

# HEC Paris – ESCP Europe

## Mathématiques II

Mathématiques approfondies

Sujet adapté – 2019

Les parties du sujet contenant des éléments n'étant plus au programme depuis 2022 ont été adaptées pour convenir aux nouveaux programmes de la filière Mathématiques approfondies en vigueur depuis 2023.

## Table des matières

|          |                                                                                                                    |          |
|----------|--------------------------------------------------------------------------------------------------------------------|----------|
| <b>1</b> | <b>Partie I – Fonction logistique et lois logistiques</b>                                                          | <b>2</b> |
|          | <i>Thèmes abordés : fonctions d'une variable ; intégrales généralisées ; variables à densité ; Python. . . . .</i> | 2        |
| <b>2</b> | <b>Partie II – Variance de la loi logistique standard</b>                                                          | <b>3</b> |
|          | <i>Thèmes abordés : intégrales généralisées ; séries ; variables à densité. . . . .</i>                            | 3        |
| <b>3</b> | <b>Partie III – Estimation à partir de données binaires</b>                                                        | <b>4</b> |
|          | <i>Thèmes abordés : probabilités discrètes ; estimations ; Python. . . . .</i>                                     | 4        |
| <b>4</b> | <b>Partie IV – Régression logistique</b>                                                                           | <b>5</b> |
|          | <i>Thèmes abordés : fonctions de plusieurs variables ; estimations ; applications linéaires. . . . .</i>           | 5        |

## Problème – Régression logistique

La régression logistique permet de modéliser l'influence qu'exercent des facteurs exogènes sur une variable binaire, c'est-à-dire une variable ne pouvant prendre que deux valeurs. Outre son domaine d'application privilégié qu'est l'apprentissage automatique (*machine learning*), la régression logistique est couramment utilisée aussi bien en médecine qu'en actuariat et en économétrie.

### 1 Partie I – Fonction logistique et lois logistiques

*Thèmes abordés : fonctions d'une variable ; intégrales généralisées ; variables à densité ; Python.*

On appelle *fonction logistique* la fonction  $\Lambda$  définie sur  $\mathbb{R}$  par

$$\forall x \in \mathbb{R}, \quad \Lambda(x) = \frac{1}{1 + e^{-x}}.$$

1. a) Montrer que  $\Lambda$  est une bijection de  $\mathbb{R}$  sur  $]0, 1[$ , dont la bijection réciproque est la fonction  $L$  définie par

$$\forall x \in ]0, 1[, \quad L(x) = \ln \left( \frac{x}{1-x} \right).$$

- b) Calculer la dérivée de la fonction  $\Lambda$ .
  - c) Justifier l'existence d'un unique réel  $x_0$  tel que  $\Lambda(x_0) = x_0$ .
  - d) Établir, pour tout  $x \in \mathbb{R}$ , l'inégalité  $|\Lambda(x) - x| \leq |x - x_0|$ .
2. Le script Python suivant, dont la première ligne définit la fonction  $\Lambda$ , permet de calculer une valeur approchée de  $x_0$  par la méthode de dichotomie.

```
def Lambda(x):
    return 1 / (1 + exp(-x))
a = 0
b = 1
eps = ...
while b - a > eps:
    c = (a + b) / 2
    if Lambda(c) > c:
        ...
    else:
        ...
x0 = (a + b) / 2
```

- a) Compléter les deux instructions manquantes et justifier le choix des valeurs affectées aux variables `a` et `b`.
  - b) Quelle valeur maximale peut-on affecter à la variable `eps` pour être assuré que l'erreur d'approximation commise ne dépasse pas  $10^{-4}$  ?
  - c) Que peut-on dire de la valeur numérique obtenue par l'instruction `Lambda(x0)-x0` ?
3. On note  $\lambda$  la dérivée de la fonction  $\Lambda$ .
  - a) Vérifier que  $\lambda$  est une densité de probabilité.

- b) Préciser la parité de la fonction  $\lambda$  ; donner l'allure de sa courbe représentative dans un repère orthogonal et en déterminer les points d'inflexion.

On dit qu'une variable aléatoire  $Z$  suit la *loi logistique standard* si elle admet la fonction  $\lambda$  pour densité. Pour tout couple  $(r, s) \in \mathbb{R} \times \mathbb{R}_+^*$ , on dit qu'une variable aléatoire  $Y$  suit la *loi logistique*  $\mathcal{L}(r, s)$  si la variable aléatoire  $Z = \frac{Y-r}{s}$  suit la loi logistique standard.

4. a) Justifier qu'une variable aléatoire qui suit une loi logistique  $\mathcal{L}(r, s)$  admet des moments de tout ordre et en indiquer l'espérance.
- b) En utilisant la méthode d'inversion, écrire une fonction Python `grandlogis(n,p,r,s)` fournissant, pour tout couple  $(n, p)$  d'entiers strictement positifs, un tableau à  $n$  lignes et  $p$  colonnes dont les coefficients sont des simulations de variables aléatoires indépendantes suivant la loi logistique  $\mathcal{L}(r, s)$ .
- c) Décrire un procédé permettant de calculer une valeur approchée de la variance de la loi logistique standard à l'aide de la fonction `grandlogis`.
5. Soient  $U_1$  et  $U_2$  deux variables aléatoires indépendantes suivant chacune la loi exponentielle de paramètre 1.
- a) Montrer que la variable aléatoire  $Z = \ln\left(\frac{U_1}{U_2}\right)$  suit la loi logistique standard (on pourra utiliser un changement de variable exponentiel, c'est-à-dire de la forme  $t = e^x$ ).
- b) En déduire une fonction Python permettant de simuler une variable aléatoire suivant la loi logistique standard à l'aide d'un générateur de variables aléatoires uniformes sur  $]0, 1[$ .

## 2 Partie II – Variance de la loi logistique standard

*Thèmes abordés : intégrales généralisées ; séries ; variables à densité.*

Les questions faisant intervenir les nombres complexes ont été retirées. Le résultat intermédiaire nécessaire à la suite de cette partie est admis.

On admet que, pour tout  $n \in \mathbb{N}^*$ ,

$$\sum_{k=1}^n \frac{1}{\sin^2\left(\frac{k\pi}{2n+1}\right)} = \frac{2n(n+1)}{3}.$$

8. Soit  $x \in ]0, \frac{\pi}{2}[$ .

- a) Justifier les inégalités suivantes :

$$\sin(x) \leq x \leq \tan(x) \quad \text{et} \quad \frac{1}{\sin^2(x)} - 1 \leq \frac{1}{x^2} \leq \frac{1}{\sin^2(x)}.$$

- b) En utilisant le résultat admis ci-dessus, en déduire, pour tout  $n \in \mathbb{N}^*$ , l'encadrement :

$$\frac{n(2n-1)}{3} \leq \frac{(2n+1)^2}{\pi^2} \sum_{k=1}^n \frac{1}{k^2} \leq \frac{2n(n+1)}{3}.$$

- c) Établir l'égalité  $\sum_{k=1}^{+\infty} \frac{1}{k^2} = \frac{\pi^2}{6}$ .

9. Soit  $Z$  une variable aléatoire suivant la loi logistique standard.

a) À l'aide d'une intégration par parties, justifier que la variance de  $Z$ , notée  $V(Z)$ , vérifie l'égalité

$$V(Z) = 4 \int_0^{+\infty} \frac{xe^{-x}}{1+e^{-x}} dx.$$

b) Établir, pour tout  $n \in \mathbb{N}$ , l'égalité

$$\int_0^{+\infty} \frac{xe^{-x}}{1+e^{-x}} dx = \sum_{k=0}^n (-1)^k \int_0^{+\infty} xe^{-(k+1)x} dx + I_n,$$

$$I_n = (-1)^{n+1} \int_0^{+\infty} \frac{xe^{-(n+2)x}}{1+e^{-x}} dx.$$

c) Montrer que l'intégrale  $I_n$  tend vers 0 lorsque  $n$  tend vers  $+\infty$  et en déduire l'égalité

$$\int_0^{+\infty} \frac{xe^{-x}}{1+e^{-x}} dx = \sum_{k=0}^{+\infty} \frac{(-1)^k}{(k+1)^2}.$$

d) En utilisant la formule établie en 8.c), déduire de l'égalité précédente que la variance de  $Z$  est égale à  $\frac{\pi^2}{3}$ .

10. a) Établir la convergence des deux intégrales  $\int_0^{+\infty} \ln(x)e^{-x} dx$  et  $\int_0^{+\infty} (\ln(x))^2 e^{-x} dx$ .

b) On pose  $I = \int_0^{+\infty} \ln(x)e^{-x} dx$  et  $J = \int_0^{+\infty} (\ln(x))^2 e^{-x} dx$ . En utilisant le résultat de la question 5.a), calculer  $J - I^2$ .

### 3 Partie III – Estimation à partir de données binaires

*Thèmes abordés : probabilités discrètes ; estimations ; Python.*

Dans cette partie,  $\theta$  est un paramètre réel inconnu et  $F$  désigne la fonction de répartition d'une variable aléatoire à densité dont une densité  $f$  est continue et strictement positive sur  $\mathbb{R}$ . Soit  $(Y_n)_{n \in \mathbb{N}^*}$  une suite de variables aléatoires indépendantes définies sur un espace probabilisé  $(\Omega, \mathcal{A}, \mathbb{P}_\theta)$  suivant chacune la loi de Bernoulli de paramètre  $F(\theta)$ .

11. Justifier que  $F$  est une bijection de  $\mathbb{R}$  sur  $]0, 1[$ . On note  $F^{-1}$  sa bijection réciproque.

12. Pour tout  $n \in \mathbb{N}^*$ , on pose  $\bar{Y}_n = \frac{1}{n} \sum_{j=1}^n Y_j$ . Montrer que la suite  $(\sqrt{n}(\bar{Y}_n - F(\theta)))_{n \in \mathbb{N}^*}$  converge en loi vers une variable aléatoire suivant une loi normale centrée dont on précisera la variance.

13. Pour tout  $n \in \mathbb{N}^*$  et tout  $\omega \in \Omega$ , on pose

$$T_n(\omega) = \begin{cases} F^{-1}(\bar{Y}_n(\omega)) & \text{si } 0 < \bar{Y}_n(\omega) < 1, \\ 0 & \text{sinon.} \end{cases}$$

De plus, pour tout  $n \in \mathbb{N}^*$ , on note  $E_n$  l'événement  $[0 < \bar{Y}_n < 1]$ .

a) Calculer  $\mathbb{P}_\theta(E_n)$  et trouver la limite de cette probabilité lorsque  $n$  tend vers  $+\infty$ .

b) Soient  $x \in \mathbb{R}$  et  $n \in \mathbb{N}^*$ .

- (i) Établir l'égalité ensembliste  $\{\omega \in E_n \mid T_n(\omega) \leq x\} = [\bar{Y}_n \leq F(x)] \cap E_n$  et montrer que  $[T_n \leq x]$  est un élément de la tribu  $\mathcal{A}$ .
- (ii) Justifier l'encadrement

$$\mathbb{P}_\theta([\bar{Y}_n \leq F(x)] \cap E_n) \leq \mathbb{P}_\theta([T_n \leq x]) \leq \mathbb{P}_\theta([\bar{Y}_n \leq F(x)] \cap E_n) + 1 - \mathbb{P}_\theta(E_n).$$

- c) Montrer que, pour tout  $x \neq \theta$ ,

$$\lim_{n \rightarrow +\infty} \mathbb{P}_\theta([T_n \leq x]) = \begin{cases} 0 & \text{si } x < \theta, \\ 1 & \text{si } x > \theta. \end{cases}$$

- d) En déduire que  $(T_n)_{n \in \mathbb{N}^*}$  est une suite convergente d'estimateurs du paramètre  $\theta$ .

14. Pour tout  $n \in \mathbb{N}^*$  et tout  $\omega \in \Omega$ , on pose

$$U_n(\omega) = \begin{cases} \frac{T_n(\omega) - \theta}{\bar{Y}_n(\omega) - F(\theta)} & \text{si } \bar{Y}_n(\omega) \neq F(\theta), \\ \frac{1}{f(\theta)} & \text{si } \bar{Y}_n(\omega) = F(\theta). \end{cases}$$

On admet sans démonstration que, pour tout  $n \in \mathbb{N}^*$ ,  $U_n$  est une variable aléatoire sur  $(\Omega, \mathcal{A}, \mathbb{P}_\theta)$ .

- a) Soit  $\varepsilon > 0$ . Pour tout  $n \in \mathbb{N}^*$ , on note  $B_n(\varepsilon)$  l'événement  $\left[ \left| U_n - \frac{1}{f(\theta)} \right| \leq \varepsilon \right]$ .
- (i) Établir l'existence d'un réel  $\alpha > 0$  tel que  $\forall x \in [\theta - \alpha, \theta + \alpha]$ ,  $\left| \frac{1}{f(x)} - \frac{1}{f(\theta)} \right| \leq \varepsilon$ .
- (ii) Pour un tel  $\alpha$ , justifier l'inclusion  $[|T_n - \theta| \leq \alpha] \cap E_n \subset B_n(\varepsilon)$ , où  $E_n$  a été défini à la question 13.
- b) Montrer que la suite  $(U_n)_{n \in \mathbb{N}^*}$  converge en probabilité vers  $\frac{1}{f(\theta)}$ .
- c) **Rappel (lemme de Slutsky).** Si  $X_n$  converge en loi vers  $X$  et si  $Y_n$  converge en probabilité vers une constante  $c$ , alors  $X_n Y_n$  converge en loi vers  $cX$ . En déduire que la suite  $(\sqrt{n}(T_n - \theta))_{n \in \mathbb{N}^*}$  converge en loi vers une variable aléatoire suivant une loi normale centrée dont on précisera la variance.

## 4 Partie IV – Régression logistique

*Thèmes abordés : fonctions de plusieurs variables ; estimations ; applications linéaires.*

Dans toute cette partie,  $p$  désigne un entier supérieur ou égal à 2. Pour tout couple  $(n, m) \in (\mathbb{N}^*)^2$ , on note  $\mathcal{M}_{n,m}(\mathbb{R})$  l'ensemble des matrices à  $n$  lignes et  $m$  colonnes à coefficients réels et  ${}^t M$  la transposée de toute matrice  $M \in \mathcal{M}_{n,m}(\mathbb{R})$ . Pour tout  $m \in \mathbb{N}^*$ , le produit scalaire usuel de deux vecteurs  $u$  et  $v$  de  $\mathbb{R}^m$  est noté  $\langle u, v \rangle$ . Si  $U$  et  $V$  sont les matrices colonnes représentant  $u$  et  $v$  dans la base canonique, le produit scalaire  $\langle u, v \rangle$  est donc l'unique coefficient de la matrice  ${}^t U V$ . On rappelle que les fonctions  $\Lambda$  et  $L$  ont été définies dans la partie I.

Dans cette partie, on note  $Y$  une variable aléatoire de Bernoulli, dite *variable endogène*, dont la loi dépend du niveau de  $p$  *facteurs exogènes*. L'influence de ces facteurs sur la loi de  $Y$  est résumée par la fonction  $b$  qui associe à un vecteur  $x \in \mathbb{R}^p$  la probabilité  $b(x)$  que  $Y$  soit égale à 1 lorsque les niveaux des facteurs sont donnés par les composantes du vecteur  $x$ .

Dans le *modèle de régression logistique* envisagé dans cette partie, la fonction  $b$  est supposée de la forme

$$b : x \longmapsto \Lambda(\langle \alpha, x \rangle),$$

où  $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_p)$  est un vecteur de  $\mathbb{R}^p$  dont les composantes  $\alpha_1, \alpha_2, \dots, \alpha_p$  sont des paramètres inconnus qui représentent les degrés d'influence des divers facteurs exogènes sur la variable endogène  $Y$ .

Pour estimer les paramètres du modèle, on dispose de  $k$  vecteurs  $x^{(1)}, x^{(2)}, \dots, x^{(k)}$  de  $\mathbb{R}^p$  ( $k \in \mathbb{N}^*$ ) et, pour tout  $i \in \llbracket 1, k \rrbracket$ , d'une suite  $(Y_{i,n})_{n \in \mathbb{N}^*}$  de variables aléatoires indépendantes suivant chacune la loi de Bernoulli de paramètre  $b(x^{(i)}) = \Lambda(\langle \alpha, x^{(i)} \rangle)$ . Pour chaque indice fixé  $i$  et pour tout  $n \in \mathbb{N}^*$ , les variables aléatoires  $Y_{i,1}, Y_{i,2}, \dots, Y_{i,n}$  définissent donc un  $n$ -échantillon associé à la loi de la variable endogène lorsque les niveaux des facteurs exogènes sont les composantes  $x_1^{(i)}, x_2^{(i)}, \dots, x_p^{(i)}$  du vecteur  $x^{(i)}$  dans la base canonique de  $\mathbb{R}^p$ .

15. On note respectivement  $A$  et  $M$  la matrice du vecteur  $\alpha$  et la matrice de la famille  $(x^{(1)}, x^{(2)}, \dots, x^{(k)})$  dans la base canonique de  $\mathbb{R}^p$  :

$$A = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{pmatrix} \in \mathcal{M}_{p,1}(\mathbb{R}) \quad \text{et} \quad M = \begin{pmatrix} x_1^{(1)} & \cdots & x_1^{(k)} \\ \vdots & & \vdots \\ x_p^{(1)} & \cdots & x_p^{(k)} \end{pmatrix} \in \mathcal{M}_{p,k}(\mathbb{R}).$$

On suppose que le rang de la matrice  $M$  est égal à  $p$ .

- Montrer que la matrice  $M^t M$  est inversible.
  - Montrer que pour toute matrice  $H \in \mathcal{M}_{k,1}(\mathbb{R})$ , la matrice  $U \in \mathcal{M}_{p,1}(\mathbb{R})$  pour laquelle l'unique coefficient de la matrice  ${}^t(MU - H)({}^t(MU - H))$  est le plus petit possible, est la matrice  $(M^t M)^{-1} M H$ .
  - Expliquer pourquoi les lois des variables aléatoires  $Y_{i,n}$  ne suffiraient pas à définir le vecteur  $\alpha$  si le rang de  $M$  n'était pas égal à  $p$ .
16. Pour tout  $n \in \mathbb{N}^*$  et tout  $i \in \llbracket 1, k \rrbracket$ , on pose  $\bar{Y}_{i,n} = \frac{1}{n} \sum_{j=1}^n Y_{i,j}$  et, pour tout  $\omega \in \Omega$ ,

$$T_{i,n} = \begin{cases} L(\bar{Y}_{i,n}(\omega)) & \text{si } 0 < \bar{Y}_{i,n}(\omega) < 1, \\ 0 & \text{sinon.} \end{cases}$$

- Soient  $(c_1, c_2, \dots, c_k) \in \mathbb{R}^k$ . En utilisant les résultats de la partie III, montrer que la suite

$$\left( \sum_{i=1}^k c_i T_{i,n} \right)_{n \in \mathbb{N}^*}$$

est une suite convergente d'estimateurs du paramètre

$$\sum_{i=1}^k c_i \langle \alpha, x^{(i)} \rangle.$$

- Pour tout  $n \in \mathbb{N}^*$  et tout  $\omega \in \Omega$ , on pose

$$H_n(\omega) = \begin{pmatrix} T_{1,n}(\omega) \\ T_{2,n}(\omega) \\ \vdots \\ T_{k,n}(\omega) \end{pmatrix} \quad \text{et} \quad \begin{pmatrix} A_{1,n}(\omega) \\ A_{2,n}(\omega) \\ \vdots \\ A_{p,n}(\omega) \end{pmatrix} = (M^t M)^{-1} M H_n(\omega).$$

Montrer que, pour tout  $j \in \llbracket 1, p \rrbracket$ , la suite  $(A_{j,n})_{n \in \mathbb{N}^*}$  est une suite convergente d'estimateurs de  $\alpha_j$ .