absence de censure à chaque fois que le nombre de données est pair : on convient alors que $\tilde{m}_n = (t_{n/2} + t_{n/2+1})/2$).

Dans le cas étudié au § IV.2.*c*, on trouve que $\hat{m} = 4$ mois (idem pour $\tilde{m}$ à partir des valeurs vraies de l'échantillon).

### IV.2.*h*. Estimation de la moyenne

Rappelons d'abord que nous avons décalé toutes les valeurs de telle sorte que $\mathbb{P}(T < 0) = 0$.

La moyenne $\mu$ est définie par

$$\mu = \mathbb{E}(T) = \int_{t=0}^{\infty} t \, f(t) \, \mathrm{d}t.$$

Or $f(t) = -\,\mathrm{d}S/\mathrm{d}t$. En intégrant par parties, on obtient

$$\mu = \Big[ -t \, S(t) \Big]_{t=0}^{\infty} + \int_{t=0}^{\infty} S(t) \, \mathrm{d}t = \int_{t=0}^{\infty} S(t) \, \mathrm{d}t.$$

$\hat{S}_n(t)$ est constante et vaut $\hat{S}_n(\tau'_i)$ sur $[\tau'_i, \tau'_{i+1}[$. Un estimateur de $\mu$ est donc

$$\hat{\mu}_n = \sum_{i=0}^{n'} \hat{S}_n(\tau'_i) \, (\tau'_{i+1} - \tau'_i),$$

en posant $\tau'_0 = 0$, $\tau'_{n'+1} = \infty$ et $0 \times \infty = 0$. $\hat{\mu}_n$ n'est pas définie si $\hat{S}_n(\tau'_{n'}) \neq 0$, c.-à-d. si le dernier point n'est pas détecté. On convient généralement de poser $\hat{S}_n(\tau'_{n'}) = 0$, mais on sous-estime alors la moyenne.

Avec cette convention, on obtient dans le cas étudié au § IV.2.*c* que $\hat{\mu} \approx 5{,}6$ mois ($\approx \tilde{\mu} = 5{,}64$ mois à partir des valeurs vraies, mais c'est un coup de chance).

### IV.2.*i*. Estimation de la variance et du biais par la méthode du bootstrap

Il existe des formules compliquées mais peu satisfaisantes pour estimer la variance et le biais de la loi. En pratique, il est plus commode de faire des simulations. On peut notamment utiliser la méthode du bootstrap, c.-à-d. qu'on rééchantillonne aléatoirement l'échantillon observé. Rappelons succinctement le principe de cette méthode :

- Pour chaque simulation $q$, on tire $n$ nombres aléatoires $t_i^{(q)}$ selon la loi $\hat{S}_n(t)$ [※2] ;
- On pose $\tau_i^{(q)} = \min\{t_i^{(q)}, s_i\}$ et $\delta_i^{(q)} = \mathbb{1}(t_i^{(q)} \leqslant s_i)$ ;
- On calcule l'estimateur de Kaplan $\hat{S}_n^{(q)}(t)$ à partir des $\tau_i^{(q)}$ et $\delta_i^{(q)}$ ;
- On compare la distribution des $\hat{S}_n^{(q)}(t)$ à $\hat{S}_n(t)$ et on en déduit une estimation de la variance et du biais de la loi.

La même méthode permer de calculer la variance et le biais de quantités dérivées telles que la moyenne, la médiane ou la dispersion intrinsèque.

**Remarques**

- $\hat{S}_n(t)$ étant constante sur tout intervalle $[\tau_i, \tau_{i+1}[$ et discontinue en tout point correspondant à une détection, on ne tire que des valeurs égales aux $\tau_{i,\,\delta_i=1}$. Si ceux-ci sont trop espacés, il peut être souhaitable de lisser $\hat{S}_n(t)$ ou d'« éparpiller » aléatoirement les $\tau_i^{(q)}$ autour des $\tau_i$ détectés.
- On peut tenir compte des incertitudes sur les $\tau_i$ en leur ajoutant une erreur tirée aléatoirement.
- La procédure décrite ci-dessus est applicable au cas d'une censure de type I. Dans le cas d'une censure aléatoire, il est souhaitable de tirer aussi les seuils $s_i$ de manière aléatoire (cf. Maller & Zhou, p. 239). On peut, là encore, le faire par la méthode du bootstrap.

## V. Estimation paramétrique. Méthode du maximum de vraisemblance

Nous supposons ici la *forme* de la loi connue. Notons $\vec{\theta} = (\theta_1, \ldots, \theta_k)$ les *paramètres* de la loi et $S(t\,;\vec{\theta})$ la probabilité $\mathbb{P}(T > t \mid \vec{\theta})$.

### V.1. En l'absence de censure

En l'absence de données censurées (et en négligeant les incertitudes observationnelles), la vraisemblance, c.-à-d. la probabilité que la situation observée se produise connaissant $\vec{\theta}$, donc d'obtenir simultanément $T_i$ dans $[t_i, t_i + \mathrm{d}t_i[$ pour tout $i \in [\![1, n]\!]$, est la quantité

$$\begin{aligned}\mathcal{L} &= \mathbb{P}(T_1 \in [t_1, t_1 + \mathrm{d}t_1[ \mid \vec{\theta}) \times \cdots \times \mathbb{P}(T_n \in [t_n, t_n + \mathrm{d}t_n[ \mid \vec{\theta}) \\ &= f(t_1\,;\vec{\theta})\,\mathrm{d}t_1 \times \cdots \times f(t_n\,;\vec{\theta})\,\mathrm{d}t_n,\end{aligned}$$

car les $T_i$ sont des variables aléatoires indépendantes.

La méthode du maximum de vraisemblance consiste à adopter comme estimateur $\tilde{\vec{\theta}}_n$ le vecteur $\vec{\theta}$ maximisant la vraisemblance ou, ce qui est plus commode numériquement, minimisant $L \coloneqq -\ln \mathcal{L}$.

Dans le cas où $f(t)$ est une loi normale de moyenne $\mu = \theta_1$ et d'écart-type (c.-à-d. de dispersion intrinsèque) $\sigma = \theta_2$,

$$\mathcal{L} \propto \left(\frac{1}{\sqrt{2\,\pi}\;\sigma}\right)^n \exp\!\left(-\sum_{i=1}^{n} \frac{(t_i - \mu)^2}{2\,\sigma^2}\right),$$

2. Il suffit de tirer un nombre aléatoire $a_i$ de manière uniforme dans $[0, 1]$ et de chercher $t_i^{(q)}$ tel que $\hat{S}_n(t_i^{(q)}) = a_i$.

soit

$$L = \sum_{i=1}^{n} \frac{(t_i - \mu)^2}{2\,\sigma^2} + n \ln \sigma + \mathrm{c^{te}}.$$

$L$ est minimale pour $\tilde{\mu}_n$ et $\tilde{\sigma}_n$ solutions du système $\{\partial L/\partial\mu = 0, \partial L/\partial\sigma = 0\}$, soit

$$\tilde{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} t_i \qquad \text{et} \qquad \tilde{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n} (t_i - \tilde{\mu}_n)^2.$$

On obtient les mêmes valeurs de $\tilde{\mu}_n$ et $\tilde{\sigma}_n$ que par la méthode des moments. Rappelons que $\tilde{\sigma}_n$ sous-estime $\sigma$ car il est calculé à partir de $\tilde{\mu}_n$ au lieu de $\mu$ (un estimateur non biaisé de $\sigma^2$ est $\sum_{i=1}^{n} (t_i - \tilde{\mu}_n)^2/(n-1)$).

## V.2. En présence de censure

Si $s_i$ est le seuil de censure pour la donnée n$^{\mathrm{o}}$ $i$,

$$\mathbb{P}(T_i > s_i) = S(s_i\,;\vec{\theta}) = 1 - F(s_i\,;\vec{\theta}) = \int_{t=s_i}^{\infty} f(t\,;\vec{\theta})\,\mathrm{d}t.$$

La vraisemblance est donc

$$\begin{aligned}\mathcal{L} &= \Big[\prod_{i=1,\,\delta_i=1}^{n} \mathbb{P}(T_i \in [\tau_i,\, \tau_i + \mathrm{d}\tau_i[\mid \vec{\theta})\Big] \prod_{i=1,\,\delta_i=0}^{n} \mathbb{P}(T_i > \tau_i \mid \vec{\theta}) \\ &\propto \prod_{i=1}^{n} f(\tau_i\,;\vec{\theta})^{\delta_i}\, S(\tau_i\,;\vec{\theta})^{1-\delta_i}.\end{aligned}$$

Dans le cas d'une loi normale de moyenne $\mu = \theta_1$ et d'écart-type $\sigma = \theta_2$,

$$S(t) = \frac{1}{2}\,\mathrm{erfc}\,\frac{t-\mu}{\sqrt{2}\,\sigma},$$

où erfc est la fonction d'erreur complémentaire.

On ne peut pas donner de formule littérale pour les valeurs $\check{\mu}_n$ et $\check{\sigma}_n$ minimisant $L$, mais on peut aisément les calculer numériquement à l'aide de bibliothèques telles que *Numerical recipes*.

## V.3. Maximum de vraisemblance et estimation non paramétrique

La méthode du maximum de vraisemblance n'est pas réservée à l'estimation paramétrique : on peut montrer que l'estimateur de Kaplan & Meier est l'estimateur non paramétrique maximisant la vraisemblance.

Ordonnons les $k$ valeurs détectées $x_i$ distinctes par ordre croissant et posons $x_0 = 0$ et $x_{k+1} = \infty$. Notons $n_i$ le nombre de valeurs $\tau_j$ supérieures ou égales à $x_i$, $d_i$ le nombre de détections égales à $x_i$, $\gamma_i$ le nombre de valeurs censurées dans $[x_i, x_{i+1}[$ et $y_i^{(j)}$ ($j \in [\![1, \gamma_i]\!]$) les valeurs censurées comprises dans cet intervalle.

Pour une détection,

$$\mathbb{P}(T_i = x_i) = (S[x_i^-] - S[x_i]).$$

(Maximiser la vraisemblance de manière non paramétrique va clairement rendre $S(t)$ discontinue, ce qui n'est pas possible de manière paramétrique quand la loi de probabilité adoptée est continue.)

La vraisemblance vaut donc

$$\mathcal{L} = \Big(\prod_{j=1}^{\gamma_0} S[y_0^{(j)}]\Big) \times \Big([S(x_1^-) - S(x_1)]^{d_1} \prod_{j=1}^{\gamma_1} S[y_1^{(j)}]\Big) \times \cdots \times \Big([S(x_k^-) - S(x_k)]^{d_k} \prod_{j=1}^{\gamma_k} S[y_k^{(j)}]\Big).$$

$S$ étant une fonction décroissante comprise entre 0 et 1, le maximum est obtenu en prenant $S(y_0^{(j)}) = 1$ et, pour tout $i > 0$, $S(x_i^-) = S(x_{i-1})$ et $S(y_i^{(j)}) = S(x_{i+1}^-)$.

Posons $P_i = S(x_i)$ $(= S(y_i^{(j)}) = S(x_{i+1}^-))$ et $p_i = P_i/P_{i-1}$. Il faut désormais maximiser

$$\mathcal{L} = \prod_{i=1}^{k} (P_{i-1} - P_i)^{d_i} P_i^{\gamma_i}.$$

On a $P_i = p_1 \cdots p_i$, donc

$$\mathcal{L} = \prod_{i=1}^{k} (p_1 \cdots p_{i-1})^{d_i} (1 - p_i)^{d_i} (p_1 \cdots p_i)^{\gamma_i} = \prod_{i=1}^{k} (p_1 \cdots p_i)^{d_i+\gamma_i} p_i^{-d_i} (1 - p_i)^{d_i}.$$

On a

$$\prod_{i=1}^{k} (p_1 \cdots p_i)^{d_i+\gamma_i} = \prod_{i=1}^{k} p_i^{\sum_{j=i}^{k} (d_j+\gamma_j)}.$$

$d_j + \gamma_j = n_j - n_{j+1}$, donc $\sum_{j=i}^{k} (d_j + \gamma_j) = n_i - n_{k+1} = n_i$ et

$$\mathcal{L} = \prod_{i=1}^{k} p_i^{n_i-d_i} (1 - p_i)^{d_i},$$

soit

$$L = -\sum_{i=1}^{k} ([n_i - d_i] \ln p_i + d_i \ln[1 - p_i]).$$

Le minimum de $L$ est obtenu en cherchant les $\hat{p}_i$ solutions du système $\{(\partial L/\partial p_i = 0)_{i \in [\![1,\, k]\!]}\}$, soit $\hat{p}_i = (n_i - d_i)/n_i$, c'est-à-dire le résultat obtenu pour l'estimateur de Kaplan & Meier. Celui-ci est donc un estimateur de maximum de vraisemblance.

# VI. Comparaison d'échantillons, corrélation et régression linéaire avec des données censurées

Cf. références.

# VII. Limites supérieures en astrophysique

Voir Kashyap *et al.* (2010) pour une clarification de la distinction entre borne supérieure de l'intervalle de confiance et limite supérieure sur le flux d'un objet.

# VIII. Références

### Site du cours

`http://www2.iap.fr/users/fioc/enseignement/analyse_de_survie/` .

### Livres et articles


- Droesbeke, J.-J., Fichet, B. & Tassi, Ph. (rédacteurs ; 1989) : *Analyse statistique des durées de vie. Modélisation des données censurées* ; éd. Economica.
- Efron, B. (1967) : *The two sample problem with censored data* ; Proceedings of the 5$^{\text{th}}$ Berkeley symposium on mathematical statistics and probabilities 4, 831.
- Feigelson, E. D. & Nelson, P. I. (1985) : *Statistical methods for astronomical data with upper limits. I. Univariate distributions* ; ApJ 293, 192.

- Feigelson, E. D. & Jogesh Babu, R. (2012) : *Modern statistical methods for astronomy – With R application* ; Cambridge university press.
- Isobe, T., Feigelson, E. D. & Nelson, P. I. (1986) : *Statistical methods for astronomical data with upper limits. II. Correlation and regression* ; ApJ 306, 490.
- Kaplan, E. L. & Meier, P. (1958) : *Nonparametric estimation from incomplete observations* ; Journal of the American statistical association 53, 457.
- Kashyap, V. L. *et al.* (2010) : *On computing upper limits to source intensities* ; ApJ 719, 900.
- Maller, R. A. & Zhou, X. (1996) : *Survival analysis with long-term survivors* ; éd. Wiley & sons.
- Miller, R. G. (1983) : *What price Kaplan-Meier?* ; Biometrics 39, 1077.
- Schmitt, J. H. (1985) : *Statistical analysis of astronomical data containing upper bounds : general methods and examples drawn from X-ray astronomy* ; ApJ 293, 191.


### Codes de statistiques de l'université Penn State

`http://www.astro.psu.edu/statcodes/`

(voir en particulier ASURV).

### Foire aux questions du Statistical consulting center for astronomy de l'université Penn State

`http://www.stat.psu.edu/~mga/scca/` .