BANQUE COMMUNE D'EPREUVES

CONCOURS D'ADMISSION DE 2012

Conception : C.C.I.P.

OPTION SCIENTIFIQUE

Code épreuve : 283

MATHEMATIQUES II

Mercredi 9 mai 2012, 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.
Ils ne doivent faire usage d'aucun document : **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

---

- Toutes les variables aléatoires qui interviennent dans ce problème sont réelles et définies sur un même espace probabilisé $(\Omega, \mathcal{A}, P)$, où P peut dépendre de paramètres réels inconnus a, b, σ etc ; elles admettent toutes une espérance et une variance : si J désigne l'une de ces variables aléatoires, on note $E(J)$ son espérance et $V(J)$ sa variance.
  Si J_1, J_2 et $J_1 + J_2$ sont des variables aléatoires à densité, on admet alors l'existence de la covariance de J_1 et J_2, notée $\mathrm{Cov}(J_1, J_2)$, qui est définie par la formule : $\mathrm{Cov}(J_1, J_2) = \frac{1}{2}\big(V(J_1 + J_2) - V(J_1) - V(J_2)\big)$.
  On admet que les covariances de variables aléatoires à densité vérifient les mêmes règles de calcul que celles des variables aléatoires discrètes.
- Pour tout (k, ℓ) de $(\mathbb{N}^*)^2$, on note $\mathcal{M}_{k,\ell}(\mathbb{R})$ l'ensemble des matrices à k lignes et ℓ colonnes à coefficients réels ; on note $\mathcal{M}_k(\mathbb{R})$ l'ensemble des matrices carrées d'ordre k.
- On note ${}^tQ$ la transposée d'une matrice Q.
- Dans tout le problème, n désigne un entier supérieur ou égal à 3.

L'objet du problème est l'étude de quelques propriétés du modèle de régression linéaire élémentaire.

Partie I. Quelques résultats statistiques et algébriques

On considère une population d'individus statistiques dans laquelle on étudie deux caractères quantitatifs $\mathcal{X}$ et $\mathcal{Y}$. On extrait de cette population, un échantillon de n individus sélectionnés selon des valeurs choisies du caractère $\mathcal{X}$ et numérotés de 1 à n.
Pour tout i de $[\![1, n]\!]$, les réels x_i et y_i sont les observations respectives de $\mathcal{X}$ et de $\mathcal{Y}$ pour l'individu i de l'échantillon. On suppose que les réels $x_1, x_2, \ldots, x_n$ ne sont pas tous égaux.
Soit a et b deux paramètres réels. On pose pour tout i de $[\![1, n]\!]$: $u_i = y_i - (ax_i + b)$. $\quad (*)$

1. On note $\overline{x}$ (resp. $\overline{y}$) et s_x^2 (resp. s_y^2), la moyenne empirique et la variance empirique de la série statistique $(x_i)_{1 \leqslant i \leqslant n}$ (resp. $(y_i)_{1 \leqslant i \leqslant n}$) ; on rappelle que : $\overline{x} = \dfrac{1}{n}\sum_{i=1}^{n} x_i$ et $s_x^2 = \dfrac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$.

   a) Montrer que $s_x^2 > 0$.

   b) Établir les formules : $\sum_{i=1}^{n}(x_i - \overline{x})y_i = \sum_{i=1}^{n}(x_i y_i) - n\overline{x}\,\overline{y}$ et $\sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n}(x_i^2) - n\overline{x}^2$.

Tournez la page S.V.P.

c) On pose pour tout i de $[\![1,n]\!]$: $\alpha_i = \dfrac{(x_i - \overline{x})}{ns_x^2}$. Montrer que : $\sum_{i=1}^{n} \alpha_i = 0$, $\sum_{i=1}^{n} \alpha_i x_i = 1$ et $\sum_{i=1}^{n} {\alpha_i}^2 = \dfrac{1}{ns_x^2}$.

2. On pose : $y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \in \mathcal{M}_{n,1}(\mathbb{R})$, $u = \begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \in \mathcal{M}_{n,1}(\mathbb{R})$, $\theta = \begin{pmatrix} a \\ b \end{pmatrix} \in \mathcal{M}_{2,1}(\mathbb{R})$ et $M = \begin{pmatrix} x_1 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{pmatrix} \in \mathcal{M}_{n,2}(\mathbb{R})$.

Les n relations $(*)$ s'écrivent sous la forme matricielle suivante : $y = M\theta + u$.

a) Quel est le rang de la matrice M ?

b) Calculer la matrice ${}^tMM$ et justifier son inversibilité.

3. L'espace vectoriel $\mathbb{R}^n$ est muni de sa structure euclidienne canonique. Soit $\mathcal{F}$ le sous-espace vectoriel engendré par les vecteurs $(x_1, x_2, \ldots, x_n)$ et $(1, 1, \ldots, 1)$ de $\mathbb{R}^n$. On note K la matrice du projecteur orthogonal de $\mathbb{R}^n$ sur $\mathcal{F}$ dans la base canonique de $\mathbb{R}^n$ et $G = I - K$, où I désigne la matrice identité de $\mathcal{M}_n(\mathbb{R})$.

a) On cherche les matrices $\theta = \begin{pmatrix} a \\ b \end{pmatrix}$ de $\mathcal{M}_{2,1}(\mathbb{R})$ qui minimisent $\sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} \big(y_i - (ax_i + b)\big)^2$.

Montrer que ce problème admet une unique solution $\hat{\theta} = \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix}$ et qu'elle vérifie la relation : ${}^tMM\hat{\theta} = {}^tMy$.

b) Montrer que : $\hat{a} = \sum_{i=1}^{n} \alpha_i y_i$ et $\hat{b} = \overline{y} - \hat{a}\overline{x}$.

c) Exprimer K en fonction de M et ${}^tM$.

d) Soit $\hat{u}$ la matrice-colonne de $\mathcal{M}_{n,1}(\mathbb{R})$ de composantes $\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_n$ définie par $\hat{u} = y - M\hat{\theta}$.
Montrer que : $\hat{u} = Gy = Gu$.

e) En déduire les égalités : ${}^t\hat{u}\hat{u} = \sum_{i=1}^{n} \hat{u}_i^2 = {}^tyGy = {}^tuGu$.

Partie II. Le modèle de régression linéaire

Le contexte et les notations sont ceux de la partie I. Dans cette partie, on cherche à modéliser les fluctuations aléatoires du caractère $\mathcal{Y}$ sur l'échantillon.

Les hypothèses du modèle de régression linéaire élémentaire sont les suivantes :
- les réels a et b sont des paramètres inconnus ;
- pour tout i de $[\![1,n]\!]$, la valeur x_i du caractère $\mathcal{X}$ est connue et la valeur y_i du caractère $\mathcal{Y}$ est la réalisation d'une variable aléatoire Y_i ;
- pour tout i de $[\![1,n]\!]$, Y_i est la somme d'une composante déterministe $ax_i + b$, fonction affine de la valeur choisie x_i, et d'une composante aléatoire U_i ;
- les variables aléatoires $U_1, U_2, \ldots, U_n$ sont mutuellement indépendantes, de même loi, possèdent une densité, et pour tout i de $[\![1,n]\!]$: $E(U_i) = 0$ et $V(U_i) = \sigma^2$, où le paramètre inconnu σ est strictement positif.

Le modèle de régression linéaire s'écrit alors : pour tout i de $[\![1,n]\!]$, $Y_i = ax_i + b + U_i$ (1).
L'objectif consiste à estimer les paramètres inconnus a, b et σ^2 du modèle (1).

On pose pour tout $n \geqslant 3$: $\overline{Y_n} = \dfrac{1}{n}\sum_{i=1}^{n} Y_i$ et $\overline{U_n} = \dfrac{1}{n}\sum_{i=1}^{n} U_i$.

4. On note A_n et B_n les deux variables aléatoires définies par : $A_n = \sum_{i=1}^{n} \alpha_i Y_i$ et $B_n = \overline{Y_n} - A_n\overline{x}$, où le réel α_i a été défini dans la question 1.c).

a) Montrer que A_n et B_n sont des estimateurs sans biais de a et b respectivement.

b) Établir les formules suivantes : $V(A_n) = \dfrac{\sigma^2}{ns_x^2}$ et $V(B_n) = \left(1 + \dfrac{\overline{x}^2}{s_x^2}\right)\dfrac{\sigma^2}{n}$.

c) Calculer $\mathrm{Cov}(A_n, B_n)$.

5. Dans cette question uniquement, l'entier n n'est plus fixé. On suppose l'existence de $\lambda = \lim\limits_{n\to+\infty} \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i$ et

$\mu^2 = \lim\limits_{n\to+\infty} \dfrac{1}{n}\sum\limits_{i=1}^{n}(x_i - \overline{x})^2$, avec $(\lambda, \mu) \in \mathbb{R} \times \mathbb{R}_+^*$.

Montrer que les deux suites $(A_n)_{n\geqslant 3}$ et $(B_n)_{n\geqslant 3}$ convergent en probabilité vers a et b respectivement.

6.a) On pose pour tout i de $[\![1, n]\!]$: $\widehat{U}_i = Y_i - A_n x_i - B_n$. Calculer $E(\widehat{U}_i)$.

b) Établir l'égalité : $\sum\limits_{i=1}^{n} \widehat{U}_i^2 = \sum\limits_{i=1}^{n}(U_i - \overline{U}_n)^2 - n s_x^2 (A_n - a)^2$.

c) Calculer $E\left(\sum\limits_{i=1}^{n} \widehat{U}_i^2\right)$. En déduire un estimateur sans biais de σ^2.

Partie III. Hypothèse de normalité et prévision

Le contexte et les notations de cette partie sont ceux des parties I et II. De plus, on suppose dans cette partie que pour tout i de $[\![1, n]\!]$, la variable aléatoire U_i suit une loi normale $\mathcal{N}(0, \sigma^2)$.

On pose : $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$ et $U = \begin{pmatrix} U_1 \\ \vdots \\ U_n \end{pmatrix}$. Le modèle (1) de la partie II s'écrit alors matriciellement : $Y = M\theta + U$.

Soit $W_1, W_2, \ldots, W_q$ ($q \in \mathbb{N}^*$), q variables aléatoires réelles définies sur $(\Omega, \mathcal{A}, P)$. On définit le vecteur aléatoire $(W_1, W_2, \ldots, W_q)$ à valeurs dans $\mathbb{R}^q$, en associant à tout ω de Ω le vecteur $\big(W_1(\omega), W_2(\omega), \ldots, W_q(\omega)\big)$ de $\mathbb{R}^q$.
On dit que le vecteur aléatoire $(W_1, W_2, \ldots, W_q)$ est *normal* si pour tout q-uplet $(\rho_1, \rho_2, \ldots, \rho_q)$ de nombres réels, différent de $(0, 0, \ldots, 0)$, la variable aléatoire $\sum\limits_{i=1}^{q} \rho_i W_i$ suit une loi normale de variance non nulle.

Dans le cas où le vecteur $(W_1, W_2, \ldots, W_q)$ est normal, *on admet* que les variables aléatoires $W_1, W_2, \ldots, W_q$ sont mutuellement indépendantes si et seulement si pour tout (i, j) de $[\![1, q]\!]^2$ avec $i \neq j$, $\mathrm{Cov}(W_i, W_j) = 0$.

7.a) Montrer que le vecteur aléatoire $(Y_1, Y_2, \ldots, Y_n)$ est normal mais que le vecteur $(Y_1 - \overline{Y}_n, Y_2 - \overline{Y}_n, \ldots, Y_n - \overline{Y}_n)$ ne l'est pas.

b) Déterminer la loi de chacune des variables aléatoires A_n et B_n. Le vecteur aléatoire (A_n, B_n) est-il normal ?

8. Soit S une matrice inversible de $\mathcal{M}_n(\mathbb{R})$. On note T la matrice-colonne des composantes du vecteur aléatoire $(T_1, T_2, \ldots, T_n)$ telle que $T = SU$.

a) Montrer que le vecteur $(T_1, T_2, \ldots, T_n)$ est normal.

b) On suppose que la matrice S est orthogonale. Montrer que $T_1, T_2, \ldots, T_n$ sont mutuellement indépendantes.

9. Soit $\widehat{U}_1, \widehat{U}_2, \ldots, \widehat{U}_n$ les variables aléatoires qui ont été définies dans la question 6.

On note $\widehat{U}$ la matrice-colonne de composantes $\widehat{U}_1, \widehat{U}_2, \ldots, \widehat{U}_n$ définie par $\widehat{U} = Y - M\begin{pmatrix} A_n \\ B_n \end{pmatrix}$.

a) Montrer que $\widehat{U} = GU$, où la matrice G a été définie dans la question 3.

b) Justifier l'existence d'une matrice orthogonale R de $\mathcal{M}_n(\mathbb{R})$ et d'une matrice diagonale D de $\mathcal{M}_n(\mathbb{R})$, telles que $G = RD\,{}^tR$. Quels sont les éléments diagonaux de D ?

c) Soit Z la matrice-colonne de composantes $Z_1, Z_2, \ldots, Z_n$ définie par $Z = {}^tRU$. Quelle est la loi de $\sum\limits_{i=1}^{n-2} Z_i^2$?

d) En déduire que la variable aléatoire $\sum\limits_{i=1}^{n} \widehat{U}_i^2$ suit la loi $\Gamma\left(2\sigma^2, \dfrac{n-2}{2}\right)$.

e) Soit p un réel donné vérifiant $0 < p < 1$. Établir l'existence d'un réel c_n ne dépendant pas des paramètres inconnus a, b et σ^2, tel que $P\left(\left[\sum\limits_{i=1}^{n} \widehat{U}_i^2 \geqslant c_n \sigma^2\right]\right) = p$.

Tournez la page S.V.P.

Dans les questions 10 et 11, on suppose qu'une $(n+1)$-ième valeur de $\mathcal{X}$, notée x_{n+1}, est choisie mais que la valeur correspondante y_{n+1} de $\mathcal{Y}$ est inconnue. On suppose que y_{n+1} est la réalisation d'une variable aléatoire Y_{n+1} qui vérifie $Y_{n+1} = ax_{n+1} + b + U_{n+1}$, où les variables aléatoires $U_1, U_2, \ldots, U_{n+1}$ sont mutuellement indépendantes et de même loi $\mathcal{N}(0, \sigma^2)$.

10. On pose pour tout n-uplet $r = (r_1, r_2, \ldots, r_n)$ de $\mathbb{R}^n$: $\widehat{Y}_{n+1}^{(r)} = \sum_{i=1}^{n} r_i Y_i$.

L'ensemble $\{\widehat{Y}_{n+1}^{(r)} ; r \in \mathbb{R}^n\}$ est l'ensemble des "prédicteurs linéaires" de Y_{n+1}.

a) Soit g la fonction définie sur $\mathbb{R}^n$ à valeurs réelles, telle que pour tout $r = (r_1, r_2, \ldots, r_n)$ de $\mathbb{R}^n$, $g(r_1, r_2, \ldots, r_n) = \sum_{i=1}^{n} r_i^2$. On rappelle que pour tout i de $[\![1, n]\!]$: $\alpha_i = \dfrac{(x_i - \overline{x})}{n s_x^2}$.

Montrer que la fonction g admet un minimum absolu sous les contraintes $\sum_{i=1}^{n} r_i = 1$ et $\sum_{i=1}^{n} x_i r_i = x_{n+1}$, atteint en l'unique point $r^* = (r_1^*, r_2^*, \ldots, r_n^*)$, où pour tout i de $[\![1, n]\!]$, $r_i^* = \dfrac{1}{n} + (x_{n+1} - \overline{x})\alpha_i$.

b) Montrer que parmi les prédicteurs linéaires $\widehat{Y}_{n+1}^{(r)}$ de Y_{n+1}, qui vérifient $E(\widehat{Y}_{n+1}^{(r)}) = E(Y_{n+1})$ pour tout (a, b) de $\mathbb{R}^2$, $\widehat{Y}_{n+1}^{(r^*)}$ est celui qui a la plus petite variance.

Vérifier que $\widehat{Y}_{n+1}^{(r^*)} = A_n x_{n+1} + B_n$.

11.a) Déterminer la loi de la variable aléatoire $Y_{n+1} - (A_n x_{n+1} + B_n)$.

b) On note Φ la fonction de répartition de la loi $\mathcal{N}(0, 1)$. Soit p un réel donné vérifiant $\dfrac{1}{2} < p < 1$.

Justifier l'existence d'un réel d_n, que l'on exprimera à l'aide de Φ^{-1}, ne dépendant pas de a, b et σ^2, tel que $P(\{|Y_{n+1} - (A_n x_{n+1} + B_n)| \leqslant d_n \sigma\}) = p$.

c) En déduire, à l'aide de la question 9.e), un intervalle dont les bornes ne dépendent que des $(Y_i)_{1 \leqslant i \leqslant n}$, des $(x_i)_{1 \leqslant i \leqslant n+1}$, de c_n et d_n, qui contienne Y_{n+1} avec une probabilité supérieure ou égale à $2p - 1$.

S'agit-il d'un intervalle de confiance au sens usuel du terme ?