

Introduction à la statistique inférentielle : loi des grands nombres et théorème de la limite centrée

1 Études de cas

Ce paragraphe présente quatre expériences numériques dans lesquelles on va mettre en évidence quelques grands principes, qui assurent le lien entre les domaines des probabilités (variables aléatoires) et la statistique inférentielle (estimation). Les paragraphes qui suivent illustrent donc quelques uns des plus importants résultats de la théorie des probabilités : il s'agit d'étudier systématiquement le comportement d'une variable aléatoire résultant de la répétition d'une expérience aléatoire un grand nombre de fois en suivant.

1.1 Expérience n°1

1.1.1 Protocole expérimental et analyse des résultats obtenus

Activité. Le lancer d'une pièce (non truquée) à Pile ou Face est une expérience aléatoire : il n'est jamais possible de prédire, avant un lancer, le côté de la pièce qu'on obtiendra à l'issue du lancer.

On a réalisé des séries de 20 lancers, et on s'est intéressé au nombre moyen de fois où l'on a obtenu "Pile" :

Série n°	1	2	3	4	5	6	7	8	9	10
Nombres de "Pile"	13	10	11	12	9	9	12	5	9	15
Fréquences empiriques	0,65	0,50	0,55	0,60	0,45	0,45	0,60	0,25	0,45	0,75

Il est donc clair que, même si la probabilité théorique d'obtenir Pile est de $\frac{1}{2}$, sur une série de 20 lancers, les résultats Pile ou Face obtenus sont imprévisibles, et on n'observe pas l'apparition d'exactly 10 Pile et 10 Face lors d'une série de 20 lancers¹ ... Les valeurs calculées des fréquences empiriques sont erratiques, et parfois éloignées de la probabilité théorique de $\frac{1}{2}$.

Si l'on considère maintenant des séries de 200 lancers, simulé sur ordinateur, on observe les résultats suivants :

Série n°	1	2	3	4	5	6	7	8	9	10
Nombres de "Pile"	104	104	105	97	88	107	100	92	97	111
Fréquences empiriques	0,520	0,520	0,525	0,485	0,440	0,535	0,500	0,460	0,485	0,555

Sur des séries plus longues, on observe que les fréquences empiriques se sont beaucoup rapprochées de la fréquence théorique de $\frac{1}{2}$. L'apparition de cette fréquence "limite" ne doit cette fois plus rien au hasard : c'est un phénomène qui est lui tout-à-fait systématique (et qui doit être relié au fait que la probabilité théorique d'apparition de Pile est $\frac{1}{2}$).

1.1.2 Synthèse

Le principal phénomène à retenir peut se résumer dans un premier temps à quelques aspects qualitatifs simples :
- une expérience aléatoire est une expérience dont les résultats peuvent être divers, et dont aucun ne peut être prédit avec certitude avant l'expérience ;

¹Notons que la probabilité de cet événement se calcule en modélisant l'expérience au moyen d'une variable aléatoire suivant une loi Binomiale $\mathcal{B}(20; \frac{1}{2})$: on a $\mathbb{P}(X = 10) = C_{20}^{10} \left(\frac{1}{2}\right)^{10} \left(\frac{1}{2}\right)^{10} \simeq 0,176$, ce qui permet de dire que dans pratiquement 5 cas sur 6 la composition obtenue n'est pas équilibrée.

- la répétition d'une même expérience aléatoire conduit, apparemment, à des formes de régularité de certains résultats indirects de l'expérience, s'ils sont convenablement choisis et traités.

C'est la mise en évidence de ces phénomènes de régularité, et leur description, qui est l'objet des paragraphes suivants.

1.2 Expérience n°2

1.2.1 Protocole expérimental

On simule le lancer d'un dé par ordinateur, avec un tableur par exemple. On constitue des séries de lancers, qui sont appelés échantillons, dont la taille est petit à petit augmentée : les séries étudiées sont donc de plus en plus longues. Puis pour chacun des échantillons constitués, on calcule la moyenne des faces amenées par le dé dans la série considérée.

1.2.2 Analyse des résultats obtenus

Les résultats obtenus sont reportés sur le document 1. On peut faire les remarques suivantes :

- au départ, pour de petits échantillons, cette moyenne est imprévisible, et son comportement est erratique ;
- au fur et à mesure que la taille de l'échantillon grandit, la moyenne calculée se stabilise aux environs de 3,5 ;
- la stabilisation de la moyenne est plus rapide au début, mais se ralentit par la suite.

Si X est la variable aléatoire correspondant au numéro de la face sortie, remarquons que la loi de X (qui régit donc un lancer est donnée par

Face	1	2	3	4	5	6
Probabilité associée	1/6	1/6	1/6	1/6	1/6	1/6

que $\mathbb{E}(X) = 3,5$, et donc que la moyenne théorique des numéros amenés est 3,5.

1.2.3 Synthèse

Soit X une variable aléatoire d'espérance $\mathbb{E}(X)$. Considérons n répétitions (indépendantes) de l'expérience aléatoire qui conduit à observer n réalisations de X : $X_1, X_2, \dots, X_n$.

Considérons la valeur moyenne M_n des X_i ($1 \leq i \leq n$) sur cette série de réalisation :

$$M_n = \frac{X_1 + X_2 + \dots + X_n}{n}.$$

Alors M_n est une variable aléatoire, et $\lim_{n \rightarrow +\infty} M_n = \mathbb{E}(X)$, ou encore :

$$\lim_{n \rightarrow +\infty} \frac{X_1 + X_2 + \dots + X_n}{n} = \mathbb{E}(X).$$

En clair, lorsque $n \rightarrow +\infty$, il y a convergence de la moyenne empiriques des n réalisations de X vers $\mathbb{E}(X)$.

1.3 Expérience n°3

1.3.1 Protocole expérimental

On simule le lancer d'un dé par ordinateur, toujours à l'aide d'un tableur. On constitue un grand nombre d'échantillons, dont la taille est variable (de 5, 10 ou 20 lancers) : les échantillons de taille identique sont comptabilisés selon leur moyenne (c'est à dire que pour toutes les valeurs possibles des moyennes, on compte de nombre d'échantillon présentant cette moyenne).

1.3.2 Analyse des résultats obtenus

Les résultats obtenus sont reportés sur le document 2. On peut faire les remarques suivantes :

- tous les échantillons ne sont pas de même moyenne, leur répartition par moyenne identique est diffuse ;
- lorsque la taille des échantillons est faible, les différences de moyennes entre eux sont plus grandes que lorsqu'il s'agit d'échantillon de plus grande taille. Cela traduit l'idée que l'augmentation de la taille des échantillons diminue les différences de composition des échantillons ;
- lorsque la taille des échantillons augmente, leur répartition quant à leur moyenne tend à se rapprocher de celle donnée par une loi normale ;
- la moyenne de cette loi normale est constante (autour de 3,5), mais son écart-type diminue avec la taille des échantillons considérés.

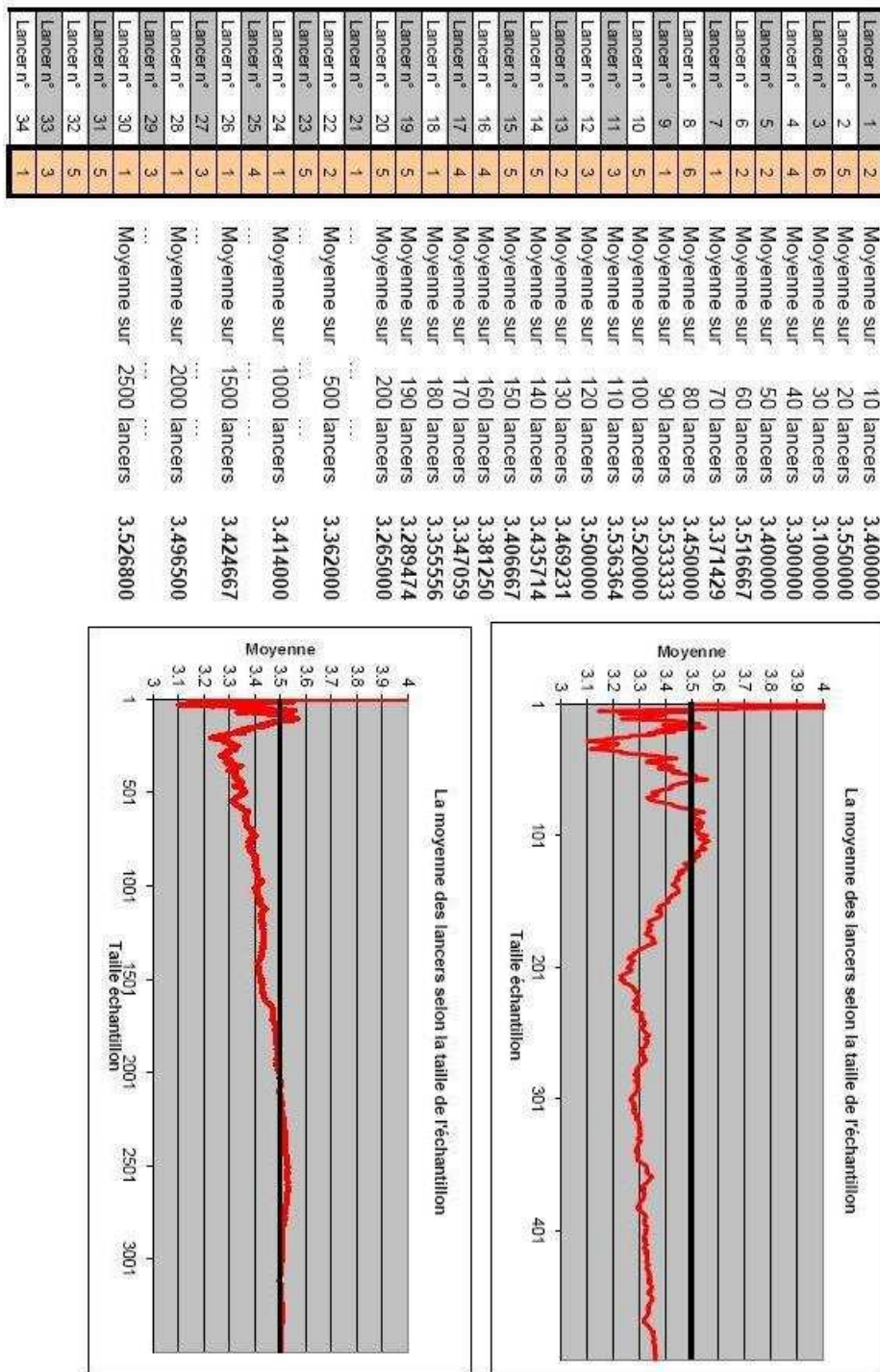


FIG. 1 – Document 1

1.3.3 Synthèse

Soit X une variable aléatoire d'espérance m .

Intéressons-nous maintenant à n répétitions ($n \in \mathbb{N}$) de l'expérience aléatoire associée à la variable X : chaque répétition conduit à une réalisation X_i ($1 \leq i \leq n$) de X . On obtient ainsi une série de n variables aléatoires $X_1, X_2, \dots, X_n$ (indépendantes). La répartition des échantillons de taille n selon la moyenne calculée $M_n = \frac{X_1 + X_2 + \dots + X_n}{n}$ des réalisations de X sur chaque échantillon tend à se rapprocher de celle donnée une loi normale de moyenne $\mathbb{E}(X)$, et d'écart-type décroissant lorsque $n \rightarrow +\infty$.

Fluctuation d'échantillonnage

Echantillon n°	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Lancer n° 1	4	2	3	4	5	2	6	7	8	9	10	11	12	0	1	4	2	6	4	0	3	0	3	0	4	3	2	2	0	2
Lancer n° 2	0	2	0	0	0	0	0	0	4	3	2	0	0	0	4	3	1	0	0	4	0	3	2	4	0	2	3	4	0	2
Lancer n° 3	4	2	2	3	1	3	4	4	2	0	1	0	5	1	3	0	2	1	2	1	4	1	4	0	1	0	3	3	1	0
Lancer n° 4	0	0	0	0	0	3	0	4	4	6	5	4	6	4	3	0	0	1	5	4	0	1	2	5	3	5	1	4	0	3
Lancer n° 5	0	4	2	5	3	1	4	5	3	4	4	4	4	5	2	3	0	4	3	0	2	2	3	1	3	1	2	0	5	5
Lancer n° 6	0	1	2	1	2	1	0	0	0	2	4	1	0	5	5	4	0	2	3	3	2	0	2	0	0	0	4	4	2	0
Lancer n° 7	4	1	4	2	2	2	4	1	0	2	1	5	0	1	4	2	3	0	4	1	0	1	2	0	3	4	1	5	2	2
Lancer n° 8	1	2	0	4	2	2	0	6	1	2	1	1	2	1	1	5	0	1	2	2	1	4	3	1	0	2	0	2	3	3
Lancer n° 9	0	2	1	1	0	4	1	3	3	0	6	1	1	2	0	1	1	5	1	4	1	1	5	1	2	5	4	2	4	2
Lancer n° 10	2	0	3	1	3	2	0	0	2	5	3	4	2	1	5	1	0	5	0	0	5	0	5	1	0	6	5	3	4	3
Lancer n° 11	2	0	4	4	1	2	5	2	4	3	1	0	4	3	4	2	5	0	4	4	1	2	1	4	0	0	3	2	5	0
Lancer n° 12	1	0	0	1	4	1	6	2	5	0	1	0	6	2	5	4	3	4	1	3	3	1	2	3	2	4	4	1	0	1
Lancer n° 13	0	0	2	3	4	0	2	0	5	3	2	1	0	2	1	3	0	2	1	3	1	1	3	0	5	4	4	0	4	0
Lancer n° 14	0	1	1	2	3	1	1	4	2	1	2	4	4	4	1	0	5	3	2	1	0	4	3	4	4	0	5	1	0	2
Lancer n° 15	4	0	1	4	4	2	2	3	2	3	2	3	5	0	4	0	5	2	5	4	1	5	4	1	5	4	1	3	2	1
Lancer n° 16	2	5	2	2	2	4	2	0	1	4	4	0	0	5	3	1	4	2	4	5	3	0	2	1	4	2	2	1	3	0
Lancer n° 17	4	0	1	0	1	5	4	1	1	1	2	0	2	1	5	1	3	4	0	6	4	1	0	5	4	4	2	5	0	3
Lancer n° 18	0	3	3	1	1	1	3	2	3	5	0	5	0	4	4	3	2	0	1	0	2	1	0	5	0	6	2	4	1	3
Lancer n° 19	3	1	5	2	2	1	5	3	0	0	1	3	1	0	2	1	0	2	5	1	0	4	5	4	0	3	2	4	3	5
Lancer n° 20	0	0	0	0	3	5	1	4	0	0	4	0	0	0	1	1	0	0	4	1	4	5	3	2	0	2	3	1	1	4

Moyenne par échantillon																														
Echantillon n°	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
de taille 5 :	4.00	3.20	3.33	4.80	3.56	3.00	4.78	4.60	3.44	3.60	3.44	4.20	4.33	2.80	3.78	2.80	3.33	4.20	3.33	4.60	2.67	2.80	3.78	2.40	4.33	2.00	3.00	4.40	3.11	3.40
de taille 10 :	3.80	3.30	3.30	3.40	3.50	3.10	4.90	4.40	3.30	3.50	3.40	3.30	4.10	2.40	3.90	2.70	3.60	4.00	3.60	3.80	2.90	3.10	3.90	2.50	4.50	3.30	3.20	3.80	3.20	3.30
de taille 20 :	3.90	3.96	3.10	3.10	3.00	2.96	4.00	3.75	3.36	3.55	2.96	4.25	3.95	2.75	3.55	2.80	3.40	4.20	3.75	3.50	3.00	3.16	3.76	3.00	4.40	3.56	3.36	3.25	3.26	3.20

Moyenne par échantillon

Echantillon n°	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
de taille 5 :	4.00	3.20	3.33	4.80	3.56	3.00	4.78	4.60	3.44	3.60	3.44	4.20	4.33	2.80	3.78	2.80	3.33	4.20	3.33	4.60	2.67	2.80	3.78	2.40	4.33	2.00	3.00	4.40	3.11	3.40
de taille 10 :	3.80	3.30	3.40	3.50	3.10	4.90	4.40	3.40	3.30	3.50	3.40	3.80	4.10	2.40	3.90	2.70	3.60	4.00	3.60	2.90	3.10	3.90	2.50	4.50	3.30	3.20	3.80	3.20	3.30	3.30
de taille 20 :	3.90	3.95	3.10	3.10	3.00	2.95	4.00	3.75	3.35	3.55	2.95	4.25	3.95	2.75	3.55	2.80	3.40	4.20	3.75	3.50	3.00	3.15	3.75	3.00	4.40	3.55	3.35	3.25	3.25	3.20

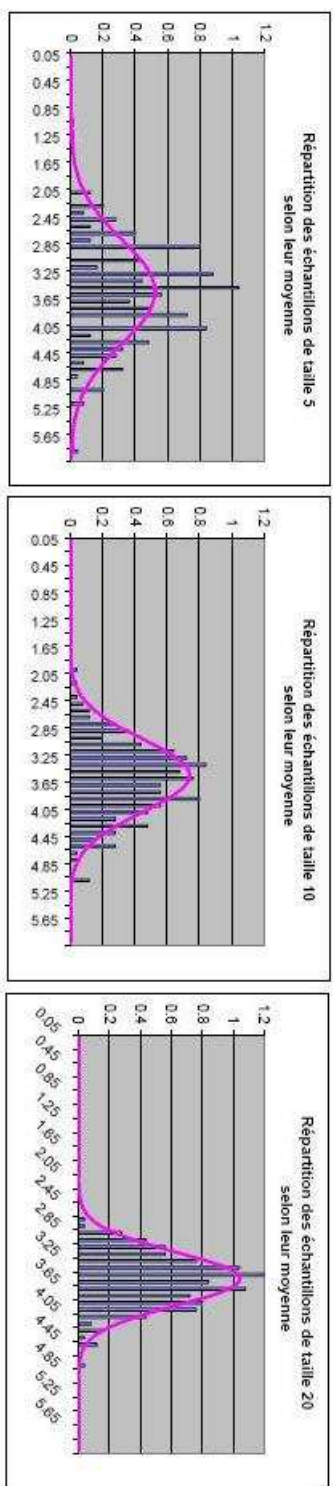


FIG. 2 – Document 2

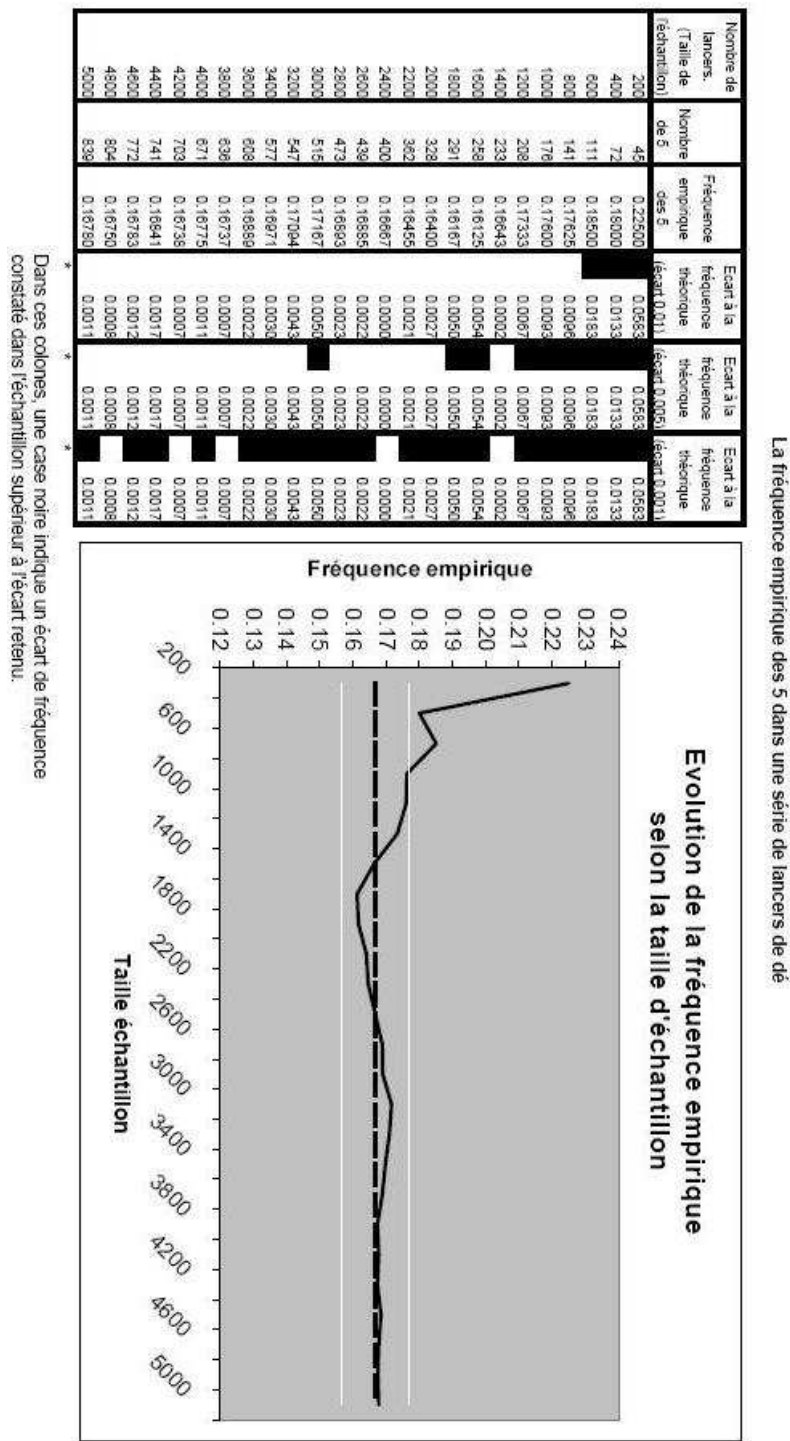
1.4 Expérience n°4

1.4.1 Protocole expérimental

On simule le lancer d'un dé par ordinateur, toujours à l'aide d'un tableur. On constitue des échantillons de taille croissante, et pour chacun d'eux, on observe la fréquence d'apparition du chiffre 6.

1.4.2 Analyse des résultats obtenus

Les résultats obtenus sont reportés sur le document 3. On peut faire les remarques suivantes :



Dans ces colonnes, une case noire indique un écart de fréquence constaté dans l'échantillon supérieur à l'écart retenu.

FIG. 3 – Document 3

- lorsque la taille des échantillons augmente, la fréquence d'apparition du chiffre 5 dans la série observée tend vers un nombre compris entre 0,16 et 0,17, alors que la fréquence théorique d'apparition du chiffre 5 est $\frac{1}{6} \simeq 0,166$;
- la fréquence théorique d'apparition du chiffre 5 peut être retrouvée, si on ne la connaissait pas, en réalisant la répétition, un grand nombre de fois (et par ordinateur), du lancer du dé ;
- la convergence ne se fait pas de manière monotone pour tout écart par rapport à la fréquence théorique, il peut arriver, à n'importe quel moment, un échantillon conduisant à une fréquence d'apparition du chiffre 5 qui dépasse cet

écart (même si c'est de moins en moins probable lorsque la taille de l'échantillon augmente).

1.4.3 Synthèse

Lorsqu'on est face à un phénomène aléatoire qui conduit à l'observation d'une variable aléatoire discrète dont la loi est inconnue (en tout cas, les probabilités de réalisation de chaque valeur), une simulation faisant apparaître des échantillons de grande taille de réalisations des issues de l'expérience permet de déterminer empiriquement la loi de probabilité.

2 Loi (faible) des grands nombres

2.1 Formulation heuristique de la loi (faible) des grands nombres

Le phénomène mis ici en évidence, à savoir la convergence des fréquences empiriques observées vers la probabilité théorique d'apparition, et ce lorsque l'expérience est répétée un nombre de fois de plus en plus grand, porte le nom de "loi (faible) des grands nombres". Celle-ci peut s'énoncer simplement de la manière suivante :

Loi faible des grands nombres : *La répétition un grand nombre de fois d'une même expérience aléatoire produit une stabilisation des fréquences d'apparition empiriques de chacune des issues possibles de cette expérience aléatoire. De plus, les écarts entre la fréquence théorique d'apparition de chacune des issues possibles de l'expérience aléatoire et les fréquences empiriques diminuent avec la taille des séries aléatoires utilisés pour les calculs empiriques.*

Voyons maintenant la formulation mathématique de ce résultat.

2.2 Formulation mathématique de la loi (faible) des grands nombres

2.2.1 Inégalité de Bienaymé-Tchebychev

Nous débutons ce paragraphe par un résultat classique sur les variables aléatoires, que nous admettrons.

Proposition 1 (Inégalité de Bienaymé-Tchebychev) *Soit X une variable aléatoire admettant une espérance et une variance. Alors, pour tout $t > 0$, on a :*

$$\mathbb{P}(|X - \mathbb{E}(X)| > \sigma_X t) \leq \frac{1}{t^2}.$$

L'objet de l'exemple suivant est de fournir un cas dans lequel on peut établir la loi faible des grands nombres avec les résultats figurant au programme des BTS.

Exemple 1 *Cas d'une variable aléatoire suivant une loi de Bernoulli.*

Soit X une variable aléatoire suivant une loi de Bernoulli $\mathcal{B}(1; p)$.

(On peut envisager les exemples suivants :

i. *X est la variable qui vaut 1 si le lancer d'un dé à 6 faces amène le "5", et 0 sinon. Alors $X \rightsquigarrow \mathcal{B}(1; \frac{1}{5})$.*

ii. *X est la variable qui vaut 1 si le lancer d'une pièce amène le pile "P", et 0 sinon. Alors $X \rightsquigarrow \mathcal{B}(1; \frac{1}{2})$.*

On pourrait les multiplier ...)

Considérons une série de n lancers successifs, que nous supposons indépendants. Le résultat (échec ou réussite) de la n -ième expérience est noté X_n . La fréquence empirique de succès dans la série de longueur n est donnée par la variable aléatoire $F_n = \frac{X_1 + \dots + X_n}{n}$, puisque $S_n = X_1 + \dots + X_n$ compte le nombre de succès dans la série. On note que S_n suit une loi Binomiale $\mathcal{B}(n; p)$, les X_i étant supposées indépendantes. Donc, pour tout $n \in \mathbb{N}$ fixé,

$$\mathbb{E}(F_n) = \mathbb{E}\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{1}{n}(\mathbb{E}(X_1) + \dots + \mathbb{E}(X_n)) = \frac{1}{n}(p + \dots + p) = \frac{1}{n}np = p,$$

et, par indépendance des X_i , on a

$$\mathbb{V}(F_n) = \mathbb{V}\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{1}{n^2}(\mathbb{V}(X_1) + \dots + \mathbb{V}(X_n)) = \frac{1}{n^2}(p(1-p) + \dots + p(1-p)) = \frac{1}{n}p(1-p),$$

et donc $\sigma(F_n) = \sqrt{\frac{p(1-p)}{n}}$.

En appliquant l'inégalité de Bienaymé-Tchebychev ci-dessus, on obtient, pour tout $t > 0$, la relation

$$\mathbb{P}\left(|F_n - \mathbb{E}(X)| > t\sqrt{\frac{p(1-p)}{n}}\right) \leq \frac{1}{t^2},$$

soit encore

$$\mathbb{P}\left(|F_n - \mathbb{E}(X)| > t\sqrt{\frac{p(1-p)}{n}}\right) \leq \frac{1}{t^2},$$

ce qui s'écrit aussi

$$\mathbb{P}\left(F_n \notin \left[\mathbb{E}(X) - t\frac{\sigma_X}{\sqrt{n}}; \mathbb{E}(X) + t\frac{\sigma_X}{\sqrt{n}}\right]\right) \leq \frac{1}{t^2}.$$

Autrement dit, pour tout $t > 0$, il existe une taille N d'échantillon au-delà de laquelle toute fréquence empirique F_n calculée sur un échantillon de taille $n \geq N$ s'"éloigne" de p avec une probabilité inférieure à $\frac{1}{t^2}$.

2.2.2 La loi des grand nombres pour les fréquences empiriques

L'exemple précédent constitue la "preuve" du théorème suivant :

Théorème 2 (Loi faible des grands nombres) Soit X une variable aléatoire de Bernoulli de paramètre p . Pour tout $n \in \mathbb{N}$, on note F_n la moyenne empirique du nombre de fois où $X = 1$ sur un échantillon de taille n . Alors, avec une probabilité aussi proche de 1 qu'on le souhaite, il existe une taille N d'échantillon au-delà de laquelle toute fréquence empirique F_n calculée sur un échantillon de taille $n \geq N$ est aussi proche de p qu'on le souhaite.

Exemple 2 Supposons que l'on s'intéresse à une série de lancers d'une pièce de monnaie. On note F_n le nombre moyen de lancers qui amènent "Pile" sur une série de n lancers successifs indépendants.

Dans ces conditions, on peut employer la formule obtenue dans l'exemple précédent le théorème, pour $X \rightsquigarrow \mathcal{B}\left(1; \frac{1}{2}\right)$.

Posons $t = 10$. Alors, puisque $\mathbb{E}(X) = \frac{1}{2}$, et $\sigma_X = \frac{1}{2}$,

$$\mathbb{P}\left(F_n \notin \left[\frac{1}{2} - \frac{10}{2} \frac{1}{\sqrt{n}}; \frac{1}{2} + \frac{10}{2} \frac{1}{\sqrt{n}}\right]\right) \leq \frac{1}{100}, \text{ soit donc } \mathbb{P}\left(F_n \in \left[\frac{1}{2} - \frac{5}{\sqrt{n}}; \frac{1}{2} + \frac{5}{\sqrt{n}}\right]\right) \geq \frac{99}{100}.$$

Pour $n = 200$ lancers successifs, on a $\mathbb{P}(F_n \in [0, 15; 0, 85]) \geq \frac{99}{100}$ (c'est-à-dire que $F_n \in [0, 15; 0, 85]$ avec une certitude de 99%).

Pour $n = 1.000$ lancers successifs, on a $\mathbb{P}(F_n \in [0, 34; 0, 66]) \geq \frac{99}{100}$ (c'est-à-dire que $F_n \in [0, 34; 0, 66]$ avec une certitude de 99%).

Pour $n = 10.000$ lancers successifs, on a $\mathbb{P}(F_n \in [0, 45; 0, 55]) \geq \frac{99}{100}$ (c'est-à-dire que $F_n \in [0, 45; 0, 55]$ avec une certitude de 99%).

Pour $n = 100.000$ lancers successifs, on a $\mathbb{P}(F_n \in [0, 48; 0, 52]) \geq \frac{99}{100}$ (c'est-à-dire que $F_n \in [0, 48; 0, 52]$ avec une certitude de 99%).

Ainsi, avec une certitude de 99%, il existe toujours un nombre de répétitions de l'expérience aléatoire qui conduit à l'observation d'une fréquence empirique aussi proche de $\frac{1}{2}$ qu'on le souhaite.

Exemple 3 Supposons que l'on s'intéresse au même exemple que le précédent, mais qu'on se contente de certitudes à 95%. Dans ces conditions, on peut employer la formule obtenue dans l'exemple précédent le théorème, pour $X \rightsquigarrow \mathcal{B}\left(1; \frac{1}{2}\right)$. Posons $t_0 = \frac{10}{\sqrt{5}}$. Alors, puisque $\mathbb{E}(X) = \frac{1}{2}$, et $\sigma_X = \frac{1}{2}$,

$$\mathbb{P}\left(F_n \notin \left[\frac{1}{2} - \frac{t_0}{2\sqrt{n}}; \frac{1}{2} + \frac{t_0}{2\sqrt{n}}\right]\right) \leq \frac{1}{5}, \text{ soit donc } \mathbb{P}\left(F_n \in \left[\frac{1}{2} - \frac{2,23}{\sqrt{n}}; \frac{1}{2} + \frac{2,23}{\sqrt{n}}\right]\right) \geq \frac{95}{100}.$$

Pour $n = 100$ lancers successifs, on a $\mathbb{P}(F_n \in [0, 28; 0, 72]) \geq \frac{95}{100}$ (c'est-à-dire que $F_n \in [0, 28; 0, 72]$ avec une certitude de 95%).

Pour $n = 200$ lancers successifs, on a $\mathbb{P}(F_n \in [0, 34; 0, 66]) \geq \frac{95}{100}$ (c'est-à-dire que $F_n \in [0, 34; 0, 66]$ avec une certitude de 95%).

Pour $n = 1.000$ lancers successifs, on a $\mathbb{P}(F_n \in [0, 43; 0, 57]) \geq \frac{95}{100}$ (c'est-à-dire que $F_n \in [0, 43; 0, 57]$ avec une certitude de 95%).

Pour $n = 10.000$ lancers successifs, on a $\mathbb{P}(F_n \in [0, 48; 0, 52]) \geq \frac{95}{100}$ (c'est-à-dire que $F_n \in [0, 48; 0, 52]$ avec une certitude de 95%).

Pour $n = 100.000$ lancers successifs, on a $\mathbb{P}(F_n \in [0, 49; 0, 51]) \geq \frac{95}{100}$ (c'est-à-dire que $F_n \in [0, 49; 0, 51]$ avec une certitude de 95%).

Exemple 4 Supposons que l'on s'intéresse toujours aux lancers successifs d'une pièce à Pile ou Face. Nous évaluons, avec ce théorème, la probabilité que la moyenne empirique F_n soit comprise dans l'intervalle $[0, 4; 0, 6]$ pour des séries de lancers de plus en plus longues :

Pour $n = 100$ lancers successifs, on a $\mathbb{P}(F_n \in [0, 4; 0, 6]) \geq \frac{1}{4}$ en posant $t = 2$ (c'est-à-dire que $F_n \in [0, 4; 0, 6]$ avec une certitude de 75%).

Pour $n = 200$ lancers successifs, on a $\mathbb{P}(F_n \in [0, 4; 0, 6]) \geq \frac{1}{8}$ en posant $t = 2\sqrt{2}$ (c'est-à-dire que $F_n \in [0, 4; 0, 6]$ avec une certitude de 87,5%).

Pour $n = 1.000$ lancers successifs, on a $\mathbb{P}(F_n \in [0, 4; 0, 6]) \geq \frac{1}{40}$ en posant $t = 2\sqrt{10}$ (c'est-à-dire que $F_n \in [0, 4; 0, 6]$ avec une certitude de 97,5%).

Pour $n = 10.000$ lancers successifs, on a $\mathbb{P}(F_n \in [0, 4; 0, 6]) \geq \frac{1}{400}$ en posant $t = 20$ (c'est-à-dire que $F_n \in [0, 4; 0, 6]$ avec une certitude de 99,75%).

Pour $n = 100.000$ lancers successifs, on a $\mathbb{P}(F_n \in [0, 4; 0, 6]) \geq \frac{1}{4}$ en posant $t = 20\sqrt{10}$ (c'est-à-dire que $F_n \in [0, 4; 0, 6]$ avec une certitude de 99,975%).

Ainsi, lorsque la longueur de la série des observations grandit, on obtient une valeur moyenne empirique qui reste de plus en plus souvent dans un intervalle centré sur la valeur moyenne théorique.

Exemple 5 Un sondage politique "brut" fournit un exemple de situation décrite par le phénomène précédent : supposons que la proportion de votant pour un candidat C au second tour d'une élection présidentielle soit proche de $\frac{1}{2}$. Pour chaque personne interrogée, on réalise une expérience de Bernoulli, de paramètre p , dans laquelle les réponses successives des sondés sont indépendantes (le sondé vote pour C ou pas ...). Le nombre moyen empirique de votants pour C parmi n personnes interrogées est F_n . On sait que, pour tout $t > 0$, $\mathbb{P}\left(F_n \notin \left[p - \frac{t}{2\sqrt{n}}; p + \frac{t}{2\sqrt{n}}\right]\right) \leq \frac{1}{t^2}$, d'après l'étude précédent le théorème.

Posons $t = 2\sqrt{5}$. Alors, pour un sondage portant sur $n = 1.000$ personnes, on a :

$\mathbb{P}\left(F_n \notin \left[p - \frac{t}{2\sqrt{n}}; p + \frac{t}{2\sqrt{n}}\right]\right) \leq \frac{5}{100}$: il y a donc 95% de chances que le résultat estimé par sondage soit le bon,

avec une marge d'erreur de l'ordre de $\pm \frac{t}{2\sqrt{n}} = \pm \frac{\sqrt{5}}{\sqrt{1.000}} \simeq \pm 0,07$, soit ± 7 points ...

On est donc prié de manipuler avec d'infinies précautions les résultats des sondages ... (même si les intervalles obtenus par les sondeurs font que ceux-ci sont fiables à 95% avec une erreur de ± 3 points, ainsi que cela sera vu dans les applications du théorème de la limite centrée).

2.2.3 La loi des grand nombres pour les moyennes empiriques

3 Théorème de la limite centrée

Ce théorème porte, pour les Anglo-Saxons, le nom de "Central Limit Theorem". Il est parfois abusivement appelé, pour cette raison, "Théorème Central Limite" en français, ce qui ne signifie pas grand'chose ...

3.1 Mise en évidence du théorème de la limite centrée

Exemple 6 Considérons une variable aléatoire X admettant une espérance m et une variance $\mathbb{V}(X) \neq 0$. On note $\sigma \neq 0$ l'écart-type de X .

Intéressons-nous maintenant à n répétitions ($n \in \mathbb{N}$) de l'expérience aléatoire associée à la variable X : chaque

répétition conduit à une réalisation X_i ($1 \leq i \leq n$) de X . On obtient ainsi une série de n variables aléatoires $X_1, X_2, \dots, X_n$, que nous supposons indépendantes. On considère, pour tout $n \in \mathbb{N}$, la moyenne empirique de $X_1, \dots, X_n$, soit donc la variable aléatoire Z_n définie par la relation $Z_n = \frac{X_1 + X_2 + \dots + X_n}{n}$. Lorsque $n \rightarrow +\infty$, on observe que Z_n suit de plus en plus exactement une loi Normale $\mathcal{N}\left(m; \frac{\sigma}{\sqrt{n}}\right)$.

3.2 Formulation mathématique du théorème de la limite centrée

Théorème 3 Soit X une variable aléatoire admettant une espérance m , une variance $\mathbb{V}(X) \neq 0$ (et donc un écart-type $\sigma \neq 0$).

Soit $(X_i)_{i \in \mathbb{N}}$ une suite de variables aléatoires indépendantes et identiquement distribuées selon la loi de X . Alors, pour $n \in \mathbb{N}$ suffisamment grand, $Z_n = \frac{X_1 + X_2 + \dots + X_n}{n}$ suit approximativement une loi Normale $\mathcal{N}\left(m; \frac{\sigma}{\sqrt{n}}\right)$.

Remarque 1 " n suffisamment grand" signifie en pratique $n \geq 30$.

Preuve

◀ Nous admettons la partie difficile de ce théorème, qui est le fait que Z_n finisse par suivre une loi Normale lorsque $n \rightarrow +\infty$.

Pour le reste, notons que si $n \in \mathbb{N}^*$ est un entier quelconque, on a d'une part

$$Z_n = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{X_1}{n} + \frac{X_2}{n} + \dots + \frac{X_n}{n}$$

et par conséquent

$$\mathbb{E}(Z_n) = \mathbb{E}\left(\frac{X_1}{n} + \frac{X_2}{n} + \dots + \frac{X_n}{n}\right) = \frac{\mathbb{E}(X_1)}{n} + \frac{\mathbb{E}(X_2)}{n} + \dots + \frac{\mathbb{E}(X_n)}{n},$$

par linéarité de l'espérance. Mais comme $\mathbb{E}(X_1) = \mathbb{E}(X_2) = \dots = \mathbb{E}(X_n) = \mathbb{E}(X) = m$, on a finalement $\mathbb{E}(Z_n) = \frac{nm}{n} = m$. On a également

$$\mathbb{V}(Z_n) = \mathbb{V}\left(\frac{X_1}{n} + \frac{X_2}{n} + \dots + \frac{X_n}{n}\right) = \frac{\mathbb{V}(X_1)}{n^2} + \frac{\mathbb{V}(X_2)}{n^2} + \dots + \frac{\mathbb{V}(X_n)}{n^2},$$

par propriété de la variance vis-à-vis de variables aléatoires indépendantes. Comme $\mathbb{V}(X_1) = \mathbb{V}(X_2) = \dots = \mathbb{V}(X_n) = \mathbb{V}(X) = \sigma^2$, on a finalement $\mathbb{V}(Z_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$, et donc $\sigma_{Z_n} = \frac{\sigma}{\sqrt{n}}$. ▶

Remarque 2 Dans le cas où X suit une loi normale, les variables aléatoires $X_1, X_2, \dots, X_n$ suivent aussi une loi normale, de même que $\frac{X_1}{n}, \frac{X_2}{n}, \dots, \frac{X_n}{n}$; en supposant que les $(X_i)_{i \in \mathbb{N}}$ sont indépendantes, on peut affirmer que Z_n suit aussi une loi normale (la somme indépendante de lois normales est une loi normale, ainsi qu'on l'a admis dans un chapitre antérieur!).

Exemple 7 On s'intéresse à l'expérience aléatoire consistant à lancer un dé à 6 faces équilibré. On note X la variable aléatoire égale à la face amenée. La loi de X est bien connue, mais notons que $\mathbb{E}(X) = 3,5$ et que $\sigma_X \simeq 1,708$.

• Étude des échantillons de taille 1.

Ces échantillons sont au nombre de 6, et les valeurs moyennes de X calculées sur de tels échantillons se répartissent selon la loi

Z_1	1	2	3	4	5	6
Probabilité associée	1/6	1/6	1/6	1/6	1/6	1/6

(en effet, si l'échantillon de taille 1 obtenu conduit à observer $X_1 = 1$, alors $Z_1 = 1$, etc.).

De plus $\mathbb{E}(Z_1) = 3,5$ et $\sigma_{Z_1} \simeq 1,708$.

• Étude des échantillons de taille 2.

Ces échantillons sont au nombre de 36, et les valeurs moyennes de X calculées sur de tels échantillons se répartissent selon la loi

Z_2	1	1,5	2	2,5	3	3,5	4	4,5	5	5,5	6
Probabilité associée	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36

(en effet, si l'échantillon de taille 2 obtenu conduit à observer $X_1 = 1$ et $X_2 = 4$, alors $Z_2 = \frac{X_1 + X_2}{2} = 2,5$, etc.).

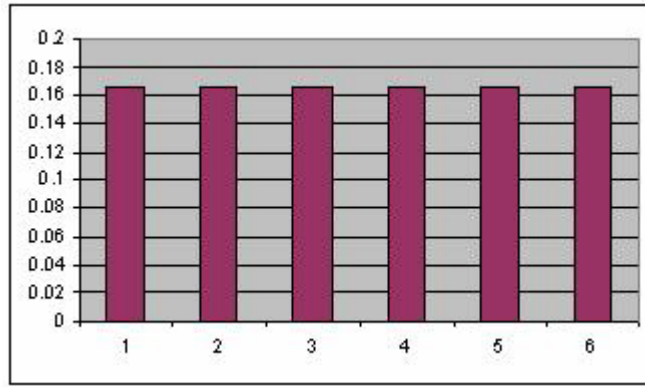


FIG. 4 – Répartition des valeurs moyennes de X calculées sur des échantillons de taille 1.

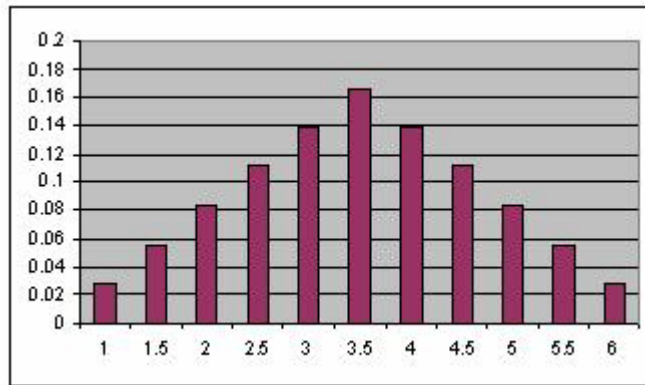


FIG. 5 – Répartition des valeurs moyennes de X calculées sur des échantillons de taille 2.

De plus $\mathbb{E}(Z_2) = 3,5$ et $\sigma_{Z_2} \simeq 1,208$. On a bien $1,208 \simeq \frac{1,708}{\sqrt{2}}$.

• Étude des échantillons de taille 3.

Ces échantillons sont au nombre de $6^3 = 216$, et les valeurs moyennes de X calculées sur de tels échantillons se répartissent selon la loi

1	4/3	5/3	2	7/3	8/3	3	10/3	11/3	4	13/3	14/3	5	16/3	17/3	6
1/6 ³	3/6 ³	6/6 ³	10/6 ³	15/6 ³	21/6 ³	25/6 ³	27/6 ³	27/6 ³	25/6 ³	21/6 ³	15/6 ³	10/6 ³	6/6 ³	3/6 ³	1/6 ³

(en effet, si l'échantillon de taille 3 obtenu conduit à observer $X_1 = 1$, $X_2 = 4$ et $X_3 = 5$, alors $Z_3 = \frac{X_1 + X_2 + X_3}{3} = \frac{10}{3}$, etc.).

De plus $\mathbb{E}(Z_3) = 3,5$ et $\sigma_{Z_3} \simeq 0,986$. On a bien $0,986 \simeq \frac{1,708}{\sqrt{3}}$.

• Étude des échantillons de taille n .

Le théorème de la limite centrée permet, pour n grand (en pratique $n \geq 30$), de prévoir la répartition approximative de Z_n .

Le corollaire suivant est une conséquence du phénomène de centrage-réduction d'une variable aléatoire suivant une loi $\mathcal{N}(m; \sigma)$. il justifie la terminologie de théorème de la limite centrée :

Corollaire 4 Sous les mêmes hypothèses, pour $n \in \mathbb{N}$ suffisamment grand, $\frac{Z_n - m}{\frac{\sigma}{\sqrt{n}}}$ suit approximativement une loi Normale $\mathcal{N}(0; 1)$.

Remarque 3 " n suffisamment grand" signifie en pratique $n \geq 30$.

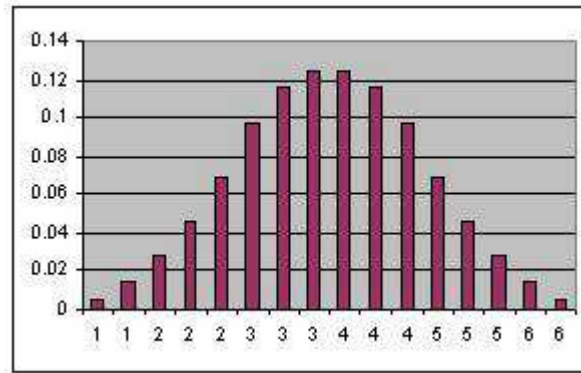


FIG. 6 – Répartition des valeurs moyennes de X calculées sur des échantillons de taille 3.

Table des matières

1 Études de cas	1
1.1 Expérience n°1	1
1.1.1 Protocole expérimental et analyse des résultats obtenus	1
1.1.2 Synthèse	1
1.2 Expérience n°2	2
1.2.1 Protocole expérimental	2
1.2.2 Analyse des résultats obtenus	2
1.2.3 Synthèse	2
1.3 Expérience n°3	2
1.3.1 Protocole expérimental	2
1.3.2 Analyse des résultats obtenus	2
1.3.3 Synthèse	3
1.4 Expérience n°4	4
1.4.1 Protocole expérimental	4
1.4.2 Analyse des résultats obtenus	5
1.4.3 Synthèse	6
2 Loi (faible) des grands nombres	6
2.1 Formulation heuristique de la loi (faible) des grands nombres	6
2.2 Formulation mathématique de la loi (faible) des grands nombres	6
2.2.1 Inégalité de Bienaymé-Tchebychev	6
2.2.2 La loi des grand nombres pour les fréquences empiriques	7
2.2.3 La loi des grand nombres pour les moyennes empiriques	8
3 Théorème de la limite centrée	8
3.1 Mise en évidence du théorème de la limite centrée	8
3.2 Formulation mathématique du théorème de la limite centrée	9