Statistiques Fondamentales

Simon Coste

2025-02-17

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.

- **Horaires.** Les cours ont lieu
  - lundi 13h45 - 16h45, SG1021
  - jeudi 8h30 - 10h30, SG1021
  - vendredi 10h45 - 12h45 puis 13h30 - 15h30, SG2015
- **Agenda.** Les cours reprennent le 13 janvier. Il y a six semaines de cours ; le programme approximatif est le suivant :
  - S1 : Estimation, intervalles de confiance (chapitres 1 à 5).
  - S2 : Tests (chapitres 6 à 8)
  - S3 : Tests, exercices, **examen partiel le jeudi 30**.
  - S4 : Modèles linéaires et MCO (chapitres 9 à 11).
  - S5 : Modèles exponentiels et théorie de l'information (chapitres 13 à 17)
  - S6 : Compléments et révisions.

  Il y aura une interrogation par semaine.

  Il y a un partiel jeudi 30 janvier.

  L'examen final a lieu **Lundi 3 mars de 14h à 17h.**
- **Calcul de la note finale**. Trois quantités entrent en jeu : la note de contrôle continu $c \in [0, 10]$, la note du partiel $p \in [0, 20]$, et la note que vous obtiendrez à l'examen, $e \in [0, 20]$. La note finale $f \in [0, 20]$ est donnée par la formule

$$f = 20^{\theta} \left(\frac{\max\{p, e\} + e}{2} \right)^{1-\theta}$$

  où $\theta = 0.3 \times c/10$. C'est une moyenne géométrique qui ne pénalise personne, et qui ne peut que tirer votre note vers le haut si vous avez une note de contrôle continu c strictement positive.

Utiliser ce site

Chaque chapitre de ce livre contient une page dédiée au cours théorique, et une page d'exercices.

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* rend l'analyse classique moins puissante. Il peut s'agir de problèmes dans lesquels des données indépendantes ont été générées par le même phénomène. Dans cette section, nous allons développer un exemple qui illustre bien 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 Une histoire d'influenceurs

Dans le monde des applications mobiles, une donnée clé pour de nombreux modèles de type freemium est la *conversion*. Il ne suffit pas que des utilisateurs téléchargent l'app et utilisent sa version gratuite : encore faut-il qu'ils décident de payer, c'est-à-dire de convertir leur accès gratuit en abonnement payant. C'est ce qu'on appelle « convertir ». À de nombreux égard, cette métrique est essentielle pour faire croître un projet. Or, avant même de *convertir* des utilisateurs, il est évidemment essentiel d'en *acquérir*. C'est le rôle du marketing ; aujourd'hui, une grande partie du marketing moderne passe par du marketing de contenu : on paie des influenceurs pour qu'ils parlent de l'app ou créent du « contenu » qui la mentionne.

Lorsque j'ai lancé ma première app mobile, mon amie Carlita était influenceuse, et possédait une communauté assez large et proche de mon business. J'avais signé avec elle un contrat de marketing : Carlita devait publier une vingtaine de contenus par mois (stories, reels) parlant de mon app. C'était mon seul et unique canal d'acquisition de clients.

Estimation. Entre janvier et mars, Carlita avait généré 4023 acquisitions, et sur ces acquisitions, il y eut 460 conversions. À supposer que la communauté de Carlita soit stable et qu'une proportion fixe p des nouveaux utilisateurs convertissent effectivement le mois suivant, on pouvait donc estimer p par $460/4023 \approx 11.5\%$.

Précision. Au fil du temps, Carlita générait environ 10000 nouveaux utilisateurs gratuits par semestre. Je pouvais donc m'attendre à ce que dans la foulée, $11.5\% \times 10000 = 1150$ d'entre eux convertissent leur abonnement gratuit en abonnement payant, ce qui me permettait de prévoir tout un tas de choses (besoins en coût de serveurs, cashflow futur, etc). Seulement, pour des raisons comptables, il était indispensable que le nombre de conversions semestrielles ne tombe pas en dessous de 1000. À quel point ce nombre de 11.5% était-il donc précis ?

Une modélisation rapide permet de s'en rendre compte. Chaque utilisateur peut être vu comme une variable de Bernoulli de paramètre p : $X_i = 1$ si conversion, 0 sinon. L'estimation ci-dessus est la moyenne empirique des conversions, $\bar{X}_n = (X_1 + \cdots + X_n)/n$ avec $n = 4023$. La loi de cette variable est tout simplement $\mathrm{Bin}(n, p)/n$, et on sait qu'avec probabilité supérieure à 99%, l'estimée $\bar{X}_n$ et la vraie valeur p ne diffèrent pas de plus de 0.7% : l'inégalté de Bienaymé-Chebychev dit que $\mathbb{P}(|\bar{X}_n - p| > 0.007) < \frac{\mathrm{Var}(\bar{X}_n)}{0.007^2}$, et l'on peut borner ceci (pourquoi ?) par

$$\frac{1}{4n0.007^2} = 99\%.$$

Autrement dit, j'étais sûr à 99% que le vrai taux de conversion de Carlita se situait quelque part entre 10.8% et 12.2%. En moyenne, il était donc peu probable que sur 10000 utilisateurs que Carlita pouvait emmener sur mon app chaque semestre, seulement moins de 1000 convertissent.

Ce raisonnement donne ce qu'on appelle un *intervalle de confiance* pour p.

Tests. Un jour, une webagence vint me démarcher. Elle me proposait de signer un contrat avec un de *ses* influenceurs, appelons-le Szeequie, qui avait un taux de conversion supposé meilleur que Carlita, à savoir 12%. L'agence avait estimé ce chiffre sur le dernier mois de Szeequie : il avait généré 10000 acquisitions pour 1200 conversions, sur des apps similaires aux miennes. Évidemment, une question se pose : Szeequie a généré bien plus d'acquisitions que Carlita, l'estimation de 12% est donc plus précise que celle des 11.5%. Mais à quel point ? Peut-on *vraiment* dire que la communauté de Szeequie convertit mieux que celle de Carlita ?

Reprenons le calcul ci-dessus. On a vu qu'il y a 99% de chances pour que p_{carlita} soit compris entre 10.7% et 12.2%. Le même calcul montre qu'il y a 99% de chances pour que p_{szeequie} soit compris entre 11.99% et 12.01%.

Autrement dit, en écartant des événements trop rares (qui arrivent moins de 1% des fois), les deux intervalles se chevauchent : je ne suis pas **absolument sûr** que Szeequie est absolument meilleur que Carlita. Peut-être que p_{carlita} est égal à 12.2% mais qu'elle n'avait pas eu de chance lorsque j'ai fait ma première estimation. En revanche, si Carlita se met à acquérir pour moi beaucoup d'utilisateurs (disons, 100000) et que l'estimation de son taux de conversion ne change pas et reste 11.5%, alors les deux estimations seront plus précises et me permettront d'être sûr que je devrais plutôt travailler avec Szeequie.

En attendant, Carlita reste mon amie : tant que je ne suis pas sûr que d'autres sont vraiment meilleurs, je préfère travailler avec elle.

Ce mode de raisonnement porte le nom de *test d'hypothèse*, et ce n'est pas pour rien que je mentionne mon amitié avec Carlita : dans tous les problèmes statistiques, il y a des préférences, des non-dits et des choix qui rendent les choses non symétriques. On veut avant tout être sûr de rejeter une hypothèse (dite « nulle »), et surtout pas la rejeter à tort.

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 Θ est appelé espace des paramètres. On écrira fréquemment $\mathbb{E}_\theta$ ou Var_θ pour désigner des espérances, variances, etc., calculées avec la loi P_θ.

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_θ 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_θ.

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 paramètre, disons $\theta_\star$, tel que les observations $x_1, \dots, x_n$ sont des réalisations de loi $p_{\theta_\star}$. Si ce n'est pas le cas, il existe certainement un $\theta_\star$ tel que $p_{\theta_\star}$ est « la loi la plus proche possible » de la vrai loi qui a généré les x_i. 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 Θ 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 $(\mathcal{X}, \mathscr{F}, (P_\theta)_\theta)$, alors une statistique est une fonction mesurable de $(\mathcal{X}, \mathscr{F})$.

1) Le premier point important est qu'une statistique ne peut pas prendre θ en argument. Elle ne doit pas du tout dépendre de θ.

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_θ, on peut définir la valeur moyenne de T sous le modèle P_θ 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 calcul de la statistique nécessite seulement des multiplications et des additions (comme pour une moyenne empirique), c'est bien ; mais parfois, elle nécessite la résolution d'un problème algorithmique trop difficile (typiquement NP-complet).

Si le but est de deviner la valeur de θ à partir des observations d'une variable aléatoire de distribution P_θ, il est naturel de considérer des statistiques à valeurs dans Θ. 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 θ est une statistique à valeurs dans Θ.

En fait, on n'est pas obligés de vouloir estimer précisément θ. Peut-être qu'on veut estimer quelque chose qui dépend de θ, mais qui n'est pas θ, comme $\cos(\theta)$ ou θ^2 ; disons, une fonction $\varphi(\theta)$. Dans ce cas, un estimateur de $\varphi(\theta)$ sera simplement une statistique à valeurs dans l'espace où vit $\varphi(\theta)$.

La tradition statistique consiste à coiffer d'un chapeau tout paramètre que l'on veut estimer, et à l'affubler éventuellement d'un indice (typiquement n) pour rappeler des paramètres d'intérêt, comme la taille de l'échantillon. On utilisera par exemple $\widehat{\theta}_n$ pour désigner un estimateur de θ calculé sur un échantillon de taille n. Cette notation cache le fait que $\widehat{\theta}$ est bien une variable aléatoire.

Exemple 1.1. Soit $X = (X_1, \dots, X_n)$ des variables iid gaussiennes. Le modèle statistique sous-jacent est $\mathcal{X} = \mathbb{R}^n$, $\mathscr{F}$ la tribu borélienne usuelle, et la famille de lois que l'on considère est la famille $N(\mu, \sigma^2)^{\otimes n}$. La moyenne empirique

$$\widehat{\mu} = \frac{X_1 + \cdots + X_n}{n}$$

est un estimateur, associé à la fonction mesurable $T : \mathbb{R}^n \to \mathbb{R}$ définie par $T(x_1, \dots, x_n) = n^{-1}(x_1 + \dots + x_n)$. On a donc $\widehat{\mu}_n = T(X_1, \dots, X_n)$. C'est bien une variable aléatoire.

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 θ déterminent le hasard qui génère les observations x_i. La statistique bayésienne consiste à renverser le point de vue, et à rendre le paramètre

θ 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 θ ou un autre paramètre qui dépend de θ. 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 θ lui-même, mais un paramètre $\phi = \phi(\theta)$ qui dépend de θ. Dans ce cas, si $\hat{\phi}$ est un estimateur de ϕ alors le biais est défini par $\mathbb{E}_\theta[\hat{\phi} - \phi]$ et le risque quadratique par $\mathbb{E}_\theta[|\hat{\phi} - \phi|^2]$.

Théorème 2.1 (Décomposition biais-variance)**.** *Le risque quadratique* $\mathbb{E}_\theta[|\hat{\theta} - \theta|^2]$ *est égal à*

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

Démonstration. En notant x l'espérance de $\hat{\theta}$, on voit que le risque quadratique est égal à $\mathbb{E}[|\hat{\theta} - x - (\theta - x)|^2]$. Le carré se développe en trois termes : le premier, $\mathbb{E}[|\hat{\theta} - x|^2]$, est la variance de $\hat{\theta}$. Le second, $-2\mathbb{E}[(\hat{\theta} - x)(\theta - x)]$, est égal à $-2(\theta - x)\mathbb{E}[\hat{\theta} - x]$, c'est-à-dire 0. Le dernier, $\mathbb{E}[(\theta - x)^2]$, est égal à $(\theta - x)^2$, c'est-à-dire $(\theta - \mathbb{E}[\hat{\theta}])^2$: c'est bien le carré du biais.

□

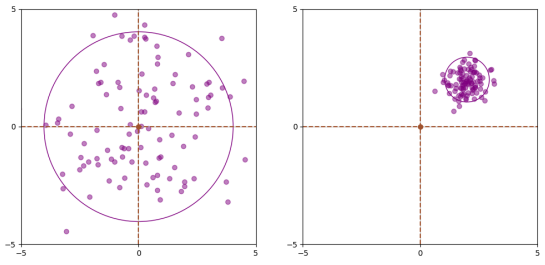


Figure 2.1: À gauche, RQ élevé mais biais nul ; à droite, RQ faible mais biais non nul.

2.2 Convergence

La dépendance du biais ou du risque quadratique (ou d'autres indicateurs) vis à vis de la taille de l'échantillon est une question importante : pour une suite d'expériences donnée, on veut que ces indicateurs tendent vite vers zéro. Quelles sont les meilleures vitesses envisageables, et comment les obtenir ?

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 $\varepsilon > 0$,

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

On dit qu'elle converge $\mathbb{P}$-presque sûrement vers X si

$$\mathbb{P}(\lim X_n = X) = 1.$$

Définition 2.2 (convergence d'un estimateur)**.** Une suite d'estimateurs $(\hat{\theta}_n)$ est convergente pour l'estimation de θ lorsque, pour tout $\theta \in \Theta$, sous P_θ, la suite $(\hat{\theta}_n)$ converge en probabilité vers θ ; autrement dit, lorsque pour tout $\varepsilon > 0$,

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

La suite est fortement convergente si, pour tout θ, la convergence a lieu P_θ-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, parce que beaucoup d'estimateurs sont des sommes de variables indépendantes.

Définition 2.3 (normalité asymptotique)**.** Soit θ un paramètre à estimer, et $\hat{\theta}_n$ une suite d'estimateurs de θ. 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 Σ est une matrice de covariance qui dépend peut-être de θ. Pour éviter les cas dégénérés, on demande à ce que Σ soit non-nulle.

La normalité asymptotique sera utilisée pour construire des intervalles de confiance et des tests. Aussi, prouver que des estimateurs sont asymptotiquement normaux est une tache importante, qui est grandement simplifiée par les outils suivants.

2.4 Trois outils sur la normalité asymptotique

Vous devez maîtriser sur le bout des doigts le Théorème Central-Limite. Le Chapitre 23 de l'appendice revient sur différentes formes qu'il peut prendre.

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 σ^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 $N(0,\sigma^2)$.

Un autre lemme, dit « 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 φ,*

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

Démonstration. Fixons une fonction test φ 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, φ 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 $2M$. Le terme $|\mathbb{E}A|$ est donc plus petit que

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

- 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 que $\mathbb{E}A \to 0$.

□

On termine par la « delta-méthode » : si une suite d'estimateurs est asymptotiquement normale, leur image par n'importe quelle fonction lisse g l'est encore, et on sait calculer la moyenne et la variance limites.

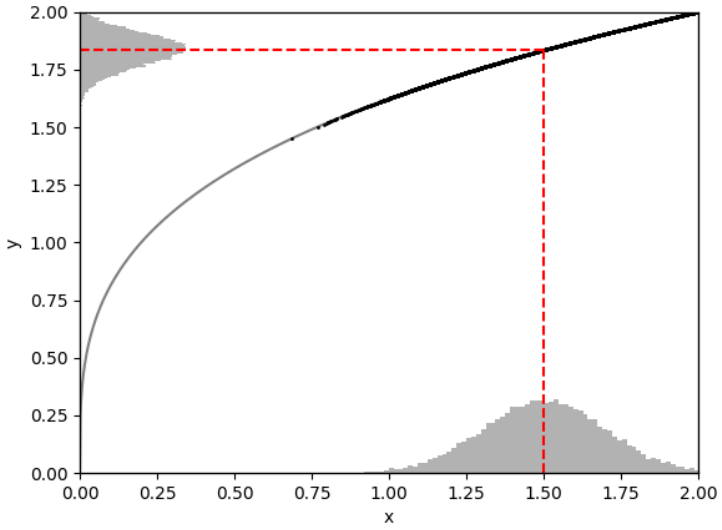


Figure 2.2: Une petite gaussienne autour de 1.5 est transformée par une fonction $f \in \mathscr{C}^1$ en une petite gaussienne autour de $f(1.5)$, de variance normalisée par $f'(1.5)$.

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 α (de dérivée non nulle en α), alors la variable aléatoire*

$$\sqrt{n}(g(X_n) - g(\alpha))$$

converge en loi vers

$$N(0, g'(\alpha)^2\sigma^2).$$

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. L'approximation au premier ordre de g au point α dit que $\sqrt{n}(g(X_n) - g(\alpha))$ est à peu près égale à $\sqrt{n}(X_n - \alpha)g'(\alpha)$, et comme $g'(\alpha)$ est une constante, ce terme converge bien en loi vers $N(0, \sigma^2 g'(\alpha)^2)$. Il suffit donc de montrer que le terme de reste r_n qui complète le « à peu près » de la phrase précédente tend lui-même vers 0. On va montrer qu'il tend bien vers 0 en probabilité : l'application du Lemme de Slutsky ci-dessus permettra de conclure.

Soit donc $s > 0$. On veut montrer que $\mathbb{P}(|r_n| > s)$ tend vers zéro. On se donne un $\varepsilon > 0$, et on introduit l'événement $E_n(b) = \{|\sqrt{n}(X_n - \alpha)| < b\}$, où b est une constante arbitraire.

Par définition de la dérivabilité de g en α, il y a une fonction r telle que $g(x) - g(\alpha) - (x - \alpha)g'(\alpha)$ est égal à $r(x)$ et telle que $r(x)/x \to 0$ quand $x \to 0$. Au vu des définitions ci-dessus, on peut même écrire que $r_n = \sqrt{n}r(X_n - \alpha)$. Il y a un δ tel que si $|x| < \delta$ alors $|r(x)|/|x| < s/b$.

Sur E_n, dès que n est plus grand que $(b/\delta)^2$, alors $|X_n - \alpha| < \delta$ et donc $|r_n| < (s/b)\sqrt{n}|X_n - \alpha|$, ce qui est plus petit que s. On vient de montrer que sur l'événement E_n, dès que n est grand il devient impossible d'avoir $|r_n| > s$. Par conséquent,

$$\overline{\lim}\mathbb{P}(|r_n| > s) \leqslant \overline{\lim}\mathbb{P}(\overline{E_n}).$$

Or, par définition de la convergence en loi, $\mathbb{P}(\overline{E_n})$ converge vers la probabilité $p(b)$ qu'une variable de loi $N(0, \sigma^2)$ soit plus grande en valeur absolue que b. Cette probabilité converge vers 0 quand b est très grand, donc on peut choisir b de sorte que $p(b) < \varepsilon$. Ainsi,

$$\overline{\lim}\mathbb{P}(|r_n| > s) \leqslant \varepsilon.$$

Cela entraîne bien la convergence vers 0 de $\mathbb{P}(|r_n| > s)$.

□

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 μ et une variance σ^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. *Si les X_i sont indépendants (ou simplement décorrélés), l'estimateur $\hat{\sigma}_n^2$ est sans biais.*

Démonstration. Il suffit de calculer l'espérance de $(n-1)\hat{\sigma}_n^2$, ce qui revient à calculer la somme des $\mathbb{E}[X_i^2 - 2X_i\bar{X}_n + \bar{X}_n^2]$. Il y a trois éléments dans cette expression : $\mathbb{E}[X_i^2], -2\mathbb{E}[X_i\bar{X}_n]$ et $\mathbb{E}[\bar{X}_n^2]$.

Le premier terme est égal à σ^2. Le second terme vaut $-2\sum_j \mathbb{E}[X_iX_j]/n$, et tous les termes avec $i \neq j$ sont nuls car les X_k sont décorrélés. Il ne reste que le terme $j = i$, à savoir $-2\sigma^2/n$. Enfin, $\mathbb{E}[\bar{X}_n^2]$ est la variance de $\bar{X}_n$, c'est-à-dire σ^2/n.

En additionnant tout, on obtient $\sigma^2 - \sigma^2/n$ et comme il y avait n termes dans la somme initiale, on obtient $(n-1)\mathbb{E}[\hat{\sigma}_n^2] = n\sigma^2 - \sigma^2$ et donc $\mathbb{E}[\hat{\sigma}_n^2] = \sigma^2$.

□

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 (qui est un cas particulier de la première). 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, dont la moyenne n'est pas le paramètre θ lui-même, mais plutôt une *fonction* de θ :

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

C'est cette fonction φ qu'on appelle *moment*. Typiquement,

- la moyenne d'une loi $\mathscr{E}(\theta)$ n'est pas θ 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, P_θ-presque sûrement on a

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

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

3.2 Estimateur des moments

Si la fonction φ est inversible et si $\bar{T}_n$ appartient presque sûrement à l'ensemble de définition de φ^{-1}, alors $\varphi^{-1}(\bar{T}_n)$ est bien définie. Pour qu'en plus cette quantité converge presque sûrement vers θ, il faut s'assurer que φ^{-1} est continue. C'est par exemple le cas lorsque l'ensemble des paramètres Θ est un ouvert, et que φ 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 φ 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 θ. 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 φ^{-1} en $\varphi(\theta)$ n'est autre que l'inverse de la matrice jacobienne de φ en θ.

□

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 λ, 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 ?
6. Dans la delta-méthode, on demande à ce que la dérivée de g au point limite soit non nulle. Pourquoi ?
7. On se place dans le cadre de la delta-méthode, avec une suite X_n telle que $\sqrt{n}(X_n - \alpha)$ converge en loi vers $N(0,1)$. Si on a une fonction g telle que $g'(\alpha) = 0$ mais $g''(\alpha) \neq 0$, comment renormaliser $X_n - \alpha$ pour qu'on ait encore la normalité asymptotique ?

Exercices

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

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

1. Déterminer un estimateur de θ 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 β connu. Proposer un estimateur de α par la méthode des moments.
2. On suppose à présent que les deux paramètres α, β sont inconnus. Proposer un estimateur de (α, β) par la méthode des moments.

Exercice 3.6 (Lois de Gumbel)**.** La loi de Gumbel (centrale) de paramètre β 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 β.

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 ρ et donner ses propriétés.

Exercice 3.8 (Estimation des espèces manquantes de Fisher-Corbet (examen 24))**.** Un naturaliste capture des papillons dans une forêt lointaine. Il effectue deux collectes de même durée. Soit S le nombre total d'espèces distinctes dans cette forêt ; on note X_k^1 (respectivement X_k^2) le nombre de spécimens de l'espèce k qui ont été capturées pendant la première collecte (respectivement, la seconde). On fait l'hypothèse selon laquelle les X_k^i sont des variables aléatoires indépendantes avec $X_k^i \sim$ Poisson(θ_k) où θ_k est un paramètre positif qui représente la rareté de l'espèce k.

1. On pose N_r le nombre d'espèces dont exactement r spécimens ont été capturés **pendant la première collecte**. Par exemple, N_3 est le nombre d'espèces dont on a capturé 3 spécimens et N_0 est le nombre d'espèces qu'on n'a pas capturées.
   a. Calculer l'espérance et la variance des N_r.
   b. Quelle est la probabilité qu'aucun spécimen de l'espèce k n'ait été capturé pendant la première collecte, mais qu'au moins un ait été capturé pendant la seconde ?
2. En utilisant le développement $e^x - 1 = \sum_{t=1}^{\infty} x^t/t!$, montrer que le nombre d'espèces N qui n'ont pas été capturées pendant la première collecte mais qui ont été capturées pendant la seconde collecte vérifie
$$\mathbb{E}[N] = \sum_{t=0}^{\infty}(-1)^t \sum_{k=1}^{S} e^{-\theta_k} \frac{\theta_k^{t+1}}{(t+1)!}.$$
3. En déduire que
$$\hat{N} := N_1 - N_2 + N_3 - N_4 + N_5 - \cdots$$
est un estimateur sans biais de N *même lorsque* S *n'est pas connu.*
4. Expliquer pourquoi ce résultat est étonnant. Quelles pourraient être les limitations de cet estimateur ?

Concours

J'ai choisi un polynôme $p \in \mathbb{R}[X]$. Puis, j'ai tiré aléatoirement $n = 1000$ réalisations iid d'une loi $N(0,1)$, appelées ε_i, et pour chacune j'ai calculé $x_i = p(\varepsilon_i)$. Ces $\{x_i\}$ sont les observations que je vous donne dans ce fichier.

Votre objectif est de deviner le mieux possible le polynôme p, c'est-à-dire son degré et ses coefficients.

Règles :

- la solution doit m'être envoyée par mail **avant dimanche 20 janvier, 23h59**.
- l'objet du mail doit être la suite des coefficients que vous avez estimés. Par exemple, si vous pensez que le polynôme est $p(x) = 1 + x^2 - x^5 + 3x^7$, vous écrivez 1,0,1,0,0,-1,0,3 dans l'objet du mail, sans rien d'autre.
- vous devez me montrer que vous avez compris ce que vous faites et l'exprimer dans le cadre du cours.
- envoyez moi le code ou les calculs que vous avez faits.

Ceux qui auront correctement estimé le degré et/ou le polynôme auront comme récompense des M&Ms.

4 Intervalles de confiance

4.1 Principe

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

Pour simplifier, on supposera d'abord que θ 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 θ,

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

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 α 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 $X_1, \ldots, X_n$ de loi $N(\mu, \sigma^2)$. On va donner des intervalles de confiance pour l'estimation des paramètres μ et σ dans plusieurs cas de figure.

4.2.1 Estimation de μ

Lorsque σ est connue.

La moyenne empirique $\bar{X}_n$ est un estimateur sans biais de μ. 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 μ. 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 $z_{0.99} \approx 2.57$. Or, l'inégalité

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

équivaut à[1]

$$\mu \in \left[\bar{X}_n \pm \frac{z_{0.99}\sigma}{\sqrt{n}}\right]. \tag{4.3}$$

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 σ 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 μ 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 σ est inconnue.

Lorsque σ n'est pas connue, les bornes A, B ci-dessus ne sont pas des statistiques, car elles dépendent de σ. On peut estimer σ sans biais (cf Théorème 2.5) 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 Équation 4.1, on remplace σ 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 gaussienne : c'est 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 α de $\mathscr{T}(k)$. Alors,

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

[1] on rappelle que $(XY)^{-1} = Y^{-1}X^{-1}$.

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

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

est équivalente à

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

et les deux côtés de cet intervalle sont des statistiques; en les notant A, B, on a bien trouvé un intervalle de confiance de niveau α, 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 μ lorsque σ n'est pas connue est donné par*

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

4.2.2 Estimation de σ

Supposons maintenant qu'on désire estimer la variance σ^2.

Lorsque μ est connue.

En supposant que μ est connue, l'estimateur des moments le plus naturel pour estimer σ^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 éloignés possibles[2] tels que $\mathbb{P}(k_{n,\alpha}^- < \chi^2(n) < k_{n,\alpha}^+) = 1 - \alpha$. Ainsi,

$$P_{\sigma^2}(k_{n,\alpha}^- < \frac{n\tilde{\sigma}_n^2}{\sigma^2} < k_{n,\alpha}^+) = 1 - \alpha.$$

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

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

est un intervalle de confiance de niveau α pour σ^2.

Lorsque μ est inconnue.

[2] à titre d'exemple, toute valeur de z plus grande que 10 donne un $\theta(z)$ plus grand que 99.99%.

Cette fois, on utilise l'estimateur de Théorème 2.5, 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 (cela sera montré dans Section 5.2.2). Le même raisonnement que ci-dessus donne l'intervalle de confiance de niveau $1-\alpha$ suivant :

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

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 $p \in]0,1[$. La moyenne empirique, $\hat{p}_n = (X_1 + \cdots + X_n)/n$ est un estimateur est non biaisé de p et son risque quadratique $p(1-p)/n$. 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 $\mathrm{Bin}(n,p) - p$, on pourra construire des intervalles de confiance de niveau $1-\alpha$. Ces quantiles peuvent être calculés par des méthodes numériques, mais il y a plus simple.

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 α. 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_α est le quantile symétrique d'ordre α 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 α, 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 μ, 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. L'inégalité de Hoeffding est beaucoup plus fine que l'inégalité de Bienaymé-Tchebychev, mais elle ne s'applique qu'aux variables qui sont bornées ou sous-gaussiennes.

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 $X_1, \dots, X_n$ 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 α pour l'estimation de la moyenne des X_i.

5 Outils pour les Intervalles de Confiance

5.1 Quantiles

5.1.1 Quantiles des lois continues

Si X est une variable aléatoire continue sur $\mathbb{R}$, un quantile d'ordre $\beta \in]0,1[$, noté q_β, 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 bijection continue. Les quantiles symétriques z_β 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 β est de la forme $1-\alpha$ avec α 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.

β	90%	95%	98%	99%	99.9%	99.99999%
z_β	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, $\mathbb{P}(X > x)$ est compris entre*

$$\left(1 - \frac{1}{x^2}\right) \frac{1}{x} \frac{e^{-x^2/2}}{\sqrt{2\pi}} \tag{5.1}$$

et

$$\frac{1}{x}\frac{e^{-x^2/2}}{\sqrt{2\pi}}. \tag{5.2}$$

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. Le nombre $\mathbb{P}(X > x)$ est exactement égal à $(2\pi)^{-1/2}\int_x^\infty e^{-t^2/2}dt$. En multipliant et en divisant l'exponentielle dans l'intégrale par t et en faisant une intégration par parties, on peut écrire ceci sous la forme

$$\frac{e^{-x^2/2}}{x\sqrt{2\pi}} - \int_x^\infty \frac{e^{-t^2/2}}{t^2\sqrt{2\pi}}dt.$$

Comme l'intégrale I à droite est positive, tout ce terme est bien plus petit que Équation 5.1. Par ailleurs, en refaisant la même astuce, on peut écrire I sous la forme

$$\frac{e^{-x^2/2}}{x^3\sqrt{2\pi}} - 2\int_x^\infty \frac{e^{-t^2/2}}{t^3\sqrt{2\pi}}dt.$$

Si J est la nouvelle intégrale à droite, elle est positive ; on a donc montré que $\mathbb{P}(X > x)$ est aussi égal à

$$\left(1 - \frac{1}{x^2}\right)\frac{1}{x}\frac{e^{-x^2/2}}{\sqrt{2\pi}} + J$$

et donc, est plus grand que Équation 5.2.

□

5.1.2 Quantiles généraux

Dans le cas général où la fonction de répartition $F(x) = \mathbb{P}(X \leqslant x)$ n'est pas continue (par exemple pour les lois discrètes), on peut toujours définir les quantiles de n'importe quel ordre $\beta \in [0, 1]$ en prenant la plus petite valeur q telle que $\mathbb{P}(X \leqslant q) \geqslant \beta$. Autrement dit,

$$q(\beta) = \inf\{t : F(t) \geqslant \beta\}.$$

5.2 Calculs de lois

5.2.1 Lois Gamma

Une variable aléatoire de loi Gamma de paramètres $\lambda > 0, \alpha > 0$ possède une densité $\gamma_{r,\alpha}(x)$ égale à

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

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 par un changement de variables :

$$\varphi_{\lambda,\alpha}(t)=\left(1-\frac{it}{\lambda}\right)^{-\alpha}. \tag{5.3}$$

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

La loi du chi-deux est la loi du carré d'une gaussienne standard X. Calculons sa densité ; pour toute fonction-test φ, $\mathbb{E}[\varphi(X^2)]$ 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 :

$$\begin{aligned}\mathbb{E}[e^{it(X_1^2+\dots+X_n^2)}]&=\mathbb{E}[e^{itX_1^2}]^n &(5.4)\\ &=(1-2it)^{-\frac{n}{2}}. &(5.5)\end{aligned}$$

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

$$\frac{1}{2^{n/2}\Gamma(n/2)}e^{-x/2}x^{n/2-1}\mathbf{1}_{x>0}. \tag{5.6}$$

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 φ 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

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

La densité $t_n(u)$ de T_n est donc donnée par

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

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

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

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 $X_1, \dots, X_n$ 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 $X = (X_1, \dots, X_n)$ est gaussien. Posons $Z = (X_1 - \bar{X}_n, \dots, X_n - \bar{X}_n)$. 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 $\mathscr{V} = \{x \in \mathbb{R}^n : x_1 + \dots + x_n = 0\}$. Soit $(f_i)_{i=2,\dots,n}$ 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,

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

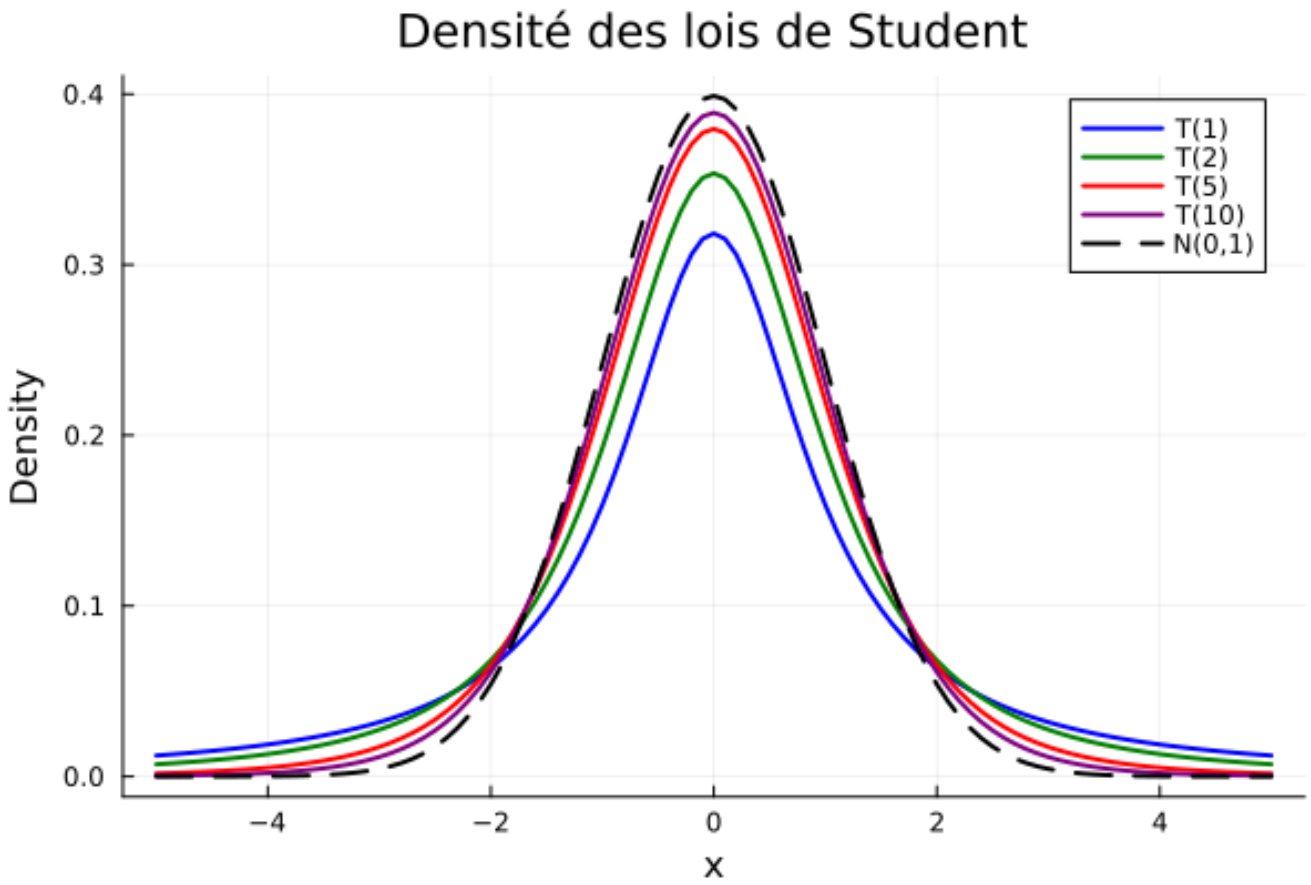


Figure 5.1: Densité de plusieurs lois de Student.

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 $X_1, \ldots, X_n$ 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 $S_n = X_1 + \ldots + X_n$. Pour tout $t > 0$,*

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geqslant t) \leqslant e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}} \tag{5.7}$$

et

$$\mathbb{P}(|S_n - \mathbb{E}[S_n]| \geqslant t) \leqslant 2e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}. \tag{5.8}$$

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,*

$$\mathbb{E}[e^{t(X - \mathbb{E}[X]}] \leqslant e^{\frac{t^2(b-a)^2}{8}}. \tag{5.9}$$

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 5.9, 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 ξ. 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 λ 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 φ_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 λ 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 5.7. On déduit Équation 5.8 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 5.9. 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 μ d'un modèle gaussien où la variance σ^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 de Poisson de paramètre θ.

1. Étudier $\bar{X}_n$.
2. Montrer que $\sqrt{\bar{X}_n}$ converge en probabilité vers $\sqrt{\theta}$.
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 θ 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 θ que l'on notera $\hat{\theta}_n$. Étudier ses propriétés (risque quadratique, convergence) et l'utiliser pour construire un premier intervalle de confiance $I_1(\alpha)$ non-asymptotique pour θ de niveau $1-\alpha$.
2. Construire un intervalle de confiance asymptotique $I_2(\alpha)$ pour θ à partir de $\hat{\theta}_n$.
3. Montrer que l'estimateur $\theta_n^\star := \min_{1\le i\le n} X_i$ est meilleur que $\hat{\theta}_n$ 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 λ.

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 Exercice 3.2, lorsque les $X_1, \dots, X_n$ sont des tirages *sans remise* dans une urne à N éléments.

Exercice 5.6 (Examen 24)**.** Soit (E_i) une suite de variables aléatoires iid de loi exponentielle de paramètre $c > 0$, et soit (X_i) une suite de variables aléatoires indépendantes de loi de Poisson, avec $X_i \sim \text{Poisson}(E_i)$. Autrement dit, X_i suit une loi de Poisson de paramètre E_i qui est lui-même aléatoire et dépend de c. L'objectif est d'estimer c.

1. Calculer $\mathbb{P}(X_i = n)$ pour tout n. Quelle est la loi de X_i ?
2. Proposer un estimateur convergent de c.
3. Proposer un intervalle de confiance asymptotique de c de niveau de risque α.

6 Test d'hypothèses

Si l'on essaie d'estimer le rendement μ 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 $(\mathcal{X}, \mathscr{F}, (P_\theta)_{\theta\in\Theta})$, c'est supposer que θ 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 θ. 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 θ 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 $\theta \in H_0 \Rightarrow \mathsf{accepter}$ (avec grande probabilité), donc $\mathsf{rejeter} \Rightarrow \theta \notin H_0$ (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 σ^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 μ soit effectivement égal à zéro ! On pose donc

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

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

Supposons maintenant que σ 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

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

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 μ 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 Φ 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 α. 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

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

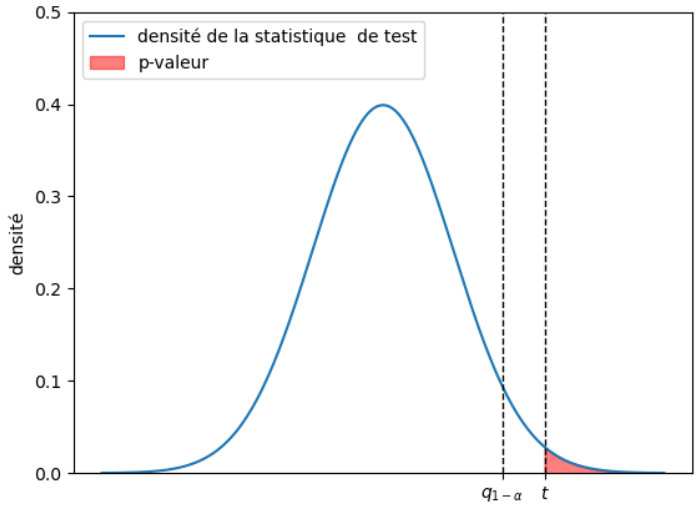


Figure 6.1: p-valeur d'un test dont la statistique d'intérêt est t.

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 événément 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

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

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 $\mathrm{d_{TV}}(P_n, Q) \to 0$ alors P_n converge en loi vers Q mais l'inverse n'est pas vrai.

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

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

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

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

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

[1] on rappelle que $(XY)^{-1} = Y^{-1}X^{-1}$.

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 ν. 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 ν.

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 α, β 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 α et le test est de niveau de confiance supérieur à $1-\alpha$.

7.5 Variation totale et distance informationnelle

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 l'entropie et 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 d_{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 ρ 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.

□

On peut se demander s'il existe une inégalité dans l'autre sens, permettant de borner la distance KL par quelque chose qui dépend de la distance en variation totale. La réponse est oui, et dépend d'une quantité importante en théorie de l'information, la *varentropie*.

Définition 7.4. La varentropie de Q par rapport à P est la variance de $\ln(\mathrm{d}P/\mathrm{d}Q)$ par rapport à P :

$$\mathrm{Varent}(P|Q) = \int \ln\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right)^2 dP - \left(\int \ln\left(\frac{\mathrm{d}P}{\mathrm{d}Q}\right) dP\right)^2.$$

La varentropie mesure à quel point l'information est concentrée autour de l'entropie.

Théorème 7.4 (Borne entropique inférieure)**.**

$$\mathrm{d}_{\mathrm{KL}}(P \mid Q) \leqslant \frac{1+\sqrt{\mathrm{Varent}(\mathrm{P} \mid \mathrm{Q})}}{1-\mathrm{d}_{\mathrm{TV}}(P,Q)}$$

Démonstration. Introduisons l'événement $E = \{x : \ P(x) \geqslant Q(x)e^d\}$ où $d = \mathrm{d}_{\mathrm{KL}} - \sqrt{\mathrm{Varent}}/(1-\mathrm{d}_{\mathrm{TV}})$. Il est clair que

$$P(E) \geqslant e^d Q(E).$$

L'espérance sous P de la variable aléatoire $X = \ln \mathrm{d}P/\mathrm{d}Q$ est d_{KL}, et sa variance est Varent ; le complémentaire de E est $\{X < d\} = \{X - \mathrm{d}_{\mathrm{KL}} < -\sqrt{\mathrm{Varent}}/(1-\mathrm{d}_{\mathrm{TV}})\}$, donc l'inégalité de Bienaymé-Tchebychev dit que

$$P(\bar{E}) \leqslant (1-\mathrm{d}_{\mathrm{TV}})^2.$$

En combinant les deux inégalités, on voit que

$$P(E) - Q(E) \geqslant (1-(1-\mathrm{d}_{\mathrm{TV}})^2)(1-e^{-d}).$$

Or, le membre de gauche est lui-même plus petit que d_{TV}. Par conséquent,

$$\mathrm{d}_{\mathrm{TV}} \geqslant (1-(1-\mathrm{d}_{\mathrm{TV}})^2)(1-e^{-d}).$$

Il est très simple de réarranger les termes de cette inégalité, pour finalement trouver qu'elle est équivalente à

$$e^d \leqslant 1 + \frac{1}{1-\mathrm{d}_{\mathrm{TV}}}.$$

En revenant à la définition de d et en utilisant $\ln(1+x) \leqslant x$, on voit immédiatement que

$$\mathrm{d}_{\mathrm{KL}} \leqslant \frac{1}{1-\mathrm{d}_{\mathrm{TV}}} + \frac{\sqrt{\mathrm{Varent}}}{1-\mathrm{d}_{\mathrm{TV}}}$$

ce qui est l'inégalité demandée.

□

8 Tests du χ_2

Les tests du χ_2 sont une famille de tests qui visent, pour la plupart, à tester si un échantillon discret a été généré par une loi précise ; on parle parfois de test d'ajustement.

8.1 Loi multinomiale

Soit Ω 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 $\mathbf{p} = (p_1, \dots, p_k)$ tels que les p_i sont positifs et de somme 1. On observe n tirages iid selon une même loi sur Ω. 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 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 $N \sim \mathrm{Mult}(n, \mathbf{p})$. Le vecteur $\sqrt{n}(\frac{N}{n} - \mathbf{p})$ converge en loi vers $N(0, \Sigma)$, 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 χ^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, il suffit de regarder l'épaisseur des 26 sections du dictionnaire pour s'en rendre compte.

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 $D^{-1}D^2D^{-1} - D^{-1}\mathbf{p}_0\mathbf{p}_0D^{-1}$ c'est-à-dire

$$I_k - \sqrt{\mathbf{p}}_0\mathbf{p}_0^\top.$$

L'appendice Chapitre 21 rappelle pourquoi cette matrice est une matrice de projection orthogonale.

On a montré que $D^{-1}\sqrt{n}(N/n - \mathbf{p}_0)$ converge en loi vers

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

La statistique qui va nous servir à faire des tests est la norme au carré de $D^{-1}\sqrt{n}(N/n - \mathbf{p}_0)$. En manipulant cette expression, on obtient sa forme usuelle, le *contraste du* χ_2.

Définition 8.1 (Contraste du χ_2)**.** Dans le contexte ci-dessus, le *contraste du* χ_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 21). 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 χ_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).

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.). On cherche 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.

Ici, le modèle statistique sera donc $(\mathbf{p}^{\otimes n} : \mathbf{p} \in S_{k,h})$, où $S_{k,h}$ est l'ensemble des $\mathbf{p} = (p_{i,j})$ 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

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

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 : 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 distinctes. Montrer que $\mathrm{d}_{\mathrm{TV}}(P^{\otimes n}, Q^{\otimes n})$ tend vers 1.
- 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 σ est connu.
- Même question lorsque σ est inconnu (attention : dans ce cas les hypothèses ne sont pas simples mais composites. Il faut utiliser le rapport des maximums de vraisemblance, que l'on verra plus tard).
- 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 μ_1 et μ_2 sont connues.
- Même question lorsque μ_1 et μ_2 ne sont pas connues (même remarque que précédemment).

- 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 λ dans un échantillon de n variables aléatoires de Poisson de paramètre λ.

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 α 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 ε ?

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 θ est un paramètre réel.

1. Donner un intervalle de confiance pour θ 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 ρ_θ 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 α.
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 χ_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 χ_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 ?

Exercice 8.11 (Tests multiples (examen 24))**.** Soient $x_1, \dots, x_n$ des variables aléatoires indépendantes gaussiennes dans $\mathbb{R}^d$, de lois respectives $N(\mu_i, \sigma_i^2 I_d)$, où $\mu_i \in \mathbb{R}^d$ et $\sigma_i^2 > 0$ sont inconnus. On note $H_{0,i}$ l'hypothèse « $\mu_i = 0$ ».

1. Construire un test de $H_{0,i}$ de niveau de confiance $1-\alpha$. On le notera $T_\alpha^{(i)}$.
2. On effectue chacun des tests $T_\alpha^{(i)}$, indépendamment les uns des autres. Si tous les μ_i sont simultanément nuls, quelle est la probabilité qu'aucune des hypothèses $H_{0,i}$ ne soit rejetée ? Quelle est la limite de cette probabilité lorsque $n \to \infty$?
3. Dorénavant, on notera H_0 l'hypothèse selon laquelle tous les μ_i sont nuls. Quel niveau de confiance $1-\delta$ faudrait-il choisir pour que, sous H_0, la probabilité de l'événement « aucun des tests $T_\delta^{(i)}$ ne rejette $H_{0,i}$ » soit égale à $1-\alpha$?
4. On s'intéresse maintenant à la propriété suivante : « sous H_0, la probabilité de l'événement "plus de k% des tests $T_\delta^{(i)}$ ont rejetés $H_{0,i}$" est inférieure à α ».

1. Sous H_0, quelle est la loi du nombre d'hypothèses $H_{0,i}$ rejetées par $T_\delta^{(i)}$?
2. Trouver un δ (dépendant de k, α, n) tel que la famille de tests $T_\delta^{(i)}$ vérifie la propriété ci-dessus. (indice : Wassily).

Travail pratique[1]

Faites un test du chi-deux (indépendance ou adéquation) de votre choix. Vous devez vous-même trouver une question qui vous intéresse. Cela peut être n'importe quoi :

- vérifier la loi de Benford dans je ne sais quel document ou site ou journal ;
- se demander si le département de région d'une personne et ses opinions politiques sont indépendantes ;
- tester une corrélation simple entre la météo à NY et la positivé du rendement du S&P500 (« il pleut à Wall Street » vs « les marchés sont dans le vert ») ;
- se demander si les naissances en France sont uniformément réparties sur les 365 jours de l'année ;
- tester si ChatGPT ou je ne sais quel LLM est capable de générer des échantillons de lois précises (loi uniforme, loi de Poisson, etc). Je serai curieux d'avoir le résultat.
- Etc.

Trouvez donc une question intéressante, trouvez des données pertinentes (c'est peut-être la partie la moins simple), formulez la question sous la forme d'un test statistique, et réalisez le test avec votre langage de programmation favori (qui n'est pas R, soyez de bon goût). Écrivez quelques lignes pour me dire où vous avez eu les données et comment vous vous adaptez au formalisme du cours. Une page max (snippet de code compris) devrait suffire.

Pas besoin d'aller chercher très loin, un jeu de données simple suffit ! Amusez-vous bien.

Vous devez m'envoyer ça avant le 9 février 23h59, avec pour objet du mail votre nom suivi de "chi-deux".

[1] on rappelle que $(XY)^{-1} = Y^{-1}X^{-1}$.

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 α, β 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 α, β 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 positive, par conséquent cette fonction possède au moins un minimum local $(\hat{\alpha}, \hat{\beta})$, et ce minimiseur est un 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,

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

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

En manipulant ces équations et en introduisant $\bar{x}, \bar{x}, \bar{x}\bar{x}, \bar{x}\bar{y}$ les moyennes empiriques respectives des $x_i, y_i, x_i^2, x_i y_i$, on voit que les conditions de premier ordre deviennent $\alpha + \beta\bar{x} - \bar{y} = 0$, et $\alpha(x_1 + \cdots + x_n) + \beta(x_1^2 + \cdots + x_n^2) - (x_1 y_1 + \cdots + x_n y_n) = 0$, 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 α puis β :

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

Le coefficient β 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. 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**[1].

On cherchera donc des nombres θ_i tels que y_i est aussi proche que possible de

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

oùe paramètre θ, sera toujours vu **comme le vecteur colonne**[2] des θ_i.

Remarque : où est passée la constante ? *Dans l'équation ci-dessus, on a l'impression que le terme constant, qui correspondait à α 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 θ_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 $\mathscr{V}_X = \{X\theta : \ \theta \in \mathbb{R}^d\}$.

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

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

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

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

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 21. 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 $\hat{\theta} - \theta$ est égal à

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

Remarque générale. Si, en dimension 1, on cherche à trouver le θ 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 θ 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 $\hat{\theta}$: 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 deux 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. Comme tous les outils statistiques puissants, la régression linéaire et le coefficient R2 donnent toujours lieu à des utilisations fantastiques : la régression suivante, par exemple, possède un R^2 respectable.

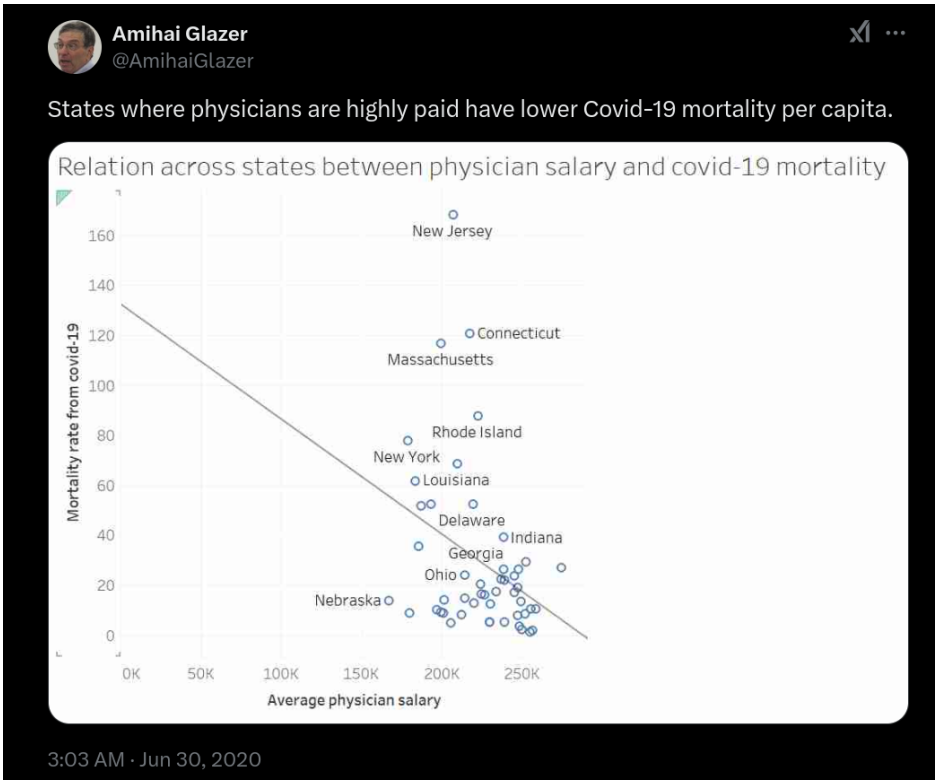


Figure 9.1: Cet ajustement affine est vraiment très très mauvais.

9.4 Théorème de cohérence

On cherche à faire une régression linéaire de Y sur deux ensembles de variables explicatives $X \in \mathbb{R}^{n,d}$ et $Z \in \mathbb{R}^{n,k}$. L'estimateur des MCO qui utilise toutes ces variables est

$$(\hat{\theta}, \hat{\mu}) \in \arg\min_{\theta,\mu} |Y - X\theta - Z\mu|^2. \tag{9.4}$$

Si j'avais omis la variable explicative Z et que j'avais effectué la régression de Y vers X, j'aurais obtenu

$$\hat{\beta} = \arg\min_{\beta} |Y - X\beta|^2.$$

Les deux estimateurs $\hat{\theta}$ et $\hat{\beta}$ capturent l'effet de X sur Y : ils devraient donc être égaux.

C'est faux en général.

Si l'on omet les variables Z, il faut « enlever Z de Y et de X », c'est-à-dire projeter sur l'orthogonal de l'espace des colonnes de Z. On notera Y_Z et X_Z les projections de Y et X sur cet espace. C'est le théorème de *cohérence*, dû à Frish et Waugh : omettre des variables explicatives dans un modèle aboutit nécessairement à un estimateur biaisé si ces variables ne sont pas enlevées du modèle.

Théorème 9.2 (Théorème de Frish-Waugh)**.** *L'estimateur $\hat{\theta}$ obtenu dans la régression Équation 9.4 de Y sur X et Z est égal au $\hat{\theta}_X$ obtenu dans la régression de Y_Z sur X_Z,*

$$\hat{\theta}_X = \arg\min_{\theta} |Y_Z - X_Z\theta|^2.$$

Démonstration. La matrice des variables explicatives dans la régression « totale » est $[X, Z]$, de taille $(n, d+k)$. La formule donnant Équation 9.4 est donc

$$\begin{bmatrix}\hat{\theta}\\ \hat{\mu}\end{bmatrix} = \begin{bmatrix}X^\top X & X^\top Z\\ Z^\top X & Z^\top Z\end{bmatrix}^{-1}\begin{bmatrix}X^\top Y\\ Z^\top Y\end{bmatrix}$$

La formule d'inversion des matrices par bloc (Théorème 22.1) dit que

$$\begin{bmatrix}X^\top X & X^\top Z\\ Z^\top X & Z^\top Z\end{bmatrix}^{-1} = \begin{bmatrix}S^{-1} & -S^{-1}X^\top Z(Z^\top Z)^{-1}\\ * & *\end{bmatrix}$$

où $S = X^\top X - X^\top Z(Z^\top Z)^{-1}Z^\top X$ est le complément de Schur. Par conséquent, $\hat{\theta}$ est donné par

$$S^{-1}X^\top Y - S^{-1}X^\top Z(Z^\top Z)^{-1}Z^\top Y$$

ce qui se factorise en

$$S^{-1}X^\top(I - Z(Z^\top Z)^{-1}Z^\top)Y.$$

On reconnaît la matrice de projection $P = I - Z(Z^\top Z)^{-1}Z^\top$ sur l'orthogonal des colonnes de V (voir Chapitre 21). De même, on voit que S se factorise en

$$X^\top(I - Z(Z^\top Z)^{-1}Z^\top)X$$

ce qui vaut $X^\top PX$; et comme $P^2 = P$, on peut même écrire $S = (PX)^\top(PX)$. Comme $PX = X_Z$ et $PY = Y_Z$ (ce sont de simples notations), on a

$$\hat{\theta} = (X_Z^\top X_Z)^{-1}X^\top Y_Z.$$

C'est bien l'expression de $\hat{\theta}_X$.

□

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 θ et σ^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_{θ,σ^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 11.3 appliqué à notre problème : en effet, $X\hat{\theta}$ est la projection orthogonale de Y sur l'espace vectoriel $\mathrm{Im}(X)$, tandis que $\hat{\varepsilon}$ est la projection orthogonale de Y sur le sous-espace orthogonal à $\mathrm{Im}(X)$.

□

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 σ^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 ε_i sont iid, centrés, et de même variance σ^2 — sous cette dernière hypothèse, on parle de modèle *homoscédastique*.

Sous ces hypothèses, $\hat{\theta}$ est toujours un estimateur sans biais de θ : cela se voit directement en prenant l'espérance de Équation 9.3. De plus, la loi de θ n'est plus gaussienne, mais θ 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] :

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

où Σ_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 Σ_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 23.3). 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 θ ; ici, θ 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 $\kappa_{d,1-\alpha}$ (on notera simplement κ) désigne le quantile d'ordre $1-\alpha$ de cette loi, on en déduit que ξ 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

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

Malheureusement, cette région nécessite de connaître σ. 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 χ_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 $f = f_{d,n-d,1-\alpha}$ est le quantile d'ordre $1-\alpha$ d'une loi $\mathscr{F}_{d,n-d}$, alors la région*

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

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

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

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

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

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

La matrice Σ, 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 $\mathbb{E}[e^{i\langle t, X\rangle}]$ d'un vecteur gaussien $N(\mu, \Sigma)$ est égale à

$$\exp\left\{i\langle t, \mu\rangle - \frac{\langle (t-\mu), \Sigma(t-\mu)\rangle}{2}\right\}. \tag{11.1}$$

Le théorème suivant synthétise toutes les propriétés de stabilité des lois gaussiennes.

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.*

Démonstration. La transformée de Fourier de MX est $\mathbb{E}[\exp(i\langle t, MX\rangle)]$, ce qui vaut également $\mathbb{E}[\exp(i\langle M^\top t, X\rangle)]$. En appliquant la formule Équation 11.1, puis en remettant le $M^\top$ à droite des produits scalaires et en réarrangeant les termes, on obtient

$$\exp\left\{i\langle t, M\mu\rangle - \frac{\langle (t-\mu), M\Sigma M^\top(t-\mu)\rangle}{2}\right\}.$$

C'est bien la transformée de Fourier de $N(M\mu, M\Sigma M^\top)$.

Pour le second point, il suffit de vérifier que la transformée de Fourier $\mathbb{E}[e^{i\langle t, X+Y\rangle}]$ est égale à $\mathbb{E}[e^{i\langle t, X\rangle}]\mathbb{E}[e^{i\langle t, Y\rangle}]$. C'est un simple calcul découlant de Équation 11.1.

□

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 μ en deux blocs

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \tag{11.2}$$

et sa covariance Σ 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*

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

et de covariance

$$\Sigma_{1,1} - \Sigma_{2,1}\Sigma_{2,2}^{-1}\Sigma_{1,2}. \tag{11.3}$$

La formule Équation 11.3 est appelée *complément de Schur*. Son inverse est égal au bloc inférieur-droit de l'inverse de Σ comme expliqué dans Chapitre 22.

Démonstration. Je préfère laisser mon lecteur confiant méditer lui-même sur le fait que la loi conditionnelle d'une partie X d'un vecteur gaussien contre l'autre partie Y du même vecteur est elle-même une loi gaussienne, dont il suffit par conséquent de calculer l'espérance et la variance pour l'identifier. On peut le faire par une technique puissante : on cherche à « enlever Y de X ». Plus précisément, on va essayer d'écrire

$$X = Z + AY$$

où Z est indépendante de Y et A est une matrice à trouver. Comme $\mathrm{Cov}(X, Y) = \Sigma_{1,2}$ et qu'on cherche Z, A tels que $\mathrm{Cov}(Z, Y) = 0$, on voit vite que A doit vérifier $\Sigma_{1,2} = A\Sigma_{2,2}$, et donc que

$$Z = X - \Sigma_{1,2}\Sigma_{2,2}^{-1}Y.$$

Comme Z et Y sont décorrélées et conjointement gaussiennes, elles sont indépendantes, et $\mathbb{E}[Z|Y] = \mathbb{E}[Z] = \mu_1 - A\mu_2$. Ainsi, $\mathbb{E}[X|Y] = \mathbb{E}[Z] - AY$ et on en déduit tout de suite Équation 11.2. Pour la variance conditionnelle, $\mathrm{Var}(X|Y) = \mathrm{Var}(Z) + A\mathrm{Var}(Y)A^\top$ et Équation 11.3 en découle facilement.

□

11.3 Théorème de Cochran

Le fait que dans un modèle gaussien, la quantité $(\bar{X}_n - \mu)/\hat{\sigma}_n$ suive une loi de Student, ou que dans le modèle linéaire gaussien, les résidus sont indépendants de l'estimateur des moindres carrés, sont tous les deux des applications du 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, $\mathbb{E}[\langle X, e\rangle\langle X, f\rangle]$ se développe en

$$\sum_{i,j} e_i f_j \mathrm{Cov}(X_i, X_j)$$

et comme les X_i sont iid, cela vaut $\langle e, f\rangle$. 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 χ_2 indépendantes de paramètres $p = \dim E, q = \dim F$. La loi de leur rapport X/Y apparaît souvent dans des problèmes de 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 $f_{p,q}(x)$ sur $[0, \infty[$ est donnée par*

$$\frac{1}{Z_{p,q}} \frac{\left(\frac{px}{px+q}\right)^{\frac{p}{2}} \left(1 - \frac{px}{px+q}\right)^{\frac{q}{2}}}{x} \tag{11.4}$$

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

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

Le calcul est facile, puisque les lois du χ_2 ont une densité connue donnée par Équation 5.6. Soit φ 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 11.4.

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 θ_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 ℓ_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 θ_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 θ vérifie une équation linéaire. Le test de significativité est un test de contrainte linéaire : en notant δ_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 θ vérifie l'identité

$$C\theta = c$$

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, on voit que l'expression

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

suit une loi $\chi_2(r)$. Maintenant, si l'on estime le terme σ^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 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

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

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 d'un test de contraintes linéaires au sens du paragraphe précédent : il y a $d-1$ 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 $C\theta = 0$. Dans la statistique de Wald associée à cette contrainte, la matrice $C(X^\top X)^{-1}C^\top$ est le bloc B_X obtenu à partir de $(X^\top X)^{-1}$ en enlevant la première ligne et la première colonne (qui correspondent à la constante). Il se trouve que ce test possède une expression plus simple.

Théorème 12.3.

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

Démonstration. Il existe une formule qui permet d'inverser des matrices par blocs : la formule de Schur (Théorème 22.1). Commençons par décomposer X sous la forme[1] $[\mathbf{1}, X_0]$, avec X_0 de taille $n \times (d-1)$ et $\mathbf{1}$ le vecteur constant égal à 1. La matrice $X^\top X$ vaut

$$\begin{bmatrix} n & u \\ v & X_0^\top X_0 \end{bmatrix}$$

où u est le vecteur ligne $\mathbf{1}^\top X_0$ et v est le vecteur colonne $X_0^\top \mathbf{1}$. La formule de Schur dit que l'inverse de cette matrice est

$$\begin{bmatrix} * & * \\ * & (X_0^\top X_0 - vu/n)^{-1} \end{bmatrix}$$

La matrice uv/n s'écrit $X_0^\top E X_0$, où $E = \mathbf{1}\mathbf{1}^\top/n$ est la projection orthogonale sur le sous-espace vectoriel engendré par $\mathbf{1}$. Comme $(I-E)^2 = I-E$, on peut écrire $(X_0^\top X_0 - vu/n)^{-1}$ sous la forme $X_1^\top X_1$, où $X_1 = (I-E)X_0$. La statistique de Wald s'écrit alors

$$\frac{\langle \hat\theta_1, X_1^\top X_1 \hat\theta_1 \rangle}{(d-1)\hat\sigma_n^2} \tag{12.2}$$

où $\hat\theta_1 = C\hat\theta$ est composé des entrées de $\hat\theta$ sauf la première. On reconnaît, en haut, la norme $|X_1\hat\theta_1|^2$.

Il suffit de montrer que l'expression dans Équation 12.2 est égale à $(n-d)R^2/(d-1)(1-R^2)$. On rappelle que $R^2 = 1 - |\hat Y - Y|^2/|\bar y\mathbf{1} - Y|^2$. L'expression $R^2/(1-R^2)$ devrait donc valoir

$$\frac{|\bar y\mathbf{1} - Y|^2 - |\hat Y - Y|^2}{|\hat Y - Y|^2}.$$

Au vu de Équation 12.2 et du fait que $\hat\sigma_n^2 = |\hat Y - Y|^2/(n-d)$, il suffit de montrer que $|X_1\hat\theta_1|^2$ et $|\bar y\mathbf{1} - Y|^2 - |\hat Y - Y|^2$ sont égales.

Le vecteur $\bar y\mathbf{1}$ est la projection orthogonale de Y sur $\mathbf{1}$. Ainsi, on a

$$Y = \hat\varepsilon + \bar y\mathbf{1} + (\hat Y - \bar y\mathbf{1})$$

et ces trois vecteurs sont orthogonaux. Le théorème de Pythagore dit que $|\bar y\mathbf{1} - Y|^2 - |\hat Y - Y|^2$ est égal à $|\hat Y - \bar y\mathbf{1}|^2$. On reconnaît la projection orthogonale de $Y - \bar y\mathbf{1} = (I-E)Y$ sur l'espace vectoriel engendré

[1] on rappelle que $(XY)^{-1} = Y^{-1}X^{-1}$.

par les colonnes de la matrice $X_1 = (I - E)X_0$. Si l'on note $\tilde{\theta}_1$ l'estimateur des MCO de la régression de $(I - E)Y$ vers $(I - E)X_0$, il suffit donc de montrer que $\tilde{\theta}_1 = \hat{\theta}_1$ pour terminer la démonstration. C'est exactement le théorème de Frish-Waugh Théorème 9.2.

□

Exercices

Questions

- Retrouver la formule des MCO par une méthode analytique.
- Construire un intervalle de confiance de niveau $1-\alpha$ pour le coefficient θ_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 $n \to \infty$.
- Réfléchir au tweet suivant : random points ALWAYS produce a slope.

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 ℓ^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 θ ne soient pas trop grands au sens ℓ^2.

1. Montrer que la solution au problème est donnée par $\hat{\theta}_\lambda = (X^\top X + \lambda I_d)^{-1} X^\top Y$ sans aucune contrainte de rang sur X.
2. Calculer la loi de $\hat{\theta}_\lambda$ 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 (sans constantes).

Exercice 12.5 (Théorème de Frish-Waugh)**.**

1. À quelle condition sur les variables X, Z a-t-on $\hat{\theta}_X = \hat{\theta}$?
2. Montrer que centrer les variables explicatives d'un modèle revient à enlever la constante.

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 σ^2 connue. Proposer un test simple de l'hypothèse nulle.
2. Même question lorsque σ 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 ε_i ne sont plus iid, mais possèdent une covariance Σ non scalaire.

1. On suppose que l'on connaît Σ. Montrer que

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

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

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

   est un estimateur asymptotiquement normal et sans biais de θ.

Exercice 12.8. On considère le modèle de régression linéaire $y_i = b_0 + b_1 x_i + \varepsilon_i$ où $i = 1, \dots, n$ et les ε_i sont des variables aléatoires indépendantes $\mathcal{N}(0, \sigma^2)$ et b_0, b_1 et σ^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 $Y_0 = b_0 + b_1 x_0 + \eta$, avec η une erreur d'observation de loi $N(0, \sigma^2)$ indépendante des ε_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 σ, 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 σ^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 θ qui soit sans biais[2]. 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.

Exercice 12.10 (Régression polynomiale (examen 25-bis))**.** On dispose de n variables $x_1, \dots, x_n \in \mathbb{R}$, et n variables à expliquer $y_1, \dots, y_n \in \mathbb{R}$. L'objectif est de trouver le meilleur polynôme possible de degré d qui relie les x_i aux y_i.

1. Soit $\mathscr{F}_d$ l'ensemble des polynômes à coefficients réels, de degré inférieur ou égal à d. On veut trouver $f \in \mathscr{F}_d$ tel que
   $$f \in \arg\min_{g \in \mathscr{F}_d} \sum_{i=1}^{n} (y_i - g(x_i))^2.$$
   Exprimer ce problème de minimisation comme un problème de moindres carrés et exprimer la solution f en fonction des x_i et des y_i (on pourra faire des hypothèses raisonnables sur les x_i).
2. On suppose qu'il existe un polynôme $f_\star$ de degré au plus d tel que $y_i = f_\star(x_i) + \varepsilon_i$ où les ε_i sont des variables aléatoires indépendantes de loi $N(0, \sigma^2)$, avec σ^2 une variance inconnue. On note $\widehat{f}$ l'estimateur trouvé à la question précédente. Quelle est la loi des coefficients de $\widehat{f}$? Donner aussi l'estimateur de σ^2 et sa loi.
3. Tester l'hypothèse selon laquelle $f_\star$ est en réalité de degré strictement inférieur à d.
4. Maintenant, on suppose que les x_i sont de dimension 2. On souhaite faire une régression polynomiale de degré d comme ci-dessus des x_i vers les y_i. Quelle est la dimension de cette nouvelle régression ? Tester l'hypothèse selon laquelle les y_i dépendent seulement des $|x_i|^2$ (difficile).

[2]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 θ, et où Z est une constante qui ne dépend que de θ. Ces modèles appartiennent à la famille des *modèles exponentiels*.

13.1 Définitions

Soit ν une mesure de référence (σ-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 ν : 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 ν) 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

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

est appelé *log-vraisemblance*. Lorsqu'il est bien défini, le gradient en θ de la log-vraisemblance

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

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 ν 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 θ 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 ν 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 Θ (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

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

et sa matrice hessienne[1] *par*

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

Démonstration. 1. Prenons un petit δ tel que $\theta + \delta$ et $\theta - \delta$ sont dans Θ. 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 δ dans une boule autour de θ, 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 ∇^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 θ et son intégrale est finie dès que θ est dans Θ. Son gradient en θ 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 δ dans un voisinage de θ, 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 θ.
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 μ 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_θ-presque sûrement, disons égale à un certain α, et donc que ν-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_θ 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_θ 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 ν-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 d_{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 ν 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 ℓ ou sur Θ 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 θ tend vers le bord de Θ, 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.

Le maximum de vraisemblance ("maximum likelihood") est un bon prétendant au titre de concept le plus important des statistiques modernes. Son histoire est intéressante : voir cet article.

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 η est le *pas* de la montée de gradient.

14.4 Exemple d'EMV

Pour illustrer le propos, regardons l'exemple classique de l'estimation de μ 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.

14.5 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 Θ_0, Θ_1 sont deux régions distinctes de l'espace des paramètres Θ. Dans ces cas, nous pouvons définir deux maximums de vraisemblance, un par hypothèse : par exemple,

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

Dans le cas où les régions Θ_0, Θ_1 sont constituées d'un seul élément, disons θ_0, θ_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 14.2. 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.

14.6 Limitations de l'EMV

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, que l'on verra dans le chapitre suivant) 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** μ : 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 θ est grande comme en deep learning. L'estimation par maximum de vraisemblance sera alors quasiment infaisable.

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 θ qui est « contenue » dans la distribution p_θ.

Définition 15.1 (Information de Fisher dans les modèles exponentiels)**.** Dans le modèle exonentiel associé à T, 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 :

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

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

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

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, l'information de Fisher coïncide avec la variance du score. C'est cette dernière définition qu'on retient en général, car elle n'est pas propre aux modèles exponentiels.

Définition 15.3 (Information de Fisher dans les modèles généraux)**.** Dans un modèle statistique général $(p_\theta)_{\theta\in\Theta}$, l'information de Fisher est définie comme la variance du score :

$$I(\theta) = \mathrm{Var}_\theta(\nabla_\theta \ln p_\theta(X)),$$

pourvu que tout soit bien défini (la densité doit être dérivable, etc.).

On rappelle que le gradient de la fonction $\ln p_\theta : \mathbb{R}^d \to \mathbb{R}$ est vu comme un vecteur colonne. Par conséquent, la variance de la variable aléatoire $\nabla_\theta \ln p_\theta(X)$ (qui est centrée) est la matrice

$$\mathbb{E}_\theta[\nabla_\theta \ln p_\theta(X) \nabla_\theta \ln p_\theta(X)^\top].$$

15.2 Borne de Cramér-Rao

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

Lorsque le paramètre θ 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.

La borne $I(\theta)$ est très petite lorsque l'information est très grande, ce qui est intuitif : si les observations portent beaucoup d'information sur le paramètre, le risque quadratique du meilleur estimateur peut certainement être très petit. Inversement, si $I(\theta)$ est très grande, cela signifie que les observations ne portent pas beaucoup d'information sur le paramètre, et donc que le risque quadratique sera nécessairement assez grand.

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

□

15.3 Interprétation

L'information de Fisher d'un modèle dépend crucialement de *comment* on paramétrise le modèle, et c'est ce qui a donné lieu à la terrible confusion qui m'a paralysé en cours jeudi matin.

Prenons l'exemple de la loi de Bernoulli, et considérons d'abord sa paramétrisation usuelle,

$$p_\theta(x) = \theta^x(1-\theta)^{1-x}$$

avec $x \in \{0,1\}$, et θ est dans $]0,1[$. La fonction de score est donnée par $(x\ln\theta + (1-x)\ln(1-\theta))'$, soit $\frac{x}{\theta} - \frac{1-x}{1-\theta}$, ou encore

$$\frac{x}{\theta(1-\theta)} + \frac{1}{1-\theta},$$

[1] Rappelons que (cf Section 21.4) 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.

et donc l'information de Fisher de ce modèle est égale à la variance de $X/\theta(1-\theta) + 1/(1-\theta)$, qui vaut $1/\theta(1-\theta)$:

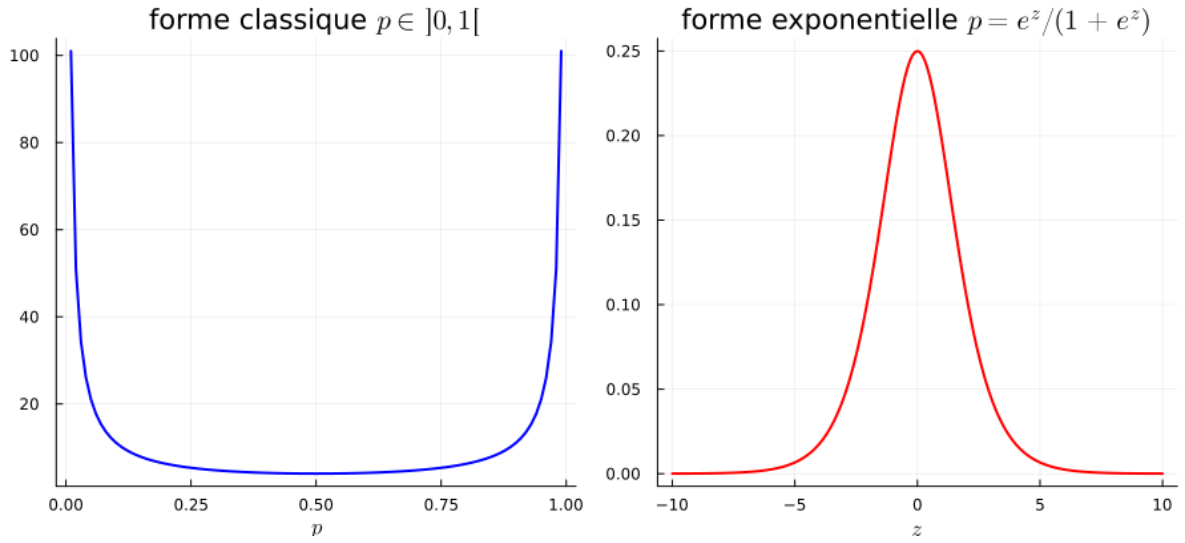
$$I(\theta) = \frac{1}{\theta(1-\theta)}. \tag{15.2}$$

Maintenant, nous allons paramétriser le modèle de Bernoulli avec les logits : plutôt que de considérer p, nous allons paramétriser $z = \ln(\theta/(1-\theta))$. La densité de Bernoulli s'écrit alors

$$p_z(x) = \frac{e^{zx}}{1+e^z}.$$

Comme x vaut 0 ou 1, on voit que la probabilité d'obtenir 1 est égale à $\theta = \theta(z) = e^z/(1+e^z)$. Il s'agit donc d'une autre paramétrisation de p, mais celle-ci a l'avantage d'exprimer les lois de Bernoulli sous forme exponentielle, avec $T(x) = x$. L'information de Fisher dans cette autre paramétrisation du modèle de Bernoulli est donnée par la formule $\mathrm{Var}_z(X) = \frac{e^z}{(1+e^z)^2}$, soit

$$\tilde{I}(z) = \theta(z)(1-\theta(z)). \tag{15.3}$$



Les formules Équation 15.2 et Équation 15.3 (qui sont bien celles que j'ai obtenues en cours ce matin) sont toutes les deux correctes, mais elles ne coïncident pas. Pourquoi ?

La réponse est que $I(\theta)$ et $\tilde{I}(z)$ quantifient la quantité d'information **sur le paramètre estimé** qui et contenue dans les observations, et que le paramètre n'est pas le même dans les deux cas.

- pour le modèle Équation 15.2, le paramètre est θ. Dans ce cas, l'information contenue dans les observations est maximale lorsque θ est proche de 0 ou 1.

- pour le modèle Équation 15.3, le paramètre est z. Dans ce cas, c'est le contraire : on voit que l'information contenue dans les observations est maximale lorsque z est proche de 0, ce qui correspond à $\theta = 1/2$.

Considérons les deux modèles en parallèle, dans le régime où la probabilité de succès est très proche de 1, c'est-à-dire θ proche de 1 ou z très grand[2].

[2] à titre d'exemple, toute valeur de z plus grande que 10 donne un $\theta(z)$ plus grand que 99.99%.

- dans le premier cas, une infime variation de θ peut entraîner une grande variation du modèle : par exemple, passer de 99% à 99.999%(une variation de moins de 1%) se verra tout de suite sur les observations.
- dans le second cas, même de grandes variations de z n'auront aucune incidence sur les données. Passer de $z = 10$ à $z = 100$ revient à passer d'une probabilité de succès de 99.99% à 99.99999%.

Autrement dit, la paramétrisation a radicalement changé l'information, parce que le paramètre dans chacun des deux modèles n'est pas le même et ne représente pas la même chose.

Exercices

Questions

- On dispose d'une observation x d'une variable aléatoire $N(\mu, 1)$. Quel est l'EMV de μ ?
- On dispose d'une observation x d'une variable aléatoire $\mathrm{Ber}(p)$. Quel est l'EMV de μ ?
- Calculer l'EMV de μ et σ^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 θ. Quel est l'EMV pour σ^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 $p-1$) 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[3]. des x_i Montrer que l'EMV (lorsqu'il existe) est l'unique θ 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[4] 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 θ dans un modèle statistique. Montrer que $f(\hat{\theta})$ est l'estimateur du maximum de vraisemblance de $f(\theta)$, pour n'importe quelle fonction θ 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 λ 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.

[3] C'est-à-dire la mesure de probabilité définie par $n^{-1} \sum \delta_{x_i}$.

[4] 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 λ ni m ne sont connus.

Exercice 15.3. Soit (p_θ) un modèle statistique, avec θ dans un ouvert U. Soit ϕ un difféomorphisme de V vers un ouvert U. On considère le modèle statistique $(p_{\phi(\mu)})$, où μ est dans V. On note $I_1(\theta)$ la matrice d'information de Fisher du modèle (p_θ), et $I_2(\mu)$ celle du modèle $(p_{\phi(\mu)})$.

1. Montrer que $I_2(\mu) = J_\phi(\mu) I(\phi(\mu)) J_\phi(\mu)^\top$, où $J_\phi(\mu)$ est la matrice jacobienne de ϕ en μ.
2. Retrouver les formules pour l'information de Fisher des lois de Bernoulli dans la section Section 15.3.

Exercice 15.4. 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 $\mathcal{N}(\mu, \sigma^2)$.
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 15.5. On se donne un échantillon $(X_1, \dots, X_n)$ de loi $\Gamma(\alpha, \beta)$[5].

1. On suppose le paramètre β connu. Proposer un estimateur de α par la méthode des moments.
2. On suppose à présent que les deux paramètres α, β sont inconnus. Proposer un estimateur de (α, β) par la méthode des moments.
3. Toujours dans le cas où α, β sont inconnus, donner le système d'équation que satisfont les estimateurs de (α, β) par la méthode du maximum de vraisemblance.

Exercice 15.6. 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 θ obtenu par méthode des moments. Est-il consistant? Proposer un intervalle de confiance asymptotique de niveau de confiance α.
3. Soit T_n l'estimateur du maximum de vraisemblance de θ. 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 α.
4. Comparer les estimateurs $\widehat{\theta}_n$ et T_n sur la base des longueurs moyennes des intervalles de confiance asymptotiques associés.

Exercice 15.7. 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 λ.
2. Calculer son risque quadratique.

[5] On rappelle que sa densité est $\frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x)$ sur $[0, \infty[$

Exercice 15.8. Dans une urne contenant 1000 tickets, 20 sont marqués θ et 980 sont marqués 10θ, où θ 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 σ-finie? Donner un estimateur qui s'apparenterait à un maximum de vraisemblance $\widehat{\theta}$ de θ (maximiser $P_\theta(\{X\})$), puis montrer que $\mathbb{P}(\widehat{\theta} = \theta) \geq 0,98$.
2. On renumérote les tickets marqués 10θ 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 $\widetilde{\theta}$ et montrer que $\mathbb{P}(\widetilde{\theta} < 10\theta) = 0,02$.

Exercice 15.9 (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 $Y_i \sim \mathrm{Ber}(p(x_i))$ 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 θ.
   ii) Écrire l'équation vérifiée par l'EMV de θ.
   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 Φ 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 ℓ_k. Étant donné un code, le nombre moyen de bits d'information nécessaire pour coder une variable aléatoire X sur $\mathcal{X}$ est donc

$$\mathbb{E}[\ell_X] = \sum_{i=1}^{k} \ell_i \mathbb{P}(X = a_i).$$

Exemple 16.1. Un roman contient environ 80000 mots, donc environ 400000 signes (en comptant les espaces). Un signe peut être l'une des 26 lettres de l'alphabet (en majuscule ou minuscule), un signe de ponctuation, un chiffre, une lettre accentuée… La plupart des romans occidentaux ont besoin d'un alphabet de ~200 lettres. Le codage binaire basique lettre-par-lettre nécessiterait donc environ $\lceil \log_2 200 \rceil \approx 8$ bits d'information par lettre, soit 3200000 bits. On compte plutôt en paquet de 8 bits, appelés bytes : le codage du roman nécessiterait donc environ 400000 bytes soit 400kB.

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 $\ln(2)$.

Le théorème de codage suivant ne s'applique qu'aux variables discrètes.

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

$$\lim \frac{\mathbb{E}[\ell_n^\star]}{n} = \mathrm{Ent}(P).$$

L'entropie d'une loi est donc en le nombre moyen de bits d'information nécessaire à son codage le plus économe : asymptotiquement, ce codage est le *codage entropique* qui assigne à chaque $a \in \mathcal{X}$ un code de longueur proche de $\log_2(\mathbb{P}(X = a))$. Cependant, il existe d'autres codages qui sont asymptotiquement meilleurs que le codage ci-dessus, et qui sont meilleurs même à distance finie : par exemple, le codage de Huffman.

Démonstration. On énumère les $n > 2$ éléments de $\mathcal{X}$ en $a_1, \dots, a_n$, et on note $p_i = P(X = a_i)$. Soit ℓ un codage et soit ℓ_i la longueur du code $\ell(a_i)$. La longueur moyenne de ce code est

$$E[\ell_X] = \sum_{i=1}^{n} p_i \ell_i.$$

□

Ainsi, $E[\ell_X] - \mathrm{Ent}(P)$ vaut

$$\sum_{i=1}^{n} p_i \left(\ell_i + \log_2(p_i)\right)$$

c'est-à-dire, après une petite manipulation,

$$-\sum_{i=1}^{n} p_i \log_2 \left(\frac{1}{p_i 2^{\ell_i}}\right).$$

La fonction $-\log_2$ étant convexe, l'inégalité de Jensen dit que la quantité ci-dessus est plus grande que

$$-\log_2 \left(\sum_{i=1}^{n} p_i \frac{1}{p_i 2^{\ell_i}}\right)$$

qui se simplifie en

$$-\log_2 \left(\sum_{i=1}^{n} \frac{1}{2^{\ell_i}}\right).$$

Il suffit alors de montrer que la somme s à l'intérieur du $\log_2$ est plus petite que 1 : c'est cette inégalité qui est connue sous le nom de *lemme de Kraft*. Elle entraîne immédiatement que $-\log_2(s) \geqslant 0$, et donc $E[\ell_X] \geqslant \mathrm{Ent}(P)$.

Démonstration du lemme de Kraft. Mettons la constante s à une puissance entière k : en écrivant s^k comme

$$\left(\sum_{a \in \mathcal{X}} \frac{1}{2^{\ell(a)}}\right)^k$$

et en développant, on trouve

$$\sum_{a^1, \dots, a^k \in \mathcal{X}} \frac{1}{2^{\ell(a^1) + \cdots + \ell(a^k)}}$$

c'est-à-dire

$$\sum_{A\in\mathcal{X}^k}\frac{1}{2^{\ell(A)}}.$$

Notons c_i le nombre de mots de k lettres dont le code est lui-même de longueur i. La quantité précédente vaut

$$\sum_{i=1}^{+\infty} c_i\frac{1}{2^i}.$$

Vu que le code est injectif, deux mots différents ont deux codes différents : or, il y a 2^i codes possibles de longueur i, donc $c_i \leqslant 2^i$. De plus, si N est la longueur de la lettre la plus longue à coder, il est clair que le mot le plus long à coder aura une longueur de kN. Ainsi, c_i devient nul dès que $i > kN$. On en déduit que la somme ci-dessus est plus petite que

$$\sum_{i=1}^{kN} 2^i\frac{1}{2^i} = kN.$$

On vient de montrer que, pour n'importe quel nombre entier k non nul,

$$s^k \leqslant kN.$$

Si le nombre s était strictement plus grand que 1, le terme de droite aurait une croissance exponentielle en k, tandis que le terme de gauche n'aurait qu'une croissance linéaire. C'est impossible, et on doit donc avoir $s \leqslant 1$.

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 » $\log_2 Q$: d_{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.

[1] À savoir $\ln(2)$.

16.4 Information de Fisher et entropie

Dans le cas général d'une loi de densité f par rapport à une mesure de référence ν, l'entropie est donnée par

Définition 16.3 (Entropie)**.** L'entropie d'une loi de densité f par rapport à la mesure de référence ν est donnée par

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

Lorsque ν est la mesure de comptage sur un ensemble fini ou discret $\mathcal{X}$, on retrouve la formule de l'entropie discrète ci-dessus.

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

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

□

Comme dans ce cadre, l'entropie est concave, la courbure en θ est un indicateur de à quel point l'entropie forme un « pic » autour de θ.

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 cherche 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**.

Le théorème de Boltzmann-Gibbs doit que les lois d'entropie maximale vérifiant des contraintes de moyennes sont *exactement* 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 α, β sont réels. En réalité, β doit être positif sinon ce n'est pas une loi de probabilité. De plus, il est facile de voir que si α 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 $H(p)$ est le nombre de $[0, \infty]$ défini par

$$H(p) = -\sum_{i=1}^{n} p_i \ln p_i.$$

Ce nombre vaut $+\infty$ si et seulement si l'un des p_i est nul. Le fait que p soit une loi de probabilité se traduit en pratique par des contraintes supplémentaires, à savoir $s(p) = 0$ où

$$-1 + \sum_{i=1}^{n} p_i = 0.$$

[1] Après discussion avec mes collègues, il y a consensus sur la nécessité d'utiliser un critère de compacité L1 de type Dunford-Pettis.

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 sera supposée non vide (sinon, le problème n'a pas de solution) ; de plus, la fonction H est concave. En effet, sa matrice hessienne au point p est égale à $\mathrm{diag}(-1/p)$, qui est bien définie négative. Le problème Équation 17.1 possède donc une solution. Pour la trouver, on utilise les outils classiques 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 λ et un certain μ. Comme les p_i somment à 1, le nombre μ 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 λ qui vérifie

$$E_\lambda[T(X)] = c.$$

L'existence d'un tel λ 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.

[2] Si l'un des p_i est nul, $H(p) = +\infty$, donc on peut se restreindre aux $p_i > 0$.
[3] Penser à la contrainte absurde $E_\lambda[X^2] = -1$.

Problèmes et sujets

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

$$\begin{aligned} H_0 &: \quad X_i = Y_i \text{ en loi pour tout } i, \\ H_1 &: \quad \text{il existe } i \neq j \text{ tels que } X_i \neq Y_i \text{ en loi.} \end{aligned}$$

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 :

$$\frac{n}{15} \left(T_n^{(3)} \right)^2 + \frac{n}{96} \left(T_n^{(4)} - 3 \right)^2 \to \chi_2^2$$

[4] Le quantile d'ordre 0.975 d'une $\mathscr{B}(12, 0.5)$ est 9.

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 ?

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 χ_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.

Sujets passés d'examens

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
- Partiel 2024 avec sa correction intégrée, assez proche de ce que j'attend de vous pour le partiel 2025.
- Examen 2024 avec sa correction.
- Second examen 2024, avec sa correction.

[5] Loi hypergéométrique.

- Partiel 2025 avec aussi sa correction intégrée.

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

La fonction de répartition empirique des X_i est

$$F_n(t) = \frac{1}{n}\#\{i : X_i \leqslant t\}.$$

La loi des grands nombres montre 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,* $\lim_{n\to\infty} |F_n - F|_\infty = 0$.

Ce théorème dit que F_n est un estimateur « convergent au sens de la norme uniforme » de F. On va utiliser ce théorème pour faire des tests : 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 $X_{(0)} = -\infty$. La quantité $|F_n - F|_\infty$ est calculable grâce à la représentation suivante :

$$|F_n - F|_\infty = \sup_{i\in\{0,\dots,n-1\}} \left|\frac{i}{n} - F(X_{(i)})\right| \vee \left|\frac{i}{n} - F(X_{(i+1)})\right|. \tag{18.1}$$

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.

[1] On rappelle que $|g|_\infty = \sup_{x\in\mathbb{R}} |g(x)|$.

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)**.**

1) $\sqrt{n}|F_n - F|_\infty$ *converge en loi vers* $|B|_\infty$*, où* $(B_t)_{t\in[0,1]}$ *est un pont Brownien standard*[1]*. 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*

$$1 - 2\sum_{k=0}^{\infty} (-1)^k e^{-2x^2(k+1)^2}.$$

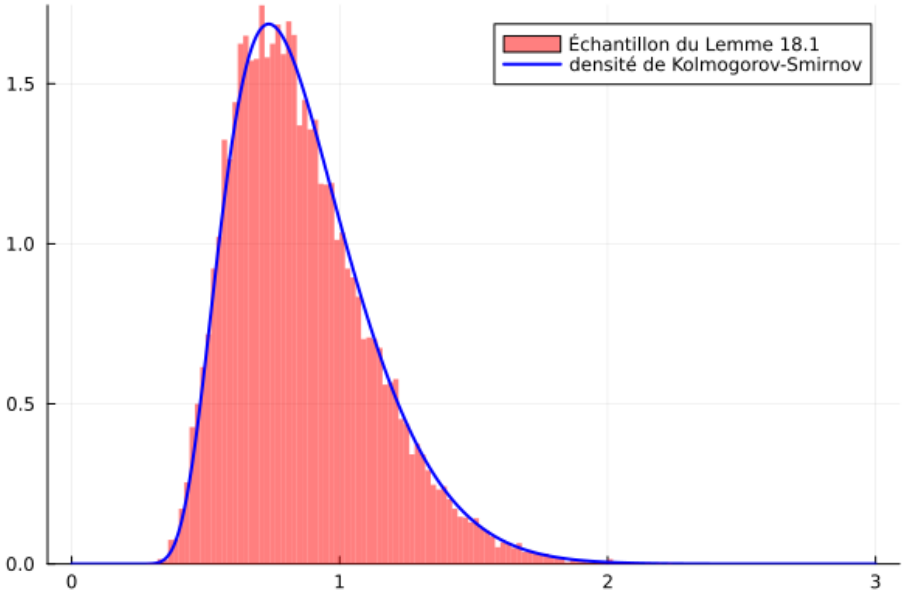
2) Si $X_1, \dots, X_n$ *suivent une loi de fonction de répartition* $G \neq F$*, alors* $\sqrt{n}|F_n - F|_\infty \to \infty$ *presque sûrement.*

Démonstration. La démonstration du premier point nécessite des outils plus avancés en théorie des probabilités (théorème de Donsker).

Pour le second point, si $G \neq F$, alors il existe un x tel que $G(x) \neq F(x)$, donc $|F(x) - G(x)| = c > 0$. Or, par la loi des grands nombres, $F_n(x)$ converge vers $G(x)$ presque sûrement. Comme $|F_n - F|_\infty$ est plus grand que $|F_n(x) - G(x)|$ qui converge vers $c > 0$, on voit que dès que n est assez grand, $\sqrt{n}|F_n - F|_\infty$ est plus grand que $\sqrt{n}c/2$ et donc tend vers ∞.

□

[1] on rappelle que $(XY)^{-1} = Y^{-1}X^{-1}$.



Ce théorème permet de construire des tests aymptotiques de niveau exactement $1-\alpha$ en utilisant les quantiles de la loi de Kolmogorov-Smirnov, qui ont été tabulés. Je note q_β le quantile usuel, $\mathbb{P}(\mathrm{KS} < q_\beta) = \beta$. L'essentiel de la masse de cette loi est supportée sur l'intervalle $[0.5, 2]$.

β	0.1%	1%	95%	98%	99%	99.99%
q_β	0.44	0.57	1.36	1.52	1.63	2.22

Exercices

Rien pour l'instant.

20 Fisher VS Bayes

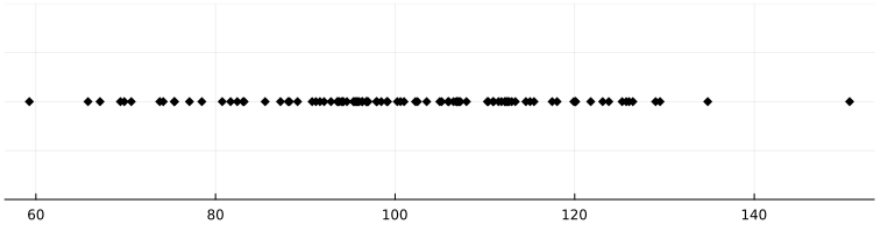
Les test d'intelligence comme le QI ont été conçus pour être standardisés : la répartition du QI d'un test sur une large population ressemble à une gaussienne centrée en 100, et d'écart-type 15. Cela signifie que si l'on prend un individu au hasard, la probabilité que son QI soit entre 85 et 115 est de 68%.

Dans une population de $n = 1000$ individus, on note x_i le QI de la personne i. Pour mesurer les x_i, on décide de faire passer un test à chaque personne : le résultat de ce test, y_i, est une estimation de x_i : n se doute bien qu'un seul test ne suffit pas à déterminer exactement x_i. Cependant, il est raisonnable de penser que y_i est centrée en x_i, et d'écart-type pas trop grand, disons 5.

20.1 Le problème du point de vue de Fisher

En statistique *fréquentiste*, les y_i sont des réalisations de variables iid $N(x_i, 5)$ et on cherche à estimer x_i. Toutes les méthodes vues en cours indiquent que dans ce cas, l'estimateur de référence de x_i est y_i.

Imaginons que les tests y_i aient une répartition comme suit :



La personne "la plus intelligente" a obtenu $y_i = 147$. On peut estimer son QI à 150, mais il y a quand même un doute : dans la population, le QI x_i est sensé être à peu distribué selon $N(100, 15^2)$. Un QI de 147 représente à peu près une déviation de 3.3 écarts-types : un événement de cet ordre a une probabilité d'environ 0.1%. La probabilité pour que parmi 100 personnes, au moins une ait un QI plus grand que 145 est donc $1 - 0.001^{100} \approx 0.3 \approx 10\%$. C'est peu, sans être improbable.

Mais il y a une autre possibilité : peut-être que la personne i a un QI plus proche de 140, mais que ce jour-là, elle a eu de la chance. En fait, la probabilité pour qu'une personne ait un x_i qui dévie de plus de 2 écarts-types (donc, $x_i > 130$) est de 5%, et la probabilité pour que le jour du test, cette personne dévie de 2 écarts-types (donc $y_i - x - i > 20$), est encore de 5%. Cela fait une probabilité d'environ 0.25%.

Lemme 20.1. *Il est plus probable que la personne "la plus intelligente" de la salle soit 1) assez intelligente avec un x_i de 130, et 2) qu'elle ait eu de la chance le jour du test - en tout cas, c'est 2 à 3 fois plus probable que le fait qu'elle ait un QI x_i supérieur à 150.*

La théorie statistique "à la Fisher" comme nous l'avons vue dans ce cours ne permet pas d'intégrer une connaissance *a priori* sur le paramètre μ à estimer. Dans notre cas, la connaissance a priori qu'on a sur les μ_i est qu'ils sont eux-mêmes aléatoires, distribués selon $N(100, 15^2)$: il est donc presque impossible qu'un des μ_i soit plus grand que, disons, 200.

La statistique bayésienne permet d'intégrer cette connaissance a priori et de *corriger* l'estimation naïve de Fisher.

20.2 Estimation bayésienne

20.2.1 Formule de Tweedie

Théorème 20.1 (Formule de Tweedie)**.** *Soit X une variable aléatoire de densité ϱ et soit $\varepsilon \sim N(0, \sigma^2)$. On pose $Y = X + \varepsilon$, qui est une version bruitée de X. Si f est la densité de Y, alors*

$$\mathbb{E}[X|Y] = Y + \sigma^2 \nabla_y \ln f(Y).$$

Démonstration. La loi jointe du couple (X, Y) est $\varrho(x)g_\sigma(y-x)$, où g_σ est la densité de $N(0, \sigma^2)$. La loi de Y est la convolution $\varrho * g_\sigma = \int g(x)g_\sigma(y-x)dx$. Enfin, la densité conditionnelle de X sachant Y est donnée par la formule de Bayes,

$$\frac{\varrho(x)g_\sigma(y-x)}{\int g(x)g_\sigma(y-x)dx}.$$

L'espérance conditionnelle $\mathbb{E}[X|Y=y]$ vaut donc

$$\int \frac{x\varrho(x)g_\sigma(y-x)}{\int g(x)g_\sigma(y-x)dx}.$$

Dans l'intégrale du haut, on peut artificiellement écrire $x = x - y + y$ afin d'obtenir

$$\int \frac{(x-y)\varrho(x)g_\sigma(y-x)dx}{\int g(x)g_\sigma(y-x)dx} + y\int \frac{\varrho(x)g_\sigma(y-x)dx}{\int g(x)g_\sigma(y-x)dx}.$$

Le second terme est égal à y, et en dérivant sous l'intégrale, on voit que le premier est égal à $\partial_y \ln(\varrho * g_\sigma)(y)$.

□

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)
- The Epic Story of Maximum Likelihood
- 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
- Information THeory, même chose.

21 Algèbre linéaire

21.1 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 ℓ 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 ℓ 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 ℓ, 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

$$X^\top X = \sum_{i=1}^{n} \mathbf{x}_i^\top \mathbf{x}_i.$$

21.2 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 21.1. *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*

$$M = \sum_{i=1}^{n} \lambda_i u_i u_i^\top.$$

21.3 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 21.2. *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.

□

21.4 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 21.1. *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 21.1. 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.

22 Formules d'inversion

22.1 Inversion par blocs

On considère une matrice M qui s'écrit par blocs de la façon suivante :

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{22.1}$$

Les formules de Schur permettent d'écrire l'inverse de la matrice M, s'il existe, en fonction des blocs. Essayons d'inverser la matrice M par un pivot de Gauss. Si D est inversible, en multipliant les colonnes des deux blocs à droite par $-D^{-1}C$ et en les enlevant aux colonnes des blocs de gauche, on arrive à éliminer le bloc C :

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix} \times \begin{bmatrix} I & 0 \\ -D^{-1}C & I \end{bmatrix} = \begin{bmatrix} A - BD^{-1}C & B \\ 0 & D \end{bmatrix}. \tag{22.2}$$

On voit que le terme $A - BD^{-1}C$ va jouer un rôle important, puisque s'il est inversible, alors M le sera aussi.

Définition 22.1. Le complément de Schur est

$$S = A - BD^{-1}C.$$

Continuons le pivot de Gauss. En faisant la même chose pour éliminer le bloc B par manipulation des lignes, on voit que

$$\begin{bmatrix} I & -BD^{-1} \\ 0 & I \end{bmatrix} \times \begin{bmatrix} S & B \\ 0 & D \end{bmatrix} = \begin{bmatrix} S & 0 \\ 0 & D \end{bmatrix}.$$

On a donc démontré que

$$\begin{bmatrix} I & -BD^{-1} \\ 0 & I \end{bmatrix} \times \begin{bmatrix} A & B \\ C & D \end{bmatrix} \times \begin{bmatrix} I & 0 \\ -D^{-1}C & I \end{bmatrix} = \begin{bmatrix} S & 0 \\ 0 & D \end{bmatrix}$$

L'inverse d'une matrice de la forme

$$\begin{bmatrix} I & X \\ 0 & I \end{bmatrix}$$

est

$$\begin{bmatrix} I & -X \\ 0 & I \end{bmatrix}$$

donc on obtient la décomposition

$$M = \begin{bmatrix} I & BD^{-1} \\ 0 & I \end{bmatrix} \times \begin{bmatrix} S & 0 \\ 0 & D \end{bmatrix} \times \begin{bmatrix} I & 0 \\ D^{-1}C & I \end{bmatrix}.$$

Enfin, pour inverser M il suffit d'inverser chacune de ces matrices et d'inverser l'ordre[1], donc

$$M^{-1} = \begin{bmatrix} I & 0 \\ -D^{-1}C & I \end{bmatrix} \times \begin{bmatrix} S^{-1} & 0 \\ 0 & D^{-1} \end{bmatrix} \times \begin{bmatrix} I & -BD^{-1} \\ 0 & I \end{bmatrix}.$$

En faisant cette multiplication matricielle, on obtient la forme générale de l'inverse d'une matrice par blocs. En regroupant tous ces calculs, on obtient le théorème suivant.

Théorème 22.1 (Théorème d'inversion de Schur)**.** *Si D est inversible et si $A - BD^{-1}C$ est inversible, alors la matrice*

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}$$

est inversible et son inverse est donné par

$$M^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix}.$$

En regardant de plus près les identités qui viennent de la démonstration du théorème, on peut tirer beaucoup de choses non triviales. D'abord, la formule Équation 22.2 montre que le rang de M est égal au rang d'une matrice triangulaire supérieure, qui est lui-même la somme des rangs des blocs diagonaux, et ceci est valable même si S ou M ne sont pas inversibles :

$$\text{rang}(M) = \text{rang}(A - BD^{-1}C) + \text{rang}(D).$$

La même formule permet de voir immédiatement que le déterminant de M se calcule très simplement :

$$\det(M) = \det(D)\det(A - BD^{-1}C).$$

22.2 Inversion des perturbations de rang faible

Soit $A \in \mathbb{R}^{n,n}, B \in \mathbb{R}^{r,r}, X \in \mathbb{R}^{n,r}, Y \in \mathbb{R}^{r,n}$. La matrice XBY a un rang plus petit que r.

$$(A + XBY)^{-1} = A^{-1} - A^{-1}X(B^{-1} + YA^{-1}X)^{-1}YA^{-1} \tag{22.3}$$

Par exemple, si u, v sont deux vecteurs de taille n, cette formule permet de calculer l'inverse de $A + uv$, qui est une perturbation de rang 1 de la matrice A :

$$(A + uv^\top)^{-1} = A^{-1} - \frac{1}{1 + \langle v, A^{-1}u\rangle}A^{-1}uv^\top A^{-1}.$$

22.3 Formule de la résolvante

$$Y^{-1} - X^{-1} = X^{-1}(X - Y)Y^{-1}$$

[1] on rappelle que $(XY)^{-1} = Y^{-1}X^{-1}$.

23 50 nuances de TCL

23.1 La version classique

Soit (X_i) une suite de variables aléatoires iid possédant une moyenne μ et une variance σ^2. On note $\bar{X}_n$ leur moyenne empirique,

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{23.1}$$

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 23.1. *La variable aléatoire*

$$\frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}}$$

converge en loi vers $N(0,1)$.

Démonstration. Si φ est la transformée de Fourier commune de la loi des $Y_i = (X_i - \mu)/\sigma$ et ψ celle de Équation 23.1, 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 $N(0,1)$.

□

23.2 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 23.2. *Si ces variables vérifient la* condition de Lindeberg, *à savoir que pour tout $\delta > 0$,*

$$\frac{1}{\varsigma_n^2}\sum_{i=1}^{n}\mathbb{E}[|X_i-\mu_i|^2\mathbf{1}_{|X_i-\mu_i|>\delta\varsigma_n}]\to 0 \tag{23.2}$$

alors la variable aléatoire

$$\frac{\bar{X}_n-\bar{\mu}_n}{\varsigma_n}$$

converge en loi vers $N(0,1)$.

23.3 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 ε_i des variables aléatoires iid de variance σ^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

$$\sqrt{n}\frac{\sum_{i=1}^{n}x_i\varepsilon_i}{\sigma s_n} \tag{23.3}$$

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:

$$\max_{i=1,\dots,n}\frac{|x_i|^2}{s_n^2}\to 0. \tag{23.4}$$

Théorème 23.3. *Sous les hypothèses précédentes, Équation 23.3 converge en loi lorsque $n\to\infty$ vers une $N(0,1)$.*

Démonstration. La démonstration repose sur Théorème 23.2 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 23.4.

En regroupant tout ceci, on voit que Équation 23.2 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 23.4, ce terme tend vers zéro.

□