

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

Joseph Salmon

Septembre 2014

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

## Estimation

Modèle statistique

Estimateur

Joseph Salmon

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

## Estimation

Modèle statistique

Estimateur

## Estimation

Modèle statistique

Estimateur

Joseph Salmon

## Rappel

- ▶ On observe des réalisations  $(y_1, \dots, y_n)$  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

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

## Estimation

Modèle statistique

Estimateur

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

## Estimation

Modèle statistique

Estimateur

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

## Modèle statistique

Un modèle statistique est une famille de lois de probabilité

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

indexées par un ensemble de paramètres  $\Theta$ .

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

## Estimation

Modèle statistique

Estimateur

Joseph Salmon

## Modèle paramétrique

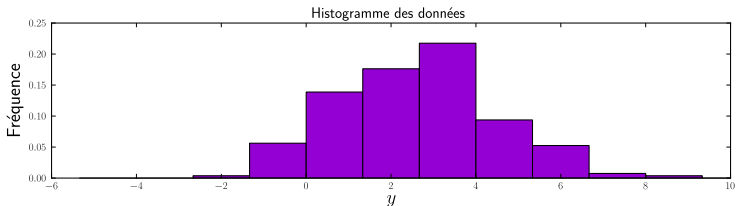
Un modèle paramétrique est une **famille de lois** de probabilité  $\mathcal{M} = \{\mathbb{P}_\theta : \theta \in \Theta\}$  indexée par un ensemble fini, disons  $p$ , de paramètres :  $\Theta \subset \mathbb{R}^d$

Rem: le modèle est indexé par un **nombre ou un vecteur réel**.  $d$  est la dimension du modèle.

**Exemple:**

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

# Exemple



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

## Estimation

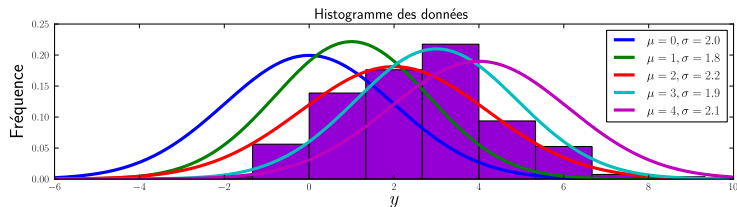
Modèle statistique

Estimateur

- ▶  $(y_1, \dots, y_n)$  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



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

## Estimation

Modèle statistique

Estimateur

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

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

## Estimation

Modèle statistique

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**  $(y_1, \dots, y_n)$ , dans le but d'approcher  $g(\theta)$ .

### Estimateur : définition

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

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



# Propriétés d'un estimateur : le biais

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

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

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

## Estimateur sans biais

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

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

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

## Estimation

Modèle statistique

Estimateur

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

En effet,

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

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

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

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

## Estimation

Modèle statistique

Estimateur

# Estimateur sans biais de la variance

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

## Estimation

Modèle statistique

Estimateur

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

### Théorème

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

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

# Propriétés d'un estimateur : la variance

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

## Estimation

Modèle statistique

Estimateur

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

$$\text{Var}(\hat{g}) = \text{Var}(\hat{g}(y_1, \dots, y_n)) = \mathbb{E}(\hat{g} - \mathbb{E}(\hat{g}))^2$$

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

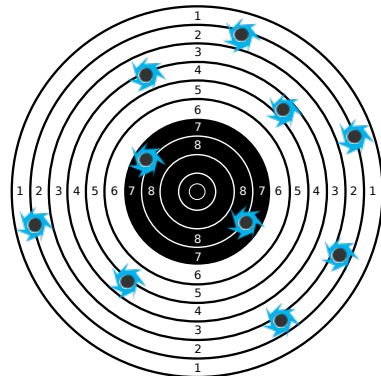
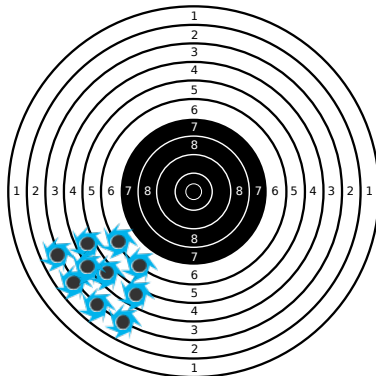
# Biais ou variance ?

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

## Estimation

Modèle statistique

Estimateur



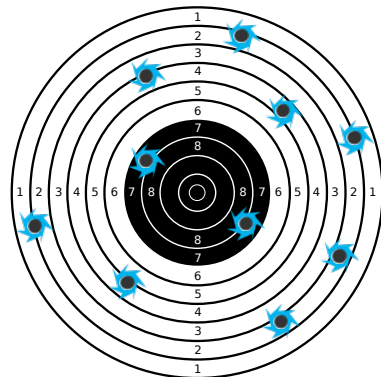
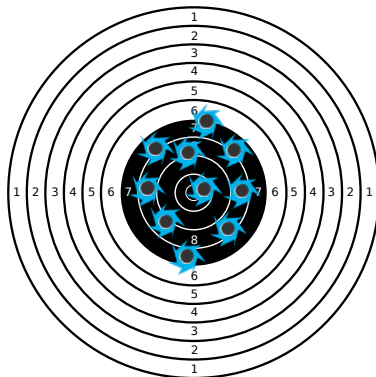
# Biais ou variance ?

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

## Estimation

Modèle statistique

Estimateur



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

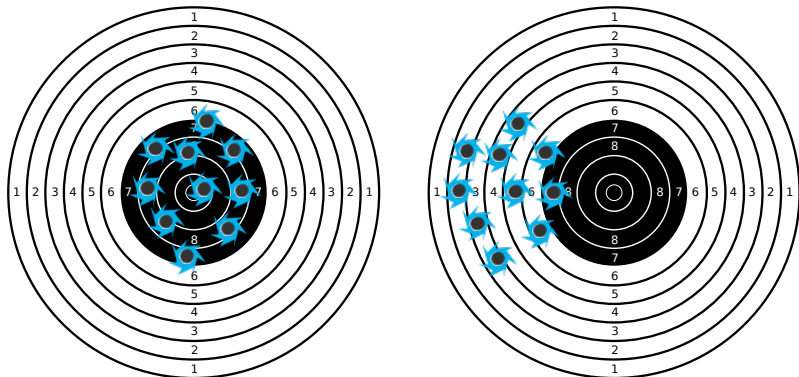
# Biais ou variance ?

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

## Estimation

Modèle statistique

Estimateur



- Si  $\hat{g}_0$  et  $\hat{g}_1$  ont la même variance, alors on préfère 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 :

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

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

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

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

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

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

## Estimation

Modèle statistique

Estimateur