

# Statistique pour Mathématiciens

Victor M. Panaretos

Section de Mathématiques – EPFL

`victor.panaretos@epfl.ch`



# Organisation du cours

- Cours mardi 13.15–15.00
- Exercices lundi 10.15–12.00
- Référence principale (**libre accès en ligne**, et disponible à la Librairie La Fontaine, RLC) :

**Panaretos, V.M. (2016). *Statistique pour Mathématiciens*. PPUR.**

- Page web : moodle
- Test bonus le **15 avril, 13.15**
- Examen final écrit.
- La note finale  $N$  sera calculée selon l'algorithme
  - $F = 0.75 \times E + 0.25 \times \max\{E, T\}$
  - $E$  = examen final,  $T$  = test bonus
  - on arrondi  $F$  pour obtenir  $N$

# Introduction

Commençons par les maths :

Mathématiques  $\leftarrow$  μαθηματικά



$\simeq$  apprendre

Une manière :

- 1 d'exprimer une grande variété de notions complexes avec précision et cohérence
- 2 de "*légitimer les conquêtes de notre intuition*"<sup>1</sup> - apprendre, comprendre et conclure correctement

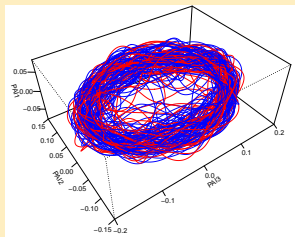
et la statistique ?

utiliser les maths  
pour  
extraire des informations  
à partir de  
données  
en présence d'  
incertitude.

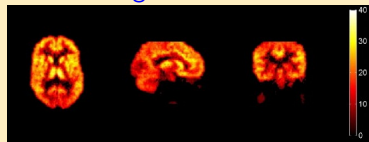
Habituellement, on pense à des ensembles de nombres lorsqu'on parle de données, mais...

...en fait, tous les objets qui peuvent être exprimées mathématiquement sont potentiellement des “données”

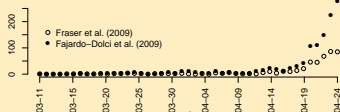
## Biologie structurale



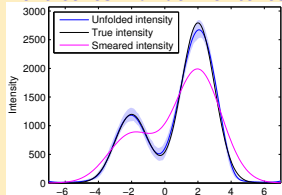
## Imagerie médicale



## Contrôle d'Epidemiques



## Particules Fondamentales



Les probabilités nous aident pour la partie **incertitude**

- C'est la discipline mathématique qui étudie les phénomènes aléatoires (où *stochastiques*)
- Elle consiste en une base sur laquelle on peut construire des modèles qui acceptent la présence d'incertitude

Les probabilités nous donnent un cadre de travail dans lequel on peut comprendre et quantifier l'effet que la présence d'incertitude a sur notre extraction d'informations à partir des données.

# Notre cadre générale

- 1 Nous disposons d'une **distribution  $F(x; \theta)$**  qui dépend d'un paramètre **inconnu**  $\theta \in \mathbb{R}^p$ .
- 2 Nous **observons la réalisation de  $n$  variables aléatoires  $X_1, \dots, X_n$** , indépendantes et identiquement distribuées, qui suivent cette distribution. Mais **nous ne connaissons toujours pas la vraie valeur de  $\theta$  qui a généré les  $X_i$  !**
- 3 Nous voulons utiliser les  $n$  observations (les réalisations de  $X_1, \dots, X_n$ ) afin de faire des **assertions concernant la vraie valeur de  $\theta$** , et de quantifier l'incertitude associée à ces assertions.

Semble trop simpliste ?

→ Contient l'essence de la plupart des idées utilisées dans des problèmes plus complexes !

→ Plusieurs situations plus complexes peuvent souvent être réduites à ce cas simple en utilisant les mathématiques de façon adéquate.

# Quels types d'assertions peut-on faire sur la vraie valeur de $\theta$ ?

Les trois problèmes statistiques que nous allons considérer sont :

- ❶ **Estimation.** Etant donné un échantillon  $X_1, \dots, X_n$  tiré d'une distribution  $F_\theta$  qui dépend d'un paramètre inconnu  $\theta$ , comment peut-on construire un estimateur, i.e une fonction de l'échantillon, dont le but est d'estimer  $\theta$  ?
- ❷ **Tests d'hypothèses.** Etant donnée une valeur plausible  $\theta_0$  pour  $\theta$  (ou plusieurs valeurs plausibles formant un ensemble  $\Theta_0$ ), est-ce que, sur la base de l'échantillon  $X_1, \dots, X_n$ , cette valeur (ou cet ensemble) est un bon indicateur de la vraie valeur de  $\theta$  ?
- ❸ **Intervalles de confiance.** Plutôt que de tenter d'estimer la valeur précise du paramètre  $\theta$  qui a généré notre échantillon  $X_1, \dots, X_n$ , est-ce qu'on peut construire un ensemble de valeurs sous la forme d'un intervalle, qui aura une grande probabilité de contenir le vrai paramètre  $\theta$  ?

Avant d'attaquer ces problèmes statistique, il nous faut développer l'arrière-plan :

- (A) **Modèles probabilistes** : quels modèles, pourquoi, comment les manipuler, comment les choisir, formes abstraites (pour obtenir des résultats qui sont valables pour tous les modèles considérés).
- (B) **Théorie d'échantillonnage** : la relation entre les données et les modèles probabilistes, et le comportement probabiliste des données (de l'échantillon).

Enfin, comme annoncé, nous allons nous intéresser aux trois problèmes :

- (C) **Estimation.**
- (D) **Tests d'hypothèses.**
- (E) **Intervalles de confiance.**

# Modèles Probabilistes

# Nomenclature

Dans le cadre de ce cours, un modèle de probabilité sera la distribution (aussi appelée loi ou fonction de répartition)  $F$  d'une variable aléatoire  $X$  qui prend des valeurs dans le sous-ensemble  $\mathcal{X} \subseteq \mathbb{R}$  de la droite des réels :

$$F(x) = \mathbb{P}[X \leq x], \quad x \in \mathbb{R}.$$

- Ecrivons  $X \sim F$  pour dire que  $F$  est la distribution de  $X$ .
- Si  $\{X_i\}_{i \in I}$  sont de variables aléatoires indépendantes et identiquement distribuées selon la distribution  $F$ , écrivons  $X_i \stackrel{iid}{\sim} F$ .
- $\mathcal{X}$  est appelé l'*espace échantillon*,  $\Theta$  est appelé l'*espace des paramètres*.
- La distribution  $F$  dépendra typiquement d'un ou de plusieurs paramètres,  $\theta = (\theta_1, \dots, \theta_p)^\top \in \Theta \subseteq \mathbb{R}^p$  (dépendamment du contexte, une différente lettre grecque ou latine peut être utilisée).
- Afin d'indiquer que la distribution  $F$  dépend du paramètre  $\theta$ , nous allons souvent écrire  $F_\theta$  ou  $F(x; \theta)$ . Par conséquent :  $F(x; \theta) = \mathbb{P}_\theta[X \leq x]$ .

# Modèles réguliers discrets

Afin de spécifier un modèle de probabilité discret, nous devons définir :

- 1 L'espace échantillon  $\mathcal{X}$  des valeurs possibles que peut prendre la variable aléatoire discrète  $X$ , c'est-à-dire un ensemble discret

$$\mathcal{X} = \{x : \mathbb{P}[X = x] > 0\}.$$

- 2 La valeur de la fonction de masse  $f(x; \theta)$ , en tant que fonction de  $x \in \mathcal{X}$  et de  $\theta \in \Theta$ .

On considèrera seulement de modèles telles que  $\mathcal{X} \subseteq \mathbb{Z}$ .

Rappelons quelques modèles discrètes de base, et pourquoi il sont importants.

## Définition (Distribution de Bernoulli)

*On dit qu'une variable aléatoire  $X$  suit une distribution de Bernoulli de paramètre  $p \in [0, 1]$ , noté  $X \sim \text{Bern}(p)$ , si*

- ❶  $\mathcal{X} = \{0, 1\}$ ,
- ❷  $f(x; p) = p1\{x = 1\} + (1 - p)1\{x = 0\}$ .

*L'espérance, la variance et la fonction génératrice des moments (FGM) de  $X \sim \text{Bern}(p)$  sont données par*

$$\mathbb{E}[X] = p, \quad \text{Var}[X] = p(1 - p), \quad M(t) = 1 - p + pe^t.$$

## Définition (Distribution binomiale)

On dit qu'une variable aléatoire  $X$  suit une distribution binomiale de paramètres  $p \in [0, 1]$  et  $n \in \mathbb{N}$ , noté  $X \sim \text{Binom}(n, p)$ , si

- ❶  $\mathcal{X} = \{0, 1, 2, \dots, n\}$ ,
- ❷  $f(x; p) = \binom{n}{x} p^x (1 - p)^{n-x}$ .

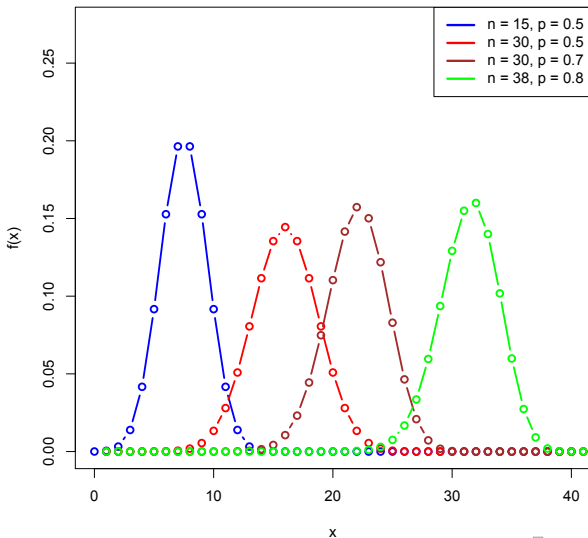
La moyenne, la variance et la fonction génératrice des moments de  $X \sim \text{Binom}(n, p)$  sont données par

$$\mathbb{E}[X] = np, \quad \text{Var}[X] = np(1 - p), \quad M(t) = (1 - p + pe^t)^n.$$

si  $X = \sum_{i=1}^n Y_i$  où  $Y_i \stackrel{iid}{\sim} \text{Bern}(p) \implies X \sim \text{Binom}(n, p)$

# Loi Binomiale

Binomial Distribution PMF



## Définition (Distribution géométrique)

Une variable aléatoire  $X$  suit une distribution géométrique de paramètre  $p \in (0, 1]$ , noté  $X \sim \text{Geom}(p)$ , si

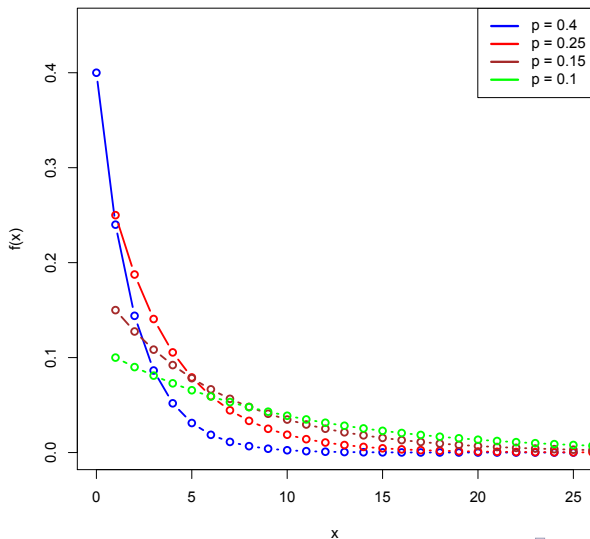
- ❶  $\mathcal{X} = \{0\} \cup \mathbb{N}$ ,
- ❷  $f(x; p) = (1 - p)^x p$ .

La moyenne, la variance et la fonction génératrice des moments de  $X \sim \text{Geom}(p)$  sont données par

$$\mathbb{E}[X] = \frac{1 - p}{p}, \quad \text{Var}[X] = \frac{(1 - p)}{p^2}, \quad M(t) = \frac{p}{1 - (1 - p)e^t}, \quad t < -\log(1 - p)$$

Si  $\{Y_i\}_{i \geq 1}$  sont telles que  $Y_i \stackrel{iid}{\sim} \text{Bern}(p)$  et  $T = \min\{k \in \mathbb{N} : Y_k = 1\} - 1$   
 $\implies T \sim \text{Geom}(p)$

Geometric Distribution PMF



## Définition (Distribution binomiale négative)

Une variable aléatoire  $X$  suit une distribution binomiale négative de paramètres  $p \in (0, 1]$  et  $r > 0$ , noté  $X \sim \text{NegBin}(r, p)$ , si

❶  $\mathcal{X} = \{0\} \cup \mathbb{N}$ ,

❷  $f(x; p, r) = \binom{x+r-1}{x} (1-p)^x p^r$ .

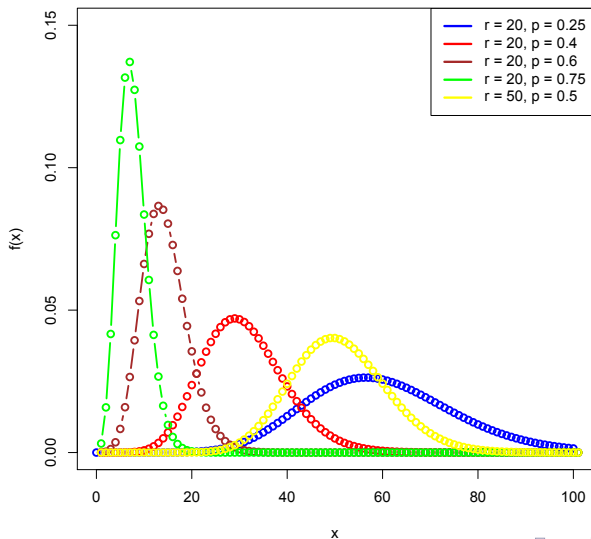
La moyenne, la variance et la fonction génératrice des moments de  $X \sim \text{NegBin}(r, p)$  sont données par

$$\mathbb{E}[X] = r \frac{1-p}{p}, \quad \text{Var}[X] = r \frac{1-p}{p^2}, \quad M(t) = \frac{p^r}{[1 - (1-p)e^t]^r}, \quad t < -\log(1-p).$$

Si  $X = \sum_{i=1}^r Y_i$  où  $Y_i \stackrel{iid}{\sim} \text{Geom}(p) \implies X \sim \text{NegBin}(r, p)$ .

# Loi Binomiale Négative

Negative Binomial Distribution PMF



## Définition (Distribution de Poisson )

Une variable aléatoire  $X$  suit une distribution de Poisson de paramètre  $\lambda > 0$ , noté  $X \sim \text{Poisson}(\lambda)$ , si

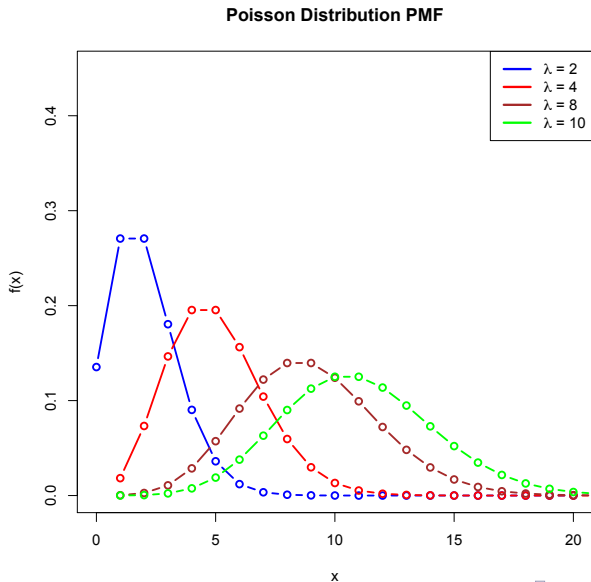
❶  $\mathcal{X} = \{0\} \cup \mathbb{N}$ ,

❷  $f(x; \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$ .

La moyenne, la variance et la fonction génératrice des moments de  $X \sim \text{Poisson}(\lambda)$  sont données par

$$\mathbb{E}[X] = \lambda, \quad \text{Var}[X] = \lambda, \quad M(t) = \exp\{\lambda(e^t - 1)\}.$$

Informellement,  $\text{Binom}(n, p) \rightarrow \text{Poisson}(\lambda)$  lorsque  $n \rightarrow \infty$  et  $p = \lambda/n$



# Modèles réguliers continus

Afin de spécifier un modèle de probabilité continu, nous devons :

- 1 Définir la fonction de densité de probabilité,  $f(x; \theta)$ , en tant que fonction de  $x \in \mathcal{X}$  et de  $\theta \in \Theta$ .
- 2 Spécifier son support (l'ensemble sur lequel  $f(x; \theta) > 0$ ), si ce n'est pas a priori claire.

Rappelons quelques modèles continus de base, et pourquoi il sont importants.

## Définition (Distribution Uniforme)

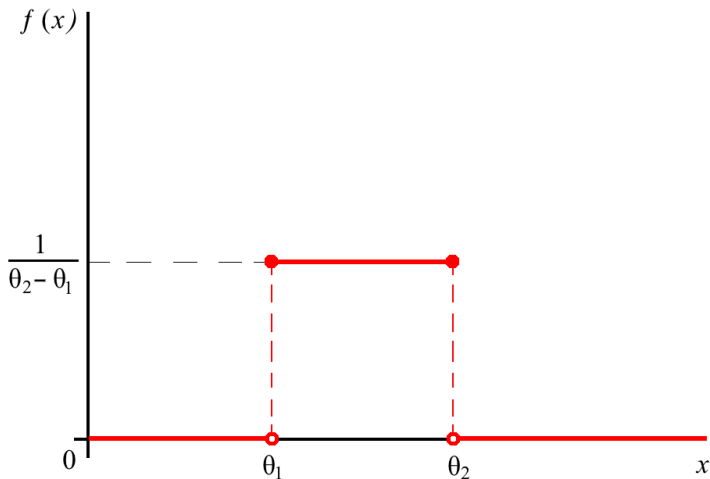
*Une variable aléatoire  $X$  suit une distribution uniforme de paramètres  $-\infty < \theta_1 < \theta_2 < \infty$ , noté  $X \sim \text{Unif}(\theta_1, \theta_2)$ , si*

$$f_X(x; \theta) = \begin{cases} (\theta_2 - \theta_1)^{-1} & \text{si } x \in (\theta_1, \theta_2), \\ 0 & \text{sinon.} \end{cases}$$

*La moyenne, la variance et la fonction génératrice des moments de  $X \sim \text{Unif}(\theta_1, \theta_2)$  sont données par*

$$\mathbb{E}[X] = (\theta_1 + \theta_2)/2, \text{ Var}[X] = (\theta_2 - \theta_1)^2/12, M(t) = \frac{e^{t\theta_2} - e^{t\theta_1}}{t(\theta_2 - \theta_1)}, t \neq 0, M(0) = 1.$$

# Densité uniforme



## Définition (Distribution exponentielle)

Une variable aléatoire  $X$  suit une distribution exponentielle de paramètre  $\lambda > 0$ , noté  $X \sim \text{Exp}(\lambda)$ , si

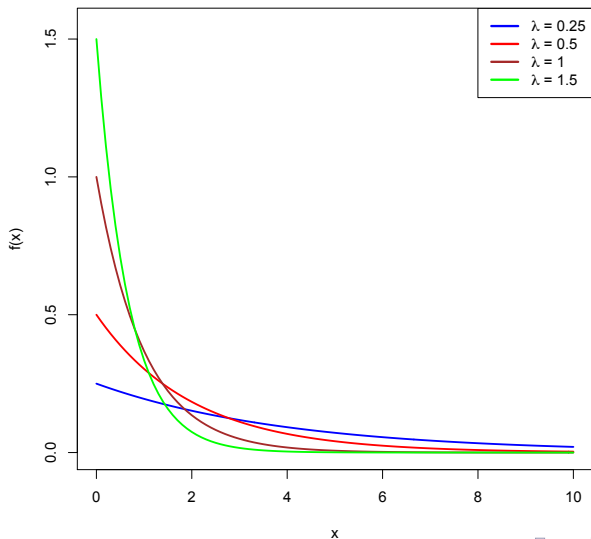
$$f_X(x; \lambda) = \begin{cases} \lambda e^{-\lambda x}, & \text{si } x \geq 0 \\ 0 & \text{si } x < 0. \end{cases}$$

La moyenne, la variance et la fonction génératrice des moments  $X \sim \text{Exp}(\lambda)$  sont données par

$$\mathbb{E}[X] = \lambda^{-1}, \quad \text{Var}[X] = \lambda^{-2}, \quad M(t) = \frac{\lambda}{\lambda - t}, \quad t < \lambda.$$

# Densité exponentielle

Exponential Distribution PDF



## Définition (Distribution gamma)

Une variable aléatoire  $X$  suit une distribution gamma de paramètres  $r > 0$  et  $\lambda > 0$  (respectivement le paramètre de forme et le paramètre d'intensité), noté  $X \sim \text{Gamma}(r, \lambda)$ , si

$$f_X(x; r, \lambda) = \begin{cases} \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}, & \text{si } x \geq 0 \\ 0 & \text{si } x < 0. \end{cases}$$

La moyenne, la variance et la fonction génératrice des moments de  $X \sim \text{Gamma}(r, \lambda)$  sont données par

$$\mathbb{E}[X] = r/\lambda, \quad \text{Var}[X] = r/\lambda^2, \quad M(t) = \left( \frac{\lambda}{\lambda - t} \right)^r, \quad t < \lambda.$$

# Loi Khi Carré (ou Khi Deux)

## Définition (Distribution khi carré)

Une variable aléatoire  $X$  suit une distribution khi carré de paramètre  $k \in \mathbb{N}$  (appelé le nombre de degrés de liberté), noté  $X \sim \chi_k^2$ , si  $X \sim \text{Gamma}(k/2, 1/2)$ . En d'autres mots,

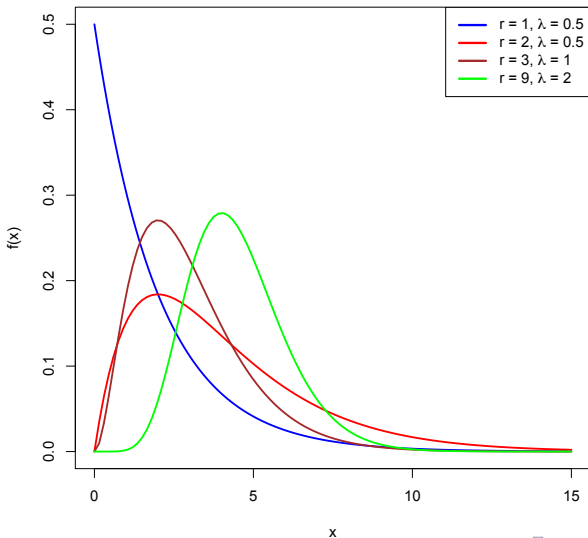
$$f_X(x; k) = \begin{cases} \frac{1}{2^{k/2} \Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}, & \text{si } x \geq 0 \\ 0 & \text{si } x < 0. \end{cases}$$

La moyenne, la variance et la fonction génératrice des moments de  $X \sim \chi_k^2$  sont données par

$$\mathbb{E}[X] = k, \quad \text{Var}[X] = 2k, \quad M(t) = (1 - 2t)^{-k/2}, \quad t < \frac{1}{2}.$$

# Densité gamma

Gamma Distribution PDF



# Loi Normale (où Loi de Gauss)

## Définition (Distribution normale)

Une variable aléatoire  $X$  suit une distribution normale de paramètres  $\mu \in \mathbb{R}$  et  $\sigma^2 > 0$  (respectivement le paramètre moyenne et le paramètre variance), noté  $X \sim N(\mu, \sigma^2)$ , si

$$f_X(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2} \left(\frac{x - \mu}{\sigma}\right)^2\right\}, \quad x \in \mathbb{R}.$$

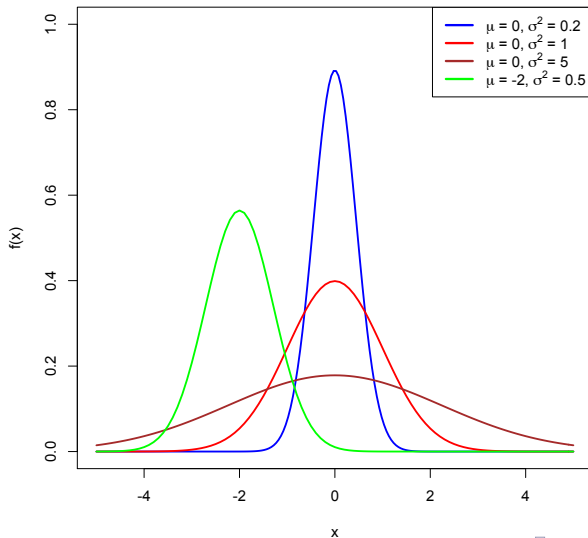
La moyenne, la variance et la fonction génératrice des moments de  $X \sim N(\mu, \sigma^2)$  sont données par

$$\mathbb{E}[X] = \mu, \quad \text{Var}[X] = \sigma^2, \quad M(t) = \exp\{t\mu + t^2\sigma^2/2\}.$$

Dans le cas spécial  $Z \sim N(0, 1)$ , nous utilisons la notation  $\varphi(z) = f_Z(z)$  et  $\Phi(z) = F_Z(z)$ , et nous les appelons respectivement la *fonction de densité normale centrée réduite* (ou *fonction de densité normale standard*) et la *fonction de répartition normale centrée réduite* (ou *fonction de répartition normale standard*).

# Densité normale

Normal Distribution PDF



# ... et on arrête jamais !

La liste n'arrête pas...

...la distribution **Pareto**, la distribution de **Weibull**, la distribution **log-normale**, la distribution **inverse-gamma**, la distribution inverse-gaussienne, la distribution **normale-gamma**, la distribution **beta**...

## Vers un cas général

- ❶ On veut développer une théorie statistique dont les propriétés seront valables pour plusieurs modèles, indépendamment de leur structure spécifique.
- ❷ Peut-on définir une classe (*une famille*) des modèles générale, telle qu'elle nous permette d'étudier les méthodes statistiques dans un cadre général ?
- ❸ Si oui, alors n'importe quelle propriété prouvée pour le cas général sera aussi valide pour les cas spéciaux !
- ❹ Les questions en dessus motivent la définition des **familles exponentielles**.

## Définition (Les familles exponentielles de distributions)

Une classe de distributions de probabilités régulières sur  $\mathcal{X} \subseteq \mathbb{R}$  est une famille exponentielle de distributions à «  $k$ -paramètre » si sa fonction de densité (ou fonction de masse) admet la représentation

$$f(x) = \exp \left\{ \sum_{i=1}^k \phi_i T_i(x) - \gamma(\phi_1, \dots, \phi_k) + S(x) \right\}, \quad x \in \mathcal{X} \quad (2.1)$$

où :

- ❶  $\phi = (\phi_1, \dots, \phi_k)$  est un paramètre de dimension  $k$  dans  $\mathbb{R}^k$  ;
- ❷  $T_i : \mathcal{X} \rightarrow \mathbb{R}$ ,  $i = 1, \dots, k$ ,  $S(x) : \mathcal{X} \rightarrow \mathbb{R}$ , et  $\gamma : \mathbb{R}^k \rightarrow \mathbb{R}$ , sont des fonctions à valeurs réelles ;
- ❸ Le support de  $f$  (l'ensemble  $\mathcal{X}$  sur lequel  $f$  est positive) ne dépend pas de  $\phi$ .

- Le paramètre  $\phi$  est appelé le **paramètre naturel**.

# Forme Naturelle vs Forme Usuelle

$$\exp \left\{ \sum_{i=1}^k \phi_i T_i(x) - \gamma(\phi) + S(x) \right\} = \exp \left\{ \sum_{i=1}^k \eta_i(\theta) T_i(x) - d(\theta) + S(x) \right\}.$$

où  $\eta : \Theta \rightarrow \mathbb{R}^k$  est une fonction injective deux fois différentiable, tel que

$$\phi = \eta(\theta)$$

et donc  $\gamma(\phi) = \gamma(\eta(\theta)) = d(\theta)$ , pour  $d = \gamma \circ \eta$ .

- **Forme naturelle** : typiquement meilleure pour faire la **théorie**.
- **Forme usuelle** : typiquement meilleure dans le cadre des **applications**.

## Exemple (Famille exponentielle binomiale)

Soit  $X \sim \text{Binom}(n, p)$ . Observons que :

$$\binom{n}{x} p^x (1-p)^{n-x} = \exp \left\{ \log \left( \frac{p}{1-p} \right) x + n \log(1-p) + \log \binom{n}{x} \right\}.$$

Définissons :

$$\phi = \log \left( \frac{p}{1-p} \right), \quad T(x) = x,$$

$$S(x) = \log \binom{n}{x}, \quad \gamma(\phi) = n \log(1 + e^\phi) = -n \log(1 - p).$$

Ainsi, si  $n$  est maintenu fixe et que seulement  $p$  a le droit de varier, le support de  $f$  ne dépend pas de  $\phi$  et on a une famille exponentielle à 1-paramètre. Ici le paramètre usuel est une bijection deux fois différentiable du paramètre naturel  $\phi$  :

$$p = \frac{e^\phi}{1 + e^\phi} \quad \& \quad \phi = \underbrace{\log \left( \frac{p}{1-p} \right)}_{=\eta(p)}.$$

Ici  $p \in (0, 1)$ , mais  $\phi \in \mathbb{R}$ .



## Exemple (Famille exponentielle gaussienne)

Soit  $X \sim N(\mu, \sigma^2)$ . Nous pouvons alors écrire :

$$\begin{aligned} f(x; \mu, \sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left( \frac{x - \mu}{\sigma} \right)^2 \right\} \\ &= \exp \left\{ -\frac{1}{2\sigma^2} x^2 + \frac{\mu}{\sigma^2} x - \frac{1}{2} \log(2\pi\sigma^2) - \frac{\mu^2}{2\sigma^2} \right\}. \end{aligned}$$

Définissons :

$$\phi_1 = \frac{\mu}{\sigma^2}, \quad \phi_2 = -\frac{1}{2\sigma^2},$$

$$T_1(x) = x, \quad T_2(x) = x^2, \quad S(x) = 0, \quad \gamma(\phi_1, \phi_2) = -\frac{\phi_1^2}{4\phi_2} + \frac{1}{2} \log \left( -\frac{\pi}{\phi_2} \right),$$

et observons que le support de  $f$  est toujours  $\mathbb{R}$ , indépendamment des valeurs du paramètre. Nous obtenons donc que la distribution  $N(\mu, \sigma^2)$  est une famille exponentielle à 2-paramètres. □

# Modèles de probabilité transformés

# Modèles de probabilité transformés

- Souvent : nous avons un modèle pour un phénomène aléatoire  $X$
- Mais nous sommes plutôt intéressés par un autre aspect de ce phénomène, disons  $g(X)$ , où  $g$  est une fonction connue.

## Exemple

Supposons que  $R$  est une variable aléatoire positive représentant le rayon de couverture d'une antenne Wireless et considérons que  $R \sim Unif[a, b]$ , pour  $0 < a < b$ .

Quelle est la distribution de l'aire de couverture  $A = \pi R^2$  ?



## Modèles de probabilité transformés

Comment la distribution d'une variable aléatoire  $X$  est transformée, lorsque la variable aléatoire  $X$  est transformée ?

## Lemme

Soit  $X$  une variable aléatoire discrète, et  $Y = g(X)$ . Alors, l'espace échantillon de  $Y$  est  $\mathcal{Y} = g(\mathcal{X})$  et

$$F_Y(y) = \mathbb{P}[g(X) \leq y] = \sum_{x \in \mathcal{X}} f_X(x) 1\{g(x) \leq y\}, \quad \forall y \in \mathcal{Y} \quad (3.1)$$

$$f_Y(y) = \mathbb{P}[g(X) = y] = \sum_{x \in \mathcal{X}} f_X(x) 1\{g(x) = y\}, \quad \forall y \in \mathcal{Y}. \quad (3.2)$$

- Preuve = énoncé !
- Cas continu : plus compliqué :
  - ❶ Si  $g$  pas monotone : au cas-par-cas.
  - ❷ Si  $g$  est monotone : on a des résultats généraux.

## Example (La normale standard au carré a une distribution $\chi_1^2$ )

Soit  $Z \sim N(0, 1)$ . Nous voulons trouver la distribution de  $Y = Z^2$ . Notez que  $F_Y(y) = \mathbb{P}[Y \leq y] = 0$  si  $y < 0$ . Pour  $y \geq 0$  nous avons :

$$\begin{aligned} F_Y(y) &= \mathbb{P}[Z^2 \leq y] = \mathbb{P}[|Z| \leq \sqrt{y}] = \mathbb{P}[-\sqrt{y} \leq Z \leq \sqrt{y}] \\ &= \Phi(\sqrt{y}) - \Phi(-\sqrt{y}) = \Phi(\sqrt{y}) - (1 - \Phi(\sqrt{y})) = 2\Phi(\sqrt{y}) - 1. \end{aligned}$$

Nous pouvons aussi trouver la densité en dérivant :

$$\begin{aligned} f_Y(y) &= 2 \frac{d}{dy} \Phi(\sqrt{y}) = 2 \frac{d}{d\sqrt{y}} \Phi(\sqrt{y}) \frac{d}{dy} \sqrt{y} \\ &= 2\phi(\sqrt{y}) \frac{y^{-1/2}}{2} = 2 \frac{1}{\sqrt{2\pi}} e^{-y/2} \frac{y^{-1/2}}{2} \\ &= \frac{1}{\sqrt{2}\sqrt{\pi}} e^{-y/2} y^{-1/2} = \frac{1}{2^{1/2}\Gamma(1/2)} y^{1/2-1} e^{-y/2}. \end{aligned}$$

Notez que la dernière expression est la densité d'une distribution  $\chi_1^2$ . Alors :

$$\boxed{Z \sim N(0, 1) \implies Z^2 \sim \chi_1^2.} \quad (3.3)$$

## Lemme

Soit  $X$  une variable aléatoire continue sur  $\mathcal{X} \subseteq \mathbb{R}$  et soit  $g : \mathcal{X} \rightarrow \mathbb{R}$  une

- ① monotone,
- ② continûment dérivable,
- ③ de dérivée jamais nulle.

Soit  $Y = g(X)$ . Alors, l'espace échantillon de  $Y$  est  $\mathcal{Y} = g(\mathcal{X})$  et

- Si  $g$  est croissante, alors

$$F_Y(y) = F_X(g^{-1}(y)).$$

- Si  $g$  est décroissante, alors

$$F_Y(y) = 1 - F_X(g^{-1}(y)).$$

Dans les deux cas, nous aurons :

$$f_Y(y) = \left| \frac{\partial}{\partial y} g^{-1}(y) \right| f_X(g^{-1}(y)), \quad y \in \mathcal{Y}.$$

## Corollaire (Transformations affines)

Soit  $X$  une variable aléatoire et  $Y = g(X)$ . Si  $g(x) = ax + b$ ,  $a \neq 0$ , alors

$$\forall y \in \mathcal{Y}, \quad F_Y(y) = \begin{cases} F_X\left(\frac{y-b}{a}\right) & a > 0, \\ 1 - F_X\left(\frac{y-b}{a}\right) + \mathbb{P}\left(X = \frac{y-b}{a}\right) & a < 0, \end{cases}$$

avec  $\mathbb{P}\left(X = \frac{y-b}{a}\right) = 0$  si  $X$  est une variable aléatoire continue. Ainsi, pour  $y \in \mathcal{Y}$  :

❶  $f_Y(y) = |a|^{-1} f_X\left(\frac{y-b}{a}\right)$ , si  $X$  est continue,

❷  $f_Y(y) = f_X\left(\frac{y-b}{a}\right)$ , si  $X$  est discrète.

## Lemme (Transformations affines de la distribution normale)

Soit  $X \sim N(\mu, \sigma^2)$ ,  $a \neq 0$ . Alors  $aX + b \sim N(a\mu + b, a^2\sigma^2)$ . Par conséquent, si  $X \sim N(\mu, \sigma^2)$ , alors

$$F_X(x) = \Phi\left(\frac{x - \mu}{\sigma}\right),$$

où  $\Phi$  est la fonction de répartition standard,

$\Phi(u) = \int_{-\infty}^u (2\pi)^{-1/2} \exp\{-z^2/2\} dz$ , qui est, on le rappelle, la fonction de répartition d'une variable aléatoire  $Z \sim N(0, 1)$ .

## Théorème (Transformations multidimensionnelles)

Soit  $g : \mathbb{R}^n \rightarrow \mathbb{R}^n$  une bijection différentiable,

$$g(\mathbf{x}) = (g_1(\mathbf{x}), \dots, g_n(\mathbf{x})), \quad \mathbf{x} = (x_1, \dots, x_n)^\top \in \mathbb{R}^n.$$

Soit  $X = (X_1, \dots, X_n)^\top$  ayant la distribution conjointe  $f_X(\mathbf{x})$ ,  $\mathbf{x} \in \mathbb{R}^n$ , et définissons  $Y = (Y_1, \dots, Y_n)^\top = g(X)$ . Alors, si  $\mathcal{Y}^n = g(\mathcal{X}^n)$ , nous avons

$$f_Y(\mathbf{y}) = f_X(g^{-1}(\mathbf{y})) \left| \det \left[ J_{g^{-1}}(\mathbf{y}) \right] \right|, \quad \text{pour } \mathbf{y} = (y_1, \dots, y_n)^\top \in \mathcal{Y}^n,$$

et zero sinon, lorsque  $J_{g^{-1}}(\mathbf{y})$  est bien défini. Ici,  $J_{g^{-1}}(\mathbf{y})$  est la matrice Jacobienne de  $g^{-1}$ , i.e. la fonction à valeur dans l'espace des matrices de dimension  $(n, n)$ ,

$$J_{g^{-1}}(\mathbf{y}) = \begin{bmatrix} \frac{\partial}{\partial x_1} g_1^{-1}(\mathbf{y}) & \cdots & \frac{\partial}{\partial x_n} g_1^{-1}(\mathbf{y}) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} g_n^{-1}(\mathbf{y}) & \cdots & \frac{\partial}{\partial x_n} g_n^{-1}(\mathbf{y}) \end{bmatrix}.$$

## Example (Convolution de densités)

Soient  $X$  et  $Y$  deux variables aléatoires continues **indépendantes**, avec densités  $f_X$  et  $f_Y$ . La densité de la variable  $X + Y$  égale la *convolution* de  $f_X$  et  $f_Y$  :

$$f_{X+Y}(u) = \int_{-\infty}^{+\infty} f_X(u-v)f_Y(v)dv.$$

Définissons  $g : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ ,  $(x, y) \xrightarrow{g} (x+y, y)$   $(u, v) \xrightarrow{g^{-1}} (u-v, v)$ .

La jacobienne de l'inverse est

$$\begin{pmatrix} 1 & 0 \\ -1 & 1 \end{pmatrix}$$

dont la déterminante absolue vaut 1. Il s'ensuit que

$$f_{X+Y,Y}(u, v) = f_{X,Y}(u-v, v) \stackrel{\text{ind.}}{=} f_X(u-v)f_Y(v),$$

et on intègre par rapport à  $v$  pour trouver la marginale  $f_{X+Y}$  :

$$f_{X+Y}(u) = \int_{-\infty}^{+\infty} f_X(u-v)f_Y(v)dv.$$

# Application : Sommes des variables aléatoires normales

## Exercice

Soient  $X_1 \sim N(\mu_1, \sigma_1^2)$  et  $X_2 \sim N(\mu_2, \sigma_2^2)$  deux variables aléatoires indépendantes. Montrez que

$$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

## Corollaire

Soient  $X_1, \dots, X_n$  de variables aléatoires indépendantes telles que  $X_i \sim N(\mu_i, \sigma_i^2)$ , et soit  $S_n = \sum_{i=1}^n X_i$ . Alors,

$$S_n \sim N\left(\sum_{i=1}^n \mu_i, \sum_{i=1}^n \sigma_i^2\right).$$

# Sélection de modèle

# Comment choisir le bon modèle probabiliste ?

**Comment** choisir un modèle ?

et

**Pourquoi** la distribution supposée est un bon modèle pour le phénomène considéré ?

En termes très généraux, la sélection d'un modèle est basée sur :

- ❶ la théorie scientifique et des expériences préalables ;
- ❷ des principes philosophiques ;
- ❸ une analyse exploratoire des données ;
- ❹ une combinaison de (1), (2) et (3).

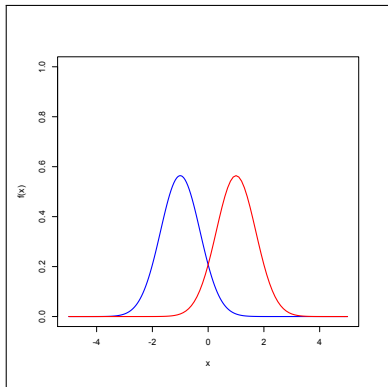
# Analyse exploratoire des données

Parfois  $\longrightarrow$  modèle de probabilité ne peut pas être choisi sans équivoque au moyen de lois physiques et/ou de principes scientifiques. Quoi faire ?

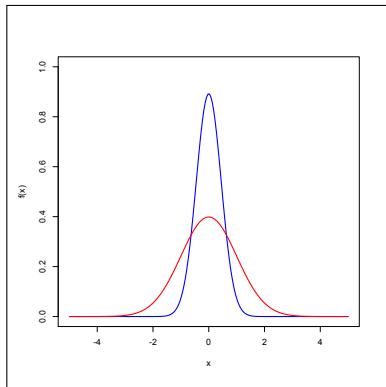
Si on a observations  $x_1, \dots, x_n$ , on peut les utiliser pour choisir entre plusieurs choix, ou au moins exclure certains choix.

Comment ? – en essayant d'apprécier certaines caractéristiques importantes que nous devrions prendre en considération quand on fait un choix de modèle :

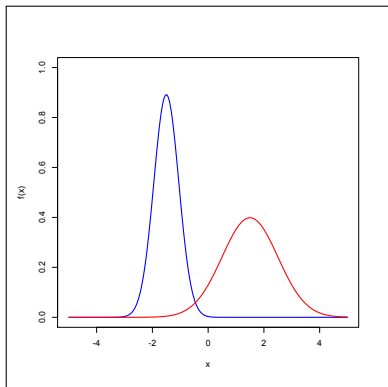
- 1 Position.
- 2 Dispersion.
- 3 Comportement des Queues.
- 4 Symétrie/Asymétrie.



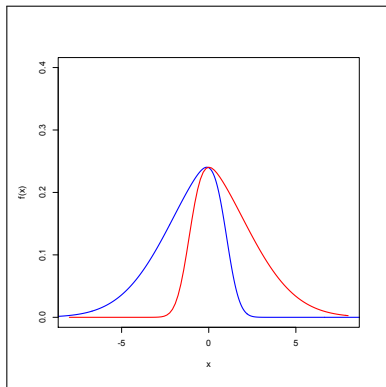
(a) Deux densités de positions différentes.



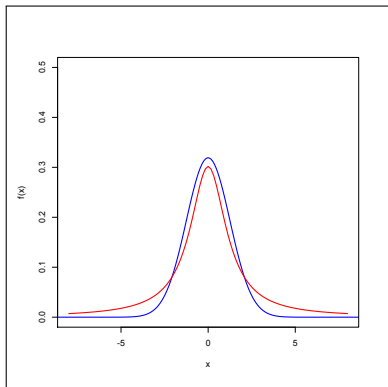
(b) Deux densités de dispersions différentes.



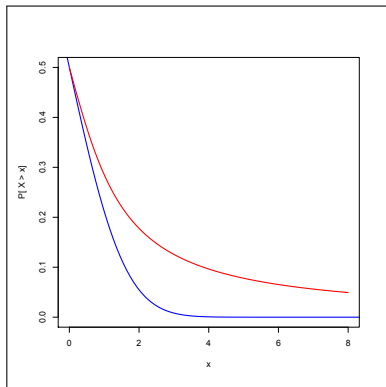
(c) Deux densités qui diffèrent par leur position et leur dispersion.



(d) Deux densités asymétriques : une avec une asymétrie positive (rouge), et une avec une asymétrie négative (bleu).



(e) Une densité à queue lourde (rouge) et une densité à queue légère (bleu).



(f) Graphique de la fonction  $x \mapsto \int_x^\infty f(y)dy$  pour les deux densités de gauche.

Pour apprécier les 4 caractéristiques importantes, on considéra des résumés :

❶ Numériques.

❷ Graphiques.

Tout d'abord, quelques notations utiles :

## Echantillon ordonné

si  $x_1, \dots, x_n$  sont  $n$  valeurs réelles, nous dénotons par  $x_{(j)}$  la  $j^{\text{e}}$  valeur de l'échantillon, lorsque ces valeurs sont placées en ordre croissant (tel que  $x_{(1)} = \min\{x_1, \dots, x_n\}$  et  $x_{(n)} = \max\{x_1, \dots, x_n\}$ ). Notez que ceci signifie que

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n-1)} \leq x_{(n)}.$$

## Exemple

Afin d'illustrer la notation, supposons que  $n = 4$  et que nous avons

$$x_1 = 5, x_2 = 12, x_3 = 2, x_4 = 12.$$

Nous écrivons alors  $x_{(1)} = 2$ ,  $x_{(2)} = 5$  et  $x_{(3)} = x_{(4)} = 12$ . Dans ce cas, nous avons donc  $x_{(1)} = x_3$ ,  $x_{(2)} = x_1$ ,  $x_{(3)} = x_{(4)} = x_2 = x_4$ .

## Définition (Moyenne et Médiane Empirique.)

Soit  $x_1, \dots, x_n$  une collection de nombres réels, appelé un échantillon. Nous définissons :

- ❶ La **moyenne empirique** comme suit

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

- ❷ La **médiane empirique** comme suit

$$M = \begin{cases} x_{(\frac{n+1}{2})} & \text{si } n \text{ est impair,} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} & \text{sinon.} \end{cases}$$

## Définition (Variance empirique et DAM)

Soit  $x_1, \dots, x_n$  une collection de nombres réels, appelé un échantillon. Nous définissons :

- 1 La **variance empirique** comme suit

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2,$$

(l'écart-type empirique est défini comme suit  $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ ).

- 2 La **DéviatiOn Absolue par rapport à la Moyenne (DAM)** comme suit

$$DAM = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|.$$

## Définition (Quartiles, EIQ et valeurs aberrantes)

Soit  $x_1, \dots, x_n$  un échantillon de  $n$  valeurs réelles, et soit

$$x_{(1)}, \dots, M, \dots, x_{(n)}$$

l'échantillon ordonné, où  $M$  est la médiane. Nous définissons :

- 1 Le **premier quartile**,  $Q_1$ , comme étant la médiane du sous-échantillon ordonné  $x_{(1)}, x_{(2)}, \dots, M$ .
- 2 Le **second quartile**,  $Q_2$ , comme étant la médiane  $M$ ,  $Q_2 = M$ .
- 3 Le **troisième quartile**,  $Q_3$ , comme étant la médiane du sous-échantillon ordonné  $M, \dots, x_{(n-1)}, x_{(n)}$ .
- 4 L'**écart interquartile (EIQ)** comme étant  $EIQ = Q_3 - Q_1$ .
- 5 Une **valeur aberrante (anglais : outlier)** est une observation qui n'appartient pas à l'intervalle  $\left[Q_1 - \frac{3}{2}EIQ, Q_3 + \frac{3}{2}EIQ\right]$ .

## Définition (Coefficient de dissymétrie empirique)

Soit  $x_1, \dots, x_n$  un échantillon de  $n$  valeurs réelles. Nous définissons le **coefficient de dissymétrie** de cet échantillon comme

$$SK = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3}{\left( \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{3/2}}.$$

Si le numérateur et le dénominateur sont égaux à zéro (ce qui peut se produire dans un échantillon discret), alors  $SK$  est indéfini.

## Définition (Histogramme)

Soient  $x_1, \dots, x_n$  une collection de  $n$  valeurs réelles et  $h > 0$  une constante. Soit  $\{I_j\}_{j \in \mathbb{Z}}$  une partition régulière de  $\mathbb{R}$  contenant des intervalles de longueur  $h > 0$ ,

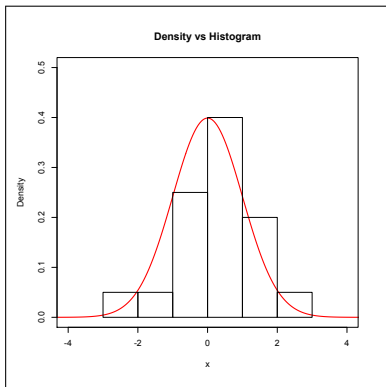
$$I_j = [\kappa + (j - 1)h, \kappa + jh), \quad j \in \mathbb{Z}$$

où  $\kappa \in \mathbb{R}$  est un certain nombre réel fixe. L'histogramme de  $x_1, \dots, x_n$  avec des intervalles de longueur  $h > 0$  et d'origine  $\kappa$  est défini comme étant le graphique de la fonction :

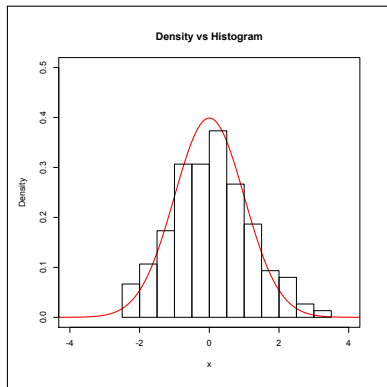
$$y \mapsto \text{hist}_{x_1, \dots, x_n}(y) = \frac{1}{h} \sum_{j \in \mathbb{Z}} 1\{y \in I_j\} \frac{1}{n} \sum_{i=1}^n 1\{x_i \in I_j\}.$$

Deux remarques :

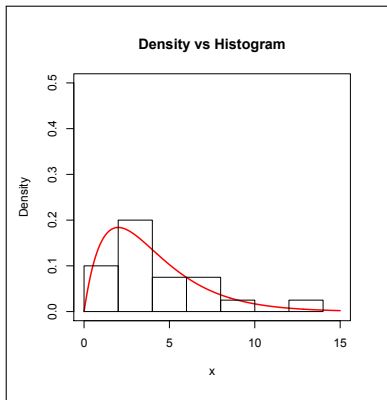
- $\int_{I_j} \text{hist}_{X_1, \dots, X_n}(y) dy$  nous donne la proportion des observations de l'échantillon qui appartiennent à  $I_j$ .
- $\mathbb{E} \left[ \int_{I_j} \text{hist}_{X_1, \dots, X_n}(y) dy \right] = \frac{1}{n} \sum_{i=1}^n \mathbb{P}[X_i \in I_j] = \int_{I_j} f(y) dy.$



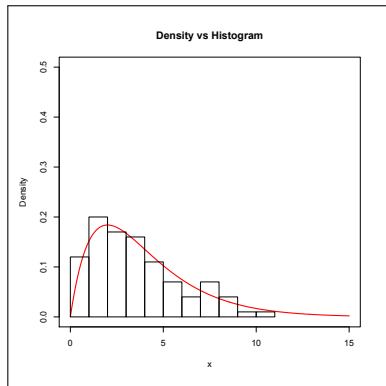
(g) Densité d'une  $N(0, 1)$  (en rouge) et l'histogramme d'un échantillon aléatoire de taille 20 tiré d'une  $N(0, 1)$  (en noir)).



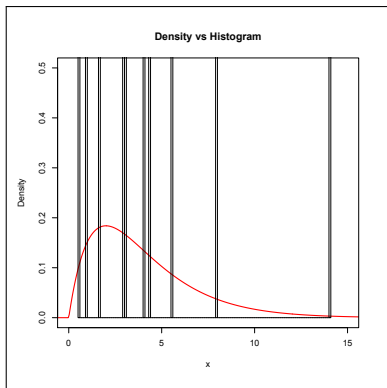
(h) Densité d'une  $N(0, 1)$  (en rouge) et l'histogramme d'un échantillon aléatoire de taille 100 tiré d'une  $N(0, 1)$  (en noir)).



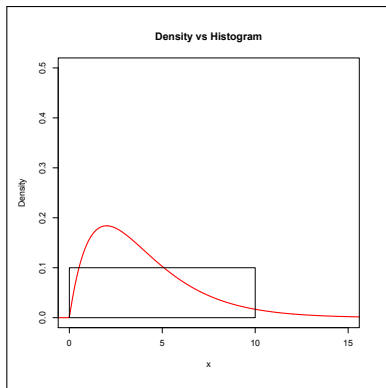
(i) Densité d'une  $\chi^2_2$  (en rouge) et l'histogramme d'un échantillon aléatoire de taille 20 tiré d'une  $\chi^2_2$  (en noir)).



(j) Densité d'une  $\chi^2_2$  (en rouge) et l'histogramme d'un échantillon aléatoire de taille 100 tiré d'une  $\chi^2_2$  (en noir)).



(k) Densité d'une  $\chi^2_2$  (en rouge) et l'histogramme d'un échantillon aléatoire de taille 20 tiré d'une  $\chi^2_2$  (en noir)) lorsque la largeur des intervalles  $h$  est très petite.



(l) Densité d'une  $\chi^2_2$  (en rouge) et l'histogramme d'un échantillon aléatoire de taille 20 tiré d'une  $\chi^2_2$  (en noir)) lorsque la largeur des intervalles  $h$  est très grande.

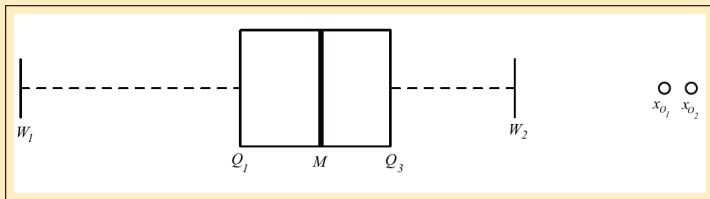
# Résumés graphiques : boxplot

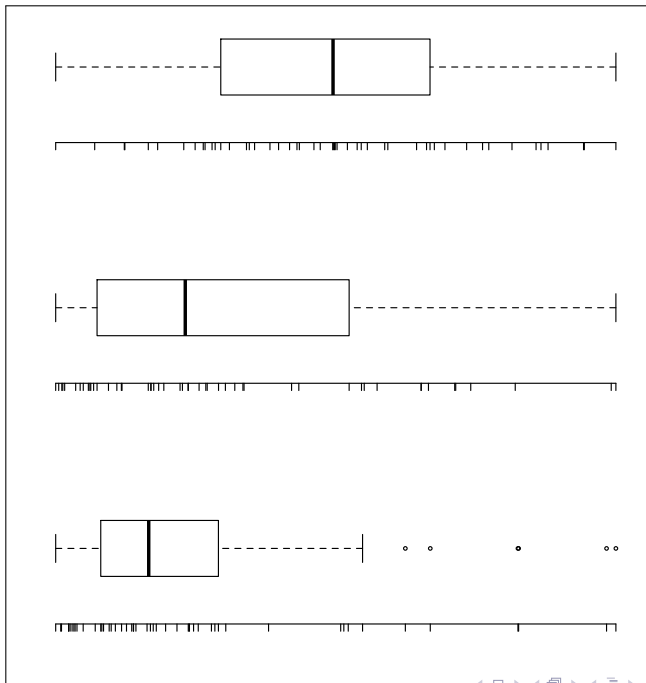
## Définition (Boîte à moustaches (anglais : boxplot))

Soit  $x_1, \dots, x_n$  une collection de  $n$  valeurs réelles. Soient :

- 1  $M$  la médiane,  $Q_1$  le premier quartile, et  $Q_3$  le troisième quartile de  $\{x_1, \dots, x_n\}$ .
- 2  $W_1 = \min_{1 \leq j \leq n} \{x_j : x_j \geq Q_1 - 1.5 \times \text{EIQ}\}$  &  
 $W_2 = \max_{1 \leq j \leq n} \{x_j : x_j \leq Q_3 + 1.5 \times \text{EIQ}\}$ .
- 3  $O = \{i \in \{1, \dots, n\} : x_i \notin [W_1, W_2]\}$ .

La boîte à moustaches de  $x_1, \dots, x_n$  est une annotation des valeurs  $M$ ,  $Q_1$ ,  $Q_3$ ,  $W_1$ ,  $W_2$ , et  $\{x_j : j \in O\}$  sur la droite réelle. La figure suivante est une annotation standard :





# Echantillonnage

# Retour au cadre général

- 1 Il y a une **distribution  $F(x; \theta)$**  qui dépend d'un **paramètre inconnu  $\theta \in \mathbb{R}^p$** .
- 2 Nous **observons la réalisation de  $n$  variables aléatoires  $X_1, \dots, X_n$** , indépendantes et identiquement distribuées, qui suivent cette distribution.
- 3 Nous voulons utiliser les  $n$  observations (les réalisations de  $X_1, \dots, X_n$ ) afin de faire des **assertions concernant la vraie valeur de  $\theta$** .

Puisque tout ce que nous avons en main est l'échantillon, **nous travaillerons essentiellement avec une fonction de l'échantillon**, disons  $T(X_1, \dots, X_n)$

Il faut, donc, comprendre le comportement probabiliste d'une telle fonction  $T(X_1, \dots, X_n)$ . Ceci est appelé *théorie d'échantillonnage*.

## Définition (Statistique)

*Soit  $\mathcal{X}$  un espace échantillon. Une statistique est une fonction  $T : \mathcal{X}^n \rightarrow \mathbb{R}$ .*

- Une statistique  $T : \mathcal{X}^n \rightarrow \mathbb{R}$  réduit une collection de  $n$  nombres à une seule valeur.
- Cependant, pour certains modèles, il est possible de choisir une statistique  $T$  telle que  $T(X_1, \dots, X_n)$  soit aussi informative au sujet de  $\theta$  que  $(X_1, \dots, X_n)$ .

## Définition (Exhaustivité)

Soit  $X_1, \dots, X_n \stackrel{iid}{\sim} f_\theta$ . Une statistique  $T : \mathcal{X}^n \rightarrow \mathbb{R}$  est appelée exhaustive pour le paramètre  $\theta$ , si

$$\mathbb{P}[X_1 \leq x_1, \dots, X_n \leq x_n | T = t]$$

ne dépend pas de  $\theta$ , pour tout  $(x_1, \dots, x_n)^\top \in \mathbb{R}^n$  et pour tout  $t \in \mathbb{R}$ .

- Si une telle statistique existe, la seule connaissance de  $T$  suffit pour faire des inférences sur  $\theta$ .

## Exemple (Estimer le biais d'une pièce de monnaie)

Soit  $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Bern}(\theta)$ , et  $T(X) = \sum_{i=1}^n X_i$ . Pour  $\mathbf{x} \in \{0, 1\}^n$ ,

$$\begin{aligned}\mathbb{P}[X = \mathbf{x} | T = t] &= \frac{\mathbb{P}[X = \mathbf{x}, T = t]}{\mathbb{P}[T = t]} = \frac{\mathbb{P}[X = \mathbf{x}]}{\mathbb{P}[T = t]} 1_{\{\sum_{i=1}^n x_i = t\}} \\ &= \frac{\theta^{\sum_{i=1}^n x_i} (1 - \theta)^{n - \sum_{i=1}^n x_i}}{\binom{n}{t} \theta^t (1 - \theta)^{n-t}} 1_{\{\sum_{i=1}^n x_i = t\}} \\ &= \frac{\theta^t (1 - \theta)^{n-t}}{\binom{n}{t} \theta^t (1 - \theta)^{n-t}} 1_{\{\sum_{i=1}^n x_i = t\}} = \binom{n}{t}^{-1} 1_{\{\sum_{i=1}^n x_i = t\}}.\end{aligned}$$

- $T$  est alors exhaustive pour  $p$ . Cela signifie qu'afin d'obtenir des informations concernant  $p$ , tout ce qui est important est de connaître le nombre total de « faces » ; en effet, l'ordre précis dans lequel sont apparues ces « faces » n'est pas pertinent dans ce cas-ci :

0 0 1 1 1 0 1    VS    1 0 0 0 1 1 1    VS    1 0 1 0 1 0 1

Comment vérifier q'une statistique est exhaustive ?

## Théorème (Critère de Fisher-Neyman (ou Critère de factorisation))

*Supposons que  $(X_1, \dots, X_n)$  a une fonction de densité/masse conjointe  $f_{X_1, \dots, X_n}(x_1, \dots, x_n; \theta)$ ,  $\theta \in \Theta$ . Une statistique  $T : \mathcal{X}^n \rightarrow \mathbb{R}$  est exhaustive pour  $\theta$  si et seulement si il existe des fonctions  $g : \mathbb{R} \times \Theta \rightarrow \mathbb{R}$  et  $h : \mathbb{R}^n \rightarrow \mathbb{R}$  telles que*

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n; \theta) = g(T(x_1, \dots, x_n), \theta)h(x_1, \dots, x_n).$$

## Exemple (Estimer le biais d'une pièce de monnaie)

Soit  $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Bern}(p)$ . Alors,

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n f_{X_i}(x_i) = p^{\sum_{i=1}^n 1\{x_i=1\}} (1-p)^{n - \sum_{i=1}^n 1\{x_i=1\}}.$$

Ainsi, le critère de Fisher-Neyman est satisfait avec

$$T(X_1, \dots, X_n) = \sum_{i=1}^n 1\{X_i = 1\} = \sum_{i=1}^n X_i$$

$$g(t, p) = p^t (1-p)^{n-t}$$

$$h(x_1, \dots, x_n) = 1.$$

Il s'ensuit que  $\sum_{i=1}^n X_i$  est exhaustive pour  $p$ .



## Définition (Distribution d'échantillonnage)

Soient  $X_1, \dots, X_n \stackrel{iid}{\sim} F$  et  $T : \mathcal{X}^n \rightarrow \mathbb{R}$  une statistique. La distribution d'échantillonnage de  $T$  sous la distribution  $F$  est la distribution de probabilité

$$F_T(t) = \mathbb{P}[T(X_1, \dots, X_n) \leq t], \quad t \in \mathbb{R}.$$

## Notation

Nous allons très souvent écrire simplement  $T$  au lieu de  $T(X_1, \dots, X_n)$ .

Dans cette notation, la distribution d'échantillonnage de  $T$  sous  $F$  est  $F_T(t) = \mathbb{P}[T \leq t]$ .

# Echantillonnage

Dans la définition de la distribution d'échantillonnage de  $T$ , nous avons spécifié sous quelle distribution  $F$  celle-ci se produit.

↪ Changer la loi des  $X_i$  (pour une certaine distribution  $G$  plutôt que  $F$ ) aura pour conséquence de changer aussi la distribution d'échantillonnage de  $T$ .

Il faut, donc, examiner précisément cette dépendance :

- 1 Examiner certaines formes spéciales de  $T$  et de  $F$
- 2 Dans des situations générales, tenter de donner des moyens d'établir une distribution approximative
- 3 Nous allons nous concentrer sur des statistiques  $T$  exhaustives et des modèles  $F$  constituant des familles exponentielles.

# Echantillonnage d'une distribution normale

Commençons avec un cas spécial, qui est quand-même d'importance majeure :

- La moyenne et la variance empirique de variables aléatoires normales

## Theorem (Théorème de Student-Fisher sur l'échantillonnage gaussien)

Soit  $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$ , et  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ ,  $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ .  
Alors,

- 1 La distribution conjointe de  $X_1, \dots, X_n$  a pour fonction de densité :

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right\}.$$

- 2 La moyenne empirique est distribuée comme suit :  $\bar{X} \sim N(\mu, \sigma^2/n)$ .
- 3 Les variables aléatoires  $\bar{X}$  et  $S^2$  sont indépendantes.
- 4 La variable aléatoire  $S^2$  satisfait  $\frac{n-1}{\sigma^2} S^2 \sim \chi_{n-1}^2$ .

## Corollaire (Moments pour l'échantillonnage d'une loi normale)

Soit  $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$ , alors

$$\mathbb{E}[\bar{X}] = \mu, \quad \text{Var}(\bar{X}) = \frac{\sigma^2}{n}, \quad \mathbb{E}[S^2] = \sigma^2, \quad \text{Var}(S^2) = \frac{2\sigma^4}{n-1}.$$

(c'est pourquoi nous utilisons un facteur  $(n-1)^{-1}$  au lieu de  $n^{-1}$  dans la définition de  $S^2$ )

## Théorème (La statistique de Student et sa loi d'échantillonnage)

Soit  $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$ , alors

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Ici  $t_{n-1}$  représente la distribution de Student avec  $n-1$  degrés de liberté.

## Définition (Distribution t de Student )

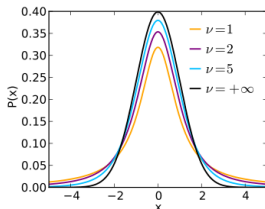
Une variable aléatoire  $X$  suit une distribution t de Student de paramètre  $k \in \mathbb{N}$  (appelé nombre de degrés de liberté), noté  $X \sim t_k$ , si

$$f_X(x; k) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right) \sqrt{k\pi}} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}},$$

La moyenne et la variance de  $X \sim t_k$  sont données par

$$\mathbb{E}[X] = 0, \quad \text{Var}[X] = \frac{k}{k-2},$$

pour  $k > 2$ . La moyenne n'est pas définie pour  $k = 1$  et la variance est non-définie pour  $k \leq 2$ . Pour tout  $k \in \mathbb{N}$ , la FGM n'est pas définie.



# Echantillonnage de familles exponentielles

Que se passerait-il si la distribution à partir de laquelle nous échantillonnons n'était pas normale, mais.....

binomiale

Poisson

géométrique...

Plus généralement : que se passe-t-il si l'échantillon  $X_1, \dots, X_n$  vient d'une certaine famille exponentielle ? Soit  $X_1, \dots, X_n \stackrel{iid}{\sim} f$ , où

$$f(x) = \exp \left\{ \sum_{i=1}^k \phi_i T_i(x) - \gamma(\phi_1, \dots, \phi_k) + S(x) \right\}, \quad x \in \mathcal{X}.$$

- 1 Est-il possible de trouver la distribution conjointe de l'échantillon  $(X_1, \dots, X_n)$  ?
- 2 Est-il possible de trouver les moments exacts de certaines statistiques clés ?
- 3 Est-il possible de trouver la distribution d'échantillonnage exacte de certaines statistiques importantes ?

## Proposition (Echantillonnage d'une famille exponentielle)

Soit  $X_1, \dots, X_n \stackrel{iid}{\sim} f$ , où

$$f(x) = \exp \{ \phi T(x) - \gamma(\phi) + S(x) \}, \quad x \in \mathcal{X}$$

avec  $\phi \in \Phi \subseteq \mathbb{R}$ , est une densité ayant la forme d'une famille exponentielle.

Alors :

- 1 La densité conjointe de  $(X_1, \dots, X_n)$  a la forme d'une famille exponentielle à 1-paramètre, donnée par

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \exp \left\{ \phi \tau(x_1, \dots, x_n) - n\gamma(\phi) + \sum_{i=1}^n S(x_i) \right\}, \quad x_i \in \mathcal{X},$$

où

$$\tau(x_1, \dots, x_n) = \sum_{i=1}^n T(x_i).$$

- 2 Si  $\Phi$  est ouvert, alors  $\gamma$  est infiniment dérivable, et en plus

$$\mathbb{E}[\tau(X_1, \dots, X_n)] = n\gamma'(\phi) < \infty \quad \text{et} \quad \text{Var}[\tau(X_1, \dots, X_n)] = n\gamma''(\phi) < \infty.$$

## Corollaire

*Sous les mêmes conditions,  $\tau$  est exhaustive pour  $\phi$  (si  $\phi = \eta(\theta)$  pour une certaine injection  $\eta(\cdot)$ ), alors il est clair que  $\tau$  est aussi exhaustive pour  $\theta$ .*

# Distributions d'Echantillonnage Approximative

# Distributions d'Echantillonnage Approximative

La distribution d'échantillonnage de la statistique  $\tau(X_1, \dots, X_n)$  **ne peut pas toujours être déterminée exactement** lorsque l'échantillon est tiré d'une famille exponentielle à un paramètre.

Par conséquent  $\longrightarrow$  tenter de l'approximer en supposant que  $n \rightarrow \infty$

Mais il faut définir « la distribution  $F_{\tau(X_1, \dots, X_n)}$  est approximée par une certaine distribution  $G$  »

- 1 Voyons  $F_{\tau(X_1, \dots, X_n)}$  comme séquence indexée par la taille de l'échantillon  $n$ .
- 2 Alors « approximation par  $G$  » doit être formalisée par une forme de convergence de  $F_n$  à  $G$  lorsque  $n \rightarrow \infty$ .

# Convergence en loi (ou Convergence faible)

## Définition (Convergence en loi (ou convergence faible))

Soit  $\{F_n\}_{n \geq 1}$  une séquence de fonctions de répartition et  $G$  une fonction de répartition sur  $\mathbb{R}$ . Nous disons que  $F_n$  converge en loi vers  $G$ , et écrivons  $F_n \xrightarrow{d} G$ , si et seulement si

$$F_n(x) \xrightarrow{n \rightarrow \infty} G(x),$$

pour tout les  $x$  qui sont des points de continuité de  $G$  (i.e. tous les  $x_0$  tels que  $\lim_{x \rightarrow 0} G(x_0 + x) = G(x_0)$ ).

à noter : si  $G$  est continue, la convergence est **uniforme**.

## Exemple (Le maximum de variables aléatoires uniformes)

Soient  $X_1, \dots, X_n \stackrel{iid}{\sim} Unif(0, 1)$ ,  $M_n = \max\{X_1, \dots, X_n\}$ , et  $Q_n = n(1 - M_n)$ .

$$\mathbb{P}[Q_n \leq x] = \mathbb{P}[M_n \geq 1 - x/n] = 1 - \left(1 - \frac{x}{n}\right)^n \xrightarrow{n \rightarrow \infty} 1 - e^{-x}.$$

Notez que la limite est la fonction de répartition d'une variable aléatoire  $Exp(1)$ .

# Convergence en loi : commentaires

- 1 Convergence en loi  $\equiv$  convergence ponctuelle de la séquence de fonctions de répartition, à l'exception qu'il n'est pas nécessaire d'avoir une convergence ponctuelle aux points de discontinuité de la limite.
- 2 Lorsque  $F_n(x) = \mathbb{P}[X_n \leq x]$  pour une séquence de variables aléatoires  $\{X_n\}_{n \geq 1}$  et  $G(x) = \mathbb{P}[Z \leq x]$  pour une autre variable aléatoire  $Z$ , nous allons abuser de la notation et écrire

$$X_n \xrightarrow{d} Z.$$

- 3 Notre but d'approximation de la loi d'échantillonnage se transforme à trouver une variable aléatoire  $Z$  dont la distribution explicite est connue, et telle que

$$\tau_n \xrightarrow{d} Z$$

# Convergence en probabilité

## Définition

Lorsqu'une séquence de variables aléatoires  $\{X_n\}$  est telle que  $\mathbb{P}[|X_n - Y| > \epsilon] \xrightarrow{n \rightarrow \infty} 0$  pour tout  $\epsilon > 0$  et pour une certaine variable aléatoire  $Y$ , nous disons que  $X_n$  converge en probabilité vers  $Y$ , et écrivons  $X_n \xrightarrow{p} Y$ .

- $X_n \xrightarrow{p} Y \implies X_n \xrightarrow{d} Y$
- L'inverse n'est généralement pas vrai.
- Cependant, si  $Y = c \in \mathbb{R}$  est une constante et si  $\{X_n\}_{n \geq 1}$  est une séquence telle que  $X_n \xrightarrow{d} c$ , nous avons :

## Lemme

Soient  $\{X_n\}_{n \geq 1}$  une séquence de variables aléatoires prenant des valeurs dans  $\mathbb{R}$ , et  $c \in \mathbb{R}$  une certaine constante, alors

$$X_n \xrightarrow{d} c \iff \mathbb{P}[|X_n - c| > \epsilon] \xrightarrow{n \rightarrow \infty} 0, \quad \forall \epsilon > 0.$$

La preuve est laissée en exercice.

# Distributions d'Echantillonnage Approximative

La statistique exhaustive pour un échantillon iid  $X_1, \dots, X_n$  tiré d'une famille exponentielle à un paramètre

$$f(x) = \exp\{\phi T(x) - \gamma(\phi) + S(x)\}$$

est de la forme  $\tau(X_1, \dots, X_n) = \sum_{i=1}^n T(X_i)$ , où

$$\mathbb{E}[\tau(X_1, \dots, X_n)] = n\gamma'(\phi) < \infty \quad \text{et} \quad \text{Var}[\tau(X_1, \dots, X_n)] = n\gamma''(\phi) < \infty.$$

Définissons

$$\overline{T}_n = \frac{1}{n} \tau(X_1, \dots, X_n) = \frac{1}{n} \sum_{i=1}^n T(X_i)$$

alors nous remarquons que nous sommes en présence d'une variables aléatoire qui :

- est en fait la moyenne de  $n$  variables aléatoires iid,
- qui a une moyenne finie  $\gamma'(\phi)$  et une variance finie  $\gamma''(\phi)/n$ .

Comment approximer la loi de  $\overline{T}_n$  au cas général ?

# Les Deux Grands Théorèmes

## Théorème (Loi faible des grands nombres)

Soit  $Y_1, \dots, Y_n$  des variables aléatoires iid telles que  $\mathbb{E}[Y_i] = \mu < \infty$  et  $\text{Var}[Y_i] = \sigma^2 < \infty$ . Soit  $\bar{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i$ , alors

$$\bar{Y}_n \xrightarrow{p} \mu.$$

En fait, la même conclusion reste vraie lorsque nous imposons une condition plus faible que celle de la variance finie, i.e. que  $\mathbb{E}|X_i| < \infty$ .

## Théorème (Théorème central limite)

Soit  $Y_1, \dots, Y_n$  des variables aléatoires i.i.d. telles que  $\mathbb{E}[Y_i] = \mu < \infty$  and  $\text{Var}[Y_i] = \sigma^2 < \infty$  et soit  $\bar{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i$ , alors

$$\sqrt{n}(\bar{Y}_n - \mu) \xrightarrow{d} N(0, \sigma^2).$$

# Distribution d'Echantillonnage Approximative pour Familles Exponentielles

## Corollaire

Soit  $X_1, \dots, X_n \stackrel{iid}{\sim} f$ , où

$$f(x) = \exp\{\phi T(x) - \gamma(\phi) + S(x)\}, \quad x \in \mathcal{X}$$

avec  $\phi \in \Phi \subseteq \mathbb{R}$  et soit

$$\overline{T}_n = \frac{1}{n} \sum_{i=1}^n T(X_i) = n^{-1} \tau(X_1, \dots, X_n).$$

Si  $\Phi$  est ouvert et  $\gamma$  est doublement différentiable, alors

$$\sqrt{n}(\overline{T}_n - \gamma'(\phi)) \xrightarrow{d} N(0, \gamma''(\phi)).$$

# Distributions approximatives pour les fonctions de sommes

## Théorème (Théorème de Slutsky)

Soit  $X$  une variable aléatoire telle que  $\mathbb{P}[X \in \mathcal{X}] = 1$ , et  $g : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$  une fonction continue en  $\mathcal{X} \times c$ , où  $c \in \mathbb{R}$ . Si  $X_n \xrightarrow{d} X$  et  $Y_n \xrightarrow{p} c$ , alors,  $g(X_n, Y_n) \xrightarrow{d} g(X, c)$  lorsque  $n \rightarrow \infty$ .

## Remarque (Théorème de l'application continue)

Notez un cas spécial important : si  $X$  est une variable aléatoire telle que  $\mathbb{P}[X \in \mathcal{X}] = 1$ , et  $g : \mathbb{R} \rightarrow \mathbb{R}$  est continue en  $\mathcal{X}$ , alors

$$X_n \xrightarrow{d} X \implies g(X_n) \xrightarrow{d} g(X).$$

## Théorème (La méthode delta)

Soit  $Z_n := a_n(X_n - \theta) \xrightarrow{d} Z$  où  $a_n, \theta \in \mathbb{R}$  pour tout  $n$  et  $a_n \uparrow \infty$ . Soit  $g : \mathbb{R} \rightarrow \mathbb{R}$  dérivable en  $\theta$ , alors  $a_n(g(X_n) - g(\theta)) \xrightarrow{d} g'(\theta)Z$ , lorsque  $g'(\theta) \neq 0$ .

# Nouveaux théorèmes limites partir des plus vieux

**ATTENTION :** On ne peut pas remplacer la constante déterministe  $c$  avec une variable aléatoire  $Y$  dans le théorème de Slutsky.

Le théorème central limite nous dit que si  $Y_1, \dots, Y_n$  sont des variables aléatoires iid de moyennes  $\mu$  et de variances  $\sigma^2 < \infty$ , alors  $\sqrt{n}(\bar{Y}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$ .

- ❶ Grâce à la méthode delta, nous obtenons de plus que

$$\sqrt{n}(g(\bar{Y}_n) - g(\mu)) \xrightarrow{d} N(0, \sigma^2[g'(\mu)]^2),$$

pour toutes les fonctions continues et dérivables  $g$ .

- ❷ Maintenant considérons  $W_n$  une séquence de variables aléatoires telle que  $W_n \xrightarrow{p} \sigma$ . Il est facile d'utiliser le théorème de Slutsky afin de conclure que

$$\sqrt{n} \left( \frac{g(\bar{Y}_n) - g(\mu)}{W_n} \right) \xrightarrow{d} N(0, [g'(\mu)]^2).$$

# Estimation ponctuelle

# Le problème d'estimation dans notre cadre générale

- 1 Il y a une **distribution**  $F(x; \theta)$  qui dépend d'un **paramètre inconnu**  $\theta \in \mathbb{R}^p$ .
- 2 Nous **observons la réalisation de  $n$  variables aléatoires**  $X_1, \dots, X_n$ , indépendantes et identiquement distribuées, qui suivent cette distribution. Mais **nous ne connaissons toujours pas la vraie valeur de  $\theta$  qui a généré les  $X_i$  !**
- 3 **Problème d'estimation ponctuelle** : Comment utiliser les  $n$  observations (les réalisations de  $X_1, \dots, X_n$ ) afin de **déterminer la vraie valeur de  $\theta$** .

Comment ? Mais avec un estimateur, bien-sur !

## Définition (Estimateur ponctuel)

*Une statistique prenant des valeurs dans  $\Theta$  est appelée un estimateur ponctuel. Réciproquement, un estimateur ponctuel est une statistique  $T : \mathcal{X}^n \rightarrow \Theta$ .*

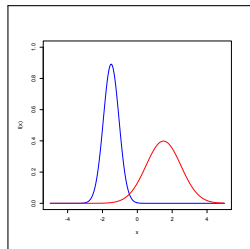
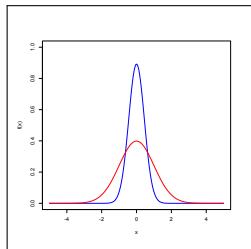
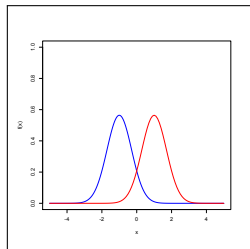
## Remarque

*Puisque l'objectif d'un estimateur est de fournir une estimation du vrai  $\theta$  qui a généré les données, nous le dénotons typiquement  $\hat{\theta}$ . Notez de plus que  $\theta$  est un paramètre déterministe tandis que  $\hat{\theta}$  est une variable aléatoire.*

# Mais... quel estimateur ?

- N'importe quelle fonction dont l'image est incluse dans  $\Theta$  pourrait être un estimateur.
- Laquelle devons-nous choisir ?

# Critères pour comparer des estimateurs



# Critères pour comparer des estimateurs

Il y a plusieurs critères différents que l'on peut utiliser, mais les statisticiens considèrent typiquement deux caractérisations de base de la concentration : la **moyenne et la variance de  $\hat{\theta}$** .

*Pourquoi ?*

- 1 Interprétation facile.
- 2 Théorème centrale limite.
- 3 Inégalités de concentration

**Il s'avère que l'erreur quadratique moyenne prend en compte la moyenne et la variance en même temps.**

# Erreur quadratique moyenne

## Définition (Erreur quadratique moyenne)

Soit  $\hat{\theta}$  un estimateur du paramètre  $\theta$  d'un modèle paramétrique  $\{F_\theta : \theta \in \Theta\}$ ,  $\Theta \subseteq \mathbb{R}$ . L'Erreur Quadratique Moyenne (EQM) de  $\hat{\theta}$  est définie comme suit

$$EQM(\hat{\theta}, \theta) = \mathbb{E} \left[ (\hat{\theta} - \theta)^2 \right].$$

## Lemme (Décomposition biais-variance)

L'erreur quadratique moyenne d'un estimateur admet la décomposition

$$EQM(\hat{\theta}, \theta) = \left( \mathbb{E}[\hat{\theta}] - \theta \right)^2 + \mathbb{E} \left[ (\hat{\theta} - \mathbb{E}[\hat{\theta}])^2 \right] = \text{biais}^2(\hat{\theta}, \theta) + \text{Var}[\hat{\theta}].$$



## Lemme

Soit  $\hat{\theta}$  un estimateur de  $\theta \in \mathbb{R}^p$  tel que  $\text{Var}[\hat{\theta}] < \infty$ . Alors, pour tout  $\epsilon > 0$ ,

$$\mathbb{P}[\|\hat{\theta} - \theta\| > \epsilon] \leq \frac{EQM(\hat{\theta}, \theta)}{\epsilon^2}$$

- Notez que  $EQM(\hat{\theta}_n, \theta) \xrightarrow{n \rightarrow \infty} 0 \implies \hat{\theta}_n \xrightarrow{p} \theta$ .
- Lorsqu'un estimateur possède une telle propriété, nous disons que cet estimateur est consistant.

## Définition (Consistance)

Un estimateur  $\hat{\theta}_n$  de  $\theta$ , construit à l'aide d'un échantillon de taille  $n$ , est consistant si  $\hat{\theta}_n \xrightarrow{p} \theta$  lorsque  $n \rightarrow \infty$ .

## Remarque

Notez que la convergence de l'EQM vers zéro implique la consistance, mais que la réciproque est généralement fausse.

# Limitations sur la précision ?

- Nous pouvons utiliser l'erreur quadratique moyenne afin de comparer deux estimateurs, et ainsi obtenir une idée de leur performance
- Mais y-a t'il une *meilleure erreur quadratique moyenne réalisable* pour un problème donné ?
- Ce problème est très difficile, car il est équivalent au problème consistant à trouver un estimateur uniformément optimal : un estimateur  $T_*$  tel que

$$EQM(T_*, \theta) \leq EQM(T, \theta)$$

pour tout  $\theta \in \Theta$  et pour tous les estimateurs  $T$ .

- Pour apprécier la difficulté du problème, supposons que  $\Theta = \mathbb{R}$  et considérons l'estimateur  $S(X_1, \dots, X_n) = 0$  :
  - C'est un estimateur ridicule, car il n'utilise pas les données, mais quand même quand la vérité est  $\theta = 0$ , alors  $EQM(S, 0) < EQM(T, \theta)$  pour tout  $T$  – aucun autre estimateur peut battre  $S$  a cet endroit de l'espace  $\Theta$ .
  - Même une montre cassée donne l'heure exacte deux fois par jour...

## Théorème (Borne de Cramér-Rao)

Soit  $X_1, \dots, X_n$  un échantillon iid tiré d'un modèle paramétrique régulier  $f(\cdot; \theta)$ ,  $\Theta \subseteq \mathbb{R}$  et soit  $T : \mathcal{X}^n \rightarrow \Theta$  un estimateur de  $\theta$ , pour tout  $n$ . Supposons que :

❶  $\text{Var}(T) < \infty$ , pour tout  $\theta \in \Theta$ .

❷  $\frac{\partial}{\partial \theta} \left[ \int_{\mathcal{X}^n} T(x_1, \dots, x_n) f_{X_1, \dots, X_n}(x_1, \dots, x_n; \theta) dx \right] = \int_{\mathcal{X}^n} T(x_1, \dots, x_n) \frac{\partial}{\partial \theta} f_{X_1, \dots, X_n}(x_1, \dots, x_n; \theta) dx$ .

❸  $\frac{\partial}{\partial \theta} \left[ \int_{\mathcal{X}^n} f_{X_1, \dots, X_n}(x_1, \dots, x_n; \theta) dx \right] = \int_{\mathcal{X}^n} \frac{\partial}{\partial \theta} f_{X_1, \dots, X_n}(x_1, \dots, x_n; \theta) dx$ .

Si nous dénotons le biais de  $T$  par  $\beta(\theta) = \mathbb{E}(T) - \theta$ , alors  $\beta(\theta)$  est dérivable et

$$\text{Var}(T) \geq \frac{(\beta'(\theta) + 1)^2}{n \int_{\mathcal{X}^n} \left( \frac{\partial}{\partial \theta} \log f(x; \theta) \right)^2 f(x; \theta) dx}.$$

- La condition 1. n'est en réalité pas nécessaire, mais si elle n'est pas vérifiée le théorème ne nous apprend pas grand chose...
- On appelle la quantité positive  $\int_{\mathcal{X}} \left( \frac{\partial}{\partial \theta} \log f(x; \theta) \right)^2 f(x; \theta) dx = \mathbb{E} \left( \frac{\partial}{\partial \theta} \log f(X_1; \theta) \right)^2$  l'information de Fisher,  $I(\theta)$ .
- Même si le biais est égal à zéro, la variance sera bornée inférieurement par  $1/[n \times I(\theta)]$ .

# La méthode du maximum de vraisemblance

# Motivation

La statistique comme “probabilité inverse”. → Considerons le cas discret.

## Point de vue Probabilités

Si on se dispose d'un paramètre  $\theta \in \Theta$ , alors pour tout  $(x_1, \dots, x_n) \in \mathcal{X}^n$ , on peut évaluer

$$(x_1, \dots, x_n) \mapsto \mathbb{P}_\theta[X_1 = x_1, \dots, X_n = x_n]$$

c'est à dire, comment se varie la probabilité comme fonction de l'échantillon (=du résultat).

## Point de vue Statistiques

Si on se dispose d'un échantillon  $(x_1, \dots, x_n) \in \mathcal{X}^n$ , alors pour tout  $\theta \in \Theta$  on peut évaluer

$$\theta \mapsto \mathbb{P}_\theta[X_1 = x_1, \dots, X_n = x_n]$$

c'est à dire, comment se varie la probabilité comme fonction du paramètre (=du modèle).

Intuition : on imagine que les  $\theta$  plausibles à partir du connaissance de l'échantillon sont ceux qui rendent notre échantillon assez probable...

# Maximum de vraisemblance : cas discret

## Définition (La vraisemblance pour une collection discrète iid)

Soit  $X_1, \dots, X_n$  une collection de variables aléatoires discrètes, indépendantes et identiquement distribuées de fonction de masse  $f(x; \theta)$ , où  $\theta \in \mathbb{R}^p$ . La vraisemblance de  $\theta$  est définie par

$$L : \Theta \rightarrow [0, 1]$$

$$L(\theta) = \prod_{i=1}^n f(X_i; \theta).$$

Remarques :

- 1 La vraisemblance est une fonction aléatoire
- 2 La vraisemblance est, en effet, la fonction  $\prod_{i=1}^n f(X_i; \theta)$  vue comme fonction de  $\theta$
- 3 La vraisemblance **n'est pas "la probabilité de  $\theta$ "**
- 4 La vraisemblance  $L(\theta)$  est la réponse à la question : *quelle est la probabilité de l'échantillon observé lorsque le paramètre est égal à  $\theta$*

# Maximum de vraisemblance : cas discret

Lorsque  $\theta$  est inconnu, il semble que l'estimation la plus adaptée serait une valeur  $\hat{\theta}$  pour laquelle ce que nous observons est le plus probable — une valeur qui serait compatible avec nos observations empiriques

## Définition (Estimateur du maximum de vraisemblance)

Soit  $X_1, \dots, X_n$  un échantillon aléatoire iid tiré d'une distribution  $F_\theta$  de fonction de masse  $f(x; \theta)$  et soit  $\hat{\theta}$  tel que

$$L(\theta) \leq L(\hat{\theta}), \quad \forall \theta \in \Theta.$$

Alors  $\hat{\theta}$  est appelé un estimateur du maximum de vraisemblance (EMV) de  $\theta$ .

- Lorsqu'il existe un unique maximum à la fonction de vraisemblance, nous parlons de l'estimateur du maximum de vraisemblance  $\hat{\theta} = \arg \max_{\theta \in \Theta} L(\theta)$

# Maximum de vraisemblance : cas continu

**Et le cas continu ?** On utilisera la même définition, avec la densité au lieu de la fonction de masse, même si on va perdre l'interprétation en de termes de probabilités !

## Définition (La vraisemblance pour une collection continue iid)

Soit  $X_1, \dots, X_n$  une collection de variables aléatoires continues, indépendantes et identiquement distribuées de fonction de densité  $f(x; \theta)$ , où  $\theta \in \mathbb{R}^p$ . La vraisemblance de  $\theta$  est définie par

$$L : \Theta \rightarrow [0, +\infty)$$

$$L(\theta) = \prod_{i=1}^n f(X_i; \theta).$$

Remarques :

- 1 Notons que maintenant la vraisemblance prend de valeurs dans  $\mathbb{R}_+$  entier.
- 2 Puisque  $F(x + \epsilon/2; \theta) - F(x - \epsilon/2; \theta) \approx \epsilon f(x; \theta)$  lorsque  $\epsilon \downarrow 0$ , nous pouvons voir  $\epsilon^n L(\theta)$  comme étant la probabilité approximative d'un échantillon "proche" à ce que nous avons observé.

## Définition (La vraisemblance pour une collection iid)

Soit  $X_1, \dots, X_n$  une collection de variables aléatoires indépendantes et identiquement distribuées de fonction de densité/masse  $f(x; \theta)$ , où  $\theta \in \mathbb{R}^p$ . La vraisemblance de  $\theta$  est définie par

$$L(\theta) = \prod_{i=1}^n f(X_i; \theta).$$

## Définition (Estimateur du maximum de vraisemblance)

Soit  $X_1, \dots, X_n$  un échantillon aléatoire iid tiré d'une distribution  $F_\theta$  de fonction de densité/masse  $f(x; \theta)$  et soit  $\hat{\theta}$  tel que

$$L(\theta) \leq L(\hat{\theta}), \quad \forall \theta \in \Theta.$$

Alors  $\hat{\theta}$  est appelé un estimateur du maximum de vraisemblance (EMV) de  $\theta$ .

# Détermination de l'EMV

- Notons que l'EMV est défini indirectement, comme l'optimum d'une fonction objective. Alors comment le déterminer ?
- Lorsque la vraisemblance est une fonction dérivable de  $\theta$ , le maximum de la fonction  $L(\theta)$  doit être une solution de l'équation

$$\nabla_{\theta} L(\theta) = 0,$$

- Avant de déclarer qu'une solution  $\hat{\theta}$  de cette équation est un EMV, nous devons d'abord vérifier que c'est bien un maximum (et non un minimum!).
- Si la vraisemblance est deux fois dérivable, ceci peut être fait en vérifiant que

$$-\nabla_{\theta}^2 L(\theta) \big|_{\theta=\hat{\theta}} \succ 0,$$

i.e que  $(-1)$  multiplié par la matrice hessienne est définie positive.

- Lorsque le paramètre est de dimension un, ceci se réduit à vérifier que la seconde dérivée est négative lorsqu'elle est évaluée à la solution de l'équation de vraisemblance.

# Détermination de l'EMV – La logVraisemblance

- Afin de résoudre  $\nabla_{\theta} L(\theta) = 0$ , il faut déterminer la dérivée d'un produit de  $n$  fonctions, ce qui peut être un calcul fastidieux.
- Afin d'éviter ceci, nous nous concentrons habituellement à maximiser la *log-vraisemblance*

$$\ell(\theta) := \log L(\theta)$$

au lieu de la vraisemblance.

- Puisque la fonction log est monotone, la vraisemblance et la log-vraisemblance ont les maximums et les minimums pour les mêmes  $\theta$ .
- L'avantage de la log-vraisemblance est que nous travaillons avec une somme de  $n$  fonctions plutôt qu'un produit,

$$\ell(\theta) = \log \left( \prod_{i=1}^n f(X_i; \theta) \right) = \sum_{i=1}^n \log f(X_i; \theta).$$

- Encore une fois, si la fonction log-vraisemblance est deux fois dérivable, un EMV  $\hat{\theta}$  de  $\theta$  satisfera

$$\nabla_{\theta} \ell(\theta)|_{\theta=\hat{\theta}} = 0 \quad \& \quad - \nabla_{\theta}^2 \ell(\theta)|_{\theta=\hat{\theta}} \succ 0.$$

## Exemple (EMV pour la loi de Bernoulli)

Soit  $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Bern}(p)$  et supposons que nous voulons utiliser la méthode du maximum de vraisemblance afin de construire un estimateur de  $p \in (0, 1)$ . La vraisemblance est :

$$L(p) = \prod_{i=1}^n f(X_i; p) = \prod_{i=1}^n p^{X_i} (1-p)^{1-X_i} = p^{\sum_{i=1}^n X_i} (1-p)^{n - \sum_{i=1}^n X_i}.$$

En prenant le logarithme de chaque côté de l'équation, nous obtenons la fonction de log-vraisemblance

$$\ell(p) = \log p \sum_{i=1}^n X_i + \log(1-p) \left( n - \sum_{i=1}^n X_i \right).$$

Nous pouvons noter que cette fonction est deux fois dérivable par rapport à  $p$  et calculer

$$\frac{d}{dp} \ell(p) = p^{-1} \sum_{i=1}^n X_i - (1-p)^{-1} \left( n - \sum_{i=1}^n X_i \right).$$

## Exemple (EMV pour la loi de Bernoulli, suite)

Résoudre l'équation  $\ell'(p) = 0$  en fonction de  $p$  est équivalent à résoudre

$$p^{-1} \sum_{i=1}^n X_i - (1-p)^{-1} \left( n - \sum_{i=1}^n X_i \right) = 0,$$

et nous pouvons voir que cette dernière équation à un unique racine donnée par  $\frac{1}{n} \sum_{i=1}^n X_i = \overline{X}$ . Appelons cette racine  $\hat{p}$ , nous devons maintenant vérifier qu'elle correspond bien à un maximum. Notez que

$$\frac{d^2}{dp^2} \ell(p) = -p^2 \sum_{i=1}^n X_i - (1-p)^{-2} \left( n - \sum_{i=1}^n X_i \right),$$

et que cette expression est toujours non-positive, car  $0 \leq \sum_{i=1}^n X_i \leq n$  presque sûrement et  $p \in (0, 1)$ . Ainsi

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

est l'unique EMV de  $p$ .



## Exemple (EMV pour la loi exponentielle)

Soit  $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Exp}(\lambda)$  et supposons que nous voulons utiliser la méthode du maximum de vraisemblance afin de construire un estimateur de  $\lambda \in (0, \infty)$ . La vraisemblance est :

$$L(\lambda) = \prod_{i=1}^n f(X_i; \lambda) = \prod_{i=1}^n \lambda e^{-\lambda X_i} = \lambda^n \exp \left\{ -\lambda \sum_{i=1}^n X_i \right\}.$$

En prenant le logarithme de chaque côté de l'équation, nous obtenons la fonction de log-vraisemblance

$$\ell(\lambda) = n \log \lambda - \lambda \sum_{i=1}^n X_i.$$

Nous pouvons noter que cette fonction est deux fois dérivable par rapport à  $\lambda$  et calculer

$$\frac{d}{d\lambda} \ell(\lambda) = n\lambda^{-1} - \sum_{i=1}^n X_i.$$

## Exemple (EMV pour la loi exponentielle, suite)

Résoudre l'équation  $\ell'(\lambda) = 0$  en fonction de  $\lambda$  nous donne l'unique racine

$$\left( \frac{1}{n} \sum_{i=1}^n X_i \right)^{-1} = 1/\overline{X}.$$

Appelons celle-ci  $\hat{\lambda}$ , nous devons maintenant vérifier qu'elle correspond bien à un maximum. Notez que

$$\frac{d^2}{d\lambda^2} \ell(\lambda) = -\frac{n}{\lambda^2}$$

et que cette expression est toujours négative, car  $\lambda > 0$ . Ainsi

$$\hat{\lambda} = \left( \frac{1}{n} \sum_{i=1}^n X_i \right)^{-1} = 1/\overline{X}$$

est l'unique EMV de  $\lambda$ .



## Exemple (EMV pour la loi gaussienne)

Soit  $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$  et supposons que nous voulons utiliser la méthode du maximum de vraisemblance afin de construire un estimateur de  $\theta = (\mu, \sigma^2) \in \mathbb{R} \times (0, \infty)$ . La vraisemblance est :

$$L(\mu, \sigma^2) = \prod_{i=1}^n f(X_i; \mu, \sigma^2) = \left( \frac{1}{\sqrt{2\pi\sigma^2}} \right)^n \exp \left\{ -\frac{\sum_{i=1}^n (X_i - \mu)^2}{2\sigma^2} \right\}.$$

En prenant le logarithme de chaque côté de l'équation,

$$\ell(\mu, \sigma^2) = -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu)^2.$$

Noter que les dérivés secondes par rapport à  $\mu$  et  $\sigma^2$  existent et

$$\begin{aligned} \frac{\partial}{\partial \mu} \ell(\mu, \sigma^2) &= \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu) \\ \frac{\partial}{\partial \sigma^2} \ell(\mu, \sigma^2) &= -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (X_i - \mu)^2. \end{aligned}$$

## Exemple (EMV pour la loi gaussienne, suite)

Résoudre l'équation  $\nabla_{(\mu, \sigma^2)} \ell(\mu, \sigma^2) = 0$  en fonction de  $(\mu, \sigma^2)$  donne un système de deux équations à deux inconnues. L'unique solution de ce système est

$$\left( \bar{X}, n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2 \right).$$

Appelons cette solution  $(\hat{\mu}, \hat{\sigma}^2)$ , nous devons maintenant vérifier qu'elle correspond bien à un maximum. Notez que

$$\frac{\partial^2}{\partial \mu^2} \ell(\mu, \sigma^2) = -\frac{n}{\sigma^2}, \quad \frac{\partial^2}{\partial (\sigma^2)^2} \ell(\mu, \sigma^2) = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^n (X_i - \mu)^2$$

$$\frac{\partial^2}{\partial \mu \partial \sigma^2} \ell(\mu, \sigma^2) = \frac{\partial^2}{\partial \sigma^2 \partial \mu} \ell(\mu, \sigma^2) = -\frac{\sum_{i=1}^n (X_i - \mu)}{\sigma^4} = \frac{n\mu - n\bar{X}}{\sigma^4}.$$

En évaluant ces dérivés secondes en  $(\hat{\mu}, \hat{\sigma}^2)$ , nous obtenons

$$\left. \frac{\partial^2}{\partial \mu^2} \ell(\mu, \sigma^2) \right|_{(\mu, \sigma^2) = (\hat{\mu}, \hat{\sigma}^2)} = -\frac{n}{\hat{\sigma}^2}, \quad \left. \frac{\partial^2}{\partial (\sigma^2)^2} \ell(\mu, \sigma^2) \right|_{(\mu, \sigma^2) = (\hat{\mu}, \hat{\sigma}^2)} = -\frac{n}{2\hat{\sigma}^4}$$

## Example

$$\left. \frac{\partial^2}{\partial \mu \partial \sigma^2} \ell(\mu, \sigma^2) \right|_{(\mu, \sigma^2) = (\hat{\mu}, \hat{\sigma}^2)} = \left. \frac{\partial^2}{\partial \sigma^2 \partial \mu} \ell(\mu, \sigma^2) \right|_{(\mu, \sigma^2) = (\hat{\mu}, \hat{\sigma}^2)} = \frac{n \hat{\mu} - n \hat{\mu}}{\hat{\sigma}^4} = 0.$$

Nous obtenons que la matrice

$$\left[ - \nabla_{(\mu, \sigma^2)}^2 \ell(\mu, \sigma^2) \right]_{(\mu, \sigma^2) = (\hat{\mu}, \hat{\sigma}^2)}$$

est diagonale. Afin de montrer qu'elle est définie positive, il suffit de montrer que les éléments de sa diagonale sont positifs. C'est bien le cas ici, puisque  $\hat{\sigma}^2$  est positif avec probabilité 1. Ainsi l'unique EMV de  $(\mu, \sigma^2)$  est donné par

$$(\hat{\mu}, \hat{\sigma}^2) = \left( \overline{X}, \frac{1}{n} \sum_{i=1}^n (X_i - \overline{X})^2 \right).$$



Notons que l'estimateur EMV de  $\sigma^2$  est biaisé.

# Equivariance de l'EMV

- Il y a des situations où nous ne sommes pas intéressés à estimer  $\theta$ , mais plutôt une transformation  $\phi = g(\theta)$  de celui-ci.
- Si la fonction  $g$  est une bijection, nous n'avons pas besoin de répéter le processus entier d'estimation

## Proposition (Equivariance bijective de l'EMV)

Soit  $\{f(\cdot; \theta) : \theta \in \Theta\}$  un modèle paramétrique où  $\Theta \subseteq \mathbb{R}^p$ . Supposons que  $\hat{\theta}$  soit un EMV de  $\theta$ , sur la base de l'échantillon  $X_1, \dots, X_n$  tiré de  $f(x; \theta)$ . Soit  $g : \Theta \rightarrow \Phi \subseteq \mathbb{R}^p$  une fonction bijective, alors,  $\hat{\phi} = g(\hat{\theta})$  est un EMV de  $\phi = g(\theta)$ .

## Exemple

Soit  $X_1, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, 1)$ , et supposons que nous sommes intéressés par l'estimation de  $\mathbb{P}[X_1 \leq x]$ , pour un  $x \in \mathbb{R}$  donné. Notons que

$$\mathbb{P}[X_1 \leq x] = \mathbb{P}[X_1 - \mu \leq x - \mu] = \Phi(x - \mu),$$

où  $\Phi$  est la fonction de répartition normale standard. La fonction  $\mu \mapsto \Phi(x - \mu)$  est une bijection, car  $\Phi$  est monotone; donc, l'EMV de  $\mathbb{P}[X_1 \leq x]$  est  $\Phi(x - \hat{\mu})$ , où  $\hat{\mu}$  est l'EMV de  $\mu$  (par l'exemple précédent  $\hat{\mu} = \bar{X}$ ). □

# Equivariance de l'EMV

## Exemple (Paramètre usuel vs naturel dans les familles exponentielles)

Soit  $X_1, \dots, X_n \stackrel{iid}{\sim} f$ , avec

$$f(x) = \exp \{ \phi T(x) - \gamma(\phi) + S(x) \}, \quad x \in \mathcal{X}$$

où  $\phi \in \Phi \subseteq \mathbb{R}$  est le paramètre naturel. Supposons maintenant que nous pouvons aussi écrire  $\phi = \eta(\theta)$ , où  $\theta \in \Theta$  est le paramètre usuel et  $\eta : \Theta \rightarrow \Phi$  est une certaine fonction bijective et dérivable (et donc  $\gamma(\phi) = \gamma(\eta(\theta)) = d(\theta)$ , pour  $d = \gamma \circ \eta$ ). Avec cette notation, la fonction de densité/masse de la famille exponentielle prend la forme :

$$\exp \{ \phi T(x) - \gamma(\phi) + S(x) \} = \exp \{ \eta(\theta) T(x) - d(\theta) + S(x) \}.$$

La proposition précédente implique que si  $\hat{\theta}$  est l'EMV de  $\theta$ , alors  $\eta(\hat{\theta})$  est l'EMV de  $\phi = \eta(\theta)$ . La réciproque est elle aussi vraie : si  $\hat{\phi}$  est l'unique EMV de  $\phi$ , alors  $\eta^{-1}(\hat{\phi})$  est l'unique EMV de  $\theta = \eta^{-1}(\phi)$ . □

# EMV dans les familles exponentielles

Ce n'était pas par hasard que l'EMV existait et était unique dans les exemples traités : c'est un phénomène général chez les familles exponentielles.

## Proposition (EMV pour la famille exponentielle à 1-paramètre)

*Soit  $X_1, \dots, X_n$  un échantillon iid tiré d'une distribution dont la fonction de densité/masse appartient à une famille exponentielle à 1-paramètre,*

$$f(x; \phi) = \exp\{\phi T(x) - \gamma(\phi) + S(x)\}, \quad x \in \mathcal{X}, \phi \in \Phi$$

*avec  $T$  une fonction non constante et l'espace des paramètres  $\Phi \subset \mathbb{R}$  un ouvert. Alors l'EMV  $\hat{\phi}$  de  $\phi$  est unique lorsqu'il existe, et est donnée par l'unique solution par rapport à  $u$  de l'équation*

$$\gamma'(u) = \overline{T}.$$

*Ici,*

$$\overline{T} = \frac{1}{n} \sum_{i=1}^n T(X_i) = \frac{1}{n} \tau(X_1, \dots, X_n).$$

# Propriétés de l'EMV quand $n \rightarrow \infty$

# Le cas Gaussien

Commençons avec le cas spécifique Gaussien.

## EMV au cas Gaussien

l'estimateur du maximum de vraisemblance pour le paramètre  $(\mu, \sigma^2)$  d'une distribution gaussienne, basé sur un échantillon iid  $X_1, \dots, X_n$ , est

$$(\hat{\mu}_n, \hat{\sigma}_n^2) = \left( \frac{1}{n} \sum_{i=1}^n X_i, \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \right) = \left( \bar{X}_n, \frac{n-1}{n} S_n^2 \right),$$

- L'EMV de  $\mu$ ,  $\hat{\mu}_n$ , est non-biaisé pour tout  $n$ .
- Pour tout  $n$ , sa distribution est normale, avec variance égale à  $\sigma^2/n$ .
- Ainsi, l'erreur quadratique moyenne est exactement  $\sigma^2/n$ , et ce, peu importe la vraie valeur de  $\mu$ .
- Il s'ensuit que  $\hat{\mu}_n$  est un estimateur consistant.

# Le cas Gaussien

- L'EMV de  $\sigma^2$ ,  $\hat{\sigma}_n^2$  est biaisé pour tout  $n$ , son biais étant égal à :

$$\text{biais}(\hat{\sigma}_n^2, \sigma^2) = \mathbb{E}[\hat{\sigma}_n^2] - \sigma^2 = \mathbb{E}\left[\frac{n-1}{n}S^2\right] - \sigma^2 = \frac{n-1}{n}\sigma^2 - \sigma^2 = -\frac{1}{n}\sigma^2.$$

- Ainsi,  $\hat{\sigma}_n^2$  sous-estime  $\sigma^2$ , même si asymptotiquement, le biais se réduit à zéro.
- La distribution de  $\hat{\sigma}_n^2$  est la même que celle d'une variable aléatoire khi carré multipliée par  $\sigma^2/n$ , i.e.

$$\frac{n}{\sigma^2}\hat{\sigma}_n^2 \sim \chi_{n-1}^2.$$

- Par conséquent, l'erreur quadratique moyenne de  $\hat{\sigma}_n^2$  est

$$EQM(\hat{\sigma}_n^2, \sigma^2) = \text{biais}^2(\hat{\sigma}_n^2, \sigma^2) + \text{Var}[\hat{\sigma}_n^2] = \frac{(2n-1)\sigma^4}{n^2}.$$

- Il s'ensuit que  $\hat{\sigma}_n^2$  est un estimateur consistant.

# Le cas général

- Il n'est habituellement pas possible de déterminer de façon exacte la distribution d'échantillonnage de l'EMV. Par contre, nous devons **recourir à des approximations en utilisant la notion de convergence en loi**
- Mais nous avons vu que, pour les familles exponentielles à un-paramètre,  $\overline{T}_n \stackrel{d}{\approx} N(\gamma'(\phi), n^{-1}\gamma''(\phi))$ .
- Alors comme le EMV satisfait  $\gamma'(\hat{\phi}) = \overline{T}_n$ , si la solution de l'équation dépend de  $\overline{T}_n$  de façon « dérivable », alors la méthode delta pourrait être utilisée
- En fait, c'est **exactement le cas !** On utilisera :

## Théorème de la fonction inverse

Soit  $h(x) : \mathbb{R} \rightarrow \mathbb{R}$  une fonction continûment dérivable, avec une dérivée différente de zéro au point  $x_0 \in \mathbb{R}$ . Alors,

- ❶ il existe un  $\epsilon > 0$  tel que  $h^{-1} \in C^1(h(x_0) - \epsilon, h(x_0) + \epsilon)$ .
- ❷  $(h^{-1})'(y) = [h'(h^{-1}(y))]^{-1}$  pour  $|y - h(x_0)| < \epsilon$ .

## Théorème

Soit  $X_1, \dots, X_n$  un échantillon iid tiré d'une distribution dont la fonction de densité/masse  $f(x; \phi_0)$  appartient à une famille exponentielle à 1-paramètre non-dégénérée,

$$f(x; \phi) = \exp\{\phi T(x) - \gamma(\phi) + S(x)\}, \quad x \in \mathcal{X}, \phi \in \Phi.$$

telle que

- 1 L'espace des paramètres  $\Phi \subset \mathbb{R}$  est un ensemble ouvert  
(qui implique que  $\gamma(\cdot)$  est deux fois continûment dérivable).
- 2 La fonction  $T$  n'est pas une constante sur le support de  $f$

Soit  $\hat{\phi}_n$  l'estimateur du maximum de vraisemblance  $\phi_0$ , dont on suppose l'existence, alors

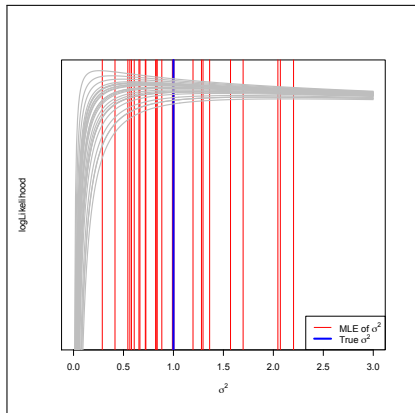
$$0 < \frac{1}{\gamma''(\phi_0)} < \infty \quad \text{et} \quad \sqrt{n}(\hat{\phi}_n - \phi_0) \xrightarrow{d} N\left(0, \frac{1}{\gamma''(\phi_0)}\right).$$

- Pour des grandes valeurs de  $n$ , l'EMV  $\hat{\phi}$  est approximativement  $N(\phi_0, [n\gamma''(\phi_0)]^{-1})$ .
- Biais asymptotique = zéro.
- Et la variance ? Notons que

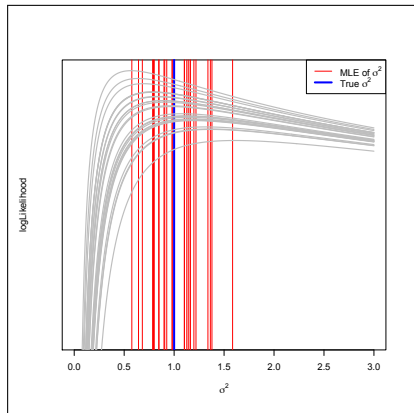
$$\begin{aligned}
 \mathbb{E}[(\ell'(\phi))^2] &= \mathbb{E} \left\{ \left[ \frac{\partial}{\partial \phi} (\phi \tau(X_1, \dots, X_n) - n\gamma(\phi)) \right]^2 \right\} \\
 &= \mathbb{E} [(\tau(X_1, \dots, X_n) - n\gamma'(\phi))^2] \\
 &= \text{Var}[\tau(X_1, \dots, X_n)] \\
 &= n\gamma''(\phi).
 \end{aligned}$$

- L'EMV atteint asymptotiquement la borne de Cramér-Rao ! .
- l'estimateur du maximum de vraisemblance de  $\phi$  a une performance quasiment optimale (pour  $n$  grand !)

# Pourquoi $1/[n\gamma''(\phi)]$ ?

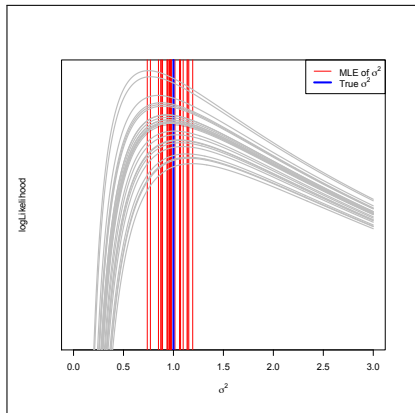


(p) Fonctions de log-vraisemblance pour le paramètre de variance correspondant à 25 réplifications d'un échantillon iid  $N(0,1)$  de taille 10.

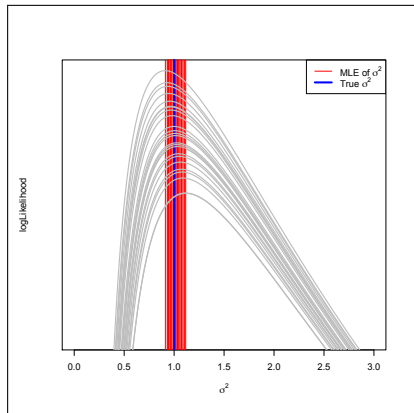


(q) Fonctions de log-vraisemblance pour le paramètre de variance correspondant à 25 réplifications d'un échantillon iid  $N(0,1)$  de taille 50.

# Pourquoi $1/[n\gamma''(\phi)]$ ?



(r) Fonctions de log-vraisemblance pour le paramètre de variance correspondant à 25 réplifications d'un échantillon iid  $N(0,1)$  de taille 150.



(s) Fonctions de log-vraisemblance pour le paramètre de variance correspondant à 25 réplifications d'un échantillon iid  $N(0,1)$  de taille 450.

## Corollaire (Consistance de l'EMV dans les familles exponentielles)

*Dans le même cadre et les mêmes conditions que pour le théorème précédent, nous avons*

$$\hat{\phi}_n \xrightarrow{p} \phi_0, \quad \text{lorsque } n \rightarrow \infty.$$

Et le paramètre usuel ?

# Convergence faible du EMV

## Corollaire

Soit  $X_1, \dots, X_n$  un échantillon iid tiré d'une distribution dont la fonction de densité/masse  $f(x; \theta_0)$  appartient à une famille exponentielle non-dégénérée à 1-pramètre

$$f(x; \theta) = \exp\{\eta(\theta) T(x) - d(\theta) + S(x)\}, \quad x \in \mathcal{X}, \theta \in \Theta.$$

Supposons que

- ❶ L'espace des paramètres  $\Theta \subset \mathbb{R}$  est un ensemble ouvert.
- ❷ La fonction  $\eta(\cdot)$  est une bijection  $C^2$  entre  $\Theta$  et  $\Phi = \eta(\Theta)$ .
- ❸ La fonction  $T$  n'est pas une constante sur le support de  $f$ .

Soit  $\hat{\theta}_n$  l'estimateur du maximum de vraisemblance de  $\theta_0$ , alors

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N\left(0, \frac{[\eta'(\theta_0)]}{d''(\theta_0)\eta'(\theta_0) - d'(\theta_0)\eta''(\theta_0)}\right).$$

**Exercice :** Prouvez ce corollaire.

- Biais asymptotique = zéro.
- Et la variance ? Si  $\phi = \eta(\theta)$  et  $\gamma(\phi) = d(\eta^{-1}(\phi))$ , notons

$$\begin{aligned}
 \mathbb{E}[(\ell'(\theta))^2] &= \mathbb{E} \left[ \left( \frac{\partial \ell(\theta)}{\partial \eta(\theta)} \frac{\partial \eta(\theta)}{\partial \theta} \right)^2 \right] = (\eta'(\theta))^2 \mathbb{E}[(\ell'(\phi))^2] \\
 &= (\eta'(\theta))^2 \text{Var}[\tau(X_1, \dots, X_n)] \\
 &= n(\eta'(\theta))^2 \frac{d''(\theta)\eta'(\theta) - d'(\theta)\eta''(\theta)}{[\eta'(\theta)]^3} \\
 &= n \frac{d''(\theta)\eta'(\theta) - d'(\theta)\eta''(\theta)}{[\eta'(\theta)]},
 \end{aligned}$$

- L'EMV atteint asymptotiquement la borne de Cramér-Rao dans le cas usuel aussi !

# Autres méthodes d'estimation

# Motivation

- Pourquoi utiliser d'autres méthodes si l'EMV est quasiment optimal pour grand  $n$  ?
- Une raison est que, parfois, le EMV n'est pas explicitement disponible.

## Exemple (EMV pour la loi de Cauchy)

Supposons que  $X_1, \dots, X_n$  sont des variables aléatoires iid suivant une *distribution de Cauchy* dont la fonction de densité est

$$f(x; \theta) = \frac{1}{\pi(1 + (x - \theta)^2)}, \quad x \in \mathbb{R}.$$

La fonction de log-vraisemblance dans ce cas est

$\ell(\theta) = -\sum_{i=1}^n \log[1 + (X_i - \theta)^2] - n \log(\pi)$ . L'EMV doit satisfaire  $\ell'(\hat{\theta}) = 0$ , ou de façon équivalente

$$\sum_{i=1}^n \frac{2(X_i - \hat{\theta})}{1 + (X_i - \hat{\theta})^2} = 0.$$

L'équation ci-dessus ne peut pas être résolue explicitement afin de trouver l'EMV.

**Solution numérique ! Point de départ ?**



## Itération de Newton-Raphson

Supposons que nous ayons une valeur initiale  $\hat{\theta}_{(0)}$  qui est près du vrai maximum  $\hat{\theta}$ . Puisque  $\hat{\theta}$  est le maximum global, il satisfait  $\ell'(\hat{\theta}) = 0$ . Supposons maintenant que  $\ell$  soit telle qu'il est possible de faire un développement en série de Taylor. Nous aurions alors :

$$0 = \ell'(\hat{\theta}) = \ell'(\hat{\theta}_{(0)}) + (\hat{\theta} - \hat{\theta}_{(0)})\ell''(\hat{\theta}_{(0)}) + \frac{1}{2}(\hat{\theta} - \hat{\theta}_{(0)})^2\ell'''(\theta_*),$$

où  $\theta_* = \lambda\hat{\theta} + (1 - \lambda)\hat{\theta}_{(0)}$  pour un certain  $\lambda \in [0, 1]$ . En supposant maintenant que  $|\hat{\theta} - \hat{\theta}_{(0)}|$  est petit, nous obtenons que le terme  $(\hat{\theta} - \hat{\theta}_{(0)})^2$  est négligeable par rapport au terme  $(\hat{\theta} - \hat{\theta}_{(0)})$ . Alors, lorsque  $\ell'''$  est bornée, nous pouvons écrire

$$\ell'(\hat{\theta}_{(0)}) + (\hat{\theta} - \hat{\theta}_{(0)})\ell''(\hat{\theta}_{(0)}) \simeq 0,$$

ce qui suggère que

$$\hat{\theta} \simeq \hat{\theta}_{(0)} - \frac{\ell'(\hat{\theta}_{(0)})}{\ell''(\hat{\theta}_{(0)})}.$$

La procédure peut maintenant être itérée en définissant  $\hat{\theta}_{(1)} = \hat{\theta}_{(0)} - \frac{\ell'(\hat{\theta}_{(0)})}{\ell''(\hat{\theta}_{(0)})}$ , ...

Comment peut-on trouver une valeur initiale  $\hat{\theta}_{(0)}$  raisonnable ?

### Exemple (EMV pour la loi de Cauchy, suite)

Notez que la densité  $f(x; \theta)$  est symétrique par rapport à  $\theta$ ,

$$f(x; \theta) = \frac{1}{\pi(1 + (x - \theta)^2)}, \quad x \in \mathbb{R}.$$

Une valeur initiale potentielle pour  $\theta$  est donc la médiane de  $X_1, \dots, X_n$ , celle-ci peut être utilisée afin d'initialiser une itération de Newton-Raphson.  $\square$

Dans d'autres cas, les choses peuvent ne pas être si claires.

## Exemple (EMV de la distribution gamma)

Soit  $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Gamma}(r, 1)$  et supposons que nous voulons estimer le paramètre  $r$  par la méthode du maximum de vraisemblance. La vraisemblance est

$$L(r) = \prod_{i=1}^n \frac{1}{\Gamma(r)} X_i^{r-1} e^{-X_i},$$

avec la log-vraisemblance correspondante

$$\ell(r) = -n \log \Gamma(r) + (r-1) \sum_{i=1}^n \log X_i - \sum_{i=1}^n X_i.$$

En dérivant et en posant l'expression obtenue égale à zéro, nous obtenons que l'EMV  $\hat{r}$  doit satisfaire

$$\frac{\Gamma'(\hat{r})}{\Gamma(\hat{r})} = \frac{1}{n} \sum_{i=1}^n \log X_i.$$

Cette équation ne peut pas être résolue explicitement. Pire encore, il n'y a pas de valeur plausible immédiate pour  $r$  lorsqu'on examine la forme de la densité.  $\square$

# Méthode des moments

Motivation :

- 1 Trouver un estimateur qu'on peut déterminer explicitement.
- 2 L'estimateur doit être assez bon (proche à  $\theta$ ) mais pas nécessairement optimal.

## Heuristique des moments

- 1 Soient  $X_1, \dots, X_n \stackrel{iid}{\sim} f_{\theta_0}$  et supposons que  $\mathbb{E}|X_1| < \infty$ .
- 2 LGN  $\implies \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{p} \mathbb{E}[X_1] = \int_{-\infty}^{+\infty} xf(x; \theta_0) dx = m(\theta_0)$
- 3 En d'autres mots :  $\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{p} m(\theta_0)$
- 4 Alors pour grand  $n$  on aura  $\frac{1}{n} \sum_{i=1}^n X_i \simeq m(\theta_0)$
- 5 Alors si  $\hat{\theta}$  est près de  $\theta$ , nous nous attendons à ce qu'il satisfasse

$$\frac{1}{n} \sum_{i=1}^n X_i \simeq m(\hat{\theta}).$$

## Définition (la méthode des moments - Cas pour un seul paramètre)

Soit  $X_1, \dots, X_n$  un échantillon aléatoire iid tiré d'une distribution  $F_\theta$  de fonction de densité/masse  $f(x; \theta)$ . Supposons que  $\mathbb{E}|X_1| < \infty$  pour tout  $\theta \in \Theta \subseteq \mathbb{R}$ . Soit  $\hat{\theta}$  tel que

$$\frac{1}{n} \sum_{i=1}^n X_i = m(\hat{\theta}),$$

où

$$m(\theta) = \int_{-\infty}^{+\infty} xf(x; \theta) dx, \quad \theta \in \mathbb{R}.$$

Alors  $\hat{\theta}$  est appelé l'estimateur par la méthode des moments (MoM) de  $\theta$ .

# Méthode des moments - Commentaires

- La méthode des moments dit que nous devons poser le premier moment théorique égal au premier moment empirique observé.
- Ceci nous donne une équation dont l'inconnue est le paramètre à estimer ; en résolvant cette équation par rapport à cet inconnue, nous obtenons un estimateur de  $\theta$ , qui est l'estimateur par la *méthode des moments*.
- Cette équation est habituellement plus facile à résoudre que l'équation obtenue en posant la dérivée de la log-vraisemblance égale à zéro, car la plutôt que d'avoir une équation de la forme

$$g(X_1, \dots, X_n, \theta) = 0,$$

nous avons un problème généralement plus facile de la forme

$$g(\theta) = h(X_1, \dots, X_n).$$

(séparation des variables)

## Exemple (Estimateur par la MoM pour la loi uniforme)

Soit  $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Unif}(0, \theta)$ , et supposons que nous voulons estimer  $\theta \in \mathbb{R}_+$ . Dans ce cas, nous avons qu'un seul paramètre, alors l'estimateur par la MoM de  $\theta$ , disons  $\hat{\theta}$ , doit être tel que

$$\frac{1}{n} \sum_{i=1}^n X_i = m(\hat{\theta}).$$

Dans ce cas,

$$m(\theta) = \int_0^\theta \frac{x}{\theta} dx = \frac{\theta}{2}.$$

Ainsi, l'estimateur par la méthode des moments est

$$\hat{\theta} = \frac{2}{n} \sum_{i=1}^n X_i.$$

Comparez avec l'EVM qui est égal à  $X_{(n)}$ . □

## Définition (Méthode des moments — Cas pour plusieurs paramètres)

Soit  $X_1, \dots, X_n$  un échantillon aléatoire iid tiré d'une distribution  $F_\theta$  de fonction de densité/masse  $f(x; \theta)$ . Supposons que  $\mathbb{E}|X_1|^p < \infty$ , pour tout  $\theta \in \Theta \subseteq \mathbb{R}^p$ . Soit  $\hat{\theta}$  tel que

$$\frac{1}{n} \sum_{i=1}^n X_i^k = m_k(\hat{\theta}), \quad k = 1, \dots, p$$

où

$$m_k(\theta) = \int_{-\infty}^{+\infty} x^k f(x; \theta) dx, \quad \theta \in \mathbb{R}^p, \quad k = 1, \dots, p.$$

Alors  $\hat{\theta}$  est appelé l'estimateur par la méthode des moments (MoM) de  $\theta$ .

## Example (Estimateur par la MoM pour la loi gamma)

Supposons que  $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Gamma}(r, \lambda)$  et que nous voulons estimer le vecteur  $(r, \lambda)^\top$ . Les équations des deux premiers moments sont :

$$\frac{1}{n} \sum_{i=1}^n X_i = m_1(\hat{r}, \hat{\lambda}) \quad \text{et} \quad \frac{1}{n} \sum_{i=1}^n X_i^2 = m_2(\hat{r}, \hat{\lambda}).$$

De plus, nous avons vu que

$$m_1(r, \lambda) = r/\lambda \quad \text{et}$$

$$m_2(r, \lambda) = \mathbb{E}^2[X_1] + \text{Var}[X_1] = r^2/\lambda^2 + r/\lambda^2 = r(r+1)/\lambda^2.$$

En résolvant le système des équations des moments par rapport aux paramètres inconnus, nous obtenons les estimateurs

$$\hat{r} = \frac{n \bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \quad \text{et} \quad \hat{\lambda} = \frac{n \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$



Inconvénient de la méthode de moments  $\longrightarrow$  il n'est pas garanti qu'elle fonctionne tout le temps...

...pour un problème à  $p$  paramètres, nous avons besoin de l'existence d'un  $p^{\text{e}}$  moment absolu !

### Exemple (L'échec de la MoM dans le cas de la loi de Cauchy)

Soit  $X_1, \dots, X_n$  des variables aléatoires iid suivant une *distribution de Cauchy* avec fonction de densité

$$f(x; \theta) = \frac{1}{\pi(1 + (x - \theta)^2)}, \quad x \in \mathbb{R}.$$

Notez que

$$m_1(0) = \frac{1}{\pi} \int_{-\infty}^{+\infty} \frac{x}{1 + x^2} dx = \infty.$$

Ainsi les équations des moments ne sont pas définies et la méthode des moments ne fonctionne donc pas. □

En général : lorsque la fonction génératrice des moments existe, alors la méthode des moments est bien définie.

# Parenthèse

# Quantiles

**Question** : étant donné  $x \in \mathbb{R}$ , quelle est la probabilité  $\mathbb{P}[X \leq x]$  que  $X$  soit plus petit ou égal à  $x$  ? **Réponse** : fonction de répartition

**Question opposée** :

étant donnée une probabilité  $\alpha \in (0, 1)$ , quel est le  $x \in \mathbb{R}$  tel que  $\mathbb{P}[X \leq x] = \alpha$  ?

**Réponse** : Souvent pas unique – motive la définition des *quantiles*.

## Définition (Fonction quantile et quantiles)

Soient  $X$  une variable aléatoire prenant des valeurs dans  $\mathcal{X} \subseteq \mathbb{R}$ , et  $F_X$  sa fonction de répartition. Nous définissons la fonction quantile de  $X$  comme étant la fonction

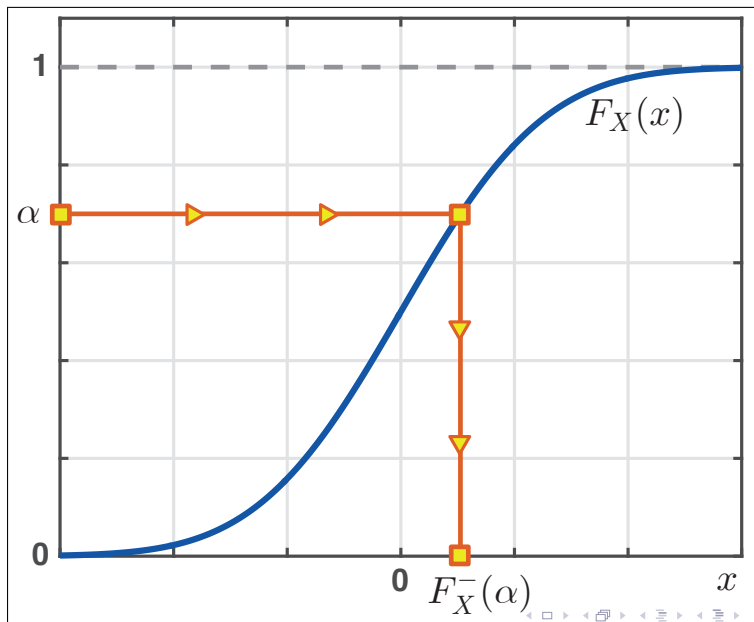
$$F_X^- : (0, 1) \rightarrow \mathbb{R} \qquad F_X^-(\alpha) = \inf\{t \in \mathbb{R} : F_X(t) \geq \alpha\}.$$

Pour une valeur de  $\alpha \in (0, 1)$  donné, nous appelons le nombre réel

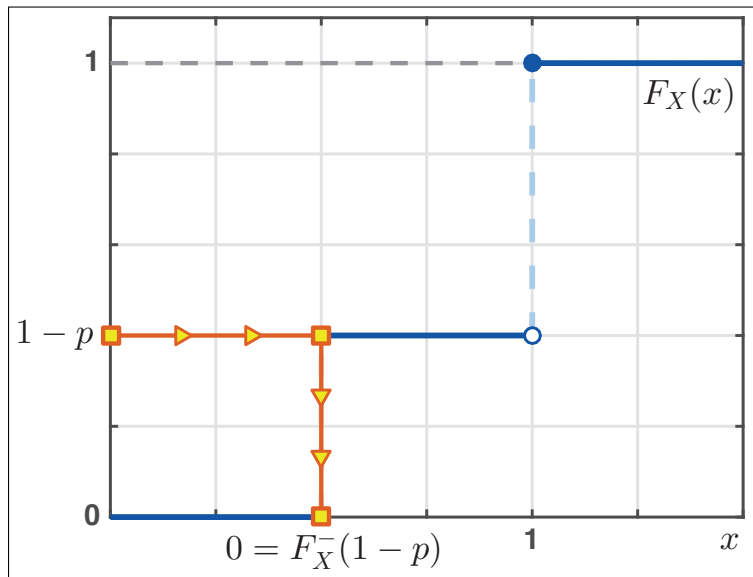
$$q_\alpha = F_X^-(\alpha)$$

le  $\alpha$ -quantile de  $X$  (ou, de façon équivalente, de  $F_X$ ).

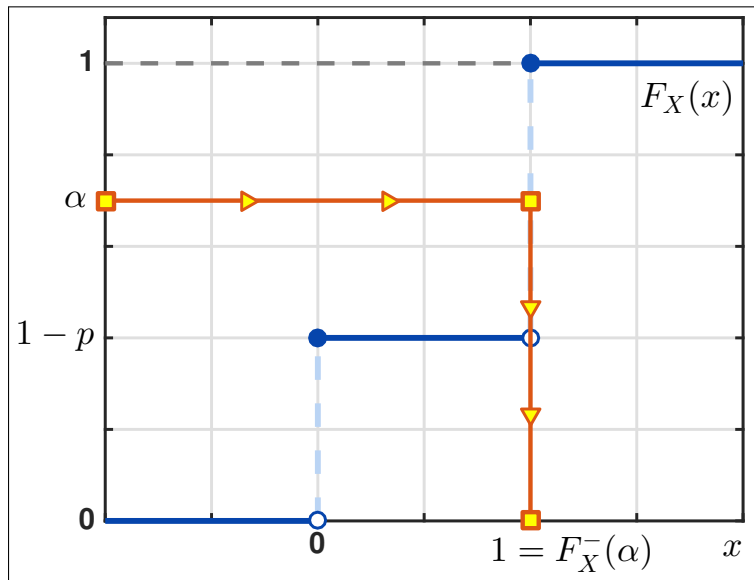
# Quantiles



# Quantiles



# Quantiles



# Tests d'hypothèse

# Le problème d'estimation dans notre cadre générale

- 1 Il y a une **distribution  $F(x; \theta)$**  qui dépend d'un **paramètre inconnu  $\theta \in \mathbb{R}^p$** .
- 2 Nous **observons la réalisation de  $n$  variables aléatoires  $X_1, \dots, X_n$** , indépendantes et identiquement distribuées, qui suivent cette distribution. Mais **nous ne connaissons toujours pas la vraie valeur de  $\theta$  qui a généré les  $X_i$  !**
- 3 **Problème de tests d'hypothèse** : Comment utiliser les  $n$  observations (les réalisations de  $X_1, \dots, X_n$ ) afin de **décider si  $\theta \in \Theta_0$  ou  $\theta \in \Theta_1$**  pour  $\Theta_0 \cap \Theta_1$  donnés.

Au lieu d'estimer la valeur précise du paramètre on s'intéresse plutôt à juger si il fait partie d'un sous-ensemble particulier ou non (par exemple, si il dépasse ou non une certaine borne)

## Exemple (Lancé d'une pièce de monnaie)

- Considérons une situation où nous voulons vérifier si une pièce de monnaie est équilibrée ou biaisée.
- Nous pouvons lancer la pièce  $n$  fois et enregistrer le résultat de chaque lancé.
- Nous souhaitons alors utiliser ces résultats afin de décider si la probabilité d'obtenir « face » est égale à  $1/2$  ou différente de  $1/2$ .
- Nous ne sommes pas vraiment intéressés à savoir la valeur exacte : au lieu de concentrer nos efforts à déterminer la valeur précise, on veut utiliser l'échantillon de manière efficace pour décider si la pièce est équilibrée ou biaisée.
- Nous pourrions formaliser ce problème en disant que  $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Bern}(p)$  et que nous voulons décider si  $p \in \{\frac{1}{2}\}$  ou  $p \in (0, 1) \setminus \{\frac{1}{2}\}$ .



# Le Cadre

Afin de rendre les choses plus concrètes :

- 1 Nous savons que le paramètre appartient à l'un des deux ensembles suivants :  $\Theta_0$  ou  $\Theta_1$ , avec  $\Theta_0 \cap \Theta_1 = \emptyset$ .
- 2 Nous voulons utiliser l'échantillon  $X_1, \dots, X_n$  que nous avons à disposition afin de décider à quel ensemble il appartient.
- 3 Cette situation se produit très souvent en science lorsqu'il y a deux hypothèses scientifiques concurrentes pour un même phénomène :

- 1 l'*hypothèse nulle*  $H_0$  qui dit que  $\theta \in \Theta_0$ ,

$$H_0 : \theta \in \Theta_0,$$

et

- 2 l'*hypothèse alternative* qui postule plutôt que  $\theta \in \Theta_1$ ,

$$H_1 : \theta \in \Theta_1.$$

## Exemple (Recherche du boson de Higgs)

- Une des plus grandes questions du dernier quart de siècle en physique : savoir si le fameux *boson de Higgs* existait ou non.
- En utilisant le Modèle standard de la physique des particules, nous pouvons calculer combien de diphotons seraient produits en moyenne s'il n'y avait pas de boson de Higgs. **Appelons ce nombre  $b$ .**
- De façon similaire, nous pouvons calculer combien de diphotons de plus seraient produits en moyenne si le boson de Higgs existait. **Dénotons ce nombre par  $s$ .**
- Par des moyens de caractérisation on sait que les événements correspondant à l'observation de diphotons suivent une distribution de Poisson avec une certaine moyenne, disons  $\mu$ .

Ainsi, l'hypothèse nulle (qui correspond à l'état de la nature si le boson de Higgs n'existait pas) est

$$H_0 : \mu = b,$$

et l'hypothèse alternative concurrente (qui décrit l'état de la nature si le boson de Higgs existait) est

$$H_1 : \mu = b + s.$$

# Fonctions de test

Notre décision sera basée sur l'échantillon, on aura donc :

## Définition (Fonction de test)

Une fonction de test  $\delta$  est n'importe quelle fonction  $\delta : \mathcal{X}^n \rightarrow \{0, 1\}$ .

On obtient 0 ou 1 dépendamment de si l'échantillon satisfait une certaine condition ou non :

$$\delta(X_1, \dots, X_n) = \begin{cases} 1, & \text{si } T(X_1, \dots, X_n) \in C, \\ 0, & \text{si } T(X_1, \dots, X_n) \notin C, \end{cases}$$

où

- $T$  est une statistique appelée *statistique de test* et
- $C$  est un sous-ensemble de l'image de  $T$ , appelé *région critique*.

De façon plus compacte :

$$\delta(X_1, \dots, X_n) = 1_{\{T(X_1, \dots, X_n) \in C\}}.$$

# Fonctions de test et types d'erreur

Notez que  $\delta$  est toujours une variable aléatoire de Bernoulli,

$$\delta = \begin{cases} 1, & \text{avec probabilité } \mathbb{P}[T(X_1, \dots, X_n) \in C], \\ 0, & \text{avec probabilité } \mathbb{P}[T(X_1, \dots, X_n) \notin C]. \end{cases}$$

- Alors une bonne fonction de test doit être telle que sa loi est concentré autour de la bonne décision.
- Est-ce qu'il y a un critère pareil à l'erreur quadratique moyenne pour quantifier cette concentration ?

# Types d'erreur

- Dans les tests d'hypothèse, il y a deux états possibles de la nature, et deux décisions possibles que l'on peut prendre.
- Ainsi, les erreurs qui peuvent être commises sont données par le tableau suivante :

Décision / Vérité	$H_0$	$H_1$
0	Pas d'erreur	<b>Erreur de type II</b>
1	<b>Erreur de type I</b>	Pas d'erreur

- Ainsi une bonne règle de décision devrait être concentrée autour de  $i$ , lorsque  $H_i$  est vraie, pour  $i \in \{0, 1\}$ .

# Types d'erreur

- Par un léger abus de notation, nous pouvons considerer une sorte de « erreur quadratique moyenne »,

$$EQM(\delta, H_i) = \mathbb{E}_\theta[(\delta - i)^2], \quad i \in \{0, 1\}.$$

- Puisque  $\delta$  est une variable de Bernoulli et que  $i$  prend des valeurs dans  $\{0, 1\}$ , nous avons que

$$\begin{aligned} EQM(\delta, H_i) = \mathbb{E}_\theta[(\delta - i)^2] &= \mathbb{E}_\theta[|\delta - i|] = \begin{cases} \mathbb{E}_\theta[\delta], & \text{si } \theta \in \Theta_0, \\ 1 - \mathbb{E}_\theta[\delta], & \text{si } \theta \in \Theta_1. \end{cases} \\ &= \begin{cases} \mathbb{P}_\theta[\delta = 1], & \text{si } \theta \in \Theta_0, \\ 1 - \mathbb{P}_\theta[\delta = 1], & \text{si } \theta \in \Theta_1. \end{cases} \\ &= \begin{cases} \mathbb{P}_\theta[\delta = 1], & \text{si } \theta \in \Theta_0, \\ \mathbb{P}_\theta[\delta = 0], & \text{si } \theta \in \Theta_1. \end{cases} \end{aligned}$$

## Définition (Les probabilités d'erreurs)

Soient  $H_0 : \theta \in \Theta_0$  et  $H_1 : \theta \in \Theta_1$  deux hypothèses à tester. La probabilité de commettre une erreur de type I est définie comme la fonction  $h : \Theta_0 \rightarrow [0, 1]$ ,

$$h(\theta) = \mathbb{P}_\theta[\delta = 1], \quad \theta \in \Theta_0.$$

La probabilité de commettre une erreur de type II est définie comme la fonction  $g : \Theta_1 \rightarrow [0, 1]$ ,

$$g(\theta) = \mathbb{P}_\theta[\delta = 0], \quad \theta \in \Theta_1.$$

## Remarque

Le fait que les deux probabilités d'erreurs soient des fonctions de  $\theta$  nous indique que nos erreurs dépendent du vrai état de la nature : il sera plus facile de distinguer entre  $\Theta_0$  et  $\Theta_1$  pour certains valeurs du vrai  $\theta$  que pour d'autres.

## Remarque (Avertissement sur les probabilités d'erreurs)

Notez que  $h(\theta) \neq 1 - g(\theta)$  puisque les deux fonctions ne sont pas définies sur le même domaine. C'est une erreur commune qu'il faut éviter.

## Remarque (Erreur de type I vs erreur de type II)

- *Dans plusieurs contextes pratiques, les deux hypothèses sont asymétriques : faire une sorte d'erreur est beaucoup plus grave que faire une erreur de l'autre type.*
- *Le type d'erreur le plus sérieux est appelé le type I et l'autre est l'erreur de type II. Ainsi, dans toutes les situations pratiques,  $H_0$  est l'hypothèse dont le rejet erroné (i.e. lorsque  $H_0$  est en fait vraie), est le plus dommageable.*

# Compte-Rendu

- 1 On veut décider entre  $\{H_0 : \theta \in \Theta\}$  et  $\{H_1 : \theta \in \Theta_1\}$  sur la base de  $X_1, \dots, X_n \stackrel{iid}{\sim} f_\theta$ .
- 2 On va utiliser une fonction de test  $\delta(X_1, \dots, X_n) = 1\{T(X_1, \dots, X_n) \in C\}$ , définie à l'aide d'une statistique de test  $T$  et d'une région critique  $C$ .
- 3 Afin de choisir de bonnes fonctions de test, il faut essayer de minimiser les probabilités des deux types d'erreur,

$$h(\theta) = \mathbb{P}_\theta[\delta = 1], \quad \theta \in \Theta_0.$$

$$g(\theta) = \mathbb{P}_\theta[\delta = 0], \quad \theta \in \Theta_1.$$

- 4 Est-il toujours possible de rendre ces deux probabilités petites pour tous les paramètres  $\theta$  contenus dans les ensembles  $\Theta_0$  et  $\Theta_1$  respectivement ?
- 5 **Malheureusement, la réponse est non**

Voici pourquoi : soit  $\delta(X_1, \dots, X_n) = 1\{T(X_1, \dots, X_n) \in C\}$  et supposons que nous voulons diminuer sa probabilité d'erreur de type I,

$$h(\theta) = \mathbb{P}_\theta[\delta = 1], \quad \theta \in \Theta_0,$$

pour tous les  $\theta \in \Theta_0$ .

Pour cela, remplacer  $C$  par un ensemble  $C_* \subset C$ , en obtenant

$$\delta_* = 1\{T(X_1, \dots, X_n) \in C_*\}.$$

Observez que,  $\forall \theta \in \Theta_0$ ,

$$\mathbb{P}_\theta[\delta_* = 1] = \mathbb{P}[T(X_1, \dots, X_n) \in C_*] \leq \mathbb{P}[T(X_1, \dots, X_n) \in C] = \mathbb{P}_\theta[\delta = 1]$$

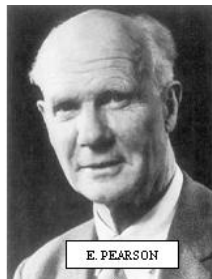
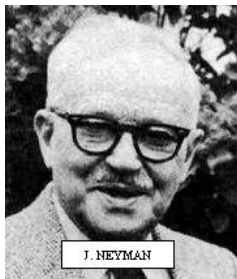
Notez cependant que  $C_* \subset C \implies C_*^c \supset C^c$  et alors  $\forall \theta \in \Theta_1$

$$\mathbb{P}_\theta[\delta_* = 0] = \mathbb{P}[T(X_1, \dots, X_n) \notin C_*] \geq \mathbb{P}[T(X_1, \dots, X_n) \notin C] = \mathbb{P}_\theta[\delta = 0].$$

En essayant de diminuer la probabilité de l'erreur de type I, nous avons augmenté celle de l'erreur de type II !

# Le cadre Neyman-Pearson

# Cadre de Neyman-Pearson



Le paradigme fondamental du *cadre de Neyman-Pearson* est informellement que :

- ➊ puisque l'erreur de type I est la plus importante, nous devons premièrement fixer la probabilité de l'erreur de type I à un certain niveau
- ➋ Une fois ce niveau fixé, nous pouvons nous concentrer sur le problème d'obtenir une petite probabilité de l'erreur de type II

## Définition (Cadre de Neyman-Pearson)

Soient  $H_0 : \theta \in \Theta_0$  et  $H_1 : \theta \in \Theta_1$  deux hypothèses à tester.

- 1 Fixer un  $\alpha \in (0, 1)$  et l'appeler seuil (ou niveau) de signification du test.
- 2 Considérer seulement les  $\delta : \mathcal{X}^n \rightarrow \{0, 1\}$  qui respectent ce seuil,

$$\mathcal{D}(\Theta_0, \alpha) = \left\{ \delta : \mathcal{X}^n \rightarrow \{0, 1\} \mid \sup_{\theta \in \Theta_0} \mathbb{P}_\theta[\delta = 1] \leq \alpha \right\}.$$

- 3 A l'intérieur de la classe  $\mathcal{D}(\Theta_0, \alpha)$ , comparer les fonctions de test en considérant laquelle a la plus petite probabilité d'une erreur de type II

$$g(\theta) = \mathbb{P}_\theta[\delta = 0], \quad \theta \in \Theta_1.$$

De façon équivalente, on compare les fonctions de test en considérant laquelle a la plus grande puissance

$$\beta(\theta) = 1 - g(\theta) = \mathbb{P}_\theta[\delta = 1], \quad \theta \in \Theta_1.$$

# Méthodes pour tester des hypothèses

# Type de méthode $\leftrightarrow$ Type d'hypothèses

La façon de construire des fonctions de test dépend fortement du type d'hypothèse à tester

- 1 **Simple vs simple** ( $H_0 : \theta = \theta_0$ ,  $H_1 : \theta = \theta_1$ , pour un certain  $\theta_0 \neq \theta_1$  donné).
- 2 **Unilatéral gauche vs unilatéral droit** : ( $H_0 : \theta \leq \theta_0$ ,  $H_1 : \theta > \theta_0$ , pour un certain  $\theta_0$  donné).
- 3 **Unilatéral droit vs unilatéral gauche**. ( $H_0 : \theta \geq \theta_0$ ,  $H_1 : \theta < \theta_0$ , pour un certain  $\theta_0$  donné).
- 4 **Simple vs bilatéral** : ( $H_0 : \theta = \theta_0$ ,  $H_1 : \theta \neq \theta_0$ , pour un certain  $\theta_0$  donné).

En résumé,

$$\underbrace{\begin{Bmatrix} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{Bmatrix}}_{\text{simple vs simple}} \text{ ou } \underbrace{\begin{Bmatrix} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{Bmatrix} \text{ ou } \begin{Bmatrix} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{Bmatrix}}_{\text{unilatéral vs unilatéral}} \text{ ou } \underbrace{\begin{Bmatrix} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{Bmatrix}}_{\text{simple vs bilatéral}}$$

# Pourquoi faire cette catégorisation ?

Il s'avère que dans certain cas, il existe une fonction de test optimale

↪ Alors si c'est le cas, on a pas besoin de considérer autre chose !

Spécifiquement :

- (a) **Simple vs simple** : Dans ce cas, nous allons être capable de trouver des **tests optimaux**, et ce, indépendamment du modèle de probabilité sous-jacent.
- (b) **Unilatéral** : Dans ce cas, nous allons être capable de trouver des **tests optimaux pour des classes spécifiques de modèles**, plus spécifiquement pour la famille exponentielle.
- (c) **Bilatéral**. Dans ce cas, nous allons démontrer, qu'en général, il n'existe pas de **tests optimaux**. Nous allons néanmoins proposer deux méthodes générale, inspirée par le concept de vraisemblance.

$$\underbrace{\left\{ \begin{array}{l} H_0 : \theta = \theta_0 \\ H_1 : \theta = \theta_1 \end{array} \right\}}_{\text{simple vs simple}} \text{ ou } \underbrace{\left\{ \begin{array}{l} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{array} \right\}}_{\text{unilatéral}} \text{ ou } \underbrace{\left\{ \begin{array}{l} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{array} \right\}}_{\text{unilatéral}} \text{ ou } \underbrace{\left\{ \begin{array}{l} H_0 : \theta = \theta_0 \\ H_1 : \theta \neq \theta_0 \end{array} \right\}}_{\text{simple vs bilatéral}}$$

# Tests Optimaux

Avant de commencer, il nous reste de définir la notion d'optimalité d'un test.

## Définition (Tests optimaux)

*Une fonction de test  $\delta$  pour  $H_0 : \theta \in \Theta_0$  vs  $H_1 : \theta \in \Theta_1$  est appelée optimale au seuil  $\alpha$  (ou uniformément plus puissante au seuil  $\alpha$ ) si les deux conditions suivantes sont respectées.*

- ❶  $\delta \in \mathcal{D}(\Theta_0, \alpha)$ , c'est à dire,  $\sup_{\theta \in \Theta_0} \mathbb{P}_\theta[\delta = 1] \leq \alpha$ .
- ❷  $\mathbb{P}_{\theta_1}[\psi = 1] \leq \mathbb{P}_{\theta_1}[\delta = 1]$  pour tout  $\theta_1 \in \Theta_1$  et pour tout  $\psi \in \mathcal{D}(\Theta_0, \alpha)$ .

**Observation utile** : Comme  $\delta$  est toujours une variable Bernoulli on a

$$\mathbb{P}_\theta[\delta = 1] = \mathbb{E}_\theta[\delta], \quad \forall \theta \in \Theta_0 \cup \Theta_1.$$

simple vs simple

# simple vs simple : lemme fondamental de Neyman-Pearson

## Lemme (Neyman-Pearson)

Supposons que  $X = (X_1, \dots, X_n)$  a la fonction de densité/masse conjointe  $f_X(\mathbf{x}; \theta)$  et que nous voulons tester

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta = \theta_1,$$

à un certain seuil  $\alpha \in (0, 1)$ , pour  $\theta_0 \neq \theta_1$  donnés. Si la variable aléatoire

$$\Lambda(X) = \frac{f_X(X_1, \dots, X_n; \theta_1)}{f_X(X_1, \dots, X_n; \theta_0)} = \frac{L(\theta_1)}{L(\theta_0)},$$

est telle qu'il existe  $Q > 0$  satisfaisant

$$\mathbb{P}_{\theta_0}[\Lambda > Q] = \alpha,$$

alors le test dont la fonction de test est donnée par

$$\delta(X) = 1\{\Lambda(X) > Q\},$$

est un test optimal de  $H_0$  versus  $H_1$  à au niveau de signification  $\alpha$ .

## Example

Soit  $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Exp}(\lambda)$  et soient  $\lambda_1 > \lambda_0$  deux constantes. Considérons le problème consistant à tester la paire d'hypothèses :

$$\begin{cases} H_0 : & \lambda = \lambda_0 \\ H_1 : & \lambda = \lambda_1. \end{cases}$$

La vraisemblance est

$$f(X_1, \dots, X_n; \lambda) = \prod_{i=1}^n \lambda e^{-\lambda X_i} = \lambda^n e^{-\lambda \sum_{i=1}^n X_i}.$$

Par le lemme de Neyman-Pearson, nous savons que nous devons baser notre test sur la statistique

$$\Lambda(X_1, \dots, X_n) = \frac{f(X_1, \dots, X_n; \lambda_1)}{f(X_1, \dots, X_n; \lambda_0)} = \left( \frac{\lambda_1}{\lambda_0} \right)^n \exp \left[ (\lambda_0 - \lambda_1) \sum_{i=1}^n X_i \right],$$

et rejeter l'hypothèse nulle si  $\Lambda > Q$ , pour  $Q$  tel que  $\mathbb{P}_{\lambda_0}[\Lambda(X_1, \dots, X_n) > Q] = \alpha$ , lorsqu'un tel  $Q$  existe.

## Exemple (suite)

Notons que  $\Lambda(X_1, \dots, X_n)$  est une fonction décroissante de  $\tau(X_1, \dots, X_n) = \sum_{i=1}^n X_i$  (puisque  $\lambda_0 < \lambda_1$ ). Ainsi,

$$\Lambda(X_1, \dots, X_n) > Q \iff \tau(X_1, \dots, X_n) \leq q,$$

pour un certain  $q$ , tel que

$$\alpha = \mathbb{P}_{\lambda_0}[\Lambda > Q] \iff \alpha = \mathbb{P}_{\lambda_0}[\tau(X_1, \dots, X_n) \leq q].$$

Sous la distribution nulle, nous savons que  $\tau(X_1, \dots, X_n)$  suit une distribution gamma de paramètres  $n$  et  $\lambda_0$ .

Ainsi, il existe un  $q$  tel que  $\alpha = \mathbb{P}_{\lambda_0}[\tau(X_1, \dots, X_n) \leq q]$ , et ce  $q$  est donné par le  $q_\alpha$ -quantile de la distribution  $\text{gamma}(n, \lambda_0)$ .

En résumé, le test optimal consiste à rejeter  $H_0$  au seuil  $\alpha$  si la statistique  $\tau(X_1, \dots, X_n)$  est inférieure au  $\alpha$ -quantile d'une distribution  $\text{gamma}(n, \lambda_0)$ .  $\square$

Le test dépend sur la statistique exhaustive! **Ce n'est pas une coïncidence**

## Example (Test simple vs simple pour les familles exponentielles)

Soit  $X_1, \dots, X_n \stackrel{iid}{\sim} f(x; \theta)$ , où

$$f(x; \theta) = \exp\{\eta(\theta)T(x) - d(\theta) + S(x)\}$$

avec  $\eta$  une fonction croissante. Supposons que nous voulons tester  $H_0 : \theta = \theta_0$  contre  $H_1 : \theta = \theta_1$ . Sans perte de généralité, supposons que  $\theta_0 < \theta_1$ .

Le lemme de Neyman-Pearson nous dit que nous devons chercher une statistique de test de la forme

$$\delta = 1\{L(\theta_1)/L(\theta_0) > Q\} = 1\{\log L(\theta_1) - \log L(\theta_0) > \log Q\}.$$

Grâce à la forme de  $f(x; \theta)$  (famille exponentielle), nous obtenons que

$$\begin{aligned}\delta &= 1\left\{(\eta(\theta_1) - \eta(\theta_0)) \sum_{i=1}^n T(X_i) - n(d(\theta_1) - d(\theta_0)) > \log Q\right\} \\ &= 1\left\{\sum_{i=1}^n T(X_i) > \frac{\log Q + n(d(\theta_1) - d(\theta_0))}{\eta(\theta_1) - \eta(\theta_0)}\right\}.\end{aligned}$$

## Exemple (Test simple vs simple pour les familles exponentielles)

Notez que  $\eta(\theta_1) - \eta(\theta_0) > 0$ , puisque  $\eta$  est croissante, et  $n(d(\theta_1) - d(\theta_0))$  est une constante.

Nous pouvons alors simplement écrire

$$\delta = 1\{\tau(X_1, \dots, X_n) > q\}.$$

- 1 Si  $\tau$  est une variable aléatoire continue, alors  $q$  va être le  $(1 - \alpha)$ -quantile de  $G_0(t) = \mathbb{P}_{\theta_0}[\tau(X_1, \dots, X_n) \leq t]$ , i.e. le  $(1 - \alpha)$ -quantile de la distribution d'échantillonnage de  $\tau(X_1, \dots, X_n)$ , lorsque l'on utilise le paramètre  $\theta_0$
- 2 Si nous avons plutôt que  $\eta$  est une fonction décroissante, alors pour  $\theta_0 < \theta_1$ , nous avons que  $\eta(\theta_1) - \eta(\theta_0) < 0$ . Dans ce cas, nous pouvons voir que la statistique de test optimal devient

$$\delta = 1\{\tau(X_1, \dots, X_n) \leq q\}.$$

Cette fois-ci, si  $\tau$  est continue et que nous voulons un test avec un seuil  $\alpha$ ,  $q$  doit être le  $\alpha$ -quantile de  $G_0(t) = \mathbb{P}_{\theta_0}[\tau(X_1, \dots, X_n) \leq t]$ .

## Exemple (Test simple vs simple pour les familles exponentielles)

Nous pouvons observer que la forme du test dépend :

- 1 du comportement de  $\eta$  (si elle est croissante ou décroissante), et
- 2 de si  $\theta_0 < \theta_1$  ou  $\theta_0 > \theta_1$ .

Le tableau suivant résume les formes de statistique de test pour les différents cas possibles.

Dans chaque cas,  $q_s$  représente le  $s$ -quantile de la distribution

$$G_0(t) = \mathbb{P}_{\theta_0}[\tau(X_1, \dots, X_n) \leq t].$$

	$\theta_0 < \theta_1$	$\theta_0 > \theta_1$
$\eta(\cdot)$ croissante	$1\{\tau(X_1, \dots, X_n) > q_{1-\alpha}\}$	$1\{\tau(X_1, \dots, X_n) \leq q_\alpha\}$
$\eta(\cdot)$ décroissante	$1\{\tau(X_1, \dots, X_n) \leq q_\alpha\}$	$1\{\tau(X_1, \dots, X_n) > q_{1-\alpha}\}$

Une observation intéressante est que la fonction de test ne dépend pas de la valeur précise de  $\theta_1$ , mais seulement de si  $\theta_1 < \theta_0$  ou  $\theta_1 > \theta_0$ . □

# Existence d'un test NP pour chaque $\alpha$

Notons que  $G_0(t) = \mathbb{P}_{\theta_0}[\tau(X_1, \dots, X_n) \leq t]$  n'est pas toujours une distribution continue. Ceci signifie qu'il se peut que nous ne soyons pas capable de trouver un test optimal pour tous les  $\alpha$  !

## Exemple

Soit  $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Poisson}(\mu)$  et considérons la paire d'hypothèses

$$H_0 : \mu = \mu_0 \quad vs \quad H_1 : \mu = \mu_1.$$

Notons que c'est la paire d'hypothèses que nous avons vu dans l'exemple du boson de Higgs, si nous posons  $\mu_0 = b$  et  $\mu_1 = b + s$ .

Ceci est un exemple avec une famille exponentielle à 1-paramètre, il donc facile de voir que la statistique exhaustive est

$$\tau(X_1, \dots, X_n) = \sum_{i=1}^n X_i,$$

et que la fonction  $\eta(\cdot)$  est strictement croissante ( $\eta(\cdot) = \log(\cdot)$ ).

## Exemple

Puisque  $\mu_1 > \mu_0$ , nous obtenons, par notre travail, que la statistique de test optimale, dictée par le cadre de Neyman-Pearson, est la suivante :

$$\delta(X_1, \dots, X_n) = 1 \left\{ \sum_{i=1}^n X_i \geq q_{1-\alpha} \right\},$$

lorsqu'il existe un  $q_{1-\alpha}$  tel que  $G_0(q_{1-\alpha}) = \mathbb{P}_{\mu_0}[\tau(X_1, \dots, X_n) \leq q_{1-\alpha}] = 1 - \alpha$ .

Puisque les variables aléatoires  $X_i$  sont indépendantes et qu'elles suivent une loi de Poisson, c'est un exercice simple de montrer que

$$\tau(X_1, \dots, X_n) \stackrel{H_0}{\sim} \text{Poisson}(n\mu_0).$$

Puisque c'est une distribution discrète, les seuls  $\alpha$  pour lesquels ce sera le cas sont

$$e^{-n\mu_0}, e^{-n\mu_0}(1 + n\mu_0), e^{-n\mu_0} \left( 1 + n\mu_0 + \frac{(n\mu_0)^2}{2} \right), \dots \text{ et ainsi de suite}$$

- Cependant, une observation intéressante est que lorsque  $n$  augmente, cette suite de valeurs devient de plus en plus dense *près de l'origine*.

# Existence d'un test NP pour chaque $\alpha$

Même si  $G_0(t) = \mathbb{P}_{\theta_0}[\tau(X_1, \dots, X_n) \leq t]$  n'est pas :

- une distribution continue
- ou n'est pas exactement connue,

on a déjà montré que (sous de conditions),

$$\sqrt{n} \left( n^{-1} \tau(X_1, \dots, X_n) - \frac{d'(\theta)}{\eta'(\theta)} \right) \xrightarrow{d} N \left( 0, \frac{d''(\theta)\eta'(\theta) - d'(\theta)\eta''(\theta)}{[\eta'(\theta)]^3} \right).$$

Cette dernière expression nous suggère d'approximer la distribution  $G_0(t) = \mathbb{P}_{\theta_0}[\tau(X_1, \dots, X_n) \leq t]$  par une distribution

$$N \left( n \frac{d'(\theta_0)}{\eta'(\theta_0)}, n \frac{d''(\theta_0)\eta'(\theta_0) - d'(\theta_0)\eta''(\theta_0)}{[\eta'(\theta_0)]^3} \right),$$

lorsque  $n$  est suffisamment grand.

qui est une loi continue, et donc on peut choisir un  $q$  approximatif pour tout  $\alpha$

# Cas unilatéral

## Théorème (Tests unilatéraux optimale pour les familles exponentielles)

Soit  $X_1, \dots, X_n$  un échantillon iid tiré d'une famille exponentielle à 1-paramètre avec fonction de densité

$$f(x; \theta) = \exp\{\eta(\theta)T(x) - d(\theta) + S(x)\}, \quad x \in \mathcal{X}, \theta \in \Theta \subseteq \mathbb{R},$$

avec

- ❶  $\Theta$  un ouvert.
- ❷  $\eta(\cdot)$  est strictement croissante et continûment dérivable,

Si  $\tau = \sum_{i=1}^n T(X_i)$  est une variable aléatoire continue, alors :

- ❶ Pour  $\alpha \in (0, 1)$ , la statistique de test  $\delta = 1\{\tau \geq q_{1-\alpha}\}$  est Uniformément la Plus Puissante (UPP) pour tester

$$\left\{ \begin{array}{l} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{array} \right\}$$

au seuil  $\alpha$ . Ici,  $q_{1-\alpha}$  est le  $(1 - \alpha)$ -quantile de  $G_0(t) = \mathbb{P}_{\theta_0}[\tau \leq t]$ .

- ❷ Pour  $\alpha \in (0, 1)$ , la statistique de test  $\delta = 1\{\tau \leq q_\alpha\}$  est uniformément la plus puissante pour tester

$$\left\{ \begin{array}{l} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{array} \right\}$$

au seuil  $\alpha$ . Ici,  $q_\alpha$  est le  $\alpha$ -quantile de  $G_0(t) = \mathbb{P}_{\theta_0}[\tau \leq t]$ .

## Remarque

Si  $\eta(\cdot)$  est strictement décroissante, alors définissons

$$\eta_1(\cdot) = -\eta(\cdot) \quad \& \quad T_1 = -T.$$

Nous avons une famille exponentielle

$$f(x; \theta) = \exp\{\eta_1(\theta) T_1(x) - d(\theta) + S(x)\}, \quad x \in \mathcal{X}, \theta \in \Theta \subseteq \mathbb{R},$$

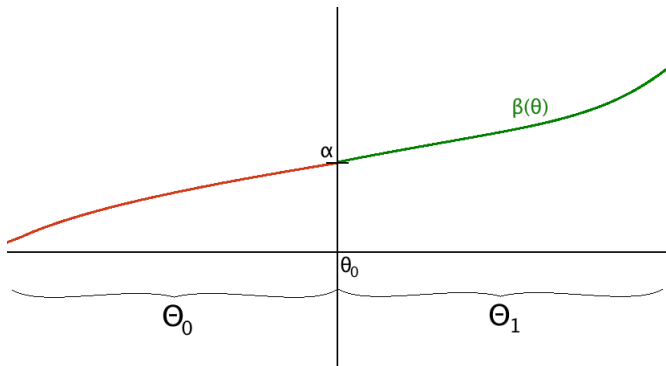
avec  $\eta_1(\cdot)$  strictement croissante.

Dans le tableau suivant, nous avons résumé le forme de la statistique de test, qui dépend de la direction des hypothèses et de la monotonicité de  $\eta$ .

	$\begin{pmatrix} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{pmatrix}$	$\begin{pmatrix} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{pmatrix}$
$\eta(\cdot)$ croissante	$1\{\tau(X_1, \dots, X_n) > q_{1-\alpha}\}$	$1\{\tau(X_1, \dots, X_n) \leq q_\alpha\}$
$\eta(\cdot)$ décroissante	$1\{\tau(X_1, \dots, X_n) \leq q_\alpha\}$	$1\{\tau(X_1, \dots, X_n) > q_{1-\alpha}\}$

# Cas unilatéral

- 1 Notez que la forme du test est exactement la même que celle du test pour la famille exponentielle d'une paire d'hypothèses simple vs simple
- 2 Cela est possible car pour une famille exponentielle, la forme du test de Neyman-Pearson ne dépend pas de la valeur précise de  $\theta_1$ , mais seulement de si  $\theta_1 < \theta_0$  ou  $\theta_1 > \theta_0$ , et de la valeur de  $\theta_0$ .
- 3 Ceci n'est pas vrai en général, mais ça l'est pour les familles exponentielles à 1-paramètre, en raison de leur forme spéciale.



$$\theta \mapsto \mathbb{E}_{\theta}[\delta]$$

# Cas bilatéral

# Cas bilatéral

Aucun espoir pour trouver des tests optimales dans ce cas :

- Pour que  $\delta : \mathcal{X}^n \rightarrow \{0, 1\}$  soit optimal en même temps pour :

(1)  $H_0 : \theta = \theta_0$  vs  $H_1 : \theta = \theta_1$ , pour tout  $\theta_1 > \theta_0$

et

(2)  $H_0 : \theta = \theta_0$  vs  $H_1 : \theta = \theta_1$ , pour tout  $\theta_1 < \theta_0$

- Mais la forme du test optimal est différente dans les deux cas !
- Rappelons le cas d'une famille exponentielle :

	$\theta_0 < \theta_1$	$\theta_0 > \theta_1$
$\eta(\cdot)$ croissante	$1\{\tau(X_1, \dots, X_n) > q_{1-\alpha}\}$	$1\{\tau(X_1, \dots, X_n) \leq q_\alpha\}$
$\eta(\cdot)$ décroissante	$1\{\tau(X_1, \dots, X_n) \leq q_\alpha\}$	$1\{\tau(X_1, \dots, X_n) > q_{1-\alpha}\}$

- On abandonne donc l'exigence d'un test optimal et on cherche pour de tests raisonnables.

# de nouveau la vraisemblance...

## d'où commencer ?

- ↪ Peut-être generaliser les tests de la forme Neyman-Pearson ?
- ↪ Peut-être utiliser un estimateur de vraisemblance pour juger si  $\theta_0$  est proche au EMV  $\hat{\theta}$  ?

- 1 le premier nous mène vers les test du rapport de vraisemblance
- 2 le deuxième vers le test de Wald

# Test du rapport de vraisemblance

# Test du rapport de vraisemblance

## Définition (Test du rapport de vraisemblance)

Soit  $X_1, \dots, X_n \stackrel{iid}{\sim} f(x; \theta)$ , qui nous donne la vraisemblance

$$L(\theta) = \prod_{i=1}^n f(X_i; \theta),$$

et soient  $H_0 : \theta \in \Theta_0$  et  $H_1 : \theta \in \Theta_1$  deux hypothèses à tester. Le rapport de vraisemblance est défini comme suit

$$\Lambda(X_1, \dots, X_n) = \frac{\sup_{\theta \in \Theta_1} L(\theta)}{\sup_{\theta \in \Theta_0} L(\theta)}.$$

Le test du rapport de vraisemblance (TRV) au seuil  $\alpha \in (0, 1)$  est défini comme étant le test dont la fonction de test est :

$$\delta(X_1, \dots, X_n) = 1\{\Lambda(X_1, \dots, X_n) > Q\},$$

où  $Q > 0$  est tel que  $\sup_{\theta \in \Theta_0} \mathbb{P}_\theta[\Lambda(X_1, \dots, X_n) > Q] = \alpha$ , lorsqu'il existe.

# TRV pour d'hypothèses bilatérales

Lorsque

$$H_0 : \theta = \theta_0 \quad \& \quad H_1 : \theta \neq \theta_0,$$

nous avons

$$\Theta_0 = \{\theta_0\} \quad \& \quad \Theta_1 = \mathbb{R} \setminus \{\theta_0\},$$

et donc, si  $L$  est une fonction continue de  $\theta$  et qu'elle atteint son supremum,

$$\Lambda(X_1, \dots, X_n) = \frac{\sup_{\theta \in \Theta_1} L(\theta)}{\sup_{\theta \in \Theta_0} L(\theta)} = \frac{\sup_{\theta \in \mathbb{R} \setminus \{\theta_0\}} L(\theta)}{L(\theta_0)} = \frac{\sup_{\theta \in \mathbb{R}} L(\theta)}{L(\theta_0)} = \frac{L(\hat{\theta})}{L(\theta_0)},$$

où  $\hat{\theta}$  est l'estimateur du maximum de vraisemblance de  $\theta$ .

Donc, pour les cas qui nous concernent :

Le test du rapport de vraisemblance (TRV) de  $H_0 : \theta = \theta_0$  vs  $H_0 : \theta \neq \theta_0$  au seuil  $\alpha \in (0, 1)$  est défini comme étant le test dont la fonction de test est :

$$\delta(X_1, \dots, X_n) = 1\{L(\hat{\theta})/L(\theta_0) > Q\},$$

où  $Q > 1$  est tel que  $\sup_{\theta \in \Theta_0} \mathbb{P}_{\theta}[L(\hat{\theta})/L(\theta_0) > Q] = \alpha$ , lorsqu'il existe.

## Example

Soient  $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$  avec  $\sigma^2$  connue. Considérons

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0.$$

Puisque l'EMV de  $\mu$  est  $\bar{X}$ , nous avons

$$\begin{aligned} L(\bar{X}) &= \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 \right\}, \\ L(\mu_0) &= \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} \exp \left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^n (X_i - \mu_0)^2 \right\}. \end{aligned}$$

Par conséquent,

$$\Lambda(X_1, \dots, X_n) = \frac{L(\bar{X})}{L(\mu_0)} = \exp \left\{ -\frac{1}{2\sigma^2} \left[ \sum_{i=1}^n (X_i - \bar{X})^2 - \sum_{i=1}^n (X_i - \mu_0)^2 \right] \right\}.$$

Notons que

$$\sum_{i=1}^n (X_i - \mu_0)^2 = \sum_{i=1}^n (X_i - \bar{X} + \bar{X} - \mu_0)^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2,$$

## Example (continuation)

Il s'ensuit que le rapport de vraisemblance se réduit à

$$\Lambda(X_1, \dots, X_n) = \exp \left\{ \frac{n}{2\sigma^2} (\bar{X} - \mu_0)^2 \right\}.$$

Nous pouvons en déduire que  $\Lambda(X_1, \dots, X_n)$  est une fonction croissante de

$$S(X_1, \dots, X_n) = \left( \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right)^2.$$

Notons que lorsque  $H_0$  est vraie,  $S \sim \chi_1^2$ . Ainsi,

$$\delta = 1\{S(X_1, \dots, X_n) > \chi_{1,1-\alpha}^2\},$$

où  $\chi_{1,1-\alpha}^2$  dénote le  $(1 - \alpha)$ -quantile d'une distribution  $\chi_1^2$ .

Notons que ceci est équivalent à rejeter l'hypothèse nulle si et seulement si

$$\left| \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \right| > z_{1-\alpha/2},$$

où  $z_{1-\alpha/2}$  est le  $(1 - \alpha/2)$ -quantile d'une distribution  $N(0, 1)$ .

# Presence d'un paramètre de nuisance

Supposons que  $X_1, \dots, X_n \stackrel{iid}{\sim} f(x; \theta, \xi)$ , où  $\theta \in \mathbb{R}$  et  $\xi \in \mathbb{R}^p$  sont deux paramètres inconnus. Nous pouvons être intéressés à tester

$$H_0 : \theta = \theta_0 \quad \text{vs} \quad H_1 : \theta \neq \theta_0$$

au seuil  $\alpha > 0$ , pour un certain  $\theta_0 \in \mathbb{R}$ , **sans faire aucune référence au (et sans se soucier du) paramètre  $\xi$** . Observez que ce paire d'hypothèses est équivalent à

$$H_0 : (\theta, \xi) \in \{\theta_0\} \times \mathbb{R}^p \quad \text{vs} \quad H_1 : (\theta, \xi) \in \{\mathbb{R} \setminus \{\theta_0\}\} \times \mathbb{R}^p$$

Dans ce cas, si  $L$  est continue, le rapport de vraisemblance est donné par

$$\Lambda = \frac{\sup_{\theta \in \mathbb{R} \setminus \{\theta_0\}, \xi \in \mathbb{R}^p} L(\theta, \xi)}{\sup_{\theta \in \{\theta_0\}, \xi \in \mathbb{R}^p} L(\theta, \xi)} = \frac{\sup_{\theta \in \mathbb{R}, \xi \in \mathbb{R}^p} L(\theta, \xi)}{\sup_{\xi \in \mathbb{R}^p} L(\theta_0, \xi)} = \frac{L(\hat{\theta}, \hat{\xi})}{\sup_{\xi \in \mathbb{R}^p} L(\theta_0, \xi)},$$

où  $(\hat{\theta}, \hat{\xi})$  est l'EMV de  $(\theta, \xi)$ . Le test du rapport de vraisemblance au seuil  $\alpha \in (0, 1)$  sera encore une fois défini comme étant le test dont la fonction de test est

$$\delta(X_1, \dots, X_n) = 1\{\Lambda(X_1, \dots, X_n) > Q\},$$

où  $Q > 0$  est tel que  $\mathbb{P}_{\theta_0}[\Lambda(X_1, \dots, X_n) > Q] = \alpha$ , lorsqu'il existe.

## Exemple (Test bilatéral pour les moyennes de lois gaussiennes)

Soit  $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$ , où  $\mu$  et  $\sigma^2$  sont inconnus. Considerons

$$H_0 : \mu = \mu_0 \quad \text{vs} \quad H_1 : \mu \neq \mu_0$$

au seuil  $\alpha > 0$ , pour une certaine valeur fixée  $\mu_0 \in \mathbb{R}$ . Nous devons déterminer

$$\Lambda(X_1, \dots, X_n) = \frac{L(\hat{\mu}, \hat{\sigma}^2)}{\sup_{\sigma^2 > 0} L(\mu_0, \sigma^2)},$$

où  $(\hat{\mu}, \hat{\sigma}^2)$  est l'EMV de  $(\mu, \sigma^2)$ . Pour le dénominateur, nous pouvons calculer que

$$\frac{\partial}{\partial \sigma^2} \ell(\mu_0, \sigma^2) = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (X_i - \mu_0)^2.$$

Nous concluons que

$$\arg \sup_{\sigma^2 > 0} L(\mu_0, \sigma^2) = \frac{1}{n} \sum_{i=1}^n (X_i - \mu_0)^2.$$

## Example (Test bilatéral pour les moyennes de lois gaussiennes)

En d'autres mots, le supremum du dénominateur est

$$\begin{aligned} & \left[ \frac{1}{2\pi(1/n) \sum_{i=1}^n (X_i - \mu_0)^2} \right]^{n/2} \exp \left\{ -\frac{\sum_{i=1}^n (X_i - \mu_0)^2}{(2/n) \sum_{i=1}^n (X_i - \mu_0)^2} \right\} \\ &= \left[ \frac{ne^{-1}}{2\pi \sum_{i=1}^n (X_i - \mu_0)^2} \right]^{n/2}. \end{aligned}$$

Au numérateur, rappelons que l'EMV de  $(\mu, \sigma^2)$  est  $(\bar{X}, \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2)$

$$\begin{aligned} L(\hat{\mu}, \hat{\sigma}^2) &= \left[ \frac{1}{2\pi(1/n) \sum_{i=1}^n (X_i - \bar{X})^2} \right]^{n/2} \exp \left\{ -\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{(2/n) \sum_{i=1}^n (X_i - \bar{X})^2} \right\} \\ &= \left[ \frac{ne^{-1}}{2\pi \sum_{i=1}^n (X_i - \bar{X})^2} \right]^{n/2}. \end{aligned}$$

Par conséquent le rapport de vraisemblance est

$$\Lambda(X_1, \dots, X_n) = \frac{L(\hat{\mu}, \hat{\sigma}^2)}{\sup_{\sigma^2 > 0} L(\mu_0, \sigma^2)} = \left[ \frac{\sum_{i=1}^n (X_i - \mu_0)^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]^{n/2}.$$

## Example (continuation)

Nous pouvons simplifier cette expression encore plus en observant que

$$\sum_{i=1}^n (X_i - \mu_0)^2 = \sum_{i=1}^n (X_i - \bar{X} + \bar{X} - \mu_0)^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2,$$

puisque les termes croisés s'annulent. En utilisant ce fait, nous pouvons écrire

$$\Lambda = \left[ \frac{\sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right]^{n/2} = \left\{ 1 + \frac{n(\bar{X} - \mu_0)^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right\}^{n/2}.$$

Observez maintenant que

$$\Lambda > Q \iff \underbrace{\frac{n(\bar{X} - \mu_0)^2}{\sum_{i=1}^n (X_i - \bar{X})^2 / (n-1)}}_{T^2} > \underbrace{(n-1)(Q^{2/n} - 1)}_{:=C} \iff \underbrace{\left| \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \right|}_{|T|} > \sqrt{C}.$$

Le test du rapport de vraisemblance est donc

$$\delta(X_1, \dots, X_n) = 1\{\Lambda > Q\} = 1\left\{ \left| \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \right| > \sqrt{C} \right\}.$$

## Example (continuation)

Ici  $\sqrt{C}$  doit être choisi afin que

$$\mathbb{P}_{H_0} \left[ \left| \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \right| > \sqrt{C} \right] = \alpha.$$

Mais, lorsque  $H_0$  est vraie, nous avons que

$$T \sim t_{n-1},$$

où  $t_{n-1}$  représente une distribution de Student avec  $n - 1$  degrés de liberté

Ceci nous donne que

$$\sqrt{C} = t_{n-1, 1-\alpha/2},$$

où  $t_{n-1, 1-\alpha/2}$  est le  $(1 - \alpha/2)$  quantile d'une distribution  $t_{n-1}$ . En conclusion, le TRV est

$$\delta = 1 \left\{ \left| \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \right| > t_{n-1, 1-\alpha/2} \right\}.$$



- Dans le cas Gaussien, nous avons pu trouver la bonne valeur critique  $Q$  pour que le TRV  $\delta = 1\{\Lambda > Q\}$  respecte le seuil  $\alpha$ .
- Que faire dans d'autre cas ?
- Par exemple : que faire dans le cadre d'un test bilatérale pour une famille exponentielle quelconque ?
- Mais, bien-sur, nous allons de nouveau recourir à des approximations.

# Valeurs critiques approximatives pour le TRV

## Théorème

Soit  $X_1, \dots, X_n$  un échantillon iid tiré d'une distribution de fonction de densité/masse  $f(x; \theta)$  qui appartient à une famille exponentielle non-dégénérée à,

$$f(x; \theta) = \exp\{\eta(\theta)T(x) - d(\theta) + S(x)\}, \quad x \in \mathcal{X}, \theta \in \Theta$$

Supposons que :

- 1 L'espace des paramètres  $\Theta \subset \mathbb{R}$  est un ensemble ouvert.
- 2 La fonction  $\eta : \Theta \rightarrow \Phi = \eta(\Theta)$  est une bijection de classe  $C^2$ .

Soit  $\hat{\theta}_n$  l'estimateur du maximum de vraisemblance de  $\theta$ , et soit  $\theta_0 \in \Theta$  un élément fixe de l'espace des paramètres, tel que  $\eta'(\theta_0) \neq 0$ . Si

$\Lambda(X_1, \dots, X_n) = L(\hat{\theta}_n)/L(\theta_0)$  est le rapport de vraisemblance, alors

$$2 \log \Lambda(X_1, \dots, X_n) = 2(\ell(\hat{\theta}_n) - \ell(\theta_0)) \xrightarrow{d} \chi_1^2,$$

lorsque  $\{H_0 : \theta = \theta_0\}$  est vraie.

Remarque : les suppositions garantissent que  $d$  est  $C^2$  à  $\theta_0$ , voir Remarque 2.15.

# Valeurs critiques approximatives pour le TRV

## Comment utiliser ce résultat ?

- 1 Choisissons  $Q = \exp \left\{ \frac{\chi_{1,1-\alpha}^2}{2} \right\}$ , où  $\chi_{1,1-\alpha}^2$  représente le  $(1 - \alpha)$ -quantile d'une distribution  $\chi_1^2$ .
- 2 Alors, comme  $2 \log \Lambda \xrightarrow{d} \chi_1^2$ , on aura

$$\begin{aligned} \mathbb{P}_{\theta_0}[\Lambda > Q] &= \mathbb{P}_{\theta_0}[\log \Lambda > \log Q] \\ &= \mathbb{P}_{\theta_0}[2 \log \Lambda > \chi_{1,1-\alpha}^2] \\ &\xrightarrow{n \rightarrow \infty} \alpha \end{aligned}$$

En conclusion, le TRV est approximativement (pour grand  $n$ ) équivalent à :

$$1 \left\{ 2(\ell(\hat{\theta}_n) - \ell(\theta_0)) > \chi_{1,1-\alpha}^2 \right\}$$

# Tests de Wald

# Utiliser la théorie d'estimation ponctuelle pour de tests bilatéraux ?

- 1 On veut tester  $\{H_0 : \theta = \theta_0\}$  vs  $\{H_1 : \theta \neq \theta_0\}$
- 2 Si on a un estimateur ponctuel  $\hat{\theta}(X_1, \dots, X_n)$  du vrai paramètre, alors on peut comparer la valeur nulle  $\theta_0$  avec la valeur observée de  $\hat{\theta}(X_1, \dots, X_n)$ .
- 3 Si ces deux valeurs sont séparées par une distance « significative », alors il est clair que nous devrions rejeter  $H_0 : \theta = \theta_0$  en faveur de  $H_1 : \theta \neq \theta_0$ .
- 4 De quelle taille doit-elle être une distance pour qu'on la considère comme « significative » ?
- 5 Cette distance ne peut pas être exprimée en terme absolue !
- 6 ...car nous devons tenir compte de la variabilité de  $\hat{\theta}$

Exprimer la distance en terme de la variance de  $\hat{\theta}$ . Ceci nous donne une statistique de test de la forme :

$$\frac{|\hat{\theta} - \theta_0|}{\sqrt{\text{Var}(\hat{\theta})}}$$

# Test de Wald

Le seul problème est : **souvent on ne connaît pas la variance de  $\hat{\theta}$**

(car elle peut dépendre de la vraie valeur du paramètre  $\theta$ )

## Définition (Test de Wald)

Soient  $X_1, \dots, X_n \stackrel{iid}{\sim} f(\cdot; \theta)$  et  $\hat{\theta}$  un estimateur de  $\theta$  basé sur l'échantillon  $X_1, \dots, X_n$ . Un test de Wald pour la paire d'hypothèses

$$\{H_0 : \theta = \theta_0, H_1 : \theta \neq \theta_0\}$$

au seuil  $\alpha$  est un test dont la fonction de test est

$$\delta(X_1, \dots, X_n) = 1 \left\{ \frac{(\hat{\theta} - \theta_0)^2}{\widehat{\text{Var}}(\hat{\theta})} > Q \right\},$$

où  $\mathbb{P}_{\theta_0} \left[ \frac{(\hat{\theta} - \theta_0)^2}{\widehat{\text{Var}}(\hat{\theta})} > Q \right] = \alpha$ , lorsqu'un tel  $Q$  existe.

# Test de Wald basé sur la vraisemblance

Nous savons que l'EMV a une performance asymptotique quasiment optimale.

Alors :

- 1 Choisissons comme l'EMV pour jouer le rôle de  $\hat{\theta}$
- 2 Et pour  $\widehat{\text{Var}}(\hat{\theta})$  ?

Quand  $n \rightarrow \infty$ , la variance de l'EMV dans une famille exponentielle est

$$\approx \frac{1}{n} \frac{[\eta'(\theta_0)]}{d''(\theta_0)\eta'(\theta_0) - d'(\theta_0)\eta''(\theta_0)}$$

Alors définissons

$$\widehat{\text{Var}}(\hat{\theta}) := \frac{1}{n} \frac{[\eta'(\hat{\theta}_n)]}{d''(\hat{\theta}_n)\eta'(\hat{\theta}_n) - d'(\hat{\theta}_n)\eta''(\hat{\theta}_n)}$$

est posons  $\widehat{J}_n = \widehat{\text{Var}}^{-1}(\hat{\theta})$ . Le test de Wald devient :

## Test de Wald basé sur la vraisemblance

$$1\{\widehat{J}_n(\hat{\theta}_n - \theta_0)^2 > Q\}$$

où  $\mathbb{P}_{\theta_0} [\widehat{J}_n(\hat{\theta} - \theta_0)^2 > Q] = \alpha$ , lorsqu'un tel  $Q$  existe.

## Théorème (Valeurs critiques approximatives pour les tests Wald)

Soit  $X_1, \dots, X_n$  un échantillon iid tiré d'une distribution ayant une fonction de densité/masse  $f(x; \theta)$  appartenant à une famille exponentielle non-dégénérée à 1-paramètre,

$$f(x; \theta) = \exp\{\eta(\theta) T(x) - d(\theta) + S(x)\}, \quad x \in \mathcal{X}, \theta \in \Theta.$$

Supposons que :

- 1 L'espace des paramètres  $\Theta \subset \mathbb{R}$  est un ensemble ouvert.
- 2 La fonction  $\eta(\cdot)$  est une bijection deux fois continûment dérivable entre  $\Theta$  et  $\Phi = \eta(\Theta)$ .

Soient  $\hat{\theta}_n$  l'estimateur du maximum de vraisemblance de  $\theta$ , et  $\widehat{J}_n = n \frac{d''(\hat{\theta}_n) \eta'(\hat{\theta}_n) - d'(\hat{\theta}_n) \eta''(\hat{\theta}_n)}{[\eta'(\hat{\theta}_n)]}$ . Soit  $\theta_0 \in \Theta$  un élément fixe de l'espace des paramètres tel que  $\eta'(\theta_0) \neq 0$ . Alors,

$$\widehat{J}_n (\hat{\theta}_n - \theta_0)^2 \xrightarrow{d} \chi_1^2,$$

lorsque  $\{H_0 : \theta = \theta_0\}$  est vraie.

Remarque : les suppositions garantissent que  $d$  est  $C^2$  à  $\theta_0$ , voir Remarque 2.15.

- Le résultat ci-dessus peut être utilisé afin de déterminer la valeur critique d'un test de Wald avec un seuil  $\alpha$ .

La fonction de test de Wald au seuil  $\alpha$ , est approximativement (pour grand  $n$ ) équivalent à

$$1 \left\{ \widehat{J}_n (\hat{\theta}_n - \theta_0)^2 > \chi_{1,1-\alpha}^2 \right\},$$

où  $\chi_{1,1-\alpha}^2$  représente le  $(1 - \alpha)$ -quantile d'une distribution  $\chi_1^2$ .

- En d'autres termes, pour de grandes valeurs de  $n$ , la valeur critique approximative devrait être  $Q \approx \chi_{1,1-\alpha}^2$ .

# $p$ -valeur de Fisher

# N.-P. et la pratique

Neyman–Pearson : une théorie mathématique Élégante et raisonnable.

**Mais** : parfois il y a des problèmes pratiques :

- ❶ Il n'est pas toujours clair à priori quel est le « bon » seuil de signification à utiliser.
  - ↪ il se peut que, pour les même données,  $H_0$  soit rejetée pour  $\alpha = 0.05$ , mais pas pour  $\alpha = 0.01$  !
- ❷ Une fois que le seuil est fixé, nous utilisons un test optimal (s'il est disponible), et nous prenons une décision basée sur nos données. Le problème maintenant est que nous n'avons pas d'indications claires afin de savoir à quel point notre décision était « sûre » ou « marginale »
  - ↪ Les scientifiques souhaitent parfois non seulement pouvoir prendre une décision, mais aussi pouvoir quantifier la confiance qu'ils ont dans cette décision.

# $p$ -valeur de Fisher : une approche duale à celle de N.-P.



- 1 Plutôt que de prendre une décision explicite (i.e.  $\delta = 0$  ou  $\delta = 1$ ), définissons une mesure qui indique à quel point les données supportent l'hypothèse nulle.
- 2 Nous laissons par la suite le scientifique juger s'il y a oui ou non assez d'évidences contre  $H_0$ .

# $p$ -valeur de Fisher : une approche duale à celle de N.-P.

## Définition ( $p$ -valeur)

Soient  $X_1, \dots, X_n \stackrel{iid}{\sim} f(\cdot; \theta)$  et  $H_0 : \theta \in \Theta_0$  une hypothèse de la forme :

$$\{H_0 : \theta = \theta_0\} \quad \text{ou} \quad \{H_0 : \theta \leq \theta_0\} \quad \text{ou} \quad \{H_0 : \theta \geq \theta_0\}.$$

Soit  $\delta_\alpha$  une fonction de test pour  $H_0$ , ayant l'une des deux formes suivantes :

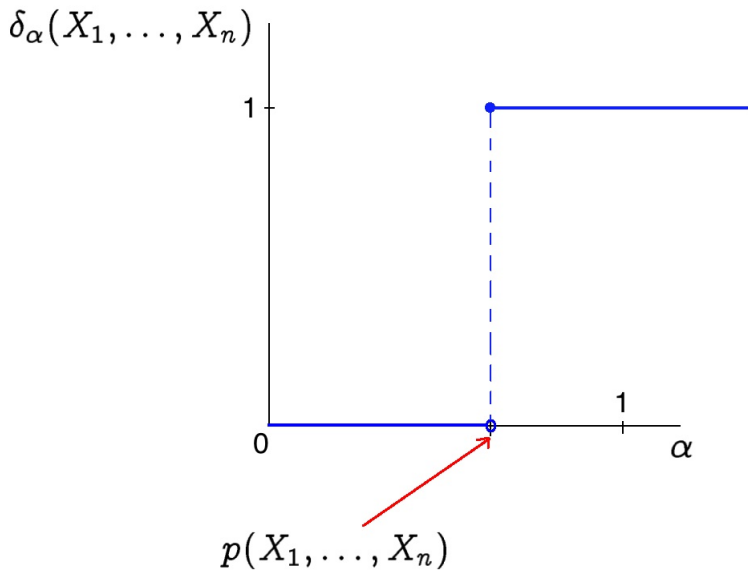
$$\delta_\alpha(X_1, \dots, X_n) := 1\{T(X_1, \dots, X_n) > q_{1-\alpha}\}$$

ou

$$\delta_\alpha(X_1, \dots, X_n) := 1\{T(X_1, \dots, X_n) \leq q_\alpha\},$$

où  $T$  est une certaine statistique de test, et  $q_z$  est le  $z$ -quantile de la distribution  $G_0(t) = \mathbb{P}_{\theta_0}[T(X_1, \dots, X_n) \leq t]$ . Alors

$$p(X_1, \dots, X_n) := \inf\{\alpha \in (0, 1) : \delta_\alpha(X_1, \dots, X_n) = 1\}.$$



# Calculs des valeurs- $p$

- La définition de la  $p$ -valeur semble un peu compliquée, il est donc naturel de se demander s'il est possible de la calculer dans des exemples concrets.
- Cela est en effet le cas lorsque l'hypothèse nulle est d'une des formes que nous avons considérées jusqu'à présent. Les calculs sont en fait plutôt simples

## Lemme (Calculs des valeurs- $p$ )

*Dans le même contexte que celui de la définition précédente, si  $G_0$  est continue<sup>a</sup> et strictement croissante, alors*

- ❶ Si  $\delta_\alpha$  est de la forme  $\delta_\alpha(X_1, \dots, X_n) := 1\{T(X_1, \dots, X_n) > q_{1-\alpha}\}$ , alors

$$p(X_1, \dots, X_n) = 1 - G_0(T(X_1, \dots, X_n)).$$

- ❷ Si  $\delta_\alpha$  est de la forme  $\delta_\alpha(X_1, \dots, X_n) := 1\{T(X_1, \dots, X_n) \leq q_\alpha\}$ , alors

$$p(X_1, \dots, X_n) = G_0(T(X_1, \dots, X_n)).$$

---

a. continuité n'est pas nécessaire

# Calculs des valeurs- $p$ – Interprétation

Le lemme nous donne une autre façon de comprendre les valeurs- $p$ .

Concentrons nous sur le cas (1), où nous rejetons pour des grandes valeurs de  $T$ .

- 1 Notez que  $1 - G_0(T(X_1, \dots, X_n))$  est égal à la probabilité d'observer quelque chose d'aussi grand, ou même plus grand que ce que nous avons observé, lorsque  $H_0$  est vraie.
- 2 Ainsi, lorsque la  $p$ -valeur est petite, nous avons en fait observé quelque chose qui serait très improbable si  $H_0$  était en effet vraie.
- 3 Nous nous attendons alors à ce que  $H_0$  soit fausse.

## Remarque (Avertissement)

*Une erreur commune est d'interpréter la  $p$ -valeur comme la probabilité que  $H_0$  soit vraie. Ceci est faux, et n'a en fait pas de sens, car le paramètre  $\theta$  n'est pas une variable aléatoire.*

## Exemple (Calculs des valeurs- $p$ , cas normal)

Soit  $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, 1)$  et considérons la paire d'hypothèses :

$$H_0 : \mu = 0 \quad \text{vs} \quad H_1 : \mu \neq 0.$$

Rappelons que le TRV pour cette paire est donné par :

$$\delta(X_1, \dots, X_n) = 1 \left\{ \left( \frac{\bar{X}}{1/\sqrt{n}} \right)^2 > \chi_{1,1-\alpha}^2 \right\},$$

où  $\chi_{1,1-\alpha}^2$  est le  $(1 - \alpha)$ -quantile d'une distribution  $\chi_1^2$ .

Nous pouvons donc définir la  $p$ -valeur correspondante comme étant :

$$1 - G_{\chi_1^2}(n\bar{X}^2)$$

(notons que  $G_{\chi_1^2}$  est une fonction monotone croissante de  $(0, \infty)$  à  $(0, 1)$  puisque la fonction de densité d'une  $\chi_1^2$  est strictement positive sur tout  $(0, \infty)$ ).  $\square$

# Lien entre Neyman-Pearson et Valeurs- $p$

Est-ce qu'il y a un lien entre les approches de Fisher et de Neyman & Pearson en ce qui concerne les tests d'hypothèse ?

Il y a une relation particulièrement simple et élégante :

## Corollaire

*Dans le même contexte que celui du dernier lemme, soit  $\alpha_0 \in (0, 1)$  et supposons que  $G_0$  est continue et strictement croissante. Si nous définissons*

$$\psi(X_1, \dots, X_n) := 1\{p(X_1, \dots, X_n) \leq \alpha_0\},$$

*alors  $\psi(X_1, \dots, X_n) = \delta_{\alpha_0}(X_1, \dots, X_n)$ . En d'autres mots, si nous rejetons l'hypothèse nulle lorsque la  $p$ -valeur est plus petite que  $\alpha_0$ , alors notre test se réduit à  $\delta_{\alpha_0}$ .*

# Intervalle de Confiance

# Rappel : notre cadre général

- 1 On dispose d'une **distribution  $F(x; \theta)$**  qui dépend d'un **paramètre inconnu  $\theta \in \mathbb{R}^p$** .
- 2 Nous **observons la réalisation de  $n$  variables aléatoires  $X_1, \dots, X_n$** , indépendantes et identiquement distribuées, qui suivent cette distribution. Mais **nous ne connaissons toujours pas la vraie valeur de  $\theta$  qui a généré les  $X_i$  !**
- 3 Nous voulons utiliser les  $n$  observations (les réalisations de  $X_1, \dots, X_n$ ) afin de faire des **assertions concernant la vraie valeur de  $\theta$** , et afin de quantifier l'incertitude associée à ces assertions.

# Quelle sorte d'affirmations concernant la vraie valeur de $\theta$ ?

- 1 **Estimation.** Etant donné un échantillon  $X_1, \dots, X_n$  tiré d'une distribution  $F_\theta$  qui dépend d'un paramètre inconnu  $\theta$ , comment peut-on construire un estimateur, i.e une fonction de l'échantillon dont le but est d'estimer  $\theta$  ?
- 2 **Tests d'hypothèses.** Etant donné une valeur plausible  $\theta_0$  pour  $\theta$  (ou plusieurs valeurs plausibles formant un ensemble  $\Theta_0$ ), est-ce que, sur la base de l'échantillon  $X_1, \dots, X_n$ , cette valeur (ou cet ensemble) est un bon indicateur de la vraie valeur de  $\theta$  ?
- 3 **Intervalles de confiance.** Plutôt que de tenter d'estimer la valeur précise du paramètre  $\theta$  qui a généré notre échantillon  $X_1, \dots, X_n$ , est-ce qu'on peut construire un ensemble de valeurs sous la forme d'un intervalle, qui aura une grande probabilité de contenir le vrai paramètre  $\theta$  ?

## Définition (Intervalle de confiance bilatéral)

Soient  $X_1, \dots, X_n \stackrel{iid}{\sim} f(x; \theta)$ , où  $\theta \in \Theta \subseteq \mathbb{R}$ , un échantillon aléatoire et  $\alpha \in (0, 1)$  une constante. Soient  $L(X_1, \dots, X_n)$  et  $U(X_1, \dots, X_n)$  deux statistiques, appelées respectivement la limite inférieure et la limite supérieure, telles que

$$\inf_{\theta \in \Theta} \mathbb{P}_{\theta} \left[ L(X_1, \dots, X_n) \leq \theta \leq U(X_1, \dots, X_n) \right] = 1 - \alpha.$$

Alors, l'intervalle aléatoire

$$\left[ L(X_1, \dots, X_n), U(X_1, \dots, X_n) \right],$$

est appelé un intervalle de confiance bilatéral pour  $\theta$  avec un seuil de confiance  $(1 - \alpha)$ .

## Définition (Intervalle de confiance unilatéral)

Soient  $X_1, \dots, X_n \stackrel{iid}{\sim} f(x; \theta)$ , où  $\theta \in \Theta \subseteq \mathbb{R}$ , un échantillon aléatoire et  $\alpha \in (0, 1)$  une constante. Soit  $L(X_1, \dots, X_n)$  une statistique telle que

$$\inf_{\theta \in \Theta} \mathbb{P}_{\theta} \left[ L(X_1, \dots, X_n) \leq \theta \right] = 1 - \alpha.$$

Alors, l'intervalle aléatoire

$$\left[ L(X_1, \dots, X_n), +\infty \right)$$

est appelé un intervalle de confiance unilatéral à gauche pour  $\theta$  avec un seuil de confiance  $(1 - \alpha)$ . De façon analogue, si  $U(X_1, \dots, X_n)$  satisfait

$$\inf_{\theta \in \Theta} \mathbb{P}_{\theta} \left[ U(X_1, \dots, X_n) \geq \theta \right] = 1 - \alpha,$$

alors l'intervalle aléatoire

$$\left( -\infty, U(X_1, \dots, X_n) \right]$$

est appelé un intervalle de confiance unilatéral à droite pour  $\theta$  au seuil  $(1 - \alpha)$ .

# Interpretation

- Il faut faire attention lorsqu'on interprète un intervalle de confiance.
- Remarquez que

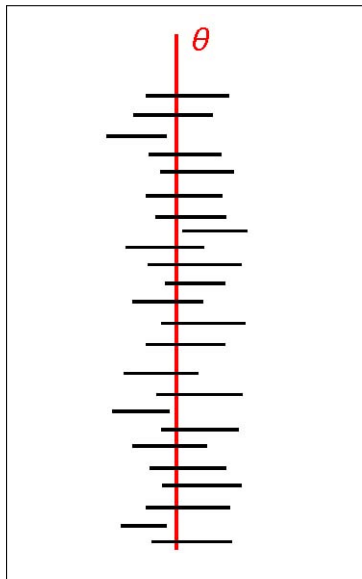
$$\inf_{\theta \in \Theta} \mathbb{P}_{\theta} \left[ L(X_1, \dots, X_n) \leq \theta \leq U(X_1, \dots, X_n) \right] = 1 - \alpha,$$

est une affirmation équivalente à

$$\inf_{\theta \in \Theta} \mathbb{P}_{\theta} \left\{ \theta \in \left[ L(X_1, \dots, X_n), U(X_1, \dots, X_n) \right] \right\} = 1 - \alpha.$$

- Toutefois, la deuxième façon d'écrire l'affirmation peut nous amener à une mauvaise interprétation de ce que signifie un intervalle de confiance.
- En effet, c'est l'intervalle  $[L, U]$  qui est aléatoire et non le paramètre  $\theta$ .
- Dire que « la probabilité que le paramètre tombe à l'intérieur de l'intervalle est au moins  $1 - \alpha$  » est faux : le paramètre ne va ou ne tombe nul part, il est fixe !
- C'est l'intervalle qui peut changer pour différentes valeurs de l'échantillon  $X_1, \dots, X_n$ , et qui peut donc couvrir ou non le paramètre.
- Il faut donc dire « la probabilité que l'intervalle couvre le paramètre  $\theta$  est au moins  $(1 - \alpha)$  ».

# Interpretation



- Une façon différente de clarifier la situation est de remarquer que :

$$\begin{aligned}\mathbb{P}_\theta \left[ L(X_1, \dots, X_n) \leq \theta \leq U(X_1, \dots, X_n) \right] &= \\ &= \mathbb{P}_\theta \left[ \{L(X_1, \dots, X_n) \leq \theta\} \cap \{U(X_1, \dots, X_n) \geq \theta\} \right],\end{aligned}$$

où le côté droit de l'expression accentue le fait que l'affirmation s'applique aux bornes aléatoires de confiance  $L$  et  $U$ , plutôt qu'au paramètre déterministe  $\theta$ .

- Afin d'éviter toute confusion, il est préférable d'écrire  $\mathbb{P}_\theta \{[L, U] \ni \theta\}$  que  $\mathbb{P}_\theta \{\theta \in [L, U]\}$ .

# Exemple (presque le “seul exemple”)

## Exemple (Intervalle de confiance pour la moyenne d'une loi normale)

Soit  $X_1, \dots, X_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$ , où  $\mu$  est inconnu et  $\sigma^2$  est connu. Nous voulons construire un intervalle bilatéral pour  $\mu$ . Nous standardisons pour obtenir :

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

Ainsi, si  $z_{\frac{\alpha}{2}}$  et  $z_{1-\frac{\alpha}{2}}$  sont les  $\alpha/2$  et  $1 - \alpha/2$  quantiles (respectivement) de la distribution  $N(0, 1)$ , nous avons :

$$\mathbb{P} \left[ z_{\frac{\alpha}{2}} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\frac{\alpha}{2}} \right] = 1 - \alpha.$$

En manipulant l'expression à l'intérieur de la probabilité, nous obtenons :

$$\mathbb{P} \left[ z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right] = 1 - \alpha$$

## Example (Cas Gaussien, suite)

$$\Leftrightarrow \mathbb{P} \left[ -\bar{X} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right] = 1 - \alpha$$

$$\Leftrightarrow \mathbb{P} \left[ \bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \geq \mu \geq \bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right] = 1 - \alpha$$

$$\Leftrightarrow \mathbb{P} \left[ \bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right] = 1 - \alpha.$$

L'égalité ci-dessus est vraie quelque soit la vraie valeur de  $\mu \in \mathbb{R}$ . Donc si

$$L(X_1, \dots, X_n) = \bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \quad \& \quad U(X_1, \dots, X_n) = \bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}},$$

alors  $[L, U]$  est un intervalle de confiance au seuil  $1 - \alpha$ . Par symétrie de  $N(0, 1)$ ,

$$\left[ \underbrace{\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}}_{L(X_1, \dots, X_n)}, \quad \underbrace{\bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}}_{U(X_1, \dots, X_n)} \right]$$

## Exemple (Cas Gaussien, suite)

Observez que l'intervalle est symétrique autour de  $\bar{X}$ , le EMV de  $\mu$ . Pour mettre l'accent sur ce fait, on l'écrit souvent sous la forme

$$\bar{X} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$$

Nous pouvons ainsi faire quelques observations importantes :

- La longueur de l'intervalle de confiance est  $2z_{1-\alpha/2}\sigma/\sqrt{n}$ , ce qui dépend de  $\sigma^2$ ,  $n$  et  $\alpha$ .
- Le paramètre  $\sigma^2$  échappe à notre contrôle, puisque c'est la variance de la distribution  $N(\mu, \sigma^2)$  sous-jacente.
- Nous pouvons cependant contrôler la taille de l'échantillon  $n$  et le seuil de confiance  $1 - \alpha$ . En augmentant  $n$ , la longueur de l'intervalle est ré-échelonnée par un facteur de  $1/\sqrt{n}$ .
- D'un autre côté, diminuer  $\alpha$  (i.e. augmenter la confiance  $1 - \alpha$ ) a pour effet d'augmenter la longueur de l'intervalle : plus nous voulons avoir de la confiance dans notre intervalle et plus l'intervalle sera grand (notons que la longueur de l'intervalle tend vers l'infini lorsque  $\alpha \rightarrow 0$ ).

## Exemple (Cas Gaussien, suite)

Maintenant, considérons le problème consistant à trouver un intervalle de confiance unilatéral à droite. En utilisant le fait que  $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$ , nous pouvons écrire

$$\Rightarrow \mathbb{P} \left[ \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha} \right] = 1 - \alpha.$$

En manipulant l'expression, nous obtenons

$$\mathbb{P} \left[ \bar{X} + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \geq \mu \right] = 1 - \alpha,$$

et l'intervalle

$$\left( -\infty, \bar{X} + z_{1-\alpha} \frac{\sigma}{\sqrt{n}} \right).$$

est un intervalle de confiance unilatéral à droite avec au seuil  $1 - \alpha$ . De façon similaire, un intervalle de confiance unilatéral à gauche avec un seuil  $1 - \alpha$  est donné par

$$\left[ \bar{X} - z_{1-\alpha} \frac{\sigma}{\sqrt{n}}, +\infty \right).$$

# Pivots et Pivots Approximatifs

- ❶ Quelle était l'idée essentielle derrière cette construction ?
- ❷ Comment construire des intervalles plus généralement ?

La construction semble un peu ad-hoc, car l'étape cruciale était le resultat

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

qui nous a permis d'écrire

$$\mathbb{P}_{\mu} \left[ z_{\alpha/2} \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq z_{1-\alpha/2} \right] = 1 - \alpha,$$

qui était valide pour toute valeur de  $\mu$ . Nous étions alors capable de manipuler l'expression à l'intérieur de la probabilité afin d'obtenir notre intervalle.

Il semblerait que ce soit le concept auquel nous devrions nous intéresser dans un cadre plus abstrait...

## Définition (Pivot)

Soit  $X_1, \dots, X_n \stackrel{iid}{\sim} f(x; \theta)$ . Une fonction

$$g : \mathcal{X}^n \times \Theta \rightarrow \mathbb{R},$$

est appelée un pivot si

- 1  $\theta \mapsto g(x_1, \dots, x_n, \theta)$  est continue pour tout  $(x_1, \dots, x_n) \in \mathcal{X}^n$ .
- 2  $\mathbb{P}[g(X_1, \dots, X_n, \theta) \leq x]$  ne dépend pas de  $\theta$ .

Si nous sommes capables de trouver un pivot pour  $\theta$ , dont la distribution est connue, nous sommes alors capables de trouver les quantiles  $q_1$  et  $q_2$  tels que

$$\mathbb{P}[q_1 \leq g(X_1, \dots, X_n, \theta) \leq q_2] = 1 - \alpha.$$

Si  $g$  a une forme nous permettant de manipuler l'inégalité à l'intérieur de la probabilité on a espoir d'obtenir un intervalle explicite.

Sinon, nous pouvons toutefois tenter de déterminer de façon numérique l'ensemble

$$\{\theta \in \Theta : q_1 \leq g(X_1, \dots, X_n, \theta) \leq q_2\},$$

# Pivots et Pivots Approximatifs

Du point de vue conceptuel, tout va bien. Cependant, il y a deux défis auxquels nous faisons maintenant face :

- 1 Comment trouver des pivots en général ?
- 2 Comment déterminer la distribution d'un pivot ?

Pour répondre à 2, nous définissons :

## Définition (Pivot approximatif)

Soit  $X_1, \dots, X_n \stackrel{iid}{\sim} f(x; \theta)$ . Une fonction

$$g : \mathcal{X}^n \times \Theta \rightarrow \mathbb{R},$$

est appelée un pivot approximatif si

- 1 Pour tout  $n \in \mathbb{N}$ ,  $\theta \mapsto g(x_1, \dots, x_n, \theta)$  est continue pour tout  $(x_1, \dots, x_n) \in \mathcal{X}^n$ .
- 2 Nous avons

$$g(X_1, \dots, X_n, \theta) \xrightarrow{d} Y,$$

où  $Y$  est une variable aléatoire dont la distribution ne dépend pas de  $\theta$ .

# Pivots et Pivots Approximatifs

- Si nous connaissons la distribution asymptotique d'un pivot approximatif, nous pouvons construire un intervalle de confiance approximatif.
- Soit  $Y$  est une variable aléatoire continue. Si  $q_1$  et  $q_2$  sont les quantiles de  $F_Y$  tels que

$$\mathbb{P}[q_1 \leq Y \leq q_2] = 1 - \alpha.$$

- Alors nous avons par définition de la convergence en loi,

$$g(X_1, \dots, X_n, \theta) \xrightarrow{d} Y$$

$$\implies \mathbb{P}[q_1 \leq g(X_1, \dots, X_n, \theta) \leq q_2] \xrightarrow{n \rightarrow \infty} \mathbb{P}[q_1 \leq Y \leq q_2] = 1 - \alpha.$$

- Nous pouvons ainsi utiliser le pivot approximatif afin de construire un intervalle de confiance approximatif.

## Exemple (Moyenne d'une distribution générale)

Soit  $X_1, \dots, X_n$  une collection de variables aléatoires iid de moyenne inconnue  $\mu = \mathbb{E}[X]$  et de variance inconnue  $\mathbb{E}[(X_1 - \mu)^2] = \sigma^2 < \infty$ . On cherche un pivot approximatif afin de construire un intervalle pour  $\mu$ .

- Par le théorème central limite, nous avons  $\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N(0, \sigma^2)$ .
- Par la loi forte des grands nombres,  $S_n^2 = \sum_{i=1}^n (X_i - \mu)^2 / (n - 1) \xrightarrow{p} \sigma^2$ .

Maintenant, nous pouvons utiliser le théorème de Slutsky afin de conclure que

$$g(X_1, \dots, X_n, \mu) = \frac{\bar{X} - \mu}{S/\sqrt{n}} \xrightarrow{d} Y \sim N(0, 1),$$

et nous avons donc trouvé un pivot approximatif. On obtient, maintenant :

$$\begin{aligned} \mathbb{P}\left[\bar{X} - z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}} \leq \mu \leq \bar{X} - z_{\frac{\alpha}{2}} \frac{S}{\sqrt{n}}\right] &= \mathbb{P}\left[z_{\alpha/2} \leq \frac{\bar{X} - \mu}{S/\sqrt{n}} \leq z_{1-\alpha/2}\right] \\ &= \mathbb{P}\left[z_{\alpha/2} \leq g(X_1, \dots, X_n, \mu) \leq z_{1-\alpha/2}\right] \\ &\xrightarrow{n \rightarrow \infty} \mathbb{P}\left[z_{\alpha/2} \leq Y \leq z_{1-\alpha/2}\right] = 1 - \alpha. \end{aligned}$$

Qui donne l'intervalle approximatif  $\bar{X} \pm z_{1-\frac{\alpha}{2}} \frac{S}{\sqrt{n}}$ .

# Pivots Approximatifs pour les familles exponentielles

Que se passe-t-il si l'on s'intéresse pas à une moyenne, mais à un paramètre général ?

On verra qu'il est possible de trouver des pivots approximatifs dans le cas d'une famille exponentielle. Nous considérons deux types d'intervalles de confiance découlant de deux types de pivots :

- 1 Intervalles de Wald.
- 2 Intervalles du rapport de vraisemblance.

## Proposition (Pivots approximatifs de Wald)

Soit  $X_1, \dots, X_n$  un échantillon iid tiré d'une distribution avec une fonction de densité/masse  $f(x; \theta)$  appartenant à une famille exponentielle non-dégénérée,

$$f(x; \theta) = \exp\{\eta(\theta) T(x) - d(\theta) + S(x)\}, \quad x \in \mathcal{X}, \theta \in \Theta.$$

Supposons que

- 1 L'espace des paramètres  $\Theta \subset \mathbb{R}$  est un ensemble ouvert.
- 2 La fonction  $\eta(\cdot)$  est une bijection deux fois continûment dérivable entre  $\Theta$  and  $\Phi = \eta(\Theta)$  telle que  $\eta' \neq 0$ .

Soit  $\hat{\theta}_n$  l'EMV de  $\theta$ , et  $\hat{J}_n = n \frac{d''(\hat{\theta}_n)\eta'(\hat{\theta}_n) - d'(\hat{\theta}_n)\eta''(\hat{\theta}_n)}{[\eta'(\hat{\theta}_n)]}$ . Définissons

$$g(X_1, \dots, X_n, \theta) := \hat{J}_n^{1/2}(\hat{\theta}_n - \theta).$$

Alors

$$g(X_1, \dots, X_n, \theta) \xrightarrow{d} N(0, 1),$$

et  $g(X_1, \dots, X_n, \theta)$  est donc un pivot approximatif pour  $\theta$ .

## Intervalles de confiance approximatifs de Wald

En utilisant la même notation que celle de la proposition précédente, on voit que le tableau suivant contient les intervalles de confiance approximatifs avec seuil  $(1 - \alpha)$  pour  $\theta$  :

Confiance approximative $1 - \alpha$	$L(X_1, \dots, X_n)$	$U(X_1, \dots, X_n)$
Bilatéral	$\hat{\theta} - z_{1-\alpha/2} \hat{J}_n^{-1/2}$	$\hat{\theta} + z_{1-\alpha/2} \hat{J}_n^{-1/2}$
Unilatéral à gauche	$\hat{\theta} - z_{1-\alpha} \hat{J}_n^{-1/2}$	$+\infty$
Unilatéral à droite	$-\infty$	$\hat{\theta} + z_{1-\alpha} \hat{J}_n^{-1/2}$

## Proposition (Pivots approximatifs du rapport de vraisemblance)

Soit  $X_1, \dots, X_n$  un échantillon iid tiré d'une distribution avec un fonction de densité/masse  $f(x; \theta)$  appartenant à une famille exponentielle non-dégénérée,

$$f(x; \theta) = \exp\{\eta(\theta) T(x) - d(\theta) + S(x)\}, \quad x \in \mathcal{X}, \theta \in \Theta.$$

Supposons que :

- 1 L'espace des paramètres  $\Theta \subset \mathbb{R}$  est un ensemble ouvert.
- 2 La fonction  $\eta(\cdot)$  est une bijection deux fois continûment dérivable entre  $\Theta$  and  $\Phi = \eta(\Theta)$  telle que  $\eta' \neq 0$ .

Soient  $\hat{\theta}_n$  l'EMV de  $\theta$ , et  $g(X_1, \dots, X_n, \theta) = 2(\ell(\hat{\theta}) - \ell(\theta))$ . Alors,

$$g(X_1, \dots, X_n, \theta) \xrightarrow{d} \chi_1^2,$$

et  $g(X_1, \dots, X_n, \theta)$  est donc un pivot approximatif pour  $\theta$ .

# Pivots du rapport de vraisemblance

- Notons que le pivot approximatif du rapport de vraisemblance  $g(X_1, \dots, X_n, \theta) = 2(\ell(\hat{\theta}) - \ell(\theta))$  n'a pas nécessairement une forme que l'on peut manipuler afin d'obtenir un intervalle de confiance explicite.
- Cependant, nous pouvons trouver de façon numérique l'intervalle de confiance approximatif, en déterminant l'ensemble

$$\{\theta \in \Theta : g(X_1, \dots, X_n, \theta) \leq q_{1-\alpha}(\chi_1^2)\},$$

où  $q_{1-\alpha}(\chi_1^2)$  set le  $(1 - \alpha)$ -quantile d'une distribution  $\chi_1^2$ .

# La dualité avec les tests d'hypothèse

# La dualité avec les tests d'hypothèse

Il semble y avoir de liens entre les intervalles de confiance et les tests d'hypothèse :

- Estimation par intervalle : trouver une région qui contient le paramètre.  
Tests d'hypothèses, est-ce qu'une région donnée contient le paramètre ?
- Tests d'hypothèses : seuil donné par  $\alpha$ .  
Estimation par intervalle : confiance  $1 - \alpha$ .
- Tests d'hypothèse : tests du rapport de vraisemblance et des tests de Wald.  
Estimation par intervalle : intervalles de Wald et du rapport de vraisemblance.

Est-il possible que nous soyons en train de regarder les deux côtés d'une même pièce de monnaie ?

## Théorème (Théorème de la dualité)

Soient  $X_1, \dots, X_n \stackrel{iid}{\sim} f(x; \theta)$  une variable aléatoire et  $\theta \in \Theta \subseteq \mathbb{R}$ .

- ❶ Si  $[L(X_1, \dots, X_n), U(X_1, \dots, X_n)]$  est un intervalle de confiance bilatéral avec seuil  $(1 - \alpha)$  pour  $\theta$ , alors le test dont la fonction de test est

$$\delta(X_1, \dots, X_n) = 1\{\theta_0 \notin [L(X_1, \dots, X_n), U(X_1, \dots, X_n)]\}$$

est un test de  $\{H_0 : \theta = \theta_0\}$  contre  $\{H_1 : \theta \neq \theta_0\}$  avec un seuil de signification égal à  $\alpha$ .

- ❷ Réciproquement, supposons que pour n'importe quel  $\theta_0 \in \Theta$ ,  $\delta(X_1, \dots, X_n; \theta_0)$  est une fonction de test pour la paire d'hypothèses  $\{H_0 : \theta = \theta_0\}$  et  $\{H_1 : \theta \neq \theta_0\}$  avec une probabilité d'erreur de type I égale à  $\alpha$ . Alors,

$$R(X_1, \dots, X_n) := \{\vartheta \in \Theta : \delta(X_1, \dots, X_n; \vartheta) = 0\}$$

est une région de confiance avec seuil  $(1 - \alpha)$  pour  $\theta$ .

- Lorsque nous suivons la procédure décrite dans la deuxième partie du théorème afin d'obtenir une région  $R$  à partir d'une fonction de test, nous parlons d'inverser un test.
- Notez que dans la partie (2), nous disons que  $R(X_1, \dots, X_n)$  est une région et non un intervalle.
- Pour certaines formes de  $\delta$  et pour certains modèles  $f(x; \theta)$ , la région  $R(X_1, \dots, X_n)$  est bel et bien un intervalle.

## Exemple (Cas Gaussien)

Le TRV au niveau  $\alpha$  pour  $\{H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu \neq \mu_0\}$  dans le cas  $N(\mu, \sigma^2)$  ( $\sigma$  inconnu) était :

$$\delta(X_1, \dots, X_n) = 1 \left\{ \left| \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \right| > t_{\{n-1, 1-\alpha/2\}} \right\}$$

L'intervalle de confiance au niveau  $1 - \alpha$  était :

$$\bar{X} \pm t_{\{n-1, 1-\alpha/2\}} S / \sqrt{n}.$$

# Dualité unilatérale ?

- Pour des résultats unilatéraux, une direction est très facile.
- Si  $(-\infty, U]$  est un intervalle unilatéral à droite avec seuil  $(1 - \alpha)$  pour  $\theta$ , alors  $\delta = 1\{U < \theta_0\}$  est un test avec un seuil  $\alpha$  pour  $\{H_0 : \theta \geq \theta_0\}$  vs  $\{H_1 : \theta < \theta_0\}$
- L'obtention d'un intervalle unilatéral à partir d'un test unilatéral dépend de la forme de la fonction de test ainsi que de la forme du modèle considéré...
- Mais ça marche bien dans le cas d'une famille exponentielle !

## Proposition (Intervalles unilatéraux à partir de tests unilatéraux)

Soit  $X_1, \dots, X_n$  un échantillon aléatoire iid tiré d'une famille exponentielle à 1-paramètre avec une fonction de densité/masse

$$f(x; \theta) = \exp\{\eta(\theta) T(x) - d(\theta) + S(x)\}, \quad x \in \mathcal{X}, \theta \in \Theta \subseteq \mathbb{R},$$

telle que  $\eta(\cdot)$  est strictement croissante et dérivable, et que  $d(\cdot)$  est dérivable. Supposons que  $\tau = \sum_{i=1}^n T(X_i)$  est une variable aléatoire continue, et que sa loi  $\mathbb{P}_\theta[\tau \leq t] = G(t; \theta)$  est continue par rapport à  $\theta$ .

- Soit  $\delta(X_1, \dots, X_n; \theta_0)$  le test UPP de

$$\begin{cases} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{cases}$$

au seuil  $\alpha$ , tel que défini avant. Alors, la région

$$R(X_1, \dots, X_n) = \{\vartheta \in \Theta : \delta(X_1, \dots, X_n; \vartheta) = 0\},$$

est un intervalle unilatéral à gauche avec seuil  $(1 - \alpha)$  de la forme  $[L(X_1, \dots, X_n), +\infty)$ .

- Soit  $\delta(X_1, \dots, X_n; \theta_0)$  le test UPP de

$$\begin{cases} H_0 : \theta \geq \theta_0 \\ H_1 : \theta < \theta_0 \end{cases}$$

au seuil  $\alpha$ , tel que défini avant. Alors, la région

$$R(X_1, \dots, X_n) = \{\vartheta \in \Theta : \delta(X_1, \dots, X_n; \vartheta) = 0\}.$$

est un intervalle unilatéral à droite avec seuil  $(1 - \alpha)$  de la forme  $(-\infty, U(X_1, \dots, X_n)]$ .

Remarques :

- En termes non techniques : sous certaines conditions, inverser un test unilatéral pour une famille exponentielle va nous donner un intervalle de confiance unilatéral.
- Observez de plus qu'il s'agit de tests unilatéraux optimaux pouvant être utilisés afin d'obtenir des intervalles de confiance.
- Puisque les tests sont optimaux, est-ce que les intervalles sont aussi optimaux ? Mais qu'entendons-nous par intervalles de confiance optimaux ?

# Optimalité dans l'estimation par intervalle

# Optimalité dans l'estimation par intervalle

Comment pouvons-nous définir la notion d'optimalité ? Il semble que n'importe quelle définition d'optimalité devrait satisfaire les deux critères suivants :

- ❶ Intuitivement, les intervalles de confiance optimaux devraient être le plus « petit » possible en moyenne, tout en respectant leur seuil de confiance : plus l'intervalle est petit et plus la localisation du paramètre est précise.
- ❷ Mathématiquement, nous avons vu qu'il existe une dualité naturelle entre les intervalles de confiance et les tests d'hypothèse.
  - Ainsi, toute notion d'optimalité pour des intervalles de confiances devrait être duale à la notion d'optimalité pour les tests d'hypothèse.
  - En d'autres mots, inverser un test d'hypothèse optimal devrait nous donner un intervalle de confiance optimal.

Puisque nous avons vu qu'en général il n'y a pas de test optimal pour une paire d'hypothèses bilatérale, le deuxième critère élimine tout espoir d'obtenir un intervalle bilatéral optimal. Qu'en est-il des intervalles unilatéraux ?

## Définition (Intervalles à gauche uniformément plus précis)

Soient  $[L(X_1, \dots, X_n), +\infty)$  et  $[M(X_1, \dots, X_n), +\infty)$  deux intervalles de confiance unilatéraux avec seuil  $(1 - \alpha)$  pour  $\theta$ . Si pour tout  $\theta \in \Theta$ ,

$$\mathbb{P}_\theta[\theta - L \geq \epsilon] \leq \mathbb{P}_\theta[\theta - M \geq \epsilon], \quad \forall \epsilon > 0,$$

alors on dit que  $[L(X_1, \dots, X_n), +\infty)$  est plus précis que  $[M(X_1, \dots, X_n), +\infty)$ .

Si  $[L, +\infty)$  est plus précis que tout intervalle à gauche au seuil  $(1 - \alpha)$ , alors il est appelé l'intervalle de confiance unilatéral à gauche uniformément plus précis

## Définition (Intervalles à droite uniformément plus précis)

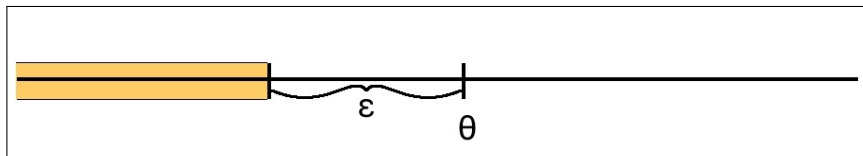
Soient  $(-\infty, U(X_1, \dots, X_n)]$  et  $(-\infty, M(X_1, \dots, X_n)]$  deux intervalles de confiance unilatéraux avec seuil  $(1 - \alpha)$  pour  $\theta$ . Si pour tout  $\theta \in \Theta$ ,

$$\mathbb{P}_\theta[U - \theta \geq \epsilon] \leq \mathbb{P}_\theta[M - \theta \geq \epsilon], \quad \forall \epsilon > 0,$$

alors on dit que  $(-\infty, U(X_1, \dots, X_n)]$  est plus précis que  $(-\infty, M(X_1, \dots, X_n)]$ .

Si  $(-\infty, U]$  est plus précis que tout intervalle à droite au seuil  $(1 - \alpha)$ , alors il est appelé l'intervalle de confiance unilatéral à droite uniformément plus précis.

# Optimality = Tightness



## Intuition

Etant donné que  $L$  tombe à gauche au moins 95% des fois, on veut de plus qu'il soit toujours plus probable que ca soit proche à  $\theta$  que pour une autre borne  $M$  (pour tout  $\theta$ !)

- 1 Nous constatons que notre définition satisfait notre premier critère : intuitivement, la notion d'optimalité est équivalente à la notion de « plus petit » intervalle de confiance.
- 2 La proposition qui suit nous montre qu'elle respecte aussi (au moins pour le cas des familles exponentielles) notre deuxième critère concernant la dualité.

# Intervalles optimales

## Proposition (Tests UPP $\Rightarrow$ intervalles UMA chez les familles exp.)

Soit  $X_1, \dots, X_n$  un échantillon aléatoire iid tiré d'une distribution exponentielle à 1-paramètre avec fonction de densité/masse

$$f(x; \theta) = \exp\{\eta(\theta)T(x) - d(\theta) + S(x)\}, \quad x \in \mathcal{X}, \theta \in \Theta \subseteq \mathbb{R},$$

telle que  $\eta(\cdot)$  est strictement croissante et dérivable, et  $d(\cdot)$  est dérivable. Supposons que  $\tau = \sum_{i=1}^n T(X_i)$  est une variable aléatoire continue dont loi la loi  $\mathbb{P}_\theta[\tau \leq t] = G(t; \theta)$  est continue en  $\theta$ .

Pour n'importe quel  $\theta_0 \in \Theta$ , définissons  $\delta(X_1, \dots, X_n; \theta_0)$  comme étant le test UPP

$$\begin{cases} H_0 : \theta \leq \theta_0 \\ H_1 : \theta > \theta_0 \end{cases}$$

au seuil  $\alpha$ . Alors, la région,

$$R(X_1, \dots, X_n) = \{\vartheta \in \Theta : \delta(X_1, \dots, X_n; \vartheta) = 0\},$$

est un intervalle de confiance unilatéral à gauche uniformément plus précis avec seuil  $(1 - \alpha)$ .