# 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

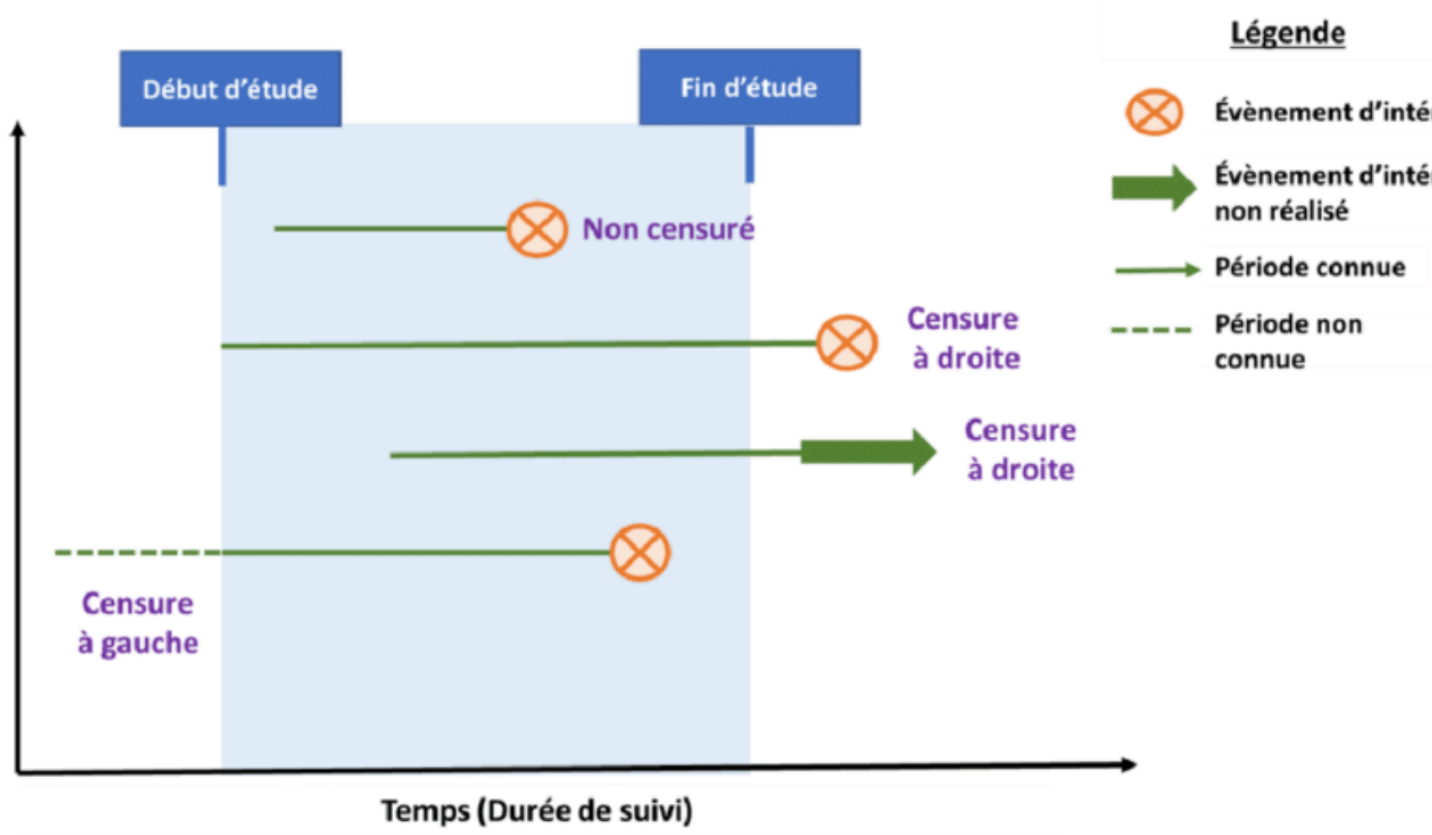
**Donnée censuré ≠ donnée manquante** :

- **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

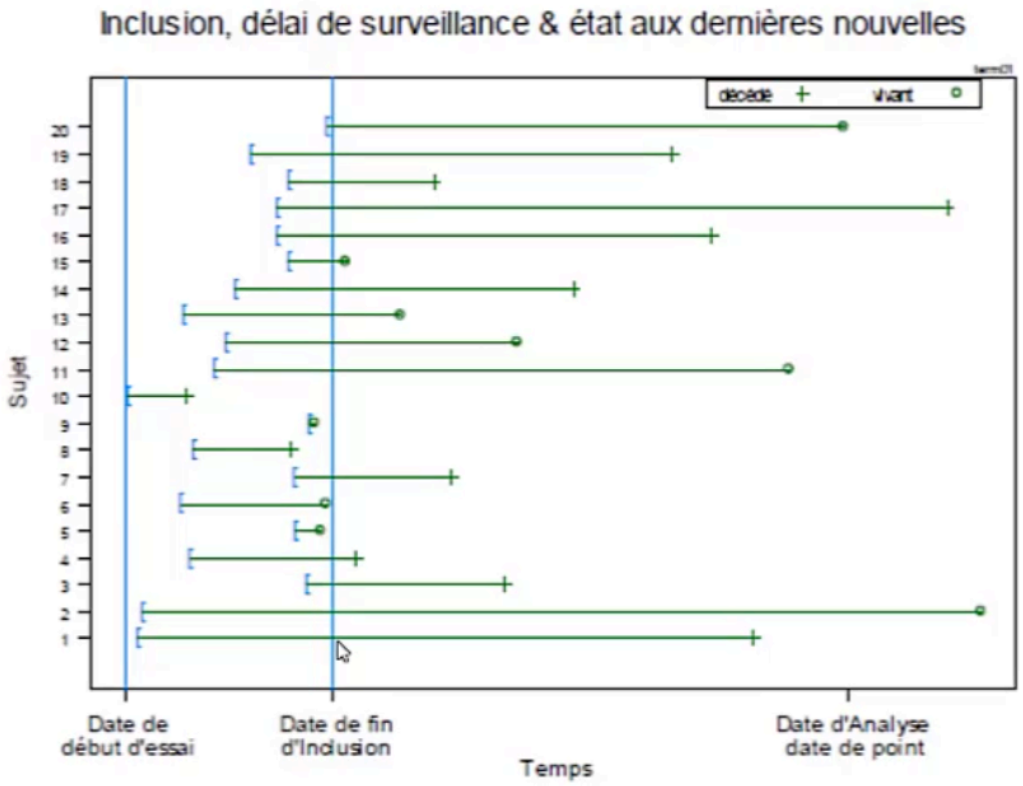
- **Date d'origine** : point de départ, varie selon patient, pour le calcul des durées de survie (ex : date de diagnostic, date de traitement, date d'inclusion dans l'étude)
- **Date des dernières nouvelles** : date de la dernière information connue sur le patient (ex : date de décès, date de la dernière 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

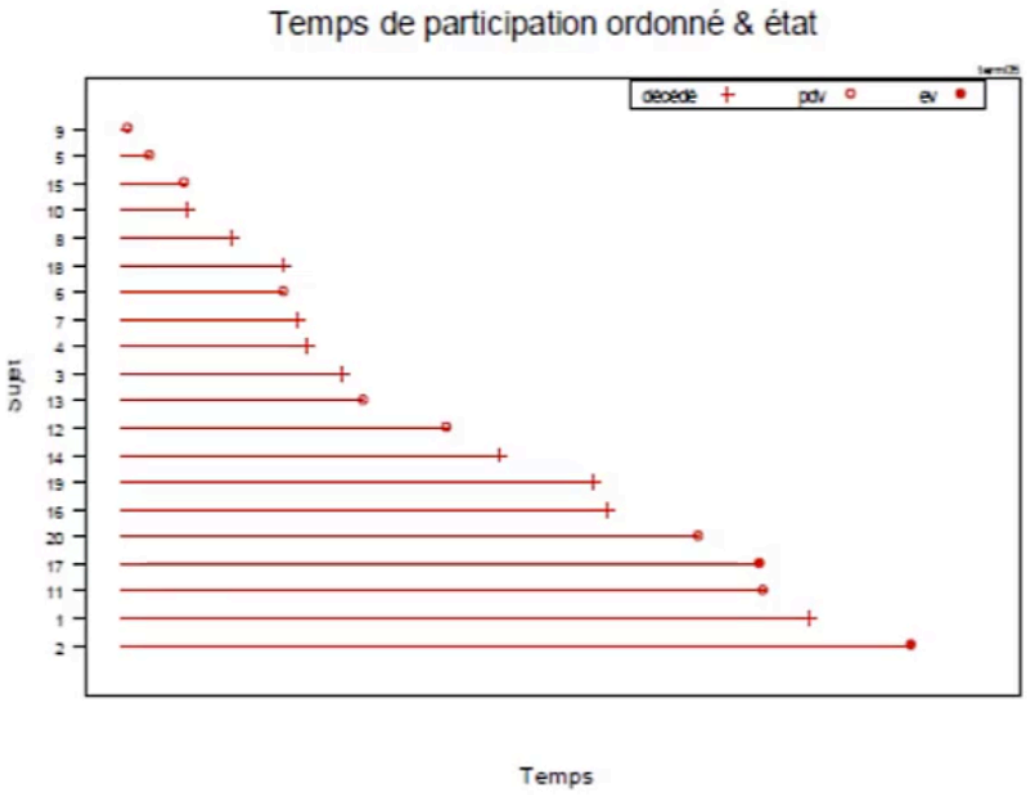
Pour chaque sujet, on doit disposer de :

- **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

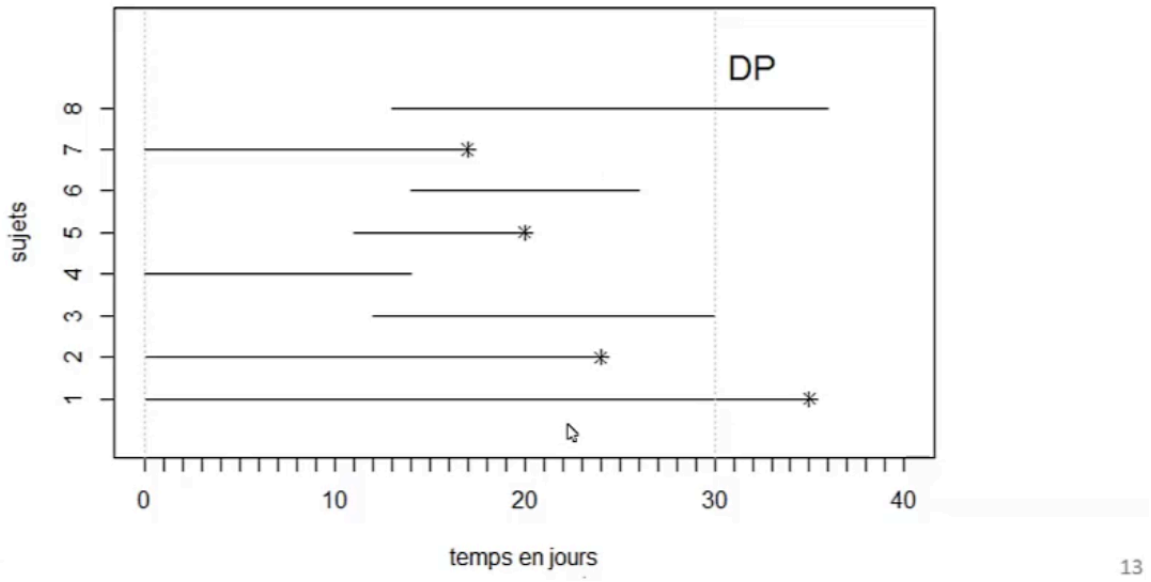


# Exemple : données simulées



### 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 $\Delta t$ 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 $\Delta t$ 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 :

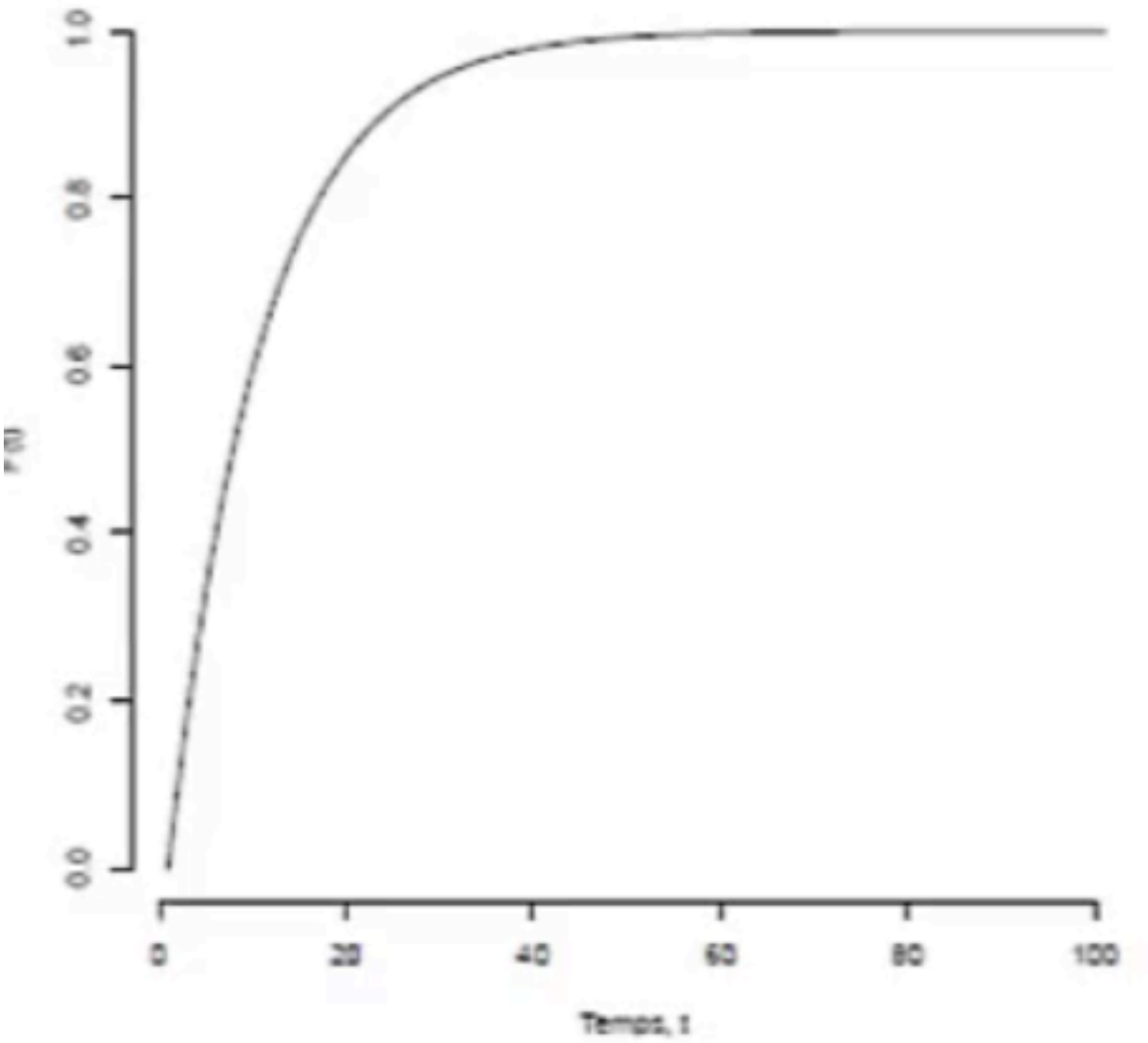
1. Différence entre :
    - La probabilité que l'évènement $T$ se produise avant un certain temps $t + \Delta t$ (un intervalle de temps très petit autour de $t$)
    - et la probabilité que l'évènement $T$ se produise avant un certain temps $t$
2. Diviser le résultat par la taille de l'intervalle de temps $\Delta t$
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 $\Delta t$ 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.



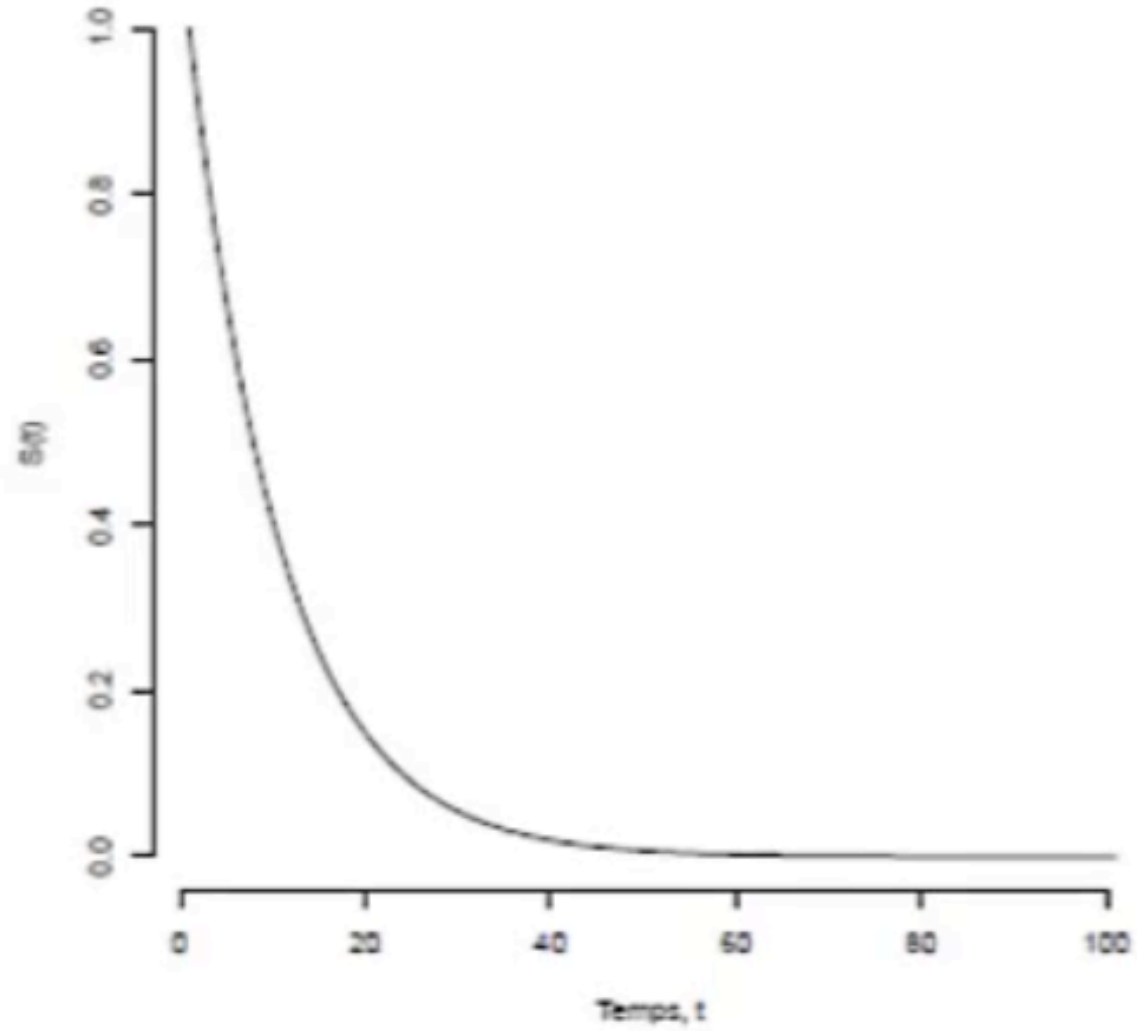

**Propriétés :**

- 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 $t$.
- $\Delta t$ : 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 $\Delta t$ 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 $\Delta t$

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

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

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

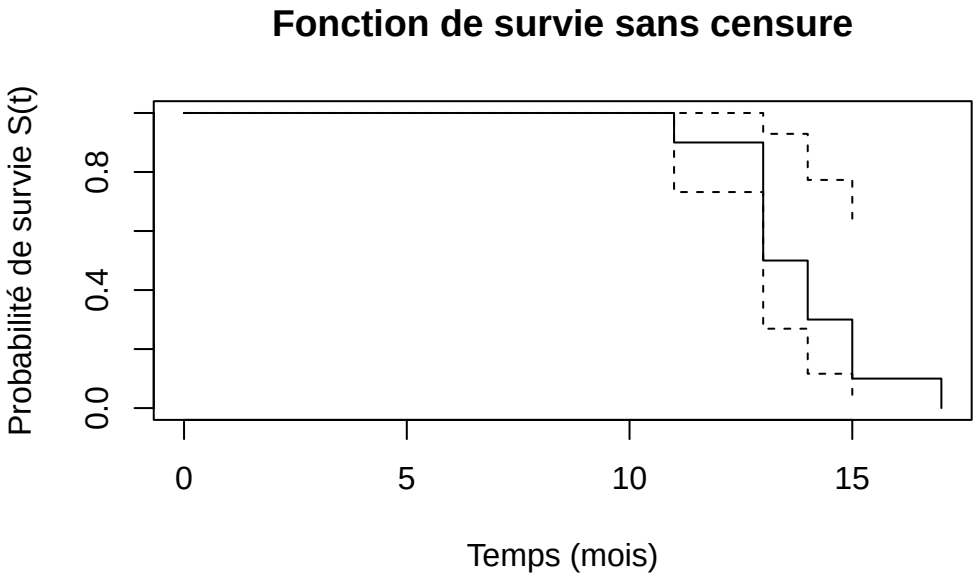
en tableau :

• 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 |
| 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.

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

```
[1] 38 22  2 19 35 22
```

```
table(T)
```

```
T
 0  1  2  3  5  6  7  8 10 11 12 14 16 17 18 19 20 21 22 23 24 25 26 28 30 31
 2  2  4  3  3  1  1  6  1  1  1  2  1  2  2  5  2  2  3  3  2  1  1  2  2  1
32 33 34 35 36 37 38 39 40 41 43 45 46 47 49
 2  2  1  3  1  1  4  1  1  1  2  1  2  1  1
```

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

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

```
[1] 41
```

Il faut aussi ordonner les valeurs uniques.

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

```
 [1]  0  1  2  3  5  6  7  8 10 11 12 14 16 17 18 19 20 21 22 23 24 25 26 28 30
[26] 31 32 33 34 35 36 37 38 39 40 41 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  TRUE FALSE  TRUE  TRUE  TRUE
```

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

```
[1] 0.7125
```

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

```
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.

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

```
[1] 0.7125
```

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

```
[1] 0.5125
```

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.

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

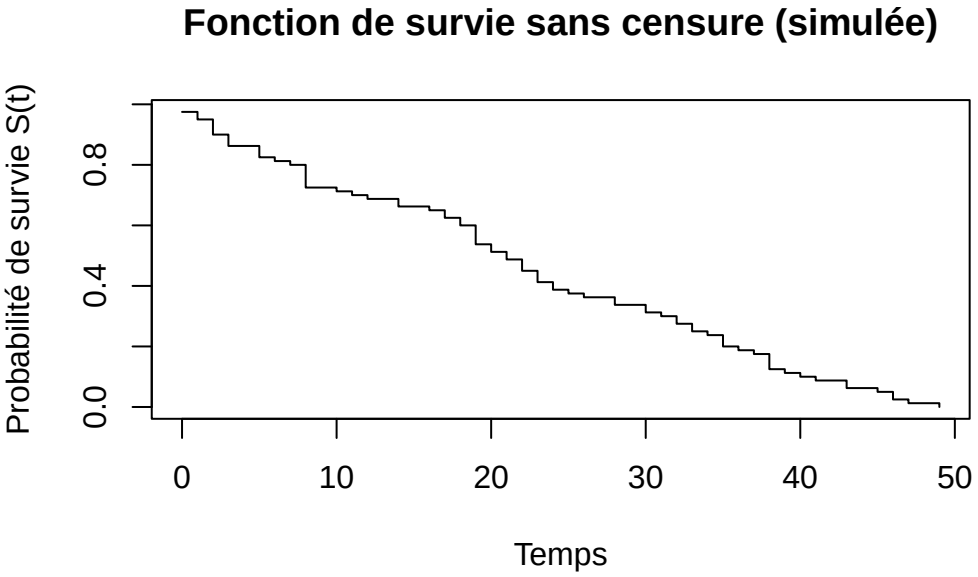
```
 [1] 0.9750 0.9500 0.9000 0.8625 0.8250 0.8125 0.8000 0.7250 0.7125 0.7000
[11] 0.6875 0.6625 0.6500 0.6250 0.6000 0.5375 0.5125 0.4875 0.4500 0.4125
[21] 0.3875 0.3750 0.3625 0.3375 0.3125 0.3000 0.2750 0.2500 0.2375 0.2000
[31] 0.1875 0.1750 0.1250 0.1125 0.1000 0.0875 0.0625 0.0500 0.0250 0.0125
[41] 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

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

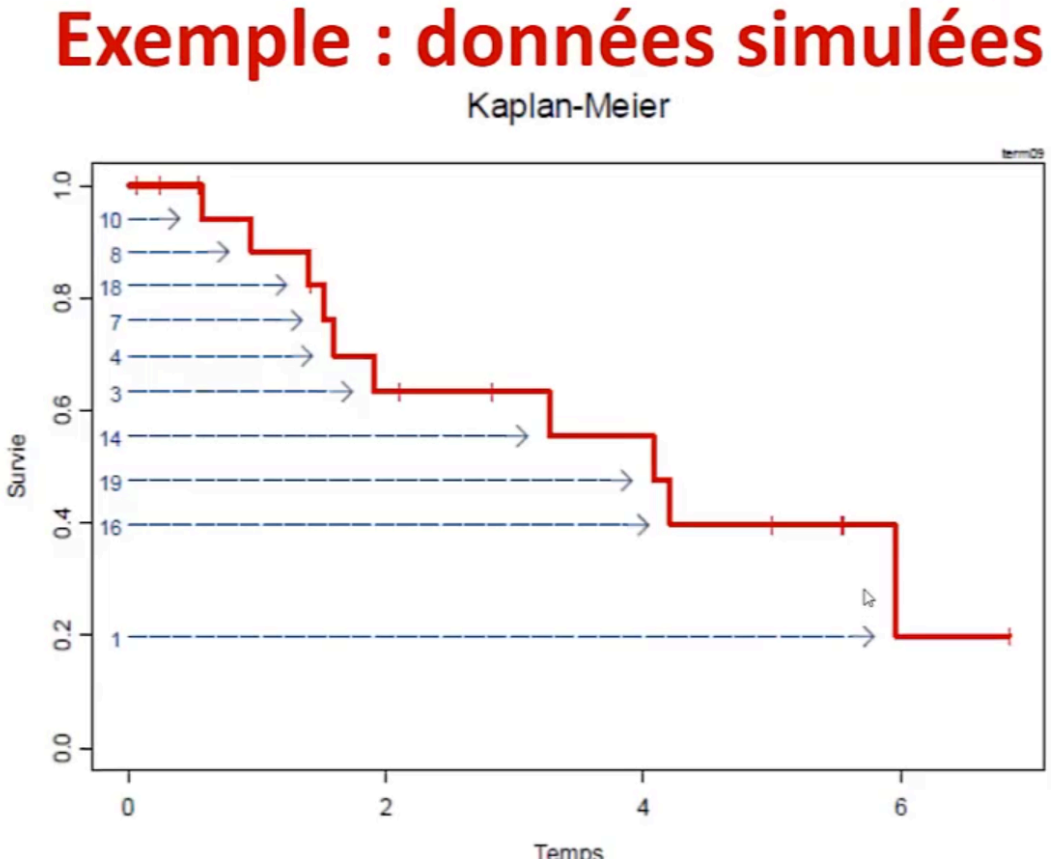
$S_k$ = probabilité cumulée de survie jusqu'au temps $t_k$ = $S(t_k)$

| $t_k$ (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 conditionnelle 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 |

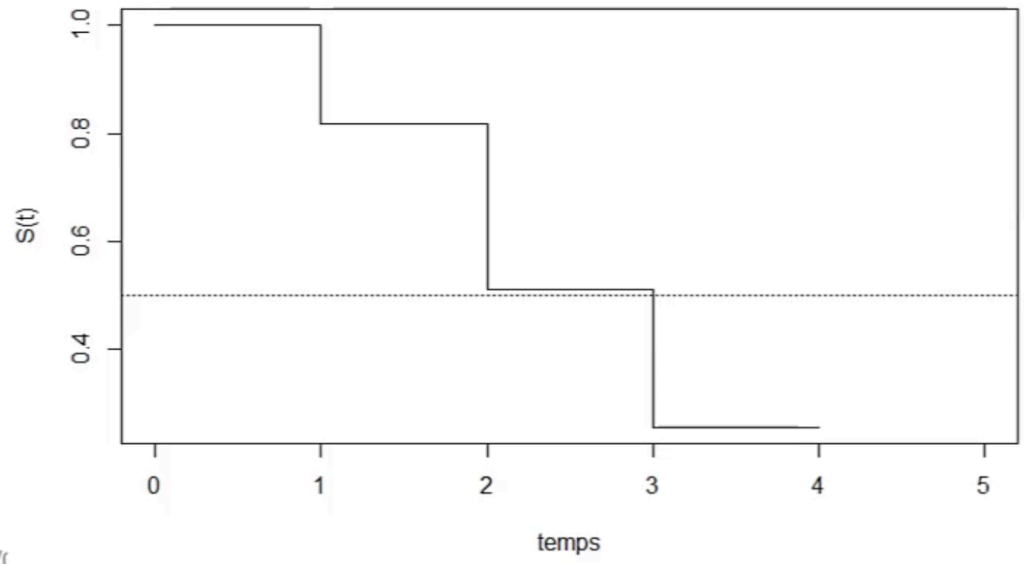
Survie médiane à 3 semaines

# 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
```

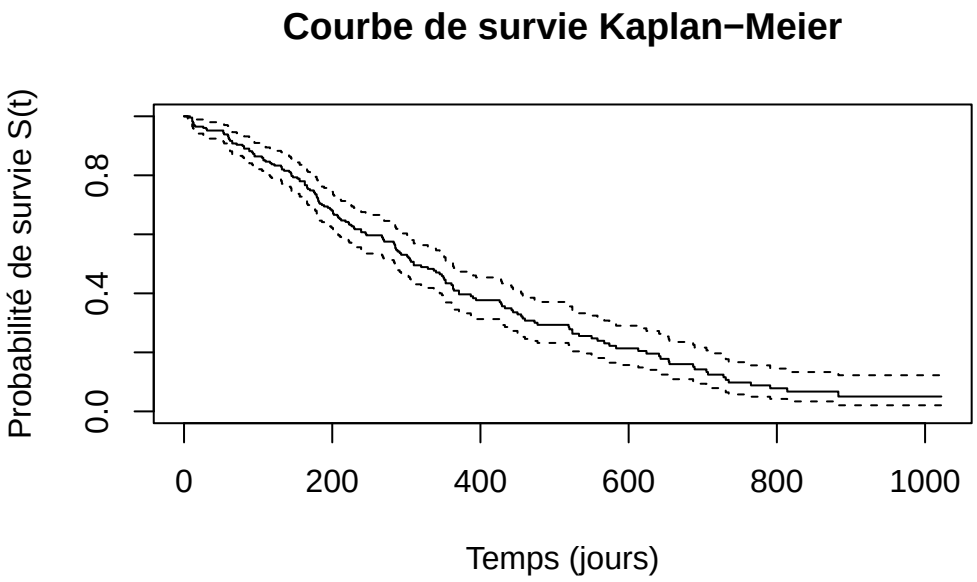
Syntaxe :

- `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 |

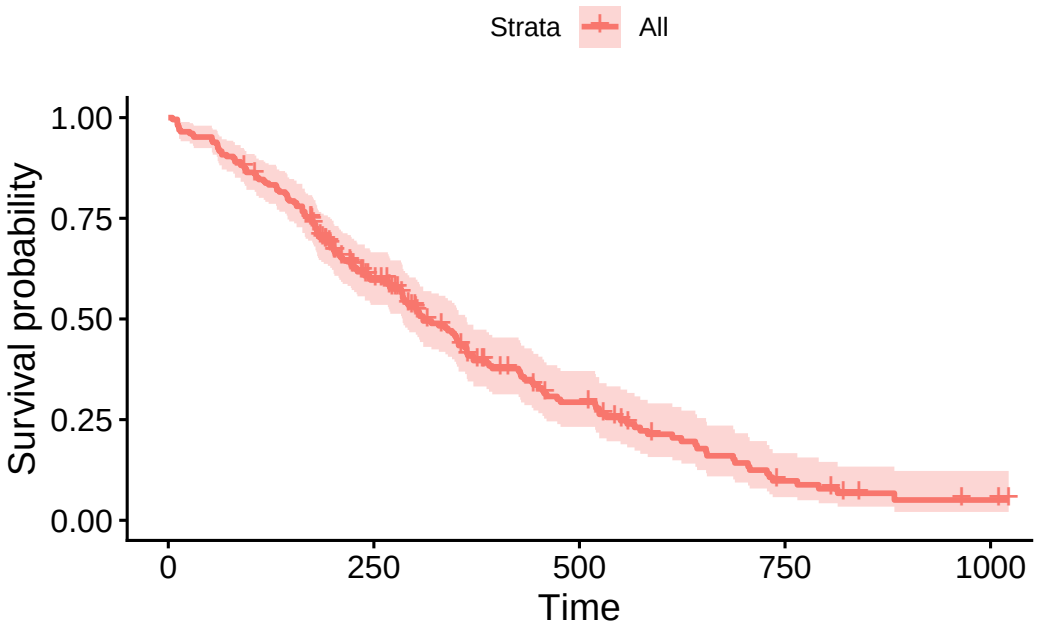
| | | | | | | |
|---|---|---|---|---|---|---|
| 183 | 156 | 1 | 0.7035 | 0.03041 | 0.6464 | 0.766 |
| 186 | 154 | 1 | 0.6989 | 0.03056 | 0.6416 | 0.761 |
| 189 | 152 | 1 | 0.6943 | 0.03070 | 0.6367 | 0.757 |
| 194 | 149 | 1 | 0.6897 | 0.03085 | 0.6318 | 0.753 |
| 197 | 147 | 1 | 0.6850 | 0.03099 | 0.6269 | 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
```

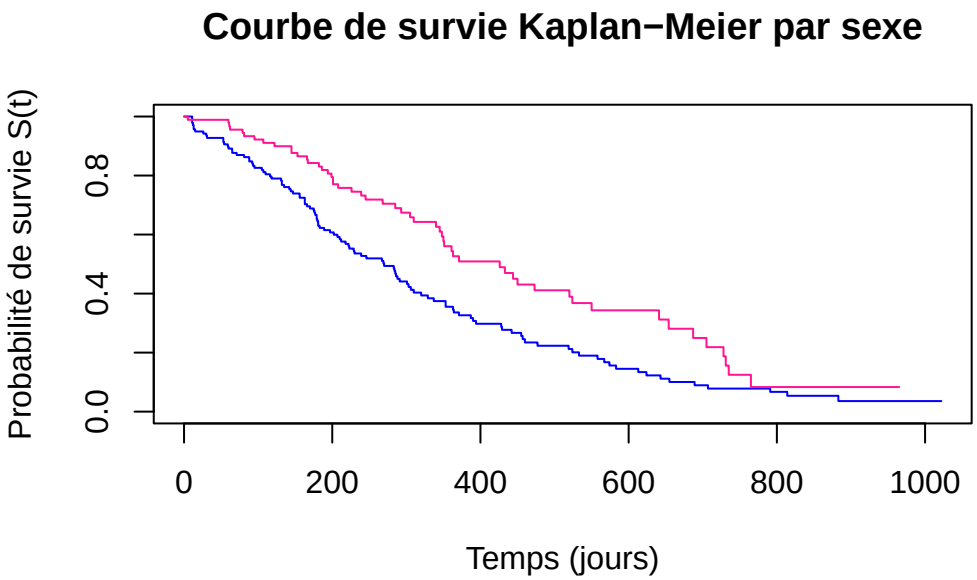
Syntaxe :

- `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 :**

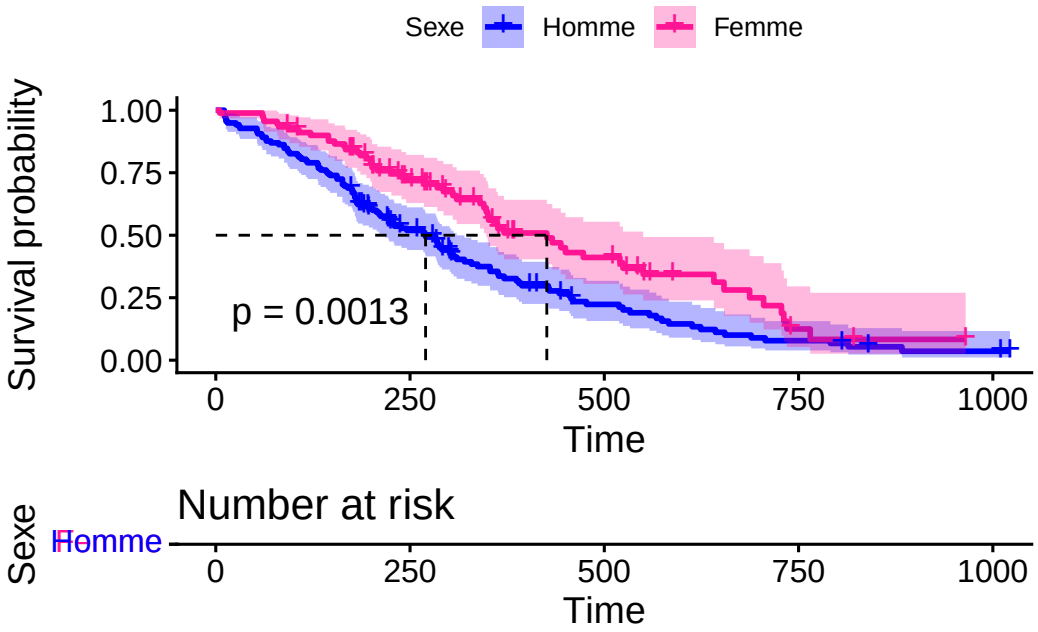
```
plot(
    fit_sex,
    xlab = "Temps (jours)",
    ylab = "Probabilité de survie S(t)",
```

```
    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)
```

```r
# 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`.

---

# 6 Modèle de Cox : analyse multivariée de survie

Modèle de Cox = modèle semi-paramétrique = modélisation de l'effet des variables explicatives sur la fonction de risque sans faire d'hypothèse sur la forme de la fonction de risque.

- Décrire et tester la dépendance
  - Entre une réponse (fonction de risque instantané $h(t)$ ou $\lambda(t)$)
  - et un vecteur de variables explicatives (covariables qualitatives ou quantitatives) $Z = (Z_1, Z_2, ..., Z_p)$

$h(t|Z_1, ..., Z_p) = h_0(t) \times \exp(\beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_p Z_p)$

- $h_0(t)$ : fonction de risque de base (baseline hazard function) = fonction de risque instantané lorsque toutes les covariables $Z_i$ sont égales à 0.
- $\exp(\beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_p Z_p)$ : effet multiplicatif des P covariables sur la fonction de risque instantané.
- $\beta_i$ : coefficient associé à la covariable $Z_i$, représentant l'effet de cette covariable sur le risque instantané.
  - Relation entre $\beta_i$ et le hazard ratio (HR) : $HR = \exp(\beta_i)$ (donc $\beta_i = \log(HR)$)

**2 hypothèses principales** : **proportionnalité des risques** et **log-linéarité**.

## 6.A Hypothèse de proportionnalité des risques

**Hypothèse de proportionnalité des risques** : les rapports des risques entre les individus restent constants dans le temps, c'est à dire que l'effet des covariables sur le risque instantané est multiplicatif et ne dépend pas du temps.

$\frac{h(t|Z_k=1)}{h(t|Z_k=0)} = \exp(\beta k) = constante$

Quand on compare 2 groupes (exemple : fumeur vs non-fumeur), le rapport de leurs risques instantanés reste le même tout au long du suivi, et ce rapport de risques = hazard ratio (HR).

Dire que les risques sont proportionnels signifie :

- à chaque instant t, le groupe A a par exemple 1,5 fois plus de risque de l'événement que le groupe B ;
- ce facteur 1,5 ne change pas au cours du temps.

Exemple : On suit 100 patients opérés.

Variable explicative : fumeur (1) vs non-fumeur (0).

Supposons : HR = 2.

Cela veut dire :

- à tout moment du suivi, un fumeur a le double du risque instantané d'avoir une complication,
- même si le risque global diminue avec le temps (fin de la période postopératoire aiguë), le ratio reste stable entre fumeur et non fumeur.

Si cette hypothèse n'est pas vraie (ex : au début les fumeurs sont à très haut risque, mais plus tard

le risque redevient identique), alors le modèle de Cox classique n'est plus adapté. (*Dans ce cas, on peut utiliser des modèles de Cox avec effets temporels*).

> **i** Note

### 6.B Hypothèse de log-linéarité

**Hypothèse de log-linéarité** : **concerne les variables quantitatives**.

L'effet des covariables est linéaire sur le logarithme du risque instantané et pas directement sur le risque.

C'est à dire que chaque unité d'augmentation de la covariable $Z_i$ entraîne une augmentation constante du logarithme du risque instantané.

$\rightarrow$ Chaque augmentation d'une unité de la variable □ produit le même pourcentage de variation du risque.

$\log(h(t|Z_1, ..., Z_p)) = \log(h_0(t)) + \beta_1 Z_1 + \beta_2 Z_2 + ... + \beta_p Z_p$

Exemple :

Variable explicative : âge (en années).

On suppose que $\beta = 0,05$ (avec $\beta$ le coefficient associé à l'âge dans le modèle de Cox).

Cela veut dire :

- chaque année supplémentaire □ multiplie le risque instantané par $exp(0,05) \approx 1,05$
- donc +5 % de risque par année d'âge.
  - Si on passe de 50 à 51 ans : +5 %.
  - Si on passe de 70 à 71 ans : encore +5 %.
- L'effet est proportionnel, constant.

> **i** Note

### 6.C Interprétation des coefficients du modèle de Cox

Modèle de Cox : $h(t \mid Z) = h_0(t) \times \exp(\beta Z)$

où :

- $h_0(t)$ est le risque instantané de base,
- $Z$ est la covariable (discrète ou continue),
- $\beta$ est le coefficient associé,
- $\exp(\beta)$ est le hazard ratio (HR).

### 6.C.1 Variable discrète

**Variable discrète** (ex : sexe, fumeur/non-fumeur)** :

Groupe de référence A avec $Z_A = 0$ et groupe comparé B avec $Z_B = 1$.

- $Z_A = 0 : h_A(t) = h_0(t)$ (risque instantané du groupe de référence))
- $Z_B = 1 : h_B(t) = h_0(t) \times \exp(\beta)$ (risque instantané du groupe comparé))

Le hazard ratio est :

$HR = \frac{h_B(t)}{h_A(t)} = \frac{h_0(t) \times \exp(\beta)}{h_0(t)} = \exp(\beta)$

Points importants :

- Le HR ne dépend pas du risque de base $h_0(t)$ (mais seulement de $\beta$)
- Il est supposé constant et multiplicatif dans le temps (proportionnalité des risques).
- Il s'interprète comme un rapport d'incidence instantané (ce n'est pas une probabilité).

Exemple : On étudie la mortalité après chirurgie.

Variable : fumeur (1) vs non-fumeur (0).

Le modèle donne $\beta = 0{,}69$.

Alors :

$HR = \exp(0{,}69) \approx 2$

Interprétation :

- À tout instant du suivi, un fumeur a deux fois plus de risque instantané de mourir qu'un non-fumeur.
- L'effet est multiplicatif et constant au cours du temps.
- Ce n'est pas une probabilité ; c'est un rapport de taux instantanés.

### 6.C.2 Variable continue

**Variable continue** (ex : âge, pression artérielle)** :

Le modèle utilise :

$h(t \mid X) = h_0(t) \times \exp(\beta X)$

Ici, $\exp(\beta)$ correspond au HR pour une unité d'augmentation de la variable.

- Le Hazard Ratio est exprimé en unité de $x$ (ex HR pour 1 an d'âge)
- Hypothèse de **modélisation** :
  - Chaque unité d'augmentation de la variable □ même pourcentage de variation du risque instantané
  - L'effet d'un an de plus sur le risque instanté ne dépend pas de l'âge (il est indépendant du niveau de la variable)
  - L'effet est constant dans le temps (proportionnalité des risques)

Exemple : Variable : âge (en années).

Supposons $\beta = 0{,}05$.
Alors : $HR = \exp(0{,}05) \approx 1{,}051$
Interprétation :

- Pour chaque année supplémentaire, le risque instantané augmente de 5,1 %.
- Passer de 50 □ 51 ans : +5,1 %.
- Passer de 70 □ 71 ans : encore +5,1 %.
- Le modèle considère donc une pente constante sur le logarithme du risque.

Cela découle directement de l'écriture :
$\log h(t) = \log h_0(t) + \beta X$
qui impose une relation linéaire sur le log-risque.

### 6.C.3 Interprétation du HR

- **HR = 1** : pas d'effet de la covariable sur le risque instantané.
- **HR > 1** : la covariable est associée à une augmentation du risque instantané.
  - Ex : HR = 2 □ doublement du risque instantané.
  - Ex : HR = 1,5 □ augmentation de 50 % du risque instantané.
- **HR < 1** : la covariable est associée à une diminution du risque instantané.

**Modèle univariable** : HR brut
**Modèle multivariable** : HR ajusté (pour les autres covariables du modèle)

> **!** Important
>
> **Ne pas confondre** :
>
> - Valeur vraie dans la population (interprétable directoement)
> - Estimation dans l'échantillon (avec intervalle de confiance et test statistique)

### 6.C.4 Résumé

### 6.C.5 Tableau de synthèse : interprétation des coefficients du modèle de Cox

| Variable | Forme du modèle | HR = $exp(\beta)$ | Interprétation | Exemple |
|---|---|---|---|---|
| **Discrète binaire** (0/1) | `$h(t\|Z) = h0(t) * exp(\beta Z)$` | $exp(\beta)$ | Rapport de risque instantané entre Z=1 et Z=0, constant dans le temps | $\beta = 0.69$ □ HR = 2 : risque doublé |

| Variable | Forme du modèle | HR = $exp(\beta)$ | Interprétation | Exemple |
|---|---|---|---|---|
| **Discrète multi-catégorielle** | Référence + indicateurs | $exp(\beta_k)$ | Comparaison de chaque catégorie à la référence | Réf = non-fumeur ; fumeur HR=2 ; ex-fumeur HR=1.3 |
| **Continue (par unité)** | $`h(t\|X) = h0(t) * exp(\beta X)`$ | $exp(\beta)$ | Variation du risque instantané pour +1 unité de X | $\beta = 0.05$ □ HR = 1.051 : +5.1 % par unité |
| **Continue (pour ΔX unités)** | $`h(t\|X) = h0(t) * exp(\beta X)`$ | $exp(\beta \cdot \Delta X)$ | Variation du risque pour ΔX unités | $\Delta X = 10$ ans : HR = exp$(0.05 \times 10) = 1.65$ |
| **Log-linéarité** | Linéarité sur log(h) | — | Chaque unité □ même pourcentage de variation du risque | Effet constant : 50□ 51 = 70□ 71 |
| **Proportionnalité des risques** | HR constant dans le temps | — | L'effet ne change pas avec le temps | Courbes "parallèles" en risque instantané |

## 6.D Tests statistiques pour les coefficients du modèle de Cox

Le modèle de Cox estime un ensemble de coefficients $\beta$ . Pour tester si une covariable a un effet significatif sur le risque instantané, deux tests principaux existent :

1. Le test de Wald
2. Le test du rapport de vraisemblance (Likelihood Ratio Test, LRT)

### 6.D.1 Hypothèses

- **Hypothèse nulle** $H_0$ : la covariable n'a pas d'effet sur le risque instantané ($\beta = 0$ □ $HR = 1$)
- **Hypothèse alternative** $H_1$ : la covariable a un effet sur le risque instantané ($\beta \neq 0$ □ $HR \neq 1$)

### 6.D.2 Tests disponibles : Wald et Rapport de vraisemblance (Likelihood Ratio Test, LRT)

#### 6.D.2.1 Test de Wald

Le modèle de Cox fournit une estimation de l'écart type du coefficient $\hat{\beta}$.

Le test de Wald utilise l'estimation du coefficient $\hat{\beta}$ et son écart type (son incertitude) pour calculer une statistique de test.

Calcul de la statistique de Wald :

$Z = \frac{\hat{\beta}}{SE(\hat{\beta})}$

Avec :

- $\hat{\beta}$ : estimateur du coefficient de la covariable
- $SE(\hat{\beta})$ : écart type de l'estimateur

Sous l'hypothèse nulle, $Z$ suit une loi normale centrée réduite, parce que pour de grands échantillons, l'estimateur $\hat{\beta}$ est asymptotiquement normal, c'est à dire que sa distribution approche une loi normale lorsque la taille de l'échantillon augmente.

On compare donc la valeur absolue de $Z$ à la table de la loi normale pour obtenir une p-value.

ex : si $|Z| > 1.96$, p-value < 0.05

- Si $|Z|$ est grand □ la variable est significative.
- Si $|Z| < 1.96$ □ p-value > 0.05 □ variable non significative.

Exemple : Supposons que le coefficient pour fumeur soit :

- $\hat{\beta}$ = 0.69
- $SE(\hat{\beta})$ = 0.30

$Z = 0.69/0.30 = 2.30$

□ p-value < 0.05 □ effet significatif.

> **!** Important
>
> Le test de Wald donne une p-value pour chaque coefficient du modèle, parce qu'il teste chaque coefficient séparément.

**Pourquoi le test de Wald donne une p-value par coefficient ?**

- Parce qu'il teste chaque coefficient séparément :
- $H_0 : \beta_i = 0$
- et chaque coefficient $\beta_i$ est testé individuellement avec son propre estimateur $\hat{\beta}_i$ et son propre écart type $SE(\hat{\beta}_i)$.
- Donc, si le modèle contient :
  - 1 variable quantitative □ 1 coefficient □ 1 p-value
  - 1 variable qualitative à 3 modalités □ 2 coefficients □ 2 p-values
  - p variables □ p p-values

---

#### 6.D.2.2 Test du rapport de vraisemblance (Likelihood Ratio Test, LRT)

= modèles emboîtés que l'on compare

= **modèle complet** = avec la covariable *vs* **modèle réduit** = sans la covariable

On compare la log-vraisemblance des deux modèles (log-vraisemblance = mesure de l'adéquation du modèle aux données).

Calcul de la statistique du rapport de vraisemblance :

$LR = 2 \times (\log L_2 - \log L_1)\ \chi^2(\text{p } ddl)$

Avec :

- $L_1$ : la log-vraisemblance du modèle réduit (le moins riche)
- $L_2$ : la log-vraisemblance du modèle complet (le plus riche)
- $p$ : nombre de paramètres ajoutés entre les deux modèles (nombre de degrés de liberté)

Sous $H_0$, le test du rapport de vraisemblance suit une loi du chi carré avec $p$ degrés de liberté.

*Quand on dit que le test est fait sous* $H_0$, cela signifie que l'on suppose que la covariable n'a pas d'effet sur le risque instantané.

On compare la valeur de $LR$ à la table du chi carré pour obtenir une p-value.

**Pourquoi le LRT donne une seule p-value par variable ?**

Car il ne teste pas $\beta_1, \beta_2, \beta_3$ individuellement mais l'ensemble du bloc associé à la variable :

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

Donc :

- Une variable binaire □ 1 coefficient □ 1 p-value
- Une variable qualitative à 4 niveaux □ 3 coefficients □ 1 p-value globale
- Un bloc de variables (ex : spline, interaction, polynôme) □ 1 p-value globale

#### 6.D.2.3 Comparaison des deux tests

**Nombre de p-values**

- Wald □ plusieurs p-values (une par coefficient)
- LRT □ une seule p-value par variable, même si cette variable comporte plusieurs coefficients.

> **Note**
>
> Variable : statut fumeur (oui/non)
>
> - Wald : 1 coefficient □ 1 p-value
> - LRT : 1 comparaison modèle complet vs réduit □ 1 p-value
>
> Variable : IMC en 3 catégories (normal, surpoids, obésité) codée en 2 indicatrices car pour un facteur à $k$ modalités, un modèle de Cox crée $k-1$ variables indicatrices (dummy variables). Donc pour 3 catégories □ 2 coefficients $\beta_1$ et $\beta_2$.
>
> - Wald : 2 coefficients □ 2 p-values
> - LRT : 1 comparaison modèle complet vs réduit □ 1 p-value

**Quel test utiliser ?**

Le test du rapport de vraisemblance est plus utilisé pour tester un ensemble de covariables simultanément (modèle multivariable).

À la limité d'une taille d'échantillon très grande, ces deux tests donnent des résultats similaires.

> **!** Important
>
> Le top : rapport de vraisemblance pour chaque variable, rapporté dans un tableau

**Tableau comparatif**

| Caractéristique | Test de Wald | Test du rapport de vraisemblance (LRT) |
|---|---|---|
| **Ce qui est testé** | Chaque coefficient séparément | L'ensemble des coefficients d'une même variable |
| **Nombre de p-values obtenues** | Une p-value par coefficient | Une seule p-value par variable (ou par bloc de variables) |
| **Ex : variable binaire** | 1 coefficient □ 1 p-value | 1 degré de liberté □ 1 p-value |
| **Ex : variable qualitative à 4 niveaux** | 3 coefficients □ 3 p-values | Test global sur les 3 coefficients □ 1 p-value |
| **Interprétation** | Effet individuel de chaque niveau / coefficient | Apport global de la variable au modèle |
| **Type de test** | Normal (approximation asymptotique) | Khi-deux |
| **Fiabilité** | Peut être instable si les SE sont mal estimés | Plus robuste en théorie |
| **Usage typique** | Lecture coefficient par coefficient | Comparaison modèles (réduit vs complet) |

# 7 TP

## 7.A Modèle de Cox avec R

- Charger la librairie `survival`.
  - `Surv` : créer un objet de survie
  - `survfit` : ajuster un modèle de survie (Kaplan Meier)
  - `survdiff` : test de Log-Rank
  - `coxph` : ajuster un modèle de Cox
- Conditions du df : "survival object" faisable
  - Temps de suivi
  - État de l'événement

### 7.A.1 Exemple avec le jeu de données `lung`

#### 7.A.1.1 Charger et afficher les premières lignes du jeu de données `lung`

```
# Charger la librairie survival
library(survival)
# Afficher les premières lignes du jeu de données
head(lung)
```

```
  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
```

#### 7.A.1.2 Créer un objet de type `Surv` avec les colonnes `time` et `status`

```
# Créer un objet Surv
surv_obj <- Surv(time = lung$time, event = lung$status)
surv_obj
```

```
  [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+
```

#### 7.A.1.3 Tester l'effet du sexe (`sex`) sur la survie avec un log-rank test

```r
# Test de Log-Rank pour le sexe
logrank_sex <- survdiff(surv_obj ~ lung$sex)
print(logrank_sex)
```

```
Call:
survdiff(formula = surv_obj ~ 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
```

#### 7.A.1.4 Ajuster un modèle de Cox avec `sex` comme covariable

Fonction `coxph` (pour Cox Proportional Hazards) :

```r
# Ajuster un modèle de Cox avec le sexe comme covariable
cox_model_sex <- coxph(surv_obj ~ lung$sex)
summary(cox_model_sex)
```

```
Call:
coxph(formula = surv_obj ~ lung$sex)

  n= 228, number of events= 165

            coef exp(coef) se(coef)      z Pr(>|z|)
lung$sex -0.5310    0.5880   0.1672 -3.176  0.00149 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

         exp(coef) exp(-coef) lower .95 upper .95
lung$sex     0.588      1.701    0.4237     0.816

Concordance= 0.579  (se = 0.021 )
Likelihood ratio test= 10.63  on 1 df,   p=0.001
Wald test            = 10.09  on 1 df,   p=0.001
Score (logrank) test = 10.33  on 1 df,   p=0.001
```

---

Résultats :

`coef` :

- estimation du coefficient $\hat{\beta}$ pour la variable `sex` (-0.5310) =
- = $h(t|\text{sex}) = h_0(t)\,\exp(\beta \times \text{sex})$

  - avec `sex` = 1 pour les femmes et 0 pour les hommes.
  - coefficient $\beta$ correspond à l'impact du sexe (femme vs homme) sur le logarithme du risque instantané.

- `exp(coef)` : estimation du hazard ratio (HR) pour les femmes par rapport aux hommes (0.5880)
  - exp$(\hat{\beta}) = HR$
  - = exp$(-0.5310) \approx 0.588$
  - Interprétation : à tout instant du suivi, les femmes ont environ 41.2 % (1 - 0.588) de risque instantané en moins de mourir par rapport aux hommes.
- `se(coef)` : écart type de l'estimateur du coefficient (0.1672) =
  - incertitude sur l'estimation de $\hat{\beta}$.
  - Utilisé pour le test de Wald (formule : $Z = \hat{\beta}/SE(\hat{\beta})$).
  - pour obtenir l'IC de l'HR : $IC_{95\%} =$ exp$(\hat{\beta} \pm 1.96 \times SE(\hat{\beta}))$
- `z` : statistique de test de Wald (-3.176) =
  - calculée comme $z = \hat{\beta}/SE(\hat{\beta})$.
  - Utilisée pour obtenir la p-value associée au test de Wald.
  - Valeur absolue élevée de `z` indique un effet significatif.
  - Plus bas, il est écrit 10.09, qui est le carré de -3.176 (car le test du chi carré est le carré du test de Wald).
- `Pr(>|z|)` : p-value associée au test de Wald (0.00149) =
  - probabilité d'observer une statistique de test aussi extrême que `z` sous l'hypothèse nulle ($\beta = 0$).
  - p-value < 0.05 indique que l'effet du sexe sur le risque instantané est statistiquement significatif.
- `exp(-coef)` : inverse du hazard ratio (1.701) =
  - exp$(-\hat{\beta}) = 1/HR$
  - Interprétation : à tout instant du suivi, les hommes ont environ 70.1 % (1.701 - 1) de risque instantané en plus de mourir par rapport aux femmes.
- `lower .95` et `upper .95` : bornes inférieure et supérieure de l'intervalle de confiance à 95 % pour le hazard ratio (0.4237, 0.816) =
  - Calculé comme $IC_{95\%} =$ exp$(\hat{\beta} \pm 1.96 \times SE(\hat{\beta}))$.
  - Indique la précision de l'estimation du HR.
  - Si l'IC ne contient pas 1, l'effet est statistiquement significatif.
- `Concordance` : mesure de la capacité prédictive du modèle (0.579) =
  - Valeur entre 0.5 (aucune capacité prédictive) et 1 (prédiction parfaite).

  - Indique dans quelle mesure le modèle classe correctement les individus en fonction de leur risque.

- `Likelihood ratio test` : statistique du test du rapport de vraisemblance (10.63) avec sa p-value (0.001) =
  - Compare le modèle complet (avec `sex`) au modèle réduit (sans `sex`).
  - p-value < 0.05 indique que l'ajout de `sex` améliore significativement le modèle.
- `Wald test` : statistique du test de Wald (10.09) avec sa p-value (0.001) =
  - Teste l'effet individuel de `sex` sur le risque instantané.
  - p-value < 0.05 indique que l'effet de `sex` est statistiquement significatif.
  - Intérêt par rapport à `z` et `Pr(>|z|)` : c'est la même information mais exprimée en chi carré.
    * Exprimé en $chi^2$ pour faciliter la comparaison avec d'autres tests (comme le LRT).
    * Mais en vrai, c'est juste le carré de `z`.

> **Note**
>
> **Pourquoi le test de Wald global renvoie un chi-2 ?**
> Parce que le carré d'une variable normale centrée réduite suit une loi du chi-2 à 1 degré de liberté.
> C'est un théorème fondamental :
> $(\mathcal{N}(0,1))^2 \sim \chi^2(1)$
> Donc :
> $W = z^2$
> Dans l'output :
> $z = -3.176$
> $ Wald □ ^2 = (–3.176)^2 = 10.09$
> Donc :
>
> - Le test individuel affiche z (test indidividuel = test de Wald pour le coefficient)
> - Le test global affiche z² = chi-2 (= test de Wald global pour la variable)
>
> Même p-value, présentation différente.

> **Note**
>
> **Qu'est-ce que** $N$ dans $N(0,1)$ ?
> $N(0,1)$ signifie loi Normale centrée réduite, c'est-à-dire :
>
> - moyenne = 0
> - variance = 1
> - forme en cloche parfaitement définie
>
> Quand on écrit :

$z \sim \mathcal{N}(0, 1)$
cela veut dire : la statistique $z$ suit une loi normale de moyenne 0 et de variance 1.
Dans ce contexte $N$ désigne la loi Normale : $N(\mu, \sigma^2)$ = loi normale de moyenne $\mu$ et variance $\sigma^2$.
Donc $N$ = normal distribution.
**Pourquoi** $z$ suit une loi normale $N(0, 1)$ ?
Dans un modèle de Cox, l'estimateur du coefficient $\hat{\beta}$ est asymptotiquement normal :
$\hat{\beta} \approx \mathcal{N}(\beta, SE(\hat{\beta})^2)$
Sous l'hypothèse nulle $H_0$ : $\beta = 0$, cela devient :
$\hat{\beta} \approx \mathcal{N}(0, SE^2)$
Si on divise par l'écart-type (on le fait pour standardiser la variable, c'est à dire pour lui donner une moyenne 0 et une variance 1) :
$z = \frac{\hat{\beta}}{SE(\hat{\beta})}$
Alors $z$ suit :
$z \sim \mathcal{N}(0, 1)$
Donc :

- $z$ = coefficient standardisé
- $z$ = suit une loi normale centrée réduite
- C'est pour ça qu'on calcule $Pr(> |z|)$ avec la loi normale

NB : diviser par l'écart-type est une opération classique en statistique pour standardiser une variable et obtenir une variable sans unité, ce qui permet de comparer des variables entre elles.
Ça permet aussi d'obtenir une statistique de test qui suit une loi connue (ici la loi normale centrée réduite car on divise par l'écart-type).

#### 7.A.1.5 Obtenir l'intervalle de confiance à 95 % pour le hazard ratio

**Coefficients** $\beta$ et leur exponentielle (HR)

```
# Calculer les coefficients ß
coefficients(cox_model_sex)
```

```
  lung$sex
-0.5310235
```

```
# Exponentielle (hazard ratio)
exp(coefficients(cox_model_sex))
```

```
 lung$sex
0.5880028
```

**Pour l'intervalle de confiance à 95 % :**

```r
# Obtenir l'intervalle de confiance à 95 % pour les coefficients
confint(cox_model_sex)
```

```
              2.5 %     97.5 %
lung$sex -0.8586875 -0.2033595
```

```r
# Obtenir l'intervalle de confiance à 95 % pour le hazard ratio
exp(confint(cox_model_sex))
```

```
             2.5 %    97.5 %
lung$sex 0.4237178 0.8159848
```

**Risque de base** : fonction de risque de base estimée

```r
# Obtenir la fonction de risque de base estimée
base_hazard <- basehaz(cox_model_sex, centered = FALSE)
head(base_hazard)
```

```
       hazard time
1 0.008907762    5
2 0.035855686   11
3 0.044934060   12
4 0.063237497   13
5 0.072463620   15
6 0.081740068   26
```

# 8 TP 2

Base de données `gehan` (disponible dans la librairie `MASS`) : 42 patients leucémiques avec 2 groupes de traitement 6-MP vs controls avec les durées de rémission

1. Combien y a-t-il d'évènements et de censures dans chacun des groupes ?
2. Combien y-a-t-il de sujets exposés au risque de rechute à 12 et 18 semaines dans chaque groupe ?
3. Un test du $Chi^2$ de Pearson serait-il approprié pour comparer les proportions d'évènements entre les deux groupes ? Justifiez votre réponse.
4. Quelle est la survie sans rechute à 6 et 12 semaines dans les deux groupes de traitement ?
5. Quelle est la médiane de survie sans rechute dans chaque groupe de traitement ?
6. Estimez et tracez les courbes de survie de Kaplan-Meier pour les deux groupes de traitement.
7. Quel test pour comparer les courbes de survie entre les deux groupes de traitement ? Justifiez votre choix.

## 8.A Réponse

Import de la librairie et des données

```
# Charger la librairie survival
library(MASS)
# Afficher les premières lignes des données
head(gehan,20)
```

```
   pair time cens   treat
1     1    1    1 control
2     1   10    1    6-MP
3     2   22    1 control
4     2    7    1    6-MP
5     3    3    1 control
6     3   32    0    6-MP
7     4   12    1 control
8     4   23    1    6-MP
9     5    8    1 control
10    5   22    1    6-MP
11    6   17    1 control
12    6    6    1    6-MP
13    7    2    1 control
14    7   16    1    6-MP
15    8   11    1 control
16    8   34    0    6-MP
17    9    8    1 control
18    9   32    0    6-MP
19   10   12    1 control
20   10   25    0    6-MP
```

Colonnes :

- `pair` : numéro de la paire de patients appariés
- `time` : temps de suivi
- `cens` : indicateur de censure (0 = censure, 1 = événement)
- `treat` : groupe de traitement (control ou 6-MP)

Donc c'est bizarre !

- Moins d'observations censurées (0) que d'observations non censurées (1) !
- Donc c'est probablement l'inverse ! :
  - `cens` = 1 □ censure
  - `cens` = 0 □ événement
- `pair` : numéro de la paire de patients appariés mais on va l'ignorer ici

### 8.A.1 Nombre d'évènements et de censures dans chaque groupe

Fonction `table` pour obtenir un tableau croisé des événements et censures par groupe de traitement.

```
# Résumé des événements et censures par groupe de traitement
table(gehan$treat, gehan$cens, dnn = c("Traitement", "Censure"))
```

```
          Censure
Traitement  0  1
   6-MP    12  9
   control  0 21
```

### 8.A.2 Nombre de sujets exposés au risque de rechute à 12 et 18 semaines dans chaque groupe

Stratégie :

1. Créer un objet `Surv` qui contiendra les temps de suivi et l'état de censure en un seul objet
    - Utiliser la fonction `Surv` du package `survival`
    - Syntaxe : `Surv(data$time, data$event)`

```
surv_object_gehan <- Surv(time = gehan$time, event = gehan$cens)
surv_object_gehan
```

```
 [1]  1  10  22   7   3  32+ 12  23   8  22  17   6   2  16  11  34+  8  32+ 12
[20] 25+  2  11+  5  20+  4  19+ 15   6   8  17+ 23  35+  5   6  11  13   4   9+
[39]  1   6+  8  10+
```

2. Utiliser la fonction `survfit` pour calculer la survie
    - Syntaxe : `survfit(Surv_object ~ group_variable, data = dataset)`

```
surv_fit_gehan <- survfit(surv_object_gehan ~ gehan$treat, data = gehan)
surv_fit_gehan
```

```
Call: survfit(formula = surv_object_gehan ~ gehan$treat, data = gehan)

                     n events median 0.95LCL 0.95UCL
gehan$treat=6-MP    21      9     23      16      NA
gehan$treat=control 21     21      8       4      12
```

3. Extraire le nombre de sujets à risque à 12 et 18 semaines en utilisant la fonction `summary`
    - Syntaxe : `summary(survfit_object, times = c(times_of_interest))`

```r
summary_gehan_12_18 <- summary(surv_fit_gehan, times = c(12, 18))
summary_gehan_12_18
```

```
Call: survfit(formula = surv_object_gehan ~ gehan$treat, data = gehan)

                gehan$treat=6-MP
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   12     12       5    0.753  0.0963        0.586        0.968
   18      9       2    0.627  0.1141        0.439        0.896

                gehan$treat=control
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
   12      6      17   0.1905  0.0857       0.0789        0.460
   18      2       2   0.0952  0.0641       0.0255        0.356
```

### 8.A.3 $Chi^2$ de Pearson approprié ?

Non, le test du $Chi^2$ de Pearson n'est pas approprié pour comparer les proportions d'évènements entre les deux groupes dans ce contexte de données de survie.

Justification :

- Le test du $Chi^2$ de Pearson compare les proportions d'événements entre des groupes à un moment fixe, sans tenir compte du temps de suivi ni de la censure.
- Dans les données de survie, les temps de suivi varient entre les individus, et certains individus peuvent être censurés (c'est-à-dire que l'événement d'intérêt n'a pas été observé avant la fin du suivi).
- Le test du $Chi^2$ de Pearson ne prend pas en compte ces aspects temporels et de censure, ce qui peut biaiser les résultats.
- Pour comparer les courbes de survie entre les groupes, il est plus approprié d'utiliser le test de Log-Rank, qui tient compte des temps de suivi et de la censure.

### 8.A.4 Survie sans rechute à 6 et 12 semaines dans les deux groupes de traitement

Utiliser la fonction `summary` pour extraire les estimations de survie à 6 et 12 semaines.

```r
summary_gehan_6_12 <- summary(surv_fit_gehan, times = c(6, 12))
summary_gehan_6_12
```

```
Call: survfit(formula = surv_object_gehan ~ gehan$treat, data = gehan)

                gehan$treat=6-MP
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6     21       3    0.857  0.0764        0.720        1.000
   12     12       2    0.753  0.0963        0.586        0.968
```

```
                gehan$treat=control
 time n.risk n.event survival std.err lower 95% CI upper 95% CI
    6     12       9    0.571  0.1080       0.3945        0.828
   12      6       8    0.190  0.0857       0.0789        0.460
```

Résultats :

- À 6 semaines :
  - Groupe 6-MP : survie sans rechute = 85.7 % (IC 95 % : 72.0 % - 100 %)
  - Groupe contrôle : survie sans rechute = 57.1 % (IC 95 % : 39.5 % - 82.8 %)
- À 12 semaines :
  - Groupe 6-MP : survie sans rechute = 75.3 % (IC 95 % : 58.6 % - 96.8 %)
  - Groupe contrôle : survie sans rechute = 19.0 % (IC 95 % : 7.9 % - 46.0 %)

### 8.A.5 Médiane de survie sans rechute dans chaque groupe de traitement

La médiane est contenue dans l'objet crée avec `Surv` : `surv_fit_gehan`.

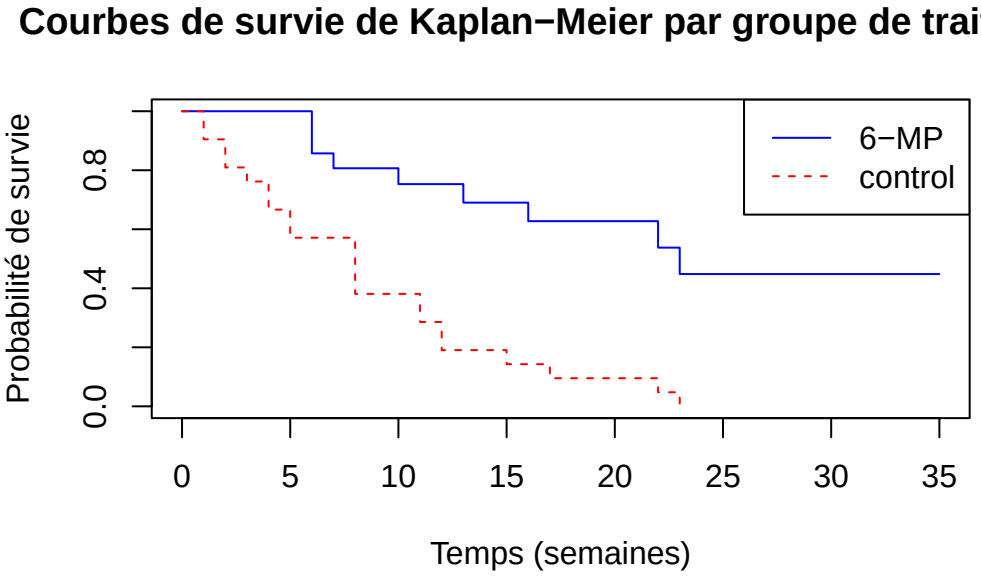
```r
# Obtenir la médiane de survie pour chaque groupe
surv_fit_gehan
```

```
Call: survfit(formula = surv_object_gehan ~ gehan$treat, data = gehan)

                     n events median 0.95LCL 0.95UCL
gehan$treat=6-MP    21      9     23      16      NA
gehan$treat=control 21     21      8       4      12
```

### 8.A.6 Estimer et tracer les courbes de survie de Kaplan-Meier pour les deux groupes de traitement

Tracer les courbes de survie de Kaplan-Meier pour les deux groupes de traitement

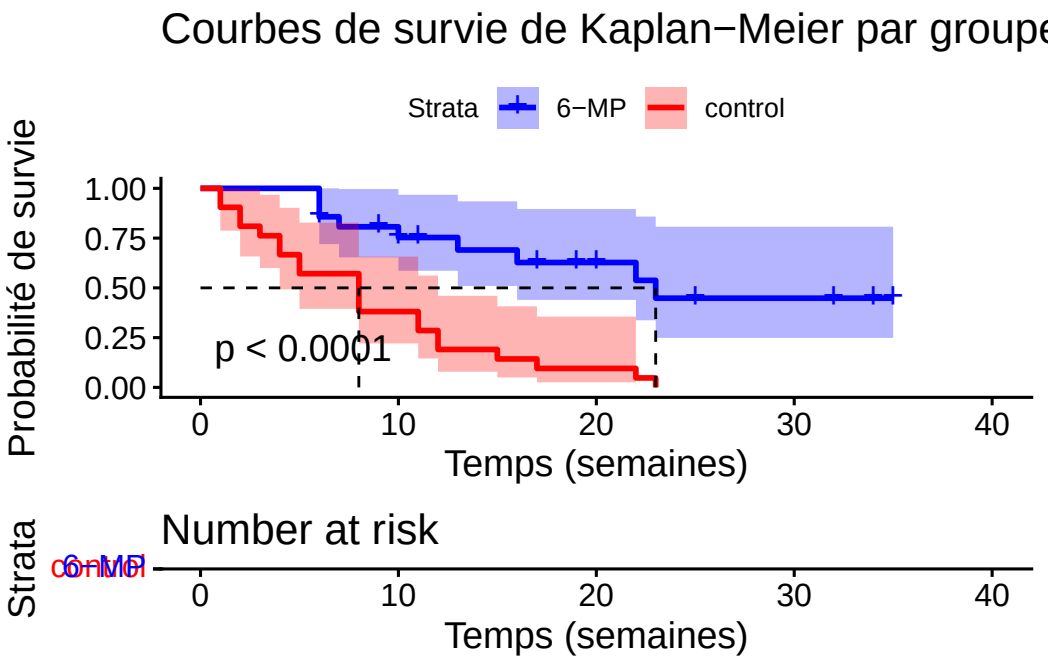
```r
# Tracer les courbes de survie
plot(surv_fit_gehan, col = c("blue", "red"), lty = 1:2,
       xlab = "Temps (semaines)", ylab = "Probabilité de survie",
       main = "Courbes de survie de Kaplan-Meier par groupe de traitement")
legend("topright", legend = levels(gehan$treat), col = c("blue", "red"), lty = 1:2)
```




Utiliser `ggsurvplot` pour une visualisation améliorée

```
# Charger la librairie survminer pour une meilleure visualisation
library(survminer)
# Tracer les courbes de survie avec ggsurvplot
ggsurvplot(surv_fit_gehan, data = gehan,
           pval = TRUE, conf.int = TRUE,
           risk.table = TRUE,
           xlab = "Temps (semaines)",
           ylab = "Probabilité de survie",
           title = "Courbes de survie de Kaplan-Meier par groupe de traitement",
           legend.labs = levels(gehan$treat),
           palette = c("blue", "red"),
           surv.median.line = "hv")
```

```
Ignoring unknown labels:
* colour : "Strata"
```




**8.A.7 Quel test pour comparer les courbes de survie entre les deux groupes de traitement ? Justifiez votre choix.**

Le test approprié pour comparer les courbes de survie entre les deux groupes de traitement est le **test de Log-Rank**.

Justification :

- Le test de Log-Rank est spécifiquement conçu pour comparer les courbes de survie entre deux ou plusieurs groupes dans le contexte des données de survie.
- Il prend en compte les temps de suivi variables et la censure des données, ce qui est essentiel dans les analyses de survie.
- Le test de Log-Rank évalue si les différences observées dans les courbes de survie sont statistiquement significatives en comparant le nombre d'événements observés à celui attendu dans chaque groupe à chaque instant de temps.

```
# Effectuer le test de Log-Rank
logrank_test <- survdiff(surv_object_gehan ~ gehan$treat)
print(logrank_test)
```

```
Call:
survdiff(formula = surv_object_gehan ~ gehan$treat)

                     N Observed Expected (O-E)^2/E (O-E)^2/V
gehan$treat=6-MP    21        9     19.3      5.46      16.8
gehan$treat=control 21       21     10.7      9.77      16.8

 Chisq= 16.8  on 1 degrees of freedom, p= 4e-05
```

# 9 TP 3

Base de données `Melanoma` dans le package `MASS` : étude de survie de patients atteints de mélanome avec différentes variables explicatives.

## 9.A Questions

1. Description des données
2. Estimation d'une courbe de survie de Kaplan-Meier globale
3. Influence du sexe sur la survie
4. Test de Log-Rank pour tester la différence de survie en foncton du sexe, en stratifiant sur le niveau d'ulcération

## 9.B Réponse

### 9.B.1 Description des données

Importer les données et la librairie nécessaire

```
# Charger la librairie MASS
library(MASS)
# Afficher les premières lignes des données Melanoma
head(Melanoma)
```

```
  time status sex age year thickness ulcer
1   10      3   1  76 1972      6.76     1
2   30      3   1  56 1968      0.65     0
3   35      2   1  41 1977      1.34     0
4   99      3   0  71 1968      2.90     0
5  185      1   1  52 1965     12.08     1
6  204      1   1  28 1971      4.84     1
```

Colonnes :

- `time` : temps de suivi en mois
- `status` : état de l'événement (1 = décès dû au mélanome, 2 = alive, 3 = décès dû à d'autres causes)
- `sex` : sexe (0 = femme, 1 = homme)
- `age` : âge au moment du diagnostic
- `year` : année du diagnostic
- `thickness` : épaisseur de la tumeur en mm
- `ulcer` : présence d'ulcération (0 = non, 1 = oui)

Truc : pour l'analyse de survie, on va recoder `status` en binaire (1 et 3 = décès toute cause, 2 = censuré)

```r
Melanoma$status2 <- ifelse(Melanoma$status == 2, 0, 1)
```

### 9.B.2 Estimation d'une courbe de survie de Kaplan-Meier globale

1. Créer un objet `Surv` avec les temps de suivi et l'état de l'événement

```r
Surv(time = data$time, event = data$status == 1)
```

```r
surv_object_melanoma <- Surv(time = Melanoma$time, event = Melanoma$status2 == 1)
surv_object_melanoma
```

```
  [1]   10    30    35+   99   185   204   210   232   232   279   295   355
 [13]  386   426   469   493   529   621   629   659   667   718   752   779
 [25]  793   817   826   833   858   869   872   967   977   982  1041  1055
 [37] 1062  1075  1156  1228  1252  1271  1312  1427  1435  1499+ 1506  1508+
 [49] 1510+ 1512+ 1516  1525  1542+ 1548  1557+ 1560  1563+ 1584  1605+ 1621
 [61] 1627+ 1634+ 1641+ 1641+ 1648+ 1652+ 1654+ 1654+ 1667  1678+ 1685+ 1690
 [73] 1710+ 1710+ 1726  1745+ 1762+ 1779+ 1787+ 1787+ 1793+ 1804+ 1812+ 1836+
 [85] 1839+ 1839+ 1854+ 1856+ 1860  1864+ 1899+ 1914+ 1919+ 1920+ 1927+ 1933
 [97] 1942+ 1955+ 1956+ 1958+ 1963+ 1970+ 2005+ 2007+ 2011+ 2024+ 2028+ 2038+
[109] 2056+ 2059+ 2061  2062  2075+ 2085  2102+ 2103  2104+ 2108  2112+ 2150+
[121] 2156+ 2165+ 2209+ 2227+ 2227+ 2256  2264+ 2339+ 2361+ 2387+ 2388  2403+
[133] 2426+ 2426+ 2431+ 2460+ 2467  2492+ 2493+ 2521+ 2542+ 2559+ 2565  2570+
[145] 2660+ 2666+ 2676+ 2738+ 2782  2787+ 2984+ 3032+ 3040+ 3042  3067+ 3079+
[157] 3101+ 3144+ 3152+ 3154  3180+ 3182  3185+ 3199+ 3228+ 3229+ 3278+ 3297+
[169] 3328+ 3330+ 3338  3383+ 3384+ 3385+ 3388+ 3402+ 3441+ 3458  3459+ 3459+
[181] 3476+ 3523+ 3667+ 3695+ 3695+ 3776+ 3776+ 3830+ 3856+ 3872+ 3909+ 3968+
[193] 4001+ 4103+ 4119+ 4124+ 4207+ 4310+ 4390+ 4479+ 4492+ 4668+ 4688+ 4926+
[205] 5565+
```

2. Calculer la courbe de survie de Kaplan-Meier globale avec `survfit`

`survfit`permet d'ajuster un modèle de survie de Kaplan-Meier.

```r
surv_fit_melanoma <- survfit(surv_object_melanoma ~ 1, data = Melanoma)
surv_fit_melanoma
```

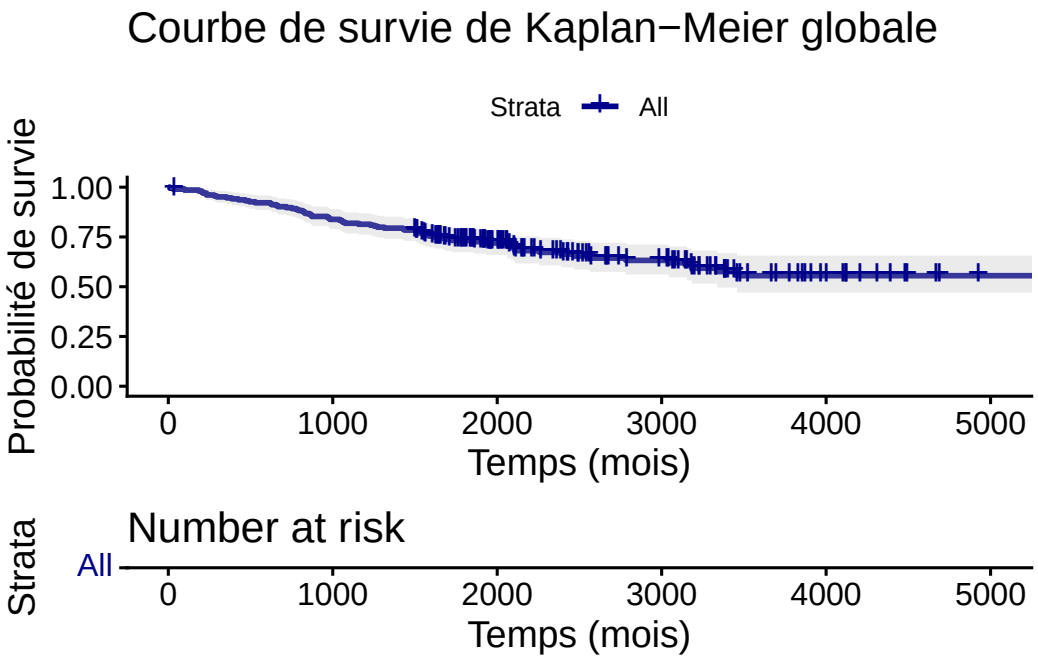
```
Call: survfit(formula = surv_object_melanoma ~ 1, data = Melanoma)

       n events median 0.95LCL 0.95UCL
[1,] 205     71     NA    3338      NA
```

3. Tracer la courbe de survie avec 'ggsurvplot' pour une meilleure visualisation

```r
# Charger la librairie survminer pour une meilleure visualisation
library(survminer)
# Tracer la courbe de survie globale
```

```
ggsurvplot(surv_fit_melanoma,
           data = Melanoma,
           conf.int = TRUE,
           risk.table = TRUE,
           xlab = "Temps (mois)",
           ylab = "Probabilité de survie",
           title = "Courbe de survie de Kaplan-Meier globale",
           palette = "darkblue"
       )
```

```
Ignoring unknown labels:
* fill : "Strata"
Ignoring unknown labels:
* fill : "Strata"
Ignoring unknown labels:
* fill : "Strata"
Ignoring unknown labels:
* fill : "Strata"
Ignoring unknown labels:
* colour : "Strata"
```




### 9.B.3 Influence du sexe sur la survie

1. Calculer les courbes de survie de Kaplan-Meier par sexe

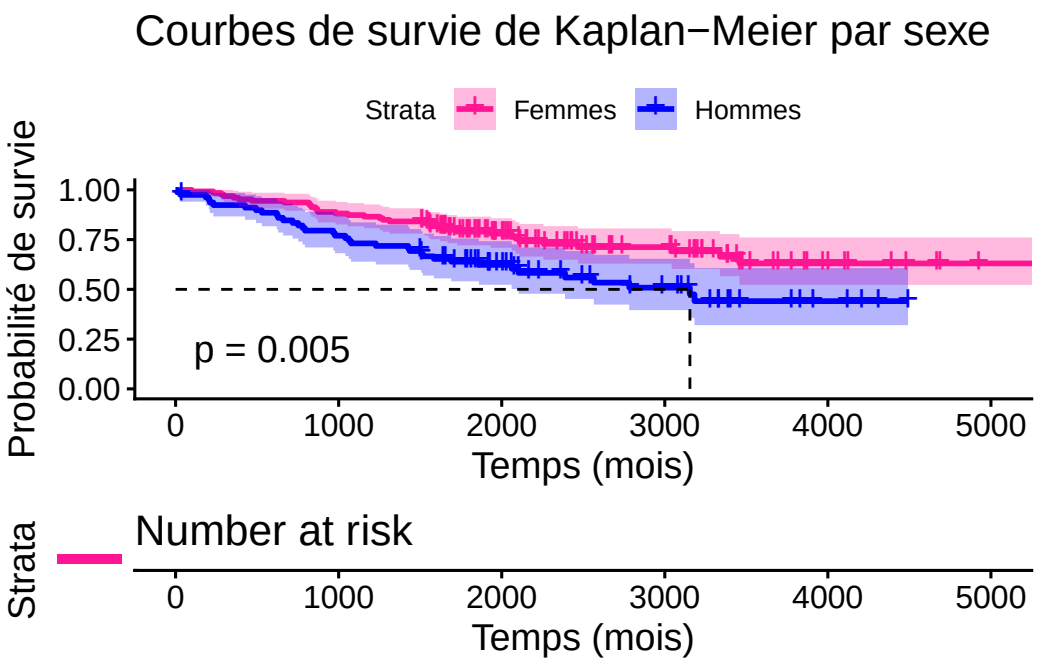
```
surv_fit_sex <- survfit(surv_object_melanoma ~ Melanoma$sex, data = Melanoma)
surv_fit_sex
```

```
Call: survfit(formula = surv_object_melanoma ~ Melanoma$sex, data = Melanoma)

                 n events median 0.95LCL 0.95UCL
Melanoma$sex=0 126     35     NA      NA      NA
Melanoma$sex=1  79     36   3154    2103      NA
```

2. Tracer les courbes de survie par sexe

```
# Tracer les courbes de survie par sexe
ggsurvplot(
        surv_fit_sex,
        data = Melanoma,
        pval = TRUE,
        conf.int = TRUE,
        risk.table = TRUE,
        xlab = "Temps (mois)",
        ylab = "Probabilité de survie",
        title = "Courbes de survie de Kaplan-Meier par sexe",
        legend.labs = c("Femmes", "Hommes"),
        palette = c("deeppink", "blue"),
        surv.median.line = "hv",
        risk.table.height = 0.25,
        risk.table.y.text.col = TRUE,
        risk.table.y.text = FALSE,
        censor = TRUE
        )
```

```
Ignoring unknown labels:
* colour : "Strata"
```



### 9.B.4 Test de Log-Rank et Modèle de Cox pour comparer les courbes de survie entre les groupes

1. Effectuer le test de Log-Rank pour comparer les courbes de survie entre les groupes

```
# Effectuer le test de Log-Rank
logrank_test_sex <- survdiff(surv_object_melanoma ~ Melanoma$sex)
print(logrank_test_sex)
```

```
Call:
survdiff(formula = surv_object_melanoma ~ Melanoma$sex)

                 N Observed Expected (O-E)^2/E (O-E)^2/V
Melanoma$sex=0 126       35     46.3      2.75       7.9
Melanoma$sex=1  79       36     24.7      5.14       7.9

 Chisq= 7.9  on 1 degrees of freedom, p= 0.005
```

2. Ajuster un modèle de Cox avec le sexe comme covariable

```
# Ajuster un modèle de Cox avec le sexe comme covariable
cox_model_sex <- coxph(surv_object_melanoma ~ Melanoma$sex)
summary(cox_model_sex)
```

```
Call:
coxph(formula = surv_object_melanoma ~ Melanoma$sex)

  n= 205, number of events= 71

               coef exp(coef) se(coef)     z Pr(>|z|)
Melanoma$sex 0.6559    1.9269   0.2376 2.761  0.00577 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

             exp(coef) exp(-coef) lower .95 upper .95
Melanoma$sex     1.927      0.519      1.21      3.07

Concordance= 0.585  (se = 0.031 )
Likelihood ratio test= 7.51  on 1 df,   p=0.006
Wald test            = 7.62  on 1 df,   p=0.006
Score (logrank) test = 7.9  on 1 df,   p=0.005
```

### 9.B.5 Si on regarde en ajustant sur l'épaisseur de la tumeur

```r
# Ajuster un modèle de Cox avec le sexe et l'épaisseur de la tumeur comme covariables
cox_model_sex_thickness <- coxph(surv_object_melanoma ~ Melanoma$sex + Melanoma$thickness)
summary(cox_model_sex_thickness)
```

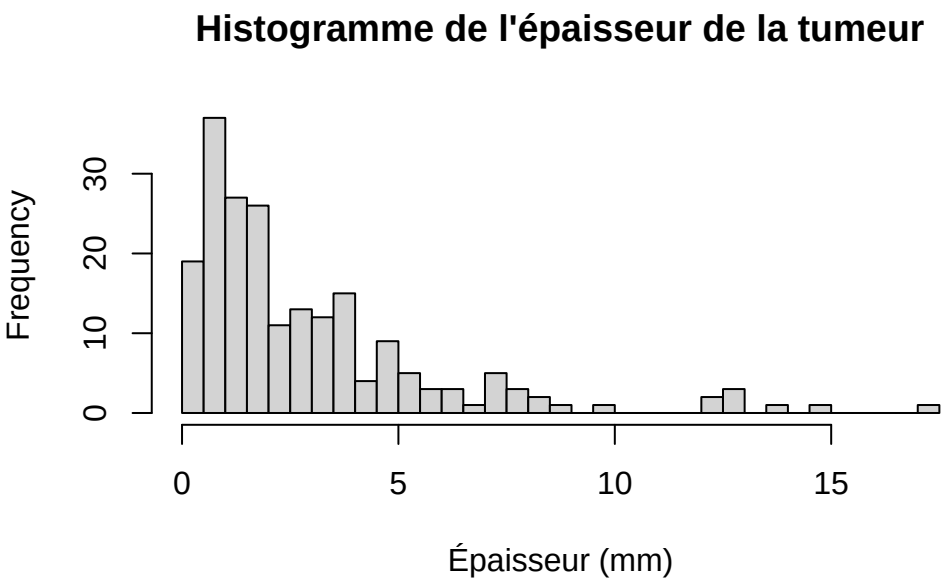
```
Call:
coxph(formula = surv_object_melanoma ~ Melanoma$sex + Melanoma$thickness)

  n= 205, number of events= 71

                      coef exp(coef) se(coef)     z Pr(>|z|)
Melanoma$sex       0.56284   1.75565  0.23787 2.366    0.018 *
Melanoma$thickness 0.14883   1.16048  0.03011 4.942 7.72e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                   exp(coef) exp(-coef) lower .95 upper .95
Melanoma$sex           1.756     0.5696     1.101     2.798
Melanoma$thickness     1.160     0.8617     1.094     1.231

Concordance= 0.696  (se = 0.034 )
Likelihood ratio test= 26.11  on 2 df,   p=2e-06
Wald test            = 30.98  on 2 df,   p=2e-07
Score (logrank) test = 34.19  on 2 df,   p=4e-08
```

Résultats :

- Le coefficient pour `Melanoma$sex` est 0.56284, ce qui signifie que les hommes ont un risque instantané de décès dû au mélanome environ 75.6 % plus élevé que les femmes, après ajustement pour l'épaisseur de la tumeur.
  - HR = exp(0.56284) = 1.75565
- Le coefficient pour `Melanoma$thickness` est 0.14883, indiquant que pour chaque augmentation de 1 mm de l'épaisseur de la tumeur, le risque instantané de décès dû au mélanome augmente d'environ 16.0 %.
  - HR = exp(0.14883) = 1.16048 en plus par mm (variable quantitative donc HR par unité)

Mais le truc c'est que l'épaisseur de la tumeur est très variable !!

```r
hist(Melanoma$thickness, breaks = 30, main = "Histogramme de l'épaisseur de la tumeur", xlab =
```




Mais pour les dermatos : ce qui compte c'est si l'épaisseur est > ou < 1 mm (critère pronostique important)

Donc on peut créer une variable binaire `thickness_bin` :

```
Melanoma$thickness_bin <- ifelse(Melanoma$thickness > 1, 1, 0)
head(Melanoma, 3)
```

```
  time status sex age year thickness ulcer status2 thickness_bin
1   10      3   1  76 1972      6.76     1       1             1
2   30      3   1  56 1968      0.65     0       1             0
3   35      2   1  41 1977      1.34     0       0             1
```

```
table(Melanoma$thickness_bin)
```

```
  0   1
 56 149
```

Puis ajuster le modèle de Cox avec cette variable binaire :

```
# Ajuster un modèle de Cox avec le sexe et l'épaisseur binaire de la tumeur comme covariables
cox_model_sex_thickness_bin <- coxph(surv_object_melanoma ~ Melanoma$sex + Melanoma$thickness_b
summary(cox_model_sex_thickness_bin)
```

```
Call:
coxph(formula = surv_object_melanoma ~ Melanoma$sex + Melanoma$thickness_bin)

  n= 205, number of events= 71

                       coef exp(coef) se(coef)     z Pr(>|z|)
Melanoma$sex          0.6813    1.9764   0.2379 2.864  0.00419 **
```

```
Melanoma$thickness_bin 1.1275    3.0879   0.3571 3.157  0.00159 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

                       exp(coef) exp(-coef) lower .95 upper .95
Melanoma$sex               1.976     0.5060     1.240     3.151
Melanoma$thickness_bin     3.088     0.3238     1.533     6.218

Concordance= 0.651  (se = 0.031 )
Likelihood ratio test= 20.51  on 2 df,   p=4e-05
Wald test            = 17.75  on 2 df,   p=1e-04
Score (logrank) test = 19.02  on 2 df,   p=7e-05
```

---

Résultats :

- Le coefficient pour `Melanoma$sex` est 0.6813 et HR = exp(0.6813) = 1.9764 donc risque instantané de décès dû au mélanome environ 97.6 % plus élevé pour les hommes par rapport aux femmes.
- Le coefficient pour `Melanoma$thickness_bin` est 1.1275 et HR = exp(1.1275) = 3.0879 donc risque instantané de décès dû au mélanome environ 208.8 % plus élevé pour les patients avec une épaisseur de tumeur > 1 mm par rapport à ceux avec une épaisseur ≤ 1 mm.

---

Pour savoir quel est le meilleur modèle entre `thickness` en quantitatif et binaire, on va comparer par rapport au modèle avec `sex` seul

1. On redéfinit les modèles

```r
surv_object_melanoma <- Surv(time = Melanoma$time, event = Melanoma$status2 == 1)
cox_model_sex <- coxph(surv_object_melanoma ~ Melanoma$sex)
cox_model_sex_thickness <- coxph(surv_object_melanoma ~ Melanoma$sex + Melanoma$thickness)
cox_model_sex_thickness_bin <- coxph(surv_object_melanoma ~ Melanoma$sex + Melanoma$thickness_b
```

2. Comparaison avec le test du rapport de vraisemblance

```r
# Comparaison des modèles avec le test du rapport de vraisemblance
anova(cox_model_sex, cox_model_sex_thickness)
```

```
Analysis of Deviance Table
 Cox model: response is  surv_object_melanoma
 Model 1: ~ Melanoma$sex
 Model 2: ~ Melanoma$sex + Melanoma$thickness
   loglik  Chisq Df Pr(>|Chi|)
1 -346.73
2 -337.43 18.598  1  1.614e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(cox_model_sex, cox_model_sex_thickness_bin)
```

```
Analysis of Deviance Table
 Cox model: response is  surv_object_melanoma
 Model 1: ~ Melanoma$sex
 Model 2: ~ Melanoma$sex + Melanoma$thickness_bin
   loglik  Chisq Df Pr(>|Chi|)
1 -346.73
2 -340.23 12.997  1   0.000312 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

On a l'impression qu'en mettant la variable en QUANTITATIF , c'est mieux !

- $Chi^2$ plus grand (18.598 > 12.997)
- p-value plus petite

Mais en vrai : ce sont des considérations uniquement statistiques !

Si la variable qui compte c'est si l'épaisseur est > ou < 1 mm, on prendra le modèle avec la variable binaire

### 9.B.6 Interaction entre sexe et ulcération

#### Modèle de Cox avec interaction entre sexe et ulcération

```r
# Ajuster un modèle de Cox avec interaction entre sexe et ulcération
surv_cox_interaction <- Surv(Melanoma$time, Melanoma$status2)
cox_model_interaction <- coxph(surv_cox_interaction ~ Melanoma$sex + strata(Melanoma$ulcer))
summary(cox_model_interaction)
```

```
Call:
coxph(formula = surv_cox_interaction ~ Melanoma$sex + strata(Melanoma$ulcer))

  n= 205, number of events= 71

               coef exp(coef) se(coef)     z Pr(>|z|)
Melanoma$sex 0.4995    1.6479   0.2396 2.084   0.0371 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

             exp(coef) exp(-coef) lower .95 upper .95
Melanoma$sex     1.648     0.6068      1.03     2.636

Concordance= 0.568  (se = 0.031 )
Likelihood ratio test= 4.31  on 1 df,   p=0.04
Wald test            = 4.34  on 1 df,   p=0.04
Score (logrank) test = 4.43  on 1 df,   p=0.04
```

### Log rank et courbe de Kaplan-Meier stratifiée sur le niveau d'ulcération

1. Effectuer le test de Log-Rank stratifié sur le niveau d'ulcération

Syntaxe : `survdiff(Surv_object ~ group_variable + strata(stratification_variable))`

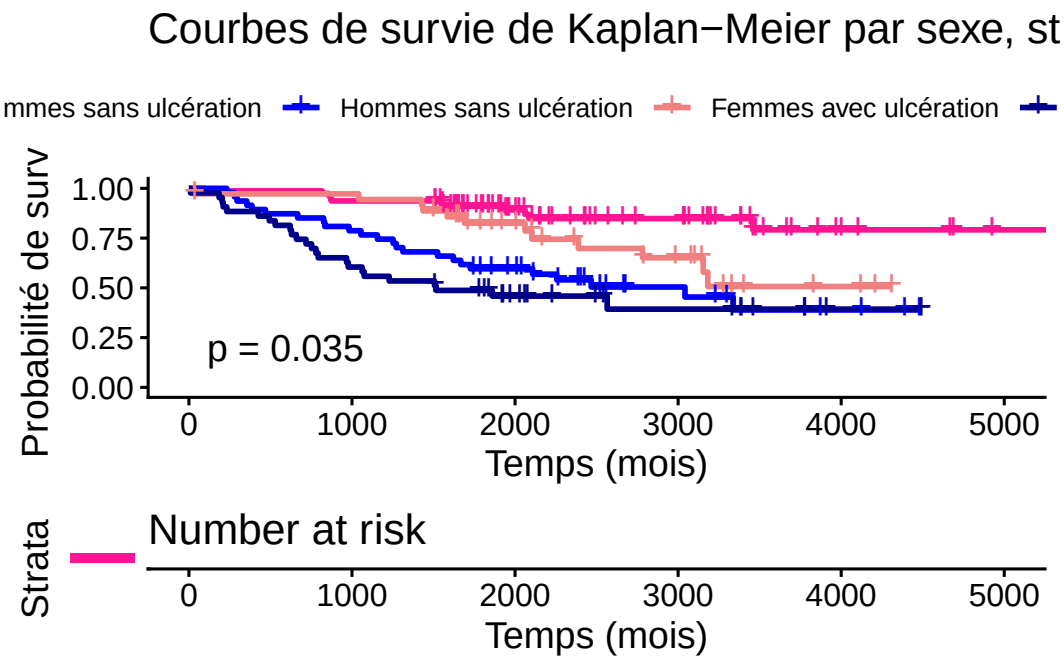
```
# Effectuer le test de Log-Rank stratifié sur le niveau d'ulcération
logrank_test_stratified <- survdiff(surv_object_melanoma ~ Melanoma$sex + strata(Melanoma$ulcer
print(logrank_test_stratified)
```

```
Call:
survdiff(formula = surv_object_melanoma ~ Melanoma$sex + strata(Melanoma$ulcer))

                 N Observed Expected (O-E)^2/E (O-E)^2/V
Melanoma$sex=0 126       35     43.6      1.68      4.43
Melanoma$sex=1  79       36     27.4      2.67      4.43

 Chisq= 4.4  on 1 degrees of freedom, p= 0.04
```

2. Tracer les courbes de survie de Kaplan-Meier par sexe, stratifiées sur le niveau d'ulcération

```
# Tracer les courbes de survie par sexe, stratifiées sur le niveau d'ulcération
ggsurvplot(
        survfit(surv_object_melanoma ~ Melanoma$sex + strata(Melanoma$ulcer), data = Melanoma),
        data = Melanoma,
        pval = TRUE,
        conf.int = FALSE,
        risk.table = TRUE,
        xlab = "Temps (mois)",
        ylab = "Probabilité de survie",
        title = "Courbes de survie de Kaplan-Meier par sexe, stratifiées sur le niveau d'ulcéra
        legend.labs = c("Femmes sans ulcération", "Hommes sans ulcération", "Femmes avec ulcéra
        palette = c("deeppink", "blue", "lightcoral", "darkblue"),
        risk.table.y.text.col = TRUE,
        risk.table.y.text = FALSE,
        censor = TRUE
        )
```

```
Ignoring unknown labels:
* colour : "Strata"
```


Courbes de survie de Kaplan−Meier par sexe, st
mmes sans ulcération
Hommes sans ulcération
Femmes avec ulcération
Probabilité de surv
1.00
0.75
0.50
0.25
0.00
p = 0.035
0
1000
2000
3000
4000
5000
Temps (mois)
Strata
Number at risk
0
1000
2000
3000
4000
5000
Temps (mois)