RAPPORT DE CORRECTION

DE MATHÉMATIQUES II Option S

Conception HEC Paris – ESCP Europe

SOMMAIRE

Le sujet

BANQUE COMMUNE D'ÉPREUVES

Code sujet : 283

Conception : HEC Paris – ESCP Europe

OPTION SCIENTIFIQUE

MATHÉMATIQUES II

Jeudi 2 mai 2019, de 8 h. à 12 h.

La présentation, la lisibilité, l'orthographe, la qualité de la rédaction, la clarté et la précision des raisonnements entreront pour une part importante dans l'appréciation des copies.
Les candidats sont invités à ***encadrer*** *dans la mesure du possible les résultats de leurs calculs.*
Aucun document n'est autorisé. ***L'utilisation de toute calculatrice et de tout matériel électronique est interdite.*** *Seule l'utilisation d'une règle graduée est autorisée.*
Si au cours de l'épreuve, un candidat repère ce qui lui semble être une erreur d'énoncé, il la signalera sur sa copie et poursuivra sa composition en expliquant les raisons des initiatives qu'il sera amené à prendre.

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 qui est l'apprentissage automatique (machine learning), la régression logistique est couramment utilisée aussi bien en médecine qu'en actuariat et en économétrie.

Partie I. Fonction logistique et lois logistiques

On appelle *fonction logistique* la fonction Λ définie sur $\mathbf{R}$ par : $\forall x \in \mathbf{R},\ \Lambda(x) = \dfrac{1}{1+\mathrm{e}^{-x}}$.

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

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

b) Calculer la dérivée de la fonction Λ.

c) Justifier l'existence d'un unique réel x_0 tel que : $\Lambda(x_0) = x_0$.

d) Établir pour tout $x \in \mathbf{R}$, l'inégalité : $|\Lambda(x) - x| \leqslant |x - x_0|$.

2. Le script *Scilab* suivant, dont la ligne (1) définit la fonction Λ, permet de calculer une valeur approchée de x_0 par la méthode de dichotomie.

```
(1) deff('y=Lambda(x)','y=1/(1+exp(-x))');
(2) a=0;
(3) b=1;
(4) eps= ·····;
(5) while b-a>eps;
(6)   c=(a+b)/2;
(7)   if Lambda(c)>c then ·····; else b= ·····; end;
(8) end;
(9) x0=(a+b)/2
```

a) Compléter la ligne `(7)` et justifier le choix des valeurs affectées en lignes `(2)` et `(3)` aux variables `a` et `b`.

b) Quelle valeur maximale peut-on affecter en ligne `(4)` 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 `(10)` suivante ?

```
(10) Lambda(x0)-x0
```

3. On note λ la dérivée de la fonction Λ.

a) Vérifier que λ est une densité de probabilité.

b) Préciser la parité de la fonction λ ; donner l'allure de sa courbe représentative dans le plan rapporté à 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 λ pour densité.
Pour tout couple $(r, s) \in \mathbf{R} \times \mathbf{R}_+^*$, on dit qu'une variable aléatoire Y suit la loi logistique $\mathcal{L}(r, s)$ si la variable aléatoire Z définie par $Z = \dfrac{Y - r}{s}$ 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 n'importe quel ordre et en indiquer l'espérance.

b) En utilisant la méthode d'inversion, écrire le script d'une fonction *Scilab*, `function S=grandlogis(n,p,r,s)`, fournissant pour tout couple `(n,p)` d'entiers strictement positifs, une matrice `S` 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. Soit U_1 et U_2 deux variables aléatoires indépendantes suivant chacune la loi exponentielle de paramètre 1.

a) Montrer que la variable aléatoire $Z = \ln\left(\dfrac{U_1}{U_2}\right)$ suit la loi logistique standard (on pourra utiliser un changement de variable exponentiel, c'est-à-dire de la forme $t = \mathrm{e}^x$).

b) En déduire un nouveau script *Scilab* permettant de simuler une variable aléatoire suivant la loi logistique standard à l'aide de la fonction `grand`.

Partie II. Variance de la loi logistique standard

- Pour tout couple $(a, b) \in \mathbf{R}^2$, on note $\mathcal{I}m(z)$ la partie imaginaire b du nombre complexe $z = a + ib$.
- Pour tout polynôme $P = \sum_{k=0}^{d} a_k X^k \in \mathbf{R}[X]$ de degré $d \in \mathbf{N}$, les termes non nuls $a_k X^k$ sont appelés les monômes de P et les a_k leurs coefficients.
- Dans la factorisation $P = a_d \prod_{k=1}^{d} (X - z_k)$ de P dans $\mathbf{C}[X]$ (lorsque $d \neq 0$), la somme $\sum_{k=1}^{d} z_k$ est appelée la somme des racines complexes de P, que les nombres complexes $z_1, z_2, \ldots, z_d$ soient distincts ou non.

Pour tout $n \in \mathbf{N}$, on pose : $P_n = \sum_{k=0}^{n} (-1)^k \binom{2n+1}{2k+1} (X-1)^{n-k}$.

6.a) Expliciter les polynômes P_0 et P_1.

b) Pour tout $n \in \mathbf{N}^*$, préciser le degré du polynôme P_n et donner les coefficients de ses deux monômes de plus hauts degrés.

c) Utiliser le résultat précédent pour montrer que pour tout $n \in \mathbf{N}^*$, la somme des racines complexes de P_n est égale à $\dfrac{2n(n+1)}{3}$.

7. Soit $x \in \mathbf{R}$ et $n \in \mathbf{N}$.

a) Justifier les égalités suivantes :

$$\sin\big((2n+1)x\big) = \mathcal{I}m\Big(\big(\cos(x)+i\sin(x)\big)^{2n+1}\Big) = \sum_{k=0}^{n}(-1)^k\binom{2n+1}{2k+1}\cos^{2(n-k)}(x)\times\sin^{2k+1}(x)\,.$$

b) En déduire, pour tout $x \in]0,\pi[$, la relation : $\dfrac{\sin\big((2n+1)x\big)}{\sin^{2n+1}(x)} = P_n\Big(\dfrac{1}{\sin^2(x)}\Big)$.

c) À l'aide du résultat de la question 6.c), montrer que pour tout $n \in \mathbf{N}^*$, on a :

$$\sum_{k=1}^{n}\frac{1}{\sin^2\Big(\dfrac{k\,\pi}{2n+1}\Big)} = \frac{2n(n+1)}{3}\,.$$

8. Soit $x \in \left]0,\dfrac{\pi}{2}\right[$.

a) Justifier les inégalités suivantes : $\quad \sin(x) \leqslant x \leqslant \tan(x) \quad$ et $\quad \dfrac{1}{\sin^2(x)} - 1 \leqslant \dfrac{1}{x^2} \leqslant \dfrac{1}{\sin^2(x)}$.

b) En utilisant le résultat de la question 7.c), en déduire, pour tout $n \in \mathbf{N}^*$, l'encadrement :

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

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

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 :

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

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

$$\int_0^{+\infty}\frac{x\,\mathrm{e}^{-x}}{1+\mathrm{e}^{-x}}\,\mathrm{d}x = \sum_{k=0}^{n}(-1)^k\int_0^{+\infty}x\,\mathrm{e}^{-(k+1)x}\,\mathrm{d}x + I_n, \quad \text{où} \quad I_n = (-1)^{n+1}\int_0^{+\infty}\frac{x\,\mathrm{e}^{-(n+2)x}}{1+\mathrm{e}^{-x}}\,\mathrm{d}x\,.$$

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

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

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

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

b) On pose $I = \displaystyle\int_0^{+\infty}\ln(x)\,\mathrm{e}^{-x}\,\mathrm{d}x$ et $J = \displaystyle\int_0^{+\infty}\big(\ln(x)\big)^2\,\mathrm{e}^{-x}\,\mathrm{d}x$.

En utilisant le résultat de la question 5.a), calculer $J - I^2$.

Partie III. Estimation à partir de données binaires

Dans cette partie, θ 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 $\mathbf{R}$.

Soit $(Y_n)_{n\in\mathbf{N}^*}$ une suite de variables aléatoires indépendantes définies sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbf{P}_\theta)$ suivant chacune la loi de Bernoulli de paramètre $F(\theta)$.

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

12. Pour tout $n \in \mathbf{N}^*$, on pose : $\overline{Y}_n = \dfrac{1}{n}\displaystyle\sum_{j=1}^{n} Y_j$.

Montrer que la suite $\left(\sqrt{n}\left(\overline{Y}_n - F(\theta)\right)\right)_{n\in\mathbf{N}^*}$ converge en loi vers une variable aléatoire suivant une loi normale centrée dont on précisera la variance.

13. Pour tout $n \in \mathbf{N}^*$ et tout $\omega \in \Omega$, on pose : $T_n(\omega) = \begin{cases} F^{-1}\left(\overline{Y}_n(\omega)\right) & \text{si } 0 < \overline{Y}_n(\omega) < 1 \\ 0 & \text{sinon} \end{cases}$.

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

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

b) Soit $x \in \mathbf{R}$ et $n \in \mathbf{N}^*$.

(i) Établir l'égalité ensembliste $\{\omega \in E_n \,/\, T_n(\omega) \leqslant x\} = [\overline{Y}_n \leqslant F(x)] \cap E_n$ et montrer que $[T_n \leqslant x]$ est un élément de la tribu $\mathcal{A}$.

(ii) Justifier l'encadrement :

$$\mathbf{P}_\theta\left([\overline{Y}_n \leqslant F(x)] \cap E_n\right) \leqslant \mathbf{P}_\theta\left([T_n \leqslant x]\right) \leqslant \mathbf{P}_\theta\left([\overline{Y}_n \leqslant F(x)] \cap E_n\right) + 1 - \mathbf{P}_\theta(E_n)\,.$$

c) Montrer que pour tout $x \neq \theta$, on a : $\lim\limits_{n\to+\infty} \mathbf{P}_\theta\left([T_n \leqslant x]\right) = \begin{cases} 0 & \text{si } x < \theta \\ 1 & \text{si } x > \theta \end{cases}$.

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

14. Pour tout $n \in \mathbf{N}^*$ et tout $\omega \in \Omega$, on pose : $U_n(\omega) = \begin{cases} \dfrac{T_n(\omega) - \theta}{\overline{Y}_n(\omega) - F(\theta)} & \text{si } \overline{Y}_n(\omega) \neq F(\theta) \\ \dfrac{1}{f(\theta)} & \text{si } \overline{Y}_n(\omega) = F(\theta) \end{cases}$.

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

a) Soit $\varepsilon > 0$.

Pour tout $n \in \mathbf{N}^*$, on note $B_n(\varepsilon)$ l'événement $\left[\left|U_n - \dfrac{1}{f(\theta)}\right| \leqslant \varepsilon\right]$.

(i) Établir l'existence d'un réel $\alpha > 0$ tel que : $\forall x \in [\theta - \alpha, \theta + \alpha]$, $\left|\dfrac{1}{f(x)} - \dfrac{1}{f(\theta)}\right| \leqslant \varepsilon$.

(ii) Pour un tel α, justifier l'inclusion : $\left[|T_n - \theta| \leqslant \alpha\right] \cap E_n \subset B_n(\varepsilon)$, où E_n a été défini dans la question 13.

b) Montrer que la suite $(U_n)_{n\in\mathbf{N}^*}$ converge en probabilité vers $\dfrac{1}{f(\theta)}$.

c) En déduire que la suite $\left(\sqrt{n}(T_n - \theta)\right)_{n\in\mathbf{N}^*}$ converge en loi vers une variable aléatoire suivant une loi normale centrée dont on précisera la variance.

Partie IV. Régression logistique

- Dans toute cette partie, p désigne un entier supérieur ou égal à 2.
- Pour tout couple $(n, m) \in (\mathbf{N}^*)^2$, on note $\mathcal{M}_{n,m}(\mathbf{R})$ l'ensemble des matrices à n lignes et m colonnes à coefficients réels et ${}^tM$ la transposée de toute matrice $M \in \mathcal{M}_{n,m}(\mathbf{R})$.
- Pour tout $m \in \mathbf{N}^*$, le produit scalaire usuel de deux vecteurs u et v de $\mathbf{R}^m$ 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 ${}^tUV$.
- *On rappelle* que les fonctions Λ et L ont été définies dans la partie I.

Le sujet

Le problème avait pour objet l'étude de la régression logistique.

La régression logistique permet de modéliser l'influence qu'exercent des facteurs exogènes sur une variable binaire (quantitative ou qualitative), c'est-à-dire une variable ne pouvant prendre que deux valeurs ou deux modalités.

Le domaine d'application privilégié de la régression logistique est l'apprentissage automatique (machine learning) mais ce modèle est couramment utilisé aussi bien en médecine qu'en actuariat et en économétrie.

Ce problème mobilisait des connaissances en analyse (étude de fonction, théorème des accroissements finis, méthode de dichotomie de recherche des solutions d'une équation, analyse complexe et polynômes, intégration par parties), en statistique et probabilités (densité, méthode d'inversion pour simuler des variables aléatoires suivant la loi logistique, méthode de convolution, estimation, convergence en loi et en probabilité, théorème de Slutsky) et en algèbre (problème des moindres carrés et projection orthogonale).

Le barème

Les quatre parties du problème comptaient respectivement pour 34%, 29%, 25 % et 12% des points de barème.

Le poids des questions de *Scilab* était assez élevé puisqu'il représentait 12% des points de barème.

Les questions les plus cotées étaient : 3.b), 4.b), 5.a), 9.c), 14.a) (ii) et 16.a) et totalisaient 21 % des points de barème.

Remarques de correction

D'une façon générale, la maîtrise du cours est insuffisante : les connaissances apprises sont souvent mal comprises et restituées assez mécaniquement et le jury a l'impression que beaucoup de candidats travaillent par réflexe conditionné.

Il apparaît également une assez forte corrélation positive entre la présentation des copies et leur contenu.

On observe toujours une très mauvaise maîtrise des opérations élémentaires et des notions de base : par exemple, addition et multiplication sont régulièrement confondues.

Rappelons qu'il est nécessaire de justifier une réponse en expliquant pourquoi on aboutit à telle conclusion plutôt qu'à une autre (cf. questions 2.a), 2.b) et 2.c)).

On peut établir une dichotomie entre les questions selon un critère de sélectivité en séparant les questions *peu sélectives*, c'est-à-dire celles traitées par une majorité de candidats ou au contraire, celles traitées par une minorité de candidats et les questions *sélectives*, c'est-à-dire celles dans lesquelles la répartition des points attribués est assez uniforme.

- *Questions peu sélectives*

Les questions 1.a), 1.b), 3.a), 5.b), 6.a), 8.b) et 8.c) sont en général *bien traitées*: il s'agit de questions très classiques sans difficulté de rédaction ou de calcul que les candidats ont rencontrées maintes fois durant leurs deux années de préparation.

Inversement, les questions 2.b), 3.b), 6.b), 7.a), 7.b), 7.c), 9.d), 10.b), 13.a), 13.b)(i), 13.d), 14.b) et 15.a) sont en général *mal traitées*.

Parmi ces questions mal traitées, certaines sont *assez souvent abordées* et révèlent un certain nombre de points du cours mal compris.
- 2.b) : il fallait lire précisément l'énoncé (qui ne demandait pas une « valeur convenable » mais la « meilleure valeur possible ») et surtout rédiger une réponse justifiée.
- 3.b) : étude de fonction, technique autrefois classique et maintenant en « déshérence » complète ; un très grand nombre de candidats confondent la notion de *point d'inflexion* avec celle de *point critique*.
- 6.b) : calculs algébriques sur les polynômes ; question excessivement technique pour la grande majorité des candidats.
- 7.a) : question ultra-classique ; si la formule de Moivre est connue, l'application de la formule du binôme est souvent erronée, ce qui montre qu'une proportion importante de candidats apprennent des formules qu'ils ne comprennent pas et ne craignent pas de présenter des calculs « malhonnêtes » (puisque le résultat annoncé par le sujet apparaît comme la conclusion du calcul).
- 7.b) : la plupart des candidats perdent des points en omettant de préciser qu'ils ne divisent pas par zéro !

Les autres questions, rarement et mal traitées, abordent des sujets qui ne sont pas du tout compris : l'application du théorème spectral dans la question 15.a), la question 1.c) dans laquelle une minorité de candidats utilisent la fonction H : $x \to H(x) = \Lambda(x) - x$ alors que la plupart des candidats « déduisent » le résultat de la monotonie de Λ ou encore la question 1.d) très souvent abandonnée.

- *Questions sélectives*

Les questions sont d'autant plus sélectives que ces questions sont fréquemment traitées.

Les *questions les plus sélectives* sont les suivantes :
- 2.a) : compréhension de l'algorithme de dichotomie, qualités d'expression pour justifier le choix initial de *a* et *b*.
- 4.a) : convergence d'une intégrale impropre.
- 5.a) : détermination d'une densité par un calcul de produit de convolution, rigueur dans l'application d'un théorème.
- 11) : rigueur dans l'application d'un théorème (théorème de la bijection).
- 12) : rigueur dans l'application d'un théorème (théorème limite central).

Les questions suivantes ont été *moins sélectives* car plus rarement traitées par les candidats.
- 4.b) : connaissance d'une méthode de simulation de variables aléatoires, traduction en langage *Scilab* des résultats précédents.
- 4.c) : traduction en langage *Scilab* d'une formule mathématique simple (variance).
- 9.b) : identification d'une formule simple (somme géométrique)
- 9.c) : maîtrise d'une technique de calcul d'ordre de grandeur (encadrement d'une intégrale)
- 13.b)(ii) : question qui n'a presque pas été abordée.

Conseils aux futurs candidats

Pour ce qui concerne la forme, le jury conseille aux futurs candidats de lire attentivement le texte préliminaire qui précède toute épreuve écrite de mathématiques, dans lequel il est précisé notamment, que la lisibilité et la qualité de la rédaction entrent pour une part non négligeable dans l'appréciation des copies : un correcteur ne s'attarde pas à essayer de « décrypter » une copie illisible. Par contre, une copie propre et claire ne peut qu'avantager son auteur. Le jury rappelle également que les abréviations dans les copies doivent être proscrites et il conseille de bien numéroter les questions et d'encadrer les résultats.

De plus, les raisonnements doivent être clairs et précis, les affirmations étant étayées par une argumentation solide. Par exemple, le recours trop fréquent à des phrases du type « il est clair que… » doit être évité au profit d'une justification correcte fondée sur un apprentissage rigoureux et une très bonne maîtrise du cours.

Le jury recommande aux futurs candidats de prendre le temps de lire l'ensemble du sujet, non seulement pour s'en imprégner, mais aussi pour pointer les questions qui paraissent faciles à résoudre, lesquelles ne se situent pas nécessairement dans la première partie du sujet.

La recherche d'une solution à une question ne doit pas dépasser quatre à cinq minutes. Au-delà de ce délai, en cas d'échec, le candidat doit admettre le résultat de cette question (si la réponse figure dans l'énoncé), passer à la question suivante sans éprouver un sentiment de déstabilisation ou de découragement. Autrement dit, le jury recommande aux futurs candidats de faire preuve d'une grande ténacité.

Statistiques

Sur les 3316 candidats, la moyenne est de 10,5 avec un écart-type de 4,8, un peu plus « serré » que celui du concours 2018 mais suffisamment élevé pour classer les candidats de manière satisfaisante.

Le nombre de candidats ayant obtenu une note supérieure ou égale à 16 est de 474, soit 14,3% des candidats présents.

On compte 34 candidats qui se voient attribuer la note maximale de 20 en baisse sensible par rapport au concours 2018.

La note médiane est de 11,2 et les premier et troisième quartiles sont égaux à 7,2 et 14,2 respectivement.

La note maximale de 20 était attribuée aux candidats ayant obtenu au moins 52% des points du barème.