Université de Poitiers

UFR Sciences fondamentales et appliquées
Département de Mathématiques
2024-2025

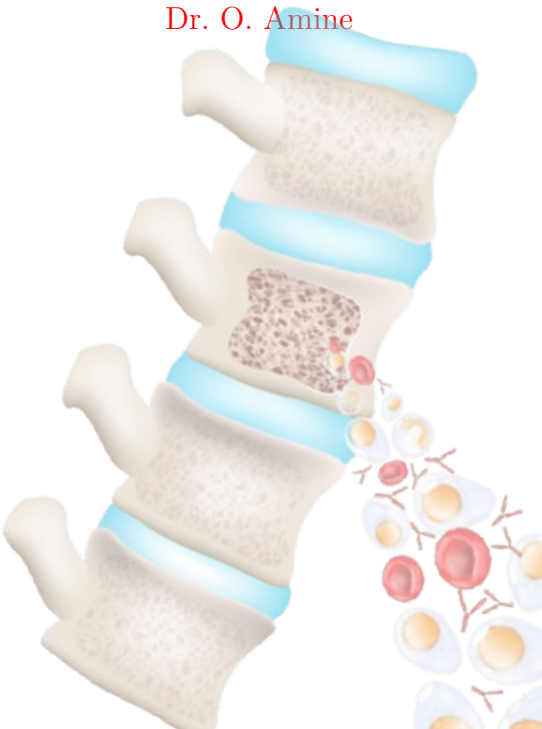
Compte Rendu

Analyse de Survie sur des Patients Atteints de Myélome

Kaplan-Meier et Nelson-Aalen

Élève :
Soufiane Lmezouari

Enseignant :
Dr. S.Yousri
Dr. O. Amine



9 mars 2025

Table des matières

Table des figures

Liste des tableaux

CHAPITRE 1

TP 1

Dans le cadre de ce travail pratique, nous nous intéressons à l'étude des **durées de survie** des patients atteints de myélome multiple, une maladie grave caractérisée par une prolifération maligne des plasmocytes dans la moelle osseuse. Ce type de cancer engendre une série de complications cliniques, notamment des atteintes rénales, sanguines, nerveuses, ainsi que des perturbations biologiques détectables à travers divers marqueurs, tels que l'indice de Bence-Jones.

L'objectif principal de ce TP est de comprendre les facteurs influençant la survie des patients atteints de myélome. À cette fin, nous analyserons une base de données regroupant des informations cliniques et biologiques, telles que l'âge, le sexe, le taux de calcium, et la présence ou non de protéinurie de Bence-Jones.

1.1 Importation des données

Pour répondre aux différentes questions, nous commencerons par récupérerle fichier nommé myel_comp.txt, contenant notre jeu de données

```
library(dplyr)
data <- read.table("myel_comp.txt", header =TRUE)
# Selectionner uniquement les colonnes AGE, SEXE, BENCE_J, CALCIUM
dat <- data %>% select(AGE, SEXE, BENCE_J, CALCIUM)
head(dat)
```

```
  AGE SEXE BENCE_J CALCIUM
> 1  67    1       0      10
> 2  38    1       1      18
> 3  81    1       1       8
> 4  75    1       0      12
> 5  57    1       1       9
> 6  46    0       0      10
```

La base de données a été importée et les variables d'intérêt ont été sélectionnées : `Age`, `Sexe`, `calcium`, et `indice de Bence-Jones`.

1.2 Représentation des données :

1.2.1 Statistiques descriptives

À l'aide de la commande **summary** on obtient :

```
> summary(dat)
```

Statistiques	AGE	SEXE	BENCE_J	CALCIUM
Min.	38.00	0.0000	0.0000	7.00
1st Qu.	51.00	0.0000	0.0000	9.00
Median	60.00	1.0000	0.0000	10.00
Mean	60.15	0.5846	0.3538	10.12
3rd Qu.	67.00	1.0000	1.0000	10.00
Max.	82.00	1.0000	1.0000	18.00

TABLE 1.1 – Résumé statistique des variables

⋉ pour la variable sexe La médiane est de 1, ce qui suggère que plus de la moitié des individus sont de sexe masculin, et La moyenne est de 0,5846, indiquant une légère prédominance masculine dans l'échantillon.

⋉ pour la variable BENCE_JLa médiane est de 0, ce qui signifie que plus de la moitié des individus n'ont pas la caractéristique mesurée par BENCE_J. et La moyenne est de 0,3538, indiquant que cette caractéristique est présente chez environ 35% des individus.

⋉ pour la variable CALCIUMLe niveau de calcium varie entre 7 et 18. et La médiane et la moyenne sont proches (10,00 et 10,12 respectivement), ce qui suggère une distribution relativement symétrique des niveaux de calcium.

1.2.2 Représenatation graphique :

1.2.2.1 Représentation de l'age en fonction de l'indice de sexe

Ce boxplot illustre la répartition des âges en fonction du Sexe (Femme 0 ou Homme 1)

```
> ggplot(dat, aes(x = SEXE, y = AGE, fill = SEXE)) +
> geom_boxplot() + scale_fill_manual(values = colors_sexe) + labs(x =
   "Sexe", y = "Age") + theme_minimal()
```

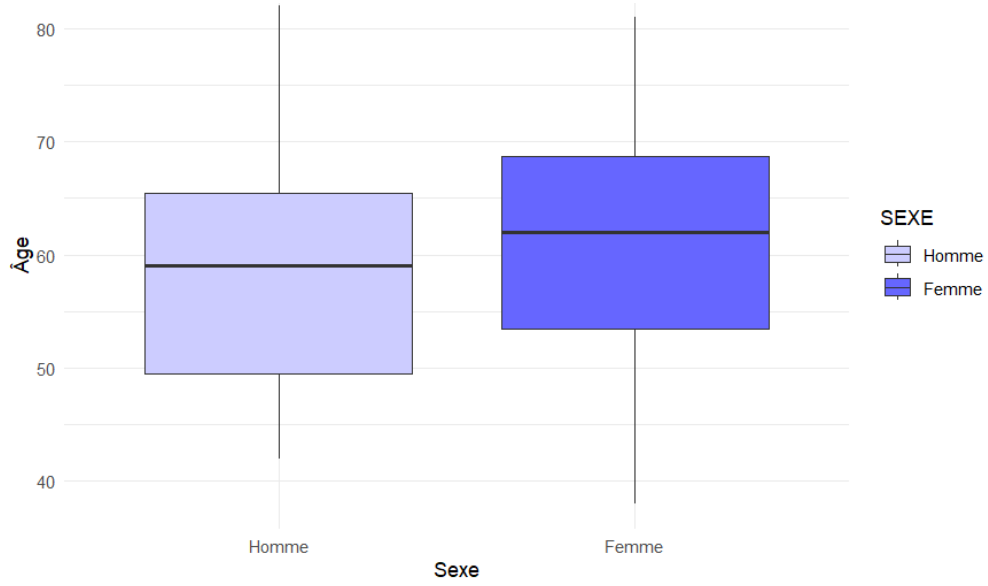


FIGURE 1.1 – Boxplot de l'âge en fonction du sexe

l'âge moyen des hommes est inférieur à celui des femmes, avec une différence d'environ 3 ans entre les deux sexes.

1.2.2.2 Représentation de l'age en fonction de l'indice de Bence Jones

```
ggplot(dat, aes(x =BENCE_J, y =AGE, fill =BENCE_J)) +
geom_boxplot() + scale_fill_manual(values =colors_bence_j) + labs(x ="BENCE_J", y
    ="Age") + theme_minimal()
```

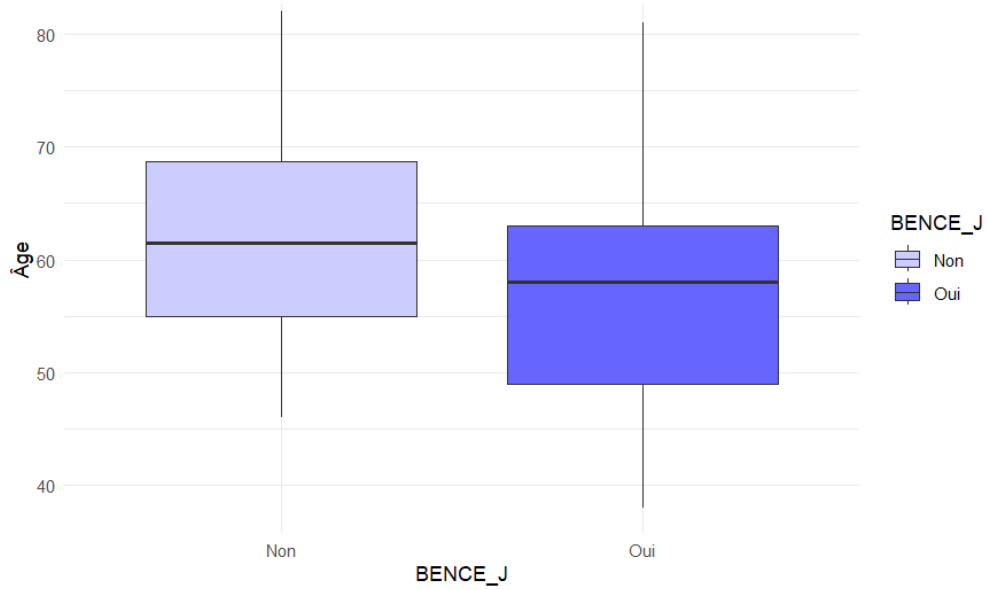


FIGURE 1.2 – Boxplot de l'âge par BENCE_J

Les personnes sans indice de Bence Jones ont un âge moyen plus élevé que celles chez qui cette protéine est détectée.

L'écart de 5 ans entre les moyennes suggère une possible corrélation entre l'âge et la présence de cette protéine, les individus plus âgés étant davantage susceptibles de l'excréter dans leur urine.

1.2.3 Comparer la moyenne d'âge entre hommes et femmes

Calcul de la moyenne d'âge des hommes et des femmes

```
> homme_age_moyen <- mean(dat$AGE[dat$SEXE == 1], na.rm = TRUE)
> femme_age_moyen <- mean(dat$AGE[dat$SEXE == 0], na.rm = TRUE)
```

Sexe	Moyenne d'âge
Hommes	61.47368
Femmes	58.29630

TABLE 1.2 – Moyenne d'âge des hommes et des femmes

Les résultats montrent que Les femmes ont une moyenne d'âge inférieure de 3,17 ans par rapport aux hommes. ce que le boxplot suggère : les femmes ont tendance à être légèrement plus jeunes que les hommes, surtout lorsqu'on observe les distributions par catégories.

1.2.4 L'âge moyen et l'indice de Bence-Jones ?

On commence tout d'abord par tester la normalité.

```
shapiro_test <- shapiro.test(dat$AGE)
shapiro_test
```

```
# W = 0.98132, p-value = 0.4318
```

▷ p-value > 0.05 : Les données suivent une distribution normale.

Alors on peut porsuivre avec une test de student :

```
t.test(AGE ~ BENCE_J, data =dat, var.equal =TRUE)
```

```
# Résultat du test t de deux échantillons
# data:  AGE by BENCE_J
# t = 1.8014, df = 63, p-value = 0.07642
```

La p-value de 0.07642 est trop élevée pour rejeter l'hypothèse nulle. Par conséquent, on ne peut pas conclure que les âges moyens diffèrent de manière significative entre les groupes "sans $BENCE\_J$" et "avec $BENCE\_J$"

1.2.5 Le lien entre les variables Sexe et Bence-Jones

Test de normalité :

```
shapiro.test(data_sexe_0$BENCE_J) # Sexe =0
shapiro.test(data_sexe_1$BENCE_J) # Sexe =1
```

```
# W = 0.57562, p-value = 1.084e-07
# W = 0.62109, p-value = 1.1e-08
```

Les résultats des tests de Shapiro-Wilk indiquent que les données des deux groupes (BENCE_J pour les groupes 0 et 1) ne suivent pas une distribution normale

Alors on va appliquer le test de Mann-Whitney pour comparer BENCE_J entre les sexes :

```
wilcox.test(BENCE_J ~ SEXE, data =dat)
```

```
# W = 462.5, p-value = 0.4216
```

p-value = 0.4216 : Comme la p-value est supérieure à 0.05, il n'y a pas de preuve suffisante pour conclure qu'il y a une différence significative entre les sexes en ce qui concerne les niveaux de BENCE_J

1.2.6 âge et consomation de calcium

On peut commencer par explorer la realtion à l'aide d'une visualisation Boxplot :

.

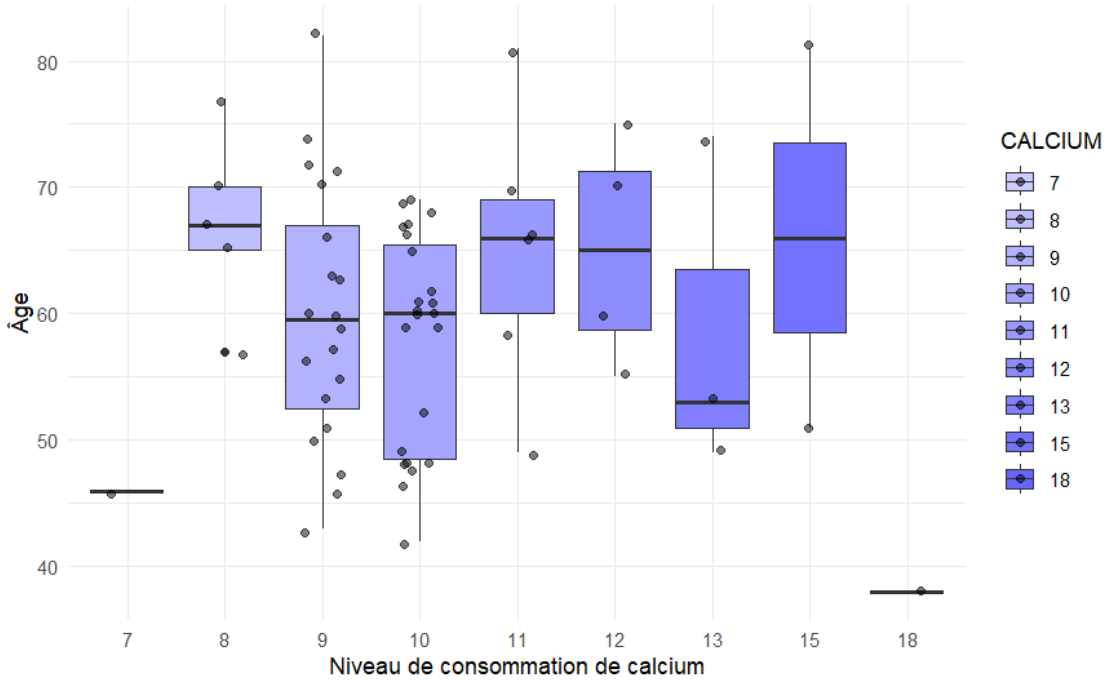


FIGURE 1.3 – **Ages moyens entre les niveaux de consommation de calcium**

Comme on peut le voir sur la Figure 1.3, les âges moyens sont plus élevés chez les patients ayant un niveau de consommation de calcium de 8 ou 15. De plus, l'écart-type le plus élevé se trouve pour le niveau de consommation de calcium de 15, ce qui indique que la dispersion des âges autour de la moyenne est plus grande pour ce groupe. En revanche, les âges moyens les plus faibles sont observés chez les patients ayant un niveau de consommation de calcium de 7 ou 18, soit aux deux extrémités des niveaux étudiés. Cela suggère que les patients de ces groupes tendent à être plus jeunes en moyenne.

1.3 Analyse de survie

1.3.1 Calcul des statistiques descriptives pour la survie

Dans cette section, nous nous concentrons sur les patients atteints de myélome, identifiés par une valeur supérieure à zéro dans la variable P_MYEL.

Nous présentons tout d'abord les statistiques descriptives des temps de survie des patients diagnostiqués avec un myélome.

```
> summary(dat2$TEMPS)
```

Statistique	Valeur
Min.	0.25
1er Quartile (25%)	6.00
Médiane	14.00
Moyenne	23.01
3e Quartile (75%)	34.00
Max.	91.00

TABLE 1.3 – Résumé des statistiques descriptives de la variable TEMPS

```
> t.test(dat2$TEMPS)
```

Statistique	Valeur
t-value	7.8286
df	64
p-value	6.465e-11
Intervalle de confiance à 95%	[17.13651, 28.87887]
Moyenne estimée	23.00769

TABLE 1.4 – Résultats du test t de Student pour la variable TEMPS

Le test affiche une statistique t de 7.8286 avec une valeur de p très faible ($p < 6.465e\text{-}11$), indiquant une différence significative entre la moyenne observée dans l'échantillon et la valeur théorique de 0.

L'intervalle de confiance à 95% pour la moyenne est [17.13, 28.87], ce qui assure un haut niveau de confiance quant à la position de la moyenne réelle de la population dans cette plage.

En résumé, la variable TEMPS est significativement différente de zéro, et sa distribution décroissante suggère que la majorité des valeurs sont concentrées dans les niveaux inférieurs, avec quelques valeurs plus élevées.

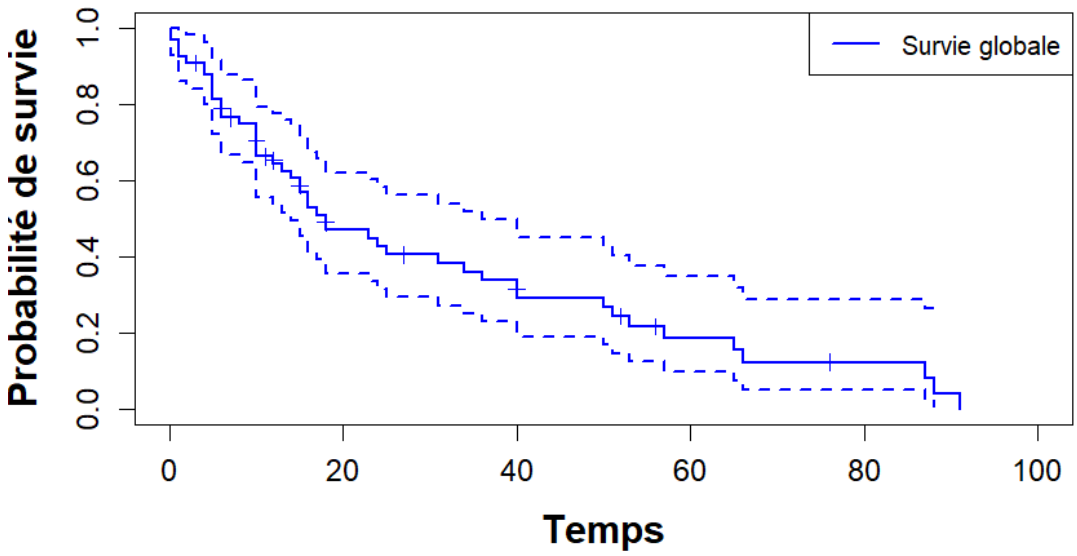


FIGURE 1.4 – Distribution des temps de survie des patients atteints de myélome

1.3.2 Distribution de la variable de l'indice de Bence Jones

Nous allons diviser les données en deux parties : l'une contenant des données avec une valeur de 1 (pour la présence de l'indicateur), et l'autre avec une valeur de 0 (pour l'absence de l'indicateur).

```
dat_avec =dat2 %>% filter(BENCE_J =="1")
dat_sans =dat2 %>% filter(BENCE_J =="0")
```

Avec la commande summary on obtient les différentes partitions :

```
summary(dat_avec$TEMPS)
summary(dat_sans$TEMPS)
```

Avec

Statistique	Valeur
Min.	0.25
1er Quartile (25%)	8.00
Médiane	34.00
Moyenne	34.10
3e Quartile (75%)	52.00
Max.	91.00

VS

Sans

Statistique	Valeur
Min.	0.25
1er Quartile (25%)	6.00
Médiane	11.50
Moyenne	16.93
3e Quartile (75%)	18.00
Max.	76.00

TABLE 1.5 – Impact de l'indice Bence-Jonessur les temps de survie : Analyse descriptive

```
> t.test(dat_sans$TEMPS)
        # t =6.3747, df =41, p-value =1.268e-07
> confidence interval:
          [11.56956 22.29949]
> mean of x
         16.93452
```

Pour les patients présentant l'indice de Bence Jones, la médiane du temps de survie est de 34 unités de temps, tandis que la moyenne est de 16.93452 unités. Par ailleurs, le test t révèle que la moyenne du temps de survie diffère de manière significative de zéro (p-value = 1.268e-07), avec un intervalle de confiance à 95% compris entre [11.56956, 22.29949].

```
> t.test(dat_sans$TEMPS)
       # t =5.5067, df =22, p-value =1.559e-05
> confidence interval:
       [21.25628 46.93937]
> mean of x
       34.09783
```

Dans le cas de l'absence de l'indice de Bence Jones, la médiane du temps de survie est de 11.50 unités de temps, tandis que la moyenne est de 34.09783 unités. Le test t indique que la moyenne du temps de survie diffère significativement de zéro (p-value = 1.559e-05), avec un intervalle de confiance à 95% allant de [21.25628 , 46.93937].

Ces résultats montrent que, dans les deux groupes, les moyennes des temps de survie sont significativement différentes de zéro, ce qui suggère que les patients des deux groupes ont survécu pendant une période notable. Cependant, la moyenne du temps de survie semble être plus élevée pour les patients présentant l'indice de Bence Jones par rapport à ceux n'ayant pas cet indice.

Les courbes de survie suivantes permet de confirmer cette hypothèse :

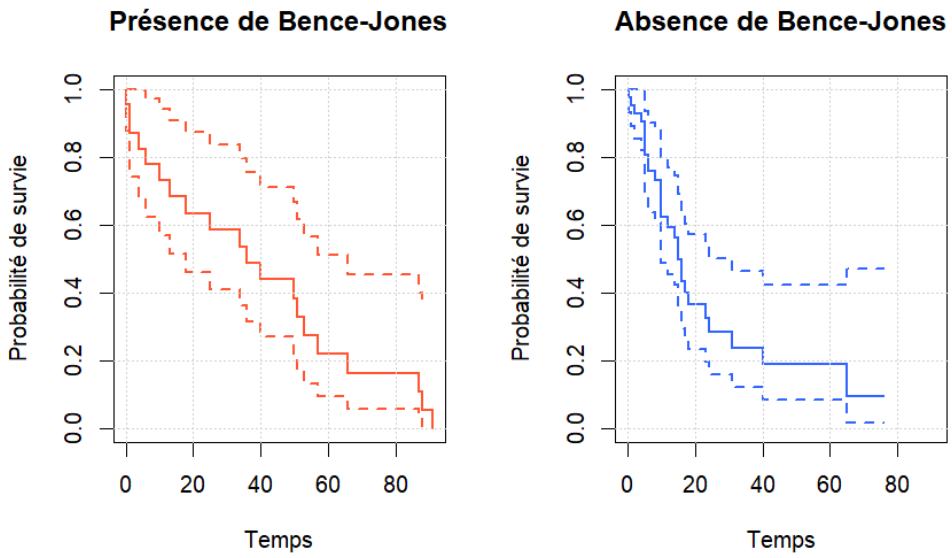


FIGURE 1.5 – Comparaison des courbes de survie des patients selon la présence de Bence-Jones

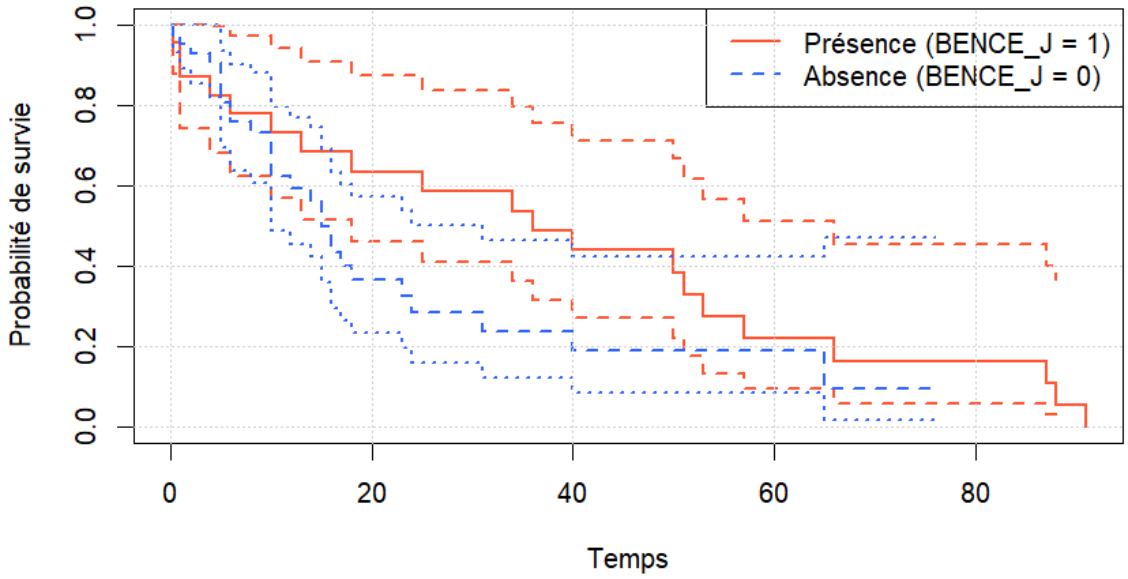


FIGURE 1.6 – Courbe de survie des patients en fonction de l'indice de Bence Jones

La courbe de survie des patients sans l'indice de Bence Jones est inférieure à celle des individus présentant cet indice. Cela indique que les individus avec l'indice de Bence Jones ont de meilleures chances de vivre plus longtemps.

En d'autres termes, la présence de l'indice de Bence Jones semble être associée à une augmentation de la probabilité de survie ou à une durée de survie plus longue. Cette observation suggère que les individus porteurs de cet indice bénéficient d'une probabilité de survie plus élevée par rapport à ceux qui ne l'ont pas.

1.3.3 Tester l'hypothèse Les courbes de survie sont égales dans les 2 groupes ?

Pour tester cette hypothèse, on va utilisez le **test de Log-rank**

les hypothèses pour le test du Log-rank sont les suivantes :

Hypothèse nulle ($\mathcal{H}_0$) :
Les courbes de survie des deux groupes sont identiques. il n'y a pas de différence significative entre les groupes en termes de survie.

$$\mathcal{H}_0 : S_1(t) = S_2(t), \quad \forall t \tag{1.1}$$

Hypothèse alternative ($\mathcal{H}_a$) :
Les courbes de survie des deux groupes sont différentes. il existe une différence significative dans les probabilités de survie entre les groupes à un moment donné.

$$\mathcal{H}_a : S_1(t) \neq S_2(t) \quad \text{pour au moins un } t \tag{1.2}$$

```
survdiff(Surv(TEMPS, DECES) ~ BENCE_J, data =dat2)
```

```
Call:
               N Observed Expected (O-E)^2/E (O-E)^2/V
> BENCE_J=0   42       28     23.4     0.885      2.07
> BENCE_J=1   23       20     24.6     0.845      2.07

> Chisq= 2.1  on 1 degrees of freedom, p= 0.1
```

le test de chi-deux (χ^2) donne une valeur unique de 2.1 avec 1 degré de liberté, et une p-value de 0.01. Comme la p-value < 0.05, nous avons suffisamment de preuves pour rejeter l'hypothèse nulle, qui stipule qu'il n'existe pas de différence significative entre les courbes de survie des deux groupes.

En d'autres termes, cela suggère qu'il existe une différence significative de survie entre les patients présentant ou non l'indice de Bence Jones.

En résumé, le test du log-rank montre que la présence de l'indice de Bence Jones est liée à des différences significatives dans les taux de survie des patients.

Nous pourrions mis en place les tests suivants :

☐ Analyse de régression de Cox : Cette analyse permettrait de mieux comprendre comment la présence de l'indice de Bence Jones, ainsi que d'autres variables, influent sur la survie des patients.

☐Test de Gehan-Breslow-Wilcoxon : Similaire au test du log-rank, ce test accorde plus de poids aux événements survenant tôt dans le suivi. Il est particulièrement utile lorsqu'on souhaite se concentrer sur les événements précoces.

☐Test de Mantel-Haenszel : Ce test évalue l'association entre une variable binaire (comme la présence/absence de l'indice de Bence Jones) et le temps de survie, tout en contrôlant les effets d'autres variables potentiellement influentes.

1.3.4 Association entre la survie et la consommation de calcium

1.3.4.1 Test direct de l'association

Utiliser un modèle de Cox pour tester directement l'association entre la variable CALCIUM et la survie.

Nous allons utiliser un modèle de Cox à risques proportionnels pour tester directement l'association entre la variable CALCIUM et la survie.

```
cox_calcium <- coxph(Surv(TEMPS, DECES) ~ CALCIUM, data =data)
summary(cox_calcium)
```

Variable	coef	exp(coef)	se(coef)	z	p-value
CALCIUM	0.1082	1.1143	0.1011	1.07	0.285

TABLE 1.6 – Résultats du modèle de Cox pour la variable CALCIUM

```
> n= 65, number of events= 48
```

Nombre total d'observations (n) : 65 et Nombre d'événements (décès) : 48

```
            coef exp(coef) se(coef)    z Pr(>|z|)
CALCIUM   0.1082    1.1143   0.1011 1.07    0.285
```

Une augmentation d'une unité dans le niveau de CALCIUM est associée à une augmentation du risque de décès d'un facteur de 10.2, mais cette relation n'est pas significative statistiquement (p-value = 0.285), ce qui suggère qu'il n'y a pas suffisamment de preuves pour conclure que CALCIUM a un impact réel sur la survie des patients

1.3.4.2 Stratification de la variable CALCIUM

Diviser la variable CALCIUM en groupes (strates) et comparer les courbes de survie entre ces groupes.

☐ 1) On Va diviser la variable CALCIUM en trois groupes :

```
nb_strates <- 3
dat <- dat3 %>%
mutate(strate =ntile(CALCIUM, nb_strates))
```

☐ 2) Comparer les courbes de survie entre les strates :

Utilisons un test de log-rank pour comparer les courbes de survie entre les groupes.

```
logrank_test <- survdiff(Surv(TEMPS, DECES) ~ STRATE, data =data)
print(logrank_test)
```

```
                N Observed    Expected    (O-E)^2/E     (O-E)^2/V
strate=1 22        18          16.0        0.240         0.382
strate=2 22        18          14.7        0.739         1.173
strate=3 21        12           17.3       1.602         2.734

#  Chisq= 2.8  on 2 degrees of freedom, p= 0.2
```

Les résultats du test montrent que, pour les trois strates, les valeurs observées et attendues de survie ne diffèrent pas de manière significative, avec une statistique du chi carré de 2,8 sur 2 degrés de liberté et une p-valeur de 0,2, ce qui indique que la différence entre les courbes de survie n'est pas statistiquement significative.

☐ 3) Représentation graphique des courbes de survie :
Visualisons des courbes de survie pour chaque strate.

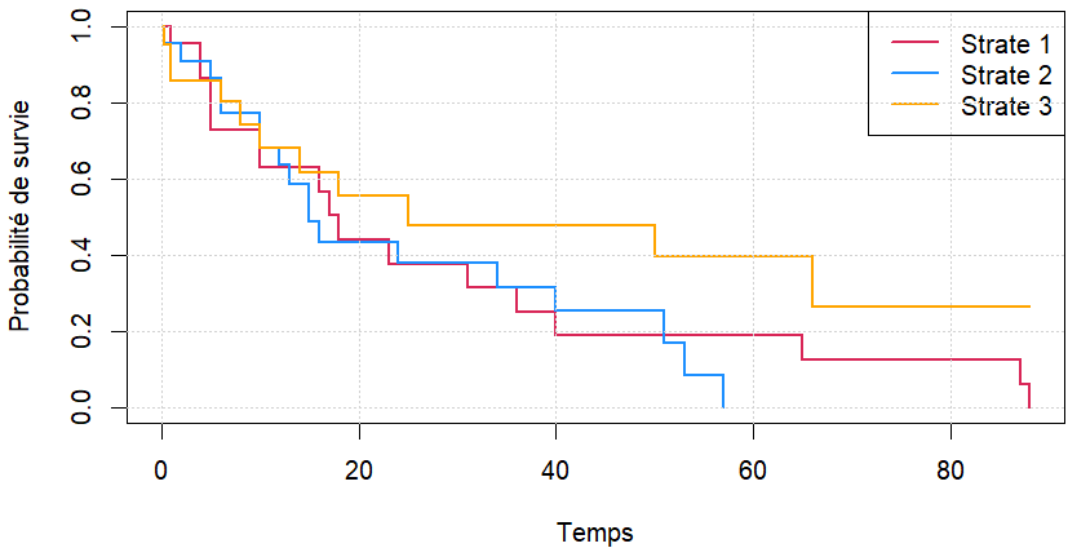


FIGURE 1.7 – Distibution des temps de survie des différentes strates

Sur le graphique précédent, on observe que les strates 1 et 2 sont relativement proches, ce qui indique que les niveaux de calcium dans ces deux groupes sont similaires ou légèrement différents. De plus, les strates 1 et 2 ont des niveaux plus bas que la strate 3, suggérant que cette dernière présente des niveaux de calcium plus élevés.

1.4 l'effet sur la survie des variables PQ, INFJ0, SEXE, FRACTURE

1.4.1 Association entre la survie et PQ

```
survdiff(Surv(TEMPS, DECES)~ PQ, data =dat3)

       Observed   Expected (O-E)^2/E (O-E)^2/V

PQ=0  9          8     4.46     2.815       3.29
PQ=1 56         40    43.54     0.288       3.29

       # Chisq= 3.3  on 1 degrees of freedom, p= 0.07
```

Les résultats du test montrent que chez les patients sans PQ (PQ=0), événements ont été observés, tandis que 8 étaient attendus.

Chez les patients avec PQ (PQ=1), 56 événements ont été observés, alors que 40 étaient attendus.

Le test du chi-carré donne une statistique de 3.3 avec une p-valeur de 0,07, ce qui suggère qu'il y a une différence significative dans les taux de survie entre les patients avec et sans PQ.

1.4.2 Association entre la survie et le sexe

```
survdiff(Surv(TEMPS, DECES)~ SEXE, data =dat3)

    # Chisq= 0.4  on 1 degrees of freedom, p= 0.5
```

Le test du chi-carré donne une statistique de 0.4 avec une p-valeur de 0.5, ce qui suggère qu'il n'y pas une différence significative dans les taux de survie entre les patients de sexe masculin et féminin.

1.4.3 Association entre la survie et INFJ0

```
survdiff(Surv(TEMPS, DECES)~ INFJ0, data =dat3)

       #   Chisq= 0.9  on 1 degrees of freedom, p= 0.3
```

Le test du chi-carré donne une statistique de 0.9 avec une p-valeur de 0.3, ce qui suggère qu'il n'y pas une différence significative dans les taux de survie entre lespatients avec et sans INFJ0

1.4.4 Association entre la survie et les fractures

```
survdiff(Surv(TEMPS, DECES)~ FRACTURE, data =dat3)

       #  Chisq= 0.9  on 1 degrees of freedom, p= 0.3
```

Le test du chi-carré donne une statistique de 0.9 avec une p-valeur de 0.3, ce qui suggère qu'il n'y pas une différence significative dans les taux de survie entre lespatients avec et sans fracture.

1.5 étudier l'effet sur la survie de variables quantitatives

pour étudier l'effet des variable quantitatives nous pourraions utiliser les trois méthodes suivantes :

- **Kaplan-Meier**.
- **Analyse de survie avec des modèles à effets mixtes** :Ces modèles permettent de modéliser à la fois les effets fixes, qui sont les influences constantes des variables explicatives sur la survie, et les effets aléatoires, qui tiennent compte de la variabilité individuelle non expliquée par les variables explicatives
- **Régression logistique pour la survie** : Nous pouvons recourir à des modèles paramétriques tels que le modèle de Weibull ou le modèle log-logistique pour modéliser la relation entre les variables explicatives et la survie.

CHAPITRE 2

TP 2

2.1 Exercice 1 : Kaplan Meier

2.1.1 Appliction 1 :

Nous avons des données représentant les durées de rémission (en semaines) de patients atteints de leucémie aiguë et traités par 6-MP. Certaines observations sont censurées (indiquées par 0), ce qui signifie que la rémission était toujours en cours au dernier suivi.

```
library(survival)
temps <- c(6, 6, 6, 6, 7, 9, 10,
    10, 11, 13, 16, 17, 19, 20, 22
    , 23, 25, 32, 32, 34, 35)
censure <- c(0, 0, 0, 1, 0, 1, 0,
     1, 1, 0, 0, 1, 1, 1, 0, 0, 1,
     1, 1, 1, 1)
```

TABLE 2.1 – **Caractéristiques de la distribution des temps de réalisation**

Individus (t_i)	**Time** (D_i)	**Status**
1	6	1
2	6	1
3	6	1
4	6	0
5	7	1
6	9	0
7	10	1
8	10	0
9	11	1
10	13	1
11	16	1
12	17	0
13	19	0
14	20	0
15	22	1
16	23	1
17	25	0
18	32	0
19	32	0
20	34	0
21	35	0

```r
surv_obj <- Surv(temps, 1- censure)
km_fit <- survfit(surv_obj ~ 1)
summary(km_fit)
```

TABLE 2.2 – Résultats de l'estimation de Kaplan-Meier

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
7	17	1	0.807	0.0869	0.653	0.996
10	15	1	0.753	0.0963	0.586	0.968
13	12	1	0.690	0.1068	0.510	0.935
16	11	1	0.627	0.1141	0.439	0.896
22	7	1	0.538	0.1282	0.337	0.858
23	6	1	0.448	0.1346	0.249	0.807

Le tableau 2.2 présente les caractéristiques de la distribution des temps de survie des patients. La colonne **Time** répertorie les durées de survie distinctes non censurées. La colonne **n.event** indique le nombre d'événements observés (rechutes) pour chaque durée de survie. La colonne **n.risk** représente le nombre de patients encore en observation juste avant le temps correspondant.
Les valeurs de **Survival** traduisent la probabilité de survie estimée à chaque instant t_i.

```r
plot(km_fit, xlab ="Temps (semaines)", ylab ="Probabilite de survie",
col ="blue", lwd =2)
```

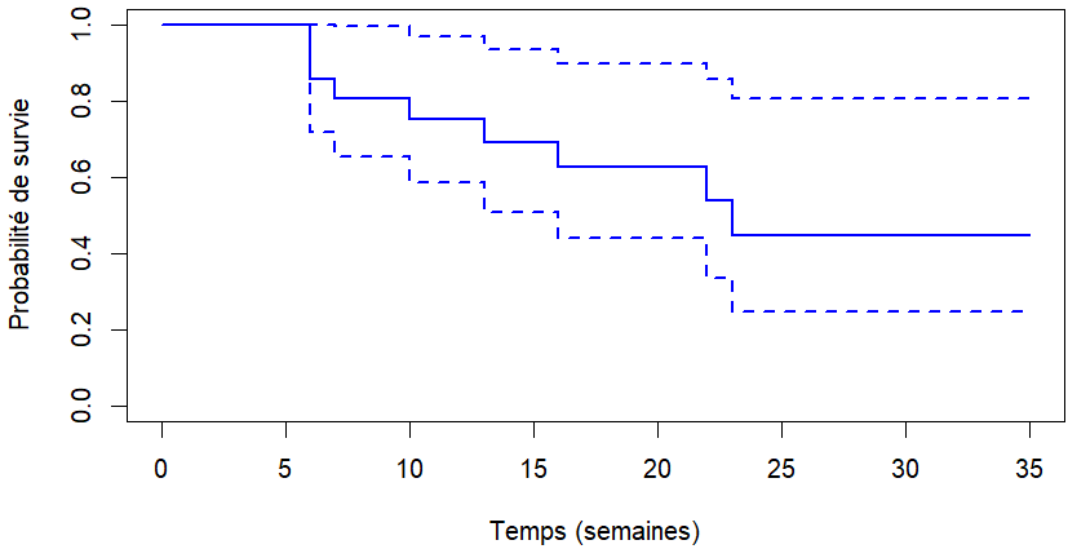


FIGURE 2.1 – Courbe de Kaplan-Meier leucémie

2.1.2 Applicatin 2 : Comparaison des estimateurs de survie

Durées de survie (mois) avec censure (*) : 1, 3, 4*, 5, 7*, 8, 9, 10*, 11, 13*.
Événements (décès) aux temps : 1, 3, 5, 8, 9, 11.

2.1.2.1 Estimateur de Kaplan-Meier

$$S(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

```
km_fit <- survfit(surv_data ~ 1)
summary(km_fit)
```

TABLE 2.3 – **Kaplan-Meier Estimate**

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	10	1	0.900	0.0949	0.7320	1.00
3	9	1	0.800	0.1265	0.5868	1.00
5	7	1	0.686	0.1515	0.4447	1.00
8	5	1	0.549	0.1724	0.2963	1.00
9	4	1	0.411	0.1756	0.1782	0.95
10	3	1	0.274	0.1620	0.0862	0.873
11	2	1	0.137	0.1264	0.0225	0.834

2.1.2.2 Estimateur de Fleming-Harrington

$S(t) = e^{-H(t)}$, où $H(t) = \sum_{t_i \leq t} \frac{d_i}{n_i}$

```
fh_fit <- survfit(surv_data ~ 1, type ="fleming-harrington")
summary(fh_fit) # Affiche S(t) =exp(-H(t))
```

TABLE 2.4 – **Fleming-Harrington Estimate**

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1	10	1	0.905	0.0905	0.7438	1.00
3	9	1	0.810	0.1210	0.6041	1.00
5	7	1	0.702	0.1451	0.4680	1.00
8	5	1	0.575	0.1653	0.3270	1.00
9	4	1	0.448	0.1706	0.2120	0.945
10	3	1	0.321	0.1624	0.1189	0.865
11	2	1	0.195	0.1384	0.0482	0.785

Extraire les résultats à des temps spécifiques (exemple pour t = 11)

```
cat("Kaplan-Meier a t=11 :", summary(km_fit, times =11)$surv, "\n")
cat("Fleming-Harrington a t=11 :", summary(fh_fit, times =11)$surv, "\n")
```

```
> Kaplan-Meier à t=11 : 0.2057143
> Fleming-Harrington à t=11 : 0.2714525
```

Survie estimée pour Fleming-Harrington légèrement supérieure. par exemple a t=11 : 27.14% vs 20.57% à t=11)

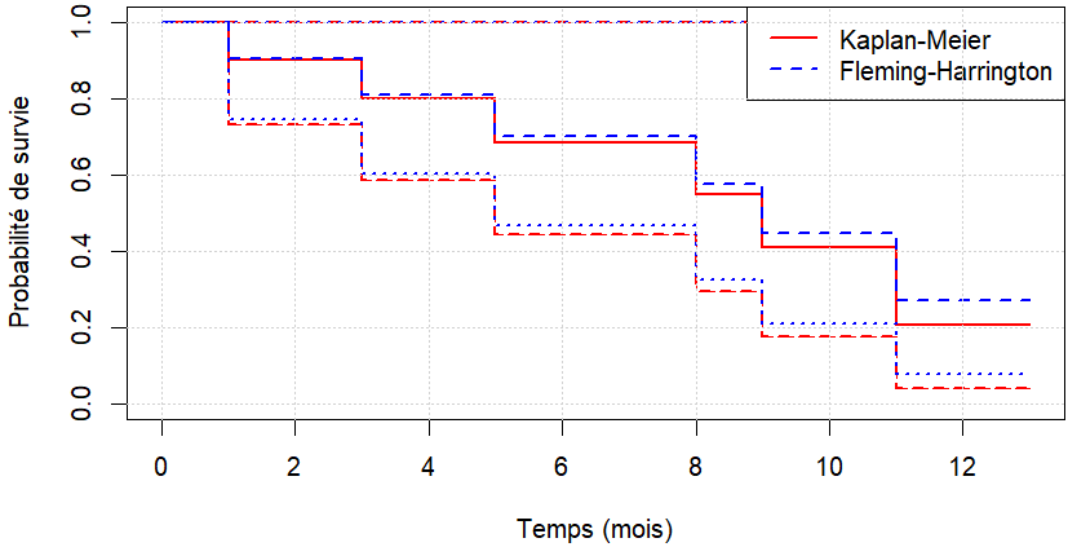


FIGURE 2.2 – courbes des survies Kaplan-Meier et Fleming-Harrington

```
print(km_fit)
```

```
      n  events  median  0.95LCL  0.95UCL
> 10      7       9       5          NA
```

```
print(fh_fit)
```

```
      n  events  median  0.95LCL  0.95UCL
> 10      7       9       5          NA
```

il est notable que les deux méthodes produisent des estimations identiques. Cette situation peut se produire dans des contextes spécifiques, notamment avec des échantillons de taille réduite ou des données présentant un faible taux de censure. En effet, sous certaines conditions favorables, les algorithmes de calcul peuvent converger vers des résultats similaires.

2.1.2.3 Application 3 :

```
data(aml)
fit <- survfit(Surv(time, status) ~ 1, data =aml)
summary(fit)
```

Time	n.risk	n.event	Survival	Std.Err	Lower 95% CI	Upper 95% CI
5	23	2	0.9130	0.0588	0.8049	1.000
8	21	2	0.8261	0.0790	0.6848	0.996
9	19	1	0.7826	0.0860	0.6310	0.971
12	18	1	0.7391	0.0916	0.5798	0.942
13	17	1	0.6957	0.0959	0.5309	0.912
18	14	1	0.6460	0.1011	0.4753	0.878
23	13	2	0.5466	0.1073	0.3721	0.803
27	11	1	0.4969	0.1084	0.3240	0.762
30	9	1	0.4417	0.1095	0.2717	0.718
31	8	1	0.3865	0.1089	0.2225	0.671
33	7	1	0.3313	0.1064	0.1765	0.622
34	6	1	0.2761	0.1020	0.1338	0.569
43	5	1	0.2208	0.0954	0.0947	0.515
45	4	1	0.1656	0.0860	0.0598	0.458
48	2	1	0.0828	0.0727	0.0148	0.462

TABLE 2.5 – Caractéristiques de la distibution des temps de rélisation

```
plot(fit)
```

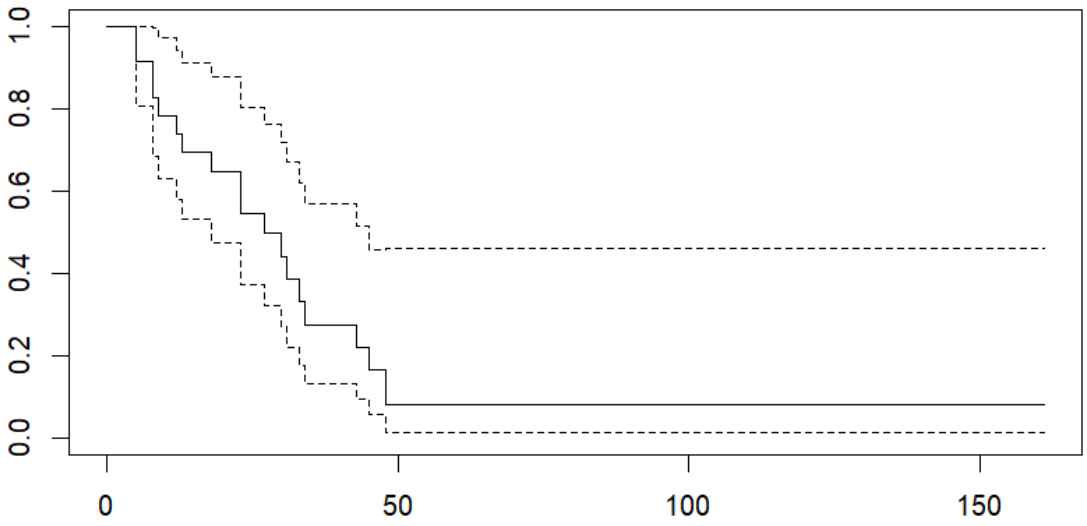


FIGURE 2.3 – Courbe de survie pour les donnees aml

Cette courbe présente une chute abrupte jusqu'au temps de survie de 50, avec une probabilité résiduelle de 0,1. Ce déclin rapide révèle qu'une majorité d'individus subissent un événement critique avant ou peu après ce seuil, entraînant une survie limitée au-delà de 50 unités de temps.

2.2 Exercice 2 :

Par la suite de cet exercice, nous allons utiliser le jeu de données chomage-tp.rda

```
new_env <- new.env()
load("chomage_tp.rda", envir =new_env)
dat <- data.frame(new_env$Dataset)
summary(dat)
```

```
     temps                   statut
 H:101   Min.   : 0.004806   censure  : 84
 F:101   1st Qu.: 0.822927   evenement:118
 Median : 1.837516
 Mean   : 2.839574
 3rd Qu.: 4.039206
 Max.   :14.690237
```

on construire une courbe de survie de Kaplan Meier, et la fonction dehasard cumulé, graphiquement.

```
dat$statut <- ifelse(dat$statut =="evenement", 1, 0)
fit <- survfit(Surv(temps, statut) ~ 1, data =dat)
ggsurvplot(fit, conf.int =TRUE, risk.table =TRUE, pval =TRUE,
data =dat, surv.median.line ="hv", palette ="jco")
```

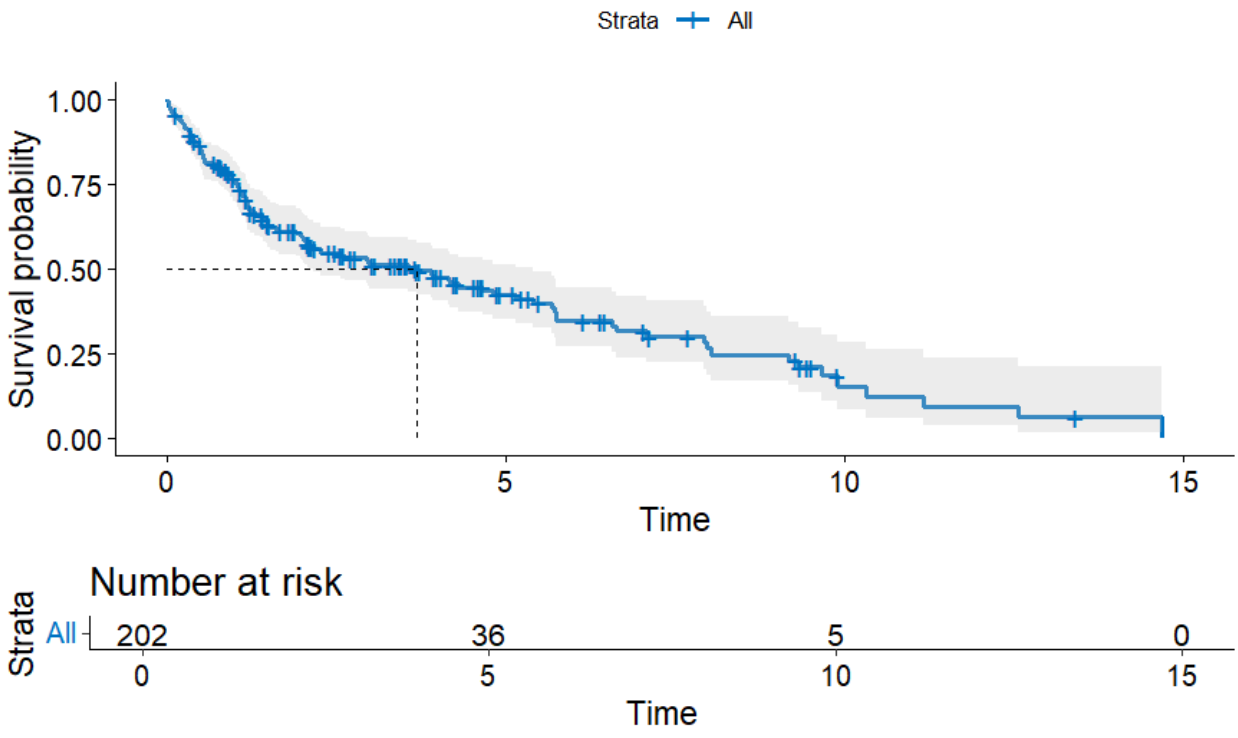


FIGURE 2.4 – Courbe de survie de Kaplan Meier data-chomage

La probabilité de rester au chômage (survie) diminue rapidement dans les premières pé-

riodes (jusqu'à temps=5), passant de 100% à environ 25%. Après temps=10, la courbe se stabilise, suggérant que les individus restants ont une probabilité faible de sortir du chômage.

Nombre à risque : 202 individus au départ, mais seulement 5 à temps=10, ce qui confirme une attrition rapide (sorties du chômage ou censure).

```
hazard_cumulative <- 1- exp(-fit$cumhaz)
times <- fit$time
plot(times, hazard_cumulative, type ="s", col ="red", lty =2,
xlab ="Temps", ylab ="Fonction de hasard cumulee",
main ="Fonction de hasard cumulee")
```

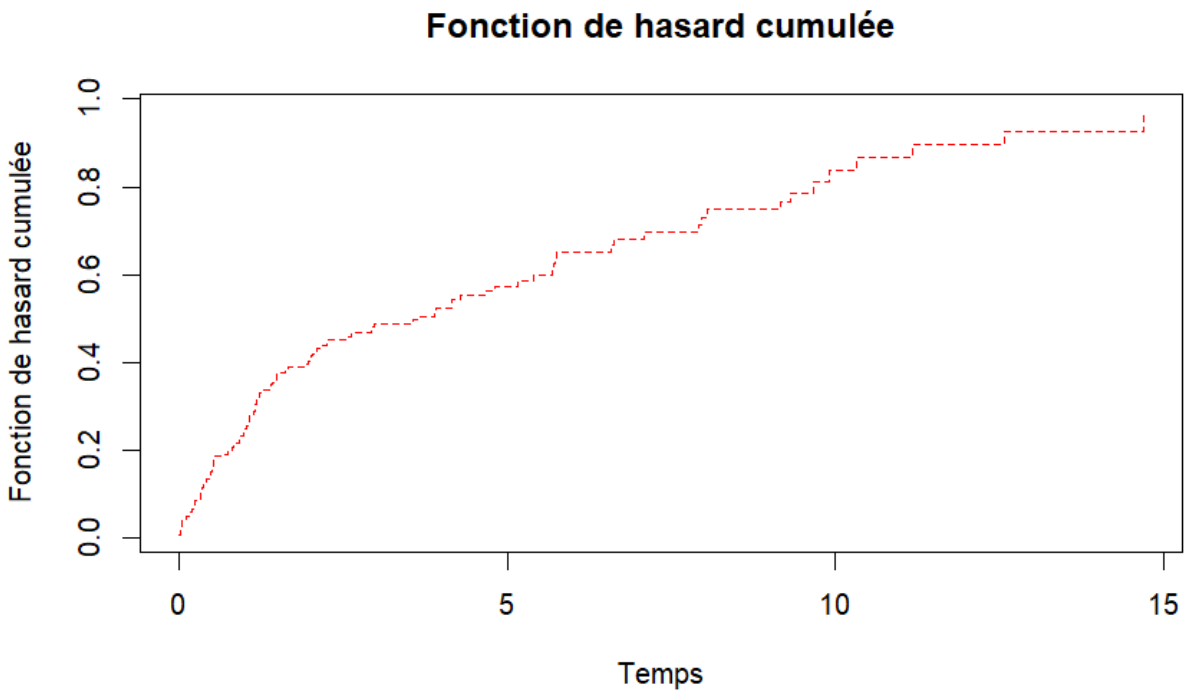


FIGURE 2.5 – Courbe de hasard cumulé des données chomage

La pente la plus raide entre temps=0 et temps=5 correspond à la phase où le risque de quitter le chômage est le plus élevé, après temps=10, la courbe s'aplatit, indiquant que le risque résiduel devient faible.

2.2.0.1 Méthode de Nelson Aalen

La méthode de Nelson-Aalen :

```
fit_aalen <- survfit(Surv(temps, statut) ~ 1, data =dat, stype =2)
plot(fit_aalen, col ="blue", lty =2, lwd =2,
xlab ="Temps",
ylab ="Probabilite de survie",
main ="Courbe de survie (Nelson Aalen)")
```

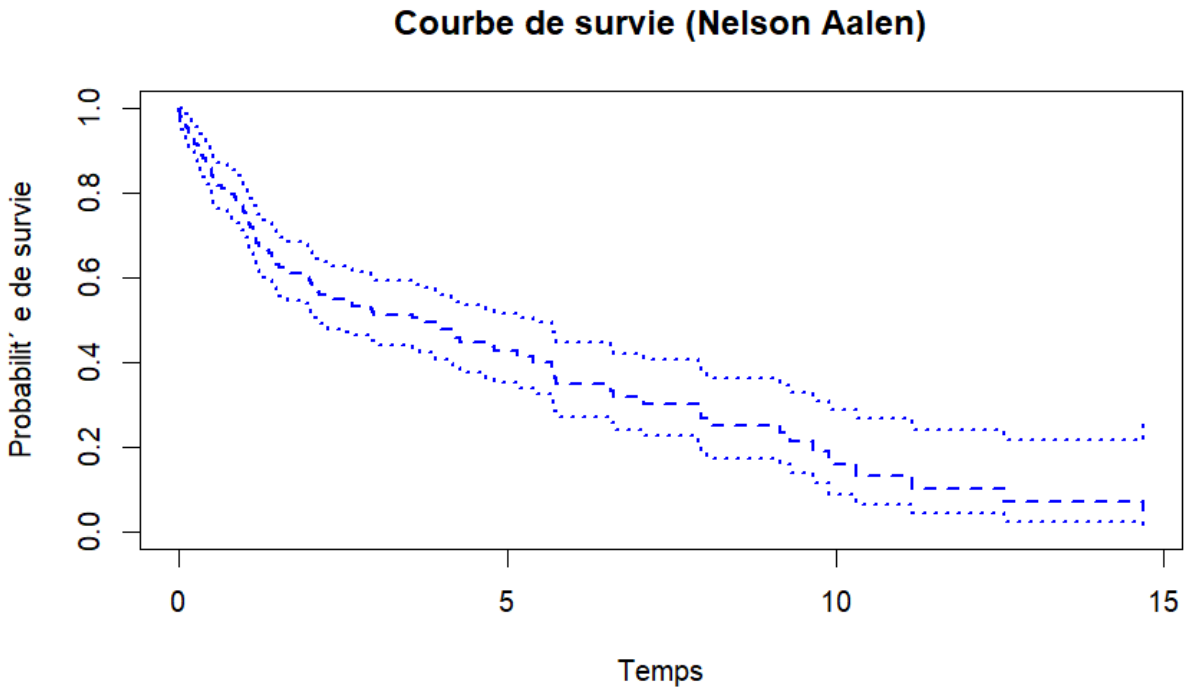


FIGURE 2.6 – Courbe de survie (Nelson Aalen)

Cette courbe est similaire à la courbe de Kaplan-Meier. Dans ce cas, le résultat restera le même.

2.2.0.2 Comparaison de genres

On va comparer les courbes de survie en fonction du genre des individus(homme ou femme). Le graphique, ci-dessous, présente les deux courbes de survie par genre.

```
fit_male <- survfit(Surv(temps, statut) ~ 1, data =dat[dat$genre =="H", ])
fit_female <- survfit(Surv(temps, statut) ~ 1, data =dat[dat$genre =="F", ])

fit_combined <- survfit(Surv(temps, statut) ~ genre, data =dat)
ggsurvplot(fit_combined, data =dat,
conf.int =TRUE, risk.table =TRUE, pval =TRUE,
surv.median.line ="hv", palette =c("blue", "red"),
legend.labs =c("Hommes", "Femmes"))
```

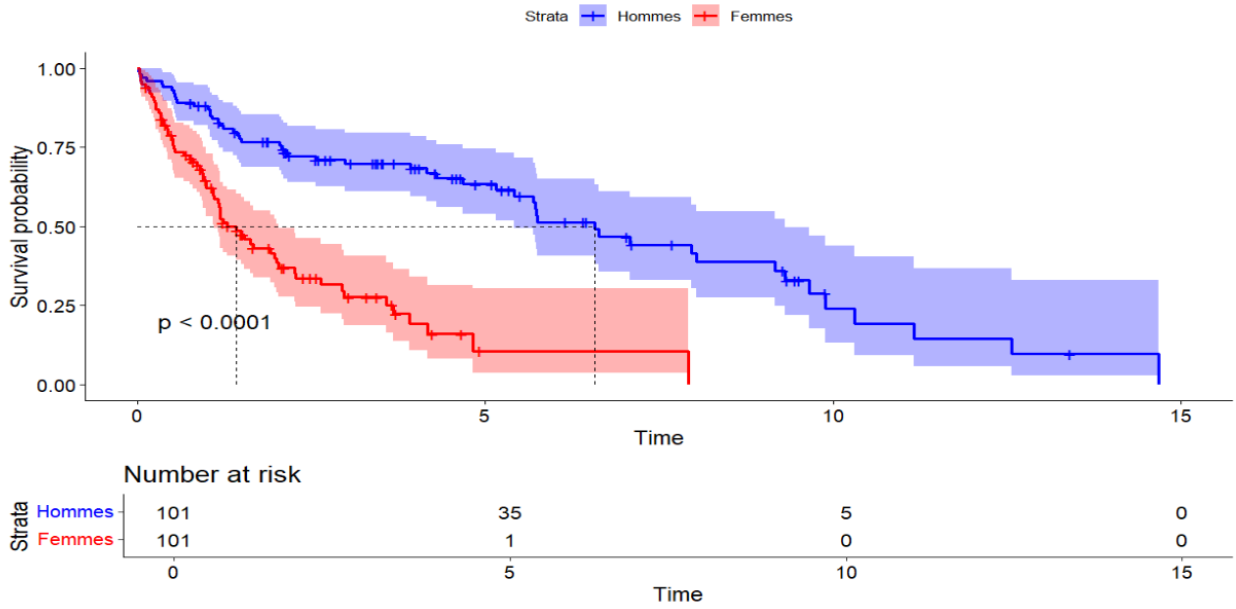


FIGURE 2.7 – Courbes de survie comparées entre hommes et femmes

Les deux courbes présentent une différence marquée : celle des femmes est nettement inférieure à celle des hommes. Par ailleurs, leur durée de survie moyenne est plus courte. Ainsi, les femmes ont une probabilité de survie réduite et un risque de décès plus élevé et plus précoce que les hommes.

```
log_rank_test <- survdiff(Surv(temps, statut) ~ genre, data =dat)
print(log_rank_test)
```

```
            Observed Expected (O-E)^2/E (O-E)^2/V
genre=H  101       52     81.8      10.9      40.7
genre=F  101       66     36.2      24.6      40.7

 # Chisq= 40.7  on 1 degrees of freedom, p= 2e-10
```

log-rank donne une p-valeur de 2e-10, ce qui est très significatif (p <0.05). Cela signifie qu'il existe une différence statistiquement significative entre les courbes de survie des hommes et des femmes.

2.3 Exercice 3 : Processus de départ des enfants

Cet exercice propose l'analyse du fichier leaving-home.rde, tiré du Panel Suisse des Ménages. Ce jeu de données porte sur les trajectoires de décohabitation familiale auprès d'un échantillon de 5 560 individus. Il comprend les variables suivantes :

— Un identifiant unique
— Le sexe des individus

— L'âge lors de l'enquête
— Le statut de formation (en cours ou non)
— La catégorie socioprofessionnelle du père
— Le prestige professionnel paternel (mesuré par l'échelle de Treiman)
— La région de résidence
— La langue principale
— L'âge au départ du domicile parental (pour ceux ayant déjà quitté le foyer)

2.3.1 Courbe de survie de "partir de chez ses parents"

Apres avoir recode certaines modalités de variable

```
plot(fit, xlab ="Age",
ylab ="Probabilite de rester chez ses parents",
main ="Courbe de survie - Partir de chez ses parents")
```

```
fit <- survfit(Surv(agdeppar, c_deppar) ~ 1, data =event_data)
summary(fit)
```

2.3.2 Courbe de survie des langues parlées

On va analyser les courbes de survie séparées pour les répondants ayant été interviewés dans chaque langue. Pour commencer, nous présentons les résultats d'un test de Wilcoxon-Mann-Whitney afin de comparer les distributions des âges entre les différentes langues

```
data: ages_par_langue[[1]] and ages_par_langue[[2]]
W =1866292, p-value =0.3832
alternative hypothesis: true location shift is not equal to 0
```

Nous observons que la valeur de la statistique W est de 1 866 292 et que la p-value est de 0,3832, ce qui est supérieur au seuil de signification habituel de 0,05. Par conséquent, nous n'avons pas suffisamment de preuves pour rejeter l'hypothèse nulle. Cela suggère qu'il n'y a pas de différence significative dans la distribution des âges entre les individus parlant différentes langues

2.3.2.1 Allemand

```
donnees_allemand <- DF %>% filter(langue =="allemand")
fit_allemand <- survfit(Surv(agdeppar, c_deppar) ~ 1, data =donnees_allemand)
plot(fit_allemand, col="orange", xlab ="Age", ylab ="Probabilité de rester chez
    ses parents")
```

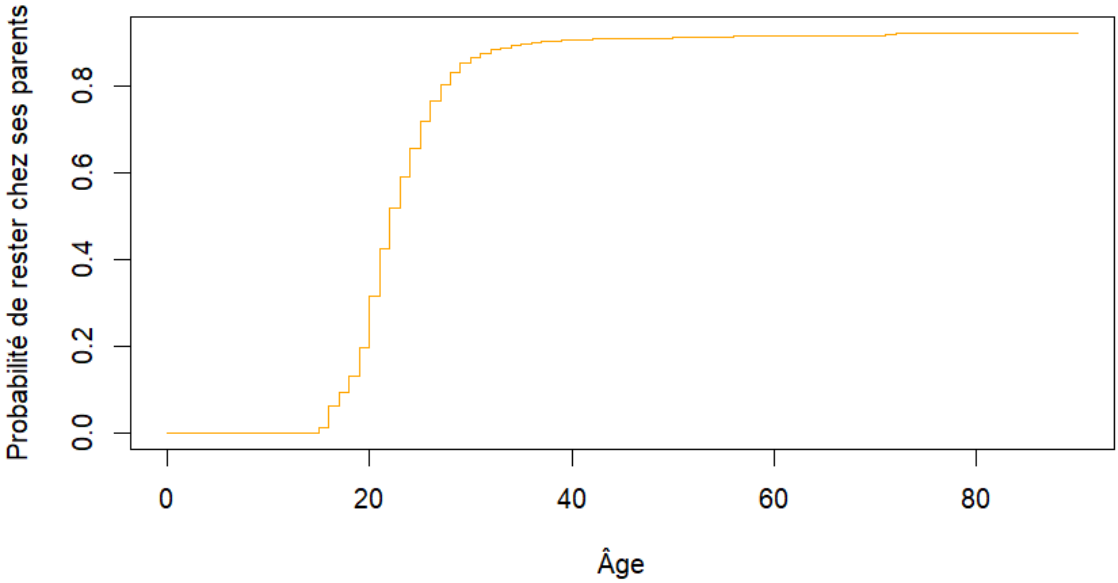


FIGURE 2.8 – **Courbe de survie - Allemand**

2.3.2.2 Francais

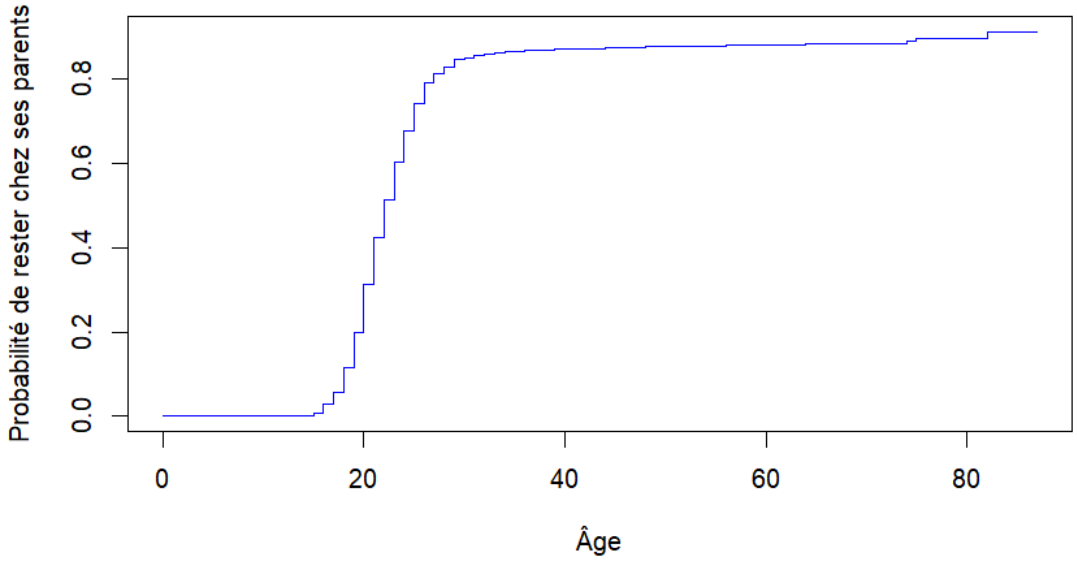


FIGURE 2.9 – **Courbe de survie - Francais**

2.3.2.3 Italien

```
donnees_italien <- DF %>% filter(langue =="italien")
fit_italien <- survfit(Surv(agdeppar, c_deppar) ~ 1, data =donnees_italien)
plot(fit_italien, col="red", xlab ="Age",
ylab ="Probabilité de rester chez ses parents")
```

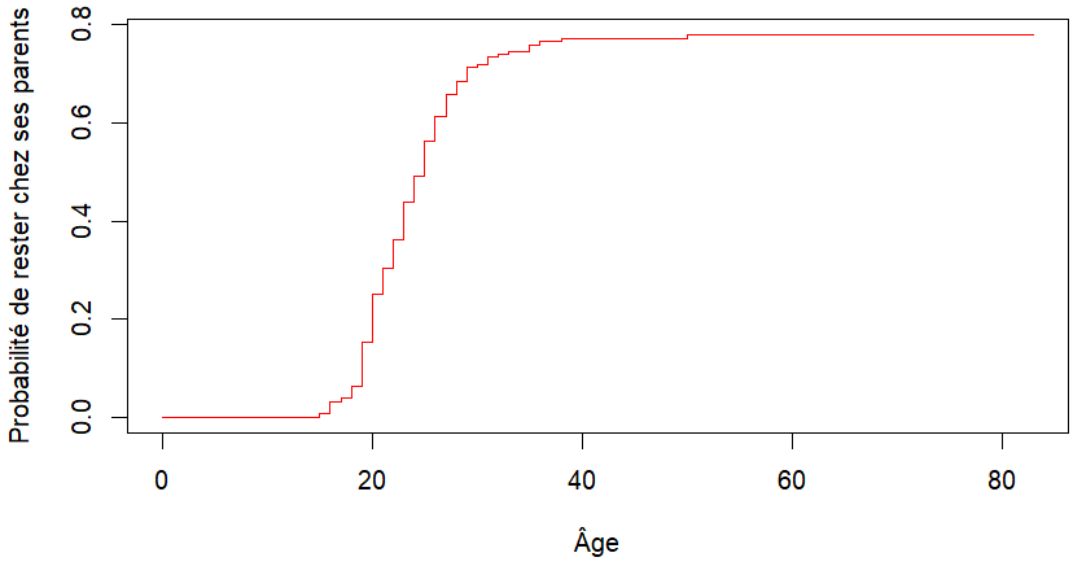


FIGURE 2.10 – **Courbe de survie - Italien**

Les trois graphiques semblent visuellement très similaires. Cependant, on observe une légère différence dans l'espacement des courbes, en particulier entre la courbe bleue et les autres. Cela pourrait indiquer que les départs du foyer parental se produisent à des âges spécifiques moins fréquemment pour les enfants français que pour leurs homologues allemands et italiens. De manière générale, les enfants français tendent à rester plus longtemps chez leurs parents avant de devenir indépendants.
Par ailleurs, bien que le rythme global des courbes de survie soit similaire, la différence dans l'espacement des paliers pourrait refléter des variations d'ordre culturel ou socio-économique entre les trois groupes