Tests Statistiques

Master 1: 2020-2021

Prof. Abdelhakim Necir

Département de Mathématiques (Univ. biskra)

07 Décembre 2020

Variable aléatoire

Soit X une variable aléatoire définie sur un espace de probabilité $(\Omega, \mathcal{A}, P)$ ayant une fonction de distribution

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

F peut être continu ou discontinue
F s'appelle aussi la fonction de répartition
F est en général inconnue que l'on souhaite définir, estimer, ...

Les paramètres usuels caractérisant la v.a. X sont:

- Espérance mathématique (la moyenne théorique):

$$\mathbf{E}[X] = \int x dF(x).$$

- Variance:

$$\begin{aligned} \mathbf{Var}[X] &= \mathbf{E}[X^2] - (\mathbf{E}[X])^2 \\ &= \int x^2 dF(x) - \left(\int x dF(x)\right)^2. \end{aligned}$$

Quantile d'ordre $0 < \alpha < 1$ (ou le fractile d'ordre $0 < \alpha < 1$) :

$$Q_\alpha = F^{\leftarrow}(\alpha) = \inf\{x : F(x) \geq \alpha\},$$

où $F^{\leftarrow}$ désigne l'inverse généralisée. En d'autres termes Q_α est la solution de l'équation

$$\alpha = F(Q_\alpha) \iff \alpha = P(X \leq Q_\alpha).$$

Échantillonnage

Soit X_1, X_2, ..., X_n un échantillons aléatoire de taille n de la variable aléatoire X, c'est à dire une suite de v.a. indépendantes et de même loi que celle de X.
A partir de cet échantillon on peut estimer plusieurs paramètres statistiques de X :

- La moyenne empirique:

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

- La variance empirique:

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

- Quantile empirique d'ordre $0 < \alpha < 1$:

$$\widehat{Q}_\alpha = X_{[\alpha n]:n},$$

où $X_{1:n} < X_{2:n} < ... < X_{n:n}$ désigne les statistiques d'ordres associées à X_1, $X_2, ..., X_n$ et $[\cdot]$ est la partie entière supérieure.

Exemple

Soit 10, 11, 13, 15, 11.5, 10.7, 12 un échantillon observé de taille $n = 7$ d'une v.a X représentant les notes de 7 étudiants:

- Moyenne empirique observée:

$$\overline{x}_7 = \frac{1}{7-1}(10 + 11 + 13 + 15 + 11.5 + 10.7 + 12) = 11.886.$$

- Variance empirique observée:

$$\begin{aligned} s_7^2 &= \frac{1}{7-1}((10 - 11.886)^2 + (11 - 11.886)^2 \\ &\quad + (13 - 11.886)^2 + (15 - 11.886)^2 \\ &\quad + (11.5 - 11.886)^2 + (10.7 - 11.886)^2 + (12 - 11.886)^2) \\ &= 2.80 \end{aligned}$$

Échantillonnage

- Quantile empirique observé d'ordre $\alpha = 0.5$ (la médiane):

$$\widehat{q}_{0.5} = x_{[7/2]:7}.$$

Nous avons

$$(x_1, x_2, ..., x_7) = (10, 11, 13, 15, 11.5, 10.7, 12)$$

donc l'échantillon ordonné est

$$(x_{1:7}, x_{2:7}, x_{3:7}, x_{4:7}, x_{5:7}, x_{6:7}, x_{7:7}) = (10, 10.7, 11, 11.5, 12, 13, 15)$$

Nous avons $[7/2] = [3.5] = 4$, donc $x_{[7/2]:7} = x_{4:7} = 11.5$.

En langage R:

```
> X=c(10,11,13,15,11.5,10.7,12)
> mean(X)
[1] 11.88571
> var(X)
[1] 2.808095
> quantile(X,1/2)
50%
11.5
> 2.80
```

Convergence

Convergence en probabilité:
On dit qu'une suite de v.a. X_n converge en probabilité vers X, et on écrit

$$X_n \xrightarrow{P} X \text{ quand } n \to \infty,$$

si: pour tout $\epsilon > 0$ la suite $P\left(|X_n - X| > \epsilon\right) \to 0$ quand $n \to \infty$.
Loi des grands nombres:

$$\overline{X}_n \xrightarrow{P} \mathbf{E}\left[X\right], \text{ quand } n \to \infty.$$

Convergence en loi (ou en distribution):
On dit qu'une suite de v.a. X_n converge en loi vers X, et on écrit

$$X_n \overset{\mathcal{D}}{\to} X, \text{ quand } n \to \infty,$$

si pour tout x

$$P(X_n \leq x) \to P(X \leq x), \text{ quand } n \to \infty.$$

Théorème Central limite: si $\mathbf{Var}(X) < \infty$, alors

$$\sqrt{n}\frac{\overline{X}_n - \mathbf{E}[X]}{\sqrt{\mathbf{Var}(X)}} \overset{\mathcal{D}}{\to} \mathcal{N}(0,1), \text{ quand } n \to \infty,$$

où $\mathcal{N}(0,1)$ désigne la variable aléatoire normale (gaussienne) centrée réduite.

La loi normale:
Une v.a. Z suit la loi normale de paramètres $\mu \in \mathbb{R}$ et $\sigma^2 \neq 0$ si sa densité:

$$f_{(\mu,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \quad x \in \mathbb{R}.$$

- Si $\mu = 0$ on dit que Z est centrée.
- Si $\sigma^2 = 1$ on dit que Z est réduite
- Si $\mu = 0$ et $\sigma^2 = 1$ à la fois on dit que Z est centrée-réduite.
- Si Z suit la loi normale de paramètres $\mu \in \mathbb{R}$ et $\sigma^2 \neq 0$, alors $(Z - \mu)/\sigma$ est centrée-réduite.
- Si Z est centrée-réduite alors $\sigma Z + \mu$ suit la loi normale de paramètres $\mu \in \mathbb{R}$ et $\sigma^2 \neq 0$.

Soit Z suit la loi normale de paramètres $\mu \in \mathbb{R}$ et $\sigma^2 \neq 0$

- $\mathbf{E}[Z] = \mu$ et $\mathbf{Var}[X] = \sigma^2$.

On dit qu'une suite de v.a. X_n converge presque sûrement vers X, et on écrit

$$X_n \stackrel{p.s}{\rightarrow} X \text{ quand } n \rightarrow \infty,$$

si:

$$P\left(\omega : X_n\left(\omega\right) \rightarrow X\left(\omega\right)\right) = 1.$$

La convergence presque sûre implique la convergence en probabilité.
Loi forte des grands nombres:

$$\overline{X}_n \stackrel{ps}{\rightarrow} \mathbf{E}\left[X\right], \text{ quand } n \rightarrow \infty.$$

Convergence

Fonction de répartition empirique:

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1(X_i \leq x).$$

- En tout x point de continuité de F, on a $F_n(x) \xrightarrow{P} F(x)$.
- $\mathbf{E}[F_n(x)] = F(x)$ et $\mathbf{Var}[F_n(x)] = n^{-1}F(x)(1 - F(x))$.
- En tout x point de continuité de F :

$$\sqrt{n}\left(F_n(x) - F(x)\right)/\sqrt{F(x)(1 - F(x))} \xrightarrow{\mathcal{D}} N(0,1)$$

- **Théorème de Glivenko-Cantelli**:

$$\sup_{x\in R}|F_n(x) - F(x)| \xrightarrow{p.s} 0, \text{ quand } n \to \infty.$$

Nous avons

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} 1\left(X_{i:n} \leq x\right).$$

Ceci peut être réecrite comme suit:

$$F_n(x) = \begin{cases} 0 & \text{si } x < X_{1:n} \\ \frac{i}{n} & \text{si } X_{i:n} \leq x < X_{i+1:n},\ i = 1, ..., n-1 \\ 1 & \text{si } x \geq X_{n:n} \end{cases}.$$

En particulier

$$F_n\left(X_{j:n}\right) = \frac{j}{n},\ j = 1, ..., n.$$

Loi uniforme sur [a,b]

Une variable aléatoire U est uniformément distribuée sur $[a, b]$, $b \neq 0$ si sa densité est:

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{si } x \in [a, b] \\ 0 & \text{si } x \notin [a, b] \end{cases}.$$

Sa fonction de répartition est

$$F(x) = \int_{-\infty}^{x} f(s)\, ds = \begin{cases} 0 & x < a \\ \frac{x-a}{b-a} & \text{si } x \in [a, b[\\ 1 & \text{si } x \geq b \end{cases}.$$

- Nous avons

$$\mathbf{E}[U] = \frac{a+b}{2}, \mathbf{Var}[U] = \frac{(b-a)^2}{12}.$$

Loi uniforme sur [a,b]

Une variable aléatoire U est uniformément distribuée sur $[0,1]$, si sa densité est:

$$f(x) = \begin{cases} 1 & \text{si } x \in [0,1] \\ 0 & \text{si } x \notin [0,1] \end{cases}.$$

Sa fonction de répartition est

$$F(x) = \int_{-\infty}^{x} f(s)\,ds = \begin{cases} 0 & x < 0 \\ x & \text{si } x \in [0,1[\\ 1 & \text{si } x \geq 1 \end{cases}.$$

- $\mathbf{E}[U] = \frac{1}{2}$, $\mathbf{Var}[U] = \frac{1}{12}$.