CONCENTRATION, LOI DES GRANDS NOMBRES

I. Moyenne d'un échantillon

1) Définition

Exemple :
On lance un dé à six faces et on considère la variable aléatoire X qui prend la valeur 1 si le dé s'arrête sur un chiffre pair et la valeur 0 sinon.
X suit donc une loi de Bernoulli de paramètre $\frac{1}{2}$.
On répète deux fois de suite cette expérience. On considère alors l'échantillon (X_1, X_2) de taille 2 de variables aléatoires X_1 et X_2 suivant la même loi que X.
Il est ainsi possible d'évaluer le résultat d'une telle expérience en étudiant la variable aléatoire moyenne de X_1 et X_2.
On appelle M_2 la variable aléatoire moyenne de l'échantillon (X_1, X_2).
Alors M_2 peut prendre les valeurs suivantes :

Valeur de X_1	Probabilité de X_1	Valeur de X_2	Probabilité de X_2	Probabilité de (X_1, X_2)	Valeur de M_2	Probabilité de M_2
0	$\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$	$\frac{0+0}{2} = 0$	$\frac{1}{4}$
0	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$	$\frac{0+1}{2} = \frac{1}{2}$	$\frac{1}{4} + \frac{1}{4} = \frac{1}{2}$
1	$\frac{1}{2}$	0	$\frac{1}{2}$	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$	$\frac{1+0}{2} = \frac{1}{2}$	
1	$\frac{1}{2}$	1	$\frac{1}{2}$	$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$	$\frac{1+1}{2} = 1$	$\frac{1}{4}$

Et on a ainsi la loi de probabilité de M_2 :

k	0	$\frac{1}{2}$	1
$P(M_2 = k)$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

Définition : Soit $(X_1, X_2, \dots, X_n)$ un échantillon de taille n de variables aléatoires indépendantes suivant une même loi.
La variable aléatoire moyenne M_n de l'échantillon est donnée par :

$$M_n = \frac{1}{n}(X_1 + X_2 + \cdots + X_n).$$

2) Propriétés

Exemple :
On reprend l'exemple précédent.
- Calculons l'espérance de M_2 :

$$E(M_2) = 0 \times \frac{1}{4} + \frac{1}{2} \times \frac{1}{2} + 1 \times \frac{1}{4} = \frac{1}{2}$$

On retrouve l'espérance de la variable X.
On comprend intuitivement que l'espérance de la variable aléatoire moyenne d'un échantillon $(X_1, X_2, \dots, X_n)$ est égale à l'espérance de la variable aléatoire X associée à cet échantillon.

- Calculons la variance de M_2 :

$$V(M_2) = \frac{1}{4} \times \left(0 - \frac{1}{2}\right)^2 + \frac{1}{2} \times \left(\frac{1}{2} - \frac{1}{2}\right)^2 + \frac{1}{4} \times \left(1 - \frac{1}{2}\right)^2 = \frac{1}{8}$$

Alors que :

$$V(X) = \frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

Ainsi la variance de la variable aléatoire moyenne est plus faible que la variance de la variable d'origine.
De plus, la dispersion de la variable aléatoire moyenne diminue au fur et à mesure que la taille de l'échantillon n augmente.
En effet, si l'échantillon devient plus grand, le nombre de situations pouvant donner des valeurs proches de l'espérance augmente.
Ainsi, les valeurs prises par la moyenne deviennent de plus en plus probables dans un voisinage de l'espérance.

Propriété : Soit une variable aléatoire X et soit un échantillon $(X_1, X_2, \dots, X_n)$ de taille n de variables aléatoires indépendantes suivant la même loi que X.

$$E(M_n) = E(X) \qquad V(M_n) = \frac{1}{n} V(X) \qquad \sigma(M_n) = \frac{1}{\sqrt{n}} \sigma(X)$$

Méthode : Calculer l'espérance, la variance et l'écart-type d'une variable aléatoire moyenne

Vidéo https://youtu.be/o67OOavrbHQ

On considère la variable aléatoire X qui prend, de façon équiprobable, les valeurs –4, 0, 1, 3 et 6.
M_{50} est la variable aléatoire moyenne d'un échantillon de taille 50 de la loi de X.
Calculer l'espérance, la variance et l'écart type de M_{50}.

Par équiprobabilité, on établit le tableau de la loi de probabilité de X.

k	–4	0	1	3	6
$P(X=k)$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$

On a ainsi :

$$E(X)=\frac{1}{5}\times(-4)+\frac{1}{5}\times 0+\frac{1}{5}\times 1+\frac{1}{5}\times 3+\frac{1}{5}\times 6=1{,}2$$

$$V(X)=\frac{1}{5}\times(-4-1{,}2)^2+\frac{1}{5}\times(0-1{,}2)^2+\frac{1}{5}\times(1-1{,}2)^2+\frac{1}{5}\times(3-1{,}2)^2$$
$$+\frac{1}{5}\times(6-1{,}2)^2=10{,}96$$

$\sigma(X)=\sqrt{10{,}96}\approx 3{,}31$

On en déduit :

$E(M_{50})=E(X)=1{,}2$

$$V(M_{50})=\frac{1}{50}V(X)=\frac{10{,}96}{50}=0{,}2192$$

$$\sigma(M_{50})=\frac{1}{\sqrt{50}}\sigma(X)\approx\frac{3{,}31}{\sqrt{50}}\approx 0{,}468$$

II. Inégalité de Bienaymé-Tchebychev

Propriété : Soit une variable aléatoire X. Pour tout réel strictement positif δ, on a :

$$P(|X-E(X)|\geq\delta)\leq\frac{V(X)}{\delta^2}$$

Méthode : Appliquer l'inégalité de Bienaymé-Tchebychev

Vidéo https://youtu.be/4XMvq1FnYwU

Soit une variable aléatoire X qui suit la loi binomiale de paramètres $n=20$ et $p=0{,}1$.
1) Appliquer l'inégalité de Bienaymé-Tchebychev avec $\delta=2\sigma(X)$. Interpréter.
2) Recommencer avec $\delta=3\sigma(X)$, puis $\delta=4\sigma(X)$. Que constate-t-on ?
3) a) Simuler N valeurs de la variable aléatoire X par une fonction en Python dans le but d'estimer la probabilité $P(|X-2|\geq 2\sigma(X))$.
On testera le programme pour différentes valeurs de N.
b) Au regard des résultats obtenus par le programme, peut-on penser que l'inégalité de Bienaymé-Tchebychev a un caractère optimal ?

1) - $E(X)=20\times 0{,}1=2$ $\quad V(X)=20\times 0{,}1\times 0{,}9=1{,}8$ $\quad \sigma(X)=\sqrt{1{,}8}$

Ainsi, on obtient :

$$P(|X - E(X)| \geq 2\sigma(X)) \leq \frac{V(X)}{\left(2\sigma(X)\right)^2}$$

Ou encore : $$P(|X - 2| \geq 2\sqrt{1{,}8}) \leq 0{,}25$$

La probabilité que l'écart de X à $E(X)$ soit supérieur à $2\sigma(X)$ est majorée par 0,25.

2) – pour $\delta = 3\sigma(X)$:

$$P(|X - E(X)| \geq 3\sigma(X)) \leq \frac{V(X)}{\left(3\sigma(X)\right)^2}$$

Ou encore : $$P(|X - 2| \geq 3\sqrt{1{,}8}) \leq \frac{1}{9}$$

– pour $\delta = 4\sigma(X)$:

$$P(|X - E(X)| \geq 4\sigma(X)) \leq \frac{V(X)}{\left(4\sigma(X)\right)^2}$$

Ou encore : $$P(|X - 2| \geq 4\sqrt{1{,}8}) \leq 0{,}0625$$

– On peut en déduire que les écarts de X à $E(X)$ de quelques σ deviennent improbables.

3) a)

```
import random as rd
import math

def simulX():
    a=0
    for expe in range(20):
        if rd.randint(1,100)<=10:
            a=a+1
    return a

def proba(N):
    echant=[simulX() for i in range(N)]
    c=0
    d=2*math.sqrt(1.8)
    for e in echant:
        if abs(e-2)>=d:
            c=c+1
    return c/N
```

```
>>> proba(1000)
0.038
>>> proba(10000)
0.0454
>>> proba(100000)
0.04178
>>> proba(100000)
0.04516
```

b) On constate qu'un écart à $E(X)$ supérieur à $2\sigma(X)$ est de probabilité souvent inférieure 0,05 (0,038 ; 0,0454 ; 0,04178 ; 0,04516) alors que l'inégalité de Bienaymé-Tchebychev nous donne pour cette même probabilité une majoration par 0,25. L'inégalité est donc loin d'être optimale.

III. Inégalité de concentration

Propriété : Soit la variable aléatoire moyenne M_n d'un échantillon de taille n de la variable aléatoire X. Pour tout réel strictement positif δ, on a :

$$P(|M_n - E(X)| \geq \delta) \leq \frac{V(X)}{n\,\delta^2}$$

Méthode : Appliquer l'inégalité de concentration pour déterminer la taille d'un échantillon

Vidéo https://youtu.be/7Nk9U-zwWOA

Soit une variable aléatoire X qui suit la loi de Bernoulli de paramètre 0,2.
On considère un échantillon de n variables aléatoires suivant la loi de X.
On appelle M_n la variable aléatoire moyenne associée à cet échantillon.
Déterminer la taille n de l'échantillon tel que la probabilité que la moyenne M_n appartienne à l'intervalle $]0{,}03\,;0{,}37[$ soit supérieure à 0,95.

On cherche à calculer n tel que $P(0{,}03 < M_n < 0{,}37) \geq 0{,}95$
Dans l'idée d'appliquer l'inégalité de concentration, on fait apparaitre l'espérance de X dans l'inégalité.
Or, $E(X) = p = 0{,}2$
Ainsi, on cherche n tel que : $P(0{,}03 - 0{,}2 < M_n - 0{,}2 < 0{,}37 - 0{,}2) \geq 0{,}95$

Soit : $P(-0{,}17 < M_n - 0{,}2 < 0{,}17) \geq 0{,}95$

Soit encore : $P(|M_n - 0{,}2| < 0{,}17) \geq 0{,}95$

Et donc, en considérant l'évènement contraire :

$$1 - P(|M_n - 0{,}2| \geq 0{,}17) \geq 0{,}95$$
$$P(|M_n - 0{,}2| \geq 0{,}17) \leq 0{,}05$$

En prenant $\delta = 0{,}17$ dans l'inégalité de concentration, on a :

$$P(|M_n - E(X)| \geq \delta) \leq 0{,}05, \text{ avec } \frac{V(X)}{n\,\delta^2} = 0{,}05.$$

Or, $V(X) = p(1-p) = 0{,}2 \times 0{,}8 = 0{,}16$
On cherche donc un entier n tel que : $\frac{0{,}16}{n\,0{,}17^2} \leq 0{,}05$
Et donc : $n \geq \frac{0{,}16}{0{,}05 \times 0{,}17^2} \approx 110{,}7$
Pour $n \geq 111$, la probabilité que la moyenne M_n appartienne à l'intervalle $]0{,}03\,;0{,}37[$ est supérieure à 0,95.

IV. Loi des grands nombres

Propriété : Soit la variable aléatoire moyenne M_n d'un échantillon de taille n de la variable aléatoire X. Pour tout réel strictement positif δ, on a :

$$\lim_{n \to +\infty} P(|M_n - E(X)| \geq \delta) = 0$$

Remarque : La loi des grands nombres traduit le fait que plus la taille de l'échantillon d'une variable aléatoire X est grande, plus l'écart entre la moyenne de cet échantillon et l'espérance de la variable aléatoire X est faible.

Méthode : Simuler des valeurs d'une variable aléatoire moyenne dans le but d'observer la loi des grands nombres

On considère la variable aléatoire X qui prend ses valeurs de manière équiprobable parmi les entiers 1 à 5.
Et on nomme M_n la variable aléatoire moyenne d'un échantillon de taille n de la variable aléatoire X.
a) Simuler 500 valeurs de la variable aléatoire M_n par une fonction en Python dans le but d'estimer la probabilité $P(|M_n - E(X)| \geq \sigma(M_n))$.
Tester le programme pour des valeurs de n de plus en plus grande.
b) Que constate-t-on ?

$$a)\ E(X) = \frac{1}{5} \times 1 + \frac{1}{5} \times 2 + \frac{1}{5} \times 3 + \frac{1}{5} \times 4 + \frac{1}{5} \times 5 = 3$$

$$V(X) = \frac{1}{5} \times (1-3)^2 + \frac{1}{5} \times (2-3)^2 + \frac{1}{5} \times (3-3)^2 + \frac{1}{5} \times (4-3)^2 + \frac{1}{5} \times (5-3)^2 = 2$$

$$\sigma(X) = \sqrt{2} \text{ donc } \sigma(M_n) = \frac{\sqrt{2}}{\sqrt{n}}$$

Ainsi : $P\big(|M_n - E(X)| \geq \sigma(M_n)\big) = P(|M_n - 3| \geq \frac{\sqrt{2}}{\sqrt{n}})$.

```
import random as rd
import math

def simulMn(n):
    S=[rd.randint(1,5) for i in range(n)]
    Mn=sum(S)/n
    return Mn

def echantMn(n):
    echant=[simulMn(n) for i in range(500)]
    c=0
    d=math.sqrt(2/n)
    for e in echant:
        if abs(e-3)>=d:
            c=c+1
    return c/500
```

```
>>> echantMn(5)
0.31
>>> echantMn(5)
0.282
>>> echantMn(10)
0.036
>>> echantMn(20)
0.008
>>> echantMn(30)
0.002
>>> echantMn(50)
0.0
>>> echantMn(100)
0.0
>>> echantMn(1000)
0.0
```

b) Plus la taille de l'échantillon est grande, plus l'écart entre la moyenne de cet échantillon et l'espérance de la variable aléatoire X est faible.