Statistique mathématique

Magalie Fromont

Deuxième année · ENS Rennes, promotion 2019
Notes prises par Téofil Adamski (version du 18 mars 2021)

Chapitre 1

Problèmes de statistique inférentielle

1.1 Introduction

Il existe trois grands types de problèmes de la statistique inférentielle à travers deux exemples concrets :

– le problème d'estimation ponctuelle ;
– le problème d'estimation par régions de confiance ;
– le problème de tests d'hypothèses.

Ces problèmes formalisent des questions concrètes se posant lors d'une expérience aléatoire. Leurs solutions peuvent servir à analyser les phénomènes sous-jacents ou une prise de décision.

Dans ce cas, on va considérer des expériences aléatoires conduisant à l'observation x d'une variable aléatoire X qui, le plus souvent, sera un n-échantillon $(X_1, \dots, X_n)$ ou un couple (Y, Z) d'échantillons indépendants. En général, les variables aléatoires seront à valeurs dans $\mathbf{R}^d$. Bien évidemment, d'autres cadres seront possibles : des variables aléatoires dans un espace de dimension infinie, un processus stochastique, etc.

Les questions concrètes que nous envisagerons se traduiront en question sur la loi d'une variable aléatoire relevant de l'un (au moins) des trois grands problèmes cités ci-dessus.

À l'inverse de ceux qu'on peut faire en théorie des probabilités, on ne va pas considérer une unique loi de probabilité mais une famille $\mathscr{P}$ de loi potentielle. Le plus souvent, cette famille $\mathscr{P}$ est indexée par un espace Θ fini ou de dimension infinie. Dans toute la suite, on considère une variable aléatoire X définie sur une espace probabilisé $(\Omega, \mathscr{A}, \mathbb{P})$.

Définition 1.1. Le *modèle statistique* correspondant à X est le triplet $(\mathscr{X}, \mathscr{B}, \{P_\theta\}_{\theta\in\Theta})$ où
– l'ensemble $\mathscr{X}$ est l'ensemble des valeurs possibles de X, appelé l'*espace des observation* ;
– la classe de parties $\mathscr{B}$ est une tribu sur $\mathscr{X}$;
– la loi P_θ est la loi de X dépendant du paramètre θ.

Ce modèle posé, on vise, sur l'observation x seule, à inférer ou induire des caractéristique de la loi de probabilité P_θ. Dans un problème d'estimation, on cherche à estimer ces caractéristiques sans *a priori*. Dans un problème de test, on cherche à valider ou invalider une hypothèses émise sur ces caractéristiques. Le cours abordera

– des notions statistiques de base utiles pour poser et analyser les différents problèmes (statistique, estimateur, région de confiance, test, etc) et
– des notions de probabilités utiles pour résoudre ces problème, à savoir
  ◦ les lois de probabilité usuelles,
  ◦ les théorèmes limites et la théorie de la convergence des suites de variables aléatoires, préalables indispensables pour les résultats de statistique inférentielle asymptotique,
  ◦ les inégalités de concentrations, préalables indispensables pour les résultats de statistique inférentielle non asymptotique.

1.2 Duo de mise en bouche

1.2.1 Un exemple avec des M&M's

On s'intéresse à la proportion θ de M&M's de couleur chaude comme rouge, jaune ou orange dans les paquets des célèbres bonbons à la cacahuète. Pour l'obtenir, on mène l'expérience aléatoire suivante : on prend n bonbons au hasard et on relève le nombre de bonbons rouge, jaune ou orange. On va considérer les trois problèmes de statistique inférentielle.

Quel modèle statistique ?

La variable aléatoire qui nous considérons sera un n-échantillons $X \coloneqq (X_1, \dots, X_n)$ de variables aléatoires indépendantes et identiquement distribuées de loi $\mathrm{Ber}(\theta)$ modélisant la couleur (chaude ou froide) de n bonbons pris au hasard telle que, pour tout $i \in [\![1, n]\!]$, on ait $X_i = 1$ si et seulement si le i-ième bonbon est de couleur chaude.

Suite à notre expérience, on note $x \coloneqq (x_1, \dots, x_n)$ l'observation de X. De plus, le modèle statistique associé sera le triplet

$$\mathfrak{M} = (\mathscr{X}, \mathscr{B}, \{P_\theta\}_{\theta\in\Theta}) \coloneqq (\{0,1\}^n, \mathscr{P}(\{0,1\}^n), \{\mathrm{Ber}(\theta)^{\otimes n}\}_{\theta\in[0,1]}).$$

Comment peut-on interpréter ce modèle ? Quelles sont ses limites ? Déjà, le modèle induit plusieurs hypothèses :

– les variables aléatoires X_i sont de même loi (cela dit que la production de bonbons est stable) ;
– elles sont indépendantes (il n'y a pas d'influence de la production d'un bonbon sur une autre).

Ce modèle semble alors réaliste.

Par quelle valeur, construite sur x, peut-on approcher ou estimer θ ?

Sur la base de l'observation $x \coloneqq (x_1, \dots, x_n)$, une valeur estimée du paramètre θ est donnée par

$$T(x) \coloneqq \frac{x_1 + \cdots + x_n}{n}.$$

Cela définit une fonction $T\colon \{0,1\}^n \longrightarrow \overline{\Theta} \coloneqq \{0, 1/n, \dots, 1\}$ qui est mesurable : c'est une statistique sur le modèle $\mathfrak{M}$ ou, puisque $\overline{\Theta} \subset \Theta \coloneqq [0,1]$, un estimateur de θ. Toute image $T(x)$ avec $x \in \{0,1\}^n$ est appelée une valeur estimée ou estimation de θ. On note usuellement $\hat{\theta}_n \coloneqq T(X)$ que l'on qualifie, par abus de langage, d'estimateur également.

Comment se justifie la construction de $\hat{\theta}_n$?

Sur la base d'un n-échantillon $(X_1, \dots, X_n)$ de variables aléatoires définie sur un même espace de probabilité $(\Omega, \mathscr{A}, \mathbb{P})$, à valeurs dans $(\mathbf{R}^d, \mathscr{B}(\mathbf{R}^d))$ et de loi P, on peut construire une mesure de probabilité aléatoire

$$\hat{P}_n \coloneqq \frac{\delta_{X_1} + \cdots + \delta_{X_n}}{n},$$

appelée la mesure empirique associé à ce n-échantillon. La loi forte des grands nombres qui nous dit que, pour tout événement B, on a presque sûrement

$$\hat{P}_n(B) \longrightarrow P(B).$$

Remarquons que le « presque sûrement » dépend de l'événement B.

On se place en dimension $d = 1$. Soit F la fonction se répartition de P. Pour tout $t \in \mathbf{R}$, on pose

$$\hat{F}_n(t) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{]-\infty,t]}(X_i)$$

la fonction de répartition empirique associée au n-échantillon. Le théorème de Glivenko-Cantelli donne la convergence presque sûre

$$\sqrt{n} \sup_{t\in\mathbf{R}} |\hat{F}_n(t) - F(t)| \longrightarrow 0$$

et l'inégalité de Dvorezky-Kiefer-Wolfowitz s'écrit

$$\mathbb{P}(\sqrt{n}\sup_{t\in\mathbf{R}}|\hat{F}_n(t)-F(t)|>s)\leqslant 2\exp(-2s^2),\quad s\in\mathbf{R}.$$

Pour un espace $\mathscr{X}$ séparable, avec probabilité 1, on a la convergence en loi $\hat{P}_n \xrightarrow{\text{loi}} P$. Pour bien comprendre ce résultat de convergence étroite presque sûre, on peut introduire des distances métrisant cette topologie. Ici, on peut écrire

$$\hat{\theta}_n=\mathbb{E}_{\hat{P}_n}[X_i]=\int t\,\mathrm{d}\hat{P}_n(t)\quad\text{et}\quad\theta=\mathbb{E}_{\hat{P}_\theta}[X_i]=\int t\,\mathrm{d}\hat{P}_\theta(t).$$

L'estimateur $\hat{\theta}_n$ est dit par insertion ou estimateur *plug-in*. Lorsque le paramètre à estimer est un moment de la loi P_θ ou s'exprime en fonction des moments de cette loi, on parle d'estimateur des moments.

Remarquons que la spécification de la loi P_θ n'est pas utile pour la méthode des moments. Cette dernière est donc une méthode d'estimation non paramétrique.

Une autre justification de sa construction

On prend en compte ici plus spécifiquement le modèle considéré $\mathfrak{M}$. Pour une observation $x\coloneqq(x_1,\dots,x_n)\in\{0,1\}^n$ et un paramètre $\theta\in\Theta$, la quantité

$$L_n(x,\theta)\coloneqq P_\theta(X_1=x_1,\dots,X_n=x_n)=P((x_1,\dots,x_n))=\theta^{x_1+\cdots+x_n}(1-\theta)^{n-(x_1+\cdots+x_n)}$$

représente le degré de vraisemblance de l'observation x pour P_θ. On vérifie que la quantité $T(x)$ est la valeur minimisant la fonction $L_n(x,\cdot)$ sur Θ, c'est-à-dire pour la quelle l'observation x est la plus vraisemblable pour la loi P_θ. La preuve de ce résultat passe classiquement par la considération de la log-vraisemblance $\ell_n\coloneqq\ln L_n$, un argument de concavité et la recherche d'un point critique.

L'estimateur $\hat{\theta}_n$ est appelé un estimateur du maximum de vraisemblance. La méthode de maximisation de la vraisemblance est, dans ce cas précis, une méthode d'estimation paramétrique.

Peut-on faire confiance à ces estimations ?

On se demande comment l'estimateur $\hat{\theta}_n$ approche le paramètre θ. Pour des propriétés asymptotiques, on utilise, pour cela, les différents théorèmes usuels de probabilité : selon les cas,

– la loi faible des grands nombres assure qu'il s'agit d'un estimateur faiblement consistant ;
– la loi forte des grands nombres assure qu'il s'agit d'un estimateur fortement consistant ;
– le théorème centrale-limite assure qu'il s'agit d'un estimateur asymptotique normal, la qualité ou performance de cette estimateur est jugée par la vitesse de convergence $\sqrt{n}$ et la variance $\theta(1-\theta)$ puisque le théorème donne

$$\sqrt{n}(\hat{\theta}_n-\theta)\xrightarrow{\text{loi}}\mathscr{N}(0,\theta(1-\theta))$$

dans notre exemple ;
– enfin, en notant $\mathbb{E}_\theta$ l'espérance et $\operatorname{Var}\theta$ la variance sous la loi P_θ, on a

$$\mathbb{E}_\theta[(\hat{\theta}_n-\theta)^2]=\operatorname{Var}_\theta[\hat{\theta}_n]=\frac{\theta(1-\theta)}{n}\longrightarrow 0$$

et on dit que la risque quadratique de l'estimateur tend vers zéro à la vitesse $1/n$.

Pour des propriétés non asymptotiques, grâce à cette dernière relation, l'erreur quadratique moyenne $\mathbb{E}_\theta[(\hat{\theta}-\theta)^2]^{1/2}$ de l'estimateur est majorée par $1/2\sqrt{n}$. La question qui vient alors naturellement est la suivante : ce risque et cette erreur quadratique moyenne sont-ils optimaux ou d'ordre optimal ? Et en quel sens ? Quel est le critère à prendre comme référence ?

On peut remarquer qu'un estimateur constant $T\coloneqq\theta_0$ est de risque quadratique nul sous θ_0 alors qu'il est de risque élevé pour d'autres paramètres θ. Pour tout estimateur T, on introduit le risque maximal de T sur Θ

$$\mathscr{R}(T,\Theta)\coloneqq\sup_{\theta\in\Theta}\mathbb{E}_\theta[(T(X)-\theta)^2]$$

et le risque minimax sur Θ

$$\mathrm{m}\mathscr{R}(\Theta)\coloneqq\inf_T\mathscr{R}(T,\Theta)$$

où l'infimum est pris sur l'ensemble des estimateurs. Un estimateur réalisant le risque minimax sur Θ est lui aussi qualifié de minimax sur Θ. Ce risque dépend de l'espace Θ car on pourra être amené à le restreindre afin d'assurer l'existence d'un estimateur minimax. De même, on pourra réduire la classe d'estimateurs.

Pour des propriétés non asymptotique, on peut utiliser le théorème central limite ou l'inégalité de Bienaymé-Tchebychev pour estimer la quantité

$$P_\theta(\sqrt{n}\,|T(X)-\theta|>t).$$

On verra également des inégalités de concentration comme celles de Hoeffding ou de Cramèr-Chernov. Celles-ci permettent alors de juger la qualité de l'estimateur $\hat{\theta}_n$ mais aussi de construire des « fourchettes d'estimation » du paramètre θ.

Test d'hypothèses

Un · e étudiant · e de l'ENS Rennes pense que la marche de bonbons discrimine les couleurs froides et qu'en réalité le paramètre θ vaut $\theta_1 \coloneqq 2/3$. On souhaite confronter les deux hypthèses : celle de la marque et celle de l'étudiant · e.

Quelle règle de décision peut-on construire sur la base de l'observation ?

On peut considérer l'estimateur ponctuel $\hat{\theta}_n$ ou un intervalle de confiance. Pour prendre en compte les variations de l'estimateur, on se donne une marge de sécurité : pour une certaine valeur critique s, on décide que la marque à tort, *i. e.* de rejeter l'hypothèse $\theta=\theta_0 \coloneqq 1/2$ au profit de l'hypothèse $\theta=\theta_1$ dès que $\hat{\theta}_n > s$. Cette valeur critique s est fixée à partir d'une évaluation de risque de se tromper avec la règle de décision correspondante.

Qu'entend-on par le risque de se tromper ?

Deux types d'erreurs sont possibles : décider que la marque a tort ou décider que la marque a raison, *i. e.* accepter l'une des deux hypothèses. L'évaluation du reste de commettre chacune de ces erreurs se fait dans le cas du modèle statistique considéré. On associe respectivement à ces deux types d'erreurs le risque de première (respectivement deuxième) espèce

$$P_{\theta_i}(T(x)\leqslant s).$$

1.2.2 Un exemple alcoolique

Le ministère de la santé étudie régulièrement les données annuelles nationales sur la consommation quotidienne moyenne d'alcool par personne fournies par l'INSEE. En particulier, il évalue la nécessite de prendre des mesures contre la consommation d'alcool.

On modélise la consommation quotidienne d'alcool pur (en gramme) par personne de plus de 15 ans par une variable aléatoire de loi $\mathscr{N}(\theta,\sigma^2)$ avec $\sigma \coloneqq 2$.

En janvier 1991, la loi Évin limite la publicité pour les boissons alcoolisées. Son objectif est de faire baisser le paramètre θ, qui valait 35, en dessous de 33. Que peut-on conclure sur la base de l'observation des consommations quotidiennes moyennes de 1991 à 1994 ?

Année	1991	1992	1993	1994
Consommation θ	34,7	34,4	33,7	33,3

Choix et critique du modèle statistique

Posons notre modèle statistique. Comme nos observations se font sur quatre année, la variable aléatoire considéré va être un 4-échantillon $X \coloneqq (X_1,X_2,X_3,X_4)$ où chaque variable aléatoire X_i modélise la consommation d'alcool sur l'année $1990+i$. Ici, on considère l'observation x_0 donnée par le tableau. On pose $n \coloneqq 4$. Le modèle statistique est le triplet

$$(\mathbf{R}^n, \mathscr{B}(\mathbf{R}^n), \{\mathscr{N}(\theta,\sigma^2)^{\otimes n}\}_{\theta\in\mathbf{R}}).$$

Cela impose les hypothèses suivantes :

– les variables aléatoires X_i sont indépendantes, cette hypothèse est critiquable sauf si les populations sondées sont différentes et potentiellement indépendantes ;
– elles sont de même loi et, en particulier, de mêmes variances et espérances, cela suppose que l'effet de la loi Évin est immédiat et durable ;
– elles sont gaussiennes, cela semble raisonnable grâce au théorème central limite puisqu'elles sont des moyennes, mais le support d'une loi gaussiennes est la droite entière $\mathbf{R}$ alors que les variables aléatoires sont positives.

Estimateur

Trouver un estimateur ponctuelle. Sur la base d'une observation $x \coloneqq (x_1, \dots, x_n) \in \mathbf{R}^n$, on peut estimer le paramètre θ comme précédemment, c'est-à-dire pour la quantité

$$T(x) \coloneqq \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Par exemple, avec l'observation x_0 donnée par le tableau, on a $T(x_0) = 34{,}025$. On pose l'estimateur $\hat{\theta}_n \coloneqq T(X)$.

Justifions ce choix. Si on prend en compte la spécification de la loi $P_\theta \coloneqq \mathscr{N}(\theta, \sigma^2)$, on considère la vraisemblance définie par la relation

$$L_n(x, \theta) \coloneqq \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\Big(-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \theta)^2\Big).$$

En introduisant la log-vraisemblance, on vérifie que la quantité $T(x)$ maximise la fonction $L_n(x, \cdot)$.

Chapitre 2

Compléments de probabilité

2.1 Inégalité classique

On considère une variable aléatoire Z et des variable aléatoires indépendantes X_1, …, X_n définies sur un même espace probabilisé $(\Omega, \mathscr{A}, \mathbb{P})$. Toutes les variables aléatoires considérées seront à valeurs réelles. On note $S_n = \sum_{i=1}^{n} X_i$. On va donner des inégalités dont le but est de préciser la concentration de la somme S_n autour de son espérance lorsque cette dernière existe. L'inégalité de Markov servira de base.

2.1.1 Rappels et notation

La densité de la loi centrée réduite est la fonction $\phi\colon \mathbf{R} \longrightarrow \mathbf{R}$ définie par

$$\phi(x) \coloneqq \frac{1}{\sqrt{2\pi}} \exp(-x^2/2).$$

On note $\Phi\colon \mathbf{R} \longrightarrow [0,1]$ sa fonction de répartition et $\overline{\Phi} \coloneqq 1 - \Phi$ sa fonction de survie.

Proposition 2.1. Pour tout $x > 0$, on a

$$\frac{1}{x\sqrt{2\pi}} \left(1 - \frac{1}{x^2}\right) \exp(-x^2/2) \leqslant \overline{\Phi}(x) \leqslant \frac{1}{x\sqrt{2\pi}} \exp(-x^2/2).$$

Preuve La minoration se trouve en intégrant par parties. La majoration est évidente. □

Corollaire 2.2. On suppose que les variables aléatoires X_i suivent la loi normale centrée réduite. Pour tout $x \geqslant 0$, on a

$$\frac{\sqrt{n}}{x}\sqrt{\frac{2}{\pi}} \left(1 - \frac{n}{x^2}\right) \exp(-x^2/2n) \leqslant \mathbb{P}(|S_n - \mathbb{E}[S_n]| \geqslant x) \leqslant \frac{\sqrt{n}}{x}\sqrt{\frac{2}{\pi}} \exp(-x^2/2n).$$

On rappel maintenant les inégalités de Markov et de Bienaymé-Tchebychev.

Proposition 2.3 *(inégalité de Markov).* Pour tout $p > 0$, tout $x > 0$ et toute variable aléatoire Z, on a

$$\mathbb{P}(|Z|^p \geqslant x) \leqslant \frac{\mathbb{E}[|Z|^p]}{x}.$$

Proposition 2.4 *(inégalité de Bienaymé-Tchebychev).* Pour tout $x > 0$ et toute variable aléatoire Z de carré intégrable, on a

$$\mathbb{P}(|Z| \geqslant x) \leqslant \frac{\mathbb{E}[Z^2]}{x^2}.$$

2.1.2 Méthode de Cramér-Chernoff

Définition 2.5. La *transformée de Laplace* d'un variable aléatoire Z est la fonction

$$L_Z\colon \left| \begin{array}{l} \mathbf{R} \longrightarrow \mathbf{R} \cup \{+\infty\}, \\ \lambda \longmapsto \mathbb{E}[\exp(\lambda z)] \end{array} \right.$$

et sa *transformée de Cramér* est la fonction

$$\Lambda_Z^* : \left| \begin{array}{l} \mathbf{R} \longrightarrow \mathbf{R} \cup \{+\infty\}, \\ t \longmapsto \sup_{\lambda \geqslant 0} (\lambda t - \Lambda_Z(\lambda)) \end{array} \right. \quad \text{avec} \quad \Lambda_Z(\lambda) \coloneqq \ln L_Z(\lambda).$$

Proposition 2.6 *(inégalité de Cramér-Chernoff).* Pour tout $x \geqslant 0$, on a

$$\mathbb{P}(Z \geqslant x) \leqslant \exp(-\Lambda_Z^*(x)).$$

Applications.

– Soit Z une variable aléatoire centrée sous-gaussienne de facteur de variance $v > 0$, *i. e.* vérifiant

$$\Lambda_Z(\lambda) \leqslant v\lambda^2/2, \quad \lambda \geqslant 0.$$

Alors pour tous $t \in \mathbf{R}$ et $x > 0$, on a

$$\Lambda_Z^*(t) \geqslant t^2/2v \qquad \text{et} \qquad \mathbb{P}(Z > x) \vee \mathbb{P}(Z < -x) \leqslant \exp(-x^2/2v)$$

– Soit Z une variable aléatoire centrée sous-gamma à droite de facteur de variance $v > 0$ et de facteur d'échelle $c > 0$, *i. e.* vérifiant

$$\Lambda_Z(\lambda) \leqslant \frac{v\lambda^2}{2(1-c\lambda)}, \quad \lambda \in [0, 1/c[.$$

Pour $x > 0$, on pose $h_1(x) \coloneqq 1 + x - \sqrt{1+2x}$. Alors pour tout $t > 0$, on a

$$\Lambda_Z^*(t) \geqslant \frac{v}{c^2} h_1\Big(\frac{ct}{v}\Big).$$

2.1.3 Autres inégalités

Proposition 2.7 *(inégalité de Hœffding).* On suppose que les variables aléatoires X_i sont bornées par des réels $a_i, b_i \in \mathbf{R}$ avec $a_i < b_i$. Alors pour tout $x > 0$, on a

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geqslant x) \leqslant \exp\Big(-\frac{2x^2}{\sum_{i=1}^n (b_i - a_i)^2}\Big).$$

Proposition 2.8 *(inégalité de Bennett).* Soit $c > 0$. On suppose que, pour tout $i \in [\![1, n]\!]$, on a

$$X_i \leqslant c \quad \text{et} \quad \mathbb{E}[X_i^2] < +\infty.$$

Posons $v \coloneqq \sum_{i=1}^n \mathbb{E}[X_i^2]$. Alors pour tout $x > 0$, on a

$$\mathbb{P}(S_n - \mathbb{E}[S_n] \geqslant x) \leqslant \exp\Big(-\frac{v}{c^2} h\Big(\frac{cx}{v}\Big)\Big) \leqslant \exp\Big(-\frac{x^2}{2v + 2cx/3}\Big)$$

avec $h(x) \coloneqq (1+x)\ln(1+x) - x$.

Proposition 2.9 *(inégalité de Bernstein).* Soient $v, c > 0$. On suppose

$$\sum_{i=1}^n \mathbb{E}[X_i^2] \leqslant v \quad \text{et} \quad \sum_{i=1}^n \mathbb{E}[(X_i^+)^k] \leqslant \frac{vk!c^{k-2}}{2}, \quad k \geqslant 3;$$

Alors pour tous $\lambda \in [0, 1/c[$ et $x \geqslant 0$, on a

$$\Lambda_{S_n - \mathbb{E}[S_n]}(\lambda) \leqslant \frac{v\lambda^2}{2(1-c\lambda)} \quad \text{et} \quad \mathbb{P}(S_n - \mathbb{E}[S_n] \geqslant x) \leqslant \exp\Big(-\frac{x^2}{2v + 2cx}\Big).$$

2.2 Complément sur les lois usuelles

Lemme 2.10 *(Stein).* Soient X un variable aléatoire suivant la loi $\mathscr{N}(0,1)$ et $g \colon \mathbf{R} \longrightarrow \mathbf{R}$ une fonction absolument continue telle que $\mathbb{E}[|Xg(X)| < +\infty$. Alors $g'(X)$ est intégrable et

$$\mathbb{E}[g'(X)] = \mathbb{E}[Xg(X)].$$

THÉORÈME 2.11 *(central limite vectoriel).* Soit $(X_n)_{n\in\mathbf{N}}$ une suite de vecteurs aléatoires de $\mathbf{R}^d$ indépendants et identiquement distribués dont toutes les coordonnées sont de carré intégrable, d'espérance $\mu \in \mathbf{R}^d$ et de matrice de covariance $\Sigma \in \mathscr{M}_d(\mathbf{R})$. Alors

$$\sqrt{n}\Big(\frac{1}{n}\sum_{k=1}^{n} X_k - \mu\Big) \xrightarrow{\text{loi}} \mathscr{N}_d(0,\Sigma).$$

Loi du khi-deux

DÉFINITION 2.12. Soit X un vecteur aléatoire suivant la loi $\mathscr{N}_d(0, I_d)$. La loi de la variable aléatoire $\|X\|_2^2$ est la *loi du khi-deux* à d degrés de liberté, notée $\chi^2(d)$.

PROPOSITION 2.13. Soit K un variable aléatoire suivant la loi $\chi^2(d)$. Alors

$$\mathbb{E}[K] = d \quad \text{et} \quad \mathrm{Var}[K] = 2d.$$

PROPOSITION 2.14. Soit $X \sim \mathscr{N}_d(\mu,\Sigma)$. Alors $(X-\mu)'\Sigma^{-1}(X-\mu) \sim \chi^2(d)$.

Preuve Il suffit de remarquer qu'on a $\Sigma^{-1/2}(X-\mu) \sim \mathscr{N}_d(0, I_d)$. □

THÉORÈME 2.15 *(Cochran).* Soit $X \sim \mathscr{N}_n(0, I_n)$. Soient $V_1, \dots, V_p \subset \mathbf{R}^n$ des sous-espaces vectoriels orthogonaux et de dimensions respectives $d_1, \dots, d_p \in \mathbf{N}^*$ tels que

$$\mathbf{R}^n = V_1 \oplus \cdots \oplus V_p.$$

Pour $i \in [\![1,p]\!]$, on note $\Pi_i \in \mathscr{M}_n(\mathbf{R})$ la matrice de projection sur V_i. Alors
- on a $\|X\|^2 = \|\Pi_1 X\|^2 + \cdots + \|\Pi_p X\|^2$;
- les vecteurs aléatoires $\Pi_i X$ sont gaussiens et indépendants ;
- pour tout $i \in [\![1,p]\!]$, on a $\|\Pi_i X\|^2 \sim \chi^2(d_i)$.

Preuve Pour chaque sous-espace vectoriel V_i, on en fixe une base orthonormée $(e_{i,1}, \dots, e_{i,d_i})$. Le premier point découle immédiatement du théorème de Pythagore. Montrons le deuxième point. Il est clair que les projections $\Pi_i X$ sont des vecteurs gaussiens. Ainsi la vecteur $(\Pi_1 X, \dots, \Pi_p X)$ est gaussien. Pour montrer l'indépendance des vecteurs $\Pi_i X$, il suffit donc de montrer qu'ils ne sont pas corrélés et c'est bien le cas puisque, pour tous indices distincts $i, j \in [\![1,p]\!]$, on a

$$\begin{aligned}\mathrm{Cov}[\Pi_i X, \Pi_j X] &= \mathbb{E}[(\Pi_i X - \mathbb{E}[\Pi_i X])(\Pi_j X - \mathbb{E}[\Pi_j X])'] \\ &= \mathbb{E}[(\Pi_i X)(\Pi_j X)'] \\ &= \mathbb{E}[\langle \Pi_i X, \Pi_j X\rangle] = 0\end{aligned}$$

par orthogonalité.

Montrons le troisième point. Soit $i \in [\![1,p]\!]$. Avec le choix des bases, on peut écrire

$$\|\Pi_i X\|^2 = \sum_{j=1}^{d_i} \langle X, e_{i,j}\rangle^2.$$

Il suffit alors de montrer que les variables aléatoires $\langle X, e_{i,j}\rangle$ sont indépendantes et suivent la loi normale centrée réduite. Soit $j \in [\![1,d_i]\!]$. Comme X est un vecteur aléatoire, la variable aléatoire $\langle X, e_{i,j}\rangle$ suit une loi gaussienne d'espérance

$$\mathbb{E}[X' e_{i,j}] = \mathbb{E}[X'] e_{i,j} = 0$$

et de variance

$$\mathrm{Var}[X' e_{i,j}] = \sum_{k=1}^{n} \mathrm{Var}[X_k]^2 e_{i,j,k} = \|e_{i,j}\|^2 = 1.$$

D'où $\langle X, e_{i,j}\rangle \sim \mathscr{N}(0,1)$. Il reste à montrer que les variables aléatoires $\langle X, e_{i,j}\rangle$ sont indépendantes. Montrons qu'elles sont décorrélées. Pour tous indices distincts $j, j' \in [\![1,d_i]\!]$, on a

$$\mathrm{Cov}[X' e_{i,j'}, X' e_{i,j'}] = \mathbb{E}[(X' e_{i,j})(X' e_{i,j'})]$$

$$= \sum_{k=1}^{n} \sum_{k'=1}^{n} \mathbb{E}[(X_k X_{k'}) e_{i,j,k} e_{i,j',k'}]$$
$$= \sum_{k=1}^{n} \mathbb{E}[X_k^2] e_{i,j,k} e_{i,j',k} = 0.$$

Finalement, on a déduit $\|\Pi_i X\|^2 \sim \chi^2(d_i)$. □

DÉFINITION 2.16. – Soient $U \sim \mathcal{N}(0,1)$ et $K \sim \chi^2(d)$. La loi de la variable aléatoire $Y/\sqrt{K/d}$ est la *loi de Student* à d degrés de liberté, notée $\mathscr{T}(d)$.
– Soient $K_1 \sim \chi^2(d_1)$ et $K_2 \sim \chi^2(d_2)$. La loi de la variable aléatoire $(K_1/d_1)/(K_2/d_2)$ est la *loi de Fischer-Snedecor* à d_1 et d_2 degrés de liberté, notée $\mathscr{F}(d_1, d_2)$.

Chapitre 3

Modèles statistiques, notion d'estimation

Dans tout le chapitre, on fixe un espace probabilisé $(\Omega, \mathscr{A}, \mathbb{P})$ et on considère un modèle statistique $\mathfrak{M} \coloneqq (\mathscr{X}, \mathscr{B}, \{P_\theta\}_{\theta\in\Theta})$. On peut lui associer une variable aléatoire $X\colon \Omega \longrightarrow \mathscr{X}$ qui suit une des lois P_θ. On cherche à estimer le paramètre inconnu θ. Pour cela, on s'intéressera à des observations $x \in \mathscr{X}$ de cette variable aléatoire.

3.1 Estimateurs

Définition 3.1. Le modèle statistique $\mathfrak{M}$ est
– *identifiable* si l'application $\theta \in \Theta \longmapsto P_\theta$ est injective ;
– *paramétrique* si $\Theta \subset \mathbf{R}^d$ pour un entier $d \in \mathbf{N}^*$.

▷ Exemples. Les modèles

$$(\{0,1\}^n, \mathscr{P}(\{0,1\}^n), \{\mathrm{Ber}(\theta)^{\otimes n}\}_{\theta\in\{0,1\}}) \quad \text{et} \quad (\mathbf{R}^n, \mathscr{B}(\mathbf{R}^n), \{\mathscr{N}(\theta,4)^{\otimes n}\}_{\theta>0}).$$

sont identifiables et paramétriques.

Définition 3.2. – Une *statistique* sur $(\mathscr{X}, \mathscr{B})$ est une fonction mesurable sur $(\mathscr{X}, \mathscr{B})$.
– Un *estimateur* de $\psi(\theta)$ sur $(\mathscr{X}, \mathscr{B})$ est une statistique à valeurs dans un sur-ensemble $\overline{\psi(\Theta)} \supset \psi(\Theta)$.
– Pour un observation x d'une variable aléatoire $X \sim P_\theta$, si T est un estimateur de $\psi(\theta)$, la valeur $T(x)$ est appelée la *valeur estimée* ou l'*estimation* de $\psi(\theta)$.

On rappel qu'un n-échantillon est une n-uplet $(X_1, \dots, X_n)$ de variables aléatoires indépendantes et identiques distribuées.

Définition 3.3. Le modèle $\mathfrak{M}$ est *modèle d'échantillonnage* si, pour tout $\theta \in \Theta$, il existe une probabilité p_θ telle que $P_\theta = p_\theta^{\otimes n}$.

Définition 3.4. La *mesure empirique* associée à un n-échantillon $(X_1, \dots, X_n)$ est le mesure de probabilité $\hat{P}_n$ définie par

$$\forall \omega \in \Omega, \quad \hat{P}_n(\omega) = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i(\omega)}.$$

Pour $B \in \mathscr{B}(\mathbf{R}^d)$ et $\omega \in \Omega$, on note $\hat{P}_n(\omega, B) \coloneqq \hat{P}_n(\omega)(B)$.

◇ Remarques. – Pour tout $B \in \mathscr{B}(\mathbf{R}^d)$, l'application $\hat{P}_n(\cdot, B)$ est une variable aléatoire discrète telle que $n\hat{P}_n(\cdot, B) \sim \mathrm{Bin}(n, P(B))$ où la probabilité P est celle de l'échantillon.
– Pour tout $\omega \in \Omega$, l'application $\hat{P}_n(\omega, \cdot)$ est une probabilité sur $\mathbf{R}^d$.

Définition 3.5. La *fonction de répartition empirique* d'un n-échantillon $(X_1, \dots, X_n)$ de variables aléatoires réelles est l'application

$$\hat{F}_n : \begin{vmatrix} \Omega \longrightarrow [0,1]^{\mathbf{R}}, \\ \omega \longmapsto [x \longmapsto F_n(\omega, x) \coloneqq \hat{P}_n(\omega,]-\infty, x]). \end{vmatrix}$$

◇ Remarque. – Pour tout $x \in \mathbf{R}$, on a $n\hat{F}_n(x) \sim \mathrm{Bin}(n, F(x))$ où la fonction F est la fonction de répartition associée à la loi de l'échantillon.
– Pour tout $\omega \in \Omega$, la fonction $\hat{F}_n(\omega, \cdot)$ est la fonction de répartition de la loi $\hat{P}_n(\omega, \cdot)$.

THÉORÈME 3.6 *(Glivenko-Cantelli).* Soit $(X_i)_{i\in\mathbf{N}^*}$ une suite de variables aléatoires réelles indépendantes et identiquement distribuées. On note F la fonction de répartition de X_1. Pour $n \in \mathbf{N}^*$, on note $\hat{F}_n$ la fonction de répartition empirique associée au n-échantillon $(X_1, \dots, X_n)$. Alors il existe un sous-ensemble $\Omega' \subset \Omega$ de probabilité 1 tel que

$$\forall \omega \in \Omega', \quad \|\hat{F}_n(\omega) - F\| \longrightarrow 0.$$

THÉORÈME 3.7 *(Varadarajan).* Soit $(X_i)_{i\in\mathbf{N}^*}$ une suite de variables aléatoires réelles indépendantes et identiquement distribuées. On note P la loi de X_1. Pour $n \in \mathbf{N}^*$, on note $\hat{P}_n$ la mesure empirique associée au n-échantillon $(X_1, \dots, X_n)$. Alors il existe un sous-ensemble $\Omega' \subset \Omega$ de probabilité 1 tel que

$$\forall \omega \in \Omega', \quad \hat{P}_n(\omega) \xrightarrow{\text{loi}} P.$$

DÉFINITION 3.8. Soit $X \coloneqq (X_1, \dots, X_n)$ un n-échantillon tel que $X \sim P_\theta = p_\theta^{\otimes n}$. Si le paramètre d'intérêt $\psi(\theta)$ s'écrit sous la forme $\psi(\theta) = \mathscr{F}(p_\theta)$, un estimateur définie par $\widehat{\psi(\theta)} = \mathscr{F}(\hat{P}_n)$ est un *estimateur par insertion* de $\psi(\theta)$

▷ EXEMPLE. L'inverse généralisé d'une fonction de répartition est la fonction

$$F^{-1}(u) \coloneqq \inf\{x \in \mathbf{R} \mid F(x) \geqslant u\}, \quad u \in [0,1].$$

Elle vérifie les propriétés suivantes :

- elle est croissante et continue à gauche ;
- pour tout $u \in [0,1]$, on a $F(x) \geqslant u \Leftrightarrow x \geqslant F^{-1}(u)$;
- pour tout $u \in [0,1]$, on a $F \circ F^{-1}(u) \geqslant u$;
- si F est continue, alors $F \circ F^{-1} = \mathrm{Id}_{[0,1]}$;
- si F est injective, alors $F^{-1} \circ F = \mathrm{Id}_{\mathbf{R}}$;
- si F est continue et $X \sim P$, alors $F(X) \sim \mathscr{U}([0,1])$;
- si $U \sim \mathscr{U}([0,1])$, alors $F^{-1}(U) \sim P$.

Dans le modèle statistique $\mathfrak{M}$, on considère un n-échantillon X tel que $X \sim P_\theta = p_\theta^{\otimes n}$. On note F_θ la fonction de répartition de p_θ. Alors $\hat{F}_n(X)$ est un estimateur par insertion de F_θ. Donnons d'autres exemples. La moyenne empirique et la variance empirique

$$\overline{X} \coloneqq \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad V_n(X) \coloneqq \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$$

sont des estimateurs par insertion de $\mathbb{E}_\theta[X_i]$ et $\mathrm{Var}_\theta[X_i]$.

DÉFINITION 3.9. Soit $X \coloneqq (X_1, \dots, X_n)$ un n-échantillon tel que $X \sim P_\theta = p_\theta^{\otimes n}$. Un *estimateur des moments* de θ est une solution du système d'équations

$$\mathbb{E}_\theta[X_1] = \frac{1}{n}\sum_{i=1}^{n} X_i^j, \quad j \in [\![1, d]\!].$$

3.2 Vraisemblance

DÉFINITION 3.10. Le modèle $\mathfrak{M}$ est *dominé* par une mesure σ-finie ν sur l'espace $(\mathscr{X}, \mathscr{B})$ si, pour tout $\theta \in \Theta$, on a $P_\theta \ll \nu$.

DÉFINITION 3.11. On suppose que le modèle $\mathfrak{M}$ est dominée par une mesure σ-finie ν. Une *vraisemblance* est une fonction mesurable $L\colon \mathscr{X} \times \Theta \longrightarrow [0, +\infty[$ telle que

$$\forall x \in \mathscr{X},\ \forall \theta \in \Theta, \quad L(x,\theta) = \frac{\mathrm{d}P_\theta}{\mathrm{d}\nu}(x).$$

La *log-vraisemblance* est la fonction $\ell \coloneqq \ln L$. Un *estimateur du maximum de vraisemblance* de θ est

une statistique T telle que

$$\forall x \in \mathscr{X}, \quad T(x) \in \arg\max_{\theta\in\Theta} L(x,\theta).$$

▷ Exemple. Si on considère n variables aléatoires X_i suivant une loi $\mathrm{Ber}(\theta)$, alors la mesure dominante est la mesure de comptage sur $\{0,1\}^n$ et la vraisemblance s'écrit

$$L(X_1,\dots,X_n,\theta) = \prod_{i=1}^{n} \theta^{X_i}(1-\theta)^{1-X_i} = \theta^{X_1+\cdots+X_n}(1-\theta)\theta^{n-X_1-\cdots-X_n}.$$

Proposition 3.12. Soient L la vraisemblance du modèle $\mathfrak{M}$ dominée par une mesure ν et X un variable aléatoire. Alors pour tout $\theta \in \Theta$, on a

$$L(X,\theta) > 0 \qquad P_\theta\text{-presque sûrement.}$$

Définition 3.13. Le modèle $\mathfrak{M}$ avec $\Theta \subset \mathbf{R}^d$ et dominé par une mesure ν est *exponentiel* de dimension $d' \in \mathbf{N}^*$ s'il existe deux fonctions mesurables $h\colon \Theta \longrightarrow \mathbf{R}$ et $g\colon \Theta \longrightarrow \mathbf{R}^{d'}$ et deux statistiques $T\colon (\mathscr{X},\mathscr{B}) \longrightarrow \mathbf{R}^{d'}$ et $c\colon (\mathscr{X},\mathscr{B}) \longrightarrow]0,+\infty[$ telles que

$$L(x,\theta) = \frac{\mathrm{d}P_\theta}{\mathrm{d}\nu}(x) = c(x)\exp(\langle g(x), T(x)\rangle - h(\theta)).$$

◇ Remarque. Un tel modèle est alors dominé par la mesure ν' donnée par $\mathrm{d}\nu'(x) = c(x)\,\mathrm{d}\nu(x)$ et il est paramétrable par la quantité $\theta' \coloneqq g(\theta)$. On peut alors écrire

$$\frac{\mathrm{d}P_\theta}{\mathrm{d}\nu'}(x) = \exp(\langle g(x), T(x)\rangle - h(\theta')).$$

Définition 3.14. Le modèle $\mathfrak{M}$ avec $\Theta \subset \mathbf{R}^d$ et dominé par une mesure ν est *exponentiel canonique* de dimension $d' \in \mathbf{N}^*$ s'il existe une statistique $T\colon (\mathscr{X},\mathscr{B}) \longrightarrow \mathbf{R}^{d'}$ telle que

$$\lambda(\theta) \coloneqq \int_{\mathscr{X}} \exp(\langle \theta, T(x)\rangle)\,\mathrm{d}\nu(x) < +\infty$$

et

$$L(x,\theta) = \frac{\mathrm{d}P_\theta}{\mathrm{d}\nu}(x) = \frac{1}{\lambda(\theta)}\exp(\langle \theta, T(x)\rangle).$$

Proposition 3.15. Le modèle $\mathfrak{M}$ est exponentiel canonique de dimension d' par rapport à une mesure ν est identifiable si et seulement si, pour tout $\theta \in \Theta$, la fonction $x \in \mathscr{X} \longmapsto \langle \theta, T(x)\rangle$ n'est pas ν-presque sûrement constante.

Dans la suite, on suppose que le modèle est dominé par une mesure σ-finie ν. On note L sa vraisemblance et ℓ sa log-vraisemblance. De plus, on suppose que toutes les lois P_θ ont le même support et que l'ensemble Θ est un ouvert de $\mathbf{R}^{d'}$.

Définition 3.16. On suppose que, pour tout $\theta \in \Theta$,

– pour P_θ-presque tout $x \in \mathscr{X}$, le gradient $\nabla_\theta \ell(x,\theta)$ existe ;
– la fonction $(\nabla_\theta \ell(\cdot,\theta))^2$ est P_θ-intégrable.

L'*information de Fischer* du modèle $\mathfrak{M}$ est la fonction

$$I\colon \left|\begin{array}{l}\Theta \longrightarrow \mathscr{M}_{d'}(\mathbf{R}),\\ \theta \longmapsto \mathrm{Var}_\theta[\nabla \ell(X,\theta)].\end{array}\right.$$

Proposition 3.17. On se place sous les hypothèses précédentes. De plus, on suppose que, pour toute fonction $h\colon X \longrightarrow \mathbf{R}$ P_θ intégrable et pour tout $i \in [\![1,d']\!]$, on a

$$\frac{\partial}{\partial\theta_i}\int_{\mathscr{X}} h(x)L(x,\theta)\,\mathrm{d}\nu(x) = \int_{\mathscr{X}} h(x)\frac{\partial}{\partial\theta_i}L(x,\theta)\,\mathrm{d}\nu(x).$$

Alors

– $\mathbb{E}_\theta[\nabla\ell(X,\theta)] = 0$;

– le coefficient (i,j) de la matrice $I(\theta)$ vaut

$$I(\theta)_{i,j} = \mathbb{E}\Big[\frac{\partial}{\partial\theta_i}L(x,\theta)\frac{\partial}{\partial\theta_j}L(x,\theta)\Big].$$

PROPOSITION 3.18. On suppose que le modèle $\mathfrak{M}$ satisfait les points suivants :
– on écrit $\mathscr{X} = \mathscr{X}_1^n$;
– pour tout $\theta \in \Theta$, on écrit $P_\theta = p_\theta^{\otimes n}$ où la mesure p_θ est une loi sur $\mathscr{X}_1$;
– le modèle $(\mathscr{X}_1, \{p_\theta\}_{\theta\in\Theta})$ est dominé par une mesure σ-finie ν_1, de vraisemblance L_1, de log-vraisemblance ℓ_1 et d'information de Fischer I_1.

Alors le modèle $\mathfrak{M}$ est
– dominé par la mesure $\nu \coloneqq \nu_1^{\otimes n}$,
– de vraisemblance $L(x,\theta) = \prod_{j=1}^n L_1(x_j,\theta)$,
– de log-vraisemblance $\ell(x,\theta) = \sum_{j=1}^n \ell(x_j,\theta)$.

Chapitre 4

Estimateurs ponctuels

On considère une suite $(X_i)_{i\in\mathbf{N}^*}$ de variables aléatoires réelles indépendantes et identiquement distribuées à valeurs dans un ensemble $\mathscr{X}_1 \subset \mathbf{R}^d$. Fixons un entier $n \in \mathbf{N}^*$. Soit $(\mathscr{X}, \{p_\theta\}_{\theta\in\Theta})$ un modèle statistique associé à $X \coloneqq (X_1, \dots, X_n)$. Soit $\psi\colon \Theta \longrightarrow \mathbf{R}^{d''}$ une fonction quelconque. On souhaite estimer le paramètre d'intérêt $\psi(\theta) \in \mathbf{R}^{d''}$.

4.1 Critères de performances asymptotiques

Définition 4.1. Un estimateur $\hat{\psi} \coloneqq T(X)$ est

- un estimateur *faiblement consistant* de $\psi(\theta)$ si, pour tout $\theta \in \Theta$, on a $\hat{\psi} \xrightarrow{P_\theta} \psi(\theta)$;
- un estimateur *fortement consistant* de $\psi(\theta)$ si, pour tout $\theta \in \Theta$, on a $\hat{\psi} \xrightarrow{P_\theta\text{-ps}} \psi(\theta)$;
- un estimateur *asymptotiquement sans biais* de $\psi(\theta)$ si, pour tout $\theta \in \Theta$, on a $\mathbb{E}_\theta[\hat{\psi}] \longrightarrow \psi(\theta)$.

Définition 4.2. Soit $(v_n)_{n\in\mathbf{N}^*}$ une suite réel positive qui tend vers l'infini. L'estimateur $\hat{\psi}$ est de *vitesse* v_n si, pour tout $\theta \in \Theta$, il existe une loi λ_θ non dégénérée sur $\mathbf{R}^{d''}$ telle que

$$v_n(\hat{\psi} - \psi(\theta)) \xrightarrow{\text{loi}/P_\theta} \lambda_\theta.$$

Si toutes les lois λ_θ sont normales, l'estimateur est dit *asymptotiquement normal*.

◇ Remarque. Par le lemme de Slutsky, un estimateur $\hat{\psi}$ de vitesse v_n et de loi limite λ_θ est faiblement consistant.

Théorème 4.3 *(méthode delta).* On suppose que $\Theta \subset \mathbf{R}^{d'}$ et que la fonction ψ est de classe $\mathscr{C}^1$. Soit $\hat{\theta}$ un estimateur de θ de vitesse v_n, *i. e.*

$$v_n(\hat{\theta} - \theta) \xrightarrow{\text{loi}/P_\theta} \lambda_\theta, \quad \theta \in \mathbf{R}^{d'}.$$

Alors pour tout $\theta \in \mathbf{R}^{d'}$, on a

$$v_n(\psi(\hat{\theta}) - \psi(\theta)) \xrightarrow{\text{loi}/P_\theta} \mathrm{J}_\psi(\theta)\lambda_\theta$$

où la matrice $\mathrm{J}_\psi(\theta)$ est la jacobienne de la fonction ψ au point θ.

Preuve Soit $\theta \in \mathbf{R}^{d'}$. On va utiliser la formule de Taylor avec le reste intégral et on obtient

$$v_n(\psi(\hat{\theta}) - \psi(\theta)) = v_n h_\theta(\hat{\theta})(\hat{\theta} - \theta)$$

où

$$h_\theta(t) \coloneqq \int_0^1 \mathrm{J}_\psi(\theta + u(t-\theta))\,\mathrm{d}u.$$

Il suffit de montrer

$$h_\theta(\hat{\theta}) \xrightarrow{\text{loi}/P_\theta} h_\theta(\theta) = \mathrm{J}_\psi(\theta) \qquad (*)$$

puisque le lemme de Slutsky nous donnera alors la conclusion. Comme $\hat{\theta} \xrightarrow{\text{loi}/P_\theta} \theta$ par notre hypothèse et la fonction h_θ est continue puisque la fonction ψ est de classe $\mathscr{C}^1$, on obtient bien la convergence $(*)$. Ceci montre le théorème. □

Proposition 4.4. On suppose que les variables aléatoires X_i admettent un moment d'ordre 2. Notons

$$\mu \coloneqq \mathbb{E}_\theta[X_i] \quad \text{et} \quad \sigma^2 \coloneqq \mathrm{Var}_\theta[X_i].$$

1. La moyenne empirique $m_1(X) \coloneqq \overline{X}$ est un estimateur fortement consistant et asymptotiquement

normal de μ. Plus précisément, on a

$$\sqrt{n}(\overline{X} - \mu) \xrightarrow{\text{loi}/P_\theta} \mathscr{N}(0, \sigma^2).$$

2. La variance empirique $m'_2(X) \coloneqq V_n(X)$ est un estimateur fortement consistant et asymptotiquement sans bien de σ^2. De plus, si les variables aléatoires X_i admettent un moment d'ordre 4, alors il est asymptotique normal et on a

$$\sqrt{n}(V_n(X) - \sigma^2) \xrightarrow{\text{loi}/P_\theta} \mathscr{N}(0, \mathrm{Var}_\theta[(X_1 - \mu)^2]).$$

PROPOSITION 4.5. On note F_θ la fonction de répartition de la loi P_θ. Soit $x \in \mathbf{R}$. Alors

1. on a $n\hat{F}_n(x) \sim \mathrm{Bin}(n, F_\theta(x))$;
2. l'estimateur $\hat{F}_n(x)$ de $F_\theta(x)$ est fortement consistant, asymptotiquement sans bias et asymptotiquement normal ; on a

$$\sqrt{n}(\hat{F}_n(x) - F_\theta(x)) \xrightarrow{\text{loi}/P_\theta} \mathscr{N}(0, F_\theta(x)(1 - F_\theta(x))).$$

THÉORÈME 4.6 *(Kolmogorov-Smirnov).* Pour tout $\theta \in \Theta$, on a

$$\sqrt{n}\|\hat{F}_n - F_\theta\|_\infty \xrightarrow{\text{loi}/P_\theta} K$$

où la variable aléatoire K suit la loi de fonction de répartition F_K définie par la relation

$$F_K(x) = \left(1 + 2\sum_{k=1}^{+\infty}(-1)^k e^{-2k^2x^2}\right)\mathbb{1}_{x>0}.$$

THÉORÈME 4.7. Soit $u \in]0,1[$.

1. Si F_θ est strictement croissante au voisinage de $F_\theta^{-1}(u)$, alors $\hat{F}_n^{-1}(u)$ est un estimateur fortement consistant de F_θ^{-1}.
2. Si F_θ est dérivable en $F_\theta^{-1}(u)$, de dérivée $f_\theta(F_\theta^{-1}) > 0$, alors $\hat{F}_n^{-1}(u)$ est un estimateur asymptotiquement sans biais et asymptotiquement normal de $F_\theta^{-1}(u)$; on a

$$\sqrt{n}(F_n^{-1}(u) - F_\theta^{-1}(u)) \xrightarrow{\text{loi}/P_\theta} \mathscr{N}\left(0, \frac{u(1-u)}{f_\theta(F_\theta^{-1}(u))^2}\right).$$

Preuve 1. On utilise le lemme de Borel-Cantelli. Soit $\varepsilon > 0$. Alors

$$P_\theta(|\hat{F}_n^{-1}(u) - F_\theta^{-1}(u)| > \varepsilon) = P_\theta(\hat{F}_n^{-1}(u) < F_\theta^{-1}(u) - \varepsilon) + P_\theta(\hat{F}_n^{-1}(u) > F_\theta^{-1}(u) + \varepsilon).$$

Comme $F \circ F^{-1} \geqslant \mathrm{Id}$ pour une fonction de répartition F, le premier terme se majore et on trouve

$$\begin{aligned}
P_\theta(\hat{F}_n^{-1}(u) < F_\theta^{-1}(u) - \varepsilon) &\leqslant P_\theta(n\hat{F}_n(F_\theta^{-1}(u) - \varepsilon) > n\hat{F}_n(F_n^{-1}(u))) \\
&\leqslant P_\theta(n\hat{F}_n(F_\theta^{-1}(u) - \varepsilon) > nu) \\
&\leqslant P_\theta(n\hat{F}_n(F_\theta^{-1}(u) - \varepsilon) - nF_\theta(F_\theta^{-1}(u) - \varepsilon) > nu - nF_\theta(F_\theta^{-1}(u) - \varepsilon)) \\
&\leqslant \exp(-2n(u - F_\theta(F_\theta^{-1}(u) - \varepsilon))^2)
\end{aligned}$$

par l'inégalité de Hoeffding. Ce dernier terme est le terme général d'une série convergente. Par ailleurs, on a

$$\begin{aligned}
P_\theta(\hat{F}_n^{-1}(u) > F_\theta^{-1}(u) + \varepsilon) &= P_\theta(nu > n\hat{F}_n(F_\theta^{-1}(u) + \varepsilon)) \\
&= P_\theta(n(F_n(F_\theta^{-1}(u) + \varepsilon) - F_\theta(F_\theta^{-1}(u) + \varepsilon)) < n(u - F_\theta(F_\theta^{-1}(u) + \varepsilon))) \\
&\leqslant \exp(-2n(u - F_\theta(F_\theta^{-1}(u) + \varepsilon))^2).
\end{aligned}$$

Ce dernier terme est aussi le terme général d'une série convergente. Finalement, la série

$$\sum_{n \in \mathbf{N}} P_\theta(|\hat{F}_n^{-1}(u) - F_\theta^{-1}(u)| > \varepsilon)$$

converge. Par le lemme de Borel-Cantelli, on obtient le résultat.

□

THÉORÈME 4.8. Sous les hypothèse
- le modèle $(\mathscr{X}_1, \{p_\theta\})$ est identifiable ;
- Θ est compact ;
- pour tout $x_1 \in \mathscr{X}_1$, la fonction $\theta \longmapsto \ell_1(x_1, \theta)$ est continue ;
- pour tout $\theta \in \Theta$, il existe un voisinage V de θ et $H \in \mathrm{L}^1(p_\theta)$ tels que
$$\forall x_1 \in \mathscr{X}_1, \quad \sup_{\theta' \in V} |\ell_1(x_1, \theta')| \leqslant H(x_1),$$

tout estimateur du maximum de vraisemblance $\hat{\theta}$ est fortement consistant.

4.2 Critères de performances non asymptotiques

DÉFINITION 4.9. Une statistique $T \colon X \longrightarrow \mathbf{R}^{d''}$ est d'ordre $p \in \mathbf{N}^*$ si
$$T \in \mathrm{L}^p(P_\theta) \quad \text{pour tout } \theta \in \Theta.$$

DÉFINITION 4.10. Le *biais* d'un estimateur $\hat{\psi} = T(X)$ de $\psi(\theta)$ d'ordre 1 est la quantité
$$B_\theta(\hat{\psi}, \psi(\theta)) \coloneqq \mathbb{E}_\theta[\hat{\psi}] - \psi(\theta), \quad \theta \in \Theta.$$
L'estimateur $\hat{\psi}$ est dit sans biais si $B_\theta(\hat{\psi}, \psi(\theta)) = 0$ pour tout $\theta \in \Theta$. Dans le cas contraire, on dit qu'il est biaisé.

◇ REMARQUE. Il se peut qu'aucun estimateur sans biais n'existe. Par exemple, on peut prendre le modèle statistique
$$(\llbracket 0, n \rrbracket, \mathscr{P}(\llbracket 0, n \rrbracket), \{\mathrm{Bin}(n, 1/\theta)\}_{\theta \geqslant 1}).$$

DÉFINITION 4.11. Une *fonction de perte ou coût* pour l'estimation du $\psi(\theta)$ est, en notant $\Psi \coloneqq \psi(\Theta)$, une fonction $c \colon \Psi \longrightarrow \mathbf{R}_+$ mesurable telle que $c(\psi_1, \psi_2) = 0$ si $\psi_1 = \psi_2$.

Le *risque* associé à cette fonction de perte ou de coût c est la quantité
$$R^c_\theta(\hat{\psi}, \psi(\theta)) \coloneqq \mathbb{E}_\theta[c(\hat{\psi}, \psi(\theta))], \quad \theta \in \Theta.$$
Le *risque maximal* sur un sous-ensemble $\Theta_0 \subset \Theta$ est la quantité
$$R^c(\hat{\theta}, \Theta_0) = \sup_{\theta \in \Theta_0} R^c_\theta(\hat{\psi}, \psi(\theta)).$$
Le *risque minimax* sur un sous-ensemble $\Theta_0 \subset \Theta$ est la quantité
$$mR^c(\Theta_0) = \sup_{\hat{\psi}} R^c(\hat{\psi}, \Theta_0).$$

◇ REMARQUE. Lorsque l'exposant c ne sera pas indiqué dans la notation du risque, le coût sera l'application $(\psi_1, \psi_2) \longmapsto \|\psi_1 - \psi_2\|_2^2$ et le risque sera alors qualifié de quadratique. De plus, avec ce coût, on peut écrire la décomposition
$$R_\theta(\hat{\psi}, \psi(\theta)) = \|B_\theta(\hat{\psi}, \psi(\theta)\|^2 + \mathbb{E}_\theta[\|\hat{\psi} - \mathbb{E}_\theta[\hat{\psi}]\|^2]$$

DÉFINITION 4.12. Un estimateur $\hat{\psi}$ est préférable à un autre estimateur $\tilde{\psi}$ si
$$\forall \theta \in \Theta, \quad R^c_\theta(\hat{\psi}, \psi(\theta)) \leqslant R^c_\theta(\tilde{\psi}, \psi(\theta))$$

DÉFINITION 4.13. Un estimateur d'ordre 2 est dit de variance uniformément minimale parmi les estimateurs sans biais (VUMSB) de $\psi(\theta)$ s'il est sans biais et préférable à tout autre estimateur sans biais d'ordre 2 de $\psi(\theta)$.

PROPOSITION 4.14. Soient $\hat{\psi}$ et $\tilde{\psi}$ deux estimateurs VUMSB de $\psi(\theta)$. Alors pour tout $\theta \in \Theta$, on a
$$\hat{\psi} = \tilde{\psi} \qquad P_\theta\text{-presque sûrement.}$$

Preuve Soit $\theta \in \Theta$. Alors

$$\begin{aligned}\mathbb{E}_\theta[\|\hat{\psi} - \tilde{\psi}\|^2] &= \mathbb{E}_\theta[\langle \hat{\psi} - \tilde{\psi}, \hat{\psi} - \tilde{\psi}\rangle] \\ &= \mathbb{E}_\theta[\langle \hat{\psi} - \psi(\theta), \hat{\psi} - \tilde{\psi}\rangle] - \mathbb{E}_\theta[\langle \tilde{\psi} - \psi(\theta), \hat{\psi} - \tilde{\psi}\rangle].\end{aligned}$$

Pour $\alpha \in \mathbf{R}$, on considère l'estimateur $\hat{\psi}_\alpha \coloneqq \hat{\psi} + \alpha(\hat{\psi} - \tilde{\psi})$ de $\psi(\theta)$. Il est d'ordre 2 et sans biais. Comme $\hat{\psi}$ est VUMSB, on a

$$\begin{aligned}R_\theta(\hat{\psi}, \psi(\theta)) &\leqslant R_\theta(\hat{\psi}_\alpha, \psi(\theta)) \\ &= R_\theta(\hat{\psi}, \psi(\theta)) + \alpha^2 \mathbb{E}_\theta[\|\hat{\psi} - \tilde{\psi}\|^2] + 2\alpha \mathbb{E}_\theta[\langle \hat{\psi} - \psi(\theta), \hat{\psi} - \tilde{\psi}\rangle].\end{aligned}$$

Cela implique $\mathbb{E}_\theta[\langle \hat{\psi} - \psi(\theta), \hat{\psi} - \tilde{\psi}\rangle] = 0$. On trouve la même égalité pour l'estimateur $\tilde{\psi}$. Ceci conduit à l'égalité $\mathbb{E}_\theta[\|\hat{\psi} - \tilde{\psi}\|^2] = 0$ et conclut la preuve. □

◇ Remarque. Un estimateur biaisé peut être préférable à un estimateur sans biais. Par exemple, soit $X \coloneqq (X_1, \dots, X_n)$ un n-échantillon de loi uniforme $\mathscr{U}([0, \theta])$. Un estimateur du maximum du vraisemblance est $\hat{\theta} \coloneqq X_{(n)}$. Mais l'estimateur sans biais $\tilde{\theta} \coloneqq \frac{n+1}{n}\theta$ est moins bon.

Exhaustivité

Définition 4.15. Une statistique T est *exhaustive* si la loi conditionnelle de X sachant $T(X)$ sous P_θ ne dépend pas de θ.

Notation. On note $T(P_X)$ la mesure image de la loi P_X par l'application T. La loi $P_{X|T(X)=t}$ est notée $P_\theta(\cdot \mid T(X) = t)$.

Théorème 4.16 *(de factorisation de Neyman-Fisher).* On suppose que le modèle est dominé par une mesure σ-finie ν et de vraisemblance L. Une statistique $T\colon \mathscr{X} \longrightarrow \mathbf{R}^{d''}$ est exhaustive si et seulement s'il existe deux fonctions $g\colon \mathscr{X} \longrightarrow \mathbf{R}_+$ et $h\colon \mathbf{R}^{d''} \times \Theta \longrightarrow \mathbf{R}_+$ mesurables telles que, pour ν-presque tout $x \in \mathscr{X}$, on ait

$$\forall \theta \in \Theta, \quad L(x, \theta) = g(x)h(T(x), \theta).$$

◇ Remarque. La statistique naturelle associée à un modèle exponentielle est donc exhaustive.

Théorème 4.17 *(Rao-Blackwell).* Soient T une statistique exhaustive et $\hat{\psi}$ un estimateur de $\psi(\theta)$ d'ordre 2. Alors l'estimateur $\tilde{\psi} \coloneqq \mathbb{E}_\theta[\hat{\psi} \mid T(X)]$ de $\psi(\theta)$ est de même bias que $\hat{\psi}$ et préférable à $\hat{\psi}$.

Preuve Soit $\theta \in \Theta$. Par l'exhaustivité de T, la variable aléatoire $\tilde{\psi}$ ne dépend pas de θ et elle est mesurable, donc il s'agit bien d'un estimateur. Comme $\mathbb{E}_\theta[\tilde{\psi}] = \mathbb{E}_\theta[\hat{\psi}]$, il a le même biais que $\hat{\psi}$. Montrons qu'il est préférable à ce dernier. On a

$$\mathbb{E}_\theta[\|\hat{\psi} - \mathbb{E}_\theta[\hat{\psi}]\|^2] = \mathbb{E}_\theta[\|\hat{\psi} - \tilde{\psi}\|^2] + \mathbb{E}_\theta[\|\tilde{\psi} - \mathbb{E}_\theta[\tilde{\psi}]\|^2] + 2\mathbb{E}_\theta[\langle \hat{\psi} - \tilde{\psi}, \tilde{\psi} - \mathbb{E}_\theta[\tilde{\psi}]\rangle].$$

Or comme $\mathbb{E}_\theta[\hat{\psi} - \tilde{\psi} \mid T(X)] = 0$, on trouve

$$\mathbb{E}_\theta[\langle \hat{\psi} - \tilde{\psi}, \tilde{\psi} - \mathbb{E}_\theta[\tilde{\psi}]\rangle \mid T(X)] = \langle \mathbb{E}_\theta[\hat{\psi} - \tilde{\psi} \mid T(X)], \tilde{\psi} - \mathbb{E}_\theta[\tilde{\psi}]\rangle = 0.$$

D'où

$$\mathbb{E}_\theta[\|\hat{\psi} - \mathbb{E}_\theta[\hat{\psi}]\|^2] = \mathbb{E}_\theta[\|\hat{\psi} - \tilde{\psi}\|^2] + \mathbb{E}_\theta[\|\tilde{\psi} - \mathbb{E}_\theta[\tilde{\psi}]\|^2].$$

Grâce à la décomposition biais-variance et comme les estimateurs ont le même biais, on en déduit

$$R_\theta(\hat{\psi}, \psi(\theta)) \geqslant R_\theta(\tilde{\psi}, \psi(\theta)).$$ □

Complétude

Définition 4.18. Une statistique $T\colon \mathscr{X} \longrightarrow \mathbf{R}^{d''}$ est complète si, pour toute fonction $f\colon \mathbf{R}^{d''} \to \mathbf{R}$ telle que $f \circ T$ est d'ordre 1, on a

$$(\forall \theta \in \Theta,\ \mathbb{E}_\theta[h \circ T(X)] = 0) \quad \Longrightarrow \quad (\forall \theta \in \Theta,\ h \circ T(X) = 0 \quad P_\theta\text{-presque sûrement})$$

THÉORÈME 4.19 *(Lehmann-Scheffé).* Soit $\hat{\psi}$ un estimateur sans biais d'ordre 2 de $\psi(\theta)$. Soit T un statistique exhaustive complète. Alors l'estimateur $\mathbb{E}_\theta[\hat{\psi} \mid T(X)]$ est l'unique estimateur VUMSB de $\psi(\theta)$.

PROPOSITION 4.20. On suppose que le modèle est exponentiel de dimension 1, dominé par une mesure σ-finie ν et de vraisemblance

$$L(x, \theta) = c(x) \exp()$$

Chapitre 5

Estimation par régions de confiance

5.1 Régions de confiance non asymptotiques

Soit $(\mathscr{X}, \mathscr{B}, \{P_\theta\}_{\theta\in\Theta})$ un modèle statistique associé à une variable aléatoire $X\colon \Omega \longrightarrow \mathscr{X}$ suivant une loi P_θ où le paramètre $\theta \in \Theta$ est inconnu.

Définition 5.1. Une *région de confiance* pour le paramètre $\psi(\theta)$ est une famille $(C(x))_{x\in\mathscr{X}}$ de partie de l'image $\psi(\Theta)$ telle que, pour tout $\theta \in \Theta$, l'ensemble $\{x \in \mathscr{X} \mid \psi(\theta) \in C(x)\}$ soit mesurable ; on notera alors

$$P_\theta(\psi(\theta) \in C(X)) \coloneqq P_\theta(\{x \in \mathscr{X} \mid \psi(\theta) \in C(x)\}).$$

Soit $\alpha \in [0, 1]$. Cette région est de *niveau de confiance* $1 - \alpha$ si

$$\inf_{\theta\in\Theta} P_\theta(\psi(\theta) \in C(X)) \geqslant 1 - \alpha.$$

Par abus de langage, les parties $C(x)$ avec $x \in \mathscr{X}$ sera aussi appelées des régions de confiance pour le paramètre $\psi(\theta)$.

Définition 5.2. On suppose que la fonction ψ est à valeurs réelles. Un *intervalle de confiance* pour le paramètre $\psi(\theta)$ avec une région de confiance pour ce même paramètre dont les parties sont des intervalles

Définition 5.3. La *probabilité de recouvrement* d'une région de confiance $(C(x))_{x\in\mathbf{C}}$ pour le paramètre $\psi(\theta)$ est la fonction

$$\theta \in \Theta \longmapsto P_\theta(\psi(\theta) \in C(X)).$$

Dans la suite, on fixe un réel $\alpha \in [0, 1]$. Remarquons qu'une région de confiance dont la probabilité de recouvrement est minorée par $1 - \alpha$ est de niveau de confiance $1 - \alpha$. Pour construire des régions de confiance, on peut utiliser diverses inégalités de concentration ou les racines pivotales.

Définition 5.4. Une *racine pivotale* pour le paramètre $\psi(\theta)$ est une fonction

$$R\colon \mathscr{X} \times \psi(\Theta) \longrightarrow \psi(\Theta)$$

telle que, pour tout $\theta \in \Theta$,

- la fonction $R(X, \psi(\theta))$ soit une variable aléatoire ;
- sa loi est indépendante du paramètre θ.

Théorème 5.5. Soient R une racine pivotale pour le paramètre $\psi(\theta)$ et Q la loi commune des variables aléatoires $R(X, \psi(\theta))$. Soit $\Psi_\alpha \subset \psi(\Theta)$ un événement tel que $Q(\Psi_\alpha) \geqslant 1 - \alpha$. Alors la région de confiance $(C(x))_{x\in\mathscr{X}}$ avec

$$C(x) \coloneqq \{t \in \psi(\Theta) \mid R(x, t) \in \Psi_\alpha\} \quad \text{avec} \quad x \in \mathscr{X}$$

est de niveau de confiance $1 - \alpha$.

Critère de performance

Définition 5.6. Un loi Q de densité f par rapport à la mesure de Lebesgue sur $\mathbf{R}^{d''}$ est *unimodale* si, pour tout $c > 0$, l'ensemble $\{f \geqslant c\} \subset \mathbf{R}^{d''}$ est convexe.

▷ Exemples. Les lois uniformes et les lois normales sur $\mathbf{R}^d$ sont unimodales.

PROPOSITION 5.7. Un loi Q de densité f par rapport à la mesure de Lebesgue sur $\mathbf{R}$ est unimodale si et seulement s'il existe un réel $m \in \mathbf{R}$, appelé un *mode* de la loi Q, tel que la fonction f soit croissante sur $]-\infty, m[$ et décroissante sur $]m, +\infty[$.

PROPOSITION 5.8. Soit Q une loi unimodale de densité f par rapport à la mesure de Lebesgue λ sur $\mathbf{R}^{d''}$. Soit $I \subset \mathbf{R}^{d''}$ un sous-ensemble convexe tel que $\{f > c\} \subset I \subset \{f \geqslant c\}$ pour un certain réel $c > 0$. Alors pour tout sous-ensemble $J \subset \mathbf{R}^{d''}$ tel que $Q(J) \geqslant Q(I)$, on a $\lambda(J) \geqslant \lambda(I)$.

DÉFINITION 5.9. La *probabilité de faux recouvrement* d'une région de confiance $(C(x))_{x \in X}$ pour le paramètre $\psi(\theta_0)$ est la fonction

$$\theta \in \Theta \setminus \{\theta_0\} \longmapsto P_\theta(\psi(\theta_0) \in C(X)).$$

5.2 Régions de confiance asymptotiques

DÉFINITION 5.10. Une région de confiance asymptotique pour le paramètre $\psi(\theta)$ est une suite $(C_n)_{n \in \mathbf{N}}$ de régions de confiance pour ce paramètre. Elle est de niveau de confiance $1 - \alpha$ si

$$\forall \theta \in \Theta, \quad \liminf_{n \to +\infty} P_\theta(\psi(\theta) \in C_n(X)) \geqslant 1 - \alpha.$$

De même, on peut définir la notion de racine pivotale asymptotique $(R_n)_{n \in \mathbf{N}}$.