

Entropie d'une variable aléatoire

Convention Puisque $\lim_{x \rightarrow 0} x \ln x = 0$, on adoptera la convention : $0 \ln 0 = 0$

I. Inégalités de convexité

On note S l'ensemble des suites $(p_n)_{n \geq 1}$ à termes positifs telle que $\sum_1^\infty p_n = 1$.

1. On considère une fonction l concave et de classe C^1 sur un intervalle I .

On se donne une suite $(p_n) \in S$ et une suite (q_n) d'éléments de I telle que les séries $\sum l(q_n)p_n$ et $\sum q_n p_n$ soient convergentes.

(a) Etablir pour tout couple (x, y) d'éléments de I l'inégalité : $l(x) - l(y) \leq l'(y)(x - y)$

(b) Montrer pour tout $y \in I$: $\sum_0^\infty l(q_n)p_n \leq l(y) + l'(y)(\sum_0^\infty q_n \cdot p_n - y)$

(c) En déduire l'inégalité : $\sum_1^\infty l(q_n)p_n \leq l(\sum_1^\infty q_n p_n)$

2. Dans cette question, (p_n) et (q_n) sont deux éléments de S . On suppose que la suite p_n est à termes strictement positifs et tels que la série de terme général $q_n \ln(\frac{p_n}{q_n})$ est convergente (pour tout n tel que $q_n = 0$ on pose $q_n \ln \frac{p_n}{q_n} = 0$)

(a) Démontrer l'inégalité $\sum_1^\infty q_n \ln \frac{p_n}{q_n} \leq 0$

(on peut montrer, ce n'est pas demandé, que cette inégalité est aussi vraie en sommant sur $[1, N]$ si les suites p_n et q_n sont nulles à partir du rang N).

(b) On suppose que $\sum_1^\infty q_n \ln \frac{p_n}{q_n} = 0$.

i. Si la suite (q_n) n'a aucun terme nul, montrer que $\sum_{n=1}^\infty q_n (\ln \frac{p_n}{q_n} + 1 - \frac{p_n}{q_n}) = 0$

ii. Démontrer que les suites (p_n) et (q_n) sont égales.

3. (optionnel) Le cas continu : Soit I un intervalle. On note \mathcal{F} l'ensemble des fonctions continues f et positives sur \mathbb{R} telles que $\int_{\mathbb{R}} f(t)dt = 1$.

Soit $f \in \mathcal{F}$. Soit l une fonction concave C^1 sur un intervalle I et r une fonction continue à valeurs dans I .

En adaptant la démonstration faite précédemment, démontrer, sous réserve de convergence des intégrales, l'inégalité :

$$\int_{\mathbb{R}} l(r(x))f(x)dx \leq l\left(\int_{\mathbb{R}} r(x)f(x)dx\right)$$

II. Entropie d'une variable aléatoire discrète.

Dans toute la suite du problème, (Ω, \mathcal{A}, P) désigne un espace probabilisé.

On note $\mathcal{L}(\Omega)$ l'ensemble des variables aléatoires sur Ω qui sont à valeurs dans \mathbb{N}^* et qui possèdent une espérance.

1. Définition de l'entropie.

Soit $X \in \mathcal{L}(\Omega)$. On note $p_n = P(X = n)$

(a) Etablir pour tout n l'inégalité : $p_n |\ln p_n| \leq \max(np_n, ne^{-n})$

(b) En déduire que la série de terme général $n \ln p_n$ est convergente. On pose alors $H(X) = - \sum_1^\infty p_n \ln p_n$.

$H(X)$ s'appelle l'entropie de X

(c) Une variable aléatoire qui possède une entropie a-t-elle toujours une espérance ?

(d) Quelle est la valeur minimale que peut prendre l'entropie d'une variable aléatoire ?

2. Entropie maximale : Cas des variables aléatoires finies.

Dans cette question on fixe un entier N et on considère une variable aléatoire qui prend ses valeurs l'ensemble $[1, N]$.

En utilisant un résultat de la première partie, démontrer que l'entropie de X est majorée par $\ln N$.

(on pourra par exemple introduire une variable aléatoire Y suivant la loi uniforme sur $[1, N]$)

3. Entropie maximale : cas des variables discrètes

Soit X une variable aléatoire suivant une loi géométrique de paramètre p .

Rappeler l'expression de $p_n = P(X = n)$ et calculer en fonction de p son entropie.

L'entropie est-elle majorée sur $\mathcal{L}(\Omega)$?

4. Soit $Y \in \mathcal{L}(\Omega)$. On note $q_n = P(Y = n)$

(a) Démontrer qu'il existe une variable aléatoire géométrique X ayant la même espérance. On note $p_n = P(X = n)$.

(b) Exprimer $H(Y) - H(X)$ à l'aide de $\sum_1^\infty q_n \ln \frac{p_n}{q_n}$ et en déduire son signe.

Comment s'interprète ce résultat en termes de maximum ?

III. Moyenne géométrique de variables aléatoires.

1. Soit (X_n) une suite de variables aléatoires réelles sur (Ω, \mathcal{A}, P) . On suppose qu'il existe une variable aléatoire X telle que :

$$\forall \epsilon > 0, \lim_{\epsilon \rightarrow 0} P(|X_n - X| > \epsilon) = 0$$

(on dit que la suite (X_n) converge vers X en probabilités)

On pose $Z_n = e^{X_n}$ et $Z = e^X$.

(a) Justifier brièvement que Z_n et Z sont bien des variables aléatoires.

Soient ϵ et α deux réels strictement positifs.

(b) Justifier l'existence de s tel que $P(|X| \geq s) < \alpha$.

(c) Etablir l'inégalité : $P(|Z_n - Z| \geq \epsilon) \leq P(|X| \geq s) + P(|X_n - X| \geq \ln(1 + \epsilon e^{-s}))$

indication : on comparera du point de vue de l'inclusion les événements $E = [|Z_n - Z| \geq \epsilon]$, et $A \cup B$, avec $A = [|X| \geq s]$ et $B = [|X_n - X| \geq \ln(1 + \epsilon e^{-s})]$

(d) Démontrer que la suite (Z_n) converge en probabilités vers Z

2. On suppose que (T_n) est une suite de variables aléatoires strictement positives indépendantes et de même loi. On suppose également que la variable aléatoire $\ln(T_1)$ possède une variance.

On pose $P_n = (T_1 \cdot T_2 \cdot \dots \cdot T_n)^{\frac{1}{n}}$

Démontrer la suite P_n converge en probabilités vers une variable aléatoire constante.

Exprimer cette constante à l'aide de la variable aléatoire T_1 .

IV. Optimisation d'un taux de rendement

On considère une succession de courses hippiques entre N chevaux participants, numérotés $1, 2, \dots, N$.

On note p_k la probabilité de victoire du cheval k et on suppose qu'aucun p_k n'est nul.

Pour tout k on note c_k la cote du cheval k . Ainsi un parieur qui a misé un montant m_k sur le k ième cheval recevra la somme $m_k c_k$ en cas de victoire de ce cheval, et perdra sa mise dans les autres cas. On suppose qu'au cours du temps les cotes des différents chevaux sont fixes.

A l'occasion de la première course, le parieur dispose d'une somme initiale égale à $R_0 > 0$ qu'il souhaite répartir en totalité sur les différents chevaux dans les proportions respectives f_1, f_2, \dots, f_N . Les f_i sont donc des nombres réels strictement positifs de somme 1. A l'issue de cette première course, le parieur se retrouve avec une somme $R_1 = R_0 M_1$, avec $M_1 > 0$.

A l'occasion de la seconde course, le parieur réinvestit la totalité de la somme R_1 , dans les mêmes proportions f_1, \dots, f_N , et donc après cette deuxième course, il possède une somme $R_2 = R_1 M_2$ avec $M_2 > 0$ et ainsi de suite.....

Les variables aléatoires M_n sont supposées indépendantes (les différentes courses sont indépendantes) et de même loi (le parieur suit la même règle de pari à toutes les courses).

La fortune du parieur après la n ième course est donc égale à $R_n = R_0 M_1 \dots M_n$.

On définit pour tout $n > 0$ le taux de rendement moyen par $T_n = \left(\frac{R_n}{R_0}\right)^{\frac{1}{n}} - 1$.

1. Montrer que la variable aléatoire T_n converge en probabilités vers une constante τ . Cette constante est appelée le taux de rendement asymptotique des paris.

La stratégie du parieur va donc consister à déterminer les proportions f_1, \dots, f_N qui maximisent τ .

2. (a) Vérifier que $\tau = e^{[\sum_1^n p_k \ln f_k c_k]} - 1$.

(b) En déduire la stratégie optimale du parieur et la valeur de τ associée.

- (c) On suppose dans cette question que $\sum_{k=1}^N \frac{1}{c_k} = 1$ (cette hypothèse signifie en pratique que la totalité des sommes pariées est redistribuée aux gagnants). Vérifier que $\sum_{k=1}^N p_k \ln c_k p_k \geq 0$.

Dans quel cas le parieur ne dispose t'il alors d'aucune stratégie lui permettant d'assurer un taux de rendement asymptotique strictement positif ?