Suites Variables Aléatoires et Théorèmes Limites

Frédérique Leblanc

données : $x_1, ..., x_n$ observations d'une variable X pour n individus.
formalisation : une réalisation (un tirage) d'un vecteur $(X_1, ..., X_n)$ aléatoire.
Echantillon : On appelle échantillon aléatoire de taille n d'une variable X, la suite des variables indépendantes $X_1, ..., X_n$ et de même loi que X. Une réalisation d'un échantillon aléatoire noté $x_1, ..., x_n$ s'appelle un échantillon (de données).

Exemples :

1. La suite $(x_1, ..., x_n) = (49, 53, 50, 49, ..., 75, 78)$ des poids à 20 ans (X) observés sur un groupe de $n = 32$ étudiants.On **supposera** (i.e. on posera **le modèle**) :

$$X \quad \text{suit une loi} \quad \mathcal{N}(\mu, \sigma^2) \quad \text{avec} \quad (\mu, \sigma) \quad \text{inconnus}$$

2. La suite $(x_1, ..., x_n) = (1, 1, 1, 1, ..., 0, 0)$ des sexes (X prend la valeur 1 pour les filles) observés sur le même groupe de $n = 32$ étudiants. On **supposera** que :

$$X \quad \text{suit une loi} \quad \mathcal{B}(p) \quad \text{avec} \quad p \quad \text{inconnu}$$

où p : la proportion inconnue de filles dans la population de laquelle est extrait l'échantillon.

Inférence statistique :

1. **Modéliser** : choisir une loi pour décrire X. Choix usuels utilisés dans ce cours : si X est continue la loi normale et si X est binaire la loi de Bernoulli (modèle naturel pour un problème sur une probabilité) $\Longrightarrow$ deux (μ et σ^2) ou un (p) **paramètres inconnus** selon le modèle.
2. **Estimer** : proposer une fonction de l'échantillon aléatoire, appelé estimateur et qui appliqué au jeu de données (on remplacera X_i par x_i) donnera une **estimation**.
3. **Propriétés Probabilistes des Estimateurs** : moyenne (espérance), variance et loi
4. **IC et tests** : donner des marges ou risques d'erreurs en estimation ou sur une décision

Definitions :

Soit $X_1, ..., X_n$ un échantillon de X. Notons $\mu = E(X)$ et $\sigma^2 = V(X)$. On appelle moyenne et variance empiriques (aléatoires) les quantités suivantes :

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{et} \quad S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i^2 - \bar{X}_n)^2.$$

Variance empirique corrigée :

$$S_n'^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i^2 - \bar{X}_n)^2.$$

Propriétés :

1. $$S_n'^2 = \frac{n}{n-1} S^2 \quad \text{et} \quad S^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \bar{X}_n^2.$$

2. $$E(\bar{X}) = \mu, \quad E(S^2) = \frac{n-1}{n} \sigma^2 \quad \text{et} \quad E(S'^2) = \sigma^2$$

3. $$V(\bar{X}) = \frac{\sigma^2}{n}$$

Dans ce cas s'ajoutent aussi les propriétés suivantes sur les lois de $\bar{X}$ et S^2 ou S'^2 :

Lois :

1. $$\bar{X}_n \quad \text{suit une loi} \quad \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$
2. $$\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^2} \quad \text{suit une loi du} \quad \mathcal{X}^2_n$$
3. $$\frac{nS^2}{\sigma^2} = \frac{(n-1)S'^2}{\sigma^2} \quad \text{suit une loi du} \quad \mathcal{X}^2_{n-1}$$

Ex : démontrer les points 1 et 2 en utilisant la propriété 4 de la loi de gauss (voir diapo 19 du CM2) pour établir le premier point et la déf. la var. $\mathcal{X}^2_n$ pour le point 2.

Grâce aux propriétés de l'échantillon gaussien on montre (à faire en ex.) que pour tout $\alpha \in [0,1]$:

$$P\left(\bar{X}_n \in \left[\mu - \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}} ; \mu + \frac{\sigma}{\sqrt{n}} u_{1-\frac{\alpha}{2}}\right]\right) = 1 - \alpha$$

Rappel notation : $u_{1-\frac{\alpha}{2}} = \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$

- L'intervalle s'appelle intervalle de fluctuation de niveau $1-\alpha$ (ou risque α) pour la moyenne empirique.
- il est fonction des paramètres du modèle (et pas de $\bar{X}_n$)
- si l'échantillon de donnée produit $\bar{x}_n$ est dans l'intervalle on dira que l'échantillon est conforme (ou dans la norme) au niveau $1-\alpha$

Loi des grands nombres : $X_1, ..., X_n$ échantillon de X où on note $E(X) = \mu$

$$\bar{X}_n \to \mu$$

Ex : n répétition d'un lancé de pièce de monnaie avec $x_i = 1$ si pile : $x_1, ..., x_n$ alors la fréquence d'apparition de pile qui est donnée par $\bar{x}_n$ tend vers $p = E(X)$ où $p = 1/2$ si pièce non truquée. ici la variable qui modélise est X de loi $\mathcal{B}(p)$.

Si la loi de X n'est pas normale l'intervalle précédent ne peut être utilisé pour *n petit* mais pour *n assez grand* c'est un inter. de fluc. de niveau approx. $1 - \alpha$ grâce au TCL (Theorem Central Limit)

Le Théorème Central Limite (TCL)

Soit $X_1, ..., X_n$ un échantillon de la variable X qui admet une espérance et variance finie notées $\mu = E(X)$ et $\sigma^2 = V(X) < +\infty$, alors

$$\frac{\sqrt{n}(\bar{X}_n - E(X))}{\sqrt{V(X)}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \longrightarrow \mathcal{N}(0,1) \quad \text{lorsque} \quad n \to +\infty$$

C'est ce théorème qui permet d'approcher une loi binômiale par une loi normale :

$$\mathcal{B}(n,p) \approx \mathcal{N}(np, np(1-p))$$

Dans le cas Bernoulli : X_i de loi $\mathcal{B}(p)$ donc $\mu = E(X) = p$ et $\sigma^2 = V(X) = p(1-p)$ le TCL donne alors pour $np > 10$ et $n(1-p) > 10$:

$$\frac{\sqrt{n}(\bar{X}_n - p)}{\sqrt{p(1-p)}} \text{ suit approx la loi } \mathcal{N}(0,1)$$

d'où l'intervalle de fluctuation de niveau approx. $1-\alpha$ pour $\bar{X}_n$:

$$\left[p - \frac{\sqrt{p(1-p)}}{\sqrt{n}} u_{1-\frac{\alpha}{2}} ; p + \frac{\sqrt{p(1-p)}}{\sqrt{n}} u_{1-\frac{\alpha}{2}} \right]$$

$\bar{x}_n$ est souvent noté f_n car il représente une fréquence lorsque $x_i \in \{0,1\}$