Modélisation statistique

Université Hassiba Benbouali de Chlef

Problématique statistique

- Point de départ : des observations (des nombres réels)

$$x_1, \dots, x_n.$$

- Modélisation statistique :
  - les observations sont des réalisations

$$X_1(\omega), \dots, X_n(\omega) \text{ de v.a.r. } X_1, \dots, X_n.$$

  - La loi $\mathbb{P}^{(X_1,\dots,X_n)}$ de $(X_1, \dots, X_n)$ est inconnue, mais appartient à une famille donnée

$$\boxed{\{\mathbb{P}_\theta^n, \theta \in \Theta\}.}$$

- Problématique : à partir de « l'observation » $X_1, \dots, X_n$, peut-on retrouver $\mathbb{P}_\theta^n$? et donc θ ?

Problématique statistique (suite)

- θ est le paramètre et Θ l'ensemble des paramètres.
- Estimation : à partir de $X_1, \ldots, X_n$, construire $\varphi_n(X_1, \ldots, X_n)$ qui « approche au mieux » θ.
- Test : à partir de $X_1, \ldots, X_n$, établir une décision $\varphi_n(X_1, \ldots, X_n) \in \{\text{ensemble de décisions}\}$ concernant θ pouvant être vraie ou fausse.

Expérience statistique

- Un modèle statistique est un objet mathématique associé à l'observation de données issues d'un phénomène aléatoire.
- Une expérience statistique consiste à recueillir une observation x d'un élément aléatoire X, à valeurs dans un espace $\mathcal{X}$ et dont on ne connaît pas exactement la loi de probabilité $\mathbb{P}$.

Modèle statistique

Définition 2.1

Le modèle statistique (ou la structure statistique) associé à cette expérience est le triplet $(\mathcal{X}, \mathcal{A}, \mathcal{P})$, où :

- $\mathcal{X}$ est l'espace des observations, ensemble de toutes les observations possibles.
- $\mathcal{A}$ est la tribu des évènements observables associée.
- $\mathcal{P}$ est une famille de lois de probabilités possibles définie sur $\mathcal{A}$.

Modèle statistique (suite)

Exemples

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

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

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

$$\mathcal{X} = \mathbb{R} \quad \mathcal{A} = \mathcal{B}(\mathbb{R}) \quad \mathcal{P} = \{\mathcal{N}(m,\sigma^2),\ (m,\sigma^2) \in \mathbb{R}\times]0,+\infty[\}$$

Modèle statistique (suite)

L'intérêt de cette notion de modèle statistique est qu'elle permet de traiter avec le même formalisme tous les types d'observations possibles.

- On dit que le modèle est discret quand $\mathcal{X}$ est fini ou dénombrable. Dans ce cas, la tribu $\mathcal{A}$ est l'ensemble des parties de $\mathcal{X}$: $\mathcal{A} = \mathcal{P}(\mathcal{X})$. C'est le cas quand l'élément aléatoire observé X a une loi de probabilité discrète.
- On dit que le modèle est continu quand $\mathcal{X} \subset \mathbb{R}^p$ et $\forall \mathbb{P} \in \mathcal{P}$, $\mathbb{P}$ admet une densité (par rapport à la mesure de Lebesgue) dans $\mathbb{R}^p$. Dans ce cas, $\mathcal{A}$ est la tribu des boréliens de $\mathcal{X}$ (tribu engendrée par les ouverts de $\mathcal{X}$) : $\mathcal{A} = \mathcal{B}(\mathcal{X})$.

Modèle paramétrique ou non paramétrique

- Un modèle paramétrique est un modèle où l'on suppose que le type de loi de X est connu, mais qu'il dépend d'un paramètre θ inconnu, de dimension d. Alors, la famille de lois de probabilité possibles pour X peut s'écrire $\mathcal{P} = \left\{\mathbb{P}_\theta,\ \theta \in \Theta \subset \mathbb{R}^d\right\}$.
- Un modèle non paramétrique est un modèle où $\mathcal{P}$ ne peut pas se mettre sous la forme ci-dessus. Par exemple, $\mathcal{P}$ peut être :
  - l'ensemble des lois de probabilité continues sur $\mathbb{R}$,
  - l'ensemble des lois de probabilité dont le support est $[0, 1]$,
  - l'ensemble des lois de probabilité sur $\mathbb{R}$ symétriques par rapport à l'origine,
  - etc ...

Théorème de Radon-Nikodym

Théorème 2.2

Si $\nu \ll \mu$, il existe une fonction positive

$$x \mapsto p(x) \overset{notation}{=} \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(x),$$

définie $\mu-$p.p., $\mu-$ intégrable, telle que

$$\nu\left[A\right] = \int_A p(x)\mu(\mathrm{d}x) = \int_A \frac{\mathrm{d}\nu}{\mathrm{d}\mu}(z)\mu(\mathrm{d}x), \quad A \in \mathcal{A}.$$

Modèle dominée

Définition 2.3

Un modèle statistique $(\mathcal{X}, \mathcal{A}, \{\mathbb{P}_\theta, \theta \in \Theta\})$ est dominée par la mesure σ-finie μ définie sur $\mathcal{A}$ si

$$\forall \theta \in \Theta : \mathbb{P}_\theta \ll \mu.$$

On appelle densités de la famille $\{\mathbb{P}_\theta, \theta \in \Theta\}$ la famille de fonctions (définies $\mu-$ p.p.)

$$x \mapsto \frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mu}(x), \ x \in \mathcal{X}, \ \theta \in \Theta.$$

Modèle identifiable

Définition 2.4

Soit $(\mathcal{X}, \mathcal{A}, \{\mathbb{P}_\theta, \theta \in \Theta\})$ un modèle statistique, il est dit identifiable si l'application $\theta \longmapsto \mathbb{P}_\theta$ définie sur Θ est injective, c'est-à-dire deux paramètres différents correspondent à deux lois distinctes.

Le modèle gaussien

Le modèle $\{\mathcal{N}(m, \sigma^2) \;:\; m \in \mathbb{R},\; \sigma \in]0, +\infty[\}$ est identifiable. Par contre, le modèle alternatif $\{\mathcal{N}(m, \sigma^2) \;:\; m \in \mathbb{R},\; \sigma \neq 0\}$ ne l'est pas puisque $\mathcal{N}\left(m, \sigma^2\right) = \mathcal{N}\left(m, (-\sigma)^2\right)$.

Définition 2.5

Une application $\mathcal{L} \ : \ (\theta, x) \to \mathbb{R}^+$ telle que, pour tout $\theta \in \Theta$, $x \mapsto \mathcal{L}(\theta, x)$ est une densité de $\mathbb{P}_\theta$ relativement à μ : $\mathcal{L}(\theta, \cdot) = \dfrac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mu}$, est appelée une vraisemblance du modèle.

Cas fondamentaux

- mesure de Lebesgue sur $\mathbb{R}^p$: $\mathcal{L}(\theta, \cdot)$ est une densité de probabilité usuelle
- mesure de comptage sur un ensemble dénombrable : $\mathcal{L}(\theta, \cdot)$ est la probabilité pour que x soit observé

Définition 2.6

Un modèle statistique $(\mathcal{X}, \mathcal{A}, \{\mathbb{P}_\theta, \theta \in \Theta\})$ est homogène s'il existe une mesure dominante μ telle que la vraisemblance $\mathcal{L}(\theta, x)$ associée est strictement positive pour μ-presque tout x (ou de manière équivalente, $\forall(\theta, \theta') \in \Theta^2$, $\mathbb{P}_\theta \ll \mathbb{P}_{\theta'}$).

Famille exponentielle

Une classe importante de modèles statistiques est la classe des modèles de la famille exponentielle.

Définition 2.7

On dit que la famille $\{\mathbb{P}_\theta,\ \theta \in \Theta\}$ est une famille exponentielle de dimension p relativement à la mesure dominante μ s'il existe des applications mesurables :

- $\alpha\ :\ \Theta \longrightarrow \mathbb{R}^p, \quad \beta\ :\ \Theta \longrightarrow \mathbb{R}^+$
- $T\ :\ \mathcal{X} \longrightarrow \mathbb{R}^p, \quad \xi\ :\ \mathcal{X} \longrightarrow \mathbb{R}^+$

telles qu'une vraisemblance du modèle statistique s'écrive :

$$\forall(\theta, x) \in \Theta \times \mathcal{X}, \quad \mathcal{L}(\theta, x) = \beta(\theta)\xi(x) \exp\left\{\langle T(x), \alpha(\theta)\rangle\right\} \qquad (1)$$

Famille exponentielle (suite)

Une telle écriture impose que :

$$C_\mu(\alpha(\theta)) := \int_{\mathcal{X}} \xi(x) \exp\left\{\langle T(x), \alpha(\theta)\rangle\right\} \mathrm{d}\mu(x) < +\infty$$

Notons que : $\beta(\theta) = \frac{1}{C_\mu(\alpha(\theta))}$

- T est une statistique naturelle.
- $\alpha(\theta)$ est le paramètre naturel.
- l'espace naturel des paramètres est l'ensemble $N_\mu := \{a \in \mathbb{R}^p,\ C_\mu(a) < +\infty\}$.

Exemples

Le modèle Binomiale

$\mathcal{X} = \{0, \dots, n\}$, la mesure dominante μ est $\sum_{x=0}^{n} \delta_x$, $\Theta = [0, 1]$ et on a

$$\mathcal{L}(p, x) = C_n^x p^x (1-p)^{n-x} = (1-p)^n C_n^x \exp\left\{ x \ln\left(\frac{p}{1-p} \right) \right\}.$$

La loi Binomiale appartient à la famille exponentielle avec $\beta(p) = (1-p)^n$, $\xi(x) = C_n^x$, $T(x) = x$ et $\alpha(p) = \ln\left(\frac{p}{1-p} \right)$.

Exemples (suite)

Le modèle de Poisson

$\mathcal{X} = \mathbb{N}$, la mesure dominante μ est $\mu = \sum_{k \in \mathbb{N}} \delta_k$, $\Theta =]0, \infty[$ et on a

$$\mathcal{L}(\lambda, x) = \frac{\lambda^x \mathrm{e}^{-\lambda}}{x!} = \mathrm{e}^{-\lambda} \frac{1}{x!} \exp\{x \ln \lambda\}.$$

La loi de poisson appartient à la famille exponentielle avec $\beta(\lambda) = \mathrm{e}^{-\lambda}$, $\xi(x) = \frac{1}{x!}$, $T(x) = x$ et $\alpha(\lambda) = \ln \lambda$.

Exemples (suite)

Le modèle normal $\mathcal{N}(m, \sigma^2)$

$\theta = (m, \sigma^2) \in \Theta = \mathbb{R}\times]0, +\infty[$ et μ la mesure de Lebesgue sur $\mathbb{R}$:

$$\begin{aligned} \mathcal{L}(\theta, x) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\{-\frac{1}{2\sigma^2}(x-m)^2\} \\ &= \beta(\theta)\xi(x) \exp\left\{\langle T(x), \alpha(\theta)\rangle\right\} \end{aligned}$$

avec $\beta(\theta) = \dfrac{\mathrm{e}^{-\frac{m^2}{2\sigma^2}}}{\sqrt{2\pi\sigma^2}}$, $\xi(x) = 1$, $T(x) = (x, x^2)$ et $\alpha(\theta) = \left(\dfrac{m}{\sigma^2}, -\dfrac{1}{2\sigma^2}\right)$.

Exemples (suite)

Le modèle Gamma

$\mathcal{X} = \mathbb{R}^+$, la mesure dominante μ est la mesure de Lebesgue sur $\mathbb{R}^+$, $\theta = (a, \lambda) \in \Theta =]0, \infty[^2$.

$$\mathcal{L}(\theta, x) = \frac{\lambda^a}{\Gamma(a)} \exp\left\{-\lambda x + (a-1)\ln x\right\}.$$

La loi de Gamma $\Gamma(a, \lambda)$ appartient à la famille exponentielle avec : $\beta(\theta) = \frac{\lambda^a}{\Gamma(a)}$, $\xi(x) = 1$, $T(x) = (x, \ln x)$ et $\alpha(\theta) = (-\lambda, a-1)$.

Exemples (suite)

Le modèle de Weibull

$$\mathcal{L}(\theta, x) = \frac{\lambda x^{\lambda-1}}{\eta^\lambda} \mathrm{e}^{-\left(\frac{x}{\eta}\right)^\lambda} = \frac{\lambda}{\eta^\lambda} \exp\left\{-\frac{x^\lambda}{\eta^\lambda} + (\lambda - 1)\ln x\right\}$$

Le terme x^λ fait que $\frac{x^\lambda}{\eta^\lambda}$ ne peut pas être mis sous la forme $T(x)\alpha(\eta, \lambda)$, donc la loi de Weibull n'appartient pas à la famille exponentielle.

Forme canonique

A partir de l'écriture (1) de la vraisemblance, on peut considérer la nouvelle mesure dominante $\nu = \xi \cdot \mu$ et la vraisemblance devient

$$\forall(\theta, x) \in \Theta \times \mathcal{X}, \quad \mathcal{L}(\theta, x) = \frac{\exp\left\{\langle T(x), \alpha(\theta)\rangle\right\}}{\int_{\mathcal{X}} \exp\left\{\langle T(x), \alpha(\theta)\rangle\right\} \mathrm{d}\nu}$$

On obtient via une réécriture exponentielle du dénominateur

$$\forall(\theta, x) \in \Theta \times \mathcal{X}, \quad \mathcal{L}(\theta, x) = \exp\left\{\langle T(x), \alpha(\theta)\rangle - \ln C_\nu(\alpha(\theta))\right\}$$

avec

$$C_\nu(a) := \int_{\mathcal{X}} \exp\left\{\langle T(x), a\rangle\right\} \mathrm{d}\nu$$

Forme canonique (suite)

Enfin, le reparamétrage $\lambda = \alpha(\theta)$ donne une vraisemblance de la forme :

$$\forall(\lambda, x) \in \alpha(\Theta) \times \mathcal{X}, \quad \mathcal{L}(\lambda, x) = \exp\left\{\langle T(x), \lambda\rangle - \ln C_\nu(\lambda)\right\}$$

Forme canonique (suite)

Définition 2.8

On dit que la famille $\{\mathbb{P}_\theta,\ \theta \in \Theta\}$ est une famille exponentielle de dimension p relativement à la mesure dominante μ s'il existe une application mesurable : $T\ :\ \mathcal{X} \longrightarrow \mathbb{R}^p$ telle que

$$\forall \theta \in \Theta \quad C_\mu(\theta) := \int_{\mathcal{X}} \exp\left\{\langle T(x), \theta\rangle\right\} \mathrm{d}\mu(x) < +\infty$$

et une vraisemblance du modèle statistique s'écrive :

$$\forall x \in \mathcal{X}, \quad \mathcal{L}(\theta, x) = \exp\left\{\langle T(x), \theta\rangle - \Psi_\mu(\theta)\right\}$$

avec $\Psi_\mu(\theta) = \ln C_\mu(\theta)$.

Forme canonique (suite)

Remarque

- L'ensemble $N_\mu := \{a \in \mathbb{R}^p,\ C_\mu(a) < +\infty\}$ est appelé l'ensemble naturel des paramètres (il contient Θ).
- La famille est dite complète si $N_\mu = \Theta$ et régulière s'il s'agit d'un ouvert.
- La forme de la vraisemblance d'une loi de la famille exponentielle canonique implique que le support de la loi doit être indépendant de θ.

Forme canonique (suite)

Théorème 2.9

1. *L'ensemble naturel des paramètres $N_\mu := \{a \in \mathbb{R}^p,\ C_\mu(a) < +\infty\}$ d'une famille exponentielle canonique est un convexe de $\mathbb{R}^p$ est la fonction $\Psi(\cdot)$ est convexe.*
2. *Pour tout θ dans l'intérieur de N_μ, la fonction*

$$\Psi_\mu(\theta) = \ln C_\mu(\theta) = \ln \int_{\mathcal{X}} \exp\left\{\langle T(x), \theta\rangle\right\} \mathrm{d}\mu(x)$$

   est indéfiniment dérivable et toute dérivation s'obtient par la dérivation sous le signe somme.

Forme canonique (suite)

Par exemple,

$$\begin{aligned}
\nabla\Psi_\mu(\theta) &= \exp(-\Psi(\theta)) \int_{\mathcal{X}} T(x) \exp\left\{\langle T(x), \theta\rangle\right\} \mathrm{d}\mu(x) \\
&= \int_{\mathcal{X}} T(x)\mathcal{L}(\theta, x)\mathrm{d}\mu(x) = \mathbb{E}\left[T(X)\right] \qquad (2) \\
\frac{\partial\Psi_\mu(\theta)}{\partial\theta_i\theta_j} &= \mathrm{Cov}(T_i(X), T_j(X)) = \mathrm{Var}\left[T(X)\right]_{i,j} \qquad (3)
\end{aligned}$$

Forme canonique (suite)

Le modèle Binomial

Pour l'écrire sous la forme canonique, on introduit la nouvelle mesure dominante $\nu = \sum_{x=0}^{n} C_n^x \delta_x$ et le reparamétrage $\theta = \ln\left(\frac{p}{1-p}\right)$, on obtient :

$$\mathcal{L}(\theta, x) = \exp\left\{x\theta - n\ln(1+\mathrm{e}^{\theta})\right\}$$

avec $T(x) = x$ et $\Psi(\theta) = n\ln(1+\exp\theta)$.

Forme canonique (suite)

Le modèle Binomial (suite)

On a

$$\begin{aligned}
\mathbb{E}[X] &= \Psi'_\mu(\theta) = n\frac{\exp\theta}{1+\exp\theta} = np \\
\mathrm{Var}[X] &= \Psi''_\mu(\theta) = n\frac{\exp\theta}{(1+\exp\theta)^2} = n\frac{\exp\theta}{1+\exp\theta}\frac{1}{1+\exp\theta} \\
&= np(1-p)
\end{aligned}$$

Forme canonique (suite)

Le modèle normal

On utilise le reparamétrage $\theta = (\theta_1, \theta_2) = \left(\frac{m}{\sigma^2}, -\frac{1}{2\sigma^2}\right)$,
$\Theta = \mathbb{R}\times]-\infty, 0[$ et on obtient $T(x) = (x, x^2)$ et
$\Psi_\mu(\theta) = \frac{1}{2}\ln(\pi) - \frac{1}{2}\left(\frac{\theta_1^2}{2\theta_2} + \ln(-\theta_2)\right)$

$$\begin{aligned}
\mathbb{E}\left[X\right] &= \frac{\partial \Psi_\mu(\theta)}{\partial \theta_1} = -\frac{1}{2}\frac{\theta_1}{\theta_2} = -\frac{1}{2}\frac{m}{\sigma^2}(-2\sigma^2) = m \\
\mathbb{E}\left[X^2\right] &= \frac{\partial \Psi_\mu(\theta)}{\partial \theta_2} = -\frac{1}{2}\left(-\frac{\theta_1^2}{2\theta_2^2} + \frac{1}{\theta_2}\right) = m^2 + \sigma^2
\end{aligned}$$

Échantillonnage

Définition 3.1

On appelle échantillon de taille n de loi $\mathbb{P}_\theta$ tout vecteur $(X_1, \ldots, X_n)$, où les $(X_i)_{1 \leq i \leq n}$ sont i.i.d. de loi $\mathbb{P}_\theta$.

- On observe un n-échantillon de v.a.r. $X_1, \ldots, X_n$.
- La loi des X_i appartient à $\{\mathbb{P}_\theta, \theta \in \Theta\}$, famille de probabilités sur $\mathbb{R}$, dominée par une mesure (σ-finie) μ sur $\mathbb{R}$.

Exemple 1

Considérons les durées de vie, supposée indépendantes et de même loi exponentielle, de n ampoules électriques :

- Pour tout $i, x_i \in \mathbb{R}_+$, donc l'espace des observations $\mathcal{X} = \mathbb{R}_+^n$. Alors la tribu associée est $\mathcal{A} = \mathcal{B}(\mathbb{R}_+^n)$
- Le modèle est continue. Comme on admet que la loi est exponentielle, mais que son paramètre est inconnu, l'ensemble des lois de probabilité possible pour chaque X_i est $\{\mathrm{Exp}(\lambda), \lambda > 0\}$.

Finalement, le modèle statistique associé est :

$$\left(\mathbb{R}_+^n,\ \mathcal{B}(\mathbb{R}_+^n),\ \{\mathrm{Exp}(\lambda)^{\otimes n}, \lambda > 0\}\right)$$

Exemple 2

On s'intéresse à la proportion inconnue de pièces défectueuses. Pour l'estimer, on prélève indépendamment n pièces dans la production et on les contrôle :

- Pour tout $i, x_i \in \{0,1\}$, par conséquent l'espace des observations $\mathcal{X} = \{0,1\}^n$. Il est fini, donc le modèle est discret et $\mathcal{A} = \mathcal{P}(\{0,1\}^n)$
- Les X_i sont indépendants et de même loi de Bernoulli $\mathcal{B}(p)$, où $p = \mathbb{P}(X_i = 1)$ est la probabilité qu'une pièce soit défectueuse.

Finalement, le modèle statistique associé est :

$$\left(\{0,1\}^n,\ \mathcal{P}(\{0,1\}^n),\ \mathcal{B}(p)^{\otimes n}\right)$$

Statistiques

Définition 4.1

Dans un modèle statistique $(\mathcal{X}, \mathcal{A}, \{\mathbb{P}_\theta, \theta \in \Theta\})$, une statistique est une application mesurable T de $(\mathcal{X}, \mathcal{A})$ dans un espace $\mathcal{Y}$ muni d'une tribu $\mathcal{B}$.

Remarque

- Une application T de $(\mathcal{X}, \mathcal{A})$ dans $(\mathcal{Y}, \mathcal{B})$ est mesurable si et seulement si $\forall B \in \mathcal{B}$, l'évènement $T^{-1}(B) = \{T(X) \in B\}$ est dans $\mathcal{A}$, c'est-à-dire $\forall A,\ T(A) = B \Rightarrow A \in \mathcal{A}$.
- Concrètement, cela signifie que l'on peut calculer la probabilité de tout évènement de la forme $\{T(X) \in B\}$, donc T ne doit pas dépendre de paramètres inconnus.

Fonction de vraisemblance

- La famille $\{\mathbb{P}_\theta, \theta \in \Theta\}$ est dominée par une mesure σ-finie μ. On se donne, pour $\theta \in \Theta$

$$f_\theta(x) = \frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{d}\mu}(x),\ x \in \mathbb{R}.$$

Fonction de vraisemblance du n-échantillon associée à la famille $\{f_\theta(\cdot), \theta \in \Theta\}$:

$$\boxed{\theta \mapsto \mathcal{L}(\theta, X_1, \ldots, X_n) = \prod_{i=1}^{n} f_\theta(X_i)}$$

- C'est une fonction aléatoire (définie μ-presque partout).

Modèle de Bernoulli

- On observe $X_1, \ldots, X_n$ i.i.d. de loi $\mathcal{B}(p)$ avec $p \in [0,1]$,

$$\mathbb{P}(X = x) = p^x (1-p)^{1-x}, \quad x \in \{0,1\}.$$

- La fonction de vraisemblance associée s'écrit

$$\begin{aligned} p \mapsto \mathcal{L}(p, X_1, \ldots, X_n) &= \prod_{i=1}^{n} p^{X_i} (1-p)^{1-X_i} \\ &= p^{\sum_{i=1}^{n} X_i} (1-p)^{n - \sum_{i=1}^{n} X_i}. \end{aligned}$$

Modèle de Poisson

- On observe $X_1, \ldots, X_n$ i.i.d. de loi $Poi(\lambda)$ avec $\lambda > 0$,

$$\mathbb{P}(X = x) = e^{-\lambda} \frac{\lambda^x}{x!}, \quad x = 0, 1, 2, \ldots .$$

- La fonction de vraisemblance associée s'écrit

$$\begin{aligned} \lambda \mapsto \mathcal{L}(\lambda, X_1, \ldots, X_n) &= \prod_{i=1}^{n} e^{-\lambda} \frac{\lambda^{X_i}}{X_i!} \\ &= \frac{1}{\prod_{i=1}^{n} X_i!} e^{-n\lambda} \lambda^{\sum_{i=1}^{n} X_i} . \end{aligned}$$

Modèle Gaussien

- On observe $X_1, \dots, X_n$ i.i.d. de loi normale $\mathcal{N}(m, \sigma^2)$ avec $\theta = (m, \sigma^2) \in \Theta = \mathbb{R} \times]0, +\infty[$,

$$f_\theta(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{\sigma^2}(x-m)^2\right\}.$$

- La fonction de vraisemblance associée s'écrit

$$\begin{aligned}\mathcal{L}((m, \sigma^2), X_1, \dots, X_n) &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{\sigma^2}(X_i - m)^2\right\} \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - m)^2\right\}\end{aligned}$$

Modèle uniforme $\mathcal{U}[0, \theta]$

- On observe $X_1, \ldots, X_n$ i.i.d. de loi uniforme $\mathcal{U}[0, \theta]$ avec $\theta \in]0, +\infty[$,
$$f_\theta(x) = \frac{1}{\theta} \mathbb{1}_{[0,\theta]}(x).$$
- La fonction de vraisemblance associée s'écrit
$$\mathcal{L}(\theta, X_1, \ldots, X_n) = \prod_{i=1}^{n} \frac{1}{\theta} \mathbb{1}_{[0,\theta]}(X_i) \quad = \frac{1}{\theta^n} \mathbb{1}_{\{\max\limits_{1 \leq i \leq n} X_i \leq \theta\}}$$

Définition 4.2

On appelle statistique associée à un n-échantillon du modèle $(\mathcal{X}, \mathcal{A}, \{\mathbb{P}_\theta, \theta \in \Theta\})$, une fonction mesurable T_n sur $\mathcal{X}^n$ à valeurs dans $\mathbb{R}^p$ et indépendante de θ.

Statistiques empiriques

- Moyenne empirique : $\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i$
- Variance empirique : $S_n^2 := \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}_n\right)^2$

Statistiques d'ordre

Soit $(X_1, \ldots, X_n)$ un n-échantillon. A toute réalisation $(x_1, \ldots, x_n)$ on peut associer le vecteur $(x_{(1)}, \ldots, x_{(n)})$ obtenu en ordonnant les x_i par ordre croissant

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$$

La statistique correspondante $(X_{(1)}, \ldots, X_{(n)})$ est appelée le vecteur des statistiques d'ordre et $X_{(i)}$ est la i-ème statistique d'ordre.

Soit $\mathbb{F}_X(x)$ la fonction de répartition de X. Dans ce cas on a, par exemple,

$$
\begin{aligned}
\mathbb{F}_{X_{(n)}}(x) &= \mathbb{P}\left(X_{(n)} \leq x\right) = \left[\mathbb{F}_X(x)\right]^n \\
\mathbb{F}_{X_{(1)}}(x) &= \mathbb{P}\left(X_{(1)} \leq x\right) = 1 - \left[1 - \mathbb{F}_X(x)\right]^n \\
\mathbb{F}_{X_{(r)}}(x) &= \mathbb{P}\left(X_{(r)} \leq x\right) = \sum_{k=r}^{n} C_n^k \left[\mathbb{F}_X(x)\right]^k \left[1 - \mathbb{F}_X(x)\right]^{n-k}
\end{aligned}
$$

La fonction de répartition empirique

Fonction de répartition empirique associée au n-échantillon $(X_1, \dots, X_n)$:

$$\mathbb{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \leq x\}}, \quad x \in \mathbb{R}.$$

Quantiles empiriques

Quantile empirique d'ordre p :

$$\widehat{q}_{n,p} = \begin{cases} X_{(k)} & \text{si} \quad p \in ((k-1)/n, k/n) \\ \frac{1}{2}\left(X_{(k)} + X_{(k+1)}\right) & \text{si} \quad p = k/n \end{cases}$$

pour $k = 1, \dots, n$, où les $X_{(i)}$ sont les statistiques d'ordre associées à l'échantillon $(X_1, \dots, X_n)$.

Statistique Libre

Définition 4.3

Pour un modèle statistique $(\mathcal{X}, \mathcal{A}, \{\mathbb{P}_\theta, \theta \in \Theta\})$, une statistique T sur ce modèle est dite libre si sa loi ne dépend pas de du paramètre θ.

Modèle gaussien $\{\mathcal{N}(\theta, 1),\ \theta \in \mathbb{R}\}$

nS_n^2 est une statistique libre pour θ.

Statistique Exhaustive

Définition 4.4

Une statistique T est exhaustive pour θ si et seulement si la loi de probabilité conditionnelle de X sachant $\{T = t\}$ ne dépend pas de θ.

Exemple

Soit $(X_1, \ldots, X_n)$, où les X_i sont i.i.d. de loi de Bernoulli $\mathcal{B}(p)$.

On veut montrer que $T(X_1, \ldots, X_n) = \displaystyle\sum_{i=1}^{n} X_i$ est une statistique exhaustive pour p. On écrit :

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n | T = t) = \frac{\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n, T = t)}{\mathbb{P}(T = t)}$$

On sait que $\displaystyle\sum_{i=1}^{n} X_i \sim \mathcal{B}(n, p)$, alors :

$$\mathbb{P}(T = t) = \mathbb{P}\left(\sum_{i=1}^{n} X_i = t\right) = C_n^t p^t (1-p)^{n-t}$$

Exemple (suite)

$$
\begin{aligned}
&= \mathbb{P}(X_1 = x_1, \dots, X_{n-1} = x_{n-1}, X_n = t - \sum_{i=1}^{n} x_i) \\
&= \mathbb{P}(X_1 = x_1) \times \mathbb{P}(X_2 = x_2) \times \cdots \times \mathbb{P}(X_n = t - \sum_{i=1}^{n} x_i) \\
&= p^{\sum_{i=1}^{n-1} x_i} (1-p)^{n-1-\sum_{i=1}^{n-1} x_i} \, p^{t-\sum_{i=1}^{n-1} x_i} (1-p)^{1-t+\sum_{i=1}^{n-1} x_i} \\
&= p^t (1-p)^{n-t}
\end{aligned}
$$

Finalement,

$$
\mathbb{P}(X_1 = x_1, \dots, X_n = x_n | T = t) = \frac{p^t (1-p)^{n-t}}{C_n^t p^t (1-p)^{n-t}} = \frac{1}{C_n^t}
$$

Exemple (suite)

qui ne dépend pas de p, alors $T = \sum_{i=1}^{n} X_i$ est une statistique exhaustive pour le paramètre p.

Théorème de factorisation de Fisher-Neyman

Théorème 4.5

Pour qu'une statistique T soit exhaustive pour θ, il faut et il suffit qu'il existe deux fonctions mesurables g et h telles que :

$$\forall x \in \mathcal{X},\ \forall \theta \in \Theta, \quad \mathcal{L}(\theta, x) = g(T(x); \theta) h(x).$$

Modèle de Bernoulli

On a

$$\begin{aligned} \mathcal{L}(p, X_1, \ldots, X_n) &= p^{\sum_{i=1}^{n} X_i} (1-p)^{n - \sum_{i=1}^{n} X_i} \\ &= g\left(T(X_1, \ldots, X_n), p\right) h(X_1, \ldots, X_n) \end{aligned}$$

Alors, $T(X_1, \ldots, X_n) = \sum_{i=1}^{n} X_i$ est une statistique exhaustive pour le paramètre p.

Modèle Gaussien

$$\begin{aligned}\mathcal{L}((m,\sigma^2),X_1,\ldots,X_n) &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i-m)^2\right\} \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n X_i^2 + \frac{m}{\sigma^2}\sum_{i=1}^n X_i - \frac{nm^2}{2\sigma^2}\right\}.\end{aligned}$$

Alors $T(X_1,\ldots,X_n) = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ est exhaustive pour le paramètre (m,σ^2).

Propriété

Si T est exhaustive et si $T = \varphi \circ S$, alors S est exhaustive.

Échantillon de loi Normale

$T = \left(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2\right) = \varphi(\bar{X}_n, S_n^2)$, donc $(\bar{X}_n, S_n^2)$ est une statistique exhaustive pour (m, σ^2). En effet :

$$T = (n\bar{X}_n, n(S_n^2 + \bar{X}_n^2))$$

Remarque

Si T est exhaustive, $\varphi \circ T$ ne l'est pas forcément !

Théorème de Darmois

Théorème 4.6

Dans un modèle d'échantillon $(\mathcal{X}, \mathcal{A}, \{\mathbb{P}_\theta, \theta \in \Theta\})$, *où le support de la loi des observations ne dépend pas de* θ, *il existe une statistique exhaustive si et seulement si cette loi appartient à la famille exponentielle. Alors*

$$T(X) = \left(\sum_{i=1}^{n} T_1(X_i), \ldots, \sum_{i=1}^{n} T_d(X_i) \right)$$

est une statistique exhaustive.

Échantillon de loi Bernoulli

La loi de Bernoulli appartient à la famille exponentielle avec $T(x) = x$. Alors, $T(X_1, \ldots, X_n) = \sum_{i=1}^{n} T(X_i) = \sum_{i=1}^{n} X_i$ est une statistique exhaustive pour le paramètre p.

Statistique Complète

Définition 4.7

Une statistique T est complète ou totale si et seulement si pour toute fonction mesurable φ, on a : $\mathbb{E}[\varphi(T)] = 0, \ \forall \theta \in \Theta \Rightarrow \varphi = 0$ presque partout sur le support de la loi de T, c'est-à-dire partout sauf sur un ensemble de mesure nulle.

Exemple

Soit $(X_1, \dots, X_n)$, où les X_i sont i.i.d. de loi de Bernoulli $\mathcal{B}(p)$.

On sait que $T(X_1, \dots, X_n) = \sum_{i=1}^{n} X_i$ est une statistique exhaustive pour p. Est-elle complète ?

On sait que $T = \sum_{i=1}^{n} X_i$ est de loi binomiale $\mathcal{B}(n, p)$, donc :

$$\mathbb{E}\left[\varphi(T)\right] = \sum_{k=0}^{n} \varphi(k)\mathbb{P}(X = k) = \sum_{k=0}^{n} \varphi(k) C_n^k p^k (1-p)^{n-k}$$

Il faut montrer que

$$\sum_{k=0}^{n} \varphi(k) C_n^k p^k (1-p)^{n-k} = 0,\ \forall p \in [0,1] \Rightarrow \forall k \in \{0, \dots, n\},\ \varphi(k) = 0$$

Exemple (suite)

Or

$$\sum_{k=0}^{n} \varphi(k) C_n^k p^k (1-p)^{n-k} = (1-p)^n \sum_{k=0}^{n} \varphi(k) C_n^k \left(\frac{p}{1-p}\right)^k$$

Soit $\theta = \dfrac{p}{1-p}$. On a :

$$\begin{aligned} & \sum_{k=0}^{n} \varphi(k) C_n^k p^k (1-p)^{n-k} = 0,\ \forall p \in [0,1] \\ \Rightarrow\ & \sum_{k=0}^{n} \varphi(k) C_n^k \theta^k = 0, \forall \theta \in \mathbb{R}^+ \end{aligned}$$

Exemple (suite)

C'est un polynôme de degré n en θ qui est identiquement nul, donc tous ses coefficients sont nuls.
Par conséquent, $\forall k \in \{0, \ldots, n\}$, $\varphi(k)C_n^k = 0$ et donc $\forall k \in \{0, \ldots, n\}$, $\varphi(k) = 0$, ce qui prouve que $T(X_1, \ldots, X_n) = \sum_{i=1}^{n} X_i$ est une statistique complète.

Théorème 4.8

Dans un modèle d'échantillon où la loi des observations appartient à la famille exponentielle, si $\alpha(\theta)$ est bijective, alors la statistique exhaustive $\sum_{i=1}^{n} T(X_i)$ est complète.