Suite de variables aléatoires et Théorèmes limites

Chap. 1&2&3 du polycopié

Chargés de cours
V. Léger & F. Leblanc (resp. UE)

données : $x_1, \ldots\ldots, x_n$ observations d'une variable X pour n individus.
formalisation : une réalisation (un tirage) d'un vecteur $(X_1, \ldots, X_n)$ aléatoire.
Echantillon : On appelle échantillon aléatoire de taille n d'une variable X, la suite des variables indépendantes $X_1, \ldots, X_n$ et de même loi que X. Une réalisation d'un échantillon aléatoire noté $x_1, \ldots, x_n$ s'appelle un échantillon de données.
Par abus de langage : on dira échantillon pour $x_1, \ldots, x_n$ (suite déterministe) et pour $X_1, \ldots, X_n$ (suite aléatoire). Le contexte indiquera qu'il s'agit de $X_1, \ldots, X_n$ (modèle) ou de $x_1, \ldots, x_n$ (les 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 $\Longrightarrow$ deux paramètres (μ et σ^2) pour le modèle normal ou un seul (p) pour le modèle de Bernoulli. Dans tous les cas les **paramètres** du modèle sont **inconnus**.
2. **Estimer** : proposer une fonction de $X_1, ..., X_n$ appelé estimateur et qui appliqué au jeu de données (on remplacera X_i par x_i) produira une **estimation**.
3. **Propriétés Probabilistes des Estimateurs** : moyenne (espérance), variance et loi
4. **IC et tests** : Estimation par intervalles et décision sur le(s) paramètre(s) inconnu(s) avec évaluations des risques d'erreurs.

Definitions :

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

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

Variance empirique corrigée :

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

Quel que soit le modèle (loi de X) on a les propriétés suivantes sur espérances et variances des statistiques usuelles :

Propriétés : $\forall n \geq 1$:

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

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

3. $$V(\bar{X}_n) = \frac{\sigma^2}{n} \quad \text{et} \quad V(S_n^2) = \mathcal{O}\left(\frac{1}{n}\right)$$

Echantillon gaussien (normal):
Si X suit la loi $\mathcal{N}(\mu, \sigma^2)$, on établit les lois suivantes pour des transformations utiles de ces statistiques :

1. $$\bar{X}_n \quad \text{suit une loi} \quad \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

2. $$\frac{\sum_{i=1}^{n}(X_i - \mu)^2}{\sigma^2} \quad \text{suit une loi du} \quad \mathcal{X}_n^2$$

3. $$\frac{nS_n^2}{\sigma^2} = \frac{(n-1)S_n'^2}{\sigma^2} \quad \text{suit une loi du} \quad \mathcal{X}_{n-1}^2$$

Ex : démontrer 1) et 2) par récurrence et en utilisant la propriété 4 de la loi normale (voir diapo 26 de Proba et VA) pour établir 1) et la déf. d'un $\mathcal{X}_n^2$ pour 2). On admettra 3).

La loi de la moyenne empirique permet d'obtenir ses fluctuations symétriques autour de $E(X)$.

Fluctuations de la moyenne emp. : cas Gaussien

En effet, pour tout $\alpha \in [0, 1]$:

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

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

- L'intervalle s'appelle intervalle de fluctuation de niveau $1 - \alpha$
- il est fonction des paramètres du modèle (et pas de $\bar{X}_n$)
- si $\bar{x}_n \in IF(\mu, \alpha)$ alors l'éch. est dit conforme au niveau $1 - \alpha$

Loi des grands nombres : Soit $X_1, ..., X_n$ un échantillon de X telle que $E(X) = \mu < \infty$ alors

$$\bar{X}_n \overset{p.s.}{\mapsto} \mu \quad n \to \infty$$

En théorie des probabilités la convergence d'une suite aléatoire est définie de différentes façons. Ici on parle de convergence presque sure, dans le sens où la probabilité qu'une suite de réalisations de $\bar{X}_n$ ne converge pas vers la limite μ, est nulle.

Illustration de ce résultat : on tire un échantillon d'une $\mathcal{B}(1/2)$ (jeu de pile ou face) et on représente la suite déterministe des moyennes des k premières valeurs de l'échantillon.

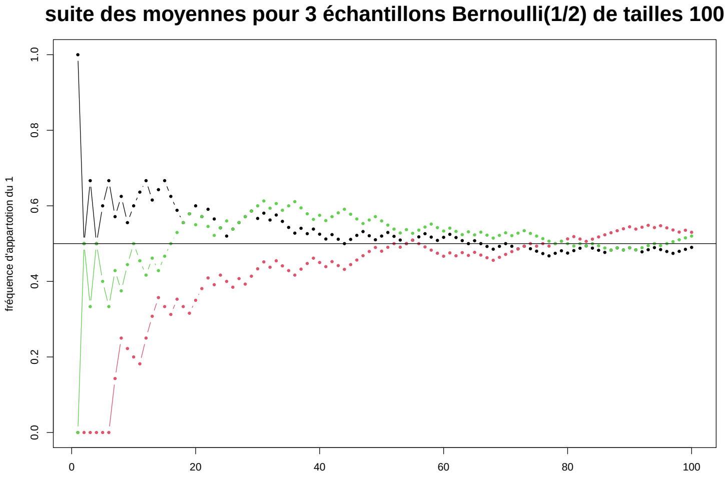
C'est aussi dans ce modèle (Bernoulli) la fréquence d'apparition du 1 dans la suite $(x_1, \ldots, x_k)$ notée f_k :

$$f_k = \bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$$

Avec R :

- on simule d'abord un échantillon de taille $n = 100$: `x=rbinom(100,1,0.5)` (rappel : $\mathcal{B}(1, p) = \mathcal{B}(p)$)
- on calcule les moyennes successives : `moy=cumsum(x)/(1:100)`
- on représente le nuages des points $(k, \bar{x}_k)_{k=1,\ldots,100}$

Dans la figure suivante : les nuages des 100 points $(k, \bar{x}_k)$ pour 3 tirages de la suite $X_1, ..., X_{100}$ avec X_i de loi $\mathcal{B}(1/2)$



Si la loi de X n'est pas normale l'intervalle $IF(\mu, \alpha)$ peut être utilisé pour *n assez grand*. On parlera alors d'Intervalle de fluctuation de niveau approx. $1 - \alpha$.

Ce résultat est une conséquence du Théorème de la limite centrale (TCL : Theorem Central Limit) qui donne la convergence en loi de la moyenne empirique centrée et réduite.

Convergence en loi : on dit qu'une suite de variables aléatoires Y_n converge en loi vers Y si :

$$\forall y, \quad P(Y_n \leq y) \to P(Y \leq y) \quad \text{lorsque} \quad n \to \infty$$

et on note $Y_n \overset{\mathcal{L}}{\mapsto} Y$.

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

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

soit

$$P\left(\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \leq u\right) \to \Phi(u) \quad \text{lorsque} \quad n \to +\infty$$

Dans le cas Bernoulli : X_i de loi $\mathcal{B}(p)$ on $\mu = E(X) = p$ et $\sigma^2 = V(X) = p(1-p)$ et 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 (autour de $E(X) = p$) de niveau approx. $1-\alpha$ pour $\bar{X}_n$:

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

Si $\bar{x}_n = f_n \in IF(p,\alpha)$, on dira que l'échantillon est conforme au niveau $1-\alpha$