Université Pierre et Marie Curie **IFMA/MPE**
2016/2017

Statistique de base

Ce document rassemble des résultats élémentaires de statistiques. Il reprend une grande partie des notes du cours de Bertrand Michel, et s'appuie sur le livre *Statistique en action* de Rivoirard et Stoltz (Vuibert, 2009).

1 Le Modèle statistique

Etant donnée une certaine expérience aléatoire, le statisticien construit d'abord un modèle statistique censé modéliser cette expérience. L'observation $\boldsymbol{Y}$ est le résultat de cette expérience. Si $\boldsymbol{y}$ est une réalisation de $\boldsymbol{Y}$, on aimerait s'aider de cette information pour en déduire la loi de $\boldsymbol{Y}$. Si nous ne faisons aucune hypothèse sur la loi de $\boldsymbol{Y}$, on dit que le modèle est non-paramétrique. Si on suppose que la loi de $\boldsymbol{Y}$ est de forme connue mais dépend d'un nombre fini de paramètres réels qui sont inconnus, on dira que le modèle est paramétrique.

Soit $(E, \mathcal{E})$ un espace mesurable.

Définition. *On appelle modèle statistique la donnée de* $(E, \mathcal{E}, (\mathrm{P}_\theta)_{\theta\in\boldsymbol{\Theta}})$ *où* $(\mathrm{P}_\theta)_{\theta\in\boldsymbol{\Theta}}$ *désigne une famille de lois de probabilités sur* $(E, \mathcal{E})$. *On notera* E_θ *l'espérance associée à* P_θ *et* V_θ *la variance.*

Si $\Theta \subset \mathbb{R}^d$, on dit que le modèle est *paramétrique*. Sinon le modèle est dit *non-paramétrique*.

Définition. *Une observation* $\boldsymbol{Y}$ *est une variable aléatoire à valeurs dans* $(E, \mathcal{E})$ *dont la loi appartient à la famille de lois* $(\mathrm{P}_\theta)_{\theta\in\boldsymbol{\Theta}}$.

Définition. *Lorsque l'observation* $\boldsymbol{Y}$ *a la forme* $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ *avec* $(Y_i)_{1\leq i\leq n}$ *indépendantes et identiquement distribuées, on parlera d'échantillon. Dans ce cas,* $\mathrm{P}_\theta = \mathrm{p}_\theta^{\otimes \mathrm{n}}$ *où* p_θ *est la loi de* Y_1.

Exemples.
Sur $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ on considère $\mathrm{P}_\theta = \mathcal{N}(\mu, \sigma^2)$ avec $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_+$. L'observation $\boldsymbol{Y}$ est la donnée d'une variable Gaussienne $\mathcal{N}(\mu, \sigma^2)$.

Modèle d'un lancer de pièces. On lance une pièce n fois. On a $E = \{0, 1\}^n$, $\mathcal{E}$ la tribu triviale, p_θ est la loi de Bernoulli de paramètre $\theta \in \boldsymbol{\Theta} = [0, 1]$ et $\mathrm{P}_\theta = \mathrm{p}_\theta^{\otimes \mathrm{n}}$. L'observation $\boldsymbol{Y}$ est un échantillon de Bernoulli.

Soit une certaine variable aléatoire sans atomes à valeurs dans $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ dont on cherche à connaître la loi. Dans ce cas, on peut prendre pour $\boldsymbol{\Theta}$ l'ensemble des lois sans atomes ou de manière équivalente l'ensemble des fonctions θ continues croissantes de $\mathbb{R}$ sur $[0, 1]$ avec $\theta(-\infty) = 0$ et $\theta(+\infty) = 1$, puis P_θ est la loi de fonction de répartition θ. Ce modèle est non-paramétrique.

2 Estimateurs

2.1 Premières propriétés

Soit g une fonction de $\mathbf{\Theta}$ dans $\mathbb{R}^p$.

Définition. *Un estimateur $\hat{g}$ de $g(\theta)$ est toute application mesurable en l'observation $\boldsymbol{Y}$ ne dépendant pas de θ.*

On peut donc écrire $\hat{g} = h(\boldsymbol{Y})$ pour une certaine fonction mesurable $h : (E, \mathcal{E}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}^p))$.

Définition.

- *La fonction de biais d'un estimateur $\hat{g}$ est (si elle existe) la fonction $\theta \in \mathbf{\Theta} \to b(\theta) := \mathrm{E}_\theta[\hat{g}] - \theta$.*
- *L'erreur quadratique d'un estimateur $\hat{g}$ est la fonction $\theta \in \mathbf{\Theta} \to R(\theta) := \mathrm{E}_\theta[(\hat{g} - g(\theta))^2]$. Sa décomposition biais-variance s'écrit*

$$R(\theta) = \mathrm{V}_\theta(\hat{g}) + b(\theta)^2$$

Définition. *On dit qu'un estimateur $\hat{g}$ est* **sans biais** *si $\mathrm{E}_\theta[\hat{g}] = g(\theta)$ pour tout $\theta \in \mathbf{\Theta}$.*

La définition d'un estimateur sans biais contient que $\hat{g}$ est P_θ-intégrable pour tout $\theta \in \mathbf{\Theta}$. Un estimateur sans biais donne donc en moyenne la bonne valeur de ce qu'il estime, ce qui est satisfaisant. Cependant, c'est un critère parfois contraignant (dans certains cas un estimateur sans biais n'existe pas, dans d'autres cas un estimateur "naturel" est avec biais). L'erreur quadratique mesure la dispersion des valeurs données par l'estimateur autour de $g(\theta)$, donc sa précision.

Soit un échantillon $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ et $\hat{g}_n = h_n(Y_1, \ldots, Y_n)$ un estimateur de $g(\theta)$. Les propriétés suivantes permettent de connaître l'erreur commise par l'estimation lorsque le nombre de répétitions devient grand.

Définition. *On dit que l'estimateur $\hat{g}_n$ est*

- **asymptotiquement sans biais** *si $\lim_{n\to+\infty} \mathrm{E}_\theta[\hat{g}_n] = g(\theta)$ pour tout $\theta \in \mathbf{\Theta}$.*
- **consistant** *si pour tout $\theta \in \mathbf{\Theta}$, $\hat{g}_n$ converge vers $g(\theta)$ en probabilité.*
- **fortement consistant** *si la convergence a lieu p.s.*
- **asymptotiquement normal** *si, pour tout $\theta \in \mathbf{\Theta}$, $a_n(\hat{g}_n - g(\theta))$ converge en loi vers une loi Gaussienne centrée, pour une certaine suite déterministe a_n (dépendant éventuellement de θ) telle que $\lim_{n\to+\infty} a_n = +\infty$.*

2.2 Estimateurs classiques

On donne ici deux méthodes pour trouver des estimateurs.

2.2.1 Méthode des moments

Soit $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ un échantillon. Supposons que la quantité $g(\theta)$ que l'on cherche à estimer est donnée par $E_\theta[\phi(Y_1)]$ pour une certaine fonction mesurable ϕ (ne dépendant pas de θ) avec $\phi(Y_1)$ P_θ-intégrable pour tout $\theta \in \boldsymbol{\Theta}$. Dans ce cas,

$$\hat{g}_n = \frac{1}{n}\sum_{i=1}^{n} \phi(Y_i).$$

est un estimateur fortement consistant par la loi des grands nombres. Si la variance de $\phi(Y_1)$ sous P_θ est finie pour tout $\theta \in \boldsymbol{\Theta}$, alors $\hat{g}_n$ est asymptotiquement normal par le théorème central limite, avec $a_n = \sqrt{n}$.

Exemple. On peut ainsi estimer le moment d'ordre p (s'il existe) de p_θ $\nu_p := \mathrm{E}_\theta[Y_1^p]$ par

$$\hat{\nu}_{p,n} := \frac{1}{n}\sum_{k=1}^{n} Y_k^p.$$

Exemple. Un estimateur sans biais et fortement consistant de la variance est

$$\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{k=1}^{n}(Y_k - \overline{Y}_n)^2.$$

2.2.2 Méthode du maximum de vraisemblance

Dans cette partie, on prend $g(\theta) = \theta$ cad que l'on souhaite estimer θ. Grossièrement, étant donnée notre observation $\boldsymbol{y}$, le principe de la méthode est de trouver la valeur de θ qui maximise la probabilité de voir l'observation $\boldsymbol{y}$.

Revenons à notre modèle statistique $(E, \mathcal{E}, (\mathrm{P}_\theta)_{\theta\in\boldsymbol{\Theta}})$. On se place dans le cas paramétrique $\boldsymbol{\Theta} \subset \mathbb{R}^d$. On suppose que les lois $(\mathrm{P}_\theta, \theta \in \boldsymbol{\Theta})$ du modèle statistique sont dominées par une même mesure μ supposée σ-finie. Le théorème de Radon-Nikodym implique que P_θ a alors une densité que l'on note $L_\theta(\boldsymbol{y})$ par rapport à μ, c'est-à-dire que pour tout ensemble mesurable $A \in \mathcal{E}$, on a $\mathrm{P}_\theta(A) = \int_{\boldsymbol{y}\in A} L_\theta(\boldsymbol{y}) d\mu(\boldsymbol{y})$. Cette densité $L_\theta(\boldsymbol{y})$ est appelée la vraisemblance du modèle.

Définition. *L'estimateur du maximum de vraisemblance noté $\hat{\theta}_{\mathrm{EMV}}$ est, s'il en existe, un θ qui maximise la vraisemblance*

$$\theta \to L_\theta(\boldsymbol{Y}).$$

Dans le cas d'un échantillon $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, on supposera que la loi p_θ de Y_1 (donc aussi des Y_i) a une densité $s_\theta(y_1)$ par rapport à une mesure de référence σ-finie $\mu(dy_1)$. La vraisemblance s'écrira alors

$$L_\theta(\boldsymbol{y}) = L_\theta(y_1, \ldots, y_n) = \prod_{i=1}^{n} s_\theta(y_i)$$

et l'estimateur du maximum de vraisemblance sera le θ qui maximise la fonction $\theta \to L_\theta(\boldsymbol{Y})$.

Sous des hypothèses de régularité, on peut montrer que l'estimateur du maximum de vraisemblance est asymptotiquement normal.

2.3 Estimateur de variance minimale

Supposons que l'on veuille estimer $g(\theta)$. On aimerait comparer plusieurs estimateurs. Une idée naturelle est de comparer l'erreur commise par ces estimateurs.

Définition. *Un estimateur sans biais de variance minimale (ESBVM) est un estimateur $\hat{g}_{ESBVM}$ sans biais qui minimise l'erreur quadratique. En d'autres termes,*

$$\mathrm{V}_\theta(\hat{g}_{ESBVM}) \leq \mathrm{V}_\theta(\hat{g})$$

pour tout $\theta \in \boldsymbol{\Theta}$ et tout $\hat{g}$ estimateur sans biais de $g(\theta)$. S'il existe, un tel estimateur est nécessairement unique.

Remarque. Rien ne nous dit qu'on ne peut pas trouver un estimateur biaisé avec une erreur quadratique plus petite.

La suite de cette partie montre que l'on a une borne inférieure sur la variance des estimateurs sans biais, appelée *borne de Cramer-Rao*. Cette borne inférieure dépend de la quantité d'information apportée par les observations telle que mesurée par *l'information de Fisher*.

On se place désormais dans le cas paramétrique. On a donc $\boldsymbol{\Theta} \subset \mathbb{R}^p$ et supposons pour simplifier que $p = 1$.

Définition. *L'information de Fisher est définie par*

$$I(\theta) := \mathrm{E}_\theta \left[(\partial_\theta \ln L_\theta(\boldsymbol{Y}))^2 \right]$$

lorsque cette quantité existe. Elle peut se réécrire sous certaines hypothèses de régularité

$$I(\theta) = -\mathrm{E}_\theta \left[\partial_\theta^2 \ln L_\theta(\boldsymbol{Y}) \right].$$

Dans le cas d'un échantillon $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, on obtient (sous hypothèses de régularité) que l'information de Fisher du modèle est donnée par

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

avec $I_1(\theta) = \mathrm{E}_\theta \left[(\partial_\theta \ln s_\theta(Y_1))^2 \right]$, qui s'écrit encore $-\mathrm{E}_\theta \left[\partial_\theta^2 \ln s_\theta(Y_1) \right]$.

Intuitivement, l'information de Fisher est une mesure de l'information contenue dans l'observation $\boldsymbol{Y}$. Plus $I(\theta)$ est élevée, meilleure sera l'information et donc plus précis pourront être les estimateurs. Cette heuristique se retrouve dans le théorème suivant qui dit que l'erreur commise par un estimateur sans biais est bornée inférieurement par l'inverse de l'information de Fisher.

Théorème (Borne de Cramer-Rao). *Soit g une fonction $\boldsymbol{\Theta} \to \mathbb{R}$ dérivable, et $\hat{g} = h(\boldsymbol{Y})$ un estimateur sans biais de $g(\theta)$. Sous certaines hypothèses de régularité, on a*

$$\mathrm{V}_\theta(\hat{g}) \geq \frac{(g'(\theta))^2}{I(\theta)}.$$

La borne inférieure de cette inégalité est appelée borne de Cramer-Rao.

Un estimateur sans biais qui atteint la borne de Cramer-Rao est nécessairement de variance minimale. Un tel estimateur est dit *efficace*.

Dans le cas d'un échantillon, si $\hat{g}_n$ est sans biais et vérifie quand $n \to +\infty$, $\mathrm{V}_\theta(\hat{g}_n) \sim \frac{g'(\theta)^2}{I_n(\theta)}$, on dit que l'estimateur est *asymptotiquement efficace*. Sous des hypothèses de régularité, l'estimateur du maximum de vraisemblance est asymptotiquement efficace.

L'existence d'un estimateur efficace est en fait liée à la forme de la vraisemblance du modèle.

Théorème. *(Sous certaines hypothèses de régularité). Un estimateur efficace existe si et seulement si la vraisemblance vérifie*

$$\ln L_\theta(\boldsymbol{y}) = a(\boldsymbol{y})\alpha(\theta) + b(\boldsymbol{y}) + \beta(\theta).$$

Dans ce cas, $g(\theta) = -\frac{\beta'(\theta)}{\alpha'(\theta)}$ admet un estimateur efficace qui est $\hat{g} := a(\boldsymbol{Y})$. C'est l'unique paramètre (à une transformation linéaire près) admettant un estimateur efficace.

3 Intervalles de confiance

3.1 Définition

Soit $(E, \mathcal{E}, (\mathrm{P}_\theta)_{\theta\in\boldsymbol{\Theta}})$ un modèle statistique et $\boldsymbol{Y}$ une observation. On souhaite estimer le paramètre $g(\theta)$.

Définition. *Soit $\alpha \in [0,1]$. On appelle région de confiance au niveau $1-\alpha$ de $g(\theta)$ un ensemble $\hat{\mathcal{C}}$ construit mesurablement par rapport à $\boldsymbol{Y}$, ne dépendant pas de θ, et tel que pour tout $\theta \in \boldsymbol{\Theta}$,*

$$\mathrm{P}_\theta\left(g(\theta) \in \hat{\mathcal{C}}\right) \geq 1 - \alpha.$$

Remarque. Dire que $\hat{\mathcal{C}}$ est construit mesurablement signifie que pour tout $\theta \in \boldsymbol{\Theta}$, l'évènement $\{g(\theta) \in \hat{\mathcal{C}}\}$ est mesurable. Si l'inégalité est en fait une égalité dans la définition précédente, on parle de niveau exact.

Remarque. Lorsque $\hat{\mathcal{C}}$ est un intervalle, on parlera plutôt d'intervalle de confiance.

Définition. *Dans le cas d'un échantillon $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$, on appelle région (resp. intervalle) de confiance asymptotique au niveau $1-\alpha$ de $g(\theta)$ un ensemble (resp. un intervalle) $\hat{\mathcal{C}}_n$ construit mesurablement par rapport à $\boldsymbol{Y}$, ne dépendant pas de θ, et tel que pour tout $\theta \in \boldsymbol{\Theta}$,*

$$\liminf_{n\to+\infty} \mathrm{P}_\theta\left(g(\theta) \in \hat{\mathcal{C}}_n\right) \geq 1 - \alpha.$$

3.2 Méthode du pivot

La méthode du pivot consiste à trouver une fonction $f(\boldsymbol{y}, g(\theta))$ mesurable en $\boldsymbol{y} \in E$ dont la loi sous P_θ ne dépend pas de θ. On cherche ensuite a, b tels que $P_\theta(f(\boldsymbol{Y}, g(\theta)) \in [a, b]) \geq 1 - \alpha$. La région de confiance est alors déterminée par $\hat{\mathcal{C}} := \{g(\theta) \,:\, f(\boldsymbol{Y}, g(\theta)) \in [a, b]\}$.

Exemple. On veut estimer la moyenne μ d'une loi Gaussienne $\mathcal{N}(\mu, \sigma^2)$ où σ^2 est connue. Pour cela on a accès à un échantillon $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. Un estimateur est donné par

$$\hat{\mu}_n = \overline{Y}_n := \frac{1}{n}\left(Y_1 + \ldots + Y_n\right).$$

On sait que $\frac{\hat{\mu}_n - \mu}{\sigma/\sqrt{n}}$ suit une loi Gaussienne $\mathcal{N}(0,1)$. En notant z_β le quantile d'ordre β de la loi Gaussienne centrée réduite, on obtient $\hat{\mathcal{C}}_n := [\hat{\mu}_n \pm \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}]$ pour un intervalle de confiance bilatère de niveau $1-\alpha$. Un intervalle de confiance unilatère serait $\hat{\mathcal{C}}_n :=]-\infty; \hat{\mu}_n + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}]$ (à gauche) ou $[\hat{\mu}_n - \frac{\sigma}{\sqrt{n}} z_{1-\alpha}; +\infty[$ (à droite).

Exemple. Soit un échantillon $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ de v.a. moyenne μ et variance σ^2 toutes deux inconnues. On ne suppose plus que les v.a. suivent une loi Gaussienne. Le théorème central limite dit que $\frac{\hat{\mu}_n - \mu}{\sigma/\sqrt{n}}$ converge en loi vers $\mathcal{N}(0,1)$. On estime σ par son estimateur $\hat{\sigma}_n$. Le lemme de Slutsky entraine que $\frac{\hat{\mu}_n - \mu}{\hat{\sigma}_n/\sqrt{n}}$ converge en loi vers $\mathcal{N}(0,1)$. Cela permet d'obtenir l'intervalle de confiance asymptotique de niveau $1-\alpha$ pour μ:

$$\left[\hat{\mu}_n - \frac{\hat{\sigma}_n}{\sqrt{n}} z_{1-\alpha/2}; \hat{\mu}_n + \frac{\hat{\sigma}_n}{\sqrt{n}} z_{1-\alpha/2}\right].$$

3.3 Utilisation d'une inégalité de probabilité

Supposons que l'on veuille estimer la moyenne $\mu(\theta)$ d'une loi de probabilité dont on sait que la variance est bornée par une constante connue M^2. Pour cela on a accès à un échantillon $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$. Un estimateur sans biais est donné par

$$\hat{\mu}_n = \overline{Y}_n = \frac{1}{n}\left(Y_1 + \ldots + Y_n\right).$$

L'inégalité de Bienaymé-Tchebychev implique que

$$\mathrm{P}_\theta\left(|\hat{\mu}_n - \mu| > t\right) \leq \frac{\mathrm{V}_\theta(\hat{\mu}_n)}{t^2} \leq \frac{M^2}{nt^2}.$$

En choisissant t tel que $\frac{M^2}{nt^2} = \alpha$, on obtient l'intervalle de confiance de niveau $1-\alpha$ de μ suivant

$$\left[\hat{\mu}_n - \frac{M}{\sqrt{\alpha n}}; \hat{\mu}_n + \frac{M}{\sqrt{\alpha n}}\right].$$

Exemple. Intervalle de confiance pour le paramètre d'un échantillon de Bernoulli.

3.4 Intervalle de confiance et réalisation d'un intervalle de confiance

Il est important de garder à l'esprit qu'un intervalle de confiance est une quantité aléatoire. Dans la pratique, c'est-à-dire pour les données étudiées, les bornes de l'intervalle de confiance construit sont la réalisation de l'intervalle de confiance pour les observations. Par exemple, dans le cas de l'échantillon Gaussien i.i.d., on a vu que

$$\left[\hat{\mu}_n \pm \frac{\hat{\sigma}_n t_{0,975}}{\sqrt{n}}\right]$$

est un intervalle de confiance de la moyenne au niveau 95 %. Si $n = 100$, $\hat{\mu}_n = 7,45$ et $\hat{\sigma}_n = 1,21$, la réalisation de cet intervalle vaut $[6,76; 7,24]$. Par abus, on dira parfois que $[6,76; 7,24]$ est l'intervalle de confiance de la moyenne, mais en toute rigueur cela n'a pas de sens : soit la moyenne est dans $[6,76; 7,24]$, soit elle ne l'est pas, la probabilité que la moyenne y soit est donc 0 ou 1!

3.5 Intervalles de confiance simultanés

Supposons que l'on souhaite construire une région de confiance pour un vecteur

$$g(\theta) = (g_1(\theta), g_2(\theta), \ldots, g_k(\theta)).$$

Supposons que l'on dispose pour chaque $g_i(\theta)$ pris séparément d'un intervalle de confiance $\hat{I}_j$ de niveau de confiance $1 - \frac{\alpha}{k}$. Alors $\hat{I}_1 \times \hat{I}_2 \times \ldots \times \hat{I}_k$ est une région de confiance pour le vecteur $g(\theta)$ de niveau $1 - \alpha$ (méthode de Bonferroni). En effet, pour tout j,

$$\mathrm{P}_\theta\left(g_j(\theta) \notin \hat{I}_j\right) \leq \frac{\alpha}{k}.$$

D'où $\mathrm{P}_\theta\left(\bigcup_{j=1}^{k}\{g_j(\theta) \notin \hat{I}_j\}\right) \leq k\frac{\alpha}{k}$ et donc

$$\mathrm{P}_\theta\left(\bigcap_{j=1}^{k}\{g_j(\theta) \in \hat{I}_j\}\right) \geq 1 - \alpha.$$

Notons que cette méthode n'est intéressante que pour de petites valeurs de k. Pour de grandes valeurs de k, la région de confiance $\hat{I}_{1,1-\frac{\alpha}{k}} \times \hat{I}_{2,1-\frac{\alpha}{k}} \times \ldots \times \hat{I}_{k,1-\frac{\alpha}{k}}$ a une probabilité beaucoup plus grande que $1 - \alpha$: elle encadre $g(\theta)$ beaucoup trop largement.