Le logiciel `R` ne possède pas de base des fonctions pour faire l'analyse de survie. Le package **survival** fournit pas mal de fonctions pour cela. Avant de pouvoir l'utiliser il faut déjà l'installer (si ce n'est pas déjà le cas).

```
> install.packages("survival")
```

Une fois installé, il faut charger cette librairie (si l'on veut utiliser les fonctionnalités supplémentaires offertes par ce package).

```
> library(survival)
```

Puisque nous parlons d'analyse de survie, nous allons très probablement travailler avec des données censurées. Comment le package **survival** gère-t-il la censure ?

L'idée repose sur une fonction qui crée un objet de classe **Surv**. Il faudra donc créer des objets de cette classe afin de pouvoir faire nos analyses de survie.

La création d'objets de la classe **Surv** se fait via la fonction `Surv`.

Exercice 1 (Savoir gérer la censure (à droite))**.** a) Allez voir la documentation de la fonction `Surv` pour comprendre son fonctionnement.

b) Générez un échantillon fictif de taille 100 selon une loi exponentielle où certaines observations seront traitées comme censurées à droite.
Pour la question b), vous afficherez l'échantillon ainsi créé.

Nous allons travailler sur des données réelles. Pour cela

1. installez le package `KMsurv` qui contient quelques jeux de données, comme ça on va pouvoir s'amuser ;
2. chargez cette librairie.

Exercice 2 (Cancer de la langue)**.**
Nous allons travailler sur des données de cancer de la langue...

a) Importez le jeu de données `tongue` et renseignez vous sur ce dernier.

b) Lisez l'aide de la fonction `survfit` afin de savoir comment obtenir une estimation de la fonction de survie via Kaplan–Meier.

c) Estimez cette dernière pour le groupe 1 (haploïde) puis pour le groupe 2 (diploïde).

d) Si lors de la question précédente vous avez fait appel deux fois à la fonction `survfit`, essayez d'utiliser une formule R afin de ne l'appeler qu'une seule fois ;-)

e) Que renvoient les fonctions `summary` et `print` lorsque vous les appliqués aux estimations obtenues lors de la question précedente. Et la fonction `plot` ?

Exercice 3 (Intervalles de confiance)**.**
Nous savons qu'un intervalle de confiance à $(1-\alpha)\%$ pour un paramètre $\theta \in \mathbb{R}$ et dont l'estimateur $\hat{\theta}$ est asymptotiquement normal est donné par

$$\left[\hat{\theta} - z_{1-\alpha/2}\sqrt{\mathrm{Var}(\hat{\theta})}, \hat{\theta} - z_{1-\alpha/2}\sqrt{\mathrm{Var}(\hat{\theta})}\right].$$

a) Rappelez l'expression d'un intervalle de confiance pour $S(t)$ via l'estimateur de Kaplan–Meier // Nelson–Aalen.

b) Cet intervalle de confiance est-il toujours compris dans l'intervalle $[0, 1]$?

c) En remarquant que pour tout $x \in \mathbb{R}$, $\exp\{-\exp(x)\} \in [0, 1]$, trouvez une approche permettant de construire un intervalle de confiance toujours compris dans $[0, 1]$.

d) Allez jeter un oeil à l'argument `conf.type` de la fonction `survfit` et essayez de comprendre son utilité.

e) A quoi sert l'option `conf.type = "log-log"` de la fonction `survfit` ? Utilisez cette option sur les données de cancer de la langue et comparez avec vos résultats obtenus lors de la question précédente.

f) Recommencez la question précédente avec cette fois ci `conf.type = "log"`. A quoi correspond cette option ?

Exercice 4. Diploide // Haploide : Est ce important ?

a) Lisez l'aide de la fonction `survdiff` et comprenez bien son utilité.

b) Répondez alors à la question du titre de cet exercice.

Exercice 5 (Impact du taux de censure)**.**
Nous allons ici nous intéresser à l'impact du taux de censure sur la qualité de nos estimations.

a) Faite une fonction qui simule n_1 durées de survie T_* selon une loi exponentielle de paramètre $\lambda = 0.1$ et n_2 censure toujours de loi exponentielle mais de paramètre $\theta = 0.07$. Au final votre fonction renvoie donc un échantillon de taille $n = n_1 + n_2$.

b) Nous allons travailler avec $n = 100$ et faire varier n_1 et n_2 sous cette contrainte. Considérons pour commencer le cas $n_1 = 90$ et $n_2 = 10$ et l'estimation de $S(10)$. Un critère bien connu pour juger de la qualité d'un estimateur est l'erreur quadratique moyenne

$$\mathbb{E}\left[\left\{\hat{S}(10) - S(10)\right\}^2\right],$$

que nous estimons facilement par Monte–Carlo par

$$\frac{1}{N}\sum_{k=1}^{N}\left\{\hat{S}(10) - S(10)\right\}^2,$$

pour un entier N assez grand, e.g., $N = 500$.
Faite une fonction qui calcule cette erreur quadratique moyenne mais aussi le biais et la variance de l'estimateur pour ce cas particulier.

c) Recommencez en faisant varier le taux de censure, e.g., $n_2 = 0, 10, \ldots, 90$. Et représentez graphiquement les résultats obtenus. Quels constats pouvez vous faire ? Est-ce cohérent ?

Exercice 6 (Tiens au fait. . .)**.**
Reprendre le même exercice que précédemment mais en considérant 3 estimateurs différents :

a) L'estimateur de Kaplan–Meier ;

b) L'estimateur de Kaplan–Meier mais en tenant compte que des données réellement observées, i.e., on omet les données censurée ;

c) L'estimateur de Kaplan–Meier mais en traitant les données censurées comme des données réellement observées.

Quels constats pouvez vous faire ?