Statistique : Estimation : modèle statistique, notion d'estimateur, biais / variance.

Joseph Salmon

Septembre 2014

Modèle statistique : contexte

Rappel

- On observe des réalisations $(y_1, \ldots, y_n)$ de variables aléatoires inconnues (éventuellement vectorielles)
- On suppose ici que les variables sont indépendantes et identiquement distribuées (*i.i.d.*) selon une loi $\mathbb{P}_Y$

- Selon la situation, la loi $\mathbb{P}_Y$ a certaines caractéristiques.
  - **Exemple**: "Pile ou face" : on sait que $\mathbb{P}_Y = \text{Bernoulli}(p)$ pour un certain $p \in [0, 1]$ inconnu

- Reformulation : on a une famille de lois candidates pour $\mathbb{P}_Y$,
  - **Exemple**: la famille des lois de Bernoulli

Modèle statistique

- La loi cible $\mathbb{P}_Y$ est indéxée par un paramètre $\theta \in \Theta$: $\mathbb{P}_Y = \mathbb{P}_\theta$ pour un θ inconnu, et Θ est l'ensemble d'indexation
  - **Exemple**: "Pile ou face" $\quad \theta = p, \quad \Theta = [0, 1]$

Modèle statistique

Un modèle statistique est une famille de lois

$$\mathcal{M} = \{\mathbb{P}_\theta : \theta \in \Theta\}$$

indexées par un ensemble de paramètres Θ.

Modèle statistique paramétrique

Modèle paramétrique

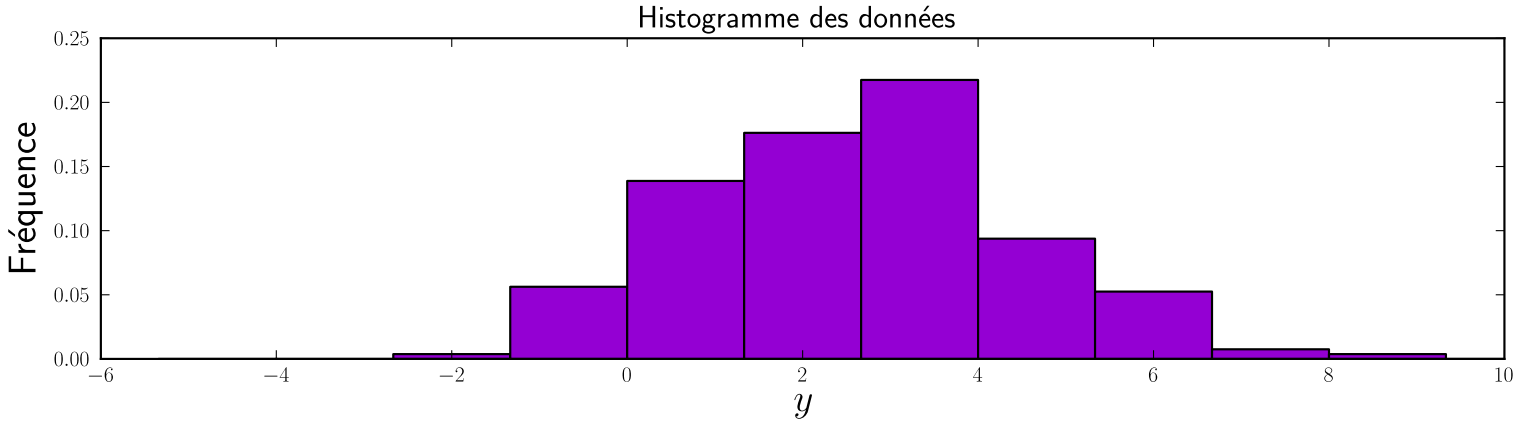
Un modèle paramétrique est une famille de lois $\mathcal{M} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ indexée par un ensemble fini, disons p, de paramètres : $\Theta \subset \mathbb{R}^p$

Rem: le modèle est indéxé par un nombres ou un vecteur réel. p est la dimension du modèle

Exemple:

- "Pile ou face" (Bernoulli) $\theta = p$; $\Theta = [0, 1]$.
- Modèle gaussien : $\theta = (\mu, \sigma^2)$, $\Theta = \mathbb{R} \times \mathbb{R}^{+*}$.

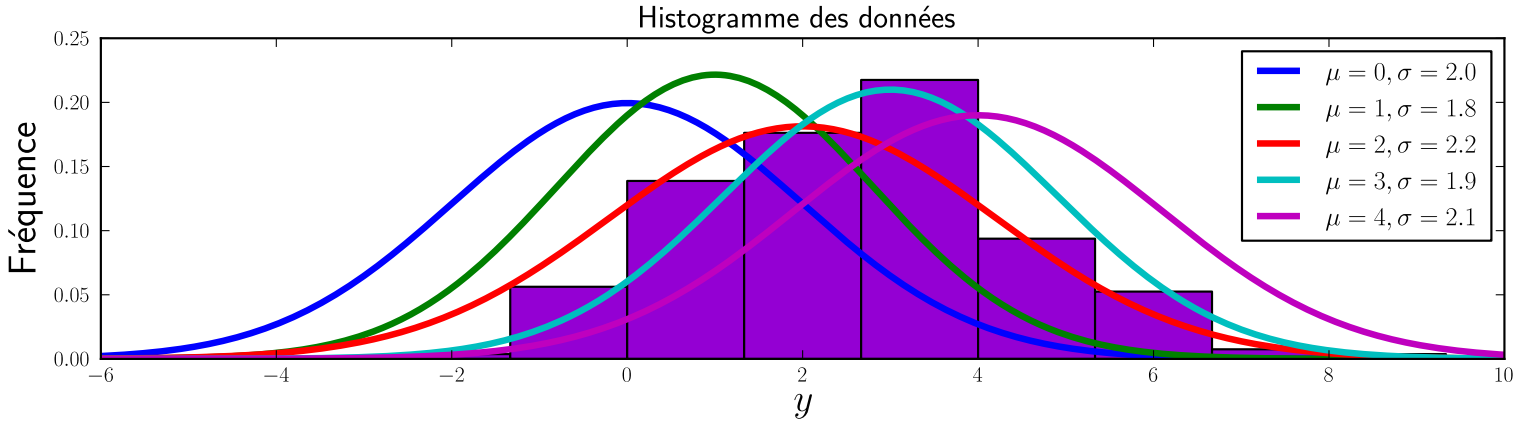
Exemple



- $(y_1, \ldots, y_n)$ un échantillon.
- Si on remarque que l'échantillon est 'symétrique' par rapport à sa moyenne empirique, avec un histogramme 'en cloche' :

Quel modèle choisir ?

Exemple : cas gaussien



- Il peut être raisonnable de chercher la loi des données dans un modèle gaussien. Le paramètre du modèle est dans ce cas $\theta = (\mu, \sigma^2)$, $\theta \in \Theta = \mathbb{R} \times \mathbb{R}^{+*}$

Estimateur

- *Objectif :* Estimer une quantité $g = g(\theta)$ qui ne dépend que de la loi $\mathbb{P}_\theta$ des observations.

  g est une constante inconnue déterministe *i.e.,* non aléatoire.

  Exemple: l'espérance, un quantile, la variance, etc.

- *Intuition :* Un estimateur $\hat{g}$ est calculé à partir de l'échantillon $(y_1, \ldots, y_n)$, dans le but d'approcher $g(\theta)$.

Estimateur : définition

Un **estimateur** $\hat{g}$ est une fonction des observations :

$$\hat{g} : (y_1, \ldots, y_n) \mapsto \hat{g}(y_1, \ldots, y_n)$$

Propriétés d'un estimateur : le biais

- le biais d'un estimateur $\hat{g}$ est l'espérance de l'erreur

$$\boxed{\text{Biais}(\hat{g}) = \mathbb{E}(\hat{g}(y_1, \ldots, y_n)) - g} \qquad \text{(dépend de } \theta\text{)}.$$

Estimateur sans biais

Un estimateur $\hat{g}$ de g est non biaisé si, quel que soit $\theta \in \Theta$,

$$\mathbb{E}(\hat{g}(y_1, \ldots, y_n)) = g(\theta)$$

Rem: Le biais est une mesure de l'erreur systématique d'une méthode. La vrai quantité d'intérêt est plutôt la valeur absolue du biais.

Estimateur sans biais de l'espérance

- L'espérance 'théorique' dépend de la loi $\mathbb{P}_\theta$.
- on cherche à estimer $g(\theta) = \mathbb{E}(Y)$

Théorème

La moyenne empirique $\hat{g}(y_1, \ldots, y_n) = \overline{y}_n = \frac{1}{n}\sum_{i=1}^n y_i$ est un estimateur sans biais de l'espérance $\mathbb{E}(Y)$

En effet,

$$\mathbb{E}(\hat{g}(y_1, \ldots, y_n)) = \mathbb{E}(\frac{1}{n}\sum_{i=1}^n y_i) = \frac{1}{n}\sum_{i=1}^n \mathbb{E}(y_i) = \mathbb{E}(Y)$$

car $\mathbb{E}(y_i) = \mathbb{E}(Y)$ (caractère *i.i.d.* des y_i)

Rem: L'estimateur $\hat{g}(y_1, \ldots, y_n) = y_1$ est aussi un estimateur sans biais de l'espérance

Estimateur sans biais de la variance

- La variance 'théorique' dépend de la loi $\mathbb{P}_\theta$.
- on cherche à estimer $g(\theta) = \mathrm{Var}(Y)$

Théorème

L'estimateur $\hat{g}(y_1, \ldots, y_n) = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \overline{y}_n)^2$ est un estimateur sans biais de la variance $\mathrm{Var}(Y)$

Rem: Attention au terme $n-1$. La variance empirique (avec un facteur 1/n) est en effet biaisée

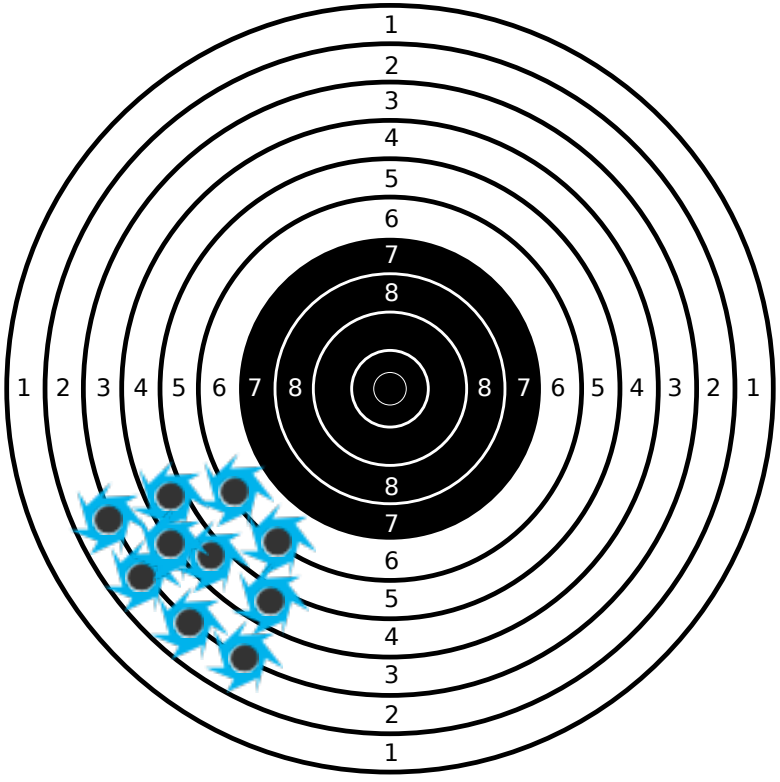
Propriétés d'un estimateur : la variance

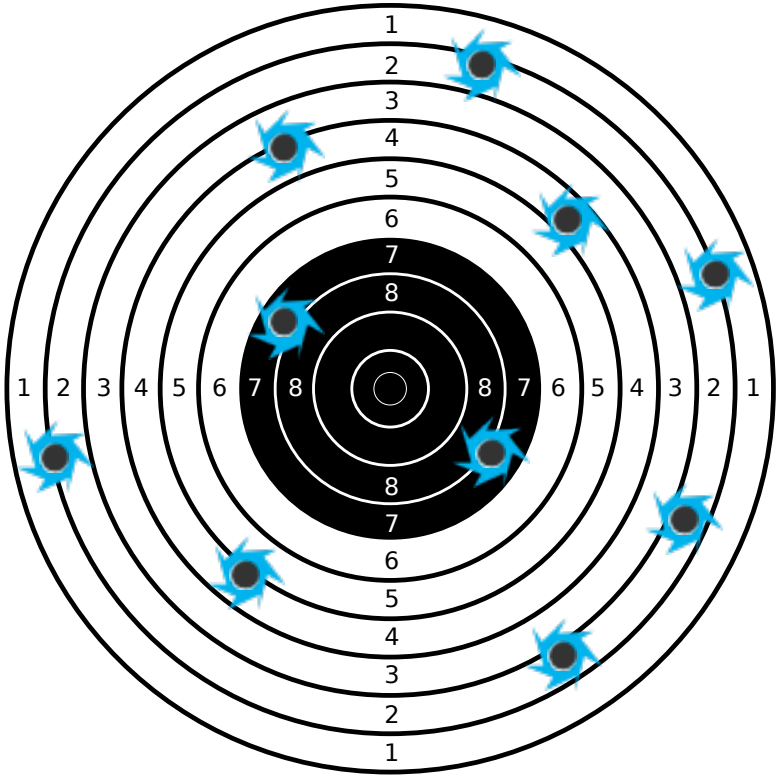
- la Variance d'un estimateur est sa variance théorique :

$$\boxed{\mathrm{Var}(\hat{g}) = \mathrm{Var}(\hat{g}(y_1, \ldots, y_n)) = \mathbb{E}(\hat{g} - \mathbb{E}(\hat{g}))^2} \qquad \text{(dépend de } \theta\text{)}.$$

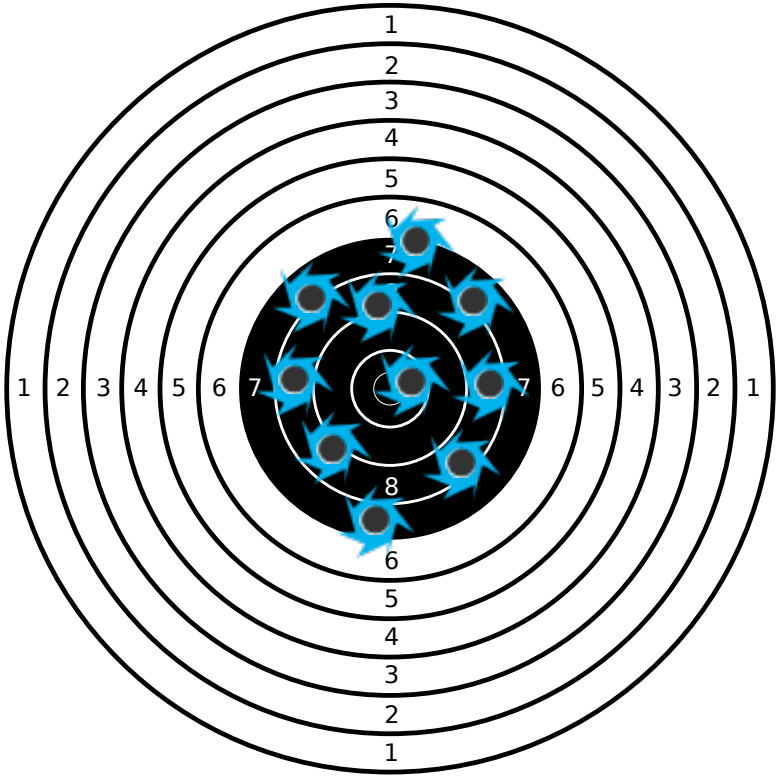
Rem: La variance mesure donc la dispersion d'un estimateur autour de sa moyenne

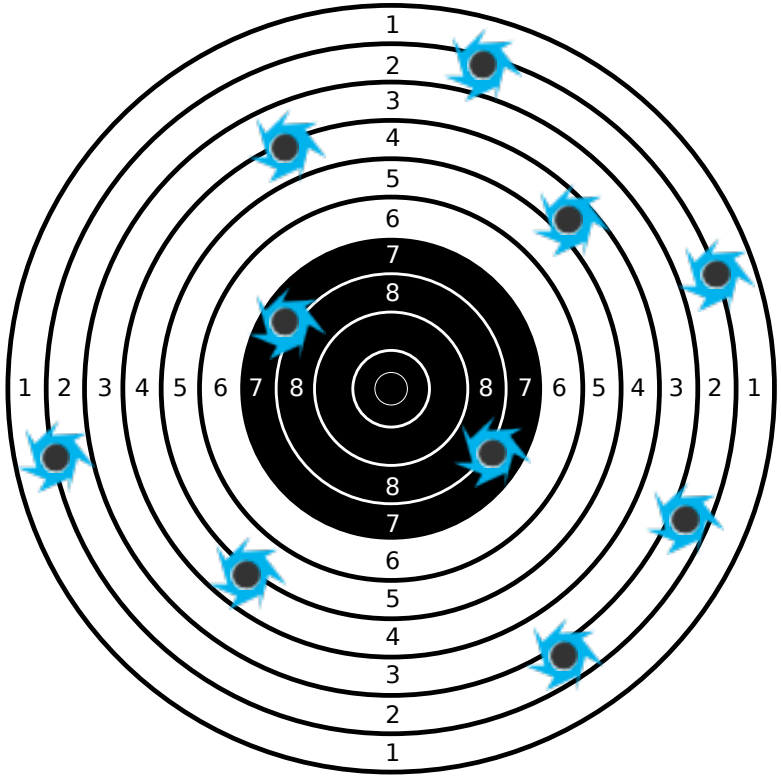
Biais ou variance ?





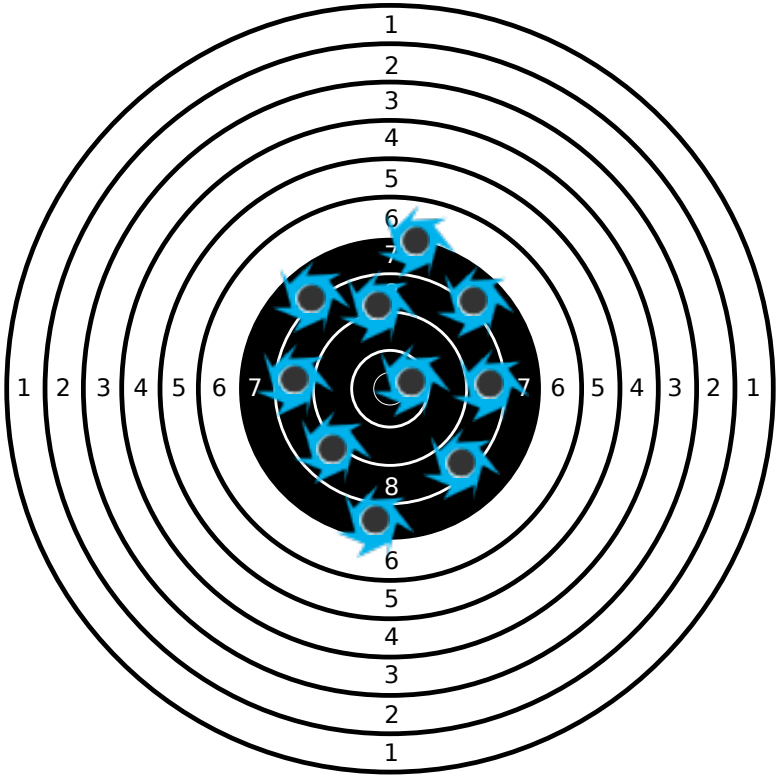
Biais ou variance ?

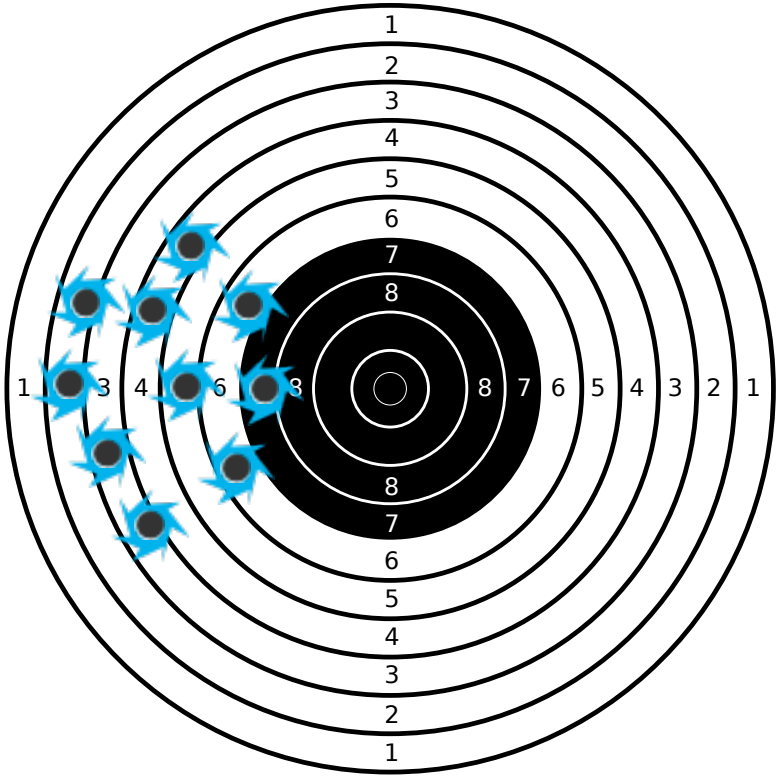




- Si $\hat{g}_0$ et $\hat{g}_1$ sont sans biais, on préfére celui de plus faible variance.

Biais ou variance ?





- Si $\hat{g}_0$ et $\hat{g}_1$ ont la même variance, alors on préfére celui de biais le plus faible.

Risque quadratique / compromis biais-variance

Risque quadratique d'un estimateur de g

Le risque quadratique d'un estimateur $\hat{g}$ est l'espérance de son erreur au carré :

$$R(\hat{g}) = \mathbb{E}\left[(\hat{g} - g)^2\right]$$

- On fait apparaître le biais $B = \mathbb{E}[\hat{g}] - g$ et on développe.

$$\begin{aligned} R(\hat{g}) &= \mathbb{E}\left[(\hat{g} - \mathbb{E}(\hat{g}) + B)^2\right] \\ &= \mathbb{E}\left[(\hat{g} - \mathbb{E}(\hat{g}))^2 + B^2 + 2B(\hat{g} - \mathbb{E}(\hat{g}))\right] \\ &= \mathrm{Var}(\hat{g}) + B^2 + 2B \underbrace{\mathbb{E}\left[\hat{g} - \mathbb{E}(\hat{g})\right]}_{=0} \end{aligned}$$

$$\boxed{\text{Risque } (\hat{g}) = \text{Variance}(\hat{g}) + \left(\text{Biais}(\hat{g})\right)^2}$$

Règle de choix : prendre l'estimateur dont le risque est le plus petit

Références I