# Statistiques Fondamentales

Simon Coste

2024-02-21

# Table des matières

# Organisation

Bienvenue sur la page du cours de Statistiques Fondamentales (STF8) du master Mathématiques Fondamentales et Appliquées de l'Université Paris-Cité. Les notes de cours sont accessibles à tous. De nombreux auteurs s'y sont succédés ; je suis le dernier en date, mais les versions précédentes ont été travaillées par Clément Levrard, Stéphane Boucheron, Stéphane Gaïffas, Pierre Youssef.

- Les CM ont lieu les jeudi à (8h30 - 10h30), et les vendredi (10h45 - 12h45) **sauf le premier cours qui a lieu lundi 8 janvier à 10h45-12h45**.
- Les TD ont lieu lundi (13h45 - 16h45) et vendredi (13h30 - 15h30), de lundi 8 janvier à vendredi 16 février.
- Il y aura deux contrôles de 2h, le vendredi 26 janvier et lundi 12 février.
- L'examen a lieu le 1er mars de 13h30 à 16h30.
- Il y aura une interro de 5 minutes chaque semaine le jeudi.

## Utiliser ce site

Chaque chapitre de ce livre contient une page dédiée au cours théorique, et contiendra dans un futur proche une page d'exercices.

La saveur du cours est essentiellement mathématique et nous n'aurons pas de TP d'info ; cependant, je vous recommande vraiment d'essayer d'appliquer tout ça via votre langage de programmation favori, c'est-à-dire ~~Python~~ ~~R~~ ~~SAS~~ ~~C++~~ Julia. J'essaierai autant que possible de fournir des mini-jeux de données avec des petits challenges pour appliquer ce que vous apprenez en cours.

Ces notes sont mises en lignes et totalement accessibles via Quarto. Si vous savez comment utiliser `git`, n'hésitez pas à corriger toutes les erreurs que vous pourriez voir (et Dieu sait qu'elles seront nombreuses) via des pull requests.

# 1 Introduction

Les outils des statistiques furent créés pour analyser des phénomènes quantitatifs dans lesquels la présence de *bruit* ou de *hasard* rendait l'analyse classique moins opérante. Il peut typiquement s'agir de problèmes dans lesquels de nombreuses données indépendantes ont été générées par le même phénomène. Dans cette section, nous allons développer un exemple pour bien comprendre les questions qui se posent et la façon de les résoudre, puis nous poserons quelques bases qui nous permettront d'utiliser le langage des mathématiques.

## 1.1 Un exemple pour fixer les idées

Une grande enseigne de distribution possède $n = 100$ magasins identiques, qui génèrent chaque année un chiffre d'affaire annuel (CA, en millions d'euros). Ce chiffre oscille autour d'une valeur de référence $\mu$. Cette valeur n'est pas observée ; ce qui est observé, ce sont tous les chiffres d'affaires des $n$ magasins, qui fluctuent tous autour de la vraie valeur $\mu$. Ces fluctuations proviennent de nombreuses sources : erreurs comptables, perturbations des ventes dues aux fournisseurs ou aux prix, etc. Ce qu'on observe, c'est donc des chiffres $x_1, \dots, x_n$ qui ne sont pas tous égaux ; comment avoir une idée de la véritable valeur de $\mu$ ?

**Estimation.** Évidemment, la moyenne empirique

$$\bar{x}_n = \frac{x_1 + \cdots + x_n}{n}$$

vient naturellement à l'esprit. En faisant le calcul, on trouve $\bar{x}_n \approx 21,6$. Cette valeur est une *estimation* du CA moyen $\mu$. Ce chiffre peut être utilisé par l'enseigne, par exemple pour jauger la rentabilité d'un possible plan d'ouverture de nouveaux magasins.

**Précision**. On pourrait se demander à quel point cette estimation est précise ou, disons, essayer de quantifier l'erreur possible qu'on fait si l'on dit que $\mu$ est égal à 21,6 millions d'euros. Cela nécessite de faire quelques hypothèses sur le hasard qui génère les fluctuations des $x_i$ autour de $\mu$. Ces fluctuations observées au cours de l'année proviennent de l'agrégation de toutes les fluctuations quotidiennes, lesquelles sont à peu près indépendantes, et pour cette raison on peut supposer (pour commencer) que ces fluctuations sont gaussiennes et ont à peu près la même variance, disons $\sigma^2 = 1$. Comme on a supposé que les $x_i$ sont des réalisations d'une loi gaussienne $N(\mu, 1)$, alors on sait que $\bar{x}_n$ est la réalisation d'une loi $N(\mu, 1/n)$, ou encore que $\bar{x}_n - \mu$ est la réalisation d'une gaussienne centrée de variance $1/n$. Les lois gaussiennes sont bien connues ; par exemple, avec probabilité supérieure à 99%, une gaussienne $N(0, \sigma^2)$ est comprise entre les valeurs $-2,96\sigma$ et $2,96\sigma$. Autrement dit, il y a 99% de chances pour que le nombre $|\bar{x} - \mu|$, qui représente l'erreur d'estimation, soit plus petite que $2,96/\sqrt{n} = 2,96/10 \approx 0,3$.

Ce dernier raisonnement peut être vu d'une autre façon. Dire que $\bar{x}_n$ et $\mu$ ne diffèrent pas de plus de $0,3$, c'est équivalent à dire que $\mu$ appartient à l'intervalle $[\bar{x} - 0,3, \bar{x} + 0,3]$. En d'autres termes, avec une probabilité supérieure à 99%, le vrai CA $\mu$ de chaque magasin se situe entre $21,3$ et $21,9$. Cela laisse tout de même une chance de 1% que le paramètre $\mu$ ne soit pas dans cette région.

**Tests**. Il existe encore un autre point de vue sur ce problème. Par exemple, le conseil d'administration de la firme veut s'assurer que le dirigeant a bien tenu sa promesse selon laquelle le CA de chaque magasin était supérieur à 21 millions d'euros. La valeur exacte de $\mu$ n'est pas le plus important : ce qui nous intéresse maintenant, c'est plutôt d'être sûrs que $\mu$ n'est pas inférieur au seuil de 21. Le dirigeant, fin statisticien, effectue alors un raisonnement par l'absurde *en probabilité*. Supposons que le CA $\mu$ soit effectivement égal à 21 (ou même, inférieur). Alors, par les mêmes calculs que ci-dessus, cela voudrait dire qu'avec 99% de chances, $\bar{x}_n$ et 21 ne devraient pas différer de plus de $0,3$ ; autrement dit, que $\bar{x}_n$ devrait se situer entre $20,7$ et $21,3$. Ce n'est pas le cas, puisque $\bar{x}_n = 21,6$. Si $\mu$ est réellement plus petit que 21, alors ce qu'on a observé est extrêmement peu probable. Par contraposée probabiliste, il est raisonnable de rejeter l'hypothèse selon laquelle $\mu$ est inférieur à 21.

---

Les trois points de vue donnés ci-dessus sont en quelque sorte les piliers de l'analyse statistique. L'estimation consiste à deviner une valeur cachée dans du bruit ; les intervalles de confiance consistent à donner une région dans laquelle se trouve cette valeur ; les tests d'hypothèse permettent de raisonner de façon logique sur cette valeur.

L'objectif du cours de statistiques de quantifier l'incertitude liée au hasard dans chacun de ces objectifs. Comme dans les exemples donnés ci-dessus, c'est un ensemble de méthodes scientifiques qui s'appuient sur la théorie des probabilités ; dans ce cours, on fera des *hypothèses* sur le hasard qui est en jeu, et on en tirera des conséquences *probables* sur le modèle sous-jacent. En théorie des probabilités, le jeu est plutôt inverse : partant d'un modèle probabiliste fixé, on essaie de déterminer quel sera le comportement des réalisations de ce modèle. Il semble difficile de faire l'un sans l'autre.

## 1.2 Qu'est-ce qu'un problème statistique ?

Il n'y aurait pas de statistiques s'il n'y avait pas de monde réel, et comme chacun sait, le monde réel est principalement composé de quantités aléatoires.

Un problème statistique tire donc toujours sa source d'un ensemble d'observations, disons $n$ observations notées $x_1, \dots, x_n$ ; cet ensemble d'observations est appelé un *échantillon*. L'hypothèse de base de tout travail statistique consiste à supposer que cet échantillon suit une certaine loi de probabilité ; l'objectif est de trouver laquelle. Évidemment, on ne va pas partir de rien : il faut bien faire des hypothèse minimales sur cette loi. Ce qu'on appelle un *modèle statistique* est le choix d'une famille de lois de probabilités que l'on suppose pertinentes.

**Définition 1.1.** Formellement, choisir un modèle statistique revient à choisir trois choses :

- $\mathcal{X}$, l'espace dans lequel vit notre échantillon ;
- $\mathscr{F}$, une tribu sur $\mathcal{X}$, pour donner du sens à ce qui est observable ou non ;
- $(P_\theta)_{\theta \in \Theta}$, une famille de mesures de probabilités sur $\mathcal{X}$ indexée par $\theta \in \Theta$, où $\Theta$ est appelé espace des paramètres. On écrira fréquemment $\mathbb{E}_\theta$ ou $\mathrm{Var}_\theta$ pour désigner des espérances, variances, etc., calculées avec la loi $P_\theta$.

En pratique, dans ce cours, on aura toujours un échantillon $(x_1, \dots, x_n)$ où les $x_i$ vivent dans un même espace, disons $\mathbb{R}^d$ pour simplifier. On devrait donc écrire $\mathcal{X} = \mathbb{R}^{d \times n}$ ; et l'on fera toujours l'hypothèse que ces observations sont indépendantes les unes des autres, et que ces observations ont la même loi de probabilité. Autrement dit, on se donnera toujours une mesure $p_\theta$ sur $\mathbb{R}^d$ et on supposera que la loi de notre échantillon est $P_\theta = p_\theta^{\otimes n}$. Dans ce cadre, les observations $x_i$ sont des *réalisations* de variables aléatoires $X_i$ iid de loi $p_\theta$.

Il faut prendre garde à distinguer les variables aléatoires $X_i$, qui sont des objets théoriques, de leurs réalisations $x_i$, qui, elles, sont bel et bien observées.

**Définition 1.2.** On dit qu'un modèle statistique est identifiable si $\theta \neq \theta'$ entraîne $P_\theta \neq P_{\theta'}$.

Si l'on a bien choisi notre modèle statistique, alors il existe un « vrai » paramètre, disons $\theta_\star$, tel que les observations $x_1, \dots, x_n$ sont des réalisations de loi $p_{\theta_\star}$. L'objectif est alors de trouver $\theta_\star$ ou quelque information que ce soit le concernant.

Dans un modèle identifiable, la statistique inférentielle (classique) permet de faire trois choses :

- Trouver une valeur approchée du vrai paramètre $\theta_\star$ (estimation ponctuelle).
- Donner une zone de $\Theta$ dans laquelle le vrai paramètre $\theta_\star$ a des chances de se trouver (intervalle de confiance).
- Répondre à des questions binaires sur $\theta_\star$, par exemple « $\theta_\star$ est-il positif ? ».

## 1.3 Qu'est-ce qu'un estimateur ?

**Définition 1.3.** Une *statistique* est une fonction mesurable des observations. Plus formellement, si le modèle statistique fixé est $(\mathcal{X}, \mathscr{F}, P)$, alors une statistique est n'importe quelle fonction mesurable de $(\mathcal{X}, \mathscr{F})$.

1) Le premier point important est qu'une statistique ne peut pas prendre $\theta$ en argument. Ses valeurs ne doivent dépendre du paramètre $\theta$ qu'au travers de $P_\theta$.

2) Le second point important est que, si $X$ est une variable aléatoire et $T$ une statistique, alors $T(X)$ est une variable aléatoire. On peut donc définir des quantités théoriques liées à $T$: typiquement, si $X$ a pour loi $P_\theta$, on peut définir la valeur moyenne de $T$ sous le modèle $P_\theta$ comme

$$\mathbb{E}_\theta[T(X)] = \int_{\mathcal{X}} T(x) P_\theta(dx)$$

ou encore sa variance $\mathbb{E}_\theta[T(X)^2] - (\mathbb{E}_\theta[T(X)])^2$, etc. On peut aussi calculer la valeur de cette statistique sur l'échantillon dont on dispose, c'est-à-dire $T(x_1, \dots, x_n)$. Par exemple, la moyenne empirique d'un $n$-échantillon réel est la fonction $T : (a_1, \dots, a_n) \to n^{-1}(a_1 + \cdots + a_n)$. Si les $x_i$ sont des réalisations des variables aléatoires $X_i$, alors $T(x_1, \dots, x_n)$ est une réalisation de la variable aléatoire $T(X_1, \dots, X_n)$.

3) Ce qui ne se voit pas dans la définition, c'est qu'une bonne statistique devrait être facilement calculable ; à la place de *statistique*, on peut penser à *algorithme* : une bonne statistique doit pouvoir être calculée facilement par un algorithme ne prenant en entrée que les échantillons $x_i$.

Si le but est de deviner la valeur de $\theta$ à partir des observations, il est naturel de considérer des statistiques à valeurs dans $\Theta$. C'est précisément la définition d'un estimateur.

**Définition 1.4.** Dans le modèle $(\mathcal{X}, \mathcal{A}, (P_\theta)_{\theta\in\Theta})$, un estimateur de $\theta$ est une statistique à valeurs dans $\Theta$.

En fait, on n'est pas obligés de vouloir estimer précisément $\theta$. Peut-être qu'on veut estimer quelque chose qui dépend de $\theta$, mais qui n'est pas $\theta$ ; disons, une fonction $\varphi(\theta)$. Dans ce cas, un estimateur de $\varphi(\theta)$ sera simplement une statistique à valeurs dans l'espace où vit $\varphi(\theta)$.

## 1.4 Points de vue

**Inférence paramétrique.** La plupart des expériences/modèles statistiques que nous rencontrerons dans ce cours, seront de nature dite *paramétrique*, autrement dit indexés par des parties de $\mathbb{R}^d$. Le mot "paramètre" est en lui-même trompeur : on parle souvent de paramètre d'une distribution pour désigner ce qui devrait plutôt s'appeler une fonctionnelle. Par exemple, la moyenne, la covariance d'une distribution sur $\mathbb{R}^d$ sont des paramètres de cette distribution. Les quantiles, l'asymétrie, la kurtosis sont d'autres paramètres.

**Statistique non paramétrique.** Tous les modèles ne sont pas paramétriques au sens ci-dessus : dans de nombreux développements des statistiques, par exemple en estimation de densité, on travaille sur des modèles plus riches qui n'admettent pas de paramétrisation naturelle par une partie d'un espace euclidien de dimension finie. C'est ce qu'on appelle l' *estimation non-paramétrique*. Nous y reviendrons au dernier chapitre.

**Statistique bayésienne.** En statistique paramétrique, les paramètres $\theta$ déterminent le hasard qui génère les observations $x_i$. La statistique bayésienne consiste à renverser le point de vue, et à rendre le paramètre $\theta$ lui-même aléatoire ; sa loi, appelée *prior*, mesure "le degré de connaissance a priori" qu'on en a. La règle de Bayes explique comment cette loi est modifiée par les observations. C'est un point de vue qui ne sera pas abordé dans ce cours.

# 2 Estimation de paramètre

On fixe un modèle statistique $(\mathcal{X}, \mathscr{F}, (P_\theta))$, et l'on cherche à estimer le paramètre $\theta$ ou un autre paramètre qui dépend de $\theta$. Dans ce chapitre, on explique comment juger la qualité d'un estimateur, et l'on donne une technique générale pour construire de bons estimateurs dans des situations assez naturelles.

## 2.1 Précision d'un estimateur

**Définition 2.1** (Biais , risque quadratique)**.**

- Le biais de $\hat{\theta}$ est la quantité $\mathbb{E}_\theta[\hat{\theta} - \theta]$. L'estimateur est dit *sans biais* s'il est de biais nul.
- Le risque quadratique de $\hat{\theta}$ est la quantité $\mathbb{E}_\theta[|\hat{\theta} - \theta|^2]$.

En pratique, on peut vouloir estimer non pas $\theta$ lui-même, mais un paramètre $\psi = \psi_\theta$ qui dépend de $\theta$, comme $\cos(\theta)$ ou $|\theta|$ par exemple. Dans ce cas, si $\hat{\psi}$ est un estimateur de $\psi$ alors le biais est défini par $\mathbb{E}_\theta[\hat{\psi} - \psi_\theta]$ et le risque quadratique par $\mathbb{E}_\theta[|\hat{\psi} - \psi_\theta|^2]$.

La dépendance du risque quadratique vis à vis de la taille de l'échantillon est une question importante : pour une suite d'expériences donnée, quelles sont les meilleures vitesses envisageables, et comment les obtenir ?

**Théorème 2.1** (Décomposition biais-variance)**.**

$$\mathbb{E}_\theta[|\hat{\theta} - \theta|^2] = \underbrace{\mathrm{Var}_\theta(\hat{\theta})}_{variance} + \underbrace{\mathbb{E}_\theta[\hat{\theta} - \theta]^2}_{carré\ du\ biais} .$$

## 2.2 Convergence

Rappelons brièvement deux notions de convergence des variables aléatoires. Une suite de variables aléatoires $X_n$ à valeurs dans $\mathbb{R}^d$ converge en probabilité vers une variable aléatoire $X$ à valeurs dans $\mathbb{R}^d$ si pour tout $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0 .$$

**Définition 2.2** (consistance d'un estimateur)**.** Une suite d'estimateurs $(\hat{\theta}_n)$ est convergente pour l'estimation de $\theta$ lorsque, pour tout $\theta \in \Theta$, sous $P_\theta$, la suite $(\hat{\theta}_n)$ converge en probabilité vers $\theta$ ; autrement dit, lorsque

$$\forall \varepsilon > 0, \qquad \lim_n P_\theta(|\hat{\theta}_n - \theta| > \varepsilon) = 0.$$

La suite est fortement convergente si, pour tout $\theta$, la convergence a lieu $P_\theta$-presque sûrement. $$

On voit parfois le mot *consistant* utilisé au lieu de *convergent*. Je pense que c'est un anglicisme.

## 2.3 Normalité asymptotique

Lorsqu'un estimateur est convergent, on peut se demander à quoi ressemblent ses fluctuations autour de sa valeur limite. Le théorème central limite indique que le comportement asymptotique gaussiens est relativement fréquent, et beaucoup d'estimateurs sont des sommes de réalisations de variables indépendantes.

**Définition 2.3** (normalité asymptotique)**.** Soit $\theta$ un paramètre à estimer, et $\hat{\theta}_n$ une suite d'estimateurs de $\theta$. On dit que ces estimateurs sont *asymptotiquement gaussiens* (ou *normaux*) si, après les avoir renormalisés convenablement, ils convergent en loi vers une loi gaussienne. Autrement dit, s'il existe une suite $a_n$ de nombres réels tels que

$$a_n(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{\text{loi}} N(0, \Sigma)$$

où $\Sigma$ est une matrice de covariance qui dépend peut-être de $\theta$ — pour éviter les cas dégénérés, on demande à ce que $\Sigma$ soit non-nulle.

## 2.4 Trois outils sur la normalité asymptotique

La normalité asymptotique n'est pas intéressante en elle-même ; l'idée est plutôt de chercher le comportement asymptotique de la statistique recentrée pour pouvoir en déduire des garanties en terme de risque asymptotique ou d'intervalle de confiance. Nous utiliserons cela de nombreuses fois dans la suite ; la normalité asymptotique sera par exemple utilisée dans la construction des intervalles de confiance. Aussi, prouver que des estimateurs sont asymptotiquement normaux est une tache importante, qui est grandement simplifiée par les deux outils suivants. On commence par rappeler le théorème centra-limite.

**Théorème 2.2** (Théorème Central-Limite)**.** *Soit $(X_i)$ une suite de variables aléatoires réelles, indépendantes et identiquement distribuées. On suppose que ces variables ont une variance $\sigma^2$ finie. Alors, la variable aléatoire*

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[X]\right)$$

*converge en loi vers une loi $N(0, \sigma^2)$.*

Le Lemme de Slutsky sera fréquemment utilisé pour combiner convergence en loi et convergence en probabilité.

**Théorème 2.3** (Lemme de Slutsky)**.** *Soit $(X_n)$ une suite de variables aléatoire qui converge en loi vers $X$ et $(Y_n)$ une suite de variables aléatoires qui converge en probabilité (ou en loi) vers une constante $c$. Alors, le* couple *$(X_n, Y_n)$ converge en loi vers $(X, c)$ ; autrement dit, pour toute fonction continue bornée $\varphi$,*

$$\mathbb{E}[\varphi(X_n, Y_n)] \to \mathbb{E}[\varphi(X, c)].$$

*Démonstration.* Fixons une fonction test $\varphi$ continue à support compact, donc bornée par un certain $M$. Il faut montrer que $\mathbb{E}[\varphi(X_n, Y_n) - \varphi(X, c)]$ tend vers zéro. L'intégrande est égal à la somme de $A = \varphi(X_n, Y_n) - \varphi(X_n, c)$ et de $B = \varphi(X_n, c) - \varphi(X, c)$.

Comme $X_n$ tend en loi vers $X$ et que $t \to \varphi(t, c)$ est continue bornée, l'espérance de $B$ tend vers zéro. Il faut donc montrer que l'espérance de $A$ tend vers zéro. On fixe un $\varepsilon > 0$.

- Par le théorème de Heine, $\varphi$ est uniformément continue : il existe $\delta > 0$ tel que $|(x,y)-(x',y')| < \delta$ entraîne que $|\varphi(x,y) - \varphi(x',y')| < \varepsilon/2$.
- On introduit l'événement $\{|Y_n - c| \leqslant \delta\}$. Par le point précédent, sur cet événement on a $|A| < \varepsilon/2$. Hors de cet événement, on peut toujours borner $|A|$ par $2M$. On a donc

$$|\mathbb{E}A| \leqslant \mathbb{P}(|Y_n - c| \leqslant \delta)\varepsilon/2 + \mathbb{P}(|Y_n - c| > \delta)2M.$$

- Comme $Y_n$ converge en probabilité vers $c$, lorsque $n$ est assez grand on a $\mathbb{P}(|Y_n - c| > \delta) < \varepsilon/4M$.
- En regroupant tout ce qui a été dit, on obtient bien $|\mathbb{E}A| \leqslant \varepsilon$ dès que $n$ est assez grand, ce qui montre bien que $\mathbb{E}A \to 0$.

□

**Théorème 2.4** (Delta-méthode)**.** *Soit $(X_n)$ une suite de variables aléatoires réelles telle que $\sqrt{n}(X_n - \alpha)$ converge en loi vers $N(0,\sigma^2)$. Pour toute fonction $g : \mathbb{R} \to \mathbb{R}$ dérivable en $\alpha$ (de dérivée non nulle en $\alpha$), on a*

$$\sqrt{n}(g(X_n) - g(\alpha)) \xrightarrow[n\to\infty]{loi} N(0, g'(\alpha)^2\sigma^2).$$

Plus généralement, si les $X_n$ sont à valeurs dans $\mathbb{R}^d$ et que $\sqrt{n}(X_n - \alpha) \to N(0,\Sigma)$, alors pour toute application $g : \mathbb{R}^d \to \mathbb{R}^k$, la suite $\sqrt{n}(g(X_n) - g(\alpha))$ converge en loi vers

$$N(0, Dg(\alpha)\Sigma Dg(\alpha)^\top)$$

où $Dg(x)$ est la matrice jacobienne de $g$ en $x$.

*Démonstration.* À écrire.

□

## 2.5 Deux estimateurs importants

Deux estimateurs sont omniprésents en statistique : la moyenne empirique et la variance empirique. Ils sont pertinents dans n'importe quel modèle où les observations sont des réalisations de variables iid possédant une moyenne $\mu$ et une variance $\sigma^2$.

La moyenne empirique est définie par

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Il est évident que $\mathbb{E}[\bar{X}_n] = \mathbb{E}[X] = \mu$. Cet estimateur est donc toujours sans biais, et son risque quadratique est égal à sa variance, c'est-à-dire $\frac{\sigma^2}{n}$.

L'estimateur de la variance empirique est défini comme

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

**Théorème 2.5.** *L'estimateur $\hat{\sigma}_n^2$ est sans biais.*

*Démonstration.* À écrire.

□

# 3 La méthode des moments

Il existe plusieurs techniques générales pour *construire* des estimateurs. Deux se démarquent : la méthode des moments, et la méthode du maximum de vraisemblance. La méthode des moments est naturelle et donne des estimateurs avec de bonnes propriétés, mais elle est moins automatique que la méthode du maximum de vraisemblance que nous verrons plus tard.

## 3.1 Qu'est-ce qu'un *moment* ?

Dans un modèle statistique, supposons qu'on dispose d'une statistiques intégrable $T$ (pas forcément réelle), dont la moyenne n'est pas le paramètre $\theta$ lui-même, mais plutôt une *fonction* de $\theta$ :

$$\mathbb{E}_\theta[T(X)] = \varphi(\theta).$$

C'est cette fonction $\varphi$ qu'on appelle *moment*. Typiquement,

- la moyenne d'une loi $\mathscr{E}(\theta)$ n'est pas $\theta$ mais $1/\theta$.
- la moyenne d'une loi log-normale de paramètres $(0, \sigma^2)$ est $e^{\sigma^2/2}$.

Prenons la moyenne empirique associée à cet estimateur, $\bar{T}_n$. Par la loi des grands nombres,

$$\bar{T}_n = \frac{1}{n}\sum_{i=1}^{n} T(X_i) \to \varphi(\theta) \qquad P_\theta - ps,$$

ce qui permet d'estimer $\varphi(\theta)$. Peut-on alors estimer $\theta$ ?

## 3.2 Estimateur des moments

Si la fonction $\varphi$ est inversible et si $\bar{T}_n$ appartient presque sûrement à l'ensemble de définition de $\varphi^{-1}$, alors $\varphi^{-1}(\bar{T}_n)$ est bien définie. Pour qu'en plus cette quantité converge presque sûrement vers $\theta$, il faut s'assurer que $\varphi^{-1}$ est continue. C'est par exemple le cas lorsque l'ensemble des paramètres $\Theta$ est un ouvert, et que $\varphi$ est un difféomorphisme sur son image — une situation si fréquente qu'elle mérite son propre théorème, et si agréable qu'elle garantit que l'estimateur associé est asymptotiquement normal.

**Théorème 3.1** (Estimation par moments)**.** *Sous l'hypothèse mentionnée ci-dessus (la fonction $\varphi$ est un difféomorphisme), l'estimateur*

$$\hat{\theta}_n = \varphi^{-1}(\bar{T}_n)$$

*est presque sûrement bien défini pour tout $n$ suffisamment grand ; il est également consistant pour l'estimation de $\theta$. En outre, si $T$ est de carré intégrable, cet estimateur est asymptotiquement normal, au sens où $\sqrt{n}(\hat{\theta}_n - \theta)$ converge en loi vers une gaussienne centrée de matrice de variance*

$$(D\varphi(\theta))^{-1}\mathrm{Var}_\theta(T)(D\varphi(\theta)^\top)^{-1}.$$

*Démonstration.* La première partie a essentiellement été démontrée un peu plus haut. Pour la seconde, il faut d'abord remarquer que si $T$ est de carré intégrable, alors $\sqrt{n}(\bar{T}_n - \varphi(\theta))$ converge vers une loi $N(0, \mathrm{Var}_\theta(T))$ par le TCL. Une simple application de la delta-méthode (Théorème 2.4) donne alors le résultat, puisque la matrice jacobienne de $\varphi^{-1}$ en $\varphi(\theta)$ n'est autre que l'inverse de la matrice jacobienne de $\varphi$ en $\theta$.

□

# Exercices

## Questions

1. Montrer que la convergence en loi vers une constante implique la convergence en probabilité.
2. Montrer que, si un modèle statistique n'est pas identifiable, alors il ne peut exister aucun estimateur convergent.
3. Trouver un couple de variables aléatoires $(X_n, Y_n)$ tel que $X_n$ converge en loi et $Y_n$ converge en loi, mais le couple ne converge pas en loi.
4. On observe un échantillon de lois de Poisson de paramètre $\lambda$, que l'on estime par la moyenne empirique. Calculer le risque quadratique de cet estimateur.
5. Quelle est la loi d'une somme de lois de Bernoulli indépendantes ? L'écart-type ?

## Exercices

**Exercice 3.1** (Variance empirique)**.** On se donne $Y_1, \dots, Y_n$, i.i.d de moyenne $\mu$ et variance $\sigma^2$.

1. On suppose $\mu$ connu. Donner un estimateur non biaisé de $\sigma^2$.
2. On suppose $\mu$ inconnu. Calculer l'espérance de $\sum_{i=1}^n (Y_i - \bar{Y}_n)^2$. En déduire un estimateur non biaisé de $\sigma^2$.

**Exercice 3.2** (Estimation de masse)**.** Au cours de la seconde guerre mondiale, l'armée alliée notait les numéros de série $X_1, \dots, X_n$ de tous les tanks nazis capturés ou détruits, afin d'obtenir un estimateur du nombre total $N$ de tanks produits.

1. Proposer un modèle pour le tirage de $X_1, \dots, X_n$.
2. Calculer l'espérance de $\bar{X}_n$. En déduire un estimateur non biaisé de $N$. Indication: la loi de $n$ tirages sans remise est échangeable.
3. Étudier la loi de $X_{(n)}$ et en déduire un estimateur non biaisé de $N$.
4. Proposer deux intervalles de confiance de niveau $1-\alpha$ de la forme $[aS, bS]$ avec $a, b \in \mathbb{R}$ et $S$ une statistique. On pourra utiliser le fait que l'inégalité de Hoeffding s'applique également aux tirages sans remise.

Selon Ruggles et Broodie (1947, JASA), la méthode statistique a fourni comme estimation une production moyenne de 246 tanks/mois entre juin 1940 et septembre 1942. Des méthodes d'espionnage traditionnelles donnaient une estimation de 1400 tanks/mois. Les chiffres officiels du ministère nazi des Armements ont montré après la guerre que la production moyenne était de 245 tanks/mois.

**Exercice 3.3** (Lois uniformes (1))**.** On considère $(X_1, \dots, X_n)$ un échantillon de loi uniforme sur $]\theta, \theta+1[$.

1. Donner la densité de la loi de la variable $R_n = X_{(n)} - X_{(1)}$, où $X_{(1)} = \min(X_1, \dots, X_n)$ et $X_{(n)} = \max(X_1, \dots, X_n)$.
2. Étudier les différents modes de convergence de $R_n$ quand $n \to \infty$.
3. Étudier le comportement en loi de $n(1 - R_n)$ quand $n \to \infty$.

**Exercice 3.4** (Lois uniformes (2))**.** Soit $X_1, \dots, X_n$ un échantillon de loi $\mathscr{U}([0, \theta])$, avec $\theta > 0$. On veut estimer $\theta$.

1. Déterminer un estimateur de $\theta$ à partir de $\bar{X}_n$.
2. On considère l'estimateur $X_{(n)} = \max_{1 \leq i \leq n} X_i$. Déterminer les propriétés asymotptiques de ces estimateurs.
3. Comparer les performances des deux estimateurs.

**Exercice 3.5** (Lois Gamma)**.** La loi Gamma $\Gamma(\alpha, \beta)$ de paramètres $\alpha, \beta > 0$ a pour densité

$$x \mapsto \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha - 1} \exp(-\beta x), \quad x > 0.$$

On se donne un échantillon $(X_1, \dots, X_n)$ de loi $\Gamma(\alpha, \beta)$ et on chercche à estimer les paramètres.

1. On suppose le paramètre $\beta$ connu. Proposer un estimateur de $\alpha$ par la méthode des moments.
2. On suppose à présent que les deux paramètres $\alpha, \beta$ sont inconnus. Proposer un estimateur de $(\alpha, \beta)$ par la méthode des moments.

**Exercice 3.6** (Lois de Gumbel)**.** La loi de Gumbel (centrale) de paramètre $\beta$ a pour fonction de répartition $F(x) = e^{-e^{-x/\beta}}$. On observe un échantillon de lois de Gumbel et l'on cherche à estimer $\beta$.

1. Calculer la densité des lois de Gumbel, ainsi que leur moyenne et variance [indice : 0.57721...]
2. En déduire un estimateur convergent dont on calculera le risque quadratique et les propriétés asymptotiques.

**Exercice 3.7** (Lois de Yule-Simon)**.** Une variable aléatoire $X$ suit la loi de Yule-Simon de paramètre $\rho > 0$ lorsque $\mathbb{P}(X = n) = \rho B(n, 1 + \rho)$, où $n \geqslant 1$ et $B$ est la fonction beta.

1. Montrer que si $\rho > 1$, alors $\mathbb{E}[X] = \rho/(\rho - 1)$.
2. Trouver un estimateur de $\rho$ et donner ses propriétés.

# 4 Intervalles de confiance

## 4.1 Principe

Dans un modèle statistique, l'estimation du paramètre d'intérêt $\theta$ par intervalles de confiance consiste à spécifier un intervalle calculable à partir des données, et qui contient $\theta$ avec grande probabilité : en d'autres termes, une *région de confiance* pour $\theta$.

Pour simplifier, on supposera d'abord que $\theta$ est un paramètre réel.

**Définition 4.1** (intervalle de confiance)**.** Un intervalle de confiance de niveau $1-\alpha$ est un intervalle $I = [A, B]$ dont les bornes $A, B$ sont des statistiques, et tel que pour tout $\theta$,

$$P_\theta(\theta \in I) \geqslant 1 - \alpha.$$

Un intervalle de confiance de niveau asymptotique $1-\alpha$ est une *suite* d'intervalles $I_n = [A_n, B_n]$ dont les bornes $A_n, B_n$ sont des statistiques, et tels que pour tout $n$,

$$P_\theta(\theta \in I_n) \geqslant 1 - \alpha.$$

Le terme « niveau » désigne $1-\alpha$ ; la vocation de ce nombre est d'être proche de 1, typiquement 99%. Le nombre $\alpha$ est parfois appelé « erreur », « marge d'erreur » ou encore « niveau de risque » ; la vocation de ce nombre est d'être proche de zéro, typiquement 1%.

Il n'y a rien d'autre à savoir sur les intervalles de confiance ; tout l'art de la chose consiste à savoir les construire. Commençons par des exemples essentiels à plusieurs titres : le cas d'un échantillon gaussien, et le cas de lois de Bernoulli.

## 4.2 Exemples gaussiens

On dispose de variables aléatoires $X_1, \dots, X_n$ de loi $N(\mu, \sigma^2)$. On va donner des intervalles de confiance pour l'estimation des paramètres $\mu$ et $\sigma$ dans plusieurs cas de figure.

### 4.2.1 Estimation de $\mu$

**Lorsque $\sigma$ est connue.**

Nous avons déjà vu que la moyenne empirique $\bar{X}_n$ est un estimateur sans biais de $\mu$. Or, nous savons aussi la loi *exacte* de $\bar{X}_n$, qui est $N(\mu, \sigma^2/n)$. Autrement dit,

$$\frac{\sqrt{n}}{\sigma}(\bar{X}_n - \mu) \sim N(0, 1). \tag{4.1}$$

Dans cette équation, on a trouvé une variable aléatoire dont la loi ne dépend plus de $\mu$. Il est donc possible de déterminer un intervalle dans lequel elle fluctue à l'aide des quantiles de la loi normale, qui sont étudiés dans Section 5.1. Si l'on se donne une marge d'erreur $\alpha = 1\%$, alors

$$\mathbb{P}((\sqrt{n}/\sigma)|\bar{X}_n - \mu| > z_{0.99}) = 1\%$$

où $z_{0.99} \approx 2.32$. Or,

$$\frac{\sqrt{n}}{\sigma}|\bar{X}_n - \mu| > z_{0.99} \tag{4.2}$$

est équivalent à

$$\bar{X}_n - \frac{z_{0.99}\sigma}{\sqrt{n}} \leqslant \mu \leqslant \bar{X}_n + \frac{z_{0.99}\sigma}{\sqrt{n}}. \tag{4.3}$$

Le passage de Équation 4.2 à Équation 4.3 est souvent appelé *pivot* et sert à passer d'un intervalle de fluctuation à un intervalle de confiance.

Nous avons donc les deux bornes de notre intervalle de confiance :

$$A = \bar{X}_n - \frac{z_{0.99}\sigma}{\sqrt{n}}$$

$$B = \bar{X}_n + \frac{z_{0.99}\sigma}{\sqrt{n}}.$$

Ces deux quantités sont bien des statistiques, car $\sigma$ est connu. De plus, nous venons de montrer que $P_\mu(\mu \in [A, B]) = 99\%$. Ici, le choix de la marge d'erreur $\alpha = 1\%$ ne jouait aucun rôle particulier ; ainsi, un intervalle de confiance de niveau $1 - \alpha$ pour l'estimation de $\mu$ est donné par

$$\left[\bar{X}_n - \frac{z_{1-\alpha}\sigma}{\sqrt{n}} \;\; ; \;\; \bar{X}_n + \frac{z_{1-\alpha}\sigma}{\sqrt{n}}\right]. \tag{4.4}$$

**Lorsque $\sigma$ est inconnue.**

Lorsque $\sigma$ n'est pas connu, les bornes $A, B$ ci-dessus ne sont pas des statistiques, car elles dépendent de $\sigma$. Heureusement, on peut estimer $\sigma$ sans biais via l'estimateur

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

Que se passe-t-il si, dans la statistique Équation 4.1, on remplace $\sigma$ par son estimation $\hat{\sigma}_n^2$ ? On obtient la statistique dite *de Student*,

$$T_n = \frac{\sqrt{n}}{\sqrt{\hat{\sigma}_n^2}}(\bar{X}_n - \mu). \tag{4.5}$$

Sa loi n'est plus une loi gaussienne, mais une loi de Student à $n-1$ paramètres de liberté $\mathscr{T}(n-1)$: le calcul de la densité est fait en détails dans Section 5.2.3 - Section 5.2.4. Les quantiles des lois de Student ont été calculés avec précision. On notera $t_{k,\alpha}$ le quantile symétrique de niveau $\alpha$ de $\mathscr{T}(k)$. Alors,

$$P_{\mu,\sigma^2}(|T_n| > t_{n-1,\alpha}) \leqslant \alpha.$$

Par le même raisonnement que tout à l'heure, l'inégalité

$$\left|\frac{\sqrt{n}}{\hat{\sigma}_n}(\bar{X}_n - \mu)\right| > t_{n-1,\alpha}$$

est équivalente à

$$\bar{X}_n - \frac{t_{n-1,\alpha}\hat{\sigma}_n}{\sqrt{n}} \leqslant \mu \leqslant \bar{X}_n + \frac{t_{n-1,\alpha}\hat{\sigma}_n}{\sqrt{n}}.$$

et les deux côtés de ces inégalités sont des statistiques; en les notant $A, B$, on a bien trouvé un intervalle de confiance de niveau $\alpha$, c'est-à-dire tel que $P_{\mu,\sigma^2}(\mu \in [A, B]) = \alpha$. Cet intervalle de confiance est d'une grande importance en pratique et mérite son propre théorème. Il est dû à William Gosset.

**Théorème 4.1** (Intervalle de Student)**.** *Un intervalle de confiance de niveau $1-\alpha$ pour l'estimation de $\mu$ lorsque $\sigma$ n'est pas connue est donné par*

$$\left[\bar{X}_n - \frac{t_{n-1,1-\alpha}\hat{\sigma}_n}{\sqrt{n}} \quad ; \quad \bar{X}_n + \frac{t_{n-1,1-\alpha}\hat{\sigma}_n}{\sqrt{n}}\right].$$

### 4.2.2 Estimation de $\sigma$

Supposons maintenant qu'on désire estimer la variance $\sigma^2$.

**Lorsque $\mu$ est connue.**

En supposant que $\mu$ est connue, l'estimateur des moments le plus naturel pour estimer $\sigma^2$ est évidemment

$$\tilde{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2.$$

Comme les $(X_i - \mu)/\sigma$ sont des variables aléatoires gaussiennes centrées réduites, l'estimateur $\tilde{\sigma}_n^2 \times (n/\sigma^2)$ est une somme de $n$ gaussiennes standard indépendantes. La loi de cette statistique est connue : c'est une loi du chi-deux à $n$ paramètres de liberté comme démontré dans Section 5.2.2. Cette loi n'est pas symétrique, puisqu'elle est supportée sur $[0, \infty[$. On note souvent $k_{n,\alpha}^-$ et $k_{n,\alpha}^+$ les nombres les plus éloignées possible (ils exisent) tels que $\mathbb{P}(k_{n,\alpha}^- < \chi^2(n) < k_{n,\alpha}^+) = 1 - \alpha$. Ainsi,

$$P_{\sigma^2}(k_{n,\alpha}^- < \frac{n\tilde{\sigma}_n^2}{\sigma^2} < k_{n,\alpha}^+) = \alpha.$$

En pivotant comme dans les exemples précédents, on obtient que l'intervalle

$$\left[\frac{n\tilde{\sigma}_n^2}{k_{n,\alpha}^+} \quad ; \quad \frac{n\tilde{\sigma}_n^2}{k_{n,\alpha}^-}\right]$$

est un intervalle de confiance de niveau $\alpha$ pour $\sigma^2$.

**Lorsque $\mu$ est inconnue.**

Cette fois, on utilise l'estimateur déjà évoqué plus tôt, à savoir

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

La loi de $(n-1)\hat{\sigma}_n^2/\sigma^2$ est encore une loi du chi-deux, mais à $n-1$ paramètres de liberté. Ainsi, le même raisonnement que ci-dessus donne l'intervalle de confiance de niveau $\alpha$ suivant :

$$\left[\frac{(n-1)\hat{\sigma}_n^2}{k_{n-1,\alpha}^+} \quad ; \quad \frac{(n-1)\hat{\sigma}_n^2}{k_{n-1,\alpha}^-}\right].$$

## 4.3 Exemples asymptotiques

### 4.3.1 Estimation du paramètre $p$ dans un modèle de Bernoulli.

Soient $X_1, \dots, X_n$ des variables indépendantes de loi $\mathscr{B}(p)$, dont on cherche à estimer le paramètre $p \in ]0,1[$. Un estimateur naturel est donné par la moyenne empirique, $\hat{p}_n = (X_1 + \cdots + X_n)/n$. Cet estimateur est non biaisé et son risque quadratique est égal à $p(1-p)/n$. De plus, la loi de $\hat{p}_n$ est connue : $n\hat{p}_n \sim \mathrm{Bin}(n,p)$. Par conséquent, si l'on connaît les quantiles de $\mathscr{B}(n,p)-p$, on pourra construire des intervalles de confiance de niveau $1-\alpha$. Ces quantiles peuvent être calculés par des méthodes numériques, mais il existe des façons plus simples de faire.

**Inégalité BT.** L'inégalité de Bienaymé-Tchebychev dit que

$$P_p(|\hat{p}_n - p| > t) \leqslant \frac{p(1-p)}{nt^2}. \tag{4.6}$$

Si l'on choisit

$$t = \sqrt{\frac{p(1-p)}{n\alpha}},$$

cette probabilité est plus petite que $\alpha$. En pivotant, on en déduit que l'intervalle $[\hat{p_n} \pm \sqrt{p(1-p)/n\alpha}]$ contient $p$ avec une probabilité supérieure à $1-\alpha$. Mais les bornes de cet intervalle ne sont pas des statistiques, car elles dépendent de $p$ ! Fort heureusement, on sait que $p$ est entre 0 et 1, ce qui entraîne que $p(1-p)$ est plus petit que $1/4$, donc l'intervalle ci-dessus est contenu dans l'intervalle plus grand

$$\left[\hat{p}_n \pm \frac{1}{2\sqrt{n\alpha}}\right].$$

Ce dernier est bien un intervalle de confiance de niveau $1-\alpha$ pour l'estimation de $p$.

**TCL.** On a mentionné que les quantiles des lois binomiales pourraient être calculés ; or, ils peuvent également être approchés grâce au théorème central-limite. Celui-ci dit que

$$\frac{\sqrt{n}(\hat{p}_n - p)}{\sqrt{p(1-p)}} \to N(0,1). \tag{4.7}$$

Si $z_\alpha$ est le quantile symétrique d'ordre $\alpha$ de $N(0,1)$, alors on en déduit que

$$\mathbb{P}\left(\left|\frac{\sqrt{n}(\hat{p}_n - p)}{\sqrt{p(1-p)}}\right| > z_\alpha\right) \to \alpha.$$

En pivotant, on voit alors que l'intervalle

$$\left[\hat{p}_n \pm z_\alpha \sqrt{p(1-p)/n}\right]$$

contient $p$ avec une probabilité *qui tend lorsque* $n \to \infty$ *vers* $1-\alpha$. Là encore, cet intervalle n'est pas un intervalle de confiance. On pourrait utiliser deux techniques.

1. Comme tout à l'heure, l'intervalle ci-dessus est contenu dans l'intervalle plus grand $[\hat{p}_n \pm z_\alpha/2\sqrt{n}]$ qui est un intervalle de confiance *asymptotique* de niveau $1-\alpha$.

2. Il y a plus fin. Nous savons par la loi des grands nombres que $\hat{p}_n \to p$ en probabilité. Ainsi, $\sqrt{\hat{p}_n(1-\hat{p}_n)} \to \sqrt{p(1-p)}$ en probabilité. Le lemme de Slutsky nous assure alors que dans Équation 4.8, on peut remplacer le dénominateur par $\sqrt{\hat{p}_n(1-\hat{p}_n)}$ pour obtenir

$$\frac{\sqrt{n}(\hat{p}_n - p)}{\sqrt{\hat{p}_n(1-\hat{p}_n)}} \to N(0,1). \tag{4.8}$$

Le reste du raisonnement est identique, et l'on obtient l'intervalle de confiance asymptotique de niveau $1-\alpha$ suivant :

$$\left[\hat{p}_n \pm z_\alpha \sqrt{\frac{\hat{p}_n(1-\hat{p}_n)}{n}}\right]$$

**Hoeffding.** L'inégalité de Bienaymé-Tchebychev n'est pas très fine. Il existe de nombreuses autres inégalités de concentration : l'inégalité de Hoeffding (Théorème 5.4) concerne les variables bornées, comme ici où les $X_i$ sont dans $[0,1]$ . Cette inégalité dit que

$$\mathbb{P}(|\hat{p}_n - p| > t) \leqslant 2e^{-2nt^2}.$$

Le choix

$$t = \sqrt{\frac{1}{2n}\ln\left(\frac{2}{\alpha}\right)}$$

donne une probabilité inférieure à $\alpha$, et fournit donc l'intervalle de confiance **non-asymptotique** de niveau $1-\alpha$ suivant :

$$\left[\bar{X}_n \pm \frac{\ln(2/\alpha)}{\sqrt{2n}}\right].$$

### 4.3.2 Estimation de moyenne dans un modèle non-gaussien.

Les deux techniques ci-dessus n'ont rien de spécifique au cas de variables de Bernoulli. En fait, elles s'appliquent à tout modèle statistique iid dont on cherche à estimer la moyenne $\mu$, pourvu que la variance existe.

La première méthode utilisant Bienaymé-Tchebychev nécessite de borner la variance. Cela peut se faire dans certains cas, mais pas dans tous.

La seconde méthode s'applique systématiquement en utilisant l'estimateur de la variance empirique $\hat{\sigma}_n^2$. En effet, la convergence

$$\frac{\sqrt{n}}{\hat{\sigma}_n}(\bar{X}_n - \mu) \to N(0,1)$$

est toujours vraie d'après le théorème de Slutsky.

**Théorème 4.2.** *Soient $X_1, \ldots, X_n$ des variables iid possédant une variance. L'intervalle*

$$\left[\bar{X}_n \pm \frac{z_\alpha \hat{\sigma}_n}{\sqrt{n}}\right]$$

*est un intervalle de confiance asymptotique de niveau $\alpha$ pour l'estimation de la moyenne des $X_i$.*

# 5 Outils pour les IC

## 5.1 Quantiles

Si $X$ est une variable aléatoire sur $\mathbb{R}$, un quantile d'ordre $\beta \in ]0,1[$, noté $q_\beta$, est un nombre tel que $\mathbb{P}(X \leqslant q_\beta) = \beta$. Lorsque $X$ est continue, un tel nombre existe forcément, car la fonction de répartition $F(x) = \mathbb{P}(X \leqslant x)$ est une surjection continue. Les quantiles symétriques $z_\beta$ sont, eux, définis par $\mathbb{P}(|X| \leqslant z_\beta) = \beta$.

Si la loi de $X$ est de surcroît symétrique, les quantiles symétriques s'expriment facilement en fonction des quantiles classiques. En effet, $\mathbb{P}(|X| \leqslant z)$ est égal à $\mathbb{P}(X \leqslant z) - \mathbb{P}(X \leqslant -z)$. Or, si la loi de $X$ est symétrique, alors $\mathbb{P}(X \leqslant -z) = 1 - \mathbb{P}(X \leqslant z)$, et donc

$$\mathbb{P}(|X| \leqslant z) = 2\mathbb{P}(X \leqslant z) - 1.$$

Il suffit alors de choisir pour $z$ le quantile $q_{\frac{1+\beta}{2}}$ pour obtenir $\mathbb{P}(|X| \leqslant z) = \beta$. Lorsque $\beta$ est de la forme $1-\alpha$ avec $\alpha$ petit (comme les niveaux des intervalles de confiance), on trouve alors $z_{1-\alpha} = q_{1-\alpha/2}$.

Les quantiles s'obtiennent en inversant la fonction de répartition : lorsque celle-ci est une bijection sur $]0,1[$, alors $q_\beta = F^{-1}(\beta)$. En règle générale, il n'y a pas de forme fermée. Par exemple, pour une loi gaussienne standard,

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} du$$

qui elle-même n'a pas d'écriture plus simple. Fort heureusement, les outils de calcul numérique permettent d'effectuer ces calculs avec une grande précision. La table suivante donne les quantiles symétriques de la gaussienne.

| $\beta$ | 90% | 95% | 98% | 99% | 99.9% | 99.99999% |
|---|---|---|---|---|---|---|
| $z_\beta$ | 1.64 | 1.96 | 2.32 | 2.57 | 3.2 | 5.32 |

Voir aussi la règle 1-2-3. Il existe de nombreuses tables de quantiles pour les lois usuelles.

**Théorème 5.1** (Queues de distribution de la gaussienne)**.** *Si $x$ est plus grand que 1,*

$$\left(\frac{1}{x} - \frac{1}{x^3}\right) \frac{e^{-x^2/2}}{\sqrt{2\pi}} \leqslant \mathbb{P}(X > x) \leqslant \frac{1}{x} \frac{e^{-x^2/2}}{\sqrt{2\pi}}$$

*En particulier, si $x$ est grand, $\mathbb{P}(X \geqslant x) \sim e^{-x^2/2}/x\sqrt{2\pi}$ avec une erreur d'ordre $O(e^{-x^2/2}/x^3)$.*

À titre d'exemple, pour $x = 2.32$ cette approximation donne 98.83%, ce qui est remarquablement proche de 98%. Pour $x = 2.57$ on trouve 99.42%.

*Démonstration.* À écrire.

□

## 5.2 Calculs de lois

### 5.2.1 Lois Gamma

Une variable aléatoire suit une loi Gamma de paramètres $\lambda > 0, \alpha > 0$ lorsque sa densité est donnée par

$$\gamma_{r,\alpha}(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} e^{-\lambda x} x^{\alpha-1} \mathbf{1}_{x>0}.$$

Les lois Gamma rassemblent les lois exponentielles ($\Gamma(\lambda, 1) = \mathscr{E}(\lambda)$) et les lois du chi-deux qu'on verra ci-dessous ($\Gamma(1/2, n/2) = \chi_2(n)$). La transformée de Fourier $\varphi_{\lambda,\alpha}$ d'une loi $\Gamma(\lambda, \alpha)$ se calcule facilement par un changement de variables :

$$\varphi_{\lambda,\alpha}(t) = \left(1 - \frac{it}{\lambda}\right)^{-\alpha}. \tag{5.1}$$

Cette identité montre également que si $X_1, \dots, X_n$ sont des variables indépendantes de loi $\Gamma(\lambda, \alpha_i)$, alors leur somme est une variable de loi $\Gamma(\lambda, \alpha_1 + \dots + \alpha_n)$.

### 5.2.2 Loi du chi-deux

Soit $X$ une loi gaussienne standard. Calculons la densité de $X^2$ ; pour toute fonction-test $\varphi$, $\mathbb{E}[\varphi(X^2)]$ est donné par

$$\frac{1}{\sqrt{2\pi}} \int e^{-x^2/2} \varphi(x^2) dx.$$

Cette intégrale est symétrique, donc on peut ajouter un facteur 2 et intégrer sur $[0, \infty[$. En posant $u = x^2$, on obtient alors la valeur

$$\frac{2}{\sqrt{2\pi}} \int_0^\infty e^{-u/2} \varphi(u) \frac{1}{2\sqrt{u}} du.$$

On reconnaît la densité d'une loi Gamma de paramètres $(1/2, 1/2)$. Cette loi s'appelle *loi du chi-deux* et on la note $\chi_2(1)$. Sa tranformée de Fourier est donnée par

$$\mathbb{E}[e^{itX^2}] = \frac{1}{\sqrt{1 - 2it}}.$$

Soient maintenant $X_1, \dots, X_n$ des variables de loi $N(0, 1)$ indépendantes. Chaque $X_i^2$ est une $\chi_2(1)$ ; leur somme a pour loi la convolée $n$ fois de $\chi_2(1)$. Calculons sa tranformée de Fourier :

$$\mathbb{E}[e^{it(X_1^2 + \dots + X_n^2)}] = \mathbb{E}[e^{itX_1^2}]^n \tag{5.2}$$

$$= (1 - 2it)^{-\frac{n}{2}}. \tag{5.3}$$

On reconnaît la transformée de Fourier d'une loi $\Gamma(1/2, n/2)$ ; cette loi s'appelle *loi du chi-deux à $n$ paramètres de liberté* et elle est notée $\chi_2(n)$. Sa densité est donnée par

$$\frac{1}{2^{n/2}\Gamma(n/2)} e^{-x/2} x^{n/2-1} \mathbf{1}_{x>0}. \tag{5.4}$$

### 5.2.3 Loi de Student

Soit $X$ une variable de loi $N(0,1)$ et $Y_n$ une variable de loi $\chi_2(n)$ indépendante de $X$. On va calculer la loi de $T_n = X/\sqrt{Y_n/n}$. Soit $\varphi$ une fonction test ; l'espérance $\mathbb{E}[\varphi(T_n)]$ est égale à

$$\frac{1}{Z_n\sqrt{2\pi}} \int_0^\infty \int_{-\infty}^\infty \varphi\left(\frac{x}{\sqrt{y/n}}\right) e^{-\frac{x^2}{2}} e^{-\frac{y}{2}} y^{\frac{n}{2}-1} dxdy$$

où $Z_n = 2^{n/2}\Gamma(n/2)$. Dans l'intégrale en $x$, on effectue le changement de variable $u = x/\sqrt{y/n}$ afin d'obtenir

$$\frac{1}{Z_n\sqrt{2\pi}} \int_0^\infty \int_{-\infty}^\infty \varphi(u) e^{-\frac{yu^2}{2n}} e^{-\frac{y}{2}} y^{\frac{n}{2}-1} \sqrt{\frac{y}{n}} dxdy.$$

La densité de $T_n$ est donc donnée par

$$t_n(u) = \frac{1}{Z_n\sqrt{2\pi n}} \int_0^\infty e^{-\frac{yu^2}{2n}-\frac{y}{2}} y^{\frac{n+1}{2}-1} dy.$$

Le changement de variables $z = y(1+u^2/n)/2$ nous ramène à

$$t_n(u) = \frac{1}{Z_n\sqrt{2\pi n}} \left(\frac{2}{1+\frac{u^2}{n}}\right)^{\frac{n+1}{2}} \int_0^\infty e^{-z} z^{\frac{n+1}{2}-1} dz.$$

On reconnaît $\Gamma((n+1)/2)$ à droite. La densité $t_n(x)$ est donc

$$t_n(x) = \frac{1}{\sqrt{n\pi}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \left(\frac{1}{1+\frac{x^2}{n}}\right)^{\frac{n+1}{2}}.$$

Cette loi s'appelle *loi de Student* de paramètre $n$ ; on dit parfois *à $n$ degrés de liberté*. Elle est notée $\mathscr{T}(n)$. La loi de Student de paramètre $n = 1$ est tout simplement une loi de Cauchy.

### 5.2.4 Loi de la statistique de Student

Soient $X_1, \ldots, X_n$ des variables gaussiennes $N(\mu, \sigma^2)$ indépendantes, et soit $T_n = \sqrt{n}(\bar{X}_n - \mu)/\sqrt{\hat{\sigma}_n^2}$, où

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

**Théorème 5.2.**

$$T_n \sim \mathscr{T}(n-1).$$

*Démonstration.* On va montrer 1° que $\bar{X}_n$ et $\sqrt{\hat{\sigma}_n^2/\sigma^2}$ sont indépendantes, et 2° que $\sqrt{\hat{\sigma}_n^2/\sigma^2}$ a bien la même loi que $\sqrt{Y_{n-1}/(n-1)}$ où $Y_{n-1}$ est une $\chi_2(n-1)$. Dans la suite, on supposera que $\mu = 0$ et $\sigma = 1$, ce qui n'enlève rien en généralité.

*Premier point.* Le vecteur $X = (X_1, ..., X_n)$ est gaussien. Posons $Z = (X_1 - \bar{X}_n, ..., X_n - \bar{X}_n)$. Le couple $(\bar{X}_n, Z_n)$ est linéaire en $X$, donc ce couple est aussi un vecteur gaussien. Or, la covariance de ses deux éléments est nulle. Par exemple, $\mathrm{Cov}(\bar{X}_n, Z_1)$ est égale à $\mathrm{Cov}(\bar{X}_n, X_1) - \mathrm{Var}(\bar{X}_n)$, ce qui par linéarité donne $1/n - 1/n = 0$. Ainsi, $\bar{X}_n$ et $Z$ sont deux variables conjointement gaussiennes et décorrélées : elles sont donc indépendantes. Comme $\hat{\sigma}_n$ est une fonction de $Z$, elle est aussi indépendante de $\bar{X}_n$.

*Second point.* $Z$ est la projection orthogonale de $X$ sur le sous-espace vectoriel $\mathscr{V} = \{x \in \mathbb{R}^n : x_1 + ... + x_n = 0\}$. Soit $(f_i)_{i=2,...,n}$ une base orthonormale de $\mathscr{V}$, de sorte que $Z = \sum_{i=2}^n \langle f_i, X\rangle f_i$. Par l'identité de Parseval,

$$|Z|^2 = \sum_{i=2}^{n} |\langle f_i, X\rangle|^2.$$

Or, les $n-1$ variables aléatoires $G_i = \langle f_i, X\rangle$ sont des gaussiennes standard iid. En effet, on vérifie facilement que $\mathrm{Cov}(G_i, G_j) = \langle f_i, f_j\rangle = \delta_{i,j}$. On en déduit donc que $|Z|^2$ suit une loi $\chi_2(n-1)$.

□

La seconde partie de la démonstration est un cas particulier du théorème de Cochran, que nous verrons dans le chapitre sur la régression linéaire.

## 5.3 Inégalités de concentration

Les outils de base pour construire des intervalles de confiance dans des circonstances générales (non gaussiennes) sont les inégalités de concentration. Une inégalité de concentration pour une variable aléatoire intégrable $X$ consiste à borner $\mathbb{P}(|X - \mathbb{E}[X]| > x)$ par quelque chose de petit quand x est grand : on cherche à contrôler la probabilité pour que les réalisations de la variable aléatoire $X$ soient éloignées de leur valeur moyenne $\mathbb{E}[X]$ de plus de $x$.

## 5.4 Inégalité de Bienaymé-Tchebychev

**Théorème 5.3.** *Soit $X$ une variable aléatoire de carré intégrable. Alors,*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geqslant x) \leqslant \frac{\mathrm{Var}(X)}{x^2}.$$

*Démonstration.* Élever au carré les deux membres de l'inégalité dans $\mathbb{P}$, puis appliquer l'inégalité de Markov à la variable aléatoire positive $|X - \mathbb{E}X|^2$ dont l'espérance est $\mathrm{Var}(X)$.

□

## 5.5 Inégalité de Hoeffding

**Théorème 5.4** (Inégalité de Hoeffding)**.** *Soient $X_1, \dots, X_n$ des variables aléatoires indépendantes, pas forcément de même loi. On suppose que chaque $X_i$ est à valeurs dans un intervalle borné $[a_i, b_i]$ et on pose $S_n = X_1 + \dots + X_n$. Pour tout $t > 0$,*

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geqslant t) \leqslant e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}} \tag{5.5}$$

*et*

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geqslant t) \leqslant 2e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}. \tag{5.6}$$

La démonstration se fonde sur le lemme suivant.

**Lemme 5.1** (lemme de Hoeffding)**.** *Soit $X$ une variable aléatoire à valeurs dans $[a, b]$. Pour tout $t$,*

$$\mathbb{E}[e^{t(X - \mathbb{E}[X])}] \leqslant e^{\frac{t^2 (b-a)^2}{8}}. \tag{5.7}$$

*Démonstration.* Soit $X$ une variable aléatoire, que par simplicité on supposera centrée et à valeurs dans l'intervalle $[a, b]$ ($a$ est forcément négatif). En écrivant

$$x = a \times \frac{b - x}{b - a} + b \times \left(1 - \frac{b - x}{b - a}\right)$$

et en utilisant la convexité de la fonction $x \mapsto e^{tx}$, on obtient $e^{tX} \leqslant (b - X)e^{ta}/(b - a) + (1 - (b - x)/(b - a))e^{bt}$), puis en prenant l'espérance et le fait que $X$ est centrée et en simplifiant,

$$\mathbb{E}[e^{tX}] \leqslant \frac{be^{ta} - ae^{tb}}{b - a}.$$

Notons $f(t)$ le terme à droite ; pour montrer Équation 5.7, il suffit de montrer que $\ln f(t) \leqslant t^2(b - a)^2/8$. La formule de Taylor dit que

$$\ln f(t) = \ln f(0) + t(\ln f)'(0) + \frac{t^2}{2}(\ln f)''(\xi)$$

pour un certain $\xi$. Or, $\ln f(0) = \ln 1 = 0$, $(\ln f)'(0) = f'(0)/f(0) = 0$, et il suffit donc de montrer que $(\ln f)''(t)$ est toujours plus petit que $(b - a)^2/4$ pour conclure. Un simple calcul montre que $\ln f(t) = \ln(b/(b - a)) + ta + \ln(1 - ae^{t(b-a)}/b)$, et donc

$$(\ln f)''(t) = \frac{(a/b)(b - a)e^{t(b-a)}}{(1 - ae^{t(b-a)}/b)^2}.$$

L'inégalité $uv/(u - v)^2 \leqslant 1/4$ appliquée à $u = a/b$ et $v = e^{t(b-a)}$ permet alors de conclure.

□

*Preuve de l'inégalité de Hoeffding.* En remplaçant $X_k$ par $X_k - \mathbb{E}[X_k]$, on peut supposer que tous les $X_i$ sont centrés et étudier seulement $\mathbb{P}(S_n > t)$. Écrivons $\mathbb{P}(S_n > t) = \mathbb{P}(e^{\lambda S_n} > e^{\lambda t})$, où $\lambda$ est un nombre positif que l'on choisira plus tard. L'inégalité de Markov borne cette probabilité par $\mathbb{E}[e^{\lambda S_n}]e^{-\lambda t}$. Comme les $X_i$ sont indépendantes, $\mathbb{E}[e^{tS_n}]$ est le produit des $e^{\varphi_k(\lambda)}$ où $\varphi_k(t) = \ln \mathbb{E}[e^{itX_k}]$. En appliquant le lemme de Hoeffding à chaque $\varphi_k$, on borne $\mathbb{P}(S_n > t)$ par

$$\exp\left(\sum_{i=1}^{n} \frac{(b_i - a_i)^2\lambda^2}{8} - t\lambda\right).$$

Le minimum en $\lambda$ du terme dans l'exponentielle est atteint au point $4t/\sum(a_i - b_i)^2$ et la valeur du minimum est le terme dans l'exponentielle de Équation 5.5. On déduit Équation 5.6 par une simple borne de l'union.

La démonstration de l'inégalité de Hoeffding ne dépend pas directement du fait que $X$ est bornée, mais plutôt de Équation 5.7. Toutes les variables aléatoires qui vérifient une inégalité de type $\mathbb{E}[e^{tX}] \leqslant e^{ct^2}$ pour une constante $c$ peuvent donc avoir leur propre inégalité de Hoeffding.

# Exercices

## Questions

1. Soit $X_n$ une variable aléatoire de loi de Student de paramètre $n$. Montrer que $X_n$ converge en loi vers $N(0,1)$.
2. Soit $X_n \sim \chi_2(n)$. La suite $(X_n)$ est-elle asymptotiquement normale ?
3. Donner un intervalle de confiance de la forme $[A, +\infty[$ pour la moyenne d'un échantillon gaussien.
4. Même question pour la variance dans un modèle gaussien centré.
5. Dans l'estimation de la moyenne $\mu$ d'un modèle gaussien où la variance $\sigma^2$ est connue, montrer que l'intervalle de confiance obtenu (Équation 4.4) est le plus grand possible de niveau $1-\alpha$.
6. Démontrer le théorème Théorème 5.1 sur l'asymptotique des queues de distribution de la loi gaussienne.
7. Montrer la borne suivante sur les quantiles de loi gaussienne standard: $q_\beta < \sqrt{\ln \frac{1}{\beta\sqrt{2\pi}}}$ (pour tout $1/2 < \beta < 1$).

8. Comparer les queues de distribution des lois $N(0,1), \chi_2(n)$ et $\mathscr{T}(n)$.
9. Expliquer à votre grand-mère la différence entre un intervalle de fluctuation et un intervalle de confiance.
10. L'intervalle de confiance de niveau $1-\alpha$ pour la moyenne d'un modèle $N(\mu,1)$ avec $n$ observations est $I_n = [\bar{X}_n \pm z_\alpha/\sqrt{n}]$. Supposons qu'on obtienne une nouvelle observation indépendante des autres, disons $Z$. La probabilité $\mathbb{P}(Z \in I_n)$ est-elle plus grande ou plus petite que $1-\alpha$ ?
11. Comparer la longueur des intervalles de confiance obtenus par les différentes méthodes de la section Section 4.3.1.
12. Le quantile d'ordre 97,5% d'une variable $X \sim \text{Bin}(12, 1/2)$ est 9. Trouver $c$ tel que $\mathbb{P}(|X-6| > c) = 95\%$.

## Exercices

**Exercice 5.1** (Lois de Poisson)**.** On suppose que l'on observe $X_1, \dots, X_n$ i.i.d de loi $\mathscr{P}(\theta)$.

1. Étudier $\bar{X}_n$.
2. Montrer que $\sqrt{\bar{X}_n} \xrightarrow[n\to\infty]{\mathbb{P}} \sqrt{\theta}$.
3. Donner deux intervalles de confiance au niveau 98% pour $\sqrt{\theta}$, et les comparer.

**Exercice 5.2** (Lois uniformes)**.** Soit $X_1, \dots, X_n$ des variables aléatoires iid de loi $\mathscr{U}[0,\theta]$. Donner un intervalle de confiance non asymptotique pour $\theta$ en utilisant l'estimateur $\hat{\theta}_n = \max_{i=1,\dots,n} X_i$.

**Exercice 5.3** (Lois exponentielles décalées)**.** Soit $X_1, \dots, X_n$ des variables aléatoires iid de densité $e^{\theta - x}\mathbf{1}_{x>\theta}$, où $\theta > 0$.

1. Calculer $\mathbb{E}_\theta[X_1]$ et en déduire un estimateur de $\theta$ que l'on notera $\widehat{\theta}_n$. Étudier ses propriétés (risque quadratique, convergence) et l'utiliser pour construire un premier intervalle de confiance $I_1(\alpha)$ non-asymptotique pour $\theta$ de niveau $1-\alpha$.
2. Construire un intervalle de confiance asymptotique $I_2(\alpha)$ pour $\theta$ à partir de $\widehat{\theta}_n$.
3. Montrer que l'estimateur $\theta_n^\star := \min_{1\le i\le n} X_i$ est meilleur que $\widehat{\theta}_n$ au sens du risque quadratique, puis l'utiliser pour construire un intervalle de confiance $I_3(\alpha)$ de niveau $1-\alpha$.
4. Comparer les longueurs de tous ces différents intervalles de confiance.

**Exercice 5.4** (Lois exponentielles)**.** Soit $X_1, \dots, X_n$ des variables aléatoires iid exponentielles de paramètre $\lambda > 0$.

1. Quelle est la loi de $S_n = X_1 + \cdots + X_n$ ?
2. Construire un intervalle de confiance de niveau $1-\alpha$ pour $\lambda$.

**Exercice 5.5** (Inégalité d'Azuma)**.** Montrer que l'inégalité de Hoeffding reste valable lorsque les $X_i$ ne sont plus supposés indépendants, mais que la suite $S_k = X_1 + \cdots + X_k$ est une martingale. Indice : $\mathbb{E}[e^{\lambda S_{n+1}}] = \mathbb{E}[e^{\lambda S_n}\mathbb{E}[e^{\lambda X_{n+1}}|S_n]]$.

Ce raffinement s'appelle *inégalite de Hoeffding-Azuma*. C'est celui que nous avons utilisé dans l'exercice (**ex-tanks?**), lorsque les $X_1, \dots, X_n$ sont des tirages *sans remise* dans une urne à $N$ éléments.

# 6 Test d'hypothèses

Si l'on essaie d'estimer le rendement $\mu$ d'un actif financier, on cherche implicitement à savoir si l'on va investir ou pas. Cette décision dépendra de notre estimation : pour faire simple, on peut considérer que si nous estimons que le rendement est positif ($\hat{\mu} > 0$), alors il faut investir. Sinon, on n'investira pas.

Les tests d'hypothèses visent à formaliser cela. Faire une *hypothèse* dans un modèle statistique $(\mathcal{X}, \mathscr{F}, (P_\theta)_{\theta \in \Theta_0})$, c'est supposer que $\theta$ appartient à une certaine région de $H_0 \subset \Theta$. Les *tests* visent à construire des procédures pour tester une hypothèse nulle, que l'on notera $H_0$, contre une hypothèse alternative, notée $H_1$.

Dans le cadre ci-dessus, on peut se placer dans un modèle où les rendements sont $\mathcal{N}(\mu, \sigma^2)$. On veut tester l'hypothèse nulle $H_0 : \ \mu \in ]-\infty, 0]$ contre l'hypothèse alternative $H_1 : \ \mu \in ]0, +\infty[$.

**Définition 6.1.** Un test est un événement qui, s'il survient, nous incite à rejeter l'hypothèse nulle. Cet événement sera noté rejeter et son complémentaire sera noté accepter.

- L'erreur de première espèce est la probabilité de rejeter l'hypothèse nulle à tort : $\alpha = \sup_{\theta \in H_0} P_\theta(\text{rejeter})$. Le **niveau d'un test** est $1 - \alpha$. C'est la probabilité d'accepter l'hypothèse nulle à raison.
- L'erreur de seconde espèce est la probabilité de ne pas rejeter l'hypothèse nulle, à tort : $\beta = \sup_{\theta \in H_1} P_\theta(\text{accepter})$. La **puissance d'un test** est $1 - \beta$. C'est la probabilité de « détecter » l'hypothèse alternative à raison.
- L'affinité d'un test est la somme des erreurs de première et seconde espèce. On parle aussi de l'*erreur totale*.

Par « événement », on veut bien dire « un élément de $\mathscr{F}$ », c'est-à-dire qui n'est déterminé que par les observations et pas par $\theta$. Formellement on écrit souvent qu'un test est une statistique, disons $T$, à valeurs dans $\{0, 1\}$. L'événement $\{T = 1\}$ est rejeter, l'événement $\{T = 0\}$ est accepter.

Un des grands objectifs de la statistique mathématique est de construire des familles de tests qui, pour un niveau de confiance $1 - \alpha$ fixé, ont la plus grande puissance possible ; autrement dit, **trouver un événement hautement improbable sous l'hypothèse nulle, et hautement probable sous l'hypothèse alternative**.

Comme on verra dans les exemples, le rôle des deux hypothèses n'est pas interchangeable. Maximiser le niveau et la puissance ne reviennent pas au même. Le choix des hypothèses $H_0$ et $H_1$ n'est pas anodin : l'hypothèse $H_0$ est une hypothèse que l'on cherche implicitement à réfuter.

1. Si $\theta \in H_0$ quel qu'il soit, les probablités pour qu'un certain événement rejeter sont infimes – disons, 1%.
2. Si cet événement arrive, par contraposée, on est amenés à rejeter l'hypothèse selon laquelle $\theta$ est dans $H_0$.

C'est pour cela que les tests sont une forme de logique statistique. Le raisonnement de base une contraposée : en logique, $A \Rightarrow B$ est équivalent à $\neg B \Rightarrow \neg A$. En statistiques, on pourrait écrire $\theta \in H_0 \Rightarrow$ accepter (avec grande probabilité), donc rejeter $\Rightarrow \theta \notin H_0$ (probablement).

## 6.1 Exemples de tests gaussiens

On se place dans un modèle où $X_1, \dots, X_n$ sont des gaussiennes $N(\mu, \sigma^2)$. Nous avons déjà vu plusieurs fois que $\bar{X}_n \sim N(\mu, \sigma^2/n)$.

### 6.1.1 Construction du test

On cherche à réfuter l'hypothèse selon laquelle ces variables aléatoires sont centrées ; autrement dit, on posera $H_0 = \{\mu = 0\}$. Sous cette hypothèse, nos variables aléatoires sont donc des variables $N(0, \sigma^2)$.

**Supposons dans un premier temps que $\sigma^2$ est connue.** Sous $H_0$, on a donc

$$\frac{\sqrt{n}\bar{X}_n}{\sigma} \sim N(0,1)$$

et par conséquent, $P_0(|\bar{X}_n| < z_{1-\alpha}\sigma/\sqrt{n}) = 1 - \alpha$. Autrement dit, sous l'hypothèse $\mu = 0$, on devrait observer l'événement

$$\bar{X}_n \in \left[\pm \frac{z_{1-\alpha}\sigma}{\sqrt{n}}\right]$$

avec probabilité élevée $1 - \alpha$. Si cet événement n'est pas observé, il est alors très douteux que $\mu$ soit effectivement égal à zéro ! On pose donc

$$\text{rejeter}_\alpha = \{|\bar{X}_n| > z_{1-\alpha}\sigma/\sqrt{n}\}.$$

Le niveau de ce test est bien $1 - \alpha$ : nous l'avons construit pour cela.

**Supposons maintenant que $\sigma$ n'est pas connue.** En l'estimant via $\hat{\sigma}_n$, nous savons que (toujours sous l'hypothèse selon laquelle $\mu = 0$)

$$\frac{\sqrt{n}\bar{X}_n}{\hat{\sigma}_n} \sim \mathscr{T}(n-1).$$

On reproduit alors le raisonnement ci-dessus : comme $\mathbb{P}(|\bar{X}_n| < t_{n-1,1-\alpha}\hat{\sigma}_n/\sqrt{n}) = \alpha$ où $t_{n-1,1-\alpha}$ est le quantile symétrique de $\mathscr{T}(n-1)$, on voit que l'événement

$$\text{rejeter}_\alpha = \{|\bar{X}_n| > t_{n-1,1-\alpha}\hat{\sigma}_n/\sqrt{n}\}$$

est bien un test de niveau $1 - \alpha$.

### 6.1.2 Calcul de la puissance et hypothèse alternative

Nous n'avons pas encore eu besoin de spécifier une hypothèse alternative, mais nous allons en avoir besoin pour calculer la puissance du test. Pour commencer, on va supposer que, si $\mu$ n'est pas nulle, alors elle ne peut être égale qu'à 1. Autrement dit, $H_1 = \{1\}$. Ce genre d'hypothèse alternative ne peut évidemment avoir de pertinence qu'en fonction du problème réel sous-jacent !

Sous l'hypothèse alternative, donc, nous savons que $\bar{X}_n \sim N(1, \sigma^2)$. La puissance du test est définie par $1 - \beta$ où $\beta = P_1(\mathsf{accepter}_\alpha)$ c'est-à-dire

$$\begin{aligned}
\beta &= P_1(|\bar{X}_n| \leqslant z_{1-\alpha}\sigma/\sqrt{n}) && (6.1)\\
&= P_1\left(-\frac{z_{1-\alpha}\sigma}{\sqrt{n}} \leqslant \bar{X}_n \leqslant \frac{z_{1-\alpha}\sigma}{\sqrt{n}}\right) && (6.2)\\
&= P_1\left(-\frac{z_{1-\alpha}\sigma}{\sqrt{n}} - 1 \leqslant \bar{X}_n - 1 \leqslant \frac{z_{1-\alpha}\sigma}{\sqrt{n}} - 1\right) && (6.3)\\
&= \Phi(-\sqrt{n}/\sigma + z_{1-\alpha}) - \Phi(-\sqrt{n}/\sigma + z_{1-\alpha}). && (6.4)
\end{aligned}$$

où $\Phi(x) = \mathbb{P}(N(0,1) \leqslant x)$. Cette expression ne peut pas plus se simplifier, mais on peut quand même la borner par $F(-\sqrt{n}/\sigma + z_{1-\alpha})$. Lorsque $x$ est grand, nous avons vu (Théorème 5.1) que $F(x) < e^{-x^2}/|x|\sqrt{2\pi}$. Ainsi, l'erreur de première espèce est bornée par $O(e^{-n/\sigma^2/2}/\sqrt{n})$. Cela tend extrêmement vite vers 0 ; en fait, dès que $n$ est plus grand que 10 et $\sigma = 1$, cette erreur est inférieure à 0.1%, donc dans ce cas le test aura une puissance supérieure à 99.9%.

Que se serait-il passé si notre hypothèse alternative n'avait pas été $\mu = 1$ mais $\mu = m$ pour n'importe quel $m \neq 0$ ? Dans ce cas, on aurait eu $H_1 = \mathbb{R} \setminus \{0\}$. L'erreur de première espèce aurait alors été $\beta = \sup_{m\neq 0} \beta_m$ où

$$\beta_m = P_m(\mathsf{accepter}_\alpha).$$

On revoyant les calculs ci-dessus, on voit que

$$\beta_m = \Phi(-m\sqrt{n}/\sigma + z_{1-\alpha}) - \Phi(-m\sqrt{n}/\sigma + z_{1-\alpha}).$$

En particulier,

$$\lim_{m\to 0} \beta_m = \Phi(-z_{1-\alpha}) - \Phi(-z_{1-\alpha}) = 1 - \alpha$$

par continuité de $\Phi$ et par définition de $z_{1-\alpha}$. Ainsi, $1 - \beta = \alpha$ : pour cette seconde hypothèse alternative, la puissance de notre test… est extrêmement faible.

**Cela vient du fait que notre hypothèse alternative contient des situations quasiment indiscernables de notre hypothèse nulle.** Par exemple, il est quasiment impossible de distinguer $\mu = 0$ de $\mu = 10^{-100}$ par exemple. Cet exemple illustre la dissymétrie entre $H_0$ et $H_1$.

## 6.2 La notion de $p$-valeur

La construction d'un test dépend du niveau de risque $\alpha$. Si le niveau de risque acceptable est de plus en petit, alors l'événement $\mathsf{rejeter}_\alpha$ devrait être de moins en moins probable. D'ailleurs, $\mathsf{rejeter}_0 = \emptyset$ et $\mathsf{accepter}_0 = \Omega$ : si l'on ne tolère aucun niveau de risque de première espèce, c'est qu'on ne veut pas rejeter l'hypothèse nulle.

Très souvent, si $\alpha < \beta$, on a même

$$\mathsf{rejeter}_\alpha \subset \mathsf{rejeter}_\beta.$$

**Définition 6.2.** La $p$-valeur d'une famille croissante de tests est le plus petit niveau de risque qui nous amène à rejeter l'hypothèse nulle compte tenu des observations. Formellement,

$$p_\star = \inf\{\alpha > 0 : \ \mathsf{rejeter}_\alpha\} = \sup\{\alpha > 0 : \ \mathsf{accepter}_\alpha\}.$$

**La $p$-valeur dépend des observations.** C'est une observation cruciale : la $p$-valeur n'est pas une propriété intrinsèque d'un test. Sur deux ensembles différents d'observations, la $p$-valeur ne sera pas la même en général.

**Calcul de $p$-valeur.** Dans de nombreux tests, la construction d'un test se fonde sur une statistique, disons $S$, qui sous l'hypothèse nulle suit une loi particulière (par exemple, $\sqrt{n}\bar{X}_n/\hat{\sigma}_n \sim \mathscr{T}(n-1)$ sous l'hypothèse $X_i \sim N(\mu, \sigma^2)$ avec $\mu = 0$ dans le cas d'un test de Student). Si le test est de la forme $S < q_{1-\alpha}$, ce qui équivaut à $F(S) < 1 - \alpha$. La $p$-valeur est donnée par

$$p_\star = \sup\{\alpha > \ 0 : S < q_{1-\alpha}\} = \sup\{\alpha : F(S) < 1 - \alpha\} = 1 - F(S).$$

# 7 Théorie des tests simples

## 7.1 La distance en variation totale

Lorsqu'on cherche à tester une hypothèse de type loi $= P$ contre une hypothèse de type loi $= Q$ (c'est-à-dire, deux hypothèses simples), on en revient à chercher un événement très improbable sous la loi $P$, et très probable sous la loi $Q$. On peut se demander en toute généralité quels sont les événements pour lesquels ces probabilités diffèrent le plus, c'est-à-dire les événements $A$ qui maximisent $P(A)-Q(A)$. Cela mène directement à la définition de la *variation totale*.

**Définition 7.1** (distance en variation totale)**.** Soient $P, Q$ deux mesures de probabilité sur un même espace $(\mathcal{X}, \mathscr{F})$. Leur distance en variation totale est

$$\mathrm{d}_{\mathrm{TV}}(P,Q) = \sup_{A\in\mathscr{F}} P(A) - Q(A).$$

La distance en variation totale est un objet important en probabilités, qui possède de nombreuses propriétés. Parmi elles, voici les plus importantes.

1. C'est une distance sur l'espace des mesures de probabilité.
2. Elle génère une topologie plus fine que celle de la convergence en loi ; autrement dit, si $\mathrm{d}_{\mathrm{TV}}(P_n, Q) \to 0$ alors $P_n$ converge en loi vers $Q$ mais l'inverse n'est pas vrai.

**Proposition 7.1.** *Soit $\nu$ une mesure telle que $P$ et $Q$ sont absolument continues[1] par rapport à $\nu$, de densités respectives $p$ et $q$ par rapport à $\nu$. Alors, $\mathrm{d}_{\mathrm{TV}}(P,Q)$ est égale à chacune des quantités suivantes :*

$$\int_{\mathcal{X}} (p(x)-q(x))_+ \mathrm{d}\nu$$
$$\frac{1}{2}\int_{\mathcal{X}} |p(x)-q(x)|\mathrm{d}\nu. \tag{7.1}$$

*De plus, notons $E$ l'ensemble mesurable $\{x \in \mathcal{X} : p(x) > q(x)\}$. Alors,*

$$\mathrm{d}_{\mathrm{TV}}(P,Q) = P(E) - Q(E). \tag{7.2}$$

L'hypothèse selon laquelle $P, Q$ sont a.c. par rapport à $\nu$ est toujours vérifiée pour $\nu = (P+Q)/2$, et n'est donc pas restrictive.

[1] Si $P$ est absolument continue par rapport à $Q$ (ce qu'on note $P \ll Q$), alors la dérivée de Radon-Nikodym existe, et c'est une fonction mesurable positive $f$ (unique à un ensemble $Q$-négligeable près) qui vérifie $P(A) = \int f(x)\mathbf{1}_{x\in A}\mathrm{d}Q$. On appelle cette fonction *densité* de $P$ par rapport à $Q$.

*Démonstration.* Pour tout événement $A \in \mathscr{F}$, la différence $P(A) - Q(A)$ est égale à $\int_A p(x) - q(x) \mathrm{d}\nu$, qui peut elle-même s'écrire sous la forme

$$\int_{A \cap E} (p - q) \mathrm{d}\nu + \int_{A \cap \bar{E}} (p - q) \mathrm{d}\nu.$$

Le second terme est négatif, puisque si $x \notin E$ alors $p(x) \leqslant q(x)$. Ainsi, $P(A) - Q(A)$ est plus petit que le premier terme, lequel est à son tour plus petit que $\int_E (p - q) d\nu = P(E) - Q(E)$. Cela montre directement Équation 7.2. Au passage, il est évident que

$$\int_E (p(x) - q(x)) \mathrm{d}\nu = \int_{\mathcal{X}} (p(x) - q(x))_+ \mathrm{d}\nu,$$

ce qui montre la première égalité de Équation 7.1. La seconde égalité résulte de la première, puisque comme $p$ et $q$ sont des densités de probabilité, on a forcément $\int (p - q)_+ = \int (p - q)_-$.

□

Dans la suite, on supposera toujours que les diverses lois possèdent toutes une densité par rapport à une mesure de référence $\nu$. C'est le cas dans de très nombreux modèles — pas tous, hélas. Les lettres majuscules désigneront les mesures, tandis que les lettres minuscules désigneront leurs densités.

## 7.2 Test optimal au sens de l'affinité

L'affinité d'un test est la somme de ses erreurs de première et seconde espèce : c'est la probabilité de « se tromper » en général, quelle que soit l'hypothèse.

**Théorème 7.1.** *Soit $\mathfrak{T}$ l'ensemble des tests possibles de l'hypothèse $H_0 : P = P_0$ contre l'hypothèse alternative $H_1 : P = P_1$. Alors, le test possédant la meilleure affinité possible parmi tous les tests possibles vérifie*

$$\inf_{T \in \mathfrak{T}} \{\alpha_T + \beta_T\} = 1 - \mathrm{d}_{\mathrm{TV}}(P_0, P_1).$$

*En particulier, le test optimal pour l'affinité est donné par la région de rejet*

$$\mathsf{rejeter}_\star = \{p_0(x) < p_1(x)\}.$$

*Démonstration.* Soit $T$ n'importe quel test. Son affinité est $P_1(\{T = 0\}) + P_0(\{T = 1\})$. En passant au complémentaire dans le second terme, on obtient

$$1 - (P_0(\{T = 0\}) - P_1(\{T = 0\})).$$

Cette quantité est forcément plus petite que $1 - \mathrm{d}_{\mathrm{TV}}(P_0, P_1)$ par la définition même de la variation totale. De plus, cette borne est atteinte en choisissant le test $T$ donné dans l'énoncé, d'où l'égalité.

□

**Commentaire.** Le théorème précédent semble donner au problème de la construction de tests une réponse définitive : il donne le test optimal au sens de l'affinité, test qui est élémentaire et intuitif. En effet, si $P_0, P_1$ sont les deux lois et si $(x_1, \dots, x_n)$ est l'échantillon observé, alors on rejette l'hypothèse nulle si la probabilité de cette observation est plus grande sous $P_1$ que sous $P_0$ : autrement dit, si

$$\frac{p_1(x_1, \dots, x_n)}{p_0(x_1, \dots, x_n)} > 1.$$

Le terme de droite s'appelle **rapport de vraisemblance**. Pourtant, ce test ne permet pas de contrôler l'erreur de première espèce. Il peut tout à fait exister d'autres tests qui ont un niveau plus élevé. Il est donc naturel de se demander si, parmi les tests ayant un niveau fixé $1 - \alpha$, il existe un autre critère d'optimalité.

## 7.3 Théorème de Neyman-Pearson

On se place toujours dans un cadre où les deux lois $P_0$ et $P_1$ possèdent deux densités $p_0, p_1$ par rapport à une mesure commune $\nu$.

**Définition 7.2.** Un *test du rapport de vraisemblance* est un test dont la région de rejet est de la forme

$$\mathsf{rejeter} = \left\{ \frac{p_1(x)}{p_0(x)} > z \right\} \tag{7.3}$$

pour un certain $z > 0$.

Le test optimal au sens de l'affinité est un test de rapport de vraisemblance ($z = 1$).

**Théorème 7.2** (Théorème de Neyman-Pearson)**.** *Tout test de même niveau qu'un test du rapport de vraisemblance est moins puissant que celui-ci.*

*Démonstration.* On suppose que la région de rejet de $T_\star$ est de la forme Équation 7.3. Soit $T$ un autre test de même niveau que $T_\star$. La quantité

$$\int_X (T(x) - T_\star(x))(p_1(x) - z p_0(x)) \mathrm{d}\nu$$

est forcément négative ou nulle : en effet, si $T_\star(x) = 1$, alors $T(x) - T_\star(x) = T(x) - 1 \leqslant 0$, mais $p_1(x)$ est plus grand que $z p_0(x)$, donc $(p_1(x) - z p_0(x)) \geqslant 0$. De même, si $T(x) = 0$, alors cette fois ce terme est négatif. Dans les deux cas, la fonction dans l'intégrale est toujours le produit de deux nombres de signes opposés : elle est donc négative. Or, en développant cette intégrale, on constate qu'elle vaut aussi

$$P_1(T = 1) - P_1(T_\star = 1) - z P_0(T = 1) + z P_0(T_\star = 1).$$

Tout ceci n'est rien d'autre que $\beta_\star - \beta - z(\alpha - \alpha_\star)$, où $\alpha, \beta$ désignent les deux types d'erreurs du test $T$ et $\alpha_\star, \beta_\star$ celles de $T_\star$. Mais nous avons supposé que $\alpha = \alpha_\star$ : des deux termes ci-dessus, ne reste que le premier, à savoir $\beta_\star - \beta$, qui est bien négatif comme demandé.

□

## 7.4 Un exemple de test de rapport de vraisemblance

Plaçons-nous dans un modèle de Bernoulli : on a des variables aléatoires $X_1, \dots, X_n$ iid de loi $\mathrm{Ber}(p)$, et l'on souhaite tester une valeur $p_0$ de $p$ contre une valeur $p_1 \neq p_0$ à partir d'une réalisation $x_1, \dots, x_n$ du modèle.

Ici, les lois sont discrètes : elles possèdent une densité par rapport à la mesure de comptage. La probabilité d'observer $x_1, \dots, x_n$ dans le modèle avec paramètre $p$ est égale à

$$\prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^s(1-p)^{n-s}$$

où $s = x_1 + \dots + x_n$. Ainsi, le rapport des vraisemblances $r$ est égal à

$$\frac{p_1^s(1-p_1)^{n-s}}{p_0^s(1-p_0)^{n-s}} = \left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)^s \left(\frac{1-p_1}{1-p_0}\right)^n.$$

Le théorème de Neyman-Pearson dit qu'un test de la forme $r > z$ est plus puissant que tous les tests ayant le même niveau. Or, cette région de rejet peut encore s'écrire

$$s \ln\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right) > \ln(z) - n \ln\left(\frac{1-p_1}{1-p_0}\right).$$

**Dans le cas où** $p_0 < p_1$, alors par croissance $p_1/(1-p_1)$ est plus grand que $p_0/(1-p_0)$, et donc cette région de rejet peut encore s'écrire

$$\frac{s}{n} > \frac{\ln(z)/n - \ln((1-p_1)/(1-p_0))}{\ln\left(\frac{p_1(1-p_0)}{p_0(1-p_1)}\right)}.$$

Cette écriture n'a rien d'intéressant en soi. Tout ce qui compte, c'est que la région de rejet *optimale au sens de Neyman-Pearson* est de la forme $\{\bar{X}_n > z'\}$ où $z'$ correspond au terme de droite ci-dessus.

**Dans le cas où** $p_0 > p_1$, alors le même raisonnement donne une région de rejet de la forme $\{\bar{X}_n < z'\}$.

La détermination de $z'$ dépendra du niveau de confiance que l'on veut se donner. L'erreur de première espèce est $P_{p_0}(\bar{X}_n > z')$, qui est la probabilité qu'une binomiale $\mathrm{Bin}(n, p_0)$ soit plus grande que $nz'$. En choisissant pour $nz'$ le quantile de niveau $1-\alpha$ de cette loi, la probabilité ci-dessus est plus petite que $\alpha$ et le test est de niveau de confiance supérieur à $1-\alpha$.

## 7.5 Une borne sur la variation totale

*Ce chapitre n'a pas été vu en cours et n'est pas au programme.*

La construction du test optimal au sens de l'affinité nécessite le calcul de la distance en variation totale, laquelle peut être notoirement difficile : - d'abord, parce que la formule Équation 7.1 peut être impossible à calculer même si $P$ et $Q$ sont connues ; - ensuite, parce que $Q$ elle-même peut parfois être très difficile à calculer (le calcul peut être de complexité exponentielle).

En pratique, on peut chercher à *borner* cette distance par d'autres quantités plus faciles à calculer. Parmi ces quantités, la *divergence de Kullback-Leibler* joue un rôle extrêmement important, notamment pour son lien avec le maximum de vraisemblance que nous verrons plus tard.

**Définition 7.3.** Soient $P$ et $Q$ deux mesures, $P$ étant absolument continue par rapport à $Q$. Alors,

$$\mathrm{d}_{\mathrm{KL}}(P \mid Q) = \int \ln\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) dP.$$

Si $P$ n'est pas absolument continue par rapport à $Q$, on pose simplement $\mathrm{d}_{\mathrm{KL}}(P \mid Q) = +\infty$.

La notation $\mathrm{d}P/\mathrm{d}Q$ désigne la densité de $P$ par rapport à $Q$. Formellement, c'est la dérivée de Radon-Nikodym. Dans le cas de variables aléatoires continues sur $\mathbb{R}^d$, c'est le rapport des densités de $P$ et de $Q$.

La divergence $\mathrm{d}_{\mathrm{KL}}$ n'est pas une distance, et c'est pour cela qu'on l'appelle *divergence* et qu'on la note avec une barre plutôt qu'une virgule : elle n'est pas symétrique en général. Cependant, elle est toujours positive (éventuellement égale à $+\infty$ même si $P \ll Q$), et n'est nulle que si $P = Q$.

**Théorème 7.3** (Borne de Bretagnole-Huber-Pinsker)**.**

$$\mathrm{d}_{\mathrm{TV}}(P, Q) \leqslant \sqrt{1 - e^{-\mathrm{d}_{\mathrm{KL}}(P|Q)}}. \tag{7.4}$$

**Remarque.** Il est facile de vérifier que $\sqrt{1-e^{-x}} \leqslant \sqrt{x}$ lorsque $x > 0$. Ainsi, Équation 7.4 entraîne la borne plus simple $\mathrm{d}_{\mathrm{TV}} \leqslant \sqrt{\mathrm{d}_{\mathrm{KL}}}$. La borne *classique* de Pinsker améliore légèrement ce résultat, puisqu'elle dit que $\mathrm{d}_{\mathrm{TV}} \leqslant \sqrt{\mathrm{d}_{\mathrm{KL}}/2}$.

*Démonstration.* Si $P$ n'est pas absolument continue par rapport à $Q$, alors $\mathrm{d}_{\mathrm{KL}}(P \mid Q) = +\infty$ et la borne demandée est vraie. Sinon, on note $\rho$ la densité de $P$ par rapport à $Q$, de sorte que $\mathrm{d}_{\mathrm{KL}}(P \mid Q) = -\int \ln \rho(x) \mathrm{d}P$. On définit ensuite $v = (\rho - 1)_+$ et $w = (\rho - 1)_-$, de sorte que $vw$ vaut toujours 0, et donc $(1+v)(1-w) = 1 - w + v = \rho$. En particulier, $\mathrm{d}_{\mathrm{KL}}(P \mid Q)$ vaut

$$\int (-\ln(1+v)) dP + \int (-\ln(1-w)) dP.$$

Or, les deux fonctions $x \mapsto -\ln(1+x)$ et $x \mapsto -\ln(1-x)$ sont concaves sur leurs ensembles de définition. Ainsi, l'inégalité de Jensen entraîne d'une part

$$\int (-\ln(1+v)) \mathrm{d}P \leqslant -\ln\left(1 + \int v \mathrm{d}P\right)$$

et d'autre part

$$\int (-\ln(1-w)) \mathrm{d}P \leqslant -\ln\left(1 - \int w \mathrm{d}P\right).$$

Or, la formule Équation 7.1 montre que $\int v \mathrm{d}P = \mathrm{d}_{\mathrm{TV}}(P, Q)$, et de même pour $\int w \mathrm{d}P$. En additionnant les deux inégalités ci-dessus, on obtient Alors

$$-\mathrm{d}_{\mathrm{KL}} \leqslant -\ln\left((1 + \mathrm{d}_{\mathrm{TV}})(1 - \mathrm{d}_{\mathrm{TV}})\right)$$

soit $-\mathrm{d}_{\mathrm{KL}} \leqslant -\ln(1 - \mathrm{d}_{\mathrm{TV}}^2)$, c'est-à-dire Équation 7.4.

□

# 8 Tests du $\chi_2$

Les tests du $\chi_2$ sont une vaste famille de tests qui visent, pour la plupart, à tester si un échantillon (souvent discret) a été généré par une loi précise ; on parle parfois de test d'ajustement.

## 8.1 Loi multinomiale

Soit $\Omega$ un ensemble fini à $k$ éléments, disons pour simplifier $\{1, \dots, k\}$. On notera $S_k$ l'ensemble des lois de probabilités sur cet ensemble, c'est-à-dire les $k$-uplets $\mathbf{p} = (p_1, \dots, p_k)$ de nombres positifs dont la somme vaut 1. On observe $n$ tirages indépendants et identiquement distribués selon une même loi sur $\Omega$. Formellement, le modèle statistique est donné par $(\mathbf{p}^{\otimes n} : \mathbf{p} \in S_k)$.

On note $N_j$ le nombre d'observations égales à $j$. Le vecteur $N = (N_1, \dots, N_k)$ suit alors une loi multinomiale de paramètres $n$ et $\mathbf{p}$, donnée par

$$\mathbb{P}(N = (n_1, \dots, n_k)) = \frac{n!}{n_1! \dots n_k!} \prod_{j=1}^{k} p_j^{n_j},$$

où $\sum_{j=1}^{k} n_j = n$. Cette loi sera notée $\mathrm{Mult}(n, \mathbf{p})$.

**Théorème 8.1.** *Soit $N \sim \mathrm{Mult}(n, \mathbf{p})$. Alors, $\sqrt{n}(\frac{N}{n} - \mathbf{p})$ converge en loi lorsque $n \to \infty$ vers $\mathcal{N}(0, \Sigma)$, où*

$$\Sigma = \mathrm{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^\top. \tag{8.1}$$

*Démonstration.* On commence par remarquer que $N = \sum_{i=1}^{n} Z_i$, où $Z_i = (\mathbf{1}_{X_i=1}, \dots, \mathbf{1}_{X_i=k})$. Les $Z_i$ sont iid de moyenne $\mathbf{p}$. Les covariances des entrées $i$ et $j$ de $Z_k$ sont données par

$$\mathbb{E}[\mathbf{1}_{X_k=i}\mathbf{1}_{X_k=j}] - p_i p_j = \delta_{i,j} p_i - p_i p_j,$$

ce qui montre que la matrice de covariance des $Z_k$ est Équation 8.1. Il suffit alors d'appliquer le TCL.

□

**Remarque.** On considère que cette approximation normale est correcte dès que $\mathbb{E}[N_j]$ est plus grand que 5 pour tout $j$.

## 8.2 Test d'adéquation

Le test du $\chi^2$ d'adéquation consiste à tester l'hypothèse nulle

$$H_0 : \mathbf{p} = \mathbf{p}_0 \tag{8.2}$$

contre l'hypothèse alternative

$$H_1 : \mathbf{p} \neq \mathbf{p}_0, \tag{8.3}$$

pour une valeur de $\mathbf{p}_0$ fixée au préalable. À partir de maintenant, on supposera implicitement que toutes les entrées de $\mathbf{p}_0$ sont non nulles — cela garantira que les limites en loi trouvées ci-dessous ne sont pas dégénérées.

**Exemple 8.1.** On peut se demander si, dans la langue courante, les 21 lettres de l'alphabet ont à peu près la même probabilité d'apparaître comme première lettre d'un mot. Cela revient à tester si $\mathbf{p}_0 = (1/26, \dots, 1/26)$, hypothèse qui est évidemment fausse.

Qu'en est-il des 9 chiffres ? On peut vouloir tester si, dans n'importe quel document (journal, site internet, article scientifique), ces 9 chiffres apparaissent à peu près uniformément en tant que premier chiffre d'un nombre. Cela reviendrait à tester $\mathbf{p}_0 = (1/9, \dots, 1/9)$.

Ce n'est pas le cas et cette hypothèse est très fréquemment réfutée : le premier chiffre significatif d'un nombre est bien plus souvent 1 ($\approx 30\%$ des cas) que 9 ($\approx 5\%$ cas). Ce phénomène s'appelle *loi de Benford*.

Le théorème Théorème 8.1 dit que $\sqrt{n}(\frac{N}{n} - \mathbf{p}) \approx N(0, \Sigma)$. Notons $\sqrt{\mathbf{p}_0} = (\sqrt{p_1}, \dots, \sqrt{p_k})$ et $D = \mathrm{diag}(\sqrt{\mathbf{p}}_0)$. Sous $H_0$, $D^{-1}\sqrt{n}(\frac{N}{n} - \mathbf{p}_0)$ converge en loi vers $D^{-1}N(0, \Sigma) = N(0, D^{-1}\Sigma(D^{-1})^\top)$. Que vaut cette matrice de covariance ?

D'abord, comme $D$ est diagonale, $D^{-1}$ l'est aussi et $(D^{-1})^\top$ vaut $D^{-1}$. De plus, $D^2$ est égal à $\mathrm{diag}(\mathbf{p}_0)$. Enfin, en faisant la multiplication on voit vite que $D^{-1}\mathbf{p}_0 = \sqrt{\mathbf{p}}_0$. Ainsi, on voit que $D^{-1}\Sigma D^{-1}$ vaut également

$$D^{-1}D^2D^{-1} - D^{-1}\mathbf{p}_0\mathbf{p}_0D^{-1} = I_k - \sqrt{\mathbf{p}}_0\mathbf{p}_0^\top.$$

En regroupant tout cela, on obtient donc que $D^{-1}\sqrt{n}(N/n - \mathbf{p}_0)$ converge en loi vers

$$N(0, I_k - \sqrt{\mathbf{p}_0}\sqrt{\mathbf{p}_0}^T).$$

La statistique qui va nous servir à faire des tests est tout simplement la norme au carré de $D^{-1}\sqrt{n}(N/n - \mathbf{p}_0)$. En manipulant légèrement cette expression, on obtient sa forme usuelle, le *contraste du* $\chi_2$.

**Définition 8.1** (Contraste du $\chi_2$)**.** Dans le contexte ci-dessus, le *contraste du* $\chi_2$ associé à la loi $\mathbf{p}$ est la statistique

$$D_n(\mathbf{p}) = \sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j}.$$

Pour faire des tests, il suffit donc de trouver la loi asymptotique de cette statistique.

**Théorème 8.2.** *Sous l'hypothèse nulle Équation 8.2, la statistique $D_n$ converge en loi vers $\chi_2(k-1)$. De plus, sous l'hypothèse alternative Équation 8.3, $D_n$ tend vers $+\infty$ presque sûrement.*

*Démonstration.* Comme $|\sqrt{\mathbf{p}_0}|$ vaut 1, la matrice $\pi_0 = I_k - \sqrt{\mathbf{p}_0}\sqrt{\mathbf{p}_0}^T$ est la matrice de projection sur l'orthogonal du vecteur $\sqrt{\mathbf{p}_0}$ (je vous renvoie à l'appendice Chapitre 20). Le théorème de Cochran (Théorème 11.3) implique alors que la statistique $D_n$, qui est égale à

$$\left|\mathrm{diag}(1/\sqrt{\mathbf{p}_0})\sqrt{n}\left(\frac{N}{n} - \mathbf{p}_0\right)\right|^2, \tag{8.4}$$

converge en loi vers la norme de la projection d'une gaussienne $N(0, I_k)$ sur un sous-espace de dimension $k-1$, c'est-à-dire une loi $\chi_2(k-1)$. Sous l'hypothèse alternative, il y a au moins un $p_i$ non nul tel que $p_i \neq (p_0)_i$. Ainsi, Équation 8.4 est plus grand que $n(N_i/n - (p_0)_i)^2/p_i$, mais $N_i$ suit une loi $\mathrm{Bin}(n, p_i)$ et donc $N_i/n$ converge en probabilité vers $p_i$. Il est alors clair que $n(N_i/n - (p_0)_i)$ converge vers $+\infty$.

□

Un test de niveau $1-\alpha$ pour l'hypothèse Équation 8.2 est alors donné par la région de rejet

$$\{D_n(\mathbf{p}_0) > \kappa_{k-1,1-\alpha}\}$$

où $\kappa_{k-1,1-\alpha}$ est le quantile d'ordre $1-\alpha$ d'une $\chi^2(k-1)$. Si $\mathbf{p}$ n'est pas égal à $\mathbf{p}_0$, le contraste $D_n$ tend vers l'infini, donc le test sera forcément dans la zone de rejet : si l'hypothèse alternative est simple, la puissance du test tend donc vers 1.

## 8.3 Test d'indépendance

Les tests du $\chi_2$ d'indépendance sont omniprésents en sciences humaines. Dans ces tests, on observe des variables aléatoires qui sont des *couples* à valeur dans deux espaces discrets ; disons, pour simplifier, que cet espace est $\Omega = \{1, \dots, k\} \times \{1, \dots, h\}$. Les observations $(x_i, y_i)$ sont des réalisations d'une variable aléatoire $(X, Y)$. Ici, le modèle statistique sera donc $(\mathbf{p}^{\otimes n} : \mathbf{p} \in S_{k,h})$, où $S_{k,h}$ est l'ensemble des $\mathbf{p} = (p_{i,j}, i \in \{1, \dots, k\}, j \in \{1, \dots, h\})$ qui sont des lois de probabilité.

Si $\mathbf{p}$ est la loi de $(X, Y)$, alors $X$ et $Y$ sont indépendantes si et seulement si $\mathbf{p}$ peut s'écrire sous la forme $p_{i,j} = p^x_i p^y_j$, où $\mathbf{p}^x \in S_k$ et $\mathbf{p}^y \in S_h$. L'ensemble de ces lois sera noté $I_{k,h}$ (« I » pour « Indépendant » ). Les tests d'indépendance visent à tester l'hypothèse nulle

$$H_0 : \ \mathbf{p} \in I_{k,h} \tag{8.5}$$

contre l'hypothèse alternative

$$H_1 : \ \mathbf{p} \notin I_{k,h}.$$

**Exemple 8.2.** On récolte des données sur le groupe socio-professionnel (GSP) et le genre. Chaque observation correspond à une personne, possédant deux attributs : `genre`, valant 0 ou 1, et `GSP`, valant l'une des 6 groupes définis par l'INSEE (Agriculteur, artisan, cadre, etc.). Le test ci-dessus vise à déterminer si les deux modalités sont indépendantes, c'est-à-dire si la proportion d'hommes et de femmes dans chaque groupe ne diffère pas significativement en fonction du groupe.

La procédure pour effectuer un tel test nécessite plusieurs étapes.

Si $\mathbf{p}$ était effectivement la loi de deux variables indépendantes $\mathbf{p}^x$ et $\mathbf{p}^y$, alors ses marginales seraient précisément $\mathbf{p}^x$ et $\mathbf{p}^y$, que l'on pourrait facilement estimer. Pour chaque $i$ et chaque $j$, les estimateurs $\hat{\mathbf{p}}^x$ et $\hat{\mathbf{p}}^y$ définis par

$$\hat{p}_i^x = \frac{\sum_{j=1}^h N_{i,j}}{n}$$

et

$$\hat{p}_j^y = \frac{\sum_{i=1}^k N_{i,j}}{n}$$

sont effectivement des estimateurs sans biais et convergents des quantités $p_i^x, p_j^y$. De plus, sous l'hypothèse nulle, $\hat{p}_i^x \hat{p}_i^y$ serait effectivement un estimateur convergent de $p_{i,j}$.

De plus, si $\mathbf{p}$ était effectivement de la forme $\hat{\mathbf{p}}^x \hat{\mathbf{p}}^y$, alors la moyenne théorique des éléments de classe $(i,j)$ serait $n\hat{p}_i^x \hat{p}_j^y$. Cette quantité, notée $\check{N}_{i,j}$, s'appelle *effectif théorique*. Nous pouvons maintenant construire la statistique qui nous servira à tester tout cela.

**Définition 8.2** (Statistique de Pearson)**.** La statistique de Pearson est définie par

$$C_n = \sum_{i=1}^k \sum_{j=1}^h \frac{(N_{i,j} - \check{N}_{i,j})^2}{\check{N}_{i,j}}.$$

Cette statistique possède une loi limite connue, encore en vertu du théorème de Cochran. Noter que la statistique de Pearson possède une expression alternative,

$$C_n = \sum \sum \frac{n(\hat{p}_{i,j} - \hat{p}_i^x \hat{p}_j^y)^2}{\hat{p}_i^x \hat{p}_j^y}.$$

**Théorème 8.3** (Loi de la statistique de Pearson)**.** *Sous l'hypothèse nulle Équation 8.5, $C_n$ converge en loi vers*

$$\chi_2((k-1)(h-1)).$$

*De plus, pour n'importe quelle loi $\mathbf{p}_1$ qui n'est pas dans $I_{k,h}$, $C_n \to +\infty$ presque sûrement.*

*Démonstration.* C'est une conséquence un peu plus technique du théorème de Cochran.

□

Tout cela permet encore une fois d'obtenir des tests très efficacement : en abrégeant $\kappa_{1-\alpha} = \kappa_{(k-1)(h-1),1-\alpha}$, on obtient que $\mathbb{P}(C_n > \kappa_{1-\alpha}) \to \alpha$. Ainsi, la région de rejet

$$\{C_n > \kappa_{1-\alpha}\}$$

fournit un test de niveau asymptotique $1-\alpha$. La seconde partie du théorème dit que si la véritable loi sous-jacente n'est effectivement pas la loi de deux variables indépendantes, alors ce test sera systématiquement rejeté — autrement dit, si l'hypothèse alternative est simple, la puissance de ce test tend vers 1.

# Exercices

## Questions

- Quelles sont les erreurs du test consistant à toujours accepter l'hypothèse nulle ?
- Quelles sont les erreurs du test consistant à toujours refuser l'hypothèse nulle ?
- Montrer que la distance en variation totale entre deux mesures de densités $p, q$ peut aussi s'écrire $\int (p/q - 1)_+ \mathrm{d}p$.
- Montrer que si $\mathrm{d}_{\mathrm{KL}}(P_n \mid Q) \to 0$, alors $P_n$ converge en loi vers $Q$.
- Calculer la distance en variation totale entre deux lois de Bernoulli de paramètres respectifs $p$ et $q$.
- Calculer la distance en variation totale entre une loi $\mathrm{Bin}(n, p)$ et une loi $N(\mu, \sigma^2)$.
- Soient $P, Q$ deux mesures. Montrer que $\mathrm{d}_{\mathrm{KL}}(P^{\otimes n} \mid Q^{\otimes n}) = n\mathrm{d}_{\mathrm{KL}}(P \mid Q)$.

## Tests élémentaires

Pour tous les cas suivants, il faut savoir réaliser rapidement un test puissant, voire même optimal au sens qu'il vous plaira.

- Tester $\mu = \mu_0$ contre $\mu = \mu_1$ dans un échantillon $N(\mu, \sigma^2)$ lorsque $\sigma$ est connu.
- Même question lorsque $\sigma$ est inconnu.
- Soient $X_1, \dots, X_n$ un échantillon iid $N(\mu_1, \sigma_1^2)$ et $Y_1, \dots, Y_m$ ($m$ et $n$ ne sont pas forcément égaux) un échantillon iid de loi $N(\mu_2, \sigma_2^2)$. Tester $\sigma_1 = \sigma_2$ lorsque $\mu_1$ et $\mu_2$ sont connues.
- Même question lorsque $\mu_1$ et $\mu_2$ ne sont pas connues.
- Donner la forme d'un test sur la valeur de $p$ pour une réalisation d'une loi $\mathrm{Bin}(n, p)$ et calculer son niveau asymptotique quand $n \to \infty$.
- Donner la forme d'un test sur la valeur de $\lambda$ dans un échantillon de $n$ variables aléatoires de Poisson de paramètre $\lambda$.

## Exercices

**Exercice 8.1.** Soient $X_1, \dots, X_n$ des variables indépendantes de loi $\chi_2(p)$. On cherche à tester l'hypothèse nulle $p = 1$ contre l'hypothèse alternative $p = 2$.

1. Écrire la forme de la région de rejet des tests de rapport de vraisemblance.
2. Essayer de calculer le niveau de ce test ; si ce n'est pas possible, essayer de le borner.

**Exercice 8.2.** Soient $X_1, \dots, X_n$ des variables indépendantes de loi $N(0, \sigma^2)$. Proposer un test de niveau $\alpha$ de l'hypothèse $\sigma^2 = 1$ contre l'hypothèse $\sigma^2 = 1 + \varepsilon$, et estimer sa puissance. Comment varie-t-elle en fonction de $n$ et de $\varepsilon$ ?

**Exercice 8.3.** Soient $X_1, \dots, X_n$ des variables indépendantes de même loi $P$. On cherche à tester l'hypothèse nulle $P = N(0,1)$ contre l'hypothèse alternative $P = \mathscr{T}(n)$.

1. Donner le test optimal au sens de l'affinité.
2. Donner un autre test, de niveau $1-\alpha$, et calculer sa puissance.
3. Comparer ces deux tests, en particulier dans le régime où $n$ est grand.

**Exercice 8.4.** Montrer que le nombre de lancers nécessaire pour distinguer une pièce équilibrée ($p = 1/2$) d'une pièce légèrement déséquilibrée ($p_1 = 1/2 + \varepsilon$) est d'ordre $1/\varepsilon^2$.

**Exercice 8.5.** On note $p$ la probabilité qu'un enfant né vivant soit un garçon. On suppose que les enfants sont de sexe indépendants, et que cette probabilité est la même pour toutes les grossesses.

1. Il y a eu en France métropolitaine en 2015 $n = 760\,421$ naissances. , dont 389 181 garçons. Tester l'hypothèse $p = \frac{1}{2}$ contre l'alternative pertinente.
2. En 1920, il y a eu 838 137 naissances dont 432 044 garçons. Tester l'hypothèse $p_{2015} = p_{1920}$.

**Exercice 8.6.** Soient $X_1, \dots, X_n$ i.i.d de loi $N(\theta, 1)$, où $\theta$ est un paramètre réel.

1. Donner un intervalle de confiance pour $\theta$ au niveau de risque 5% de la forme $[\hat{\theta}_n, +\infty[$.
2. En déduire un test de niveau 5% pour les hypothèses $H_0 : \theta = 0$ et $H_1 : \theta > 0$.
3. Donner le modèle de l'expérience statistique. Donner l'expression du test de rapport de vraisemblance $T$ pour les hypothèses $H_0 : \theta = 0$ et $H_1 : \theta = \mu$, où $\mu > 0$. Quel test retrouve-t-on?
4. Construire le test de rapport de vraisemblance au niveau 5% pour les hypothèses $H_0 : \theta = 0$ et $H_1 : \theta > 0$.

**Exercice 8.7** (Test sur des lois uniformes)**.** On se donne $X_1, \dots, X_n$ iid de loi $\mathscr{U}(0, \theta)$, et on note $M_n = \max_{j=1,\dots,n} X_i$.

1. Écrire la fonction de répartition de $M_n$, puis en déduire un test $T$ de niveau $1-\alpha$ pour les hypothèses $H_0 : \theta = 1$ contre $H_1 : \theta < 1$.
2. Donner le test du rapport de vraisemblance pour pour les hypothèses $H_0 : \theta = 1$ contre $H_1 : \theta = \theta_0$, où $\theta_0 < 1$. Calculer sa puissance.
3. On cherche à tester $H_0 : \theta = 1$ contre $H_1 : \theta < 1$. Comme la seconde hypothèse est composite, on ne peut pas directement appliquer le test du rapport de vraisemblance ; à la place, on utilise un test du *maximum* de vraisemblance, qui est de la forme

$$\frac{\sup_{\theta<1} \rho_\theta(x_1, \dots, x_n)}{\rho_1(x_1, \dots, x_n)} > z$$

   où $\rho_\theta$ est la densité d'un échantillon iid de lois $\mathscr{U}[0, \theta]$. Calculer le supremum dans cette expression, et en déduire la région de rejet.
4. Montrer que la puissance de $T$ vaut $\alpha$.
5. En utilisant la même technique, construire le test du rapport de maximum de vraisemblance pour pour les hypothèses $H_0 : \theta = 1$ contre $H_1 : \theta > 1$, noté $T'$, au niveau $1-\alpha$. Calculer sa puissance.
6. Donner un test de niveau $1-\alpha$ pour $H_0 : \theta = 1$ contre $H_1 : \theta > 1$, plus puissant que $T'$ pour n'importe quel $\theta > 1$.

**Exercice 8.8.** Une réalisation d'une variable aléatoire $X \sim \mathrm{Bin}(20, p)$ donne $X = 8$.

1. Proposer un test du rapport de vraisemblance de l'hypothèse nulle $p = p_0 = 1/2$ contre l'hypothèse alternative $p = p_1 = 1/3$. Donner l'expression de la $p$-valeur du test.
2. On tire des variables aléatoires iid de Bernoulli jusqu'à obtenir 8 succès. Écrire la loi de probabilité du nombre de lancers $N$.
3. Il se trouve que le nombre de lancers nécessaires pour cela était $N = 20$. Proposer un test du rapport de vraisemblance de l'hypothèse nulle $p = p_0 = 1/2$ contre l'hypothèse alternative $p = p_1 = 1/3$. Donner l'expression de la $p$-valeur du test.
4. Pourquoi les deux $p$-valeurs sont-elles différentes, alors que les deux tests sont identiques ?

**Exercice 8.9** (Test d'adéquation du $\chi_2$)**.** On lance 60 fois un dé et on obtient les résultats suivants :

| **Face** $k$ | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| **Effectif** $N_k$ | 10 | 13 | 8 | 12 | 9 | 8 |

Le dé est-il bien équilibré ? À titre indicatif, le quantile d'une loi $\chi^2(5)$ d'ordre 95% est 11.07.

**Exercice 8.10** (Test d'indépendance du $\chi_2$)**.** On cherche à savoir si les variables « être riche » et « être heureux » sont indépendantes. On interroge un grand échantillon de personnes à ce sujet, et l'on récolte les données suivantes :

| | riche | pauvre |
|---|---|---|
| **heureux** | 344 | 700 |
| **triste** | 257 | 705 |

L'argent fait-il le bonheur ?

# Annales de partiel

Je ne garantis pas que les notations et les concepts utilisés dans ces annales soient en phase avec le cours de cette année !

- Partiel 2020 et sa correction
- Examen 2020 ; ne pas regarder le deuxième exercice.
- Partiel 2023 et sa correction

# 9 Moindres carrés

Les modèles *linéaires* sont les modèles les plus simples dans lesquels on raisonne en termes d'*entrées* et de *sorties*. Dans ces modèles, on dispose de variables $x_i$, dites *explicatives*, et de variables $y_i$, dites *à expliquer*, et l'on suppose qu'il existe une fonction inconnue $f$ telle que

$$y_i \approx f(x_i),$$

et que l'on voudrait estimer. Les modèles linéaires consistent à supposer que $f$ est affine. Les modèles plus complexes, commes les réseaux de neurones, placent $f$ dans des classes plus riches. L'objectif de la méthode des moindres carrés est de trouver la meilleure approximation de $f$ possible dans la classe des fonctions affines.

## 9.1 Ajustement affine en une dimension.

On suppose qu'il existe entre les données $x_i$ et $y_i$ une relation de la forme $y_i \approx \alpha + \beta x_i$ où $\alpha, \beta$ sont deux nombres réels. Ici, $\approx$ signifie que la relation n'est pas parfaite : peut-être par exemple que les sorties sont bien égales à $\alpha + \beta x_i$, mais que les observations $y_i$ ont été polluées par du bruit ou des erreurs. Nous verrons cela plus tard.

Pour l'heure, nous voulons chercher les meilleurs $\alpha, \beta$ possibles. On calcule la distance entre le nuage de points $(x_i, y_i)$ et la droite d'équation $y = \alpha + \beta x$. Cette distance au carré est donnée par

$$L(\alpha, \beta) = \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2.$$

On cherchera donc les $(\hat{\alpha}, \hat{\beta})$ qui minimisent cette distance. La fonction $L$ est manifestement une fonction quadratique qui tend vers $+\infty$ lorsque $(\alpha, \beta) \to \infty$, par conséquent cette fonction possède un unique minimiseur $(\hat{\alpha}, \hat{\beta})$, et ce minimiseur est le seul point en lequel les dérivées partielles s'annulent (*conditions de premier ordre*) : $\partial_\alpha L(\hat{\alpha}, \hat{\beta}) = 0$ et $\partial_\beta L(\hat{\alpha}, \hat{\beta}) = 0$. Or,

$$\partial_\alpha L(\alpha, \beta) = \sum_{i=1}^{n} (\alpha + \beta x_i - y_i)$$

$$\partial_\beta L(\alpha, \beta) = \sum_{i=1}^{n} x_i(\alpha + \beta x_i - y_i).$$

Les conditions de premier ordre deviennent donc $n\alpha + \beta(x_1 + ... + x_n) - (y_1 + \cdots + y_n) = 0$ soit encore $\alpha + \beta\bar{x} - \bar{y} = 0$, et d'autre part $\alpha(x_1 + \cdots + x_n) + \beta(x_1^2 + \cdots + x_n^2) - (x_1 y_1 + \cdots + x_n y_n) = 0$, soit $\alpha\bar{x} + \beta\overline{xx} - \overline{xy} = 0$, où $\overline{xx}$ est la moyenne des carrés des $x_i$ et $\overline{xy}$ la moyenne des $x_i y_i$. En résolvant ces équations, on trouve d'abord $\alpha$ puis $\beta$ :

$$\beta = \frac{\overline{xy} - \bar{x}\bar{y}}{\overline{xx} - \bar{x}\bar{x}}, \qquad \alpha = \bar{y} - \hat{\beta}\bar{x}.$$

Le coefficient $\beta$ n'est rien d'autre que la covariance empirique des $x_i$ et des $y_i$, normalisé par la variance empirique des $x_i$.

L'inégalité de Cauchy-Schwartz dit que $|\overline{xy} - \bar{x}\bar{y}| \leqslant \tilde{\sigma}_x \tilde{\sigma}_y$, où l'on a noté

$$\tilde{\sigma}_x^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$$

l'estimateur naïf de la variance[1]. L'inégalité n'est une égalité que si $x$ et $y$ sont effectivement colinéaires, c'est-à-dire si $y_i = \hat{\alpha} + x_i\hat{\beta}$ pour tous les $i$. La qualité de l'ajustement affine est donc bien mesurée par la quantité

$$R^2 = \frac{\overline{xy} - \bar{x}\bar{x}}{\tilde{\sigma}_x \tilde{\sigma}_y}.$$

## 9.2 Moindres carrés ordinaires

Dans le cadre général le nombre $d$ de variables explicatives est plus grand que 1. On notera $\mathbf{x} = (x_1, \dots, x_d)$ un élément de $\mathbb{R}^d$ ; les variables explicatives seront alors $\mathbf{x}_1, \dots, \mathbf{x}_n$. Avec mes notations, ces vecteurs sont **des vecteurs lignes**[2].

On cherchera donc des nombres $\theta_i$ tels que $y_i$ est aussi proche que possible de

$$\theta_1 \mathbf{x}_{i,1} + \cdots + \theta_d \mathbf{x}_{i,d} = \mathbf{x}_i \theta,$$

où le paramètre $\theta$, sera toujours vu **comme le vecteur colonne**[3] des $\theta_i$.

**Remarque : où est passée la constante ?** *Dans l'équation ci-dessus, on a l'impression que le terme constant, qui correspondait à $\alpha$ dans l'exemple en dimension 1, a disparu. Ce n'est pas le cas : intégrer la constante au modèle revient à considérer que la variable constante égale à 1 fait partie des variables explicatives. En pratique, cela revient à poser, par exemple, $\mathbf{x}_{i,1} = 1$ pour tout $i$. Ainsi, la constante correspondra toujours à $\theta_1$.*

On pose $X$ la matrice $n \times d$ dont la $i$-ème ligne est $\mathbf{x}_i$ et $Y$ le vecteur colonne des $y_i$. La matrice dont les lignes sont composées des nombres réels $\mathbf{x}_i\theta$ n'est autre que la matrice $X\theta$. De façon générale, pour n'importe quel $\theta \in \mathbb{R}^d$, la distance entre le nuage de points $(X, Y)$ et la droite d'équation $Y = X\theta$ est alors $|Y - X\theta|$. On pourrait reproduire la méthode analytique ci-dessus pour trouver les paramètres optimaux, à savoir

$$\hat{\theta} = \arg\min_{\theta} |Y - X\theta|^2. \tag{9.1}$$

Cependant, une interprétation géométrique simplifie la tâche : le $\hat{\theta}$ qui minimise Équation 9.1 est précisément celui qui garantit que $X\hat{\theta}$ est la projection orthogonale de $Y$ sur le sous-espace vectoriel $\mathscr{V}_X = \{X\theta : \theta \in \mathbb{R}^d\}$.

[1] Si $P$ est absolument continue par rapport à $Q$ (ce qu'on note $P \ll Q$), alors la dérivée de Radon-Nikodym existe, et c'est une fonction mesurable positive $f$ (unique à un ensemble $Q$-négligeable près) qui vérifie $P(A) = \int f(x)\mathbf{1}_{x\in A}\mathrm{d}Q$. On appelle cette fonction *densité* de $P$ par rapport à $Q$.

[2] Ils sont de dimension $(1, d)$ si on les voit comme des matrices

[3] Donc, de dimension $(d, 1)$ cette fois.

**Théorème 9.1.** *Si $d \leqslant n$ et si $X$ est de rang $d$, alors*

$$\hat{\theta} = (X^\top X)^{-1} X^\top Y. \tag{9.2}$$

*Démonstration.* La projection orthogonale sur le sous-espace vectoriel engendré par les colonnes d'une matrice $X$ est la matrice $X(X^\top X)^{-1}X^\top$, comme démontré dans l'appendice Chapitre 20. Ainsi, la projection de $Y$ sur ce sous-espace est $X(X^\top X)^{-1}X^\top Y$, et c'est aussi (par définition de l'argmin) $X\hat{\theta}$. Comme $X$ est injective en vertu du théorème du rang, on en déduit le résultat.

□

L'expression Équation 9.2 possède de nombreuses expressions alternatives. Parmi elles, on pourra noter que

$$\hat{\theta} = \theta + \left( \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{x}_i \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^\top \varepsilon_i. \tag{9.3}$$

**Remarque générale.** Si, en dimension 1, on cherche à trouver le $\theta$ qui résout l'équation $y = x\theta$, on trouve évidemment $\theta = y/x$, c'est-à-dire qu'on *divise* $y$ par $x$, ou encore qu'on multiplie $y$ par l'inverse de $x$. En dimension supérieure, quand on veut résoudre en $\theta$ l'équation $Y = X\theta$, c'est pareil. Le problème, c'est qu'on ne sait pas forcément *inverser* $X$. La formule Équation 9.2 dit que même si $X$ n'est pas inversible, on peut quand même "diviser par $X$" : c'est pour cela que la matrice $(X^\top X)^{-1}X^\top$ est appelée *pseudo-inverse à gauche* de $X$ – parfois associée au nom de Moore-Penrose. Multiplier $Y$ par $(X^\top X)^{-1}X^\top$ donne $(X^\top X)^{-1}X^\top Y = \hat{\theta}$ : ce vecteur ne vérifie par forcément $X\hat{\theta} = Y$, mais parmi tous les vecteurs possibles, c'est celui qui rend $X\theta$ le plus proche possible de $Y$.

## 9.3 Résidus et $R^2$

Le vecteur $\hat{Y} = X\hat{\theta}$ est appelé *vecteur des prédictions*. Le vecteur $\hat{\varepsilon} = Y - \hat{Y} = Y - X\hat{\theta}$ est appelé *vecteur des résidus*. Si ce dernier est nul ou très petit, cela veut dire que les $Y$ sont presque parfaitement des fonctions linéaires des $X$.

**Définition 9.1.** Dans le cas d'une régression Équation 9.2 *avec constante*, le coefficient de détermination est défini par

$$R^2 = \frac{\sum_{i=1}^{n} |\hat{y}_i - \bar{y}|^2}{\sum_{i=1}^{n} |y_i - \bar{y}|^2}.$$

C'est un nombre entre 0 et 1.

Le numérateur est la variance empirique des prédictions $\hat{y}_i$. Le dénominateur est la variance empirique des observations. Dans les des cas, il s'agit de la norme carrée d'un vecteur ($\hat{Y}$ et $Y$) projeté sur l'espace des vecteurs de moyenne nulle. Comme $\hat{Y}$ est déjà une projection de $Y$ sur un certain sous-espace, on a forcément $|\hat{Y} - \bar{y}| \leqslant |Y - \bar{y}|$, donc le coefficient $R^2$ est toujours entre 0 et 1.

Plus le coefficient de détermination est proche de 1, meilleure est la régression – attention, cet indicateur possède de nombreuses limites.

# 10 Modèles linéaires

## 10.1 Modèle gaussien

À ce stade, nous n'avons fait aucune hypothèse statistique ni probabiliste sur le modèle : les $\mathbf{x}_i, y_i$ étaient donnés tels quels. Le *modèle linéaire gaussien* avec variables explicatives $\mathbf{x}_1, \dots, \mathbf{x}_n$ exogènes consiste à supposer que $Y = X\theta + \varepsilon$, où $\varepsilon = N(0, \sigma^2 I_n)$. Formellement, le modèle est indexé par $\theta$ et $\sigma^2$, et donné par

$$P_{\theta,\sigma^2} = N(X\theta, \sigma^2 I_n).$$

Dans ce modèle, la loi de l'estimateur Équation 9.2 est connue. Par simplicité, je note $H = X(X^\top X)^{-1}X^\top$ la matrice de projection orthogonale sur l'espace vectoriel engendré par les colonnes de $X$, qui est de dimension $d$.

**Théorème 10.1** (Loi de $\hat{\theta}$)**.** *Sous le modèle linéaire gaussien $P_{\theta,\sigma^2}$,*

$$\hat{\theta} \sim N(\theta, \sigma^2(X^\top X)^{-1}),$$

$$\frac{|\hat{\varepsilon}|^2}{\sigma^2} \sim \chi_2(n-d),$$

*et ces deux variables aléatoires sont indépendantes.*

*Démonstration.* Ce n'est rien de plus que le théorème de Cochran appliqué à notre problème : en effet, le vecteur des résidus est la projection orthogonale de $Y$ sur le sous-espace orthogonal à l'espace des colonnes de $X$.

□

La variable aléatoire $|\hat{\varepsilon}|^2$ est souvent appelée *Somme des Carrés des Résidus* (SCR). Le théorème précédent implique que

$$\hat{\sigma}_n^2 = \frac{|\hat{\varepsilon}|^2}{n-d}$$

est un estimateur sans biais de $\sigma^2$. et ces deux variables aléatoires sont indépendantes. En particulier, $(n-d)\hat{\sigma}_n^2/\sigma^2 \sim \chi_2(n-d)$.

## 10.2 Modèle linéaire général

Il est possible de ne pas faire d'hypothèses gaussiennes sur le modèle. Dans ce cadre plus général, on supposera que $Y = X\theta + \varepsilon$, où les $\varepsilon_i$ sont iid, centrés, et de même variance $\sigma^2$ — sous cette dernière hypothèse, on parle de modèle *homoscédastique*.

Sous ces hypothèses, $\hat{\theta}$ est toujours un estimateur sans biais de $\theta$ : cela se voit directement en prenant l'espérance de Équation 9.3. De plus, la loi de $\theta$ n'est plus gaussienne, mais $\theta$ est asymptotiquement normal sous des hypothèses supplémentaires sur $X$. Ces hypothèses sont les suivantes.

On suppose que les variables explicatives $\mathbf{x}_i$ vérifient la propriété suivante[1] :

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{x}_i = \Sigma_x, \tag{10.1}$$

où $\Sigma_x$ est inversible. Cette propriété s'écrit aussi $X^\top X/n \to \Sigma_x$.

**Théorème 10.2.** *Sous les hypothèses précédentes, $\sqrt{n}(\hat{\theta} - \theta)$ converge en loi lorsque $n \to \infty$ vers $N(0, \sigma^2 \Sigma_x^{-1})$.*

*Démonstration.* Rappelons que $\hat{\theta}$ peut s'écrire $\theta + (X^\top X/n)^{-1} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^\top \varepsilon_i$. Pour montrer que $\sqrt{n}(\hat{\theta} - \theta)$ converge, il suffit donc de démontrer que

$$\sqrt{n} \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i^\top \varepsilon_i \tag{10.2}$$

converge en loi vers $N(0, \Sigma_x^2)$ : comme le terme $(X^\top X/n)^{-1}$ converge vers $\Sigma_x^{-1}$ par hypothèse, la limite de $\sqrt{n}(\hat{\theta} - \theta)$ sera bien $N(0, \Sigma_x^{-1}\Sigma_x\Sigma_x^{-1}) = N(0, \Sigma_x^{-1})$. Malheureusement, on ne peut pas directement appliquer le TCL classique à Équation 10.2 : en effet, les variables aléatoires $X_i = \mathbf{x}_i^\top \varepsilon_i$ ne sont pas identiquement distribuées. On doit pour cela appliquer une version plus générale du TCL, que j'ai écrite en appendice (Théorème 21.3). Pour appliquer ce théorème en toute rigueur, on a besoin d'une hypothèse supplémentaire sur les $\mathbf{x}_i$ que je n'ai pas mentionnée — c'est une hypothèse technique[2].

□

## 10.3 Ellipsoïde de confiance

Les deux théorèmes énoncés ci-dessus permettent de définir des régions de confiance associées à $\theta$ ; ici, $\theta$ n'est plus un nombre réel mais un vecteur, d'où le terme de *région* et plus simplement d'*intervalle*.

[1] On rappelle que $\mathbf{x}_i$ est un vecteur ligne de taille $d$, et donc que les matrices $\mathbf{x}_i^\top \mathbf{x}_i$ sont bien des matrices carrées de taille $d \times d$.

[2] Il faut que la quantité $\max_{j=1,\dots,n} |\mathbf{x}_j|^2 / \sum |\mathbf{x}_i|^2$ tende vers zéro lorsque $n \to \infty$. Cela revient à dire que toute l'information apportée par les $x_i$ n'est pas concentrée sur une seule observation ou sur un très petit nombre d'observations.

### Préliminaire : la variance est connue

Commençons par construire une région probable pour un vecteur gaussien $\xi \sim N(0, I_d)$. Nous savons que $|\xi|^2 \sim \chi_2(d)$. Si $\kappa_{d,1-\alpha}$ désigne le quantile d'ordre $1-\alpha$ de cette loi, on en déduit que $\xi$ est de norme inférieure à $\sqrt{\kappa_{d,1-\alpha}}$ avec probabilité $1-\alpha$ ; autrement dit,

$$\mathbb{P}(0 \in B(\xi, \sqrt{\kappa})) = 1 - \alpha.$$

Il est immédiat d'en déduire que si $\xi \sim N(\mu, I_d)$, alors comme $\xi - \mu \sim N(0, I_d)$, on a

$$\mathbb{P}(\mu \in B(\xi, \sqrt{\kappa})) = 1 - \alpha.$$

Maintenant, en toute généralité, si $\xi \sim N(\mu, \Sigma)$, alors $\Sigma^{-1/2}(\xi - \mu) \sim N(0, I_d)$. On en déduit donc que

$$\mathbb{P}(\mu \in \Sigma^{1/2} B(\xi, \sqrt{\kappa})) = 1 - \alpha.$$

La région de confiance est donc l'image de la boule $B(\xi, \sqrt{\kappa})$ par la matrice symétrique $\Sigma^{1/2}$ : c'est une ellipse. Par ailleurs, l'ensemble $\Sigma B(x, \delta)$ peut aussi s'écrire $\{y \in \mathbb{R}^d : \ |\Sigma^{1/2}(x-y)|^2 \leqslant \delta\}$.

En combinant ce résultat avec la loi de $\hat{\theta}$ donnée dans Théorème 10.1, on obtient la région de confiance

$$\left\{ \theta \in \mathbb{R}^d : \left| \frac{1}{\sigma} (X^\top X)^{1/2} (\hat{\theta} - \theta) \right|^2 < \kappa_{d,1-\alpha} \right\}.$$

Malheureusement, cette région nécessite de connaître $\sigma$. Lorsqu'on ne le connaît pas, il faut l'estimer.

### Cas général

Toujours sous le modèle linéaire gaussien, nous avons vu que la loi de $|\sigma^{-1}(X^\top X)^{1/2}(\hat{\theta} - \theta)|^2$ est une $\chi_2(d)$, et que la loi de $(n-d)\hat{\sigma}_n^2 \sigma^{-2}$ est une $\chi_2(n-d)$. Par conséquent, la variable aléatoire

$$\frac{|(X^\top X)^{1/2}(\hat{\theta} - \theta)|^2}{\hat{\sigma}_n^2}$$

a pour loi le rapport de lois du $\chi_2$ indépendantes de paramètres $d$ et $n-d$. Cette loi est connue : elle est égale à $d$ fois une *loi de Fisher* dont les propriétés sont données dans Section 11.4. Cela donne directement le théorème suivant.

**Théorème 10.3** (Ellipsoïde de confiance)**.** *Soit $\hat{\theta}$ l'estimateur des MCO dans un modèle linéaire gaussien. Si $f_{d,n-d,1-\alpha}$ est le quantile d'ordre $1-\alpha$ d'une loi $\mathscr{F}_{d,n-d}$, alors la région*

$$\left\{ \theta \in \mathbb{R}^d : \frac{|(X^\top X)^{1/2}(\hat{\theta} - \theta)|^2 / d}{\hat{\sigma}_n^2} < f_{d,n-d,1-\alpha} \right\}$$

*est une région de confiance de niveau $1-\alpha$ pour $\theta$.*

*Lorsque le modèle n'est pas gaussien, mais qu'il vérifie les hypothèses de la section Section 10.2, le même résultat est valable mais le niveau de confiance de la région ci-dessus est* asymptotiquement *égal à $1-\alpha$.*

# 11 Outils gaussiens

## 11.1 Vecteurs gaussiens

Un vecteur aléatoire $X$ à valeurs dans $\mathbb{R}^n$ est un vecteur gaussien de loi $N(\mu, \Sigma)$ si sa densité est donnée par

$$\frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} \exp\left\{-\frac{1}{2}\left\langle x-\mu, \Sigma^{-1}(x-\mu)\right\rangle\right\}.$$

Ici, le vecteur $\mu \in \mathbb{R}^n$ est appelé *moyenne* de $X$ parce que

$$\mathbb{E}[X] = \mu.$$

La matrice $\Sigma$, qui est toujours supposée symétrique et à valeurs propres strictement positives (on dit *définie positive*), est appelée *matrice de covariance*, parce que

$$\mathbb{E}[(X-\mu)(X-\mu)^\top] = \Sigma.$$

De même que la transformée de Fourier d'une variable gaussienne réelle $N(m, \sigma^2)$ est égale à $e^{imt-\frac{t^2\sigma^2}{2}}$, la transformée de Fourier d'un vecteur gaussien $N(\mu, \Sigma)$ est

$$\mathbb{E}[e^{i\langle t, X\rangle}] = \exp\left\{i\langle t, \mu\rangle - \frac{\langle (t-\mu), \Sigma(t-\mu)\rangle}{2}\right\}.$$

**Théorème 11.1.**

1. *Toute fonction linéaire d'un vecteur gaussien est encore un vecteur gaussien. Si $M$ est une matrice et $X \sim N(\mu, \Sigma)$,*
$$MX \sim N(M\mu, M\Sigma M^\top).$$

2. *Si le couple $(X, Y)$ forme un vecteur gaussien, alors $X$ et $Y$ sont indépendants si et seulement si leur covariance $\mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])^\top]$ est la matrice nulle.*

## 11.2 Conditionnement gaussien

Soit $(X, Y)$ un vecteur gaussien de dimension $n+m$, avec $X \in \mathbb{R}^n$ et $Y \in \mathbb{R}^m$. On peut écrire sa moyenne $\mu$ en deux blocs

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

et sa covariance $\Sigma$ en quatre blocs

$$\Sigma = \begin{bmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{2,1} & \Sigma_{2,2} \end{bmatrix}$$

où , par symétrie, $\Sigma_{2,1} = \Sigma_{1,2}^\top$.

**Théorème 11.2.** *La loi de $X$ conditionnellement à $Y$ est une loi gaussienne de moyenne*

$$\mu_1 + \Sigma_{1,2}\Sigma_{2,2}^{-1}(X_2 - \mu_2)$$

*et de covariance*

$$\Sigma_{1,1} - \Sigma_{2,1}\Sigma_{2,2}^{-1}\Sigma_{1,2}.$$

L'expression *loi conditionnelle* signifie ici que, pour toute fonction test $\varphi : \mathbb{R}^n \to \mathbb{R}$, l'espérance conditionnelle $\mathbb{E}[\varphi(X) \mid Y]$, qui est une variable aléatoire $Y$-mesurable, vaut

$$\frac{1}{(2\pi \det(S^{-1}))^{n/2}} \int_{\mathbb{R}^n} \varphi(x) e^{-\frac{\langle X-1-(\mu_1+\Sigma_{1,2}\Sigma_{2,2}^{-1}(X_2-\mu_2))S^{-1}(X_1-(\mu_1+\Sigma_{1,2}\Sigma_{2,2}^{-1}(X_2-\mu_2)))\rangle}{2}} dx$$

où $S = \Sigma_{1,1} - \Sigma_{2,1}\Sigma_{2,2}^{-1}\Sigma_{1,2}$.

## 11.3 Théorème de Cochran

**Théorème 11.3** (Théorème de Cochran)**.** *Soit $X \sim N(0, I_n)$ et soient $E_1, \dots, E_k$ des sous-espaces orthogonaux de $\mathbb{R}^n$ tels que $\mathbb{R}^n = \oplus_{j=1}^k E_j$. On note $\pi_j(X)$ la projection orthogonale de $X$ sur $E_j$. Alors, la famille $(\pi_j(X))_{j=1,\dots,k}$ est une famille de vecteurs gaussiens indépendants. De plus,*

$$|\pi_j(X)|^2 \sim \chi_2(\dim E_j).$$

*Démonstration.* Pour chaque $E_i$, notons $d_i$ sa dimension et choisissons-lui une base orthonormale $e_1^i, \dots, e_{d_i}^i$. La projection orthogonale de $X$ sur $E_i$ est $\pi_i(X) = \sum_{t=1}^{d_i} \langle X, e_t^i\rangle e_t^i$. Notons $X_t^i = \langle X, e_t^i\rangle$. Le vecteur $(X_t^i)$ (avec $i = 1, \dots, k$ et $t = 1, \dots, d_i$), qui contient bien $d_1 + \cdots + d_k = n$ éléments, est une fonction linéaire du vecteur gaussien centré $X$, donc est lui-même un vecteur gaussien centré. Calculons sa covariance : de façon générale, si $e, f$ sont deux vecteurs fixés,

$$\mathbb{E}[\langle X, e\rangle\langle X, f\rangle] = \sum_{i,j} e_i f_j \mathrm{Cov}(X_i, X_j) = \langle e, f\rangle.$$

Il est alors immédiat que la matrice de covariance du vecteur gaussien $(X_t^i)$ n'est autre que la matrice $(\langle e_t^i, e_s^j\rangle)$, c'est-à-dire l'identité puisque les $(e_t^i)$ forment une base orthonormale de $\mathbb{R}^n$. Il en résulte les deux points de l'énoncé.

1. Les $\pi_i(X)$ sont des variables indépendantes, puisque fonctions linéaires de variables indépendantes entre elles.
2. La formule de Parseval dit que
$$|\pi_i(X)|^2 = \sum_{t=1}^{d_i} |X_t^i|^2$$
ce qui est bien une somme de $d_i$ gaussiennes $N(0,1)$ indépendantes, donc une $\chi_2(d_i)$.

□

## 11.4 Loi de Fisher

Si $N$ est un vecteur et $X, Y$ sont les projections de $N$ sur deux sous-espaces vectoriels orthogonaux, le théorème de Cochran dit que $X$ et $Y$ sont des lois du $\chi_2$ indépendantes de paramètres $p = \dim E, q = \dim F$. La loi de leur rapport $X/Y$ est connue et fréquemment utilisée en statistiques.

**Théorème 11.4.** *Soient $X, Y$ deux variables aléatoires indépendantes, de lois respectives $\chi_2(p)$ et $\chi_2(q)$. La loi du rapport $(X/p)/(Y/q)$ s'appelle loi de Fisher de paramètres $p, q$, et on la note $\mathscr{F}_{p,q}$. Sa densité est donnée par*

$$f_{p,q}(x) = \frac{\mathbf{1}_{x>0}}{Z_{p,q}} \frac{\left(\frac{px}{px+q}\right)^{\frac{p}{2}} \left(1 - \frac{px}{px+q}\right)^{\frac{q}{2}}}{x} \tag{11.1}$$

*où la constante $Z_{p,q}$ est $B(p/2, q/2)$, c'est-à-dire*

$$Z_{p,q} = \int_0^1 u^{\frac{p}{2}-1}(1-u)^{\frac{q}{2}-1} du.$$

Le calcul est facile, puisque les lois du $\chi_2$ ont une densité connue donnée par Équation 5.4. Soit $\varphi$ une fonction test et soit $F = (X/p)/(Y/q)$. Alors, $\mathbb{E}[\varphi(F)]$ vaut

$$\frac{1}{C_p C_q} \int_0^\infty \int_0^\infty \varphi\left(\frac{uq}{vp}\right) e^{-\frac{u}{2}-\frac{v}{2}} u^{\frac{p}{2}-1} v^{\frac{q}{2}-1} du dv$$

avec $C_n = 2^{n/2}\Gamma(n/2)$. Dans l'intégrale en $v$, on pose $x = uq/vp$, de sorte que l'intégrale ci-dessus devient

$$\frac{(p/q)^{\frac{p}{2}}}{C_p C_q} \int_0^\infty \varphi(x) x^{\frac{p}{2}-1} \int_0^\infty e^{-\frac{vpx}{2q}-\frac{v}{2}} v^{\frac{p}{2}-1} v^{\frac{q}{2}} dv dx.$$

On reconnaît dans l'intégrale en $v$ une fonction Gamma, égale à

$$\frac{\Gamma(p/2+q/2)}{\left(\frac{px+q}{2q}\right)^{\frac{p+q}{2}}}.$$

L'espérance $\mathbb{E}[\varphi(F)]$ vaut donc

$$\frac{(p/q)^{p/2}\Gamma\left(\frac{p+q}{2}\right)}{C_p C_q (2q)^{\frac{p+q}{2}}} \int_0^\infty \varphi(x) \frac{x^{\frac{p}{2}-1}}{(px+q)^{\frac{p+q}{2}}} dx.$$

En simplifiant, on trouve exactement la densité donnée par Équation 11.1.

# 12 Tests linéaires

Les modèles linéaires sont si riches, si puissants, et si fréquemment utilisés dans toutes les sciences quantitatives, que la question de *tester* si les paramètres estimés sont pertinents est rapidement devenue une discipline en elle-même, appelée *économétrie*.

## 12.1 Significativité d'un coefficient

Dans une régression de la forme $y_i = \theta_1 \mathbf{x}_{i,1} + \cdots + \theta_d \mathbf{x}_{i,d}$, si le $j$-ème coefficient $\theta_j$ est nul, alors cela veut dire que la $j$-ème variable explicative n'a *aucun effet* sur la variable expliquée : en effet, $\mathbf{x}_{i,j}$ pourrait avoir une toute autre valeur sans modifier la sortie $y_i$. Pour cette raison, le test d'une hypothèse de type $\theta_j = 0$ s'appelle *test de significativité*.

Dans un modèle gaussien comme Section 10.1, nous savons que $\hat{\theta} \sim N(\theta, \sigma^2 (X^\top X)^{-1})$. Notons $\ell_j^2$ le $j$-ème coefficient diagonal de la matrice $(X^\top X)^{-1}$ ; ce nombre est fréquemment appelé *levier*. Il est explicitement calculable, car il ne dépend que des données d'entrée $\mathbf{x}_t$ ; de plus, $\hat{\theta}_j \sim N(\theta_j, \sigma^2 \ell_j^2)$, et l'on en déduit (comme dans un test de Student) que sous l'hypothèse nulle $\theta_j = 0$, la statistique

$$\frac{\hat{\theta}_j}{\ell_j \hat{\sigma}_n}$$

suit une loi de Student $\mathscr{T}(n-d)$. Il est fréquent d'utiliser la notation

$$\hat{\sigma}(\hat{\theta}_j) = \ell_j \hat{\sigma}_n$$

car c'est un estimateur de la variance de $\hat{\theta}_j$.

**Théorème 12.1.** *Soit $t_{n-d,1-\alpha}$ le quantile symétrique d'ordre $1-\alpha$ de la loi $\mathscr{T}(n-d)$. Dans un modèle gaussien, le test ayant pour région de rejet*

$$\left\{ \frac{|\hat{\theta}_j|}{\hat{\sigma}(\hat{\theta}_j)} > t_{n-d,1-\alpha} \right\}$$

*est un test de significativité de $\theta_j$ au niveau $1-\alpha$.*

*Lorsque le modèle n'est pas gaussien mais vérifie les conditions de Section 10.2, ce test est asymptotiquement de niveau $1-\alpha$.*

La statistique $|\hat{\theta}_j|/\hat{\sigma}(\hat{\theta}_j)$ qui apparaît ci-dessus est appelée `t` *de Student*. Les outils usuels de statistique donnent fréquemment la valeur de cette statistique pour chaque coefficient d'une régression, ainsi que la $p$-valeur du test qui est égale à

$$1 - F_{n-d}(\mathtt{t}),$$

où $F_{n-d}$ est la fonction de répartition d'une loi $\mathscr{T}(n-d)$. Cette quantité est fréquemment notée `Prob>t`.

## 12.2 Test de contraintes linéaires

Les tests de contraintes linéaires consistent à tester si $\theta$ vérifie une équation linéaire. Le test de significativité est un test de contrainte linéaire : en notant $\delta_j$ le vecteur avec des zéro partout sauf en $j$, il s'agit du test de $\langle \delta_j, \theta \rangle = 0$. On pourrait cependant vouloir tester beaucoup d'autres contraintes de ce type : par exemple, savoir si l'influence de la variable $i$ et de la variable $j$ sont identiques se traduit par $\theta_i = \theta_j$, ou encore $\langle \delta_i - \delta_j, \theta \rangle = 0$.

Formellement, un test de contrainte linéaire consiste à tester si $\theta$

$$C\theta = c.$$

où $C$ une matrice $r \times d$ et $c$ un vecteur de taille $r$. Comme $C$ possède $r$ lignes, cela signifie que l'on teste les $r$ contraintes $\langle C_i, \theta \rangle = c_i$, où $C_i$ est la $i$-ème ligne de $C$.

Sous cette hypothèse nulle,

$$C\hat{\theta} - c \sim N(0, \sigma^2 C(X^\top X)^{-1} C^\top).$$

En multipliant par la matrice $[\sigma^{-2} C(X^\top X)^{-1} C^\top]^{-1/2}$ puis en prenant la norme au carré et en simplifiant l'expression, on voit que

$$\frac{1}{\sigma^2} \langle C\hat{\theta} - c, [C(X^\top X)^{-1} C^\top]^{-1} (C\hat{\theta} - c) \rangle \sim \chi^2(r).$$

Maintenant, si l'on estime le terme $\sigma^2$ comme d'habitude et que l'on utilise le théorème Théorème 10.1, on obtient la loi de la statistique de test (une loi de Fisher), résumée dans le théorème suivant.

**Théorème 12.2** (Test de contraintes linéaires)**.** *Sous l'hypothèse nulle $C\theta = c$, on a*

$$\frac{\langle C\hat{\theta} - c, [C(X^\top X)^{-1} C^\top]^{-1} (C\hat{\theta} - c) \rangle / r}{\hat{\sigma}_n^2} \sim \mathscr{F}_{r, n-d}. \tag{12.1}$$

La statistique Équation 12.1 est couramment appelée *statistique de Wald* associée au système linéaire $C\theta = c$. Formellement, la région de rejet du test de niveau $1 - \alpha$ de ce l'hypothèse nulle $C\theta = c$ est donc donnée par

$$\left\{ \frac{\langle C\hat{\theta} - c, [C(X^\top X)^{-1} C^\top]^{-1} (C\hat{\theta} - c) \rangle / r}{\hat{\sigma}_n^2} > f_{1-\alpha}^{r, n-d} \right\}$$

où $f$ est le quantile d'ordre $1 - \alpha$ de la loi $\mathscr{F}_{r, n-d}$.

## 12.3 Test de significativité globale de Fisher

La significativité *globale* de la régression consiste à tester si tous les coefficients sont significatifs, sauf la constante. Il s'agit donc du test de l'hypothèse nulle

$$\theta_2 = \cdots = \theta_d = 0.$$

Il s'agit bien d'un test de contraintes linéaires au sens du paragraphe précédent : il y a exactement $d-1$ contraintes linéaires. En notant $C$ la matrice de taille $(d-1,d)$

$$C = \begin{pmatrix} 0 & 1 & & 0 \\ \vdots & & & \\ 0 & \dots & & 1 \end{pmatrix},$$

on teste bien $C\theta = 0$. La matrice $C(X^\top X)^{-1}C^\top$ n'est autre que le bloc $B_X$ obtenu à partir de $(X^\top X)^{-1}$ en lui enlevant la première ligne et la première colonne (qui correspondent à la constante). La statistique de test devient alors

$$\frac{\langle \hat{\theta'}, B_X^{-1}\hat{\theta'}\rangle}{(d-1)\hat{\sigma}_n^2}. \tag{12.2}$$

Cette quantité peut sembler difficile à calculer : elle ne l'est pas, et s'exprime à l'aide du coefficient de détermination Définition 9.1.

**Théorème 12.3.**

$$\frac{R^2}{1-R^2}\frac{n-d}{d-1} \sim \mathscr{F}_{d-1,n-d}.$$

*Démonstration.* Il suffit de montrer que l'expression dans Équation 12.2 est égale à $(n-d)R^2/(d-1)(1-R^2)$.

□

# Exercices

## Questions

- Retrouver la formule des MCO par une méthode analytique.
- Construire un intervalle de confiance de niveau $1-\alpha$ pour le coefficient $\theta_j$ d'une régression.
- Soit $X$ une matrice. Pourquoi les nombres $\ell_j^2 = ((X^\top X)^{-1})_{j,j}$ sont-ils toujours des nombres positifs ?
- Écrire explicitement les deux leviers dans un modèle linéaire simple en une dimension.
- Concrètement, comment s'interprète la condition "la matrice des variables explicatives $X$ est de rang $d$" ? Qu'est-ce qui ne va pas lorsque ce n'est pas le cas ?
- Au lieu de faire un test de significativité sur un coefficient d'une régression linéaire (Théorème 12.1), tester $\theta_j = x$ pour n'importe quel $x$ (pas forcément 0).
- Dans un ajustement affine sans constante $y_i = \beta x_i$, monter que $\hat{\beta} = \sum_{i=1}^n p_i y_i$ où $p_i = x_i/|x|^2$.
- Calculer la limite en loi de $\mathscr{F}_{r,n}$ lorsque $r$ est fixé et $n \to \infty$.
- Calculer la limite en loi de $\mathscr{F}_{n,n}$ lorsque $n \to \infty$.

## Exercices

**Exercice 12.1** (Les limites du coefficient de détermination)**.**

1. Construire un jeu de variables explicatives $x_i$ et expliquées $y_i$ tel que l'ajustement affine des $x$ vers les $y$ possède un $R^2$ égal à 0 (aucune significativité linéaire), mais tel que les $y_i$ sont parfaitement déterminés par les $x_i$ (c'est-à-dire tels qu'il y a une fonction $f$ avec $f(x_i) = y_i$ pour tout $i$).
2. Montrer qu'ajouter des variables explicatives dans un modèle augmente le coefficient de détermination.

**Exercice 12.2** (Pénalité $\ell^2$ (régression *ridge*))**.** Dans une régression de type $Y = X\theta + \varepsilon$, on s'intéresse au problème

$$\arg\min |Y - X\theta|^2 + \lambda|\theta|^2.$$

Il s'agit du problème des moindres carrées, mais où la présence de la pénalité $\lambda|\theta|^2$ impose que les coefficients de $\theta$ ne soient pas trop grands au sens $\ell^2$.

1. Montrer que la solution au problème est donnée par $\widehat{\theta}_\lambda = (X^\top X + \lambda I_d)^{-1} X^\top Y$ sans aucune contrainte de rang sur $X$.
2. Calculer la loi de $\widehat{\theta}_\lambda$ lorsque les résidus sont gaussiens. Quel est son biais ?

**Exercice 12.3.** Dans un modèle linéaire $Y = X\theta + \varepsilon$, on cherche à tester une unique contrainte linéaire, à savoir $\langle z, \theta\rangle = c$ où $z \in \mathbb{R}^d$ et $c \in \mathbb{R}$.

1. Montrer que dans ce cas, la statistique de Wald s'écrit

$$|\langle z, \hat{\theta}\rangle - c|^2/\hat{\sigma}_n^2\langle z, (X^\top X)^{-1}z\rangle.$$

   Écrire le test associé à cette statistique.
2. Trouver un estimateur $\hat{\tau}^2$ de la variance de $\langle z, \hat{\theta}\rangle - c$. Sous l'hypothèse nulle, quelle est la loi de $(\langle z, \hat{\theta}\rangle - c)/\hat{\tau}_n$ ? En déduire un test associé à cette statistique.
3. Montrer que les deux tests sont équivalents.

**Exercice 12.4.** Soient $(x_1, y_1), \dots, (x_n, y_n)$ des points dans $\mathbb{R}^2$. Comparer l'ajustement affine des $x$ vers les $y$, et l'ajustement affine des $y$ vers les $x$.

**Exercice 12.5** (Théorème de Frish-Waugh)**.** Soit $Y$ un vecteur de variables à expliquer de taille $n$, et soient $X, Z$ deux matrices de variables explicatives, de dimensions $(n, d)$ et $(n, f)$. Soient $\hat{\theta}$ et $\hat{\theta}_X$ les estimateurs des moindres carrés des deux régressions $Y = X\theta + Z\mu + \varepsilon$ d'une part, et $PY = PX\theta + P\varepsilon$ d'autre part, où $P$ est la matrice de projection sur le sous-espace vectoriel orthogonal aux colonnes de $Z$. Montrer que $\hat{\theta} = \hat{\theta}_X$.

**Exercice 12.6** (Test de Chow)**.** On dispose de deux jeux de données, disons $(X_1^1, Y_1^1), \dots, (X_n^1, Y_n^2)$ et $(X_1^2, Y_1^2), \dots, (X_m^2, Y_m^2)$. Dans les deux régressions $Y^1 = X^1\theta^1 + \varepsilon^1$ et $Y^2 = X^2\theta^2 + \varepsilon^2$, on souhaite tester si $\theta^1 = \theta^2$.

1. On suppose dans un premier temps que les erreurs $\varepsilon^1, \varepsilon^2$ ont la même loi, avec une variance $\sigma^2$ connue. Proposer un test simple de l'hypothèse nulle.
2. Même question lorsque $\sigma$ n'est pas connue.
3. Même question lorsque $\varepsilon^1, \varepsilon^2$ n'ont pas la même variance.

**Exercice 12.7.** On se place dans un modèle linéaire *gaussien* de la forme $Y = X\theta + \varepsilon$, mais on suppose que les entrées de $\varepsilon_i$ ne sont plus iid, mais possèdent une covariance $\Sigma$ non scalaire.

1. Si l'on connaît $\Sigma$, on pose $Y' = \Sigma^{-1/2}Y$ et $X' = \Omega^{-1/2}X$. Montrer que

$$\hat{\theta}_{\text{MCG}} = (X^\top\Sigma^{-1}X)^{-1}X^\top\Sigma^{-1}Y$$

   est un estimateur sans biais de $\theta$, appelé *estimateur des moindres carrés généralisés*.
2. On suppose qu'on dispose d'un estimateur $\hat{\Sigma}$ de $\Sigma$. Montrer que

$$(X^\top\hat{\Sigma}^{-1}X)^{-1}X^\top\hat{\Sigma}^{-1}Y$$

   est un estimateur asymptotiquement normal et sans biais de $\theta$.

**Exercice 12.8.** On considère le modèle de régression linéaire $y_i = b_0 + b_1x_i + \varepsilon_i$ où $i = 1, \dots, n$ et les $\varepsilon_i$ sont des variables aléatoires indépendantes $\mathcal{N}(0, \sigma^2)$ et $b_0, b_1$ et $\sigma^2$ sont inconnus.

1. Donner les estimateurs des moindres carrés ordinaires $\hat{b}_0, \hat{b}_1$ et $\hat{\sigma}^2$ et leur loi jointe.
2. On dispose d'une nouvelle observation, disons $y_0$, pour laquelle la valeur de $x_0$ de la variable explicative est inconnue. L'objectif est d'estimer $x_0$. On suppose que $y_0$ est une réalisation d'une variable aléatoire $Y_0$ s'écrivant $Y_0 = b_0 + b_1x_0 + \eta$, avec $\eta$ une erreur d'observation de loi $N(0, \sigma^2)$ indépendante des $\varepsilon_i$. On suppose en outre que la variable que l'on cherche, $x_0$, n'est pas trop éloignée des autres $x_i$ : $|x_0 - \bar{x}| \leq 1$.

i) Quelle est la loi de $Y_0 - \widehat{b}_0 - \widehat{b}_1 x_0$ ?
ii) En utilisant l'estimateur $\widehat{\sigma}$ de $\sigma$, déterminer un intervalle de confiance de niveau $1-\alpha$ pour $x_0$.

3. On dispose maintenant de $m$ observations $y_{0,1}, \dots, y_{0,m}$ correspondant à la valeur $x_0$ inconnue ; ce sont des réalisations de copies indépendantes $Y_{0,j}$ de $Y_0$.

   i) Montrer que
   $$\tilde{\sigma}^2 = \frac{(n-2)\widehat{\sigma}^2 + \sum_{j=1}^{m} (Y_{0j} - \bar{Y}_0)^2}{n+m-3}$$
   est un estimateur sans biais de $\sigma^2$. Quelle est sa loi ?
   ii) Quelle est la loi de $\bar{Y}_0 - \widehat{b}_0 - \widehat{b}_1 x_0$ ?
   iii) A l'aide de $\tilde{\sigma}^2$ et de $\bar{Y}_0$, donner un intervalle de confiance pour $x_0$ de niveau $1-\alpha$.
   iv) Aurait-on pu construire un intervalle de confiance pour $x_0$ à l'aide de $\widehat{\sigma}^2$ et de $\bar{Y}_0$ ?

**Exercice 12.9** (Théorème de Gauss-Markov)**.** On se place dans un modèle linéaire gaussien $Y = X\theta + \varepsilon$. L'objectif est de montrer que $\widehat{\theta}$, l'estimateur des moindres carrés, est le meilleur estimateur linéaire de $\theta$ qui soit sans biais[1]. Soit donc $\tilde{\theta}$ un autre estimateur linéaire sans biais, disons $\tilde{\theta} = MY$.

1. Montrer que $(M - (X^\top X)^{-1} X^\top) X = 0$.
2. Calculer la matrice de variance de $\tilde{\theta}$ en fonction de $M - (X^\top X)^{-1} X^\top$ et conclure.

[1] BLUE, *best linear unbiased estimator*.

# 13 Modèles exponentiels

## Exemples

Jusqu'ici, nous avons vu de nombreux exemples de modèles statistiques. Dans la plupart des cas, il s'agissait de modèles de la loi de $n$ variables aléatoires indépendantes et identiquement distribuées selon une même loi $P$ (le modèle était donc $P^{\otimes n}$). Cette loi $P$ possède souvent une densité par rapport à une mesure de référence. Par exemple, la loi gaussienne a une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}$ :

$$\frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}} = \frac{1}{\sqrt{2\pi} e^{\frac{\mu^2}{2}}} e^{-\frac{x^2}{2}} e^{x\mu}.$$

La loi exponentielle a une densité par rapport à la mesure de Lebesgue sur $\mathbb{R}_+$ :

$$\frac{1}{1/\lambda} e^{-\lambda x}$$

Les lois discrètes ont une densité par rapport à la mesure de comptage : la loi de Bernoulli, par exemple, s'écrit

$$p^n (1-p)^{1-n} = \frac{e^{n \ln(p/(1-p))}}{(1-p)^{-1}}$$

avec $n$ valant zéro ou 1, ou encore la loi de Poisson

$$e^{-\lambda} \frac{\lambda^n}{n!} = \frac{1}{e^{e^{\ln(\lambda)}}} \frac{1}{n!} e^{-\ln(\lambda) n}$$

ou enfin la loi géométrique

$$p^n (1-p) = \frac{e^{n \ln(p)}}{(1-p)^{-1}}.$$

Dans tous ces exemples, j'ai volontairement écrit la densité de façon inhabituelle : tous ces modèles peuvent s'écrire sous la forme

$$\frac{1}{Z(\theta)} h(x) e^{\theta f(x)}$$

où $f$ et $g$ sont des fonctions qui ne dépendent pas de $\theta$, et où $Z$ est une constante qui ne dépend que de $\theta$. Ces modèles appartiennent à la famille des *modèles exponentiels*.

## 13.1 Définitions

Soit $\nu$ une mesure de référence ($\sigma$-finie) sur $\mathbb{R}^d$.

Soit $\Theta \subset \mathbb{R}^p$ (l'espace des paramètres) et soit $T : \mathbb{R}^d \to \mathbb{R}^p$ une fonction mesurable.

On suppose que pour tout $\theta \in \Theta$, la fonction $x \mapsto e^{\langle \theta, T(x) \rangle}$ est intégrable par rapport à $\nu$ : son intégrale

$$Z_\theta = \int e^{\langle \theta, T(x) \rangle} \nu(dx)$$

est appelée *fonction de partition*.

**Définition 13.1.** Le modèle exponentiel associé à $T$ est la famille de densités (par rapport à $\nu$) définie par

$$p_\theta(x) = \frac{e^{\langle \theta, T(x) \rangle}}{Z_\theta}.$$

Lorsqu'on fixe un $x$ dans l'espace des observations, la fonction

$$\theta \mapsto p_\theta(x)$$

est appelée *vraisemblance* et son logarithme

$$\ell_\theta(x) = \ln p_\theta(x)$$

est appelé *log-vraisemblance*. Lorsqu'il est bien défini, le gradient (en $\theta$) de la log-vraisemblance

$$\nabla \ln p_\theta(x)$$

est appelé *fonction de score* du modèle. Ces termes ne sont pas propres aux modèles exponentiels. On omet fréquemment de noter la dépendance en les observations, qu'on suppose fixées une bonne fois pour toutes.

## 13.2 Retour sur des exemples

**Exemple 13.1** (Gaussiennes)**.** La densité de $N(\mu, \sigma^2)$ s'écrit

$$\frac{e^{-\frac{\mu^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}} e^{\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2}.$$

La mesure $\nu$ est la mesure de Lebesgue sur $\mathbb{R}$. Le moment $T$ est

$$T(x) = \begin{bmatrix} x \\ -x^2/2 \end{bmatrix}.$$

Le bon paramètre $\theta$ est

$$\theta = \begin{bmatrix} \mu/\sigma^2 \\ 1/\sigma^2 \end{bmatrix}.$$

La fonction de partition $\exp\left(\frac{\mu^2}{2\sigma^2}\right)\sqrt{2\pi\sigma^2}$ s'écrit donc aussi

$$Z(\theta) = \exp(\theta_1^2/2\theta_2)\sqrt{2\pi/\theta_2}.$$

L'exemple de la loi Gaussienne montre qu'en règle générale, il peut être nécessaire de « reparamétriser » le modèle pour l'écrire sous sa forme exponentielle.

**Exemple 13.2** (Poisson)**.** La mesure $\nu$ est la mesure de comptage sur $\mathbb{N}$. Le paramètre est

$$\theta = \begin{bmatrix} \ln(\lambda) \\ 1 \end{bmatrix}$$

et le moment $T$ est

$$T = \begin{bmatrix} -n \\ -\ln(n!) \end{bmatrix}$$

de sorte que la fonction de partition est $Z(\theta) = e^{e^{\ln \lambda}} = e^{e^{\theta_1}}$.

## 13.3 Régularité

On supposera dorénavant que l'espace des paramètres $\Theta$ (qui est une partie de $\mathbb{R}^p$) est un ouvert, et que $Z(\theta)$ est fini pour tout $\theta \in \Theta$. Cela sera aussi valable pour les sections suivantes.

**Proposition 13.1.**

*1) Pour tout $n$, $E_\theta[|T(X)|^n]$ est fini.*

*2) La fonction de partition d'un modèle exponentiel est infiniment différentiable.*

*3) Le gradient de la log-partition est donné par*

$$\nabla \ln Z(\theta) = E_\theta[T(X)] \tag{13.1}$$

*et sa matrice hessienne[1] par*

$$\nabla^2 \ln Z(\theta) = \mathrm{Var}_\theta(T(X)). \tag{13.2}$$

*Démonstration.* 1. Prenons un petit $\delta$ tel que $\theta + \delta$ et $\theta - \delta$ sont dans $\Theta$. Comme $Z(\theta \pm \delta) = \int e^{\langle \theta, T(x)\rangle \pm \langle \delta, T(X)\rangle} \nu(dx)$ et que $e^{|\langle \delta, T(x)\rangle|} \leqslant e^{\langle \delta, T(x)\rangle} + e^{\langle -\delta, T(x)\rangle}$, on en déduit que

$$\int e^{\langle \theta, T(x)\rangle + |\langle \delta, T(x)\rangle|} \nu(dx) < \infty.$$

Le théorème d'interversion série-intégrale à termes positif montre que ce terme est aussi égal à

$$\sum_{n=0}^{\infty} \int e^{\langle \theta, T(x)\rangle} \frac{|\langle \delta, T(x)\rangle|^n}{n!} \nu(dx).$$

Tous les termes de cette somme sont donc finis, ce qui signifie précisément que $E_\theta[|\langle \delta, T(X)\rangle|^n] < \infty$ pour tout $n$. Ceci étant valable pour tout $\delta$ dans une boule autour de $\theta$, il est immédiat d'en déduire que $E_\theta[|\langle x, T(X)\rangle|^n]$ est fini pour tout $n$ et pour tout $x$. En prenant $x = \delta_i$, on voit donc que les coordonnées de $T$ ont des moments finis à tous les ordres, et donc que $T$ possède des moments finis à tous les ordres au sens où $E_\theta[|T(X)|^n] < \infty$.

[1] On rappelle que la Hessienne d'une fonction $f : \mathbb{R}^d \to \mathbb{R}$ est la matrice de ses dérivées secondes

$$\frac{\partial^2}{\partial x_i \partial x_j} f(x_1, \dots, x_d).$$

Par abus de notation, on la note souvent $\nabla^2$, mais il serait plus juste de la noter $\nabla^\top \nabla$.

2. On a $\nabla \ln Z(\theta) = \nabla Z(\theta)/Z(\theta)$. Formellement, $\nabla Z(\theta)$ est donc égal à

$$\nabla \int e^{\langle \theta, T\rangle} = \int \nabla e^{\langle \theta, T\rangle} = \int T e^{\langle \theta, T\rangle}.$$

Il est alors clair que $\nabla \ln Z(\theta) = \int p_\theta T$. Pour justifier la dérivation sous le signe intégral, notons $f(\theta, x) = e^{\langle \theta, T(x)\rangle}$. Elle est infiniment différentiable en $\theta$ et son intégrale est finie dès que $\theta$ est dans $\Theta$. Son gradient en $\theta$ est égal à $e^{\langle \theta, T(x)\rangle} T(x)$ qui est d'intégrale finie d'après le premier point. En regardant bien la démonstration, on constate également qu'il y a une constante $c$ telle que pour tout $\delta$ dans un voisinage de $\theta$, on a $E_{\theta+\delta}[|T(X)|] < c$. Tout cela permet d'appliquer le théorème de dérivation sous le signe intégral et la formule Équation 13.1.

3. Pour la formule Équation 13.2, c'est la même chose : $\nabla^2 \ln Z(\theta) = \nabla \frac{\nabla Z(\theta)}{Z(\theta)}$. Les règles de dérivation des produits disent alors que ce terme est égal à

$$\frac{Z(\theta)\nabla^2 Z(\theta) - \nabla Z(\theta) \nabla Z(\theta)^\top}{Z(\theta)^2}.$$

Il suffit donc de calculer $\nabla^2 Z(\theta)$, qui par un argument de dérivation sous l'intégrale similaire au précédent, est égal à

$$\int e^{\langle \theta, T(x)\rangle} T(x) T(x)^\top \nu(dx).$$

La formule Équation 13.2 découle alors de la définition de la covariance.

□

## 13.4 Identifiabilité

**Théorème 13.1.** *Dans un modèle exponentiel, les points suivants sont équivalents.*

*i) Le modèle est identifiable.*
*ii) La matrice hessienne de la fonction de log-partition (Équation 13.2) est inversible en tout $\theta$.*
*iii) $\nabla \ln Z(\theta)$ est un difféomorphisme.*

*Démonstration.* Démontrons d'abord l'équivalence des deux premiers points.

La matrice hessienne de $\ln Z_\theta$ est égale à $\mathrm{Var}_\theta(T(X))$, donc elle est toujours positive. Si elle n'est pas inversible, alors elle n'est pas définie positive et il existe un $\mu$ tel que $\mu^\top \mathrm{Var}_\theta(T)\mu$ vaut zéro. Or, ce terme est aussi égal à $\mathrm{Var}_\theta(\langle \mu, T\rangle)$. Cela impliquerait que la variable aléatoire $\langle \mu, T(X)\rangle$ soit constante $P_\theta$-presque sûrement, disons égale à un certain $\alpha$, et donc que $\nu$-presque sûrement, $\langle \mu, T(x)\rangle = \alpha$. Mais alors, $p_{\theta+\mu}(x)$ peut s'écrire

$$\frac{e^{\langle \theta+\mu, T(x)\rangle}}{Z(\theta+\mu)} = \frac{e^{\langle \theta, T(x)\rangle} e^{\alpha}}{Z(\theta+\mu)}$$

c'est-à-dire

$$p_\theta(x) \times \frac{e^\alpha Z(\theta)}{Z(\theta+\mu)}.$$

Or, comme $p_\theta$ est une densité de probabilité, son intégrale vaut 1 : la constante $e^\alpha Z(\theta)/Z(\theta+\mu)$ vaut donc 1 et l'on a montré que $p_{\theta+\mu}$ et $p_\theta$ sont égales partout. Le modèle n'est donc pas identifiable.

Pour l'autre sens, il suffit de reprendre toutes ces implications à l'envers : si $p_{\theta+\mu} = p_\theta$, alors pour $\nu$-presque tout $x$ on aura $\langle \theta + \mu, T(x)\rangle = \langle \theta, T(x)\rangle$, et donc $\langle \mu, T(x)\rangle = 0$, et in fine $\mu^\top \mathrm{Var}_\theta(T(X))\mu = 0$.

Comme iii) entraîne ii), il suffit donc de montrer que i) et ii) entraînent iii). Nous allons montrer la contraposée : si iii) n'est pas vrai et que ii) n'est pas vrai, c'est fini. On peut donc supposer que iii) n'est pas vrai et que ii) est vrai, et il faut montrer que i) est faux. Le point ii) entraîne que $\nabla \ln Z$ est localement injective (théorème d'inversion locale), donc si cette application n'était pas un difféomorphisme, cela voudrait dire qu'elle n'est pas injective et qu'il y aurait donc $\theta \neq \mu$ tels que $\nabla \ln Z(\theta) = \nabla \ln Z(\mu)$. Or, un calcul montre que la divergence de Kullback-Leibler (Définition 7.3) *symétrisée*, à savoir $\mathrm{d}_{\mathrm{KL}}(P_\theta \mid P_\mu) + \mathrm{d}_{\mathrm{KL}}(P_\mu \mid P_\theta)$, est égale à

$$\langle \nabla \ln Z(\theta) - \nabla \ln Z(\mu), \theta - \mu\rangle \tag{13.3}$$

et ceci vaut donc zéro : c'est donc que chacune des deux $\mathrm{d}_{\mathrm{KL}}$ vaut zéro, puisque ces deux termes sont positifs. On en déduit alors que $P_\theta = P_\mu$, et donc le modèle n'est pas identifiable.

□

*Preuve de Équation 13.3.* En utilisant seulement les définitions, $\mathrm{d}_{\mathrm{KL}}(P_\theta \mid P_\mu)$ est égal à

$$\int p_\theta(x)(\langle \theta, T(x)\rangle - \langle \mu, T(x)\rangle)\nu(dx) - \ln Z(\theta) + \ln Z(\mu).$$

Le premier terme est égal à $\langle \theta - \mu, E_\theta[T]\rangle$. La somme $\mathrm{d}_{\mathrm{KL}}(P_\theta \mid P_\mu) + \mathrm{d}_{\mathrm{KL}}(P_\mu \mid P_\theta)$ se simplifie et se factorise donc en

$$\langle \theta - \mu, E_\theta[T] - E_\mu[T]\rangle$$

et l'identité en découle puisque $\nabla \ln Z(\theta) = E_\theta[T]$.

# 14 Maximum de vraisemblance

Dans cette section, on a fixé un modèle exponentiel[1] associé au moment $T$, et l'on dispose d'observations indépendantes $x_1, \dots, x_n$ distribuées selon ce modèle. La densité de chaque observation par rapport à la mesure de référence $\nu$ est donc par $e^{\langle \theta, T(x_i)\rangle}/Z(\theta)$. En particulier, la densité de l'échantillon $x = (x_1, \dots, x_n)$ est $p_\theta(x) := p_\theta(x_1) \cdots p_\theta(x_n)$, c'est-à-dire

$$\frac{e^{\langle \theta, \sum_{i=1}^n T(x_i)\rangle}}{Z(\theta)^n}. \tag{14.1}$$

Cela reste un modèle exponentiel associé à la fonction de moment $(x_1, \dots, x_n) \to T(x_1) + \dots + T(x_n)$ et à la fonction de partition $Z(\theta)^n$.

## 14.1 Définition

**Définition 14.1.** L'estimateur du maximum de vraisemblance (EMV) est le paramètre pour lequel la vraisemblance des observations est maximale :

$$\hat{\theta}_{\text{emv}} = \arg\max_{\theta \in \Theta} p_\theta(x). \tag{14.2}$$

Il n'est pas évident que ce maximum existe, ni que le minimiseur est unique. Il existe un théorème général garantissant son existence et son unicité.

**Proposition 14.1.** *Dans un modèle exponentiel identifiable dont l'espace des paramètres $\Theta \subset \mathbb{R}^p$ est un ouvert convexe, sous certaines hypothèses, l'estimateur Équation 14.2 existe. Dans tous les cas, s'il existe, il est unique.*

Trouver le maximum d'une fonction positive $f(x)$ et trouver le maximum de son logarithme $\ln f(x)$ reviennent au même : or, il est souvent plus facile dans les modèles exponentiels de maximiser le *logarithme* de la vraisemblance $\ell(\theta)$, qui dans un modèle de la forme Équation 14.1 s'écrit

$$\sum_{i=1}^n \langle \theta, T(x_i)\rangle - n \ln Z(\theta). \tag{14.3}$$

*Démonstration.* La démonstration du théorème ci-dessus repose sur des outils analytiques simples. Dans Équation 14.3, le premier terme est une fonction linéaire. Quant au terme $\ln Z(\theta)$, sa matrice Hessienne n'est autre (Équation 13.2) qu'une matrice de variance, donc positive : $\ln Z(\theta)$ est donc convexe, et même *strictement* si le modèle est identifiable (Théorème 13.1). Ainsi, Équation 14.3 est presque sûrement strictement concave. Cela suffit à assurer que le maximum, *s'il existe*, est unique. Quant à son existence,

[1] On se restreindra toujours aux modèles exponentiels qui satisfont les propriétés de la section précédente.

elle nécessite des hypothèses sur $\ell$ ou sur $\Theta$ et je ne vois pas l'intérêt d'en énoncer de générales : ce sera au cas par cas. Mais typiquement, on peut demander à ce que $\ell(\theta) \to -\infty$ lorsque $\theta$ tend vers le bord de $\Theta$, ce qui revient à demander que $p_\theta(x) \to 0$.

□

On omettra presque systématiquement le fait que la log-vraisemblance dépend des observations $x_i$, mais **il faut garder en tête que la vraisemblance et la log-vraisemblance sont des variables aléatoires car elles dépendent de l'échantillon**. Parfois, pour indiquer quand même que l'échantillon comporte $n$ éléments, on notera $\ell_n(\theta)$. En règle générale, Équation 14.2 est donc équivalent au problème du maximum de log-vraisemblance,

$$\hat{\theta}_{\text{emv}} = \arg\max \ell(\theta).$$

## 14.2 L'EMV et les moments

L'EMV maximise la log-vraisemblance. Lorsqu'il existe et qu'il est unique, il est donc l'unique solution de $\nabla_\theta \ell(\theta) = 0$. En dérivant Équation 14.3, cette équation s'écrit encore

$$\frac{1}{n}\sum_{i=1}^{n} T(x_i) = \nabla \ln Z(\theta).$$

Or, nous avons vu (Théorème 13.1) que si le modèle est identifiable, le terme de droite, noté $\varphi(\theta)$, est un difféormorphisme. Le maximum de vraisemblance vérifie donc l'équation des moments, $\varphi(\hat{\theta}_{\text{emv}}) = \bar{T}_n$, où $\bar{T}_n = (T(x_1) + ... + T(x_n))/n$. On peut donc appliquer le théorème des moments Théorème 3.1. L'hypothèse selon laquelle $T$ est de carré intégrable vient directement de Proposition 13.1.

**Théorème 14.1.** *Dans un modèle iid, l'estimateur du maximum de vraisemblance vérifie*

$$\hat{\theta}_{\text{emv}} = \varphi^{-1}(\bar{T}_n)$$

*où $\varphi(\theta) = \nabla \ln Z(\theta) = E_\theta[T(X)]$. Par ailleurs, cet estimateur est convergent et asymptotiquement normal : $\sqrt{n}(\hat{\theta}_{\text{emv}} - \theta)$ converge en loi vers $N(0, I(\theta)^{-1})$ où*

$$I(\theta) = \text{Var}_\theta(T).$$

*Démonstration.* L'application du théorème des moments ayant été justifiée plus haut, il suffit de vérifier que l'expression de la variance asymptotique coïncide avec $I(\theta)^{-1}$. Le Théorème 3.1 dit que $\sqrt{n}(\hat{\theta}_{\text{emv}} - \theta)$ converge vers une gaussienne centrée de variance

$$D\varphi(\theta)^{-1}\text{Var}_\theta(T)(D\varphi(\theta)^{-1})^\top.$$

Or, Équation 13.2 montre que $D\varphi(\theta) = \nabla^2 \ln Z(\theta)$ vaut également $\text{Var}_\theta(T)$, d'où la simplification.

□

Il se trouve que la matrice $\text{Var}_\theta(T)$ est centrale dans la théorie des statistiques : il s'agit de la *matrice d'information de Fisher*, que nous étudierons dans la prochaine section.

## 14.3 Problème d'optimisation

Dans les modèles exponentiels usuels où les paramètres ont peu de dimensions, il est aisé de maximiser la vraisemblance en résolvant l'équation $\nabla \ell(\theta) = 0$ par des méthodes analytiques simples. Mais hors du giron des modèles classiques, on n'utilise presque jamais la formulation abstraite de Théorème 14.1. La raison principale est que, *même dans les modèles exponentiels*, la fonction de partition $Z(\theta)$ peut être très difficile à inverser – et parfois n'est même pas connue. Par exemple, un choix aussi simple que

$$T(x) = -\begin{bmatrix} x^2 \\ x^4 \end{bmatrix}$$

donne naissance à $Z(\theta) = \int e^{-\theta_1 x^2 - \theta_2 x^4} dx$ dont la formule exacte qui s'exprime via des fonctions hypergéométriques. Même si l'on accède à $\nabla \ln Z(\theta)$, il faut encore savoir en calculer l'inverse !

Dans ces cas, on maximise directement la vraisemblance en utilisant un algorithme d'optimisation, qui fournira donc une approximation de $\hat{\theta}_{\text{emv}}$ : typiquement, une variante des algorithmes de montée de gradient[2], dont la version la plus simple est

$$\theta_{t+1} - \theta_t = \eta \nabla \ell(\theta_t)$$

où $\eta$ est le *pas* de la montée de gradient.

## 14.4 Exemple

Pour illustrer le propos, regardons l'exemple classique de l'estimation de $\mu$ dans un modèle $N(\mu, 1)$, à partir de $n$ observations indépendantes. La log-vraisemblance $\ell(\mu)$ du modèle est

$$\sum_{i=1}^{n} -\frac{(x_i - \mu)^2}{2} - \frac{n}{2} \ln(2\pi).$$

Sa dérivée $\ell'(\mu)$ est égale à

$$\sum_{i=1}^{n} (x_i - \mu).$$

Le maximum de vraisemblance existe et il est unique, car le modèle est exponentiel et identifiable. Il n'y a donc qu'un seul point critique (qui vérifie $\ell'(\mu) = 0$) et celui-ci est donné par

$$\hat{\mu}_{\text{emv}} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}_n.$$

Sans surprise, l'EMV est donc bien la moyenne empirique.

[2] Le monde de l'optimisation ayant été habitué à minimiser des fonctions, les statisticiens ont pris l'habitude d'utiliser des *descentes* de gradient pour minimiser l'opposé de la log-vraisemblance.

# 15 L'information de Fisher

## 15.1 Définitions

Nous avons vu apparaître naturellement la *variance* du moment dans un modèle exponentiel, à savoir $\mathrm{Var}_\theta(T)$. Cette quantité s'appelle *information de Fisher*, parce qu'elle quantifie l'information relative au paramètre $\theta$ qui est « contenue » dans la distribution $p_\theta$.

**Définition 15.1** (Information de Fisher)**.** La matrice d'information de Fisher $I(\theta)$ est la matrice de covariance de $T$,

$$E_\theta[T(X)T(X)^\top] - E_\theta[T(X)]E_\theta[T(X)]^\top.$$

L'information de Fisher possède de nombreuses expressions alternatives. La plus importante, outre la définition, est qu'on peut interpréter $I(\theta)$ comme la matrice de covariance du *score* du modèle.

**Définition 15.2** (Fonction de score)**.** Le score est la dérivée de la log-vraisemblance :

$$s_\theta(x) = \nabla_\theta \ln p_\theta(x).$$

Dans un modèle exponentiel $p_\theta(x) = \exp(\langle\theta, T(x)\rangle - \ln Z_\theta)$, nous avons déjà vu que

$$s_\theta(x) = T(x) - \nabla \ln Z(\theta). \tag{15.1}$$

Le score dépend des observations, et donc est une variable aléatoire. En fait, Équation 13.1 montre que l'espérance du score, $E_\theta[T(X)] - \nabla \ln Z(\theta)$, vaut précisément zéro : le score est centré. Au vu de Équation 15.1, el est donc clair que

$$I(\theta) = \mathrm{Var}_\theta(s_\theta(X)).$$

## 15.2 Lien avec l'entropie

**Définition 15.3** (Entropie)**.** L'entropie d'une loi de densité $f$ par rapport à la mesure de référence $\nu$ est donnée par

$$\mathrm{Ent}(f) = -\int f(x) \ln f(x) \nu(dx).$$

Nous reviendrons plus tard sur cette quantité. Dans le cas d'un modèle exponentiel, l'entropie est égale à $E_\theta[\ln p_\theta(X)]$ par définition. On peut interpréter $I(\theta)$ comme la *courbure moyenne de l'entropie*.

**Proposition 15.1.** $I(\theta) = -\nabla_\theta^2 \mathrm{Ent}(p_\theta)$.

*Démonstration.* Il suffit de dériver deux fois sous l'intégrale et d'utiliser Équation 13.2.

□

## 15.3 Borne de Cramér-Rao

**Théorème 15.1** (Borne de Cramér-Rao)**.** *Pour tout estimateur sans biais $\hat{\theta}$ de $\theta$, on a*[1] $I(\theta)^{-1} \preceq \mathrm{Cov}_\theta(\hat{\theta})$.

Lorsque le paramètre $\theta$ est réel, la borne de Cramér-Rao dit que le risque quadratique de n'importe quel estimateur sans biais ne peut pas être plus petit que $1/I(\theta)$. Les estimateurs sans biais qui atteignent cette borne sont appelés *efficaces*, ou asymptotiquement efficaces si leur risque quadratique converge vers cette borne.

*Démonstration.* Commençons par la dimension 1. Comme $T$ est sans biais, $\int p_\theta(x)T(x)dx = \theta$. Comme $\nabla p_\theta = p_\theta \nabla_\theta \ln p_\theta = p_\theta s_\theta$, en intervertissant intégrale et dérivée, on obtient donc $1 = \int p_\theta(x)s_\theta(x)T(x)dx = E_\theta[s_\theta(X)T(X)]$. Nous avons déjà vu que le score est centré : ainsi, ce dernier terme vaut aussi $E_\theta[s_\theta(X)(T(X)-\theta)]$. L'inégalité de Cauchy-Schwarz donne alors

$$1 \leqslant \sqrt{E_\theta[|T(X)-\theta|^2]I(\theta)},$$

qui est le résultat voulu. Pour la dimension supérieure, il suffit d'appliquer ce résultat à $\langle y, T(X)\rangle$, qui est un estimateur sans biais de $\langle y, \theta\rangle$ (ici, $y$ est n'importe quel vecteur de $\mathbb{R}^p$). L'inégalité ci-dessus, après quelques menues manipulations, devient

$$\langle y, I(\theta)^{-1}y\rangle \leqslant \langle y, \mathrm{Cov}_\theta(T)y\rangle,$$

qui montre bien que $I(\theta)^{-1} \preceq \mathrm{Cov}_\theta(T)$.

□

## 15.4 Tests fondés sur l'EMV

L'idée principale de l'EMV (maximiser la vraisemblance) est utilisable pour effectuer des tests. Typiquement, on peut vouloir tester l'hypothèse nulle

$$H_0 : \ \theta \in \Theta_0$$

contre l'hypothèse alternative

$$H_1 : \theta \in \Theta_1$$

où $\Theta_0, \Theta_1$ sont deux régions distinctes de l'espace des paramètres $\Theta$. Dans ces cas, nous pouvons définir deux maximums de vraisemblance, un par hypothèse : par exemple,

$$L_0 = \sup_{\theta\in\Theta_0} p_\theta(x).$$

[1] Rappelons que (cf Section 20.4) lorsque $A, B$ sont des matrices symétriques, $A \preceq B$ équivaut à ce que $\langle y, Ay\rangle \leqslant \langle y, By\rangle$ pour tout $y$.

Dans le cas où les régions $\Theta_0, \Theta_1$ sont constituées d'un seul élément, disons $\theta_0, \theta_1$, ces maximums de vraisemblance sont simplement $p_{\theta_0}(x)$ et $p_{\theta_1}(x)$. Dans tous les cas, on peut associer à chaque hypothèse un EMV, par exemple

$$\hat{\theta}_0 = \arg\max_{\theta \in \Theta_0} p_\theta(x),$$

qui s'il existe vérifie $L_0 = p_{\hat{\theta}_0}(x)$.

**Définition 15.4.** Les *tests du rapport de vraisemblance* pour les hypothèses qui ne sont pas forcément simples sont les tests dont la région de rejet est de la forme

$$\left\{ \frac{\sup_{\theta \in \Theta_1} p_\theta(X)}{\sup_{\theta \in \Theta_0} p_\theta(X)} > z \right\}.$$

Lorsque les EMV pour chaque hypothèse existent, cette région de rejet s'écrit donc également

$$\left\{ \frac{p_{\hat{\theta}_0}(X)}{p_{\hat{\theta}_1}(X)} > z \right\}.$$

Malheureusement, il n'y a pas d'équivalent du théorème de Neyman-Pearson (Théorème 7.1, Théorème 7.2) lorsque les hypothèses ne sont pas simples.

## 15.5 Limitations

L'estimation par maximum de vraisemblance, par sa portée théorique autant que pratique, est une référence difficilement contournable. Au vu de son importance, il est de bon aloi d'en cerner les limites.

1. Si la loi $P$ qui a généré les observations n'appartient pas au modèle, l'estimateur n'a aucune chance d'être convergent, même s'il constitue quand même la meilleure estimation possible *dans ce modèle*. Le choix du modèle statistique reste donc un problème fondamental.
2. L'apparente optimalité (au sens de Cramér-Rao) de l'EMV n'est qu'asymptotique. À distance finie, il peut y avoir des estimateurs biaisés ayant un meilleur risque quadratique. Pire, dans des modèles exponentiels élémentaires comme le modèle gaussien $N(\mu, I_p)$ où l'on cherche à estimer une moyenne $\mu \in \mathbb{R}^p$ à partir d'une réalisation, il existe un estimateur dont le risque quadratique est strictement meilleur que l'EMV **quel que soit** $\mu$ : c'est le paradoxe de James-Stein sur lequel nous reviendrons peut-être.
3. Tous les modèles ne sont pas exponentiels. Même si l'estimation par maximum de vraisemblance reste pertinente en général, elle peut aussi donner des résultats peu cohérents, surtout lorsqu'elle est utilisée pour faire des tests (voir par exemple Exercice 8.7).
4. Enfin, même dans les modèles exponentiels, la fonction de partition $Z(\theta)$ peut être inaccessible, en particulier lorsque la dimension de $\theta$ est grande comme en deep learning. L'estimation par maximum de vraisemblance sera alors quasiment infaisable.

# Exercices

## Questions

- On dispose d'une observation $x$ d'une variable aléatoire $N(\mu, 1)$. Quel est l'EMV de $\mu$ ?
- On dispose d'une observation $x$ d'une variable aléatoire $\mathrm{Ber}(p)$. Quel est l'EMV de $\mu$ ?
- Calculer l'EMV de $\mu$ et $\sigma^2$ dans un modèle iid $N(\mu, \sigma^2)$ avec $n$ observations.
- Montrer que dans un modèle linéaire gaussien $Y = X\theta + \varepsilon$, l'estimateur des moindres carrés ordinaires est l'EMV de $\theta$. Quel est l'EMV pour $\sigma^2$ ?
- Dans un modèle exponentiel, construire une région (ellipsoïde) de confiance asymptotique pour le paramètre.
- Soit $\mathcal{X}$ un ensemble fini à $p$ éléments et soient $x_1, \dots, x_n$ des réalisations iid d'une même loi $P$ sur $\mathcal{X}$. On cherche $P$. Montrer que ce problème peut se formuler comme la recherche d'un paramètre (indice : il est de dimension $p$) dans un modèle exponentiel et écrire l'EMV.
- Dans un modèle exponentiel $p_\theta(x) = e^{-\langle \theta, T(x) \rangle} / Z(\theta)$, on a des observations iid $x_1, \dots, x_n$. On note $\hat{\mu}_n$ la loi empirique[2]. des $x_i$ Montrer que l'EMV (lorsqu'il existe) est l'unique $\theta$ tel que

$$\int_{\mathcal{X}} T(x) d\hat{\mu}_n = \int_{\mathcal{X}} T(x) p_\theta(x) dx.$$

- Montrer que la densité gaussienne multidimensionnelle $N(0, \Sigma)$ peut aussi s'écrire

$$\frac{\exp\left(-\langle \theta, xx^\top \rangle_F / 2\right)}{\sqrt{(2\pi)^n \det(\Sigma)}}$$

où $\langle A, B \rangle_F = \mathrm{trace}(AB^\top)$ est le produit scalaire[3] sur l'espace des matrices, et où $\theta = \Sigma^{-1}$.

## Exercices

**Exercice 15.1.** Soit $\hat{\theta}$ l'estimateur du maximum de vraisemblance d'un paramètre $\theta$ dans un modèle régulier. Montrer que $f(\hat{\theta})$ est l'estimateur du maximum de vraisemblance de $f(\theta)$, pour n'importe quelle fonction $\theta$ raisonnable.

**Exercice 15.2.** On observe un échantillon iid $(X_1, \dots, X_n)$ de lois de Laplace, c'est-à-dire de densité $x \mapsto \lambda e^{-\lambda|x-m|}/2$, où $\lambda > 0$ et $m \in \mathbb{R}$.

1. En supposant $\lambda$ connu, proposer un estimateur de $m$ par la méthode des moments, et un estimateur par la méthode du maximum de vraisemblance. Étudier leurs propriétés et les comparer.

[2] C'est-à-dire la mesure de probabilité définie par $n^{-1} \sum \delta_{x_i}$.

[3] Ce produit scalaire est appelé *produit de Frobenius* et correspond à la norme $L^2$ sur l'espace des matrices : $\|A\|_F^2 = \sum_{i,j} |A_{i,j}|^2$.

2. Même question lorsque ni $\lambda$ ni $m$ ne sont connus.

**Exercice 15.3.** Calculer l'estimateur du maximum de vraisemblance et étudier ses propriétés dans les cas suivants :

1. On observe $X_1, \dots, X_n$ de loi de Poisson de paramètre $\lambda > 0$.
2. On observe $X \sim \mathrm{Bin}(n, p)$ où $n$ est connu et $p \in ]0, 1[$.
3. On observe $X_1, \dots, X_n$ de loi $\mathscr{N}(\mu, \sigma^2)$.
4. On observe $X_1, \dots, X_n$ de loi de Pareto $\mathrm{PL}(\alpha, 1)$, dont la densité est $\alpha x^{-\alpha-1}$ sur $[1, \infty[$.

**Exercice 15.4.** On se donne un échantillon $(X_1, \dots, X_n)$ de loi $\Gamma(\alpha, \beta)$[4].

1. On suppose le paramètre $\beta$ connu. Proposer un estimateur de $\alpha$ par la méthode des moments.
2. On suppose à présent que les deux paramètres $\alpha, \beta$ sont inconnus. Proposer un estimateur de $(\alpha, \beta)$ par la méthode des moments.
3. Toujours dans le cas où $\alpha, \beta$ sont inconnus, donner le système d'équation que satisfont les estimateurs de $(\alpha, \beta)$ par la méthode du maximum de vraisemblance.

**Exercice 15.5.** Soit $(X_1, \dots, X_n)$ un $n$-échantillon de la loi uniforme sur $[-\theta, \theta]$, avec $\theta > 0$.

1. Décrire le modèle statistique associé.
2. Proposer un estimateur $\widehat{\theta}_n$ de $\theta$ obtenu par méthode des moments. Est-il consistant? Proposer un intervalle de confiance asymptotique de niveau de confiance $\alpha$.
3. Soit $T_n$ l'estimateur du maximum de vraisemblance de $\theta$. Montrer que pour tout réel $t$,

$$P_\theta^n \left( n(T_n - \theta) \leq t \right) \to e^{t/\theta} \mathbf{1}_{t \leq 0} + \mathbf{1}_{t > 0}$$

   quand $n$ tend vers l'infini. En déduire un intervalle de confiance asymptotique de niveau $\alpha$.
4. Comparer les estimateurs $\widehat{\theta}_n$ et $T_n$ sur la base des longueurs moyennes des intervalles de confiance asymptotiques associés.

**Exercice 15.6.** Soit $c > 0$ un paramètre fixé connu. On considère la loi de Weibull de paramètre $c$, notée $\mathscr{W}(c)$, dont la densité sur $\mathbb{R}_+$ est

$$\lambda c x^{c-1} e^{-\lambda x^c}$$

On observe un $n$-échantillon de loi $\mathscr{W}(c)$, avec $n$ plus grand que 3.

1. Calculer l'estimateur du maximum de vraisemblance $\widehat{\lambda}_n$ de $\lambda$.
2. Calculer son risque quadratique.

**Exercice 15.7.** Dans une urne contenant 1000 tickets, 20 sont marqués $\theta$ et 980 sont marqués $10\theta$, où $\theta$ est un réel strictement positif inconnu.

1. On tire un unique ticket de valeur $X$. Écrire le modèle statistique associé : est-il dominé par une mesure $\sigma$-finie? Donner un estimateur qui s'apparenterait à un maximum de vraisemblance $\widehat{\theta}$ de $\theta$ (maximiser $P_\theta(\{X\})$), puis montrer que $\mathbb{P}(\widehat{\theta} = \theta) \geq 0,98$.

[4] On rappelle que sa densité est $\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x)$ sur $[0, \infty[$

2. On renumérote les tickets marqués $10\theta$ par $a_i\theta$, $1 \leq i \leq 980$, où les $a_i$ sont des réels connus tous distincts dans $[10; 10, 1]$. Donner le nouvel estimateur du maximum de vraisemblance $\tilde{\theta}$ et montrer que $\mathbb{P}(\tilde{\theta} < 10\theta) = 0, 02$.

**Exercice 15.8** (Régression logistique)**.** On observe des couples $(\mathbf{x}_i, y_i)$ où les $\mathbf{x}_i$ sont des variables explicatives (vecteurs ligne de dimension $d$) et les $y_i$ sont des variables valant 0 ou 1.

1. Avant toute chose, expliquer pourquoi une régression linéaire des $\mathbf{x}_i$ sur les $y_i$ n'aurait pas beaucoup de sens.
2. On suppose dorénavant que les $y_i$ sont des réalisations indépendantes de $Y_i \sim \text{Ber}(p(x_i))$ et on suppose que la fonction $p : \mathbb{R}^{1,d} \to ]0, 1[$ s'écrit sous forme *logistique* (« modèle logit ») :

$$p(\mathbf{x}) = \frac{e^{\mathbf{x}\theta}}{1 + e^{\mathbf{x}\theta}}.$$

   où $\theta \in \mathbb{R}^d$.

   i) Écrire ce modèle sous forme exponentielle avec pour paramètre $\theta$.
   ii) Écrire l'équation vérifiée par l'EMV de $\theta$.
   iii) Se convaincre qu'elle ne possède pas d'expression exacte.

3. Mêmes questions lorsque la fonction $p$ s'écrit sous forme « probit », $p(\mathbf{x}) = \Phi(\mathbf{x}\theta)$ avec $\Phi$ la fonction de répartition de $N(0, 1)$.

# 16 Entropie et information

Soient $X_i$ des variables aléatoires définies sur un même ensemble fini $\mathcal{X} = (a_1, \dots, a_k)$. On note $p_k = \mathbb{P}(X = a_k)$. On observe un échantillon de $n$ réalisations des $X_i$, disons $(x_1, \dots, x_n)$. Avant de s'intéresser à l'information que portent ces observations sur la loi $\mathbb{P}$, qui est le problème *statistique* auquel nous nous sommes intéressé, on peut se demander quelle est la quantité d'information portée par les $x_i$ *tout court*, sans avoir à faire d'hypothèses sur la loi de $X$. Par exemple, si toutes les observations sont identiques, l'échantillon transporte peu d'information – en tout cas, moins que si les observations sont un peu plus variées. La question posée est la suivante : si l'on voulait coder les observations de sorte que deux observations différentes soient codées de deux façons différentes, et de sorte qu'en moyenne le codage des informations soit le plus compressé possible, quel code utiliserait-on ?

## 16.1 La notion de code

Il faut raisonner en termes de bits d'information : on veut coder chaque élément observé par une suite de 0 et de 1.

Par exemple, supposons que $\mathcal{X}$ ne possède que deux éléments ; on peut coder $x_i = 0$ ou 1 et voir notre échantillon comme une suite binaire de longueur $n$. On aura donc besoin de au plus $n$ bits d'information pour encoder l'échantillon ; un « bit » est la donnée d'un 0 ou d'un 1.

Si maintenant $\mathcal{X}$ contient 4 éléments, on peut coder ses élements par des suites de deux bits : $00, 01, 10, 11$. Le codage d'une observation nécessitera toujours deux bits. Pour encoder notre échantillon, on aura besoin de $n$ fois deux bits d'information au plus, donc $2n$ bits. Par exemple, l'échantillon $(a, a, a, b, a, d, d, a, a, a)$ sera codé 00000001001111000000, donc 20 bits.

Plus généralement, si $\mathcal{X}$ contient $k$ éléments, on peut chacun les coder avec au plus $\log_2\lceil k \rceil$ bits et il faudra $n \log_2\lceil k \rceil$ pour encoder l'échantillon.

Dans ces exemples, chaque codage d'un échantillon nécessitait le même nombre de bits. Ce codage-là ne nécessite aucune information particulière sur les $x_i$. Mais maintenant, dans le cas où $\mathcal{X}$ comporte 4 éléments $\{a, b, c, d\}$, supposons que le premier élément soit beaucoup plus fréquent que les autres ; autrement dit,

| | a | b | c | d |
|---|---|---|---|---|
| $\mathbb{P}(X = ...)$ | 97% | 1% | 1% | 1% |

Dans ce cas, la plupart des observations n'auront que des $a$. Donc si j'assigne à l'observation $a$ un code *d'un seul bit*, disons 0, et que j'assigne aux trois autres des codes plus longs comme $10, 110, 111$, alors le coût moyen de codage d'une observation sera

$$1 \times 97\% + 2 \times 1\% + 3 \times 1\% + 3 \times 1\% = 1.05$$

Autrement dit, en moyenne, je n'aurai pas besoin de beaucoup plus d'un bit d'observation par observation, alors que le codage élémentaire ci-dessus en nécessitait exactement deux. Avec ce code, l'échantillon $(a, a, a, b, a, d, d, a, a, a)$ devient codé par 000100111111000, donc 15 bits.

**Définition 16.1.** Un *code binaire* de $\mathcal{X}$ assigne à chaque élément $a$ de $\mathcal{X}$ un mot binaire $w(a)$. Le code d'un élément ne doit pas être le début du code d'un autre élément.

Par exemple, je n'ai pas le droit de coder l'ensemble à quatre éléments avec le code $w(1) = 1, w(2) = 10, w(3) = 11, w(4) = 010$. En effet, dans ce cas je ne serai pas en mesure de discerner si 1010 veut dire $w(1)w(4)$ ou bien $w(2)w(2)$.

La longueur du $k$-ème élément $w(k)$ est notée $\ell_k$. Étant donné un code, le nombre moyen de bits d'information nécessaire pour coder une variable aléatoire $X$ sur $\mathcal{X}$ est donc

$$\mathbb{E}[\ell_X] = \sum_{i=1}^{k} \ell_i \mathbb{P}(X = a_i).$$

## 16.2 Le théorème de Shannon

Quel est la plus petite quantité moyenne d'informations nécessaire à coder une réalisation de $X$ ? La réponse est *l'entropie de la loi de $P$*.

**Définition 16.2** (Entropie)**.** Soit $X$ une variable aléatoire de loi $P$, possédant une densité $f$ par rapport à une mesure de référence. L'entropie $\mathrm{Ent}(P)$ est

$$-\int_{\mathcal{X}} f(x) \log_2 f(x) \nu(dx).$$

Lorsque $X$ est une variable discrète prenant la valeur $a_k$ avec probabilité $p_k$, l'entropie est donc égale à

$$-\sum_k p_k \log_2 p_k.$$

En théorie de l'information, on prend plutôt le logarithme en base 2 plutôt que le logarithme népérien. L'entropie ne diffère alors que d'une constante $\ln(2)$.

**Théorème 16.1** (Théorème de codage de Shannon)**.** *Tout codage vérifie* $\mathbb{E}[\ell_X] \geqslant \mathrm{Ent}(P)$.

*Il existe un codage appelé* codage entropique *vérifiant* $\ell^\star(a) = \lceil \log_2(p_k) \rceil$*, et donc* $\mathbb{E}[\ell^\star_X] \leqslant \mathrm{Ent}(P) + 1$.

En particulier, le codage optimal d'un échantillon iid de taille $n$ vérifie

$$\lim \frac{\mathbb{E}[\ell^\star_n]}{n} = \mathrm{Ent}(P).$$

L'entropie d'une loi est donc en le nombre moyen de bits d'information nécessaire à son meilleur codage, et le meilleur codage en question est le *codage entropique* qui assigne à chaque $a \in \mathcal{X}$ un code de longueur proche de $\log_2(\mathbb{P}(X = a))$. On a donc compressé au mieux la variable $X$.

*Démonstration.* J'écrirai plus tard la preuve via le lemme de Kraft.

□

## 16.3 L'entropie relative

Soit $P$ une loi de probabilité. Le codage optimal de $P$ est $\log_2 P$. On pourrait aussi utiliser le codage d'une autre loi de probabilité, disons $Q$ : la quantité

$$\sum_{k=1}^{n} p_k \log_2 \frac{p_k}{q_k}$$

peut aussi se voir comme

$$E[\log_2 q_X] - E[\ell^\star_X].$$

Il s'agit donc bien d'une « redondance d'information » : le terme de droite est la quantité minimale d'information dont on a besoin pour coder $X$ en moyenne (via le codage entropique), et le terme de gauche est la quantité d'information dont on use pour coder $X$ en moyenne par le code $S$. À une constante près[1], la divergence de Kullback-Leibler

$$\mathrm{d}_{\mathrm{KL}}(P \mid Q) = -\int P \ln Q - \left(-\int P \ln P\right)$$

indique donc à quel point il est optimal (ou pas) d'encoder les réalisations de la loi $P$ grâce au « code » $\log_2 Q$ : $\mathrm{d}_{\mathrm{KL}}$ quantifie la *redondance* de $Q$ par rapport à $P$. Puisque le codage optimal est celui qui code $x$ par $\log_2 P(x)$, le codage utilisant $\log_2 Q(x)$ n'a pas structuré au mieux l'information disponible dans $P$ est pourrait être allégé.

## 16.4 Retour sur l'information de Fisher

Dans un modèle exponentiel, nous avons vu que l'information de Fisher est la courbure de l'entropie (Proposition 15.1). Comme dans ce cadre, l'entropie est concave, la courbure en $\theta$ est un indicateur de à quel point l'entropie forme un « pic » autour de $\theta$.

[1] À savoir $\ln(2)$.

# 17 Principe d'entropie maximale

Ce chapitre fait le pont entre la physique statistique et la statistique, et montre essentiellement qu'il s'agit de deux points de vue sur la même théorie.

Dans une expérience statistique, on observe des échantillons $x_1, \dots, x_n$ qui sont iid selon une certaine $P$ loi qui est inconnue, et qu'on cherche à estimer. Si l'on n'a aucune information sur cette loi et qu'on ne veut pas faire d'hypothèses, il faudra estimer sa densité à l'aide d'outils non-paramétriques que nous verrons plus tard ; mais dans de nombreux cas, on préfère *supposer* que la loi appartient à une certaine classe, typiquement la classe des modèles exponentiels associés à un certain moment $T : \mathcal{X} \to \mathbb{R}$.

*Pourquoi avoir choisi les modèles exponentiels ?*

## 17.1 Hasard et information

En pratique, les observations dont on dispose viennent ne viennent jamais de phénomènes dont on ne connaît rien. Il y a toujours un savoir implicite qui impose des contraintes sur $P$. Par exemple, on peut savoir que $P$ est supportée sur un ensemble compact ; ou encore, qu'elle a une variance finie. Souvent, ces contraintes s'écrivent sous la forme de moyennes : on peut savoir que la moyenne d'une certaine statistique $T(X)$ vaut exactement $c$, c'est-à-dire

$$E_{X\sim P}[T(X)] = c.$$

Typiquement, si l'on cherche une loi centrée réduite, on chorche $P$ parmi les lois qui vérifient $E[X] = 0$ et $E[X^2] = 1$.

Il est donc nécessaire de restreindre l'ensemble dans lequel on cherche $P$, et ne considérer que les lois vérifiant ces contraintes (les « lois admissibles »). Or, les contraintes comme celles ci-dessus sont vérifiées par énormément de lois. Laquelle choisir ? Si la seule information sur $P$ est la contrainte $E[T(X)] = c$, on veut choisir parmi celle qui est « la plus aléatoire possible » : autrement dit, **la loi d'entropie maximale**. Or, le théorème de Boltzmann-Gibbs doit que les lois d'entropie maximale vérifiant des contraintes de moyennes sont *exeactement* les lois exponentielles.

**Théorème 17.1** (Principe d'entropie maximale de Boltzmann-Gibbs)**.** *Soit $T : \mathbb{R}^d \to \mathbb{R}^p$. Les densités de probabilité $f$ sur $\mathbb{R}^n$ qui maximisent l'entropie $-\int f(x) \ln f(x) \nu(dx)$ sous la containte $E_{X\sim f}[T(X)] = c$, si elles existent, sont exactement de la forme $e^{\langle \theta, T(x)\rangle}/Z(\theta)$.*

Le principe d'entropie de Boltzmann-Gibbs s'applique aussi aux lois empiriques et permet de retrouver exactement l'EMV. En effet, supposons que dans une expérience statistique, on n'aie pas accès à l'échantillon $x_1, \dots, x_n$, mais seulement à des « moyennes d'observables » :

$$\frac{1}{n}\sum_{i=1}^{n} T(x_i) = \bar{T}_n.$$

Il est alors naturel de chercher $P$ parmi toutes les lois qui vérifient la contrainte $E_{X\sim P}[T(X)] = \bar{T}_n$, et *rien de plus*. Le principe de Boltzmann-Gibbs dit que les lois qui maximisent l'entropie sous cette contrainte sont exactement les lois exponentielles associées à $T$, et pour peu que le modèle soit identifiable (cf Théorème 13.1), il n'y a qu'un seul paramètre qui garantit que $E_\theta[T(X)] = \bar{T}_n$. Ce paramètre est évidemment $\hat{\theta}_{\text{emv}}$.

**Exemple 17.1** (Loi gaussienne)**.** Quelle est la densité $f$ qui maximise l'entropie, sous la contrainte d'être centrée et réduite ? Ici, cette contrainte s'écrit $E[X] = 0$ et $E[X^2] = 1$, donc le moment associé est $T(x) = (x, x^2)^\top$. Le théorème dit que cette loi s'écrit sous la forme $e^{-\alpha x - \beta x^2}/Z(\alpha, \beta)$, où $\alpha, \beta$ sont réels. En réalité, $\beta$ doit être positif sinon ce n'est pas une loi de probabilité. De plus, il est facile de voir que si $\alpha$ n'est pas nul, alors l'espérance n'est pas nulle. La loi d'entropie maximale est donc proportionnelle à $e^{-\beta x^2}$ : il s'agit évidemment d'une loi gaussienne, et le seul paramètre qui garantit que la variance est 1 est donné par $\beta = 1/2$.

## 17.2 Démonstration

La démonstration générale de Théorème 17.1 nécessite des outils de calcul des variations, puisqu'il s'agit d'un problème s'optimisation en dimension infinie. Ce n'est pas difficile formellement, mais garantir l'existence d'un maximiseur peut s'avérer technique[1]. En revanche, l'esprit de la preuve est très simple : on écrit le lagrangien du problème contraint. Lorsque l'espace d'états $\mathcal{X}$ est fini, c'est très simple. On supposera donc que $\mathcal{X} = \{1, \dots, n\}$, et on cherche une loi de probabilité sur $\mathcal{X}$, disons $p = (p_1, \dots, p_n)$, qui vérifie la contrainte de moments $m(p) = 0$, où

$$m(p) = \sum_{i=1}^{n} p_i T(i) - c$$

et qui maximise l'entropie $H(p)$, où

$$H(p) = -\sum_{i=1}^{n} p_i \ln p_i.$$

Le fait que $p$ soit une loi de probabilité se traduit en pratique par des contraintes supplémentaires, à savoir $s(p) = 0$ où

$$-1 + \sum_{i=1}^{n} p_i = 0.$$

Il s'agit d'un problème d'optimisation sous contraintes dans[2] $[0,1]^n$ :

$$\begin{cases} \max_{p\in[0,1]^n} H(p) \\ m(p) = 0 \\ s(p) = 0 \end{cases} \tag{17.1}$$

Les deux contraintes sont linéaires et leur intersection n'est évidemment pas vide ; de plus, la fonction $H$ est concave. En effet, sa matrice hessienne au point $p$ est égale à $\text{diag}(-1/p)$, qui est bien définie négative. Le problème Équation 17.1 possède donc une solution. Pour la trouver, on utilise les outils classiques

[1] Après discussion avec mes collègues, il y a consensus sur la nécessité d'utiliser un critère de compacité L1 de type Dunford-Pettis.

[2] Si l'un des $p_i$ est nul, $H(p) = +\infty$, donc on peut se restreindre aux $p_i > 0$.

de l'optimisation sous contraintes. Par facilité, j'exclurai les cas où ce maximum est atteint au bord du domaine.

Le Lagrangien de ce problème s'écrit

$$\mathscr{L}(p, \lambda, \mu) = H(p) + \lambda m(p) + \mu s(p).$$

Les conditions du premier ordre (conditions KKT) pour l'existence d'un minimiseur local s'écrivent alors $\nabla\mathscr{L} = 0$, soit $\nabla_p\mathscr{L} = 0, \nabla_\lambda\mathscr{L} = 0$ et $\nabla_\mu\mathscr{L} = 0$. Le première identité se traduit par les équations suivantes :

$$\partial_{p_i}\mathscr{L} = -(\ln(p_i) + 1) + \lambda T(i) + \mu = 0,$$

soit $p_i = e^{-\lambda T(i)+\mu-1}$ pour un certain $\lambda$ et un certain $\mu$. Comme les $p_i$ somment à 1, le nombre $\mu$ est immédiatement déterminé par l'équation $e^{\mu-1} = \sum_{i=1}^n e^{-\lambda T(i)}$. Les points critiques du système sont donc exactement

$$\left(\frac{e^{-\lambda T(i)}}{Z(\lambda)}, \dots, \frac{e^{-\lambda T(i)}}{Z(\lambda)}\right)$$

où $Z(\lambda) = \sum_{i=1}^n e^{-\lambda T(i)}$. **Autrement dit, s'il y a une solution au problème, il s'agit forcément d'une loi dans le modèle exponentiel associé à** $T$. Maintenant, la contrainte doit être réalisée, c'est-à-dire que l'on doit trouver un $\lambda$ qui vérifie

$$E_\lambda[T(X)] = c.$$

L'existence d'un tel $\lambda$ n'est pas forcément vérifiée[3]. Pour qu'il y ait une solution et une seule, on peut par exemple faire des hypothèses sur $T$ qui garantissent que le modèle est identifiable, de sorte que $E_\lambda[T(X)]$ est un difféormorphisme.

[3]Penser à la contrainte absurde $E_\lambda[X^2] = -1$.

# Problèmes

**Exercice 17.1** (Test du signe)**.** On observe $n$ couples aléatoires $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ indépendants mais pas nécessairement de même loi. On suppose de plus que les variables $X_i$ et $Y_i$ sont indépendantes et qu'elles ont une loi diffuse pour tout $i \in \{1, \dots, n\}$. On considère le test des hypothèses

$$\begin{array}{ll} H_0 : & X_i = Y_i \text{ en loi pour tout } i, \\ H_1 : & \text{il existe } i \neq j \text{ tels que } X_i \neq Y_i \text{ en loi.} \end{array}$$

1. Montrer que $P(X_i = Y_i) = 0$ et en déduire que sous $H_0$, on a $P(X_i > Y_i) = \frac{1}{2}$.
2. On pose $N = \sum_{i=1}^{n} \mathbf{1}_{X_i > Y_i}$. Quelle est la loi de $N$ sous $H_0$ ?
3. En déduire que le test défini par la région de rejet

$$\left\{ \left| N - \frac{n}{2} \right| \geq c \right\}$$

permet de construire un test de niveau inférieur à $\alpha \in ]0, 1[$ de $H_0$ contre $H_1$ pour un choix $c = c(\alpha) > 0$ que l'on précisera. Parmi tous les choix possibles de $c(\alpha)$, lequel préférer ?

4. Les moyennes générales de la première et de la deuxième année de cinquième de 12 redoublants ont été relevées:

| Élève | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Année 1 | 12.0 | 9.5 | 13.0 | 10.0 | 8.5 | 11.0 | 7.8 | 14.0 | 5.0 | 12.0 | 12.0 | 8.6 |
| Année 2 | 6.1 | 14.0 | 7.3 | 7.3 | 13.0 | 17.0 | 14.0 | 9.2 | 12.0 | 14.0 | 8.8 | 8.8 |

Le redoublement a-t-il une influence sur la moyenne générale ? [4]

**Exercice 17.2** (Test de gaussianté de Jarque-Bera)**.** Soit $(X_1, \dots, X_n)$ un $n$-échantillon de loi inconnue $F$ ayant au moins un moment d'ordre 4 et de moyenne nulle et de variance non nulle.

1. On pose, pour $k = 1, \dots, 4$,

$$T_n^{(k)} = \frac{\frac{1}{n} \sum_{i=1}^{n} X_i^k}{\left( \frac{1}{n} \sum_{i=1}^{n} X_i^2 \right)^{k/2}}.$$

Montrer que si $F$ est une distribution gaussienne, on a la convergence en loi suivante :

$$\frac{n}{15} \left( T_n^{(3)} \right)^2 + \frac{n}{96} \left( T_n^{(4)} - 3 \right)^2 \to \chi_2^2$$

2. En déduire un test de l'hypothèse nulle $H_0$ : « $F$ est gaussienne » contre l'alternative $H_1$ : « $F$ n'est pas gaussienne ».
3. Le test est-il convergent ?

[4] Le quantile d'ordre 0.975 d'une $\mathscr{B}(12, 0.5)$ est 9.

**Exercice 17.3** (Test exact de Fisher)**.** On reprend l'exercice Exercice 8.10, mais cette fois la table de contingence des observations est la suivante :

| | riche | pauvre |
|---|---|---|
| **heureux** | 1 | 9 |
| **triste** | 11 | 3 |

On cherche à tester si l'argent et le bonheur sont deux dimensions indépendantes (hypothèse nulle).

1. Un test du $\chi_2$ d'indépendance est-il adapté cette fois ?
2. On suppose que le total de chaque ligne et de chaque colonne est fixé. Montrer que sous l'hypothèse nulle, la vraisemblance d'une table de contingence de la forme

$$t = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

est égale à

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{24}{a+c}}.$$

On a supposé que $a+b = 10, c+d = 14, a+c = 11, b+d = 12$. Pour la table ci-dessus, on trouve $p = 0.001346$.

3. La notion de quantile a peu de sens pour une loi comme ci-dessus[5]. On remplace donc cette notion par la suivante : si $p(t)$ est la probabilité, sous l'hypothèse nulle, d'observer une table $t$, alors on ordonne toutes les tables possible $t_1, \dots, t_2, \dots$ par probabilité croissante. On pose $n_\alpha = \sup\{k : p(t_1) + \cdots + p(t_k) < \alpha\}$. Montrer que le test dont la région de rejet est $\{t_1, \dots, t_{n_\alpha}\}$ est un test de niveau de confiance au moins $1 - \alpha$ de l'hypothèse nulle.

[5]Loi hypergéométrique.

# 18 Estimation de densité

Soient $X_1, \dots, X_n$ des variables iid, de fonction de répartition $F$. Le problème de l'estimation de densité est celui d'estimer la densité ou la fonction de répartition de $F$ à partir de réalisations des $X_i$ : c'est un exemple typique d'estimation non-paramétrique. Dans toute la suite, on se placera dans le cas où la fonction de répartition est *continue*.

## 18.1 La répartition empirique

L'objet central est la fonction de répartition empirique,

$$F_n(t) = \frac{1}{n}\#\{i : X_i \leqslant t\}.$$

La loi des grands nombres montre immédiatement que, $\mathbb{P}$-presque sûrement, $F_n(t)$ converge vers $\mathbb{P}(X_i \leqslant t) = F(t)$. On peut étendre ce résultat simultanément à une quantité dénombrable de $t$ (par exemple, $\mathbb{Q}$) mais pas à tous. De plus, ce résultat ne dit pas si la *fonction* $F_n$ est proche de la fonction $F$, au sens de la norme uniforme par exemple. Le théorème suivant, parfois appelé *théorème fondamental de l'estimation*, confirme que c'est le cas au sens de la norme uniforme[1].

**Théorème 18.1** (Théorème de Glivenko-Cantelli)**.** *$\mathbb{P}$-presque sûrement, lorsque $n \to \infty$ on a $|F_n - F|_\infty \to 0$.*

Ce théorème dit essentiellement que $F_n$ est un estimateur « convergent au sens de la norme uniforme » de $F$. On pourra également utiliser ce théorème pour faire des tests : typiquement, si $|F_n - F|$ n'est pas suffisamment proche de zéro, on rejettera l'hypothèse selon laquelle les $X_i$ ont $F$ pour fonction de répartition. Le critère exact est appelé *test de Kolmogorov-Smirnov* et sera vu dans la section suivante.

### Calculabilité et loi

Soit $(X_{(1)}, \dots, X_{(n)})$ l'échantillon trié en ordre croissant. Par convention, on pose $X_{(0)} = -\infty$. Alors, la quantité $|F_n - F|_\infty$ est aisément calculable grâce à la représentation suivante :

$$|F_n - F|_\infty = \sup_{i \in \{0,\dots,n-1\}} \left| \frac{i}{n} - F(X_{(i)}) \right| \vee \left| \frac{i}{n} - F(X_{(i+1)}) \right|. \tag{18.1}$$

[1] On rappelle que $|g|_\infty = \sup_{x \in \mathbb{R}} |g(x)|$.

*Démonstration.* La fonction $F$ est croissante, et la fonction $\hat{F}_n$ est constante par morceaux sur tous les intervalles $[X_{(i)}, X_{(i+1)}[$. En effet, à chaque $X_{(i)}$, elle saute de la valeur à gauche $(i-1)/n$ a la valeur à droite $i/n$. Ainsi, le maximum de $F - \hat{F}_n$ sur l'intervalle $[X_{(i)}, X_{(i+1)}[$ est forcément atteint à une des deux bornes, et vaut donc soit $|i/n - F(X_{(i)})|$, soit $|i/n - F(X_{(i+1)})|$, selon celui qui est le plus grand. Le supremum de $|F - \hat{F}_n|$ sur $\mathbb{R}$ étant aussi le supremum des supremums sur tous ces intervalles, la représentation ci-dessus est vraie.

□

**Lemme 18.1.** *Si $F$ est continue, $|F_n - F|_\infty$ a la même loi que*

$$\sup_{i\in\{0,\dots,n-1\}} \left|\frac{i}{n} - U_{(i)}\right| \vee \left|\frac{i}{n} - U_{(i+1)}\right|$$

*où les $U_{(i)}$ sont des lois uniformes sur $[0,1]$, indépendantes, et triées dans l'ordre croissant.*

*Démonstration.* Lorsque $X$ est une variable aléatoire dont la fonction de répartition $F$ est continue et strictement croissante, $F(X)$ suit une loi $\mathscr{U}[0,1]$. En effet, si $t \in [0,1]$, alors $\mathbb{P}(F(X) < t)$ est égal à $\mathbb{P}(X \leqslant F^{-1}(t))$, c'est-à-dire $F(F^{-1}(t)) = t$. Lorsque $F$ est seulement continue, la même démonstration est vraie, mais il faut remplacer l'inverse $F^{-1}(t)$ par la « transformation quantile », à savoir $F^{\leftarrow}(t) = \inf\{x : F(x) \geqslant t\}$. Les $F(X_i)$ sont donc des variables iid de loi $\mathscr{U}[0,1]$, ce qui conclut la démonstration compte tenu de Équation 18.1.

□

En particulier, la loi de $|F_n - F|_\infty$ ne dépend pas de $F$ : on dit que cette statistique est *libre*.

## Démonstration du théorème de Glivenko-Cantelli

Notons $q_j$ le $j$-ème quantile d'ordre $N$ de $F$ (on choisira l'entier $N$ plus tard). Soit $x$ entre $q_j$ et $q_{j+1}$. Par croissance, $F_n(x)$ est entre $F_n(q_j)$ et $F_n(q_{j+1})$, et $F(x)$ est entre $j/N$ et $(j+1)/N$. Ainsi, $F_n(x) - F(x)$ est plus grand que

$$F_n(q_j) - \frac{j}{N} - \frac{1}{N}$$

et plus petit que

$$F_n(q_{j+1}) - \frac{j}{n} = F_n(q_{j+1}) - \frac{j+1}{N} + \frac{1}{N}.$$

Quoi qu'il arrive, $|F_n(x) - F(x)|$ est plus petit que le plus grand des $|F_n(q_j) - j/N|$ augmenté de $1/N$, donc $|F_n - F|$ aussi. Pour n'importe quel $t > 0$, nous pouvons utiliser la borne de l'union afin de borner $\mathbb{P}(|F_n - F|_\infty > t)$ par

$$\sum_{j=0}^{N} \mathbb{P}\left(\frac{1}{N} + |F_n(q_j) - j/N| > t\right). \tag{18.2}$$

Or, $nF_n(q_j)$ suit une loi $\mathrm{Bin}(n, j/N)$, donc si[2] $t - 1/N > 0$ on peut utiliser l'inégalité de Hoeffding :

$$\mathbb{P}\left(\frac{1}{N} + |F_n(q_j) - j/N| > t\right) \leqslant 2\exp\left\{-2n\left(t - \frac{1}{N}\right)^2\right\}.$$

[2] Que se passe-t-il si $t \leqslant 1/N$ ?

En choisissant par exemple $N = \lceil 2/N \rceil$, le terme entre parenthèses est plus grand que $t/2$, et la borne devient elle-même plus petite que $2e^{-nt^2/4}$. Le terme Équation 18.2 est alors plus petit que $N2e^{-nt^2/4}$, c'est-à-dire

$$\mathbb{P}(|F_n - F|_\infty > t) \leqslant \frac{2}{t} e^{-nt^2/4}. \tag{18.3}$$

Si l'on choisit une suite $t_n$ qui tend vers 0, et telle que $\sum_n e^{-nt_n^2/4}/t_n < \infty$, alors le lemme de Borel-Cantelli permet de conclure : presque sûrement, à partir d'un certain rang, on a $|F - F_n|_\infty \leqslant t_n$, et donc $|F_n - F|_\infty \to 0$.

## 18.2 Inégalité DKW

Le théorème de Glivenko-Cantelli possède une version beaucoup plus puissante car elle est entièrement quantitative, appelée *inégalité DKW*.

**Théorème 18.2** (Dvoretzky-Kiefer-Wolfowitz)**.** *Dans le même contexte, pour tout $t > 0$ on a*

$$\mathbb{P}(|F_n - F|_\infty > t) \leqslant 2e^{-2nt^2}. \tag{18.4}$$

Il faut comparer ce résultat avec Équation 18.3, dans lequel la borne est effectivement décroissante en $nt^2$, mais polynomialement. L'inégalité DKW donne une décroissance *exponentielle* en $nt^2$.

# 19 Test de Kolmogorov-Smirnov

On souhaite maintenant *tester* la distribution d'un échantillon $(x_1, \dots, x_n)$, c'est-à-dire tester l'hypothèse nulle : « les $x_i$ sont des réalisations d'une variable aléatoire dont la fonction de répartition est $F$ », où $F$ est une fonction de répartition fixée. Le théorème de Glivenko-Cantelli dit que $|F_n - F|_\infty$, sous l'hypothèse nulle, tend vers zéro. On rejettera donc l'hypothèse nulle si $|F_n - F|$ est trop grand ; mais à quel seuil ? La démonstration du théorème et l'inégalité DKW disent que la bonne échelle est $\sqrt{n}$ : en effet, $\mathbb{P}(|F_n - F|_\infty > \sqrt{\alpha/n}) = O(1/t^2)$. Un test dont la région de rejet est de la forme

$$\left\{ |F_n - F|_\infty > \frac{10}{\sqrt{n}} \right\}$$

aura un niveau de confiance d'ordre $1-\alpha$, ce qui fournit déjà un test non-asymptotique. L'utilisation de l'inégalité DKW permet d'avoir une région de rejet encore plus grande.

En réalité, si l'on suppose seulement que $F$ est continue, il se trouve que $\sqrt{n}|F_n - F|_\infty$ converge en loi vers une loi connue dont on connaît les quantiles.

**Théorème 19.1** (Kolmogorov-Smirnov)**.** *$\sqrt{n}|F_n - F|_\infty$ converge en loi vers $|B|_\infty$, où $(B_t)_{t\in[0,1]}$ est un mouvement Brownien standard. La loi de cette variable aléatoire positive est appelée* loi de Kolmogorov-Smirnov*, et sa fonction de répartition $\mathbb{P}(|B|_\infty \leqslant x)$ est donnée par*

$$1 - 2\sum_{k=0}^{\infty} (-1)^k e^{-2x^2(k+1)^2}.$$

Ce théorème permet de construire des tests aymptotiques de niveau exactement $1-\alpha$.

**Théorème 19.2.** *Si $X_1, \dots, X_n$ suivent une loi de fonction de répartition $G \neq F$, alors $\sqrt{n}|F_n - F|_\infty \to \infty$ presque sûrement.*

# Exercices

# Et après ?

## Statistiques Bayésiennes

## Séries temporelles

## Statistiques en grande dimension

$n, d \to \infty$ mais $p$ n'est pas trop grand : matrices aléatoires, PCA, compressed sensing, clustering, SVM

Modèles parcimonieux : ondelettes, LASSO, régression ridge

## Statistiques non-paramétriques

Histogrammes, kernel methods, fenêtres glissantes, malédiction de la dimension, VC dimension, complexité

## Machine learning

Input-output et apprentissage supervisé : classification, régressions logistiques, arbres de décision, online learning, reinforcement learning

## Deep learning et réseaux de neurones

# Références

- Statistics done Wrong
- The earth is round (p<.05)
- Statistiques mathématiques en action, pour ceux qui vont passer l'agrégation.
- Introduction à l'économétrie de Brigitte Dormont est un excellent livre, écrit en français, sur les modèles linéaires.
- En anglais, la référence sur les modèles linéaires est Econometric analysis de Greene.
- Méthodes statistiques de Philippe Tassi est un bon livre général.
- All of statistics de Larry Wasserman est un ouvrage de référence.
- Le cours de Stéphane Mallat au Collège de France qui est plus général, mais qui reprend tous les concepts.
- L'article original de Ronald Fisher de 1922 (et pas 1935 comme j'ai dit en cours), qui pose *toutes* les bases de la statistique moderne.
- Computer age statistical inference de Bradley Efron et Trevor Hastie n'est pas un livre de mathématiques, mais c'est le meilleur livre qui présente les idées et les algorithmes des statistiques avec un point de vue moderne.
- Elements of information theory, une référence sur la théorie de l'information.

# 20 Algèbre linéaire

## 20.1 Multiplication matricielle

La pratique des régressions linéaires nécessite une certaine familiarité avec la multiplication des matrices. On rappelle que si $A$ est une matrice à $\ell$ lignes et $m$ colonnes, et que $B$ est une matrice à $m$ lignes et $n$ colonnes, alors il est possible de les multiplier entre elles. Il en résulte une matrice $AB$ avec $\ell$ lignes et $n$ colonnes, dont le terme $i, j$ est égal à

$$\sum_{k=1}^{m} A_{i,k} B_{k,j}.$$

Ce terme peut aussi être vu comme $\langle A_{i,\cdot}, B_{\cdot,j} \rangle$, le produit scalaire entre la $i$-ème ligne de $A$ et la $j$-ème colonne de $B$.

De façon générale, le produit scalaire entre deux vecteurs de même taille, $\langle x, y \rangle$, est donc égal à la multiplication matricielle entre le vecteur ligne $x^\top$ et le vecteur colonne $y$.

Il est aussi possible de multiplier un vecteur ligne $x$ de taille $n$ et un vecteur colonne $y^\top$ de taille $m$, mais ici on n'a plus besoin que $n$ et $m$ soient égaux. Il en résulte une matrice de taille $n \times m$,

$$xy^\top = [x_i y_j]_{\substack{i=1,\dots,n \\ j=1,\dots,m}}.$$

Si, comme tout à l'heure, $A$ est une matrice $\ell, n$ et $B$ une matrice $m, n$, notons $a_i$ les *colonnes* de $A$ (vecteurs colonnes) et $b_i$ les *lignes* de $B$ (vecteurs lignes). Alors, on peut écrire

$$AB = \sum_{i=1}^{m} a_i b_i.$$

En particulier, pour n'importe quelle matrice $X$ de taille $n, d$ dont les lignes sont $\mathbf{x}_i$ (et donc, les colonnes de $X^\top$ sont les $\mathbf{x}_i^\top$), alors on peut écrire

$$X^\top X = \sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{x}_i.$$

## 20.2 Le théorème spectral

Grâce aux manipulations ci-dessus, le théorème de décomposition en vecteurs propres prend une forme légèrement différente. Ce théorème dit habituellement que toute matrice $M$ symétrique réelle peut s'écrire $UDU^\top$, avec $U$ la matrice de passage dans la base des vecteurs propres et $D = \mathrm{diag}(\lambda_i)$ la matrice diagonale des valeurs propres. C'est donc la même chose que l'énoncé suivant.

**Théorème 20.1.** *Soit $M$ une matrice symétrique réelle. Il existe une base orthonormale de vecteurs $u_1, \dots, u_n$ et des nombres réels $\lambda_1, \dots, \lambda_n$ tels que*

$$M = \sum_{i=1}^{n} \lambda_i u_i u_i^\top.$$

## 20.3 Projections orthogonales

Soit $v$ un vecteur non nul de $\mathbb{R}^n$. L'espace vectoriel engendré par $v$ est l'ensemble $\mathscr{V} = \{tv : t \in \mathbb{R}\}$, et son orthogonal est l'hyperplan $\mathscr{V}^\perp = \{x : \langle x, v\rangle = 0\}$. Les résultats élémentaires d'algèbre linéaire disent que tout vecteur $x$ se décompose *de façon unique* sous la forme

$$x = y + z$$

avec $y$ dans $\mathscr{V}$ et $z$ dans $\mathscr{V}^\top$. En particulier, il existe un $t$ tel que $y = tv$.

Considérons maintenant la matrice

$$P = \frac{1}{|v|^2} vv^\top \in \mathscr{M}_{n,n}.$$

Appliquons cette matrice à $x$. Par linéarité, $Px = Py + Pz$. Calculons ces deux termes.

1. $Pz = |v|^{-2} vv^\top z = |v|^{-2} v\langle v, z\rangle$. Comme $z$ est orthogonal à $v$, cela vaut 0.
2. $Py = tPv$. Par définition de $P$, ceci est donc égal à $t|v|^{-2} vv^\top v = t|v|^{-2} v|v|^2 = tv$, c'est-à-dire $y$.

Nous avons montré plusieurs choses. D'abord, l'application qui à $x$ associe $y$ est effectivement linéaire, et une de ses matrices est $P$. On dit que $P$ est la matrice de projection sur $\mathscr{V}$. De même, comme $(I - P)x = y + z - y = z$, la matrice $I - P$ est la matrice de projection sur $\mathscr{V}^\perp$.

Le cas d'un sous-espace vectoriel généré par *plusieurs* vecteurs $v_1, \dots, v_d$ linéairement indépendants se traite de la même façon. Soit $V = [v_1, \dots, v_d]$ la matrice $n \times d$ dont les colonnes sont les $v_i$. Tout à l'heure, $|v|^{-2}$ aurait pu s'écrire $(v^\top v)^{-1}$. L'analogue avec $V$ est donc naturellement $(V^\top V)^{-1}$, donnant naissance au théorème suivant.

**Théorème 20.2.** *Soient $v_1, \dots, v_d$ des vecteurs non-colinéaires de $\mathbb{R}^n$, et soit $V = [v_1, \dots, v_d]$ la matrice $n \times d$ dont les colonnes sont les $v_i$. La matrice de taille $n \times n$*

$$P_V = V(V^\top V)^{-1} V^\top$$

*est la matrice de projection orthogonale sur le sous-espace $\mathscr{V}$ engendré par les $v_i$. De plus, la matrice $I - P_V$ est la matrice de projection orthogonale sur le sous-espace $\mathscr{V}^\perp$.*

*Démonstration.* Si $x = y + z$ est la décomposition de $x$ en somme d'un élément $y \in \mathscr{V}$ et d'un élément $z \in \mathscr{V}^\perp$, alors $Px = Py + Pz$ et

$$Pz = V(V^\top V)^{-1} V^\top z.$$

Or, les $d$ lignes de $V^\top z$ sont les produits scalaires $\langle v_i, z\rangle$, qui sont tous nuls car $z$ est orthogonal à tous les $v_i$. Ainsi, $Pz = 0$.

D'autre part, comme $y$ est dans l'espace engendré par les $v_i$, il s'écrit sous la forme $t_1 v_1 + ... + t_d v_d$. Cela peut se récrire en disant que $y = Vt$, où $t$ est le vecteur colonne des $t_i$. Mais alors,

$$Py = V(V^\top V)^{-1} V^\top V t = Vt = y.$$

On conclut comme dans le cas $d = 1$ exposé ci-dessus. Il reste cependant un point de détail : nous devons nous assurer que $V^\top V$ est effectivement inversible ! C'est le cas, je le jure.

□

## 20.4 Matrices positives

Une matrice symétrique réelle est *positive* lorsque toutes ses valeurs propres sont positives ou nulles, et *définie positive* lorsqu'elles sont toutes strictement positives.

**Proposition 20.1.** *Une matrice $A$ est positive si et seulement si $\langle x, Ax \rangle$ est un nombre positif ou nul pour tout $x$.*

*Démonstration.* Décomposer $x$ dans une base orthonormale $u_1, ..., u_n$ de vecteurs propres de $A$ afin d'écrire $\langle x, Ax \rangle$ sous la forme $\sum_{i=1}^n \lambda_i \langle x, u_i \rangle^2$. L'équivalence est alors évidente.

□

**Définition 20.1.** On dit que $A$ est dominée par $B$ lorsque $B - A$ est une matrice positive. On note cela $A \preceq B$.

La proposition précédente montre immédiatement que c'est équivalent à ce que $\langle x, Ax \rangle \leqslant \langle x, Bx \rangle$ pour tout $x$.

# 21 50 nuances de TCL

## 21.1 La version classique

Soit $(X_i)$ une suite de variables aléatoires iid possédant une moyenne $\mu$ et une variance $\sigma^2$. On note $\bar{X}_n$ leur moyenne empirique,

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{21.1}$$

Sous les hypothèses sur les $X_i$, il est clair que $\mathbb{E}[\bar{X}_n] = \mu$, et que $\mathrm{Var}(\bar{X}_n) = \sigma^2/n$.

**Théorème 21.1.** *La variable aléatoire*

$$\frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}}$$

*converge en loi vers* $N(0,1)$.

*Démonstration.* Si $\varphi$ est la transformée de Fourier commune de la loi des $Y_i = (X_i - \mu)/\sigma$ et $\psi$ celle de Équation 21.1, alors

$$\psi(t) = \varphi(t/\sqrt{n})^n.$$

Comme $\varphi(x) \sim 1 - x^2/2 + o(x^2)$ par un développement de Taylor près de zéro, on voit que lorsque $n \to \infty$, alors $\psi(t) = (1 - t^2/2n + o(1/n))^n$ et ceci tend vers $e^{-t^2/2}$, qui est bien la transformée de Fourier de $N(0,1)$.

□

## 21.2 La version de Lindeberg-Lévy

On supposera maintenant les $X_i$ *indépendantes* (mais pas forcément de même loi). On pose $\bar{\mu} = \mathbb{E}[\bar{X}_n]$ et $s_n = \mathrm{Var}(\bar{X}_n)$, c'est-à-dire

$$\bar{\mu}_n = \frac{\sum_{i=1}^{n} \mu_i}{n}$$

$$\varsigma_n^2 = \frac{\sum_{i=1}^{n} \sigma_i^2}{n^2}$$

où $\mu_i = \mathbb{E}[X_i]$ et $\sigma_i^2 = \mathrm{Var}(X_i)$.

**Théorème 21.2.** *Si ces variables vérifient la* condition de Lindeberg, *à savoir que pour tout $\delta > 0$,*

$$\frac{1}{\varsigma_n^2}\sum_{i=1}^{n}\mathbb{E}[|X_i-\mu_i|^2\mathbf{1}_{|X_i-\mu_i|>\delta\varsigma_n}]\to 0 \tag{21.2}$$

*alors la variable aléatoire*

$$\frac{\bar{X}_n-\bar{\mu}_n}{\varsigma_n}$$

*converge en loi vers $N(0,1)$.*

## 21.3 Le théorème de Mann-Wald[1]

C'est un cas particulier du précédent.

Soient $(x_i)$ une suite de nombres réels, pas forcément aléatoires, et soient $\varepsilon_i$ des variables aléatoires iid de variance $\sigma^2$ et vérifiant $\mathbb{E}[|\varepsilon_i|^4] < c^2$ pour une certaine constante $c^2$. La moyenne pondérée

$$\frac{1}{n}\sum_{i=1}^{n}x_i\varepsilon_i$$

est clairement une variable aléatoire centrée, et sa variance est égale à

$$\sigma^2\frac{\sum_{i=1}^{n}x_i^2}{n}=\frac{\sigma^2 s_n^2}{n}.$$

Peut-on dire que la moyenne réduite

$$\sqrt{n}\frac{\sum_{i=1}^{n}x_i\varepsilon_i}{\sigma s_n} \tag{21.3}$$

converge en loi vers une $N(0,1)$ ? La réponse est *oui* en général : cependant, en toute rigueur, on faire une hypothèse sur les $x_i$. On demande à ce que la variance $s_n^2$ ne soit pas dominée par un petit nombre de $x_i$:

$$\max_{i=1,\dots,n}\frac{|x_i|^2}{s_n^2}\to 0. \tag{21.4}$$

**Théorème 21.3.** *Sous les hypothèses précédentes, Équation 21.3 converge en loi lorsque $n\to\infty$ vers une $N(0,1)$.*

*Démonstration.* La démonstration repose sur Théorème 21.2 appliqué aux $X_i = x_i\varepsilon_i$ : ces variables sont centrées, et leur variance est $\sigma^2 x_i^2$. En particulier,

$$s_n^2=\sigma^2\sum_{i=1}^{n}x_i^2.$$

Le terme $\mathbb{E}[|X_i|\mathbf{1}_{|X_i|>\delta s_n}]$ vaut $x_i^2\mathbb{E}[\varepsilon^2\mathbf{1}_{|\varepsilon|>\delta s_n/|x_i|}]$. Par l'inégalité de Cauchy-Schwarz, $\mathbb{E}[\varepsilon^2\mathbf{1}_{|\varepsilon|>\delta s_n/|x_i|}]$ est borné par $\sqrt{\mathbb{E}[\varepsilon^4]\mathbb{P}(|\varepsilon|>\delta s_n/|x_i|)} = \sigma^2 c\sqrt{\mathbb{P}(|\varepsilon|>\delta s_n/|x_i|)}$, qui est également plus petit que

[1] J'ai l'impression que ce nom n'est guère répandu dans la littérature, mais je l'ai trouvé dans le livre *Introduction à l'économétrie* de Brigitte Dormont.

$\sigma^2 c\sqrt{\mathbb{P}(|\varepsilon| > \delta m_n)}$ où $m_n$ est le plus petit des nombres $s_n/|x_1|, \ldots, s_n/|x_n|$, c'est-à-dire l'inverse de la racine carrée de Équation 21.4.

En regroupant tout ceci, on voit que Équation 21.2 devient plus petite que

$$\frac{\sigma^2 c}{s_n^2} \sum_{i=1}^{n} x_i^2 \sqrt{\mathbb{P}(|\varepsilon| > \delta m_n)}$$

c'est-à-dire $c \times \sqrt{\mathbb{P}(|\varepsilon| > \delta m_n)}$. Comme $m_n \to \infty$ par Équation 21.4, ce terme tend vers zéro.

□