CONCENTRATION, LOI DES GRANDS NOMBRES

Histoire des maths : Dans son ouvrage l'Ars Conjectandi (1713), Jacques Bernoulli valide, avec le théorème de la loi des grands nombres, le principe d'échantillonnage et ouvre la voie du lien entre fréquence et probabilité. Le mathématicien français Bienaymé et le mathématicien russe Tchebychev démontrent en 1867 l'inégalité qui porte leurs noms, en parlant de fréquences d'échantillons plutôt que de variables aléatoires. Ils fournissent ainsi la possibilité d'une démonstration plus simple de la loi des grands nombres.

Au début du XIXe siècle, la modélisation des erreurs de mesure va devenir centrale pour faire de la statistique une science à part entière. Lagrange et Laplace développent une approche probabiliste de la théorie des erreurs. Gauss présente l'écart-type comme une « erreur moyenne à craindre ». Le mathématicien et astronome belge Quetelet utilise les méthodes statistiques en sociologie (1830) et réfléchit à la distribution des données autour de la moyenne.

I. Inégalité de Bienaymé-Tchebychev :

1°) Inégalité de Markov (hors programme mais utile pour démontrer la suite) :

Théorème :

Soit X une variable aléatoire réelle positive ou nulle (toutes les valeurs prises par X sont positives ou nulles) d'espérance E(X). Alors, pour tout réel a strictement positif, $P(X \geqslant a) \leqslant \frac{E(X)}{a}$.

Preuve :

Soit X une variable aléatoire réelle positive ou nulle dont on note x_i les n valeurs pour i entier entre 1 et n.

Par définition de l'espérance, on $E(X) = \sum_{i=1}^{n} x_i P(X=x_i)$. Séparons cette somme en deux blocs en considérant les valeurs supérieures ou égales à a et celles strictement inférieures à a.

On obtient $E(X) = \sum_{x_i \geqslant a} x_i P(X=x_i) + \sum_{x_i < a} x_i P(X=x_i)$.

Pour tout entier i compris entre 1 et n, on sait que $x_i \geqslant 0$ (car X est positive ou nulle) et $P(X = x_i) \geqslant 0$ (par définition d'une probabilité) donc $\sum_{x_i < a} x_i \times P(X=x_i)$.

On en déduit que $E(X) \geqslant \sum_{x_i \geqslant a} x_i P(X=x_i)$.

Par ailleurs, dans cette partie de la somme, pour tout entier i compris entre 1 et n, $x_i \geqslant a$ donc

$E(X) \geqslant \sum_{x_i \geqslant a} x_i P(X=x_i) \geqslant \sum_{x_i \geqslant a} a P(X=x_i)$.

Or, $\sum_{x_i \geqslant a} a P(X=x_i) = a \sum_{x_i \geqslant a} P(X=x_i) = aP(X \geqslant a)$.

Par conséquent, $E(X) \geqslant aP(X \geqslant a)$. Puisque $a > 0$, on obtient enfin $P(X \geqslant a) \leqslant \frac{E(X)}{a}$.

Exemple :

Durant durant l'été, le refuge de haute montagne « les marmottes » héberge en moyenne 11 promeneur·ses par nuits. On note X la variable aléatoire donnant le nombre de promeneurs·ses hébergé·es lors d'une nuit choisie au hasard. D'après l'inégalité de Markov, $\mathrm{P}(\mathrm{X} \geqslant 22) \leqslant \frac{11}{22}$, soit $\mathrm{P}(\mathrm{X} \geqslant 22) \leqslant \frac{1}{2}$. Il y a donc moins d'une chance sur 2 qu'il y ait plus de 22 personnes dans le refuge durant un nuit choisie au hasard.

Remarque :

Cette propriété, appliquée à cet exemple, illustre très bien le fait que s'il y a plus de 22 personnes hébergées plus de la moitié des nuits, le nombre moyen de personnes hébergées ne peut pas être de 11, même s'il n'y a plus aucune personne hébergée le reste du temps.

2°) Inégalité de Bienaymé-Tchebychev :

Propriété :

Soit X une variable aléatoire d'espérance $E(X)$ et de variance $V(X)$.

Pour tout réel strictement positif δ, on a

$$P(|X - E(X)| \geqslant \delta) \leqslant \frac{V(X)}{\delta^2}, \text{ c'est-à-dire } P(X \notin]E(X) - \delta \,;\, E(X) + \delta[) \leqslant \frac{V(X)}{\delta^2}.$$

Vidéo d'explication et d'interprétation de cette inégalité : lien ou

Remarque :

La majoration établie dans cette inégalité est bien souvent loin d'être optimale, mais elle a l'immense avantage de s'appliquer à n'importe quelle variable aléatoire (du moment qu'elle a une espérance et une variance).

Preuve :

Comme $\delta > 0$, les inégalités $|X - E(X)| \geqslant \delta$ et $[X - E(X)]^2 \geqslant \delta^2$ sont équivalentes. De plus, la variable $[X - E(X)]^2$ est positive ou nulle.

On applique donc l'inégalité de Markov à la variable $[X - E(X)]^2$ et au réel δ^2.

Ainsi, $P([X - E(X)]^2 \geqslant \delta^2) \leqslant \dfrac{E([X - E(X)]^2)}{\delta^2}$

Or, $E([X - E(X)]^2) = V(X)$ donc $P([X - E(X)]^2 \geqslant \delta^2) \leqslant \dfrac{V(X)}{\delta^2}$

et on a bien $P(|X - E(X)| \geqslant \delta) \leqslant \dfrac{V(X)}{\delta^2}$.

Remarque :

Cela revient à $P(|X - E(X)| < \delta) = P(X \in]E(X) - \delta \,;\, E(X) + \delta[) > 1 - \dfrac{V(X)}{\delta^2}$ (en étudiant l'événement contraire.

Exemple :

Dans une usine automobile, la variable aléatoire D donnant le diamètre en millimètres d'une soupape prise au hasard dans la production a pour espérance $E(D) = 32$ et pour variance $V(L) = 0{,}04$. Pour qu'elle soit conforme, le diamètre de la soupape doit être compris entre 31,5 et 32,5 mm. Autrement dit, le diamètre ne doit pas s'écarter de plus de 0,5 mm de 32 mm, il faut donc $|D - 32| \leqslant 0{,}5$.

La probabilité qu'une pièce ne soit pas conforme est donc $P(|D - 32| \geqslant 0{,}5) \leqslant \dfrac{0{,}04}{0{,}5^2} = 0{,}16$.

La probabilité que la pièce ne soit pas conforme est donc inférieure à 16 %, ce qui signifie que la probabilité que la pièce soit conforme est supérieure à 84 %

Remarque :

Cette inégalité traduit le fait que la probabilité que X soit loin de l'espérance (i.e. δ grand) est faible.

2°) Application à $\delta = k\sigma$:

Propriété :

Soient X une variable aléatoire d'espérance $E(X)$ et d'écart type $\sigma(X) = \sigma$.

Pour tout entier naturel k, on a

$$P(|X - E(X)| \geqslant k\sigma) \leqslant \frac{1}{k^2}.$$

Preuve :

D'après l'inégalité de Bienaymé-Tchebychev, $P(|X - E(X)| \geqslant k\sigma) \leqslant \dfrac{V}{(k\sigma)^2}$.

Or $\dfrac{V(X)}{(k\sigma)^2} = \dfrac{V(X)}{k^2\sigma^2} = \dfrac{V(X)}{k^2 V(X)} = \dfrac{1}{k^2}$.

Exemple :

Dans l'exemple précédent, si le diamètre d'une soupape non conforme est compris entre 31,92 et 32,08 mm, elle peut être réusinée, sinon la pièce doit être jetée.

Comme $V(L) = 0{,}04$, $\sigma(L) = \sqrt{0{,}04} = 0{,}02$.

La probabilité que la pièce soit jetée est donc $P(|D - 32| \geqslant 0{,}8) = P(|D - 32| \geqslant 4\sigma) \leqslant \dfrac{1}{4^2} = 0{,}0625$.

La probabilité que la pièce soit jetée est donc inférieure à 6,25 %.

Remarque :

Cela revient à $P(|X - E(X)| < k\sigma) > 1 - \dfrac{1}{k^2}$ (en étudiant l'événement contraire).

II. Inégalité de concentration :

Théorème :

Soit $(X_1, X_2, X_3, \dots, X_n)$ un échantillon de taille n associé à une loi d'espérance $E(X)$ et de variance $V(X)$,

Soit M_n la variable aléatoire moyenne de cet échantillon.

Pour tout réel strictement positif δ, on a $P(|M_n - E(X)| \geqslant \delta) \leqslant \dfrac{V(X)}{n\delta^2}$,

Preuve :

La variable aléatoire M_n a pour espérance $E(X)$ et pour variance $\dfrac{V(X)}{n}$. Il suffit alors d'appliquer l'inégalité de Bienaymé-Tchebychev à M_n.

III. Loi des grands nombres :

Théorème :

Soit $(X_1, X_2, X_3, \dots, X_n)$ un échantillon de taille n associé à une loi d'espérance $E(X)$ et de variance $V(X)$,

Soit M_n la variable aléatoire moyenne de cet échantillon.

Pour tout réel strictement positif δ, on a $\lim\limits_{n \to +\infty} P(|M_n - E(X)| \geqslant \delta) = 0$.

Preuve :

L'après l'inégalité de concentration, on a on a $0 \leqslant P(|M_n - E(X)| \geqslant \delta) \leqslant \dfrac{V(X)}{n\delta^2}$.

$\lim\limits_{n \to +\infty} 0 = 0$ et $\lim\limits_{n \to +\infty} \dfrac{V(X)}{n\delta^2} = 0$, donc, d'après le théorème des gendarmes, $\lim\limits_{n \to +\infty} P(|M_n - E(X)| \geqslant \delta) = 0$.

Remarque :

De la loi des grands nombres, on déduit que $\lim_{n \to +\infty} P(|M_n - E(X)| \leqslant \delta) = 1$ (en passant par l'événement contraire), ce qui illustre le fait que la moyenne d'un échantillon se rapproche de l'espérance lorsque la taille de l'échantillon augmente, et qu'on peut la rendre aussi proche que l'on veut en augmentant la taille de l'échantillon.