

# Part. 1. Statistiques mathématiques : Estimation et tests

K. Meziani, IASD



## Plan Partie 1

### 2 Estimation paramétrique

- ① Echantillon
- ② Statistique paramétrique
- ③ Qualité et comparaison des estimateurs
- ④ Loi asymptotique des estimateurs

### 3 Tests


- ① Tests
- ② Tests asymptotiques
- ③ Intervalle de confiance
- ④ *p-value*

## 2.1. Échantillon

Matériel de départ : données collectées à partir d'une partie d'une population (de différentes tailles et nature.)


### Echantillon : point de vue "Applications"

Suite  $(x_1, \dots, x_n)$  finie d'observations/données au cours d'une expérience.

 En statistiques descriptives:  $(x_1, \dots, x_n)$  est appelée série statistique

### Echantillon : point de vue "Mathématiques"

Suite  $(X_1, \dots, X_n)$  finie de variables aléatoires.

 Dans le modèle le plus simple, les variables  $X_1, \dots, X_n$  sont **i.i.d.**, **indépendantes et identiquement distribuées**, de même loi  $F$  inconnue.

Objectif : estimer une loi inconnue (ou inférer au sujet d'une loi inconnue) à partir d'un échantillon  $X_1, \dots, X_n$  i.i.d. de même loi  $F$  inconnue. On note

$$X_i \stackrel{i.i.d}{\sim} F$$

## 2.2. Statistique paramétrique

Dans ce cours,  $F$  connue à un paramètre  $\theta \subseteq \mathbb{R}^r$  près. On parle de **statistique paramétrique**.

### Modèle statistique (paramétrique)

Soit  $r \in \mathbb{N}^*$ ,

$$\{F_\theta, \theta \in \Theta \subseteq \mathbb{R}^r\},$$

### Hypothèse d'identifiabilité

Pour tout  $\theta, \theta' \in \Theta$

$$F_\theta(\cdot) = F_{\theta'}(\cdot) \Rightarrow \theta = \theta'.$$

*Exemple : (Lancer pièce de monnaie)* On peut modéliser l'expérience par une loi de Bernoulli de paramètre  $\theta$ :

$$F := F_\theta = \mathcal{B}(\theta) \quad \text{où} \quad \theta \in [0, 1]$$

# Estimation statistique

## Estimateur du vrai $\theta^*$

Un **estimateur** de la vraie valeur  $\theta^*$  est construit à partir d'un échantillon  $(X_1, \dots, X_n)$  et est noté

$$\hat{\theta}_n := \hat{\theta}_n(X_1, \dots, X_n)$$

## Estimateur du maximum vraisemblance

- **La vraisemblance** est la loi jointe de  $(X_1, \dots, X_n)$
- Pour un échantillon i.i.d de loi  $f_\theta$ , **la vraisemblance** est

$$L(X_1, \dots, X_n, \theta) = f_\theta(X_1)f_\theta(X_2) \cdots f_\theta(X_n)$$

- Pour un échantillon donnée, **l'estimateur du maximum de vraisemblance**  $\hat{\theta}$  maximise la vraisemblance en  $\theta$

$$\hat{\theta} = \arg \max_{\theta \in \Theta} L(X_1, \dots, X_n, \theta)$$

Remarque Il existe d'autres estimateurs célèbres : estimateur des moindres carrés, estimateur des moments, ...

# Quelques estimateurs “classiques”

## Estimateurs classiques $\hat{\theta}$

- Un estimateur de  $\theta = E(X)$  est

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

- Un estimateur de  $\theta = V(X)$  est

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2 = s_X^2.$$

- Un estimateur de  $\theta = Cov(X, Y)$  est

$$\hat{\theta}_n = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)(Y_i - \bar{Y}_n) = s_{X,Y}.$$

## 2.3. Qualité et comparaison des estimateurs

- **Qu'est-ce qu'un bon estimateur?**

- Intuitivement, un estimateur est bon, s'il est proche de la vraie valeur du paramètre.

- **Difficulté** : un estimateur est une variable aléatoire:


- La notion de proximité peut avoir plusieurs interprétations.

## Biais

Un estimateur  $\hat{\theta}_n$  de  $\theta$  est **sans biais** si

$$E[\hat{\theta}_n] = \theta.$$

i.e. en moyenne  $\hat{\theta}_n$  est égal à  $\theta$ .

  $E[\hat{\theta}_n] \neq \theta$ ,  $\hat{\theta}_n$  est dit "**biaisé**".

### Remarques:

- $\bar{X}$  est un estimateur sans biais de la moyenne.
- $s^2$  est un estimateur biaisé de la variance.
- Un estimateur sans biais de la variance est

$$s_{X_C}^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2.$$



## Consistance

Un estimateur  $\hat{\theta}_n$  de  $\theta$  est **consistant** si

$$\hat{\theta}_n \xrightarrow[n \rightarrow +\infty]{P_\theta} \theta.$$

i.e. pour un  $n$  très grand  $\hat{\theta}_n$  est égal à  $\theta$ .

### Remarques:

- $\bar{X}$  est un estimateur consistant de la moyenne.
- $s_X^2$  est un estimateur consistant de la variance.

# Est-ce suffisant pour juger de la qualité d'un estimateur?

## Consistance

Si  $a_n$  une suite déterministe arbitraire telle que  $a_n \rightarrow 1$  alors

$$a_n \hat{\theta}_n \xrightarrow[n \rightarrow +\infty]{P_\theta} \theta.$$

- Pas assez informative pour nous guider dans le choix d'estimateurs.
- Les estimateurs non consistants doivent être avec certitude exclus.

## Biais

Valoir en moyenne le bon paramètre n'est pas suffisant si grande variance.

## Risque quadratique

**Risque quadratique (erreur moyenne quadratique)** de  $\hat{\theta}_n$  au point  $\theta \in \mathbb{R}$

$$R_n(\theta, \hat{\theta}_n) = E[(\hat{\theta}_n - \theta)^2].$$

Mesure la distance entre  $\hat{\theta}_n$  et un  $\theta$ .

## Structure du risque : biais et variance

$$R_n(\theta, \hat{\theta}_n) = \underbrace{\left(E[\hat{\theta}_n] - \theta\right)^2}_{\text{terme de biais}} + \underbrace{\text{Var}[\hat{\theta}_n]}_{\text{terme de variance}}.$$

## Discussion

- Un bon critère est le risque. Ainsi, un "bon" estimateur fera le bon "équilibre" entre le biais et la variance.
- **Un estimateur sans biais peut être moins efficace qu'un estimateur biaisé.**
- Tous les estimateurs raisonnables sont asymptotiquement sans biais.
- Privilégier la comparaison asymptotique d'estimateurs.

## 2.4. Loi asymptotique des estimateurs

### Théorème important

Lorsqu'un estimateur est **consistant**, on peut montrer sous certaines hypothèses

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n \rightarrow +\infty]{\text{Loi}_Q} \text{Loi normale centrée.}$$

### 3. Introduction

#### *Détection de missiles*

Une des premières applications de la théorie des tests : problème militaire de détection de missiles à l'aide de radar.

$\mathcal{H}_0$  : L'écho de radar est "grand" si un missile est présent.

$\mathcal{H}_1$  : L'écho de radar est "petit" dans le cas contraire.

## 3.1. Test

### Test

Un **test**  $\mathcal{H}_0$  vs  $\mathcal{H}_1$  est une règle qui, pour tout échantillon donné  $\mathcal{X}_n = (X_1, \dots, X_n)$ , dit si l'on rejette ou non  $\mathcal{H}_0$ . Pour cela, on définit une **région de rejet**  $R$  telle que

- $\mathcal{X}_n \in R$  on rejette  $\mathcal{H}_0$ .
- $\mathcal{X}_n \notin R$  on ne rejette pas  $\mathcal{H}_0$ .

# Risques

## Risque de première espèce

Probabilité de rejeter  $\mathcal{H}_0$  alors que  $\mathcal{H}_0$  est vraie. *Rejet sans raison de  $\mathcal{H}_0$  : un missile est présent ( $\rightarrow$  dangereux).*

## Risque de seconde espèce

Probabilité de ne pas rejeter  $\mathcal{H}_0$  alors que  $\mathcal{H}_0$  est fausse. *Fausse alerte missile ( $\rightarrow$  moins grave).*

## Détection de missiles

**Choisir de  $R$ ?** Minimiser les deux risques simultanément.

### Problème:

- Minimiser risque de première espèce  $\Leftrightarrow$  choisir  $R$  aussi petit que possible.
- Minimiser risque de seconde espèce  $\Leftrightarrow$  choisir  $R$  aussi grand que possible.



# Risques

## Risque de première espèce

Probabilité de rejeter  $\mathcal{H}_0$  alors que  $\mathcal{H}_0$  est vraie. *Rejet sans raison de  $\mathcal{H}_0$  : un missile est présent ( $\rightarrow$  dangereux).*

## Risque de seconde espèce

Probabilité de ne pas rejeter  $\mathcal{H}_0$  alors que  $\mathcal{H}_0$  est fausse. *Fausse alerte missile ( $\rightarrow$  moins grave).*

## Approche Neyman- Pearson

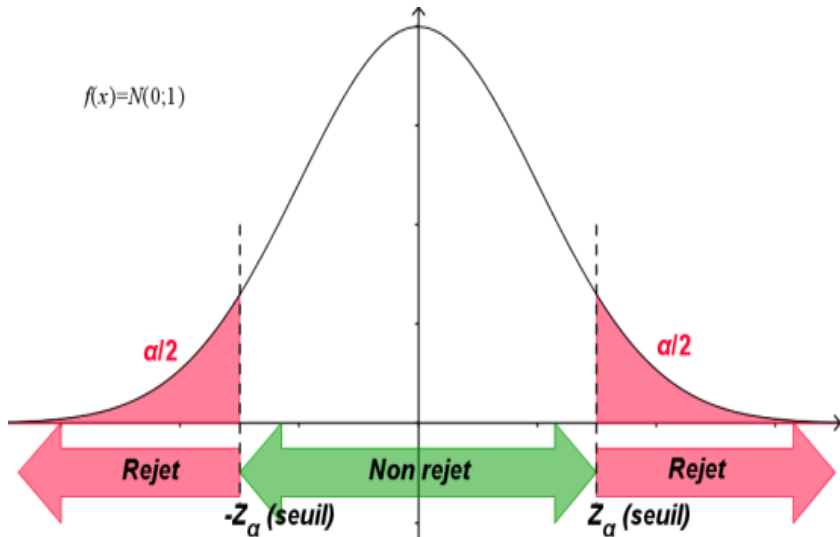
1. On fixe une borne  $\alpha \in ]0, 1[$  pour le risque de première espèce :

**Risque de première espèce  $\leq \alpha$ , avec  $\alpha$  petit (1%, 5% et 10%).**

**On dit que le test est de niveau  $\alpha$ .**

2. Ayant borné/fixé le risque de première espèce, il est naturel de chercher à minimiser le risque de seconde espèce.

## Exemple : $\mathcal{H}_0 : \theta = 0$ contre $\mathcal{H}_1 : \theta \neq 0$



## 3.2. Tests asymptotiques

### Test asymptotique d'hypothèse

Un **test asymptotique d'hypothèse**  $\mathcal{H}_0$  est une règle qui, pour tout échantillon donné  $\mathcal{X}_n = (X_1, \dots, X_n)$ , dit si l'on rejette ou non  $\mathcal{H}_0$ .

On fixe une borne  $\alpha \in ]0, 1[$  pour

$$\lim_{n \rightarrow \infty} \text{Risque de première espèce} \leq \alpha, \text{ avec } \alpha \text{ petit.}$$

### 3.3. Intervalle de confiance

#### Intervalle de confiance de niveau $1 - \alpha$ pour $\theta$

Un ensemble aléatoire  $\mathcal{C}(\mathcal{X}_n) \subseteq \mathbb{R}^r$  tel que pour tout  $\theta \in \Theta$

$$\text{Probabilité}(\theta \in \mathcal{C}(\mathcal{X}_n)) \geq 1 - \alpha.$$

#### Intervalle de confiance de niveau asymptotique $1 - \alpha$ pour $\theta$

Un ensemble aléatoire  $\mathcal{C}(\mathcal{X}_n) \subseteq \mathbb{R}^r$  tel que pour tout  $\theta \in \Theta$

$$\lim_{n \rightarrow \infty} \text{Probabilité}(\theta \in \mathcal{C}(\mathcal{X}_n)) \geq 1 - \alpha.$$

Remarque: Sous **R**, la commande `confit()` retourne un intervalle de confiance à 95%.

### 3.4. *p-value*

On résume souvent un test à sa *p-value*. (Quel niveau de test faudrait-il pour que l'intervalle de confiance contienne la valeur d'intérêt?)

#### *p-value*

Pour un échantillon  $\mathcal{X}_n$  fixé et un test donné, la *p-value* du test est le seuil  $\alpha^*$  tel que pour tout

- $\alpha > p\text{-value}$   $\alpha^*$  on rejette  $\mathcal{H}_0$ .
- $\alpha < p\text{-value}$   $\alpha^*$  on ne rejette pas  $\mathcal{H}_0$ .

Interprétation: La *p-value* peut être comprise comme l'erreur de première espèce minimale que l'on est prête à faire pour rejeter  $\mathcal{H}_0$ .

Les différents logiciels donnent la ***p*-value** du test demandé qui permet de conclure sur le résultat d'un test statistique.

En comparant à seuil  $s$  de référence choisi (traditionnellement  $s = 5\%$ )

- Si la ***p*-value**  $< s$  (il faut un risque de première espèce petit pour rejeter  $\mathcal{H}_0$ ), on rejette l'hypothèse nulle en faveur de l'hypothèse alternative. Le test est déclaré "statistiquement significatif".
- Si la ***p*-value**  $> s$  (il faut un risque de première espèce très grand pour rejeter  $\mathcal{H}_0$ ), on ne rejette pas l'hypothèse nulle, et on ne peut rien conclure quant aux hypothèses formulées.

# Exemple du test de Shapiro

Tirer au hasard 100 valeurs selon une loi normale  $\mathcal{N}(0,1)$ .

```
set.seed(2021)
X=rnorm(100,0,1)
```

Test de Shapiro: Commande sous R `shapiro.test()`

$$\mathcal{H}_0: x \sim \mathcal{N}(0,1) \text{ v.s. } \mathcal{H}_1: \overline{\mathcal{H}_0}.$$

```
shapiro.test(X)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  X
## W = 0.98262, p-value = 0.2116
```

⇒ **grande  $p$ -value, on ne rejette pas  $\mathcal{H}_0$ , l'échantillon  $X$  est issu d'une  $\mathcal{N}(0,1)$ .**

Ouvrir Rstudio et ouvrir le fichier Participant1.Rmd

Question 1: Tirer au hasard un échantillon  $Y$  de 100 valeurs selon une loi uniforme  $\mathcal{U}(0, 1)$  (commande `runif()`). (Mettre une seed)



Question 1: Tirer au hasard un échantillon  $Y$  de 100 valeurs selon une loi uniforme  $\mathcal{U}(0, 1)$  (commande `runif()`). (Mettre une seed)

```
set.seed(2021)  
Y=runif(100,0,1)
```

Question 2:  $Y$  est-il issu d'une loi normale  $\mathcal{N}(0,1)$ ? (Test de Shapiro (commande `shapiro.test()`))

Question 2:  $Y$  est-il issu d'une loi normale  $\mathcal{N}(0,1)$ ? (Test de Shapiro (commande `shapiro.test()`))

```
shapiro.test(Y)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  Y
## W = 0.93204, p-value = 6.513e-05
```

⇒ **petite  $p$ -value, on rejette  $\mathcal{H}_0$ , l'échantillon  $Y$  n'est pas issu d'une  $\mathcal{N}(0,1)$ .**