

STATISTIQUES (STA1) – RECUEIL D’EXERCICES

Luca Ganassali

ENSTA – Université Paris-Saclay

(Last update: August 5, 2026)

1. Rappels de statistiques : estimation, tests et intervalles de confiance

Objectifs : Retravailler les notions d'estimation, de tests et d'intervalles de confiance. Les exercices 1.1 à 1.3 sont à faire pendant le TD, les suivants sont à chercher de votre côté.

Exercice 1.1 (Estimation de la variance, moyenne inconnue). Soit un échantillon i.i.d. $(X_1, \dots, X_n)$ d'espérance μ et de variance $\sigma^2 > 0$ finie, toutes les deux inconnues. On s'intéresse à l'estimation de σ^2 . On note

$$\bar{X} := \frac{1}{n} \sum_{i=1}^n X_i \quad \text{et} \quad \hat{m}_2 := \frac{1}{n} \sum_{i=1}^n X_i^2.$$

On considère l'estimateur de σ^2 suivant :

$$V_n = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2.$$

1. Montrer que $V_n = \hat{m}_2 - (\bar{X})^2$.
2. V_n est-il un estimateur sans biais de σ^2 ? Sinon, proposer un estimateur sans biais qu'on notera $\hat{\sigma}^2$.

Dans toute la suite de l'exercice, on suppose de plus que les X_i suivent la loi normale $\mathcal{N}(\mu, \sigma^2)$. On admet que cela implique que $K_{n-1} := \sum_{i=1}^n (X_i - \bar{X})^2 / \sigma^2$ suit une loi du khi-deux à $n-1$ degrés de libertés, loi dont la moyenne est $n-1$ et la variance est $2(n-1)$.

3. En déduire l'expression du risque quadratique de V_n .
4. En utilisant la loi forte des grands nombres, montrer que cet estimateur est fortement consistant ?

Exercice 1.2 (Intervalles de confiance dans le modèle uniforme). Supposons que $(X_n)_{n \geq 1}$ sont i.i.d. de loi uniforme sur $[0, \theta]$, avec $\theta > 0$. Pour tout $n \geq 1$, on définit

$$M_n := \max(X_1, \dots, X_n).$$

1. Montrer que $(\frac{M_n}{\theta})^n$ est pivotale, et donner sa loi. On pourra calculer $\mathbb{P}_\theta((M_n/\theta)^n \leq u)$ pour tout $u \geq 0$.
2. Soit $\alpha \in]0, 1]$. En déduire un intervalle de confiance I de probabilité de couverture $1 - \alpha$ pour θ basé sur M_n .
3. Trouver un équivalent du diamètre de I lorsque $n \rightarrow \infty$, pour un α fixé dans $]0, 1]$.

Exercice 1.3 (Test gaussien, variance connue). On admet (ou on rappelle) dans cet exercice une propriété fondamentale des v.a. gaussiennes : si $(Z_j)_{1 \leq j \leq m}$ sont des v.a. réelles indépendantes et de loi respectives $(\mathcal{N}(\mu_j, \sigma_j^2))_{1 \leq j \leq m}$, alors pour tous réels $\alpha_0, \alpha_1, \dots, \alpha_m$, on a

$$\alpha_0 + \alpha_1 Z_1 + \dots + \alpha_m Z_m \sim \mathcal{N} \left(\alpha_0 + \sum_{j=1}^m \alpha_j \mu_j, \sum_{j=1}^m \alpha_j^2 \sigma_j^2 \right).$$

Soit $(X_1, \dots, X_{25})$ un échantillon de loi gaussienne d'espérance μ inconnue et de variance $\sigma^2 = 100$ connue.

On donne quelques quantiles de la loi $\mathcal{N}(0, 1)$:

$$q_{0.975} \sim 1.96, q_{0.95} \sim 1.65, q_{0.9} \sim 1.28, q_{0.8} \sim 0.84,$$

et quelques images de sa fonction de répartition Φ :

$$\Phi(1.21) \sim 0.89, \Phi(0.90) \sim 0.82, \Phi(0.53) \sim 0.70, \Phi(0.09) \sim 0.53.$$

1. Construire un test de niveau $\alpha = 0.10$ pour

$$\mathcal{H}_0 : \mu = 0 \quad \text{contre} \quad \mathcal{H}_1 : \mu = 1.5,$$

fondé sur la moyenne empirique $\bar{X} := \frac{1}{25} \sum_{i=1}^{25} X_i$, estimateur du paramètre μ .

2. On observe $\bar{x} = 1$. Quelle est la décision du test ? L'erreur que l'on fait peut-être ici est-elle de première espèce ? de seconde espèce ? La calculer.
3. Déterminer la taille minimum d'un échantillon dans le même cadre que ci-dessus si l'on souhaite que le test précédent ait des erreurs de première et de seconde espèce toutes deux inférieures à 0.1.

Exercice 1.4 (Des questions d'identifiabilité). On considère un modèle dans lequel l'observation X est une différence $X = Y - Z$ avec Y, Z deux variables gaussiennes indépendantes, de moyennes respectives μ_1, μ_2 et de variances respectives σ_1^2, σ_2^2 , toutes inconnues.

1. Ce modèle est-il identifiable ?
2. Supposons dans cette question que $\mu_2 = 3\mu_1 + 1$ et $\sigma_2 = 2\sigma_1$. Cela rend-il le modèle identifiable ?
3. Qu'en est-il d'un modèle où $X = Y - Z$ avec Y, Z deux variables exponentielles indépendantes à paramètres inconnus ? On pourra chercher une interprétation géométrique aux équations pour l'espérance et la variance de X .
4. Qu'en est-il d'un modèle où $X = \alpha(Y - Z)$ avec Y, Z deux variables exponentielles indépendantes à paramètres inconnus, et $\alpha \in \mathbb{R}$? Et si $\alpha > 0$?

Exercice 1.5 (Estimateur sans biais pour une Bernoulli). On considère un échantillon $(X_1, \dots, X_n)$ i.i.d. de loi de Bernoulli $\mathcal{B}(p)$.

1. Montrer que $\bar{X}$ (la moyenne empirique) est un estimateur sans biais de p .
2. Soit $g(\bar{X})$ un autre estimateur sans biais de p qui est une fonction (mesurable) de $\bar{X}$. Montrer que pour tout $x \in \mathbb{R}^+$:

$$\sum_{k=0}^n \left(g\left(\frac{k}{n}\right) - \frac{k}{n} \right) \binom{n}{k} x^k = 0$$

et en déduire que $\bar{X}$ est le seul estimateur sans biais de p qui est une fonction de $\bar{X}$.

Exercice 1.6. On reprend le cadre de l'exercice 1.3. Désormais on souhaite tester

$$\mathcal{H}_0 : \mu = 2 \quad \text{contre} \quad \mathcal{H}_1 : \mu < 2.$$

Bâtir un test approprié de niveau α donné. Exprimer la puissance de ce test en chaque $\mu < 2$ à l'aide de la fonction Φ , et commenter la dépendance de la puissance en fonction de μ , n et σ .

2. Information de Fisher. Estimation par maximum de vraisemblance

Objectifs : Calculer une information de Fisher, l'interpréter. Savoir donner la borne de Cramer-Rao. Pratiquer l'estimation par maximum de vraisemblance. Les exercices 2.1 et 2.2 sont à faire pendant le TD, les autres sont à chercher de votre côté.

Exercice 2.1 (Information de Fisher dans un modèle à deux paramètres). On considère le modèle statistique où $X_1, \dots, X_n$ sont des variables réelles indépendantes, avec pour tout $1 \leq i \leq n$, $X_i \sim \mathcal{N}(\alpha + \beta t_i, 1)$, où les constantes $(t_i)_{1 \leq i \leq n}$ sont connues et $\alpha, \beta \in \mathbb{R}$ sont des paramètres inconnus. On admettra que ce modèle est régulier.

1. Montrer que la matrice d'information de Fisher $I(\alpha, \beta)$ s'écrit

$$I(\alpha, \beta) = \begin{pmatrix} n & \sum_{i=1}^n t_i \\ \sum_{i=1}^n t_i & \sum_{i=1}^n t_i^2 \end{pmatrix}.$$

2. Déterminer une condition nécessaire et suffisante sur les $(t_i)_{1 \leq i \leq n}$ pour que $I(\alpha, \beta)$ soit inversible. Interpréter cette condition.
3. Sous la condition de la question 2, donner une borne inférieure sur la variance de tout estimateur non biaisé de α . (On prendra donc $h(\alpha, \beta) = \alpha$).
4. Supposons que l'on connaît β . Reprendre les calculs pour trouver $\mathcal{I}(\alpha)$ et donner dans ce cas une borne inférieure sur la variance de tout estimateur non biaisé de α .
5. Comment les bornes inférieures des questions 3 et 4 se comparent-elles ? Interpréter.

Exercice 2.2 (Deux estimateurs dans le modèle uniforme). On considère le modèle dans lequel $X_1, \dots, X_n$ sont des variables i.i.d. suivant une loi uniforme sur $[0, \theta]$, où $\theta > 0$ est un paramètre inconnu.

1. Trouver un estimateur $\hat{\theta}_1$ de θ en utilisant la méthode des moments appliquée au premier moment.
2. Déterminer la vraisemblance dans ce modèle, et montrer qu'il existe un unique estimateur du maximum de vraisemblance $\hat{\theta}_2$. Donner son expression.
3. Calculer le biais, la variance et le risque quadratique de $\hat{\theta}_1$ et de $\hat{\theta}_2$. On calculera la densité de $\hat{\theta}_2$.
4. Quel est le meilleur estimateur ?

Exercice 2.3 (Estimation par maximum de vraisemblance dans le cas régulier). On considère un échantillon $X = (X_1, \dots, X_n)$ i.i.d. de loi η . Pour les modèles suivants, dont on admettra qu'ils sont réguliers, déterminer l'estimateur du maximum de vraisemblance, puis énoncer le théorème central limite associé.

1. η est la loi exponentielle de paramètre¹ $\lambda > 0$, et l'on veut estimer λ .
2. η est la loi de puissance à densité sur $\mathbb{R}$ donnée par $x \mapsto \mathbf{1}_{x \geq 1} \frac{a-1}{x^a}$ avec $a > 1$ paramètre et l'on veut estimer a .

Exercice 2.4 (Information de Fisher via reparamétrisation). Soit $\{\mathbb{P}_\theta\}_{\theta \in \Theta}$ un modèle régulier unidimensionnel. Considérons une fonction différentiable bijective $h : \Lambda \rightarrow \Theta$. La famille $\{\mathbb{P}_\theta\}_{\theta \in \Theta}$ peut ainsi être reparamétrisée par $\{\mathbb{Q}_\lambda\}_{\lambda \in \Lambda}$. On note $\tilde{I}(\lambda)$ l'information de Fisher dans ce modèle reparamétrisé. Pour λ tel que $h(\lambda) = \theta$, donner $\tilde{I}(\lambda)$ en fonction de $I(\theta)$. Donner une interprétation physique de ce résultat.

¹sa densité sur $\mathbb{R}_+$ est donnée par $x \mapsto \lambda e^{-\lambda x}$.

Exercice 2.5 (Un modèle autorégressif). Nous considérons un modèle avec $(X_1, \dots, X_n)$ tels que $X_1 = \theta + Z_1$, et récursivement, $X_{j+1} = \rho(X_j - \theta) + \theta + Z_{j+1}$ avec les Z_j i.i.d. de loi $\mathcal{N}(0, \sigma^2)$. Nous supposons que $\rho \in [0, 1]$ et qu'il est connu. On admettra la régularité de ce modèle.

1. Calculer la matrice d'information de Fisher $I(\theta, \sigma^2)$ en fonction de θ, σ^2, ρ et n .
2. Donner une borne inférieure pour la variance d'un estimateur sans biais de θ . Commenter et interpréter les valeurs de cette borne lorsque $\rho = 0$, et lorsque $\rho = 1$.

3. Théorèmes asymptotiques. Tests, Théorème de Neyman-Pearson

Objectifs : Savoir appliquer la loi des grands nombres, le théorème central limite, la méthode delta et le Lemme de Slutsky. Savoir appliquer le théorème de Neyman-Pearson pour des test d'hypothèses simples. Les exercices 3.1 à 3.2 sont à faire pendant le TD, les suivants sont à chercher de votre côté.

Exercice 3.1 (Théorèmes asymptotiques). Supposons que $X_1, \dots, X_n$ sont i.i.d., centrées, de variance finie $\sigma^2 > 0$ inconnue. On note classiquement $\bar{X}$ la moyenne empirique et $\hat{\sigma}^2$ l'estimateur non biaisé de la variance.

1. Écrire $\hat{\sigma}^2$ en fonction des X_i .
2. Étudier la convergence presque sûre de $\bar{X}$ et de $\hat{\sigma}^2$.
3. Montrer que

$$\frac{\sqrt{n}\bar{X}}{\sqrt{\hat{\sigma}^2}} \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}(0, 1).$$

4. On suppose ici que $\sigma^2 = 1$ et qu'il est connu. Donner la limite en loi de

$$\sqrt{n}(e^{\bar{X}/2} - 1).$$

Exercice 3.2 (Test optimal pour la loi exponentielle). Soient $X_1, \dots, X_n$ des variables aléatoires i.i.d. de loi exponentielle de paramètre² $\lambda > 0$. On souhaite tester les hypothèses simples $\mathcal{H}_0 : \lambda = \lambda_0$ contre $\mathcal{H}_1 : \lambda = \lambda_1$ avec $\lambda_1 > \lambda_0$.

On rappelle la propriété suivante : la somme de n variables aléatoires indépendantes de loi exponentielle de paramètre $\lambda = 1$ suit une loi Gamma $\Gamma(n, 1)$, dont on notera $q_{\beta}^{\Gamma, n}$ le quantile d'ordre β .

1. Montrer que sous $\mathcal{H}_0$, $\lambda_0 X_1 \sim \text{Exp}(1)$.
2. A l'aide du Théorème de Neyman-Pearson et de la question précédente, montrer qu'il existe un test uniformément plus puissant (UPP) de niveau $\alpha \in]0, 1[$ pour ce problème, et qu'il peut s'écrire

$$\psi(X) = \mathbf{1}_{\{\sum_{i=1}^n X_i < (\lambda_0)^{-1} c(\alpha, n)\}},$$

où $c(\alpha, n)$ est une constante dépendant de α et de n est à déterminer.

3. On notera F_n la fonction de répartition de la loi $\Gamma(n, 1)$. Exprimer la puissance $\pi(\lambda_1)$ du test précédent pour toute valeur de $\lambda_1 > \lambda_0$, en fonction de F_n . Commenter sa monotonie en λ_1 .
4. Montrer par ailleurs que le test précédent est UPP(α) pour tester $\mathcal{H}_0 : \lambda = \lambda_0$ contre $\mathcal{H}'_1 : \lambda > \lambda_0$.

Exercice 3.3 (Asymptotique du maximum de vraisemblance). Reprenons le modèle où les $X_1, \dots, X_n$ sont i.i.d. et dont la loi est à densité sur $\mathbb{R}$ donnée par

$$x \mapsto \mathbf{1}_{x \geq 1} \frac{a-1}{x^a}$$

avec $a > 1$ un paramètre. On a établi que l'estimateur du maximum de vraisemblance de a est donné par

$$\hat{a}_{MV} = 1 + \left(\frac{1}{n} \sum_{i=1}^n \log X_i \right)^{-1}$$

²on rappelle que c'est loi à densité $x \mapsto \mathbf{1}_{x \geq 0} \lambda e^{-\lambda x}$.

dans l'exercice 2.3.

1. Etudier la distribution de $\log X_1$ en calculant sa fonction de répartition, et en déduire que $\hat{a}_{MV}$ est fortement consistant.
2. Calculer l'information de Fisher $I_1(a)$ de X_1 .
3. A l'aide de la méthode Delta, montrer que

$$\sqrt{n}(\hat{a}_{MV} - a) \xrightarrow[n \rightarrow \infty]{(d)} \mathcal{N}(0, (a-1)^2).$$

4. Interpréter les résultats des questions 2 et 3 à la lumière de la notion d'efficacité vue en cours. *On pourra admettre des résultats.*

Exercice 3.4 (Test de variance symétrique à deux échantillons). *On admettra dans cet exercice que si Z_1, Z_2 sont deux variables i.i.d. gaussiennes standard, alors Z_1/Z_2 suit une loi de Cauchy de densité sur $\mathbb{R}$ donnée par $x \mapsto \frac{1}{\pi(1+x^2)}$.*

On travaille avec deux échantillons indépendants, X_1 et X_2 . On suppose que $0 < \sigma_1^2 < \sigma_2^2$, et on souhaite tester

$$\mathcal{H}_0 : X_1 \sim \mathcal{N}(0, \sigma_1^2), X_2 \sim \mathcal{N}(0, \sigma_2^2)$$

contre

$$\mathcal{H}_1 : X_1 \sim \mathcal{N}(0, \sigma_2^2), X_2 \sim \mathcal{N}(0, \sigma_1^2).$$

On souhaite travailler avec des tests symétriques ψ , tels que $\psi(x_1, x_2) = 1 - \psi(x_2, x_1)$ presque partout.

1. Montrer que si ψ est un test symétrique de taille α pour ce problème, alors sa puissance est $\pi(\psi) = 1 - \alpha$.
2. Écrire le rapport de vraisemblance L , et construire un estimateur symétrique ψ_0 basé sur L pour ce problème.
3. Montrer que la taille α et la puissance de ψ_0 vérifient

$$\alpha = 1 - \pi(\psi_0) = \frac{2}{\pi} \tan^{-1} \left(\frac{\sigma_1}{\sigma_2} \right).$$

Exercice 3.5 (Plus vite que $\sqrt{n}$?). Reprenons l'exercice 3.1, en supposant que $\sigma^2 = 1$ et qu'il est connu.

1. Quelle est la limite en loi de $\sqrt{n}(\cos(\bar{X}) - 1)$?
2. Trouver une suite $(a_n)_{n \geq 1}$ telle que $a_n(\cos(\bar{X}) - 1)$ converge en loi vers une v.a. Z qui n'est pas presque sûrement constante, dont on précisera la loi.

Exercice 3.6. Reprenons l'exercice 3.2.

1. Sous quelle condition sur a ces variables sont-elles d'espérance finie ?
2. Sous la condition de la question 1, donner un estimateur $\hat{a}_{MM}$ de a via la méthode des moments. Prouver qu'il est consistant.
3. On suppose dans toute la suite que la condition de la question 1 est vérifiée. A l'aide du TCL et de la méthode Delta, montrer que $\hat{a}_{MM}$ est asymptotiquement normal et donner ses caractéristiques.
4. Pour n grand, quel est l'estimateur à privilégier selon vous, entre $\hat{a}_{MM}$ et $\hat{a}_{MV}$?

4. Voyage en pays gaussien

Objectifs : Se familiariser avec les vecteurs gaussiens, leurs propriétés et les calculs associés. Les exercices 4.1 à 4.3 sont à faire pendant le TD, le 4.4 est à chercher de votre côté.

Exercice 4.1 (Pour commencer). Soit $Z = (X, Y)$ un vecteur aléatoire sur $\mathbb{R}^2$ ayant une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^2$ donnée par

$$f(x, y) = C \exp(-x^2 + xy - y^2/2),$$

où C est une constante de normalisation que l'on ne demande pas de calculer.

1. Montrer que Z est un vecteur gaussien dont on précisera l'espérance et la matrice de covariance. *On pourra commencer par chercher $\mu \in \mathbb{R}^2$ et $M \in \mathbb{R}^{2 \times 2}$ symétrique telle que $-x^2 + xy - y^2/2 = \frac{1}{2}(z - \mu)^T M (z - \mu)$ pour $z = (x, y)$.*
2. Quelle est la loi marginale de X ? de Y ? de $2X - Y + 5$?
3. Montrer que X et $Y - X$ sont indépendants.
4. Soit $B = \frac{\sqrt{5}}{5} \begin{pmatrix} 3 & -1 \\ -1 & 2 \end{pmatrix}$. Vérifier que $B^2 = \Sigma^{-1}$. Quelle est la loi du vecteur BZ ?

Exercice 4.2 (Un contre-exemple). Soit X variable réelle gaussienne $\mathcal{N}(0, 1)$, et S une variable aléatoire de Rademacher, i.e. telle que $\mathbb{P}(S = 1) = \mathbb{P}(S = -1) = 1/2$. On suppose S et X indépendantes. On définit le vecteur aléatoire $Z = (X, SX)$.

1. Montrer que les coordonnées de Z sont des gaussiennes $\mathcal{N}(0, 1)$. *On pourra montrer que la fonction de répartition de SX est encore celle d'une gaussienne standard.*
2. Montrer que pourtant, Z n'est pas un vecteur gaussien.

Exercice 4.3 (Indépendance entre moyenne et variance empiriques dans le modèle gaussien). On considère un échantillon $X = (X_1, \dots, X_n)$ i.i.d. de loi $\mathcal{N}(\mu, \sigma^2)$ avec μ et σ^2 inconnus.

1. On s'intéresse à la variance empirique de l'échantillon définie par $S_n := \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$. En utilisant le théorème de Cochran, déterminer la loi de S_n et montrer que S_n et $\bar{X}$ sont indépendants. On pourra introduire le vecteur e de $\mathbb{R}^n$ dont toutes les coordonnées valent 1.
2. On s'intéresse à la variance empirique débiaisée, définie par $S'_n := \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$. Quelle est la loi de $\sqrt{n} \frac{\bar{X} - \mu}{\sqrt{S'_n}}$?

Exercice 4.4 (Test des variances dans le modèle gaussien). Nous considérons deux échantillons gaussiens indépendants. Le premier, $X^{(1)}$, est de taille n_1 et i.i.d. de loi $\mathcal{N}(\mu_1, \sigma_1^2)$. Le second, $X^{(2)}$, est de taille n_2 et i.i.d. de loi $\mathcal{N}(\mu_2, \sigma_2^2)$. Tous les paramètres sont supposés inconnus. On note S'_1 (resp. S'_2) la variance empirique débiaisée pour l'échantillon $X^{(1)}$ (resp. $X^{(2)}$). On veut tester

$$\mathcal{H}_0 : \sigma_1^2 = \sigma_2^2 \quad \text{contre} \quad \mathcal{H}_1 : \sigma_1^2 \neq \sigma_2^2.$$

1. Sous l'hypothèse nulle, quelle est la loi de $\frac{S'_1}{S'_2}$?
2. En déduire un test de niveau $\alpha \in]0, 1[$ répondant au problème.
3. Exprimer sa puissance en fonction de σ_1^2, σ_2^2 et une certaine fonction de répartition.

5. Modèles linéaire et linéaire gaussien : définitions, estimateur des moindres carrés, test de Student

Objectifs : Savoir faire mener les calculs de la régression linéaire, interpréter les résultats, donner des intervalles de confiance sur les coefficients, tester la nullité d'un coefficient. L'exercice 5.1 est à faire pendant le TD, les autres sont à chercher de votre côté.

Exercice 5.1 (Les eucalyptus). On souhaite expliquer la hauteur y (en mètres) d'un eucalyptus en fonction de sa circonférence x (en centimètres) à 1 mètre 30 du sol, et de la racine carrée de celle-ci. On a relevé $n = 1429$ mesures de couples (x_i, y_i) , le nuage de points étant représenté sur la figure 1 ci-contre.

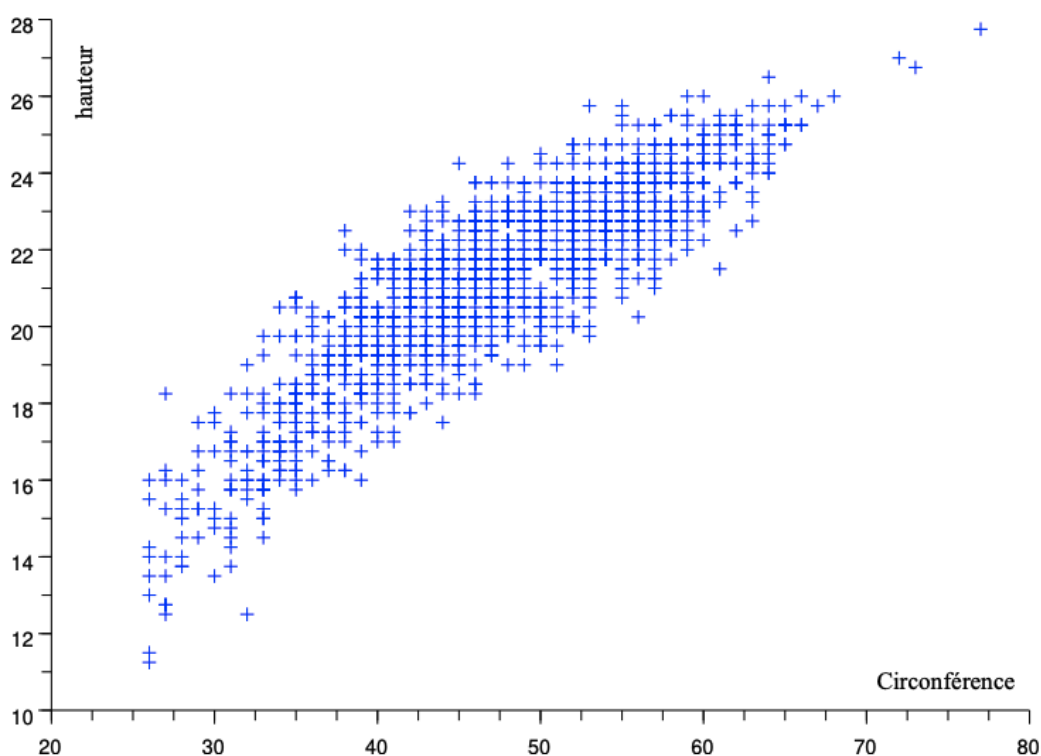


Figure 1 – Données de hauteurs d'eucalyptus (m) en fonction de leur circonférence (cm).

On propose donc le modèle linéaire suivant : pour tout $1 \leq i \leq n$, $Y_i = \beta_1 + \beta_2 X_i + \beta_3 \sqrt{X_i} + \varepsilon_i$, où les ε_i sont gaussiennes i.i.d. $\mathcal{N}(0, \sigma^2)$. On pose

$$X = \begin{bmatrix} 1 & x_1 & \sqrt{x_1} \\ \vdots & \vdots & \vdots \\ 1 & x_n & \sqrt{x_n} \end{bmatrix} \quad \text{et} \quad Y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

Nous avons observé

$$X^T X = \begin{bmatrix} ? & ? & 9792 \\ ? & 3306000 & ? \\ ? & 471200 & 67660 \end{bmatrix}, \quad X^T Y = \begin{bmatrix} 30310 \\ 1462000 \\ 209700 \end{bmatrix} \quad \text{et} \quad Y^T Y = 651900.$$

1. Compléter la matrice $X^T X$.
2. Que vaut la circonférence moyenne empirique $\bar{x}$?

3. Les calculs donnent en arrondissant :

$$(X^T X)^{-1} = \begin{bmatrix} 4.646 & 0.101 & -1.379 \\ 0.101 & 0.002 & -0.030 \\ -1.379 & -0.030 & 0.411 \end{bmatrix} \quad \text{et} \quad (X^T X)^{-1} X^T Y = \begin{bmatrix} -16.8 \\ -0.30 \\ 7.62 \end{bmatrix}.$$

Que vaut ici l'estimateur des moindres carrés $\hat{\beta}$? Représenter la courbe de régression obtenue sur la Figure 1 via le calcul de quelques points.

4. Vérifier que pour tout modèle linéaire identifiable, $Y^T X \hat{\beta} = \|X \hat{\beta}\|^2$.
5. En déduire la valeur de l'estimateur de σ^2 débiaisé et en donner une intervalle de confiance de niveau 95%. On utilisera l'approximation suivante : quand m est grand, une variable $\chi^2(m)$ est proche d'une variable $\mathcal{N}(m, 2m)$. On donne le quantile gaussien standard à 97.5% : $q = 1.96$.
6. Donner un intervalle de confiance pour β_3 de probabilité de couverture 95%. On approchera la loi $\mathcal{T}(m)$ par la loi $\mathcal{N}(0, 1)$, quand m est grand.
7. Tester l'hypothèse $\beta_2 = 0$ au niveau de risque 10%. On fera les mêmes approximations que précédemment. On donne le quantile gaussien standard à 95% : $q = 1.645$. Interpréter.

Exercice 5.2 (Exemple de régression linéaire à la main). On considère le modèle linéaire qui s'écrit matriciellement $Y = \theta_0 e + \theta_1 Z + \varepsilon$, avec $Y \in \mathbb{R}^n$, e le vecteur de $\mathbb{R}^n$ dont les coordonnées valent toutes 1, $Z = (z_1, z_2, \dots, z_n)^T \in \mathbb{R}^n$, et ε un bruit centré.

1. Donner une condition nécessaire et suffisante explicite sur le vecteur Z pour que le modèle soit identifiable. Interpréter.

On se place sous la condition d'identifiabilité de la question 1. On introduit la covariance empirique entre Y et Z définie par $C(Y, Z) := \frac{1}{n} \sum_{i=1}^n z_i Y_i - \bar{Z} \times \bar{Y}$, ainsi que la variance empirique de Z définie par $V(Z) := \frac{1}{n} \sum_{i=1}^n z_i^2 - (\frac{1}{n} \sum_{i=1}^n z_i)^2$.

2. Calculer à la main l'estimateur des moindres carrés $\hat{\theta}_{MC} = (\hat{\theta}_0, \hat{\theta}_1)^T$. On écrira $\hat{\theta}_1$ en fonction de C et V , puis $\hat{\theta}_0$ en fonction de θ_0 .
3. Montrer que le point moyen, de coordonnées $(\bar{Z}, \bar{Y})$, appartient à la droite de régression obtenue.

Exercice 5.3 (Théorème de Gauss-Markov). On note $\preceq$ la relation d'ordre dans $S_p(\mathbb{R})$ définie par :

$$A \preceq B \iff B - A \in S_p^+(\mathbb{R}).$$

Pour $\hat{\theta}_1$ et $\hat{\theta}_2$ deux estimateurs sans biais de $\theta \in \mathbb{R}^p$, on dira que $\hat{\theta}_1$ est meilleur que $\hat{\theta}_2$ si $\text{Var}(\hat{\theta}_1) \preceq \text{Var}(\hat{\theta}_2)$.

Le but de cet exercice est de démontrer le *théorème de Gauss-Markov*, dont l'énoncé est le suivant : *dans un modèle linéaire identifiable, parmi tous les estimateurs de θ linéaires en Y et sans biais, l'estimateur des moindres carrés est le meilleur (au sens de l'ordre $\preceq$).*

On considère donc $\tilde{\theta}$ un autre estimateur de θ , linéaire en Y et sans biais. On l'écrit $\tilde{\theta} = CY$ avec $C = (X^T X)^{-1} X^T + D$ avec $D \in \mathbb{R}^{p \times n}$.

1. Montrer que $DX = 0$.
2. Montrer que $\text{Var}(\tilde{\theta}) = \text{Var}(\hat{\theta}_{MC}) + \sigma^2 DD^T$. Conclure.

6. Modèles linéaire et linéaire gaussien : intervalles de confiance, modèles emboîtés, lecture des résultats

Objectifs : Savoir lire les résultats d'une régression linéaire à la main, tester des modèles emboîtés, traiter le cas d'une variable quantitative. Travailler autour de l'erreur de prédiction. L'exercice 6.1 est à faire pendant le TD, le reste est à chercher de votre côté.

Exercice 6.1 (Durée de vie d'une pièce industrielle). Une entreprise de fabrication automobile cherche à expliquer et prédire la durée de vie d'une pièce de sa fabrication, en fonction de variables d'intérêt. Cette pièce, en alliage d'aluminium, est située dans le moteur des véhicules produits par l'entreprise. C'est un parallélépipède rectangle dont la base est de surface constante, mais d'épaisseur (ou hauteur) variable. Elle est de densité variable selon l'alliage d'aluminium utilisé. On dispose d'un jeu de données de $n = 71$ observations dont les variables sont les suivantes :

- **duree_vie** : durée de vie de la pièce, définie comme le nombre de km parcourus au compteur du véhicule avant de devoir changer la pièce ;
- **epaisseur** : épaisseur de la pièce, en cm ;
- **poids** : poids de la pièce, en hg ($1 \text{ hg} = 10^2 \text{ g}$) ;
- **finition** : en fin de fabrication de la pièce en aluminium, une finition lui est appliquée. On distingue trois types de finitions, notés A, B et C.

Dans cet exercice, tous les modèles linéaires rencontrés sont supposés gaussiens. On donne ci-dessous un aperçu du début du jeu de données ainsi qu'un résumé de celui-ci.

duree_vie <dbl>	epaisseur <dbl>	poids <dbl>	finition <chr>
120913.10	6.265404	7.116102	C
139490.76	4.198606	3.115564	B
108487.94	5.286604	2.315019	A
98329.29	2.989635	1.562986	A
115342.99	5.635362	3.820599	C
97965.86	2.804801	1.294772	C

Figure 2 – Aperçu du jeu de données

duree_vie	epaisseur	poids	finition
Min. : 76750	Min. : 1.889	Min. : 0.3094	Length: 71
1st Qu.: 102958	1st Qu.: 3.183	1st Qu.: 0.4542	Class : character
Median : 118248	Median : 3.997	Median : 0.7420	Mode : character
Mean : 118682	Mean : 3.935	Mean : 0.9292	
3rd Qu.: 131886	3rd Qu.: 4.726	3rd Qu.: 1.1978	
Max. : 161501	Max. : 6.444	Max. : 3.1897	

Figure 3 – Résumé du jeu de données

1. On commence par une régression linéaire classique dont le résultat est présenté sur la figure 4. Attention, certains éléments sont effacés.

Au vu de ces résultats, quelles sont les éléments expliquant significativement la durée de vie de la pièce ? Dans quel sens sont-ils corrélés à la variable réponse ? On appuiera la réponse sur des arguments quantitatifs.

2. Soucieux d'étudier les données plus en détail, on mène une analyse bivariée des variables quantitatives et on obtient les corrélations empiriques présentées sur la figure 5(a).

2.(a). Que dire quant à la redondance des variables explicatives ? Donner une explication de cette éventuelle redondance. On souhaite remplacer la variable **poids** par

```

Call:
lm(formula = duree_vie ~ ., data = data)

Residuals:
    Min       1Q   Median       3Q      Max
-26797.8  -6220.6  -128.4   5958.2  28778.1

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 100976.0     5344.9   18.892 < 2e-16
epaisseur    162.6      1299.7    0.125  0.901
poids       3266.2      522.2    6.254 3.35e-08
finitionB    20831.8     3812.8    5.464 7.64e-07
finitionC   -3387.5     3494.3   -0.969  0.336
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10980 on  degrees of freedom
Multiple R-squared:  0.7004,    Adjusted R-squared:  0.6822
F-statistic: 38.57 on  and  DF,  p-value: < 2.2e-16

```

Figure 4 – Résultat de la première régression

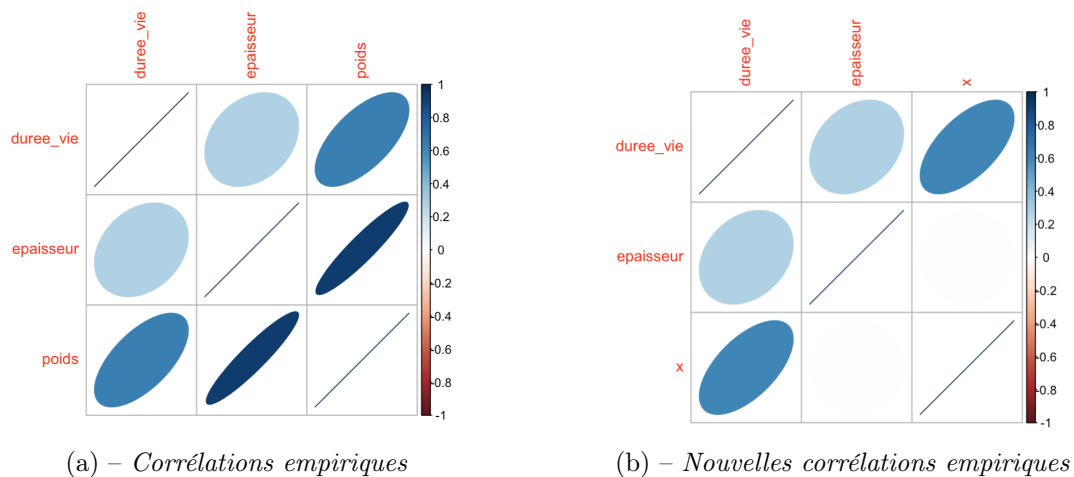


Figure 5 – Corrélations entre les variables

une variable notée x . Proposer une expression pertinente de la variable x pour supprimer la redondance dans les données.

2.(b). Nous appliquons à notre jeu de données la modification proposée en question 2.(a). : la nouvelle variable x remplace désormais la variable **poids**. Commenter la figure 5(b).

- Sur la figure 6, on présente le résultat de la régression linéaire pour ce jeu de données modifié. Compléter les trois dernières lignes du résultat affiché dans la figure 6 en expliquant votre raisonnement.
- Ce deuxième modèle (avec x) est-il plus pertinent que le premier (avec **poids**) pour expliquer les données ? Donner deux indicateurs quantitatifs pour justifier votre réponse.
- La figure 6 donne des extraits de code. Retrouver les éléments cachés à la ligne **epaisseur** de la figure 5. On ne demande pas de retrouver les étoiles. On pourra utiliser les résultats de la figure 7.
- On s'intéresse à la variable **finition**. Pour $f \in \{A, B, C\}$, on note (f) le groupe des pièces telles que **finition** = f .
 - Les groupes (A) et (B) sont-ils statistiquement équivalents ?
 - Les groupes (A) et (C) sont-ils statistiquement équivalents ?

```

Residuals:
    Min       1Q   Median       3Q      Max
-25561.9  -5573.7   -713.6   4851.0  28125.7

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    86464      5572  15.517  < 2e-16
epaisseur       3508          3704   0.947  0.00387
finitionB      20366          3392   6.004  6.69e-07
finitionC     -2867          2098  -1.367  0.40114
x              14192          6765   2.099  4.23e-09
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10640 on 10 degrees of freedom
Multiple R-squared:  0.7182,    Adjusted R-squared:  0.6781
F-statistic: 42.06 on 4 and 10 DF,  p-value: < 2.2e-16

```

Figure 6 – Résultats de la régression pour le jeu de données modifié

```

```{r}
X = as.matrix(cbind(rep(1,n),epaisseur,finition=='B',finition=='C',x))
colnames(X) <- NULL # on enlève le nom des colonnes
t(X)%*%X
```

      [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 71.00000 279.38630 23.00000 34.00000 65.97137
[2,] 279.38630 1184.43593 96.23539 131.34674 259.54139
[3,] 23.00000 96.23539 23.00000 0.00000 26.96826
[4,] 34.00000 131.34674 0.00000 34.00000 26.55636
[5,] 65.97137 259.54139 26.96826 26.55636 89.22035

```{r}
solve(t(X)%*%X)
```

      [,1]      [,2]      [,3]      [,4]      [,5]
[1,] 0.27406085 -0.045651410 -0.03859205 -0.068088863 -0.037914974
[2,] -0.04565141 0.012117616 -0.00612221 -0.001874548 0.000914027
[3,] -0.03859205 -0.006122210 0.12111580 0.071188752 -0.011453203
[4,] -0.06808886 -0.001874548 0.07118875 0.101582443 0.004045512
[5,] -0.03791497 0.000914027 -0.01145320 0.004045512 0.038842193

```

Figure 7 – Extraits de code : on a demandé $X^T X$ dans la premi partie du code, puis $(X^T X)^{-1}$ dans la seconde

- 6.(c). Les groupes (B) et (C) sont-ils statistiquement équivalents ? Justifier soigneusement votre réponse.
7. Combien y a-t-il de pièces de finition A dans l'échantillon ?
8. L'ingénieur auquel vous présentez les résultats vous déclare : "au vu de cette étude statistique, le fait d'appliquer la finition B augmente la durée de vie de la pièce. Suggérez-vous de n'utiliser plus que la finition B dans notre processus de fabrication ? "
- Que répondre à cette question au vu de l'étude réalisée ? Argumentez. Quelle nouvelle étude recommanderiez-vous éventuellement afin de vous donner plus d'éléments pour répondre à cette question ?

Exercice 6.2 (Minimisation de l'erreur de prédiction). Dans cet exercice, on s'intéresse à l'erreur de prédiction dans le modèle linéaire. On cherche à apprendre le modèle suivant pour prédire le réel Y_i en fonction d'une variable explicative réelle z_i :

$$\forall 1 \leq i \leq n, \quad Y_i = \beta_0 + \beta_1 z_i + \varepsilon_i.$$

On notera $\bar{y}$ et $\bar{z}$ les moyennes empiriques de $z = (z_1, \dots, z_n)^T$ et de $Y = (Y_1, \dots, Y_n)^T$ (attention, elles dépendent de n). On suppose que le modèle est identifiable et on note $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ l'estimateur des moindres carrés de β . On note σ^2 la variance (constante) du bruit.

Sous le même modèle, on observe une nouvelle valeur z_{n+1} de la variable explicative et on cherche à prédire la variable réponse Y_{n+1} avec l'estimateur

$$\hat{y}_{n+1} := \hat{\beta}_0 + \hat{\beta}_1 z_{n+1} = x_{n+1}^T \hat{\beta}, \quad \text{où } x_{n+1} := \begin{bmatrix} 1 \\ z_{n+1} \end{bmatrix}.$$

L'erreur de prédiction est définie par :

$$\text{err}(z_{n+1}) := \mathbb{E} [(Y_{n+1} - \hat{y}_{n+1})^2].$$

Notons que dans cette espérance, l'aléa vient du bruit ε_{n+1} dans Y_{n+1} ainsi que de l'aléa dans le $\hat{\beta}$.

1. Montrer que si S est un vecteur aléatoire de $\mathbb{R}^m$ de matrice de covariance (finie) C , alors pour tout vecteur $u \in \mathbb{R}^m$, $\text{Var}(u^T S) = u^T C u$.
2. Que vaut $\mathbb{E}[Y_{n+1} - \hat{y}_{n+1}]$?
3. Ecrire la matrice X du plan d'expérience dans ce modèle (on mettra l'intercept dans la première colonne), et montrer que

$$(X^T X)^{-1} = \frac{1}{nv(z)} \begin{bmatrix} \frac{1}{n} \sum_{i=1}^n z_i^2 & -\bar{z} \\ -\bar{z} & 1 \end{bmatrix},$$

avec $v(z) := \frac{1}{n} \sum_{i=1}^n z_i^2 - (\bar{z})^2$ la variance empirique de z .

4. En utilisant les résultats des questions 1, 2 et 3, montrer que

$$\text{err}(z_{n+1}) = \sigma^2 \left(1 + \frac{1}{n} + \frac{(z_{n+1} - \bar{z})^2}{\sum_{i=1}^n (z_i - \bar{z})^2} \right).$$

5. Pour quelle(s) valeur(s) de z_{n+1} l'erreur de prédiction est-elle minimale ? Interpréter ce résultat.
6. Quelle est la limite de l'erreur de prédiction minimale lorsque $n \rightarrow \infty$? Interpréter cette valeur limite.

Exercice 6.3 (Equivalence Student/Fisher pour la nullité d'un coefficient). Dans cet exercice souhaite montrer l'équivalence entre les tests de Student et de Fisher pour la nullité d'un paramètre. On considère donc le modèle linéaire Gaussien identifiable $Y = X\theta + \varepsilon$ pour lequel on veut tester la nullité du dernier coefficient θ_p .

1. Donner la statistique T du test de Student pour le test de nullité du dernier coefficient θ_p .
2. Donner la statistique F du test de Fisher pour les modèles emboîtés correspondants. On fera apparaître $\hat{\sigma}^2$ au dénominateur.
3. Soit T_d une variable suivant une loi de Student à d degrés de liberté. Rappeler sa définition et en déduire la loi suivie par la variable $F_d = T_d^2$.
4. On note la matrice du plan d'expérience sous forme de bloc $X = [X_0 | X_p]$, où $X_0 = [X_1 | \dots | X_{p-1}]$ est la matrice de taille $n \times (p-1)$ des $(p-1)$ premières colonnes de X , et X_p est sa dernière colonne. Ecrire la matrice $X^T X$ sous forme de 4 blocs.
5. On donne (ou on rappelle) le lemme suivant :

Lemme (Inversion par blocs). Soit A une matrice inversible s'écrivant par blocs $A = \begin{bmatrix} T & U \\ V & W \end{bmatrix}$ avec T inversible. Alors $Q = W - VT^{-1}U$ est inversible et l'inverse de A s'écrit :

$$A^{-1} = \begin{bmatrix} T^{-1} + T^{-1}UQ^{-1}VT^{-1} & -T^{-1}UQ^{-1} \\ -Q^{-1}VT^{-1} & Q^{-1} \end{bmatrix}.$$

Grâce à la formule d'inversion matricielle par blocs, en déduire que

$$[(X^T X)^{-1}]_{i,i} = (X_p^T (I_n - \Pi_0) X_p)^{-1},$$

où Π_0 est la matrice $n \times n$ de projection orthogonale sur l'espace engendré par les colonnes de X_0 .

6. En reprenant les notations du cours $\hat{Y}$ et $\hat{Y}_0$, vérifier que $\hat{Y}_0 = \Pi_0 \hat{Y}$ puis montrer que

$$\hat{Y} - \hat{Y}_0 = (\hat{\theta}_{MC})_p (I_n - \Pi_0) X_p.$$

7. En déduire que dans ce cas $F = T^2$. Conclure.