♦ Chapitre 16. Comportements asymptotiques et prise de décision

I. — Inégalités de concentration

1) Inégalités de Markov et de Bienaymé-Tchebychev

Propriété 1. — Inégalité de Markov

Soit X une variable aléatoire réelle à valeurs positives ou nulles. On suppose que X admet une espérance. Alors, pour tout réel $a > 0$,

$$\mathbf{P}(X \geqslant a) \leqslant \frac{\mathbf{E}(X)}{a}.$$

Exemple 2. On lance 100 fois de suite une pièce équilibrée. Montrer que la probabilité d'obtenir au moins 90 fois *pile* est inférieure ou égale à $\frac{5}{9}$.

Théorème 3. — Inégalité de Bienaymé-Tchebychev

Soit X une variable aléatoire réelle. On suppose que X admet une variance. Alors, pour tout réel $\varepsilon > 0$,

$$\mathbf{P}(|X - \mathbf{E}(X)| \geqslant \varepsilon) \leqslant \frac{\mathbf{V}(X)}{\varepsilon^2}.$$

Exemple 4. Montrer que l'inégalité de Bienaymé-Tchebychev permet d'améliorer de façon très significative la majoration de l'exemple précédent.

Remarque 5. En utilisant le fait que l'évènement $\{|X - \mathbf{E}(X)| < \varepsilon\}$ est l'évènement contraire de $\{|X - \mathbf{E}(X)| \geqslant \varepsilon\}$, l'inégalité de Bienaymé-Tchebychev permet de minorer la probabilité $\mathbf{P}(|X - \mathbf{E}(X)| < \varepsilon)$.

Exemple 6. Soit X une variable aléatoire réelle admettant une variance. On note σ l'écart-type de X. Montrer que

$$\mathbf{P}(\mathbf{E}(X) - 2\sigma < X < \mathbf{E}(X) + 2\sigma) \geqslant \frac{3}{4}.$$

2) Loi faible des grands nombres

Dans tout ce paragraphe, n est un entier naturel non nul et X_1, X_2, ..., X_n sont des variables aléatoires indépendantes et de même loi. On note, de plus, μ et σ respectivement l'espérance et l'écart-type de chacune de ces variables aléatoires.

Définition 7

La variable aléatoire $M_n = \dfrac{1}{n}\displaystyle\sum_{k=1}^{n} X_k$ est appelée la **moyenne empirique** des variables aléatoires X_1, X_2, ..., X_n.

Propriété 8

Soit M_n la moyenne empirique de X_1, X_2, ..., X_n. Alors

$$\mathbf{E}(M_n) = \mu \qquad \text{et} \qquad \mathbf{V}(M_n) = \frac{\sigma^2}{n}.$$

Exemple 9. On suppose que X_1, X_2, ..., X_n sont des variables aléatoires indépendantes suivant toutes la même loi géométrique de paramètre p. Déterminer l'espérance et la variance de la moyenne empirique M_n de ces n variables aléatoires.

Propriété 10. — Inégalité de concentration

Si M_n est la moyenne empirique de X_1, X_2, ..., X_n alors, pour tout réel $\varepsilon > 0$,

$$\mathbf{P}(|M_n - \mu| \geqslant \varepsilon) \leqslant \frac{\sigma^2}{n\varepsilon^2}.$$

Exemple 11. On dispose d'une urne contenant des boules dont certaines sont rouges. On note p la proportion de boules rouges dans l'urne. On effectue n tirages successifs avec remise et on note f la fréquence de bouless rouges tirées.

Comment choisir n pour que f soit une valeur approchée de p à 10^{-2} près avec une probabilité supérieure ou égale à 0,95 ?

Corollaire 12 : Loi faible des grands nombres

Si M_n est la moyenne empirique de X_1, X_2, ..., X_n alors, pour tout réel $\varepsilon > 0$,

$$\lim_{n\to+\infty} \mathbf{P}(|M_n - \mu| \geqslant \varepsilon) = 0.$$

Remarque 13. Considérons une expérience aléatoire et intéressons-nous à un évènement particulier A lié à cette expérience. On répète successivement et de façon indépendante cette expérience et on note, à la i-ème expérience, X_i la variable aléatoire égale à 1 si A est réalisé et 0 sinon. Pour tout $k \in \mathbb{N}^*$, X_k suit une loi de Bernoulli de paramètre $\mathbf{P}(\text{A})$ et, de plus, les variables X_k sont mutuellement indépendantes. Ainsi, la variable aléatoire $S_n = X_1 + X_2 + \cdots + X_n$ qui compte le nombre de fois où l'évènement A s'est produit suit une loi binomiale $\mathscr{B}(n, \mathbf{P}(\text{A}))$ et la moyenne empirique $M_n = \frac{S_n}{n}$ représente la fréquence de réalisation de A sur les n premières expériences. La loi des grands nombres assure que, pour tout $\varepsilon > 0$, la probabilité que $\mathbf{P}(\text{A})$ n'appartiennent pas $]M_n - \varepsilon\,;M_n + \varepsilon[$ tend vers 0 lorsque n tend vers $+\infty$ ce qui traduit le fait que les réalisations de l'expérience pour lesquelles M_n s'éloigne de $\mathbf{P}(\text{A})$ sont rares et ce d'autant plus que la taille de l'échantillon est grand.

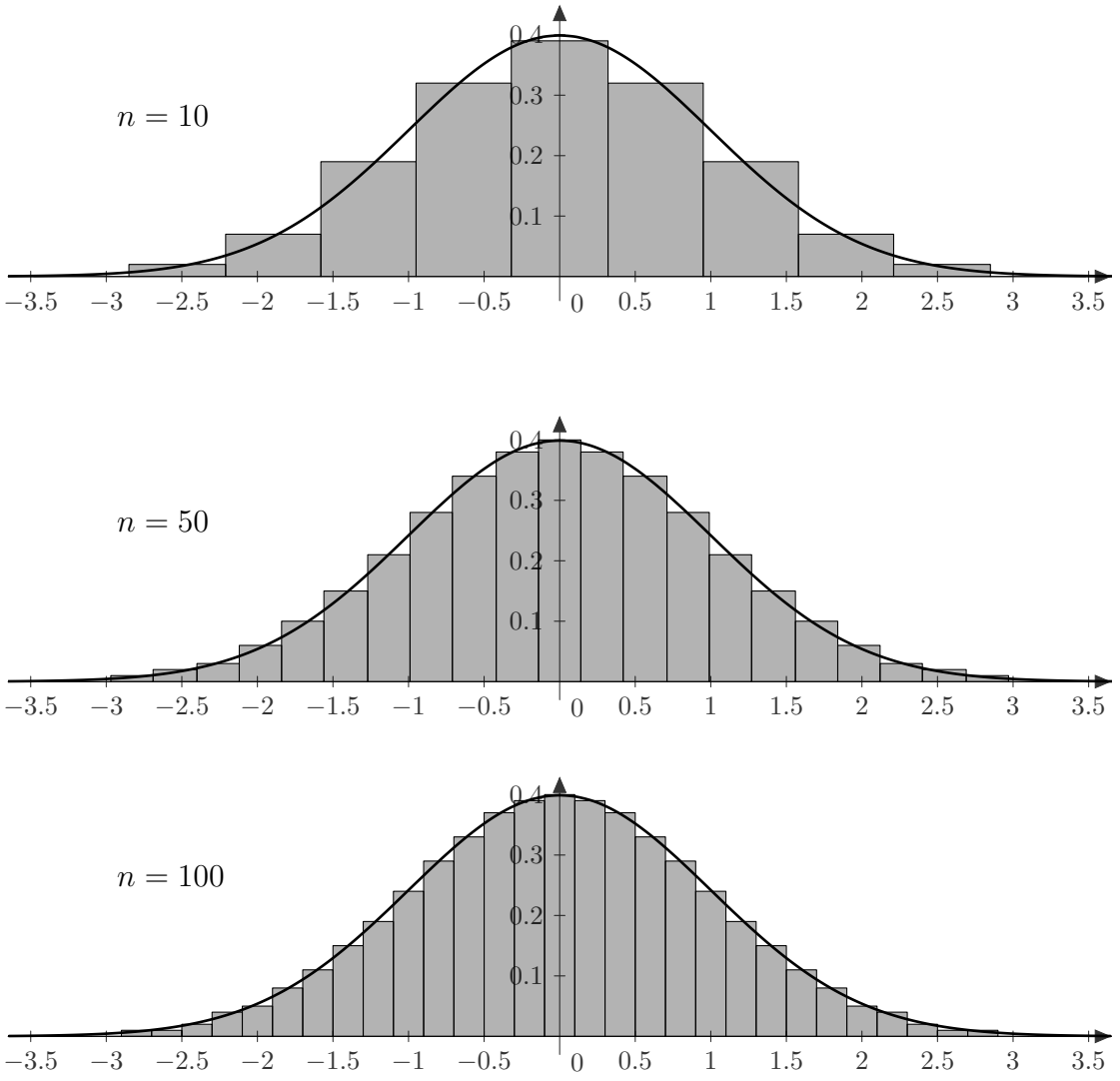
II. — Théorème central limite

1) Un cas particulier : le théorème de de Moivre-Laplace

On considère une variable aléatoire S_n qui suit une loi binomiale de paramètres n et $p \in]0\,;1[$. L'espérance de S_n est donc $\mu = np$ et l'écart-type de S_n est donc $\sigma = \sqrt{np(1-p)}$. On considère

la variable centrée réduite $S_n^* = \dfrac{X - \mu}{\sigma}$ et on observe le comportement de la loi de probabilité de S_n^* lorsque n augmente.

Dans les exemples ci-dessous, on a choisi $p = 0{,}5$.



On constate que l'histogramme représentant la loi de probabilité de S_n^* se rapproche de la courbe gaussienne centrée réduite. Ceci est la traduction graphique du théorème suivant.

Théorème 14. — Théorème de de Moivre-Laplace

Soit S_n une variable aléatoire suivant une loi binomiale de paramètres n et $p \in]0\,;1[$ et $S_n^* = \dfrac{S_n - np}{\sqrt{np(1-p)}}$. Alors, pour tout réel x,

$$\lim_{n\to+\infty} \mathbf{P}(S_n^* \leqslant x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t.$$

Autrement dit, en notant $F_{S_n^*}$ la fonction de répartition de S_n^* et F_Y la fonction de répartition d'une variable aléatoire Y suivant une loi $\mathcal{N}(0,1)$ alors, pour tout réel x,

$$\lim_{n\to+\infty} F_{S_n^*}(x) = F_Y(x).$$

Remarque 15. Ce théorème peut se comprendre ainsi : si S_n suit une loi binomiale de paramètres n et p alors, lorsque n devient grand, la loi de la variable $S_n^* = \dfrac{X_n - np}{\sqrt{np(1-p)}}$ peut être approchée par la loi $\mathcal{N}(0,1)$. On considère, en général, qu'on a une bonne approximation si

$$n \geqslant 30 \qquad np \geqslant 5 \qquad n(1-p) \geqslant 5.$$

Exemple 16. Supposons que $n = 192$ et $p = \dfrac{1}{4}$ et considérons une variable aléatoire S suivant une loi $\mathscr{B}\left(192, \dfrac{1}{4}\right)$. Alors, $np = 48 \geqslant 5$ et $n(1-p) = 144 \geqslant 5$. De plus, $\sqrt{np(1-p)} = 6$ donc $S^* = \dfrac{S-48}{\sqrt{36}} = \dfrac{S-48}{6}$. Dès lors,

$$\mathbf{P}(S \leqslant 60) = \mathbf{P}\left(\frac{S-48}{6} \leqslant \frac{60-48}{6}\right) = \mathbf{P}(S^* \leqslant 2) \approx \int_{-\infty}^{2} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{t^2}{2}} \,\mathrm{d}t \approx 0{,}977$$

et on peut vérifier que

$$\mathbf{P}(S \leqslant 60) = \sum_{k=0}^{60} \binom{192}{k} \left(\frac{1}{4}\right)^k \left(1 - \frac{1}{4}\right)^{192-k} \approx 0{,}979.$$

2) Cas général

Théorème 17. — Théorème central limite

Soit (X_n) une suite de variables aléatoires indépendantes, suivant toute la même loi et admettant une espérance μ et une variance σ^2. Pour tout $n \in \mathbb{N}$, on note $M_n = \dfrac{1}{n}\sum_{k=1}^{n} X_k$ la moyenne empirique de X_1, X_2, ..., X_n et $M_n^* = \dfrac{M_n - \mu}{\frac{\sigma}{\sqrt{n}}}$ la moyenne empirique centrée réduite.

Alors, pour n suffisamment grand, la loi de M_n^* est approximativement la loi $\mathcal{N}(0,1)$.

Exemple 18. On peut démontrer (voir l'exercice 15 du chapitre 8) que si X et Y sont deux variables aléatoires indépendantes suivant des lois de Poisson de paramètres λ et μ alors $X+Y$ suit une loi de Poisson de paramètre $\lambda + \mu$. On peut alors généraliser ce résultat par récurrence : si X_1, X_2, ..., X_n sont des variables aléatoires indépendantes suivant des lois de Poisson de paramètre λ_1, λ_2, ..., λ_n alors $S_n = X_1 + X_2 + \cdots + X_n$ suit une loi de Poisson de paramètre $\lambda_1 + \lambda_2 + \cdots + \lambda_n$.

Considérons alors $n = 100$ variables aléatoires X_1, X_2, ..., X_{100} suivant toute la même loi de Poisson de paramètre $\lambda = 1$. Alors, $S_{100} = X_1 + X_2 + \cdots + X_{100}$ suit une loi de Poisson de paramètre 100. Comme l'espérence et la variance d'une loi de Poisson de paramètre 1 sont égales à 1, la moyenne empirique centrée réduite de ces variables aléatoires est $M_{100}^* = \dfrac{\frac{S_{100}}{100} - 1}{\frac{1}{\sqrt{100}}} = \dfrac{S_{100} - 100}{10}$.

Par le théorème central limite,

$$\mathbf{P}(S_{100} \leqslant 80) = \mathbf{P}\left(\frac{S_{100}-80}{10} \leqslant \frac{80-100}{10}\right) = \mathbf{P}(M_{100}^* \leqslant -2) \approx \int_{-\infty}^{-2} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{t^2}{2}} \,\mathrm{d}t \approx 0{,}0226.$$

Or, comme S_{100} suit une loi de Poisson de paramètre 100,

$$\mathbf{P}(S_{100} \leqslant 80) = \sum_{k=0}^{80} \frac{100^k}{k!} \mathrm{e}^{-100} \approx 0{,}0227.$$

Remarque 19. Le théorème de de Moivre-Laplace est un cas particulier du théorème central limite en prenant des variables aléatoires indépendantes X_k suivant toutes le même loi de Bernoulli de paramètre p. En effet, dans ce cas, $S_n = X_1 + X_2 + \cdots + X_n$ suit une loi binomiale de paramètres n et p et

$$M_n^* = \frac{\frac{S_n}{n} - p}{\frac{\sqrt{p(1-p)}}{\sqrt{n}}} = \frac{n(\frac{S_n}{n} - p)}{n\left(\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)} = \frac{S_n - np}{\sqrt{n}\sqrt{p(1-p)}} = \frac{S_n - np}{\sqrt{np(1-p)}} = S_n^*$$

donc le théorème central limite assure que S_n^* suit approximativement une loi normale centrée réduite pour n suffisamment grand.

3) Application aux tests statistiques

Le principe général des tests que nous allons voir est le suivant. Si X_1, X_2, ..., X_n sont des variables aléatoires indépendantes ayant toutes la même loi qu'une variable X d'espérance μ et de variance σ^2 alors, d'après le théorème central limite, la moyenne empirique $M_n^* = \dfrac{M_n - \mu}{\frac{\sigma}{\sqrt{n}}}$ suit approximativement une loi normale centrée réduite. Or, les valeurs d'une telle variable sont essentiellement concentrées autour de 0 et il est rare qu'elles s'en éloignent.

Pour tester l'hypothèse H_0 : « $\mathbf{E}(X) = \mu$ » contre l'hypothèse H_1 : « $\mathbf{E}(X) \neq \mu$ », on procède alors de la manière suivante :

- on fixe un risque d'erreur α (ou, de façon équivalente, un niveau de confiance $1 - \alpha$) ;
- on détermine l'unique réel u_α tel que la probabilité qu'une variable aléatoire $X \hookrightarrow \mathcal{N}(0,1)$ n'appartienne pas à $[-u_\alpha\,;u_\alpha]$ est égale à α ;
- on calcule M_n^* et
  - si la valeur de M_n^* n'appartient pas à $[-u_\alpha\,;u_\alpha]$, on rejette H_0 avec un risque d'erreur α ;
  - sinon, on accepte l'hypothèse.

D'un point de vue pratique, les valeurs de u_α se déterminent à l'aide de tables. En voici quelques-unes :

niveau de confiance $1-\alpha$	80%	90%	95%	99%
risque d'erreur α	20%	10%	5%	1%
u_α	1,29	1,65	1,96	2,58

Exemple 20. On étudie un processus de fabrication de flacons. La contenance d'un flacon fabriqué est une variable aléatoire suivante une loi normale qui doit avoir une espérance $\mu = 100$ (en cL) et qui a un d'écart-type $\sigma = 1$. On souhaiterait savoir si la machine est bien réglée i.e. si effectivement μ est bien égale à 100. Pour cela, on prélève 50 flacons dans la production et on mesure leur contenance. La moyenne de ces 50 contenances est 100,5 en (cL).

Doit-on considérer, au seuil de confiance 99%, qu'il faut modifier les réglages de la machine ?

Exemple 21. Une maladie touche, dans la population mondiale, une personne sur 100. Dans une zone délimitée, on a étudié un échantillon de 500 personnes et on a décelé 7 personnes malades.

Doit-on considérer, au seuil de confiance 95%, que la proportion de malades dans cette zone est anormale ?

III. — Exercices

Exercice 1. On lance 3600 fois un dé équilibré. En utilisant l'inégalité de Bienaymé-Tchebychef, minorer la probabilité que le nombre d'apparitions du nombre 1 soit compris entre 480 et 720,

Exercice 2. Soit X une variable aléatoire telle que e^X admet une espérance.

1. Justifier que $\{X \geqslant t\} = \{\mathrm{e}^X \geqslant \mathrm{e}^t\}$.
2. En déduire, en utilisant l'inégalité de Markov, que $\mathbf{P}(X \geqslant t) \leqslant \mathbf{E}(e^{X-t})$.

Exercice 3. Un institut de sondage a été missionné pour estimer la proportion p de végétariens en France. Il interroge pour cela n français. Puisque le choix des sondés s'effectue sur une population très grande, on admet que l'expérience peut s'apparenter à une suite de n tirages indépendants avec remise. On note X_n la variable aléatoire égale au nombre de végétariens sondés et on souhaite quantifier à quel point la fréquence $F_n = \dfrac{X_n}{n}$ approche la proportion p.

1. Déterminer la loi de X_n.
2. En considérant la fonction $f : x \longmapsto x(1-x)$, montrer que $p(1-p) \leqslant \frac{1}{4}$.
3. Montrer que, pour tout $\varepsilon > 0$, $\mathbf{P}\left(|F_n - p| \geqslant \varepsilon\right) \leqslant \dfrac{1}{4n\varepsilon^2}$.
4. En déduire une condition sur n pour que F_n soit une approximation de p à 10^{-2} près avec une probabilité supérieure ou égale à 95%.

Exercice 4. On définit une variable aléatoire S en exécutant le programme Python suivant :

```
from random import random
S = 0
for i in range(200):
  X=random()
  S+=X
```

1. **a.** À chaque tour de boucle, quelle est la loi de X ?

   b. Rappeler son espérance et calculer sa variance.
2. Calculer l'espérance et la variance de la variable aléatoire S.
3. En utilisant l'inégalité de Bienaymé-Tchebychev, minorer la probabilité de l'évènement $\{S \in [90, 110]\}$.

Exercice 5. Soit $n \in \mathbb{N}^*$ et $p \in]0\,;1[$. On considère une variable aléatoire X suivant une loi binomiale $\mathscr{B}(n,p)$. On pose $Y = \frac{X}{n} - p$.

1. Montrer que $\mathbf{E}(|Y|)^2 \leqslant \mathbf{E}(Y^2)$.
2. En considérant $\mathbf{V}(Y)$, montrer $\mathbf{E}(Y^2) = \dfrac{p(1-p)}{n}$.
3. En appliquant l'inégalité de Markov, déduire des questions précédentes que, pour tout réel $\varepsilon > 0$,
$$\mathbf{P}\left(\left|\frac{X}{n} - p\right| \geqslant \varepsilon\right) \leqslant \frac{\sqrt{p(1-p)}}{\varepsilon\sqrt{n}}.$$

Exercice 6. Soit X une variable aléatoire suivant une loi de Poisson de paramètre $\lambda > 0$.

1. Montrer que $\mathbf{P}(X \geqslant 2\lambda) \leqslant \mathbf{P}((X - \lambda + 1)^2 \geqslant (\lambda+1)^2)$.
2. En déduire que $\mathbf{P}(X \geqslant 2\lambda) \leqslant \dfrac{1}{\lambda+1}$.

Exercice 7. Soit un entier $n \geqslant 2$. On effectue n lancers d'une pièce équilibrée et on définit, pour tout $k \in [\![1, n]\!]$, une variable aléatoire X_k égale à 1 si le k-ième lancer donne pile, et 0 sinon. On pose également

$$S_n = X_1 + \cdots + X_n \quad \text{et} \quad M_n = \frac{S_n}{n}.$$

1. Quelle est la loi suivie par X_k pour tout $k \in [\![1, n]\!]$? En déduire la loi de S_n.

2. Déterminer l'espérance et la variance de S_n et de M_n.

3. **a.** À l'aide de l'inégalité de Bienaymé-Tchebychev, minorer la probabilité que M_n se trouve entre 0,4 et 0,6

b. Au bout de combien de lancers la probabilité que M_n se trouve entre 0,4 et 0,6 est au moins égale à 95% ?

4. Dans la suite, on considère que S_n suit approximativement une loi $\mathcal{N}(\mu, \sigma^2)$.

a. Donner les valeurs des paramètres μ et σ^2.

b. On pose $S_n^* = \dfrac{S_n - \mu}{\sigma}$. Quelle est la loi de S_n^* ?

c. Vérifier que

$$\mathbf{P}(0,4 \leqslant M_n \leqslant 0,6) = \mathbf{P}\left(-0,2\sqrt{n} \leqslant S_n^* \leqslant 0,2\sqrt{n}\right).$$

d. Pour quelle valeur de n cette probabilité est-elle environ égale à 95% ?

e. Comparer avec la réponse à la question **3.b.**.

Exercice 8. On fixe un réel $a > 0$. Soit n variables aléatoires X_1, X_2, ..., X_n indépendantes qui suivent la loi uniforme sur $[0, a]$. On se propose d'estimer la valeur de a de deux manières différentes.

1. Soit $k \in [\![1, n]\!]$. Déterminer la densité, la fonction de répartition, l'espérance et la variance de X_k.

2. On pose $M_n = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$.

a. Calculer l'espérance et la variance de M_n.

b. En déduire une variable Y_n telle que $\mathbf{E}(Y_n) = a$ puis calculer $\mathbf{V}(Y_n)$.

c. Soit $\varepsilon > 0$. Démontrer que $\mathbf{P}(|Y_n - a| \geqslant \varepsilon) \xrightarrow[n \to +\infty]{} 0$.

3. On pose $U_n = \max(X_1, X_2, \ldots, X_n)$.

a. Déterminer la fonction de répartition de U_n et en déduire une densité de U_n.

b. Démontrer que U_n admet une espérance et une variance et les calculer.

c. En déduire une variable Z_n telle que $\mathbf{E}(Z_n) = a$ puis calculer $\mathbf{V}(Z_n)$.

d. Soit $\varepsilon > 0$. Démontrer que $\mathbf{P}(|Z_n - a| \geqslant \varepsilon) \xrightarrow[n \to +\infty]{} 0$.

4. Comparer les variances de Y_n et Z_n. Lequel de ces deux estimateurs devrait converger le plus vite vers a ?

Exercice 9. On dispose d'une pièce de monnaie censée être bien équilibrée. On a effectué avec cette pièce 100 séries de lancers en s'arrêtant, pour chaque série, à la première apparition de *face*. On a obtenu les résultats suivants :

rang d'apparition du premier *face*	1	2	3	4	5	6	7	8	9
nombre de séries	48	20	14	3	7	4	2	0	2

On note X_k la variable aléatoire égale au nombre de lancers nécessaires pour obtenir le premier *face* lors de la k-ième série.

1. Déterminer, pour tout $k \in [\![1, 100]\!]$, la loi de X_k et préciser son espérance et sa variance.

2. En utilisant les résultats du tableau précédent, peut-on considérer, au niveau de confiance 95%, que la pièce est équilibrée ?

Exercice 10. À Boston, en 1986, le docteur Benjamin Spock, militant contre la guerre du Vietnam, fut jugé pour incitation publique à la désertion. Le juge chargé de l'affaire était soupçonné de ne pas être équitable dans la sélection des jurés : parmi les 700 personnes qu'il avait désignées comme jurés lors de ses procès précédents, il y avait 15% de femmes alors que, sur l'ensemble de la ville, 29% des jurés éligibles étaient de femmes.

1. Au seuil de confiance 95%, peut-on considérer que le juge est impartial dans la désignation des jurés ?

2. Si ce juge avait moins d'expérience et qu'il n'avait désigné, avec les mêmes pourcentages, que 40 jurés lors de ses précédents procès, la conclusion serait-elle la même ?

Exercice 11. En utilisant l'inégalité de Bienaymé-Tchebychev, montrer que, pour tout réel $x > 0$,

$$\int_{-\infty}^{x} \mathrm{e}^{-\frac{t^2}{2}} \,\mathrm{d}t \geqslant \sqrt{2\pi}\left(1 - \frac{1}{2x^2}\right).$$

Exercice 12. On considère une variable aléatoire réelle X à valeurs strictement positives. On suppose, de plus, que X admet une espérance $m > 0$ et une variance V. On considère une suite de variables aléatoires mutuellement indépendantes $(X_n)_{n\in\mathbb{N}^*}$ ayant toutes la même loi que X et on pose, pour tout $n \in \mathbb{N}^*$, $S_n = \sum_{k=1}^{n} X_k$.

On se donne un réel $\ell > 0$. Le but de l'exercice est de démontrer que

$$\lim_{n\to+\infty} \mathbf{P}(S_n \leqslant \ell) = 0.$$

1. Soit $n \in \mathbb{N}^*$. Justifier que, pour tout $\varepsilon > 0$,

$$\mathbf{P}\left(\left|\frac{S_n}{n} - m\right| \geqslant \varepsilon\right) \leqslant \frac{V}{n\varepsilon^2}.$$

2. Soit $n \in \mathbb{N}^*$. Justifier que, pour tout $\varepsilon > 0$, $\{S_n \leqslant n(m-\varepsilon)\} \subset \left\{\left|\frac{S_n}{n} - m\right| \geqslant \varepsilon\right\}$.

3. Soit $n \in \mathbb{N}^*$. En choisissant convenablement ε, déduire des questions précédentes que,

$$\mathbf{P}\left(S_n \leqslant \frac{nm}{2}\right) \leqslant \frac{4V}{m^2 n}.$$

4. Justifier l'existence d'un entier N tel que, pour tout $n \geqslant N$, $\frac{nm}{2} \geqslant \ell$ puis conclure.