UPS TOULOUSE III **Le 15/01/2004**
MAGISTÈRE

EXAMEN
MODÉLISATION STATISTIQUE

Durée 3 heures

PROBLÈME I

7 points

L'hypermarché Euromarket souhaite commercialiser un nouveau crédit par l'intermédiaire de sa carte de paiement et cherche un moyen de vérifier les revenus des clients qui sollicitent ce crédit. Une méthode possible consiste à établir une relation théorique fiable entre le revenu et l'âge des clients. En appliquant cette formule à un client quelconque, on pourra calculer une valeur approximative de son revenu en fonction de son âge et valider ou non sa déclaration. Euromarket dispose des résultats suivants où x_i représente l'âge du client sollicitant le nouveau crédit et y_i son revenu annuel exprimé en milliers d'Euros.

x_i	36	52	55	41	37	43	40	28	49	30
y_i	249	264	267	262	250	255	262	221	266	230

x_i	53	42	28	32	44	39	35	40	46	45
y_i	267	259	224	243	268	259	250	260	268	265

On suppose que chaque y_i est une réalisation d'une variable aléatoire Y_i satisfaisant

$$Y_i = a + bx_i + cx_i^2 + \varepsilon_i, \qquad i = 1, 2, \cdots, n$$

où (ε_i) est une suite de variables aléatoires indépendantes et de même loi $\mathcal{N}(0, \sigma^2)$.

1) Donner une condition simple pour que cette régression soit identifiable.

2) Estimer les paramètres inconnus a, b, c et σ^2.

3) Proposer un intervalle de confiance à 95% pour σ^2.

4) Tester l'hypothèse selon laquelle la régression est linéaire.

5) Pour une nouvelle valeur $x^* = 60$, calculer la prédiction naturelle $\widehat{Y}^*$ de Y^* et trouver un intervalle de prévision à 95% pour Y^*.

PROBLÈME II

6 points

Un laboratoire de recherche spécialisé dans la sécurité des transports aériens veut valider ses techniques d'entrainement pour les contrôleurs aériens. Il décide de mener une étude sur cinq techniques utilisées pour l'entrainement de 40 contrôleurs soumis à des simulations de trafic intense impliquant des collisions potentielles. Il obtient les résultats suivants pour les cinq techniques d'entrainement notées $T_1, T_2, \cdots, T_5$ où un score élevé indique une meilleure attention.

T_1	10.2	10.5	9.9	11.0	9.2	10.8	9.5	10.4	
T_2	10.8	11.2	10.7	11.3	12.5	11.4	11.0	9.9	10.5
T_3	12.4	11.5	12.1	13.0	11.7	10.8	11.6	12.2	
T_4	10.5	10.8	9.8	10.7	10.4	10.2	11.5		
T_5	10.2	11.1	12.7	10.2	10.9	11.5	12.4	10.9	

Soit y_{ij} le score obtenu avec la technique d'entrainement T_i par le j^e contrôleur. On suppose que pour $i = 1, 2, \cdots, 5$, y_{ij} est une réalisation d'une variable aléatoire Y_{ij} satisfaisant

$$Y_{ij} = m_i + \varepsilon_{ij}, \qquad i = 1, 2, \cdots, n_i$$

où (ε_{ij}) est une suite de variables aléatoires indépendantes et de même loi $\mathcal{N}(0, \sigma^2)$.

1) Estimer les paramètres inconnus m et σ^2.

2) Tester l'hypothèse d'égalité des cinq techniques d'entrainement en dressant le tableau d'analyse de la variance à un facteur.

PROBLÈME III

7 points

On appelle carré latin d'ordre $n \geq 1$, un tableau Δ_n carré d'ordre n constitué de n^2 nombres entiers choisis entre 1 et n, tel que chaque ligne et chaque colonne de Δ_n soit formée par une permutation de $\{1, 2, \cdots, n\}$. Voici par exemple un carré latin d'ordre 5.

$i \setminus j$	1	2	3	4	5
1	4	5	3	1	2
2	5	1	4	2	3
3	3	4	2	5	1
4	2	3	1	4	5
5	1	2	5	3	4

On souhaite étudier le modèle d'analyse de la variance à trois facteurs sans intéraction

$$Y_{ijk} = \mu + a_i + b_j + c_k + \varepsilon_{ijk} \qquad (i,j,k) \in \Delta_n$$

où (ε_{ijk}) est une suite de variables aléatoires indépendantes et de même loi $\mathcal{N}(0, \sigma^2)$. On suppose également que les contraintes standards sur la décomposition de la moyenne sont réalisées, à savoir $a_* = b_* = c_* = 0$.

1) Déterminer les estimateurs des moindres carrés des paramètres μ, a, b, c et σ^2.

2) Montrer que l'on a la décomposition $S_T = S_a + S_b + S_c + S_R$ avec

$$S_T = \sum_{(i,j,k)\in\Delta_n} (Y_{ijk} - Y_{***})^2, \qquad S_R = \sum_{(i,j,k)\in\Delta_n} (Y_{ijk} - \widehat{\mu} - \widehat{a}_i - \widehat{b}_j - \widehat{c}_k)^2$$

$$S_a = n\sum_{i=1}^{n} \widehat{a}_i^2, \qquad S_b = n\sum_{j=1}^{n} \widehat{b}_j^2, \qquad S_c = n\sum_{k=1}^{n} \widehat{c}_k^2.$$

3) Dresser le tableau d'analyse de la variance associé à ce modèle à trois facteurs.

4) Proposer un test d'absence d'effet pour chacun des 3 facteurs.

5) On dispose de 5 classes d'étudiants et l'on souhaite étudier les performances de 5 méthodes d'enseignement. Afin de ne pas rédiger 25 sujets d'examen différents pour les 5 classes et les 5 méthodes d'enseignement, on adopte le dispositif en carré latin donné ci-dessus. Ainsi, la i^e classe pour la j^e méthode d'enseignement recevra le k^e sujet d'examen où k est le nombre entier figurant dans la case (i, j). On obtient alors les résultats suivants où y_{ijk} représente la note d'un étudiant choisi au hasard dans la i^e classe ayant suivi la j^e méthode d'enseignement et reçu le k^e sujet d'examen.

5	16	13	8	7
12	4	11	10	5
12	13	10	20	5
3	8	5	10	14
7	7	15	11	9

Par exemple, $y_{114} = 5$, $y_{125} = 16$, $y_{215} = 12$. Tester l'hypothèse d'absence d'effet pour chacun des 3 facteurs.