S7_1_Suvie

Table of contents

1 Plan du cours

1. Introduction : données de survie et censure
2. Estimation de la fonction de survie : estimateur de Kaplan-Meier
3. Comparaison de fonctions de survie : test du log-rank

4. Modélisation de la survie : modèle de Cox

2 Introduction

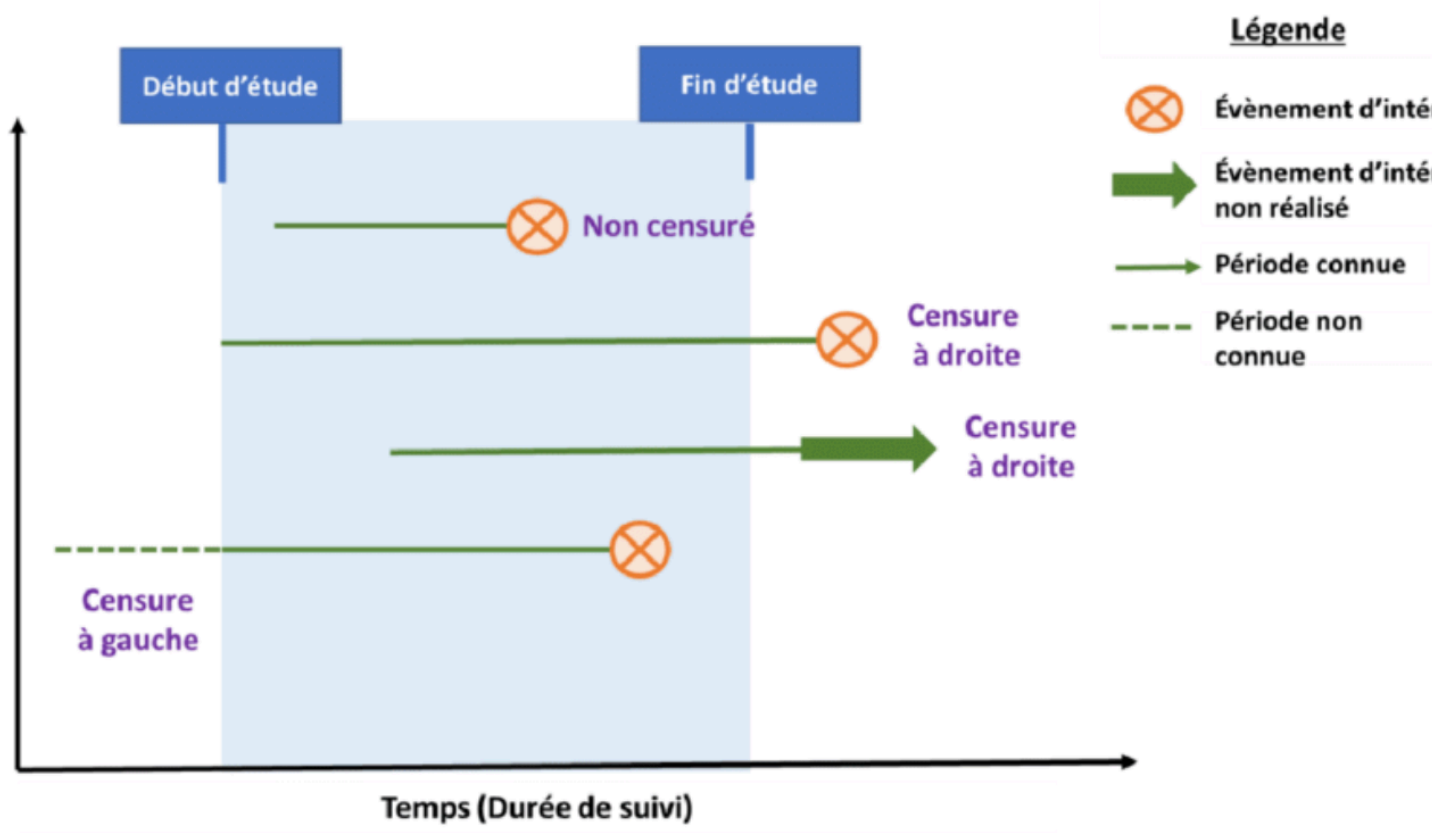
Donnée censuré ≠ donnée manquante :

- **Censurée** : on sait que l'événement d'intérêt n'est pas survenu avant un certain temps (par ex : pas de récidive jusqu'à la perte de vue)
- **Manquante** : on ne sait pas si l'événement d'intérêt est survenu ou non

2.A Données de survie

Donnée de survie = *survival time* ou *time to event data*

- Définition : délai de survenue d'un événement d'intérêt (*endpoint event*) à partir d'un temps de départ (souvent le temps 0)
- Correspond à tout délai entre deux dates d'intérêt =
  - Censure à gauche : point de départ !
  - Censure à droite : événement non survenu avant la fin de l'étude



2.B Problème de la censure

- certains ne présentent pas l'événement d'intérêt pendant la période d'étude
- tous les patients n'ont pas le même temps d'observation

On aimerait observer T_i = délai jusqu'à l'événement d'intérêt pour chaque individu i.

Mais on observe en fait :

- $min(T_i, C)$ et $1_{T_i>C}$ c'est à dire que :
  - $min(T_i, C)$: on cherche la plus petite valeur entre le délai jusqu'à l'événement d'intérêt T_i et une durée d'observation maximale fixe C
  - $1_{T_i>C}$: 1 à chaque fois que T_i est supérieur à C (censure à droite) = c'est à dire que l'événement n'est pas survenu avant la fin de l'étude
  - 0 sinon : si l'évènemet est survenu avant la censure
- ou $min(T_i, C_i)$, $1_{T_i>C_i}$

C : durée d'observation maximale fixe

C_i : durée d'observation variable selon les individus = **censure aléatoire**

Information partielle = censure à droite.

Par exemple :

- C = 3 ans (censure à 3 ans pour tous les individus)
- T_i = délai jusqu'à la récidive pour le patient i
- Si le patient i récidive à 2 ans : on observe $min(2, 3) = 2$ et $1_{2>3} = 0$ (événement observé avant la censure)
- Si le patient i ne récidive pas avant 3 ans : on observe $min(T_i, 3) = 3$ et $1_{T_i>3} = 1$ (événement non observé avant la censure)

Tip

En gros : il faut différencier

- les individus pour lesquels on observe l'événement d'intérêt avant la date de censure (on connaît leur temps de survie exact, parce qu'ils "n'ont pas survécu" jusqu'à la censure)
- les individus pour lesquels on ne sait pas si l'événement d'intérêt est survenu ou non avant la date de censure (on sait juste qu'ils ont "survécu" jusqu'à la censure)

Donc colonnes dans la BDD pour le critère de survie :

- L'ÉVÈNEMENT :
  - Survenue ou non (0/1)
  - Date et délai par rapport au temps de départ

- DURÉE DE SUIVI
  - Indépendante de la survenue ou non de l'événement d'intérêt
  - Depuis de le temps de départ

---

3 Données de survie

3.A Modélisation satistique

Analyse statistique dépend de la question de recherche :

- Est-ce que l'évènement s'est produit (pendant la période d'étude) ? = **modèle binaire** = régression logistique
- Quand l'évènement s'est-il produit ? = **modèle de survie** = régression de Cox

3.B Applications

- **Essai thérapeutique** : comparer l'efficacité de deux interventions revient à comparer les durées de survie après intervention dans les deux groupes
- **Étude épidémiologique** : estimation de l'association entre un facteur de risque et la durée de survie ou le temps de survenue d'une maladie

3.C Terminologie

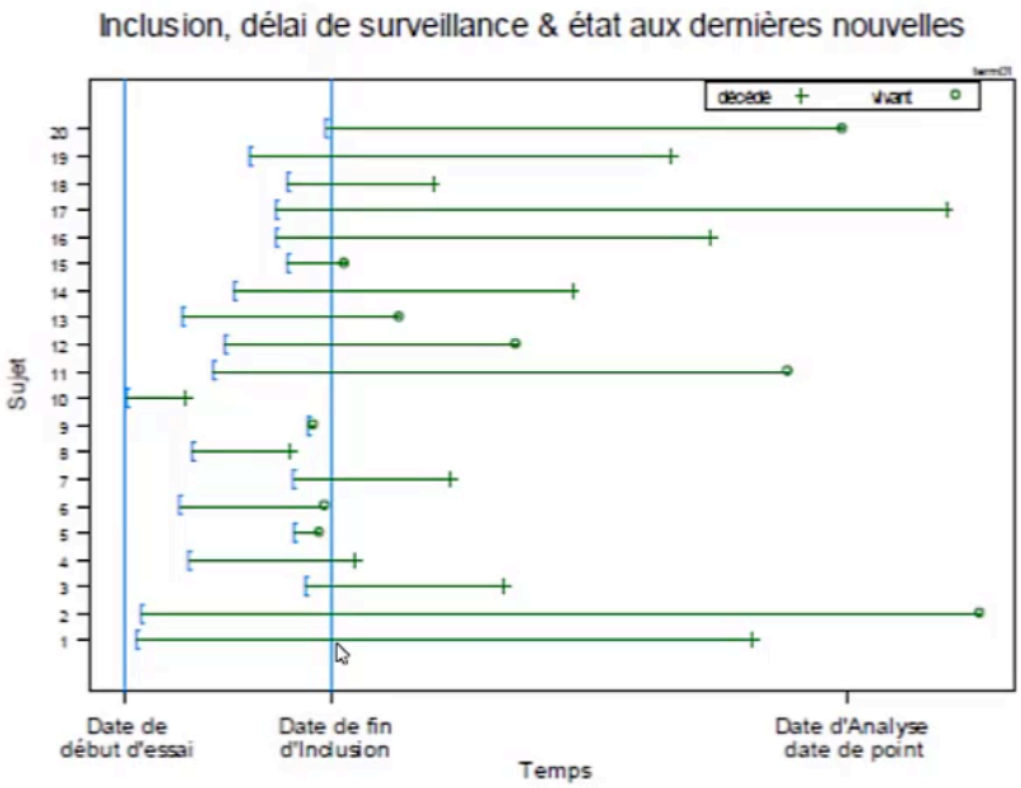
- **Date d'origine** : point de départ, varie selon patient, pour le calcul des durées de survie (ex : date de diagnostic, date de traitement, date d'inclusion dans l'étude)
- **Date des dernières nouvelles** : date de la dernière information connue sur le patient (ex : date de décès, date de la dernière consultation, date de la fin de l'étude)
- **Date de point** : commune à tous les patients, pour le calcul des durées de survie (ex : date de fin de l'étude)
- **Censure** : information incomplète, l'événement d'intérêt n'est pas survenu avant la date de point
- **Temps de participation** : variable d'étude
  - Évènement avant date de point (décès, récidive, etc.) : temps de participation = délai entre date d'origine et date de l'événement
  - Pas d'évènement avant date de point (censure) :
    - La date des dernières nouvelles est antérieure à la date de point : perdus de vue
    - La date des dernières nouvelles est égale à la date de point : censure administrative

3.D Ce qui est nécessaire

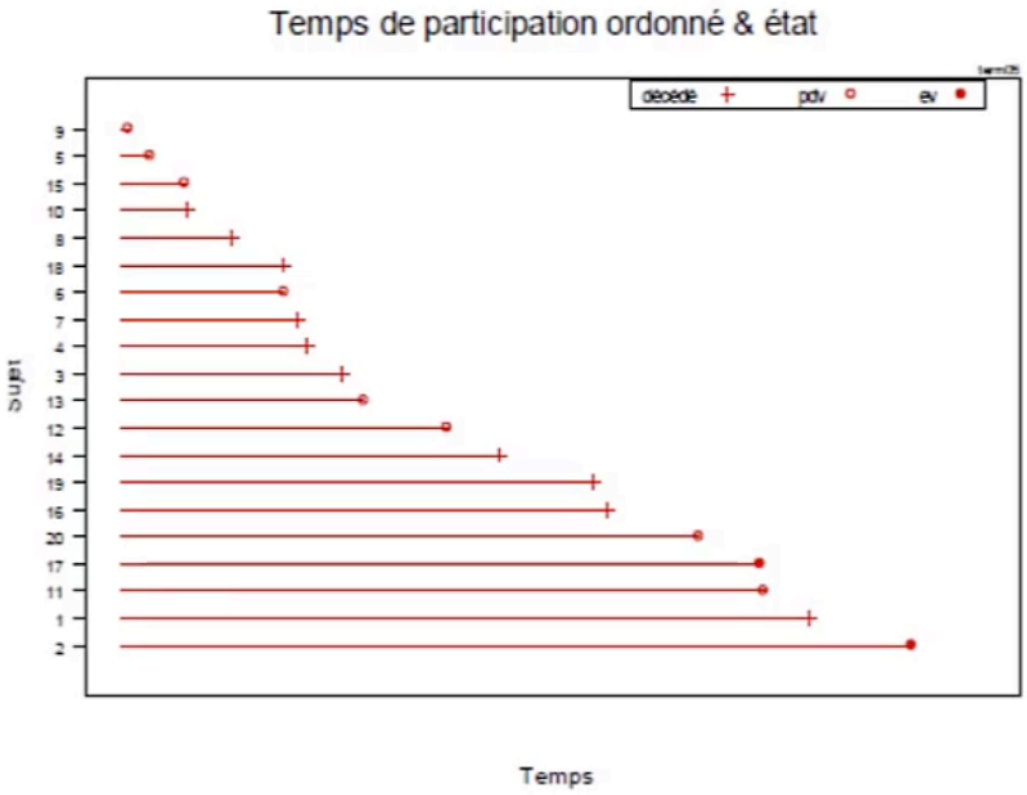
Pour chaque sujet, on doit disposer de :

- **Temps de participation** (délai entre date d'origine et date de l'événement ou date des dernières nouvelles)
- **État de l'événement** (1 = événement survenu, 0 = censuré) à la fin du temps de participation

Exemple : données simulées

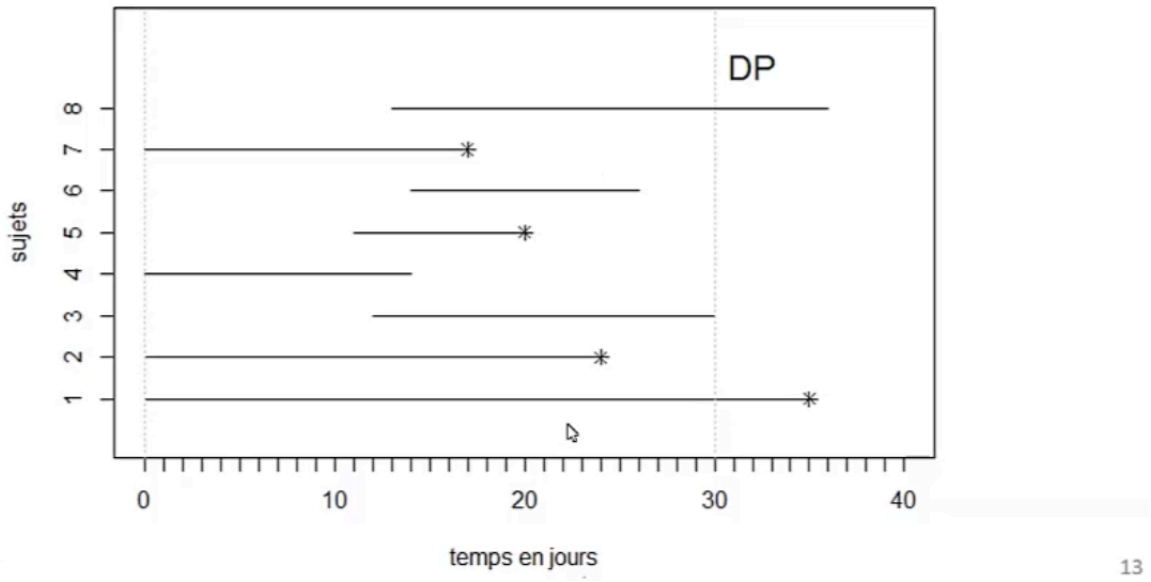


Exemple : données simulées



3.D.1 Exercice

A partir du graphique déterminer le temps de participation T_k pour chaque patient et l'état I_k en T_k.



16/11/20 13

Suivi :

- Patient 1 : 30 jours (évènement à 35 jours)
- Patient 2 : 24 jours (évènement à 24 jours)

Numéro du malade	Date d'origine (DO)	Date : Etat aux dernières nouvelles	Etat à la date de point (DP)	T_k jours	I_k
1	J0	J35 DCD	Vivant	30	0
2	J0	J24 DCD	DCD	24	1
3	J12	J30 Vivant	Vivant	18	0
4	J0	J14 Vivant	Perdu de vue	14	0
5	J11	J20 DCD	DCD	9	1
6	J14	J26 Vivant	Perdu de vue	12	0
7	J0	J17 DCD	DCD	17	1
8	J13	J36 Vivant	Vivant	12	0

3.E Loi de probabilité de T

Loi de probabilité de T = délai jusqu'à l'évenement (non observée)

Décrite par l'une de ces fonctions :

- Densité de probabilité, $f(t)$
- Fonction de répartition, $F(t)$
- Fonction de survie, $S(t)$
- Fonction de risque instantané, $h(t)$
- Fonction de risque cumulée, $H(t)$

3.E.1 Densité de probabilité

$f(t) = \lim_{\Delta t \to 0} \frac{P(T<t+\Delta t)-P(T<t)}{\Delta t}$

- $lim_{\Delta t \to 0}$: signifie que l'on regarde un intervalle de temps de plus en plus petit autour de t (on réduit l'intervalle Δt jusqu'à ce qu'il tende vers 0).
- T : variable aléatoire représentant le délai jusqu'à l'événement d'intérêt = moment où l'événement se produit.
- t : moment spécifique dans le temps où l'on évalue la densité de probabilité.
- $P(T < t)$: probabilité que l'événement d'intérêt T se produise avant le temps t.

On calcule la densité de probabilité $f(t)$ en prenant la limite lorsque l'intervalle de temps Δt autour

de t devient très petit. Cela nous permet d'estimer la probabilité que l'événement T se produise précisément à ce moment t.

En gros :

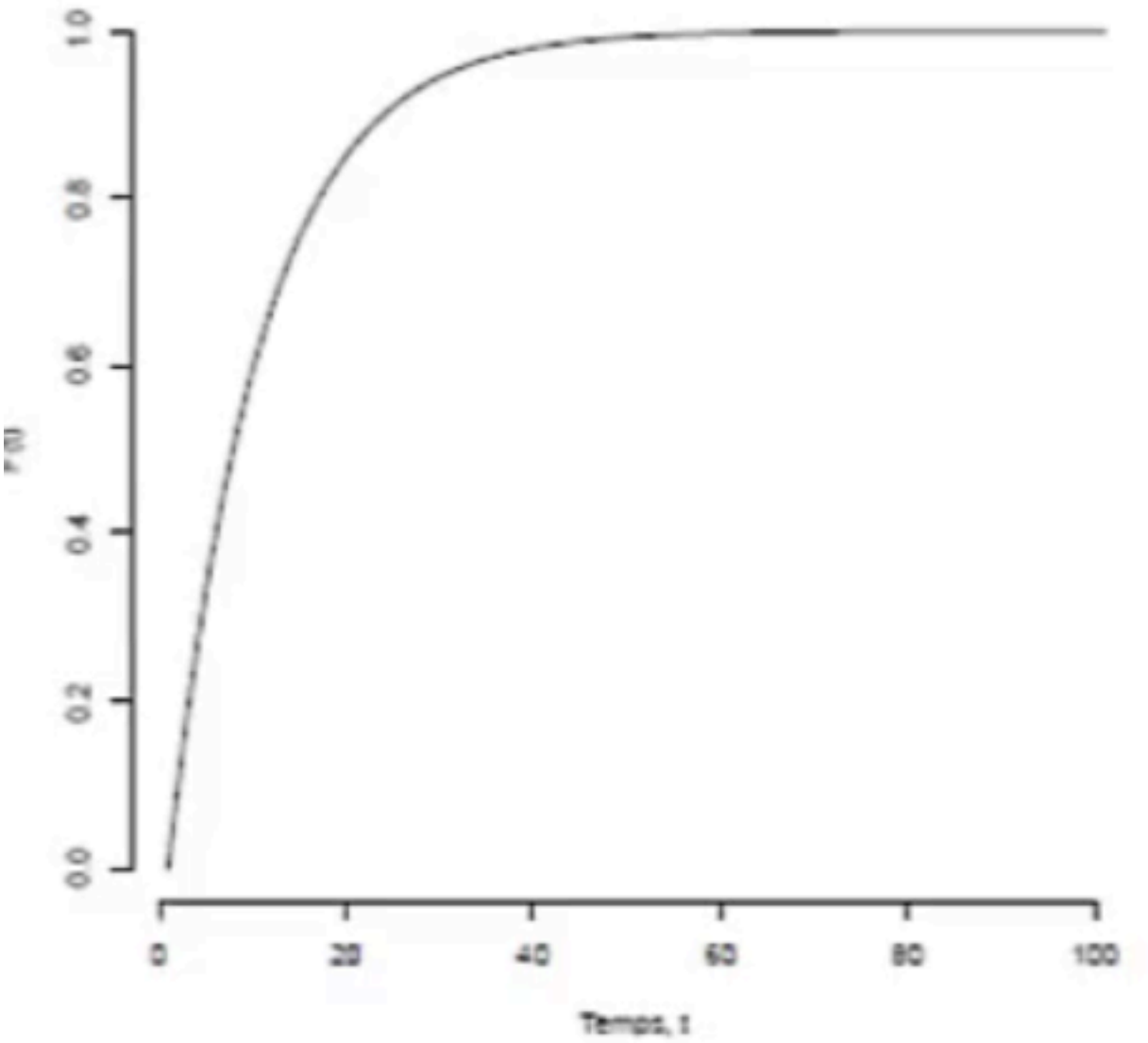
1. Différence entre :
   - La probabilité que l'évènement T se produise avant un certain temps $t + \Delta t$ (un intervalle de temps très petit autour de t)
   - et la probabilité que l'évènement T se produise avant un certain temps t
2. Diviser le résultat par la taille de l'intervalle de temps Δt
3. On obtient ainsi un "taux moyen" de probabilité par unité de temps dans cet intervalle très petit autour de t
4. En prenant la limite lorsque Δt tend vers 0, on obtient la **densité de probabilité instantanée**

3.E.2 Fonction de répartition

Probabilité que la variable aléatoire T prenne une valeur inférieure ou égale à une quantité t :

$F(t) = P(T \leq t) = \int_0^t f(u)du$

- $F(t)$: fonction de répartition, qui donne la probabilité que l'événement d'intérêt T se produise avant ou à un moment spécifique t.
- $P(T \leq t)$: probabilité que l'événement T se produise avant ou à temps t.
- $\int_0^t f(u)du$: intégrale de la densité de probabilité $f(u)$ de 0 à t, qui calcule la probabilité cumulative jusqu'à ce moment.



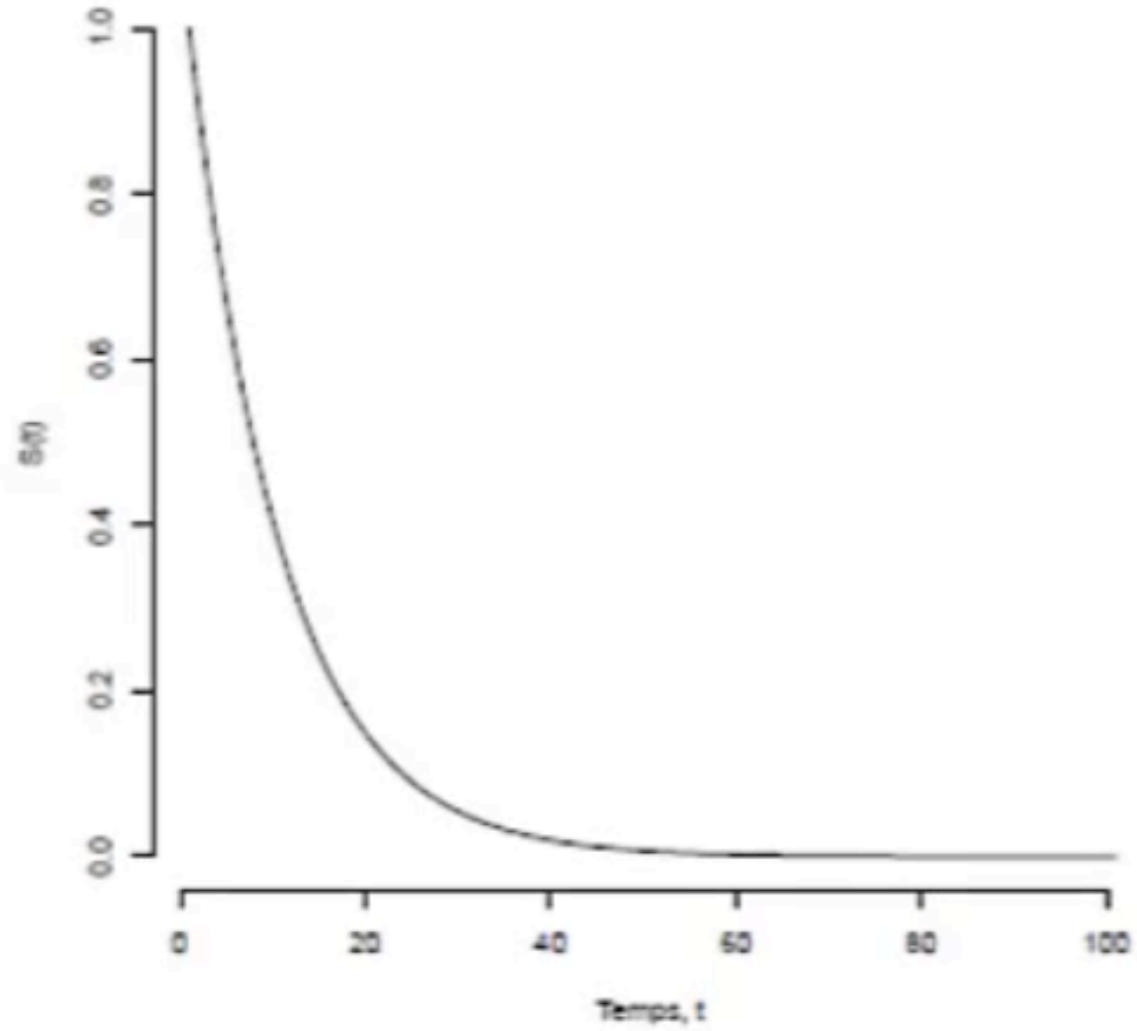
Propriétés :

- Fonction croissante (plus on attend, plus la probabilité que l'événement se soit produit augmente)
- La vitesse de croissance dépend de la densité de probabilité $f(t)$ = caractérise la loi de la variable aléatoire T
- $F(0) = 0$: si on n'attend pas, la probabilité de voir l'évènement = 0
- $\lim_{t\to\infty} F(t) = 1$: si on attend indéfiniment, la probabilité de voir l'évènement = 1
- F dérivable et $F' = f$ avec f la densité de probabilité (la dérivée de la fonction de répartition = la fonction de densité).

3.E.3 Fonction de survie

Représente la fraction d'individus encore en vie en t.

$S(t) = P(T > t) = 1 - F(t) = \int_t^{\infty} f(u)du$



Propriétés :

- Fonction décroissante (plus on attend, plus la probabilité de survie diminue)
- La vitesse de décroissance dépend de la densité de probabilité $f(t)$ = caractérise la loi de la variable aléatoire T
- $S(0) = 1$: si on n'attend pas, la probabilité de survie = 1
- $\lim_{t\to\infty} S(t) = 0$: si on attend indéfiniment, la probabilité de survie = 0
- S dérivable et $S' = -f$ avec f la densité de probabilité (la dérivée de la fonction de survie = l'opposé de la fonction de densité).

3.E.4 Fonction de risque instantané

Définition : fonction de “hazard” ou “hasard”. Représente le risque instantané de survenue de l'événement à l'instant t, conditionnellement au fait que l'individu ait survécu jusqu'à ce temps t.

Densité “**conditionnelle**” de l'événement à l'instant t sachant que l'individu est encore en vie à ce moment.

$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$

NB : la barre verticale “|” signifie “conditionnellement à”.

- $h(t)$: fonction de risque instantané, qui mesure le risque de survenue de l'événement d'intérêt à un moment spécifique t, conditionnellement au fait que l'individu ait survécu jusqu'à ce temps t.
- $P(t \leq T < t + \Delta t \mid T \geq t)$:
  - probabilité que l'événement T se produise dans l'intervalle de temps entre t et $t + \Delta t$,
  - **conditionnellement** au fait que l'individu ait survécu jusqu'à temps t.
- Δt : intervalle de temps très petit autour de t.
- $\lim_{\Delta t \to 0}$: signifie que l'on regarde un intervalle de temps de plus en plus petit autour de t (on réduit l'intervalle Δt jusqu'à ce qu'il tende vers 0).
  - Permet d'obtenir un taux instantané de risque à ce moment précis t.
  - Sinon, on obtiendrait une moyenne sur un intervalle de temps plus large.

En résumé : c'est une **densité conditionnelle** qui mesure le risque instantané de survenue de l'événement à un moment spécifique t, en tenant compte du fait que l'individu a déjà survécu jusqu'à ce temps t ET NE L'A PAS ENCORE PRÉSENTÉ !

On peut représenter plus simplement la fonction en :

$h(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)}$

C'est à dire que la fonction de risque instantané est le rapport entre la densité de probabilité et la fonction de survie.

C'est logique ! parce que :

- Fonction de densité $f(t)$: probabilité que l'événement se produise précisément à l'instant t
- Fonction de survie $S(t)$: probabilité que l'individu soit encore en vie à temps t (donc n'ait pas encore présenté l'événement)
- Donc le rapport $f(t)/S(t)$: probabilité que l'événement se produise à l'instant t sachant que l'individu est encore en vie à ce moment.
- Le signe négatif dans $-\frac{S'(t)}{S(t)}$ vient du fait que la dérivée de la fonction de survie $S'(t)$ est négative (puisque $S(t)$ est décroissante). Donc en prenant l'opposé, on obtient une valeur positive pour la fonction de risque instantané.

3.E.5 Fonction de risque cumulée

Représente le risque cumulé de survenue de l'événement jusqu'au temps t.

$H(t) = \int_0^t h(u)du$

Propriétés :

- Fonction croissante (plus on attend, plus le risque cumulé augmente)
- La vitesse de croissance dépend de la densité de probabilité $f(t)$ = caractérise la loi de la variable aléatoire T
- $H(0) = 0$: si on n'attend pas, le risque cumulé = 0
- $\lim_{t\to\infty} H(t) = \infty$: si on attend indéfiniment, le risque cumulé = infini

3.E.6 Relations entre les fonctions

Fonction	Notation	Définition
Densité de probabilité	$f(t)$	$\lim_{\Delta t\to 0} \frac{P(t<T<t+\Delta t)}{\Delta t}$
Fonction de répartition	$F(t)$	$P(T \leq t) = \int_0^t f(u)du$
Fonction de survie	$S(t)$	$P(T > t) = 1 - F(t) = \int_t^\infty f(u)du$
Fonction de risque instantané	$h(t)$	$\lim_{\Delta t\to 0} \frac{P(t\leq T<t+\Delta t \mid T\geq t)}{\Delta t}$
Fonction de risque cumulée	$H(t)$	$\int_0^t h(u)du$

$P(t < T < t + \Delta t) = P(t < T < t + \Delta t \mid T > t) \times P(T > t)$

$P(t < T < t + \Delta t)$ = la probabilité que l'événement se produise dans l'intervalle de temps entre t et $t + \Delta t$.

Mais en fait 2 probabilités en une :

1. **probabilité** que l'individu soit encore en vie au temps t : $P(T > t)$ = la probabilité que l'individu ait survécu jusqu'à temps t (donc n'ait pas encore présenté l'événement avant t).
   NB : c'est la fonction de survie $S(t)$
2. **probabilité** que l'événement se produise dans l'intervalle de temps entre t et $t + \Delta t$ PARMI LES INDIVIDUS ayant survécu jusqu'à temps t : $P(t \leq T < t + \Delta t \mid T \geq t)$.
   NB : c'est la fonction de risque instantané $h(t)$ multipliée par la taille de l'intervalle Δt

Et sachant que :

f = densité de probabilité = $\lim_{\Delta t\to 0} \frac{P(t<T<t+\Delta t)}{\Delta t}$

h = fonction de risque instantané = $\lim_{\Delta t\to 0} \frac{P(t\leq T<t+\Delta t|T\geq t)}{\Delta t}$

S = fonction de survie = $P(T > t)$

donc on en déduit :

$f(t) = h(t) \times S(t)$

Et sachant que :

f = densité de probabilité = $-S'(t)$

S = fonction de survie = $1 - F(t) = e^{-H(t)}$ car $H(t) = -\ln(S(t))$

En résumé :

Fonction	Notation	Relation avec les autres fonctions
Densité de probabilité	$f(t)$	$f(t) = h(t) \times S(t)$
Fonction de répartition	$F(t)$	$F(t) = 1 - S(t)$
Fonction de survie	$S(t)$	$S(t) = e^{-H(t)}$
Fonction de risque instantané	$h(t)$	$h(t) = \frac{f(t)}{S(t)}$
Fonction de risque cumulée	$H(t)$	$H(t) = -\ln(S(t))$

4 Estimation d'une courbe de survie : estimateur de Kaplan-Meier

4.A Objectif de l'analyse de survie

- Estimer le délai médian avant la survenue de l'événement d'intérêt
- Comparer ce délai entre plusieurs groupes de patients
- Étudier l'effet de variables explicatives sur ce délai

4.B Approches

- **Approche paramétrique** : modélisation de la fonction de risque = on fait une hypothèse sur la forme de la fonction de risque (ex : loi exponentielle)
- **Approche non paramétrique** : estimation de la fonction de survie sans faire d'hypothèse sur la forme de la fonction de risque = estimateur de Kaplan-Meier
- **Approche semi-paramétrique** : modèle multivarié : modélisation de l'effet des variables explicatives sur la fonction de risque sans faire d'hypothèse sur la forme de la fonction de risque = modèle de Cox

4.B.1 Estimation en absence de censure

En l'absence de censure, on remplace la probabilité théorique par une proportion basée sur les données observées.

- **Estimateur de la fonction de répartition** $F(t)$ à partir des données observées $\hat{F}(t) = \frac{1}{n} \sum_{i=1}^{n} 1_{(T_i \leq t)}$.
  - $1_{(T_i \leq t)}$ = indicatrice qui vaut 1 si $T_i \leq t$ (si l'évènement T a eu lieu avant t) et 0 sinon.

- $\sum_{i=1}^{n} 1_{(T_i \leq t)}$ = somme des 0 et des 1 pour tous les individus = nombre d'individus ayant présenté l'évènement avant t.
- diviser le tout par le nombre total d'individus n permet d'obtenir la proportion d'individus ayant présenté l'évènement avant t.
- Donc $\widehat{F}(t)$ = estimateur empirique = proportion d'individus ayant présenté l'évènement avant t dans la population étudiée
- $\widehat{F}(t)$ est un estimateur de la fonction de répartition $F(t)$, qui serait la proportion "vraie" d'individus ayant présenté l'évènement avant t dans la population générale / idéale
- Fonction en escalier, monotone, croissante de 0 à 1

• **Estimateur de la fonction de survie** $S(t) = 1 - \widehat{F}(t)$

$\widehat{S}(t) = 1 - \widehat{F}(t) = \frac{1}{n} \sum_{i=1}^{n} 1_{(T_i > t)}$

- $1_{(T_i > t)}$ = indicatrice qui vaut 1 si $T_i > t$ (si l'évènement T a eu lieu après t) et 0 sinon.
- $\sum_{i=1}^{n} 1_{(T_i > t)}$ = somme des 0 et des 1 pour tous les individus = nombre d'individus n'ayant pas présenté l'évènement avant t (ayant survécu au delà de t).
- diviser le tout par le nombre total d'individus n permet d'obtenir la proportion d'individus n'ayant pas présenté l'évènement avant t (ayant survécu au delà de t).
- Donc $\widehat{S}(t)$ = estimateur empirique = proportion d'individus n'ayant pas présenté l'évènement avant t (ayant survécu au delà de t) dans la population étudiée

= $\widehat{S}(t) = \frac{\text{Nombre d'individus survivant au dela de } t}{\text{Nombre total d'individus}}$

- Fonction en escalier, monotone, décroissante de 1 à 0.

4.B.1.1 Exemple

1. Données = temps de décès (en mois) de 10 patients :
   13, 13, 14, 13, 15, 11, 17, 13, 14, 15
2. Estimation de la fonction de survie

Estimateur de $S(t) = 1 - F(t)$

$\widehat{S}(t) = \frac{1}{n} \sum_{i=1}^{n} 1_{(T_i > t)} = \frac{\text{Nombre d'individus survivant au dela de } t}{\text{Nombre total d'individus}}$

en tableau :

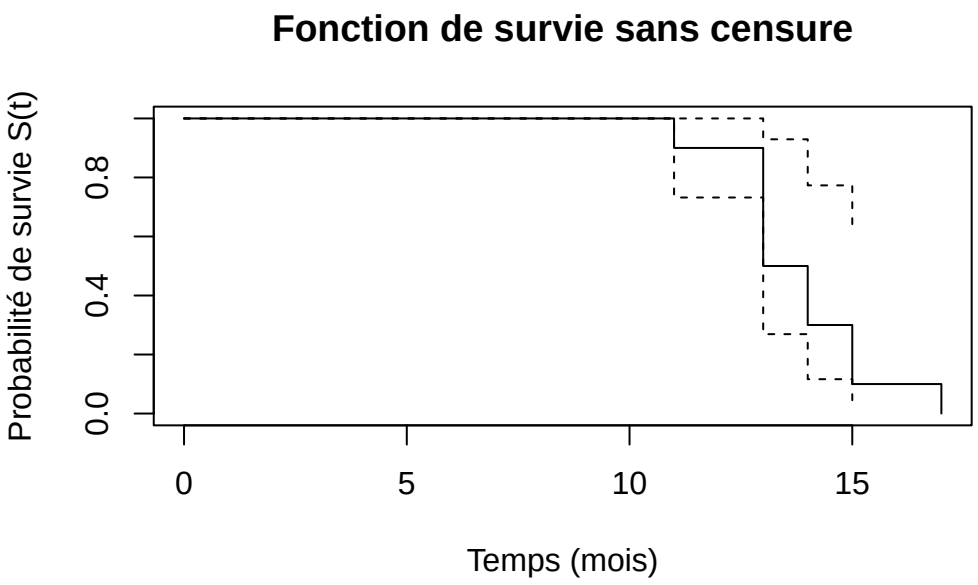
• Premier décès à 11 mois, compté dans l'intervalle 11 - 12 par convention.

t	Nb décès	Nb décès cumulé	$\widehat{S}(t)$
0	0	0	10/10 = 1.0
11	1	1	9/10 = 0.9
13	4	5	5/10 = 0.5
14	2	7	3/10 = 0.3
15	2	9	1/10 = 0.1

t	Nb décès	Nb décès cumulé	$\hat{S}(t)$
17	1	10	0/10 = 0.0

3. Tracé de la fonction de survie sans Kaplan Meier

```r
library(survival)

# 1) Données
temps <- c(13, 13, 14, 13, 15, 11, 17, 13, 14, 15)

# Comme il n'y a PAS de censure : event = 1 pour tout le monde
event <- rep(1, length(temps))

# Objet Surv
surv_object <- Surv(time = temps, event = event)

# 2) Estimation de la fonction de survie
surv_fit <- survfit(surv_object ~ 1)

# 3) Tracé de la fonction de survie
plot(
    surv_fit,
    xlab = "Temps (mois)",
    ylab = "Probabilité de survie S(t)",
    main = "Fonction de survie sans censure")
```



4.B.1.2 2e exemple avec temps simulés

On utilise `runif` qui génère des nombres aléatoires suivant une loi uniforme (uniforme = tous les intervalles de même longueur ont la même probabilité d'être choisis).

`floor` arrondit à l'entier inférieur.

```
T = floor(runif(80,0,50)) # runif génère 1000 temps de survie entre 0 et 50 qui sont arrondis à
head(T)
```

```
[1] 20  0 19  5 20 38
```

```
table(T)
```

```
T
 0  1  2  3  4  5  7  8  9 11 12 13 14 15 16 17 18 19 20 22 23 25 26 27 28 30
 3  1  1  3  2  2  2  1  2  3  1  1  1  1  1  3  2  2  4  2  3  3  2  1  1  1
31 32 33 34 36 37 38 39 40 41 42 43 45 46 47 49
 1  2  2  2  1  1  4  4  1  1  1  2  2  3  3  1
```

Le jeu de données est généré mais on a pas les valeurs uniques.

```
tt <- unique(T)
length(tt) # 39 valeurs uniques
```

```
[1] 42
```

Il faut aussi ordonner les valeurs uniques.

```
tt <- sort(tt)
tt
```

```
 [1]  0  1  2  3  4  5  7  8  9 11 12 13 14 15 16 17 18 19 20 22 23 25 26 27 28
[26] 30 31 32 33 34 36 37 38 39 40 41 42 43 45 46 47 49
```

Pour regarder à la date 10 :

- avec `head()` : R prend chacune des valeurs de T et regarde si elle est supérieure à 10 (TRUE) ou non (FALSE).
- avec `mean()` : R calcule la proportion de TRUE (donc la proportion de survivants au delà de 10), en convertissant TRUE en 1 et FALSE en 0.

```
head(T > 10)
```

```
[1]  TRUE FALSE  TRUE FALSE  TRUE  TRUE
```

```
mean(T > 10) # proportion de TRUE = proportion de survivants au delà de 10
```

```
[1] 0.7875
```

On peut faire ça pour toutes les valeurs uniques de T.

```r
S <- function(t) mean(T > t)
```

Syntaxe : `function(t) mean(T > t)`

- `function(t)` : on définit une fonction qui prend un argument `t`
- `mean(T > t)` : la fonction calcule la proportion de survivants au delà de `t`

Pas de séparation par des virgules car une seule instruction dans le corps de la fonction = R comprend que tout ce qui suit `function(t)` fait partie du corps de la fonction.

```r
S(10) # proportion de survivants au delà de 10
```

```
[1] 0.7875
```

```r
S(20) # proportion de survivants au delà de 20
```

```
[1] 0.55
```

On applique cette fonction à toutes les valeurs uniques de T avec `sapply`.

`sapply` applique une fonction (ici une fonction anonyme) à chaque élément d'un vecteur (ici `tt`).

≠ `lapply` qui renvoie une liste, `sapply` renvoie un vecteur ou une matrice.

```r
S_values <- sapply(tt, S)
S_values
```

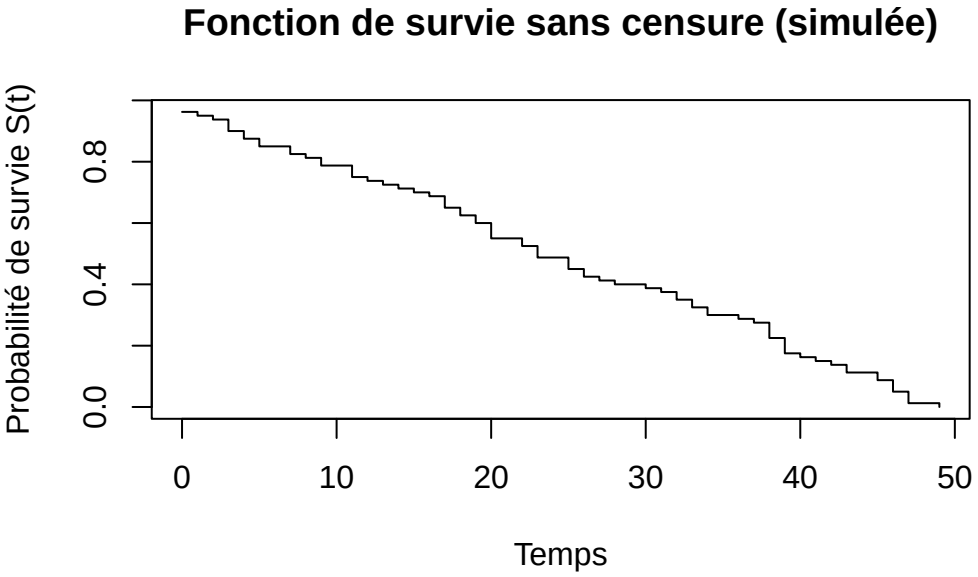
```
 [1] 0.9625 0.9500 0.9375 0.9000 0.8750 0.8500 0.8250 0.8125 0.7875 0.7500
[11] 0.7375 0.7250 0.7125 0.7000 0.6875 0.6500 0.6250 0.6000 0.5500 0.5250
[21] 0.4875 0.4500 0.4250 0.4125 0.4000 0.3875 0.3750 0.3500 0.3250 0.3000
[31] 0.2875 0.2750 0.2250 0.1750 0.1625 0.1500 0.1375 0.1125 0.0875 0.0500
[41] 0.0125 0.0000
```

On peut tracer la fonction de survie estimée.

Syntaxe de `plot` : `plot(x, y, type, xlab, ylab, main)`

- `x` : valeurs sur l'axe des x (ici `tt`)
- `y` : valeurs sur l'axe des y (ici `S_values`)
- `type="s"` : pour une fonction en escalier

```r
plot(
    tt, # x values = temps
    S_values, # y values = probabilité de survie S(t) calculée
    type="s", # type = "s" pour une fonction en escalier
    xlab="Temps",
    ylab="Probabilité de survie S(t)",
    main="Fonction de survie sans censure (simulée)")
```



4.B.2 Estimation en présence de censure : probabilités conditionnelles

Probabilités conditionnelles = on ne considère que les individus “à risque” à chaque instant.

- À chaque temps t_j où un événement est observé, on calcule la probabilité de survie conditionnelle **sachant que l'individu a survécu jusqu'à ce temps** t_j.
- On multiplie ces probabilités conditionnelles pour obtenir la probabilité de survie jusqu'à un temps t donné.

Par exemple :

Probabilité d'être en vie à 2 et 3 ans : probabilité décomposée en deux parties :

$S(2) = P(T > 2) = P(T > 2 \mid T > 1) \times P(T > 1)$

$S(3) = P(T > 3) = P(T > 3 \mid T > 2) \times P(T > 2)$

$S(3) = P(T > 3 \mid T > 2) \times S(2)$

= Probabilité de survivre entre 1 et 2 ans sachant qu'on a survécu jusqu'à 1 an multipliée par la probabilité de survivre jusqu'à 1 an.

On généralise à un ensemble de K temps ordonnés définis arbitraitement et aléatoirement

On “découpe” la période d'étude pour obtenir des petits intervalles $[t_{k-1}, t_k)$.

$S(t_k) = P(T > t_k \mid T > t_{k-1}) \times P(T > t_{k-1}) = P(T > t_k \mid T > t_{k-1}) \times S(t_{k-1})$

Parce que $S(t_{k-1}) = P(T > t_{k-1})$

Tableau de données Kaplan Meier :

N_k = nombre d'individus à risque au temps t_k (ayant survécu jusqu'à t_k)

Q_k = probabilité conditionnelle de survie au temps t_k = $P(T > t_k \mid T > t_{k-1})$

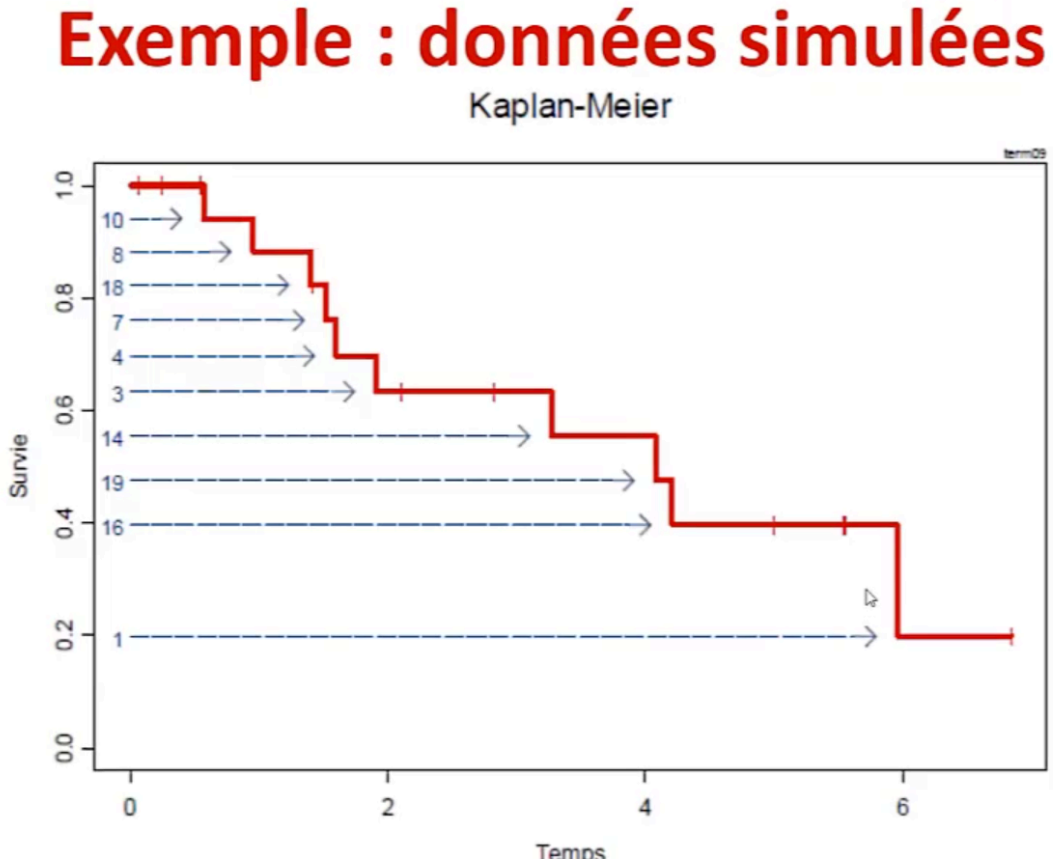
S_k = probabilité cumulée de survie jusqu'au temps t_k = $S(t_k)$

t_k (temps)	N_k (nombre à risque en pendant la période)	D_k (décès en début de période)	C_k (censurés)	Q_k (Probabilité conditionnelle de survie)	S_k (Probabilité cumulée de survie)
t_0 = 0	N_0 = n	0	0	1	1
t_1	$N_1 = N$	d_1	c_1	$Q_1 = \frac{1-d_1}{N_1}$	$S_1 = q_1$
t_2	$N_2 = N_1 - D_1 - C_1$	d_2	c_2	$Q_2 = \frac{N_2-d_2}{N_2}$	$S_2 = q_1 q_2$
...	...	...	...	...	...
t_k	$N_k = N_{k-1} - D_{k-1} - C_{k-1}$	d_k	c_k	$Q_k = \frac{N_k-d_k}{N_k}$	$S_k = q_k q_{k-1} .. q_2 q_1$

Calculs à faire :

- N_k : nombre de sujet à risques (à chaque teps diminuée du nombre de décès et de censure durant la période précédante)
- Q_k : quantités de survie conditionnelle

$(\frac{\text{Nombre de personne sur la periode} - \text{Nombre deces sur la periode}}{\text{Nombre de personne sur la periode}})$



4.B.2.1 Exemple 1

Données de survie avec censure :

1, 1, 1+, 1+, 1+, 2, 2, 2, 2+, 3, 3, 3+, 4+, 5+, avec “+” représente une censure.

Donc 14 individus au total.

Par défaut, un évènement qui a lieu au temps 1 a lieu dans l’intervalle 1 à 2, donc représenté sur la deuxième ligne.

Tableau :

Temps t_k	Nombre à risque N_k	Décès D_k	Censurés C_k	Probabilité condition-nelle de survie Q_k	Probabilité cumulée de survie S_k
0	14	0	0	1	1
1	14	2	3	1 - 2/14 = 6/7	6/7
2	9	3	1	1 - 3/9 = 6/9	6/7 * 6/9 = 4/7
3	5	2	1	1 - 2/5 = 3/5	4/7 * 3/5 = 12/35
4	2	0	1	1 - 0/2 = 1	12/35 * 1 = 12/35
5	1	0	1	1 - 0/1 = 1	12/35 * 1 = 12/35

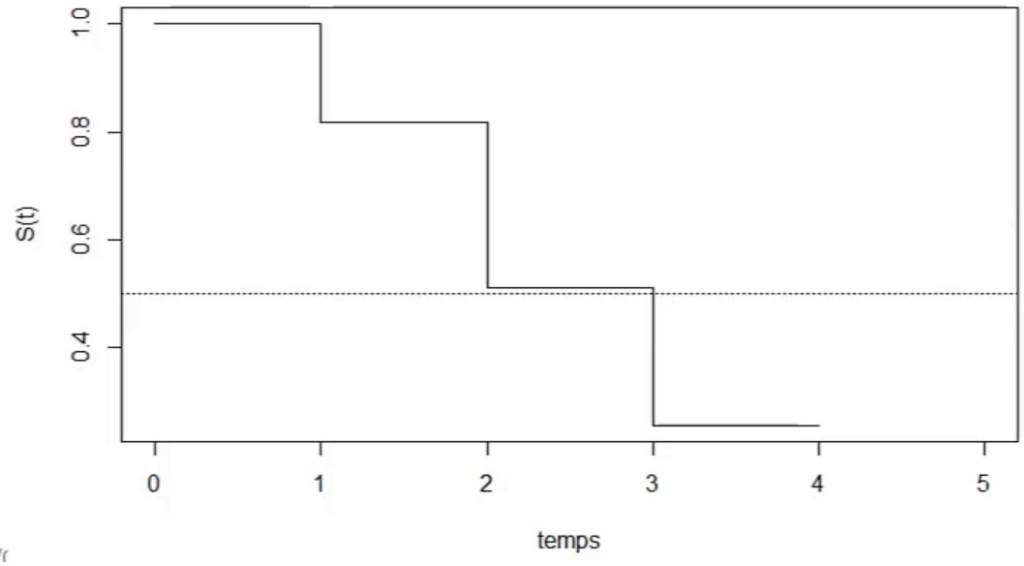
Survie médiane à 3 semaines

TP / Question 4 :

Fonction en escalier, monotone décroissante de 1 à 0 :

- Temps en abscisse
- Taux de survie cumulatif en ordonnée

Marche d’escalier à chaque production d’événement



Sur une courbe de survie, décrire la médiane de survie (la où la barre horizontale coupe la barre verticale à 0.5).

4.B.2.2 Exemple 2 avec R

Faire un objet de type surv reprenant les données de survie.

D'abord charger la librairie `survival`.

```
library(survival)
```

Utiliser un jeu de données avec censure (ex : `lung` dans la librairie `survival`).

```
head(lung) # afficher les premières lignes
```

```
  inst time status age sex ph.ecog ph.karno pat.karno meal.cal wt.loss
1    3  306      2  74   1       1       90       100     1175      NA
2    3  455      2  68   1       0       90        90     1225      15
3    3 1010      1  56   1       0       90        90       NA      15
4    5  210      2  57   1       1       90        60     1150      11
5    1  883      2  60   1       0      100        90       NA       0
6   12 1022      1  74   1       1       50        80      513       0
```

Les colonnes importantes :

- `lung$time` : temps de SUIVI (en jours) = follow-up time
- `lung$status` : état de l'événement (1 = censuré, 2 = décès) = censoring status

Ensuite : on crée un objet de type `Surv`.

Un objet de type `Surv` combine les informations de temps de suivi et d'état de l'événement.

En gros, dans une équation contenant des données de survie, il faut mettre un objet `Surv` pour représenter la variable dépendante, qui contient à la fois le temps de suivi et l'état de l'événement.

```
Surv(lung$time, lung$status)
```

```
  [1]  306   455  1010+  210   883  1022+  310   361   218   166   170   654
 [13]  728    71   567   144   613   707    61    88   301    81   624   371
 [25]  394   520   574   118   390    12   473    26   533   107    53   122
 [37]  814   965+   93   731   460   153   433   145   583    95   303   519
 [49]  643   765   735   189    53   246   689    65     5   132   687   345
 [61]  444   223   175    60   163    65   208   821+  428   230   840+  305
 [73]   11   132   226   426   705   363    11   176   791    95   196+  167
 [85]  806+  284   641   147   740+  163   655   239    88   245   588+   30
 [97]  179   310   477   166   559+  450   364   107   177   156   529+   11
[109]  429   351    15   181   283   201   524    13   212   524   288   363
[121]  442   199   550    54   558   207    92    60   551+  543+  293   202
[133]  353   511+  267   511+  371   387   457   337   201   404+  222    62
[145]  458+  356+  353   163    31   340   229   444+  315+  182   156   329
[157]  364+  291   179   376+  384+  268   292+  142   413+  266+  194   320
[169]  181   285   301+  348   197   382+  303+  296+  180   186   145   269+
[181]  300+  284+  350   272+  292+  332+  285   259+  110   286   270    81
```

```
[193]  131   225+  269   225+  243+  279+  276+  135    79    59   240+  202+
[205]  235+  105   224+  239   237+  173+  252+  221+  185+   92+   13   222+
[217]  192+  183   211+  175+  197+  203+  116   188+  191+  105+  174+  177+
```

Note

```
Surv(
    time,
    time2,
    event,
    type=c('right', 'left', 'interval', 'counting', 'interval2', 'mstate'),
    origin=0)
is.Surv(x)
```

- `time`: temps de suivi pour les données à censure à droite (pour les données à censure par intervalle, le premier argument est le temps de début de l'intervalle).
- `time2`: temps de fin de l'intervalle pour les données à censure par intervalle ou les données de processus de comptage uniquement. Les intervalles sont supposés être ouverts à gauche et fermés à droite, (début, fin].
- `event`: indicateur de statut, normalement 0=vivant, 1=décédé. D'autres choix sont TRUE/FAUX (TRUE = décès) ou 1/2 (2=décès).
- `type=c('right', 'left', 'interval', 'counting', 'interval2', 'mstate')`: type de censure
- `origin=0`: pour les données de processus de comptage, l'origine de la fonction de risque, c'est à dire le temps à partir duquel on commence à compter les événements.
- `is.Surv(x)`: fonction pour vérifier si un objet `x` est de type `Surv`.

Pour faire la table de Kaplan Meier, on utilise la fonction `survfit`.

```
fit <- survfit(Surv(lung$time, lung$status)~ 1)
fit
```

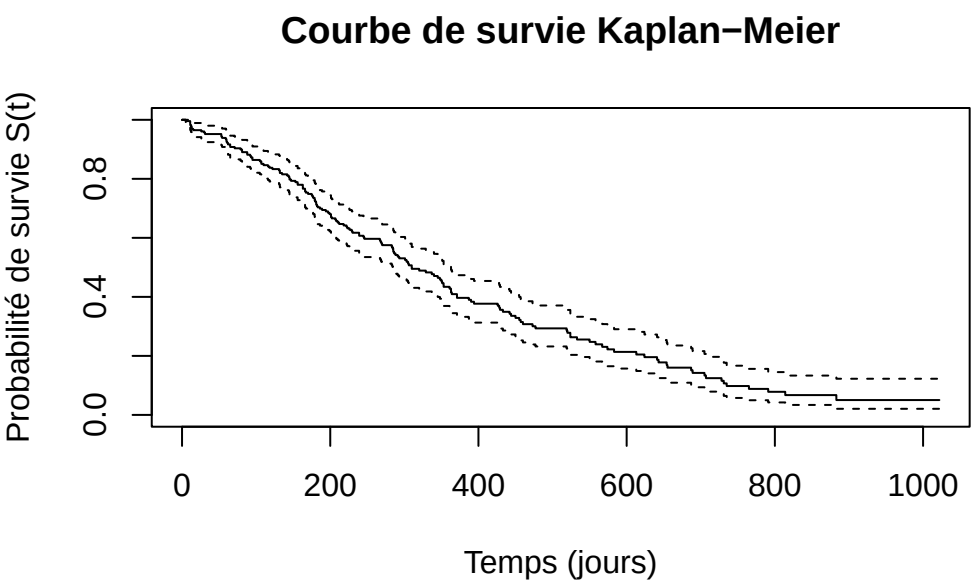
```
Call: survfit(formula = Surv(lung$time, lung$status) ~ 1)

       n events median 0.95LCL 0.95UCL
[1,] 228    165    310     285     363
```

Syntaxe :

- `survfit(formula, data, ...)`
- `formula` : formule de modélisation, où la variable dépendante est un objet `Surv` et le côté droit de la formule spécifie les variables explicatives
- `data` : jeu de données contenant les variables utilisées dans la formule.
- `~ 1` : signifie qu'il n'y a pas de variable explicative, on estime la survie globale.

Pour tracer la courbe de survie Kaplan Meier :

```
plot(
    fit,
    xlab = "Temps (jours)",
    ylab = "Probabilité de survie S(t)",
    main = "Courbe de survie Kaplan-Meier")
```



> **Note**
>
> Pour obtenir de l'aide sur les fonctions :
>
> ```
> ?survfit
> ?plot.survfit
> ```
>
> C'est `plot.survfit` qui est la méthode de traçage pour les objets de type `survfit`, l'aide ne se trouve pas dans `plot` seul car `plot` est une fonction générique qui peut être utilisée pour différents types d'objets.

Pour obtenir le résumé (la table) de l'estimation Kaplan Meier :

```
summary(fit)
```

```
Call: survfit(formula = Surv(lung$time, lung$status) ~ 1)

 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    5    228       1   0.9956 0.00438       0.9871        1.000
   11    227       3   0.9825 0.00869       0.9656        1.000
   12    224       1   0.9781 0.00970       0.9592        0.997
   13    223       2   0.9693 0.01142       0.9472        0.992
   15    221       1   0.9649 0.01219       0.9413        0.989
```

26	220	1	0.9605	0.01290	0.9356	0.986
30	219	1	0.9561	0.01356	0.9299	0.983
31	218	1	0.9518	0.01419	0.9243	0.980
53	217	2	0.9430	0.01536	0.9134	0.974
54	215	1	0.9386	0.01590	0.9079	0.970
59	214	1	0.9342	0.01642	0.9026	0.967
60	213	2	0.9254	0.01740	0.8920	0.960
61	211	1	0.9211	0.01786	0.8867	0.957
62	210	1	0.9167	0.01830	0.8815	0.953
65	209	2	0.9079	0.01915	0.8711	0.946
71	207	1	0.9035	0.01955	0.8660	0.943
79	206	1	0.8991	0.01995	0.8609	0.939
81	205	2	0.8904	0.02069	0.8507	0.932
88	203	2	0.8816	0.02140	0.8406	0.925
92	201	1	0.8772	0.02174	0.8356	0.921
93	199	1	0.8728	0.02207	0.8306	0.917
95	198	2	0.8640	0.02271	0.8206	0.910
105	196	1	0.8596	0.02302	0.8156	0.906
107	194	2	0.8507	0.02362	0.8056	0.898
110	192	1	0.8463	0.02391	0.8007	0.894
116	191	1	0.8418	0.02419	0.7957	0.891
118	190	1	0.8374	0.02446	0.7908	0.887
122	189	1	0.8330	0.02473	0.7859	0.883
131	188	1	0.8285	0.02500	0.7810	0.879
132	187	2	0.8197	0.02550	0.7712	0.871
135	185	1	0.8153	0.02575	0.7663	0.867
142	184	1	0.8108	0.02598	0.7615	0.863
144	183	1	0.8064	0.02622	0.7566	0.859
145	182	2	0.7975	0.02667	0.7469	0.852
147	180	1	0.7931	0.02688	0.7421	0.848
153	179	1	0.7887	0.02710	0.7373	0.844
156	178	2	0.7798	0.02751	0.7277	0.836
163	176	3	0.7665	0.02809	0.7134	0.824
166	173	2	0.7577	0.02845	0.7039	0.816
167	171	1	0.7532	0.02863	0.6991	0.811
170	170	1	0.7488	0.02880	0.6944	0.807
175	167	1	0.7443	0.02898	0.6896	0.803
176	165	1	0.7398	0.02915	0.6848	0.799
177	164	1	0.7353	0.02932	0.6800	0.795
179	162	2	0.7262	0.02965	0.6704	0.787
180	160	1	0.7217	0.02981	0.6655	0.783
181	159	2	0.7126	0.03012	0.6559	0.774
182	157	1	0.7081	0.03027	0.6511	0.770
183	156	1	0.7035	0.03041	0.6464	0.766
186	154	1	0.6989	0.03056	0.6416	0.761
189	152	1	0.6943	0.03070	0.6367	0.757
194	149	1	0.6897	0.03085	0.6318	0.753
197	147	1	0.6850	0.03099	0.6269	0.749

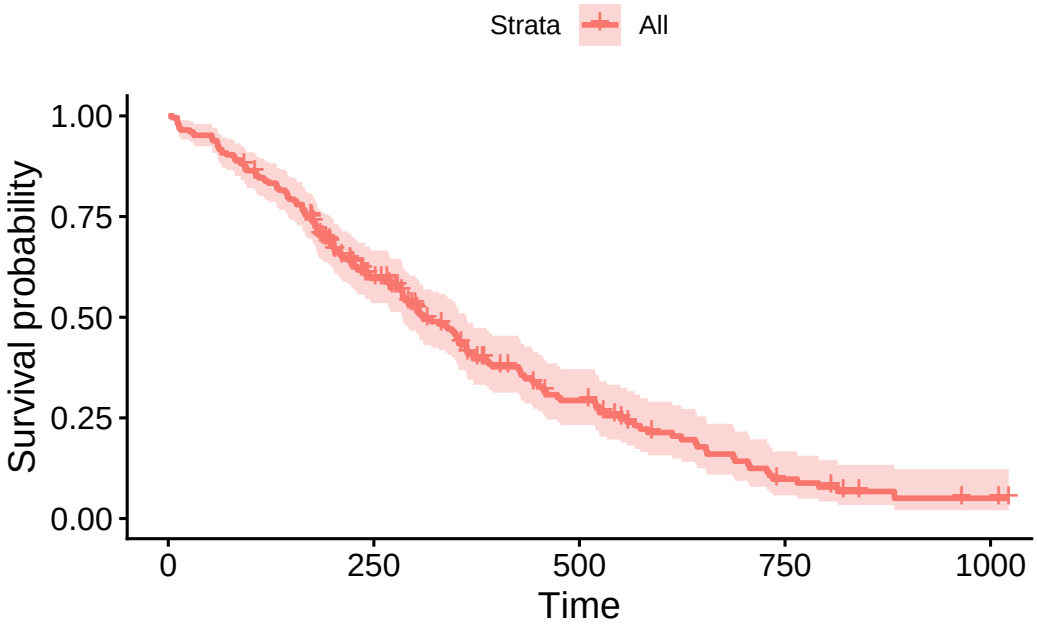
199	145	1	0.6803	0.03113	0.6219	0.744
201	144	2	0.6708	0.03141	0.6120	0.735
202	142	1	0.6661	0.03154	0.6071	0.731
207	139	1	0.6613	0.03168	0.6020	0.726
208	138	1	0.6565	0.03181	0.5970	0.722
210	137	1	0.6517	0.03194	0.5920	0.717
212	135	1	0.6469	0.03206	0.5870	0.713
218	134	1	0.6421	0.03218	0.5820	0.708
222	132	1	0.6372	0.03231	0.5769	0.704
223	130	1	0.6323	0.03243	0.5718	0.699
226	126	1	0.6273	0.03256	0.5666	0.694
229	125	1	0.6223	0.03268	0.5614	0.690
230	124	1	0.6172	0.03280	0.5562	0.685
239	121	2	0.6070	0.03304	0.5456	0.675
245	117	1	0.6019	0.03316	0.5402	0.670
246	116	1	0.5967	0.03328	0.5349	0.666
267	112	1	0.5913	0.03341	0.5294	0.661
268	111	1	0.5860	0.03353	0.5239	0.656
269	110	1	0.5807	0.03364	0.5184	0.651
270	108	1	0.5753	0.03376	0.5128	0.645
283	104	1	0.5698	0.03388	0.5071	0.640
284	103	1	0.5642	0.03400	0.5014	0.635
285	101	2	0.5531	0.03424	0.4899	0.624
286	99	1	0.5475	0.03434	0.4841	0.619
288	98	1	0.5419	0.03444	0.4784	0.614
291	97	1	0.5363	0.03454	0.4727	0.608
293	94	1	0.5306	0.03464	0.4669	0.603
301	91	1	0.5248	0.03475	0.4609	0.597
303	89	1	0.5189	0.03485	0.4549	0.592
305	87	1	0.5129	0.03496	0.4488	0.586
306	86	1	0.5070	0.03506	0.4427	0.581
310	85	2	0.4950	0.03523	0.4306	0.569
320	82	1	0.4890	0.03532	0.4244	0.563
329	81	1	0.4830	0.03539	0.4183	0.558
337	79	1	0.4768	0.03547	0.4121	0.552
340	78	1	0.4707	0.03554	0.4060	0.546
345	77	1	0.4646	0.03560	0.3998	0.540
348	76	1	0.4585	0.03565	0.3937	0.534
350	75	1	0.4524	0.03569	0.3876	0.528
351	74	1	0.4463	0.03573	0.3815	0.522
353	73	2	0.4340	0.03578	0.3693	0.510
361	70	1	0.4278	0.03581	0.3631	0.504
363	69	2	0.4154	0.03583	0.3508	0.492
364	67	1	0.4092	0.03582	0.3447	0.486
371	65	2	0.3966	0.03581	0.3323	0.473
387	60	1	0.3900	0.03582	0.3258	0.467
390	59	1	0.3834	0.03582	0.3193	0.460
394	58	1	0.3768	0.03580	0.3128	0.454

```
  426     55       1   0.3700 0.03580       0.3060        0.447
  428     54       1   0.3631 0.03579       0.2993        0.440
  429     53       1   0.3563 0.03576       0.2926        0.434
  433     52       1   0.3494 0.03573       0.2860        0.427
  442     51       1   0.3426 0.03568       0.2793        0.420
  444     50       1   0.3357 0.03561       0.2727        0.413
  450     48       1   0.3287 0.03555       0.2659        0.406
  455     47       1   0.3217 0.03548       0.2592        0.399
  457     46       1   0.3147 0.03539       0.2525        0.392
  460     44       1   0.3076 0.03530       0.2456        0.385
  473     43       1   0.3004 0.03520       0.2388        0.378
  477     42       1   0.2933 0.03508       0.2320        0.371
  519     39       1   0.2857 0.03498       0.2248        0.363
  520     38       1   0.2782 0.03485       0.2177        0.356
  524     37       2   0.2632 0.03455       0.2035        0.340
  533     34       1   0.2554 0.03439       0.1962        0.333
  550     32       1   0.2475 0.03423       0.1887        0.325
  558     30       1   0.2392 0.03407       0.1810        0.316
  567     28       1   0.2307 0.03391       0.1729        0.308
  574     27       1   0.2221 0.03371       0.1650        0.299
  583     26       1   0.2136 0.03348       0.1571        0.290
  613     24       1   0.2047 0.03325       0.1489        0.281
  624     23       1   0.1958 0.03297       0.1407        0.272
  641     22       1   0.1869 0.03265       0.1327        0.263
  643     21       1   0.1780 0.03229       0.1247        0.254
  654     20       1   0.1691 0.03188       0.1169        0.245
  655     19       1   0.1602 0.03142       0.1091        0.235
  687     18       1   0.1513 0.03090       0.1014        0.226
  689     17       1   0.1424 0.03034       0.0938        0.216
  705     16       1   0.1335 0.02972       0.0863        0.207
  707     15       1   0.1246 0.02904       0.0789        0.197
  728     14       1   0.1157 0.02830       0.0716        0.187
  731     13       1   0.1068 0.02749       0.0645        0.177
  735     12       1   0.0979 0.02660       0.0575        0.167
  765     10       1   0.0881 0.02568       0.0498        0.156
  791      9       1   0.0783 0.02462       0.0423        0.145
  814      7       1   0.0671 0.02351       0.0338        0.133
  883      4       1   0.0503 0.02285       0.0207        0.123
```

Pour faire un autre tracé avec “ggsurvplot” de la librairie “survminer” :

```
library(survminer)
ggsurvplot(fit, data = lung)
```

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
i The deprecated feature was likely used in the ggpubr package.
  Please report the issue at <https://github.com/kassambara/ggpubr/issues>.
```



On va essayer de faire la même chose en comparant les groupes de sexe (variable `sex` dans le jeu de données `lung`).

```
fit_sex <- survfit(Surv(time, status) ~ sex, data = lung)
fit_sex
```

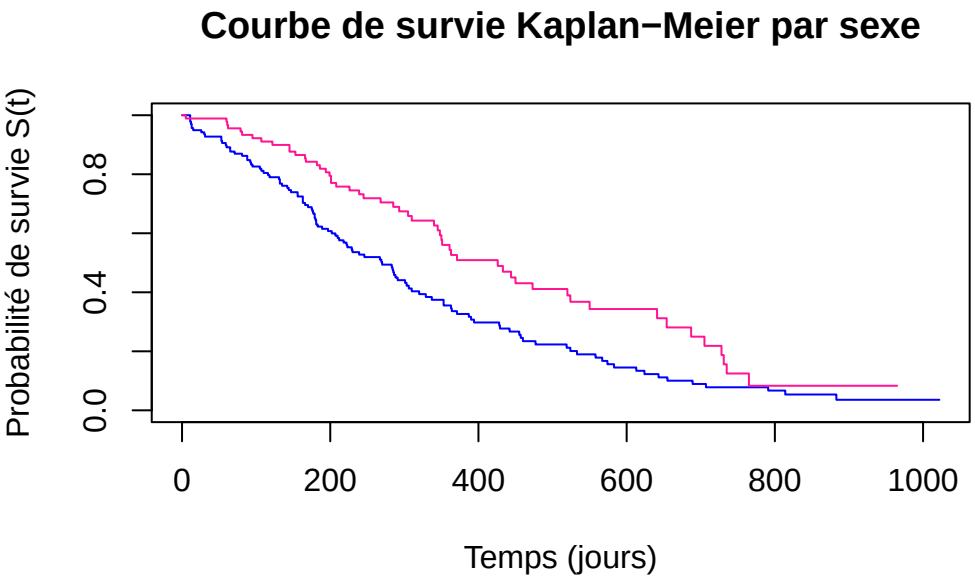
```
Call: survfit(formula = Surv(time, status) ~ sex, data = lung)

        n events median 0.95LCL 0.95UCL
sex=1 138    112    270     212     310
sex=2  90     53    426     348     550
```

Syntaxe :

- `Surv(time, status) ~ sex` : on modélise la survie en fonction de la variable explicative `sex`.

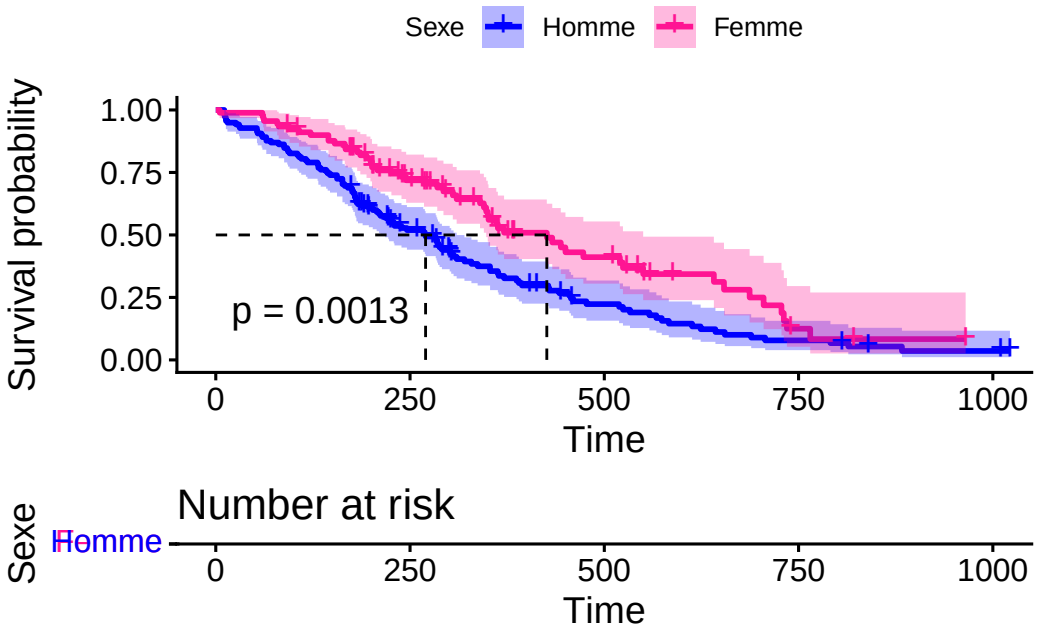
Pour tracer la courbe de survie Kaplan Meier par sexe :

```
plot(
    fit_sex,
    xlab = "Temps (jours)",
    ylab = "Probabilité de survie S(t)",
    main = "Courbe de survie Kaplan-Meier par sexe",
    col = c("blue", "deeppink"),
    conf.int = FALSE # pas d'intervalle de confiance
)
```



On le refait avec `ggsurvplot` + ajout de l'intervalle de confiance + test de log-rank.

```
ggsurvplot(
    fit_sex,
    data = lung,
    conf.int = TRUE,
    pval = TRUE,
    risk.table = TRUE,
    risk.table.col = "strata",
    legend.labs = c("Homme", "Femme"),
    legend.title = "Sexe",
    palette = c("blue", "deeppink"),
    surv.median.line = "hv"
)
```



5 Comparaison de courbes de survie : test de Log-Rank

2 cadres principaux :

- Essai comparatif : différence de délais de survie entre 2 groupes
- Étude épidémiologique (cohorte) : impact de facteurs de risque sur la survie

Comparaison des fonctions de survie dans leur ensemble.

5.A Principe

Comparaison de deux (p) fonctions de survie à partir de deux (p) échantillons indépendants.
Comparaison de deux fonctions de survie / totalité des courbes
Hypothèse nulle H_0 : les deux fonctions de survie ne sont pas différentes
Hypothèse alternative H_1 : les deux fonctions de survie sont différentes

5.B Test de Log-Rank dans R

Utilisation de la fonction `survdiff` de la librairie `survival`.

```r
# Charger les librairies nécessaires
library(survival)
library(survminer)
# Créer un objet Surv
surv_object <- Surv(time = lung$time, event = lung$status)
# Ajuster le modèle de survie par sexe
surv_fit <- survfit(surv_object ~ lung$sex)
# Effectuer le test de Log-Rank : fonction survdiff
logrank_test <- survdiff(surv_object ~ lung$sex)
# Afficher les résultats du test
print(logrank_test)
```

```
Call:
survdiff(formula = surv_object ~ lung$sex)

             N Observed Expected (O-E)^2/E (O-E)^2/V
lung$sex=1 138      112     91.6      4.55      10.3
lung$sex=2  90       53     73.4      5.68      10.3

 Chisq= 10.3  on 1 degrees of freedom, p= 0.001
```

Résultats :

- **N** : nombre d'individus dans chaque groupe (138 hommes, 90 femmes).
- **Observed** : nombre d'événements observés (décès) dans chaque groupe (112 hommes, 53 femmes).

- **Expected** : nombre d'événements attendus dans chaque groupe sous l'hypothèse nulle (91.6 hommes, 73.4 femmes).
- **(O-E)^2/E** : contribution de chaque groupe au test du chi carré basé sur la différence entre les événements observés et attendus.
- **(O-E)^2/V** : contribution de chaque groupe au test du chi carré basé sur la variance des différences observées-attendues.
- **Chisq** : statistique du test du chi carré (10.3).
- **degrees of freedom** : degrés de liberté du test (1, car deux groupes).
- **p** : valeur p associée au test (0.001).

Interprétation :

- La valeur p (0.001) est inférieure au seuil de signification habituel (0.05), ce qui indique une différence statistiquement significative entre les fonctions de survie des hommes et des femmes.

Conclusion : on rejette l'hypothèse nulle et on conclut qu'il y a une différence significative entre les fonctions de survie des hommes et des femmes dans cette étude.

> **!** Important
>
> Donc les étapes sont:
>
> 1. Créer un objet de survie `Surv` comprenant les temps de suivi et l'état de l'événement.
> 2. Ajuster un modèle de survie avec `survfit` en fonction de la variable explicative (ici le sexe).
> 3. Effectuer le test de Log-Rank avec la fonction `survdiff`.