

# Méthodes d'analyse biostatistique projet 2

(date limite : le 21 novembre)

Wen, Zehai; Li, Qingyue

2023-11-20 15:17:52-05:00

## Exercice 1a

Décrivez brièvement les variables du jeu de données `diabetic` dans la bibliothèque `survival` en R. Quelle était une question de recherche menant à cette collecte de données ?

### Solution

- **ID** : Il est utilisé pour distinguer chaque participant dans l'ensemble de données.
- **laser** : Il s'agit du type de traitement laser reçu. 1 = xenon, 2 = argon.
- **age** : Il s'agit de l'âge auquel le diabète a été diagnostiqué chez le patient.
- **eye** : Il s'agit d'un facteur avec des niveaux de gauche et de droit.
- **trt** : Il s'agit du groupe de traitement. 0 = contrôle, 1 = laser.
- **risk** : Il s'agit de classer les participants dans les groupes de risques différents. Les valeurs varient de 6 à 12 où 6 représente le groupe avec risque le plus petit.
- **time** : Il s'agit du nombre de jours du début de la recherche à la cécité ou à la dernière observation.
- **status** : Il s'agit d'une variable binaire pour indiquer si une perte de vision s'est produite au cours de la période d'étude, où 0 signifie qu'on ne perd pas de vision et 1 pour le contraire.

Ces données proviennent d'une étude d'analyse de la survie de patients atteints de rétinopathie diabétique à haut risque, conçue pour évaluer l'efficacité du traitement au laser dans le ralentissement de la progression de la cécité. ////

## Exercice 1bc

Produisez un graphique de Kaplan-Meier utilisant l'échantillon entier, le groupe de `trt = 0` et le groupe de `trt = 1` pour le temps jusqu'à devenir aveugle. Décrivez le courbe de survie dans vos mots.

**Solution** Nous avons obtenu un graphique montrant la probabilité de survie pour l'ensemble de l'échantillon depuis le début du traitement jusqu'à la cécité (définie comme une baisse de l'acuité visuelle à  $\frac{5}{200}$ ). La courbe commence à 1 (soit une probabilité de survie de 100%) et diminue progressivement. Par conséquent, il y a plus de risques de devenir aveugle au fil du temps. Sur l'axe des abscisses, le temps est indiqué en jours. L'axe des ordonnées montre la proportion de patients qui n'ont pas encore diagnostiqué de cécité à un moment donné. Le graphique en R comporte trois courbes qui représentent la probabilité de survie de l'échantillon entier (représenté en noir), du groupe non traité (représenté en rouge, `trt = 0`) et du groupe traité (représenté en bleu, `trt = 1`).

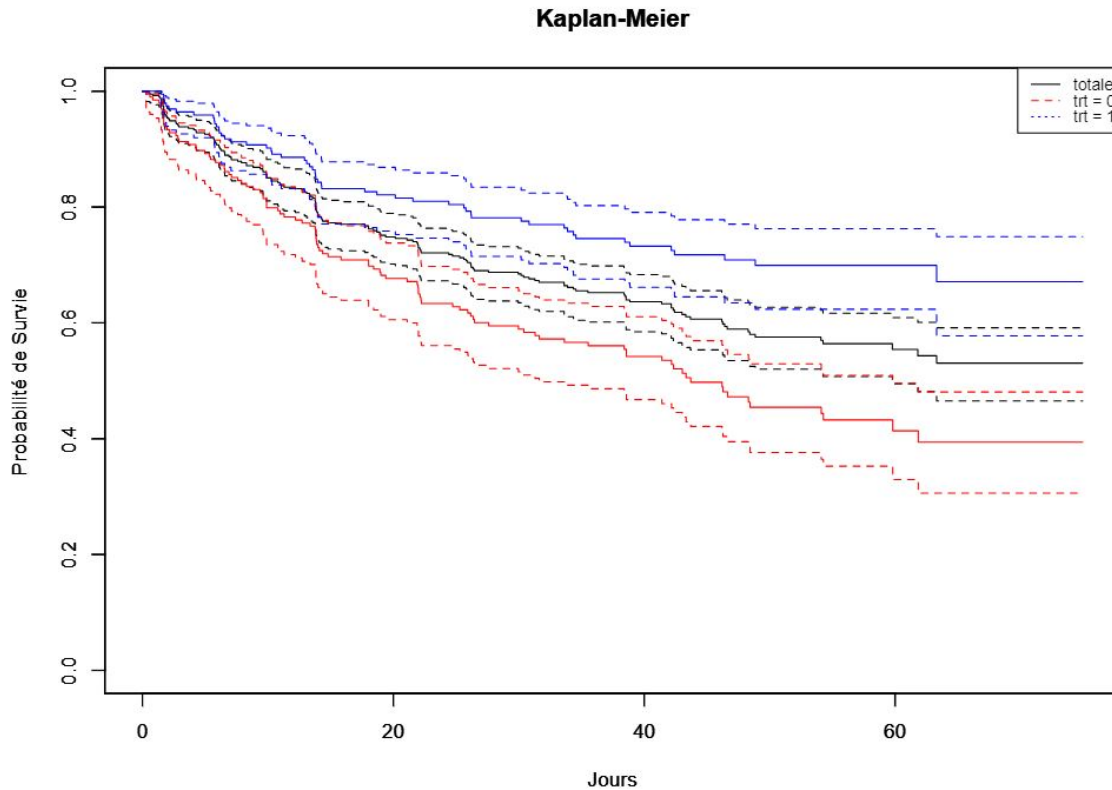
```
#Kaplan-Meier
result.km <- survfit(Surv(time, status) ~ 1, conf.type="log-log")
plot(result.km, xlab = "Jours", ylab = "Probabilite de Survie", main =
      "Kaplan-Meier")
result.kmtrt0 <- survfit(Surv(time[trt == 0], status[trt == 0]) ~ 1,
      conf.type="log-log")
par(new=TRUE)
plot(result.kmtrt0, xlab = "Jours", ylab = "Probabilite de Survie", main =
      "Kaplan-Meier", col = "red")
```

```

result.kmtrt1 <- survfit(Surv(time[trt == 1], status[trt == 1]) ~ 1,
  conf.type="log-log")
par(new=TRUE)
plot(result.kmtrt1, xlab = "Jours", ylab = "Probabilite de Survie", main =
  "Kaplan-Meier", col = "blue")
legend("topright", legend=c("totale", "trt = 0", "trt = 1"), col=c("black", "red",
  "blue"), lty=1:5, cex=0.8)
#abline ( v = 43.7 , col = 'red' , lty =2)

```

Le plot est :



Le graphique montre que la courbe de survie du groupe traité est plus élevée que celle du groupe non traité à la plupart des moments, ce qui suggère que le traitement puisse aider à retarder l'apparition de la cécité. ///

## Exercice 1d

Pour le groupe `trt = 0`, trouvez le temps médian de survie et construisez un intervalle de confiance à 95% pour le temps médian de survie.

**Solution** Le code R crée d'abord un sous-ensemble des données `subsetdata` en filtrant les patients avec `trt == 0` de l'ensemble de données `diabetic` à l'aide de la fonction `subset`. Ensuite, nous avons utilisé la fonction `survfit` pour estimer la durée de survie médiane de ce sous-ensemble et l'intervalle de confiance à 95% correspondant.

```

subset_data <- subset(diabetic, trt == 0)
fit <- survfit(Surv(time, status) ~ 1, data=subset_data, conf.type="log-log")
result.km <- fit

```

```
print(result.km)
```

La réponse est :

```
Call: survfit(formula = Surv(time, status) ~ 1, data = subset_data,
  conf.type = "log-log")
```

```
      n events median 0.95LCL 0.95UCL
[1,] 197 101 43.7 31.6 59.8
```

Donc :

1. Chez les patients n'ayant pas reçu de traitement ( $\text{trt} = 0$ ), la durée médiane de survie est de 43,7 jours.
  2. L'intervalle de confiance à 95% a une limite inférieure de 31,6 jours et une limite supérieure de 59,8 jours.
- ////

## Exercice 1e

Faites un test de Log-Rank pour comparer les deux groupes. Discutez les résultats.

**Solution** On effectue un test de Log-Rank :

```
result.logrank <- survdiff(Surv(time, status) ~ trt, data = diabetic)
print(result.logrank)
```

La réponse est :

```
Call:
survdiff(formula = Surv(time, status) ~ trt, data = diabetic)
```

```
      N Observed Expected (O-E)^2/E
trt=0 197 101 71.8 11.9
trt=1 197 54 83.2 10.3
      (O-E)^2/V
trt=0 22.2
trt=1 22.2
```

```
Chisq= 22.2 on 1 degrees of freedom, p= 2e-06
```

Comme la p-valeur est  $2e - 06$ , qui est bien inférieure à 0,05, la différence de temps avant la cécité entre les deux groupes de traitement est statistiquement significative. En outre, dans ce cas, le nombre de cécités était significativement plus élevé que prévu dans le groupe qui n'a pas reçu de traitement au laser ( $\text{trt}=0$ ), tandis que le nombre de cécités était plus faible que prévu dans le groupe qui a reçu un traitement au laser ( $\text{trt}=1$ ).

En conclusion, les résultats du test Log-Rank ont montré que le fait de recevoir ou non un traitement au laser a un effet significatif sur la courbe de survie des patients. Cela implique que le traitement au laser peut être une intervention efficace pour retarder la perte de vision chez les patients diabétiques. ////

## Exercice 1f

On se demande si d'autres variables dans le jeu de données pourraient être des facteurs confondants, et si l'on devrait stratifier le test du Log-Rank sur l'une de ces variables. À partir de statistiques descriptives, de graphiques

et/ou d'arguments adaptés au contexte de l'étude, discuter de laquelle des variables du jeu de données risque d'agir comme facteur confondant et reproduire le test du Log-Rank stratifié pour cette variable.

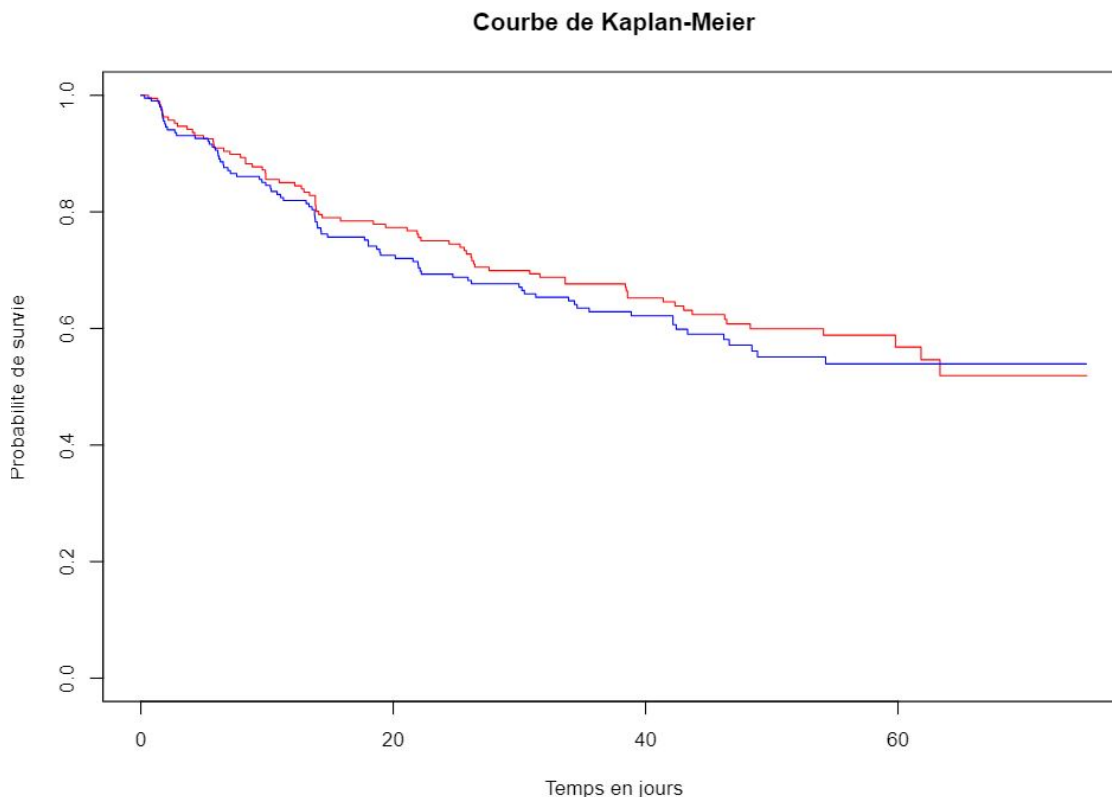
**Solution** Dans la vie quotidienne, on trouve que des personnes âgées ont plus grande probabilité de devenir aveugle. On soupçonne naturellement que l'âge soit un facteur confondant qui influence la durée jusqu'à la cécité. On utilise donc l'âge comme la variable à analyser. On va d'abord tracer les courbes de survie pour les deux groupes de l'âge. Notamment, on classe les patients en deux groupes, "Younger" et "Older", selon si l'âge du patient soit inférieur ou supérieur à l'âge médian. Pour ce faire, on commence par la graphique Kaplan-Meier pour les deux groupes, en utilisant le rouge pour représenter le groupe "Younger" et le bleu pour représenter le groupe "Older". Le graphique montre que la courbe rouge est plus élevée que la courbe bleue à la plupart des moments, ce qui signifie que la probabilité de survie est plus élevée pour le groupe "Younger" que pour le groupe "Older".

---

```
median_age <- median(diabetic$age, na.rm = TRUE)
diabetic$age_group <- ifelse(diabetic$age <= median_age, "younger", "older")
result.kmage <- survfit(Surv(time, status) ~ age_group, data = diabetic)
plot(result.kmage, main = 'Courbe de Kaplan-Meier', xlab = 'Temps en jours',
      ylab = 'Probabilite de survie', col = c("red", "blue"))
```

---

Le plot est :



Sur cette graphique on peut trouver que le groupe de patients "Younger" présente une probabilité de cécité plus faible que le groupe de patients "Older". Par conséquent, on conclut que l'âge est un facteur important. On effectue le test de Log-Rank stratifié pour les deux groupes d'âge.

---

```
survdifff(Surv(time, status) ~ trt + strata(age), data = diabetic)
```

---

Le résultat est :

Call:

```
survdifff(formula = Surv(time, status) ~ trt + strata(age), data = diabetic)
```

	N	Observed	Expected	(O-E)^2/E	(O-E)^2/V
trt=0	197	101	71.9	11.8	23.4
trt=1	197	54	83.1	10.2	23.4

Chisq= 23.4 on 1 degrees of freedom, p= 1e-06

Comme la p-valeur est  $1e-06$ , il y a une différence significative dans le délai de cécité entre les deux groupes de traitement. Même après stratification par âge, il existe toujours une différence significative dans le risque de cécité entre le groupe `trt=1` et le groupe de `trt=0`. On conclut que les patients du groupe de `trt=1` ont une probabilité de cécité plus faible que les patients du groupe `trt=0`. En conclusion, l'âge est un facteur important influençant la durée de la cécité et qu'il doit être pris en compte lors de l'élaboration des stratégies de traitement.

////

## Exercice 2a

Pour le jeu de données `rotterdam` dans la bibliothèque `survival` en R, on s'intéresse par l'association entre les variables explicatives `age`, `meno`, `sizes`, `grade`, `nodes` et `chemo`, et la variable de réponse le temps jusqu'à une récurrence du cancer au sein. Ajustez un modèle de Cox pour cela et rapporter les rapports de risque pour toutes les variables explicatives avec leurs intervalles de confiance (à 95% ?).

**Solution** L'ensemble des données comprenait 2982 patientes avec 1518 événements (récurrence du cancer du sein). La cohérence (concordance) du modèle était de 0,678, ce qui signifie que la précision de la prédiction du modèle était modérée.

Les résultats suivants montrent l'effet de chaque variable sur le risque de récurrence, exprimé en rapport de risque :

- **Âge (age)** : le rapport de risque était de 0,9859, indiquant une diminution légère (environ  $\frac{1}{0.9859} - 1 \approx 1,4\%$ ) du risque de récurrence pour chaque année d'âge supplémentaire. Cet effet est significatif (valeur  $p < 0,05$ ). L'intervalle de confiance à 95% est de 0,9792 à 0,9927.
- **Statut ménopausique (meno)** : le rapport de risque est de 1,1969, signifiant que les patientes ménopausées ont un risque de récurrence environ 20% plus élevé que les patientes non ménopausées. Cet effet est également significatif (valeur  $p < 0,05$ ). L'intervalle de confiance à 95% était compris entre 1,0069 et 1,4228.
- **Taille de la tumeur (taille 20-50 et taille >50)** : des tumeurs plus grandes (20 – 50mm et > 50mm) ont été associées à un risque accru de récurrence, avec des rapports de risques de 1,4462 et 1,8996 respectivement. Les intervalles de confiance sont respectivement de 1,2909 à 1,6202 et de 1,5988 à 2,2571. Ces effets sont significatifs (valeur  $p < 0,05$ ). Cela suggère que plus la tumeur est importante, plus son impact sur le risque de récurrence est grand, et que cet impact est relativement certain.
- **Grade pathologique (grade)** : le rapport de risque était de 1,4413, indiquant que les tumeurs de grade supérieur sont associées à un risque plus élevé de récurrence. Cet effet est aussi significatif (valeur  $p < 0,05$ ). L'intervalle de confiance à 95% est compris entre 1,2703 et 1,6353.
- **Nombre de ganglions lymphatiques atteints (nodes)** : le rapport de risque était de 1,0802, chaque ganglion supplémentaire atteint augmentant le risque de récurrence d'environ 8%. Cet effet est hautement significatif (valeur  $p < 0,001$ ). L'intervalle de confiance à 95% est compris entre 1,0706 et 1,0900.

- **Chimiothérapie (chemo)** : le rapport de risque était de 0,8902, indiquant un risque légèrement plus faible de récurrence chez les patients recevant une chimiothérapie. Toutefois, cette différence n'était pas statistiquement significative (valeur  $p = 0,0993$ ). L'intervalle de confiance à 95% est compris entre 0,7751 et 1,0222.

Ces résultats suggèrent que l'âge, le statut ménopausique, la taille de la tumeur, le classement pathologique et le nombre de ganglions lymphatiques atteints sont des facteurs prédictifs importants de la récurrence du cancer du sein.

---

```
#ex2
library(survival)
data(rotterdam)
attach(rotterdam)
#2a
resultat.cox <- coxph(Surv(rtime, recur) ~ age + meno + size + grade + nodes +
  chemo, data = rotterdam)
print(summary(resultat.cox))
```

---

La réponse est :

---

Call:

```
coxph(formula = Surv(rtime, recur) ~ age + meno + size + grade +
  nodes + chemo, data = rotterdam)
```

```
n= 2982, number of events= 1518
```

```
      coef exp(coef) se(coef) z Pr(>|z|)
age      -0.014194 0.985906 0.003481 -4.077 4.56e-05 ***
meno       0.179756 1.196925 0.088210 2.038 0.0416 *
size20-50 0.368943 1.446205 0.057965 6.365 1.95e-10 ***
size>50    0.641656 1.899624 0.087968 7.294 3.00e-13 ***
grade      0.365566 1.441329 0.064432 5.674 1.40e-08 ***
nodes      0.077162 1.080217 0.004580 16.847 < 2e-16 ***
chemo     -0.116355 0.890159 0.070594 -1.648 0.0993 .
```

---

```
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
      exp(coef) exp(-coef) lower .95 upper .95
age          0.9859    1.0143 0.9792 0.9927
meno         1.1969    0.8355 1.0069 1.4228
size20-50    1.4462    0.6915 1.2909 1.6202
size>50      1.8996    0.5264 1.5988 2.2571
grade        1.4413    0.6938 1.2703 1.6353
nodes        1.0802    0.9257 1.0706 1.0900
chemo        0.8902    1.1234 0.7751 1.0222
```

```
Concordance= 0.678 (se = 0.007 )
```

```
Likelihood ratio test= 468.6 on 7 df, p=<2e-16
```

```
Wald test            = 589.9 on 7 df, p=<2e-16
```

```
Score (logrank) test = 647.9 on 7 df, p=<2e-16
```

---

## Exercice 2b

Nommez l'hypothèse importante du modèle de Cox. Expliquez sans le faire comment on peut vérifier cette hypothèse.

**Solution** Le modèle de Cox fait une hypothèse que le rapport de risque entre les individus des groupes de traitements et du groupe de contrôle est constant pour tout temps. Cela signifie que l'ampleur relative du risque entre les individus ne change pas au fil du temps. Afin de vérifier cette hypothèse, il s'agit tout simplement d'observer si des courbes de fonctions hasard sont proportionnelles en tout temps ou non. Pour une méthode non graphique, on peut aussi utiliser les résidus de Schoenfeld. ///

## Exercice 2c

Décrivez dans vos mots les trois tests statistiques discutés en Sections 5.3.1, 5.3.2, 5.3.3 du livre "Applied Survival Analysis Using R" de Moore (2016) pour les coefficients dans le modèle de Cox. Comme dans la Section 5.3., dans un deuxième temps, discutez/comparez entre les trois tests leurs avantages et inconvénients. Cette réponse devrait consister en environ 8 à 12 lignes max au total.

**Solution** Écrire  $h_i(t_j)$  pour le hasard de patient  $i$  au temps d'échec  $t_j$ . Si l'hypothèse de proportionnalité est vraie, on peut écrire  $h_i(t_j) = h_0(t_j)\psi_i = h_0(t_j)e^{z_i\beta}$  où  $h_0$  est le hasard de référence,  $z_i \in \{0, 1\}$  selon si le patient  $i$  est dans le groupe de traitement ou le groupe de contrôle et  $\beta$  est le paramètre à estimer. La vraisemblance partielle est définie par  $L(\beta) = \prod_{k=1}^D \frac{\psi_i}{\sum_{l \in R_k} \psi_l}$  où  $D$  est le nombre d'échecs et  $R_k$  est l'ensemble des patients à risque au temps  $t_k$ . On teste l'hypothèse nulle  $H_0 : \beta = 0$ . Fixer un niveau  $\alpha \in (0, \frac{1}{2})$ . Écrire  $l = \ln(L)$ ,  $S = l'$ ,  $I = -l''$  et soit  $\hat{\beta} \in \arg \max_{\beta} L(\beta)$ . Le test de Wald rejette  $H_0$  si  $\left| \hat{\beta} \sqrt{I(\hat{\beta})} \right| > z_{\frac{\alpha}{2}}$ . Le test du score rejette  $H_0$  si  $\left| \frac{S(0)}{\sqrt{I(0)}} \right| > z_{\frac{\alpha}{2}}$ . Le test du ratio vraisemblance rejette  $H_0$  si  $\left| 2(l(\hat{\beta}) - l(0)) \right| > \chi_{\frac{\alpha}{2}, 1}^2$ . Si on ne peut pas obtenir  $\hat{\beta}$ , il faut utiliser le test du score. Supposons à partir de maintenant qu'on a obtenu  $\hat{\beta}$ . Le résultat du test du ratio vraisemblance a un avantage par rapport aux deux autres tests qu'il est invariant contre les transformations monotones. Si l'invariance n'est pas nécessaire, alors le test de Wald est plus commun et plus facile à calculer. ///

## Exercice 3a

Pour le jeu de données `datasurv.txt`, écrire  $t_1 < \dots < t_D$  pour tous les temps distincts d'échec,  $n_i$  pour le nombre de sujets à risque au temps  $t_i$  et  $d_i$  pour le nombre d'échecs au temps  $t_i$ . Comme les patients à risque inclut ceux qui ont eu un échec, on a  $n_i \geq d_i \geq 1$  pour tout  $i$ . On définit  $q_i = \frac{d_i}{n_i}$  et  $S(t) = \prod_{t_i \leq t} (1 - q_i)$  pour tout  $t \in \{1, \dots, t_D\}$ . Produisez un tableau de valeurs de  $t_i, n_i, d_i, q_i$  et  $S(t_i)$ .

**Solution** On produit le tableau automatiquement en R :

```
library(survival)
donnee <- read.delim("Desktop/u de m/STT6510/projet2/datasurv.txt", header =
  TRUE, sep = " ")
#donnee <- read.delim("datasurv.txt", header = TRUE)
attach(donnee)

#3a
#surv_objet <- Surv(time_days, event)
```



```
#print(grepl("\\+$", surv_objet[1]))

max_jour <- 500
t <- seq(max_jour)
n <- rep(0, max_jour)
d <- rep(0, max_jour)
q <- rep(0, max_jour)
S <- rep(0, max_jour)
n[1] <- sum(time_days >= 1)
S[1] <- 1

for (i in 2:max_jour) {
  n[i] <- sum(time_days >= i)
  if (i %in% time_days) {
    d[i] <- sum(event[which(time_days == i)])
  }
  q[i] <- d[i] / n[i]
  S[i] <- S[i-1] * (1 - q[i])
}

tableau_a_remettre <- data.frame(
  cbind(t[sort(time_days[event == 1])],
        n[sort(time_days[event == 1])],
        d[sort(time_days[event == 1])],
        q[sort(time_days[event == 1])],
        1 - q[sort(time_days[event == 1])],
        S[sort(time_days[event == 1])])
names(tableau_a_remettre) <- c("t", "n", "d", "q", "1-q", "S")
print(tableau_a_remettre)
```

Le résultat est :

$t$	$n$	$d$	$q$	$1 - q$	$S$
196	18	1	0.0556	0.9444	0.9444
258	13	1	0.0769	0.9231	0.8718
262	12	1	0.0833	0.9167	0.7991
375	5	1	0.2000	0.8000	0.6393
377	3	1	0.3333	0.6667	0.4262
409	2	1	0.5000	0.5000	0.2131

Table 1: Résultats de l'analyse de survie

## Exercice 3bc

En R, réalisez un graphique de  $q$  en fonction de  $t$  pour  $t \in \{1, \dots, 500\}$ . Ici,  $q(t) = q_i$  si  $t \in \{1, \dots, t_D\}$  et  $q(t) = 0$ . Sinon, Produisez en particulière le même graphique pour  $t \in \{200, \dots, 260\}$ .

Finalement, lissez cela utilisant le noyau d'Epanechnikov avec un paramètre de lissage 5. Plus précisément, on définit le noyau d'Epanechnikov comme  $K(u) = \frac{3}{4}(1 - u^2)\chi_{[-1,1]}(u)$  pour tout  $u \in \mathbb{R}$  et on définit la fonction de lissage comme :

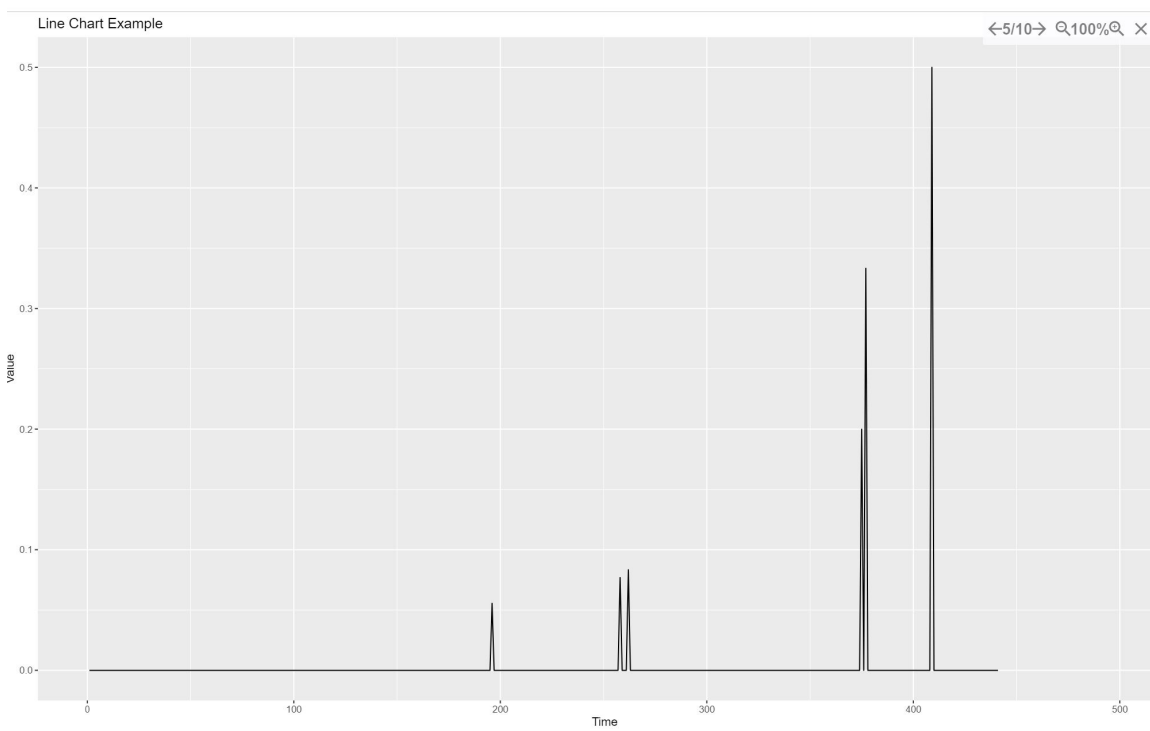
$$\hat{h}(t) = \frac{1}{b} \sum_{i=1}^D K\left(\frac{t-t_i}{b}\right) q_i$$

Ici  $b = 5$ . Produisez un graphique de  $\hat{h}$  en fonction de  $t$  pour  $t \in \{200, \dots, 260\}$ . Vérifiez les résultats en main pour  $t = \{200, 210, 220, 230, 240, 250\}$ .

**Solution** On procède encore en R :

```
library(ggplot2)

p <- ggplot(data.frame(t = t, q = q), aes(x = t, y = q)) +
  geom_line() + labs(title = "Line Chart Example", x = "Time", y = "Value")
show(p)
```



Selon la définition de  $K$ , on a  $K\left(\frac{t-t_i}{b}\right) = 0$  si  $|t_i - t| > b$ .

$$\hat{h}(200) = \frac{1}{5} \times K\left(\frac{200-196}{5}\right) \times \frac{1}{18} = \frac{1}{5} \times K(0.8) \times \frac{1}{18} = \frac{1}{5} \times (1-0.8^2) \times \frac{3}{4} \times \frac{1}{18} = \frac{3}{1000}$$

On a que  $\hat{h}(210) = \hat{h}(220) = \hat{h}(230) = \hat{h}(240) = \hat{h}(250) = 0$ . ///

## Exercice 3d

```
#3d
debut <- 200
fin <- 260
```

```

p <- ggplot(data.frame(t = t[debut:fin], q = q[debut:fin]), aes(x = t, y = q)) +
  geom_line() + labs(title = "Line Chart Example", x = "Time", y = "Value")
show(p)

epanechnikov <- function(u) {
  if (abs(u) <= 1) {
    return(0.75 * (1 - u^2))
  } else {
    return(0)
  }
}

b <- 5
D <- length(t[sort(time_days[event == 1])])
lisse <- rep(0, fin - debut)
for (i in debut:fin) {
  diff_seq <- (i - t[sort(time_days[event == 1])])/b
  epan <- rep(0, D)
  for (j in 1:D) {
    epan[j] <- epanechnikov(diff_seq[j])
  }
  lisse[i - debut + 1] <- (1/b)*sum(q[sort(time_days[event == 1])]) * epan
}

p <- ggplot(data.frame(t = t[debut:fin], lisse = lisse), aes(x = t, y = lisse)) +
  geom_line() + labs(title = "Line Chart Example", x = "Time", y = "Value")
show(p)

```

