TD 5 : vecteurs gaussiens et modèle linéaire

Les questions marquées d'un astérisque **(*)** *sont facultatives.*

Exercice 1. (p-valeurs, suite du TD 4) Quelle est la p-valeur des deux tests suivants ?

1. On compte le nombre de fois qu'on tire une boule noire lors de 300 tirages dans un urne contenant une boule noire et des boules blanches. On souhaite tester s'il y a 19 boules blanches ou s'il y en a plus. On en a tiré 9.

k	6	7	8	9	10	11	12
$\mathbb{P}(X \leq k)$	0,0066	0,016	0,0341	0,065	0,1123	0,178	0,2612

FIGURE 1 – Fonction de masse et de répartition d'une variable aléatoire $X \sim \text{Bin}(300; 0,05)$.

2. Des chercheurs ont développé une nouvelle espèce de sardines d'élevage avec pour objectif de maximiser la masse de viande obtenue par poisson. La masse produite par une sardine adulte de l'ancienne espèce était de 0,113 kg. Sur 25 sardines de la nouvelle espèce, on a mesuré une masse moyenne produite de 0,145 kg et un écart-type de 0,073 kg. On souhaite tester si la masse moyenne de la nouvelle espèce est différente de celle de l'ancienne ou si elle est supérieure.

α	0,95	0,96	0,97	0,98	0,981	0,99
`qt(`α`,24)`	1,711	1,828	1,974	2,172	2,196	2,492

FIGURE 2 – Quantiles de la loi $\mathcal{T}_{24}$.

Exercice 2. (matrices définies positives)

1. Rappeler la définition d'une matrice symétrique définie positive. D'une matrice symétrique semi-définie positive.
2. Les matrices suivantes sont-elles symétriques définies positives ? Symétriques semi-définies positives ?

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \quad \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \quad \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \quad \begin{pmatrix} 1 & -2 \\ -2 & 2 \end{pmatrix}$$

Exercice 3. Corrigez tout ce qui est faux dans le raisonnement suivant :

« *Soient X et Y deux v.a. gaussiennes d'espérance 1 telles que $Var(X) = 1$, $Var(Y) = 2$ et $Cov(X, Y) = -2$. Alors (X, Y) est un vecteur gaussien d'espérance $(1, 1)$ et de matrice de covariance*

$$\begin{pmatrix} 2 & -2 \\ -2 & 1 \end{pmatrix}$$

donc le vecteur $(X + Y, X - Y)$ est un vecteur gaussien d'espérance $(2, 0)$ et de matrice de covariance

$$\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} 2 & -2 \\ -2 & 1 \end{pmatrix} = \begin{pmatrix} 0 & -1 \\ 4 & -3 \end{pmatrix} \text{ ».}$$

(*) De telles variables existent-elles ?

Exercice 4. (Modèle linéaire) (Rappel) Un *modèle linéaire* (multivarié) est un modèle statistique dans lequel un vecteur d'observations $\mathbf{Y} = (Y_i)_{1\leq i\leq d}$ s'exprime en fonction de variables explicatives $\mathbf{X} = (X_{ij})_{1\leq i\leq d, 1\leq j\leq p}$ *via* la relation

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon,$$

où ε est un vecteur aléatoire de dimension d (le bruit) et β est un vecteur de taille p. La matrice $\mathbf{X}$ et le vecteur $\mathbf{Y}$ sont supposés observés. L'enjeu est généralement d'estimer β. Si $d = 1$, ce modèle s'écrit plus simplement

$$Y = \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon.$$

Dans les questions qui suivent, on suppose qu'on observe une variable aléatoire Y et des variables explicatives (X_1, X_2). Lesquels des modèles suivants pouvez-vous réécrire sous forme de modèle linéaire ? Précisez la transformation des X et Y effectuée.

1. $Y = w_1 X_1 + w_2 X_1^2 + w_3 X_1 X_2^2 + \varepsilon$,
2. $Y = \sqrt{w_1 + w_2 X_1 + w_3 X_2 + \varepsilon}$,
3. $Y = X_1^{w_1} X_2^{w_2} 4^{w_3} \varepsilon$,
4. $\log(Y) = w_1 X_1 + w_2 \log(X_2) + \varepsilon$,
5. On observe une suite $Y^{(1)}, \ldots, Y^{(t)}$ et $X^{(1)}, \ldots, X^{(t)}$ dépendant de l'instant t de la manière suivante : pour tout $s \in \{1, \ldots, t\}$, $Y^{(s)} = w_1 \cos(\pi s/6) + w_2 s + w_3 X^{(s)} + \varepsilon^{(s)}$.

Exercice 5. (théorème de Cochran et applications) Le théorème de Cochran s'énonce comme suit. Soit $X \sim \mathcal{N}(0, \sigma^2 I_d)$ un vecteur gaussien (*la matrice de covariance* $= \sigma^2 I_d$ *est important !*). Soit F un sous-espace vectoriel de $\mathbb{R}^d$ de dimension p. Notons P_F la projection orthogonale sur F et $P_{F^\perp}$ la projection orthogonale sur $F^\perp$. Alors

(a) $P_F X$ et $P_{F^\perp} X$ sont indépendantes, $P_F X \sim \mathcal{N}(0, \sigma^2 P_F)$ et $P_{F^\perp} X \sim \mathcal{N}(0, \sigma^2 P_{F^\perp})$;

(b) $\|P_F X\|^2$ et $\|P_{F^\perp} X\|^2$ sont indépendantes, $\frac{1}{\sigma^2}\|P_F X\|^2 \sim \chi^2(p)$ et $\frac{1}{\sigma^2}\|P_{F^\perp} X\|^2 \sim \chi^2(d-p)$.

1. Soit $n \geq 2$ un entier et $\mathbf{X} = (X_1, \ldots, X_n)$ un vecteur de v.a. i.i.d. de loi $\mathcal{N}(0, \sigma^2)$. Soit F le sous-espace vectoriel de $\mathbb{R}^n$ engendré par le vecteur $(1, 1, \ldots, 1)$.

   (a) $\mathbf{X}$ est-il un vecteur gaussien ?

   (b) Calculer la projection orthogonale de $\mathbf{X}$ sur F.

   (c) Calculer la projection orthogonale de $\mathbf{X}$ sur $F^\perp$.

   Notons $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ et $\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$.

   (d) Quelle est la loi de $(n-1)\hat{\sigma}_n^2$?

   (e) Montrer que $\bar{X}_n$ et $\hat{\sigma}_n^2$ sont indépendants.

   Soit $d \geq 1$ un entier, $Z \sim \mathcal{N}(0, 1)$ et $U \sim \chi^2(d)$ avec Z et U indépendantes. La loi de Student à d degrés de liberté $\mathcal{T}(d)$ est définie comme la loi de $\dfrac{Z}{\sqrt{U/d}}$.

   (f) Montrer que $\dfrac{\bar{X}_n}{\hat{\sigma}_n/\sqrt{n}}$ suit la loi de Student à $n-1$ degrés de liberté.

2. Suite à une panne, un laboratoire de biologie a remplacé un appareil de mesure. Il aimerait s'assurer que l'incertitude de mesure du nouvel appareil est la même que celle du précédent. L'écart-type de mesure du précédent appareil était de 0,7. Sur 100 mesures avec le nouvel appareil, on estime un écart-type de 1,12. Le laboratoire a-t-il des raisons de penser que les écart-types des deux appareils sont différents ?

ordre du quantile	0,025	0,05	0,95	0,975
$\chi^2(99)$	73,4	77,0	123,2	128,4
$\chi^2(100)$	74,2	77,9	124,3	129,6
$\chi^2(101)$	75,1	78,8	125,5	130,7

FIGURE 3 – Quelques quantiles de lois du χ^2

Exercice 6. (*) Soient $X_1, \ldots, X_n$ des variables aléatoires i.i.d. de loi $\mathcal{N}(\mu_1, \sigma_1^2)$ et $Y_1, \ldots, Y_m$ des variables aléatoires i.i.d. de loi $\mathcal{N}(\mu_2, \sigma_2^2)$ indépendantes des X_i. Notez que m et n ne sont pas forcément égaux. Soient μ_0 et σ_0 deux réels.

Quelles sont les statistiques de test, leur loi sous l'hypothèse nulle et la région de confiance pour les tests suivants ?

1. $\mu_1 = \mu_0$ vs. $\mu_1 \neq \mu_0$ (quand σ_1 connu et quand σ_1 inconnu) ;
2. $\mu_1 = \mu_2$ vs. $\mu_1 > \mu_2$ (quand σ_1 et σ_2 connus et quand σ_1 et σ_2 inconnus et supposés égaux)
   (*) Et quand les variances sont inconnues et différentes (Test t de Welch) ? Connaît-on exactement la loi de la statistique de test dans ce cas ?
3. $\sigma_1 = \sigma_0$ vs. $\sigma_1 > \sigma_0$ (quand μ_1 inconnu) ;
4. $\sigma_1 = \sigma_2$ vs. $\sigma_1 \neq \sigma_2$ (quand μ_1 et μ_2 inconnus).