Statistiques mathématiques

R. Petit

année académique 2016 - 2017

Table des matières

Introduction

En probabilités, une variable aléatoire X donnée est entièrement définie par sa loi. On peut l'exprimer par la fonction de répartition F^X ou par la fonction de densité $f^X = \frac{d}{dx}F^X$. Ces fonctions permettent de déterminer :

$$P[a \leqslant X \leqslant b] = \int_a^b f^X(x)\,dx = F^X(b) - F^X(a).$$

Ou encore :

$$E[X] = \int_{-\infty}^{+\infty} x f^X(x)\,dx.$$

Cependant, les fonctions f^X et F^X ne sont jamais connues précisément. Elles peuvent être approchées par des modélisations, mais les modèles ne sont jamais exacts. En probabilités, on cherche donc les observations sur base de la loi qui est connue, alors qu'en statistiques, on cherche à retrouver la loi sur base de n observations $X_1, \ldots, X_n$.

Nous allons nous intéresser à des *modèles statistiques* sous la forme $\left(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \mathcal{P}^{(n)}\right)$ où :

$$\mathcal{P}^{(n)} = \left\{P^{(n)}\right\} = \left\{P_\theta^{(n)} \text{ t.q. } \theta \in \Theta \subset \mathbb{R}^k\right\},$$

et donc les $P^{(i)}$ sont chacun une loi possible pour $(X_1, \ldots, X_n)$.

Ces modèles sont dits *paramétriques* car les différentes lois sont les mêmes au paramètre θ près. Nous n'étudierons que des modèles paramétriques où Θ est un espace de dimension $d \in \mathbb{N}$ finie.

Exemple 0.1. Soient $X_1, \ldots, X_n$ des variables aléatoires iid (indépendantes et identiquement distribuées).

— Si les X_i sont de loi normale $\mathcal{N}(\mu, \sigma^2)$, alors le paramètre θ est donné par :

$$\theta = \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} \in \Theta = \mathbb{R} \times \mathbb{R}^+ \subset \mathbb{R}^2 ;$$

— si les X_i sont de loi uniforme $\mathrm{Unif}(0, \theta)$, le paramètre θ est donné par $\theta \in \Theta = \mathbb{R}_0^+ \subset \mathbb{R}$;

— si les X_i sont de loi $\mathrm{Bern}(p)$, le paramètre θ est donné par $\theta = p \in \Theta = [0, 1] \subset \mathbb{R}$.

Remarque. Une loi normale $\mathcal{N}(\mu, \sigma^2)$ est déraisonnable car les valeurs observables ne vont empiriquement pas vers les infinis alors que la distribution le permet théoriquement mais n'est pas **complètement** déraisonnable car ces probabilités sont négligeables grâce à l'exponentielle de $(-x^2)$ dans la formule de la densité.

Chapitre 1

Théorie de l'échantillonnage

1.1 Terminologie et définitions

Définition 1.1. On appelle *modèle d'échantillonnage* un modèle d'observations iid.

Définition 1.2. Soit un modèle statistique $\left(E^n, \mathcal{B}(E^n), \mathcal{P}^{(n)}\right)$ où $\mathcal{P}^{(n)} = \left\{P_\theta^{(n)} \text{ t.q. } \theta \in \Theta \subset \mathbb{R}^k\right\}$. On note ici $P_\theta^{(n)}$ une loi possible pour $(X_1, \dots, X_n)$ et P_θ une loi possible pour X_i avec i fixé. On dit alors que $P_\theta^{(n)}$ est déterminé par P_θ.

Remarque. Ici, deux visions vont s'opposer et se compléter : la vision *population* qui est associée à P_θ et la version *échantillonage* (ou *empirique*), qui, elle, est associée à $P_\theta^{(n)}$.

Définition 1.3. On définit la fonction indicatrice $I_{[\cdot]}$ qui vaut 1 quand l'expression entre crochets est vraie et 0 sinon.

Définition 1.4. Soit $X_1, \dots, X_n$ une suite de n observations. On définit la i*eme statistique d'ordre* par $X_{(i)} = X_k$ t.q. $\left|\{X_j \text{ t.q. } X_j < X_k, 1 \leqslant j \leqslant n\}\right| = i$. On définit également la *statistique d'ordre* par $\left(X_{(i)}\right)_i$.

Définition 1.5. On définit les fonctions de répartitions comme suit :

— la fonction de répartition population :

$$F_\theta(x) = P_\theta[X_i \leqslant x] ;$$

— la fonction de répartition empirique :

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{[X_i \leqslant x]}.$$

Remarque. La fonction F_n empirique est une fonction en escaliers. Elle fait des sauts de hauteur $\frac{1}{n}$, et est telle que :

$$\lim_{x\to+\infty} F_n(x) = 1 \qquad \text{et} \qquad \lim_{x\to-\infty} F_n(x) = 0.$$

On peut également remarquer que $F_n(X_{(i)}) = \frac{i}{n}$. En effet, par définition de $X_{(i)}$, il y a exactement i observations inférieures à $X_{(i)}$. Dès lors, la fonction indicatrice donnera i fois la valeur 1 et $(n-i)$ fois la valeur 0. La somme donc donc i et la fonction donne $\frac{i}{n}$.

Définition 1.6. On appelle *statistique* toute fonction faisant intervenir **uniquement** des observations.

Exemple 1.1. Par exemple F_n est une statistique car seules les valeurs X_i sont utilisée, mais F_θ n'est pas une statistique car la valeur du paramètre θ apparait et n'est pas une observation.

Remarque. Une statistique peut être à valeur scalaire ($X_{(i)}$ par exemple), à valeur vectorielle ($\left(X_{(i)}\right)_{1\leqslant i\leqslant n}$ par exemple), à valeur ensembliste ($[X_i \pm \overline{X}]$ avec i fixé par exemple), ou encore à valeur fonctionnelle (F_n par exemple).
Remarque. L'objectif est de pouvoir approximer la loi régissant les populations (F_θ) à l'aide de la loi observée empiriquement. Par la loi des grands nombres, on a :

$$F_n(x) \xrightarrow[n\to+\infty]{\text{p.s. par } P_\theta} .$$

Théorème 1.7 (Théorème de Glivenko-Cantelli)**.** *Si* F_n *et* F_θ *sont repsectivement une fonction de répartition empirique et de population, alors :*

$$\sup_{x\in\mathbb{R}}\left|F_n(x) - F_\theta(x)\right| \xrightarrow[n\to+\infty]{\text{p.s.}} 0$$

1.2 Moments

Définition 1.8 (Moments pour populations)**.** On définit $\mu'_r(\theta)$ le *moment **non**-centré* d'ordre r avec $r \in \mathbb{N}^*$ par :

$$\mu'_r(\theta) \coloneqq E_\theta[X_1^r].$$

On définit également $\mu_r(\theta)$, le *moment **centré*** d'ordre r avec $r \in \mathbb{N}^*$ par :

$$\mu_r(\theta) \coloneqq E_\theta\left[\left(X_1 - \mu'_r(\theta)\right)^r\right].$$

Définition 1.9 (Moments pour échantillon)**.** On définit m'_r, le *moment **non**-centré* d'ordre r avec $r \in \mathbb{N}^*$ par :

$$m'_r \coloneqq \frac{1}{n}\sum_{i=1}^{n} X_i^r.$$

On définit également le *moment **centré*** d'ordre r avec $r \in \mathbb{N}^*$ par :

$$m_r \coloneqq \frac{1}{n}\sum_{i=1}^{n}\left(X_i - m'_r\right)^r.$$

Remarque. La loi des grands nombres dit que :

$$m'_r \xrightarrow[n\to+\infty]{\text{p.s.}} \mu'_r(\theta),$$

mais on ne peut pas dire que :

$$m_r \xrightarrow[n\to+\infty]{\text{p.s.}} \mu_r(\theta).$$

Ce n'est donc pas possible car pour m'_r, il y a une somme de variables iid alors que pour m_r, les variables sommées ne sont pas iid (mais dépendent toutes de tous les X_i).

En réalité, il y a convergence, mais on ne peut pas l'exprimer de manière triviale par la loi des grands nombres.

1.2.1 Indicateurs

On peut observer que $\mu_1'(\theta) = E_\theta[X_1]$. Pareil pour $m_1' = \overline{X}$. Le moment d'ordre 1 est donc un indice de position. On a alors $\mu \coloneqq \mu_1(\theta) = E[(X - E[X_1])] = E[X_1] - E[X_1] = 0$. Cette valeur n'est donc pas intéresante. Par contre :

$$\mu_2(\theta) = E\left[(X_1 - E[X_1])^2\right] =: \mathrm{Var}(X) \qquad \text{si} \qquad m_2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2 =: s^2.$$

Le moment d'ordre 2 est donc un indice de dispertion.
Définition 1.10. On appelle le *coefficient d'asymétrie de Fisher* la quantité :

$$\gamma_1 \coloneqq \mu_3(\theta) \cdot \left(\mu_2(\theta)\right)^{-\frac{3}{2}}.$$

Remarque. Le dénominateur $\mu_2(\theta)^{\frac{3}{2}}$ apparait afin de rendre invariant le coeficient d'asymétrie de Fisher aux transformations affines.
Définition 1.11. Le coefficient d'asymétrie de Fisher *empirique* est doné par :

$$m_3 \cdot m_2^{-\frac{3}{2}}.$$

Définition 1.12. On appelle *coefficient d'applatissement de Fisher* la quantité :

$$\gamma_2 \coloneqq \mu_4(\theta) \cdot \left(\mu_2(\theta)\right)^{-2} - 3.$$

Définition 1.13. Le coefficient d'applatissement de Fisher *empirique* est doné par :

$$m_4 \cdot m_2^{-2} - 3.$$

Remarque. Si $\gamma_2 \gneqq 0$, c'est que les évenements extrêmes sont de plus haute probabilité et si $\gamma_2 \lneqq 0$, c'est que les évenements extrêmes sont de moins haute probabilité.

À nouveau, le dénominateur y a été ajouté afin de rendre le coefficient invariant aux transformations affines. Et le terme -3 sert à annuler le coefficient d'applatissement de Fisher pour une normale $\mathcal{N}(\mu, \sigma^2)$.

1.3 Quantile

Définition 1.14. Si F_θ est inversible, alors on définit $x_\alpha(\theta) \coloneqq F_\theta^{-1}(\alpha)$, et on appelle $x_\alpha(\theta)$ un *quantile*.
Remarque. Il faut cependant faire attention car on peut avoir le cas de F_θ discontinue où on choisit $\alpha = F_\theta^{-1}$(point de discontinuité) ou alors le cas de F_θ admettant un plateau et où on choisit α sur le plateau.
Définition 1.15. On définit alors :

$$x_\alpha(\theta) \coloneqq \inf\left\{x \in \mathbb{R} \text{ t.q. } F_\theta(x) \geqslant \alpha\right\}.$$

Remarque. On donne les noms de *médiane*, *quartile*, *décile*, *percentile* pour α valant, avec k entier, respectivement $\frac{1}{2}$, $\frac{k}{4}$ avec $k < 4$, $\frac{k}{10}$ avec $k < 10$, et $\frac{k}{100}$ avec $k < 100$.
Définition 1.16. Pour les échantillons, on définit le *quantile empirique d'ordre* α par :

$$x_\alpha^{(n)} \coloneqq \inf\{x \in \mathbb{R} \text{ t.q. } F_n(x) \geqslant \alpha\}.$$

Remarque. On peut également définir des indices de position, dispertion, asymétrie, applatisement, etc. sur les quantiles plutôt que sur les moments. Ils auront des propriétés différentes et une robustesse différente aux valeurs aberrantes.