

Statistique pour mathématiciens

Yoav Zemel

(légèrement adapté du cours de V. Panaretos)

Section de Mathématiques – EPFL

yoav.zemel@epfl.ch



Organisation du cours

- Cours lundi 10.15–12.00, CM5
- Exercices mardi 13.15–15.00, MA11
- Référence principale disponible à la Librairie La Fontaine, RLC :
Panaretos, V.M. (2016). *Statistique pour Mathématiciens*. PPUR.
- Page web : Moodle
- Test le 1 mai.
- Examen final écrit.

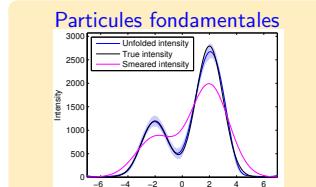
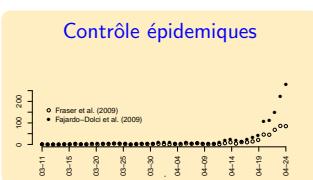
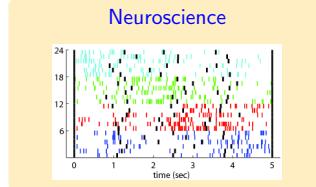
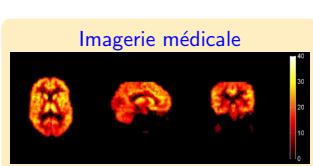
Introduction



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 exprimés mathématiquement sont potentiellement des "données"



Les probabilités nous aident pour la partie incertitude

- C'est la discipline mathématique qui étudie les phénomènes aléatoires (ou *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

- ❶ Nous disposons d'une **distribution** $F(x; \theta)$ qui dépend d'un paramètre **inconnu** $\theta \in \mathbb{R}^p$.
- ❷ 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 le vraie valeur de θ qui a générée les X_i !
- ❸ Nous voulons utiliser les n observations (les réalisations de X_1, \dots, X_n) afin de faire des **affirmations concernant la vraie valeur de θ** , et de quantifier l'incertitude associée à ces affirmations.

Semblé 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.

Yoav Zemel (EPFL)

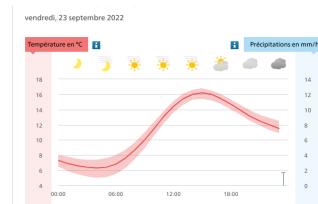
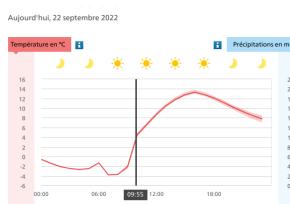
Statistique pour mathématiciens

7 / 132

Yoav Zemel (EPFL) Statistique pour mathématiciens

8 / 132

Intervalles de confiance



Yoav Zemel (EPFL)

Statistique pour mathématiciens

9 / 132

Road Map

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'échantillonage** : 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**.

Yoav Zemel (EPFL)

Statistique pour mathématiciens

10 / 132

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 $X_i \stackrel{iid}{\sim} F$.
- La distribution F dépend typiquement d'un ou de plusieurs paramètres, $\theta = (\theta_1, \dots, \theta_p)^T \in \Theta \subseteq \mathbb{R}^p$ (dépendamment du contexte, une différente lettre grecque ou latine peut être utilisée).
- \mathcal{X} est appelé l'*espace échantillon*, Θ est appelé l'*espace des paramètres*.
- Afin d'indiquer que la distribution F dépend du paramètre θ , nous allons souvent écrire F_θ ou $F(x; \theta)$. Par conséquence : $F(x; \theta) = \mathbb{P}_\theta[X \leq x]$.

Yoav Zemel (EPFL)

Statistique pour mathématiciens

11 / 132

Yoav Zemel (EPFL)

Statistique pour mathématiciens

12 / 132

Modèles probabilistes

Modèles réguliers discrets

Rappel : la **fonction génératrice des moments (FGM)** de X est

$$M(t) = M_X(t) = \mathbb{E}[\exp(tX)] \in (0, \infty], \quad t \in \mathbb{R}.$$

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

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

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

- ❷ 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 considère uniquement de modèles telles que $\mathcal{X} \subseteq \mathbb{Z}$.

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

Loi Bernoulli

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

$$\mathcal{X} = \{0, 1\},$$

$$f(x; p) = p1\{x = 1\} + (1 - p)1\{x = 0\}.$$

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

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

Loi binomiale

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

$$\mathcal{X} = \{0, 1, 2, \dots, n\},$$

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

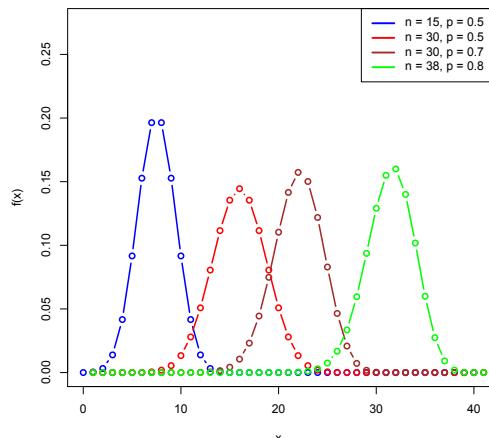
La moyenne, la variance et la fonction génératrice des moments de $X \sim \text{Binom}(n, p)$ sont données par

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

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



Loi géométrique

Définition (Distribution géométrique)

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

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

$$f(x; p) = (1 - p)^x p.$$

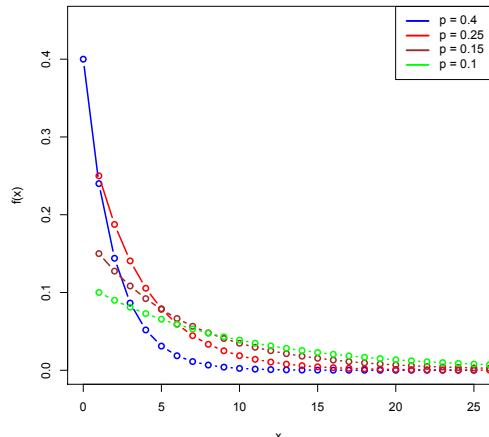
La moyenne, la variance et la fonction génératrice des moments de $X \sim \text{Geom}(p)$ sont données par

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

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

Loi géométrique

Geometric Distribution PMF



Loi binomiale négative

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

$$① \quad \mathcal{X} = \{0\} \cup \mathbb{N},$$

$$② \quad f(x; p, r) = \binom{x+r-1}{x} (1-p)^x p^r.$$

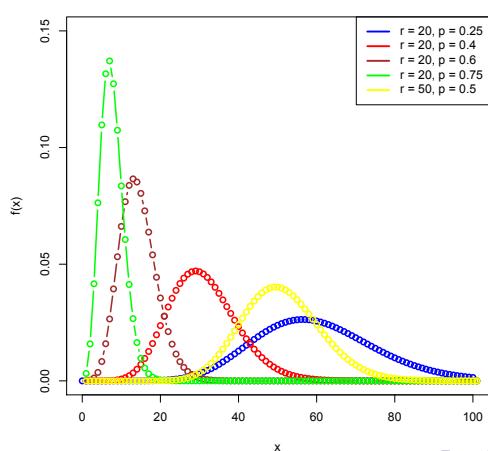
La moyenne, la variance et la fonction génératrice des moments de $X \sim \text{NegBin}(r, p)$ sont données par

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

Si $r \in \mathbb{N}$ et $X = \sum_{i=1}^r Y_i$ avec $Y_i \stackrel{iid}{\sim} \text{Geom}(p) \Rightarrow X \sim \text{NegBin}(r, p)$.

Loi binomiale négative

Negative Binomial Distribution PMF



Loi de Poisson

Définition (Distribution de Poisson)

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

$$① \quad \mathcal{X} = \{0\} \cup \mathbb{N},$$

$$② \quad f(x; \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}.$$

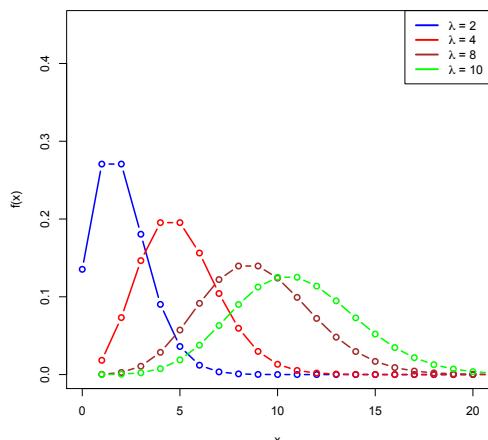
La moyenne, la variance et la fonction génératrice des moments de $X \sim \text{Poisson}(\lambda)$ sont données par

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

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

Loi de Poisson

Poisson Distribution PMF



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.

Loi uniforme

Définition (Distribution uniforme)

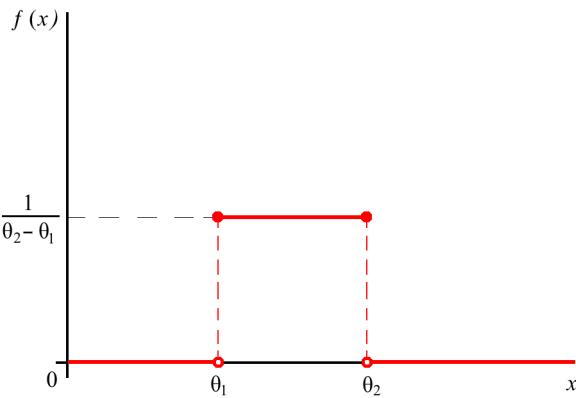
Une variable aléatoire X suit une distribution uniforme de paramètres $-\infty < \theta_1 < \theta_2 < \infty$, noté $X \sim \text{Unif}(\theta_1, \theta_2)$, si

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

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

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

Densité uniforme



Yoav Zemel (EPFL)

Statistique pour mathématiciens

25 / 132

Navigation icons

Loi exponentielle

Définition (Distribution exponentielle)

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

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

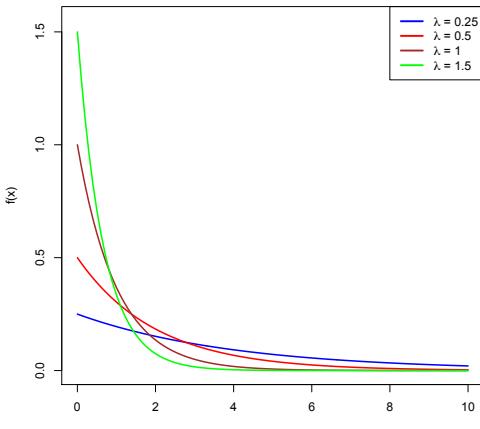
La moyenne, la variance et la fonction génératrice des moments de $X \sim \text{Exp}(\lambda)$ sont données par

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

Navigation icons

Densité exponentielle

Exponential Distribution PDF



Yoav Zemel (EPFL)

Statistique pour mathématiciens

27 / 132

Navigation icons

Loi gamma

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

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

Navigation icons

28 / 132

Loi khi carré (ou khi deux)

Définition (Distribution khi carré)

Une variable aléatoire X suit une distribution khi carré de paramètre $k > 0$ (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

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

Yoav Zemel (EPFL)

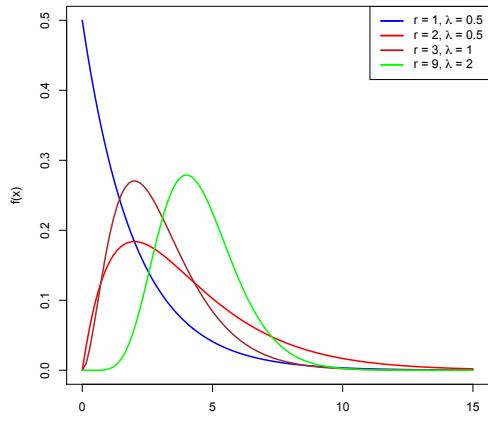
Statistique pour mathématiciens

29 / 132

Navigation icons

Densité gamma

Gamma Distribution PDF



Yoav Zemel (EPFL)

Statistique pour mathématiciens

30 / 132

Navigation icons

Loi normale (ou 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**).

Yoav Zemel (EPFL)

Statistique pour mathématiciens

31 / 132

... et on ne s'arrête jamais !

La liste ne s'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

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

Yoav Zemel (EPFL)

Statistique pour mathématiciens

33 / 132

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 :

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

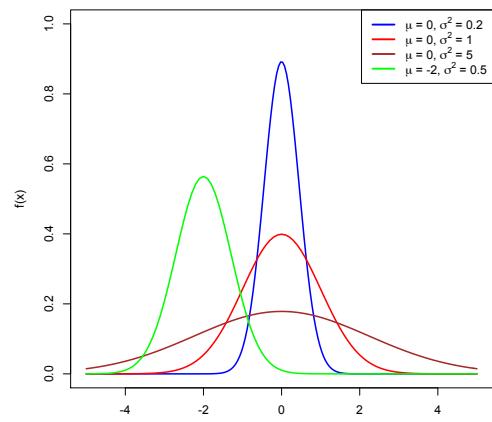
Yoav Zemel (EPFL)

Statistique pour mathématiciens

35 / 132

Densité normale

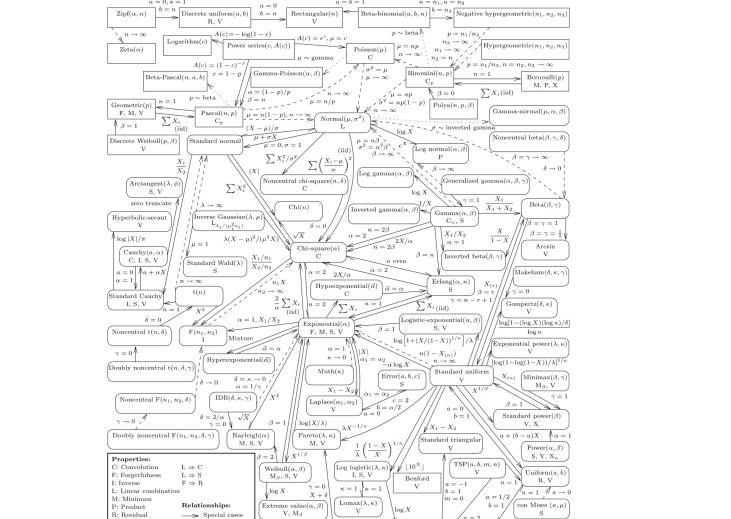
Normal Distribution PDF



Yoav Zemel (EPFL)

Statistique pour mathématiciens

32 / 132



Yoav Zemel (EPFL)

Statistique pour mathématiciens

34 / 132

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 dérivable), telle que

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

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

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

Yoav Zemel (EPFL)

Statistique pour mathématiciens

36 / 132

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 ϕ et on a une famille exponentielle à 1-paramètre. Ici le paramètre usuel est une bijection deux fois dérivable du paramètre naturel ϕ :

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

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 des paramètres. Nous obtenons donc que la distribution $N(\mu, \sigma^2)$ est une famille exponentielle à 2-paramètres. □

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 $R \sim \text{Unif}[a, b]$, 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?

Modèles de probabilité transformés : cas discret

Lemme

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

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

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

• Preuve = enoncé!

• Cas continu : plus compliqué :

① Si g pas monotone : au cas-par-cas.

② Si g est monotone : on a des résultats généraux.

Exemple (La normale standard au carré a une distribution χ^2_1)

Soit $Z \sim N(0, 1)$. Nous voulons trouver la distribution de $Y = Z^2$. Noter que $F_Y(y) = \mathbb{P}[Y \leq y] = \mathbb{P}[|Z| \leq \sqrt{y}] = \mathbb{P}[-\sqrt{y} \leq Z \leq \sqrt{y}]$

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

Nous pouvons aussi trouver la densité en dérivant :

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

Noter que la dernière expression est la densité d'une distribution χ^2_1 . Alors :

$$Z \sim N(0, 1) \implies Z^2 \sim \chi^2_1. \quad (3.3)$$

Modèles de probabilité transformés : cas continu

Lemme

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

$$F_Y(y) = 1 - F_X(g^{-1}(y)) + \mathbb{P}(X = g^{-1}(y)).$$

Dans les deux cas, si X est continue et g est

- ❶ continûment dérivable,

- ❷ et de dérivée jamais nulle,

alors

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

Yoav Zemel (EPFL)

Statistique pour mathématiciens

43 / 132

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.

Yoav Zemel (EPFL) Statistique pour mathématiciens 44 / 132

Modèles transformés : cas continu multidimensionnel

Théorème (Transformations multidimensionnelles)

Soit $g : \mathcal{X}_n \subseteq \mathbb{R}^n \rightarrow \mathbb{R}^n$ injective et continûment dérivable,

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

Soit $X = (X_1, \dots, X_n)^\top$ prenant valeurs dans \mathcal{X}_n et ayant une densité conjointe $f_X(\mathbf{x})$, $\mathbf{x} \in \mathcal{X}_n$, et définissons $Y = (Y_1, \dots, Y_n)^\top = g(X)$. Alors

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

et zero sinon, lorsque $J_{g^{-1}}(\mathbf{y})$ est bien définie (c'est-à-dire quand $\det(J_g(\mathbf{y})) \neq 0$). Ici, $J_{g^{-1}} : \mathcal{Y}_n \rightarrow \mathbb{R}^{n \times n}$ est la matrice Jacobienne de g^{-1} ,

$$J_{g^{-1}}(\mathbf{y}) = \begin{bmatrix} \frac{\partial}{\partial x_1} g_1^{-1}(\mathbf{y}) & \dots & \frac{\partial}{\partial x_n} g_1^{-1}(\mathbf{y}) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} g_n^{-1}(\mathbf{y}) & \dots & \frac{\partial}{\partial x_n} g_n^{-1}(\mathbf{y}) \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

Yoav Zemel (EPFL)

Statistique pour mathématiciens

45 / 132

Yoav Zemel (EPFL)

Statistique pour mathématiciens

46 / 132

Exemple (Convolution de densités)

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

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

Preuve. Définissons $g : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ par $g(x, y) = (x + y, y)$ de sorte que $g^{-1}(u, v) = (u - v, v)$.

La jacobienne de l'inverse est

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

donc le déterminant absolu vaut 1. Il s'ensuit que

$$f_{X+Y,Y}(u, v) = 1 \times f_{X,Y}(u - v, v) = f_X(u - v)f_Y(v),$$

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

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

Yoav Zemel (EPFL)

Statistique pour mathématiciens

47 / 132

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. Montrer 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,

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

Yoav Zemel (EPFL) Statistique pour mathématiciens 48 / 132

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 :

- ➊ la théorie scientifique et des expériences préalables ;
- ➋ des principes philosophiques (parsimonie/rasoir d'Occam, entropie) ;
- ➌ une analyse exploratoire des données ;
- ➍ une combinaison de (1), (2) et (3).

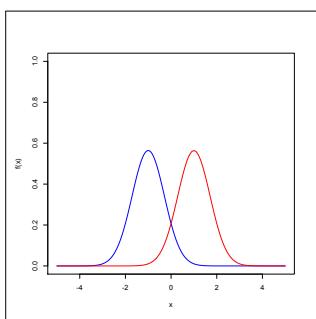
Analyse exploratoire des données

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

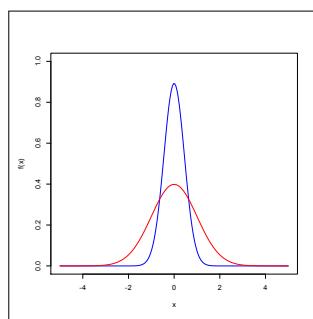
Si on a observations x_1, \dots, x_n , on peut les utiliser pour choisir entre plusieurs choix, ou au moins exclure certains choix.

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

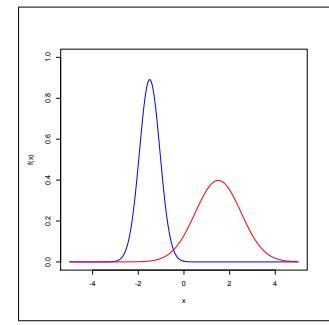
- ➊ Position.
- ➋ Dispersion.
- ➌ Comportement des queues.
- ➍ Symétrie / asymétrie.



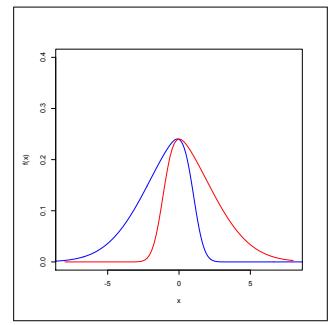
(a) Deux densités de positions différentes.



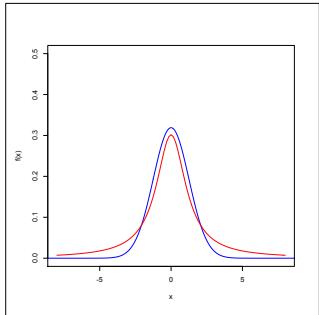
(b) Deux densités de dispersions différentes.



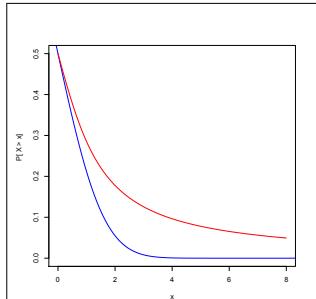
(c) Deux densités qui diffèrent par leur position et leur dispersion.



(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) Les fonctions $x \mapsto \int_x^\infty f(y) dy$ pour les deux densités de gauche.

Résumés numériques : centre

Définition (Moyenne et médiane empiriques.)

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

- ① La moyenne empirique comme suit

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

- ② La médiane empirique comme suit

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

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

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

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

- ② La déviation absolue par rapport à la moyenne (DAM)

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

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

- ① Numériques.

- ② 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^{e} valeur de l'échantillon, lorsque ces valeurs sont placées en ordre croissant (tel que $x_{(1)} = \min\{x_1, \dots, x_n\}$ et $x_{(n)} = \max\{x_1, \dots, x_n\}$). 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 $x_{(1)} = x_3$, $x_{(2)} = x_1$, $x_{(3)} = x_4 = x_2 = x_4$.

Moyenne et médiane

$$\bar{x} = \operatorname{argmin}_{\gamma \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n (x_i - \gamma)^2, \quad M \in \operatorname{argmin}_{\gamma \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n |x_i - \gamma|$$

Résumés numériques : queues

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 $[Q_1 - \frac{3}{2} EIQ, Q_3 + \frac{3}{2} EIQ]$.

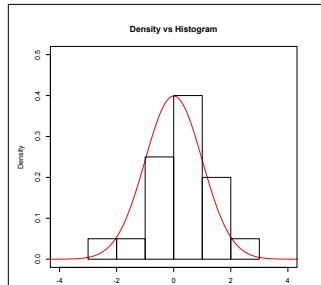
Résumés numériques : symétrie/asymétrie

Définition (Coefficient d'asymétrie empirique)

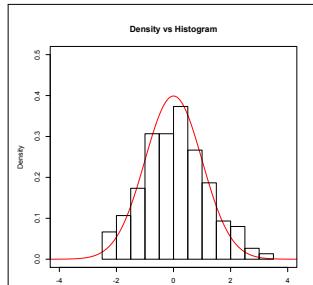
Soit x_1, \dots, x_n un échantillon de n valeurs réelles. Nous définissons le coefficient d'asymétrie de cet échantillon comme

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

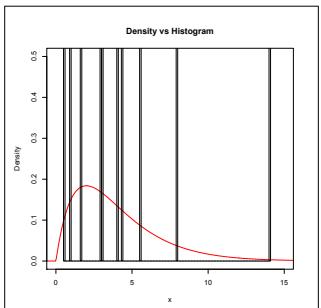
Si le numérateur et le dénominateur sont égaux à zéro, c'est-à-dire si $x_1 = \dots = x_n$ (ce qui peut se produire dans un échantillon discret), alors SK n'est pas défini.



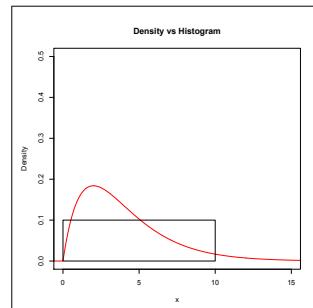
(g) Densité d'une $N(0, 1)$ (en rouge) et l'histogramme d'un échantillon aléatoire de taille 20 tiré d'une $N(0, 1)$ (en noir).



(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).



(k) Densité d'une χ_2^2 (en rouge) et l'histogramme d'un échantillon aléatoire de taille 20 tiré d'une χ_2^2 (en noir) lorsque la largeur des intervalles h est très petite.



(l) Densité d'une χ_2^2 (en rouge) et l'histogramme d'un échantillon aléatoire de taille 20 tiré d'une χ_2^2 (en noir) lorsque la largeur des intervalles h est très grande.

Résumés graphiques : histogrammes

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 $h > 0$,

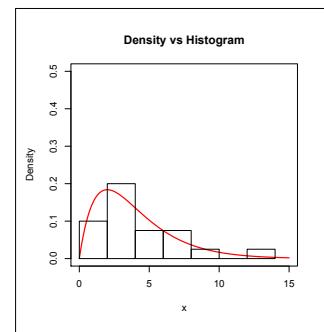
$$I_j = [\kappa + (j-1)h, \kappa + jh), \quad j \in \mathbb{Z}$$

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 κ est défini comme étant le graphique de la fonction :

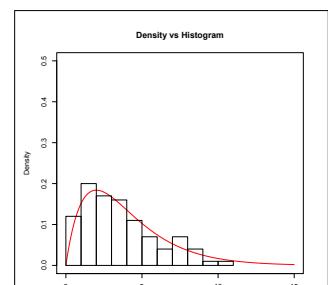
$$y \mapsto \text{hist}_{x_1, \dots, x_n}(y) = \frac{1}{h} \sum_{j \in \mathbb{Z}} 1\{y \in I_j\} \frac{1}{n} \sum_{i=1}^n 1\{x_i \in I_j\}.$$

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 .
- $\mathbb{E} \left[\int_{I_j} \text{hist}_{x_1, \dots, x_n}(y) dy \right] = \frac{1}{n} \sum_{i=1}^n \mathbb{P}[X_i \in I_j] = \int_{I_j} f(y) dy$.



(i) Densité d'une χ_2^2 (en rouge) et l'histogramme d'un échantillon aléatoire de taille 20 tiré d'une χ_2^2 (en noir).



(j) Densité d'une χ_2^2 (en rouge) et l'histogramme d'un échantillon aléatoire de taille 100 tiré d'une χ_2^2 (en noir).

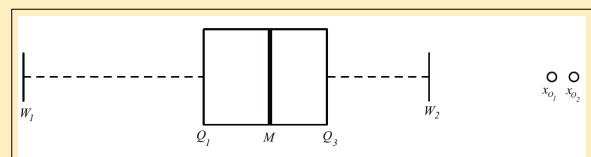
Résumés graphiques : boxplot

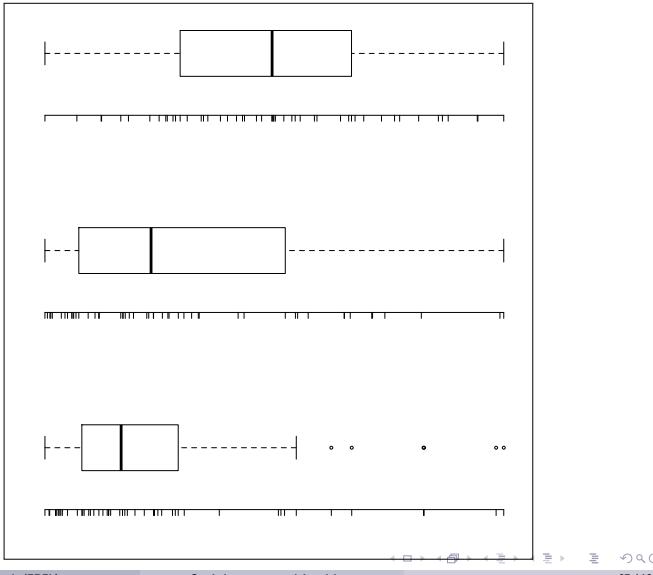
Définition (Boîte à moustaches (anglais : boxplot))

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 $W_1 = \min_{1 \leq j \leq n} \{x_j : x_j \geq Q_1 - 1.5 \times EIQ\}$ & $W_2 = \max_{1 \leq j \leq n} \{x_j : x_j \leq Q_3 + 1.5 \times EIQ\}$.
- 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 :





Echantillonage

Retour au cadre général

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

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*.

Statistiques exhaustives

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 θ si

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

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

- ➌ Si une telle statistique existe, la seule connaissance de T suffit pour faire des inférences sur θ .

Statistique

Définition (Statistique)

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

- ➌ Une statistique $T : \mathcal{X}^n \rightarrow \mathbb{R}$ réduit une collection de n nombres à une seule valeur.
- ➍ 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 θ que (X_1, \dots, X_n) .

Statistiques exhaustives

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

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

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

- ➌ 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

Critère de Fisher–Neyman

Comment vérifier qu'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 θ si et seulement s'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 $X_1, \dots, X_n \stackrel{iid}{\sim} Bern(p)$. Si un des $x_i \notin \{0, 1\}$, la fonction de masse est 0. Si $x_i \in \{0, 1\}$ pour chaque i alors

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

Ainsi, le critère de Fisher–Neyman est satisfait avec

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

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

$$h(x_1, \dots, x_n) = \prod_{i=1}^n 1\{x_i \in \{0, 1\}\}.$$

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

Critère de Fisher–Neyman : preuve (cas discret)

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

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 fonction de répartition

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

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 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)

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

- ④ 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)

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

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

(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)

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

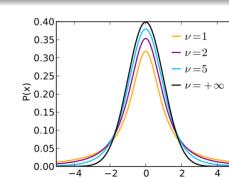
Une variable aléatoire X suit une distribution t de Student de paramètre $k > 0$ (appelé nombre de degrés de liberté), noté $X \sim t_k$, si

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

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

$$\mathbb{E}[X] = \begin{cases} 0 & k > 1 \\ \text{indéfinie} & k \leq 1 \end{cases}, \quad \text{Var}[X] = \begin{cases} \frac{k}{k-2} & k > 2 \\ \text{indéfinie} & k \leq 2 \end{cases}$$

La FGM $M_X(t)$ est infinie pour tout $t \neq 0$ et tout $k > 0$.



Echantillonage de familles exponentielles

Que se passerait-il si la distribution à partir de laquelle nous échantillonons 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 ? Soient $X_1, \dots, X_n \stackrel{iid}{\sim} f$, où

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

- ➊ Est-il possible de trouver la distribution conjointe de l'échantillon (X_1, \dots, X_n) ?
- ➋ Est-il possible de trouver les moments exacts de certaines statistiques clés ?
- ➌ Est-il possible de trouver la distribution d'échantillonage exacte de certaines statistiques importantes ?

Yoav Zemel (EPFL)

Statistique pour mathématiciens

85 / 132

Proposition (Echantillonage d'une famille exponentielle)

Soient $X_1, \dots, X_n \stackrel{iid}{\sim} f$, 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 :

- ➊ 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 $\tau(x_1, \dots, x_n) = \sum_{i=1}^n T(x_i)$. La statistique τ est donc exhaustive pour ϕ .

- ➋ Si Φ est ouvert, alors γ est infiniment dérivable, et en plus

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

Yoav Zemel (EPFL)

Statistique pour mathématiciens

86 / 132

Distributions d'Echantillonage Approximative

Distributions d'Echantillonage Approximative

La distribution d'échantillonage 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 → 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 »

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

Yoav Zemel (EPFL)

Statistique pour mathématiciens

87 / 132

Convergence en loi/distribution (ou convergence faible)

Définition (Convergence en loi (ou convergence faible))

Soit $\{F_n\}_{n \geq 1}$ une suite de fonctions de répartition et G une fonction de répartition sur \mathbb{R} . On dit que F_n converge en loi vers G (noté $F_n \xrightarrow{d} G$) si

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

pour tout x point de continuité de G ($\lim_{h \rightarrow 0} G(x+h) = G(x)$).

Si G est continue, la convergence est uniforme.

Exemple (Le maximum de variables aléatoires uniformes)

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

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

Noter que la limite est la fonction de répartition d'une variable aléatoire $\text{Exp}(1)$. □

Yoav Zemel (EPFL)

Statistique pour mathématiciens

90 / 132

Convergence en loi : commentaires

❶ Convergence en loi \equiv convergence ponctuelle de la suite de fonctions de répartition, sauf qu'il n'est pas nécessaire d'avoir une convergence ponctuelle aux points de discontinuité de la limite.

❷ Lorsque $F_n(x) = \mathbb{P}[X_n \leq x]$ pour une suite 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

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

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

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

Yoav Zemel (EPFL) Statistique pour mathématiciens 91 / 132



91 / 132

Distributions approximatives d'échantillonage

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

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

et remarquons que c'est une variable aléatoire qui :

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

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

Yoav Zemel (EPFL) Statistique pour mathématiciens 93 / 132



93 / 132

Distribution approximative d'échantillonage pour familles exponentielles

Corollaire

Soient $X_1, \dots, X_n \xrightarrow{iid} f$, 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

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

Si Φ est ouvert, alors (γ est doublement dérivable et)

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

Yoav Zemel (EPFL) Statistique pour mathématiciens 95 / 132



95 / 132

Convergence en probabilité

Définition

Lorsqu'une suite 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 suite telle que $X_n \xrightarrow{d} c$, nous avons :

Lemme

Soient $\{X_n\}_{n \geq 1}$ une suite 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.

Yoav Zemel (EPFL) Statistique pour mathématiciens 92 / 132



92 / 132

Les deux grands théorèmes

Théorème (Loi faible des grands nombres)

Soient 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 $\mathbb{E}|Y_i| < \infty$.

Théorème (Théorème central limite)

Soient 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

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

Yoav Zemel (EPFL) Statistique pour mathématiciens 94 / 132



94 / 132

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 sur $\mathcal{X} \times c$, 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)

Noter 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 sur \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 $a_n \uparrow \infty$. Si $g : \mathbb{R} \rightarrow \mathbb{R}$ est dérivable en θ , alors $a_n(g(X_n) - g(\theta)) \xrightarrow{d} g'(\theta)Z$.

Yoav Zemel (EPFL) Statistique pour mathématiciens 96 / 132



96 / 132

Nouveaux théorèmes limites à partir de 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 μ et de variances $\sigma^2 < \infty$, alors $\sqrt{n}(\bar{Y}_n - \mu) \xrightarrow{d} N(0, \sigma^2)$.

➊ 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 toute fonction g dérivable au point μ .

➋ Maintenant considérons une suite de variables aléatoires W_n 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

➊ Il y a une distribution $F(x; \theta)$ qui dépend d'un paramètre inconnu $\theta \in \mathbb{R}^p$.

➋ 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 θ qui a généré les X_i !

➌ **Problème d'estimation ponctuelle :** Comment utiliser les n observations (les réalisations de X_1, \dots, X_n) afin de déterminer la vraie valeur de θ .

Comment ? Mais avec un estimateur, bien sûr !

Définition (Estimateur ponctuel)

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

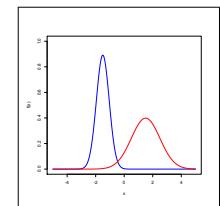
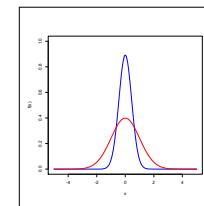
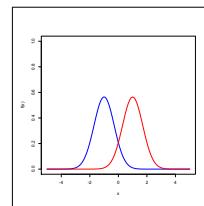
Remarque

Puisque l'objectif d'un estimateur est de fournir une estimation du vrai θ qui a généré les données, nous le dénotons typiquement $\hat{\theta}$. Noter de plus que θ 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 Θ pourrait être un estimateur.
- Laquelle devons-nous choisir ?

Critères pour comparer des estimateurs



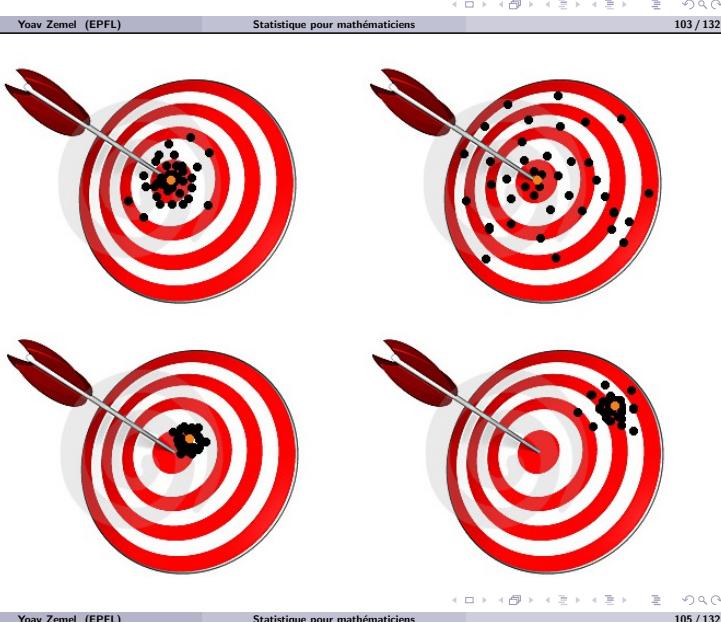
Critères pour comparer des estimateurs

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 ?

- ❶ Interprétation facile.
- ❷ Théorème centrale limite.
- ❸ Inégalités de concentration

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



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 $0 = EQM(S, 0) < EQM(T, \theta)$ pour tout $T \neq S$ – aucun autre estimateur peut battre S à cet endroit de l'espace Θ .
 - Même une montre cassée donne l'heure exacte deux fois par jour...

Erreur quadratique moyenne

Définition (Erreur quadratique moyenne)

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

$$EQM(\hat{\theta}, \theta) = \mathbb{E}_\theta \left[(\hat{\theta} - \theta)^2 \right] \in [0, \infty].$$

Lemme (Décomposition biais-variance)

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

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

Concentration et EQM

Lemme

Soit $\hat{\theta}$ un estimateur de $\theta \in \mathbb{R}^p$. Alors, pour tout $\epsilon > 0$,

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

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

Remarque

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

Borne de Cramér–Rao

Théorème (Borne de Cramér–Rao)

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

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

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

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

- a. C'est-à-dire \mathcal{X} ne dépend pas de θ

- On appelle la quantité $\int_{\mathcal{X}} \left(\frac{\partial}{\partial \theta} \log f(x; \theta) \right)^2 f(x; \theta) dx = \mathbb{E} \left(\frac{\partial}{\partial \theta} \log f(X_1; \theta) \right)^2$ l'information de Fisher, $I(\theta) \geq 0$.
- Si $I(\theta) = 0$ alors $\beta'(\theta) + 1 = 0$ et le théorème ne dit rien.
- Même si le biais est égal à zéro, la variance sera bornée inférieurement par $1/I(\theta)$.

Cramér–Rao dans les familles exponentielles

Examinons les conditions

$$\begin{aligned} \textcircled{1} \quad & \frac{\partial}{\partial \theta} \left[\int_{\mathcal{X}^n} \widehat{\theta}(x_1, \dots, x_n) f_{X_1, \dots, X_n}(x_1, \dots, x_n; \theta) dx \right] = \int_{\mathcal{X}^n} \widehat{\theta}(x_1, \dots, x_n) \frac{\partial}{\partial \theta} f_{X_1, \dots, X_n}(x_1, \dots, x_n; \theta) dx. \\ \textcircled{2} \quad & \frac{\partial}{\partial \theta} \left[\int_{\mathcal{X}^n} f_{X_1, \dots, X_n}(x_1, \dots, x_n; \theta) dx \right] = \int_{\mathcal{X}^n} \frac{\partial}{\partial \theta} f_{X_1, \dots, X_n}(x_1, \dots, x_n; \theta) dx. \end{aligned}$$

dans le cas d'une famille exponentielle (remplaçons θ par ϕ).

Proposition

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

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

avec $\phi \in \Phi \subseteq \mathbb{R}$ ouvert. Alors (2) est satisfaite. Si $\text{var}_\phi(\widehat{\phi}(X_1, \dots, X_n)) < \infty$ alors (1) est également satisfaite.

Comme la borne de Cramér–Rao parle de la variance de $\widehat{\phi}$ ce n'est pas grave de la supposer finie...

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, une fois l'échantillon est observé, les θ plausibles sont ceux qui rendent notre échantillon le plus probable possible...

Yoav Zemel (EPFL)

Statistique pour mathématiciens

115 / 132

Maximum de vraisemblance : cas discret

Lorsque θ 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_θ 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 θ .

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

Yoav Zemel (EPFL)

Statistique pour mathématiciens

117 / 132

Définition générale

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 θ 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_θ 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 θ .

Yoav Zemel (EPFL)

Statistique pour mathématiciens

119 / 132

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 θ est définie par

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

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

Remarques :

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

Yoav Zemel (EPFL)

Statistique pour mathématiciens

116 / 132

Maximum de vraisemblance : cas continu

Et le cas continu ? On utilise la même définition, avec la densité au lieu de la fonction de masse, même si on va perdre l'interprétation en 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 θ est définie par

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

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

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.

Yoav Zemel (EPFL)

Statistique pour mathématiciens

118 / 132

Détermination de l'EMV — La logVraisemblance

- Il est souvent plus simple de maximiser la log-vraisemblance

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

(équivalent parce que \log est monotone et $f > 0$) car on travaille avec une somme et non pas un produit

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

- Lorsque ℓ est une fonction dérivable de $\theta \in \mathbb{R}^p$, le maximum de la fonction doit être une solution de l'équation

$$\nabla_\theta \ell(\theta) = 0.$$

- 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

$$-\nabla_\theta^2 \ell(\theta)|_{\theta=\hat{\theta}} > 0,$$

c'est-à-dire que (-1) multiplié par la matrice hessienne est définie positive.

- Lorsque $\theta \in \mathbb{R}$ (paramètre en dimension un) : il suffit $\ell'(\theta) = 0 > \ell''(\theta)$

Yoav Zemel (EPFL)

Statistique pour mathématiciens

120 / 132

Exemple (EMV pour la loi de Bernoulli)

Soient $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Bern}(p)$ 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 a une unique racine donnée par $\frac{1}{n} \sum_{i=1}^n X_i = \bar{X}$. Appelons cette racine \hat{p} , nous devons maintenant vérifier qu'elle correspond bien à un maximum. Noter 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

$$\hat{p} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i \in [0, 1]$$

est l'unique EMV de p . □

Exemple (EMV pour la loi exponentielle)

Soient $X_1, \dots, X_n \stackrel{iid}{\sim} \text{Exp}(\lambda)$ 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 λ 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 λ nous donne l'unique racine

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

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

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

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

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

est l'unique EMV de λ . □

Par l'inégalité de Jensen, c'est un estimateur biaisé.

Exemple (EMV pour la loi gaussienne)

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

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

Les dérivées secondes par rapport à μ et σ^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 (μ, σ^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. Noter 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 dérivées secondes en $(\hat{\mu}, \hat{\sigma}^2)$, nous obtenons

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

Exemple

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

Nous obtenons que la matrice

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

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 (μ, σ^2) est donné par

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

Equivariance de l'EMV

- Il y a des situations où nous ne sommes pas intéressés à estimer θ , 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)

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

Exemple

Soient $X_1, \dots, X_n \stackrel{iid}{\sim} \mathcal{N}(\mu, 1)$, 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

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

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

Notons que l'estimateur EMV de σ^2 est biaisé.

Equivariance de l'EMV

Exemple (Paramètre usuel vs naturel dans les familles exponentielles)

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

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

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 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 un EMV de θ , alors $\eta(\hat{\theta})$ est un EMV de $\phi = \eta(\theta)$. Le réciproque est lui aussi vrai : si $\hat{\phi}$ est un EMV de ϕ , alors $\eta^{-1}(\hat{\phi})$ est un 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 ϕ est unique lorsqu'il existe, et est donné par l'unique solution par rapport à u de l'équation

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

Ici,

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