

MST2 - Statistiques

N. Brunel, P. dos Santos, E. Perinel¹

1. nicolas.brunel@ensiie.fr

Plan et bibliographie

Objectif du cours

- Objet d'étude : Un tableau de p mesures sur n "individus" (i.e. un tableau de chiffres ou de fréquences de taille $n \times p$).
- Objectifs des statistiques :
 - ① Description : Construction de résumés numériques et graphiques (pertinents, interprétables, fiables)
 - ② Modélisation
 - ③ Estimation et évaluation de la précision
 - ④ Décision entre plusieurs hypothèses
 - ⑤ Prédiction

Domaine d'applications

- Métiers de la banque, finance (économétrie : prédiction, identification de facteurs, modélisation de risque de crédit, ...),
- Sciences humaines (sociologie : étude de réseaux sociaux, économie : évaluation des politiques économiques),
- Ingénierie (analyse et optimisation de process, fiabilité, contrôle qualité, ...);
- Sciences expérimentales (biologie, chimie, écologie, ... : planification d'expériences, découvertes d'associations, de groupes d'individus);
- (e-)Marketing (datamining, webmining);
- ...

Sans ordinateur, les statistiques ne sont rien...

- Logiciels payants :
 - ▶ **SAS**,
 - ▶ **SPAD**,
 - ▶ SPSS,
 - ▶ **Matlab** (langage matriciel général pour les maths appliquées),
- Logiciels gratuits :
 - ▶ **R** (langage matriciel - calcul scientifique poussé pour les statistiques).
 - ▶ Weka (data mining / machine learning),...

Bibliographie en statistique

- G. Saporta, Probabilités, Analyse de données et statistique, Technip.
- A. Monfort, Cours de Statistique Mathématique, Economica.
- T. Hastie, Tibshirani, J. Friedman (Stanford Univ.), Elements of Statistical Learning, www-stat.stanford.edu/~hastie/Papers/ESLII.pdf.
- S. Tufféry, Data Mining et statistique décisionnelle, Technip.
- S. Weisberg, Applied Linear Regression, Wiley.
- D. Fourdrinier, Statistique Inférentielle, Dunod.
- Site de la Société Française de Statistique (SFdS, <http://www.sfds.asso.fr/>)

Plan du cours

Chapitre 1 Statistique Descriptive et Introduction à l'inférence statistique

Chapitre 2 Vraisemblance, estimation ponctuelle

Chapitre 3 Intervalles de confiance

Chapitre 4 Théorie des tests

Organisation du chapitre 1

1 Introduction à l'inférence statistique

- Les données et le cadre probabiliste
- Les outils probabilistes de la statistique
- Statistiques descriptives : graphiques
- Statistiques descriptives : indicateurs numériques
- Liaison entre deux variables
- Inférence Statistique

2 Modèles Statistiques

- Modèles et estimateurs
- Autres modèles classiques

Chap. 1 : Statistique Descriptive et Introduction à l'Inférence Statistique

1.1. Les données et le cadre probabiliste

Des données réelles quantitatives

- Les tailles (en cm), et poids (en kg), par exemple....

Des données réelles qualitatives

- Etude de l'association entre "croyance religieuse" et éducation.
 $n = 2726$. La p -value est < 0.0001 , $V^2 = 0.0127$

	Croyances religieuses			
Diplôme	Fondament.	Modéré	Libéral	Total
<Bac	178 (137.8) ¹ (4.5) ²	138 (161.5) (-2.6)	108 (124.7) (-1.9)	424
Bac	570 (539.5) (2.6)	648 (632.1) (1.3)	442 (488) (-4.0)	1660
Licence \geq	138 (208.7) (-6.8)	252 (244.5) (0.7)	252 (188.9) (6.3)	642
Total	886	1038	802	2726

Table: Enquête - 1996 - USA, 1. Fréquences théoriques sous indépendance, 2.

Résidu de Pearson Standardisé :
$$e_{ij} = \frac{n_{ij} - np_{ij}^0}{\sqrt{np_{ij}^0}}$$

Modélisation Probabiliste : Population vs Echantillon

- Les données x_1, \dots, x_n sont les réalisations répétées d'une variable aléatoire $\mathbf{X} : (\Omega, \mathcal{A}, P) \rightarrow (\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$. Soit $P_{\mathbf{X}}$ la loi de probabilité (image) définie sur $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$.
 - ▶ Le modèle probabiliste $\mathcal{P} = (\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p), P_{\mathbf{X}})$ est la POPULATION idéalisée (de taille infinie).
 - ▶ L'échantillon $\mathcal{S}_n = (x_1, \dots, x_n)$ est une sous-population finie de \mathcal{P}
- La grande question est "Comment retrouver $P_{\mathbf{X}}$ à partir de \mathcal{S}_n ?"
 - 1 A t on un a priori sur la forme de $P_{\mathbf{X}}$ (par ex. modèle paramétrique $P_{\mathbf{X}} = P_{\theta}, \theta \in \mathbb{R}$) ?
 - 2 Quelles statistiques T choisir pour décrire $P_{\mathbf{X}}$ de manière pertinente (quantiles, indicateurs de position, dispersion) ?
 - 3 Comment estimer un paramètre $\theta = T(P_{\mathbf{X}})$ de manière optimale ?
 - 4 Comment différencier deux lois $P_{\mathbf{X}}$ et $P'_{\mathbf{X}}$ à partir de \mathcal{S}_n ?

1.2. Les outils probabilistes de la statistique

Définitions

- Définition 1.1 : la **Loi** d'une variable aléatoire (v.a.) $X : (\Omega, \mathcal{A}, P) \longrightarrow (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ est la proba. P_X t.q.

$$\forall a, b / [a, b] \subset \mathbb{R}, P_X([a, b]) = P(X \in [a, b])$$

- Définition 1.2 :
 - ▶ **Densité** de la v.a. continue X (définie sur \mathbb{R}^p , relativement à la mesure de Lebesgue) : fonction $f_X \geq 0$ t.q.

$$E_f(T(X)) = \int_{\mathbb{R}^p} T(x) f_X(x) dx$$

pour fonction T continue bornée réelle.

- ▶ **Densité** variable discrète Y définie $\mathcal{Y} = \{y_1, y_2, \dots\}$: fonction $f_Y \geq 0$ t.q.

$$E_f(T(Y)) = \sum_{y_k \in \mathcal{Y}} f_Y(y_k) T(y_k)$$

pour fonction T continue bornée réelle, (i.e $f : \mathcal{Y} \longrightarrow \mathbb{R}$, t.q. $y \mapsto f(y) = P(Y = y)$).

Une loi incontournable : la loi normale

- Loi Normale **centrée et réduite** $X \sim N(0, 1)$ si $f_X(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$.
- Loi Normale (**Gaussienne**) quelconque $Y \sim N(m, \sigma^2)$ si $f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-m)^2}{2\sigma^2}\right)$ et $Y = m + \sigma X$.
- **Vecteur Gaussien** dans \mathbb{R}^p : $Y \sim N(m, \Sigma)$ si $f_Y(y) = \frac{1}{(2\pi)^{p/2}} |\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(y-m)^\top \Sigma^{-1}(y-m)\right)$; ou si $\forall a \in \mathbb{R}^p$, $a^\top Y$ gaussien.
- **Loi du Khi-2** à n ddl χ_n^2 : Si X_1, \dots, X_n i.i.d de loi $N(0, 1)$ alors $S_n^2 = X_1^2 + \dots + X_n^2 \sim \chi_n^2$ et de densité $f_S(s) = \frac{s^{n/2-1} \exp(-s/2)}{2^{n/2} \Gamma(n/2)}$.
Généralisation : **Loi Gamma** : densité p/r à Lebesgue sur \mathbb{R} : $f(x; a, b) = \frac{b^a}{\Gamma(a)} x^{a-1} \exp(-bx) 1_{\mathbb{R}^+}(x)$.

Modèle d'échantillonnage

- Définition 1.3 : Un **échantillon** $\mathcal{S}_n = (X_1, \dots, X_n)$ de taille n est un n -uplets de v.a X_i à valeurs dans $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$ de lois indépendantes et identiquement distribuées (noté **i.i.d**) et de densité commune $f(x)$. L'échantillon \mathcal{S}_n a une densité

$$f_{X_1, \dots, X_n}(x_1, \dots, x_n) = \prod_{i=1}^n f(x_i)$$

relativement à la mesure produit $dx^{\otimes n}$ définie sur $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))^{\otimes n}$.

- Définition 1.4 : Une **statistique** est toute fonction (mesurable) de l'échantillon $T(X_1, \dots, X_n)$. Le plus souvent, T est intégrable $E_f [|T(X_1, \dots, X_n)|] < \infty$.

1.4. Statistiques Descriptives : graphiques

Description directe de la probabilité P_X

- Représentation graphique de l'estimation empirique
 - ▶ Fonction de répartition empirique,
 - ▶ diagramme en bâtons : hauteur des bâtons \propto fréquences f_k
 - ▶ Histogramme,
 - ▶ Boîte à moustaches.

Histogramme et densité

- Notation : $n_k = \#\{x_i \in C_k\}$ effectif de la classe C_k , $f_k = \frac{n_k}{n}$ fréquence de la classe C_k , $F_k = \frac{\sum_{i \leq k} n_i}{n}$ fréquence cumulée (ou $\sum_{i \leq k} f_i$).
- L'histogramme est une "sorte" de dérivée de la fonction de répartition empirique qui approche la densité $f_X = F'_X$:
Soit C_1, C_2, \dots, C_K une partition (en intervalles) de $[A, B]$, avec $A \leq \min x_i \leq \max x_i \leq B$, l'estimateur de l'histogramme est

$$\hat{f}_n(x) = \sum_{k=1}^K \frac{f_k}{h_k} 1_{\{x \in C_k\}}$$

Qualité d'estimation de l'histogramme

Pour tout $x \in \mathbb{R}$, pour K fixé, et si $x \in C_k$ alors

$$E_P \left[\hat{f}_n(x) \right] = \frac{P(X \in C_k)}{h_k} \quad \text{et} \quad V(\hat{f}_n(x)) = \frac{P(X \in C_k)(1 - P(X \in C_k))}{nh_k^2}$$

Les quantiles

- Soit F_X la fonction de répartition de la variable aléatoire X .
 - ▶ Un quantile q_α est tel que $F(q_\alpha) = \alpha$ avec $0 \leq \alpha \leq 1$, et $q_\alpha = F^{-1}(\alpha) = \inf \{x | F(x) \geq \alpha\}$.
- Quantile empirique = quantile de la f.d.r empirique de l'échantillon $\mathcal{S}_n = (x_1, \dots, x_n)$
 - ▶ Quartiles Q_1, Q_2, Q_3 définis par $F(Q_1) = 0.25$, $F(Q_2) = 0.5$, $F(Q_3) = 0.75$ et $\Delta Q = Q_3 - Q_1$.

Boîte à Moustaches (ou Box-plot)

- Représentation graphique des quantiles les plus importants,
- Les moustaches = observations à une distance $\pm 1.5\Delta Q$ des quartiles Q_1, Q_3 ,

⇒ permet la détection de la masse centrale de la pop., sa dispersion et les atypiques.

1.5. Statistiques descriptives : indicateurs numériques

Indicateurs de tendance centrale

- Pour une variable aléatoire réelle des mesures de positions sont :
 - ▶ $E(X)$: espérance ou moyenne
 - ▶ $F_X^{-1}(\frac{1}{2})$: médiane (en tout généralité $\inf \{x \in \mathbb{R} | F_X(x) \geq \frac{1}{2}\}$)
 - ▶ x_0 t.q. $f'_X(x_0) = 0$: mode
- Pour un échantillon observations x_1, \dots, x_n :

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

Médiane empirique soit $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ l'échantillon ordonné alors $Me = x_{(n+1/2)}$ ou $Me \in]x_{(n)}, x_{(n+1/2)}[$ (si n pair). En toute généralité, calculé à partir de la f.d.r empirique F_n .

Interprétation

- Moyenne et Médiane sont des approximations de la variable X par une fonction déterministe (constante)
 - ▶ $E(X) = \operatorname{argmin}_m E((X - m)^2)$
 - ▶ $Me(X) = \operatorname{argmin}_m E(|X - m|)$
- Versions empiriques :
 - ▶ $\bar{x} = \operatorname{argmin}_m \sum_{i=1}^n (x_i - m)^2$
 - ▶ $Me(x) \in \operatorname{argmin}_m \sum_{i=1}^n |x_i - m|$
- Quelle est la qualité (représentativité) de ces résumés (de tendance centrale) ?
Nécessité d'évaluer la qualité de cette approximation, i.e. la dispersion autour de cette valeur.

Indicateurs de dispersion : variance et écart-type

- Définition de la variance

$$\begin{aligned}V_P(X) &= E_P \left((X - E(X))^2 \right) \\ &= E_P(X^2) - E_P(X)^2\end{aligned}$$

et écart-type (homogène aux observations)

$$\sigma_X = \sigma(X) = \sqrt{V_P(X)}.$$

- Propriétés

- ▶ $E_P \left((X - a)^2 \right) = V_P(X) + (E_P(X) - a)^2$ (formule de Huygens).
- ▶ $V_P(aX + b) = a^2 V(X)$ et $\forall a, b, \sigma(aX + b) = |a| \sigma_X$

- Estimateurs :

- ▶ Biaisé : $s_x = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$
- ▶ “Non-biaisé” : $\tilde{s}_x = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

Indicateurs d'asymétrie : Skewness

- Le skewness est le moment centré d'ordre 3 normalisé. Il mesure l'asymétrie d'une distribution

$$\gamma_1 = \frac{E[(X - E(X))^3]}{\sigma^3}$$

- Référence pour une loi normale : $\gamma_1 = 0$.
- Estimateur sans biais (sous hypothèse gaussianité) :

$$G_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\tilde{s}_x} \right)^3$$

Indicateurs d'aplatissement : Kurtosis

- Le kurtosis est le moment standardisé d'ordre 4. Il mesure l'aplatissement de la distribution

$$\gamma_2 = \frac{E[(X - E(X))^4]}{\sigma^4}$$

- Référence pour une loi normale $\gamma_2 = 3$.
- Estimateur sans biais (sous hypothèse normale) :

$$G_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{\tilde{s}_x} \right)^4 - 3 \frac{(n-1)^2}{(n-2)(n-3)}$$

1.6. Liaisons entre deux variables

Covariance de variables aléatoires

- Soient $X, Y : (\Omega, \mathcal{A}, P) \rightarrow \mathbb{R}$ deux variables aléatoires de loi $P_{X,Y}$, alors $\langle X, Y \rangle = E_P(XY)$ définit un produit scalaire dans $L^2(P)$ (mesure de similarité) et $\frac{\langle X, Y \rangle}{\|X\|^2 \|Y\|^2} = \cos(\theta_{X,Y})$.
- En statistique, la **covariance** indique si les fluctuations autour de la moyenne de X et de Y sont similaires :

$$\text{cov}_P(X, Y) = \langle X - E_P(X), Y - E_P(Y) \rangle$$

- Propriétés :
 - ▶ $\forall a, b \text{ cov}_P(aX + b, Y) = a \times \text{cov}_P(X, Y)$
 - ▶ $V_P(X + Y) = V_P(X) + V_P(Y) + 2\text{cov}_P(X, Y)$
 - ▶ $|\text{cov}_P(X, Y)| \leq \sqrt{V_P(X)V_P(Y)}$
- Corrélation $r(X, Y) = r_{XY} = \frac{\text{cov}_P(X, Y)}{(V_P(X)V_P(Y))^{1/2}}$.

Estimation empirique et graphique

- Estimation (version empirique) à partir d'un échantillon $(x_1, y_1), \dots, (x_n, y_n)$:

$$\widehat{\text{cov}}_P(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$
$$\widehat{r}_{XY} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

- Soit $\mathbf{X} = (X_1, \dots, X_p)^\top$ v.a. à valeurs dans \mathbb{R}^p , la matrice de covariance $\Sigma_{\mathbf{X}} = \text{cov}_P(X_1, \dots, X_p) = (\text{cov}_P(X_i, X_j))_{i,j} = E \left((\mathbf{X} - E(\mathbf{X}))(\mathbf{X} - E(\mathbf{X}))^\top \right)$
 - Matrice de similarité (Gram) entre variables aléatoires dans $L^2(\mathbb{R})$.
 - $\Sigma_{\mathbf{X}}$ est la matrice t.q

$$\forall \mathbf{a} = (a_1, \dots, a_p), V_P(a_1 X_1 + \dots + a_p X_p) = \mathbf{a}^\top \Sigma_{\mathbf{X}} \mathbf{a}$$

- Estimateur empirique de la covariance**

$$\widehat{\Sigma}_{\mathbf{X},n} = \frac{1}{n} \sum_{i=1}^n (\mathbf{X}_i - \bar{\mathbf{X}}) (\mathbf{X}_i - \bar{\mathbf{X}})^\top$$

Exemple : La loi normale multivariée

- Soit $\mu \in \mathbb{R}^p$ et Σ matrice symétrique définie positive alors $X \sim N(\mu, \Sigma)$ si sa densité est

$$f_X(x) = \frac{\det(\Sigma)^{-1/2}}{(2\pi)^{p/2}} \exp\left(-\frac{1}{2}(x - \mu)^\top \Sigma^{-1}(x - \mu)\right)$$

- $E_P(X) = \mu$ et $\text{cov}_P(X_1, \dots, X_p) = \Sigma$.
- X est un vecteur gaussien.

Loi jointe de deux variables qualitatives

- Soit X, Y deux variables aléatoires qualitatives (les modalités sont resp. x_1, \dots, x_p et y_1, \dots, y_q).
- La loi jointe de (X, Y) est un tableau $\mathbf{P} \in \mathbb{R}^{p \times q}$, avec $P_{kl} = P(X = x_k, Y = y_l)$.
- Les distributions de X et Y sont P_X et P_Y (appelées distribution marginales de $P_{X,Y}$) et se représente par des vecteurs

$$\begin{cases} \mathbf{p}_X = (P(X_i = x_k))_{1 \leq k \leq p} & = (p_{k.})_{1 \leq k \leq p} \\ \mathbf{p}_Y = (P(Y_i = y_l))_{1 \leq l \leq q} & = (p_{.l})_{1 \leq l \leq q} \end{cases}$$

Tableau de contingence et estimation de la loi

- Pour un échantillon $\mathcal{S}_n = ((x_1, y_1) \dots (x_n, y_n))$, on note les effectifs :
 $n_{kl} = \#\{X_i = x_k, Y_i = y_l\}$, $n_{k\cdot} = \#\{X_i = x_k\}$, $n_{\cdot l} = \#\{Y_i = y_l\}$.
- Représentation des données : **Tableau de contingence**

	y_1	y_2	\dots	y_j	\dots	y_q	
x_1	n_{11}	n_{12}				n_{1q}	$n_{1\cdot}$
\vdots							
x_i				n_{ij}			$n_{i\cdot}$
\vdots							
x_p	n_{p1}					n_{pq}	
	$n_{\cdot 1}$	$n_{\cdot 2}$		$n_{\cdot j}$		$n_{\cdot q}$	n

$$n_{i\cdot} = \sum_{j=1}^q n_{ij}, \quad n_{\cdot j} = \sum_{i=1}^p n_{ij} \quad (\text{marges du tableau}),$$
$$\sum_{i=1}^p n_{i\cdot} = \sum_{j=1}^q n_{\cdot j} = n.$$

- Estimation empirique : $\hat{p}_{ij} = \frac{n_{ij}}{n}$; distributions marginales $\hat{p}_{\cdot j} = \frac{n_{\cdot j}}{n}$ et $\hat{p}_{i\cdot} = \frac{n_{i\cdot}}{n}$; distributions conditionnelles $\hat{p}_{i|j} = \frac{n_{ij}}{n_{\cdot j}} = \frac{\hat{p}_{ij}}{\hat{p}_{\cdot j}}$.

1.3. Inférence statistique

Rappels sur les différentes convergences stochastiques

- Soit $(X_n)_{n \geq 1}$ une suite de variables aléatoires (v.a.) de (Ω, \mathcal{A}, P) dans \mathbb{R}^p , et X v.a. définie sur le même espace. On parle de :

Convergence en probabilités $X_n \xrightarrow{Proba} X$, si pour tout $\varepsilon > 0$,
 $P(\|X_n - X\| > \varepsilon) \rightarrow 0$ quand $n \rightarrow \infty$

Convergence presque-sure $X_n \xrightarrow{P-ps} X$, si $P(\{\lim_n \|X_n - X\| = 0\}) \rightarrow 1$
quand $n \rightarrow \infty$

Convergence en loi $X_n \rightsquigarrow X$, si pour toute fonction continue bornée f ,
 $E_P[f(X_n)] \rightarrow E_P[f(X)]$ quand $n \rightarrow \infty$

Les grands théorèmes de probabilité

- **Loi Forte des Grands Nombres** : Si X_1, \dots, X_n échantillon i.i.d de loi commune P_X tel que $E[|X_i|] < \infty$ alors

$$\bar{X}_n = \frac{X_1 + \dots + X_n}{n} \xrightarrow[n \rightarrow \infty]{} E[X_i] \quad P - ps$$

- **Théorème Limite Central** : Si X_1, \dots, X_n échantillon i.i.d de loi commune P_X tel que $E[X_i] = m < \infty$ et $\sigma^2 = E[X_i^2] - E[X_i]^2$ alors

$$\frac{\bar{X}_n - m}{\frac{\sigma}{\sqrt{n}}} \rightsquigarrow N(0, 1)$$

Fonction de Répartition

- (Cas $X \in \mathbb{R}$) : La probabilité P_X peut être décrite par la fonction de répartition $F_X(x) = P(X \leq x)$ (f.d.r)
 - ▶ F_X fonction croissante, positive tel que $F(-\infty) = 0$ et $F(+\infty) = 1$.
 - ▶ $E_P(T(X)) = \int_{\mathbb{R}} T(x) dF_X(x) = \int_{\mathbb{R}} T(x) f_X(x) dx$ si F_X admet une dérivée f_X .

Fonction de répartition empirique

Si s_n est un échantillon, on définit la f.d.r. empirique, pour tout $x \in \mathbb{R}$,

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^n 1_{\{x_i \leq x\}}.$$

La mesure de probabilité associée $P_n(\cdot) = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}(\cdot)$ est appelée la mesure empirique.

Théorème fondamental de la Statistique

“A partir d'un échantillon X_1, \dots, X_n i.i.d de loi F_X , on peut retrouver la “probabilité définie sur tout Ω ”

- ① Convergence ponctuelle : $\forall x \in \mathbb{R}$,

$$E_P \left[\hat{F}_n(x) \right] = F_X(x) \quad \text{et} \quad V(\hat{F}_n(x)) = \frac{F_X(x)(1 - F_X(x))}{n}$$

d'où $E \left[\left(\hat{F}_n(x) - F_X(x) \right)^2 \right] = \frac{F_X(x)(1 - F_X(x))}{n} \rightarrow 0$ et $\hat{F}_n(x) \xrightarrow{n \rightarrow \infty} F_X(x)$
(en Proba et ps).

- ② Convergence Uniforme : Théorème de Glivenko-Cantelli

$$\sup_{x \in \mathbb{R}} \left| \hat{F}_n(x) - F_X(x) \right| \xrightarrow{p.s.} 0$$

2. Modèles Statistiques

2.1. Modèles et estimateurs

Modèle Statistique et Estimateurs

- Définition 1.3 : un **modèle statistique** \mathcal{M} est une famille de famille de densités paramétrées par un espace des paramètres $\Theta \subset \mathbb{R}^q$.
 $\mathcal{M} = \{f(x; \theta), x \in \mathcal{X} | \theta \in \Theta\}$.

En statistique paramétrique, la loi est connue au paramètre θ près (de dimension finie).

- Définition 1.4 : Un estimateur de θ est une statistique de l'échantillon $\hat{\theta}(X_1, X_2, \dots, X_n)$, qui ne dépend pas du paramètre inconnu θ , et qui "approche" la valeur de θ .
- Définition 1.5 : Un **Estimateur Sans Biais** est $E_{\theta} [\hat{\theta}(X_1, \dots, X_n)] = \theta$ pour tout θ . $B(\hat{\theta}) = E_{\theta} [\hat{\theta}(X_1, \dots, X_n)] - \theta$ est le biais de l'estimateur. Un estimateur est **asymptotiquement sans biais** si $E_{\theta} [\hat{\theta}(X_1, \dots, X_n)] \rightarrow \theta$ si $n \rightarrow \infty$.
- Définition 1.6 : Un Estimateur est **convergent/consistant** : si $\hat{\theta}(X_1, \dots, X_n) \rightarrow \theta$ en P_{θ} - probabilité si $n \rightarrow \infty$.

Evaluation d'un estimateur : le risque quadratique

- Définition 1.7 : Le risque quadratique est définie à partir de la fonction de coût quadratique pour un **estimateur** $\hat{\theta} = d(X_1, \dots, X_n)$

$$\begin{aligned}\mathcal{R}(\hat{\theta}, \theta) &= E_{\theta} \left[(\hat{\theta} - \theta)^2 \right] \\ &= \underbrace{E_{\theta} \left[(E_{\theta} [\hat{\theta}] - \theta)^2 \right]}_{\text{Biais}} + \underbrace{E_{\theta} \left[(\hat{\theta} - E_{\theta} [\hat{\theta}])^2 \right]}_{\text{Variance}}\end{aligned}$$

- ▶ Généralisation au cas multivarié :

$$E_{\theta} \left[\left\| \hat{\theta} - \theta \right\|^2 \right] = E_{\theta} \left[\left\| \hat{\theta} - E_{\theta} [\hat{\theta}] \right\|^2 \right] + \sum_{i=1}^p V_{\theta}(\hat{\theta}_i).$$

- Mesure utile pour comparer et choisir des estimateurs.
- D'autres fonctions de coût sont envisageables, pour l'estimation de paramètres ou d'une variable.

Exemple de modèle : Modèle de Bernoulli et multinomial

- Modèle de Bernoulli : $\mathcal{B}(1, p)$ défini sur $\{0, 1\}$; mesure de référence $\mu(x) = \delta_0(x) + \delta_1(x)$ (somme de mesures de dirac) ; densité est $f(x; p) = p1_{\{0\}}(x) + (1 - p)1_{\{1\}}(x)$

$$\mathcal{M} = \{p1_{\{0\}}(x) + (1 - p)1_{\{1\}}(x) \mid p \in]0, 1[\}$$

- Modèle multinomial de paramètres K, n et (p_1, \dots, p_K) : n tirages (indépendants) parmi k catégories avec une probabilité p_k ; le modèle est défini sur $\{0, 1, \dots, n\}^{\otimes K}$ et $P(N_1 = n_1, \dots, N_K = n_K) = \frac{n!}{n_1! \dots n_K!} p_1^{n_1} \dots p_K^{n_K}$. On peut prendre comme mesure de référence $(\delta_0 + \dots + \delta_K)^{\otimes K}$, et la densité s'écrit

$$f(n_1, \dots, n_K; n, K, p_i) = \sum_{(n_1, \dots, n_K)} \frac{n!}{n_1! \dots n_K!} p_1^{n_1} \dots p_K^{n_K} 1_{N_1=n_1, \dots, N_K=n_K}$$

avec l'espace des paramètres le simplexe (ouvert)

$$\mathcal{S}_K = \left\{ (p_1, \dots, p_K) \in]0, 1[^K \mid \sum_{k=1}^K p_k = 1 \right\}.$$

Modèle de Poisson

- Modèle de Poisson définie \mathbb{N} avec $P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}$. La mesure de référence est la mesure de comptage $\mu = \sum_{k=0}^{\infty} \delta_{\{k\}}$ et la densité s'écrit $f(x; \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$;

$$\mathcal{M} = \{f(x; \lambda) \mid \lambda \in \mathbb{R}^+\}.$$

- Loi Normale (**Gaussienne**) quelconque $Y \sim N(m, \sigma^2)$: mesure de référence = mesure de Lebesgue, $f_Y(y; m, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-m)^2}{2\sigma^2}\right)$.
 - ▶ $\sigma = \sigma_1$ fixé (moyenne inconnue) : $\mathcal{M}_1 = \{f(x; m, \sigma_1) \mid m \in \mathbb{R}\}$
 - ▶ $\mathcal{M}_2 = \{f(x; m, \sigma) \mid m \in \mathbb{R}, \sigma^2 \in \mathbb{R}^{*+}\}$

2.2. Autres modèles classiques

Modèle position-échelle

- Dans \mathbb{R}^p , soit $m \in \mathbb{R}^p$ vecteur déterministe, et ε variable aléatoire, de fonction de répartition F_0 t.q. $F_0(0) = \frac{1}{2}$ (par ex. Normale, ou Cauchy), alors la variable

$$Y = m + \sigma\varepsilon$$

suit un modèle position-échelle (généralise la loi normale).

Modèles de régression

- Régression linéaire : $Y = b_0 + b_1 X_1 + b_2 X_2 + \varepsilon$, avec $\varepsilon \sim N(0, \sigma^2)$.
- Régression logistique $Y \sim \mathcal{B}\left(1, p(x_1, x_2) = \frac{\exp(b_0 + b_1 x_1 + b_2 x_2)}{1 + \exp(b_0 + b_1 x_1 + b_2 x_2)}\right)$.
- **Objectif** : estimation des paramètres $\mathbf{b} = (b_0, b_1, b_2)$ et prédiction de la variable Y à partir de (X_1, X_2) et d'une estimation préalable.

Modèle de mélange de lois gaussiennes

- On peut tirer 2 populations \mathcal{P}_0 ou \mathcal{P}_1 au hasard X , avec $P(X = 0) = 1 - p$ et $P(X = 1) = p$
- Sachant la population, la loi de Y sachant X est gaussienne de paramètres (m_0, σ_0^2) ou (m_1, σ_1^2) .
- La densité de Y p/r à Lebesgue :
$$f_Y(y) = (1 - p) \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{(y-m_0)^2}{2\sigma_0^2}\right) + p \frac{1}{\sqrt{2\pi}\sigma_1} \exp\left(-\frac{(y-m_1)^2}{2\sigma_1^2}\right).$$
- L'espace des paramètres est $\theta = (m_0, m_1, p)$ (on suppose que $m_0 < m_1$) dans $\mathbb{R} \times \mathbb{R} \times]0, 1[$.
- **Objectif** : estimation du paramètre et prédiction de la classe d'appartenance X , sachant Y , à partir du calcul de la probabilité a posteriori $P(X|Y = y)$.