Université Paris-Sud
Département de mathématiques

Préparation à l'agrégation
Option Probabilités-Statistiques

Devoir de Statistiques

Les questions notées (*) sont facultatives

Exercice 1.

On considère $\beta \in \mathbb{R}^d$ et Σ une matrice de covariance de taille d. Pour tout vecteur $v \in \mathbb{R}^d$, on note $\|v\|^2 = \sum_{k=1}^d v_k^2$. Soit Q une matrice carrée telle que $Q\,{}^tQ = \Sigma$.

(1) Rappeler brièvement pourquoi une telle matrice Q existe.

(2) Soit Z un vecteur gaussien de loi $\mathcal{N}(0, I_d)$. Quelle est la loi de $QZ + \beta$? En déduire une méthode de simulation de la loi $\mathcal{N}(\beta, \Sigma)$ à partir d'un d-échantillon de loi $\mathcal{N}(0, 1)$.

(3) Quel est le support de la loi $\mathcal{N}(0, \Sigma)$? Que peut-on dire quand Σ n'est pas inversible ?

(4) On suppose désormais que Σ est inversible. Soit $X \sim \mathcal{N}(\beta, \Sigma)$. Montrer que Q est inversible, et donner la loi de $\|Q^{-1}(X - \beta)\|^2$.

(5) Soit maintenant M une matrice (non nécessairement carrée) telle que ${}^tMM = \Sigma^{-1}$. En utilisant la question précédente, trouver la loi de $\|M(X - \beta)\|^2$.

(6) Soit $X^{(1)}, \ldots, X^{(n)}$ n tirages indépendants de même loi que X. Simulez avec `Scilab` un tel tirage pour $d = 2$, $\Sigma = [5, 1; 1, 1]$, $n = 1000$, et β à votre guise. Quelle est la forme du nuage de points ? Essayer également avec Σ non inversible.

(7) Quand n est grand, quel k faut-il prendre pour qu'environ 95% des points soient dans la région $\mathcal{E} = \{x \in \mathbb{R}^d, \|M(x - \beta)\|^2 \leq k\}$?

(8) On va montrer que $\mathcal{E}$ est (l'intérieur d') un ellipsoide (une ellipse quand $d = 2$). La matrice Σ^{-1} peut se diagonaliser de la façon suivante : $\Sigma^{-1} = {}^tPDP$, avec P orthogonale et D diagonale. Écrire $\|M(x - \beta)\|^2$ en fonction de P, D, x et β.

(9) Quelle équation vérifie u quand $\beta + P^{-1}u$ appartient à $\mathcal{E}$? Conclure.

(10*) Dans le cas $d = 2$, quels sont les paramètres (demi grand axe et demi petit axe) de l'ellipse, en fonction des valeurs propres de Σ ?

(11*) Dans le cas $d = 2$ donner un paramétrage de la frontière de $\mathcal{E}$, c'est à dire l'écrire comme une courbe paramétrée.

(12*) Tracer l'ellipse dans l'exemple de la question (6). On pourra utiliser la fonction `spec`.

(13) Donner la densité de la loi $\mathcal{N}(\beta, \Sigma)$. Quelles en sont les lignes de niveau ? Utiliser la fonction `contour` pour les tracer dans l'exemple de la question (6).

Exercice 2.

On considère le modèle de régression linéaire simple suivant : soit $x \in \mathbb{R}^n$ un vecteur, et $Y \in \mathbb{R}^n$ un vecteur aléatoire tel que Y_i suit une loi normale $\mathcal{N}(\beta_1 x_i + \beta_2, \sigma^2)$, où β_1, β_2 et σ^2 sont des paramètres inconnus à estimer. On observe les variables indépendantes $Y_1, \ldots, Y_n$.

On note $\beta = {}^t(\beta_1, \beta_2)$, $x = {}^t(x_1, \ldots, x_n)$, $Y = {}^t(Y_1, \ldots, Y_n)$, $\mathbb{1} = {}^t(1, \ldots, 1)$. Pour tous vecteurs $v, w \in \mathbb{R}^n$, on note $var(v) = \overline{v^2} - \overline{v}^2$, $cov(v, w) = \overline{vw} - \overline{v}\,\overline{w}$, où

$$\overline{v} = n^{-1} \sum_{i=1}^{n} v_i, \quad \overline{vw} = n^{-1} \sum_{i=1}^{n} v_i w_i, \quad \text{ et } \|v\|^2 = \sum_{i=1}^{n} v_i^2.$$

On suppose que $var(x) \neq 0$.

(1) Donner le modèle statistique.

(2) Écrire la log vraisemblance associée $L(\beta_1, \beta_2, \sigma^2)$

(3) Maximiser $(\beta_1, \beta_2) \mapsto L(\beta_1, \beta_2, \sigma^2)$ et en déduire des estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ de β_1 et β_2. On note $\hat{\beta} = {}^t(\hat{\beta}_1, \hat{\beta}_2)$.

(4) Par dérivation de $\sigma^2 \mapsto L(\beta_1, \beta_2, \sigma^2)$, trouver l'estimateur du maximum de vraisemblance de σ^2, qu'on notera S.

Pour la suite, on admettra que S est indépendant de $\hat{\beta}$, et que $nS/\sigma^2 \sim \chi^2(n-2)$.

(5) Donner la loi de $\hat{\beta}$ (on pourra trouver une matrice A telle que $\hat{\beta} = AY$), et celle de $\hat{\beta}_1$.

(6) Calculer le biais de $\hat{\beta}$.

(7) Quel est le biais de S ? Pourquoi choisit-on plutôt $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} (Y_i - \hat{\beta}_1 x_i - \hat{\beta}_2)^2$?

(8) Donner un intervalle de confiance de niveau $1 - \alpha$ pour σ^2.

(9*) Dans cette question, on suppose que pour $n = 100$, on a obtenu comme réalisation de l'estimateur $\hat{\sigma}^2$ la valeur 1.2. Tester au niveau 5% que $\sigma^2 = 1$.

(10) Donner un intervalle de confiance de niveau $1 - \alpha$ pour β_1. On rappelle que si Z suit une loi $\mathcal{N}(0, 1)$ et V une loi $\chi^2(d)$ indépendante de Z, alors $Z/\sqrt{V/d}$ suit une loi de Student $\mathcal{T}(d)$.

(11) En utilisant la méthode de Bonferroni, trouver une région de confiance pour β.

(12) En utilisant l'exercice 1, donner la loi de $\|M(\hat{\beta} - \beta)\|^2$, pour $M = \sigma^{-1}(x\ \mathbb{1})$ (matrice de taille $n \times 2$).

(13) En déduire une région de confiance elliptique pour β. On rappelle que si Z suit une loi $\chi^2(d_1)$ et V une loi $\chi^2(d_2)$ indépendante de Z, alors $(Z/d_1)/(V/d_2)$ suit une loi de Fisher $\mathcal{F}(d_1, d_2)$.

(14) Quelle est la meilleure région de confiance pour β entre celle de la question (11) et celle de la question (13) ?

(15) Trouvez une fonction de la Stixbox qui permet d'estimer β et donne des intervalles de confiance sur β_1 et β_2.