



# Statistique Inférentielle

N. Jégou

Université Rennes 2

Master 1 Mathématiques Appliquées, Statistiques

# Plan du cours

- Introduction
- Modèle Statistique
- Estimateurs - Propriétés
- Construction d'estimateurs
- Estimation par intervalles

## Bibliographie

- Pagès J., Statistique générale pour utilisateurs :  
1) Méthodologie, PUR (2010)
- Husson F. et Pagès J., Statistique générale pour utilisateurs :  
2) Exercices et corrigés, PUR (2013)
- Saporta G., Probabilités, analyse des données et statistique  
Editions TECHNIP (2011)
- Wonnacott H. et Wonnacott J., Statistique :  
économie-gestion-sciences-médecine, Economica (1999)
- Monfort A., Cours de statistique mathématique, Economica  
(1982)



## Exemple 1

- On souhaite tester l'efficacité d'un médicament  
 $n = 100$  patients atteints prennent le médicament  
A l'issue de l'étude, 72 patients sont guéris  
Quelle est la probabilité  $p$  de guérison suite au traitement ?



## Exemple 1

- On souhaite tester l'efficacité d'un médicament  
 $n = 100$  patients atteints prennent le médicament  
A l'issue de l'étude, 72 patients sont guéris  
Quelle est la probabilité  $p$  de guérison suite au traitement ?
- On est tenté de considérer  $p \approx 0.72$



## Exemple 1

- On souhaite tester l'efficacité d'un médicament  
 $n = 100$  patients atteints prennent le médicament  
 A l'issue de l'étude, 72 patients sont guéris  
 Quelle est la probabilité  $p$  de guérison suite au traitement ?
- On est tenté de considérer  $p \approx 0.72$
- Questions :  
 Quel crédit donner à cette proposition ?  
 Cette idée est-elle cohérente avec une modélisation mathématique ?  
 Le niveau de confiance est faible ? Fort ?



## Exemple 2

- Des biologistes étudient le développement de poissons  
Des poissons qui se développent correctement pèsent en moyenne 1 kg  
Ils prélèvent  $n = 20$  : leur poids moyen est 949.5 gr



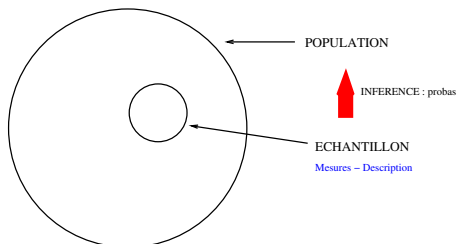
## Exemple 2

- Des biologistes étudient le développement de poissons  
Des poissons qui se développent correctement pèsent en moyenne 1 kg  
Ils prélèvent  $n = 20$  : leur poids moyen est 949.5 gr
- Questions :  
Faut-il en déduire que les poissons ne se développent pas correctement ?  
Cette valeur est-elle conforme à un développement normal ?



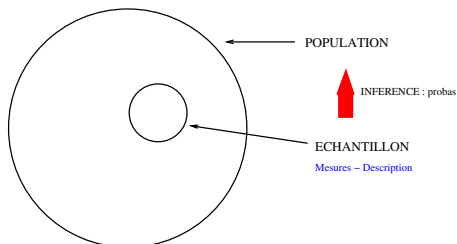
## Inférence vs descriptive

- Les données de l'**échantillon** ne nous intéressent pas en tant que telles
- Les résumer, les représenter est le domaine de la statistique descriptive



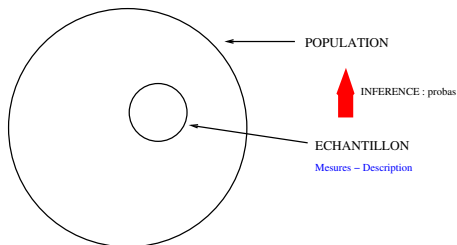
## Inférence vs descriptive

- Elles nous intéressent car elles donnent une information sur une ensemble plus vaste dont elles proviennent : la **population**
- L'opération de "remontée" de l'échantillon à la population est appelée **inférence statistique**



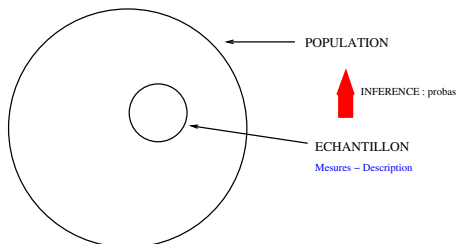
## Principe de base de l'inférence

- Si l'on prélève un nouveau jeu de données, les nouvelles observations seront différentes des précédentes
- L'inférence statistique suppose de prendre en compte l'aspect aléatoire des données



## Principe de base de l'inférence

- L'idée de base est ainsi de considérer ces observations comme issues d'un phénomène aléatoire
- L'inférence statistique s'appuie donc sur des outils probabilistes





# Echantillonnage

- La façon de recueillir ces données a une grande importance dans la pratique
- L'objet n'est pas ici de développer la stratégie selon laquelle l'échantillon a été prélevé (le plan de sondage) : ceci relève de la théorie des sondages

# Echantillonnage

- Le principe de base que nous retenons est que chaque individu constitutif de la population doit avoir la même chance de figurer dans l'échantillon
- L'échantillon doit ainsi être prélevé au hasard ; nous considèrerons le cas standard où **les tirages sont supposés indépendants** :
  - la population est de taille infinie ou bien
  - le tirage se fait avec remise

## Notations

- On considère  $n$  variables aléatoires  $X_1, \dots, X_n$
- $X_1, \dots, X_n$  sont des répliques i.i.d. d'une même variable  $X$  de loi inconnue
- Les données dont on dispose sont des réalisations de ces variables ; elles sont notées  $x_1, \dots, x_n$



## Notations

- On considère  $n$  variables aléatoires  $X_1, \dots, X_n$
- $X_1, \dots, X_n$  sont des répliques i.i.d. d'une même variable  $X$  de loi inconnue
- Les données dont on dispose sont des réalisations de ces variables ; elles sont notées  $x_1, \dots, x_n$
- **Attention !**
  - $X_i$  est une variable aléatoire
  - $x_i$  est un nombre



## Modèle statistique - Définition

- Un modèle statistique est un objet mathématique associé à l'observation de données aléatoires
- On considère d'abord l'expérience aléatoire qui consiste à recueillir **une** observation  $x$  de la variable  $X$
- $X$  est supposée être à valeurs dans un espace  $\mathcal{X}$
- On ne connaît pas la loi de probabilité  $\mathbb{P}$  de  $X$

## Modèle statistique - Définition

Un principe de la modélisation est de **supposer** que la loi de probabilité  $\mathbb{P}$  appartient à une famille  $\mathcal{P}$  de lois de probabilités possibles, d'où la définition suivante :

### Définition (Modèle statistique)

*On appelle modèle statistique tout triplet  $(\mathcal{X}, \mathcal{A}, \mathcal{P})$  où*

- *$\mathcal{X}$  est l'espace des observations, c'est-à-dire l'ensemble de tous les résultats possibles de l'expérience*
- *$\mathcal{A}$  est une tribu sur  $\mathcal{X}$*
- *$\mathcal{P}$  est une famille de probabilités sur  $(\mathcal{X}, \mathcal{A})$*

## Exemples

- La définition d'un modèle statistique repose donc sur une hypothèse concernant la famille d'appartenance de la loi de  $X$
- Cet aspect doit être gardé en mémoire : les résultats que l'on obtient ensuite ne valent que sous cette hypothèse

**Exemple 1** Hypothèse :  $X \sim \mathcal{B}(p)$  d'où le modèle associé à une observation de  $X$

$$\mathcal{X} = \{0, 1\} \quad \mathcal{A} = \mathcal{P}(\{0, 1\}) \quad \mathcal{P} = \{\mathcal{B}(p), p \in ]0, 1[ \}$$

**Exemple 2** Hypothèse :  $X \sim \mathcal{N}(\mu, \sigma^2)$  d'où le modèle associé à une observation de  $X$

$$\mathcal{X} = \mathbb{R} \quad \mathcal{A} = \mathcal{B}(\mathbb{R}) \quad \mathcal{P} = \{\mathcal{N}(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+ \}$$

## Modèle discret - Modèle continu

- Le modèle est dit **discret** lorsque  $\mathcal{X}$  est fini ou dénombrable  
Alors  $\mathcal{A}$  est la tribu formée par l'ensemble des parties de  $\mathcal{X}$  :  
 $\mathcal{A} = P(\mathcal{X})$
- Le modèle est dit **continu** lorsque  $\mathcal{X} \subset \mathbb{R}^p$  et que  $\forall P \in \mathcal{P}$ ,  $\mathbb{P}$  admet une densité dans  $\mathbb{R}^p$   
Dans ce cas,  $\mathcal{A}$  est la tribu des boréliens de  $\mathcal{X}$  :  $\mathcal{A} = \mathcal{B}(\mathcal{X})$

Dans l'exemple 1, le modèle est discret

Dans l'exemple 2, le modèle est continu

# Echantillon

Avant d'étendre la définition du modèle à  $n$  observations, on précise la notion d'échantillon.

On considère des variables i.i.d. d'où la définition que l'on prend pour un échantillon :

## Définition (Echantillon)

*Un échantillon de taille  $n$  (ou  $n$ -échantillon) est une suite  $X_1, \dots, X_n$  de  $n$  variables aléatoires indépendantes, de même loi  $\mathbb{P}$*



## Modèle produit

- Le  $n$ -échantillon définit un vecteur aléatoire  $(X_1, \dots, X_n)'$  de loi  $\mathbb{P}^{\otimes n}$
- Avec comme modèle pour une observation  $\mathcal{M} = (\mathcal{X}, \mathcal{A}, \mathcal{P})$ , le modèle associé à un  $n$ -échantillon est le **modèle produit** :

$$\mathcal{M}_n = (\mathcal{X}^n, \mathcal{A}_n, \{\mathbb{P}^{\otimes n}\})$$

avec  $\mathcal{A}_n$  une tribu sur  $\mathcal{X}^n$



## Exemples

Ainsi dans nos exemples :

	$\mathcal{X}$	$\mathcal{A}$	$\mathcal{P}$
Exemple 1	$\{0, 1\}^n$	$P(\{0, 1\}^n)$	$\{\mathcal{B}(p)^{\otimes n}, p \in ]0, 1[ \}$
Exemple 2	$\mathbb{R}^n$	$\mathcal{B}(\mathbb{R}^n)$	$\{\mathcal{N}(\mu, \sigma^2)^{\otimes n}, \mu \in \mathbb{R}, \sigma \in \mathbb{R}^+ \}$



## Modèle paramétrique - Modèle non paramétrique

Il s'agit de préciser l'hypothèse faite sur la famille d'appartenance de la loi de  $X$  :

### Définition (Modèle paramétrique - Modèle non paramétrique )

- *Si la loi de  $X$  appartient à une famille de lois indexables par un nombre fini de paramètres, le modèle est dit paramétrique. On note alors  $\mathcal{P} = \{\mathbb{P}_\theta, \theta \in \Theta\}$  où  $\Theta \in \mathbb{R}^d$  est l'espace des paramètres*
- *Si la famille d'appartenance de la loi de  $X$  n'est pas indexable par un nombre fini de paramètres, on parle alors de modèle non paramétrique*





## Paramétrique vs Non paramétrique

- Exemples 1 et 2 : modèle paramétrique
- Exemple d'hypothèse non paramétrique : la loi de  $X$  appartient à la famille des lois continues
- Avantage : on réduit le risque de mauvaise spécification du modèle
- Inconvénient : techniques d'inférence plus difficiles
- Possibilité de tester l'appartenance à une famille paramétrique

## Estimateur

- Cadre du cours : Modèle paramétrique
- $\Rightarrow$  inférence sur ce(s) paramètre(s) caractéristique(s) de la loi : estimation ponctuelle, estimation par intervalles, tests...
- Pour cela, on introduit la notion d'estimateur :

### Définition (Estimateur)

*Un estimateur de  $\theta$  est une fonction mesurable de  $(X_1, \dots, X_n)$ , indépendante de  $\theta$ , à valeurs dans un sur-ensemble de  $\Theta$*

## Estimateur

- Un estimateur est une variable aléatoire fonction des  $X_i$  :

$$\hat{\theta} = f(X_1, X_2, \dots, X_n)$$

- Par exemple :

$$X_1 \quad \inf_{i=1\dots n} \{X_i\} \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

- Exemple 1** : Le nombre moyen de guérisons est un estimateur “naturel” de la probabilité  $p$  :

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



## Estimateur

- Un estimateur est une variable aléatoire fonction des  $X_i$  :

$$\hat{\theta} = f(X_1, X_2, \dots, X_n)$$

- Par exemple :

$$X_1 \quad \inf_{i=1\dots n} \{X_i\} \quad \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

- Exemple 2** : Le poids moyen dans l'échantillon est un estimateur "naturel" du poids moyen  $\mu$  dans le lac :

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

## Big Picture

- On ne dispose que de la valeur de l'estimateur prise en les observations :  $\hat{\theta} = f(x_1, \dots, x_n)$
- **Exemple 1** :  $\hat{\rho} = 0.72$       **Exemple 2** :  $\hat{\mu} = 0.9495$  gr
- On souhaite que cette estimation soit proche du paramètre inconnu

---

<sup>1</sup>Les vitesses ne sont pas abordées ici

## Big Picture

- On ne dispose que de la valeur de l'estimateur prise en les observations :  $\hat{\theta} = f(x_1, \dots, x_n)$
- **Exemple 1** :  $\hat{p} = 0.72$       **Exemple 2** :  $\hat{\mu} = 0.9495$  gr
- On souhaite que cette estimation soit proche du paramètre inconnu
- $\Rightarrow$  Quelle confiance avoir en cette estimation ?

---

<sup>1</sup>Les vitesses ne sont pas abordées ici

## Big Picture

- On ne dispose que de la valeur de l'estimateur prise en les observations :  $\hat{\theta} = f(x_1, \dots, x_n)$
- **Exemple 1** :  $\hat{p} = 0.72$       **Exemple 2** :  $\hat{\mu} = 0.9495$  gr
- On souhaite que cette estimation soit proche du paramètre inconnu
- $\Rightarrow$  Quelle confiance avoir en cette estimation ?
- Pour le savoir, on étudie les propriétés théoriques de l'estimateur
  - Propriétés asymptotiques ( $n \rightarrow \infty$ ) : convergence, vitesse<sup>1</sup>
  - Propriétés à  $n$  fixé : biais, variance, risque quadratique

<sup>1</sup>Les vitesses ne sont pas abordées ici



## Propriétés de convergence : rappels

Principales formes de convergence pour une suite de variables aléatoires :

- Convergence en loi

la suite de variables aléatoires  $\{X_n\}$  converge en loi vers la variable aléatoire  $X$  si, pour tout réel  $x$  où la fonction de répartition  $F$  de  $X$  est continue, on a

$$\lim_{n \rightarrow \infty} F_n(x) = F(x)$$

On note alors

$$X_n \xrightarrow[\mathcal{L}]{} X$$



## Propriétés de convergence : rappels

Principales formes de convergence pour une suite de variables aléatoires :

- Convergence en loi
- Convergence en probabilité

la suite de variables aléatoires  $\{X_n\}$  converge en probabilité vers la variable aléatoire  $X$  si pour tout  $\varepsilon > 0$ ,

$$\lim_{n \rightarrow +\infty} \mathbb{P}(|X_n - X| \geq \varepsilon) = 0$$

On note alors

$$X_n \xrightarrow{\mathcal{P}} X$$



## Propriétés de convergence : rappels

Principales formes de convergence pour une suite de variables aléatoires :

- Convergence en loi
- Convergence en probabilité
- Convergence presque sûre

la suite de variables aléatoires  $\{X_n\}$  converge presque sûrement vers la variable aléatoire  $X$  si

$$\mathbb{P}\left(\lim_{n \rightarrow +\infty} X_n = X\right) = 1$$

On note alors

$$X_n \xrightarrow[ps]{} X$$

## Propriétés de convergence : rappels

Principales formes de convergence pour une suite de variables aléatoires :

- Convergence en loi
- Convergence en probabilité
- Convergence presque sûre

Conv. presque sûre  $\Rightarrow$  Conv. en probabilité  $\Rightarrow$  Conv. en loi



## Consistance

La consistance d'un estimateur paramétrique spécifie la convergence en probabilité vers le paramètre :

### Définition (Consistance)

On dit que  $\hat{\theta}$  est consistant (ou convergent) si  $\hat{\theta} \xrightarrow[\mathcal{P}]{} \theta$  c'est-à-dire si

$$\forall \theta \in \Theta, \forall \varepsilon > 0, \lim_{n \rightarrow \infty} \mathbb{P}(\|\hat{\theta} - \theta\| > \varepsilon) = 0$$

avec  $\|\cdot\|$  une norme sur l'espace des paramètres

Dans le cas de l'estimation d'un paramètre unidimensionnel, on prendra la valeur absolue

## Loi(s) des grands nombres

- La loi des grands nombres justifie l'intérêt, en terme de convergence, de faire des moyennes
- Elles stipulent la convergence d'une moyenne de variable aléatoire vers l'espérance commune
- La **loi faible** est un résultat de convergence en probabilité
- La **loi forte** assure, moyennant des hypothèses plus fortes, la convergence presque sûre

## Loi(s) des grands nombres

### Théorème (Loi des grands nombres)

- **Loi faible** : Soit  $(X_n)$  une suite de variables aléatoires indépendantes et de même espérance  $\mathbb{E}[X]$ , on a convergence en probabilité de  $(\frac{1}{n} \sum_{k=1}^n X_k)$  vers  $\mathbb{E}[X]$  :

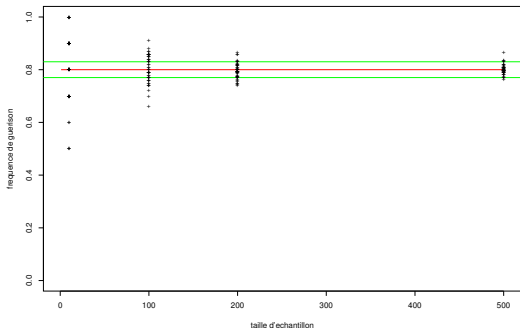
$$\forall \varepsilon > 0, \quad \lim_{n \rightarrow +\infty} \mathbb{P} \left( \left| \frac{X_1 + \dots + X_n}{n} - \mathbb{E}[X] \right| \geq \varepsilon \right) = 0$$

- **Loi forte** : Soit  $(X_n)$  une suite de variables aléatoires indépendantes, intégrables et de même loi, on a convergence presque sûre de  $(\frac{1}{n} \sum_{k=1}^n X_k)$  vers  $\mathbb{E}[X]$  :

$$\mathbb{P} \left( \lim_{n \rightarrow +\infty} \frac{X_1 + \dots + X_n}{n} = \mathbb{E}[X] \right) = 1$$

## Loi(s) des grands nombres

- Illustration sur l'Exemple 1 avec  $p = 0.8$  ;  $K = 50$  simulations pour  $n = 10, 100, 200, 500$
- Quand  $n$  augmente la probabilité que  $\bar{X}$  sorte du couloir  $p \pm \varepsilon$  tend à se réduire



## Théorème central limite

Le théorème central limite précise le comportement asymptotique de la moyenne d'échantillon puisqu'il en donne la loi limite :

### Théorème (Théorème Central Limite)

Soit  $(X_n)$  une suite de variables aléatoires i.i.d. selon une loi commune  $X$  d'espérance  $\mathbb{E}[X] = \mu$  et de variance  $V(X) = \sigma^2$ . On a alors :

$$\frac{1}{\sqrt{n}} \left( \frac{X_1 + \dots + X_n - n\mu}{\sigma} \right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, 1)$$



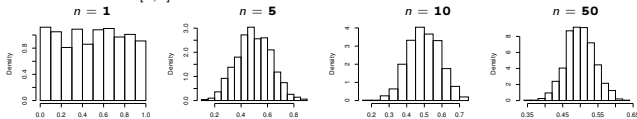
## Théorème central limite

- Comme  $\frac{1}{\sqrt{n}} \left( \frac{X_1 + \dots + X_n - n\mu}{\sigma} \right) = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ , le sens concret de ce théorème est que  $\bar{X}$  suit approximativement (pour  $n$  assez grand) une loi  $\mathcal{N}(\mu, \sigma^2/n)$
- L'aspect remarquable est que cela est vrai quelque soit la loi de  $X$  (pour peu qu'elle admette une variance)
- Pour des  $X_i \underset{\text{i.i.d}}{\sim} \mathcal{N}(\mu, \sigma^2)$  on a  $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$
- Dans ce dernier cas, il n'est plus question d'approximation

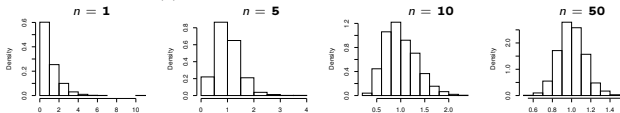


# Théorème central limite

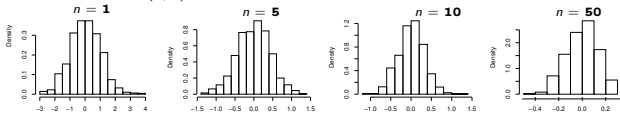
Loi de  $X$  :  $\mathcal{U}_{[0,1]}$



Loi de  $X$  :  $\exp(1)$



Loi de  $X$  :  $\mathcal{N}(0, 1)$





## Retour aux exemples

### Exemple 1

- Hypothèse :  $X_i \underset{\text{i.i.d}}{\sim} \mathcal{B}(p)$
- LGN  $\Rightarrow \hat{p}$  estimateur consistant de  $p$  :

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{\mathcal{P}} p$$

- $n\hat{p} = \sum_{i=1}^n X_i \sim \mathcal{B}(n, p)$
- Si  $n$  est jugé assez grand, TCL  $\Rightarrow$

$$\mathcal{L}(\hat{p}) \approx \mathcal{N}(p, p(1-p)/n)$$

## Retour aux exemples

### Exemple 2

- Hypothèse :  $X_i \underset{\text{i.i.d}}{\sim} \mathcal{N}(\mu, \sigma^2)$  (★)
- LGN  $\Rightarrow \bar{X}$  estimateur consistant de  $\mu$  :

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{\mathcal{P}} \mu$$

- Sous (★)  $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$
- Si (★) n'est pas vérifiée, si  $n$  est jugé assez grand, TCL  $\Rightarrow$

$$\mathcal{L}(\bar{X}) \approx \mathcal{N}(\mu, \sigma^2/n)$$

## Propriétés à $n$ fixé

- Propriétés de convergence :  $n \rightarrow \infty$
- En pratique,  $n$  est fixé

=> Nécessité d'étudier les propriétés d'un estimateur pour  $n$  fixe

- Nous présentons
  - Biais
  - Variance
  - Risque quadratique

## Biais

### Définition (biais)

Soit  $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$  un estimateur. Son espérance sous la loi  $\mathbb{P}_\theta$  est

$$\mathbb{E}_\theta[\hat{\theta}] = \int_{\mathcal{X}^n} \hat{\theta}(x) \mathbb{P}_\theta(x) dx$$

où  $x = (x_1, \dots, x_n)$

1. Le biais de  $\hat{\theta}$  en  $\theta$  est  $\mathbf{Biais}(\hat{\theta}) = \mathbb{E}_\theta[\hat{\theta}] - \theta$
2.  $\hat{\theta}$  est sans biais si pour chaque  $\theta \in \Theta$ ,  $\mathbf{Biais}(\hat{\theta}) = 0$
3.  $\hat{\theta}$  est asymptotiquement sans biais si pour chaque  $\theta \in \Theta$ ,  $\lim_{n \rightarrow \infty} \mathbf{Biais}(\hat{\theta}) = 0$



## Biais

- Un estimateur est sans biais si, en moyenne (i.e. sur tous les  $n$ -échantillons), il tombe sur le paramètre
- ... mais en pratique, on dispose d'un seul échantillon...
- Le fait que l'estimateur soit sans biais est simplement une garantie (théorique) sur le comportement en moyenne de l'estimateur

## Biais

- La moyenne d'échantillon estime sans biais l'espérance de la loi commune

$$\mathbb{E}[X_i] = \mu \Rightarrow \mathbb{E}[\bar{X}] = \mu \text{ (Linéarité de l'espérance)}$$

- Exemple 1 :  $\mathbb{E}[\hat{p}] = p$
- Exemple 2 :  $\mathbb{E}[\bar{X}] = \mu$
- Mais par exemple, sont aussi sans biais

$$X_1, \quad \frac{X_1 + X_2}{2}, \quad X_1 + \frac{X_2 + X_3}{2}, \dots$$





## Variance

La variance d'un estimateur mesure la variabilité de cet estimateur, autour de son espérance :

### Définition (Variance)

Soit  $\hat{\theta} = \hat{\theta}(X_1, \dots, X_n)$  un estimateur et  $\mathbb{E}_\theta[\hat{\theta}]$  son espérance sous la loi  $\mathbb{P}_\theta$ . Sa variance est

$$V(\hat{\theta}) = \mathbb{E}_\theta[\|\hat{\theta} - \mathbb{E}_\theta[\hat{\theta}]\|^2] = \int_{\mathcal{X}^n} \|\hat{\theta}(x) - \mathbb{E}_\theta[\hat{\theta}(x)]\|^2 \mathbb{P}_\theta(x) dx$$

où  $\|\cdot\|$  est une norme sur  $\Theta$

## Variance

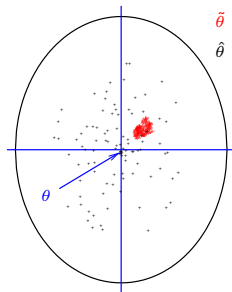
- Si  $V(\hat{\theta})$  faible, les valeurs de  $\hat{\theta}$  sont proches les unes des autres
- Avec  $V(\hat{\theta})$  faible, on est garanti que l'observation de  $\hat{\theta}$  dont on dispose (sur l'échantillon) est proche de celle que l'on aurait avec d'autres échantillons
- La variance de la moyenne d'échantillon décroît avec  $n$  :

$$V(X_i) = \sigma^2 \Rightarrow V(\bar{X}) = \frac{\sigma^2}{n} \text{ (avec des } X_i \text{ décorrélés)}$$



## Risque quadratique

- $\hat{\theta}$  estimateur sans biais de  $\theta$  mais de grande variance
- $\tilde{\theta}$  estimateur biaisé de  $\theta$  mais de petite variance



- Quel est le meilleur choix ?

## Risque quadratique

- On souhaite que les valeurs de l'estimateurs soient aussi proches possible de  $\theta$
- On souhaite donc que  $\|\hat{\theta} - \theta\|$  soit petit
- $\|\hat{\theta} - \theta\|$  est aléatoire
- On définit le **risque quadratique** (ou erreur quadratique moyenne) comme l'espérance de cette variable aléatoire :

$$\mathbb{E}_{\theta}[\|\hat{\theta} - \theta\|^2]$$

## Risque quadratique

### Définition (Risque quadratique - Décomposition biais-variance)

- Soit  $\hat{\theta}$  un estimateur d'ordre 2, le risque quadratique de  $\hat{\theta}$  sous  $\mathbb{P}_\theta$  est :

$$\mathcal{R}(\theta, \hat{\theta}) = \mathbb{E}_\theta[\|\hat{\theta} - \theta\|^2]$$

- $\mathcal{R}(\theta, \hat{\theta})$  est la somme d'un terme de biais et d'un terme de variance :

$$\mathcal{R}(\theta, \hat{\theta}) = \|\mathbb{E}_\theta[\hat{\theta}] - \theta\|^2 + \mathbb{E}_\theta[\|\hat{\theta} - \mathbb{E}_\theta(\hat{\theta})\|^2]$$

- Pour  $\theta \in \mathbb{R}$ , cette décomposition s'écrit

$$\mathcal{R}(\theta, \hat{\theta}) = \mathbf{Biais}^2(\hat{\theta}) + V(\hat{\theta})$$

## Estimateur préférable - de variance minimum

### Définition (Estimateur préférable - de variance minimum)

Soit  $\hat{\theta}$  et  $\hat{\theta}'$  deux estimateurs d'ordre 2

- On dit que  $\hat{\theta}$  est préférable à  $\hat{\theta}'$  si

$$\forall \theta \in \Theta, \mathcal{R}(\theta, \hat{\theta}) \leq \mathcal{R}(\theta, \hat{\theta}')$$

- Si  $\hat{\theta}$  est sans biais, on dit qu'il est de variance uniformément minimum parmi les estimateurs sans biais s'il est préférable à tout autre estimateur sans biais d'ordre 2

## Estimateur préférable - de variance minimum

- **Exemple 1** :  $X_i \underset{i.i.d.}{\sim} \mathcal{B}(p)$  et  $\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i$ 
  - $\hat{p}$  sans biais donc

$$\mathcal{R}(p, \hat{p}) = V(\hat{p}) = \frac{p(1-p)}{n}$$

- $\mathbb{E}[X_1] = p$  donc  $X_1$  est aussi sans biais et

$$\mathcal{R}(p, X_1) = V(X_1) = p(1-p)$$

- $\hat{p}$  préférable à  $X_1$
- **Exemple 2** : idem
- Pour déterminer les estimateurs sans biais de variance uniformément minimum : exhaustivité, information de Fisher<sup>2</sup>

---

<sup>2</sup>Notions non abordées ici

## Méthode des moments

- LGN : convergence de la moyenne d'échantillon vers l'espérance

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{\mathcal{P}} \mathbb{E}[X]$$

⇒ Si  $n$  assez grand, on a bon espoir que  $\bar{X} \approx \mathbb{E}[X]$

- De même, si  $X$  tel que  $m_k = \mathbb{E}[X^k]$  existe,

$$\text{LGN} \Rightarrow \frac{1}{n} \sum_{i=1}^n X_i^k \xrightarrow{\mathcal{P}} \mathbb{E}[X^k]$$

⇒ Si  $n$  assez grand, on a bon espoir que  $\frac{1}{n} \sum_{i=1}^n X_i^k \approx \mathbb{E}[X^k]$





## Méthode des moments

- **Idée :**
  - Exprimer  $\theta$  comme fonction des moments  $m_k$
  - Remplacer les  $m_k$  par les moments empiriques  $\frac{1}{n} \sum_{i=1}^n X_i^k$
- **Exemple 1 :**  $X \sim \mathcal{B}(p) \Rightarrow p = m_1 = \mathbb{E}[X]$   
 Estimateur des moments :  $\hat{p} = \frac{1}{n} \sum_{i=1}^n X_i$
- **Exemple 2 :**  $X \sim \mathcal{N}(\mu, \sigma^2)$  d'où

$$\begin{cases} \mu &= \mathbb{E}[X] &= m_1 \\ \sigma^2 &= \mathbb{E}[X^2] - \mathbb{E}[X]^2 &= m_2 - m_1^2 \end{cases}$$

- Estimateur des moments :

$$\hat{\mu} = \bar{X} \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2$$

## Méthode des moments

- L'estimateur des moments est défini comme la solution en  $\theta$  du système à  $p$  équations

$$\begin{cases} m_1(\theta) = \hat{m}_1 \\ \cdot \\ \cdot \\ \cdot \\ m_p(\theta) = \hat{m}_p \end{cases}$$

- Si l'application

$$M : \theta \mapsto (m_1(\theta), \dots, m_p(\theta))$$

est une bijection, alors l'estimateur des moments existe et est unique



## Maximum de vraisemblance : Exemple 1

- Soit  $(X_1 = x_1, \dots, X_n = x_n)$  les observations
- La probabilité d'observer ces données s'écrit

$$\begin{aligned}
 L(x_1, \dots, x_n; p) &= \mathbb{P}_p(X_1 = x_1, \dots, X_n = x_n) \\
 &= \prod_{i=1}^n \mathbb{P}_p(X_i = x_i) \\
 &= \prod_{i=1}^n p^{x_i} (1 - p)^{1-x_i} \\
 &= p^{\sum_{i=1}^n x_i} (1 - p)^{n - \sum_{i=1}^n x_i}
 \end{aligned}$$

## Maximum de vraisemblance : Exemple 1

- Cette probabilité est inconnue car fonction de  $p$  inconnue
- Sur les données  $\sum_{i=1}^n = 72$  donc elle s'écrit

$$L(x_1, \dots, x_n; p) = p^{72}(1 - p)^{28}$$

- Ainsi, si  $p = 0.5$ , la probabilité d'observer nos données est

$$L(x_1, \dots, x_n, p = 0.5) = 0.5^{72} \times 0.5^{28} \approx 8 \times 10^{-31}$$

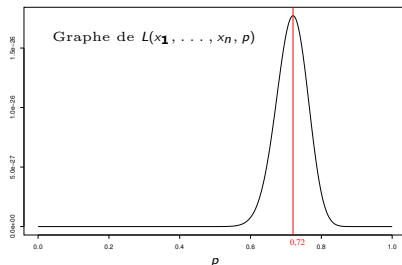
- et si  $p = 0.7$ , la probabilité d'observer nos données est

$$L(x_1, \dots, x_n, p = 0.7) = 0.7^{72} \times 0.3^{28} \approx 1.6 \times 10^{-26}$$



## Maximum de vraisemblance : Exemple 1

- Quelle valeur de  $p$  maximise la probabilité d'observer les données ?
- Représentons  $p \mapsto L(x_1, \dots, x_n; p) = p^{72}(1-p)^{28}$



- La probabilité d'observer nos données est maximum pour  $p = 0.72$

## Maximum de vraisemblance : Exemple 1

- $p = 0.72$  est la valeur du paramètre qui rend maximum la probabilité d'observer nos données :  $\Rightarrow \hat{p} = 0.72$  estimation naturelle de  $p$
- Plus généralement on montre que

$$p \mapsto L(x_1, \dots, x_n; p) = p^{\sum x_i} (1 - p)^{n - \sum x_i}$$

est maximum pour  $p = \frac{1}{n} \sum_{i=1}^n x_i$

- L'estimateur du maximum de vraisemblance pour  $p$  est ainsi

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

## Maximum de vraisemblance

### Définition (Vraisemblance)

- *Cas discret* : La vraisemblance du paramètre  $\theta$  pour la réalisation  $(x_1, \dots, x_n)$  est l'application  $L : \mathcal{X}^n \times \Theta$  définie par

$$L(x_1, \dots, x_n; \theta) = \mathbb{P}_\theta^{\otimes n}(\{x_1, \dots, x_n\}) = \prod_{i=1}^n \mathbb{P}_\theta(\{x_i\})$$

- *Cas absolument continu* : Soit  $f(\cdot, \theta)$  la densité associée à  $\mathbb{P}_\theta$ . La vraisemblance du paramètre  $\theta$  pour la réalisation  $(x_1, \dots, x_n)$  est l'application  $L : \mathcal{X}^n \times \Theta$  définie par

$$L(x_1, \dots, x_n; \theta) = \prod_{i=1}^n f(x_i, \theta)$$

## Maximum de vraisemblance

### Définition (Estimateur du maximum de vraisemblance)

*Un estimateur du maximum de vraisemblance est une statistique  $g$  qui maximise la vraisemblance, c'est-à-dire telle que*

$$\forall (x_1, \dots, x_n) \in \mathcal{X}^n : L(x_1, \dots, x_n; g(x_1, \dots, x_n)) = \sup_{\theta \in \Theta} L(x_1, \dots, x_n; \theta)$$

*L'estimateur du maximum de vraisemblance s'écrit donc sous la forme  $\hat{\theta} = g(X_1, \dots, X_n)$*



# Maximum de vraisemblance

## Propriétés

- Invariance

Soit  $\Psi : \Theta \rightarrow \mathbb{R}^k$  et  $\hat{\theta}$  l'estimateur du maximum de vraisemblance de  $\theta$ , alors  $\Psi(\hat{\theta})$  est l'estimateur du maximum de vraisemblance de  $\Psi(\theta)$

- Consistance

On suppose que  $\mathbb{P}_\theta$  admet une densité  $f(x, \theta)$ , que  $\Theta$  est un ouvert et que  $\theta \mapsto f(x, \theta)$  est différentiable, alors l'estimateur du maximum de vraisemblance  $\hat{\theta}$  est consistant



## Confiance ou pas ?

**Exemple 2** :  $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n x_i = 949.5$  avec  $n = 20$

Faut-il avoir confiance en cette estimation ?

- $\hat{\mu}$  a de bonnes propriétés : convergence, absence de biais
- $V(\hat{\mu}) = \sigma^2/n$  :  $\hat{\mu}$  a une petite variance si  $n$  grand<sup>3</sup>

**Mais**

- Il faut quantifier cette confiance en fonction de  $n$
- La confiance est d'autant plus grande que les  $x_i$  sont proches les uns des autres : il faut estimer  $\sigma^2$

⇒ L'intervalle de confiance répond à ces questions

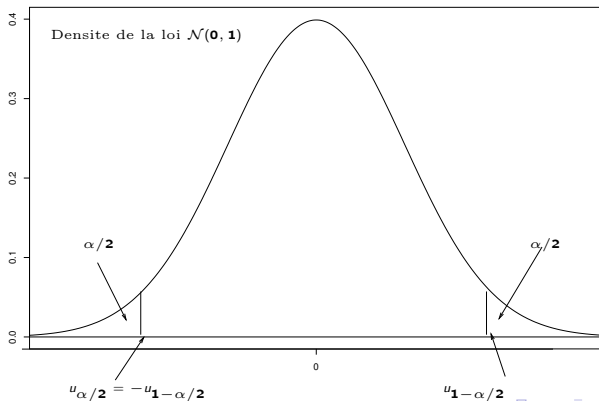
---

<sup>3</sup> $\hat{\mu}$  est de variance uniformément minimum parmi les estimateurs sans biais

## Le principe dans le cas de base

- Exemple 2 :  $X_i \underset{\text{i.i.d}}{\sim} \mathcal{N}(\mu, \sigma^2) \Rightarrow \bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$  et

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1)$$



## Le principe dans le cas de base

- Avec  $u_\alpha$  le quantile d'ordre  $\alpha$  de la loi  $\mathcal{N}(0, 1)$ , on a

$$\mathbb{P} \left( -u_{1-\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < u_{1-\alpha/2} \right) = 1 - \alpha$$

- Soit

$$\mathbb{P} \left( \bar{X} - u_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + u_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha$$

- Ce qu'on peut noter

$$\mathbb{P} \left( \mu \in \bar{X} \pm u_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}} \right) = 1 - \alpha$$

## Le principe dans le cas de base

- L'intervalle  $\left[ \bar{X} - u_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}; \bar{X} + u_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}} \right]$  a la probabilité  $1 - \alpha$  de contenir  $\mu$
- C'est une intervalle dont les bornes sont aléatoires<sup>4</sup>
- On obtient l'intervalle de confiance en remplaçant dans  $\bar{X}$  par son observation  $\bar{x}$  sur l'échantillon :

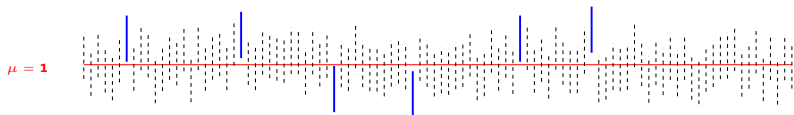
$$\begin{aligned} \text{IC} &= \left[ \bar{x} - u_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}; \bar{x} + u_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}} \right] \\ &= \bar{x} \pm u_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}} \end{aligned}$$

<sup>4</sup>Parfois appelé intervalle de probabilité



## Interprétation

- On simule  $K = 100$  échantillons de taille  $n = 20$  selon une loi  $\mathcal{N}(\mu = 1, \sigma^2 = 0.01)$
- On prend  $\alpha = 0.05 = 5\%$  et on calcule pour chaque échantillon, l'intervalle de confiance :  $\mathbf{IC} = \bar{x} \pm u_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}$



- Environ  $\alpha = 0.05 = 5\%$  des intervalles de confiance ne contiennent pas  $\mu$

## Interprétation

Que peut-on dire (avec  $\alpha = 0.05$ ) ?

- L'IC a 95% de chance de contenir  $\mu$  ? **NON !**  
L'IC n'est pas aléatoire : il contient  $\mu$ ... ou pas
- La procédure garantit 95% de réussite  $\Rightarrow$
- On peut avoir un niveau de confiance de 95% en l'intervalle calculé, d'où son nom

## Intervalle de probabilité - Intervalle de confiance

### Définition (Intervalle de probabilité)

Soit  $\alpha \in ]0, 1[$ . On appelle *intervalle de probabilité* pour  $\theta_0$  de niveau  $1 - \alpha$  tout intervalle de la forme  $[A_n, B_n]$ , où  $A_n$  et  $B_n$  sont des fonctions mesurables telles que,  $\forall \theta_0 \in \Theta$  :

$$\mathbb{P}_{\theta_0}(\theta_0 \in [A_n, B_n]) = 1 - \alpha$$

### Définition (Intervalle de confiance)

Soit  $\alpha \in ]0, 1[$ . On appelle *intervalle de confiance* pour  $\theta_0$  de niveau  $1 - \alpha$  toute réalisation  $[a_n, b_n]$  d'un intervalle de probabilité  $[A_n, B_n]$  de niveau  $1 - \alpha$



## Retour à l'exemple 2

- En résumé

$$X_i \underset{\text{i.i.d}}{\sim} \mathcal{N}(\mu, \sigma^2) \Rightarrow \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1) \Rightarrow \mathbf{IC} = \bar{x} \pm u_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

- $\bar{x} = 0.945$  est connu ; si  $\sigma^2$  est connu, on peut calculer **IC**
- **Problème** : en pratique  $\sigma^2$  est inconnu donc on l'estime
- On utilise la statistique  $\hat{\sigma}^2$  consistante et sans biais (cf. TD) :

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

- **Question** : quelle est la loi de  $T = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}}$  ?

## Vers la loi de Student

- La loi de Student est définie pour traiter ce cas classique
- Elle s'appuie sur la loi du  $\chi^2$  qui elle même sert à caractériser les lois des variables comme

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

- Ces deux lois dérivent de la loi normale qui
  - est une famille de lois classiques en estimation
  - qui donne une bonne approximation pour de nombreux résultats via le TCL

## Loi du $\chi^2$

### Définition (Loi du $\chi^2$ )

Soit  $U_1, \dots, U_p$  des variables indépendantes de même loi  $\mathcal{N}(0, 1)$ , on appelle loi du chi-deux à  $p$  degrés de liberté, notée  $\chi_p^2$ , la loi de la variable  $\sum_{i=1}^p U_i^2$

- Par exemple, pour  $X_i \underset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$ , on a :

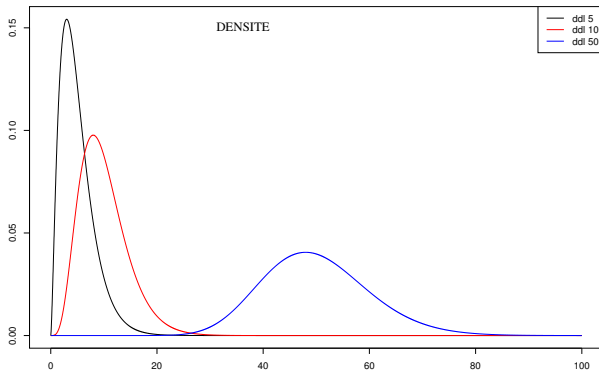
$$\frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^2 \sim \chi_n^2$$

- Surtout, sous les mêmes hypothèses :

$$(n-1) \frac{\hat{\sigma}^2}{\sigma^2} = \sum_{i=1}^n \left( \frac{X_i - \bar{X}}{\sigma} \right)^2 \sim \chi_{n-1}^2$$



# Loi du $\chi^2$



## Loi de Student

### Définition (Loi de Student)

*La loi de Student à  $p$  degrés de liberté est la loi du rapport indépendant d'une loi normale centrée-réduite et de la racine d'un  $\chi^2$  divisé par son degré de liberté  $p$ . Pour  $T$  suivant une loi de Student à  $p$  degrés de liberté, on note  $T \sim \mathcal{T}_p$*

- Comme on a

$$T = \frac{\bar{X} - \mu}{\hat{\sigma}/\sqrt{n}} = \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}} \times \frac{\sigma}{\hat{\sigma}} = \frac{(\bar{X} - \mu)}{\sigma/\sqrt{n}} \times \frac{1}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma}\right)^2}}$$

- $T$  est le rapport<sup>5</sup> d'une loi  $\mathcal{N}(0, 1)$  et de la racine d'un  $\chi^2_{n-1}$  divisé par  $n - 1$  son ddl :  $T \sim \mathcal{T}_{n-1}$

<sup>5</sup>L'indépendance du rapport est admise



## Loi de Student

La loi de Student se rapproche de la loi  $\mathcal{N}(0, 1)$  quand  $n \rightarrow \infty$  :

$$T \xrightarrow[\mathcal{L}]{} \mathcal{N}(0, 1).$$

