Contenu du module

Ce cours de SNP contient les sections suivantes:

- Estimation non paramétrique et théorie Asymptotique
- Tests non paramétriques
- Mesures d'association

References

[1] A. B. Tsybakov. Introduction à l'estimation non-paramétrique, Springer-Verlag, Berlin, 2004.

[2] D. Bosq. Nonparametric statistics for stochastic processes, Springer-Verlag, 1996.

[3] Wand, M.P., Jones, M.C. (1995). Kernel Smoothing, London: Chapman and Hall.

[4] Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis, London: Chapman and Hall.

I) Introduction

La statistique paramétrique est le cadre "classique" de la statistique. On dispose d'un échantillon $X_1, \ldots, X_n$ d'observations issu d'une population X. On veut estimer une fonction ou quantité relative à cette population (moyenne, variance, densité, distribution,...) à partir de l'échantillon $X_1, \ldots, X_n$. On suppose que la fonction à estimer est connue à un vecteur de paramètres près.

Exemple Soit $(X_1, \ldots, X_n)$ échantillon i.i.d de distribution $\mathrm{N}(m, \sigma^2)$: estimer m et σ^2, sela revient à estimer la densité

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

Il s'agit d'une estimation paramétrique.

Mais souvent ;

- On ne suppose pas de forme paramétrique pour la fonction à estimer. Exemple : on veut étudier la surface moyenne du logement Y en fonction du salaire X : $m(x) = E[Y|X = x]$
- On s'autorise toutes les formes a priori ou on ne fait aucune hypothèse sur la forme/nature/type de la distribution des variables aléatoires. Exemple : $\mathcal{F} = \{f : [0,1] \to R, \quad f \ \textit{croissante}\}$.

C'est le cas de SNP. Par exemple : X va de densité f : estimer f.

SNP : Quand l'utiliser ?

- Quand on n'arrive pas à ajuster correctement les observations avec une distribution paramétrique,
- Quand on n'a aucune idée de modèle, ou qu'on ne veut pas avoir un a priori sur le modèle,
- Quand on ne sait pas combien de composantes on veut mettre dans un modèle.
- Quand le nombre de variables est trop grand (problème de grande dimension) et qu'un modèle paramétrique est nonutilisable car il aurait de toutes façons trop de paramètres.

Avantages et inconvénients

1. Moins d'a priori sur les observations,
2. Modèles plus généraux, donc plus robuste.
3. Vitesses de convergence plus lentes, il faut plus de données pour obtenir une précision équivalente.

II) Estimation de la Fonctions de Répartition

On observe $X_1, \ldots, X_n$ ichantillon issu d'une v.a. réelle, de fonction de répartition (fdr) F :

$$F(x) = P(X \leq x).$$

L'estimateur naturel de la fdr F est la fdr empirique notée F_n définie par

$$\begin{aligned} F_n(x) &= \frac{1}{n}\{nbr\ X_i \leq x\} = \frac{1}{n}\sum\nolimits_{i=1}^{n} 1_{(X_i \leq x)} \\ &= \begin{cases} 0, & X_{(1)} > x \\ i/n, & X_{(i)} \leq x < X_{(i+1)} \\ 0, & X_{(n)} < x \end{cases} \end{aligned}$$

avec $\min(X_i) = X_{(1)} \leq \ldots \leq X_{(n)} = \max(X_i)$ les statistiques d'ordres associées à l'échantillon.

Il est clair que F_n est un estimateur non paramétrique de la fdr F.

Propriétés :

1) Biais : on a

$$E(F_n) = E\left(\frac{1}{n}\sum\nolimits_{i=1}^{n} 1_{(X_i \leq x)}\right) = P(X \leq x) = F(x)$$

Donc, $Biais(F_n) = E(F_n) - F = 0$.

2) Variance :

$$\begin{aligned} E\left(F_n^2\right) &= E\left(\frac{1}{n}\sum\nolimits_{i=1}^{n} 1_{(X_i \leq x)}\right)^2 = \frac{1}{n^2}E\left(\sum 1^2_{(X_i \leq x)} + \sum_{i \neq j} 1_{(X_i \leq x)}1_{(X_j \leq x)}\right) \\ &= \frac{1}{n}F(x) + \frac{n-1}{n}F^2(x). \end{aligned}$$

D'où,

$$Var(F_n) = E\left(F_n^2\right) - E^2(F_n) = \frac{1}{n}F(x)(1 - F(x)).$$

3) Comme $Var(F_n) \to 0$, alors $F_n \to F$ en Probabilités.

4) D'après le Thèoreme de la limite centrale:

$$\sqrt{n}\left(\frac{F_n(x) \to F(x)}{\sqrt{F(x)(1 - F(x))}}\right) \to N(0,1) \text{ en loi.}$$

5) De plus, d'après le théorème de Glivenko-Contelli :

$$\sup_x |F_n(x) - F(x)| \to 0 \text{ P.s.}$$

Exercice 1: Soit (X_n) une suite de variables aléatoires indépendante et identiquelent distribués, de fonction de survie notée G, telle que $G(x) = 1 - F(x) = P(X > x)$, où F et la fonction de distribution de X. On considère la fonction empirique G_n définie par :

$$G_n(t) = \frac{1}{n} \sum_{j=1}^{n} 1_{(X_j > t)}, \quad t \in \mathbb{R}$$

1) Quelle est la loi de $nG_n(t)$? la loi limite de $\sqrt{n}G_n(t)$?

2) Montrer la converge en probabilités de $G_n(t)$ vers $G(t)$ lorsque $n \to \infty$.

3) Montrer que cette convergence est aussi presque sûre et on norme infinie.

4) Montrer que la quantité $\|G_n(t) - G(t)\|_\infty$ est indépendante de $G(t)$.

Exercice 2: Soit F une distribution sur R et soit $\theta \in R_+$ un paramètre inconnu. On dispose d'un échantillon $X_1, \ldots, X_n$ de fonction de répartition

$$P(X \leq x) = F_\theta(x) = F(x - \theta).$$

Considérons la variable aléatoire $Y_n = \sum_{i=1}^{n} 1_{(X_i > 0)}$.
1) Pour $n \in \mathbb{N}^*$ fixé, montrer que Y_n suit une loi Binomiale de paramètres n, p à preciser.
2) Montrer que la loi limite de $\frac{1}{\sqrt{n}}(Y_n - np)$ est gaussienne. Préciser la moyenne et la variance de cette loi limite.
3) Déterminer la loi limite des deux statistiques suivantes:

$$V_n = \left(\frac{Y_n}{n}\right)^2 \quad \text{et} \quad W_n = \frac{n}{Y_n}.$$