Un sujet pour l'épreuve A
(raisonnement et calcul)

Présentation

Le texte proposé ci-après est conçu pour l'épreuve A, portant plus particulièrement sur le calcul et le raisonnement ; ce texte représente la moitié d'un sujet (l'épreuve A dure 2 heures 30).
Le thème choisi est celui d'une modélisation de prises de mesures dans la perspective d'un diagnostic ; la modélisation est ici déjà terminée, de sorte que l'enjeu du sujet est d'en tirer les conséquences par le calcul avant d'aboutir à la prise de décision espérée.

On suggère enfin que plusieurs questions proposées sont très proches des sujets qui seront abordés en classe (ajustement affine, lois gaussiennes) ; le présent texte a donc un rôle essentiellement didactique.

Les compétences mobilisées dans ce problème sont essentiellement les suivantes :
- ▷ Raisonner, démontrer, argumenter.
  Cette compétence est notamment présente dans les questions B2 et B4.
- ▷ Calculer, maîtriser le formalisme mathématique.
  Cette compétence est présente dans de nombreuses questions.
- ▷ Mobiliser des connaissances scientifiques pertinentes.
  Cette compétence est présente dans les questions B2, D1, D2.
- ▷ Communiquer à l'écrit et à l'oral : compétence présente, par nature, dans l'ensemble des questions.

On rappelle que l'emploi d'une calculatrice n'est pas autorisé dans cette épreuve.

Énoncé

A. Contexte et but de l'étude

On souhaite élaborer un test simple et rapide de recherche d'une maladie chez des patients. L'objectif est d'organiser des dépistages automatiques, afin d'identifier des individus pour lesquels on a estimé un risque. On mesure facilement un symptôme de la maladie par des prélèvements sanguins : si un individu est infecté, la concentration d'une molécule diminue à vitesse très rapide au cours du temps, alors qu'il reste stable pour un patient sain.
Notons K_t la concentration de la molécule étudiée chez un patient au temps t (mesuré en jours, $t \geq 0$) et posons $Y_t = \ln\left(\frac{K_t}{K_0}\right)$, K_0 étant une concentration de référence correspondant à l'état « normal » pour un individu « moyen ».

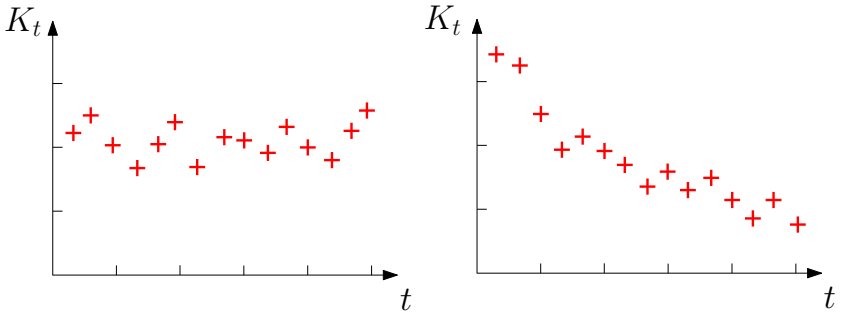

FIGURE 1 – Relevé des concentrations chez des patients sains (gauche) et infectés (droite).

Nous voulons mettre en place un procédé qui permettra d'identifier si un patient est :

▷ **sain :** le logarithme de la concentration de la molécule chez le patient est constant à des variations naturelles près

$$Y_t = b + X_t\,,$$

où t est la variable de temps et les X_t sont des variables aléatoires pour tout t ;

▷ **ou infecté :** le logarithme de la concentration évolue suivant le modèle

$$Y_t = b - at + X_t\,,$$

où a et b sont des constantes strictement positives.

Il s'agit donc de déterminer si la constante a peut être considérée comme nulle ou non. En pratique, on ne dispose que d'un nombre très limité de prélèvements (N est de l'ordre de 4 ou 5 par exemple), pour des raisons de coût mais aussi car on souhaite pouvoir décider si un patient est sain ou malade dans un délai assez court. On développe donc dans ce problème une méthode probabiliste pour répondre au problème posé.
Notation : les mesures sont effectuées aux instants $t_1 < \cdots < t_N$, et fournissent des valeurs $Y_{t_1}, \ldots, Y_{t_N}$. Pour simplifier, on pose $y_k = Y_{t_k}$.

B. Méthode par régression linéaire (ou ajustement affine)

On introduit la fonction de deux variables

$$f(\alpha, \beta) = \sum_{k=1}^{N} (y_k - \alpha t_k - \beta)^2\,.$$

Le but de la méthode consiste à minimiser $f(\alpha, \beta)$.
On pose

$$\overline{t} = \frac{1}{N}\sum_{k=1}^{N} t_k,\quad \overline{Y} = \frac{1}{N}\sum_{k=1}^{N} y_k\,,$$

$$\sigma_t^2 = \frac{1}{N}\sum_{k=1}^{N}(t_k-\overline{t})^2,\ \ \sigma_{t,Y} = \frac{1}{N}\sum_{k=1}^{N}(t_k-\overline{t})(y_k-\overline{Y}).$$

On admet provisoirement que la fonction f admet un minimum pour des valeurs respectives α_0 et β_0 de α et β.

B.1. Exprimer $\sum_{k=1}^{N} t_k^2$ et $\sum_{k=1}^{N} t_k Y_{t_k}$ en fonction de $\overline{t}, \overline{Y}, \sigma_t^2$ et $\sigma_{t,Y}$.

B.2. Montrer que

$$\alpha_0 = \frac{\sigma_{t,Y}}{\sigma_t^2} \quad \beta_0 = \overline{Y} - \alpha_0 \overline{t}.$$

Réciproquement, on prend les formules de la question **B.2** comme définitions, et on réexamine à partir de là le problème de minimisation de la fonction f. On pose dans cette optique : $\alpha = \alpha_0 + r$, $\beta = \beta_0 + s$.

B.3. Exprimer $f(\alpha,\beta) - f(\alpha_0,\beta_0)$ en fonction de r et s.

B.4. En déduire que la fonction f présente un unique minimum sur $\mathbf{R}^2$, situé au point (α_0,β_0).

On considère désormais la droite $y = \alpha_0 x + \beta_0$ comme « proche » de la droite $y = -ax + b$.

C. Moyennes et variances des estimateurs

Compte tenu d'un espacement suffisant entre les prises de sang successives, on suppose que les variables aléatoires $X_{t_1},\ldots,X_{t_N}$ sont centrées, de même variance et décorrélées entre elles, c'est à dire qu'il existe une constante σ telle que

$$\mathbb{E}[X_t] = 0, \mathbb{E}[X_t^2] = \sigma^2, \mathbb{E}[X_t X_{t'}] = 0,\ \forall t \neq t'.$$

Pour simplifier un peu, on pose $X_k = X_{t_k}$.
On pose alors

$$\overline{X} = \frac{1}{N}\sum_{k=1}^{N} X_k, \quad \sigma_{t,X} = \frac{1}{N}\sum_{k=1}^{N}(t_k-\overline{t})(X_k-\overline{X}).$$

C.1. Calculer $\mathrm{Cov}(X_k,\overline{X})$, $\mathbb{V}(\overline{X})$ et $\mathbb{V}(\overline{Y})$.

C.2. Montrer que

$$\sigma_{t,Y} = \sigma_{t,X} - a\sigma_t^2$$

et en déduire que

$$\alpha_0 = -a + \frac{\sigma_{t,X}}{\sigma_t^2}.$$

C.3. Montrer que $\mathbb{E}[\alpha_0] = -a$.

C.4. Montrer que

$$\sigma_{t,X} = \frac{1}{N}\sum_{k=1}^{N}(t_k-\overline{t})X_k$$

et en déduire que $\mathbb{V}(\alpha_0) = \frac{\sigma^2}{N\sigma_t^2}$

D. Lois des estimateurs lorsque le bruit est gaussien

On a peu de connaissances sur le mécanisme biologique qui dicte les fluctuations naturelles de la concentration de la molécule, et donc la loi des variables aléatoires X_t. Par contre, on soupçonne que la fluctuation à un temps donné est due à un grand nombre de phénomènes de très faible intensité et d'origine indépendantes.

D.1. Énoncer un théorème qui justifie que l'on puisse supposer avec une bonne approximation que pour tout t la variable aléatoire X_t est gaussienne (ou normale).

On suppose désormais que les variables aléatoires $X_{t_1},\ldots,X_{t_N}$ sont des variables gaussiennes centrées, indépendantes et de même variance σ^2.

D.2. On suppose, dans cette question, que Z_1 et Z_2 sont deux variables aléatoires gaussiennes indépendantes, de moyennes respectives m_1 et m_2 et de variances respectives σ_1^2 et σ_2^2. Donner (sans justification particulière) la loi de la variable $Z_1 + Z_2$.

D.3. En utilisant l'expression de α_0 de issue de la question **C.2** et l'expression de $s_{t,X}$ issue de la question **C.4**, en déduire la loi de α_0.

D.4. Le nombre C tel que

$$\int_{-\infty}^{C} \frac{1}{\sqrt{2\pi}} \exp(-\frac{t^2}{2}) \mathrm{d}t = 0,75$$

est approximativement $C = 0,67$. Calculer la probabilité de l'évènement

$$a \leqslant 0,67 \times \frac{\sigma}{N\sigma_t} - \alpha_0.$$

D.5. Il est admis qu'un patient n'a pas un grand risque d'infection si on enregistre chez lui une valeur a inférieure à une valeur seuil a_s. En supposant la valeur de σ connue, décrire une méthode qui permette de dire qu'avec 75% de chance un patient n'a pas un grand risque d'infection.