Statistique (MATH-F-315, Cours #2)

Thomas Verdebout

Université Libre de Bruxelles

2015

Plan de la partie Statistique du cours

1. Introduction.
2. Théorie de l'estimation.
3. Tests d'hypothèses et intervalles de confiance.
4. Régression.
5. ANOVA.

Estimateurs à dispersion minimale

Pour simplifier ici, nous considérons le cas univarié uniquement. On peut considérer qu'un estimateur $\hat{\theta}_1$ de θ est "meilleur" qu'un estimateur $\hat{\theta}_2$ si

$$\mathrm{E}_\theta\left[(\hat{\theta}_1 - \theta)^2\right] \leq \mathrm{E}_\theta\left[(\hat{\theta}_2 - \theta)^2\right] \qquad \forall \theta \in \Theta.$$

Definition

La quantité $\mathrm{E}_\theta[(\hat{\theta} - \theta)^2]$ est appelée *écart quadratique moyen* (entre $\hat{\theta}$ et θ).
Il n'existe que pour les estimateurs $\hat{\theta}$ de carré intégrable ($\mathrm{E}_\theta[\hat{\theta}^2] < \infty$).

Sur base de l'écart quadratique moyen, nous considérons donc dans la suite qu'un estimateur $\hat{\theta}_1$ de θ est donc plus performant qu'un estimateur $\hat{\theta}_2$ si son écart quadratique moyen (pris par rapport à $\boldsymbol{\theta}$) est uniformément plus petit (*uniformément* ici signifie *pour toute valeur de* θ).

Estimateurs à dispersion minimale

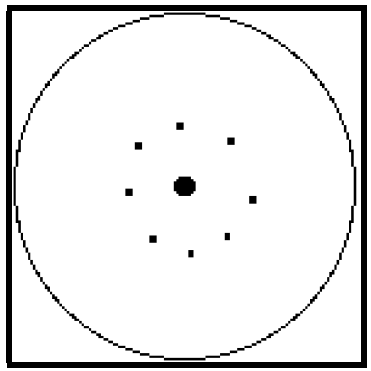
Nous avons que

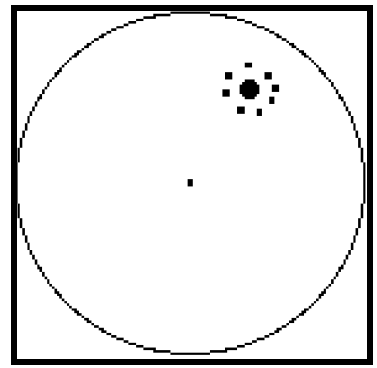
$$\begin{aligned}
\mathrm{E}_\theta[(\hat{\theta}-\theta)^2] &= \mathrm{E}_\theta[(\hat{\theta}-\mathrm{E}_\theta[\hat{\theta}]+\mathrm{E}_\theta[\hat{\theta}]-\theta)^2] \\
&= \mathrm{E}_\theta[(\hat{\theta}-\mathrm{E}_\theta[\hat{\theta}])^2]+\mathrm{E}_\theta[(\mathrm{E}_\theta[\hat{\theta}]-\theta)^2]+2\mathrm{E}_\theta[(\hat{\theta}-\mathrm{E}_\theta[\hat{\theta}])(\mathrm{E}_\theta[\hat{\theta}]-\theta)]. \\
&= \mathrm{E}_\theta[(\hat{\theta}-\mathrm{E}_\theta[\hat{\theta}])^2]+(\mathrm{E}_\theta[\hat{\theta}]-\theta)^2+2\mathrm{E}_\theta[\hat{\theta}-\mathrm{E}_\theta[\hat{\theta}]](\mathrm{E}_\theta[\hat{\theta}]-\theta).
\end{aligned}$$

Or, $\mathrm{E}_\theta[\hat{\theta}-\mathrm{E}_\theta[\hat{\theta}]]=0$. Donc

$$\begin{aligned}
\mathrm{E}_\theta[(\hat{\theta}-\theta)^2] &= \mathrm{Var}_\theta(\hat{\theta})+(\mathrm{E}_\theta[\hat{\theta}]-\theta)^2 \\
&= \mathrm{Var}_\theta(\hat{\theta})+(\mathrm{Biais}_\theta(\hat{\theta}))^2.
\end{aligned}$$

L'écart quadratique moyen est donc la variance augmentée du carré du biais (trade-off entre biais et variance):





Estimateurs à dispersion minimale

Recherchons maintenant un estimateur θ^* dont l'écart quadratique moyen soit uniformément minimum dans l'ensemble de *tous* les estimateurs de θ.

Soit θ_0 un point quelconque de Θ. Considérons l'estimateur $T(\mathbf{X}) = \theta_0$ p.s. pour tout θ (estimateur dégénéré en θ_0). On met complètement de côté l'échantillon et on prend un point fixé θ_0 dans l'espace des paramètres pour estimer θ.

Ecart quadratique moyen de cet estimateur: $(\theta_0 - \theta)^2$en θ, et donc s'annule en θ_0.

Un estimateur θ^* minimisant l'écart quadratique moyen dans l'ensemble de *tous* les estimateurs de θ devrait donc présenter un écart quadratique moyen nul en tout θ_0, ce qui est impossible.

Une possibilité raisonnable est de se restreindre aux estimateurs $\hat{\boldsymbol{\theta}}$ sans biais ($\mathrm{E}_{\boldsymbol{\theta}}[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}$ pour tout $\boldsymbol{\theta} \in \Theta$), l'écart quadratique moyen coïncide avec la variance.

La condition de non-biais élimine donc les estimateurs dégénérés.

Estimateurs à dispersion minimale

Peut-on espérer l'existence d'estimateurs à variance uniformément minimale *dans la classe des estimateurs sans biais* ?

Pour commencer, définissons le concept de vraisemblance.

- Cas discret:
  On appelle *vraisemblance* (*likelihood*) la probabilité jointe $L_{\boldsymbol{\theta}}(\mathbf{X})$ du vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_n)$ évaluée en $\mathbf{X} = (X_1, \ldots, X_n)$.

$$L_{\boldsymbol{\theta}}(\mathbf{X}) = L_{\boldsymbol{\theta}}(X_1, \ldots, X_n) = \prod_{i=1}^{n} p_{\boldsymbol{\theta}}(X_i).$$

- Cas continu:
  On appelle *vraisemblance* la densité jointe $L_{\boldsymbol{\theta}}(\mathbf{X})$ du vecteur aléatoire $\mathbf{X} = (X_1, \ldots, X_n)$ évaluée en $\mathbf{X} = (X_1, \ldots, X_n)$. Si la loi-population est de densité $f_{\boldsymbol{\theta}}$, on obtient

$$L_{\boldsymbol{\theta}}(\mathbf{X}) = L_{\boldsymbol{\theta}}(X_1, \ldots, X_n) = \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(X_i).$$

Estimateurs à dispersion minimale

Soit $L_\theta(\mathbf{X})$ une vraisemblance satisfaisant à certaines conditions de régularité: $L_\theta(\mathbf{x}) > 0\ \forall, \theta$, $\theta \mapsto L_\theta(\mathbf{x})$ dérivable sous le signe de l'expression $\int L_\theta(\mathbf{x}) d\mathbf{x} = 1$, la dérivée $\partial_\theta \log L_\theta(\mathbf{x})$ est de variance finie :

$$0 < \mathcal{I}(\theta) := \int (\partial_\theta \log L_\theta(\mathbf{x}))^2 L_\theta(\mathbf{x}) d\mathbf{x} = \mathrm{Var}_\theta\left(\partial_\theta \log L_\theta(\mathbf{X})\right) < \infty;$$

la quantité $\mathcal{I}(\theta)$ est appelée *Information de Fisher* (relative à θ). On a le résultat suivant

Théorème (Inégalité de Cramér-Rao)

Sous les conditions énoncées ci-dessus et si T est une statistique telle que (i) $\mathrm{Var}_\theta(T) < \infty$ pour tout $\theta \in \Theta$ et (ii) l'expression $\theta = \int T(\mathbf{x}) L_\theta(\mathbf{x})\ d\mathbf{x}$, on a que

$$\mathrm{Var}_\theta\left(T(\mathbf{X})\right) \geq \frac{1}{\mathcal{I}(\theta)} \qquad \text{pour tout } \theta \in \Theta.$$

Preuve.
D'abord, notons que puisque $\theta = \int T(\mathbf{x}) L_\theta(\mathbf{x})\, d\mathbf{x}$ peut être dérivée sous le signe,

$$1 = \frac{d}{d\theta} \int T(\mathbf{x}) L_\theta(\mathbf{x})\, d\mathbf{x} = \int T(\mathbf{x}) \partial_\theta \log L_\theta(\mathbf{x})\, L_\theta(\mathbf{x})\, d\mathbf{x} = \mathrm{Cov}_\theta(T(\mathbf{X}), \partial_\theta \log L_\theta(\mathbf{x})).$$

Maintenant, calculons la variance (non négative) de la variable aléatoire $S_\theta(\mathbf{X}) := T(\mathbf{X}) - (\mathcal{I}(\theta))^{-1} \partial_\theta \log L_\theta(\mathbf{X})$:

$$\begin{aligned} 0 &\leq \mathrm{Var}_\theta(S_\theta(\mathbf{X})) \\ &= \mathrm{Var}_\theta(T(\mathbf{X})) + \underbrace{(\mathcal{I}(\theta))^{-2} \mathrm{Var}_\theta(\partial_\theta \log L_\theta(\mathbf{X}))}_{(\mathcal{I}(\theta))^{-1}} - \underbrace{2(\mathcal{I}(\theta))^{-1} \mathrm{Cov}_\theta(T(\mathbf{X}),\ \partial_\theta \log L_\theta(\mathbf{X}))}_{2(\mathcal{I}(\theta))^{-1}} \end{aligned}$$

Donc

$$0 \leq \mathrm{Var}_\theta(T(\mathbf{X})) - (\mathcal{I}(\theta))^{-1},$$

ce qui établit le résultat.

Efficacité

Definition

Un estimateur $\hat{\theta}$ de θ est dit *efficace* (pour θ) si son biais est nul et que sa variance atteint uniformément la borne de Cramér-Rao : $\mathrm{Var}_\theta(\hat{\theta}) = 1/\mathcal{I}(\theta)$ pour tout $\theta \in \Theta$.

La même définition peut aussi s'exprimer de façon équivalente à partir de l'écart quadratique moyen :

Definition

Un estimateur $\hat{\theta}$ de θ est dit *efficace* (pour θ) si son écart quadratique moyen (par rapport à θ) atteint la borne de Cramér-Rao uniformément en θ :
$\mathrm{E}_\theta[(T(\mathbf{X}) - \theta)^2] = 1/\mathcal{I}(\theta)$ pour tout $\theta \in \Theta$.

Remarques

1. L'équivalence des deux définitions provient de ce que, pour un estimateur sans biais, la variance et l'écart quadratique moyen coïncident, et que, la variance étant comprise entre l'écart quadratique moyen et la borne, l'égalité de ces deux derniers implique celle de la variance et de l'écart quadratique moyen, donc l'absence de biais. Un estimateur biaisé ne saurait donc être efficace.
2. Un estimateur efficace de θ est à variance uniformément minimale dans la classe des estimateurs sans biais de θ; la réciproque n'est pas vraie, car il arrive que la borne ne puisse être atteinte.

Estimateurs à dispersion minimale

Exemple 1 (échantillon de Bernoulli)

Soient X_i i.i.d. $\mathrm{Bin}(1,p)$. On a que: $L_p(\mathbf{X}) = p^{\sum_{i=1}^n X_i}(1-p)^{n-\sum_{i=1}^n X_i}$. Donc,

$$\log L_p(\mathbf{X}) = \sum_{i=1}^n X_i(\log p) + (n - \sum_{i=1}^n X_i)\log(1-p),$$

et

$$\partial_p \log L_p(\mathbf{X}) = \sum_{i=1}^n X_i \frac{1}{p} - (n - \sum_{i=1}^n X_i)\frac{1}{1-p} = \sum_{i=1}^n X_i \left(\frac{1}{p} + \frac{1}{1-p}\right) - \frac{n}{1-p}$$

$$= \sum_{i=1}^n X_i \left(\frac{1}{p(1-p)}\right) - \frac{n}{1-p} \ .$$

Calculons l'information de Fisher ($\sum_{i=1}^n X_i \sim \mathrm{Bin}(n,p)$):

$$\mathcal{I}(p) := \mathrm{Var}_p(\partial_p \log L_p(\mathbf{X})) = \frac{1}{p^2(1-p)^2}\mathrm{Var}\left(\sum_{i=1}^n X_i\right) = \frac{np(1-p)}{p^2(1-p)^2} = \frac{n}{p(1-p)}$$

On sait que $\hat{p} = \bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ est un estimateur sans biais de θ. On a

$$\mathrm{Var}_p(\hat{p}) = \frac{p(1-p)}{n} = \big(\mathcal{I}(p)\big)^{-1}.$$

Donc $\hat{p}$ est un estimateur efficace de p.

Estimateurs à dispersion minimale

Exemple 2 (moyenne d'un échantillon gaussien).
Soient X_i i.i.d. $\mathcal{N}(\mu, \sigma^2)$, σ^2 spécifié. On a que:

$$\log L_{\mu,\sigma^2}(\mathbf{X}) = -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2}\sum_{i=1}^{n}\left(\frac{X_i - \mu}{\sigma}\right)^2.$$

Donc

$$\partial_\mu \log L_{\mu,\sigma^2}(\mathbf{X}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)$$

Calculons l'information de Fisher correspondante :

$$\begin{aligned}\mathcal{I}(\mu) := \mathrm{E}_{\mu,\sigma^2}[(\partial_\mu \log L_{\mu,\sigma^2}(\mathbf{X}))^2] &= \mathrm{Var}_{\mu,\sigma^2}(\partial_\mu \log L_{\mu,\sigma^2}(\mathbf{X})) \\ &= \frac{1}{\sigma^4}\sum_{i=1}^{n}\mathrm{E}[(X_i - \mu)^2] = \frac{1}{\sigma^4}n\sigma^2 = \frac{n}{\sigma^2}.\end{aligned}$$

Or $\bar{X}$ est un estimateur sans biais de μ, et

$$\mathrm{Var}_{\mu,\sigma^2}(\bar{X}) = \frac{\sigma^2}{n} = \left[\mathcal{I}(\mu)\right]^{-1}.$$

Donc $\bar{X}$ est un estimateur efficace de μ.

Exemple 3 (variance d'un échantillon gaussien).
De la même façon, mais en supposant cette fois μ spécifié,

$$\partial_{\sigma^2} \log L_{\mu,\sigma^2}(\mathbf{X}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} (X_i - \mu),$$

et

$$\begin{aligned} \mathrm{E}_{\mu,\sigma^2}[(\partial_{\sigma^2} \log L_{\mu,\sigma^2}(\mathbf{X}))^2] &= \mathrm{Var}_{\mu,\sigma^2} \partial_{\sigma^2} \log L_{\mu,\sigma^2}(\mathbf{X}) \\ &= \frac{1}{4\sigma^4} n \mathrm{Var}_{\mu,\sigma^2}(((X_1 - \mu)/\sigma)^2) = \frac{n}{2\sigma^4} \end{aligned}$$

puisque $((X_1 - \mu)/\sigma)^2 \sim \chi^2_1$ et que la variance d'une chi-carré à un degré de liberté est 2.
On calcule aisément que $\mathrm{Var}_{\mu,\sigma^2}(s^2) = 2(n-1)\sigma^4/n^2$ et $\mathrm{Var}_{\mu,\sigma^2}(S^2) = 2\sigma^4/(n-1)$: ni s^2 ni S^2 ne sont donc efficaces (s^2 étant biaisé ne peut l'être).
En revanche, la variance de $\hat{\sigma^2} := \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ a pour variance $2\sigma^4/n$, et $\hat{\sigma^2}$ est donc efficace.

La méthode du maximum de vraisemblance

La méthode du maximum de vraisemblance est sans doute la méthode d'estimation la plus utilisée. Elle possède de nombreuses propriétés intéressantes, notamment des propriétés de convergence, de normalité et d'efficacité asymptotiques.

Soit $\mathbf{X}$ une observation dont le comportement est caractérisé par une vraisemblance $L_{\boldsymbol{\theta}}(\mathbf{X})$, $\boldsymbol{\theta} \in \Theta$.

Definition

On appelle *estimateur maximum de vraisemblance* (en anglais, *maximum likelihood estimator* ou *MLE*) de θ toute valeur $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$ maximisant la vraisemblance $L_{\boldsymbol{\theta}}(\mathbf{X})$:

$$\hat{\boldsymbol{\theta}} = \text{Argmax}_{\boldsymbol{\theta}} L_{\boldsymbol{\theta}}(\mathbf{X})$$

ou, de façon équivalente,

$$\hat{\boldsymbol{\theta}} = \text{Argmax}_{\boldsymbol{\theta}} \log L_{\boldsymbol{\theta}}(\mathbf{X}).$$

La méthode du maximum de vraisemblance

Prenons le cas continu (le cas discret se résoud de façon similaire). Soit $\mathbf{X} = (X_1, \ldots, X_n)$, où les X_i sont i.i.d. de densité $f_{\boldsymbol{\theta}}(x)$. La vraisemblance associée est donnée par $L_{\boldsymbol{\theta}}(\mathbf{X}) = \prod_{i=1}^{n} f_{\boldsymbol{\theta}}(X_i)$. On obtient dès lors que

$$\begin{aligned}\hat{\boldsymbol{\theta}} &= \text{Argmax}_{\boldsymbol{\theta}} \log L_{\boldsymbol{\theta}}(\mathbf{X}) \\ &= \text{Argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \log f_{\boldsymbol{\theta}}(X_i)\end{aligned}$$

Si $\boldsymbol{\theta} \longmapsto f_{\boldsymbol{\theta}}(\mathbf{x})$ est différentiable et Θ ouvert, on peut rechercher $\hat{\boldsymbol{\theta}}$ parmis les solutions du système

$$\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log f_{\boldsymbol{\theta}}(X_i) = \mathbf{0},$$

un système d'équations appelées *équations de vraisemblance*.

La méthode du maximum de vraisemblance

Exemple 1 (échantillon de Bernoulli). Soient $X_1, \ldots, X_n$ i.i.d. $\mathrm{Bin}(1, p)$, $p \in (0,\ 1)$. On a

$$L_p(X_1, \ldots, X_n) = p^{\sum_{i=1}^n X_i} (1-p)^{n - \sum_{i=1}^n X_i},$$

donc

$$\log L_p(X_1, \ldots, X_n) = \sum_{i=1}^n X_i \log(p) + \left(n - \sum_{i=1}^n X_i \right) \log(1-p)$$

et

$$\begin{aligned} \frac{\partial}{\partial p} \log L_p(X_1, \ldots, X_n) &= \left(\sum_{i=1}^n X_i \right) \frac{1}{p} - \left(n - \sum_{i=1}^n X_i \right) \frac{1}{1-p} \\ &= \left(\sum_{i=1}^n X_i \right) \left(\frac{1}{p} + \frac{1}{1-p} \right) - \frac{n}{1-p}. \end{aligned}$$

La méthode du maximum de vraisemblance

Annuler cette dérivée conduit à l'équation

$$\left(\sum_{i=1}^{n} X_i\right)\left(\frac{(1-p)+p}{p(1-p)}\right) - \frac{np}{p(1-p)} = 0$$

qui s'écrit encore

$$\left(\sum_{i=1}^{n} X_i\right) - np = 0$$

ou

$$\sum_{i=1}^{n} X_i = np.$$

La solution des équations de vraisemblance est donc

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

La méthode du maximum de vraisemblance

Exemple 2 (échantillon gaussien). Soient $X_1, \ldots, X_n$ i.i.d. $\mathcal{N}(\mu, \sigma^2)$. La vraisemblance s'écrit

$$L_{\mu,\sigma^2}(X_1, \ldots, X_n) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2\right];$$

donc

$$\log L_{\mu,\sigma^2}(X_1, \ldots, X_n) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2.$$

Le système des équations de vraisemblance comprend deux équations. La première équation est relative à μ :

$$\frac{\partial}{\partial\mu}\log L_{\mu,\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu) = 0.$$

Cette équation (inconnue: μ) est satisfaite si et seulement si $\sum_{i=1}^{n}(X_i - \mu) = 0$. L'unique solution en est donc

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i =: \bar{X}.$$

La méthode du maximum de vraisemblance

La seconde équation provient de l'annulation de la dérivée par rapport à σ^2. En y remplaçant μ par $\hat{\mu}$ (méthode de substitution), on obtient

$$\frac{\partial}{\partial \sigma^2} \log L_{\hat{\mu},\sigma^2} = -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(X_i - \hat{\mu})^2 = 0,$$

qui est équivalente (car $\sigma^2 > 0$) à

$$-\frac{n}{2}\sigma^2 + \frac{1}{2}\sum_{i=1}^{n}(X_i - \bar{X})^2 = 0.$$

L'unique solution est

$$\hat{\sigma^2} = \frac{1}{n}\sum(X_i - \bar{X})^2 = s^2.$$

La méthode du maximum de vraisemblance

Si on note comme avant $\mathcal{I}_{\boldsymbol{\theta}}^{(n)}$ la matrice d' information de Fisher pour un échantillon de taille n, on obtient

$$\mathcal{I}_{\boldsymbol{\theta}}^{(n)} := \mathrm{Var}_{\boldsymbol{\theta}}\big(\frac{\partial}{\partial \boldsymbol{\theta}} \log L_{\boldsymbol{\theta}}^{(n)}(X_1, \ldots, X_n)\big) = \mathrm{Var}_{\boldsymbol{\theta}}\left(\sum_{i=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \log L_{\boldsymbol{\theta}}^{(1)}(X_i)\right)$$

$$= \sum_{i=1}^{n} \mathrm{Var}_{\boldsymbol{\theta}}\big(\frac{\partial}{\partial \boldsymbol{\theta}} \log L_{\boldsymbol{\theta}}^{(1)}(X_i)\big) = \sum_{i=1}^{n} \mathcal{I}_{\boldsymbol{\theta}}^{(1)} = n\mathcal{I}_{\boldsymbol{\theta}}^{(1)}.$$

La solution des équations de vraisemblance et l'information de Fisher sont donc liées. Sous des conditions très générales, on montre que si $\hat{\boldsymbol{\theta}}$ est solution des équations de vraisemblance,

$$n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(\mathbf{0}, (\mathcal{I}_{\boldsymbol{\theta}}^{(n)})^{-1})$$

C'est à dire $\hat{\boldsymbol{\theta}}^{(n)} \approx \mathcal{N}(\boldsymbol{\theta}, (n\mathcal{I}_{\boldsymbol{\theta}}^{(1)})^{-1})$. Notons que $(n\mathcal{I}_{\boldsymbol{\theta}}^{(1)})^{-1}$ est la borne de Cramér-Rao (pour $\boldsymbol{\theta}$ et n observations). Au sens de l'approximation ci-dessus, l'estimateur $\hat{\boldsymbol{\theta}}$ est donc normal, sans biais, et efficace. De tels estimateurs sont dits *B.A.N. (Best Asymptotically Normal)*.

Autres méthodes d'estimation

Soient $X_1, \ldots, X_n$ i.i.d. $\mathrm{P}_{\boldsymbol{\theta}}$, où $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)'$. Notons

- $\mu_k(\boldsymbol{\theta}) := \mathrm{E}[X_1^k]$, $k = 1, 2, \ldots$ les moments-population
- $m_k := \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} X_i^k$, $k = 1, 2, \ldots$ les moments empiriques correspondants.

Supposons que les moments-population existent et soient finis jusqu'à l'ordre K au moins. Ces moments sont des fonctions du paramètre $\boldsymbol{\theta}$: faisons l'hypothèse que l'application $\boldsymbol{\theta} \mapsto (\mu_1(\boldsymbol{\theta}), \mu_2(\boldsymbol{\theta}), \ldots, \mu_K(\boldsymbol{\theta}))'$ soit bijective.
La méthode des moments consiste à prendre comme estimateur de $\boldsymbol{\theta}$ la solution $\hat{\boldsymbol{\theta}}$ du système

$$\begin{cases} \mu_1(\boldsymbol{\theta}) = m_1 \\ \vdots \\ \mu_K(\boldsymbol{\theta}) = m_K \end{cases}$$

Autres méthodes d'estimation

En général, les estimateurs efficaces sont très sensibles aux observations "aberrantes". Les statistiques habituelles qui sont efficaces sous un modèle sont généralement très affectées par une petite modification du modèle sous-jacent.

Prenons l'exemple de la moyenne empirique $n^{-1} \sum_{i=1}^{n} X_i$ d'un échantillon aléatoire simple $X_1, \ldots, X_n$. Si une des observations devient arbitrairement grande, la moyenne empirique devient de la même manière dans le sens où la moyenne de l'échantillon $X_1, \ldots, X_n, \infty$ vaut ∞.

La médiane empirique souffre moins d'une situation comme celle de ci-dessus. Dans ce sens, elle peut être qualifiée de plus robuste que la médiane. Dans les années 1960, les statistiques robustes ont connu un grand essor.

Autres méthodes d'estimation

En particulier, la classe des M-estimateurs d'un paramètre $\boldsymbol{\theta}$ a été abondamment étudiée. Un M-estimateur de $\boldsymbol{\theta}$ associé à une fonction objectif $\varphi_{\boldsymbol{\theta}}$ est défini par

$$\hat{\boldsymbol{\theta}} = \text{Argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{n} \varphi_{\boldsymbol{\theta}}(X_i)$$

Un choix judicieux de la fonction $\varphi_{\boldsymbol{\theta}}$ peut permettre d'obtenir un estimateur robuste.

Notons que les estimateurs maximum de vraisemblance sont un cas particulier de M-estimateur obtenu en prenant $\varphi_{\boldsymbol{\theta}} = \log f_{\boldsymbol{\theta}}$.