Examen terminal

Durée 2 heures. Epreuve sans calculatrice ni documents, à l'exception des tables de quantiles. Toutes les réponses doivent être justifiées. Pensez à numéroter les copies si vous en rendez plusieurs.

Exercice 1 (7 points)**.**

1. Soit Z un vecteur aléatoire gaussien de loi $\mathcal{N}_q(\mu, \Sigma)$. Soit $b \in \mathbb{R}^n$ un vecteur et $A \in M_{n,q}(\mathbb{R})$ une matrice. Déterminer la loi de $b + AZ$. On justifiera en détail la formule proposée.
2. On considère un modèle linéaire de la forme suivante : pour $i = 1, \ldots, n$,

$$Y_i = \theta_1 x_i^1 + \theta_2 x_i^2 + \cdots + \theta_p x_i^p + \sigma G_i,$$

où les G_i sont des variables gaussiennes centrées réduites indépendantes et x_i^j représente la valeur de la j-ème variable explicative pour le i-ème individu.

   (a) Ecrire le système sous la forme $Y = X\theta + \sigma G$ en précisant les objets et leur taille.

   (b) Dans la suite on suppose que X est de rang p. Rappeler la formule pour l'estimateur $\widehat{\theta}$ de θ et expliquer brièvement pourquoi on l'appelle "estimateur des moindres carrés".

   (c) Donner la formule pour les estimateurs $\widehat{Y}$ et $\widehat{\sigma}^2$ ainsi que leur interprétation géométrique en fonction de Y.

   (d) En utilisant la première question, déterminer la loi de $\widehat{\theta}$ puis celle de $\widehat{\theta}_1$.

   (e) Proposez un intervalle de confiance de niveau de sûreté $1 - \alpha$ pour θ_1 en supposant que σ est connu. On justifiera la construction.

   (f) Même question lorsque la valeur de σ est inconnue.

Exercice 2 (6 points)**.** Une régression linéaire multiple portant sur quatre variables explicatives X_1, X_2, X_3, X_4 a été effectuée sur 20 observations. Les calculs ont donné une somme totale des carrés de 86996 et une somme des carrés résiduelle de 1426.

1. Est-ce que la regression est significative dans son ensemble ? Indication : effectuer un test de niveau $\alpha = 0,05$ de l'hypothèse $\theta_1 = \theta_2 = \theta_3 = \theta_4 = 0$. On détaillera soigneusement toutes les étapes.
2. Un de vos collègues prétend que les variables X_3 et X_4 sont inutiles pour prédire la variable réponse. Afin de vérifier ses dires, vous effectuez une regression utilisant seulement X_1 et X_2, ce qui donne une somme des carrés résiduelle de 24013. Effectuez un test afin de dire si l'affirmation de votre collègue est vraisemblable.
3. Expliquer comment calculer la P-valeur pour le test précécent (le calcul n'est pas demandé). Le résultat obtenu sera-t-il grand ou petit ?

4. Expliquer la structure géométrique des tests de Fisher.

Exercice 3 (9 points)**.** Un psychologue a voulu étudier la vitesse de lecture d'enfants entre 6 et 8 ans. Pour cela, il a mesuré le temps écoulé (en centièmes de seconde) entre le moment où le mot est présenté à l'enfant et le moment où il est correctement épelé. On cherche maintenant à étudier cette variable selon deux facteurs :

- le sexe (1 = garçon ; 2 = fille)
- l'âge (1 = fin de C.P. ; 2 = C.E.1)

Pour chacune des 4 configurations définies par ces 2 facteurs, 6 observations de la variable réponse ont été réalisées. Sur les 24 observations, on a estimé un temps moyen de lecture de 1184 cs, avec une variance de 80725.
On désigne par y_{ijk} le temps réalisé par le k^e enfant de sexe i et d'âge j. On suppose que pour $i \in \{1,2\}$, $j \in \{1,2\}$ et $k \in \{1,\ldots,6\}$,

$$y_{ijk} = m_{ij} + e_{ijk} = \mu + a_i + b_j + c_{ij} + e_{ijk}$$

où e_{ijk} est la réalisation d'une v.a. $E_{ijk} \sim \mathcal{N}(0,\sigma^2)$ et les E_{ijk} sont indépendantes.
La mise en œuvre de ce modèle a donné les estimations suivantes :

$$\widehat{\mu}(y) = 1184 \qquad \widehat{a}_1(y) = 50 \qquad \widehat{b}_1(y) = 220 \qquad \widehat{c}_{11}(y) = 100$$

1. Quelles sont les caractéristiques du plan d'expérience ?
2. Préciser les contraintes associées à cette paramétrisation. En déduire les estimations des autres paramètres intervenant dans cette paramétrisation.
3. A partir des données ci-dessus, trouver les estimations des paramètres m_{ij}.
4. Représenter le diagramme d'interactions. Commenter.
5. Calculer les sommes des carrés expliquées par l'interaction (SSI).
6. Expliquer en détail comment calculer, à partir des données disponibles, les quantités $SS1$, $SS2$ et SSR. On ne demande pas le calcul final de l'application numérique, qui donne $SS1 = 60000$, $SS2 = 1161600$ et $SSR = 475800$.
7. Tester l'hypothèse d'absence d'intéraction entre les deux facteurs (les ordres de grandeur suffisent pour conclure).
8. Testez l'absence d'effet principal de chacun des facteurs.
9. Proposez un sous-modèle pertinent et calculer approximativement son coefficient de détermination.