IFMA-MPE **Statistique de base**
2017-2018

Examen de rattrapage
2 heures

Exercice 1 *On se donne un échantillon $(X_1, \dots, X_n)$ de variables aléatoires indépendantes de loi exponentielle de paramètre $\theta > 0$ inconnu. La densité est donc donnée par $x \to \theta e^{-\theta x} \mathbf{1}_{\{x \geq 0\}}$.*

1. *Donner la vraisemblance du modèle.*
2. *Calculer l'estimateur du maximum de vraisemblance $\widehat{\theta}$ de θ.*
3. *Montrer que $\widehat{\theta}_n$ converge presque sûrement vers θ.*
4. *Montrer que $\sqrt{n}\left(\frac{1}{\widehat{\theta}_n} - \frac{1}{\theta}\right)$ converge en loi vers une loi que l'on précisera.*
5. *En déduire que $\sqrt{n}(\widehat{\theta}_n - \theta)$ converge aussi en loi.*
6. *On aimerait détecter des valeurs aberrantes, c'est-à-dire des observations X_i qui sont anormalement grandes. On s'intéresse donc au maximum $Z_n := \max_{1 \leq i \leq n} X_i$.*
   (a) *Calculer la probabilité $P(Z_n \leq x)$ pour $x \geq 0$.*
   (b) *Montrer que la variable aléatoire $\theta Z_n - \ln(n)$ converge en loi vers Y de fonction de répartition $F_Y(x) = e^{-e^{-x}}$.*
   (c) *On admet que $\widehat{\theta} Z_n - \ln(n)$ converge aussi en loi vers Y. On note*

$$x_\alpha := -\ln(-\ln(1-\alpha)).$$

   Trouver a_n s'exprimant en fonction de n, $\widehat{\theta}_n$ et x_α tel que $P(Z_n \leq a_n) \to 1 - \alpha$ quand $n \to +\infty$.
   (d) *En déduire une stratégie pour déceler une valeur aberrante dans les observations.*

Exercice 2
1. *Dans une grande ville, on a enregistré le nombre de naissances de filles et de garçons sur l'année 2016. On notera X le nombre de garçons, Y le nombre de filles, et $n = X + Y$ le nombre total de naissances (que l'on suppose grand). Construire un test statistique permettant de savoir si il naît en moyenne autant de filles que de garçons dans cette ville.*
2. *On considère deux lycées d'une même ville. On note $(X_i)_{1 \leq i \leq n}$ les moyennes au bac des bacheliers du premier lycée, et $(Y_i)_{1 \leq i \leq m}$ celles pour le deuxième lycée. On suppose que ces variables suivent des lois Gaussiennes $\mathcal{N}(\mu_X, \sigma_X^2)$ et $\mathcal{N}(\mu_Y, \sigma_Y^2)$ et que $\sigma_X^2 = \sigma_Y^2 = \sigma^2$ (σ^2 est inconnue). Construire un test statistique pour savoir si le premier lycée donne en moyenne de meilleurs résultats que le deuxième lycée.*

Exercice 3 **A/** *On considère une régression linéaire simple*

$$Y_k = ax_k + b + \varepsilon_k,\ k = 1 \dots n \tag{1}$$

où $(\varepsilon_k)_k$ *sont i.i.d. de loi* $\mathcal{N}(0, \sigma^2)$*. On suppose que* $\sum_{k=1}^n x_k = 0$ *(la variable explicative est centrée).*

1. *Donner l'expression des estimateurs des moindres carrés ordinaires* $\widehat{a}$ *et* $\widehat{b}$*.*
2. *On note* $\widehat{Y}_k = \widehat{a}x_k + \widehat{b}$ *et* $\overline{Y} = \frac{1}{n}\sum_{k=1}^n Y_k$*. Exprimer la somme des carrés expliquée* $SCE = \sum_{k=1}^n (\widehat{Y}_k - \overline{Y})^2$ *en fonction de* $\widehat{a}$ *et* $S := \sum_{k=1}^n x_k^2$*.*

B/ *On considère désormais une régression linéaire multiple*

$$Y_k = a_0 + a_1 x_k^{(1)} + \ldots + a_{p-1} x_k^{(p-1)} + \varepsilon_k,\ k = 1 \dots n. \tag{2}$$

On note

$$M = \begin{bmatrix} 1 & x_1^{(1)} & \dots & x_1^{(p-1)} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n^{(1)} & \dots & x_n^{(p-1)} \end{bmatrix}$$

On notera $x_k^{(0)} = 1$ *pour tout* k *entre* 1 *et* n*. On suppose que les colonnes de* M *sont* **orthogonales***, c'est-à-dire que si* $0 \leq i \neq j \leq p-1$*, alors*

$$\sum_{k=1}^n x_k^{(i)} x_k^{(j)} = 0.$$

On notera $S_j := \sum_{k=1}^n (x_k^{(j)})^2$ *pour* $j = 0, \dots, p-1$ *(donc* $S_0 = n$*).*

1. *Que vaut* $\sum_{k=1}^n x_k^{(j)}$ *pour* $j = 1, \dots, p-1$*?*
2. *Ecrire le modèle de régression sous forme matricielle, puis rappeler l'expression de l'estimateur MCO* $\widehat{\theta} = (\widehat{a}_0, \dots, \widehat{a}_{p-1})$ *en fonction de* M *et* $Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}$*.*
3. *Donner l'expression de* $\widehat{a}_j$ *pour* j *entre* 0 *et* $p-1$*.*
4. *Montrer que* $\widehat{a}_j$ *est aussi l'estimateur MCO du coefficient de* $x^{(j)}$ *dans la régression linéaire simple* $y_k = ax_k^{(j)} + b + \varepsilon_k$*.*
5. *Soit* j *un entier entre* 1 *et* $p-1$*. Rappeler le test de significativité du coefficient* a_j*. On notera* $\widehat{\varepsilon}_k$ *pour* $1 \leq k \leq n$ *les résidus de la régression* (2)*, et* $\widehat{\sigma}$ *l'estimateur de l'écart-type des erreurs dont on rappellera l'expression.*
6. *On aimerait évaluer la pertinence des variables explicatives en les retirant du modèle l'une après l'autre. On note* Ω_j *le modèle de régression qui comprend les variables explicatives* $x^{(0)}, x^{(1)}, \dots, x^{(j)}$*. Donner la statistique du F-test permettant de comparer les modèles* Ω_{j-1} *et* Ω_j*. On l'exprimera en fonction de la somme des carrés des résidus* $SCR = \sum_{k=1}^n \widehat{\varepsilon}_k^2$*, de* S_j *et de* $\widehat{a}_j$*.*
7. *Est-ce que l'ordre dans lequel on retire les variables explicatives joue un rôle ici?*