STA201- M1 Mathématiques Appliquées - Paris Saclay / ENSTA

université PARIS-SACLAY

FACULTÉ DES SCIENCES D'ORSAY

ENSTA IP PARIS

STA201

Modélisation statistique

Christine Keribin
Université Paris-Saclay

5 septembre 2021

Avant Propos

Le cours STA201 fait suite au cours MA101 dans lequel les principes de base de la statistique inférentielle ont été étudiés : estimateur, tests, intervalle de confiance dans les cas classiques d'échantillons iid avec un paramètre d'intérêt (en général univarié). Il apporte des compléments en théorie de l'estimation (estimation par maximum de vraisemblance, efficacité, tests de Wald et du rapport de vraisemblance, cadre multivarié...), illustre le cadre de la modélisation statistique et détaille en particulier le modèle linéaire.

L'enseignement comporte une part de travaux dirigés sur ordinateurs qui permettent de transformer le savoir théorique en une pratique de la modélisation de données réelles et de l'estimation de modèles avec un logiciel (logiciel R).

Objectifs Être capable, en utilisant les bases théoriques de la modélisation statistique, de :
- définir une modélisation adaptée à un jeu de données réelles
- construire l'estimateur du maximum de vraisemblance, étudier ses propriétés. étudier le risque et l'efficacité d'un estimateur
- construire les tests de Wald et du rapport de vraisemblance
- estimer un modèle statistique (linéaire) avec un logiciel (R) et interpréter les résultats obtenus ;
- prendre en compte le risque de toute décision statistique.

Prérequis Le cours de STA201 suppose connues les bases de la statistique inférencelle, notions de statistique d'un niveau L3 ou première année d'école d'ingénieur :
- biais, variance, risque, consistance d'un estimateur (loi des grands nombres)
- loi limite d'un estimateur renormalisé (théorème central limite) et approximation asymptotique
- méthodes de construction d'un estimateur
- Cadre de Neyman-Pearson pour la construction d'un test statistique. Cas du test d'un paramètre univarié. Interprétation
- Construction d'intervalles de confiance dans les cas classiques de paramètres univariés estimés dans un échantillon iid.

Structure Ce polycopié est structuré de la façon suivante :
- L'introduction rappelle la problématique de la démarche statistique, et définit en particulier le cadre du modèle dominé et la vraisemblance.
- Les propriétés des estimateurs sont étudiées au chapitre 2 (nouveautés : exhaustivité, optimalité, asymptotique de la delta-méthode),
- le chapitre 3 est dédié à l'estimateur du maximum de vraisemblance et ses propriétés. On y introduira la statistique de Wald.
- Le chapitre 4 étend le cadre de Neyman-Pearson au test dans les familles à vraisemblance monotone. Il introduit le test de Wald et celui du rapport de vraisemblances (maximales).
- Le chapitre 5 étend la notion d'intervalle de confiance à un paramètre multivarié et présente plusieurs constructions de régions de confiance.
- Enfin, le chapitre 6 présente le modèle linéaire dans lequel une variable réponse est expliquée ou prédite en fonction d'une combinaison linéaire de variables explicatives : régression linéaire multiple quand les variables explicatives sont quantitatives, Anova quand elles sont qualitatives, Ancova lorsque l'une est qualitative et l'autre quantitative.

Une annexe propose quelques rappels sur les vecteurs gaussiens, bien utiles pour l'étude du modèle linéaire, quelques rappels de convergence et des tables de lois usuelles.

Bibliographie Une bibiographie est disponible en fin de polycopié. En particulier :

- Statistique inférentielle : Lejeune (2010), Delmas (2010), Pagès (2005), Saporta (2006), Rivoirard et Stoltz (2009), Bickel et Doksum (2015), Keribin (2018) (cours MA101 pour les prérequis)
- Modèle linéaire : Cornillon et Matzner-Løber (2007), Azais et Bardet (2005), Pagès (2005), Rivoirard et Stoltz (2009), Saporta (2006)
- Logiciel R : Cornillon et autres (2008),

Chapitre 1

Introduction

Nous récoltons et utilisons des données de plus en plus nombreuses et de nature de plus en plus variée : après les données de type traditionnel (qualitatives, quantitatives, séries temporelles), c'est au tour de données moins structurées (images, vidéos, enregistrements sonores, tweets, réseaux) d'être utilisées pour comprendre des phénomènes ou améliorer des processus de décision.

Les phénomènes générant les données sont en général complexes, partiellement connus ou observés. Ils peuvent faire preuve de variabilité : sous les mêmes conditions d'expérience, le résultat n'est pas forcément le même, d'où l'incertitude générée. Dans un milieu partiellement connu, il est crucial de pouvoir calibrer le risque associé à la décision prise. Et quelles en sont les limites ? : que peut-on faire dire (ou ne pas dire) à des jeux de données ?

La statistique propose un cadre mathématique pour expliquer, prédire et décider en environnement incertain ou mal posé à partir de données observées :

— Ce médicament est-il efficace ?
— Quels sont les gènes qui entrent en compte dans le développement d'une maladie ?
— Quel est le risque de survenue d'une crue ?
— Quelle sera l'affluence aux urgences dans deux jours ?
— Reconnaître une espèce de plante à partir de son image
— Ce crédit doit-il être accordé ?
— ...

La statistique permet de définir un langage commun pour décrire les résultats d'expériences, et une méthodologie pour les traiter. C'est une branche des mathématiques appliquées qui a pris naissance au XIXème siècle, et qui se développe exponentiellement en ce moment grâce à l'essor de la puissance des ordinateurs et celui des moyens de stockage de données. Elle développe des aspects *théoriques* (définitions, propriétés, théorèmes) qui fonde la démarche, des aspects *méthodologiques* pour mettre en œuvre les procédures et une part *appliquée* pour le traitement concret de jeux de donnés. Elle s'appuie en particulier sur des outils mathématiques probabilistes et d'optimisation qu'elle enrichit en retour.

Afin de combler le fait que le phénomène n'est que partiellement observé, la démarche statistique nécessite de faire des hypothèses sur ce phénomène, que l'on résume en une seule : le jeu de données est le résultat du tirage d'une "*vraie*" loi de probabilité faisant partie d'une famille de lois $\mathcal{P}$ plus ou moins grande. Il s'agira de choisir dans cette famille celle la loi la plus adaptée. Le choix de la famille (modèle) est donc crucial, puisqu'il va déterminer le contour de l'espace dans lequel on travaille, et donc la complexité du problème (famille plus ou moins vaste) et la qualité des conclusions (peut-on être serein avec une décision prise dans un environnement complètement inadapté ?).

Ce cours développe la statistique *paramétrique* (la famille de lois est caractérisée par un paramètre de dimension finie) *fréquentiste* (le paramètre est déterministe). D'autres directions existent comme la statistique non paramétrique (par opposition à paramétrique) et la statistique bayésienne (le paramètre est lui même modélisé par une variable aléatoire).

1.1 Probabilité et statistique

L'objet mathématique de base pour décrire un phénomène aléatoire est la variable aléatoire.

Définition 1. *Une* ***variable aléatoire*** *est une application mesurable d'un espace probabilisé* $(\Omega, \mathcal{A}, \mathbb{P})$ *dans un espace mesurable* $(E, \mathcal{E})$.

La mesure de probabilité $\mathbb{P}_X$ *sur* $(E, \mathcal{E})$ *est la* ***loi*** *de la variable aléatoire* X*, image de* $\mathbb{P}$ *par* X.

$$\forall B \in \mathcal{E}, \mathbb{P}_X(B) = \mathbb{P}(\omega | X(\omega) \in B) = \mathbb{P}(X^{-1}(B))$$

Le fait d'utiliser la loi image permet de s'affranchir de l'espace Ω. On utilisera dans le cours les variables aléatoires réelles où $E = \mathbb{R}^k$ muni de sa tribu borélienne $\mathcal{B}$. Une variable aléatoire est caractérisée par sa fonction de répartition

$$\forall x \in \mathbb{R}^k, \;\; F_X(x) = \mathbb{P}_X(X_1 \leq x_1 \cap \ldots \cap X_k \leq x_k).$$

Étudier la loi d'une variable aléatoire réelle revient à étudier une probabilité sur $\mathbb{R}^k$.

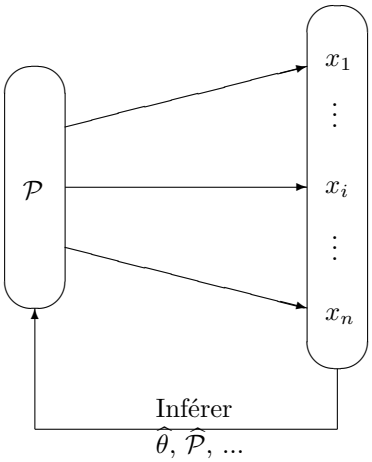
Un exemple classique, le lancer d'une pièce Le lancer d'une pièce engendre un résultat Pile ou Face. Comme il est conditionné par des changements de conditions initiales (position initiale de la main, vitesse d'impulsion, temps d'agitation, ...) impossibles à modéliser de façon déterministe, un tirage peut être représenté par une variable aléatoire qui résume les chances (probabilité) d'avoir un résultat donné : la probabilité de tirer Face (ou $x = 1$) vaut θ, celle de de tirer Pile (ou $x = 0$) vaut $1-\theta$ (on suppose que le fait de tomber sur la tranche est de probabilité nulle). La modélisation par la loi Binomiale $\mathrm{X}_i \sim \mathcal{B}(1, \theta)$ du tirage i permet de prendre en compte l'incertitude des conditions de l'expérience. Le(a) probabiliste travaille dans un cadre connu, et s'intéresse aux propriétés d'une loi de probabilité $\mathbb{P}$, par exemple au calcul de son espérance, qui donne une information de tendance,

$$\mathbb{E}(X) = \sum_{k=1}^{K} p_k x_k \text{ (v.a. discrète)} \, ; \, \mathbb{E}(X) = \int_x x f(x) dx \text{ (v.a. continue)} \, .$$

ou de sa variance qui indique la dispersion ou variabilité

$$\mathrm{var}(X) = \mathbb{E}(X - \mathbb{E}(X))^2.$$

Un joueur (statisticien) se demande maintenant si son adversaire triche au tirage de la pièce. Il ne connait donc pas la valeur de θ de son adversaire qu'il a besoin d'estimer. Il réalise une expérience statistique en notant le résultat de n tirages (réalisés dans les mêmes conditions et considérés comme indépendants) de son adversaire. Il obtient ainsi le résultat $x_1, \ldots, x_n$ d'un n-échantillon de la loi $\mathcal{B}(1, \theta)$ de paramètre θ inconnu. La statistique donne un cadre mathématique au joueur pour estimer θ à partir de l'observation partielle (n est fini) du phénomène, calculer l'incertitude de cette estimation, proposer une réponse à son interrogation sur la sincérité de l'adversaire (test) et le risque de cette réponse.



Pour résumer,

— Le rôle du probabiliste est d'étudier les propriétés des lois $\mathbb{P}$ de la famille $\mathcal{P}$.

— Le rôle du statisticien est inverse : à partir de l'observation d'une loi inconnue $\mathbb{P}$, il propose (infère) les propriétés de cette loi au sein d'une famille $\mathcal{P}$ et définit une procédure de décision pour répondre à la question posée.

1.2 Modèle statistique

La modélisation statistique est à la base de toute inférence statistique. Modéliser l'expérience, c'est proposer une loi théorique pour la variable aléatoire $X = (X_1, \ldots, X_n)$.

Définition 2. *Un **modèle statistique** est la donnée d'un espace mesurable $(\mathcal{X}^n, \mathcal{A}^n)$ muni d'une famille de lois de probabilité $\mathcal{P} = (\mathbb{P}_\theta^n)_{\theta\in\Theta}$:*

$$\mathcal{M} = (\mathcal{X}^n, \mathcal{A}^n, \mathbb{P}_\theta^n, \theta \in \Theta)$$

*Quand il existe $d \in \mathbb{N}^*$ tel que $\Theta \subset \mathbb{R}^d$, le modèle est dit **paramétrique**. Sinon, il est **non paramétrique**.*

Dans le cas paramétrique, la forme de la loi n'est connue qu'à la valeur du paramètre θ près, à inférer avec l'observation de l'échantillon. Dans le cas non paramétrique, la loi est considérée comme un élément d'un espace de dimension infinie, et il s'agira d'estimer ses coordonnées dans cette base. Nous considérerons uniquement le cas paramétrique dans ce cours.

Définition 3. *Une **observation** X est une variable aléatoire à valeurs dans $\mathcal{X}^n$ et dont la loi appartient à $\mathcal{P}$. Les **données** sont les réalisations (valeurs) $x_1, \ldots, x_n$ prises par l'échantillon $X_1, \ldots, X_n$.*

Par abus de langage, le résultat $X_i(\omega) = x_i$ de cette variable aléatoire peut parfois être également appelé observation.

Dans le cas d'un tirage indépendant dans une population infinie, ou le cas d'un échantillonnage avec remise dans une population finie, la loi de l'échantillon se factorise en produit des lois de chacune des composantes de l'échantillon.

Définition 4. *On appelle* ***n-échantillon i.i.d.*** *le modèle statistique de n composantes indépendantes et de même loi $\mathbb{P}_\theta$ (identiquement distribuées)*

$$\mathcal{M} = (\mathcal{X}^n, \mathcal{A}^n, \mathbb{P}_\theta^{\otimes n}, \theta \in \Theta).$$

La loi $\mathbb{P}_\theta$ est parfois appelée ***loi mère*** *de l'échantillon. On dit que l'échantillon est de* ***taille*** *n.*

Remarque Dans le cadre i.i.d, on dit parfois que le modèle est $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathbb{P}_\theta, \theta \in \Theta)$, puisque le processus d'échantillonnage est décrit par la caractérisation i.i.d.

Exemples Le choix de la loi mère dépend du phénomène observé : si le phénomène est binaire, la loi de Bernoulli $\mathcal{B}(1, \theta)$ est un choix naturel, comme dans l'exemple introductif :

$$\mathbb{P}(X = 1) = \theta; \;\; \mathbb{P}(X = 0) = 1 - \theta.$$

Le modèle est $\mathcal{M} = (\{0, 1\}^{\otimes n}, \mathcal{A}^n, \mathcal{B}(1, \theta)^{\otimes n}, \theta \in]0; 1[)$ ou $X_i \sim_{i.i.d.} \mathcal{B}(1, \theta)$, $\theta \in]0; 1[$.

Considérons maintenant une expérience où n mesures indépendantes d'une constante physique μ sont réalisées. On peut d'abord raisonnablement supposer que $X_i = \mu + \varepsilon_i$ $i = 1, \ldots, n$ où ε_i représente l'erreur pour l'observation i. Plusieurs types d'hypothèses sont alors envisageables (et donc plusieurs modèles) :

- $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ ne dépend pas de μ
- les ε_i sont indépendants
- les ε_i sont indépendants et identiquement distribués
- la loi de ε_i est continue (à densité), et symétrique autour de 0
- les ε_i sont i.i.d de loi gaussienne $\mathcal{N}(\mu, \sigma^2)$, σ^2 inconnu
- les ε_i sont i.i.d de loi gaussienne $\mathcal{N}(\mu, \sigma^2)$, σ^2 connu

Dans le dernier exemple, le paramètre du modèle est $\theta = \mu \in \mathbb{R}$, dans l'avant dernier $\theta = (\mu, \sigma^2) \in \mathbb{R} \times]0; +\infty[$, l'antépénultième est un exemple de modèle non paramétrique. On ne sait vraiment bien traiter que les deux derniers exemples. Si on utilise un modèle alors que ses hypothèses sont notoirement fausses sur le jeu de données, les conclusions risquent elles aussi d'être erronées. Il est donc important de pouvoir valider ces hypothèses.

1.2.1 Identifiabilité

L'inférence statistique (paramétrique) recherche dans la famille de loi paramétrée $\mathcal{P} = (\mathbb{P}_\theta^n)_{\theta \in \Theta}$ le "vrai" paramètre $\theta^\star$ avec lequel les données ont été générées. Ceci n'est possible que si la connaissance de la loi permet de déterminer de façon unique le paramètre, c'est à dire que le modèle est identifiable.

Définition 5. *Un modèle à paramétrage dans Θ est* ***identifiable*** *si*

$$\forall \theta, \theta' \in \Theta, \;\; \mathbb{P}_\theta = \mathbb{P}_{\theta'} \Rightarrow \theta = \theta'$$

Sans cette condition, on peut trouver θ et θ' dans θ tels que $\mathbb{P}_\theta = \mathbb{P}_{\theta'}$ et il est impossible de remonter à θ par la seule observation.

Exemple Soient $X_1 \sim \mathcal{N}(\alpha_1 + \beta, 1)$ et $X_2 \sim \mathcal{N}(\alpha_2 + \beta, 1)$, avec $\alpha_1, \alpha_2, \beta \in \mathbb{R}$. $\theta = (\alpha_1, \alpha_2, \beta)$ n'est pas identifiable tandis que $\theta = (\alpha_1, \alpha_2)$ l'est pour β connu.

1.2.2 Modèle dominé

En statistique, on travaille dans le cadre de modèles dominés.

Définition 6 (domination)**.** *La mesure* $\mathbb{P}$ *est* ***dominée*** *par rapport à une mesure positive* σ*-finie*[1] ξ*, si les ensembles négligeables pour* ξ *le sont également pour* $\mathbb{P}$*. On dit aussi que* $\mathbb{P}$ *est* ***absolument continue*** *par rapport à* ξ*.*

On peut alors définir une fonction $f = \frac{d\mathbb{P}}{d\xi}$ mesurable et positive définie ξ presque sûrement, dérivée de Radon-Nykodym de $\mathbb{P}$ par rapport à ξ. f est appelée **densité** de la loi de $\mathbb{P}$, et on note $\mathbb{P} = \int f d\xi$.

- modèle **continu** : $\mathbb{P}$ est absolument continue par rapport à la mesure de Lebesgue. f représente la densité au sens usuel (dérivée de la fonction de répartition).
- modèle **discret** : la densité est l'ensemble des probabilités ponctuelles $f(x) = \mathbb{P}(X = x)$ et ξ est la mesure de comptage.

Un modèle statistique est dominé quand la mesure ξ est commune à toutes les probabilités $\mathbb{P} \in \mathcal{P}$. On a :

$$\forall A \in \mathcal{A}, \;\; \mathbb{P}_\theta(A) = \int_A f_\theta(x)\, d\xi(x)$$

La notion de modèle statistique dominé permet d'englober les cas discret et continu dans un même cadre.

Exemples La loi gaussienne $\mathcal{N}(\mu, \sigma^2)$ d'espérance μ et de variance σ^2 est dominée par la mesure de Lebesgue ν, sa densité est

$$f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp - \left(\frac{(x-\mu)^2}{2\sigma^2} \right)$$

et sa fonction de répartition

$$F(x) = \int_{-\infty}^{x} f_{\mu,\sigma^2}(t) d\nu(t) = \int_{-\infty}^{x} f_{\mu,\sigma^2}(t) dt$$

Un n-échantillon iid gaussien est dominé par $\nu^{\otimes n}$, et sa densité est $\prod_{i=1}^n f_{\mu,\sigma^2}(x_i)$. La loi de Poisson est dominée par la mesure de comptage $\sum_{k\in\mathbb{N}} \delta_k$ où δ_k est le Dirac en k. Sa densité s'écrit, pour $k \in \mathbb{N}$

$$f_\lambda(k) = e^{-\lambda} \frac{\lambda^k}{k!} = \mathbb{P}(X = k)$$

1.2.3 Vraisemblance

Définition 7. *Dans un modèle paramétrique dominé, on appelle* ***vraisemblance*** *d'une réalisation* $x = (x_1, \ldots, x_n)$ *du* n*-échantillon, la fonction de* θ *telle que :*

$$\theta \mapsto L(\theta; x_1, \ldots, x_n) = f_\theta(x_1, \ldots, x_n)$$

Pour un échantillon i.i.d. : $L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^n f_\theta(x_i)$.

Attention, la vraisemblance (fonction de θ) n'est pas la densité (fonction de x), même si pour θ et x fixés elles ont les mêmes expressions.

1. Une mesure ξ sur un espace mesurable $(\mathcal{X}, \mathcal{A})$ est σ-finie si $\mathcal{X}$ est réunion d'une famille dénombrable d'ensembles mesurables de mesure finie

Exemples La vraisemblance d'une réalisation du n-échantillon i.i.d. de loi gaussienne $\mathcal{N}(\mu, \sigma^2)$ de variance connue σ^2 est la fonction de μ définie par :

$$\mu \mapsto L(\mu; x) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp - \left(\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2} \right)$$

Si σ^2 est également inconnue, le paramètre $\theta = (\mu, \sigma^2)$ est bidimensionnel et la vraisemblance est la fonction

$$\theta = (\mu, \sigma^2) \mapsto L(\theta; x) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp - \left(\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2} \right)$$

La loi de Bernoulli $\mathcal{B}(1, \theta)$ d'espérance θ admet la densité $\theta^x (1-\theta)^{1-x}$ définie par rapport à la mesure discrète $\delta_0 + \delta_1$. La vraisemblance de l'observation d'un n-échantillon i.i.d. de loi de Bernoulli $\mathcal{B}(1, \theta)$ s'écrit donc

$$L(\theta; x_1, \ldots, x_n) = \prod_{i=1}^n \theta^{x_i} (1-\theta)^{1-x_i}$$

A θ fixé, $L(\theta; X)$ (où $X \sim \mathbb{P}_\theta^{\otimes n}$) est une variable aléatoire, vraisemblance de l'observation.

Notons que la densité et donc la vraisemblance ne sont définies qu'à un ensemble de mesure nulle près, on doit parler en toute rigueur de version de densité par rapport à la mesure dominante. Mais on conviendra de prendre la version de la densité la plus régulière (continue, indéfiniment dérivable, ...), et lorsque ce choix est clair, on parlera abusivement de *la* vraisemblance.

Dans le cas de modèles i.i.d., la vraisemblance de l'échantillon est le produit des vraisemblances individuelles de chacune des observations. Il est donc souvent pratique de travailler avec la **log-vraisemblance** pour remplacer le produit par une somme :

$$\log L(\theta; X_1, \ldots, X_n) = \sum_{i=1}^n \log f_\theta(X_i) := \sum_{i=1}^n \ell_\theta(X_i)$$

et on note $\ell_\theta = \log f_\theta$.

1.3 Des modèles de plus en plus complexes

Le modèle le plus simple en statistique inférentielle est le modèle paramétrique i.i.d. avec un paramètre d'intérêt en général unidimensionel ; il s'étend facilement au cas de deux échantillons indépendants chacun étant i.i.d., mais pas forcément de même loi. Qu'en est-il de la comparaison de plus de deux échantillons, ou de l'étude d'un phénomène qui dépend de plusieurs facteurs externes ?

Les modèles de régression permettent d'étendre la modélisation statistique à des cas plus généraux, où il s'agit d'expliquer ou de prédire une variable aléatoire *réponse* Y en fonction d'une liste de variables *explicatives* $X = (X_1, \ldots, X_p)$ observées sur des individus ou des objets. C'est le cas de très nombreuses situations, par exemple :

- On observe la taille d'alevins en fonction de leur âge. Quelle est la taille moyenne d'un alevin à un âge donné ? Quelle taille aura un alevin pris au hasard à un âge donné ?
- On dispose de la vitesse du vent, de la pluviométrie, d'une image satellite de Paris. Quelle sera la concentration d'ozone demain sur Paris ?
- Est-il possible de classer automatiquement un courriel (normal, indésirable) en fonction de certains mots ou ponctuations dans le corps du message ?

— Un organisme de crédit dispose de son portefeuille de clients (CSP, soldes, découverts, incidents). Peut-il accepter un nouveau crédit pour un client existant sans risque d'incident ? et pour un nouveau client ?
— Comment choisir les destinataires d'un mailing pour maximiser le taux de réponses positives, tout en minimisant le nombre d'envois ?
— La consommation d'alcool est-elle un facteur de risque pour le cancer du foie ?

Quelle est l'information utile ?
— Les événements observés fournissent une base de connaissance pour expliquer un phénomène ou prédire de nouvelles configurations
— Il s'agit de tirer partie des *corrélations* entre les phénomènes pour en prévoir d'autres.

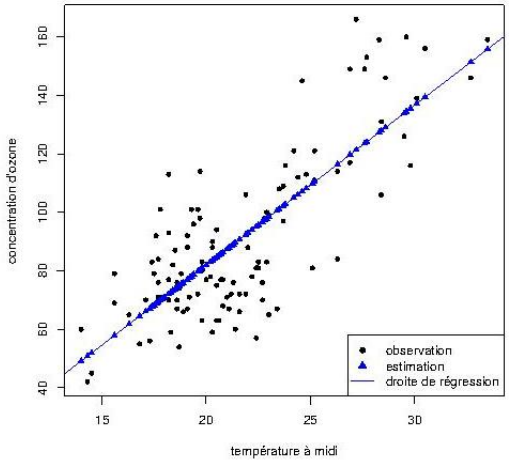


FIGURE 1.1 – Régression linéaire : Concentration maximale journalière d'ozone (en $\mu g/m^3$) en fonction de la température à midi dans la région de Rennes Cornillon et Matzner-Løber (2007)

Les modèles de régression font partie des modèles d'apprentissage supervisé. *Apprentissage* fait référence à un processus automatisable sur un ensemble de données pour en extraire de la connaissance ; *supervisé* indique que ce processus possède des observations qui servent de référence pour construire les estimateurs.

Nous étudierons le modèle de régression linéaire qui postule l'additivité d'un bruit centré i.i.d à une fonction déterministe appelé fonction de régression. La fonction de régression modélise l'espérance de Y en fonction des variables explicatives et est linéaire en un paramètre θ,

$$Y_i = \sum_{j=1}^{p} x_{ij}\theta_j + \varepsilon_i$$

C'est le cas de la modélisation de la concentration d'ozone Y_i en fonction de la température à midi t_i de la figure 1.1 : on a $x_{i1} = 1$, $x_{i2} = t_i$ et $\theta = (\theta_1, \theta_2)$ où θ_1 est l'ordonnée à l'origine de la droite de régression et θ_2 sa pente.

Nous verrons que le modèle linéaire permet également d'étendre la comparaison de deux échantillons à celle de K échantillons qui ne diffèrent que par le niveau d'un variable explicative qualitative : c'est le cas de la modélisation des poids de portées en fonction de génotype de leur mère par exemple.

1.4 Les étapes de la démarche statistique

Nous terminons ce chapitre introductif par rappeler les étapes de la démarche générale de l'inférence statistique.

A partir d'un échantillon de n observations, il s'agit de déduire -ou d'inférer- certaines propriétés du modèle probabiliste inconnu qui les a générées, puis d'utiliser ces estimations à des fins de décision ou de prédiction. Cette démarche peut être résumée par les étapes suivantes :

1. Acquérir et préparer les données, prendre en compte leur nature, effectuer une analyse descriptive et exploratoire.
2. Définir un modèle adapté à la situation observée :
   — loi de probabilité de la variable réponse,
   — équation liant l'espérance de la réponse et les covariables.
3. Estimer les paramètres du modèle grâce aux observations.
4. Vérifier l'adéquation de l'estimation aux observations (diagnostics d'ajustement, analyse des résidus, tests de loi, indépendance).
5. Vérifier la capacité de généralisation du modèle, c'est à dire le bon comportement du modèle sur des données non encore observées.
6. Utiliser le modèle à des fins de décision ou de prédiction.

Ces étapes ne s'enchaînent pas forcément de façon linéaire. En effet, le processus est itératif en fonction des résultats obtenus à chacune d'elle. Plusieurs modèles pourront être mis en compétition, que des procédures de choix de modèle permettront de départager.

Chapitre 2

Estimateurs

Pour déterminer la hauteur moyenne des plans d'un champ de maïs, un échantillon de n plans de maïs est constitué par tirage aléatoire. La hauteur d'un plan est modélisée par une variable aléatoire gaussienne $\mathcal{N}(\mu, \sigma^2)$. On cherche, à partir d'un échantillon $X = (X_1, \ldots, X_n)$, à estimer l'espérance μ. Estimer la valeur de μ (inconnu) à partir de la réalisation d'un échantillon c'est :

- Préciser le modèle
- Construire un estimateur
- Etudier ses propriétés, aussi bien à distance finie (n fixé) qu'asymptotiquement (quand n tend vers l'infini). On souhaite un estimateur qui est précis à distance finie, qui converge vite quand la taille de l'échantillon augmente.
- Comparer les estimateurs
- Identifier s'il y en a un meilleur que tous les autres, et avec quel critère : question d'optimalité

Nous développons ces aspects dans ce chapitre, qui commence par rappeler des propriétés (biais, variance, risque) à distance finie, puis présente les notions d'exhaustivité et d'information de Fisher, l'efficacité et la borne de Cramer-Rao, et l'optimalité (estimateurs UVMB). Ce chapitre termine par les propriétés asymptotiques (consistance, loi asymptotique, delta-méthode).

2.1 Estimation ponctuelle

Dans un modèle statistique paramétrique, on cherche à estimer la valeur du paramètre inconnu. L'estimation peut être ponctuelle (ce chapitre), ou par intervalle (ou région si le paramètre est multidimensionnel) de confiance (cf chapitre 5).

Définition 8. *Toute fonction $t(.)$ de l'échantillon $X = (X_1, \ldots, X_n)$ et calculable à partir de l'observation, est une* ***statistique****. En particulier, la définition de t ne dépend pas du paramètre inconnu.*

Par exemple, la moyenne empirique, $t(X) = \max_i X_i$, $t(X) = 0$, $t(X) = (X_{(1)}, \ldots, X_{(n)})$ où $X_{(1)} \leq \ldots \leq X_{(n)}$ (statistique d'ordre) sont des statistiques. Une statistique est un résumé de l'échantillon.

Quand une statistique est utilisée dans le but d'estimer un paramètre θ (ou une fonction $\nu(\theta)$ du paramètre, par exemple l'espérance de la loi mère, sa variance, un quantile,...), c'est un estimateur de θ (ou de $\nu(\theta)$).

Définition 9. *Soit $X_1, \ldots, X_n$ un n-échantillon d'une loi $\mathbb{P}_\theta \in \mathcal{P}$. Un estimateur de $\nu(\theta)$ est une variable aléatoire T_n, fonction mesurable de l'échantillon et calculable sur l'échantillon*

$$T_n = t(X_1, \ldots, X_n).$$

On la note souvent $\widehat{\nu} = T_n$.

Une estimation ponctuelle de $\nu(\theta)$ est calculée à partir d'une réalisation $t_n = t(x_1, \ldots, x_n)$ de T_n. On note en général les réalisations en lettres minuscules et la variable aléatoire en lettres majuscules. Si les notations utilisent des lettres grecques, on ne différencie pas, par abus de notation, la variable aléatoire de sa réalisation, comme dans l'exemple suivant :

Exemple L'estimateur empirique $\widehat{\mu}$ de l'espérance μ d'une loi est

$$\widehat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

On trouve aussi la notation $\bar{X}$ pour désigner l'estimateur empirique de l'espérance, et sa réalisation $\bar{x}$.

2.2 Propriétés d'un estimateur

Bien évidemment, il existe des estimateurs plus ou moins pertinents : le biais, la variance et le risque sont des critères de qualité d'un estimateur. Soit $\widehat{\nu}$ l'estimateur d'une quantité $\nu(\theta)$ dépendant de la loi $\mathbb{P}_\theta$

Définition 10. *Le **biais** est défini par*

$$Biais(\widehat{\nu}) = \mathbb{E}(\widehat{\nu} - \nu(\theta)) = \mathbb{E}(\widehat{\nu}) - \nu(\theta)$$

$\widehat{\nu}$ est sans biais ssi $Biais(\widehat{\nu}) = 0$

Définition 11. *La **variance** est* $\mathrm{var}(\widehat{\nu}) = \mathbb{E}[(\widehat{\nu} - \mathbb{E}(\widehat{\nu}))^2]$

L'analyse ne s'arrête pas à une valeur trouvée, le statisticien s'intéresse à calibrer l'erreur qu'il commet en prenant une décision.

Définition 12. *Le **risque** de l'estimateur $\widehat{\nu}$ pour une fonction de perte $\ell(\widehat{\nu}, \nu)$ de Θ dans $\mathbb{R}^+$ est*

$$R(\widehat{\nu}) = \mathbb{E}(\ell(\widehat{\nu}, \nu)),$$

*Quand la fonction de perte est $\ell(\widehat{\nu}, \nu) = (\widehat{\nu} - \nu)^2$ le risque est appelé **risque quadratique** ou **erreur quadratique moyenne** (Mean Square Error en anglais).*

Exemple L'erreur quadratique moyenne de l'estimateur empirique de l'espérance μ est

$$EQM(\bar{X}) = \mathbb{E}(\bar{X} - \mu)^2 = Var(\bar{X}) = \frac{\sigma^2}{n}.$$

Il ne dépend pas de l'espérance μ, mais décroît avec le nombre d'observations n et croît avec la variance (variabilité) du phénomène observé σ^2.

— Si la variance est connue égale à σ_0^2 ou $\leq \sigma_0^2$, alors, le risque **a priori** de $\bar{X}$ peut être estimé par $EQM(\bar{X})$. Si on veut garantir $EQM(\bar{X}) \leq \epsilon^2$, on peut le faire en choisissant au moins $n_0 = (\sigma_0/\epsilon)^2$ observations.
— Si on n'a aucune idée de la valeur de σ^2, la planification n'est pas possible, mais on peut utiliser les n expériences pour estimer σ^2, par exemple, $\widehat{\sigma}^2 = \frac{1}{n}\sum_i (X_i - \bar{X})^2$ ou $n\widehat{\sigma}^2/(n-1)$. L'estimation **a posteriori** du risque est $\widehat{\sigma}^2/n$, elle-même soumise à une erreur aléatoire.

De façon générale, EQM dépend de la variance de $\widehat{\nu}$ et de son biais (nul dans le cas de $\bar{X}$), que l'on peut voir comme "l'erreur au long cours" de $\widehat{\nu}$. En effet, en écrivant $(\widehat{\nu}-\nu) = [\widehat{\nu}-E(\widehat{\nu})]+[E(\widehat{\nu})-\nu]$, on obtient la formule de la décomposition biais-variance :

$$EQM(\widehat{\nu}) = (Biais(\widehat{\nu}))^2 + Var(\widehat{\nu}).$$

Le risque quadratique permet de définir une relation d'ordre partiel sur les estimateurs :

Définition 13. *Un estimateur δ_1 de $\nu(\theta)$* ***domine*** *l'estimateur δ_2 si, pour tout $\theta \in \Theta$,*

$$R_\theta(\delta_1, \nu) \leq R_\theta(\delta_2, \nu)$$

cette inégalité étant stricte pour au moins une valeur de ν.
Un estimateur est ***admissible*** *s'il n'existe aucun estimateur le dominant. Sinon, il est* ***inadmissible****.*

Il n'est pas possible en général de minimiser le risque quadratique uniformément : on recherchera donc des estimateurs optimaux dans des sous classes, par exemple celle des estimateurs sans biais : estimateurs UVMB, Uniformément de Variance Minimale parmi les estimateurs sans Biais, cf. section 2.5 de ce chapitre.

2.3 Exhaustivité

Comment construire un bon résumé de l'échantillon pour estimer un paramètre ? L'échantillon apporte une certaine "*information*" sur le paramètre θ. Lorsque l'on résume l'échantillon par une statistique, il ne faut pas réduire l'information de l'échantillon initial. L'information peut être définie suivant deux approches :

— une propriété de la statistique $T = t(X)$. T ne peut nous renseigner sur le paramètre θ que si sa loi dépend du paramètre θ. En particulier, si la loi conditionnelle de $X = (X_1, \ldots, X_n)$ à T fixé ne dépend pas du paramètre θ, il ne reste plus, dans l'échantillon, d'information supplémentaire à T concernant θ : on dit alors que la statistique T est **exhaustive** pour θ.
— une définition mathématique d'une quantité d'information. Il s'agit alors de chercher dans quelles circonstances cette quantité se conserve lorsque l'échantillon est résumé : **information de Fisher**

Ces deux approches sont développées dans cette section.

Définition 14. *On dit que $t(X)$ est une* ***statistique exhaustive*** *pour $\theta \in \Theta \subset \mathbb{R}^p$ si la loi de $X = (X_1, \ldots, X_n)$ conditionnellement à $t(X) = t$ n'est pas une fonction du paramètre θ*

$$\mathbb{P}_\theta(X | t(X) = t) \text{ ne dépend pas de } \theta.$$

La notion d'exhaustivité n'implique pas forcément une réduction de dimension : l'échantillon X est exhaustif. Mais c'est une réduction suffisante pour ne pas perdre d'information : on parle aussi de **statistique suffisante** (*sufficient* en anglais).

Exemple Dans le cas de l'estimation du paramètre θ de la loi de Bernoulli (cf exemple du joueur), on a

$$\mathbb{P}_\theta(X_1 = 1, \ldots, X_n = x_n) = \theta^s(1-\theta)^{n-s}, \text{ où } s = \sum_{i=1}^n x_i$$

et, avec $S = \sum_{i=1}^n X_i$,

$$\mathbb{P}_\theta(X = x|S = s) = \frac{\mathbb{P}_\theta(X = x, S = s)}{\mathbb{P}_\theta(S = s)} = \frac{\mathbb{P}_\theta(X = x, S = s)}{\sum_y \mathbb{P}_\theta(X = y, S = s)} = \frac{\mathbb{1}_{\{S=s\}}}{\binom{n}{s}}$$

La statistique S est exhaustive pour estimer θ. Le théorème de factorisation est bien utile pour vérifier si une statistique est exhaustive

Théorème 1 (de factorisation)**.** *Soit T une statistique de $(\mathcal{X}^n, \mathcal{A}^n) \to (\mathcal{Y}, \mathcal{B})$. T est exhaustive pour θ si et seulement s'il existe deux fonctions mesurables positives $g : \mathcal{Y} \to \mathbb{R}^+$ et $h : \mathcal{X}^n \to \mathbb{R}^+$ telles que la densité $f(x;\theta)$ de l'échantillon se factorise sous la forme*

$$f(x;\theta) = h(x)g(t(x);\theta) \tag{2.1}$$

où $x = (x_1, \ldots, x_n)$.

Nous proposons une preuve dans le cas discret (cf Bickel et Doksum (2015)).

Preuve
Soit $\mathcal{X}^n$ l'ensemble des réalisations possibles de X et soit $t_i = t(x_i)$. Alors, T est discret et $\sum_{i=1}^\infty \mathbb{P}_\theta[T = t_i] = 1$ pour tout θ.

Condition suffisante : supposons la factorisation suivant (2.1), on doit montrer que $\mathbb{P}_\theta[X = x_j|T = t_i]$ est indépendant de θ pour tout i et tout j. Par définition de la probabilité conditionnelle dans le cas discret, il est suffisant de monter que $\mathbb{P}_\theta[X = x_j|T = t_i]$ est indépendant de θ sur chacun des ensembles $V_i = \{\theta : \mathbb{P}_\theta[T = t_i] > 0\}$, $i \in \mathbb{N}$. Si la condition (2.1) est vérifiée,

$$\mathbb{P}[T = t_i] = \sum_{x:T(x)=t_i} f(x,\theta) = g(t_i;\theta) \sum_{x:T(x)=t_i} h(x).$$

Pour $\theta \in V_i$, la probabilité conditionnelle s'écrit

$$\begin{aligned}
\mathbb{P}_\theta[X = x_j|T = t_i] &= \mathbb{P}_\theta[X = x_j, T = t_i]/\mathbb{P}_\theta[T = t_i] \\
&= \frac{f(x_j;\theta)}{\mathbb{P}_\theta[T = t_i]} \mathbb{1}_{T(x_j)=t_i} \\
&= \frac{g(t_i;\theta)h(x_j)}{\mathbb{P}_\theta[T = t_i]} \text{ si } T(x_j) = t_i \text{ et 0 sinon}
\end{aligned}$$

Donc, si $T(x_j) \neq t_i$, $\mathbb{P}_\theta[X = x_j|T = t_i] = 0$, et si $T(x_j) = t_i$

$$\mathbb{P}_\theta[X = x_j|T = t_i] = \frac{g(t_i;\theta)h(x_j)}{\mathbb{P}_\theta[T = t_i]} = \frac{g(t_i;\theta)h(x_j)}{g(t_i;\theta)\sum_{x:T(x)=t_i} h(x)} = \frac{h(x_j)}{\sum_{x:T(x)=t_i} h(x)}$$

et donc T est exhaustive.

Réciproquement, si T est exhaustive, soit $g(t_i,\theta) = \mathbb{P}_\theta[T = t_i]$, $h(x) = \mathbb{P}[X = x_j|T = t_i]$, alors par définition de la probabilité conditionnelle

$$f(x;\theta) = \mathbb{P}_\theta[X = x, T = t(x)] = g(t(x);\theta)h(x)$$

□

Exemple

— Dans le cas d'un échantillon indépendant de loi Binomiale, $X_i \sim \mathcal{B}(n_i, \theta)$, la statistique $S = \sum_i X_i$ est exhaustive

— Dans le cas d'un n-échantillon iid de loi $\mathcal{N}(\theta, 1)$, $S = \sum_i X_i$ est également exhaustive pour estimer θ.

— Dans le cas d'un n-échantillon iid de loi uniforme sur $[0; \theta]$, $T = \max_{1 \leq i \leq n} X_i$ est une statistique exhaustive pour estimer θ. Dans les cas précédents, la statistique $T = X_1$ ne l'est pas.

Le principe de factorisation donne un moyen de reconnaître si une statistique est exhaustive, mais permet difficilement de la construire ou de savoir s'il en existe une.

Jusqu'à quel point peut-on réduire l'échantillon pour ne pas perdre d'information sur l'estimation du paramètre ? : jusqu'à obtenir une statistique minimale, c'est à dire dont on ne peut plus réduire la dimension.

Définition 15. *On dit que la statistique T_n est **exhaustive minimale** si elle est exhaustive, et si pour toute statistique exhaustive S_n, on peut trouver une fonction u telle que $T_n = u(S_n)$.*

Ainsi,

— Une statistique exhaustive minimale est fonction de n'importe quelle autre statistique exhaustive.

— Tout estimateur pertinent est fonction d'une statistique exhaustive minimale

— deux statistiques exhaustives minimales pour θ sont en liaison bijective

— si $\Theta \subset \mathbb{R}^K$, une statistique exhaustive de dimension K est en règle générale minimale (mais il n'existe pas forcément de stat exhaustive à valeur dans $\mathbb{R}^k$ pour estimer $\theta \in \mathbb{R}^k$.)

Le théorème suivant donne une condition suffisante pour qu'une statistique soit exhaustive minimale

Théorème 2. *Soit X un n-échantillon iid de densité $f(x_1, \ldots, x_n; \theta)$. Soit une statistique T. Si on a l'équivalence*

$$T(x_1, \ldots, x_n) = T(y_1, \ldots, y_n) \Leftrightarrow \theta \mapsto \frac{f(x_1, \ldots, x_n; \theta)}{f(y_1, \ldots, y_n; \theta)} \text{ ne dépend pas de } \theta \tag{2.2}$$

alors T est une statistique exhaustive minimale.

Preuve

Montrons d'abord que l'équivalence implique l'exhaustivité de T pour θ. Pour tout $t \in T(\mathcal{X})$, soit l'ensemble

$$\mathcal{T} = \{x \in \mathcal{X} | T(x) = t\}.$$

A tout élément $x \in \mathcal{X}$, on associe $x_t \in \mathcal{T}$, on a par construction $T(x) = T(x_{T(x)})$. Donc, en utilisant l'implication de (2.2), le rapport suivant

$$h(x) = \frac{f(x; \theta)}{f(x_{T(x)}; \theta)}$$

est indépendant de θ. Soit maintenant la fonction $g(T(x); \theta) = f(x_{T(x)}; \theta)$. On a

$$f(x, \theta) = h(x) g(T(x); \theta)$$

et la statistique est exhaustive par factorisation.

Montrons maintenant que la statistique exhaustive T est minimale. Soit S une autre statistique exhaustive. Par le théorème de factorisation, il existe deux fonctions $\tilde{g}$ et $\tilde{h}$ telles que $f(x;\theta) = \tilde{h}(x)\tilde{g}(S(x);\theta)$. Ainsi, pour tout x et y tels que $S(x) = S(y)$, le rapport

$$\frac{f(x;\theta)}{f(y;\theta)} = \frac{\tilde{h}(x)}{\tilde{h}(y)}$$

ne dépend pas de θ. La réciproque de (2.2) implique que $T(x) = T(y)$, et donc il existe une fonction de u telle que $T = u(S)$. Donc T est (exhaustive) minimale.

□

Exemple Pour la loi gaussienne $\mathcal{N}(\mu, \sigma^2)$ à espérance et variance inconnues, le couple

$$\left(\sum_i X_i^2, \sum_i X_i\right)$$

est exhaustif minimal. En revanche, la famille des lois de Weibull à deux paramètres de densité

$$f(x, \alpha, \lambda) = \alpha\lambda x^{\alpha-1} e^{\lambda x^\alpha} \mathbb{1}_{x>0}$$

n'admet pas de statistique exhaustive de dimension 2. Une statistique exhaustive minimale est même ici de dimension n, c'est le vecteur des statistiques d'ordre.

2.4 Information de Fisher

Sous certaines conditions de régularité, il est possible de définir une caractéristique du modèle appelée Information de Fisher, importante à deux titres :

— d'une part, on peut montrer que c'est la borne minimale de la variance d'un estimateur dans le modèle considéré (cf section 2.4.3 de ce chapitre)

— d'autre part, elle intervient dans la définition de la loi asymptotique de l'estimateur du maximum de vraisemblance (cf chapitre 3)

Si pour tout $x \in \mathcal{X}$, la log-vraisemblance d'une réalisation est différentiable (en θ), on définit le vecteur gradient de la log-vraisemblance

$$\dot{\ell}_\theta = \left(\frac{\partial}{\partial\theta_1}\ell_\theta, \ldots, \frac{\partial}{\partial\theta_p}\ell_\theta\right)'$$

où la notation prime $'$ désigne la transposée. Ce vecteur aléatoire est appelé **score**. Il s'exprime à partir du gradient de f_θ par rapport à θ, noté $\dot{f}_\theta$

$$\dot{\ell}_\theta = \frac{\dot{f}_\theta}{f_\theta}$$

Donc, si $\dot{\ell}_\theta$ est intégrable

$$\mathbb{E}_\theta[\dot{\ell}_\theta(X)] = \int \dot{\ell}_\theta(x) f_\theta d\xi(x) = \int \frac{\dot{f}_\theta}{f_\theta} f_\theta d\xi(x) = \int \dot{f}_\theta d\xi(x)$$

et si on peut échanger le signe somme et la dérivation, le score est centré : $\mathbb{E}_\theta[\dot{\ell}_\theta(X)] = 0$.

Définition 16. *Dans un modèle paramétrique dominé tel que* $\mathbb{E}_\theta[\dot{\ell}(X)] = 0$ *et* $\mathbb{E}_\theta[||\dot{\ell}(X)||^2] < \infty$*, la matrice de variance du score s'appelle* ***Information de Fisher*** *au point* $\theta \in \Theta \subset \mathbb{R}^p$

C'est une matrice de taille $p \times p$, et, le score étant centré, se réduit à l'expression

$$I(\theta) = \mathbb{E}_\theta[\dot{\ell}_\theta(X)\dot{\ell}'_\theta(X)]$$

Exemple La log-vraisemblance d'un modèle iid gaussien $\mathcal{N}(\mu, \sigma^2)$, σ^2 connu est

$$\ell_\mu(X) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{\sum_i (X_i - \mu)^2}{2\sigma^2}$$

de score

$$\dot{\ell}_\mu(X) = \frac{\sum_i (X_i - \mu)}{\sigma^2}$$

On vérifie que le score est centré : $\mathbb{E}(\dot{\ell}_\mu(X)) = 0$. L'information de Fisher est

$$I_n(\mu) = \mathbb{E}[(\dot{\ell}_\mu(X))^2] = \mathrm{var}(\dot{\ell}_\mu(X)) = \mathrm{var}\left[\frac{\sum_i (X_i - \mu)}{\sigma^2}\right] = \frac{n}{\sigma^4}\mathrm{var}(X_1) = \frac{n}{\sigma^2}$$

Dans le modèle de Bernoulli, $I_n(\theta) = n/(\theta(1-\theta))$.

Score et information de Fisher ont des propriétés intéressantes dans le cas de modèles réguliers, que nous commençons par définir.

Définition 17. *Un modèle paramétrique* $(\mathcal{X}, \mathcal{A}, \mathbb{P}_\theta)$*,* $\theta \in \Theta$ *ouvert de* R^p*, et tel que* $\mathbb{P}_\theta$ *admet une densité* $f(.;\theta)$ *par rapport à une mesure dominante* ν *est* ***régulier*** *si*

1. *Le support des lois* $f(.;\theta)$ *est indépendant de* $\theta \in \Theta$
2. $\theta \mapsto f(x;\theta)$ *est deux fois continûment différentiable sur* Θ*, pour tout* x *du support*
3. *Pour tout* $A \in \mathcal{A}$*, l'intégrale* $\int_A f(x;\theta)d\nu(x)$ *est au moins deux fois dérivable sous le signe d'intégration et on peut permuter intégration et dérivation*

Ainsi, on a, pour tout $\theta \in \Theta$ et tout $A \in \mathcal{A}$

$$\frac{\partial}{\partial\theta_j}\int_A f(x;\theta)d\nu(x) = \int_A \frac{\partial}{\partial\theta_j} f(x;\theta)d\nu(x), \quad j = 1, \ldots, p$$

$$\frac{\partial^2}{\partial\theta_j\partial\theta_k}\int_A f(x;\theta)d\nu(x) = \int_A \frac{\partial^2}{\partial\theta_j\partial\theta_k} f(x;\theta)d\nu(x), \quad j,k = 1, \ldots, p$$

Il existe des conditions suffisantes sur $f(x;\theta)$ pour obtenir la condition 3. : par exemple, l'existence pour tout θ d'un voisinage V_θ tel que

$$\sup_{\tilde{\theta}\in V_\theta} \frac{\partial}{\partial\theta_j} f(x;\tilde{\theta}) \in \mathbb{L}^1(\nu)$$

C'est le cas par exemple quand les intégrales $\int_A \frac{\partial}{\partial\theta_j} f(x;\theta)d\nu(x)$ et $\int_A \left|\frac{\partial}{\partial\theta_j} f(x;\theta)\right| d\nu(x)$ sont des fonctions continues de θ.

Les modèles gaussien ou de Bernoulli sont bien évidemment réguliers. En revanche, un échantillon iid de loi $\mathcal{U}[0,\theta]$ où $\theta > 0$ est inconnu n'est pas régulier, puisque le support de la loi dépend du paramètre inconnu.

Propriété 1. *Le score d'un modèle régulier est centré. De plus, il est* ***additif*** *: pour deux variables aléatoires indépendantes X et Y associées aux modèles $(\mathcal{X}, P_\theta)$ et $(\mathcal{Y}, Q_\theta)$*

$$\dot{\ell}_\theta(X,Y) = \dot{\ell}^P_\theta(X) + \dot{\ell}^Q_\theta(Y)$$

Cette propriété qui se déduit directement de la factorisation de la densité jointe. En particulier, pour un modèle iid, $\dot{\ell}(X,\theta) = \sum_i \dot{\ell}(X_i,\theta)$.

Propriété 2. *La matrice d'information de Fisher d'un modèle régulier est additive, symétrique, semi-définie positive et vérifie*

$$\begin{aligned} I_n(\theta) &= \mathbb{E}_\theta[\dot{\ell}_\theta(X)[\dot{\ell}_\theta(X)]'] \\ &= -\mathbb{E}_\theta[\ddot{\ell}_\theta(X)] \end{aligned}$$

où $\ddot{\ell}_\theta(X)$ désigne la matrice des dérivées secondes en θ de la log-vraisemblance

L'information de Fisher étant une matrice de variance, elle est symétrique, semi-définie et positive. L'additivité découle de celle de la log-vraisemblance pour un échantillon indépendant. Montrons l'égalité dans le cas d'un paramètre unidimensionnel

$$\begin{aligned} \mathbb{E}_\theta[\ddot{\ell}_\theta(X)] &= \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\dot{\ell}_\theta(X)\right] = \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta}\frac{\dot{f}_\theta(X)}{f_\theta(X)}\right] = \mathbb{E}_\theta\left[\frac{\ddot{f}_\theta(X)f_\theta(X) - (\dot{f}_\theta(X))^2}{(f_\theta(X))^2}\right] \\ &= \int \ddot{f}_\theta(x)d\nu(x) - \mathbb{E}_\theta\left[(\dot{\ell}_\theta(X))^2\right] \end{aligned}$$

Or la première intégrale est nulle puisqu'on peut échanger intégration et dérivation, d'où le résultat.

Exemple : Le modèle iid gaussien $\mathcal{N}(\mu,\sigma^2)$, σ^2 connu, étant régulier, on peut aussi calculer l'information de Fisher à partir de la dérivée seconde de la log-vraisemblance :

$$I_n(\mu) = -\mathbb{E}\left[\frac{-n}{\sigma^2}\right] = \frac{n}{\sigma^2}$$

2.4.1 Interprétation de l'information de Fisher

L'information de Fisher mesure l'information apportée par chaque observation sur l'estimation du paramètre du modèle :

— Si $X = (X_1, \ldots, \mathrm{X}_n)$ est un n**-échantillon iid** d'information $I_n(\theta)$, alors

$$I_n(\theta) = nI_1(\theta)$$

chaque observation apporte la même quantité d'information à la connaissance du paramètre.

— L'information de Fisher est liée à la **précision** avec laquelle le paramètre est estimé : ce point sera mis en lumière plus tard.

— L'information $I_T(\theta)$ portée par une statistique quelconque $T(X)$ est inférieure ou égale à celle apportée par l'échantillon $X = (X_1, \ldots, X_n)$

$$I_T(\theta) \leq I_n(\theta)$$

En effet, soit $T(X)$ une statistique de densité $g(t,\theta)$ qu'on substitue à l'échantillon, alors

$$L(x;\theta) = g(t,\theta)h(x,\theta|t)$$

où h est la densité de l'échantillon conditionnellement à $T=t$. On a :

$$\frac{\partial^2}{\partial\theta^2}\log L(x,\theta) = \frac{\partial^2}{\partial\theta^2}\log g(t,\theta) + \frac{\partial^2}{\partial\theta^2}\log h(x,\theta|t)$$
$$-I_n(\theta) = -I_T(\theta) - \mathbb{E}\mathbb{E}_{X|T}\left[\frac{\partial^2}{\partial\theta^2}\log h(X,\theta|T)\right]$$

Or le dernier terme est positif ou nul, puisque c'est l'espérance (en T) de l'information de Fisher de $X|T=t$.

— Il vaut mieux choisir des statistiques qui permettent de ne pas perdre d'information pour l'estimation d'un paramètre. Les statistiques exhaustives ont cette propriété.

2.4.2 Borne Fréchet-Darmois-Cramér-Rao

La variance d'un estimateur ne peut pas être inférieure à une certaine borne, qui dépend de la quantité d'information (de Fisher) apportée par l'échantillon. Le théorème suivant précise cette propriété :

Théorème 3 (FDCR)**.** *Si le modèle est régulier et l'information de Fisher $I_n(\theta)$ inversible, on a, pour tout estimateur **sans biais** T_n de θ tel que $\mathbb{E}(|T_n|) < \infty$:*

$$\mathrm{var}(T_n) \geq I_n(\theta)^{-1}$$

*Soit h une fonction différentiable sur Θ. Pour tout estimateur T_n **sans biais** de $h(\theta)$ tel que $\mathbb{E}(|T_n|) < \infty$, on a*

$$\mathrm{var}(T_n) \geq Dh(\theta)I_n(\theta)^{-1}[Dh(\theta)]'$$

où $Dh(\theta) = (\partial h(\theta)/\partial\theta_1, \ldots, \partial h(\theta)/\partial\theta_p)$

Preuve
Preuve dans le cas unidimensionnel

$$\mathrm{cov}\left[T_n, \frac{\partial}{\partial\theta}\log L(X,\theta)\right] = \mathbb{E}\left[T_n\frac{\partial}{\partial\theta}\log L(X,\theta)\right] - \mathbb{E}[T]\underbrace{\mathbb{E}\left[\frac{\partial}{\partial\theta}\log L(X,\theta)\right]}_{=0}$$
$$= \int t(x)\frac{\partial}{\partial\theta}\log L(x,\theta)L(x,\theta)d\nu(x)$$
$$= \int t(x)\frac{\partial}{\partial\theta}L(x,\theta)d\nu(x)$$
$$= \frac{\partial}{\partial\theta}\int t(x)L(x,\theta)d\nu(x)$$
$$= \frac{\partial}{\partial\theta}\mathbb{E}[T_n] = \frac{\partial}{\partial\theta}h(\theta) \qquad \text{car } T_n \text{ est sans biais}$$

De plus, par Cauchy-Schwarz (produit scalaire dans l'espace des fonctions de X de carré intégrable),

$$\mathrm{cov}\left[T_n, \frac{\partial}{\partial\theta}\log L(X,\theta)\right]^2 \leq \mathrm{var}(T_n)\,\mathrm{var}\left[\frac{\partial}{\partial\theta}\log L(X,\theta)\right]$$

soit

$$\left[\frac{\partial}{\partial\theta}h(\theta)\right]^2 \leq \mathrm{var}(T_n)I_n(\theta).$$

□

Définition 18. *La limite inférieure de la variance des estimateurs sans biais s'appelle* ***borne de Cramér-Rao***

On peut définir une inégalité de Cramér Rao pour un estimateur T_n biaisé pour estimer $h(\theta)$, de biais $b(\theta)$:

$$\mathrm{var}(T_n) \geq (Dh(\theta) + Db(\theta))I_n(\theta)^{-1}[Dh(\theta) + Db(\theta)]'$$

En particulier, pour estimer $h(\theta) = \theta$ la variance minimale d'un estimateur biaisé est nulle, et atteinte par un estimateur constant, qui est admissible mais peu intéressant.

2.4.3 Efficacité

Définition 19. *Un estimateur* ***sans biais*** T_n *est* ***efficace*** *pour estimer $h(\theta)$ s'il atteint la borne de Cramér-Rao, ie*

$$\mathrm{var}(T_n) = Dh(\theta)I_n(\theta)^{-1}[Dh(\theta)]'$$

et il est Uniformément de Variance Minimum parmi les estimateurs sans Biais (UVMB), donc optimal parmi les estimateurs sans biais.

Exemple Dans le modèle gaussien $\mathcal{N}(\mu, \sigma^2)$, σ^2 connu, la borne de Cramér-Rao pour l'estimation de μ vaut $1/I_n(\mu) = \sigma^2/n$. On ne peut donc trouver d'estimateur sans biais de variance inférieure à σ^2/n. Or, la variance de $\bar{X}$ atteint la borne de Cramér-Rao : on dit que l'estimateur $\bar{X}$ est efficace pour estimer μ dans ce modèle. Il est donc de variance minimale parmi les estimateur sans biais (Uniformément de Variance Minimale par les estimateurs sans Biais), et donc optimal.

Mais il faut relativiser la notion d'efficacité comme l'indique le théorème suivant (admis).

Théorème 4 (admis)**.** *Un estimateur sans biais T_n de $h(\theta)$, $\theta \in \Theta$, Θ intervalle de $\mathbb{R}$, atteint la borne de Cramer Rao ssi*

(a) la loi des observations appartient à une famille exponentielle (définition 20)

(b) il existe deux constantes A et B connues telles que $T_n = A\sum_i a(X_i)+B$ et $h(\theta) = \mathbb{E}_\theta(T_n)$

Ainsi, ceci restreint l'existence d'estimateurs efficaces essentiellement aux modèles exponentiels et pour estimer l'espérance de la statistique exhaustive uniquement.

Définition 20. *Un modèle dominé appartient à une* ***famille exponentielle*** *(à un paramètre) si sa densité peut s'écrire, pour tout x de son support, sous la forme*

$$f(x;\theta) = \exp\Big(a(x)\alpha(\theta) + \beta(\theta) + c(x)\Big)$$

Par exemple, la loi gaussienne à variance connue, la loi de Bernoulli, de Poisson ou la loi exponentielle dont partie de la famille exponentielle de lois.

Exemple La borne de Cramér-Rao pour l'estimation de la variance $h(\theta) = \theta(1-\theta)$ de la loi de Bernoulli $\mathcal{B}(\theta)$ vaut

$$BCR(\theta) = \left[\frac{\partial}{\partial\theta}\theta(1-\theta)\right]^2 / I_n(\theta) = \frac{(1-2\theta)^2\theta(1-\theta)}{n}$$

La loi de Bernoulli est de la famille exponentielle

$$f(x,\theta) = \exp\left(x\log\theta + (1-x)\log(1-\theta)\right) = \exp\left(x\log\frac{\theta}{1-\theta} + \log(1-\theta)\right)$$

où $a(x) = x$, $\alpha(\theta) = \log\frac{\theta}{1-\theta}$, $\beta(\theta) = \log(1-\theta)$, $c(x) = 0$. On sait qu'il n'existe pas d'estimateur non biaisé de variance égale à $BCR(\theta)$, puisque la seule fonction de θ qu'on sait estimer efficacement est $\mathbb{E}(a(X)) = \mathbb{E}(X) = \theta$.

Remarque Le théorème 4 s'étend aux familles exponentielles multidimensionnelles de la forme

$$f(x;\theta) = \exp\Big(\sum_j a_j(x)\alpha_j(\theta) + \beta(\theta) + c(x)\Big)$$

avec

$$h_j(\theta) = \mathbb{E}_\theta\left(\frac{\sum_i a_j(X_i)}{n}\right).$$

2.5 Estimation optimale

Nous avons vu dans les rappels que risque quadratique d'un estimateur se décompose en un terme de variance et un terme de biais :

$$R_\theta(\widehat{\theta}_n) = \mathrm{var}_\theta(\widehat{\theta}_n) + (b_\theta(\theta))^2$$

Il n'est en général pas possible de définir un estimateur uniformément meilleur en θ : comparer par exemple la moyenne empirique et l'estimateur constant égal à 0 par estimer l'espérance d'une loi gaussienne. On restreint donc l'étude à l'ensemble des estimateurs sans biais. Comment construire des estimateurs de variance minimale parmi les estimateurs sans biais (UVMB).

2.5.1 Estimateur sans biais de variance minimale

On commence par montrer l'unicité de l'estimateur UVMB quand il existe.

Théorème 5 (Unicité)**.** *S'il existe, un estimateur de θ sans biais et de variance minimum est unique (p.s.)*

Preuve
par l'absurde.
Supposons l'existence de deux estimateurs UVMB distincts T_1 et T_2. Soit $T_3 = \frac{T_1+T_2}{2}$. T_3 est non biaisé, et en appelant $\rho = \mathrm{cor}(T_1,T_2) = \mathrm{cov}(T_1,T_2)/\sqrt{\mathrm{var}\,T_1\,\mathrm{var}\,T_2}$ et $V = \mathrm{var}\,T_1 = \mathrm{var}\,T_2$

$$\mathrm{var}\,T_3 = \frac{1}{4}\left(\mathrm{var}\,T_1 + \mathrm{var}\,T_2 + 2\rho\sqrt{\mathrm{var}\,T_1\,\mathrm{var}\,T_2}\right) = \frac{1}{2}V(1+\rho)$$

Si $\rho < 1$, alors $\mathrm{var}(T_3) < T$, ce qui est impossible. Donc $\rho = 1$ et T_1 et T_2 sont liés linéairement et positivement. Donc il existe $\lambda > 0$ tel que $T_1 - \mathbb{E}T_1 = \lambda(T_2 - \mathbb{E}T_2)$ presque sûrement. D'où $\lambda = 1$ car $\mathrm{var}\,T_1 = \mathrm{var}\,T_2$. Et comme $\mathbb{E}T_1 = \mathbb{E}T_2$, on en déduit que $T_1 = T_2$ presque sûrement.

□

Puis on montre qu'on ne dégrade pas un estimateurs sans biais en le conditionnant par une statistique exhaustive

Théorème 6 (Rao-Blackwell)**.** *Soit T un estimateur sans biais de θ et S une statistique exhaustive pour θ. Alors*

$$\mathbb{E}(T|S)$$

est un estimateur sans biais de θ et qui domine T. Cette propriété s'étend à un estimateur T sans biais d'une fonction $h(\theta)$ de θ.

Remarque au passage sur la définition de $\mathbb{E}(T|S)$. Soit un couple (S,T) de loi de $f_{(T,S)}$ par rapport $\nu \otimes \nu'$. Pour chaque résultat s de S, la densité de T conditionnelle à $\{S = s\}$ est la fonction $f_T(.|S = s)$ égale à $f_{(T,S)}(., s)/f_S(s)$ pour $f_S(s) \neq 0$ et nulle sinon. $\mathbb{E}(T|S = s) = m(s)$ est l'espérance de la variable aléatoire T sous la loi conditionnelle à $\{S = s\}$. $\mathbb{E}(T|S)$ est la variable aléatoire $m(S)$ appelée espérance de T conditionnelle à S.

Preuve
$S = S(X)$ est une statistique exhaustive de $(X_1, \ldots, X_n)$ donc la loi conditionnelle $\{X|S(X) = s\}$ ne dépend pas de θ. L'espérance conditionnelle

$$\mathbb{E}(T|S = s) = \int t(x) f(x|S = s) d\nu(x)$$

ne dépend donc pas de θ, mais seulement de s. Donc $\mathbb{E}(T|S)$ est $S(X)$ mesurable et calculable à partir de la statistique exhaustive, c'est bien un estimateur de θ, qui est sans biais grâce au théorème de l'espérance totale

$$\mathbb{E}(T^*) = \mathbb{E}[\mathbb{E}(T|S)] = \mathbb{E}(T) = \theta$$

Le théorème de la variance totale permet d'affirmer que le conditionnement ne peut pas dégrader la variance

$$\mathrm{var}(T) = \mathrm{var}[\underbrace{\mathbb{E}(T|S)}_{T^*}] + \mathbb{E}[\underbrace{\mathrm{var}(T|S)}_{\geq 0}] \geq \mathrm{var}(T^*)$$

Si de plus $\mathbb{E}[\ \mathrm{var}(T|S)\ ] = 0$, alors $\mathrm{var}(T|S) = \mathbb{E}[(T - \mathbb{E}(T|S))^2|S] = 0$ et donc T est une fonction de S presque sûrement : l'estimateur UVMB sera donc fonction d'une statistique exhaustive.

□

Exemple Pour estimer l'espérance de la loi de Bernoulli $\mathcal{B}(1,\theta)$, on peut améliorer l'estimateur non biaisé $T = X_1$, première composante de l'échantillon, par $\bar{X}$ qui est une statistique exhaustive.

$$\mathbb{E}(X_1|\bar{X}) = \frac{1}{n}\sum_i \mathbb{E}(X_i|\bar{X}) = \mathbb{E}(\bar{X}|\bar{X}) = \bar{X}$$

Si une statistique exhaustive peut améliorer un estimateur sans biais non exhaustif par conditionnement, toute statistique exhaustive n'amène pas à obtenir un estimateur UVMB. Considérons $T = X_1$ la première composante de l'échantillon pour estimer l'espérance de la loi. C'est un estimateur non biaisé, et par ailleurs, l'échantillon tout entier est exhaustif. Or

$$\mathbb{E}(X_1|X_1, \ldots, X_n) = X_1$$

de variance supérieure à celle de $\bar{X}$. Le conditionnement n'a pas permis d'atteindre l'estimateur UVMB. C'est la notion de statistique complète qui permet de résoudre le point.

Définition 21. *On dit qu'une statistique exhaustive S est **complète** ou **totale** pour une famille de lois $f(x,\theta)$ si pour toute fonction mesurable h telle que $h(S(x)) \in \mathbb{L}_1(\mathbb{P}_\theta)$ pour tout θ, vérifiant*

$$\forall \theta \in \Theta, \ \ \mathbb{E}[h(S(x))] = 0$$

on a

$$\forall \theta \in \Theta, \ \ h(S(x)) = 0 \quad \mathbb{P}_\theta \quad p.s.$$

Par exemple, la statistique $S = \sum_i X_i$ est complète dans le modèle iid de Bernoulli $\mathcal{B}(1,\theta)$. En effet, soit $\bar{X} = S/n$ un estimateur sans biais de θ, $h(S)$ un autre estimateur sans biais de θ, où $h : \{0,\dots,n\} \to \mathbb{R}$ et f l'application définie sur $\{0,\dots,n\}$ par $f(s) = s/n - h(s)$. $\bar{X}$ et $h(S)$ étant sans biais, on a $\mathbb{E}[f(S)] = 0$, soit

$$\begin{aligned}
0 &= \sum_{k=0}^{n} f(k) \left(\begin{array}{c} n \\ k \end{array} \right) \theta^k (1-\theta)^{n-k} \\
&= \sum_{k=0}^{n} \theta^k f(k) \left(\begin{array}{c} n \\ k \end{array} \right) \left[\sum_{j=0}^{n-k} \left(\begin{array}{c} n-k \\ j \end{array} \right) (-\theta)^j \right] \\
&= \sum_{m=0}^{n} \theta^m \sum_{j,k:k+j=m} \left[\left(\begin{array}{c} n \\ k \end{array} \right) \left(\begin{array}{c} n-k \\ j \end{array} \right) (-1)^j f(k) \right]
\end{aligned}$$

Cette fonction polynomiale est nulle sur l'intervalle $]0;1[$, on a par identification que tous les coefficients sont nuls, ie pour tout $m = 1,\dots,n$,

$$\sum_{j,k:k+j=m} \left(\begin{array}{c} n-k \\ j \end{array} \right) (-1)^j f(k) = 0$$

et par récurrence que la fonction est nulle. On vérifie que $X = (X_1,\dots,X_n)$ n'est pas complète en considérant la fonction $h(X) = X_1 - X_2$.

Théorème 7 (Lehman-Scheffé)**.** *Si on dispose d'un estimateur sans biais de $h(\theta)$, et si $S(X)$ est une statistique exhaustive et complète, alors $T^* = \mathbb{E}(T|S)$ est l'unique estimateur UVMB de $h(\theta)$.*

Preuve
L'UVMB est unique et dépend de S statistique exhaustive et complète, il est de la forme $T_1 = f(S)$ ou $T_2 = g(S)$. Or il est sans biais, donc $\mathbb{E}(T_1) = \mathbb{E}(T_2) = 0$, donc $\mathbb{E}[(f-g)(S)] = 0$ donc $f - g = 0$ $\mathbb{P}_\theta-$ p.s. d'après la complétude.

□

Remarque On ne dispose pas toujours d'un estimateur sans biais. Par exemple dans le modèle de Bernoulli, $\sqrt{\theta}$ ne peut être estimé sans biais construit uniquement sur l'observation de $S = \sum_i X_i$. En effet, cela reviendrait à écrire $\sqrt{\theta}$ sous la forme d'un polynôme de θ sur l'intervalle $]0;1[$.

2.5.2 Résumé et les limites...

— Il n'y a en général pas d'estimateur uniformément meilleur.
— optimal veut dire UVMB. On sait construire l'unique estimateur optimal $\mathbb{E}(T|S)$ si on dispose d'une statistique sans biais T et d'une statistique exhaustive et complète S. Mais, il n'existe pas toujours d'estimateur sans biais, et donc pas toujours d'estimateur UVMB
— on peut déterminer une borne minimale aux estimateurs sans biais, c'est la borne de Cramér-Rao. Si la variance d'un estimateur atteint la borne de Cramér-Rao, il est efficace, et donc UVMB. Mais il n'existe pas toujours d'estimateur efficace, et un estimateur optimal (UVMB) peut ne pas être efficace. Seule une fonction particulière du paramètre d'une loi de la famille exponentielle peut être estimée efficacement
— Il peut exister des estimateurs de risque inférieur à celui d'un estimateur optimal UVMB si on accepte un léger biais, ce qui rend cet estimateur UVMB non admissible.
— Les estimateurs du maximum de vraisemblance sont asymptotiquement efficaces (et donc optimaux)

2.6 Asymptotique

Dans les sections précédentes, nous avons étudié les propriétés des estimateurs à distance finie, i.e., pour n fixé. L'étude asymptotique s'entend quand la taille de l'observation n tend vers l'infini. Par exemple, un estimateur est asymptotiquement sans biais si son biais tend vers 0. L'estimateur empirique de la variance $\sum_i (X_i - \bar{X})^2/n$ est de biais négatif $-\sigma^2/n$, mais il est asymptotiqment sans biais.

2.6.1 Consistance

La consistance d'un estimateur est la convergence de la variable aléatoire vers le paramètre à estimer (cf annexe B pour quelques définitions et propriétés de convergence du cours de probabilité).

Définition 22 (Convergences)**.** *Soit $\widehat{\nu}_n$ un estimateur de $\nu = \nu(\theta)$, défini à partir d'un n-échantillon de loi $\mathbb{P}_\theta$*

— *$\widehat{\nu}_n$ est (faiblement)* ***consistant*** *ou* ***convergent*** *ssi $\widehat{\nu}_n$ tend en probabilité vers ν quand $n \to \infty$:*

$$\forall \theta \in \Theta, \forall \varepsilon, \lim_{n\to\infty} \mathbb{P}_\theta(|\widehat{\nu}_n - \nu| > \varepsilon) = 0$$

On note : $\widehat{\nu}_n \xrightarrow{\mathcal{P}} \nu$.

— *$\widehat{\nu}_n$ est* ***fortement consistant*** *(convergence* ***presque sûre****) ssi $\forall \theta \in \Theta$,*

$$\mathbb{P}_\theta(\lim_{n\to\infty} |\widehat{\nu}_n - \nu| = 0\ ) = 1$$

On dit aussi $\widehat{\nu}_n \to \nu$ avec probabilité 1.

— *$\widehat{\nu}_n$ converge en* ***moyenne quadratique*** *vers ν ssi son risque tend vers 0 quand la taille de l'échantillon tend vers l'infini :*

$$\forall \theta \in \Theta, \lim_{n\to\infty} R_\theta(\widehat{\nu}_n, \nu) = 0.$$

On note : $\widehat{\nu}_n \xrightarrow{L^2} \nu$.

Ces définitions sont présentées sous une forme plus générale dans un cours de probabilité, où la limite de la suite d'estimateurs indexée par n peut être elle-même une variable aléatoire. Dans le cadre statistique, la limite recherchée ν est le paramètre à inférer, qui n'est pas aléatoire. En statistique, on regarde les phénomènes en moyenne, et c'est la consistance (faible) qui est en général utilisée.

La consistance presque sûre implique la consistance (en probabilité). La consistance en risque implique la consistance (en probabilité). Les réciproques sont fausses.

Il faut bien sûr privilégier les estimateurs consistants, puisqu'ils assurent qu'en probabilité ils infèrent la vraie valeur quand la taille de l'échantillon est infini. Ils infèrent donc une valeur (très) proche quand l'échantillon est (très) grand.

Le lemme de l'application continue (1 en annexe B) est utile pour étudier les estimateurs : ainsi, si $\widehat{\theta}(X)$ est un estimateur (fortement) consistant de θ, et ν une fonction continue de θ, alors $\nu(\widehat{\theta})$ est un estimateur (fortement) consistant de $\nu(\theta)$.

Les théorèmes suivant sont bien utiles pour montrer la consistance

Théorème 8 (Loi faible des grands nombres)**.** *Si $(X_1, \ldots, X_n)$ est un échantillon i.i.d d'une loi $\mathbb{P}$ de carré intégrable, de variance finie σ^2, sa moyenne empirique $\bar{X}$ converge en probabilité (et même presque sûrement), la suite $(\bar{X})$ est faiblement (fortement) consistante vers $\mathbb{E}(X)$.*

Théorème 9 (Loi faible des grands nombres de Kolmogorov)**.** *Si $(X_1, \ldots, X_n)$ est un échantillon i.i.d d'une loi $\mathbb{P}$ telle $\mathbb{E}(|X|)$ est finie, alors $\bar{X}$ converge en probabilité (et même presque sûrement), la suite $(\bar{X})$ est faiblement (fortement) consistante vers $\mathbb{E}(X)$.*

Exemple L'estimateur empirique $\bar{X}$ de l'espérance est sans biais et fortement consistant. C'est l'application directe de la loi des grands nombres. Par ailleurs, c'est un estimateur sans biais de $\mathbb{E}(X)$ et de variance

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

qui tend asymptotiquement vers 0, donc son risque tend vers 0.

2.6.2 Loi asymptotique

Le risque donne une information globale sur la vitesse de convergence de l'estimateur, en $\sqrt{1/n}$ pour l'estimateur de l'espérance. Ce n'est qu'une information partielle de la loi de l'estimateur, dont on a besoin pour construire des intervalles de confiance et des tests. Si la loi de l'estimateur $\bar{X}$ est connue dans le gaussien (puisque son espérance et sa variance le sont), il n'en est pas de même pour l'estimateur $\bar{X}$ de l'espérance d'une loi donnée.

Dans ce cas, le théorème essentiel suivant donne le comportement asymptotique de $\bar{X}$:

Théorème 10 (central limite)**.** *Si $\bar{X}$ est la moyenne empirique d'un n-échantillon d'une loi d'espérance μ et de variance finie σ^2, la suite $\sqrt{n}(\bar{X} - \mu)$ converge en loi vers $\mathcal{N}(0, \sigma^2)$.*

De façon générale, si
— la vitesse de l'estimateur est en $\sqrt{n}$,
— la convergence a lieu en loi,
— la loi limite est gaussienne,
on dit que l'estimateur est asymptotiquement normal. Un estimateur est d'autant meilleur que sa vitesse de convergence est rapide et sa loi limite concentrée autour de 0.

2.6.3 Delta-méthode

Dans le cas où on souhaite estimer $h(\theta)$, on peut utiliser la delta-méthode.

Propriété 3. *Soit h une fonction de $\mathbb{R}^p$ dans $\mathbb{R}^q$, U_n une suite de variables aléatoires de $\mathbb{R}^p$ et a_n, $n \in \mathbb{N}$, une suite déterministe de réels. On suppose que*

- *$a_n \to +\infty$*
- *Il existe un vecteur déterministe $U \in \mathbb{R}^p$, et un vecteur aléatoire L tels que $a_n(U_n - U) \xrightarrow{\mathcal{L}} L$*
- *h est une fonction différentiable en U, de différentielle notée $Dh(U)$, matrice de taille $p \times q$.*

On a alors la convergence en loi suivante :

$$a_n \left(h(U_n) - h(U)\right) \xrightarrow{\mathcal{L}} Dh(U)\ L$$

En particulier, si l'estimateur U_n de U est asymptotiquement normal $L \sim \mathcal{N}(0, V_1(\theta))$, on déduit le comportement de $h(U_n)$ en fonction de celui de U_n

$$\sqrt{n}(h(\widehat{\theta}_n) - h(\theta)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, Dh(\theta)V_1(\theta)[Dh(\theta)]')$$

où la notion prime $'$ indique la transposée.

Preuve
En écrivant

$$a_n \left(h(U_n) - h(U)\right) = a_n \left(h(U_n) - h(U)\right) - Dh(U)\ [a_n(U_n - U)] + Dh(U)\ [a_n(U_n - U)]$$

on voit d'après le lemme de Slusky [1] qu'il suffit de montrer que

$$A_n = a_n \left(h(U_n) - h(U)\right) - Dh(U)\ [a_n(U_n - U)] \xrightarrow{\mathcal{P}} 0.$$

Pour tout $\varepsilon > 0$, montrons la convergence de $\mathbb{P}(||A_n|| \geq \varepsilon)$ vers 0. Fixons ε. Puisque h est différentiable, pour tout $k > 0$, il existe $\delta > 0$ tel que pour tout $V \in \mathbb{R}^p$

$$||V - U|| \leq \delta \Rightarrow a_n||h(V) - h(U) - Dh(U)\ [V - U]|| \leq a_n k||V - U|| \leq a_n \delta$$

Posons $k = \varepsilon/m$. On a

$$\begin{aligned}
\mathbb{P}\Big(||A_n|| \geq \varepsilon\Big) &\leq \mathbb{P}\Big(\{||A_n|| \geq \varepsilon\} \cap \{||U_n - U|| \leq \delta\}\Big) + \mathbb{P}\Big(||U_n - U|| > \delta\Big) \\
&\leq \mathbb{P}\Big(a_n \frac{\varepsilon}{m}||U_n - U|| \geq A_n \geq \varepsilon\Big) + \mathbb{P}\Big(a_n||U_n - U|| > a_n\delta\Big) \\
&\leq 2\ \mathbb{P}\Big(a_n||U_n - U|| \geq \min(m, a_n\delta)\Big)
\end{aligned}$$

Puisque $a_n \to +\infty$, on a pour tout m :

$$\limsup_{n\to\infty} \mathbb{P}(||A_n|| \geq \varepsilon) \leq \limsup_{n\to\infty} 2\mathbb{P}\Big(a_n||U_n - U|| \geq m\Big) \leq 2\mathbb{P}\Big(||L|| \geq m\Big)$$

où on a utilisé le lemme de Portmanteau [2] associé à la convergence en loi de $a_n(U_n - U)$. Comme ceci est valable pour tout $m > 0$, on en déduit $\mathbb{P}(||A_n|| \geq \varepsilon) \to 0$, ce qu'il était suffisant de démonter.

□

1. Si $X_n \xrightarrow{\mathcal{L}} X$ et $Y_n \xrightarrow{\mathcal{L}} c$ où c est une constance, alors $(X_n, Y_n) \xrightarrow{\mathcal{L}} (X, c)$, d'où $X_n + Y_n \xrightarrow{\mathcal{L}} X + c$, cf annexe B

2. Pour tout ensemble fermé F, $\limsup \mathbb{P}(X_n \in F) \leq \mathbb{P}(X \in F)$ est équivalent à la convergence en loi de X_n vers X

Exemple : Asymptotique de l'estimateur de la variance d'une loi de Bernoulli. Soit $h(\theta) = \theta(1-\theta)$, et $\widehat{\theta} = \bar{X}$. On a $Dh(\theta)I_1^{-1}(\theta)Dh(\theta)' = \theta(1-\theta)(1-2\theta)^2$, soit

$$\sqrt{n}(h(\widehat{\theta}) - h(\theta)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \theta(1-\theta)(1-2\theta)^2)$$

Chapitre 3

Estimation par Maximum de Vraisemblance

Nous avons vu en première année que l'estimateur des moments est facilement construit pour des paramètres qui s'expriment en fonction des moments de la loi mère, en remplaçant l'expression du moment par celle du moment empirique correspondant. Si les hypothèses de la loi forte des grands nombres sont vérifiées, et si le paramètre est une fonction continue des moments, l'estimateur des moments est fortement consistant. Mais il peut ne pas être admissible, avec une vitesse de convergence très lente par rapport à celle d'autres estimateurs. Par ailleurs, il n'y a pas unicité de sa définition : pour estimer par la méthode des moments la variance d'une loi de Bernoulli, on peut utiliser le moment d'ordre un ou le moment d'ordre deux.

Nous étudions dans ce chapitre une autre classe d'estimateurs, l'estimateur du maximum de vraisemblance (EMV), facile à mettre en œuvre et possédant de bonnes propriétés, ce qui en fait un estimateur de choix en modélisation statistique

Ce chapitre utilise la notion de vraisemblance (définie dans le chapitre introductif) et d'information de Fisher (définie au chapitre 2).

3.1 Méthode du maximum de vraisemblance

Elle s'applique dans le cas où toutes les lois du modèle sont dominées par une mesure commune (mesure de Lebesgue, de comptage, ...). Soit f_θ la densité de la loi $\mathbb{P}_\theta$ par rapport à cette mesure commune. La méthode du maximum de vraisemblance consiste à estimer θ comme la valeur $\widehat{\theta}$ maximisant la vraisemblance des observations, ie $\prod_{i=1,\dots,n} f_\theta(X_i)$ quand les observations sont indépendantes.

3.2 Estimateur du maximum de vraisemblance

La vraisemblance en un θ donné peut être vue comme la probabilité de l'observation quand le modèle est défini avec ce θ. L'idée de l'estimateur du maximum de vraisemblance est de proposer un $\widehat{\theta}$ qui rend la vraisemblance la plus grande possible, ie, qui rend la probabilité de l'observation la plus grande possible quand on la calcule avec ce $\widehat{\theta}$.

Définition 23 (EMV)**.** *On appelle* ***estimation*** *du maximum de vraisemblance, une valeur* $\widehat{\theta}_n$

maximisant la vraisemblance

$$\widehat{\theta}_n \in \arg\max_{\theta\in\Theta} L(\theta; x).$$

$\widehat{\theta}_n = t(x_1, \ldots, x_n)$ *est une fonction des données, ce qui induit la statistique* $t(X_1, \ldots, X_n)$ *que l'on note (abusivement) avec la même notation :* $\widehat{\theta}_n = t(X_1, \ldots, X_n)$ *est appelé* ***Estimateur du Maximum de Vraisemblance***

De même, on ne fait pas de différence de notation entre $L(\theta; X)$, vraisemblance de l'échantillon $X = (X_1, \ldots, X_n)$, vue comme une fonction de variables aléatoires et $L(\theta; x)$ vraisemblance définie comme fonction déterministe pour un résultat d'observation. Quand l'échantillon est i.i.d. on utilise plutôt la log-vraisemblance, qui transforme le produit en somme :

$$\ell_n(\theta; x) = \log L(\theta; x) = \sum_{i=1}^{n} \log f_\theta(x_i)$$

Lorsque la log-vraisemblance est régulière (dérivable), on cherche les valeurs $\widehat{\theta}_n$ de θ annulant les équations de vraisemblance (équations du **score**)

$$U_n(\widehat{\theta}_n) := \dot{\ell}_n(\widehat{\theta}_n; x) = \left(\frac{\partial}{\partial\theta_k}\ell_n(\widehat{\theta}_n; x)\right)_{k=1,\ldots,dim(\theta)} = 0,$$

Puis on vérifie que $\widehat{\theta}_n$ est bien un maximum : $H_n(\theta) = \ddot{\ell}_n(\theta; x)$ est définie négative autour de $\widehat{\theta}_n$.

Exemples L'estimateur du maximum de vraisemblance de l'espérance θ d'une loi de Bernoulli est $\widehat{\theta} = \bar{X}$, obtenu en résolvant l'équation du score :

$$\frac{\sum_i X_i}{\theta} - \frac{n - \sum_i X_i}{1-\theta} = 0$$

Le hessien est aléatoire et négatif :

$$\ddot{\ell}_n(\theta; x) = -\frac{\sum_i X_i}{\theta^2} - \frac{n - \sum_i X_i}{(1-\theta)^2} < 0$$

d'espérance l'opposé de l'information de Fisher $\mathbb{E}[\ddot{\ell}_n(\theta; x)] = -n/(\theta(1-\theta))$.
L'estimateur du maximum de vraisemblance de $\theta = (\mu, \sigma^2)$ dans le modèle i.i.d. gaussien est

$$\widehat{\theta} = (\widehat{\mu}, \widehat{\sigma}^2);\ \ \widehat{\mu} = \bar{X};\ \ \widehat{\sigma}^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}$$

Il n'est pas toujours possible de calculer explicitement l'EMV. Dans ce cas, on déploie un **schéma numérique**, par exemple celui de Newton-Raphson, dont le principe est la linéarisation de U_n, fonction dont on cherche un zéro, au point courant $(\theta^{(m)}, U_n(\theta^{(m)}))$

$$0 = U_n(\theta^{(m)}) + H_n(\theta^{(m)})(\theta^{(m+1)} - \theta^{(m)})$$

Si $H_n(\theta^{(m)})$ est inversible, l'algorithme itère jusqu'à convergence l'étape suivante :

$$\theta^{(m+1)} = \theta^{(m)} - [H_n(\theta^{(m)})]^{-1} U_n(\theta^{(m)})$$

La recherche d'EMV présente certains écueils :

— la vraisemblance n'est pas forcément strictement concave pour tout θ, il peut exister des optima locaux : l'initialisation d'un algorithme de Newton doit être faite précautionneusement pour les éviter.
— l'EMV n'est pas forcément unique
— l'EMV peut ne pas exister
— La méthodologie doit être différente quand le modèle n'est pas régulier, par exemple dans le modèle uniforme : $\mathcal{U}(0, \theta)$

mais, l'EMV présente l'avantage de ne pas être affecté par un changement de mesure dominante, et il a de bonnes propriétés asymptotiques qui en font un estimateur de choix.

3.3 Propriétés

Pour estimer un paramètre, il est nécessaire que le modèle soit identifiable, c'est à dire que la fonction $\theta \mapsto f(., \theta)$ soit injective : deux paramètres différents définissent deux modèles différents.

Le paramètre d'une loi de Bernoulli est identifiable, l'espérance et la variance d'une loi gaussienne aussi. En revanche, si on considère un échantillon iid $Y_i = a(1 + \varepsilon_i)$ où $a \geq 0$ et les ε_i sont des variables gaussiennes centrées de variance σ^2 inconnue, le modèle n'est pas identifiable en σ^2 si $a = 0$.

3.3.1 Propriétés à distance finie

On suppose qu'il existe un unique EMV $\widehat{\theta}_n$ de θ.

Théorème 11. *Si T_n est une statistique exhaustive pour θ, alors $\widehat{\theta}_n$ est une fonction de T_n*

Preuve
D'après le critère de factorisation (théorème 1), on peut trouver deux fonctions h et g telles que

$$f(x = (x_1, \ldots, x_n); \theta) = h(x)g(t(x); \theta)$$

donc,

$$\max_{\theta \in \Theta} f(x; \theta) = \max_{\theta \in \Theta} g(t(x); \theta)$$

d'où l'EMV $\widehat{\theta}_n$ est tel que

$$h(T_n, \widehat{\theta}_n) \geq h(T_n, \theta)$$

et s'écrit donc sous la forme $\widehat{\theta}_n = \widehat{\theta}_n(T_n)$.

□

Remarque Mais $\widehat{\theta}_n$ lui-même n'est pas forcément exhaustif.

Théorème 12 (Invariance par reparamétrisation)**.** *Pour toute application h de Θ, $h(\widehat{\theta}_n)$ est l'EMV de $h(\theta)$.*

Preuve
Soit $\nu = h(\theta) \in h(\Theta)$, on a $\Theta = \cup_{\nu \in h(\Theta)} \{\theta | h(\theta) = \nu\}$. On définit la vraisemblance de ν par

$$L(\nu) = \sup_{\theta | h(\theta) = \nu} f(x; \theta).$$

D'où

$$\sup_{\nu \in h(\Theta)} L(\nu) = \sup_{\theta \in \Theta} f(x; \theta) = f(x; \widehat{\theta}_n)$$

Or, $\widehat{\theta}_n \in \{\theta | h(\theta) = h(\widehat{\theta}_n)\}$ et il réalise le sup de $f(x; .)$ donc

$$f(x; \widehat{\theta}_n) = \sup_{\theta | h(\theta) = h(\widehat{\theta}_n)} f(x; \theta) = L(h(\widehat{\theta}_n))$$

où la deuxième égalité provient de la définition de la vraisemblance de $\nu = h(\theta)$.

□

3.3.2 Consistance l'EMV

θ servant de variable générique, on appelle $\theta^\star$ le "vrai" paramètre, celui de la loi qui a généré les données.

Propriété 4. *Dans le modèle dominé $(\mathbb{P}_\theta)_{\theta \in \Theta}$, soit $\widehat{\theta}_n$ l'EMV obtenu à partir d'un n-échantillon i.i.d. $X_i \sim f_{\theta^\star}$. On suppose*
- *H_1 : le modèle est **identifiable***
- *H_2 : Θ est compact et pour tout $x \in \mathcal{X}$, $\theta \to f_\theta(x)$ **continue***
- *H_3 : soit $h(x) = \sup_{s \in \Theta} |\log f_s(x)|$. Pour tout $\theta \in \Theta$, $h \in L_1(\mathbb{P}_\theta)$*

alors $\widehat{\theta}_n$ est consistant.

Remarque : Le théorème indique que $\widehat{\theta}_n$ est **asymptotiquement sans biais** ; mais il peut être biaisé à distance finie. C'est par exemple le cas de l'EMV de la variance d'un loi gaussienne à espérance et variance inconnues. Nous avons vu qu'il est dans ce cas la variance empirique, estimateur biaisé de la variance.

La preuve de ce théorème est admise, mais on donne ici les idées qui la guident :
- D'après H_3, $\log f_\theta(X_1)$ est intégrable, et on peut utiliser la loi (forte) des grands nombres pour montrer que la log-vraisemblance correctement renormalisée est convergente :

$$\frac{1}{n} \log L(\theta; X) = \frac{1}{n} \sum_{i=1}^{n} \log f_\theta(X_i) \overset{LGN}{\longrightarrow} \mathbb{E}_{\theta^\star}(\log(f_\theta(X_1)))$$

- Comme le modèle est identifiable d'après H_1, $\theta^\star$ est l'unique minimum de la divergence de Kullback $K(\theta^\star, \theta) = \mathbb{E}_{\theta^\star}(\log(f_{\theta^\star}(X_1))) - \mathbb{E}_{\theta^\star}(\log(f_\theta(X_1)))$ et est donc l'unique maximum de

$$\theta \to K(\theta) = \mathbb{E}_{\theta^\star}(\log(f_\theta(X_1)))$$

- d'après H_2 et H_3, $\theta \to f_\theta(x)$ continue, Θ compact et $h \in L_1$. La convergence de la log-vraisemblance est uniforme : elle "respecte" donc la forme de la limite. Ainsi, l'estimateur $\widehat{\theta}$ maximisant $L(\theta, X)$ converge vers l'unique valeur $\theta^\star$ maximisant la fonction limite $K(\theta)$.

3.3.3 Normalité asymptotique de l'EMV

La consistance étant acquise, le théorème 13 pose le comportement de la loi limite.

Théorème 13. *Soit $\widehat{\theta}_n$ l'EMV du paramètre $\theta^\star$ calculé sur un n-échantillon i.i.d. $X_i \sim f_{\theta^\star}$. Si $\widehat{\theta}_n$ est **consistant**, si le modèle est **régulier** et $I_1(\theta^\star)$ est **inversible**, alors pour tout $\theta^\star \in \Theta^\circ$, $\sqrt{n}(\widehat{\theta}_n - \theta^\star)$ converge en loi sous $\mathbb{P}_{\theta^\star}$ vers une loi normale :*

$$\sqrt{n}(\widehat{\theta}_n - \theta^\star) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I_1(\theta^\star)^{-1})$$

où $I_1(\theta^\star) = -\mathbb{E}_{\theta^\star}[\frac{\partial^2}{\partial\theta^2} \log f_{\theta^\star}(X_1)]$

Les grandes lignes de la preuve suivent le cheminement suivant :

— Montrer que le vecteur du **score** $U_n(\theta^\star) = \nabla \ell_n(\theta^\star, X)$ correctement renormalisé est asymptotiquement gaussien :

$$\frac{1}{n} U_n(\theta^\star) = \frac{1}{n} \sum_{i=1}^{n} \nabla \log f_{\theta^\star}(X_i) \xrightarrow{LGN} \mathbb{E}_{\theta^\star}[U_1(\theta^\star)] = 0$$

$$\underset{\theta^\star}{\text{var}}[\frac{\sqrt{n}}{n} U_n(\theta^\star)] = I_1(\theta^\star); \quad \frac{\sqrt{n}}{n} U_n(\theta^\star) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I_1(\theta^\star))$$

— Montrer que la matrice **hessienne** $H_n(\theta^\star) = \nabla^2 \ell_n(\theta^\star; x)$ renormalisée en n converge vers l'opposé de l'information de Fisher d'une composante

$$\frac{1}{n} H_n(\theta^\star) \xrightarrow{LGN} -I_1(\theta^\star)$$

— Effectuer un développement de Taylor-Lagrange

$$\underbrace{\sqrt{n}\frac{U_n(\theta^\star)}{n}}_{\xrightarrow{\mathcal{L}} \mathcal{N}(0, I_1(\theta^\star))} = \sqrt{n} \underbrace{\frac{U_n(\widehat{\theta}_n)}{n}}_{=0} + \underbrace{\frac{H_n(\overline{\theta}_n)}{n}}_{\to -I_1(\theta^\star)} \sqrt{n}(\theta^\star - \widehat{\theta}_n),$$

Le modèle étant iid, toutes les composantes apportent la même information. La matrice limite $I_1(\theta^\star)$ est l'information de Fisher d'une composante de l'échantillon. C'est une matrice symétrique définie positive, on peut donc définir la matrice "racine carrée" $I_1(\theta^\star)^{1/2}$ pour réécrire le résultat sous la forme

$$\sqrt{n} I_1(\theta^\star)^{1/2}(\widehat{\theta}_n - \theta^\star) = I_n(\theta^\star)^{1/2}(\widehat{\theta}_n - \theta^\star) \xrightarrow{\mathcal{L}} \mathcal{N}(0, Id_p)$$

en utilisant l'information de Fisher de l'échantillon en entier $I_n(\theta^\star) = nI_1(\theta^\star)$. Son inverse $I_n(\theta^\star)^{-1}$ est un terme qui peut être interprété comme une approximation de la variance de l'EMV à distance finie. Comme $I_n(\theta^\star)$ dépend de $\theta^\star$ inconnu, si on sait lui substituer un estimateur consistant, le théorème de Slutsky permet d'affirmer :

Propriété 5. *Soit $\widehat{\theta}_n$ l'emv de $\theta^\star$ et soit $\widehat{V}_n$ tel que $\widehat{V}_n I_n(\theta^\star) \xrightarrow{Pr} Id_p$. Alors,*

$$\widehat{V}_n^{-1/2}(\widehat{\theta}_n - \theta^\star) \xrightarrow{\mathcal{L}} \mathcal{N}(0, Id_p)$$

Ces résultats donnent des propriétés intéressantes pour les "grands" échantillons (par ex $n \geq 30$)

— $\widehat{\theta}_n$ tend à devenir gaussien et on peut prendre pour loi approchée à distance finie la loi asymptotique

$$\widehat{\theta}_n \overset{appr}{\sim} \mathcal{N}(\theta^\star, I_n^{-1}(\theta^\star)); \;\; \mathrm{var}(\widehat{\theta}_n) \simeq I_n^{-1}(\theta^\star) = \frac{1}{n} I_1^{-1}(\theta^\star)$$

et plus généralement (Slutsky), si $\widehat{V}_n I_n(\theta^\star) \xrightarrow{Pr} Id_p$,

$$\widehat{V}_n^{-1/2}(\widehat{\theta}_n - \theta^\star) \overset{appr}{\sim} \mathcal{N}(0, Id_p)$$

— L'EMV $\widehat{\theta}_n$ dans les cas réguliers est **asymptotiquement efficace** (et donc asymptotiquement UMVB) : la dispersion asymptotique est donc minimale dans la famille des estimateurs sans biais.

Delta-méthode L'utilisation de la delta-méthode (voir section 2.6.3) permet l'extension de l'asymptotique de l'EMV à celle d'une fonction de l'EMV. Dans le cas de l'EMV d'un modèle régulier, on applique le théorème avec $U_n = \widehat{\theta}_n$, $U = \theta$, $a_n = \sqrt{n}$ et $L \sim \mathcal{N}(0, V_1(\theta))$ avec $V_1(\theta) = I_1^{-1}$. On déduit du comportement asymptotiquement normal de l'EMV celui d'une fonction différentiable de l'EMV :

$$\sqrt{n}(h(\widehat{\theta}_n) - h(\theta)) \xrightarrow{\mathcal{L}} \mathcal{N}(0, Dh(\theta) V_1(\theta) Dh(\theta)')$$

3.3.4 Statistique de Wald

La statistique de Wald est une forme pivotale de $A\widehat{\theta}$, où A est une matrice de taille compatible avec celle de θ. Elle permet de déterminer la loi asymptotique (et donc approchée à distance finie) de combinaisons linéaires du paramètre.

Théorème 14. *Soit une famille paramétrique $\mathbb{P}_\theta, \theta \in \Theta$ et A est une matrice de dimension $q \times p$ de rang r. Si $\widehat{\theta}_n$ est un estimateur asymptotiquement normal de θ (par ex, l'EMV) tel que*

$$\widehat{V}_n^{-1/2}(\widehat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, Id_p)$$

alors

$$W = [A(\widehat{\theta}_n - \theta)]'(A\widehat{V}_n A')^{-1} A(\widehat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \chi^2(r)$$

W est la statistique de **Wald**.

Chapitre 4

Tests

Nous commençons par rappeler la méthodologie de construction d'un test et la construction de test optimaux pour des hypothèses simples, et discutons le cas de variables de loi discrète. Puis nous étendons l'étude au cas d'hypothèses composites et introduisons les familles à vraisemblance monotone. Enfin, nous présentons deux tests classiques et souvent utilisés dans le cadre de l'estimation par maximum de vraisemblance : le test de Wald et le test de rapport de vraisemblance.

4.1 Construction d'un test

Construire un test statistique, c'est définir une **règle de décision** permettant de choisir entre une hypothèse, dite **nulle** H_0, et son **alternative** H_1 en utilisant les résultats d'un échantillon. Il s'agit de répondre, par exemple, aux questions suivantes : La pièce est-elle truquée ? L'utilisation de sels d'argent augmente-t-il significativement le volume des précipitations ? Le niveau de corrosion a-t-il une influence sur la durée de vie d'une pièce mécanique ? Rappelons brièvement la méthodologie de construction d'un test :

1. Définir le **modèle** statistique des données $(\Omega, \mathcal{A}, P_\theta, \theta \in \Theta)$.
2. Définir les **hypothèses** en compétition qui traduisent la question à laquelle il faut répondre :
   — l'hypothèse **hypothèse nulle** H_0, en général sous la forme $\theta \in \Theta_0 \subset \Theta$ dans le cadre paramétrique
   — l'hypothèse **alternative** H_1, en général sous la forme $\theta \in \Theta_1 \subset \Theta$, disjoint de Θ_0.
   Si Θ_0 (ou Θ_1) est réduit à un singleton, l'hypothèse est dite simple, sinon, elle est composite.
3. Construire une statistique de test $T(X)$, et calculer sa loi sous H_0. Cette statistique doit avoir une loi différente sous H_1.
4. Définir la **règle de décision** permettant de choisir entre H_0 et H_1. C'est une fonction δ de $T(X)$ qui vaut 1 sur un ensemble $\mathcal{R}_\alpha$ appelé *région critique* ou *région de rejet* de l'hypothèse nulle (choix de H_1) et 0 sur son complémentaire (choix de H_0) :

$$\delta(T) = \mathbb{1}\{T(X) \in \mathcal{R}_\alpha\}.$$

   On appelle parfois région de rejet l'événement $\{T \in \mathcal{R}_\alpha\}$ lui-même. La région de rejet est calibrée par le choix a priori du **niveau** α, permettant de maîtriser le l'erreur commise

en rejetant H_0 à tort :

$$\alpha \geq \sup_{\theta \in \Theta_0} P_\theta(\{T \in \mathcal{R}_\alpha\}).$$

5. Calculer la valeur observée t de la statistique de test T sur l'échantillon et prendre la décision.

Exemple test de Student de la moyenne d'un échantillon gaussien à variance inconnue. On teste $\mu = \mu_0$ contre $\mu = \mu_1 > \mu_0$; la statistique de test est $T = \sqrt{n}(\bar{X} - \mu_0)/\widehat{\sigma}$ qui suit une loi $\mathcal{T}(n-1)$ sous H_0. La région de rejet est $\mathcal{R}_\alpha = [q_{1-\alpha}^{\mathcal{T}(n-1)}; +\infty[$, où $q_{1-\alpha}^{\mathcal{T}(n-1)}$ est le quantile d'ordre α de la loi de Student à $n-1$ degrés de liberté ; on a $P_{H_0}(\{T \in \mathcal{R}_\alpha\}) = \alpha$.

4.1.1 Risques d'un test

Soit $\mathcal{R}$ la région de rejet d'un test de statistique T.

Définition 24. *On appelle*

— ***erreur ou risque de première espèce*** *la probabilité de rejeter H_0 alors qu'elle est vraie :*

$$\theta_0 \in \Theta_0 \mapsto \alpha(\theta_0) = P_{\theta_0}(\{T \in \mathcal{R}\}).$$

— ***erreur ou risque de seconde espèce*** *la probabilité de conserver H_0 alors qu'elle est fausse :*

$$\theta_1 \in \Theta_1 \mapsto \beta(\theta_1) = P_{\theta_1}(\{T \notin \mathcal{R}\}).$$

— ***puissance*** *la probabilité de refuser H_0 quand elle est fausse :*

$$\theta_1 \in \Theta_1 \mapsto \pi(\theta_1) = P_{\theta_1}(\{T \in \mathcal{R}\}) = 1 - \beta(\theta_1).$$

A l'issue du test, les quatre situations suivantes sont possibles

	Choix H_0	Choix H_1
H_0 vraie	$1-\alpha$ bonne décision	α : erreur première espèce mauvaise décision
H_1 vraie	β : erreur deuxième espèce mauvaise décision	$\pi = 1-\beta$: puissance bonne décision

La décision du test, à partir de la valeur observée t de la statistique de test est :

— si $t \in \mathcal{R}_\alpha$, on rejette H_0 au niveau (ou au risque) α : les observations sont significativement différentes de celles attendues sous H_0. L'erreur commise en choisissant H_1 à tort est α, c'est ainsi que le test a été construit.

— si $t \notin \mathcal{R}_\alpha$, on conserve H_0 dans le test de niveau α : les données ne sont pas significatives pour choisir H_1. C'est un cas un peu moins confortable, car l'erreur de seconde espèce β commise en prenant cette décision (conserver H_0 à tort), n'est en général pas connue et peut être assez grande.

Dans le cadre de la théorie de Neyman-Pearson, l'erreur de première espèce α est calibrée $(5\%, 1\%, ...)$. L'objectif du test est le rejet de H_0, puisque le risque α de cette décision est contrôlé. On note ainsi la dissymétrie des hypothèses H_1 et H_0 : le contrôle est fait sur le risque de première espèce α, mais pas sur celui de seconde espèce β. Parmi tous les tests de niveau α, on cherchera bien sûr celui qui permet d'avoir la plus grande puissance, c'est à dire l'erreur de seconde espèce la plus faible.

Exemple pour tester la nocivité d'un nouveau médicament, il est préférable de choisir *dangereux* pour H_0, et *inoffensif* pour H_1, afin de ne pas choisir l'hypothèse d'inoffensivité sans en calibrer le risque. Peut-être qu'on décidera dangereux avec une erreur inconnue, et donc qu'on abandonnera un médicament potentiellement efficace et sans danger, mais c'est préférable à le déclarer inoffensif sans connaître l'erreur commise dans cette décision (à quel risque il pourrait être dangereux alors qu'il a étiqueté inoffensif).

Pou tester l'efficacité d'un médicament, il est préférable de choisir *inefficace* pour H_0, et *efficace* pour H_1, afin de limiter la mise sur le marché de médicaments inefficaces.

4.1.2 P-Value

La **p-value** définit le niveau critique à partir duquel on rejetterait le test, étant donnée l'observation qu'on vient de faire et qui est prise pour seuil :

$$\text{P-value}(\omega) = P_Z\{Z \in \mathcal{R}(T(\omega))\},$$

où le seuil dans $\mathcal{R}$ est remplacé par la valeur de la statistique observée sur les données $t = T(\omega)$ et Z est une variable aléatoire de même loi que T. C'est une **variable aléatoire** de loi uniforme dans le cas d'un statistique de test univariée. Dans un test de niveau α, H_0 est rejetée si $\alpha >$ p-value, conservée si $\alpha <$ p-value :

— si $0.05 >$ p-value > 0.01, le test est significatif,
— si $0.01 >$ p-value > 0.001, le test est très significatif,
— si $0.001 >$ p-value, le test est hautement significatif.

Remarque Il ne faut pas confondre le niveau α d'un test qui est un scalaire avec la p-value qui est une variable aléatoire : son résultat dépend de l'échantillon observé.

4.1.3 Propriétés d'un test

— Le test est sans biais si sa puissance est supérieure à son niveau ($\pi > \alpha$) : la probabilité de choisir H_1 à raison est supérieure à celle de choisir H_1 à tort.
— Le test est convergent si sa puissance tend vers 1 : on détecte toujours H_1.

4.2 Tests entre deux hypothèses simples

Si on considère le test de l'espérance μ d'un modèle gaussien à variance σ^2 connue d'hypothèses simples $\mu = \mu_0$ contre $\mu = \mu_1 > \mu_0$ (μ_0 et μ_1 sont donnés et connus), il semble raisonnable de travailler avec un estimateur de μ, par exemple $\bar{X}$ qui est efficace dans ce modèle. Sa loi sous (H_0) est déterminée, c'est $\mathcal{N}(\mu_0, \sigma^2/n)$. Mais elle dépend explicitement de μ_0. On considère alors la variable $T_n = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$ qui suit sous (H_0) une gaussienne centrée réduite, loi indépendante de μ et σ : on dit que T_n est une statistique libre. Si les données sont générées sous (H_0), on a

$$\mathbb{P}_{\mu_0}(T > q^*_{1-\alpha}) = \alpha$$

où $q^*_{1-\alpha}$ est le quantile d'ordre $1-\alpha$ d'une loi gaussienne centrée réduite, d'où une région de rejet de niveau α pour T_n est $\mathcal{R} = \{t; t > q^*_{1-\alpha}\}$

On aurait pu choisir une autre région de rejet, par exemple $\tilde{\mathcal{R}} = \{t; t < q^*_\alpha\}$, de même niveau. Mais $\mathcal{R}$ est préférable car elle est de puissance plus grande :

$$\mathbb{P}_{\mu_1}(T \in \tilde{\mathcal{R}}) \leq \mathbb{P}_{\mu_1}(T \in \mathcal{R})$$

Y aurait-il pour ce test une région de rejet de même niveau et encore plus puissante que $\mathcal{R}$? Le théorème de Neyman-Pearson permet de répondre à cette question.

4.2.1 Méthode de Neyman-Pearson

Théorème 15 (Test de Neyman-Pearson de deux hyp. simples)**.** *La région critique optimale du test de $\theta = \theta_0$ contre $\theta = \theta_1$ de risque α est définie par*

$$\mathcal{R}_{opt} = \left\{ x \in \mathbb{R}^n | \frac{L(\theta_1; x)}{L(\theta_0; x)} > k_\alpha \right\}; \;\; \mathbb{P}(\{X \in \mathcal{R}_{opt}\}) = \alpha$$

$\mathcal{R}_{opt}$ dépend évidemment de α, même si la notation ne le rappelle pas. Il s'agit donc de refuser (H_0) lorsque la vraisemblance sous (H_1) est vraiment plus grande que celle sous (H_0), le degré de différence étant calibré par la constante k_α, elle-même déterminée par la contrainte du risque de première espèce α.

Preuve dans le cas continu On commence par noter que $RV(X) = L(\theta_1; x)/L(\theta_0; x)$ est bien une statistique parce que θ_1 et θ_0 sont donnés.

Supposons que k_α existe. Soit $\mathcal{R}_\alpha$ une autre région de rejet quelconque de niveau α. Alors

$$\mathbb{P}_{\theta_0}(X \in \mathcal{R}_\alpha) = \alpha$$

Il y a égalité des niveaux des deux régions :

$$\begin{aligned} \mathbb{P}_{\theta_0}(X \in \mathcal{R}_\alpha) &= \mathbb{P}_{\theta_0}(X \in \mathcal{R}_\alpha \setminus \mathcal{R}_{opt}) + \mathbb{P}_{\theta_0}(X \in \mathcal{R}_\alpha \cap \mathcal{R}_{opt}) = \alpha \\ \mathbb{P}_{\theta_0}(X \in \mathcal{R}_{opt}) &= \mathbb{P}_{\theta_0}(X \in \mathcal{R}_{opt} \setminus \mathcal{R}_\alpha) + \mathbb{P}_{\theta_0}(X \in \mathcal{R}_\alpha \cap \mathcal{R}_{opt}) = \alpha \end{aligned}$$

On en déduit $\mathbb{P}_{\theta_0}(X \in \mathcal{R}_\alpha \setminus \mathcal{R}_{opt}) = \mathbb{P}_{\theta_0}(X \in \mathcal{R}_\alpha \setminus \mathcal{R}_{opt})$, soit

$$\int_{x \in \mathcal{R}_\alpha \setminus \mathcal{R}_{opt}} L(\theta_0, x)\, dx = \int_{x \in \mathcal{R}_{opt} \setminus \mathcal{R}_\alpha} L(\theta_0, x)\, dx \tag{4.1}$$

On compare maintenant les puissances :

$$\begin{aligned} \pi(\mathcal{R}_{opt}) = \mathbb{P}_{\theta_1}(X \in \mathcal{R}_{opt}) &= \int_{x \in \mathcal{R}_{opt} \setminus \mathcal{R}_\alpha} L(\theta_1, x)\, dx + \int_{x \in \mathcal{R}_{opt} \cap \mathcal{R}_\alpha} L(\theta_1, x)\, dx \\ \pi(\mathcal{R}_\alpha) = \mathbb{P}_{\theta_1}(X \in \mathcal{R}_\alpha) &= \int_{x \in \mathcal{R}_\alpha \setminus \mathcal{R}_{opt}} L(\theta_1, x)\, dx + \int_{x \in \mathcal{R}_{opt} \cap \mathcal{R}_\alpha} L(\theta_1, x)\, dx \end{aligned}$$

Il reste à comparer le premier terme de chaque somme

$$\begin{aligned} \int_{x \in \mathcal{R}_\alpha \setminus \mathcal{R}_{opt}} L(\theta_1, x)\, dx &\leq \int_{x \in \mathcal{R}_\alpha \setminus \mathcal{R}_{opt}} k_\alpha L(\theta_0, x)\, dx = \int_{x \in \mathcal{R}_{opt} \setminus \mathcal{R}_\alpha} k_\alpha L(\theta_0, x)\, dx \text{ en utilisant (4.1)} \\ &< \int_{x \in \mathcal{R}_{opt} \setminus \mathcal{R}_\alpha} L(\theta_1, x)\, dx \end{aligned}$$

donc, $\pi(\mathcal{R}_\alpha) < \pi(\mathcal{R}_{opt})$. Soit $A(k) = \{x | L(\theta_1, x) > kL(\theta_0, x)\}$. Si $\mathbb{P}(L(\theta_1, x)/L(\theta_0, x) = k) = 0$, alors l'application $k \to \mathbb{P}_{\theta_0}(A(k))$ est une fonction monotone continue. De plus $\mathbb{P}_{\theta_0}(A(0)) = 1$ et $\lim_{k \to +\infty} \mathbb{P}_{\theta_0}(A(k)) = 0$. Par le théorème des valeurs intermédiaires, il existe k_α tel que $\mathbb{P}_{\theta_0}(A(k_\alpha)) = \alpha$. Si ce n'est pas le cas, et qu'il existe k tel que $\mathbb{P}(L(\theta_1, x)/L(\theta_0, x) = k) \neq 0$, on utilisera un test randomisé (voir le cas discret ci-dessous).

Retour à l'exemple Dans le cas gaussien à variance connue, le rapport des vraisemblances est

$$\exp\left(\frac{(\mu_1-\mu_0)\sum_i X_i - (\mu_1-\mu_0)^2}{2\sigma^2}\right)$$

Il ne dépend de l'échantillon que par l'intermédiaire de $\sum_i X_i$ statistique exhaustive. La région de rejet optimale est donc de la forme $\{x|\sum_i x_i > q\}$, ou encore $\{x|t(x) > q^*_{1-\alpha}\}$ avec $t(x) = \sqrt{n}(\bar{x}-\mu_0)/\sigma$. On retrouve le test que l'on avait construit en début de section, et qui est donc le plus puissant.

Preuve dans le cas discret Dans le cas discret, un niveau étant donné, on trouvera rarement une région critique de niveau exactement égal à α. Prenons l'exemple du modèle de Bernoulli $\mathcal{B}(1,\theta)$. On considère le test de $\theta = \theta_0$ contre $\theta = \theta_1 > \theta_0$. Le théorème de Neyman-Pearson propose une région de rejet de la forme $\{s(x)|s(x) > q_{1-\alpha}\}$ pour la statistique $S(X) = \sum_i X_i$ qui suit une loi binomiale $\mathcal{B}(n,\theta_0)$ sous (H_0). Si $\theta_0 = 0.8$ et $n = 5$, alors $\mathbb{P}_{\theta_0}(S \leq 4) = .67$ et $\mathbb{P}_{\theta_0}(S \leq 5) = 1$. On ne peut trouver de quantile dont l'ordre est exactement 0.95.

Par le même raisonnement que pour le cas continu, si $\mathcal{R} = \{x \in \mathbb{R}^n | \frac{L(\theta_1;x)}{L(\theta_0;x)} > k\}$ est de niveau $\alpha(k)$, alors il n'existe pas de test de niveau inférieur ou égal à $\alpha(k)$ dont la puissance soit supérieure à celle de $\mathcal{R}$.

On choisira donc k de façon à ce que $\alpha(k)$ soit le plus proche possible de α tout en lui étant inférieur, ce qui conduit à un test trop conservatif (sauf quand le risque peut par chance être exactement α).

Test randomisé Il est en fait possible, dans le cas d'un modèle de loi discrète, de déterminer un test de risque de première espèce exactement égal à α. Comme le problème provient du fait que les probabilités en un point ne sont pas nulles dans le cas discret, alors qu'elles le sont dans le cas continu, on peut définir la décision suivante :

- décider (H_1) pour tout x tel que $\frac{L(\theta_1;x)}{L(\theta_0;x)} > k$
- décider (H_0) pour tout x tel que $\frac{L(\theta_1;x)}{L(\theta_0;x)} < k$
- décider (H_0) avec probabilité γ (et donc (H_1) avec probabilité $1-\gamma$) si $\frac{L(\theta_1;x)}{L(\theta_0;x)} = k$

Pour tout point frontière, $\frac{L(\theta_1;x)}{L(\theta_0;x)} = k$, on réalise donc un tirage au sort. La résolution du test revient à trouver les inconnues k et γ en écrivant la contrainte de première espèce :

$$\alpha = \mathbb{P}_{\theta_0}\left(x|L(x,\theta_1) > kL(x,\theta_0)\right) \;+\; (1-\gamma)\,\mathbb{P}_{\theta_0}\left(x|L(x,\theta_1) = kL(x,\theta_0)\right)$$

Dans le cas de l'exemple de Bernoulli, $k = 5$, $\mathbb{P}_{\theta_0}(\sum_i X_i = 5) = 0.33$. Si $\sum_i X_i = 5$, on décide (H_0) avec probabilité

$$\gamma = 1 - \frac{\alpha}{\mathbb{P}_{\theta_0}(\sum_i X_i = 5)}$$

soit, pour $\alpha = 0.95$, $\gamma = 0.85$

Définition 25. *Quand un test fait appel à un tirage aléatoire, il est appelé test* ***randomisé*** *ou* ***mixte****. Quand la règle de décision ne fait pas appel à un tirage aléatoire, le test est dit* ***pur****.*

Dans un test randomisé, deux échantillons ayant la même valeur observée de la statistique de test peuvent mener à deux décisions différentes. Dans un test pur, ils conduiront toujours à la même décision. Nous ne considèrerons dans la suite que des tests purs, quitte à ce que leur niveau soit inférieur au niveau attendu.

4.2.2 Propriétés

Propriété 6. *Le test de Neyman-Pearson entre deux hypothèses simples est* ***sans biais*** *: sa puissance π est supérieure à son erreur de première espèce α :*

$$\alpha = \mathbb{P}_{\theta_0}(\mathcal{R}_{opt}) \leq \mathbb{P}_{\theta_1}(\mathcal{R}_{opt}) = \pi$$

Preuve Dans la région de rejet, $L(\theta_1, x) > k_\alpha L(\theta_0, x)$ donc

$$\int_{\mathcal{R}_{opt}} L(\theta_1, x)\, dx > k_\alpha \int_{\mathcal{R}_{opt}} L(\theta_0, x)\, dx$$

Donc, si $k_\alpha > 1$, le résultat est immédiat. Si $k_\alpha \leq 1$, nous pouvons montrer une proposition équivalente : le risque de seconde espèce β vérifie $\beta < 1 - \alpha$

$$\beta = \mathbb{P}_{\theta_1}(\overline{\mathcal{R}_{opt}});\ \ 1 - \alpha = \mathbb{P}_{\theta_0}(\overline{\mathcal{R}_{opt}})$$

Dans la région d'acceptation $\overline{\mathcal{R}_{opt}}$, on a $L(\theta_1, x) \leq k_\alpha L(\theta_0, x)$ d'où

$$1 - \pi = \beta = \int_{\overline{\mathcal{R}_{opt}}} L(\theta_1, x)\, dx \leq k_\alpha \int_{\overline{\mathcal{R}_{opt}}} L(\theta_0, x)\, dx = k_\alpha(1 - \alpha) \leq 1 - \alpha$$

ce qui achève la preuve.

On peut aussi démontrer que le test est consistant sous quelques conditions peu contraignantes.

4.2.3 Utilisation d'une statistique exhaustive

Si on dispose de plus d'une statistique **exhaustive** T, la région critique en dépend exclusivement et le test de NP se réduit à une région de rejet de la forme

$$\mathcal{R}_{opt} = \left\{ t | \frac{g(\theta_1; t)}{g(\theta_0; t)} > k_\alpha \right\};\ \ \mathbb{P}(\{T \in \mathcal{R}_{opt}\}) = \alpha$$

Exemple : nous avons vu que le test de NP de l'espérance pour deux hypothèses simples dans un modèle iid $\mathcal{N}(\mu, 1)$ est le test usuel de la moyenne à variance connue, qui s'écrit en simplement en fonction de la statistique exhaustive $\sum_i X_i$. Il est le plus puissant et sa région de rejet est indépendante de la valeur de $\mu_1 > \mu_0$ qui est l'alternative que nous avons définie dans l'exemple. Or, remarquons que la région de rejet ne dépend pas de μ_1. Ainsi pour chaque valeur de $\mu > \mu_0$, la puissance est la plus grande possible, c'est une fonction de μ_1 et vaut ici, en appelant F^* la fonction de répartition d'une loi gaussienne centrée réduite

$$\pi(\mu_1) = \mathbb{P}_{\mu_1}(X \in \mathcal{R}_{opt}) = 1 - \mathrm{F}^* \left(q^*_{1-\alpha} + \frac{\mu_0 - \mu_1}{\sigma/\sqrt{n}} \right)$$

4.3 Tests à hypothèses composites

Les hypothèses simples sont restrictives. Nous aurons plutôt envie de savoir si les données montrent une augmentation du paramètre concerné par rapport à une situation de référence par exemple. Dans le cas gaussien à variance connue, cela peut se traduire par

$$(H_0) : \mu \leq \mu_0 \text{ contre } (H_1) : \mu > \mu_0 \tag{4.2}$$

ou, pour une différence

$$(H_0) : \mu = \mu_0 \text{ contre } (H_1) : \mu \neq \mu_0 \tag{4.3}$$

Ces hypothèses sont composites ((4.2) est **unilatérale**, (4.3) est **bilatérale**), elles concernent plusieurs valeurs de μ_0 (et même une infinité dans l'exemple) et de μ_1. On peut les généraliser en les écrivant sous la forme

$$(H_0) : \theta \in \Theta_0 \text{ contre } (H_1) : \theta \in \Theta_1 \tag{4.4}$$

4.3.1 Hypothèses unilatérales

Notons que la région de rejet du test de l'espérance que nous avons vu à la section précédente ne dépend pas de la valeur de μ_1. Elle peut donc servir pour toute hypothèse simple alternative, dès que $\mu_1 > \mu_0$. C'est donc une région de rejet intéressante pour

$$(H_0) : \mu = \mu_0 \text{ contre } (H_1) : \mu(= \mu_1) > \mu_0 \tag{4.5}$$

Pour chaque μ_1, nous avons vu à la section précédente que la région de rejet de Neyman-Pearson est la plus puissante. Ainsi, le test de NP d'hypothèses (4.5) est uniformément le plus puissant parmi les tests de niveau α

Définition 26. *Un test est **uniformément plus puissant** (UPP) si, quelle que soit la valeur de $\theta_1 \in \Theta_1$ définissant l'alternative, la puissance $\pi(\theta_1) = \mathbb{P}_{\theta_1}(\mathcal{R}_\alpha)$ du test est supérieure à la puissance de tout autre test de niveau α.*

Mais remarquons que dans (4.5), seule (H_1) est composite. Est-il encore possible de définir un test UPP de l'espérance d'une gaussienne à variance connue avec les hypothèses (4.2) ? La réponse est oui, mais il faut adapter la notion de niveau telle que nous l'avons utilisée jusqu'à maintenant, où elle était synonyme de risque de première espèce :

Définition 27. *Quand l'hypothèse nulle est composite :*
- *le risque de première espèce est une fonction de θ : pour $\theta \in \Theta_0$,*

$$\alpha(\theta) = \mathbb{P}_\theta(T \in \mathcal{R})$$

- *la **taille** du test est définie par :* $\sup_{\theta \in \Theta_0} \alpha(\theta)$
- *Un test est de **niveau** α si sa taille $\leq \alpha$*

Ainsi, dans un test UPP de niveau α, toutes les valeurs du paramètre sous l'hypothèse nulle ne conduisent pas forcément à un test de risque de première espèce α, mais de risque de première espèce inférieur ou égal à α. Dans le cas de notre exemple fil rouge, pour tout $\mu \in \Theta_0 =]-\infty; \mu_0]$

$$\alpha(\mu) = \mathbb{P}_\mu(X \in \mathcal{R}_{opt}) = 1 - \mathrm{F}^* \left(q^*_{1-\alpha} + \frac{\mu_0 - \mu}{\sigma/\sqrt{n}} \right)$$

Pour $\mu \leq \mu_0$, $\alpha(\mu) \leq \alpha(\mu_0) = \alpha$ et $\mathcal{R}_{opt}$ est UPP pour (4.2). La même région de rejet est donc optimale pour des tests dont on a étendu les hypothèses. Cette propriété vient du fait que le rapport de vraisemblance qui est à la base de la définition de la région de rejet est une fonction monotone d'une statistique exhaustive. On appelle ces familles de modèles à rapport de vraisemblance monotone, et la forme de la région de rejet de différentes hypothèses composites est donné par le théorème suivant :

Théorème 16 (Lehman : rapport de vraisemblance monotone)**.** *S'il existe une statistique T exhaustive minimale telle que pour tout couple (θ_1, θ_0) le* ***rapport de vraisemblance*** *$RV = L(\theta_1; x)/L(\theta_0; x)$ soit une fonction* ***monotone*** *de T, alors il existe un test UPP pour les situations d'hypothèses unilatérales :*

- $(H_0) : \theta \leq \theta_0$ et RV est une fonction croissante de $T : \mathcal{R} = \{T > k\}$
- $(H_0) : \theta \leq \theta_0$ et RV est une fonction décroissante de $T : \mathcal{R} = \{T < k\}$
- $(H_0) : \theta \geq \theta_0$ et RV est une fonction croissante de $T : \mathcal{R} = \{T < k\}$
- $(H_0) : \theta \geq \theta_0$ et RV est une fonction décroissante de $T : \mathcal{R} = \{T > k\}$

Preuve
Cas d'un RV croissant et du test $(H_0) : \theta \leq \theta_0$ contre $(H_1) : \theta = \theta' > \theta_0$.

Dans le cas des familles à rapport de vraisemblance monotone, la région de rejet optimale $\mathcal{R}_{opt}$ de taille α de $(H_0) : \theta = \theta_0$ contre $(H_1) : \theta = \theta'$ où $\theta' > \theta_0$, ne dépend pas de θ' et est unique.

De plus, la fonction $\theta \mapsto \mathbb{P}_\theta(\mathcal{R}_{opt})$ est monotone. En effet, soit $\theta_1 < \theta_2$. Le test de $\theta = \theta_1$ contre $\theta = \theta_2 > \theta_1$ pour lequel on utilise la région $\mathcal{R}_{opt}$ précédente est de taille $a = \mathbb{P}_{\theta_1}(\mathcal{R}_{opt})$, c'est un test de Neyman-Pearson de taille a, donc $UPP(a)$ pour tester $\theta = \theta_1$ contre $\theta = \theta_2 > \theta_1$. Donc sa puissance en θ_2 est supérieure à celle de n'importe quel autre test, par exemple celui qui tirerait à pile ou face avec probabilité a, donc

$$\mathbb{P}_{\theta_2}(\mathcal{R}_{opt}) > a = \mathbb{P}_{\theta_1}(\mathcal{R}_{opt})$$

ce qui montre la monotonie.

Maintenant, on utilise cette monotonie pour dire que

$$\mathbb{P}_\theta(\mathcal{R}_{opt}) < \mathbb{P}_{\theta_0}(\mathcal{R}_{opt}) = \alpha \text{ pour tout } \theta < \theta_0$$

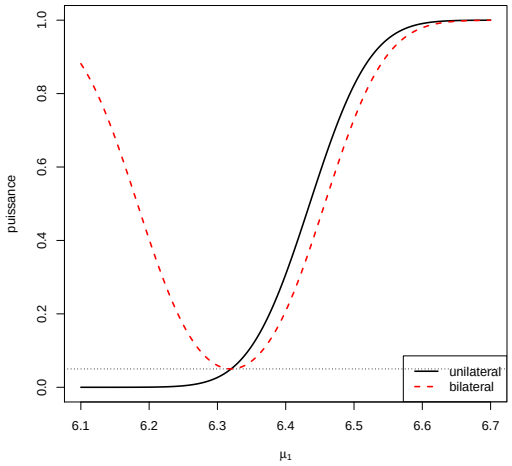
et donc $\mathcal{R}_{opt}$ est de taille α pour tester $\theta \leq \theta_0$ contre $\theta > \theta_0$. Si on considère un autre test de région $\mathcal{R}$ de taille α pour tester les mêmes hypothèses, alors $\mathcal{R}$ est de niveau α pour tester $\theta = \theta_0$ contre $\theta = \theta' > \theta_0$. Mais $\mathcal{R}_{opt}$ est optimale pour ce test donc sa puissance est plus grande $\mathbb{P}_{\theta'}(\mathcal{R}_{opt}) \geq \mathbb{P}_{\theta'}(\mathcal{R})$ pour tout $\theta > \theta$. D'où le résultat

□

Ce théorème s'applique en particulier avec la statistique exhaustive des familles exponentielles de loi, dont est issue la loi gaussienne qui nous sert d'exemple.

4.3.2 Hypothèses bilatérales

Mais l'existence d'un test UPP n'est pas assurée. Par exemple, il n'existe pas de test UPP pour tester (4.3) dans un modèle à rapport de vraisemblance monotone. En effet la région de rejet $\{T > q^*_{1-\alpha}\}$ est de puissance inférieure à la région de rejet de région de rejet $\{T < q^*_\alpha\}$ pour $\theta < \theta_0$. Comme il est raisonnable que les petites et les grandes valeurs fassent rejeter le test, une région de rejet de la forme $\mathcal{R} = \{|T| > q^*_{1-\alpha/2}\}$ sera plus adaptée. Bien qu'elle ne soit pas UPP, elle est sans biais. On peut montrer d'ailleurs montrer qu'on ne peut pas trouver d'autre région sans bais uniformément plus puissante sur cet exemple : le test est **UPP parmi les tests sans biais**.



Théorème 17. *Dans les familles exponentielles à rapport de vraisemblance monotone de statistique exhaustive T, il existe un test* ***UPP-sans biais*** *au niveau α pour tester (4.3), défini par la région de rejet*

$$\mathcal{R} = \{T \notin [c_1; c_2]\}$$

avec $\mathbb{P}_{\theta_0}(\mathcal{R}) = \alpha$. Le test maximise la puissance pour $\theta \neq \theta_0$ parmi les tests sans biais.

Remarque Même si l'usage veut qu'on définisse les constantes c_1 et c_2 en répartissant le risque de première espèce pour moitié de chaque côté, cette méthodologie ne conduit pas toujours à un test UPP-sans biais quand la loi de la statistique n'est pas symétrique. Mais elle est bien pratique et facilite les calculs.

4.4 Extensions

Quand les méthodes précédentes ont échoué, ou que les hypothèses paramétriques sont plus générales, on peut considérer les tests comme le test du rapport de vraisemblances maximales ou le test de Wald qui sont présentés dans cette section.

4.4.1 Test du rapport des vraisemblances maximales

Définition 28. *Soit une famille paramétrique $\mathbb{P}_\theta, \theta \subset \Theta$ et les hypothèses $(H_0) : \theta \in \Theta_0$ contre $(H_1) : \theta \in \Theta_1 = \Theta - \Theta_0$. On appelle* ***rapport des vraisemblances maximales****, la statistique $RV(X)$ telle que*

$$RV(X) = \frac{\sup_{\theta \in \Theta_0} L(\theta; X)}{\sup_{\theta \in \Theta} L(\theta; X)}$$

Le ***test du rapport de vraisemblance*** *(TRV) est le test défini par une région de rejet de la forme*

$$\mathcal{R} = \{RV(X) < k_\alpha \leq 1\}.$$

Par exemple pour le test bilatéral (4.3), où θ_0 peut être un paramètre vectoriel de dimension p

$$RV(X) = \frac{L(\theta_0; X)}{\sup_{\theta\in\Theta} L(\theta; X)}$$

Cette statistique compare la situation de référence avec la meilleure des situations possibles sous l'alternative. Il semble raisonnable de dire que plus $RV(X)$ est grand, et plus l'hypothèse (H_0) est vraisemblable, puis que la vraisemblance de ce modèle est de l'ordre de la plus grande des vraisemblances quand θ varie et l'échantillon est fixé : ce qui revient à remplacer θ par l'estimateur du maximum de vraisemblance au dénominateur.

4.4.2 Exemples

Pour tester (4.3) dans le modèle gaussien à variance connue, on retrouve par cette méthode une région de rejet de la forme

$$\mathcal{R}_\alpha = \{x; |\bar{x} - \mu_0| > k_\alpha\}$$

qui est la forme classique du test bilatéral UPP-sans biais.

La forme plus générale est bien adaptée quand il existe des paramètres qui restent non contraints sous (H_0) : par exemple, le test de $\mu = \mu_0$ (et σ^2 est quelconque) contre $\mu \neq \mu_0$ (et σ^2 quelconque) dans le modèle gaussien à espérance et variance inconnues, $\theta = (\mu, \sigma^2)$. Dans ce cas, l'EMV sous (H_0) est

$$\widehat{\theta}_0 = (\mu_0, \widehat{\sigma}_0^2 = \frac{1}{n}\sum_i (X_i - \mu_0)^2)$$

celui sous (H_1) est

$$\widehat{\theta} = (\bar{X}, \widehat{\sigma}^2 = \frac{1}{n}\sum_i (X_i - \bar{X})^2)$$

En en déduit le rapport des vraisemblances maximales :

$$RV = \left(\frac{\widehat{\sigma}^2}{\widehat{\sigma}_0^2}\right)^{n/2} \frac{\exp - \left(\frac{\sum_i (X_i-\mu_0)^2}{2\widehat{\sigma}_0^2}\right)}{\exp - \left(\frac{\sum_i (X_i-\bar{X})^2}{2\widehat{\sigma}^2}\right)} = \left(\frac{\widehat{\sigma}^2}{\widehat{\sigma}_0^2}\right)^{n/2}$$
$$= \left(\frac{\sum_i (X_i-\mu_0)^2}{\sum_i (X_i-\bar{X})^2}\right)^{-n/2} = \left(1 + n\frac{(\bar{X}-\mu_0)^2}{\sum_i (X_i-\bar{X})^2}\right)^{-n/2}$$

qui fait intervenir la statistique de Student $T = \sqrt{n}\frac{\bar{X}-\mu_0}{\widehat{\sigma}}$: on retrouve la région de rejet classique du test de Student $\mathcal{R}_\alpha = \{t; |t| > qt_{1-\alpha/2}\}$.

Pour le test de $\sigma^2 = \sigma_0^2$ contre $\sigma^2 \neq \sigma_0^2$ dans un modèle gaussien à espérance et variance inconnues,

— Sous (H_0), $\widehat{\theta}_{(H_0)} = (\bar{X}, \sigma_0^2)$

— Sous (H_1), $\widehat{\theta}_{(H_1)} = (\bar{X}, \widehat{\sigma}^2 = \sum_i (X_i - \bar{X})^2/n)$

On en déduit

$$\frac{L(\widehat{\theta}_{(H_0)}, x)}{L(\widehat{\theta}_{(H_1)}, x)} = \left(\frac{\widehat{\sigma}^2}{\sigma_0^2}\right)^{n/2} \exp\left(-\frac{\sum_i (X_i-\bar{X})^2}{2\sigma_0^2} + \frac{n}{2}\right) = \left(\frac{\widehat{\sigma}^2}{\sigma_0^2}\exp(1 - \frac{\widehat{\sigma}^2}{\sigma_0^2})\right)^{n/2} = \left(g(\frac{\widehat{\sigma}^2}{\sigma_0^2})\right)^{n/2}$$

La région de rejet est donc de la forme $\{g(\frac{\widehat{\sigma^2}}{\sigma_0^2}) < k_\alpha\}$. L'étude de la fonction $g(y) = y\exp(1-y)$ indique qu'elle est croissante sur $[0,1]$, puis décroissante pour $y > 1$. De plus, $g(0) = 0$ et $\lim_{y\to+\infty} g(y) = 0$. Par le théorème des valeurs intermédiaires, il existe deux réels positifs, l'un y_1 inférieur à 1, l'autre y_2 supérieur à 1 tels que $g(y_1) = g(y_2) = k_\alpha$. La région d'acceptation de (H_0) est donc

$$\overline{\mathcal{R}_\alpha} = \{y_1 \ < \frac{\widehat{\sigma}^2}{\sigma_0^2} < y_2\} = \{ny_1 \ < n\frac{\widehat{\sigma}^2}{\sigma_0^2} < ny_2\} = \{q_{\alpha_1}^{\chi^2-n} \ < \frac{n\widehat{\sigma}^2}{\sigma_0^2} < q_{1-\alpha_2}^{\chi^2-n}\}$$

On détermine α_1 et α_2 vérifiant $\alpha_1 + \alpha_2 = \alpha$ et $g(q_{\alpha_1}^{\chi^2-n}/n) = g(q_{1-\alpha_2}^{\chi^2-n}/n)$.

Quand la loi de la statistique est symétrique, $\alpha_1 = \alpha/2$ et le test est UPP-sans biais. Mais ce n'est pas le cas ici, car la loi du Khi-deux n'est pas symétrique. Le test classique avec $\alpha_1 = \alpha/2$ n'est donc pas UPP-sans biais, mais en est une bonne approximation quand n est suffisamment grand.

4.4.3 Propriétés du TRV

Le test TRV n'a pas de propriétés d'optimalité notables, mais on constate dans des situations usuelles qu'il est UPP-sans biais. Son asymptotique est bien définie sous certaines hypothèses de régularité

Théorème 18 (asymptotique du RV)**.** *Soit une famille paramétrique $\mathbb{P}_\theta, \theta \subset \Theta$. Si Θ_0 définit une sous-hypothèse linéaire de Θ, $dim(\Theta_0) = q$, $dim(\Theta) = p$, et sous les conditions de régularité de l'EMV, alors, sous (H_0)*

$$-2\log(RV) \xrightarrow{\mathcal{L}} \chi^2(p-q)$$

La région de rejet $\{-2\log(TRV) > q_{1-\alpha}^{\chi^2_{p-q}}\}$ du test de rapport de vraisemblances maximales est asymptotiquement de niveau α.

Eléments de preuve Nous traitons le cas $p = 1$ pour une hypothèse nulle simple $\Theta_0 = \{\theta_0\}$ $(q = 0)$. On développe la log-vraisemblance en série de Taylor autour de $\widehat{\theta}$, l'EMV sous (H_1) : il existe $\theta^\star \in [\theta_0, \widehat{\theta}]$ tel que

$$\log(RV) = \log L(\theta_0, x) - \log L(\widehat{\theta}, x) = (\theta_0 - \widehat{\theta})\frac{\partial}{\partial\theta}\log L(\widehat{\theta}, x) + \frac{1}{2}(\theta_0 - \widehat{\theta})^2\frac{\partial^2}{\partial\theta^2}\log L(\theta^\star, x)$$

Or, par définition de l'EMV, $\frac{\partial}{\partial\theta}\log L(\widehat{\theta}, x) = 0$, d'où

$$-2\log RV = -(\theta_0 - \widehat{\theta})^2\frac{\partial^2}{\partial\theta^2}\log L(\theta^\star, x)$$

Sous (H_0), $\theta = \theta_0$, et $\widehat{\theta}$ tend vers θ_0 p.s., donc $\theta^\star \to \theta_0$ p.s.. Pour $n \to +\infty$,

$$\begin{aligned}\frac{1}{n}\frac{\partial^2}{\partial\theta^2}\log L(\theta^\star, X) = \frac{1}{n}\sum_i \frac{\partial^2}{\partial\theta^2}\log f(X_i, \theta^\star) &\simeq \frac{1}{n}\sum_i \frac{\partial^2}{\partial\theta^2}\log f(X_i, \theta_0) \\ &\to \mathbb{E}_{\theta_0}\left[\frac{\partial^2}{\partial\theta^2}\log f(X_1, \theta_0)\right] = -I_1(\theta_0)\end{aligned}$$

donc $-\frac{\partial^2}{\partial\theta^2}\log L(\theta^\star, X) \simeq nI_1(\theta_0) = I_n(\theta_0)$, d'où

$$-2\log RV \simeq (\theta_0 - \widehat{\theta})^2 I_n(\theta_0)$$

On reconnaît de le carré de $\sqrt{I_n(\theta_0)}(\theta_0 - \widehat{\theta})$ qui tend asymptotiquement vers une gaussienne centrée réduite d'après le comportement asymptotique de l'EMV. Donc $-2\log RV \xrightarrow{\mathcal{L}} \chi^2(1)$.

4.4.4 Test de Wald

On considère un modèle paramétrique régulier de densité $f(x,\theta)$ où θ est un vecteur de $\mathbb{R}^p$ et soit A une matrice de dimension $r \times p$. On souhaite tester :

$$(H_0) : A\theta = A\theta_0 \text{ contre } (H_1) : A\theta \neq A\theta_0$$

L'estimateur du maximum de vraisemblance $\widehat{\theta}_n$ est asymptotiquement normal : sous (H_0)

$$V_n^{-1/2}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, Id_p)$$

En particulier, $V_n = [I_n(\theta_0)]^{-1} = [I_1(\theta_0)]^{-1}/n$ pour un échantillon iid

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, [I_1(\theta_0)]^{-1})$$

d'où

$$\sqrt{n}A(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, AI_1^{-1}(\theta_0)A')$$

ou encore, si $rang(A) = r$,

$$T_n = [AV_nA']^{-1/2}A(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, Id_r) \tag{4.6}$$

Propriété 7. *Si $rang(A) = r$, la statistique de **Wald** est le carré de la norme de T_n, et sa loi asymptotique sous (H_0) est*

$$W = [A(\widehat{\theta}_n - \theta_0]'(AV_nA')^{-1}A(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \chi^2(r) \tag{4.7}$$

On en déduit une région de rejet du **test de Wald** de niveau asymptotique α pour une hypothèse bilatère :

$$\mathcal{R}_\alpha = \{x | W(x) > q_{1-\alpha}^{\chi^2(r)}\}; \;\; \mathbb{P}_{(H_0)}(X \in \mathcal{R}_\alpha) \simeq \alpha$$

Cas particulier Si $r = 1$, T_n n'a qu'une composante, et on peut définir les tests directement à partir de (4.6). La région de rejet suivante est de niveau asymptotique α

$$\mathcal{R}_\alpha = \{x; |T_n(x)| > q^*_{1-\alpha/2}\} = \{x; W(x) > q_{1-\alpha}^{\chi^2(r)}\}$$

L'utilisation de T_n à la place de W permet en particulier de tester des hypothèses unilatérales comme $(H_0) : A\theta = A\theta_0$ contre $(H_1) : A\theta > A\theta_0$, avec $\mathcal{R}_\alpha = \{T_n > q^*_{1-\alpha}\}$ Le test de Wald mis sous cette forme est utilisable plus généralement pour tester toute combinaison linéaire d'un paramètre dont l'estimateur a un comportement asymptotiquement gaussien.

Cas d'une fonction non linéaire de θ On souhaite tester

$$(H_0) : h(\theta) = h(\theta_0) \text{ contre } (H_1) : h(\theta) \neq h(\theta_0)$$

et on dispose d'un emv $\widehat{\theta}$ asymptotiquement normal

$$\widehat{V}_n^{-1/2}(\widehat{\theta}_n - \theta_0) \xrightarrow{\mathcal{L}} \mathcal{N}(0, Id_p)$$

Soit h une fonction de différentielle $Dh(\theta) = A(\theta)$. Si $rang(A(\theta)) = r$, la loi de la statistique de Wald sous (H_0) se déduit de l'utilisation de la méthode delta

$$W = [h(\widehat{\theta}_n) - h(\theta_0)]'(A(\theta_0)\widehat{V}_n A(\theta_0)')^{-1}(h(\widehat{\theta}_n) - h(\theta_0)) \xrightarrow{\mathcal{L}} \chi^2(r)$$

Pour que W soit une statistique, il ne faut pas qu'il reste des paramètres inconnus dans sa définition. Si c'est le cas, on les estime

$$W = [h(\widehat{\theta}_n) - h(\theta_0)]'[A(\widehat{\theta}_n)\widehat{V}_n A(\widehat{\theta}_n)']^{-1}(h(\widehat{\theta}_n) - h(\theta_0)) \xrightarrow{\mathcal{L}} \chi^2(r)$$

La région de rejet du test de Wald est asymptotiquement de niveau α

$$\mathcal{R}_\alpha = \{x | W(x) > q_{1-\alpha}^{\chi^2(r)}\}; \quad \lim_n \mathbb{P}_{(H_0)}(W \in \mathcal{R}_\alpha) \leq \alpha$$

Si $h(\theta) \in \mathbb{R}$, on peut utiliser $T_n = (A(\theta_0)\widehat{V}_n A(\theta_0)')^{-1/2}(h(\widehat{\theta}_n) - h(\theta_0)) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$

Exemple Le test habituel de la comparaison de la moyenne de deux échantillons dont on suppose les lois de même variance est typiquement un test de Wald permettant de tester (H_0) : $\mu_1 = \mu_2$ contre $(H_0) : \mu_1 \neq \mu_2$, ou en le réécrivant : $(H_0) : \mu_1 - \mu_2 = 0$ contre $(H_0) : \mu_1 - \mu_2 \neq 0$. Si de plus, le modèle est gaussien, on dispose de lois exactes (et donc d'un test de niveau exact α) à distance finie... à suivre en TD !

4.4.5 Test de Wald ou TRV ?

Notons que ces deux tests sont asymptotiques, leur niveau n'est garanti que quand $n \to \infty$. A distance finie, la région de rejet définie avec la loi asymptotique n'est qu'approximativement de niveau α. Ces deux tests comparent des modèles emboîtés $\Theta_0 \subset \Theta_1$, où Θ_0 est défini comme la restriction de Θ_1 aux paramètres vérifiant la restriction $h(\theta) = 0$. Ils sont conçus pour avoir le même niveau asymptotique α.

- Le test de Wald est plus simple numériquement : il ne nécessite qu'une seule optimisation, mais dépend de l'estimation (délicate en pratique) de la matrice de covariance.
- le test du rapport de vraisemblance est plus compliqué numériquement : il nécessite deux optimisations, dont une sous contrainte ; mais il souvent considéré comme meilleur.
- Dans certains cas, ces sont des tests équivalents, soit asymptotiquement, soit à distance finie en utilisant les lois exactes.

En présence de décisions contradictoires sur ces deux tests, on pourra préférer le test du rapport de vraisemblance.

Chapitre 5

Intervalle et région de confiance

Un test est une réponse qui peut paraître péremptoire sur la propriété d'une loi (un paramètre ou une fonction du paramètre par exemple) observée par l'intermédiaire d'un échantillon : on accepte l'hypothèse nulle ou on la refuse. Il ne donne pas directement une idée de la variabilité ou de la précision de l'estimation du paramètre : l'intervalle de confiance va matérialiser cette information. Et nous verrons qu'il existe un lien entre ces deux notions.

5.1 Qu'est-ce qu'un intervalle de confiance

L'estimation par intervalle de confiance permet de matérialiser la variabilité de l'estimation, et donc sa fiabilité ou sa précision. On cherche par exemple une borne supérieure $\widehat{\theta}_{sup}$ pour un paramètre θ, pour laquelle on espère fortement (avec une forte probabilité) que θ soit inférieur à cette valeur $\widehat{\theta}_{sup}$, c'est à dire

$$\mathbb{P}[\theta \leq \widehat{\theta}_{sup}] = \mathbb{P}(\theta \in]-\infty, \widehat{\theta}_{sup}]) \geq 1-\alpha \tag{5.1}$$

où α est "petit".

Définition 29. *La variable aléatoire $\widehat{\theta}_{sup} = \widehat{\theta}_{sup}(X)$ définie par* (5.1)*, fonction de l'échantillon X, est appelée* ***borne supérieure de confiance*** *de niveau $1-\alpha$ ou de risque α.*

Pour un tirage amenant à $\widehat{\theta}_{sup} \geq \theta$, $\widehat{\theta}_{sup}$ est effectivement une borne supérieure de θ (que l'on ne connaît pas) ; la perte associée à cette décision est nulle. Pour un tirage amenant à $\widehat{\theta}_{sup} < \theta$, la décision va être erronée : on va dire que la borne supérieure est $\widehat{\theta}_{sup}$, alors que le vrai paramètre n'est pas dans l'intervalle $(-\infty; \widehat{\theta}_{sup}]$. C'est α qui pilote l'erreur commise : plus α est petit, plus le risque de donner une borne supérieure trop petite est faible, mais moins informatif est l'intervalle donné. A l'extrême, choisir une borne supérieure à l'infini donne une confiance maximum, mais plus du tout d'information !

Exemple Soit $X \sim \mathcal{N}(\mu, \sigma^2)$ et σ^2 connu ; $\bar{X}$ estime $\mu = E(X)$, et $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$. On a :

$$\begin{aligned} 1-\alpha &= \mathbb{P}_\mu\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \geq q^*_\alpha\right) \\ &= \mathbb{P}_\mu\left(\mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}} q^*_{1-\alpha}\right) \end{aligned}$$

soit $\widehat{\mu}_{sup} = \bar{X} + \frac{\sigma}{\sqrt{n}} q^*_{1-\alpha}$, où q^*_α est le quantile d'ordre α de la loi normale centrée réduite. Si σ^2 est inconnu, il peut être estimé par son estimateur sans biais $\widehat{\sigma}^2 = \sum_i (X_i - \bar{X})^2/(n-1)$, d'où $\widehat{\mu}_{sup} = \bar{X} + \frac{\widehat{\sigma}}{\sqrt{n}} q^{\mathcal{T}_{n-1}}_{1-\alpha}$ où $\mathcal{T}(n-1)$ est la loi de Student à $n-1$ degrés de liberté.

Définition 30. *L'intervalle de confiance bilatéral $\{\widehat{\theta}_{inf}, \widehat{\theta}_{sup}\}$ de niveau $1-\alpha$ ou de risque α est un intervalle dont les bornes aléatoires dépendent de l'échantillon, et tel que*

$$\mathbb{P}(\widehat{\theta}_{inf} \leq \theta \leq \widehat{\theta}_{sup}) \geq 1 - \alpha.$$

Les bornes de l'intervalle sont aléatoires (l'une des bornes peut être éventuellement infinie dans le cas d'un intervalle de confiance unilatéral). On ne peut dire si la vraie valeur du paramètre appartient à l'intervalle de confiance estimé. Mais si le calcul de l'intervalle est refait sur différents échantillons indépendants, la vraie valeur du paramètre sera incluse dans l'intervalle en moyenne $(1-\alpha) \times 100\%$ des cas.

Méthode de construction La méthodologie utilisée pour calculer l'IC de l'exemple est assez générale, il s'agit de la méthode dite **pivotale**. A partir d'un estimateur de θ, on calcule sa loi sous $\mathbb{P}_\theta$, et on en déduit une statistique pivotale $T_n(\widehat{\theta}, \theta)$ dont la loi ne dépend pas de θ. On exprime les bornes de l'intervalle de confiance en fonction de T_n et de ses quantiles

Dans le cas de la loi exponentielle d'espérance μ, on sait que $2\sum_i X_i/\mu$ suit une loi du Khi-deux à $2n$ degrés de liberté, d'où

$$\mathbb{P}\left(\mu \leq \frac{2\sum_i X_i}{q^{\chi^2_{2n}}_\alpha}\right) = 1-\alpha \text{ et } IC(\mu, 1-\alpha) = \left]0; \frac{2\sum_i X_i}{q^{\chi^2_{2n}}_\alpha}\right]$$

et on en déduit un IC de μ

Si la loi de la statistique pivotale n'est pas connue à distance finie, mais tend asymptotique vers une loi limite qui ne dépend pas de θ, on construit l'IC comme si la loi à distance finie était la loi limite. Pour $n \geq 30$, on peut approcher la loi du Khi-deux par une loi $\mathcal{N}(2n, 4n)$ et $q^{\chi^2_{2n}}_\alpha \simeq 2n + 2\sqrt{n} q^*_\alpha$ où q^* est le quantile de la loi gaussienne centrée réduite d'ordre α.

$$\mathbb{P}\left(\mu \leq \frac{2\sum_i X_i}{2n + 2\sqrt{n} q^*_\alpha}\right) \simeq 1-\alpha \text{ et } IC(\mu, 1-\alpha) \simeq \left]0; \frac{2\sum_i X_i}{2n + 2\sqrt{n} q^*_\alpha}\right]$$

Le niveau de l'IC construit est asymptotiquement $1-\alpha$ mais n'est qu'**approximativement** $1-\alpha$ à distance finie. L'approximation s'améliore avec n croissant. La méthode asymptotique est typiquement à utiliser avec un EMV dont ne sait pas caractériser sa loi exacte à distance finie. Bien sûr, si la loi est connue à distance finie, il vaut mieux utiliser la loi exacte !

Il existe d'autres techniques, par exemple en utilisant des inégalités de probabilités classiques, comme celle de Bienaymé-Tchebychev ou d'Hoeffding (voir Rivoirard et Stoltz (2009)).

5.1.1 IC d'une fonction de θ

La même définition s'applique pour définir un IC d'une fonction $\nu(\theta)$ du paramètre θ

Définition 31. *Soit $X = (X_1, \ldots, X_n)$ un n-échantillon de loi $\mathbb{P}_\theta$, où $\theta \in \Theta \subset \mathbb{R}$ est inconnu. Un **intervalle de confiance** de **niveau** $1-\alpha$ pour $\nu(\theta)$ est un intervalle $[\widehat{\nu}_{inf}(X), \widehat{\nu}_{sup}(X)]$ dont les bornes sont **aléatoires**, telles que, pour tout $\theta \in \Theta$*

$$P_\theta(\widehat{\nu}_{inf}(X) < \nu(\theta) < \widehat{\nu}_{sup}(X)) \geq 1 - \alpha.$$

où α est "petit".

Une réalisation $[\widehat{\nu}_{inf}(x), \widehat{\nu}_{sup}(x)]$ est obtenue à partir des données $x = (x_1, \ldots, \mathrm{x}_n)$.

Méthode de construction On peut partir d'un $IC = [\widehat{\theta}_{min}; \widehat{\theta}_{max}]$ de θ

Si ν est bijective, il suffit de d'appliquer la fonction ν à chaque borne de θ, soit

— $\left[\nu(\widehat{\theta}_{min}); \nu(\widehat{\theta}_{max})\right]$ si ν est croissante

— $[\nu(\widehat{\theta}_{max}); \nu(\widehat{\theta}_{min})]$ si ν est décroissante

Dans les autres cas, on peut utiliser la delta-méthode

Exemples Soit X suivant une loi exponentielle d'espérance μ. L'IC de $\nu(\mu) = P(X > 1) = \exp -(\frac{1}{\mu})$ se construit en notant que

$$1 - \alpha = \mathbb{P}\left(0 < \mu \leq \frac{2\sum_i X_i}{q_\alpha^{\chi^2_{2n}}}\right) = \mathbb{P}\left(0 < \exp(-\frac{1}{\mu}) \leq \exp\left(-\frac{q_\alpha^{\chi^2_{2n}}}{2\sum_i X_i}\right)\right)$$

D'où

$$IC(\nu(\mu), \alpha) = \left]0; \exp\left(-\frac{q_\alpha^{\chi^2_{2n}}}{2\sum_i X_i}\right)\right]$$

Pour construire un IC de la variance $h(\theta) = \theta(1-\theta)$ d'une loi de Bernoulli de paramètre θ, la delta-méthode amène à

$$\frac{\sqrt{n}}{\sqrt{\theta(1-\theta)(1-2\theta)^2}}(h(\widehat{\theta}) - h(\theta)) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Le dénominateur du à la variance ne permet de définir facilement un IC sans calculs supplémentaires. Le lemme de Slutsky permet d'écrire

$$\frac{\sqrt{n}}{\sqrt{\widehat{\theta}(1-\widehat{\theta})(1-2\widehat{\theta})^2}}(h(\widehat{\theta}) - h(\theta)) \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

et un intervalle de confiance bilatère de niveau asymptotique $1-\alpha$ est

$$IC(h(\theta), 1-\alpha) = \left[h(\widehat{\theta}) - q^*_{1-\alpha/2}\frac{\sqrt{\widehat{\theta}(1-\widehat{\theta})(1-2\widehat{\theta})^2}}{\sqrt{n}}; h(\widehat{\theta}) + q^*_{1-\alpha/2}\frac{\sqrt{\widehat{\theta}(1-\widehat{\theta})(1-2\widehat{\theta})^2}}{\sqrt{n}}\right]$$

avec $\mathbb{P}(h(\theta) \in IC(h(\theta), 1-\alpha)) \simeq 1-\alpha$

5.2 Région de confiance

Il est possible d'étendre la notion d'intervalle de confiance à un paramètre θ multidimensionnel.

5.2.1 Région de confiance de Bonferroni

Si $\theta = (\theta_1, \ldots, \theta_p) \in \Theta \in \mathbb{R}^p$, on cherche une région de confiance sous la forme d'un produit cartésien d'intervalles de confiance de chacune des composantes $RC(\theta) = IC(\theta_1) \times ... \times IC(\theta_p)$

$$\begin{aligned}
\mathbb{P}\left(\theta \notin RC(\theta)\right) &= \mathbb{P}(\overline{\theta_1 \in IC(\theta_1) \cap \ldots \cap \theta_p \in IC(\theta_p)}) \\
&= \mathbb{P}(\overline{\theta_1 \in IC(\theta_1)} \cup \ldots \cup \overline{\theta_p \in IC(\theta_p)}) \\
&\leq \underbrace{\mathbb{P}\left(\theta_1 \notin IC(\theta_1)\right)}_{\leq \alpha/p} + \ldots + \underbrace{\mathbb{P}\left(\theta_p \notin IC(\theta_p)\right)}_{\leq \alpha/p}
\end{aligned}$$

Si les intervalles individuels sont de niveau $1-\alpha/p$, alors $RC(\theta)$ est de niveau **simultané** $1-\alpha$:

$$\mathbb{P}\left(\theta \in RC(\theta)\right) \geq 1-\alpha$$

Propriété 8. *L'intersection de K intervalles de confiance de risque α/K forment une région de confiance de Bonferroni de risque simultané α.*

Cette procédure est en général très **conservative**, c'est à dire que le niveau de confiance de l'IC est en général supérieur au niveau réel.

Exemple Région de confiance simultanée de niveau α pour $\theta = (\mu, \sigma^2)$ dans le cas gaussien. On sait que la statistique pivotale $\sqrt{n}(\bar{X}-\mu)/\sqrt{\widehat{\sigma}^2}$ suit une loi de Student $\mathcal{T}(n-1)$ et que $(n-1)\widehat{\sigma}^2/\sigma^2$ suit une loi du Khi-deux à $n-1$ degrés de liberté. On construit deux intervalles de confiance de niveau $1-\alpha/2$

$$A_1 = IC_{1-\alpha/2}(\sigma^2) = \left[\frac{(n-1)\widehat{\sigma}^2}{q^{\chi^2_{n-1}}_{\alpha/4}} \; ; \; \frac{(n-1)\widehat{\sigma}^2}{q^{\chi^2_{n-1}}_{\alpha/4}}\right]$$

$$A_2 = IC_{1-\alpha/2}(\mu) = \left\{\mu \in \mathbb{R} \; tq \; (\bar{X}-\mu)^2 \leq \widehat{\sigma}^2 [q^{\mathcal{T}_{n-1}}_{1-\alpha/4}]^2\right\}$$

La région de confiance est matérialisée en noir sur la figure 5.1.

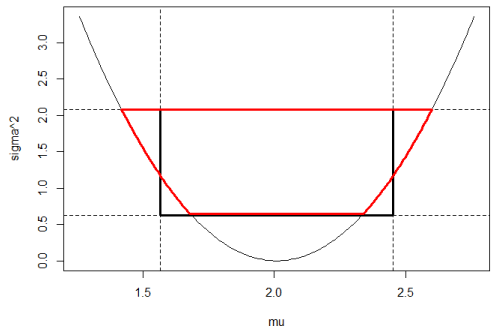


FIGURE 5.1 – Deux RC de Bonferonni de $\theta = (\mu, \sigma^2)$

On peut aussi considérer une RC simultanée en modifiant A_2 de la façon suivante :

$$\tilde{A}_2 = \left\{n(\bar{X}-\mu)^2 \leq \sigma^2 [q^*_{1-\alpha/4}]^2\right\}$$

Or, comme $\bar{X}$ et $\widehat{\sigma}^2$ sont indépendants dans le modèle iid $\mathcal{N}(\mu, \sigma^2)$, les deux intervalles sont indépendants, et le niveau de la RC vaut

$$\mathbb{P}(A_1 \cap \tilde{A}_2) = \mathbb{P}(A_1)\mathbb{P}(\tilde{A}_2) = 1-\alpha+\frac{\alpha^2}{4} > 1-\alpha$$

Si $\alpha = 5\%$, $\alpha^2/4 = 0.0006$ et le niveau confiance de la région simultanée est quasiment le niveau attendu. Les limites de cette RC sont tracées en rouge sur la figure 5.1.

5.2.2 Région de confiance de type Wald

Considérons un estimateur de $\theta \in \Theta \subset \mathbb{R}^p$ asymptotiquement normal

$$(\widehat{\theta} - \theta)' V_n(\widehat{\theta})^{-1} (\widehat{\theta} - \theta) \xrightarrow{\mathcal{L}} \chi^2(p)$$

Alors

$$RC_W(\theta) = \{\theta \in \Theta \; tq \; (\widehat{\theta} - \theta)' V_n(\widehat{\theta})^{-1} (\widehat{\theta} - \theta) \leq q^{\chi^2_p}_{1-\alpha}\}$$

est une région de confiance de niveau asymptotique $1-\alpha$, et à distance finie,

$$\mathbb{P}(\theta \in RC_W(\theta)) \leq 1 - \alpha$$

Cette méthode permet de construire une RC simultanée des composantes du paramètre. Elle s'étend à la construction d'une RC simultanée de $A\theta$ en utilisant la statistique de Wald.

Exemple Le score et le hessien du modèle gaussien $\mathcal{N}(\theta = (\mu, \sigma^2))$ s'écrivent

$$\frac{\partial}{\partial \theta} \log L(\theta, X) = \begin{pmatrix} \frac{\sum_i (X_i - \mu)}{\sigma^2} \\ -\frac{n}{2\sigma^2} + \frac{\sum_i (X_i - \mu)^2}{2\sigma^4} \end{pmatrix} \; ; \; \frac{\partial^2}{\partial \theta^2} \log L(\theta, X) = \begin{pmatrix} -\frac{n}{\sigma^2} & -\frac{\sum_i (X_i - \mu)}{\sigma^4} \\ -\frac{\sum_i (X_i - \mu)}{\sigma^4} & \frac{n}{2\sigma^4} - \frac{\sum_i (X_i - \mu)^2}{\sigma^6} \end{pmatrix}$$

On en déduit l'information de Fisher

$$I_n = -\mathbb{E}\left(\frac{\partial^2}{\partial \theta^2} \log L(\theta, X)\right) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix} = nI_1$$

D'où en utilisant l'asymptotique de l'EMV dans le modèle gaussien (statistique de Wald avec $A = Id_2$)

$$\frac{n}{\sigma^2}(\bar{X} - \mu)^2 + \frac{n}{2\sigma^2}(S_n^2 - \sigma^2)^2 \xrightarrow{\mathcal{L}} \chi^2(2)$$

et la RC simultanée de niveau asymptotique $1-\alpha$

$$\frac{n}{\sigma^2}(\bar{X} - \mu)^2 + \frac{n}{2\sigma^2}(S_n^2 - \sigma^2)^2 \leq q^{\chi^2_2} 1 - \alpha$$

Cette région est représentée en pointillé large bleu sur la figure 5.2. On peut en faire une approximation elliptique centrée autour de l'EMV en écrivant par Slutsky (cf pointillé fin rose)

$$\frac{n}{\widehat{\sigma}^2}(\bar{X} - \mu)^2 + \frac{n}{2\widehat{\sigma}^2}(S_n^2 - \sigma^2)^2 \leq q^{\chi^2_2} 1 - \alpha$$

5.3 Lien entre intervalle de confiance et test

Considérons le test de $(H_0) : \mu = \mu_0$ contre $(H_1) : \mu \neq \mu_0$ dans le modèle gaussien à variance connue

$$\alpha = \mathbb{P}_{\mu_0}\left(|\frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}| > q^*_\alpha\right) = \mathbb{P}_{\mu_0}\left(T \in \mathcal{R}_\alpha\right)$$

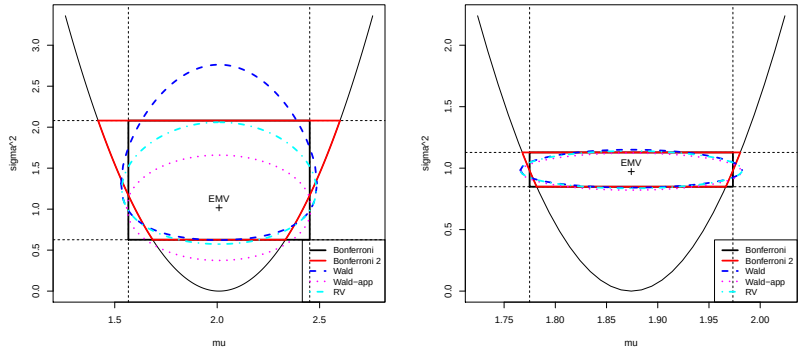


FIGURE 5.2 – Limites de la RC simultanée de type Wald en pointillé large bleu, avec une approximation utilisant Slutsky en pointillé fin rose et RC de type Rapport de Vraisemblance en pointillé irrégulier bleu turquoise. A gauche pour $n = 30$, à droite pour $n = 500$

d'où

$$\begin{aligned} 1-\alpha = \mathbb{P}_{\mu_0}\left(T \notin \mathcal{R}_\alpha\right) &= \mathbb{P}_{\mu_0}\left(|\underbrace{\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}}_{T}| \leq q^*_\alpha \right) \\ &= \mathbb{P}_{\mu_0}\left(\bar{X} - \frac{\sigma}{\sqrt{n}} q^*_{1-\alpha} \leq \mu_0 \leq \bar{X} + \frac{\sigma}{\sqrt{n}} q^*_{1-\alpha} \right) \\ &= \mathbb{P}_{\mu_0}\left(\mu_0 \in IC(\mu_0, 1-\alpha)\right) \end{aligned}$$

Ainsi, le test de région de rejet $\mathcal{R}_\alpha = \{x; t(x) < q^*_\alpha\}$ est équivalent à celui qui décide le rejet quand $\mu_0 \notin IC(\mu, 1-\alpha)$, et l'acceptation sinon.

De façon plus générale :

— si RC est une région de confiance de niveau $1-\alpha$ de θ, alors pour tout $\theta^\star$, le test qui décide de rejeter (H_0) quand $\theta^\star \notin RC$ est un test de niveau α pour tester $(H_0) : \theta = \theta^\star$ contre $(H_1) : \theta \neq \theta^\star$

— si pour tout $\theta^\star$ on dispose d'un test de niveau α de $(H_0) : \theta = \theta^\star$ contre $(H_1) : \theta \neq \theta^\star$, alors la région d'acceptation de ce test est une région de confiance de niveau $1-\alpha$ pour θ.

On voit que cela fonctionne également dans le cas unilatéral de l'exemple introductif où on a écrit

$$1-\alpha = P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \geq q^*_\alpha\right) = P\left(\mu \leq \bar{X} + \frac{\sigma}{\sqrt{n}} q^*_{1-\alpha}\right)$$

d'où

$$\alpha = \mathbb{P}_\mu \left(\underbrace{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}_{T} < q_\alpha^* \right) = \mathbb{P}_\mu \left(T \in \mathcal{R}_\alpha \right)$$
$$= \mathbb{P} \left(\mu > \bar{X} + \frac{\sigma}{\sqrt{n}} q_{1-\alpha} \right) = \mathbb{P}_\mu \left(\mu \notin IC(\mu, 1-\alpha) \right)$$

Ainsi, la région de rejet $\mathcal{R}_\alpha = \{x; t(x) < q_\alpha^*\}$ est de niveau α pour le test de $(H_0) : \mu = \mu_0$ contre $(H_1) : \mu < \mu_0$. Ce test est identique à celui rejetant (H_0) si μ_0 n'appartient pas à l'intervalle de confiance $IC(\mu, 1-\alpha)$.

5.3.1 RC du Rapport de Vraisemblance

A partir du test de RV de niveau α de

$$(H_0) : \theta = \theta^\star \text{ contre } (H_1) : \theta \in \Theta \setminus \{\theta^\star\} \subset \mathbb{R}^p,$$

on dérive la RC de θ de niveau de confiance $1-\alpha$ en utilisant la statistique du rapport de vraisemblance

$$RV = -2 \log \frac{L(\theta^\star; X)}{L(\widehat{\theta}_{emv}; X)} \xrightarrow{\mathcal{L}} \chi^2(p)$$

soit

$$IC(\theta^\star) = \left\{ \theta \in \Theta; RV \leq q_{1-\alpha}^{\chi^2(p)} \right\}$$

Cette région de confiance de $\theta = (\mu, \sigma^2)$ dans le cas iid gaussien est représentée figure5.2 sous forme d'un pointillé irrégulier bleu turquoise. Pour n grand, les IC de Wald et du rapport de vraisemblance sont équivalents.

5.4 Simulation d'un niveau

Des simulations peuvent permettre de comparer les niveaux attendus et observés des tests ou ICs. La procédure est la suivante :

- Répéter B fois de façon indépendante la simulation de l'IC de niveau $1-\alpha$ d'un paramètre $\nu(\theta)$:
  - Générer un n-échantillon $X^b = (X_1, \ldots, X_n)$ de loi F_θ
  - Calculer l'IC observé IC^b
- Le niveau de confiance observé est estimé par $1 - \widehat{\alpha} = \sum_b \mathbb{1}_{\theta^\star \in IC^b} / B$

La procédure est similaire pour estimer le niveau réel d'un test : on compte le nombre moyen de rejet du test sur B simulations indépendantes.

Ce type de simulation s'appelle **méthode de Monte-Carlo**, elle utilise l'approximation d'une espérance par une moyenne de B réalisations indépendantes (d'un IC, d'un test) dont le résultat est binaire :

- pour un IC : μ_0 appartient ou non à l'IC : $Y_b = \mathbb{1}_{\theta^\star \in IC^b} \sim \mathcal{B}(1, 1-\alpha)$
- pour un test : (H_0) rejetée ou non : $Y_b = \mathbb{1}_{t^b_{obs} \in \mathcal{R}} \sim \mathcal{B}(1, \alpha)$

L'estimation est sans biais : $\mathbb{E}(\widehat{\alpha}) = \alpha$ et fortement consistante par la loi des grands nombres. On peut utiliser deux méthodes pour calculer la variance de cette estimation :

1. $Y_b \sim \mathcal{B}(1, 1-\alpha)$, d'où $\text{var}(\widehat{\alpha}) = \alpha(1-\alpha)/n$
2. Reprendre K fois la simulation du niveau et utiliser l'estimateur de la variance d'un K-échantillon

Ces méthodes de simulations permettent de comparer des niveaux d'IC/ RC ou de tests, et de tester si le niveau observé est bien le niveau annoncé.

Chapitre 6

Modèle linéaire

Les modèles de régression permettent d'expliquer ou de prédire une variable aléatoire *réponse* Y en fonction d'une liste de variables *explicatives* $X = (X_1, \ldots, X_p)$ observées sur des individus ou des objets. C'est le cas de très nombreuses situations, que nous avons évoquées dans l'introduction. Par exemple, on souhaite expliquer le taux d'ozone Y *variable réponse* en fonction de la température à midi X_1 (figure 1.1), mais aussi de la pression X_2, de la direction du vent X_3 qui sont les *variables explicatives*.

Le modèle de régression linéaire est un outil simple pour modéliser de nombreuses situations concrètes. Il généralise très largement l'estimation de l'espérance à partir d'un échantillon, en permettant de faire varier celle-ci en fonction de conditions d'expériences. Les propriétés inférentielles (estimateurs, intervalle de confiance et tests) s'établissent avec des outils simples d'algèbre linéaire et de vecteurs gaussiens. Le modèle de régression linéaire est un outil de base de l'ingénieur et du statisticien, présent dans tous les logiciels statistiques.

Bibliogprahie La bibliographie associée est nombreuse, on peut conseiller par exemple les références suivantes : Azais et Bardet (2005), Cornillon et Matzner-Løber (2007), Cornillon et autres (2008), Pagès (2005), Rivoirard et Stoltz (2009).

Notations Soit X une matrice de dimension $n \times p$. Nous noterons x_i une ligne de X, X_j une colonne de X, c'est à dire $\mathrm{x}_i = (x_{i1}, \ldots, x_{ip})$ et $X_j = (x_{1j}, \ldots, x_{nj})'$ où la notation prime désigne la transposée d'une matrice.

6.1 Définition et hypothèses

On dispose de n couples $(\mathrm{x}_1, Y_1), \ldots, (\mathrm{x}_n, Y_n)$ où Y_i est la i-ème observation associée aux valeurs de p variables explicatives $\mathrm{x}_i = (x_{i1}, \ldots, x_{ip})$. La **régression linéaire** est un modèle à bruit additif qui postule que la loi de Y_i s'écrit comme la somme d'un terme déterministe m_i et d'une variable aléatoire ε_i

$$\begin{aligned} Y_i &= m_i + \varepsilon_i \\ &= x_{i1}\theta_1 + \ldots + x_{ip}\theta_p + \varepsilon_i \\ &= \mathrm{x}_i\theta + \varepsilon_i \end{aligned} \tag{6.1}$$

où le ε_i (appelé **bruit** ou **erreur** ou **résidu**) est une variable aléatoire d'espérance nulle et de variance ne dépendant pas de x_i. Le terme déterministe dépend des variables explicatives x_i et d'un paramètre θ. C'est une fonction appelée **fonction de régression** et elle est linéaire en le paramètre dans le cas du modèle linéaire :

$$m(\mathrm{x}_i) = \mathrm{x}_i\theta = \mathbb{E}(Y_i|\mathrm{x}_i).$$

Elle représente la valeur de l'espérance de Y_i qui a été observé sous la condition d'expérience x_i

Les hypothèses de la régression linéaire (6.1) s'expriment donc de la façon suivante :

1. L'espérance est linéaire en θ : $\mathbb{E}(Y_i|\mathrm{x}_i) = \mathrm{x}_i\theta$, où x_i est le vecteur ligne des covariables et θ est le vecteur colonne de dimension p des p **paramètres** inconnus.
2. Les **résidus sont centrés** : $\mathbb{E}(\varepsilon_i|\mathrm{x}_i) = 0$.
3. La **variance de l'erreur est constante**, indépendantes des variables explicatives : $\mathrm{var}(\varepsilon_i|\mathrm{x}_i) = \sigma^2$.
4. Les résidus sont **décorrélés** : $\mathrm{cov}(\varepsilon_i, \varepsilon_j|\mathrm{x}_i, \mathrm{x}_j) = 0$ pour $i \neq j$.

La formulation individuelle précédente peut se reformuler matriciellement par

$$Y_{n\times 1} = X_{n\times p}\theta_{p\times 1} + \varepsilon_{n\times 1}, \;\; \varepsilon_{n\times 1} \sim \mathcal{N}(0, \sigma^2 I_n), \tag{6.2}$$

avec $Y = (Y_1, \ldots, Y_n)'$ le vecteur colonne (aléatoire, de dimension n) des observations, X est la **matrice du plan d'expérience** et I_n la matrice identité de taille n. X est donnée, de taille $n \times p$; c'est la concaténation des n vecteurs lignes x_i ou celle des p variables colonnes $X_j = (x_{1j}, \ldots, x_{nj})'$.

Le modèle dépend des paramètres $\theta \in \mathbb{R}^p$ et $\sigma \in \mathbb{R}^{+*}$ et la famille de loi paramétrique est $\mathcal{P} = \mathcal{N}(X\theta, \sigma^2 I_n)$. Le paramètre de ce modèle est $\beta = (\theta, \sigma^2)$. De manière plus générale, le modèle peut être défini par

$$Y = m + \varepsilon, \;\; \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n), \;\; m \in V \text{ sous espace vectoriel de } \mathbb{R}^p. \tag{6.3}$$

On ne fait plus alors référence à X, mais seulement à son image, $V = Im(X)$. Cette présentation renvoie souvent à la terminologie **modèle linéaire** gaussien.

6.1.1 Remarques sur les hypothèses

Les résidus sont centrés

Cette hypothèse traduit le fait qu'il ne doit pas manquer de terme pertinent au modèle proposé. En effet, l'oubli d'une covariable entraîne un biais qui décentre le bruit, comme par exemple lors de l'utilisation d'une droite pour estimer un phénomène parabolique.

La variance de l'erreur est constante

Pour vérifier cette hypothèse, il faut disposer de répétitions d'observations pour chaque jeu de variables explicatives x, comme par exemple dans le cas de l'analyse de variance (ANOVA). Pour des données individuelles, la variabilité d'observations dans des conditions proches pourra donner un indicateur graphique.

Les résidus sont décorrélés

Ce postulat est en général considéré comme vérifié quand chaque observation correspond à un échantillonnage indépendant ou à une expérience physique menée dans des conditions indépendantes. En revanche, si la variable temps a de l'importance, il est plus difficilement vérifié (une évolution peut ne pas être totalement indépendante d'un passé proche), et une modélisation plus fine des résidus pourra être mise en place (séries chronologiques).

Cas d'un bruit gaussien

On parle alors de **régression linéaire gaussienne**. Cette hypothèse est importante à distance finie (n fixé), car c'est elle qui permet d'accéder aux lois des estimateurs. Si le nombre d'observations est grand, cette hypothèse n'est plus nécessaire, mais au prix de la perte de l'exactitude de la loi des estimateurs à distance finie. Notons aussi l'extension possible à un bruit non forcément homoscédastique, ni décorrélé, qui est parfois appelé plus généralement modèle linéaire gaussien, quoique cette différence de dénomination ne soit pas complètement fixée.

Intercept

En général, la première variable explicative est constante égale à 1 sur l'ensemble des observations ($\forall i, x_{i1} = 1$) ; dans ce cas, la composante du paramètre associée θ_1 s'appelle l'**intercept**.

6.1.2 Exemples

Régression simple

C'est le cas d'un modèle avec une variable explicative et un intercept. Sur la Figure 1.1, les observations de la concentration en ozone y_i sont tracées en fonction de la température à midi t_i. La droite de régression est $m(\mathrm{x}_i) = \mathrm{x}_i\theta = a + bt_i$, soit $\mathrm{x}_i = (1, t_i)$, $\theta = (a, b)$. La matrice du plan d'expérience se réduit à

$$X = \begin{pmatrix} 1 & t_1 \\ \vdots & \vdots \\ 1 & t_n \end{pmatrix}.$$

C'est Sir Galton (1822-1911) qui a introduit l'expression *régression*. Cherchant à expliquer la taille des fils par celle des pères, Sir Galton avait constaté que si un père est plus grand que la moyenne, son fils l'est probablement aussi, mais en moyenne, le fils s'en écarte moins que son père [1]. Il y avait donc *régression* vers la moyenne, la pente b étant dans cet exemple inférieure à 1. L'expression régression linéaire peut bien sûr être maintenant utilisée pour une pente supérieure à 1.

Régression multiple

La concentration en ozone peut dépendre de bien d'autres variables explicatives, par exemple, la vitesse du vent, l'hygrométrie ou la pression. L'expression *régression multiple* est généralement réservée à la prise en compte simultanée de plusieurs variables explicatives quantitatives.

Nous verrons à la section 6.6 comment prendre en compte une variable explicative qualitative, et son interaction avec une variable explicative quantitative. Certains auteurs continuent à désigner la régression linéaire avec régresseurs qualitatifs et quantitatifs comme régression multiple,

1. voir par exemple la présentation de cette anecdote le 11 décembre 2008 par Norbert Schappacher : http ://www.afhalifax.ca/magazine/wp-content/sciences/Darwin/Gradualisme/HM9.pdf

tandis que d'autres l'appelle modèle linéaire, et d'autres encore, comme dans le logiciel SAS, *modèle linéaire général*.

Régression polynômiale

La régression polynômiale est un cas particulier de régression multiple dont les régresseurs sont les puissances d'une variable explicative. Ainsi, la fonction de régression s'écrit, avec $\mathrm{x} = (1, t, \ldots, t^{p-1})$,

$$m(\mathrm{x}) = \theta_1 + \theta_2 t + \ldots + \theta_p t^{p-1}. \tag{6.4}$$

La matrice du plan d'expérience est telle que

$$X = \left(t_i^j\right)_{i=1,\ldots,n}^{j=0,\ldots,p-1}.$$

La fonction de régression $m(\mathrm{x}) = \mathrm{x}\theta$ est bien linéaire par rapport au paramètre, mais pas en la variable t. La figure 6.1 représente,à partir du même jeu de données, différentes régressions

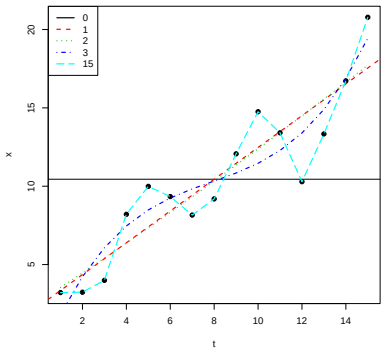


FIGURE 6.1 – Exemples de régression polynomiale pour différents degrés du polynôme

polynomiales calculées pour différents ordres du polynôme. Une problématique générale sera de savoir comment choisir l'ordre (dans le cas de la régression polynômiale) ou plus généralement le nombre et les bonnes variables pour expliquer un modèle.

Linéarité par rapport au paramètre

L'appellation *linéaire* ne veut pas dire que le lien entre les variables explicatives et la variable à expliquer est linéaire, mais que le **modèle est linéaire par rapport aux paramètres**. Nous l'avons déjà noté dans l'exemple de la régression polynômiale. De même, $Y = \alpha + \beta \exp(X) + \varepsilon$ et $\log(Y) = \log(\alpha) + \beta \log(X) + \varepsilon$ sont des cas de régression linéaire, alors que $Y = \alpha + \exp(\beta X) + \varepsilon$ ne l'est pas.

6.1.3 Identifiabilité du paramètre θ

Pour pouvoir estimer le paramètre θ, il faut que l'observation d'un échantillon donné conduise à une unique valeur de ce paramètre.

Théorème 19 (Conditions nécessaires et suffisantes d'identifiabilité)**.** *La paramétrisation* $\mathbb{E}(Y) = X\theta$ *est* ***identifiable*** *si et seulement si l'une des propriétés équivalentes suivantes est vérifiée :*

- *les colonnes de* X *sont indépendantes,*
- X *est de rang plein,*
- X *est injective,*
- $Ker(X) = \{0\}$.

Si elle ne l'est pas, tout élément c de $Ker(X)$ est tel que $\mathbb{E}(Y) = X\theta = X(\theta + c)$. La représentation de l'espérance $\mathbb{E}(Y)$ n'est donc pas unique en θ qui n'est pas interprétable ni estimable, à moins de faire des hypothèses supplémentaires.

Définition 32. *Un modèle identifiable est appelé* ***régulier*** *; un modèle non identifiable est appelé* ***singulier****.*

La dimension du modèle est la dimension de $Im(X)$*. C'est donc la dimension du paramètre* θ *si le modèle est identifiable.*

6.2 Estimation

Dans le modèle linéaire

$$Y = X\theta + \varepsilon = m + \varepsilon, \;\; \mathbb{E}(\varepsilon) = 0; \;\; \text{var}(\varepsilon) = \sigma^2 I_n.$$

estimer la relation entre la variable réponse Y et les p variables explicatives, c'est estimer le paramètre θ : l'hypothèse de linéarité de l'espérance permet de simplifier le problème initial, c'est à dire passer de l'estimation de n paramètres d'espérance $m_i,\ i = 1, \ldots, n$, à l'estimation d'un nombre fini p de paramètres $\theta_j,\ j = 1, \ldots, p$. La variance σ^2 est également un paramètre (de nuisance) qu'il faut également estimer.

On supposera dans la suite que le modèle est identifiable. On pourra se référer à la section 6.6 pour l'estimation de modèles non identifiables.

6.2.1 Estimation par moindres carrés

Lorsque la forme de loi de ε n'est pas spécifiée, un estimateur naturel est celui qui minimise l'erreur entre l'observation Y et sa prédiction $X\widehat{\theta} = \widehat{m}$. On cherche donc un estimateur de θ sous forme d'un minimiseur de la **somme des carrés résiduels**

$$SCR(\theta) = \sum_{i=1}^{n} (Y_i - \mathrm{x}_i\theta)^2 = ||Y - X\theta||^2,$$

où $||.||$ est la norme euclidienne dans $\mathbb{R}^n$, i.e. :

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} ||Y - X\theta||^2, \tag{6.5}$$

d'où le nom d'estimateur des moindres carrés (EMC). Il en est de même quand la formulation du modèle linéaire est donnée par (6.3) :

$$\widehat{m} = \arg\min_{m \in V} ||Y - m||^2.$$

L'EMC cherche m (ou θ) tel que l'erreur résiduelle soit minimale. La dérivée de $SCR(\theta)$ par rapport à θ est le vecteur :

$$\frac{\partial SCR(\theta)}{\partial \theta} = -2X'Y + 2X'X\theta$$

Comme le modèle est identifiable, la matrice $X'X$ est inversible. En effet, soit $v \in Ker(X'X)$ alors $X'Xv = 0$, donc $||Xv||^2 = v'X'Xv = 0$ soit $Xv = 0$ d'où $v = 0$ car X est injective. La matrice des dérivées secondes est déterministe, définie positive

$$\frac{\partial^2 SCR(\theta)}{\partial \theta^2} = 2X'X,$$

permettant de caractériser un minimum. On en déduit le théorème suivant

Théorème 20. *Quand le modèle est identifiable, l'EMC de θ est*

$$\widehat{\theta} = (X'X)^{-1}X'Y.$$

Une autre façon d'interpréter ce résultat est de noter que la solution du problème d'optimisation est le projecteur orthogonal H_V de Y sur $V = Im(X)$:

Propriété 9. *Soit X une matrice injective. La **projection orthogonale** sur $Im(X) \subset \mathbb{R}^n$ est*

$$H_X = X(X'X)^{-1}X'.$$

On vérifie en effet que $H'_X = H_X$, $H_X^2 = H_X$, $H_X X = X$.

Définition 33. *On appelle vecteur des **valeurs ajustées**, la projection orthogonale $\widehat{Y}$ de Y sur $Im(X)$:*

$$\widehat{m} = H_X Y = X\widehat{\theta} = \widehat{Y}.$$

*La matrice de la projection orthogonale H_X est appelée **hat matrix** car c'est elle qui "met un chapeau" sur Y.*

Enfin, une troisième méthode permet de calculer $\widehat{\theta}$, en utilisant la décomposition unique de tout vecteur $Y \in \mathbb{R}^n$ en une composante sur $Im(X)$ et une composante sur son orthogonal $Im(X)^\perp$:

$$Y = H_X Y + (I - H_X)Y.$$

La quantité $(I - H_X)Y$ étant un élément de $Im(X)^\perp$ est orthogonale à tout élément quelconque $Im(X) = \{v = X\alpha, \alpha \in \mathbb{R}^p\}$:

$$\begin{aligned} < v, (I - H_X)Y > &= 0, \forall v \in Im(X) \\ < X\alpha, (I - H_X)Y > &= 0, \forall \alpha \in \mathbb{R}^p \\ X'Y &= X'H_X Y \text{ avec } H_X Y = X\widehat{\theta} \\ X'Y &= X'X\widehat{\theta} \text{ avec } X \text{ de rang plein} \\ \widehat{\theta} &= (X'X)^{-1}X'Y \end{aligned}$$

6.2.2 Exemples

Régression linéaire simple

Le modèle s'écrit $Y = \theta_1 \mathbb{1} + \theta_2 X_2 + \varepsilon$, où $\mathbb{1}$ est l'intercept, première colonne du plan d'expérience X et $X_2 = (x_1, \ldots, x_n)'$ est la deuxième colonne contenant les valeurs de l'unique variable explicative. On a

$$X'X = \begin{pmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{pmatrix}; \quad (X'X)^{-1} = \frac{1}{n\sum_i x_i^2 - \sum_i x_i \sum_i x_i} \begin{pmatrix} \sum_i x_i^2 & -\sum_i x_i \\ -\sum_i x_i & n \end{pmatrix}$$

soit

$$\widehat{\theta} = \begin{pmatrix} \widehat{\theta}_1 \\ \widehat{\theta}_2 \end{pmatrix} = (X'X)^{-1}X'Y = \begin{pmatrix} \bar{Y} - \bar{x}\widehat{\theta}_2 \\ \frac{(\frac{1}{n}\sum_i x_i Y_i) - \bar{x}\bar{Y}}{(\frac{1}{n}\sum_i x_i^2) - \bar{x}^2} \end{pmatrix}$$

Notons que les Y_i et $\bar{Y}$ sont aléatoires, tandis que les x_i et $\bar{x}$ sont donnés. La droite de régression

$$\widehat{y} = \widehat{\theta}_1 + \widehat{\theta}_2 x$$

passe par le point moyen $(\overline{x}, \overline{Y})$. Ces formules peuvent être également obtenues en dérivant directement SCR par rapport au paramètre.

Coordonnées de $\widehat{\theta}$

Quand le modèle est identifiable, $\widehat{\theta}$ est unique et définit les coordonnées de $\widehat{Y}$ dans le repère $(X_1, \ldots, X_p)$ formé des colonnes de X :

$$\widehat{Y} = X\widehat{\theta} = \sum_{j=1}^{p} \widehat{\theta}_j X_j.$$

Mais ce repère n'est en général pas orthogonal, et $\widehat{\theta}_j$ n'est donc pas en général la projection orthogonale de Y sur X_j

$$\begin{aligned} H_{X_j}Y &= H_{X_j}H_X Y \\ &= \widehat{\theta}_1 H_{X_j}X_1 + \ldots + \widehat{\theta}_p H_{X_j}X_p \\ &= \widehat{\theta}_j X_j + \sum_{i \neq j} \widehat{\theta}_j H_{X_j}X_i. \end{aligned}$$

Par exemple, si $X_1 = \mathbb{1}$ est le vecteur dont toutes les coordonnées valent 1 (associé à l'estimation de l'intercept), alors

$$H_{X_1}(Y) = \overline{Y}X_1,$$

alors qu'en régression linéaire simple $Y = \theta_1\mathbb{1} + \theta_2 X_2 + \varepsilon$ avec $X_1 = \mathbb{1}$, l'estimateur de l'intercept est

$$\widehat{\theta}_1 = \overline{Y} - \widehat{\theta}_2\overline{x},$$

en général différent de $\overline{Y}$. Lorsque toutes les variables sont orthogonales deux à deux, $X'X$ est alors une matrice diagonale et $H_{X_j}Y = \widehat{\theta}_j X_j$.

6.2.3 Propriétés des estimateurs

Le biais et la variance de $\widehat{\theta}$ se déduisent immédiatement de la définition des estimateurs :

Propriété 10. *L'EMC $\widehat{\theta}$ de θ est sans biais :* $\mathbb{E}(\widehat{\theta}) = \theta$ *et de variance* $\mathrm{var}(\widehat{\theta}) = \sigma^2(X'X)^{-1}$.

De plus, $\widehat{\theta}$ vérifie la propriété de Gauss-Markov :

Théorème 21 (Gauss-Markov)**.** *Parmi les estimateurs* ***linéaires et sans biais*** *de θ, l'EMCO $\widehat{\theta}$ est de variance minimum.*

Démonstration. Soit $\tilde{\theta} = CY$ un autre estimateur linéaire et sans biais de θ. La condition sans biais se traduit par $E(CY) = CX\theta = \theta$, soit $CX = I_p$. Définissons maintenant $M = C - A$, avec $A = (X'X)^{-1}X'$. On vérifie que $MA' = AM' = 0$, et donc

$$\mathrm{var}(\tilde{\theta}) = \sigma^2 CC' = \sigma^2(MM' + AA') = \mathrm{var}\,\widehat{\theta} + \sigma^2 MM'$$

Comme MM' est semi définie positive ($u'M'Mu \geq 0, \forall u \in \mathbb{R}^p$), $\widehat{\theta}$ est de moindre variance que $\tilde{\theta}$.

$\Diamond$

L'estimateur $\widehat{s}^2$ est biaisé à distance finie, et asymptotiquement sans biais. En effet, soit $\widehat{\varepsilon} = Y - X\widehat{\theta}$; alors,

$$\mathbb{E}(SCR(\widehat{\theta})) = \mathbb{E}(||\widehat{\varepsilon}||^2) = \mathbb{E}(tr(\widehat{\varepsilon}'\widehat{\varepsilon})),$$

où $tr(M)$ désigne la trace de la matrice M. De plus, en notant $H_{X^\perp} = I_n - H_X$ la projection sur l'orthogonal de X,

$$\mathbb{E}(tr(\widehat{\varepsilon}'\widehat{\varepsilon})) = \mathbb{E}(tr(\widehat{\varepsilon}\widehat{\varepsilon}')) = tr(\mathbb{E}(\widehat{\varepsilon}\widehat{\varepsilon}')) = tr(\mathbb{E}(H_{X^\perp}\varepsilon\varepsilon' H'_{X^\perp})) = \sigma^2 tr(H_{X^\perp}) = \sigma^2(n-p),$$

car la trace d'un projecteur est égale à la dimension de l'espace sur lequel est faite la projection. D'où la définition suivante de $\widehat{\sigma}^2$, un estimateur non biaisé de la variance :

Propriété 11. *L'estimateur $\widehat{\sigma}^2 = \frac{SCR(\widehat{\theta})}{n-p}$ est un estimateur sans biais de la variance.*

Toutes ces propriétés s'établissent sans nécessiter l'hypothèse gaussienne, mais celle-ci sera nécessaire pour obtenir la loi des estimateurs.

6.2.4 Estimation par maximum de vraisemblance

Lorsque la loi des observations est supposée gaussienne, la méthode d'estimation naturelle de $\beta = (\theta, \sigma^2)$ est celle du maximum de vraisemblance. La vraisemblance des n variables $\mathrm{Y} = (Y_1, \ldots, Y_n)$ est leur densité jointe, vue comme fonction du paramètre β et de Y.

Les variables Y_i étant indépendantes, la loi jointe de Y est le produit des lois des Y_i, ce qui donne sous l'hypothèse de bruit gaussien :

$$\begin{aligned} L_n(\theta, \sigma^2; \mathrm{Y}) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(Y_i - \mathrm{x}_i\theta)^2\right) \\ &= \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left(-\frac{1}{2\sigma^2}\sum_i (Y_i - \mathrm{x}_i\theta)^2\right) \\ &= \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left(-\frac{1}{2\sigma^2}||Y - X\theta||^2\right), \end{aligned}$$

On retrouve la quantité $SCR(\theta) = ||Y - X\theta||^2$, somme des carrés résiduels.

Le principe de l'estimation par maximum de vraisemblance du paramètre $\beta = (\theta, \sigma^2)$ inconnu est la recherche d'une valeur $\widehat{\beta} = (\widehat{\theta}, \widehat{s}^2)$ qui rend maximum la vraisemblance (donc le logarithme de la vraisemblance) pour un échantillon donné. $\widehat{\beta}$ dépend de l'échantillon Y, c'est une variable aléatoire. Il dépend donc bien évidement de n, mais nous ne l'avons pas mentionné dans la notation utilisée pour ne pas en alourdir l'écriture :

$$\widehat{\beta} = \arg \max_{\beta \in \mathbb{R}^p \times \mathbb{R}^{+*}} \log L_n(\beta; Y).$$

L'estimateur est cherché parmi les solutions des équations de vraisemblance, qui annulent la dérivée du logarithme de la vraisemblance par rapport au paramètre :

$$\partial \log L_n(\beta; Y)/\partial\beta = 0,$$

et telles que la dérivée seconde de la log-vraisemblance soit définie négative. Or,

$$\frac{\partial L_n(\theta, \sigma^2; Y)}{\partial\theta} = -\frac{1}{\sigma^2} SCR(\theta).$$

L'EMV de θ a donc la même expression que l'EMC, mais l'EMC ne nécessite pas la connaissance du type de la loi du bruit.

Théorème 22. *L'estimateur du maximum de vraisemblance $\widehat{\beta} = (\widehat{\theta}, \widehat{s}^2)$ du modèle linéaire*

$$Y = X\theta + \varepsilon, \;\; \mathbb{E}(\varepsilon) = 0, \;\; \mathrm{cov}(\varepsilon) = \sigma^2 I_n, \;\; \theta \in \mathbb{R}^p, \;\; \sigma \in \mathbb{R}^{+*},$$

supposé identifiable (X injective) et tel que $p \leq n$, est :

$$\widehat{\theta} = (X'X)^{-1}X'Y; \quad \widehat{s}^2 = \frac{1}{n}||Y - X\widehat{\theta}||^2. \tag{6.6}$$

Démonstration. Les équations de vraisemblance amènent immédiatement à la formulation des estimateurs :

$$\frac{\partial L_n(\theta, \sigma^2; Y)}{\partial\theta} = \frac{1}{\sigma^2}\left(X'X\theta - X'Y\right)$$
$$\frac{\partial L_n(\theta, \sigma^2; Y)}{\partial\sigma^2} = -\frac{n}{\sigma^2} + \frac{1}{\sigma^4}||Y - X\theta||^2.$$

On vérifie par ailleurs que la dérivée seconde de la log-vraisemblance est définie négative. ◊

6.3 Loi des estimateurs et intervalles de confiance

L'hypothèse de bruit gaussien permet d'obtenir les lois des estimateurs à distance finie (n fixé). Remarque : Il est possible de s'en affranchir pour l'étude asymptotique, mais c'est ce point n'est pas traité dans ce cours.

Théorème 23. *L'estimateur du maximum de vraisemblance $\widehat{\beta} = (\widehat{\theta}, \widehat{s}^2)$ du modèle linéaire* ***gaussien***

$$Y = X\theta + \varepsilon, \;\; \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n), \;\; \theta \in \mathbb{R}^p, \;\; \sigma \in \mathbb{R}^{+*},$$

supposé identifiable et tel que $p \leq n$, suit la loi suivante :

$$\widehat{\theta} \sim \mathcal{N}\left(\theta, \sigma^2(X'X)^{-1}\right); \quad \frac{n}{\sigma^2}\widehat{s}^2 = \frac{n-p}{\sigma^2}\widehat{\sigma}^2 \sim \chi^2(n-p). \tag{6.7}$$

De plus, $\widehat{\theta}$ et $\widehat{\sigma}^2$ sont indépendants.

Démonstration. D'après (6.6), $\widehat{\theta} = (X'X)^{-1}X'Y$ est une fonction linéaire de Y. Par hypothèse, Y est de loi gaussienne $Y \sim \mathcal{N}_n(X\theta, \sigma^2 I_n)$. La loi de $\widehat{\theta}$ est donc gaussienne, caractérisée par son espérance $\mathbb{E}(\widehat{\theta}) = \theta$ et sa variance $\mathrm{var}(\widehat{\theta}) = \sigma^2(X'X)^{-1}$, cf la propriété (10).
Par ailleurs, $\widehat{\theta} = (X'X)^{-1}X'Y = (X'X)^{-1}X'H_XY$ est fonction de $H_XY = \widehat{Y}$ la projection de Y sur $Im(X)$ tandis que $\widehat{\sigma}^2$ est fonction de $(I - H_X)Y$ orthogonal à $Im(X)$. Les vecteurs gaussiens $(I - H_X)Y$ et $\widehat{Y}$ sont décorrélés et donc indépendants. La loi de $\widehat{\sigma}^2$ est une conséquence du théorème de Cochran. ◊

Dans le cas de la régression, $||(I-H_X)Y|^2 = ||\widehat{\varepsilon}||^2 = SCR$ suit donc une loi du $\sigma^2\chi^2(n-p,0)$: en effet, puisque l'espérance de Y a été entièrement décrite par $X\theta$, l'espérance de la projection de Y sur l'orthogonal de X est nulle, et le paramètre de décentrage est nul.
Pour rappel, la loi du $\chi^2(d)$ centré est celle de la somme de d carrés de variables normales $Z_j \sim \mathcal{N}(0,1)$ indépendantes.

6.3.1 Autres lois utiles

Les lois de combinaisons linéaires estimables de θ sont bien évidemment gaussiennes si la variance σ^2 du bruit est connue. Dans les applications, le paramètre σ^2 n'est en général pas connu mais estimé par $\widehat{\sigma}^2$. Il est intéressant de s'intéresser aux lois renormalisées de ces combinaisons linéaires.

Théorème 24. *Soit L de dimension $1\times p$ définissant une forme linéaire $L\theta$ du paramètre θ d'un modèle linéaire gaussien identifiable. Si σ^2 est estimée par $\widehat{\sigma}^2$, la loi de $L\widehat{\theta}$ est définie par :*

$$\frac{L\widehat{\theta} - L\theta}{\sqrt{\widehat{\sigma}^2 L(X'X)^{-1}L'}} \sim \mathcal{T}(n-p). \tag{6.8}$$

$\mathcal{T}(n-p)$ est la loi de Student à $n-p$ degrés de liberté dont la définition est rappelée ci-après.

La preuve est immédiate. En particulier, la statistique T_j associée à la composante $\widehat{\theta}_j$, $j = 1,\dots,p$, suit la loi

$$T_j = \frac{\widehat{\theta}_j - \theta_j}{\widehat{\sigma}\sqrt{[(X'X)^{-1}]_{jj}}} \sim \mathcal{T}(n-p)$$

où$[(X'X)^{-1}]_{jj}$ désigne le $j$ème terme diagonal de la matrice $(X'X)^{-1}$.

Pour prendre en compte simultanément plusieurs formes linéaires de θ, considérons A, matrice de taille $q\times p$ de rang $r \le q \le p$. Alors la statistique de Wald suit une loi exacte, loi de Fisher à r et $n-p$ degrés de liberté. On l'appelle d'ailleurs la statistique de Fisher dans le modèle linéaire.

$$F_n = \frac{1}{r\widehat{\sigma}^2}(A(\widehat{\theta}-\theta))'[A(X'X)^{-1}A']^{-1}A(\widehat{\theta}-\theta) \sim \mathcal{F}(r,n-p). \tag{6.9}$$

En effet, dans l'expression de F_n, $A(X'X)^{-1}A'$ est de rang r et $A\widehat{\theta} \sim \mathcal{N}(A\theta, \sigma^2 A(X'X)^{-1}A')$, d'où le numérateur est un $\chi^2(r)$. La loi du dénominateur est celle de $\widehat{\sigma}^2$.

6.3.2 Intervalle de confiance d'une espérance

Les lois précédentes permettent de déduire la forme des intervalles de confiance.

Théorème 25. *Soit L une matrice de dimension $1\times p$. Un **intervalle de confiance** de niveau $1-\alpha$ d'une forme linéaire de $L\theta$ est donné par*

$$\left[L\widehat{\theta} - t_{n-p}(1-\alpha/2)\widehat{\sigma}\sqrt{L'(X'X)^{-1}L}\ ;\ L\widehat{\theta} + t_{n-p}(1-\alpha/2)\widehat{\sigma}\sqrt{L'(X'X)^{-1}L}\right], \tag{6.10}$$

où $t_{n-p}(1-\alpha/2)$ est le quantile d'ordre $1-\alpha/2$ de la loi de Student à $n-p$ degrés de liberté.

Ces intervalles de confiance sont en particulier utilisés pour estimer une composante θ_j du paramètre, ou pour estimer l'espérance d'une observation sous une condition $L = \mathrm{x}^*$ donnée.

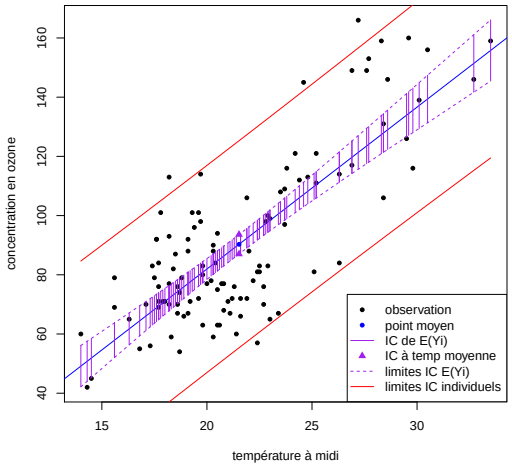


FIGURE 6.2 – Intervalles de confiance en moyenne ou en valeur individuelle du taux d'ozone pour chaque valeur de la température

Illustration Prenons l'exemple de la régression simple du taux d'ozone en fonction de la température à midi considéré en introduction. Les intervalles de confiance du taux d'ozone sont représentés sur la figure 6.2 en chaque valeur de température du jeu de données (barres verticales). L'équation (6.10) (utilisée avec $L_i = (1 \quad t_i)$, où t_i est la température de la i-ème observation) montre que les extrémités supérieures (resp.inférieures) sont placées sur une hyperbole. L'estimation est plus précise autour du point moyen qu'aux extrémités de l'intervalle d'observation.

6.3.3 Intervalle de confiance de la prévision d'une nouvelle observation

A partir des n observations, nous avons estimé $\widehat{\theta}$. Pour prédire une nouvelle observation Y_{n+1} sous la condition d'expérience $\mathrm{x}_{n+1} = (x_{n+1,1}, \ldots, x_{n+1,p})$, il est naturel d'utiliser le modèle

$$Y_{n+1} = \mathrm{x}_{n+1}\theta + \varepsilon_{n+1},$$

en substituant à θ son estimation $\widehat{\theta}$. L'observation Y_{n+1} est alors prédite par

$$\widehat{Y}^p_{n+1} = \mathrm{x}_{n+1}\widehat{\theta}.$$

L'erreur de prévision est

$$\widehat{\varepsilon}^p_{n+1} = Y_{n+1} - \widehat{Y}^p_{n+1},$$

d'espérance et variance :

$$\begin{aligned}
\mathbb{E}(Y_{n+1} - \widehat{Y}^p_{n+1}) &= 0 \\
\mathrm{var}(Y_{n+1} - \widehat{Y}^p_{n+1}) &= \mathrm{var}(\mathrm{x}_{n+1}(\widehat{\theta} - \theta) + \varepsilon_{n+1}) \\
&= x_{n+1}\,\mathrm{var}(\widehat{\theta} - \theta)\mathrm{x}'_{n+1} + \sigma^2 \\
&= \sigma^2[\mathrm{x}_{n+1}(X'X)^{-1}\mathrm{x}'_{n+1} + 1].
\end{aligned}$$

La loi de l'erreur de prévision est donc telle que

$$\frac{Y_{n+1} - \widehat{Y}^p_{n+1}}{\sigma\sqrt{\mathrm{x}_{n+1}(X'X)^{-1}\mathrm{x}'_{n+1} + 1}} \sim \mathcal{N}(0,1).$$

D'où l'intervalle de prévision d'une valeur individuelle de niveau $1-\alpha$:

$$[\mathrm{x}_{n+1}\widehat{\theta} \pm t_{n-p,1-\alpha/2}\widehat{\sigma}\sqrt{\mathrm{x}_{n+1}(X'X)^{-1}\mathrm{x}'_{n+1} + 1}]. \tag{6.11}$$

Illustration La figure 6.2 représente également les limites de chaque IC individuel, pour chaque température du domaine d'observation. Les ICs individuels sont plus larges que les ICs de la valeur moyenne conformément à (6.11).

6.3.4 Région de confiance

Une région de confiance étend la notion d'intervalle de confiance à $q < p$ combinaisons linéaires simultanées de θ.

Région de confiance de Fisher

C'est une région de confiance de Wald, pour laquelle on utilise la loi exacte à distance finie :

Théorème 26. *Soit A une matrice de dimension $q \times p$ et de rang q. Une **région de confiance** de Fisher de niveau $1-\alpha$ de $q \leq p$ combinaisons linéaires $A\theta$ est donnée par*

$$RC_\alpha(A\theta) = \{u \in \mathbb{R}^q, \frac{1}{q\widehat{\sigma}^2}(A\widehat{\theta} - u)'[A(X'X)^{-1}A']^{-1}(A\widehat{\theta} - u) \leq f_{q,n-p}(1-\alpha)\}$$

où $f_{q,n-p}(1-\alpha)$ est le quantile d'ordre $1-\alpha$ de la loi de Fisher à $(q, n-p)$ degrés de liberté.

Exemple Pour une région de confiance de (θ_1, θ_2), la matrice A est de dimension $2 \times p$ où $A_{11} = A_{22} = 1$, les autres termes étant nuls. D'où

$$RC_\alpha(\theta_1,\theta_2) = \{(\tilde{\theta}_1,\tilde{\theta}_2)' \in \mathbb{R}^2, \frac{1}{2\widehat{\sigma}^2}[\widehat{\theta}_1-\tilde{\theta}_1\ \widehat{\theta}_2-\tilde{\theta}_2][A(X'X)^{-1}A']^{-1}[\widehat{\theta}_1-\tilde{\theta}_1\ \widehat{\theta}_2-\tilde{\theta}_2]' \leq f_{2,n-p}(1-\alpha)\}$$

Si on note c_{ij} le terme général de $\widehat{\sigma}^2(X'X)^{-1}$, et $D = c_{11}c_{22} - c_{12}^2$, on obtient en développant

$$RC_\alpha(\theta_1,\theta_2) = \{(\tilde{\theta}_1,\tilde{\theta}_2); c_{22}(\widehat{\theta}_1-\tilde{\theta}_1)^2 - 2c_{12}(\widehat{\theta}_1-\tilde{\theta}_1)(\widehat{\theta}_2-\tilde{\theta}_2) + c_{11}(\widehat{\theta}_2-\tilde{\theta}_2)^2 \leq 2Df_{2,n-p}(1-\alpha)\}$$

Cette région de confiance est une ellipse qui tient compte de la corrélation entre $\widehat{\theta}_1$ et $\widehat{\theta}_2$, voir Figure 6.3.

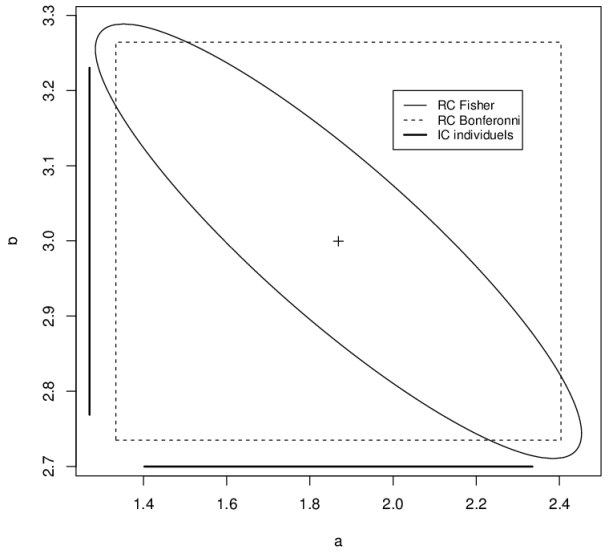
Intervalle de confiance de Bonferroni

La région de confiance de Bonferroni est le produit cartésien de deux intervalles de confiance dit de Bonferroni, obtenus en remarquant que :

$$P\left((\theta_1,\theta_2) \notin R(\theta_1,\theta_2)\right) \leq P(\theta_1 \notin IC_{1-\alpha/2}(\theta_1)) + P(\theta_2 \notin IC_{1-\alpha/2}(\theta_2)) \leq \alpha.$$

Deux intervalles de confiance de Bonferroni de risque $\alpha/2$ forment une région de confiance de risque simultané α. Sur la Figure 6.3, la région a la forme d'un rectangle. Si les composantes de l'estimateur ne sont pas fortement corrélées, alors la région définie par les intervalles de confiance de Bonferroni sont proches de la région de confiance de Wald. Sinon, on remarque sur la figure que la région de Wald est plus adaptée, parce que plus resserrée pour la même rique.

FIGURE 6.3 – Comparaison de la région de confiance de Wald (ellipse) et celle de Bonferroni (rectangle) de niveau α. Pour mémoire, les traits gras indique l'intervalle de confiance de risque α de chaque composante.



6.4 Tests dans le modèle linéaire gaussien

Les tests dans le modèle linéaire concernent ceux de l'espérance et de la variance.

6.4.1 Test de Student

Le test de Student est dédié au test d'une relation affine des composantes du paramètre $L\theta = c$, avec L matrice ligne de dimension p. La statistique naturelle est alors

$$T = \frac{L\widehat{\theta} - c}{\widehat{\sigma}\sqrt{L(X'X)^{-1}L'}} \sim_{H_0} \mathcal{T}(n-p),$$

et sa loi sous H_0 bien connue d'après (6.8). Pour le test bilatère $H_0 : L\theta = c$ contre $H_1 : L\theta \neq c$, la région de rejet du test de risque α est

$$\mathcal{R}_\alpha = \{|T| > t_{n-p,1-\alpha/2}\}.$$

Pour le test unilatéral $H_0 : L\theta = c$ contre $H_1 : L\theta < c$, la région de rejet du test de risque α est

$$\mathcal{R}_\alpha = \{T < t_{n-p,\alpha}\}.$$

Puissance Pour l'alternative $(H_1) : L\theta = c' < c$, T suit une loi de Student décentrée de facteur de non-centralité $(c'-c)/\sigma$. La puissance

$$\begin{aligned}\pi(c') &= \mathbb{P}_{L\theta=c'}\left(\frac{L\widehat{\theta} - c}{\widehat{\sigma}\sqrt{L(X'X)^{-1}L'}} < t_{n-p,\alpha}\right) = \mathbb{P}_{L\theta=c'}\left(\frac{(L\widehat{\theta} - c')/\sigma + (c'-c)/\sigma}{(\widehat{\sigma}/\sigma)\sqrt{L(X'X)^{-1}L'}} < t_{n-p,\alpha}\right) \\ &= \mathrm{F}_{\mathcal{T}(n-p,(c'-c)/\sigma)}(t_{n-p,\alpha}) > \pi(c) = \alpha\end{aligned}$$

augmente quand c' s'écarte de c. Le test est sans biais. La puissance du test n'est en général pas calculable puisqu'on ne connaît pas de façon exacte l'alternative.

Exemple Ce test est utilisé en particulier pour tester la significativité (non-nullité) d'une composante de θ. $(H_0) : \theta_j = 0$ contre $(H_1) : \theta_j \neq 0$. L est alors la forme linéaire qui ne contient que des 0 à l'exception d'un 1 pour la composante j à tester. Si on rejette (H_0), on décide que le coefficient est non nul, au risque d'erreur de première espèce α : la variable correspondante est utile dans la définition du modèle, on dit qu'elle est significative. Si on ne peut rejeter (H_0) qu'on conserve, on considèrera le coefficient nul avec un risque de seconde espèce, la variable n'apporte rien au modèle.

6.4.2 Test de Fisher

Le test de Fisher est le test d'un sous modèle linéaire du modèle (6.3) : $Y = m+\varepsilon, m \in V$, où V est un sous espace vectoriel de R^n qu'on notera dorénavant $V = \Omega$, de dimension $dim(\Omega) = p < n$. Soit ω un sous espace vectoriel de Ω tel que $dim(\omega) = q < p$: on dit que ω est **emboîté** dans Ω. On souhaite tester

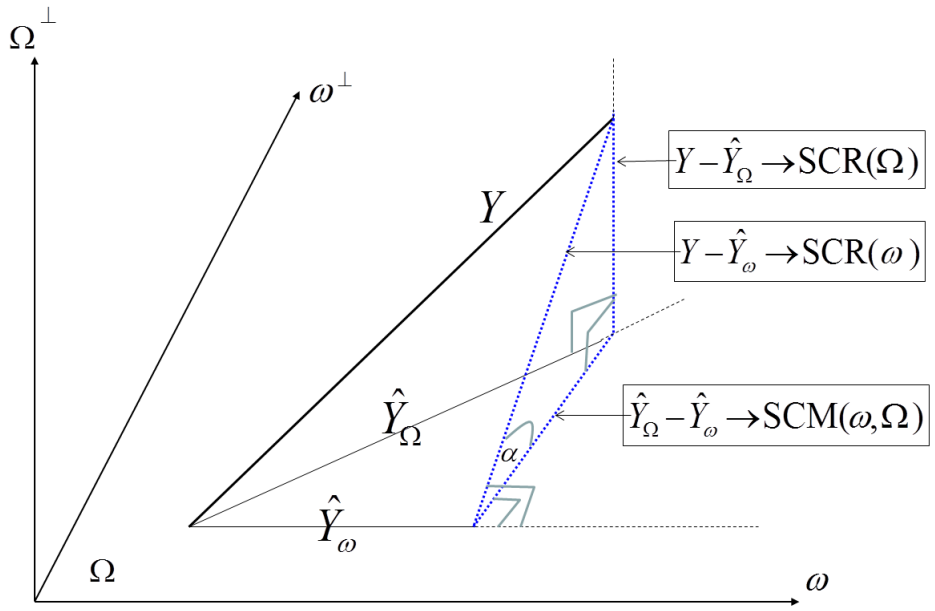
$$H_0 : m \in \omega \text{ contre } H_1 : m \in \Omega \backslash \omega.$$

Y se décompose de la façon suivante

$$\begin{aligned} Y &= (Y - H_\Omega Y) + (H_\Omega Y - H_\omega Y) + H_\omega Y \\ &= (Y - \widehat{Y}_\Omega) + (\widehat{Y}_\Omega - \widehat{Y}_\omega) + \widehat{Y}_\omega \end{aligned} \tag{6.12}$$

soit,

FIGURE 6.4 – Illustration des projections sur Ω et ω



$$Y - \widehat{Y}_\omega = (Y - \widehat{Y}_\Omega) + (\widehat{Y}_\Omega - \widehat{Y}_\omega).$$

Or, $\widehat{Y}_\Omega$ et $\widehat{Y}_\omega$ appartiennent à Ω, et sont donc orthogonaux à $Y - \widehat{Y}_\Omega$. En appliquant le théorème de Pythagore dans le triangle en pointillé Figure 6.4, il vient :

$$||Y - \widehat{Y}_\omega||^2 = ||\widehat{Y}_\Omega - \widehat{Y}_\omega||^2 + ||Y - \widehat{Y}_\Omega||^2$$
$$\mathrm{SCR}(\omega) = \mathrm{SCM}(\omega,\Omega) + \mathrm{SCR}(\Omega).$$

On reconnaît dans l'égalité précédente la somme des carrés résiduels $\mathrm{SCR}(\Omega)$ dans Ω, la somme des carrés résiduels $\mathrm{SCR}(\omega)$ dans ω, et un terme $\mathrm{SCM}(\omega, \Omega)$ appelé somme des carrés "Modèle", indiquant la perte d'ajustement résultant du choix de ω plutôt que Ω. La statistique utilisée pour le test est

$$F = \frac{(\mathrm{SCR}(\omega) - \mathrm{SCR}(\Omega))/(p-q)}{\mathrm{SCR}(\Omega)/(n-p)} = \frac{\mathrm{SCM}(\omega,\Omega)/(p-q)}{\mathrm{SCR}(\Omega)/(n-p)}. \tag{6.13}$$

dont le numérateur et le dénominateur sont des variables aléatoires indépendantes. La décomposition (6.12) en composantes orthogonales et le théorème de Cochran permettent de déduire que $\mathrm{SCR}(\Omega)/\sigma^2$ suit la loi $\chi^2(||(I_n - H_\Omega)m||^2, n-p)$, et $\mathrm{SCM}(\omega,\Omega)/\sigma^2$ la loi $\chi^2(||(H_\Omega - H_\omega)m||^2, p-q)$. Sous H_0, $H_\omega m = H_\Omega m = m$, et les lois du χ^2 centrées. F suit alors une loi de Fisher $\mathcal{F}(p-q, n-p)$. Sinon, la loi du numérateur est décentrée et F aura tendance à avoir des valeurs plus grandes que celles attendues : la modélisation Ω ajuste mieux que la modélisation ω. Le théorème suivant s'en déduit :

Théorème 27. *Soit le modèle linéaire gaussien $Y = m + \varepsilon, m \in \Omega$, $\varepsilon \sim \mathcal{N}(0, I_n)$, de dimension $dim(\Omega) < n$. Soit ω un sous espace vectoriel de Ω tel que $dim(\omega) = q < p$, et soit F la statstique de Fisher définie par (6.13). Le test de Fisher*

$$H_0 : m \in \omega \text{ contre } H_1 : m \in \Omega \backslash \omega.$$

de région de rejet

$$\mathcal{R}_\alpha = \{F > f_{p-q,n-p,1-\alpha}\},$$

où $f_{p-q,n-p,1-\alpha}$ est le quantile de la loi de Fisher $\mathcal{F}(p-q, n-p)$, est de niveau α.

Les applications sont nombreuses :
- Test de significativité globale de la régression : ω est le modèle i.i.d. et Ω le modèle de régression.
- Test de significativité d'une composante : $\omega = \{\theta | \theta_j = 0\}$. Le test de Fisher revient alors à effectuer un test bilatéral de Student de cette composante, car si $T \sim \mathcal{T}(n-p)$, alors $T^2 \sim \mathcal{F}(1, n-p)$.
- Test d'hypothèses linéaires simultanées $\omega = \{\theta | A\theta = 0\}$.

Puissance du test de Fisher La puissance en m du test de Fisher est $\mathbb{P}(F_\lambda > f_{p-q,n-p,1-\alpha})$, où F_λ est une variable de loi de Fisher décentrée à $p-q$ et $n-p$ degrés de liberté, de paramètre de non-centralité $\lambda = \|H_\Omega m - H_\omega m\|^2/\sigma^2$ (d'autant plus grand que m est éloigné de ω). On peut montrer que pour tout $t \in \mathbb{R}$, la fonction $\lambda \mapsto \mathbb{P}(F_\lambda > t)$ est croissante. La puissance du test de Fisher est donc une fonction croissante de la distance entre m et ω. Le test est sans biais.

La section suivante apporte une autre écriture du test de Fisher.

6.4.3 Test de Wald

Si le modèle emboîté ω est défini par des hypothèses affines (multidimensionnelles) $A\theta = c$, une minimisation sous contraintes amène à la relation

$$\widehat{\theta}_\omega = \widehat{\theta} + (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}(c - A\widehat{\theta}).$$

et

$$\mathrm{SCM}(\omega,\Omega) = (A\widehat{\theta}-c)'[A(X'X)^{-1}A']^{-1}(A\widehat{\theta}-c).$$

D'après (6.9), la statistique de Wald

$$W = \frac{1}{(p-q)\widehat{\sigma}^2}(A\widehat{\theta}-c)'[A(X'X)^{-1}A']^{-1}(A\widehat{\theta}-c))$$

suit sous H_0 une loi de Fisher $F(p-q,n-p)$, où $p-q$ est le rang de A. Cette statistique présente l'avantage de ne faire intervenir que l'estimation dans Ω, et c'est la statistique de Fisher si $c=0$.

Démonstration. Le lagrangien du système s'écrit

$$\mathcal{L}(\theta,\lambda) = ||Y-X\theta||^2 - \lambda'(A\theta-c)$$

d'où conditions de Lagrange :

$$\frac{\partial\mathcal{L}}{\partial\theta} = -2X'Y + 2X'X\widehat{\theta}_\omega - A'\widehat{\lambda} = 0 \tag{6.14}$$

$$\frac{\partial\mathcal{L}}{\partial\lambda} = A\widehat{\theta}_\omega - c = 0 \tag{6.15}$$

On multiple la première égalité par $A(X'X)^{-1}$:

$$\begin{aligned}
-2A(X'X)^{-1}X'Y + 2A(X'X)^{-1}X'X\widehat{\theta}_\omega - A(X'X)^{-1}A'\widehat{\lambda} &= 0\\
-2A(X'X)^{-1}X'Y + 2A\widehat{\theta}_\omega - A(X'X)^{-1}A'\widehat{\lambda} &= 0\\
-2A(X'X)^{-1}X'Y + 2c - A(X'X)^{-1}A'\widehat{\lambda} &= 0
\end{aligned}$$

puis on replace $\widehat{\lambda}$ dans 6.14

$$-2X'Y + 2X'X\widehat{\theta}_\omega - A'[A(X'X)^{-1}A']^{-1}(2c - A(X'X)^{-1}X'Y) = 0$$

et on calcule $\widehat{\theta}_c$

$$\begin{aligned}
\widehat{\theta}_\omega &= (X'X)^{-1}[X'Y + A'[A(X'X)^{-1}A']^{-1}(c - A(X'X)^{-1}X'Y)]\\
&= \widehat{\theta} + (X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}(c - A\widehat{\theta})
\end{aligned}$$

Le calcul de la somme des carrés modèle s'en déduit

$$\mathrm{SCM}(\omega,\Omega) = ||X\widehat{\theta} - X\widehat{\theta}_\omega||^2 = ||X(X'X)^{-1}A'[A(X'X)^{-1}A']^{-1}(c - A\widehat{\theta})||^2$$

◊

6.4.4 Test de la variance

Nous finissons cette présentation des tests par celui du paramètre σ^2. Soit $\sigma_0 \in \mathbb{R}^{+*}$ une valeur fixée de la variance. Le test de $H_0 : \sigma = \sigma_0$ contre $H_1 : \sigma > \sigma_0$, de région de rejet

$$\mathcal{R}_\alpha = \{\widehat{\sigma}^2 > \sigma_0^2 k_{n-p;1-\alpha}/(n-p)\}$$

où $k_{n-p;1-\alpha}$ est le quantile d'une loi $\chi^2(n-p)$, est de niveau α. Cette propriété se déduit immédiatement de la loi (6.7) de $\widehat{\sigma}^2$.

6.4.5 Application : tableau d'analyse de la variance

Soit Ω un modèle linéaire contenant un intercept et soit ω le modèle i.i.d. : on a $\omega \subset \Omega$. Les procédures logicielles d'estimation en régression linéaire fournissent en général la table suivante, ou une table de principe équivalent :

Source	DF	Sum of Squares	Mean Square	F value	Prob>F
Model	$p-1$	$\|P_\Omega Y - P_\omega Y\|^2 = SCM(\omega,\Omega)$	$\frac{SCM}{p-1}$	$\frac{SCM/(p-1)}{SCR(\Omega)/(n-p)}$	$P(Z>F)$
Error	$n-p$	$\|Y - P_\Omega Y\|^2 = SCR(\Omega)$	$\frac{SCR(\Omega)}{n-p}$		
Total	$n-1$	$\|Y - P_\omega Y\|^2 = SCR(\omega)$	$\frac{SCR(\omega)}{n-1}$		

où Z est une variable aléatoire de loi de Fisher à $p-1$ et $n-p$ degrés de liberté. $P(Z>F)$ est donc la p-value du test de Fisher de l'hypothèse échantillon i.i.d. gaussien (test de significativité globale).

6.4.6 Interprétation des traces d'un logiciel

Les logiciels statistiques effectuent les calculs précédents, comme par exemple la fonction `lm` de `R`, dédiée à l'estimation d'un modèle linéaire. Voici un exemple de trace d'une régression multiple d'une variable quantitative à expliquer en fonction de trois variables qualitatives X_1, X_2, X_3.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -20.3129    51.8246  -0.392    0.701
X1           -3.2241     3.6100  -0.893    0.386
X2            0.2334     0.2768   0.843    0.412
X3           20.3895     1.2662  16.103  7.1e-11 ***
---
Residual standard error: 4.163 on 15 degrees of freedom
F-statistic: 584.6 on 3 and 15 DF,  p-value: 9.386e-16
```

Le tableau de l'estimation des coefficients fait apparaître quatre lignes, l'une pour l'estimation du coefficient de l'intercept, les trois autres pour les coefficients des variables explicatives. La colonne `estimate` affiche les valeurs estimées $\widehat{\theta}_j$, $j = 0, \ldots, 3$ des quatre coefficients, `Std. Error` est l'écart type estimé $(\mathrm{var}(\widehat{\theta}_j))^{1/2}$, `t value` est la statistique du test de Student de significativité du coefficient (ie $\theta_j = 0$ contre $\theta_j \neq 0$), et `Pr(>|t|)` sa p-value. L'erreur standard résiduelle est l'estimation de $\widehat{\sigma}$. `F-statistic` est la valeur de la statistique de Fisher du test de significativité globale de la régression : $\theta_1 = \theta_2 = \theta_3 = 0$ contre l'un des paramètres $\theta_1$ou θ_2 ou θ_3 n'est pas nul. Le nombre de degrés de liberté du numérateur est $4 - 1 = 3$, celui du dénominateur est $n - 4 = 15$ où n est le nombre d'observations du jeu de données.

La fonction `anova` dresse des tableaux d'analyse de la variance, ici du modèle iid ($\theta_1 = \theta_2 = \theta_3 = 0$) contre le modèle d'étude à trois variables explicatives (et un intercept)

```
Model 1: y ~ 1
Model 2: y ~ 1+X1+X2+X3
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
```

```
1     18 30654
2     15   260  3     30394 584.55 9.386e-16 ***
```

On y reconnaît les informations suivantes, avec ω =`Model 1` (ici, le modèle iid avec un intercept uniquement, $q = 1$), Ω =`Model 2` (ici, le modèle d'étude avec toutes les variables $p = 4$). On retrouve les résultats du test de significativité globale de la régression

Source	Res.Df	RSS	Df	Sum of Sq	F	Prob>F
ω	$n-q$	$SCR(\omega)$				
Ω	$n-p$	$SCR(\Omega)$	$p-q$	$SCM(\omega,\Omega)$	$f_{obs} = \frac{SCM/(p-q)}{SCR(\Omega)/(n-p)}$	$\mathbb{P}(F > f_{obs})$

6.5 Validation du modèle

Une fois l'estimation réalisée, il est important de vérifier que le modèle représente correctement les données observées, et en particulier que les hypothèses de modélisation ne sont pas mises en défaut : cela passe par l'étude graphique de l'ajustement, des résidus, la prise en compte de critères, l'étude de l'influence de certaines variables et de certaines observations. Bien évidemment, des test statistiques peuvent être menés.

6.5.1 Tests statistiques

- Test de significativité globale : test de Fisher du modèle iid contre le modèle d'étude, voir son lien avec le R^2, en Section 6.5.4.
- Test d'adéquation : Il s'agit de comparer le modèle d'étude Ω à un surmodèle Ω_E peu contestable qui contient Ω. On teste alors (Ω) contre (Ω_E). Considérons par exemple un modèle ANCOVA avec une variable qualitative à I niveaux et une variable qualitative. Le modèle est de dimension $2I$. Supposons maintenant que la variable quantitative soit observée sur J niveaux, avec des données répétées. Un modèle alternatif Ω_E est le modèle ANOVA(2)
- Test d'ajout ou de suppression de régresseurs : tant que les régresseurs sont peu nombreux, tous les cas de figure peuvent s'envisager. Dès qu'ils le sont plus, des stratégies de sélection de variable doivent être mises en place.
- Test sur l'hypothèse gaussienne : test de Kolmogorov-Smirnov ou test de Shapiro-Wilks.

6.5.2 Résidus

L'étude des résidus est un outil de vérification. Les résidus (estimés) sont définis par la relation

$$\widehat{\varepsilon} = Y - \widehat{Y} = Y - X\widehat{\theta} = (I - H_X)Y = H_{X^\perp}Y = H_{X^\perp}\varepsilon.$$

Ils appartiennent donc à $Im(X)^\perp$, l'orthogonal de X, appelé espace des résidus. Les résidus sont donc toujours orthogonaux à $\widehat{Y}$ et forment un sous espace vectoriel de dimension $n - p$.

Propriété 12. *Dans le modèle linéaire gaussien (6.2), les résidus $\widehat{\varepsilon} = Y - \widehat{Y}$ possèdent les propriétés suivantes :*

- *centrés :*
$$E(\widehat{\varepsilon}) = H_{X^\perp}E(\varepsilon) = 0$$
- *hétéroscédastiques :*
$$\text{var}(\widehat{\varepsilon}) = H_{X^\perp}\sigma^2 I_n H'_{X^\perp} = \sigma^2 H_{X^\perp} = \sigma^2(I - X(X'X)^{-1}X') = \sigma^2(I - H)$$

— *décorrélés avec les valeurs estimées :* $\text{cov}(\widehat{Y}, \widehat{\varepsilon}) = 0$.
— *Si* $\mathbb{1} \in Im(X)$*, les résidus estimés sont linéairement dépendants.*

La dernière propriété découle du fait que $\widehat{\varepsilon} \perp \mathbb{1}$, donc $\sum_i \widehat{\varepsilon}_i = 0$.

Pour utiliser des résidus estimés ayant les propriétés analogues à celles du bruit dont ils sont issus, il faut que :
— les éléments non diagonaux de H_X soient suffisamment petits,
— les éléments diagonaux h_{ii} de H_X soient approximativement égaux. Une valeur élevée de h_{ii} indique une condition d'expérience, x_i isolée : l'observation sera influente dans l'estimation de θ (effet levier) ; si les données sont bien réparties, les h_{ii} sont à peu près égaux à p/n.

Les **résidus normalisés** éliminent l'hétéroscédasticité

$$r_i = \frac{\widehat{\varepsilon_i}}{\sigma\sqrt{1-h_{ii}}}.$$

L'estimation de σ inconnu par $\widehat{\sigma}$ permet obtenir les **résidus standardisés**, parfois également appelés **studentisés** : ils possèdent une variance unité permettant la détection de valeurs importantes : c'est la règle empirique d'appartenance de 95% des résidus à l'intervalle $[-2, 2]$.

$$t_i = \frac{\widehat{\varepsilon_i}}{\widehat{\sigma}\sqrt{1-h_{ii}}}.$$

Les résidus **studentisés par validation croisée** sont définis par

$$t_i^* = \frac{\widehat{\varepsilon_i}}{\widehat{\sigma}_{(i)}\sqrt{1-h_{ii}}}. \tag{6.16}$$

où $\widehat{\sigma}_{(i)}$ est l'estimée de σ basée sur des observations privées de l'observation i. Cette définition est particulièrement intéressante dans le cadre du modèle gaussien, puisqu'on peut alors déterminer leur loi.

Théorème 28. *Si la matrice X est de rang plein, si $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, et si la suppression de la ligne i ne modifie pas le rang de la matrice X, alors les résidus studentisés par validation croisée définis par (6.16) suivent une loi de Student à $n - p - 1$ degrés de liberté.*

La loi de Student a un degré de liberté de moins, puisqu'il y a une observation de moins. Les résidus sont donc homoscédastiques, et dès que le nombre d'observations est supérieur à 30, la règle empirique d'appartenance de 95% des résidus à l'intervalle $[-2, 2]$ peut s'appliquer.

6.5.3 Représentations graphiques

Différents types de graphiques permettent une étude visuelle :
— valeurs estimées en fonction de valeurs observées,
— résidus en fonction des valeurs estimées : pour tester la non-corrélation,
— résidus en fonction d'une valeur de covariable.
— résidu de l'observation $i + 1$ en fonction de celui de l'observation i pour détecter des corrélations
— diagramme quantile/quantile des résidus (droite de Henry) pour étudier le caractère gaussien de la loi, qui peut également se tester statistiquement.

Les graphes des résidus permettent de détecter des points aberrants, des biais d'estimation, ou une hétéroscédasticité. La figure (6.5) montre quelques exemples de graphes, qui ne présentent pas ici de problème particulier.

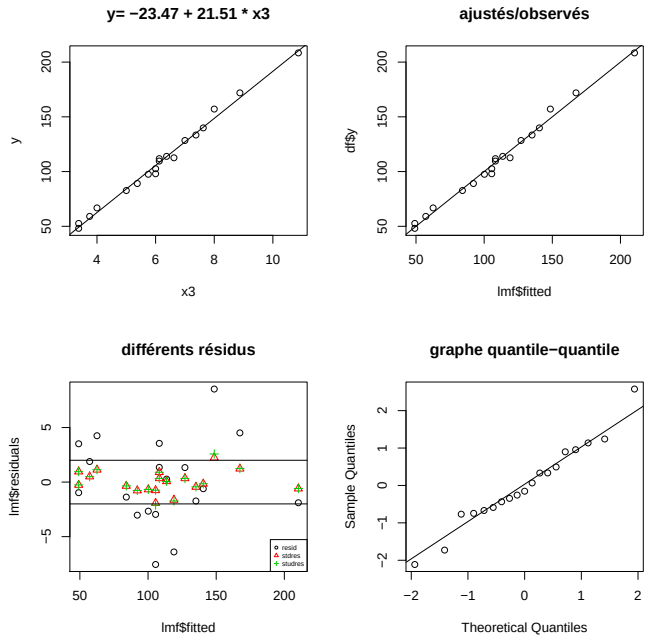


FIGURE 6.5 – Diagnostics visuels. En haut à gauche : les observations en fonction de la troisième variable X_3. En haut à droite : graphe des ajustés/observés ; en bas à gauche : la représentation de divers résidus (*resid*= résidus bruts, *stdres*= résidus studentisés, *stdres*= résidus studentisés par validation croisée) ; en bas à droite : la droite de Henry quantile théorique/quantile empirique

6.5.4 Le coefficient de détermination multiple R^2

C'est un indicateur global de la qualité de l'ajustement.

Définition 34. *Soit $\Omega = Im(X)$ un modèle contenant le régresseur $\mathbb{1}$. Le coefficient de détermination (multiple) est défini par*

$$R^2 = \frac{||\widehat{Y} - \overline{Y}\mathbb{1}||^2}{||Y - \overline{Y}\mathbb{1}||^2} = \frac{SCM(\Omega)}{SCR(\mathbb{1})} = \cos^2 \alpha. \tag{6.17}$$

Dans la décomposition (6.12), choisissons pour ω le modèle i.i.d (toutes les coordonnées de m sont égales). Le coefficient de détermination mesure le cosinus carré de l'angle entre le vecteur $\widehat{Y} - \overline{Y}$ (où $\overline{Y} = \widehat{Y}_\omega$) et le vecteur des résidus $Y - \overline{Y}\mathbb{1}$ dans le modèle i.i.d., cf Figure 6.4. En effet, si $\mathbb{1} \in Im(X)$,

$$R = \frac{||\widehat{Y} - \overline{Y}\mathbb{1}||}{||Y - \overline{Y}\mathbb{1}||} = \frac{||\widehat{Y} - \overline{Y}\mathbb{1}||^2}{||\widehat{Y} - \overline{Y}\mathbb{1}||||Y - \overline{Y}\mathbb{1}||} = \frac{< \widehat{Y} - \overline{Y}\mathbb{1}, \widehat{Y} - Y + Y - \overline{Y}\mathbb{1} >}{||\widehat{Y} - \overline{Y}\mathbb{1}||||Y - \overline{Y}\mathbb{1}||} = \frac{< \widehat{Y} - \overline{Y}\mathbb{1}, Y - \overline{Y}\mathbb{1} >}{||\widehat{Y} - \overline{Y}\mathbb{1}||||Y - \overline{Y}\mathbb{1}||}$$

Le **coefficient de corrélation** R s'interprète comme la **part de variance expliquée** par les régresseurs supplémentaires. C'est dans ce cas la corrélation empirique entre les données observées et les données prédites : $R = 1 \Leftrightarrow \widehat{Y} = Y$, $R = 0 \Leftrightarrow \widehat{Y} = \overline{Y}$.

Bien sûr, le coefficient de corrélation empirique ne donne d'indication que sur la linéarité de la relation, et il faut faire attention à son interprétation : on ne peut pas espérer un grand R^2 si les observations sont très nombreuses d'une part, et le R^2 est d'autant plus grand que le modèle comporte beaucoup de paramètres d'autre part. Utiliser un R^2 ajusté peut pondérer ce phénomène

$$R_a^2 = 1 - \frac{n-1}{n-p}\frac{||\widehat{\varepsilon}||^2}{||Y - \bar{y}\mathbb{1}||^2}.$$

On peut remarquer qu'il existe un lien entre le R^2 et la statistique du test de significativité globale de la régression : modèle iid contre modèle Ω. En effet, si on note R^2_{obs} la valeur observé du R^2 sur les données ajustée, la p-value de ce test est

$$\mathbb{P}\left(F > \frac{R^2_{obs}/(p-1)}{(1-R^2_{obs})/(n-1)}\right),$$

où $F \sim \mathcal{F}(p-1, n-p)$. Si n et p sont fixés, la p-value est donc une fonction décroissante de R^2_{obs}. Si R^2_{obs} est petit, alors la p-value est grande et on ne rejette pas le modèle iid contre le modèle d'étude, qui n'est donc pas significativement explicatif.

Mais attention, le R^2 est un **critère**, c'est une lecture brute de la qualité de l'ajustement, et la comparaison entre deux modèles via le R^2 n'est pas un test (puisqu'on ne calibre pas le risque de la décision). En effet, plus le modèle comporte de variables, meilleur sera le R^2. La comparaison des R^2 de deux modèles ne suffit donc pas pour choisir entre ces deux modèles.

6.5.5 Multicolinéarité

Certaines covariables peuvent être très corrélées, et causent alors une estimation imprécise des paramètres. En effet, on peut montrer que la variance de $\widehat{\theta}_j$ peut s'écrire

$$\text{var}(\widehat{\theta}_j) = \frac{\sigma^2}{(1-R_j^2)\sum_{i=1}^n (x_{ij} - \overline{X_j})},$$

où R_j^2 est le coefficient de détermination entre la variable X_j prise comme variable réponse, et les autres colonnes de X prises comme régresseurs. La variance de $\widehat{\theta}_j$ est donc minimisée quand $R_j = 0$, c'est à dire quand les covariables sont décorrélées : un tel choix de matrice de plan d'expérience est appelé **design orthogonal**. Dans le cas inverse, quand $R_j \to 1$, la variance explose vers l'infini. La formule de la variance produit aussi un outil de diagnostic pour mesurer le degré de colinéarité (**variance influence factor**)

$$VIF_j = \frac{1}{1-R_j^2}$$

qui est d'autant plus grand que le problème de colinéarité est sérieux. En général, on considère qu'il existe un problème de colinéarité quand $VIF_j > 10$. Pour résoudre ce problème, il existe différentes stratégies :

- omettre l'une des deux variables,
- construire a variable combinée (si possible facilement interprétable) à partir des variables en cause,
- utiliser une régression **Ridge**,
- utiliser une régression sur les composantes principales de X.

6.5.6 Influence d'une observation

Données influentes

Certaines observations ont un poids important dans la définition d'une ou plusieurs composantes de l'estimateur. Ce sont des valeurs influentes ou **points leviers**.

$$\widehat{y}_i = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j$$

h_{ii} représente le poids de l'observation sur son propre ajustement. On montre que si $h_{ii} = 1$ alors les h_{ij} de la même ligne sont nuls, donc $\widehat{y}_i$ est entièrement déterminé par y_i.
Comme $tr(H_X) = \sum h_{ii} = p$, la moyenne des h_{ii} vaut p/n. Si h_{ii} est grand, y_i influe fortement sur $\widehat{y}_i$. Un point levier est un point dont la ligne dans X est éloignée du centre de gravité des lignes de X.
Si $h_{ii} = 0$, alors $h_{ij} = 0$ pour tout j différent de i, y_i n'a pas d'influence sur $\widehat{y}_i$ qui vaut alors 0.

Valeurs aberrantes

Une donnée aberrante est un point (x_i, y_i) pour lequel la valeur associée à t_i^* est élevée (comparée au seuil donné par la loi de Student) :

$$|t_i^*| > t_{n-p-1}(1 - \alpha/2)$$

Un point peut être levier mais non aberrant, ou aberrant et non levier, ou levier et aberrant.

Distance de Cook

Elle mesure l'influence de l'observation i sur l'estimation du paramètre β

$$C_i = \frac{1}{p\widehat{\sigma}^2}(\widehat{\beta}_{(i)} - \widehat{\beta})'(X'X)^{-1}(\widehat{\beta}_{(i)} - \widehat{\beta}) = \frac{(\widehat{y}_i - x_i\widehat{\beta}_{(i)})^2}{p\widehat{\sigma}^2} = \frac{h_{ii}}{p(1-h_{ii})}t_i^2 = \frac{h_{ii}}{p(1-h_{ii})^2}\frac{\widehat{\varepsilon}_i^2}{\widehat{\sigma}^2}$$

où $\widehat{\beta}_{(i)}$ est l'estimation de β sans prendre en compte l'observation i.

- Une observation influente est donc une observation qui, enlevée, conduit à une grande variation dans l'estimation des coefficients, c'est à dire à une distance de Cook élevée.
- Le terme t_i^2 mesure le degré d'adéquation de l'observation y_i au modèle estimé $x_i\widehat{\beta}$, le second est le rapport $\text{var}(\widehat{y}_i)/\text{var}(\widehat{\varepsilon}_i) = h_{ii}/(1-h_{ii})$
- Pour juger si la distance de Cook est élevée, Cook propose le seuil $f_{n,n-p}(0.1)$ comme souhaitable et $f_{n,n-p}(0.5)$ comme préoccupant.

Les points présentant une distance de Cook élevée seront des points leviers, des points aberrants ou les deux et influenceront l'estimation puisque la distance de Cook est une distance entre $\widehat{\beta}$ et $\widehat{\beta}_{(i)}$.

6.6 Cas des variables explicatives qualitatives

Nous avons vu comment insérer les covariables explicatives quantitatives dans la matrice du plan d'expérience. Mais qu'en est-il d'une variable qualitative ? Ce point est important : on peut souhaiter expliquer le rendement de semences de maïs en fonction de leur type, un taux d'hémoglobine en fonction du médicament utilisé pour le traitement, le poids d'une portée en fonction de son génotype, ...

Définition 35. *Une variable* ***qualitative*** *(ou* ***facteur****) ne peut prendre qu'un nombre discret de valeurs appelées* ***niveaux****. Quand les niveaux peuvent être ordonnés, le facteur est dit* ***ordinal****. Sinon, on parle de facteur* ***nominal****.*

Nous présentons dans cette section trois modèles à facteurs explicatifs
- ANOVA(1) : un facteur explicatif qualitatif. Par exemple expliquer le poids des portées par le génotype de leur mère, figure 6.6
- ANCOVA : un facteur explicatif qualitatif et une variable explicative quantitative. Par exemple, expliquer le poids des portées par le génotype et le poids de la mère
- ANOVA(2) : deux facteurs explicatifs qualitatifs. Par exemple, expliquer le poids des portées par le génotype de la mère et le génotype de la portée

Le terme analyse de variance vient de la décomposition de la variance totale en variance résiduelle (résultant de la variance à l'intérieur des groupes), et variance entre les groupes (captée par le modèle).

6.6.1 ANOVA1 : Analyse de la variance à un facteur

Les niveaux sont souvent codés avec des chaînes de caractères, il est donc nécessaire de les transformer sous une forme numérique.

Exemple genotype Par exemple, le génotype est un facteur à quatre niveaux codés A, B, I, J dans le jeu de données `genotype` de la librairie `MASS` de R.

Soit I le nombre de niveaux du facteur explicatif. On peut envisager d'associer un entier i à chaque niveau ($i = 1, \dots, I$) pour rendre la variable numérique. Une régression linéaire pourrait être opérée entre la variable réponse et la variable transformée devenue artificiellement "quantitative". Mais cette régression n'a pas de sens, en particulier quand la variable est nominale. Ce qui est plus intéressant, c'est d'estimer les réponses pour chaque niveau du facteur, c'est à dire d'écrire la régression sous la forme

$$Y_{ik} = \alpha_i + \varepsilon_{ik}, \;\; i = 1, \dots, I; \;\; k = 1, \dots, n_i; \;\; n = \sum_{i=1}^{I} n_i; \;\; \varepsilon \sim \mathcal{N}(0, \sigma^2 Id_n) \tag{6.18}$$

où Y_{ik} est l'observation du k-ième individu du groupe de niveau i, groupe issu d'un échantillonnage iid d'une loi d'espérance α_i pour la réponse. Le bruit ε est iid centré de variance indépendante du niveau du facteur. Si les individus sont classés par niveau du facteur explicatif, le vecteur des observations est le vecteur colonne $Y = (Y_{11}, \dots, Y_{1n_1}, \dots, Y_{I1}, \dots, Y_{In_I})'$, le paramètre est le vecteur colonne $\theta = (\alpha_1, \dots, \alpha_I)'$, et la matrice du plan d'expérience s'écrit

$$X = \begin{pmatrix} \mathbb{1}_{n_1} & & \\ & \ddots & \\ & & \mathbb{1}_{n_I} \end{pmatrix}$$

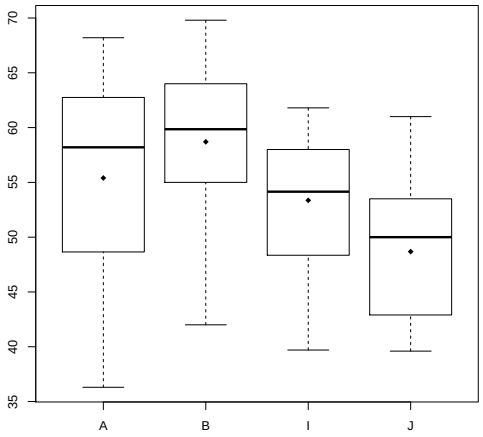


FIGURE 6.6 – Boîtes à moustaches des poids des portées en fonction du génotype de la mère. Le point représente le poids moyen du groupe

avec $\mathbb{1}_{n_i}$ un vecteur colonne de 1 de longueur n_i. On retrouve la définition d'un modèle linéaire (gaussien avec l'hypothèse gaussienne)

$$Y = X\theta + \varepsilon; \;\; \varepsilon | X \sim \mathcal{N}(0, \sigma^2 Id_n)$$

Ainsi, un facteur à I niveaux observé sur n individus est remplacé par I variables, indicatrices du niveau. On note que $Im(X)$ contient le vecteur $\mathbb{1}_n$, mais il n'apparait pas explicitement comme une des colonnes de la matrice du plan d'expérience. Or, on préfère en général qu'il y figure explicitement. On écrit alors

$$Y_{ik} = \mu + \alpha_i + \varepsilon_{ik}, \;\; i = 1, \ldots, I; \;\; k = 1, \ldots, n_i; \;\; n = \sum_{i=1}^{I} n_i. \tag{6.19}$$

Le paramètre est le vecteur colonne $\theta = (\mu, \alpha_1, \ldots, \alpha_I)'$ de dimension $I+1$, et la matrice du plan d'expérience s'écrit

$$X = \begin{pmatrix} \mathbb{1}_{n_1} & \mathbb{1}_{n_1} & & \\ \vdots & & \ddots & \\ \mathbb{1}_{n_I} & & & \mathbb{1}_{n_I} \end{pmatrix},$$

L'ajout de l'intercept a rendu le modèle non-identifiable, X n'est plus injective, de noyau engendré par le vecteur de $(1, -1, \ldots, -1)' \in \mathbb{R}^{I+1}$: par exemple, les paramètres $\tilde{\mu} = \mu + 1$ et $\tilde{\alpha}_i = \alpha_i - 1$ induisent la même loi des observations. Pour retrouver un modèle régulier, il faut restreindre l'espace du paramètre à un espace de dimension I. Ceci est fait en posant une condition d'identifiabilité (CI), ie, une condition linéaire sur les composantes de θ de la forme $L\theta = 0$, de façon que la restriction de X au sous-espace de $\mathbb{R}^{K+1}$ défini par la condition soit injective. On retrouve le modèle (6.18) si on choisit $\mu = 0$, mais ce n'est pas cette condition qui est en général retenue.

Exemples Voici quelques exemples de CI utilisées en ANOVA1 :

— Le premier niveau $\alpha_1 = 0$ est mis à zéro, soit $L = (0\ 1\ 0 \dots 0)$. C'est la CI par défaut du logiciel R, exemple pour $I = 4$

$$\begin{array}{lllll} \mathbb{E}(Y_{1k}) = m_1 = & \mu & & & \\ \mathbb{E}(Y_{2k}) = m_2 = & \mu & +\alpha_2 & & \\ \mathbb{E}(Y_{3k}) = m_3 = & \mu & & +\alpha_3 & \\ \mathbb{E}(Y_{4k}) = m_4 = & \mu & & & +\alpha_4 \end{array}$$

$\mu = m_1$ représente la valeur de la réponse pour le niveau 1,
$\alpha_i = m_i - m_1$ représente la différence de réponse entre le niveau j et le niveau 1 pris comme référence.

— Le dernier niveau $\alpha_I = 0$ est mis à zéro, soit $L = (0\ 0 \dots 0\ 1)$. C'est la CI par défaut du logiciel SAS.

$$\begin{array}{lllll} \mathbb{E}(Y_{1k}) = m_1 = & \mu & +\alpha_1 & & \\ \mathbb{E}(Y_{2k}) = m_2 = & \mu & & +\alpha_2 & \\ \mathbb{E}(Y_{3k}) = m_3 = & \mu & & & +\alpha_3 \\ \mathbb{E}(Y_{4k}) = m_4 = & \mu & & & \end{array}$$

$\mu = m_4$ représente la valeur de la réponse pour le niveau I,
$\alpha_i = m_i - m_4$ représente la différence de réponse entre le niveau i et le niveau I pris comme référence.

— $\sum_i n_i \alpha_i = 0$, soit $L = (0\ n_1\ \dots n_I)$:

$$\begin{array}{lllll} \mathbb{E}(Y_{1k}) = m_1 = & \mu & +\alpha_1 & & \\ \mathbb{E}(Y_{2k}) = m_2 = & \mu & & +\alpha_2 & \\ \mathbb{E}(Y_{3k}) = m_3 = & \mu & & & +\alpha_3 \\ \mathbb{E}(Y_{4k}) = m_4 = & \mu & -\frac{n_1}{n_4}\alpha_1 & -\frac{n_2}{n_4}\alpha_2 & -\frac{n_3}{n}\alpha_3 \end{array}$$

$\mu = (n_1 m_1 + n_2 m_2 + n_3 m_3 + n_4 m_4)/n$ représente la moyenne sur l'ensemble des observations,
a_i représente la différence de réponse entre le niveau i et la moyenne des observations.

— $\sum_i \alpha_i = 0$, soit $L = (0\ 1\ 1 \dots 1)$:

$$\begin{array}{lllll} \mathbb{E}(Y_{1k}) = m_1 = & \mu & +\alpha_1 & & \\ \mathbb{E}(Y_{2k}) = m_2 = & \mu & & +\alpha_2 & \\ \mathbb{E}(Y_{3k}) = m_3 = & \mu & & & +\alpha_3 \\ \mathbb{E}(Y_{4k}) = m_4 = & \mu & -\alpha_1 & -\alpha_2 & -\alpha_3 \end{array}$$

$\mu = (m_1 + m_2 + m_3 + m_4)/4$ représente la moyenne de la réponse sur l'ensemble des niveaux,
α_i représente la différence de réponse entre le niveau i et la moyenne de moyennes de chaque niveau. Lorsque le plan est équilibré $n_1 = \dots = n_4 = n/I$, cette contrainte est égale à la précédente.

— Le supplémentaire orthogonal est généré avec $L = (-1\ 1 \dots 1)$.

Certaines formes linéaires L ne permettent pas de restreindre suffisamment l'espace du paramètre. Par exemple, $L = (1, -1, 0, 0)$ dans le cas $I = 4$ ne fonctionne pas. On verra à la section 6.6.4 que toute CI doit vérifier $Ker(L) \cap Ker(X) = \{0\}$ pour engendrer une restriction correcte.

Dimension du modèle Dans (6.19), la dimension de θ est $I + 1$, mais pour l'estimation, on restreint à un espace de dimension $dim(Im(X)) = I$ par une CI qui fixe un degré de liberté. La dimension du modèle, c'est la dimension de $Im(X)$, pas la dimension initiale de θ.

Estimation La CI $L\theta = 0$ permet de retrouver une situation régulière dans un sous-espace restreint, l'estimateur $\widehat{\theta}$ est unique sous cette condition et s'écrit (cf section 6.6.4) :

$$\widehat{\theta} = (X'X + C'C)^{-1}X'Y = (X'X)^{-}X'Y$$

ce qui revient à utiliser un inverse généralisé Q^{-} (tel que $QQ^{-}Q = Q$) de $Q = X'X$. Ainsi, tout ce qui a été fait auparavant dans le cas de la régression linéaire multiple peut être repris en remplaçant $(X'X)^{-1}$ par $(X'X)^{-}$.

Exemple genotype (suite) Pour le jeu de données `genotype`, l'estimation avec la fonction `lm` donne la sortie suivante

```
Call:
lm(formula = Wt ~ Mother, data = genotype)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   55.400      1.910  28.999   <2e-16 ***
MotherB        3.300      2.797   1.180   0.2429
MotherI       -2.038      2.702  -0.754   0.4539
MotherJ       -6.720      2.746  -2.447   0.0175 *

Residual standard error: 7.642 on 57 degrees of freedom
Multiple R-squared:  0.1882,Adjusted R-squared:  0.1455
F-statistic: 4.405 on 3 and 57 DF,  p-value: 0.007433
```

La ligne `Intercept` contient l'estimation du paramètre d'intercept, représentant ici la moyenne des poids pour les portées de mère de génotype A, puisque la CI par défaut dans R est $\alpha_1 = 0$. Les autres coefficients sont les différences de poids par rapport à celui des portées de mère A. On en déduit les es poids moyens pour chaque portée :

```
      A       B       I       J
55.4000 58.7000 53.3625 48.6800
```

Le test sur la ligne `MotherB` (resp. `MotherI`, `MotherJ`) représente le test de nullité de la différence de poids entre les portés de mère de génotype B (resp. I, J) et celles de génotype A, ie, l'égalité des poids moyens. On peut remarquer au passage que la différence de poids entre A et J est significative, alors que celle entre A et B ne l'est pas, celle entre A et I ne l'est pas non plus. Mais les sorties ne permettent pas de conclure entre les portées de mère I et B, pour lesquelles il faudra faire un test particulier. Voyons les différences de résultats avec une autre CI

```
Call:
lm(formula = Wt ~ Mother, data = genotype, contrasts = list(Mother = "contr.SAS"))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   48.680      1.973  24.672  < 2e-16 ***
MotherA        6.720      2.746   2.447 0.017518 *
MotherB       10.020      2.840   3.528 0.000834 ***
MotherI        4.683      2.746   1.705 0.093647 .
```

```
Residual standard error: 7.642 on 57 degrees of freedom
Multiple R-squared:  0.1882,Adjusted R-squared:  0.1455
F-statistic: 4.405 on 3 and 57 DF,  p-value: 0.007433
```

Les valeurs des coefficients ne sont pas les mêmes comme attendu, mais on peut les retrouver : par exemple : $55.4 - 6.72 = 48.68$ dans la première estimation correspond à l'intercept dans la seconde (la CI de SAS met le paramètre du dernier niveau à 0). Les tests indiqués portent maintenant sur la différence de poids avec les mères de génotype J.

Il y a des informations qui ne changent pas, parce qu'elles sont intrinsèques au modèle et ne dépendent pas de la paramétrisation : ce sont la racine carrée de l'erreur résiduelle, le test de significativité globale et le coefficient de détermination (R^2). Notons au passage celui-ci est faible, ce qui indique un modèle faiblement ajusté par cette régression.

Le test de significativité globale de la régression (iid contre modèle) fait intervenir une statistique de Fisher à 3 degrés de liberté au numérateur et 57 au dénominateur. La différence de dimension entre le modèle iid et le modèle d'analyse de variance est de $4 - 1 = 3$, d'où nombre de degrés de liberté du numérateur. Au dénominateur apparait le nombre de degrés de liberté du à l'estimation de la variance du bruit σ^2 . Si $Im(X)$ est de dimension 4, son orthogonal est de dimension $n - 4 = 57$. On en déduit qu'il a $n = 61$ observations dans ce jeu de données.

6.6.2 ANCOVA : Analyse de la covariance

Un modèle de régression avec une variable explicative quantitative et une qualitative s'appelle analyse de covariance. On souhaite par exemple expliquer la perte par abrasion d'une gomme de pneu en fonction de sa dureté et du type du carbone incorporé [2]. Soit Y_{ik} l'observation de la k-ième expérience du type i, effectué avec une gomme dont la dureté est t_{ik}. Pour chaque type de carbone $i = 1, \ldots, I$, et pour chaque expérience $k = 1, \ldots, n_i$ avec ce type,

$$Y_{ik} = \alpha_i + \beta_i t_{ik} + \varepsilon_{ik}, \;\; i = 1, \ldots, I; \;\; k = 1, \ldots, n_I; \;\; n = \sum_{i=1}^{I} n_i; \;\; \varepsilon \sim \mathcal{N}(0, \sigma^2 Id_n) \qquad (6.20)$$

Le paramètre associé est $\theta = (\alpha_1, ..., \alpha_I, \beta_1, \ldots, \beta_I)$ et le modèle définit une droite par niveau du facteur, avec une variance commune du bruit. Afin de faire apparaître l'intercept sur l'ordonnée à l'origine comme en ANOVA1 et également sur la pente, on représentera le modèle plutôt sous la forme

$$Y_{ik} = \mu + \alpha_i + (\nu + \beta_i) t_{ik} + \varepsilon_{ik}, \;\; i = 1, \ldots, I; \;\; k = 1, \ldots, n_I; \;\; n = \sum_{i=1}^{I} n_i; \;\; \varepsilon \sim \mathcal{N}(0, \sigma^2 Id_n) \quad (6.21)$$

de matrice du plan d'expérience :

$$X = \begin{pmatrix} 1 & 1 & & & t_{11} & t_{11} & & \\ \vdots & \vdots & & & \vdots & \vdots & & \\ 1 & 1 & & & t_{1n_1} & t_{1n_1} & & \\ \vdots & & \ddots & & \vdots & & \ddots & \\ 1 & & & 1 & t_{I1} & & & t_{I1} \\ \vdots & & & \vdots & \vdots & & & \vdots \\ 1 & & & 1 & t_{In_I} & & & t_{In_I} \end{pmatrix}.$$

2. O.L. Davies (1947) Statistical Methods in Research and Production. Oliver and Boyd, Table 6.1 p. 119

qui est non injective. On utilise les mêmes CI que celles décrites en ANOVA1, choisies en fonction de l'interprétation souhaitée pour les coefficients. Il est possible de définir (et tester !) des modèle plus simples :

— droites parallèles $\nu + \beta_1 = \ldots = \nu + \beta_I$ si le type n'agit que sur l'ordonnée à l'origine de la droite, et non sur la pente.
— droites avec même ordonnée à l'origine si le type n'agit que sur la pente
— regroupement de droites pour deux (voire plusieurs) niveaux du facteur.

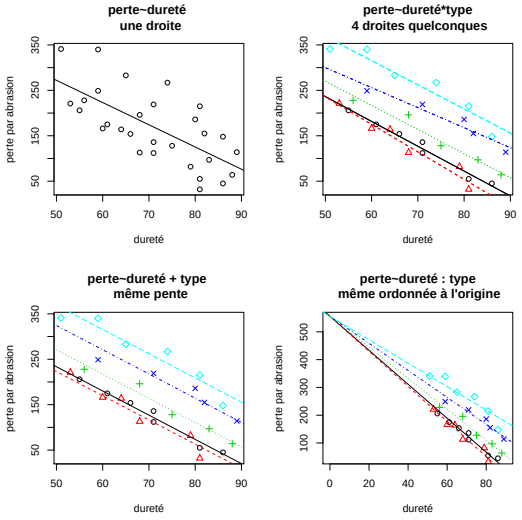


FIGURE 6.7 – 4 modèles différents : une seule droite de régression (linéaire simple), 4 droites quelconques, 4 droites parallèles, 4 droites de même ordonnées à l'origine

Dans le cas de la gomme de pneu, le test de Fisher entre un modèle avec quatre droites parallèles contre quatre droites quelconques conserve quatre droites parallèles (p-value=0.63) :

```
Model 1: perte ~ dureté + type
Model 2: perte ~ dureté * type
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     23 6518.8
2     19 5738.6  4     780.2 0.6458 0.6366
```

6.6.3 ANOVA2 : Analyse de la variance à deux facteurs

Reprenons l'exemple des portées, que l'on cherche maintenant à expliquer suivant le génotype et la mère (facteur à I niveaux) et celui de la portée (facteur à J niveaux). Il y a donc $I \times J$ cas possibles. De plus, n_{ij} individus sont observés pour chaque cas ij. Soit Y_{ijk} la k-ième portée de génotype j dont la mère est de génotype i :

$$Y_{ijk} = \mu+\alpha_i+\beta_j+\gamma_{ij}+\varepsilon_{ijk}, \quad i = 1,\ldots,I; \quad j = 1,\ldots,J; k = 1,\ldots,n_{ij}; \quad n = \sum_{i,j} n_{i,j}; \quad \varepsilon \sim \mathcal{N}(0,\sigma^2 Id_n) \tag{6.22}$$

Le paramètre est $\theta = (\mu, \alpha_1, \ldots, \alpha_I, \beta_1, \ldots, \beta_J, \gamma_{11}, \ldots, \gamma_{IJ})$, α_i indique l'effet du génotype i de la mère, β_j l'effet du génotype j de la portée et γ_{ij} modélise l'interaction entre les deux : pour

deux génotypes i et i' de la mère , l'un est affecte plus le poids de la portée de génotype j que celui de la portée de génotype j'. Avec le vecteur des individus $Y = (Y_{ijk})$ dont les coordonnées sont rangées par ordre lexicographique des indices, la matrice du plan d'expérience s'écrit :

$$X = \begin{pmatrix} \mathbb{1}_{n_{11}} & \mathbb{1}_{n_{11}} & & \mathbb{1}_{n_{11}} & & \mathbb{1}_{n_{11}} & & & & \\ \vdots & \vdots & & & \ddots & & \ddots & & & \\ \mathbb{1}_{n_{1J}} & \mathbb{1}_{n_{1J}} & & & & \mathbb{1}_{n_{1J}} & & \mathbb{1}_{n_{1J}} & & \\ \vdots & & \ddots & & \vdots & & & & \ddots & \\ \mathbb{1}_{n_{I1}} & & \mathbb{1}_{n_{I1}} & \mathbb{1}_{n_{I1}} & & & & & \mathbb{1}_{n_{I1}} & \\ \vdots & & \vdots & & \ddots & & & & & \ddots \\ \mathbb{1}_{n_{IJ}} & & \mathbb{1}_{n_{IJ}} & & & \mathbb{1}_{n_{IJ}} & & & & \mathbb{1}_{n_{IJ}} \end{pmatrix}.$$

Le modèle présenté est le modèle ANOVA2 *avec interactions* (on dit aussi modèle ANOVA2 *complet*). Si tous les γ_{kl} sont nuls, le modèle est appelé ANOVA2 *simple* (ou sans interaction).

6.6.4 Généralisation

La question de la paramétrisation de l'ANOVA1 (6.19) peut paraître arbitraire, parce que la CI $\mu = 0$ apparait naturellement quand on pose le modèle sous la forme (6.18). En revanche, le modèle ANOVA2 n'est pas identifiable, même sans intercept. Il est donc intéressant de regarder plus largement la question de l'identifiabilité.

Conditions d'identifiabilité

Si le paramètre de la loi n'est pas identifiable, il est possible de restreindre l'espace des paramètres en fixant une condition d'identifiabilité, permettant ainsi l'unicité du paramètre dans le modèle restreint.

Définition 36. *Une condition d'identifiabilité (CI) est une contrainte d'appartenance du paramètre θ à un sous espace vectoriel $\mathcal{C}$ de $\mathbb{R}^p$ tel que $X(\mathcal{C}) = Im(X) = V$ et la restriction de X à $\mathcal{C}$ est injective.*
Le paramètre d'intérêt est alors l'unique $\theta_C \in \mathcal{C}$ tel que $m = \theta_C$.

Elle s'exprime sous la forme

$$\mathcal{C} = \{C\theta = 0\},$$

où C est une matrice de dimension $dim(Ker(X)) \times p$ de rang la dimension de $Ker(X)$ telle que $Ker(C) \cap Ker(X) = \{0\}$. La restriction de X à $\mathcal{C}$ est alors injective. En effet, soit θ et $\tilde{\theta}$ de $\mathcal{C}$, tels que $X\theta = X\tilde{\theta}$ et $C\theta = C\tilde{\theta}$. On a $\tilde{\theta} - \theta \in Ker(C) \cap Ker(X) = \{0\}$ donc $\theta = \tilde{\theta}$.

Une telle CI existe toujours, par exemple en prenant $\mathcal{C}$ égal au supplémentaire orthogonal de $Ker(X)$ dans $\mathbb{R}^p$.

Estimation Utiliser une condition d'identifiabilité, c'est compléter le système initial $X\theta = Y$ par $C\theta = 0$ pour tenir compte des paramètres surabondants. On récrit le modèle

$$\tilde{Y} := \begin{bmatrix} Y \\ 0 \end{bmatrix} = \begin{bmatrix} X \\ C \end{bmatrix} \theta + \begin{bmatrix} \varepsilon \\ 0 \end{bmatrix} := \tilde{X}\theta + \tilde{\varepsilon}.$$

et on en déduit l'unique estimateur dans $\mathcal{C}$

$$\widehat{\theta}_C = (\tilde{X}'\ \tilde{X})^{-1}\tilde{X}\tilde{Y} = \left([X'\ C'] \begin{bmatrix} X \\ C \end{bmatrix}\right)^{-1} [X'\ C'] \begin{bmatrix} Y \\ 0 \end{bmatrix} = (X'X + C'C)^{-1}X'Y$$

ce qui revient à utiliser un inverse généralisé $Q^- = (X'X + C'C)^{-1}$ de $Q = X'X$ (tel que $QQ^-Q = Q$). Q^- dépend de la forme de C, qui est définie à un facteur multiplicatif près : C et λC ($\lambda > 0$) amènent à la même régularisation.

Démonstration. On peut écrire :

$$\begin{aligned} X'X(X'X + C'C)^{-1}X'X &= (X'X + C'C - C'C)(X'X + C'C)^{-1}X'X \\ &= X'X - C'C(X'X + C'C)^{-1}X'X \end{aligned}$$

Montrons que la matrice $M = C'C(X'X + C'C)^{-1}X'X$, est identiquement nulle. Soit $r > 0$ la dimension de $Ker(X)$ et p la dimension de θ. C est une matrice de dimension $r \times p$ de rang r telle que $Ker(C) \cap Ker(X) = \{0\}$ pour fixer l'ensemble des singularités de X et M est de dimension $p \times p$.

Soit $B_u = (u_1, \ldots, u_r)$ une base de vecteurs propres de $Ker(X)$. Tout vecteur de cette base appartient aussi à $Ker(M)$, car, pour tout $j = 1, \ldots, r$,

$$Mu_j = C'C(X'X + C'C)^{-1}X'(Xu_j) = 0$$

Donc B_u est un système libre de r vecteurs de $Ker(M)$.

Par ailleurs, M est symétrique parce que $X'X(X'X + C'C)^{-1}X'X$ l'est, et $Ker(C)$ est de dimension $p - r$. Soit $B_v = (v_1, \ldots, v_{p-r})$ une base de $Ker(C)$. Alors, pour tout $j = 1, \ldots, r - p$,

$$v_j'M = M'v_j = Mv_j = 0$$

Donc B_v est un système libre de $p - r$ vecteurs de $Ker(M)$. Comme $Ker(C) \cap Ker(X) = \{0\}$, vecteurs de $B_u \cup B_v$ forme une base de $Ker(M)$ d'où $M = 0$, puisque toutes ses valeurs propres sont nulles. ◊

Formes estimables

Même dans le cas singulier, certaines combinaisons linéaires des coordonnées du paramètre θ donnent toujours la même expression, quelque soit le choix de la paramétrisation induite par la non identifiabilité : ce sont les **formes estimables**. C'est le cas pour chaque $\mathrm{x}_i\theta = m_i$: chaque coordonnée de $\mathbb{E}(Y) = X\theta = m$ est donc estimable, de même que toute combinaison linéaire des coordonnées de m.

Théorème 29 (Conditions nécessaires et suffisantes d'estimabilité)**.** *Soit A une matrice $1 \times p$. Les conditions suivantes sont équivalentes :*

(i) $A\theta$ est estimable.

(ii) $Ker(X) \subset Ker(A)$.

(iii) Il existe une application linéaire T telle que $A = TX$. Dans ce cas, $A\theta = Tm$.

Toute forme estimable A est donc une combinaison linéaire des lignes de X.

Dans un modèle régulier, toute forme est estimable, et dans un modèle singulier, les composantes de θ ne le sont pas.

Démonstration. $(iii) \to (ii)$ est évident. $(i) \to (iii)$ s'obtient en notant que si $A \neq TX$, alors pour $u \in Ker(X)$, $Au \neq TXu = 0$ donc $u \notin Ker(A)$, ce qui est en contradiction avec (ii).

$(i) \to (ii)$ Si $A\theta$ est estimable, alors pour tout $u \in Ker(X)$, $Au = TXu = T0 = 0$ donc $u \in Ker(A)$.

$(ii) \to (i)$ Si $Ker(X) \subset Ker(A)$, alors pour tout $u \in Ker(X)$ non nul, $\theta \neq \theta + u$, et $A(\theta + u) = A\theta$ car u appartient aussi à $Ker(A)$ par hypothèse. Ainsi, $A\theta$ est insensible à la paramétrisation choisie. ◊

Exemple le modèle ANOVA1 équilibré est tel que $n_1 = \ldots = n_I = n/I$:

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, i = 1, \ldots, I; \quad j = 1, \ldots, n/I$$

Soit T de dimension $1 \times n$. Les formes estimables A sont telles que $TX = A$, d'où leur expression $A = (A_1 = A_2 + A_3 + \ldots + A_{I+1}, A_2, A_3, \ldots, A_{I+1})$.

Cette approche est intéressante pour le praticien, qui n'est pas forcément intéressé par la valeur d'un coefficient en lui-même, mais par certaines combinaisons linéaires : il s'agit par exemple de comparer la différence de réponse entre le niveau i et le niveau i' d'un ANOVA1 : $\alpha_i - \alpha_{i'} = \mu + \alpha_i - (\mu + \alpha_{i'})$ est estimable comme différence de deux coordonnées de $\mathbb{E}(Y)$. On vérifie également que la forme linéaire correspondante satisfait aux conditions des formes estimables.

Dimension du modèle

La dimension du modèle est la dimension de V (ou de $Im(X)$). C'est donc la dimension du paramètre θ si le modèle est identifiable. Sinon, la dimension de θ est supérieure à la dimension du modèle. Par exemple :

- La dimension de l'ANOVA1 défini en (6.19) est I, alors que la dimension initiale de θ est $I+1$. Nous avons vu à la section précédente plusieurs exemples de condition d'identifiabilité sous forme d'une relation linéaire pour restreindre la dimension de l'espace de θ à un espace de dimension I
- La dimension de l'ANCOVA définie en (6.21) est $2I$ alors que la dimension de θ est $2I+1$. Il faudra définir une contrainte d'identifiabilité sous forme d'une relation linéaire.
- L'ANOVA2 défini en (6.22) comporte $(I + 1)(J + 1)$ paramètres, alors que le modèle, si toutes les conditions d'expérience sont observées, est de dimension IJ. Il faudra donc imposer $1 + I + J$ relations linéaires de rang $1 + I + J$.

Annexe A

Vecteurs gaussiens

Ce chapitre rappelle quelques propriétés de vecteurs gaussiens et quelques lois utiles.

A.1 Définition

Définition 37 (Loi gaussienne sur $\mathbb{R}$)**.** *La loi gaussienne $\mathcal{N}(\mu, \sigma^2)$, $\mu \in \mathbb{R}$ et $\sigma^2 \in \mathbb{R}^+$ est la probabilité de densité par rapport à la mesure de Lebesgue*

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

Définition 38 (Vecteur Gaussien)**.** *Un **vecteur aléatoire** Y à valeurs dans $\mathbb{R}^n$ est **gaussien** si et seulement si toute combinaison linéaire de ses coordonnées est gaussienne, ie :*

$$\textit{Pour tout } U \in \mathbb{R}^n, \ \ \exists \mu \in \mathbb{R}^n, \sigma^2 \in \mathbb{R}^+ \textit{ t.q. } U'Y \sim \mathcal{N}(\mu, \sigma^2)$$

A.2 Propriétés

Soit Y un vecteur aléatoire gaussien de dimension n.

- Sa loi est complètement déterminée par
  - son **espérance** : $\mathbb{E}(Y) = (\mathbb{E}(Y_1), \ldots, \mathbb{E}(Y_n))' = \mu \in \mathbb{R}^n$
  - et sa**variance** : $\text{var}(Y) = (\text{cov}(Y_i, Y_j)) = \Sigma$, qui est une matrice de dimension $n \times n$.
- Si Σ est **inversible**, la densité par rapport à la mesure de Lebesgue sur $\mathbb{R}^n$ est

$$\frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma}} \exp - \frac{(y-\mu)'\Sigma^{-1}(y-\mu)}{2}$$

- Si A est une matrice $p \times n$ et $Y \sim \mathcal{N}_n(\mu, \Sigma)$,

$$AY \sim \mathcal{N}_p(A\mu, A\Sigma A')$$

Vecteurs gaussiens et indépendance

- Si Y est un vecteur gaussien et si sa variance est diagonale par blocs, alors les blocs de coordonnées correspondants forment des **vecteurs gaussiens indépendants**.

— Un n-**échantillon gaussien** iid est un **vecteur gaussien** de loi $\mathcal{N}_n(0, Id_n)$, c'est-à-dire un vecteur dont les n composantes sont des variables aléatoires indépendantes de loi gaussienne centrée réduite.
— Lorsqu'on fait un **changement de base orthonormée**, un vecteur gaussien reste un vecteur gaussien.
— La somme des carrés de n gaussiennes centrées réduite est appelée **Loi du Khi-**2 à n degrés de liberté : c'est la norme d'un vecteur iid gaussien centré réduit

A.3 Projection d'un vecteur gaussien

Théorème 30 (Cochran). *Si $Y \sim \mathcal{N}_n(\mu, I_n)$, et si $E_1 \oplus \ldots \oplus E_r = \mathbb{R}^n$ est une décomposition de $\mathbb{R}^n$ en r sous-espaces orthogonaux, alors les projections orthogonales $\Pi_1(Y), \ldots, \Pi_r(Y)$ sur ces sous-espaces sont des vecteurs gaussiens indépendants tels que, pour tout $j = 1, \ldots, r$*

$$||\Pi_j(Y)||^2 \sim \chi^2(d_j = Dim(E_j), \mu_j = ||\Pi_j(\mu)||^2).$$

Preuve Pour tout j, soit $(e_{j1}, \ldots, e_{jk})$ une base orthonormée de E_j où e_{jk} est le k-ième vecteur de la base de E_j. La décomposition sur cette base orthonormée du vecteur Y s'écrit

$$\Pi_j Y = \sum_{k=1}^{d_j} < e_{jk}, Y > e_{jk}$$

Soit U matrice de passage d'une base orthonormée de $\mathbb{R}^n$ dans base orthonormée de $\mathbb{R}^n$. Elle est telle que $UU' = Id_n$. On a $UY \sim \mathcal{N}_n(U\mu, Id_n)$. Les variables $e'_{jk}Y$ sont indépendantes quand j et k varient. Donc $\Pi_1(Y), \ldots, \Pi_r(Y)$ sont indépendantes.

Pour un sous-espace E_j, et pour $k = 1, \ldots, k_j$

$$e'_{jk}Y \sim \mathcal{N}(e'_{jk}\mu, e'_{jk}e_{jk} = 1)$$

d'où , avec $\mu_j = ||\Pi_j\mu||^2 = \sum_{k=1}^{d_j}(e'_{jk}\mu)^2$ on a

$$||\Pi_j(Y)||^2 = \sum_{k=1}^{d_j} ||e'_{jk}Y||^2 \sim \chi^2(d_j, \mu_j),$$

Application En décomposant $\mathbb{R}^n = E_1 \oplus E_1^\perp$ où E_1 est la droite engendrée par le vecteur unitaire $(1, \ldots 1)/\sqrt{n}$, on déduit du théorème de Cochran que pour un n-échantillon gaussien Y :
— $\bar{Y}$ et $\widehat{\sigma}^2$ sont indépendants
— $\sum_i (Y_i - \bar{Y})^2 \sim \sigma^2 \chi^2_{n-1}$

A.4 Quelques lois utiles

A.4.1 Loi du Khi-deux

Définition 39 (Loi du Khi-2). *Si $Y \sim \mathcal{N}_n(\mu, Id_n)$, alors*

$$K_n = ||Y||^2 \sim \chi^2(n, ||\mu||^2)$$

*Quand $||\mu||^2 = \sum_{i=1}^n \mu_i^2 \neq 0$ la loi est dite **décentrée** de facteur de décentrage $||\mu||^2$.*

On a : $\mathbb{E}(K_n) = n + ||\mu||^2$; $\text{var}(K_n) = 2(n + 2||\mu||^2)$.

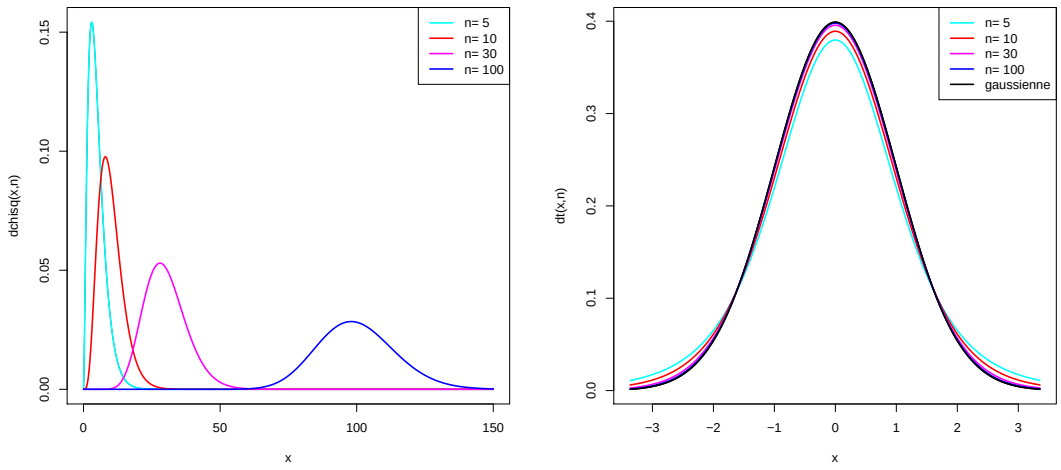


FIGURE A.1 – Quelques exemples de densités de loi du Khi-deux (à gauche) et de Student(à droite)

A.4.2 Lois de Student et Fisher

Définition 40 (Loi de Student)**.** *Soit deux variables Z et K* ***indépendantes*** *telles que $Z \sim \mathcal{N}(0,1)$ et $K \sim \chi^2(p)$. Alors, la v.a.*

$$T = \frac{Z}{\sqrt{\frac{K}{p}}} \sim \mathcal{T}(p)$$

suit une loi appelée loi de ***Student*** *à p degrés de liberté.*

On a :

$$\mathbb{E}(T) = 0; \quad \text{var}(T) = \frac{p}{p-2} \text{ pour } p > 2$$

Remarque : Si $Z \sim \mathcal{N}(\mu, 1)$ avec $\mu \neq 0$, alors la v.a. $Z/\sqrt{K/p}$ suit une loi de Student **décentrée** de facteur de **décentrage** ou de **non-centralité** $\mu = \mathbb{E}(Z)$.

Définition 41 (Loi de Fisher)**.** *Soit deux variables K_1 et K_2* ***indépendantes*** *telles que $K_1 \sim \chi^2(n_1)$ et $K_2 \sim \chi^2(n_2)$. Alors, la v.a.*

$$F = \frac{K_1/n_1}{K_2/n_2} \sim \mathcal{F}(n_1, n_2)$$

suit une loi appelée loi de ***Fisher*** *à (n_1, n_2) degrés de liberté.*

On a,

$$\mathbb{E}(F) = \frac{n_2}{n_2 - 2} \text{ pour } n_2 > 2; \quad \text{var}(F) = \frac{2n_2^2(n_1 + n_2 - 2)}{n_1(n_2-2)^2(n_2-4)} \text{ pour } n_2 > 4$$

et

$$q_{1-\alpha}^{F(n_1,n_2)} = 1/q_\alpha^{F(n_2,n_1)}$$

Remarque : Si K_1 suit une loi du Khi-deux de facteur de décentrage θ, alors la v.a. F suit une loi de Fisher **décentrée** (supérieurement) de facteur de décentrage θ.

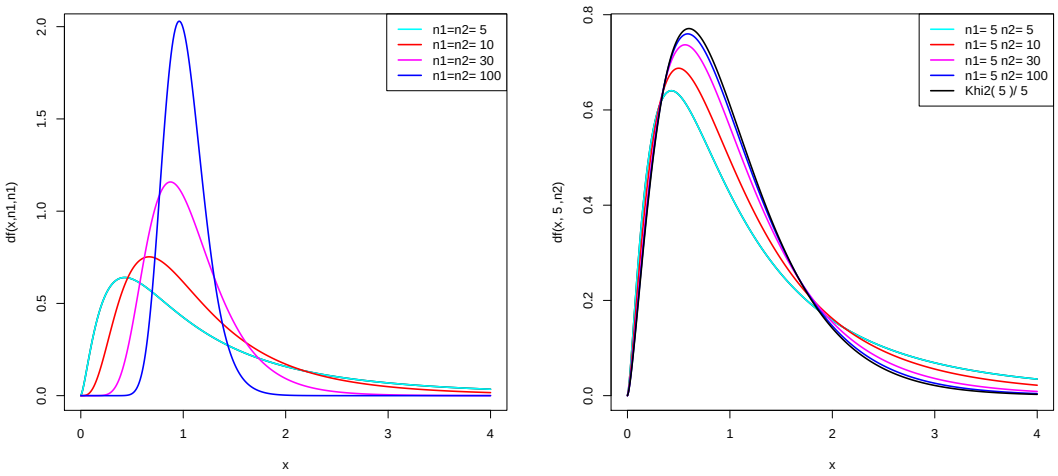


FIGURE A.2 – Quelques exemples de densités de loi de Fisher

Annexe B

Rappels de convergence

B.1 Définitions

Définition 42. *Soit X_n une suite de v.a. et X une v.a.. On étudie le comportement de X_n quand n tend vers l'infini.*

— *La suite X_n* ***converge en loi*** *vers la v.a. X si*

$$\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$$

pour tout x où la fonction de répartition de X est continue. On parle aussi de convergence en distribution ou de convergence faible. On note $X_n \xrightarrow{\mathcal{L}} X$.

— *La suite X_n* ***converge en probabilité*** *vers la v.a. X si*

$$\forall \varepsilon, \lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0$$

On note : $X_n \xrightarrow{\mathcal{P}} X$.

— *La suite X_n converge en* ***presque sûrement*** *(convergence forte) vers la v.a. X si*

$$\mathbb{P}(\lim_{n\to\infty} |X_n - X| = 0\ ) = 1$$

On dit aussi $X_n - X \to 0$ avec probabilité 1 et on note $X_n \xrightarrow{p.s.} X$.

— *La suite X_n converge en* ***moyenne quadratique*** *vers la v.a. X si*

$$\mathbb{E}[(X_n - X)^2] \to 0$$

On dit aussi que X_n converge vers X dans L^2 et on note : $X_n \xrightarrow{L^2} X$.

Définitions équivalentes de la convergence en loi ou lemme de porte-manteau

— $X_n \xrightarrow{\mathcal{L}} X$ ssi $\mathbb{E}[h(X_n)] \to \mathbb{E}[h(X)]$ pour toute fonction continue et bornée h

— $X_n \xrightarrow{\mathcal{L}} X$ ssi $\mathbb{E}[h(X_n)] \to \mathbb{E}[h(X)]$ pour toute fonction lipschitzienne et bornée h
On dit qu'une fonction $H : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ est lipschitzienne de constante $K > 0$ ssi $||h(x) - h(y)|| \leq K||x - y||$. En particulier, toute fonction continûment dérivable sur un intervalle fermé borné est lipschitzienne.

B.2 Relations entre les convergences

Théorème 31. *On a les relations suivantes :*

1. $X_n \xrightarrow{p.s.} X$ *implique* $X_n \xrightarrow{\mathcal{P}} X$
2. $X_n \xrightarrow{L^2} X$ *implique* $X_n \xrightarrow{\mathcal{P}} X$
3. $X_n \xrightarrow{\mathcal{P}} X$ *implique* $X_n \xrightarrow{\mathcal{L}} X$
4. *Soit c est une constante réelle.* $X_n \xrightarrow{\mathcal{P}} c$ *ssi* $X_n \xrightarrow{\mathcal{L}} c$

Remarque En général, les implications réciproques de 2. et 3. sont fausses.

— La convergence en loi n'implique pas celle en probabilité. Contre-exemple : Soit $X \sim \mathcal{N}(0,1)$ et $X_n = -X$ pour tout n. Alors, X_n a même loi que X pour tout n, donc $\lim_n \mathbb{P}(X_n \leq x) = \mathbb{P}(X \leq x)$ pour tout x, d'où $X_n \xrightarrow{\mathcal{L}} X$. Mais $\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|2X| > \varepsilon) \neq 0$, donc X_n ne converge pas vers 0 en probabilité

— La convergence en probabilité n'implique pas celle en moyenne quadratique, en particulier parce que la convergence en probabilité ne nécessite pas d'avoir un moment d'ordre deux fini. Contre-exemple : Soit $U_{[0;1]}$ une variable uniforme sur $[0;1]$ et $X_n = \sqrt{n}\mathbb{1}_{U \leq 1/n}$.
Alors, $\mathbb{P}(|X_n| > \varepsilon) = \mathbb{P}(U \leq 1/n) = 1/n \to 0$ donc $X_n \xrightarrow{\mathcal{P}} 0$. Mais $\mathbb{E}(X_n^2) = \int_0^{1/n} \sqrt{n}^2 dt = 1$ pour tout n

Le résultat suivant est très utile pour étudier les estimateurs :

Lemme 1 (de l'application continue)**.** *Soit $X_1, \ldots, X_n \sim \mathcal{P}_\theta$. Soit g une fonction continue (au moins en tout point x d'un ensemble A tel que $X \in A) = 1$), et soit X_n une suite de variables aléatoires :*

— *Si* $X_n \xrightarrow{p.s.} X$*, alors* $g(X_n) \xrightarrow{p.s.} g(X)$
— *Si* $X_n \xrightarrow{\mathcal{P}} X$*, alors* $g(X_n) \xrightarrow{\mathcal{P}} g(X)$
— *Si* $X_n \xrightarrow{\mathcal{L}} X$*, alors* $g(X_n) \xrightarrow{\mathcal{L}} g(X)$

Un outil bien pratique pour montrer la convergence probabilité est l'inégalité de Tchébychev

Propriété 32 (Inégalité de Bienaymé-Tchebychev)**.** *Soit T une v.a. telle que $\mathbb{E}(T^2) < +\infty$. Alors,*

$$\forall t > 0, \quad \mathbb{P}(\{|T - \mathbb{E}(T)| > t\}) \leq \frac{\mathrm{var}(T)}{t^2}$$

B.3 Convergence de couples de variables aléatoires

On peut généraliser la définition des convergences en probabilité à des couples de variables aléatoires.

Définition 43. *Le couple de variables aléatoires (X_n, Y_n) converge en probabilité vers (X, Y) si, pour tout $\varepsilon > 0$,*

$$\mathbb{P}\left(d((X_n, Y_n), (X, Y)) > \varepsilon\right) \to 0$$

où d est la distance euclidienne de $\mathbb{R}^2$

$$d((x_1, y_1), (x_2, y_2)) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

On a le résultat suivant

Propriété 33. *Si* $X_n \xrightarrow{\mathcal{P}} X$ *et* $Y_n \xrightarrow{\mathcal{P}} Y$*, alors* $(X_n, Y_n) \xrightarrow{\mathcal{P}} (X, Y)$.
Le lemme de l'application continue est encore vrai pour les vecteurs aléatoires.

On a donc équivalence entre la convergence en probabilité du couple et la convergence en probabilité de chacune des marginales

Attention Cette assertion est fausse pour la convergence en loi. La connaissance de chacune des marginales ne détermine pas en général la loi jointe du couple. Il y a des exceptions : si X_n et Y_n sont indépendantes, ou si l'une des deux marginales converge vers une constate

Propriété 34. *Quelques propriétés de la convergence en loi de couples de v.a.*
- *Si* $(X_n, Y_n) \xrightarrow{\mathcal{L}} (X, Y)$*, alors* $X_n \xrightarrow{\mathcal{L}} X$ *et* $Y_n \xrightarrow{\mathcal{L}} Y$
  La réciproque est fausse en général
- *si* X_n *et* Y_n *sont indépendants, et* X *et* Y *indépendantes, alors*
  si $X_n \xrightarrow{\mathcal{L}} X$ *et* $Y_n \xrightarrow{\mathcal{L}} Y$*, alors* $(X_n, Y_n) \xrightarrow{\mathcal{L}} (X, Y)$

Lemme 2 (Lemme de Slutsky)**.** *Si* $X_n \xrightarrow{\mathcal{L}} X$ *et* $Y_n \xrightarrow{\mathcal{L}} c$ *où* c *est une constante, alors* $(X_n, Y_n) \xrightarrow{\mathcal{L}} (X, c)$.
En appliquant cette convergence jointe à une fonction continue, on a en particulier
- $X_n + Y_n \xrightarrow{\mathcal{L}} X + c$
- $X_n Y_n \xrightarrow{\mathcal{L}} cX$
- $X_n / Y_n \xrightarrow{\mathcal{L}} X/c$

Démonstration. On utilise la caractérisation de la convergence en loi du lemme porte-manteau. Soit h une fonction lipschitzienne de constante K bornée par M.

$$|\mathbb{E}[h(X_n, Y_n)] - \mathbb{E}[h(X, c)]| \leq |\mathbb{E}[h(X_n, Y_n)] - \mathbb{E}[h(X_n, c)]| + |\mathbb{E}[h(X_n, c)] - \mathbb{E}[h(X, c)]|$$

Le second terme tend vers 0 car $X_n \xrightarrow{\mathcal{L}} X$ et en appliquant le lemme de l'application continue. On majore maintenant le premier terme

$$\begin{aligned}
|\mathbb{E}[h(X_n, Y_n)] - \mathbb{E}[h(X_n, c)]| \leq & |\left(\mathbb{E}[h(X_n, Y_n)] - \mathbb{E}[h(X_n, c)]\right) \mathbb{1}_{||Y_n - c|| > \varepsilon}| \\
& + |\left(\mathbb{E}[h(X_n, Y_n)] - \mathbb{E}[h(X_n, c)]\right) \mathbb{1}_{||Y_n - c|| \leq \varepsilon}| \\
& \leq 2 \sup_{x,y} ||h(x, y)||\ \mathbb{P}(||Y_n - c|| > \varepsilon) + K\mathbb{E}[||Y_n - c|| \mathbb{1}_{||Y_n - c|| \leq \varepsilon}] \\
& \leq 2M\ \mathbb{P}(||Y_n - c|| > \varepsilon) + K\varepsilon \mathbb{P}(||Y_n - c|| \leq \varepsilon)
\end{aligned}$$

Comme $Y_n \xrightarrow{\mathcal{L}} c$ le premier terme tend vers 0, et on majore le second terme par $K\varepsilon$ pour tout $\varepsilon > 0$. ◊

Annexe C

Tables

C.1 Table de probabilité de la loi gaussienne

t	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
0,00	0,5	0,5040	0,5080	0,5120	0,5160	0,5199	0,5239	0,5279	0,5319	0,5359
0,10	0,5398	0,5438	0,5478	0,5517	0,5557	0,5596	0,5636	0,5675	0,5714	0,5753
0,20	0,5793	0,5832	0,5871	0,5910	0,5948	0,5987	0,6026	0,6064	0,6103	0,6141
0,30	0,6179	0,6217	0,6255	0,6293	0,6331	0,6368	0,6406	0,6443	0,6480	0,6517
0,40	0,6554	0,6591	0,6628	0,6664	0,6700	0,6736	0,6772	0,6808	0,6844	0,6879
0,50	0,6915	0,6950	0,6985	0,7019	0,7054	0,7088	0,7123	0,7157	0,7190	0,7224
0,60	0,7257	0,7291	0,7324	0,7357	0,7389	0,7422	0,7454	0,7486	0,7517	0,7549
0,70	0,758	0,7611	0,7642	0,7673	0,7704	0,7734	0,7764	0,7794	0,7823	0,7852
0,80	0,7881	0,7910	0,7939	0,7967	0,7995	0,8023	0,8051	0,8078	0,8106	0,8133
0,90	0,8159	0,8186	0,8212	0,8238	0,8264	0,8289	0,8315	0,8340	0,8365	0,8389
1,00	0,8413	0,8438	0,8461	0,8485	0,8508	0,8531	0,8554	0,8577	0,8599	0,8621
1,10	0,8643	0,8665	0,8686	0,8708	0,8729	0,8749	0,8770	0,8790	0,8810	0,8830
1,20	0,8849	0,8869	0,8888	0,8907	0,8925	0,8944	0,8962	0,8980	0,8997	0,9015
1,30	0,9032	0,9049	0,9066	0,9082	0,9099	0,9115	0,9131	0,9147	0,9162	0,9177
1,40	0,9192	0,9207	0,9222	0,9236	0,9251	0,9265	0,9279	0,9292	0,9306	0,9319
1,50	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,60	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,70	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,80	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,90	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767
2,00	0,9772	0,9778	0,9783	0,9788	0,9793	0,9798	0,9803	0,9808	0,9812	0,9817
2,10	0,9821	0,9826	0,9830	0,9834	0,9838	0,9842	0,9846	0,9850	0,9854	0,9857
2,20	0,9861	0,9864	0,9868	0,9871	0,9875	0,9878	0,9881	0,9884	0,9887	0,9890
2,30	0,9893	0,9896	0,9898	0,9901	0,9904	0,9906	0,9909	0,9911	0,9913	0,9916
2,40	0,9918	0,9920	0,9922	0,9925	0,9927	0,9929	0,9931	0,9932	0,9934	0,9936
2,50	0,9938	0,9940	0,9941	0,9943	0,9945	0,9946	0,9948	0,9949	0,9951	0,9952
2,60	0,9953	0,9955	0,9956	0,9957	0,9959	0,9960	0,9961	0,9962	0,9963	0,9964
2,70	0,9965	0,9966	0,9967	0,9968	0,9969	0,9970	0,9971	0,9972	0,9973	0,9974
2,80	0,9974	0,9975	0,9976	0,9977	0,9977	0,9978	0,9979	0,9979	0,9980	0,9981
2,90	0,9981	0,9982	0,9982	0,9983	0,9984	0,9984	0,9985	0,9985	0,9986	0,9986

t	3	3,1	3,2	3,3	3,4	3,5	3,6	3,8	4	4,5
P	0,99865	0,99903	0,99931	0,99952	0,99966	0,99977	0,99984	0,99993	0,99997	1

t	1,282	1,645	1,96	2,326	2,576	3,09
P	0,9	0,95	0,975	0,99	0,995	0,999

t	-3,719	-4,2649	-4,7534	-5,1993	-5,612	-5,9978
P	10^ -4	10^ -5	10^ -6	10^ -7	10^ -8	10^ -9

C.2 Quantiles de la loi de Student

n \ α	0,75	0,900	0,950	0,975	0,990	0,995
1	1,000	3,078	6,314	12,706	31,821	63,657
2	0,817	1,886	2,920	4,303	6,965	9,925
3	0,765	1,638	2,353	3,182	4,541	5,841
4	0,741	1,533	2,132	2,776	3,747	4,604
5	0,727	1,476	2,015	2,571	3,365	4,032
6	0,718	1,440	1,943	2,447	3,143	3,707
7	0,711	1,415	1,895	2,365	2,998	3,499
8	0,706	1,397	1,860	2,306	2,896	3,355
9	0,703	1,383	1,833	2,262	2,821	3,250
10	0,700	1,372	1,812	2,228	2,764	3,169
11	0,697	1,363	1,796	2,201	2,718	3,106
12	0,696	1,356	1,782	2,179	2,681	3,055
13	0,694	1,350	1,771	2,160	2,650	3,012
14	0,692	1,345	1,761	2,145	2,624	2,977
15	0,691	1,341	1,753	2,131	2,602	2,947
16	0,690	1,337	1,746	2,120	2,583	2,921
17	0,689	1,333	1,740	2,110	2,567	2,898
18	0,688	1,330	1,734	2,101	2,552	2,878
19	0,688	1,328	1,729	2,093	2,539	2,861
20	0,687	1,325	1,725	2,086	2,528	2,845
21	0,686	1,323	1,721	2,080	2,518	2,831
22	0,686	1,321	1,717	2,074	2,508	2,819
23	0,685	1,319	1,714	2,069	2,500	2,807
24	0,685	1,318	1,711	2,064	2,492	2,797
25	0,684	1,316	1,708	2,060	2,485	2,787
26	0,684	1,315	1,706	2,056	2,479	2,779
27	0,684	1,314	1,703	2,052	2,473	2,771
28	0,683	1,313	1,701	2,048	2,467	2,763
29	0,683	1,311	1,699	2,045	2,462	2,756
30	0,683	1,310	1,697	2,042	2,457	2,750
40	0,681	1,303	1,684	2,021	2,423	2,704
60	0,679	1,296	1,671	2,000	2,390	2,660
120	0,677	1,289	1,658	1,980	2,358	2,617
	0,675	1,282	1,645	1,960	2,327	2,576

C.3 Quantiles de la loi du Khi-deux

n \	0,01	0,025	0,05	0,1	0,5	0,9	0,95	0,975	0,99	0,995
1	0,0002	0,001	0,004	0,02	0,46	2,71	3,84	5,02	6,63	7,88
2	0,02	0,05	0,10	0,21	1,39	4,61	5,99	7,38	9,21	10,6
3	0,12	0,22	0,35	0,58	2,37	6,25	7,81	9,35	11,34	12,84
4	0,30	0,48	0,71	1,06	3,36	7,78	9,49	11,14	13,28	14,86
5	0,55	0,83	1,15	1,61	4,35	9,24	11,07	12,83	15,09	16,75
6	0,87	1,24	1,64	2,20	5,35	10,64	12,59	14,45	16,81	18,55
7	1,24	1,69	2,17	2,83	6,35	12,02	14,07	16,01	18,48	20,28
8	1,65	2,18	2,73	3,49	7,34	13,36	15,51	17,53	20,09	21,95
9	2,09	2,70	3,33	4,17	8,34	14,68	16,92	19,02	21,67	23,59
10	2,56	3,25	3,94	4,87	9,34	15,99	18,31	20,48	23,21	25,19
11	3,05	3,82	4,57	5,58	10,34	17,28	19,68	21,92	24,72	26,76
12	3,57	4,40	5,23	6,30	11,34	18,55	21,03	23,34	26,22	28,3
13	4,11	5,01	5,89	7,04	12,34	19,81	22,36	24,74	27,69	29,82
14	4,66	5,63	6,57	7,79	13,34	21,06	23,68	26,12	29,14	31,32
15	5,23	6,26	7,26	8,55	14,34	22,31	25	27,49	30,58	32,8
16	5,81	6,91	7,96	9,31	15,34	23,54	26,3	28,85	32	34,27
17	6,41	7,56	8,67	10,09	16,34	24,77	27,59	30,19	33,41	35,72
18	7,01	8,23	9,39	10,86	17,34	25,99	28,87	31,53	34,81	37,16
19	7,63	8,91	10,12	11,65	18,34	27,2	30,14	32,85	36,19	38,58
20	8,26	9,59	10,85	12,44	19,34	28,41	31,41	34,17	37,57	40
21	8,90	10,30	11,59	13,24	20,34	29,62	32,67	35,48	38,93	41,4
22	9,54	11,00	12,34	14,04	21,34	30,81	33,92	36,78	40,29	42,8
23	10,20	11,70	13,09	14,85	22,34	32,01	35,17	38,08	41,64	44,18
24	10,90	12,40	13,85	15,66	23,34	33,2	36,42	39,36	42,98	45,56
25	11,50	13,10	14,61	16,47	24,34	34,38	37,65	40,65	44,31	46,93
26	12,20	13,80	15,38	17,29	25,34	35,56	38,89	41,92	45,64	48,29
27	12,90	14,60	16,15	18,11	26,34	36,74	40,11	43,19	46,96	49,64
28	13,60	15,30	16,93	18,94	27,34	37,92	41,34	44,46	48,28	50,99
29	14,30	16,00	17,71	19,77	28,34	39,09	42,56	45,72	49,59	52,34
30	15,00	16,80	18,49	20,60	29,34	40,26	43,77	46,98	50,89	53,67
40	22,20	24,40	26,51	29,05	39,34	51,81	55,76	59,34	63,69	66,77
50	29,70	32,40	34,76	37,69	49,34	63,17	67,5	71,42	76,15	79,49
60	37,50	40,50	43,19	46,46	59,34	74,4	79,08	83,3	88,38	91,95
70	45,40	48,80	51,74	55,33	69,33	85,53	90,53	95,02	100,43	104,21
80	53,50	57,20	60,39	64,28	79,33	96,58	101,88	106,63	112,33	116,32
90	61,80	65,60	69,13	73,29	89,33	107,57	113,15	118,14	124,12	128,3
100	70,10	74,20	77,93	82,36	99,33	118,5	124,34	129,56	135,81	140,17

C.4 Quantiles de la loi de Fisher

p = degré du dénominateur \ a=0,99	n = degrés du numérateur																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	inf
1	4052	5000	5403	5625	5764	5859	5928	5981	6022	6056	6106	6157	6209	6235	6261	6287	6313	6339	6366
2	98,50	99,00	99,17	99,25	99,30	99,33	99,36	99,37	99,39	99,40	99,42	99,43	99,45	99,46	99,47	99,47	99,48	99,49	99,50
3	34,12	30,82	29,46	28,71	28,24	27,91	27,67	27,49	27,35	27,23	27,05	26,87	26,69	26,60	26,50	26,41	26,32	26,22	26,13
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,55	14,37	14,20	14,02	13,93	13,84	13,75	13,65	13,56	13,46
5	16,26	13,27	12,06	11,39	10,97	10,67	10,46	10,29	10,16	10,05	9,89	9,72	9,55	9,47	9,38	9,29	9,20	9,11	9,02
6	13,75	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,72	7,56	7,40	7,31	7,23	7,14	7,06	6,97	6,88
7	12,25	9,55	8,45	7,85	7,46	7,19	6,99	6,84	6,72	6,62	6,47	6,31	6,16	6,07	5,99	5,91	5,82	5,74	5,65
8	11,26	8,65	7,59	7,01	6,63	6,37	6,18	6,03	5,91	5,81	5,67	5,52	5,36	5,28	5,20	5,12	5,03	4,95	4,86
9	10,56	8,02	6,99	6,42	6,06	5,80	5,61	5,47	5,35	5,26	5,11	4,96	4,81	4,73	4,65	4,57	4,48	4,40	4,31
10	10,04	7,56	6,55	5,99	5,64	5,39	5,20	5,06	4,94	4,85	4,71	4,56	4,41	4,33	4,25	4,17	4,08	4,00	3,91
11	9,65	7,21	6,22	5,67	5,32	5,07	4,89	4,74	4,63	4,54	4,40	4,25	4,10	4,02	3,94	3,86	3,78	3,69	3,60
12	9,33	6,93	5,95	5,41	5,06	4,82	4,64	4,50	4,39	4,30	4,16	4,01	3,86	3,78	3,70	3,62	3,54	3,45	3,36
13	9,07	6,70	5,74	5,21	4,86	4,62	4,44	4,30	4,19	4,10	3,96	3,82	3,66	3,59	3,51	3,43	3,34	3,25	3,17
14	8,86	6,51	5,56	5,04	4,69	4,46	4,28	4,14	4,03	3,94	3,80	3,66	3,51	3,43	3,35	3,27	3,18	3,09	3,01
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,67	3,52	3,37	3,29	3,21	3,13	3,05	2,96	2,87
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,55	3,41	3,26	3,18	3,10	3,02	2,93	2,84	2,75
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,46	3,31	3,16	3,08	3,00	2,92	2,83	2,75	2,65
18	8,29	6,01	5,09	4,58	4,25	4,01	3,84	3,71	3,60	3,51	3,37	3,23	3,08	3,00	2,92	2,84	2,75	2,66	2,57
19	8,18	5,93	5,01	4,50	4,17	3,94	3,77	3,63	3,52	3,43	3,30	3,15	3,00	2,92	2,84	2,76	2,67	2,58	2,49
20	8,10	5,85	4,94	4,43	4,10	3,87	3,70	3,56	3,46	3,37	3,23	3,09	2,94	2,86	2,78	2,69	2,61	2,52	2,42
21	8,02	5,78	4,87	4,37	4,04	3,81	3,64	3,51	3,40	3,31	3,17	3,03	2,88	2,80	2,72	2,64	2,55	2,46	2,36
22	7,95	5,72	4,82	4,31	3,99	3,76	3,59	3,45	3,35	3,26	3,12	2,98	2,83	2,75	2,67	2,58	2,50	2,40	2,31
23	7,88	5,66	4,76	4,26	3,94	3,71	3,54	3,41	3,30	3,21	3,07	2,93	2,78	2,70	2,62	2,54	2,45	2,35	2,26
24	7,82	5,61	4,72	4,22	3,90	3,67	3,50	3,36	3,26	3,17	3,03	2,89	2,74	2,66	2,58	2,49	2,40	2,31	2,21
25	7,77	5,57	4,68	4,18	3,85	3,63	3,46	3,32	3,22	3,13	2,99	2,85	2,70	2,62	2,54	2,45	2,36	2,27	2,17
30	7,56	5,39	4,51	4,02	3,70	3,47	3,30	3,17	3,07	2,98	2,84	2,70	2,55	2,47	2,39	2,30	2,21	2,11	2,01
40	7,31	5,18	4,31	3,83	3,51	3,29	3,12	2,99	2,89	2,80	2,66	2,52	2,37	2,29	2,20	2,11	2,02	1,92	1,81
60	7,08	4,98	4,13	3,65	3,34	3,12	2,95	2,82	2,72	2,63	2,50	2,35	2,20	2,12	2,03	1,94	1,84	1,73	1,60
120	6,85	4,79	3,95	3,48	3,17	2,96	2,79	2,66	2,56	2,47	2,34	2,19	2,03	1,95	1,86	1,76	1,66	1,53	1,38
inf	6,64	4,61	3,78	3,32	3,02	2,80	2,64	2,51	2,41	2,32	2,19	2,04	1,88	1,79	1,70	1,59	1,48	1,33	1,05

p = degré du dénominateur \ a=0,95	n = degrés du numérateur																		
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	inf
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251	252	253	254
2	18,51	19,00	19,16	19,25	19,30	19,33	19,35	19,37	19,38	19,40	19,41	19,43	19,45	19,45	19,46	19,47	19,48	19,49	19,50
3	10,13	9,55	9,28	9,12	9,01	8,94	8,89	8,85	8,81	8,79	8,74	8,70	8,66	8,64	8,62	8,59	8,57	8,55	8,53
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,91	5,86	5,80	5,77	5,75	5,72	5,69	5,66	5,63
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,77	4,74	4,68	4,62	4,56	4,53	4,50	4,46	4,43	4,40	4,37
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,00	3,94	3,87	3,84	3,81	3,77	3,74	3,70	3,67
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,64	3,57	3,51	3,44	3,41	3,38	3,34	3,30	3,27	3,23
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,35	3,28	3,22	3,15	3,12	3,08	3,04	3,01	2,97	2,93
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,14	3,07	3,01	2,94	2,90	2,86	2,83	2,79	2,75	2,71
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,98	2,91	2,85	2,77	2,74	2,70	2,66	2,62	2,58	2,54
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,85	2,79	2,72	2,65	2,61	2,57	2,53	2,49	2,45	2,41
12	4,75	3,89	3,49	3,26	3,11	3,00	2,91	2,85	2,80	2,75	2,69	2,62	2,54	2,51	2,47	2,43	2,38	2,34	2,30
13	4,67	3,81	3,41	3,18	3,03	2,92	2,83	2,77	2,71	2,67	2,60	2,53	2,46	2,42	2,38	2,34	2,30	2,25	2,21
14	4,60	3,74	3,34	3,11	2,96	2,85	2,76	2,70	2,65	2,60	2,53	2,46	2,39	2,35	2,31	2,27	2,22	2,18	2,13
15	4,54	3,68	3,29	3,06	2,90	2,79	2,71	2,64	2,59	2,54	2,48	2,40	2,33	2,29	2,25	2,20	2,16	2,11	2,07
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,42	2,35	2,28	2,24	2,19	2,15	2,11	2,06	2,01
17	4,45	3,59	3,20	2,96	2,81	2,70	2,61	2,55	2,49	2,45	2,38	2,31	2,23	2,19	2,15	2,10	2,06	2,01	1,96
18	4,41	3,55	3,16	2,93	2,77	2,66	2,58	2,51	2,46	2,41	2,34	2,27	2,19	2,15	2,11	2,06	2,02	1,97	1,92
19	4,38	3,52	3,13	2,90	2,74	2,63	2,54	2,48	2,42	2,38	2,31	2,23	2,16	2,11	2,07	2,03	1,98	1,93	1,88
20	4,35	3,49	3,10	2,87	2,71	2,60	2,51	2,45	2,39	2,35	2,28	2,20	2,12	2,08	2,04	1,99	1,95	1,90	1,84
21	4,32	3,47	3,07	2,84	2,68	2,57	2,49	2,42	2,37	2,32	2,25	2,18	2,10	2,05	2,01	1,96	1,92	1,87	1,81
22	4,30	3,44	3,05	2,82	2,66	2,55	2,46	2,40	2,34	2,30	2,23	2,15	2,07	2,03	1,98	1,94	1,89	1,84	1,78
23	4,28	3,42	3,03	2,80	2,64	2,53	2,44	2,37	2,32	2,27	2,20	2,13	2,05	2,01	1,96	1,91	1,86	1,81	1,76
24	4,26	3,40	3,01	2,78	2,62	2,51	2,42	2,36	2,30	2,25	2,18	2,11	2,03	1,98	1,94	1,89	1,84	1,79	1,73
25	4,24	3,39	2,99	2,76	2,60	2,49	2,40	2,34	2,28	2,24	2,16	2,09	2,01	1,96	1,92	1,87	1,82	1,77	1,71
30	4,17	3,32	2,92	2,69	2,53	2,42	2,33	2,27	2,21	2,16	2,09	2,01	1,93	1,89	1,84	1,79	1,74	1,68	1,62
40	4,08	3,23	2,84	2,61	2,45	2,34	2,25	2,18	2,12	2,08	2,00	1,92	1,84	1,79	1,74	1,69	1,64	1,58	1,51
60	4,00	3,15	2,76	2,53	2,37	2,25	2,17	2,10	2,04	1,99	1,92	1,84	1,75	1,70	1,65	1,59	1,53	1,47	1,39
120	3,92	3,07	2,68	2,45	2,29	2,18	2,09	2,02	1,96	1,91	1,83	1,75	1,66	1,61	1,55	1,50	1,43	1,35	1,26
inf	3,84	3,00	2,61	2,37	2,21	2,10	2,01	1,94	1,88	1,83	1,75	1,67	1,57	1,52	1,46	1,40	1,32	1,22	1,03

Bibliographie

J.-M. Azais et J.-M. Bardet. *Le modèle linéaire par l'exemple.* Dunod, Paris, 2005.

P. J. Bickel et K. A. Doksum. *Mathematical statistics : basic ideas and selected topics, volume I*, volume 117. CRC Press, 2015.

P.-A. Cornillon et autres. *Statistiques avec R.* Presses Universitaires de Rennes, Rennes, 2008.

P.-A. Cornillon et E. Matzner-Løber. *Régression Théorie et applications.* Springer, Paris, 2007.

J.-F. Delmas. *Introduction au calcul des probabilités et à la statistique.* Les Presses de l'ENSTA, 2010.

C. Keribin. *Bases de la statistique inférentielle.* poly ENSTA 1A, 2018.

M. Lejeune. *Statistique-La théorie et ses applications.* Springer, 2010.

J. Pagès. *Statistique générales pour utilisateurs.* Presses Universitaires de Rennes, Rennes, 2005.

V. Rivoirard et G. Stoltz. *Statistique en action.* Vuibert, Paris, 2009.

G. Saporta. *Probabilités, Analyse des données et Statistique, 2ème édition.* Editions TECHNIP, Paris, 2006.

W. Venables et B. Ripley. *Modern Applied Statistics with S-Plus, 3rd edition.* Springer–Verlag, New York, 2001.

Table des matières