Table des matières

Chapitre 1

Estimation non-paramétrique de la fonction de répartition et de la fonction quantile

Exercice 1.1.
Le but de cet exercice est d'illustrer les propriétés asymptotiques de la fonction de distribution empirique à partir de variables aléatoires iid distribué selon la loi exponentielle.

1. Génèrer un échantillon de taille $n = 100$ variables aléatoires i.i.d. selon une loi exponentielle L du paramètre λ à choisir.
2. Soit F la fonction de distribution de la loi L. Démontrer que $F_n(x)$ converge vers $F(x)$ presque sûrement.
3. Illustrer la convergence presque sûre de $F_n(x)$ vers $F(x)$
4. Illustrer également la normalité asymptotique associée à $F_n(x)$
5. Illustrer graphiquement la convergence de $\sup_x |F_n(x) - F(x)|$

Exercice 1.2.
Soit $X_1, \cdots, X_n$ n variables aléatoires iid de fonction de répartition F et de densité f. On note par $X_{(1)}, \cdots, X_{(n)}$ la statistique d'ordre associée et on rappelle le résultat suivant :

$$\sqrt{n}\left(X_{\lceil np \rceil} - q_p\right) \Longrightarrow \mathcal{N}\left(0, \frac{p(1-p)}{(f(q_p))^2}\right),$$

avec q_p est le quantile d'ordre p.
Utiliser ce résultat pour illustrer un intervalle de confiance pour la médiane empirique dans un cadre gaussien.

Le diagramme quantile-quantile est un outil graphique qui permet d'évaluer la pertinence de l'ajustement d'une loi de probabilités. À l'issue d'une enquête statistique, on soupçonne celle-ci de suivre une distribution classique, par exemple la loi gaussienne. À partir de la série statistique observée, on calcule alors un certain nombre de quantiles $q_{i/n+1}$. Si la série statistique suit bien la loi théorique choisie, on devrait avoir les quantiles observés $X_{(\lceil i/n+1 \rceil)}$ égaux aux quantile $q_{i/n+1} = F^{-1}(i/n+1)$ associés au modèle théorique. On place alors le nuage de point $(X_{(\lceil i/n+1 \rceil)}, q_{i/n+1})$. En abscisse se trouvent donc les quantiles théoriques et en ordonnée les quantiles observés. Si la loi théorique choisie est pertinente, les points doivent se positionner suivant la première diagonale.

Illustrer des diagrammes quantile-quantile des données issues de loi gaussienne, loi de Student et loi uniforme.

Chapitre 2

Estimation non-paramétrique de la fonction de survie

Exercice 2.1.
A partir de ce modèle exponentiel simple dont la fonction h de risque est constante au cours du temps t, tracez les courbes des fonctions de densité, de répartition, de survie, et de risque pour un organisme dont la durée moyenne de survie est égale à 2 semaines ($\mu = 2$)

Exercice 2.2.
Trouver l'estimateur de Kaplan-Meier des données suivantes (n=10)

```
Surv(timeOFevent, event)
[1] 3 6 6 8+ 9 10 14+ 16 17 18
```

+ : veut dire censuré.
A partir de l'estimation de la fonction de survie et des données, calculer l'estimation de l'écart type $\hat{\sigma}(\hat{S}(t_3))$
En déduire un intervalle de confiance à 95% pour $S(t_3)$.

Exercice 2.3.
Soit le temps T(semaine) de rémission chez deux groupes de patients atteints de leucémie :

```
temps1 <-c(6, 6, 6, 7, 10, 13, 16, 22, 23, 6, 9, 10, 11, 17, 19, 20, 25, 32, 32, 34, 35)
status1 <-c(1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)
```

Le second groupe est soumis à un placebo

```
temps2 <-c(1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23)
status2 <-c(1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)
```

1. Dans un tableur Excel, créez un tableau à trois colonnes : temps, status, et groupe afin de pouvoir y rentrer les données ci-dessus.
2. Enregistrez votre fichier au format texte .txt avec des séparateurs de type tabulation et utilisez la fonction read.table pour importer ce fichier de données dans R
3. Tracez la courbe de Kaplan-Meier sur l'ensemble des données
4. Tracez les courbes de Kaplan-Meier pour chacun des deux groupes
5. Par lecture graphique, donnez le temps médian (médiane) de rémission des patients atteints de leucémie dans chacun des deux groupes

6. Utlisez la fonction survfit pour obtenir le temps médian
7. Pourriez-vous conclure sur la significativité de cette différence entre les deux groupes par une simple lecture graphique des IC ?

Exercice 2.4.

1. Importer le fichier de données Alcool dans R. Combien y-a-t-il d'observations ? Expliquer les variables.
2. La variable censurée qu'on va étudier correspond aux délais jusqu'à rechute de la maladie alcoolique. Faire appel à la library "survival" et tracer la fonction de survie (courbe de Kaplan Meyer) de cette variable censurée (utiliser les fonctions survfit et Surv.
3. Représenter sur le même graphe le temps jusqu'à la rechute alcoolique chez les hommes et chez les femmes.
4. La médiane de survie est le moment où 50% des sujets sont survivants et 50% sont décédées.