

TD 13 - probabilités : convergence, approximation et estimation

Convergence, inégalité de Bienaymé-Tchebychev

Exercice 1 - classique - premières utilisations de l'égalité

On jette 10 000 fois une pièce équilibrée. On note X la fréquence d'apparition de « pile » obtenue.

1. Justifier que X est la moyenne empirique de variables de Bernoulli à préciser.
2. A l'aide de l'inégalité de Bienaymé-Tchebychev, montrer que :
 - a. $P(X \notin [0, 45; 0, 55]) \leq \frac{1}{100}$
 - b. $P(X \notin [0, 49; 0, 51]) \leq \frac{1}{4}$
 - c. Comment interpréter ces résultats ?
3. On suppose que 10 000 est un nombre de lancers suffisant pour considérer que la loi de X est proche de la loi normale $\mathcal{N}(m, \sigma^2)$
 - a. Calculer m et σ^2
 - b. Montrer que $P(X \in [0, 49; 0, 51]) = \Phi(2) - \Phi(-2)$
 - c. Comparer le résultat obtenu avec celui de la question 1.

Exercice 2 - convergence en loi d'une suite de variables discrètes vers une variable à densité

On considère, pour $n \in \mathbb{N}^*$, une variable aléatoire V_n suivant la loi uniforme discrète $\mathcal{U}([1, n])$

Pour $n \in \mathbb{N}^*$, on pose $X_n = \frac{1}{n}V_n$

1. Justifier que $X_n(\Omega) \subset [0, 1]$
2. Soit $x \in [0, 1]$, montrer que, pour $n \in \mathbb{N}^*$, $P(X_n \leq x) = \frac{\lfloor nx \rfloor}{n}$, puis que :

$$x - \frac{1}{n} \leq P(X_n \leq x) \leq x$$
3. Soit U une variable suivant la loi uniforme à densité $\mathcal{U}([0, 1])$
Montrer que : X_n converge en loi vers U

Exercice 3 - sommes de lois de Poisson

Soit $(X_i)_{i \in \mathbb{N}}$ une suite de variables aléatoires indépendantes suivant chacune la loi de Poisson de paramètre 1

On note $S_n = \sum_{i=1}^n X_i$ pour $n \in \mathbb{N}^*$

1. Quelle est la loi de S_n ?
2. Déterminer la variable centrée réduite S_n^* associée à S_n
3. En utilisant le théorème central limite, montrer que : $P(S_n \leq n) = P(S_n^* \leq 0) \xrightarrow{n \rightarrow +\infty} 1/2$
4. En déduire que : $e^{-n} \sum_{k=0}^n \frac{n^k}{k!} \xrightarrow{n \rightarrow +\infty} \frac{1}{2}$

Exercice 4 - classique

Soit $(X_i)_{i \in \mathbb{N}}$ une suite de variables aléatoires indépendantes de mêmes lois, d'espérance m et de variance σ^2

On pose, pour $n \in \mathbb{N}^*$: $\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$ et $S_n = \sqrt{n} \times \frac{\overline{X}_n - m}{\sigma}$

On note Φ la fonction de répartition de la loi normale centrée réduite $\mathcal{N}(0, 1)$

1. Montrer que S_n est la variable centrée réduite associée à $\overline{X}_n$
2. Déterminer une majoration de $P(|S_n| \geq k)$ pour $k \in \{1, 2, 3\}$ à l'aide de l'inégalité de Bienaymé-Tchebychev.
3.
 - a. Montrer que, pour tout $x \in \mathbb{R}$, $\Phi(x) - \Phi(-x) = 2\Phi(x) - 1$
 - b. En déduire que, si N suit la loi normale centrée réduite $\mathcal{N}(0, 1)$, alors, pour $x \geq 0$:
 - i. $P(|N| \leq x) = 2\Phi(x) - 1$
 - ii. $P(|N| \geq x) = 2(1 - \Phi(x))$
4. On suppose n assez grand pour approcher la loi de S_n par $\mathcal{N}(0, 1)$
Donner une approximation de $P(|S_n| \geq k)$ pour $k \in \{1, 2, 3\}$ et comparer avec les résultats avec la question 2.)
Données numériques : $\Phi(1) \approx 0,8413$, $\Phi(2) \approx 0,9772$, $\Phi(3) \approx 0,9987$

Estimation

Exercice 5 - classique

Soit $(X_1, \dots, X_n)$ un n -échantillon de variables aléatoires indépendantes et de même loi uniforme sur $[0, a]$.

On note : $\overline{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$, $U_n = 2\overline{X}_n$, $V_n = \max(X_1, \dots, X_n)$, $W_n = \frac{n+1}{n} V_n$

1.
 - a. Calculer $E(U_n)$ et $V(U_n)$
 - b. Montrer que : $\forall \varepsilon > 0, \quad P(|U_n - a| > \varepsilon) \xrightarrow{n \rightarrow +\infty} 0$
 - c. Justifier que U_n est un estimateur de a
2. On rappelle que : pour $x \in \mathbb{R}$, $[V_n \leq x] = \bigcap_{i=1}^n [X_i \leq x]$
 - a. Déterminer la fonction de répartition de V_n
 - b. En déduire que V_n est une variable à densité, et en préciser une densité notée f_{V_n}
 - c. Calculer $E(V_n)$ et $V(V_n)$, puis $E(W_n)$ et $V(W_n)$
 - d. A l'aide de l'inégalité de Bienaymé-Tchebychev, montrer que :
$$\forall \varepsilon > 0, \quad P(|W_n - a| > \varepsilon) \xrightarrow{n \rightarrow +\infty} 0$$
 - e. Justifier que W_n est un estimateur de a
3. Des deux estimateurs W_n et U_n de a , lequel préférer ?

Intervalles de confiance

Exercice 6 - classique aussi

Soit $(X_n)_{n \in \mathbb{N}}$ une suite de variables aléatoires indépendantes suivant la même loi exponentielle de paramètre $\frac{1}{a}$ avec $0 < a < 1$

On pose, pour tout $n \in \mathbb{N}^*$, $S_n = \frac{1}{n} \sum_{k=1}^n X_k$

1. Estimation ponctuelle

- a. Déterminer $E(S_n)$ et $V(S_n)$
- b. Justifier que : $\forall \varepsilon > 0, \lim_{n \rightarrow +\infty} \mathbf{P}(|S_n - a| > \varepsilon) = 0$
- c. Justifier que S_n peut être considéré comme un estimateur de a

2. Estimation par intervalle de confiance en utilisant l'inégalité de Bienaymé-Tchebychev

- a. En écrivant l'inégalité de Bienaymé-Tchebychev pour S_n , montrer que, pour tout $\varepsilon > 0$:

$$P(|S_n - a| \leq \varepsilon) \geq 1 - \frac{1}{n\varepsilon^2}$$

- b. En déduire une valeur de n pour laquelle $\left[S_n - \frac{1}{10}, S_n + \frac{1}{10}\right]$ est un intervalle de confiance de a au niveau de confiance de 95%

3. Estimation par intervalle de confiance en appliquant le théorème de la limite centrée.

- a. Montrer que $S_n^* = \frac{\sqrt{n}}{a}(S_n - a)$ converge en loi vers la loi normale centrée réduite.

On supposera dans la suite que n est suffisamment grand pour confondre la loi de S_n^* avec la loi normale centrée réduite $\mathcal{N}(0, 1)$ On pose $\alpha = 1,96$ et on pourra considérer que $\Phi(\alpha) = 0,975$

- b. Montrer que $P\left(|S_n - a| \leq \frac{a\alpha}{\sqrt{n}}\right) = 0,95$
- c. En déduire que $P\left(|S_n - a| \leq \frac{\alpha}{\sqrt{n}}\right) \geq 0,95$
- d. En déduire que $J_n = \left[S_n - \frac{\alpha}{\sqrt{n}}, S_n + \frac{\alpha}{\sqrt{n}}\right]$ est un intervalle de confiance de a au niveau de confiance de 95%
- e. Au vu de ce résultat, quelle valeur de n suffit-il de prendre pour que $\left[S_n - \frac{1}{10}, S_n + \frac{1}{10}\right]$ soit un intervalle de confiance de a au niveau de confiance de 95% ?
- f. Comparer l'intervalle J_n obtenu dans la question 3d) avec $\left[S_n - \frac{1}{10}, S_n + \frac{1}{10}\right]$ pour la valeur n obtenue dans la question 2.b.

Exercice 7 - classique

Pour un entier $n \geq 1$, on note $(X_1, \dots, X_n)$ un n -échantillon de variables indépendantes de même loi de Bernoulli de paramètre p , et on pose $Z_n = \frac{1}{n} \sum_{i=1}^n X_i$

- 1. Déterminer l'espérance et la variance de Z_n
- 2. Montrer que Z_n est un estimateur de p

On dit qu'un estimateur T_n de θ est sans biais si $E(T_n) = \theta$ et on dit qu'il est convergent si $\forall \varepsilon > 0, \lim_{n \rightarrow +\infty} P(|T_n - \theta| > \varepsilon) = 0$

3. Montrer que Z_n est un estimateur de p sans biais convergent.
4. On note Z_n^* la variable centrée réduite associée à Z_n
 - a. Justifier que $(Z_n^*)_{n \in \mathbb{N}^*}$ converge en loi vers une variable aléatoire suivant la loi normale centrée réduite $\mathcal{N}(0, 1)$ de fonction de répartition Φ

On supposera dans la suite que n est suffisamment grand pour identifier la loi de Z_n^* à $\mathcal{N}(0, 1)$
On donne $\Phi(2, 57) \approx 0, 9949$ et $\Phi(2, 58) \approx 0, 9951$

- b. Déterminer un réel $a > 0$ tel que $P(|Z_n^*| \leq a) \geq \frac{99}{100}$
- c. Montrer que $\sqrt{p(1-p)} \leq \frac{1}{2}$ et en déduire que : $P\left(|Z_n - p| \leq \frac{a}{2\sqrt{n}}\right) \geq \frac{99}{100}$
- d. En déduire un intervalle de confiance de p au niveau de confiance de 99%

Exercice 8

Dans cet exercice, θ désigne un réel strictement positif et n un entier naturel supérieur ou égal à 2

Pour tout k de $\mathbb{N}$, on pose : $u_k = \frac{1}{1+\theta} \left(\frac{\theta}{1+\theta}\right)^k$

1. Montrer que la suite $(u_k)_{k \in \mathbb{N}}$ définit bien une loi de probabilité.
On considère maintenant une variable aléatoire X prenant ses valeurs dans $\mathbb{N}$ et dont la loi est donnée par :

$$\forall k \in \mathbb{N}, \quad P(X = k) = u_k$$

2. On pose $Y = X + 1$. Reconnaitre la loi de Y , puis en déduire l'espérance et la variance de X
3. Dans cette question, on souhaite estimer le paramètre θ par la méthode du maximum de vraisemblance. Pour ce faire, on considère un échantillon $(X_1, X_2, \dots, X_n)$ composé de variables aléatoires indépendantes ayant toutes la même loi que X et on introduit la fonction L , de $\mathbb{R}_+^*$ dans $\mathbb{R}$, définie par :

$$\forall \theta \in \mathbb{R}_+^*, \quad L(\theta) = \prod_{k=1}^n P(X_k = x_k)$$

où $x_1, x_2, \dots, x_n$ désignent des entiers naturels éléments de $X(\Omega)$

L'objectif est de choisir la valeur de θ qui rend $L(\theta)$ maximale.

- a. Ecrire $\ln(L(\theta))$ en fonction de θ et de $S_n = \sum_{k=1}^n x_k$

- b. On considère la fonction φ , définie par :

$$\forall \theta \in]0; +\infty[, \quad \varphi(\theta) = S_n \ln(\theta) - (S_n + n) \ln(1 + \theta)$$

Montrer que la fonction φ admet un maximum, atteint en un seul réel que l'on notera $\hat{\theta}_n$ et que l'on exprimera en fonction de S_n . Que représente $\hat{\theta}_n$ pour la fonction L ?

On pose dorénavant : $T_n = \frac{1}{n} \sum_{i=1}^n X_i$

La variable T_n est appelée estimateur du maximum de vraisemblance pour θ

4.
 - a. Calculer $E(T_n)$ et $V(T_n)$ pour $n \in \mathbb{N}^*$
 - b. Montrer que, pour $\varepsilon > 0$, $\lim_{n \rightarrow +\infty} P(|T_n - \theta| > \varepsilon) = 0$

Exercice 9 - EML

Soit $(X_n)_{n \in \mathbb{N}}$ une suite de variables aléatoire indépendantes suivant la même loi exponentielle de paramètre $a > 0$ inconnu, paramètre que l'on cherche à estimer.

On pose, pour $n \in \mathbb{N}^*$, $\overline{X}_n = \frac{1}{n} \sum_{k=1}^n X_k$ et $T_n = \sqrt{n}(a\overline{X}_n - 1)$

1. Justifier avec soin que $(T_n)_{n \in \mathbb{N}^*}$ converge en loi vers une variable aléatoire N suivant la loi normale centrée réduite.
2. Montrer que : pour $n \in \mathbb{N}^*$, $T_n \in [-2, 2] \Leftrightarrow a \in \left[\frac{\sqrt{n} - 2}{\overline{X}_n \sqrt{n}} ; \frac{\sqrt{n} + 2}{\overline{X}_n \sqrt{n}} \right]$
3. On admet que $\Phi(2) \geq 0,975$. Montrer que : $\left[\frac{\sqrt{n} - 2}{\overline{X}_n \sqrt{n}} ; \frac{\sqrt{n} + 2}{\overline{X}_n \sqrt{n}} \right]$ est un intervalle de confiance de a au niveau de confiance de 95%
4. Ecrire une fonction Python qui :
 - choisit un réel $a \in]0, 5[$ quelconque
 - crée une liste de 1 000 réalisations indépendantes de la loi $\mathcal{E}(a)$ (on utilisera le résultat : si $U \hookrightarrow \mathcal{U}([0, 1])$ alors $-\frac{1}{\lambda} \ln(1 - U) \hookrightarrow \mathcal{E}(\lambda)$)
 - renvoie un couple (u, v) pour lequel $a \in [u, v]$ avec une probabilité au moins égale à 95%, ainsi que la valeur de a
L'utilisateur pourra alors vérifier que $u \leq a \leq v$

Exercice 10 - Ecricome 2017

Soit n un entier naturel non nul.

On effectue une série illimitée de tirages *indépendants* d'une boule avec remise dans une urne contenant n boules numérotées de 1 à n . Pour tout entier naturel k non nul, on note X_k la variable aléatoire égale au numéro de la boule obtenue au $k^{\text{ième}}$ tirage.

Pour tout entier naturel k non nul, on note S_k la somme des numéros obtenus lors des k premiers tirages :

$$S_k = \sum_{i=1}^k X_i$$

On considère enfin la variable aléatoire T_n égale au nombre de tirages nécessaires pour que, pour la première fois, la somme des numéros des boules obtenues soit supérieure ou égale à n

Exemple : avec $n = 10$, si les numéros obtenus aux cinq premiers tirages sont, dans cet ordre : 2, 4, 1, 5, 9, alors :

$$S_1 = 2, S_2 = 6, S_3 = 7, S_4 = 12, S_5 = 21, \text{ et } T_{10} = 4.$$

Toutes les variables aléatoires seront supposées définies sur le même espace probabilisé $(\Omega, \mathcal{A}, P)$

Partie A

1. Pour $k \in \mathbb{N}^*$, déterminer la loi de X_k ainsi que son espérance.
2.
 - a. Déterminer $T_n(\Omega)$
 - b. Calculer $P(T_n = 1)$
 - c. Montrer que : $P(T_n = n) = \left(\frac{1}{n}\right)^{n-1}$
3. Dans cette question, $n = 2$. Déterminer la loi de T_2
4. Dans cette question, $n = 3$. Déterminer la loi de T_3 et vérifier que $E(T_3) = \frac{16}{9}$

Partie B

5. Déterminer $S_k(\Omega)$
6. Soit $k \in \llbracket 1, n-1 \rrbracket$
 - a. Exprimer S_{k+1} en fonction de S_k et de X_{k+1}
 - b. En utilisant un système complet d'événements lié à la variable aléatoire S_k , démontrer alors que :

$$\forall i \in \llbracket k+1, n \rrbracket, \quad P(S_{k+1} = i) = \frac{1}{n} \sum_{j=k}^{i-1} P(S_k = j)$$

7.
 - a. Pour $k \in \mathbb{N}^*$ et $j \in \mathbb{N}^*$,
rappeler la formule du triangle de Pascal liant les nombres $\binom{j-1}{k-1}$, $\binom{j-1}{k}$ et $\binom{j}{k}$
 - b. En déduire que, pour tout $k \in \mathbb{N}^*$, et tout entier $i \geq k+1$:

$$\sum_{j=k}^{i-1} \binom{j-1}{k-1} = \binom{i-1}{k}$$

c. Pour tout entier $k \in \llbracket 1, n \rrbracket$, on note $\mathcal{H}_k$ la proposition :

$$\ll \forall i \in \llbracket k, n \rrbracket, P(S_k = i) = \frac{1}{n^k} \binom{i-1}{k-1} \gg$$

Démontrer, par récurrence, que $\mathcal{H}_k$ est vraie pour tout entier $k \in \llbracket 1, n \rrbracket$

8. a. Soit $k \in \llbracket 1, n-1 \rrbracket$. Comparer les événements $[T_n > k]$ et $[S_k \leq n-1]$

b. En déduire que : $\forall k \in \llbracket 0, n-1 \rrbracket, P(T_n > k) = \frac{1}{n^k} \binom{n-1}{k}$

9. Démontrer que $E(T_n) = \sum_{k=0}^{n-1} P(T_n > k)$ puis que : $E(T_n) = \left(1 + \frac{1}{n}\right)^{n-1}$

10. Calculer $\lim_{n \rightarrow +\infty} E(T_n)$

Partie C

Dans cette partie, on fait varier l'entier n et on étudie la convergence en loi de la suite de variables $(T_n)_{n \geq 1}$ obtenue.

11. Soit Y une variable aléatoire à valeurs dans $\mathbb{N}^*$ telle que : $\forall k \in \mathbb{N}^*, P(Y = k) = \frac{k-1}{k!}$

a. Vérifier par le calcul que : $\sum_{k=1}^{+\infty} P(Y = k) = 1$.

b. Montrer que Y admet une espérance et calculer cette espérance.

12. Pour tout entier $k \in \mathbb{N}^*$, démontrer que : $\lim_{n \rightarrow +\infty} P(T_n > k) = \frac{1}{k!}$.

13. Démontrer alors que $(T_n)_{n \geq 1}$ converge en loi vers Y

14. On importe sous Python les bibliothèques `numpy` sous l'alias `np` et `numpy.random` sous l'alias `rd`. On rappelle que l'instruction `rd.randint(1, n+1)` renvoie un entier aléatoire de $\llbracket 1, n \rrbracket$. Compléter la fonction ci-dessous qui prend en argument le nombre n de boules contenues dans l'urne afin qu'elle simule la variable aléatoire T_n :

```
def T(n):
    S = ...
    y = ...
    while ... :
        tirage = rd.randint(1, n+1)
        S += ...
        y += ...
    return
```

La suite page suivante.

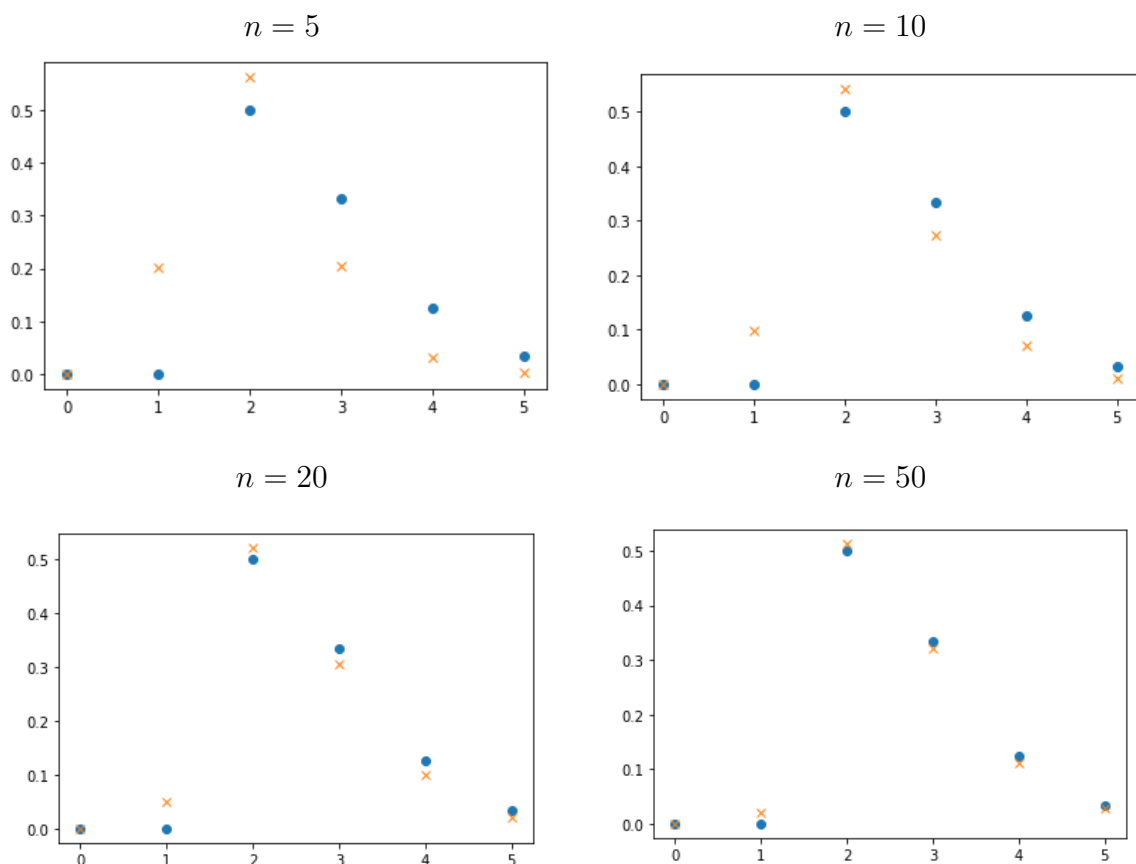
15. On suppose déclarée la fonction précédente et on écrit le script ci-dessous :

```
def freq(n):
    y = np.zeros(n+1)
    for i in range(100000):
        k = T(n)
        y[k] += 1
    return y/100000

def Loi(n):
    y = np.zeros(n+1)
    fact = 1
    for k in range(1, n+1):
        fact = fact * k
    y[k] = (k-1)/fact
    return y

N=[5, 10, 20, 50]
fig=0
for n in N:
    fig +=1
    plt.figure(fig)
    x=np.arange(6)
    y1=[Loi(n) for n in range(6)]
    y2=[freq(n) for n in range(6)]
    plt.plot(x, y1, 'o', x, y2, 'x')
```

L'exécution de ce script pour les valeurs n indiquées a permis d'obtenir les graphes ci-dessous :



- Expliquer ce que représentent les vecteurs renvoyés par les fonctions `freq` et `Loi`.
Comment ces vecteurs sont-ils représentés graphiquement dans chaque graphique obtenu ?
- Expliquer en quoi cette succession de graphiques permet d'illustrer le résultat de la question 13.