

# HLMA408: Traitement des données

Échantillonnage aléatoire

**Joseph Salmon**

<http://josephsalmon.eu>

Université de Montpellier



# Sommaire

Introduction

Loi d'échantillonnage et estimation

Statistique, ou propriété des quantités calculées sur un échantillon

Approximations gaussiennes et intervalles de confiance

# Sommaire

Introduction

Loi d'échantillonnage et estimation

Statistique, ou propriété des quantités calculées sur un échantillon

Approximations gaussiennes et intervalles de confiance

# Retour sur l'étude babies23.data

À San Francisco<sup>(1)</sup>, 1236 naissances ont été répertoriées au cours d'une année à la *Kaiser Foundation Health Plan*<sup>(2)</sup>

- ▶ Quelle est la proportion des mères qui fume du tabac?
- ▶ Parmi les fumeuses, quelle est la consommation quotidienne?
- ▶ Cette consommation influe-t-elle sur le développement de l'enfant?

Faire une enquête exhaustive est compliqué et prend du temps<sup>(3)</sup>  
⇒ sondage sur un échantillon

---

<sup>(1)</sup>D. Nolan and T. P. Speed. *Stat labs: mathematical statistics through applications*. Springer Science & Business Media, 2001.

<sup>(2)</sup>dont seulement 1226 ont donné leur consommation de tabac

<sup>(3)</sup>l'étude a pris un an!

## Retour sur l'étude babies23.data

À San Francisco<sup>(1)</sup>, 1236 naissances ont été répertoriées au cours d'une année à la *Kaiser Foundation Health Plan*<sup>(2)</sup>

- ▶ Quelle est la proportion des mères qui fume du tabac?
- ▶ Parmi les fumeuses, quelle est la consommation quotidienne?
- ▶ Cette consommation influe-t-elle sur le développement de l'enfant?

Faire une enquête exhaustive est compliqué et prend du temps<sup>(3)</sup>  
⇒ sondage sur un échantillon

---

<sup>(1)</sup>D. Nolan and T. P. Speed. *Stat labs: mathematical statistics through applications*. Springer Science & Business Media, 2001.

<sup>(2)</sup>dont seulement 1226 ont donné leur consommation de tabac

<sup>(3)</sup>l'étude a pris un an!

# Questions abordées

- ▶ Comment **estimer** une moyenne sur un échantillon ?
- ▶ Comment **mesurer l'erreur** introduite par l'**échantillonnage** ?
- ▶ Comment choisir le nombre de naissances à recenser ?

# Sommaire

Introduction

**Loi d'échantillonnage et estimation**

Statistique, ou propriété des quantités calculées sur un échantillon

Approximations gaussiennes et intervalles de confiance

## Un peu de vocabulaire

- ▶ **Population (totale)** :  $N$ , peut être grand (voire infini)
- ▶ **Échantillon** : sous partie de la population
- ▶ **Paramètre** : grandeur définie sur la population que l'on cherche à estimer
- ▶ **Taille de l'échantillon** : nombre d'individus échantillonnés,  $n$
- ▶ **Statistique** : quantité aléatoire calculée sur l'échantillon (fluctue en fonction de l'échantillon)



## Un peu de vocabulaire

- ▶ **Population (totale)** :  $N$ , peut être grand (voire infini)
- ▶ **Échantillon** : sous partie de la population
- ▶ **Paramètre** : grandeur définie sur la population que l'on cherche à estimer
- ▶ **Taille de l'échantillon** : nombre d'individus échantillonnés,  $n$
- ▶ **Statistique** : quantité aléatoire calculée sur l'échantillon (fluctue en fonction de l'échantillon)

---

---

### Définition

---

---

Une **statistique** qui permet d'estimer un paramètre s'appelle un **estimateur** (du paramètre sous-jacent)

---

---

Rem. : on utilise souvent la notation "chapeau", e.g.,  $\hat{x}_n$  pour désigner un estimateur

# Échantillonnage aléatoire simple

- ▶ Choisir  $n$  individus parmi les  $N$  de la population totale, de façon aléatoire et uniforme
- ▶ On s'interdit de choisir deux fois le même individu dans l'échantillon (tirage sans remise)
- ▶ en général  $n \ll N$

Rem. : différent de l'échantillonnage par strates (sondages, etc.)

# Dénombrement

Prenons l'exemple de la base de données `babies23.data` :

- ▶ population totale:  $N = 1226$  (valeurs manquantes éliminées)
- ▶ échantillon:  $n = 91$  (choix arbitraire ici)

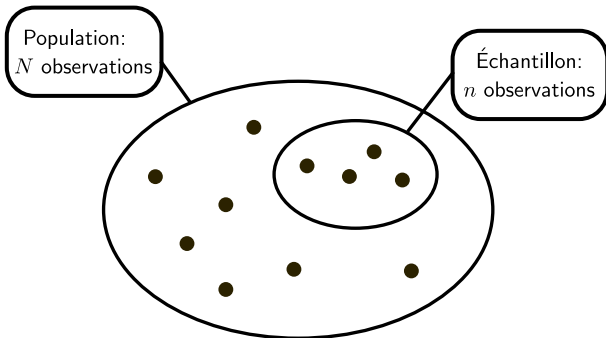
Combien y a-t-il d'échantillons possibles de taille 91 sur une population de taille 1226 ?

# Dénombrement

Rappel :  $N = 1226$ ,  $n = 91$

On veut choisir  $n$  individus distincts parmi les  $N$ :

Il y a  $\binom{1226}{91} \simeq 2.72 \times 10^{139}$  choix possibles.



# Coefficient binomial

Nombre d'échantillons possibles:

$$\frac{1226 \times 1225 \times \cdots \times 1136}{91 \times 90 \times \cdots \times 1} = \frac{1226!}{1135! \times 91!}$$

Ce dernier nombre se note<sup>(4)</sup>

$$\binom{1226}{91} \quad \text{ou} \quad C_{1226}^{91} \approx 2.72 \times 10^{139}$$

Rem. : il y a  $\binom{N}{n}$  échantillons possibles de taille  $n$  dans une population de  $N$  individus. En tirant un échantillon selon la loi uniforme, chaque échantillon a la même probabilité:

$$\frac{1}{\binom{N}{n}}$$

---

<sup>(4)</sup>  $\binom{1226}{91} =$

27163361289825550523268206145916032768087536721887922928079532500193847739178087  
753157271974542740511341704937942015497847278125901041363840

# Code Python

```
import pandas as pd
df_babies = pd.read_csv("babies23.data",
                        skiprows=38, sep='\s+')
n_samples = 91
df_extract = df_babies.sample(n=n_samples)
```

Vocabulaire : **échantillon** ( : *sample*)

Rem. : voir le notebook `Echantillonnage.ipynb`

# Échantillonnage

- ▶ L'échantillonnage aléatoire simple met une structure aléatoire sur l'échantillon
- ▶ Différents échantillons ont des propriétés statistiques différentes, liées à la méthode d'échantillonnage
- ▶ Question : quelles sont ces propriétés ici?

# Sommaire

Introduction

Loi d'échantillonnage et estimation

Statistique, ou propriété des quantités calculées sur un échantillon

Approximations gaussiennes et intervalles de confiance



## Moyenne empirique

Notons  $x_i$  la variable qui vaut 1 (ou 0) si la  $i^{\text{e}}$  mère fume (ou non)

On cherche à estimer le **paramètre** taux de tabagisme chez la mère

$$\mu := \bar{x}_N = \frac{1}{1226}(x_1 + x_2 + \cdots + x_{1226}) \quad (\text{parfois}) \text{ inconnu}$$

à partir de l'échantillon prélevé

**Exemple** : : (cf. notebook)  $\bar{x}_N = 682/1226 \approx 55.63\%$

Technique classique : prendre la moyenne sur l'échantillon, qui vaut

$$\bar{x}_n := \frac{1}{91}(x_{i_1} + x_{i_2} + \cdots + x_{i_{91}})$$

où  $i_k$  est le numéro du  $k^{\text{e}}$  individu échantillonné

**Exemple** : : (cf. notebook)  $\bar{x}_n = 52/91 \approx 57.14\%$

Rem. : les  $i_k$  sont des variables aléatoires ici

Question :  $\bar{x}_n$  (que l'on calcule) est-il éloigné de  $\mu$  (**inconnu**) ?

# Espérance

Rappel : les observations  $x_{i_1}, x_{i_2}, \dots, x_{i_{g_1}}$  sont aléatoires

Une moyenne par rapport à l'aléatoire s'appelle une **espérance**; celle du premier individu de notre échantillon vaut:

$$\begin{aligned}\mathbb{E}(x_{i_1}) &= \sum_{i=1}^N x_i \cdot \mathbb{P}(x_{i_1} = x_i) \\ &= \sum_{i=1}^N x_i \cdot \frac{1}{N} = \mu\end{aligned}$$

De même pour tous les individus de l'échantillon car ils sont supposés *i.i.d.* (**indépendants** et **identiquement** distribués)

$$\mathbb{E}(x_{i_1}) = \mathbb{E}(x_{i_2}) = \dots = \mathbb{E}(x_{i_{g_1}}) = \mu$$

## Espérance de la moyenne $\bar{x}_n$

L'espérance de la moyenne  $\bar{x}_n$  sur l'échantillon est

$$\begin{aligned}\mathbb{E}(\bar{x}_n) &= \frac{1}{91} \left( \mathbb{E}(x_{i_1}) + \mathbb{E}(x_{i_2}) + \cdots + \mathbb{E}(x_{i_{91}}) \right) \\ &= \frac{1}{91} \left( 91 \times \mu \right) = \mu.\end{aligned}$$

En "espérance" (en moyenne vis-à-vis de l'aléa de l'échantillonnage), notre estimateur est égal au paramètre  $\mu$

# Biais ( : *Bias*)

---

---

## Définition

---

---

Le **biais** d'un estimateur  $\hat{x}_n$  de  $\mu$  est noté  $\mathbb{B}(\hat{x}_n)$  et vaut

$$\mathbb{B}(\hat{x}_n) := \mathbb{E}(\hat{x}_n) - \mu$$

---

---

Interprétation :

le biais mesure l'erreur "systématique d'un estimateur"

Rem. : pour des variables *i.i.d.* et d'espérance  $\mu$ , la moyenne empirique est **sans biais** ou non biaisée  $\mathbb{B}(\bar{x}_n) = 0$

# Variance de l'estimateur

## Définition

La **variance** de l'estimateur  $\hat{x}_n$  est définie comme

$$\text{Var}(\hat{x}_n) = \mathbb{E} \left[ \left( \hat{x}_n - \mathbb{E}(\hat{x}_n) \right)^2 \right] = \mathbb{E}(\hat{x}_n^2) - (\mathbb{E}(\hat{x}_n))^2$$

Interprétation : la variance mesure la variation / dispersion d'un estimateur autour de son espérance

## Propriétés

Soit  $\alpha \in \mathbb{R}$ , et  $X, Y$  deux variables aléatoires indépendantes

$$\text{Var}(X + \alpha) = \text{Var}(X)$$

$$\text{Var}(\alpha X) = \alpha^2 \text{Var}(X)$$

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

## Variance de la moyenne empirique

Variance (en supposant les  $x_{i_k}$  indépendants) :

$$\text{Var}(\bar{x}_n) = \text{Var}\left(\frac{1}{n} \sum_{k=1}^n x_{i_k}\right)$$

## Variance de la moyenne empirique

Variance (en supposant les  $x_{i_k}$  indépendants) :

$$\text{Var}(\bar{x}_n) = \text{Var}\left(\frac{1}{n} \sum_{k=1}^n x_{i_k}\right) = \frac{1}{n^2} \text{Var}\left(\sum_{k=1}^n x_{i_k}\right)$$

## Variance de la moyenne empirique

Variance (en supposant les  $x_{i_k}$  indépendants) :

$$\begin{aligned}\text{Var}(\bar{x}_n) &= \text{Var}\left(\frac{1}{n} \sum_{k=1}^n x_{i_k}\right) = \frac{1}{n^2} \text{Var}\left(\sum_{k=1}^n x_{i_k}\right) \\ &= \frac{1}{n^2} \sum_{k=1}^n \text{Var}(x_{i_k}) = \frac{\text{Var}(x_{i_1})}{n}\end{aligned}$$



## Variance de la moyenne empirique

Variance (en supposant les  $x_{i_k}$  indépendants) :

$$\begin{aligned}\text{Var}(\bar{x}_n) &= \text{Var}\left(\frac{1}{n} \sum_{k=1}^n x_{i_k}\right) = \frac{1}{n^2} \text{Var}\left(\sum_{k=1}^n x_{i_k}\right) \\ &= \frac{1}{n^2} \sum_{k=1}^n \text{Var}(x_{i_k}) = \frac{\text{Var}(x_{i_1})}{n}\end{aligned}$$

Conclusion : la variance de la moyenne empirique est réduite d'un facteur  $n$  par rapport à celle d'une seule observation

## Variance de la moyenne empirique

Variance (en supposant les  $x_{i_k}$  indépendants) :

$$\begin{aligned}\text{Var}(\bar{x}_n) &= \text{Var}\left(\frac{1}{n} \sum_{k=1}^n x_{i_k}\right) = \frac{1}{n^2} \text{Var}\left(\sum_{k=1}^n x_{i_k}\right) \\ &= \frac{1}{n^2} \sum_{k=1}^n \text{Var}(x_{i_k}) = \frac{\text{Var}(x_{i_1})}{n}\end{aligned}$$

Conclusion : la variance de la moyenne empirique est réduite d'un facteur  $n$  par rapport à celle d'une seule observation

Rem :  $\sigma^2 := \text{Var}(x_{i_1}) = \dots = \text{Var}(x_{i_n})$

## Variance de la moyenne empirique

Variance (en supposant les  $x_{i_k}$  indépendants) :

$$\begin{aligned}\text{Var}(\bar{x}_n) &= \text{Var}\left(\frac{1}{n} \sum_{k=1}^n x_{i_k}\right) = \frac{1}{n^2} \text{Var}\left(\sum_{k=1}^n x_{i_k}\right) \\ &= \frac{1}{n^2} \sum_{k=1}^n \text{Var}(x_{i_k}) = \frac{\text{Var}(x_{i_1})}{n}\end{aligned}$$

Conclusion : la variance de la moyenne empirique est réduite d'un facteur  $n$  par rapport à celle d'une seule observation

Rem :  $\sigma^2 := \text{Var}(x_{i_1}) = \dots = \text{Var}(x_{i_n}) = \sum_{i=1}^N (x_i - \mu)^2 \mathbb{P}(x_{i_1} = x_i)$

## Variance de la moyenne empirique

Variance (en supposant les  $x_{i_k}$  indépendants) :

$$\begin{aligned}\text{Var}(\bar{x}_n) &= \text{Var}\left(\frac{1}{n} \sum_{k=1}^n x_{i_k}\right) = \frac{1}{n^2} \text{Var}\left(\sum_{k=1}^n x_{i_k}\right) \\ &= \frac{1}{n^2} \sum_{k=1}^n \text{Var}(x_{i_k}) = \frac{\text{Var}(x_{i_1})}{n}\end{aligned}$$

Conclusion : la variance de la moyenne empirique est réduite d'un facteur  $n$  par rapport à celle d'une seule observation

Rem :  $\sigma^2 := \text{Var}(x_{i_1}) = \dots = \text{Var}(x_{i_n}) = \sum_{i=1}^N (x_i - \mu)^2 \mathbb{P}(x_{i_1} = x_i)$

$$= \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

# Écart quadratique moyen

( : *Mean Squared Error, MSE*)

---

---

## Définition

---

---

L'**écart quadratique moyen** (ou **erreur quadratique moyenne**) d'un estimateur  $\hat{x}_n$  d'un paramètre  $\mu$  est donné par:  $\mathbb{E}(\hat{x}_n - \mu)^2$

---

---

Rem. : l'**écart quadratique moyen** mesure la performance d'un estimateur; plus il est petit, meilleur est l'estimateur

Propriété :

$$\mathbb{E}(\hat{x}_n - \mu)^2 = \text{Var}(\hat{x}_n) + \mathbb{B}(\hat{x}_n)^2$$

# Écart quadratique moyen

( : *Mean Squared Error, MSE*)

---

---

## Définition

---

---

L'**écart quadratique moyen** (ou **erreur quadratique moyenne**) d'un estimateur  $\hat{x}_n$  d'un paramètre  $\mu$  est donné par:  $\mathbb{E}(\hat{x}_n - \mu)^2$

---

---

Rem. : l'**écart quadratique moyen** mesure la performance d'un estimateur; plus il est petit, meilleur est l'estimateur

Propriété :

$$\mathbb{E}(\hat{x}_n - \mu)^2 = \text{Var}(\hat{x}_n) + \mathbb{B}(\hat{x}_n)^2$$

# Écart quadratique moyen

( : *Mean Squared Error, MSE*)

---

---

## Définition

---

---

L'**écart quadratique moyen** (ou **erreur quadratique moyenne**) d'un estimateur  $\hat{x}_n$  d'un paramètre  $\mu$  est donné par:  $\mathbb{E}(\hat{x}_n - \mu)^2$

---

---

Rem. : l'**écart quadratique moyen** mesure la performance d'un estimateur; plus il est petit, meilleur est l'estimateur

Propriété :

$$\mathbb{E}(\hat{x}_n - \mu)^2 = \text{Var}(\hat{x}_n) + \mathbb{B}(\hat{x}_n)^2$$

$$\mathbb{E}(\hat{x}_n - \mu)^2 = \mathbb{E}(\hat{x}_n - \mathbb{E}(\hat{x}_n) + \mathbb{E}(\hat{x}_n) - \mu)^2 = \mathbb{E}(\hat{x}_n - \mathbb{E}(\hat{x}_n) + \mathbb{B}(\hat{x}_n))^2$$

# Écart quadratique moyen

( : *Mean Squared Error, MSE*)

---

---

## Définition

---

---

L'**écart quadratique moyen** (ou **erreur quadratique moyenne**) d'un estimateur  $\hat{x}_n$  d'un paramètre  $\mu$  est donné par:  $\mathbb{E}(\hat{x}_n - \mu)^2$

---

---

Rem. : l'**écart quadratique moyen** mesure la performance d'un estimateur; plus il est petit, meilleur est l'estimateur

Propriété :

$$\mathbb{E}(\hat{x}_n - \mu)^2 = \text{Var}(\hat{x}_n) + \mathbb{B}(\hat{x}_n)^2$$

$$\begin{aligned} \mathbb{E}(\hat{x}_n - \mu)^2 &= \mathbb{E}(\hat{x}_n - \mathbb{E}(\hat{x}_n) + \mathbb{E}(\hat{x}_n) - \mu)^2 = \mathbb{E}(\hat{x}_n - \mathbb{E}(\hat{x}_n) + \mathbb{B}(\hat{x}_n))^2 \\ &= \mathbb{E}(\hat{x}_n - \mathbb{E}(\hat{x}_n))^2 + \underbrace{\mathbb{E}(\mathbb{B}(\hat{x}_n))^2}_{\text{}} + \underbrace{2\mathbb{E}((\hat{x}_n - \mathbb{E}(\hat{x}_n))\mathbb{B}(\hat{x}_n))}_{\text{}} \end{aligned}$$



# Écart quadratique moyen

( : *Mean Squared Error, MSE*)

---

---

## Définition

---

---

L'**écart quadratique moyen** (ou **erreur quadratique moyenne**) d'un estimateur  $\hat{x}_n$  d'un paramètre  $\mu$  est donné par:  $\mathbb{E}(\hat{x}_n - \mu)^2$

---

---

Rem. : l'**écart quadratique moyen** mesure la performance d'un estimateur; plus il est petit, meilleur est l'estimateur

Propriété :

$$\mathbb{E}(\hat{x}_n - \mu)^2 = \text{Var}(\hat{x}_n) + \mathbb{B}(\hat{x}_n)^2$$

$$\begin{aligned}\mathbb{E}(\hat{x}_n - \mu)^2 &= \mathbb{E}(\hat{x}_n - \mathbb{E}(\hat{x}_n) + \mathbb{E}(\hat{x}_n) - \mu)^2 = \mathbb{E}(\hat{x}_n - \mathbb{E}(\hat{x}_n) + \mathbb{B}(\hat{x}_n))^2 \\ &= \mathbb{E}(\hat{x}_n - \mathbb{E}(\hat{x}_n))^2 + \underbrace{\mathbb{E}(\mathbb{B}(\hat{x}_n))^2}_{\text{}} + 2 \underbrace{\mathbb{E}((\hat{x}_n - \mathbb{E}(\hat{x}_n))\mathbb{B}(\hat{x}_n))}_{\text{}}\end{aligned}$$

# Écart quadratique moyen

( : *Mean Squared Error, MSE*)

---

---

## Définition

---

---

L'**écart quadratique moyen** (ou **erreur quadratique moyenne**) d'un estimateur  $\hat{x}_n$  d'un paramètre  $\mu$  est donné par:  $\mathbb{E}(\hat{x}_n - \mu)^2$

---

---

Rem. : l'**écart quadratique moyen** mesure la performance d'un estimateur; plus il est petit, meilleur est l'estimateur

Propriété :

$$\mathbb{E}(\hat{x}_n - \mu)^2 = \text{Var}(\hat{x}_n) + \mathbb{B}(\hat{x}_n)^2$$

$$\begin{aligned}\mathbb{E}(\hat{x}_n - \mu)^2 &= \mathbb{E}(\hat{x}_n - \mathbb{E}(\hat{x}_n) + \mathbb{E}(\hat{x}_n) - \mu)^2 = \mathbb{E}(\hat{x}_n - \mathbb{E}(\hat{x}_n) + \mathbb{B}(\hat{x}_n))^2 \\ &= \mathbb{E}(\hat{x}_n - \mathbb{E}(\hat{x}_n))^2 + \underbrace{\mathbb{E}(\mathbb{B}(\hat{x}_n))^2}_{\mathbb{B}(\hat{x}_n)^2} + 2 \underbrace{\mathbb{E}((\hat{x}_n - \mathbb{E}(\hat{x}_n))\mathbb{B}(\hat{x}_n))}_{\mathbb{B}(\hat{x}_n)\mathbb{E}(\hat{x}_n - \mathbb{E}(\hat{x}_n))}\end{aligned}$$

# Écart quadratique moyen

( : *Mean Squared Error, MSE*)

---

---

## Définition

---

---

L'**écart quadratique moyen** (ou **erreur quadratique moyenne**) d'un estimateur  $\hat{x}_n$  d'un paramètre  $\mu$  est donné par:  $\mathbb{E}(\hat{x}_n - \mu)^2$

---

---

Rem. : l'**écart quadratique moyen** mesure la performance d'un estimateur; plus il est petit, meilleur est l'estimateur

Propriété :

$$\mathbb{E}(\hat{x}_n - \mu)^2 = \text{Var}(\hat{x}_n) + \mathbb{B}(\hat{x}_n)^2$$

$$\begin{aligned}\mathbb{E}(\hat{x}_n - \mu)^2 &= \mathbb{E}(\hat{x}_n - \mathbb{E}(\hat{x}_n) + \mathbb{E}(\hat{x}_n) - \mu)^2 = \mathbb{E}(\hat{x}_n - \mathbb{E}(\hat{x}_n) + \mathbb{B}(\hat{x}_n))^2 \\ &= \mathbb{E}(\hat{x}_n - \mathbb{E}(\hat{x}_n))^2 + \underbrace{\mathbb{E}(\mathbb{B}(\hat{x}_n))^2}_{\mathbb{B}(\hat{x}_n)^2} + 2 \underbrace{\mathbb{E}((\hat{x}_n - \mathbb{E}(\hat{x}_n))\mathbb{B}(\hat{x}_n))}_{\mathbb{B}(\hat{x}_n)\mathbb{E}(\hat{x}_n - \mathbb{E}(\hat{x}_n))=0}\end{aligned}$$

# Écart quadratique moyen

( : *Mean Squared Error, MSE*)

---

---

## Définition

---

---

L'**écart quadratique moyen** (ou **erreur quadratique moyenne**) d'un estimateur  $\hat{x}_n$  d'un paramètre  $\mu$  est donné par:  $\mathbb{E}(\hat{x}_n - \mu)^2$

---

---

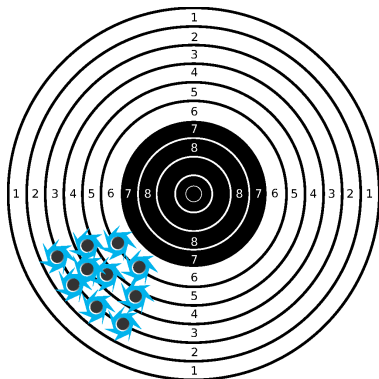
Rem. : l'**écart quadratique moyen** mesure la performance d'un estimateur; plus il est petit, meilleur est l'estimateur

Propriété :

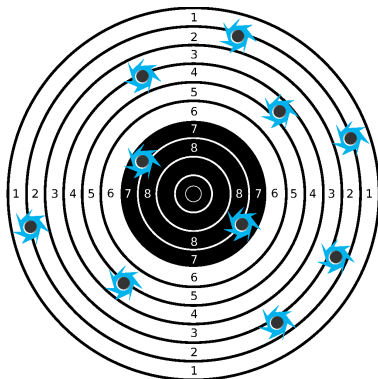
$$\mathbb{E}(\hat{x}_n - \mu)^2 = \text{Var}(\hat{x}_n) + \mathbb{B}(\hat{x}_n)^2$$

$$\begin{aligned}\mathbb{E}(\hat{x}_n - \mu)^2 &= \mathbb{E}(\hat{x}_n - \mathbb{E}(\hat{x}_n) + \mathbb{E}(\hat{x}_n) - \mu)^2 = \mathbb{E}(\hat{x}_n - \mathbb{E}(\hat{x}_n) + \mathbb{B}(\hat{x}_n))^2 \\ &= \mathbb{E}(\hat{x}_n - \mathbb{E}(\hat{x}_n))^2 + \underbrace{\mathbb{E}(\mathbb{B}(\hat{x}_n))^2}_{\mathbb{B}(\hat{x}_n)^2} + 2 \underbrace{\mathbb{E}((\hat{x}_n - \mathbb{E}(\hat{x}_n))\mathbb{B}(\hat{x}_n))}_{\mathbb{B}(\hat{x}_n)\mathbb{E}(\hat{x}_n - \mathbb{E}(\hat{x}_n))=0} \\ &= \text{Var}(\hat{x}_n) + \mathbb{B}(\hat{x}_n)^2\end{aligned}$$

# Biais ou variance?

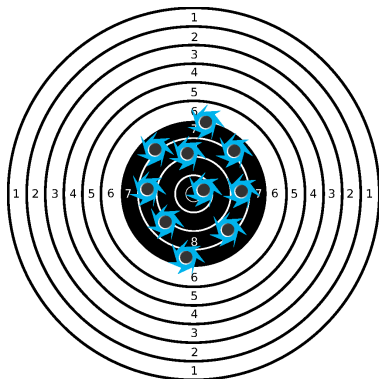


Erreurs systématiques  
(fort biais)

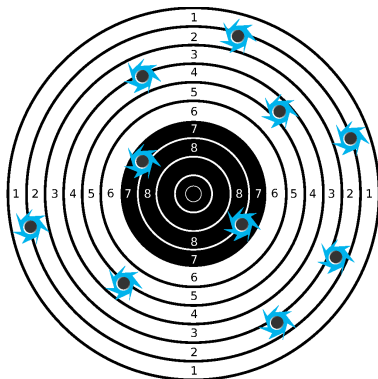


Erreurs stochastiques  
(forte variance)

# Biais ou variance?

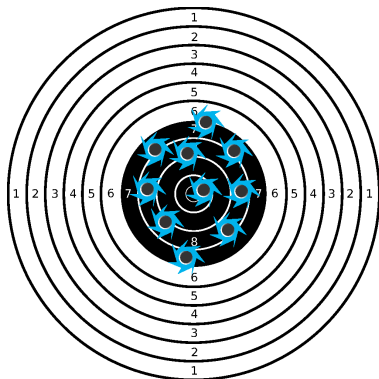


Biais nul  
faible variance

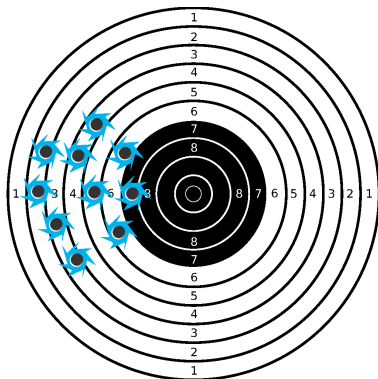


Biais nul  
forte variance

## Biais ou variance?



Biais nul  
faible variance



Biais important  
faible variance

# Estimation avec ou sans biais de $\sigma^2$

---

---

## Définition

---

---

La variance empirique est définie par

$$s_n^2(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2$$

---

---

Biais : (cas *i.i.d.*)  $\mathbb{E}(s_n^2(\mathbf{x})) = \frac{n-1}{n} \sigma^2$  (cf. calcul en TD)

Rem. : l'estimateur<sup>(5)</sup> de  $\sigma^2$ ,  $\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x}_n)^2$  est **sans biais**, mais dans le cas *i.i.d.* gaussien, son risque quadratique ( $= \frac{2\sigma^4}{n-1}$ ) est plus grand que celui de  $s_n(\mathbf{x})$  ( $= \frac{2n-1}{n^2} \sigma^4$ )

---

<sup>(5)</sup> appelé parfois l'estimateur sans biais de la variance



# Sommaire

Introduction

Loi d'échantillonnage et estimation

Statistique, ou propriété des quantités calculées sur un échantillon

Approximations gaussiennes et intervalles de confiance

# Théorème Central Limite (TCL)

Si la taille de l'échantillon est grande la moyenne empirique  $\bar{x}_n$  est distribuée approximativement suivant une loi gaussienne :

---

---

## Théorème

---

---

Soient  $x_1, \dots, x_n$  des variables aléatoires indépendantes, distribuées suivant la même loi, d'espérance  $\mu$  et de variance  $\sigma^2$  ; Alors, si  $n$  est grand ( $n \geq 30$ ), la variable

$$Z = \frac{\bar{x}_n - \mu}{\sigma/\sqrt{n}}$$

suit approximativement une loi normale centrée, réduite  $\mathcal{N}(0, 1)$ .

---

---

Rem. :  $\text{Var}(\bar{x}_n) = \frac{\sigma^2}{n}$  donc  $\text{Var}(Z) = 1$  et  $\mathbb{E}(Z) = 0$

Notation :  $q_{\mathcal{N}}(1 - \alpha)$ , quantile de la loi  $\mathcal{N}(0, 1)$  au niveau  $1 - \alpha$

## Échantillonnage aléatoire simple (retour)

Pour un tel échantillon, les variables  $x_{i_k}$ ,  $k = 1, \dots, n$  ont bien la même loi, mais ne sont pas indépendantes (tirage **sans** remise).

Si  $n \ll N$ , on peut considérer que la dépendance est tellement faible qu'elle est négligeable (et le TCL s'applique).

**Bilan** : La moyenne empirique  $\bar{x}_n$  est distribuée approximativement suivant une loi gaussienne si  $n$  (taille de l'échantillon) est grand.

## Échantillonnage aléatoire simple (retour)

Pour un tel échantillon, les variables  $x_{i_k}$ ,  $k = 1, \dots, n$  ont bien la même loi, mais ne sont pas indépendantes (tirage **sans** remise).

Si  $n \ll N$ , on peut considérer que la dépendance est tellement faible qu'elle est négligeable (et le TCL s'applique).

**Bilan** : La moyenne empirique  $\bar{x}_n$  est distribuée approximativement suivant une loi gaussienne si  $n$  (taille de l'échantillon) est grand.

## Intervalles de confiance

Si l'on souhaite estimer  $\mu$ , un premier intervalle de confiance est

$$\left[ \bar{x}_n - \frac{\sigma}{\sqrt{n}}, \bar{x}_n + \frac{\sigma}{\sqrt{n}} \right]$$

Cela nous donne un intervalle de confiance au niveau 68%:

$$\mathbb{P} \left( \mu \in \left[ \bar{x}_n - \frac{\sigma}{\sqrt{n}}, \bar{x}_n + \frac{\sigma}{\sqrt{n}} \right] \right)$$

## Intervalle de confiance

Si l'on souhaite estimer  $\mu$ , un premier intervalle de confiance est

$$\left[ \bar{x}_n - \frac{\sigma}{\sqrt{n}}, \bar{x}_n + \frac{\sigma}{\sqrt{n}} \right]$$

Cela nous donne un intervalle de confiance au niveau 68%:

$$\mathbb{P} \left( \mu \in \left[ \bar{x}_n - \frac{\sigma}{\sqrt{n}}, \bar{x}_n + \frac{\sigma}{\sqrt{n}} \right] \right) = \mathbb{P} \left( \mu - \bar{x}_n \in \left[ -\frac{\sigma}{\sqrt{n}}, +\frac{\sigma}{\sqrt{n}} \right] \right)$$

## Intervalle de confiance

Si l'on souhaite estimer  $\mu$ , un premier intervalle de confiance est

$$\left[ \bar{x}_n - \frac{\sigma}{\sqrt{n}}, \bar{x}_n + \frac{\sigma}{\sqrt{n}} \right]$$

Cela nous donne un intervalle de confiance au niveau 68 %:

$$\begin{aligned} \mathbb{P} \left( \mu \in \left[ \bar{x}_n - \frac{\sigma}{\sqrt{n}}, \bar{x}_n + \frac{\sigma}{\sqrt{n}} \right] \right) &= \mathbb{P} \left( \mu - \bar{x}_n \in \left[ -\frac{\sigma}{\sqrt{n}}, +\frac{\sigma}{\sqrt{n}} \right] \right) \\ &= \mathbb{P} \left( \underbrace{\frac{\mu - \bar{x}_n}{\sigma/\sqrt{n}}}_{\sim \mathcal{N}(0,1)} \in [-1, +1] \right) \end{aligned}$$

## Intervalle de confiance

Si l'on souhaite estimer  $\mu$ , un premier intervalle de confiance est

$$\left[ \bar{x}_n - \frac{\sigma}{\sqrt{n}}, \bar{x}_n + \frac{\sigma}{\sqrt{n}} \right]$$

Cela nous donne un intervalle de confiance au niveau 68 %:

$$\begin{aligned} \mathbb{P} \left( \mu \in \left[ \bar{x}_n - \frac{\sigma}{\sqrt{n}}, \bar{x}_n + \frac{\sigma}{\sqrt{n}} \right] \right) &= \mathbb{P} \left( \mu - \bar{x}_n \in \left[ -\frac{\sigma}{\sqrt{n}}, +\frac{\sigma}{\sqrt{n}} \right] \right) \\ &= \mathbb{P} \left( \underbrace{\frac{\mu - \bar{x}_n}{\sigma/\sqrt{n}}}_{\sim \mathcal{N}(0,1)} \in [-1, +1] \right) \\ &= \Phi(1) - \Phi(-1) = 2\Phi(1) - 1 \\ &\approx 0.68 \end{aligned}$$



## Intervalles de confiance

Si l'on souhaite estimer  $\mu$ , un premier intervalle de confiance est

$$\left[ \bar{x}_n - \frac{\sigma}{\sqrt{n}}, \bar{x}_n + \frac{\sigma}{\sqrt{n}} \right]$$

Cela nous donne un intervalle de confiance au niveau 68 %:

$$\begin{aligned} \mathbb{P} \left( \mu \in \left[ \bar{x}_n - \frac{\sigma}{\sqrt{n}}, \bar{x}_n + \frac{\sigma}{\sqrt{n}} \right] \right) &= \mathbb{P} \left( \mu - \bar{x}_n \in \left[ -\frac{\sigma}{\sqrt{n}}, +\frac{\sigma}{\sqrt{n}} \right] \right) \\ &= \mathbb{P} \left( \underbrace{\frac{\mu - \bar{x}_n}{\sigma/\sqrt{n}}}_{\sim \mathcal{N}(0,1)} \in [-1, +1] \right) \\ &= \Phi(1) - \Phi(-1) = 2\Phi(1) - 1 \\ &\approx 0.68 \end{aligned}$$

Rem. :  $\Phi$  fonction de répartition d'une gaussienne centrée-réduite

## Intervalles de confiance : précision fixée

Objectif : trouver  $\delta > 0$  pour créer un intervalle de confiance (IC), tel que  $\mu$  appartienne à l'IC avec probabilité  $1 - \alpha$  :

$$\left[ \bar{x}_n - \delta \frac{\sigma}{\sqrt{n}}; \bar{x}_n + \delta \frac{\sigma}{\sqrt{n}} \right]$$

On résout donc en  $\delta > 0$  :

$$1 - \alpha = \mathbb{P} \left( \mu \in \left[ \bar{x}_n - \delta \frac{\sigma}{\sqrt{n}}; \bar{x}_n + \delta \frac{\sigma}{\sqrt{n}} \right] \right)$$

## Intervalles de confiance : précision fixée

Objectif : trouver  $\delta > 0$  pour créer un intervalle de confiance (IC), tel que  $\mu$  appartienne à l'IC avec probabilité  $1 - \alpha$  :

$$\left[ \bar{x}_n - \delta \frac{\sigma}{\sqrt{n}}; \bar{x}_n + \delta \frac{\sigma}{\sqrt{n}} \right]$$

On résout donc en  $\delta > 0$  :

$$\begin{aligned} 1 - \alpha &= \mathbb{P} \left( \mu \in \left[ \bar{x}_n - \delta \frac{\sigma}{\sqrt{n}}; \bar{x}_n + \delta \frac{\sigma}{\sqrt{n}} \right] \right) \\ &= \mathbb{P} \left( \mu - \bar{x}_n \in \left[ -\delta \frac{\sigma}{\sqrt{n}}, +\delta \frac{\sigma}{\sqrt{n}} \right] \right) \end{aligned}$$

## Intervalles de confiance : précision fixée

Objectif : trouver  $\delta > 0$  pour créer un intervalle de confiance (IC), tel que  $\mu$  appartienne à l'IC avec probabilité  $1 - \alpha$  :

$$\left[ \bar{x}_n - \delta \frac{\sigma}{\sqrt{n}}; \bar{x}_n + \delta \frac{\sigma}{\sqrt{n}} \right]$$

On résout donc en  $\delta > 0$  :

$$\begin{aligned} 1 - \alpha &= \mathbb{P} \left( \mu \in \left[ \bar{x}_n - \delta \frac{\sigma}{\sqrt{n}}; \bar{x}_n + \delta \frac{\sigma}{\sqrt{n}} \right] \right) \\ &= \mathbb{P} \left( \mu - \bar{x}_n \in \left[ -\delta \frac{\sigma}{\sqrt{n}}, +\delta \frac{\sigma}{\sqrt{n}} \right] \right) \\ &= \mathbb{P} \left( \frac{\mu - \bar{x}_n}{\sigma/\sqrt{n}} \in [-\delta, +\delta] \right) \end{aligned}$$

## Intervalles de confiance : précision fixée

Objectif : trouver  $\delta > 0$  pour créer un intervalle de confiance (IC), tel que  $\mu$  appartienne à l'IC avec probabilité  $1 - \alpha$  :

$$\left[ \bar{x}_n - \delta \frac{\sigma}{\sqrt{n}}; \bar{x}_n + \delta \frac{\sigma}{\sqrt{n}} \right]$$

On résout donc en  $\delta > 0$  :

$$\begin{aligned} 1 - \alpha &= \mathbb{P} \left( \mu \in \left[ \bar{x}_n - \delta \frac{\sigma}{\sqrt{n}}; \bar{x}_n + \delta \frac{\sigma}{\sqrt{n}} \right] \right) \\ &= \mathbb{P} \left( \mu - \bar{x}_n \in \left[ -\delta \frac{\sigma}{\sqrt{n}}, +\delta \frac{\sigma}{\sqrt{n}} \right] \right) \\ &= \mathbb{P} \left( \frac{\mu - \bar{x}_n}{\sigma/\sqrt{n}} \in [-\delta, +\delta] \right) \\ &= \Phi(\delta) - \Phi(-\delta) = 2\Phi(\delta) - 1 \end{aligned}$$

## Intervalles de confiance : précision fixée

Objectif : trouver  $\delta > 0$  pour créer un intervalle de confiance (IC), tel que  $\mu$  appartienne à l'IC avec probabilité  $1 - \alpha$  :

$$\left[ \bar{x}_n - \delta \frac{\sigma}{\sqrt{n}}; \bar{x}_n + \delta \frac{\sigma}{\sqrt{n}} \right]$$

On résout donc en  $\delta > 0$  :

$$1 - \alpha = \mathbb{P} \left( \mu \in \left[ \bar{x}_n - \delta \frac{\sigma}{\sqrt{n}}; \bar{x}_n + \delta \frac{\sigma}{\sqrt{n}} \right] \right)$$

$$= \mathbb{P} \left( \mu - \bar{x}_n \in \left[ -\delta \frac{\sigma}{\sqrt{n}}, +\delta \frac{\sigma}{\sqrt{n}} \right] \right)$$

$$= \mathbb{P} \left( \frac{\mu - \bar{x}_n}{\sigma/\sqrt{n}} \in [-\delta, +\delta] \right)$$

$$= \Phi(\delta) - \Phi(-\delta) = 2\Phi(\delta) - 1$$

$$\iff \delta = \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right) := q_{\mathcal{N}} \left( 1 - \frac{\alpha}{2} \right) \quad (\text{quantile gaussien } 1 - \frac{\alpha}{2})$$

## Intervalles de confiance : précision fixée

Objectif : trouver  $\delta > 0$  pour créer un intervalle de confiance (IC), tel que  $\mu$  appartienne à l'IC avec probabilité  $1 - \alpha$  :

$$\left[ \bar{x}_n - \delta \frac{\sigma}{\sqrt{n}}; \bar{x}_n + \delta \frac{\sigma}{\sqrt{n}} \right]$$

On résout donc en  $\delta > 0$  :

$$1 - \alpha = \mathbb{P} \left( \mu \in \left[ \bar{x}_n - \delta \frac{\sigma}{\sqrt{n}}; \bar{x}_n + \delta \frac{\sigma}{\sqrt{n}} \right] \right)$$

$$= \mathbb{P} \left( \mu - \bar{x}_n \in \left[ -\delta \frac{\sigma}{\sqrt{n}}, +\delta \frac{\sigma}{\sqrt{n}} \right] \right)$$

$$= \mathbb{P} \left( \frac{\mu - \bar{x}_n}{\sigma/\sqrt{n}} \in [-\delta, +\delta] \right)$$

$$= \Phi(\delta) - \Phi(-\delta) = 2\Phi(\delta) - 1$$

$$\iff \delta = \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right) := q_{\mathcal{N}} \left( 1 - \frac{\alpha}{2} \right) \quad (\text{quantile gaussien } 1 - \frac{\alpha}{2})$$

IC au niveau  $1 - \alpha$ :  $\left[ \bar{x}_n - \frac{\sigma}{\sqrt{n}} q_{\mathcal{N}} \left( 1 - \frac{\alpha}{2} \right); \bar{x}_n + \frac{\sigma}{\sqrt{n}} q_{\mathcal{N}} \left( 1 - \frac{\alpha}{2} \right) \right]$

## Qu'est-ce qu'un niveau de confiance ?

Que veut dire la phrase : “le niveau de confiance de l'intervalle [...;...] est de 95 %” ?



## Qu'est-ce qu'un niveau de confiance ?

Que veut dire la phrase : “le niveau de confiance de l'intervalle [...;...] est de 95 %” ?

Si l'on prend beaucoup d'échantillons de taille  $n$ , on peut calculer autant de valeurs de  $\bar{x}_n$  que l'on a d'échantillons, et donc autant d'intervalles de confiance que d'échantillons.

La phrase ci-dessus indique que la moyenne  $\mu$  (inconnue) sur la population se trouve dans 95 % de ces intervalles de confiance.

## Qu'est-ce qu'un niveau de confiance ?

Que veut dire la phrase : “le niveau de confiance de l'intervalle [...;...] est de 95 %” ?

Si l'on prend beaucoup d'échantillons de taille  $n$ , on peut calculer autant de valeurs de  $\bar{x}_n$  que l'on a d'échantillons, et donc autant d'intervalles de confiance que d'échantillons.

La phrase ci-dessus indique que la moyenne  $\mu$  (**inconnue**) sur la population se trouve dans 95 % de ces intervalles de confiance.

$$\begin{aligned} \text{Ordre de grandeur : } \alpha = 5\% &\implies q_{\mathcal{N}}\left(1 - \frac{\alpha}{2}\right) = q_{\mathcal{N}}(0.975) \\ &: \alpha = 1\% \implies q_{\mathcal{N}}\left(1 - \frac{\alpha}{2}\right) = q_{\mathcal{N}}(0.995) \end{aligned}$$

## Qu'est-ce qu'un niveau de confiance ?

Que veut dire la phrase : “le niveau de confiance de l'intervalle [...;...] est de 95 %” ?

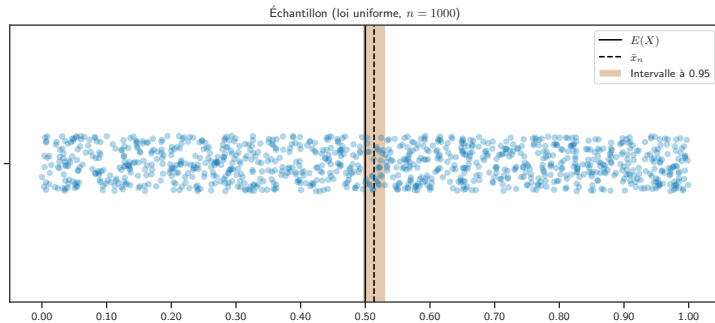
Si l'on prend beaucoup d'échantillons de taille  $n$ , on peut calculer autant de valeurs de  $\bar{x}_n$  que l'on a d'échantillons, et donc autant d'intervalles de confiance que d'échantillons.

La phrase ci-dessus indique que la moyenne  $\mu$  (inconnue) sur la population se trouve dans 95 % de ces intervalles de confiance.

$$\begin{aligned} \text{Ordre de grandeur : } \alpha = 5\% &\implies q_{\mathcal{N}}\left(1 - \frac{\alpha}{2}\right) = q_{\mathcal{N}}(0.975) \\ &: \alpha = 1\% \implies q_{\mathcal{N}}\left(1 - \frac{\alpha}{2}\right) = q_{\mathcal{N}}(0.995) \end{aligned}$$

```
>>> from scipy.stats import norm
>>> norm.ppf(0.975), norm.ppf(0.995)
(1.959963984540054, 1.6448536269514722)
```

# Exemple numérique



# Biographie du jour : Pierre-Simon de Laplace<sup>(6)</sup>



- ▶ mathématicien, astronome, physicien et homme politique français (1749-1827)
- ▶ Démontre le premier la forme générale du théorème central limite
- ▶ Introduit la distribution de Laplace
- ▶ Un des premiers démographes: estime la population française avec un **intervalle de confiance**
- ▶  $\Delta f = 0$  équation de Laplace / Laplacien (en physique)
- ▶ ...

---

<sup>(6)</sup>[https://fr.wikipedia.org/wiki/Pierre-Simon\\_de\\_Laplace](https://fr.wikipedia.org/wiki/Pierre-Simon_de_Laplace)

# Bibliographie I

- ▶ Nolan, D. and T. P. Speed. *Stat labs: mathematical statistics through applications*. Springer Science & Business Media, 2001.