NOM : _____ **Prénom :** _____ **Note :** / 20

EXAMEN — STATISTIQUES DES DUREES DE VIE — DURÉE 2h

- ☞ Aucun document n'est autorisé, la calculatrice est admise.
- ☞ Rédigez directement vos réponses sur l'énoncé.
- ☞ Attention le QCM est en mode « concours », i.e., +1 si réponse juste **avec** justification, 0 si non réponse, réponse partielle ou fasse ou **sans** justification. Seules les réponses cochées devront être justifiées.

Exercice 1. Questions de cours (QCM) **(4 points)**

a) Soit $\hat{\theta}$ un estimateur pour $\theta_0 \in (0, \infty)$ tel que

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{\mathrm{d}} N(0, \sigma^2), \qquad n \to \infty.$$

On supposera de plus que $\hat{\theta}$ est positif p.s.
Posons $\hat{\psi} = \log \hat{\theta}$ et notons $\hat{\sigma}^2$ un estimateur de σ^2. Parmi les propositions suivantes, trouvez celle(s) correspondant à un intervalle de confiance à 95% pour $\psi_0 = \log \theta_0$.

- ☐ $\hat{\psi} \pm 1.96\hat{\sigma}$
- ☐ $\hat{\psi} \pm 1.96\hat{\sigma}/\sqrt{n}$
- ☐ $\hat{\psi} \pm 1.96 \exp(-\hat{\psi})\hat{\sigma}/\sqrt{n}$
- ☐ $\hat{\psi} \pm 1.96\hat{\sigma}/(\hat{\theta}\sqrt{n})$

b) Soient $\hat{S}(t)$ l'estimation par Kaplan–Meier de la fonction de survie $S(t)$ et $t_1 < t_2 < \cdots < t_n$ des durées de survie observées. Au passage rappelons que cet estimateur est donné par

$$\hat{S}(t) = \prod_{j\,:\,T_j \leq t} \left\{ 1 - \frac{d(T_j)}{r(T_j)} \right\},$$

où pour tout $t \geq 0$,

$$d(t) = \text{nombre de décès à l'instant } t$$
$$r(t) = \text{nombre d'individus à risque à l'instant } t.$$

Soit $t > t_n$ fixé. Parmi les propositions suivantes, laquelle/lesquelles est/sont correcte(s) ?

☐ si la plus grande observation est un décès alors $\hat{S}(t) = 0$

☐ On a toujours $\hat{S}(t) > 0$

☐ si la plus grande observation est censurée alors $\hat{S}(t) > 0$

☐ si la plus grande observation est censurée alors $\hat{S}(t) = 0$

c) On s'intéresse aux données `ovarian` du package `survival` qui regroupent les données de guérison du cancer des ovaires selon deux traitements. Plus précisément, ces données contiennent les variables suivantes

futime la durée de guérison ;

fustat la variable codant la censure (à droite) ;

rx la variable indiquant le traitement reçu, i.e., 1 ou 2.

A partir de la sortie `R` suivante, dire quelles sont les affirmations correctes (on utilisera les niveaux standards).

```
> summary(ovarian)

     futime          fustat         rx
 Min.   :  59.0   Mode :logical   1:13
 1st Qu.: 368.0   FALSE:14        2:13
 Median : 476.0   TRUE :12
 Mean   : 599.5
 3rd Qu.: 794.8
 Max.   :1227.0

> survdiff(Surv(futime, fustat) ~ rx,data=ovarian)

Call:
survdiff(formula = Surv(futime, fustat) ~ rx, data = ovarian)

      N Observed Expected (O-E)^2/E (O-E)^2/V
rx=1 13        7     5.23     0.596      1.06
rx=2 13        5     6.77     0.461      1.06

 Chisq= 1.1  on 1 degrees of freedom, p= 0.303
```

☐ Un traitement est plus efficace

☐ Les deux traitements ont la même performance

☐ L'utilisation de la fonction `survdiff` est une absurdité scientifique

☐ L'utilisation de la fonction `survdiff` est totalement justifiée

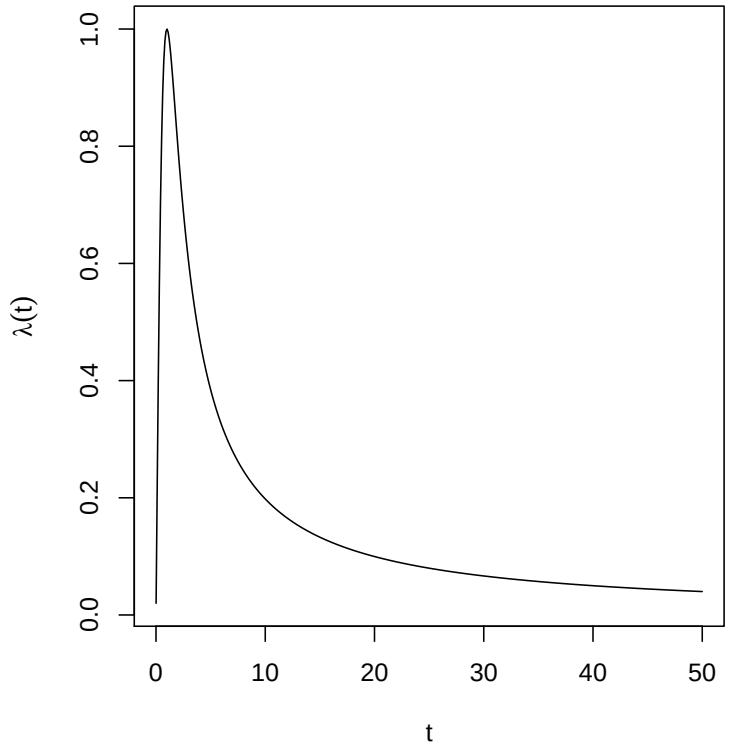

FIGURE 1 – Les médecins pensent que le taux de panne pour une maladie donnée devrait avoir cette forme.

d) A partir de la Figure 1, quels choix paramétriques pour le taux de panne feriez vous ?

- ☐ J'utiliserai un modèle de Weibull
- ☐ J'utiliserai un modèle Gamma
- ☐ J'utiliserai un modèle log-logistique
- ☐ J'utiliserai un modèle de Cox

Exercice 2. Pour voir si l'on a compris... **(8 points)**
Voici quelques rappels. L'estimateur de Kaplan–Meier de $S(t)$ ainsi que l'estimateur de sa variance sont donnés par

$$\hat{S}(t) = \prod_{j:\, T_j \le t} \left\{ 1 - \frac{d(T_j)}{r(T_j)} \right\}, \qquad \widehat{\mathrm{Var}}\{\hat{S}(t)\} \approx \hat{S}(t)^2 \sum_{j:\, T_j \le t} \frac{d(T_j)}{r(T_j)\{r(T_j) - d(T_j)\}},$$

où pour tout $t \ge 0$,

$$d(t) = \text{nombre de décès à l'instant } t, \quad r(t) = \text{nombre d'individus à risque à l'instant } t.$$

a) Montrez que la fonction

$$\begin{aligned} f\colon (0,1) &\longrightarrow \mathbb{R} \\ x &\longmapsto \log(-\log x) \end{aligned}$$

est bijective. **(1 point)**

b) Soit $\hat{\theta}$ un estimateur asymptotiquement Gaussien. Donnez une approximation pour $\mathrm{Var}\{f(\hat{\theta})\}$. **(2 points)**

c) En utilisant la fonction f, construisez un intervalle de confiance à 95% pour $S(t)$, qui sera nécessairement inclus dans l'intervalle $[0, 1]$. **(2 points)**

d) Voici un jeu de données contenant 10 durées (en minutes) avant l'apparition des premiers signes de sommeil d'étudiants du cours de Mathieu de Vendredi 8h.

$$1 \quad 3 \quad 4^+ \quad 5 \quad 7^+ \quad 8 \quad 9 \quad 10^+ \quad 11 \quad 13^+$$

i) Donnez, via Kaplan–Meier, une estimation de la fonction de survie (que vous résumerez sous forme de tableau) ; **(1 point)**

ii) Donnez, via Kaplan–Meier, une estimation pour $S(6)$ ainsi qu'un intervalle de confiance à 95%—en utilisant la question c). **(2 points)**

Exercice 3. Infection par cathéter **(8 points)**

Nous étudions la durée avant infection lors de la pose d'un cathéter pour des patients devant subir une dialyse. Les données sont sous le format suivant :

patient Numéro d'identification unique du patient ;

time Durée avant infection ;

status Variable codant la censure ;

age Âge du patient lors de la pose du cathéter ;

sex Variable codant le sexe du patient (1 : Homme, 2 : Femme) ;

disease Type d'infection (GN : néphrite glomérulaire, AN : néphrite aiguë, PKD : polykystose rénale, Other : autre)

```
> fit <- coxph(Surv(time, status) ~ age + factor(sex) + factor(disease),
+ data = kidney)
```

a) Précisez quel modèle est ajusté ici—le seul nom du modèle ne suffit pas, il faut également son expression mathématique ! **(1 point)**

b) Complétez la sortie R suivante en remplissant la colonne « z »—avec deux chiffres après la virgule. **(1 point)**

```
> fit

Call:
coxph(formula = Surv(time, status) ~ age + factor(sex) + factor(disease),
    data = kidney)


                       coef exp(coef) se(coef)         z          p
age                 0.00318   1.00319  0.01115                0.775
factor(sex)2       -1.48314   0.22692  0.35823
factor(disease)GN   0.08796   1.09194  0.40637                0.829
factor(disease)AN   0.35079   1.42020  0.39972                0.380
factor(disease)PKD -1.43111   0.23904  0.63111                0.023

Likelihood ratio = 17.6  on 5 df, p =
n= 76, number of events= 58
```

c) Vous aurez peut-être remarqué que la ligne `Likelihood ratio test=17.6 on 5 df, p =` en bas de la sortie précédente est incomplète. A quelle technique statistique correspond cette ligne ? (Précisez quel est son objectif). Complétez ainsi cette ligne incomplète en vous aidant des sorties `R` suivantes et conclure **(2 points)**

```
> pnorm(17.6 / 58)

[1] 0.6192259

> 1 - pnorm(17.6 / 58)

[1] 0.3807741

> pchisq(17.6, 5)

[1] 0.9965082

> 1 - pchisq(17.6, 5)

[1] 0.003491841
```

d) Vous aurez peut-être remarqué que dans la colonne `p`, la deuxième valeur est manquante. A quelle technique statistique correspond cette case ? (Précisez aussi son objectif). Complétez ainsi cette case incomplète en vous aidant des sorties `R` suivantes **(2 points)**

```
> pnorm(-1.48314/0.35823)
[1] 1.735097e-05
> 2 * pnorm(-1.48314/0.35823)
[1] 3.470194e-05
> 1 - pnorm(0.22692 / 0.35823)
[1] 0.2632206
> 2 * (1 - pnorm(0.22692 / 0.35823))
[1] 0.5264413
```

e) Selon vous, la variable sexe a-t-elle un effet significatif ? Et si oui, le fait d'être un homme, augmente-t-il le risque ou au contraire le diminue-t-il ? **(2 points)**