TD – Analyse de survie

Laurent Gardes

Exercice 1 *On suppose que la fonction de survie de T est celle d'une loi de Weibull de paramètres $\alpha > 0$ et $\beta > 0$ i.e., $S(t) = \exp\left(-\alpha t^{\beta}\right)$ pour tout $t \geq 0$. Calculer la fonction de hasard de T en fonction de α et β. Que remarque t'on lorsque $\beta = 1$?*

Exercice 2 *Soit T une durée de survie de loi absolument continue et telle que pour tout $t \geq 0$ on ait $\lambda(t) = \lambda > 0$. Quelle est la loi suivie par T ?*

Exercice 3 *Soit T une variable alátoire de loi absolument continue. Démontrer que pour tout $t > 0$,*

$$\ln S(t) = -\int_0^t \lambda(x)dx.$$

Exercice 4 *Soit T une variable aléatoire discrète prenant ses valeurs dans l'ensemble $\{t_1, t_2, \cdots\}$. On note $\{t_{(1)} < t_{(2)} < \cdots\}$ les valeurs rangées par ordre croissant et on pose $p_j := \mathbb{P}(T = t_{(j)})$ pour tout $j \in \mathbb{N} \setminus \{0\}$.*

i) Montrer que pour tout $j \in \mathbb{N} \setminus \{0\}$,

$$\lambda(t_{(j)}) =: \lambda_j = p_j \Big/ \sum_{k \geq j} p_k \in [0, 1].$$

ii) Montrer que $\lambda_1 = p_1$ et que pour tout $j \geq 2$,

$$p_j = \lambda_j \prod_{k=1}^{j-1} (1 - \lambda_k).$$

Exercice 5 *Soit T une durée de survie. On suppose que $\mathbb{E}(T) < \infty$.*

1) Si T est une variable aléatoire discrète prenant ses valeurs dans l'ensemble $\{t_1, t_2, \cdots\}$, montrer que

$$\mathbb{E}(T) = \int_0^{\infty} S(t)dt.$$

2) Même question si T est une variable aléatoire de loi absolument continue.

Exercice 6 *Soit T une variable aléatoire positive absolument continue de fonction de survie $S(\cdot)$ strictement décroissante. Quelle est la loi de la variable aléatoire $\Lambda(T)$?*

Exercice 7 *On dispose des réalisations $x_1, \cdots, x_n$ de n variables aléatoires indépendantes $X_1, \cdots, X_n$ et de loi commune $\mathbb{P}_X \in \{\mathbb{P}_\theta;\ \ \theta \in \Theta\}$. En utilisant la méthode du maximum de vraisemblance, calculer l'estimateur du maximum de vraisemblance de θ dans les deux cas suivants :*

i) Pour tout $t \geq 0$, $\mathbb{P}_\theta(]t, \infty[) = \exp(-\theta t)$, *avec* $\theta \in \Theta =]0, \infty[$.

ii) Pour tout $k \in \mathbb{N}$, $\mathbb{P}_\theta(\{k\}) = \exp(-\theta)\lambda^k/k!$ *avec* $\theta \in \Theta =]0, \infty[$.

Exercice 8 *On suppose que le couple aléatoire* $(T^\star, \Delta)$ *est modélisé par une censure à droite. Montrer que* $\mathbb{P}(T^\star > x) \leq \mathbb{P}(T > x)$. *Comment interprétez-vous ce résultat ?*

Exercice 9 *On se place sous un modèle avec censure aléatoire à droite lorsque la loi du couple* (T, C) *est discrète à support fini. En notant* $t^\star_{(1)} < \cdots < t^\star_{(m)} < t^\star_{(m+1)} = \infty$ *les valeurs de* T, *on rappelle que l'estimateur de Kaplan-Meier des probabilités* $p_k := \mathbb{P}[T = t^\star_{(k)}]$, $k = 1, \cdots, m+1$ *est donné par* $\widehat{p}_{n,1} = O_1/N_1$ *et, pour tout* $k = 2, \cdots, m+1$, *par*

$$\widehat{p}_{n,k} = \frac{O_j}{N_j} \prod_{k=1}^{j-1} \left(1 - \frac{O_k}{N_k}\right).$$

Vérifier que

$$\sum_{j=1}^{m+1} \widehat{p}_{n,j} = 1.$$

Vous pouvez dans un premier temps démontrer que pour $j = 2, \cdots, m$,

$$\frac{O_j}{N_j} \prod_{k=1}^{j-1} \left(1 - \frac{O_k}{N_k}\right) = \prod_{k=1}^{j-1} \left(1 - \frac{O_k}{N_k}\right) - \prod_{k=1}^{j} \left(1 - \frac{O_k}{N_k}\right).$$

Exercice 10 *Montrer qu'en cas d'absence de censure, l'estimateur de Kaplan-Meier coincide avec l'estimateur empirique.*

Exercice 11 *Proposer une autre méthode permettant de retrouver l'expression de l'estimateur de Kaplan-Meier.*

Exercice 12 *Pour* $n = 10$ *individus on dispose des données suivantes.*

0.5	0.8	0.8	1	1.3	1.3^+	1.3^+	1.4	1.4^+	1.7^+

Ces valeurs correspondent aux $n = 10$ *observations* $\{t^\star_1, \cdots, t^\star_n\}$ *de la variable aléatoire* $T^\star = \min(T, C)$. *L'exposant* + *signifie que la valeur a été censurée à droite.*

1) *Donner les valeurs observées* $d_1, \cdots, d_n$ *de la variable aléatoire* $\Delta = \mathbb{I}_{T \leq C}$ *indiquant la présence d'une censure.*
2) *Donner la valeur* m *ainsi que l'ensemble* $\{t^\star_{(1)} < \cdots < t^\star_{(m)}\}$ *des observations distinctes et ordonnées.*
3) *Pour tout* $j = 1, \cdots, m$*, calculer la valeur* $\widehat{S}_n^{(KM)}(t^\star_{(j)} \mid (\mathbf{t}^\star, \mathbf{d}))$.

Remarque – Pour faire l'exercice précédent sur **R**, il faut charger le package `survival`. La procédure est décrite ci-dessous.

```
install.packages("survival")
library(survival)
Tet<-c(0.5,0.8,0.8,1,1.3,1.3,1.3,1.4,1.4,1.7)
Delta<-c(1,1,1,1,1,0,0,1,0,0)
data<-Surv(time=Tet,event=Delta)
fit<-survfit(data~1)
summary(fit)
```

On peut récupérer les valeurs de l'estimateur de Kaplan-Meier de la façon suivante.

```
fit$surv
```

Exercice 13 *On rappelle que l'estimateur de Greenwood de la variance asymptotique de l'estimateur de Kaplan-Meier est*

$$\mathcal{V}_n^{(G)}(x) := \left[\widehat{S}_n^{(KM)}(x)\right]^2 \sum_{j:T^\star_{(j)} \leq x} \frac{O_j(\mathbf{T}^\star, \boldsymbol{\Delta})}{N_j(\mathbf{T}^\star)(N_j(\mathbf{T}^\star) - O_j(\mathbf{T}^\star, \boldsymbol{\Delta}))}$$

Expliquer comment a été obtenue son expression.

Exercice 14 *Montrer qu'en absence de censure,*

$$\mathcal{V}_n^{(G)}(x) = \frac{1}{n}\widehat{S}_n(x)\left[1 - \widehat{S}_n(x)\right],$$

où $\widehat{S}_n(\cdot)$ *est la fonction de survie empirique.*

Exercice 15 *En reprenant les données de l'exercice 12, calculer les valeurs de l'estimateur de Nelson-Aalen (de la fonction de hasard cumulé ainsi que de la fonction de survie).*

Remarque – On peut obtenir les valeurs de l'estimateur de la fonction de hasard cumulé à l'aide de **R** de la façon suivante (à faire à la suite des commandes **R** données après l'exercice 12).

```
fit<-survfit(data~1)
fit$cumhaz
```

Exercice 16 *Montrer que*

$$\begin{aligned}\mathrm{LR}(\mathbf{t}^\star,\mathbf{d}) &= \frac{1}{\mathcal{V}_n^{(LR)}}\left[\sum_{k=1}^{m}\left(O_k^{(2)}-N_k^{(2)}\frac{O_k}{N_k}\right)\right]^2\\ &= \frac{1}{\mathcal{V}_n^{(LR)}}\left[\sum_{k=1}^{m}\frac{N_k^{(1)}N_k^{(2)}}{N_k}\left(\frac{O_k^{(1)}}{N_k^{(1)}}-\frac{O_k^{(2)}}{N_k^{(2)}}\right)\right]^2\end{aligned}$$

Exercice 17 *On observe les durées de survie dans 2 groupes d'individus distincts (et donc indépendants). On dispose des données suivantes.*

Groupe 1	0.5	0.8	0.8	1.0	1.3	1.3^+	1.3^+	1.4	1.4^+	1.7^+
Groupe 2	0.5	0.6	0.8	0.9	0.9^+	1.0	1.0	1.2^+	1.4	

TABLE 1 – Observations pour les deux groupes. L'exposant + signifie que la valeur a été censurée à droite.

1) *Donner la valeur m ainsi que l'ensemble $\{t^\star_{(1)} < \cdots < t^\star_{(m)}\}$ des observations distinctes et ordonnées des deux groupes réunis.*
2) *Compléter les tableaux suivants :*

k	1	2	3	4	5	6	7	8	9
$g=1$	10	9	9	?	7	6	6	3	1
$g=2$	9	8	7	?	4	2	1	1	0

TABLE 2 – Valeurs $N_k^{(g)}$ pour $g \in \{1,2\}$ et $k \in \{1,\cdots,m\}$.

k	1	2	3	4	5	6	7	8	9
$g=1$	1	0	2	?	1	0	1	1	0
$g=2$	1	1	1	?	2	0	0	1	0

TABLE 3 – Valeurs $O_k^{(g)}$ pour $g \in \{1,2\}$ et $k \in \{1,\cdots,m\}$.

k	1	2	3	4	5	6	7	8	9
	1.05	0.53	1.69	?	1.91	0.00	0.86	1.50	?

TABLE 4 – Valeurs observées de $E_k(\mathbf{T}^\star,\boldsymbol{\Delta})$ pour $k \in \{1,\cdots,m\}$.

3) Pour effectuer ce test sur le logiciel **R** *on procède de la façon suivante :*

```
Tet1<-c(0.5,0.8,0.8,1,1.3,1.3,1.3,1.4,1.4,1.7)
Delta1<-c(1,1,1,1,1,0,0,1,0,0)
Tet2<-c(0.5,0.6,0.8,0.9,0.9,1.0,1.0,1.2,1.4)
Delta2<-c(1,1,1,1,0,1,1,0,1)
Tet<-c(Tet1,Tet2)
Delta<-c(Delta1,Delta2)
groupe<-c(rep(1,10),rep(2,9))
res<-survdiff(Surv(time=Tet,event=Delta)~groupe)
```

Le résultat est donné sous la forme ci-dessous.

```
Call:
survdiff(formula = Surv(time = Tet, event = Delta) ~ groupe)

          N Observed Expected (O-E)^2/E (O-E)^2/V
groupe=1 10        6     8.07     0.533       1.7
groupe=2  9        7     4.93     0.873       1.7

 Chisq= 1.7  on 1 degrees of freedom, p= 0.2
```

Expliquer comment ont été obtenues les valeurs ci-dessus. Quelle est la conclusion du test ?

Exercice 18 *On se place sous le modèle de censure aléatoire à droite en présence d'une covariable X. Dans le cas où la loi du vecteur aléatoire (T, C) est absolument continue, montrer que la vraisemblance du modèle s'écrit*

$$K \prod_{i=1}^{n} [\lambda(t_i^\star \mid x_i)]^{d_i} \exp\left(- \int_0^{t_i^\star} \lambda(z \mid x_i) dz \right),$$

où K est une constante ne dépendant que des lois de C et X.

Exercice 19 *On se place sous le modèle de censure aléatoire à droite en présence d'une covariable X. Dans le cas où la loi conditionnelle de T sachant $X = x$ est absolument continue, montrer que la fonction de survie conditionnelle $S(\cdot \mid X)$ est à risques proportionnels i.e.,*

$$S(t \mid x) = [S_0(t)]^{\rho(x;\theta)}.$$

si et seulement si

$$\lambda(t \mid x) = \lambda_0(t) \times \rho(x; \theta).$$

Exercice 20 *Démontrer le résultat suivant. Soit T une variable aléatoire positive prenant les valeurs $t^\star_{(1)} < t^\star_{(2)} < \cdots$ et soit $\rho \geq 0$. On a $S(t) = [S_0(t)]^\rho$ pour tout $t > 0$ si et seulement si $(1-\lambda_i) = (1-\lambda_{i,0})^\rho$ pour tout $i \geq 1$ où $\lambda_i = \lambda(t^\star_{(i)})$ et $\lambda_{i,0} = \lambda_0(t^\star_{(i)})$.*

Exercice 21 *Soit la variable aléatoire $T^\star := \min(T, C)$ où T et C sont des variables aléatoires positives et indépendantes de loi discrète à valeurs dans $\mathbb{N}$. En notant $\Lambda(\cdot)$ la fonction de hasard cumulé de T on a $\mathbb{E}[\Lambda(T^\star)] = \mathbb{P}(T \leq C)$.*

Exercice 22 *Pour le modèle de Cox, les résidus de martingale des observations $\{(t^\star_i, d_i, x_i);\ i = 1, \cdots, n\}$ sont donnés par*

$$\mathrm{MR}_i := d_i - \exp\left(\widehat{\beta}_n^\top x_i\right) \sum_{j: t^\star_{(j)} \leq t^\star_i} \widehat{\lambda}_{0,j},$$

les estimateurs $\widehat{\beta}_n$ et $\widehat{\lambda}_{0,j}$ étant données dans le cours. Démontrer que, s'il n'y a pas d'ex-œquo

$$\sum_{i=1}^{n} \mathrm{MR}_i = 0.$$

Exercice 23 (Illustration du modèle de Cox avec R)

1) Simuler un échantillon de taille $n = 200$ du triplet $(T^\star, \Delta, X)$ selon le modèle suivant : X suit une loi $\mathcal{N}(0,1)$, $T^\star := \min(T, C)$ où T et C sont deux variables aléatoires indépendantes conditionnellement à X et telles que pour tout $x \in \mathbb{R}$, la loi conditionnelle de T sachant $X = x$ est une loi exponentielle de paramètre $\lambda_0 \exp(\beta x)$ et la loi conditionnelle de C sachant $X = x$ est une loi exponentielle de paramètre $\lambda_1 \exp(\beta x)$ avec $\lambda_0 > 0$, $\lambda_1 > 0$ et $\beta \in \mathbb{R}$ (prendre par exemple $\lambda_0 = 1$, $\lambda_1 = 1/2$ et $\beta = 2$. Correction −

```
# Taille de l'echantillon #
n=200
# Choix des parametres #
beta<-2
lambda0<-1
lambda1<-0.5
# Simulation de X #
```

```
X=rnorm(n)
# Simulation des v.a. T et C #
U=runif(n)
V=runif(n)
Tps=exp(-beta*X)/lambda0*log(1/U)
C=exp(-beta*X)/lambda1*log(1/V)
# Calcul de T etoile #
Tet=c()
Tet[which(Tps<=C)]=Tps[which(Tps<=C)]
Tet[which(Tps>C)]=C[which(Tps>C)]
# Calcul de Delta #
Delta=rep(1,n)
Delta[which(Tps>C)]=0
```

2) Le vecteur $\mathbf{t}^\star$ *regroupant les* n *observations de* $T^\star$ *contient-il des ex-æquo ?*

Correction – En théorie, les lois conditionnelles de T *et* C *étant absolument continues, il n'y a pas d'ex-æquo. Pour le vérifier avec* **R***, on utilise la fonction* `unique()`

```
> Tetunique=unique(Tet)
> length(Tetunique)
[1] 200
```

Le vecteur `Tetunique` *regroupe les éléments distincts du vecteur* `Tet`*. La taille de ce vecteur est ici égale à* $n = 200$*, il n'y a donc pas d'ex-æquo.*

3) Ajuster le modèle de Cox à cet échantillon. Interpréter les résultats.

Correction –

```
> install.packages("survival")
> library(survival)
> data<-Surv(time=Tet,event=Delta)
> coxph(data~X)
Call:
coxph(formula = data ~ X)

   coef exp(coef) se(coef)     z      p
X 1.990     7.315    0.177 11.24 <2e-16

Likelihood ratio test=162.2  on 1 df, p=< 2.2e-16
n= 200, number of events= 135
```

La valeur observée de l'estimateur du maximum de vraisemblance de β

est de $\hat{\beta} = 1.990$. *La valeur* $\exp(\hat{\beta}) = 7.315$ *s'interprète de la façon suivante : si la valeur de la covariable* X *augmente de* 1, *le risque instantané est multiplié par* 7.315. *La valeur* `se(coeff)` *est l'écart-type estimé de* $\hat{\beta} = 1.990$ *et* `z` *est la valeur observée de la statistique de Wald (*`z`$=$`coef`$/$`se(coef)`*). Comme d'habitude* `p` *est la p-valeur du test de Wald (*$H_0 : \beta = 0$*). La valeur* 2×10^{-16} *est très proche de 0 ce qui nous suggère que* $\hat{\beta}$ *est significativement différent de 0 (on n'accepte pas* H_0*). Enfin,* `number of events` *est le nombre d'observations non censurées.*