Chapitre 5
Théorèmes limites et notions de statistiques

Cours proba-stats - Université Lorraine - L2 Informatique
Resp.: J. UNTERBERGER

10 janvier 2023

Table des matières

1 Loi des grands nombres **2**

2 Théorème central limite **3**

3 Application aux statistiques **4**

1 Loi des grands nombres

On observe plusieurs fois le résultat d'une expérience aléatoire répétée.
On considère alors une suite $(X_n)_{n\geq 1}$ de variables aléatoires indépendantes et identiquement distribuées, à valeurs dans $\mathbb{R}$. Si l'on connaît les n valeurs $X_1, \ldots, X_n$ et que l'on veut estimer l'espérance commune des X_i, il est naturel de regarder la moyenne (au sens usuel) de ces n nombres :

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

Cette quantité que l'on note $\overline{X}_n$ est appelée *moyenne empirique*. Il s'agit d'une v.a. d'espérance

$$\mathbb{E}\big[\overline{X}_n\big] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \mathbb{E}[X_1].$$

Supposons de plus ces variables de carré intégrable, c'est-à-dire que $\mathbb{E}[X_i^2] < +\infty$. Comme les v.a. sont indépendantes, on a :

$$\mathrm{Var}\left(\overline{X}_n\right) = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\,\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{1}{n}\,\mathrm{Var}\left(X_1\right)$$

Le fait que $\mathrm{Var}\left(\overline{X}_n\right) \to 0$ lorsque $n \to \infty$ signifie que la v.a. $\overline{X}_n$ est de plus en plus concentrée autour de sa moyenne $\mathbb{E}[X_1]$. Le théorème suivant formalise ce résultat.

Théorème (Loi faible des grands nombres)**.**

Soit $(X_n)_{n\geq 1}$ une suite de v.a. i.i.d. telles que $\mathbb{E}[X_i^2] < +\infty$. On pose $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Alors,

$$\forall \varepsilon > 0, \quad \forall n \geq 1, \quad \mathbb{P}\left(\left|\overline{X}_n - \mathbb{E}[X_1]\right| \geq \varepsilon\right) \leq \frac{\mathrm{Var}(X_1)}{n\,\varepsilon^2}.$$

En conséquence, pour tout $\varepsilon > 0$, $\mathbb{P}\left(\left|\overline{X}_n - \mathbb{E}[X_1]\right| \geq \varepsilon\right) \underset{n\to\infty}{\longrightarrow} 0$ (on dit alors que $\overline{X}_n$ *converge en probabilité* vers le nombre $\mathbb{E}[X_1]$).

Démonstration. On a :

$$\begin{aligned}
\mathbb{P}\left(\left|\overline{X}_n - \mathbb{E}[X_1]\right| \geqslant \varepsilon\right) &= \mathbb{P}\left(\left|\overline{X}_n - \mathbb{E}[\overline{X}_n]\right| \geqslant \varepsilon\right) \\
&= \mathbb{P}\left(\left|\overline{X}_n - \mathbb{E}[\overline{X}_n]\right|^2 \geqslant \varepsilon^2\right) \\
&\leqslant \frac{\mathbb{E}((\overline{X}_n - \mathbb{E}[\overline{X}_n])^2)}{\varepsilon^2} = \frac{\mathrm{Var}(\overline{X}_n)}{\varepsilon^2} = \frac{\mathrm{Var}(X_1)}{n\,\varepsilon^2}
\end{aligned}$$

en utilisant l'inégalité de Markov. □

Interprétons ce résultat dans le cas d'une suite de lancers à pile ou face.
Considérons la v.a X_i qui vaut 1 si pile sort au i-ème lancer et 0 si face sort au i-ème lancer.
Les lancers étant indépendants, la suite $(X_i)_{i\geq 1}$ est une suite de v.a. i.i.d. de Bernoulli de paramètre $p = \mathbb{P}(X_i = 1)$, la probabilité d'obtenir pile à un lancer. Notons que $\mathbb{E}[X_1] = p$ et $\mathrm{Var}(X_1) = p(1-p)$.
Par ailleurs, $\displaystyle\sum_{i=1}^{n} X_i$ est le nombre de piles obtenus pendant les n premiers lancers, et donc $\overline{X}_n$ est la proportion de piles obtenus pendant les n premiers lancers.

La loi faible des grands nombres dit que, dans un certain sens, la proportion $\overline{X}_n$ de piles obtenus lors des n premiers lancers se rapproche, lorsque n tend vers $+\infty$, de $\mathbb{E}[X_1] = p$ qui est la probabilité d'obtenir pile. Autrement dit, lorsque n est grand, on s'attend lors de n lancers à obtenir environ $p\,n$ piles. La loi faible des grands nombres donne un sens mathématique précis à cette phrase.

Il s'agit d'un résultat quantitatif : on a une majoration explicite de la probabilité que la moyenne empirique soit éloignée de plus de ε de $\mathbb{E}[X_1]$.
Par exemple, dans le cas d'une pièce équilibrée, on a $\mathbb{E}[X_1] = 1/2$ et $\mathrm{Var}(X_1) = 1/4$. Donc

$$\mathbb{P}\left(\left|\overline{X}_n - \frac{1}{2}\right| \geqslant \frac{1}{10}\right) \leqslant \frac{100}{4n}.$$

Pour $n = 500$, on obtient :

$$\mathbb{P}\left(\overline{X}_{500} \notin [0.4, 0.6]\right) \leqslant 0.05.$$

Pour $n = 2500$, on obtient :

$$\mathbb{P}\left(\overline{X}_{2500} \notin [0.4, 0.6]\right) \leqslant 0.001.$$

La loi faible des grands nombres peut en fait être améliorée (résultat admis).

Théorème (Loi forte des grands nombres)**.**

Soit $(X_n)_{n\geq 1}$ une suite de v.a. i.i.d. intégrables. On pose $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$. Alors

$$\mathbb{P}\big(\{\omega \in \Omega \ : \ \text{la suite numérique } (\overline{X}_n(\omega))_{n\geq 1} \text{ converge vers le nombre } \mathbb{E}[X_1]\}\big) = 1.$$

Autrement dit, presque sûrement, $\overline{X}_n(\omega) \to \mathbb{E}[X_1]$ lorsque $n \to \infty$ (on dit que $\overline{X}_n$ *converge presque sûrement* vers le nombre $\mathbb{E}[X_1]$).

2 Théorème central limite

En gardant les mêmes notations que précédemment, on a vu que si $\mathbb{E}(X_i) = \mu$ et $\mathrm{Var}(X_i) = \sigma^2$, alors $\mathbb{E}(\overline{X}_n) = \mu$ et $\mathrm{Var}(\overline{X}_n) = \frac{\sigma^2}{n}$.
Le théorème de la limite centrale va nous dire plus précisément que pour n suffisamment grand,

la variable $\overline{X}_n$ se comporte comme une variable gaussienne $\mathcal{N}(\mu, \frac{\sigma^2}{n})$.

Cela revient à dire que $\overline{X}_n - \mu$ se comporte selon la loi $\mathcal{N}(0, \frac{\sigma^2}{n})$.
Ou encore que $\frac{\sqrt{n}}{\sigma}(\overline{X}_n - \mu)$ se comporte selon une loi normale centrée réduite $\mathcal{N}(0, 1)$.

Théorème (Théorème de la limite centrale ou théorème central limite)**.**

Soit $(X_n)_{n\geq 1}$ une suite de v.a. i.i.d. telles que $\mathbb{E}[X_i] = \mu$ et $\mathrm{Var}(X_i) = \sigma^2$.
On pose

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

Alors, pour tous $-\infty \leq a < b \leq +\infty$,

$$\mathbb{P}\left(a \leqslant \frac{\sqrt{n}}{\sigma}\Big(\overline{X}_n - \mu\Big) \leqslant b\right) \longrightarrow_{n\to\infty} \int_a^b \frac{1}{\sqrt{2\pi}}\, e^{-x^2/2}\, dx.$$

Ce théorème met en évidence le rôle fondamental de la loi normale, qui apparaît comme limite quelles que soient les lois au départ : c'est pourquoi on la rencontre souvent dans l'étude statistique de nombreux phénomènes.

Soit $Z \sim \mathcal{N}(0,1)$. Le TCL nous dit que

$$\mathbb{P}\left(a \leqslant \frac{\sqrt{n}}{\sigma}\left(\overline{X}_n - \mu\right) \leqslant b\right) \longrightarrow_{n\to\infty} \mathbb{P}(a < Z < b).$$

On sait par exemple que $\mathbb{P}(-1.96 < Z < 1.96) \approx 0.95$.
On en déduit donc que pour n suffisamment grand,

$$\mathbb{P}\left(-1.96 \leqslant \frac{\sqrt{n}}{\sigma}\left(\overline{X}_n - \mu\right) \leqslant 1.96\right) \approx 0.95.$$

Cela peut se réécrire :

$$\mathbb{P}\left(-1.96\frac{\sigma}{\sqrt{n}} \leqslant \overline{X}_n - \mu \leqslant 1.96\frac{\sigma}{\sqrt{n}}\right) = \mathbb{P}\left(|\overline{X}_n - \mu| \leqslant 1.96\frac{\sigma}{\sqrt{n}}\right) \approx 0.95,$$

ou encore

$$\mathbb{P}\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} \leqslant \overline{X}_n \leqslant \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) \approx 0.95.$$

Reprenons l'exemple du lancer d'une pièce équilibrée ($\mu = 1/2$ et $\sigma^2 = 1/4$).
Pour $n = 500$, on obtient :

$$\mathbb{P}\left(\overline{X}_{500} \in [0.46, 0.54]\right) \approx 0.95.$$

L'intervalle est plus resérré que dans le résultat qu'on avait obtenu par l'inégalité de Markov.

Plus généralement, le TCL nous dit que pour n suffisammment grand,

on peut approcher une loi binomiale $\mathcal{B}(n,p)$ par une loi gaussienne $\mathcal{N}(np, np(1-p))$.

3 Application aux statistiques

Un référendum se prépare. Soit p la proportion des gens qui voteront OUI au référendum. On effectue un sondage auprès d'un échantillon de n personnes prises au hasard. Parmi ces personnes, 60% répondent OUI. Que peut-on en déduire sur la proportion p ? Autrement dit, quelle est la fiabilité du sondage ? En quoi dépend-elle du nombre de personnes interrogées ?

On fait implicitement un certaines hypothèses : l'échantillon sondé est pris "au hasard", c'est-à-dire que l'on suppose les n réponses indépendantes (les personnes interrogées ne s'influencent pas). Et l'on considère que chaque individu a une probabilité p de répondre OUI, $1-p$ de répondre NON, où le nombre p ne dépend pas de l'individu.

Les réponses des n personnes interrogées peuvent être modélisées par n variables aléatoires indépendantes $X_1, \ldots, X_n$ à valeur dans l'ensemble $\{0,1\}$, ayant toutes pour loi la loi de Bernoulli de paramètre p.

On a donc $\mathbb{E}(X_i) = p$ et $\mathrm{Var}(X_i) = p(1-p)$.

D'où :

$$\mathbb{P}(|\overline{X}_n - p| \geqslant a) \leqslant \frac{p(1-p)}{na^2} \leqslant \frac{1}{4na^2}.$$

Si l'on s'intéresse à la question de savoir si le référendum sera ou non accepté, la proportion de réponse positives étant de 0.6, pour 1000 personnes interrogées, la marge d'erreur que l'on se donne sera ici $a = 0.1$.
D'où :

$$\mathbb{P}(|\overline{X}_n - p| \geqslant 0.1) \leqslant \frac{1}{4000 \times 0.1^2} = 0.025.$$

Autrement dit, si notre modèle est le bon, ce qui est fort légitime, il y a moins de 2.5% de chances pour que la proportion de sondés votant OUI soit en dehors de l'intervalle $[p - 0.1, p + 0.1]$.

On suppose maintenant que seules 51% des 1000 personnes sondées ont répondu OUI. La majoration ci-dessus devient :

$$\mathbb{P}(|\overline{X}_n - p| \geqslant 0.01) \leqslant \frac{1}{4000 \times 0.01^2} = 2.5,$$

ce qui n'apporte aucune information !

Le TCL nous dit que :

$$\mathbb{P}\Big(|\overline{X}_n - p| \leqslant 1.96 \frac{\sqrt{p(1-p)}}{\sqrt{n}}\Big) \approx 0.95,$$

d'où

$$\mathbb{P}\Big(|\overline{X}_n - p| \leqslant 1.96 \frac{1}{2\sqrt{n}}\Big) \geqslant 0.95.$$

Pour $n = 1000$, cela donne :

$$\mathbb{P}\Big(|\overline{X}_n - p| \leqslant 0.03\Big) \geqslant 0.95.$$

Cela ne permet pas de conclure que le OUI va l'emporter avec une grande probabilité.

Pour $n = 10000$, cela donne :

$$\mathbb{P}\Big(|\overline{X}_n - p| \leqslant 0.01\Big) \geqslant 0.95,$$

donc le sondage nous dit qu'il y a "95% de chances pour que le OUI l'emporte".

Conclusion : un sondage donnant 51% de OUI est une indication fiable "à 95%" sur le résultat du scrutin seulement si plus de 10000 personnes environs ont été interrogées.