1. Le contexte

Considérons une variable aléatoire X de loi de Bernoulli $\mathscr{B}(p)$, où p est la proportion d'individus de la population ayant une propriété donnée, un échantillon $(X_1, X_2, ..., X_n)$ de taille n de X et la proportion (ou fréquence) d'échantillon $F_n = \dfrac{\sum_{i=1}^{n} X_i}{n}$, où $\sum_{i=1}^{n} X_i$ représente le nombre d'individus de l'échantillonnage ayant la propriété.

On sait que si $np \geqslant 10$ et $n(1-p) \geqslant 10$, alors $U = \dfrac{F_n - p}{\sqrt{\dfrac{p(1-p)}{n}}}$ suit approximativement la loi normale $\mathscr{N}(0;1)$. On détermine alors le réel u_α tel que $P(-u_\alpha < U < u_\alpha) = 1-\alpha$.
Pour $\alpha = 5\%$, on a $u_\alpha = 1.96$.
Remarque : Lorsque n est petit, on doit utiliser la loi exacte de nF_n, à savoir la loi Binomiale $\mathscr{B}(n,p)$.

Intervalle de fluctuation de la fréquence F

On suppose que l'on connait p.

On en déduit que $P\left(p - \sqrt{\dfrac{p(1-p)}{n}}u_\alpha \leqslant F_n \leqslant p + \sqrt{\dfrac{p(1-p)}{n}}u_\alpha\right) = 1-\alpha$, et donc $P(F_n \in IF_p) = 1-\alpha$,
avec

$$IF_p = \left[p - \sqrt{\frac{p(1-p)}{n}}u_\alpha \,;\, p + \sqrt{\frac{p(1-p)}{n}}u_\alpha\right]$$

intervalle de fluctuation IF_p de F_n au niveau $1-\alpha = 0.95$.

Pour tout $p \in [0,1]$, on a $0 \leqslant p(1-p) \leqslant \dfrac{1}{4}$ et donc on a l'inclusion d'événements $IF_p \subset IF'_p = \left[p - \dfrac{1}{\sqrt{n}} \,;\, p + \dfrac{1}{\sqrt{n}}\right]$ et donc $P\left(F \in IF'_p\right) = P(F \in IF_p) \geqslant 1-\alpha = 0.95$: $IF'_p = \left[p - \dfrac{1}{\sqrt{n}} \,;\, p + \dfrac{1}{\sqrt{n}}\right]$ est un intervalle de fluctuation de F de niveau (supérieur ou égal à) $1-\alpha = 0.95$.

Intervalle de confiance de la proportion p

On suppose que l'on ne connait pas p mais que l'on a une observation f_n de F_n à partir d'un échantillon.

On a $P\left(F_n - \sqrt{\dfrac{p(1-p)}{n}}u_\alpha \leqslant p \leqslant F_n + \sqrt{\dfrac{p(1-p)}{n}}u_\alpha\right) = 1-\alpha$, et donc $P(p \in IC_p) = 1-\alpha$, avec

$$IC_p = \left[F_n - \sqrt{\frac{p(1-p)}{n}}u_\alpha \,;\, F_n + \sqrt{\frac{p(1-p)}{n}}u_\alpha\right]$$

intervalle de confiance IC_p de p au niveau $1-\alpha = 0.95$.

Bien remarquer que p est fixé et que ce sont les bornes de l'intervalle, et donc l'intervalle, qui sont aléatoires ; chaque échantillon donne a priori un intervalle différent. Cela s'interprète donc en disant que 95% des échantillons fournissent un intervalle contenant p.

Pour tout $p \in [0,1]$, on a $0 \leqslant p(1-p) \leqslant \dfrac{1}{4}$, $0 \leqslant \sqrt{p(1-p)} \leqslant \dfrac{1}{2}$, $\sqrt{\dfrac{p(1-p)}{n}}u_\alpha \leqslant \dfrac{1.96}{2}\dfrac{1}{\sqrt{n}} \leqslant \dfrac{1}{\sqrt{n}}$ et donc on a l'inclusion d'événements $IC_p \subset IC'_p = \left[F_n - \dfrac{1}{\sqrt{n}} \,;\, F_n + \dfrac{1}{\sqrt{n}}\right]$ et donc $P\left(p \in IC'_p\right) = P(p \in IC_p) \geqslant 1-\alpha = 0.95$: $IC'_p = \left[F_n - \dfrac{1}{\sqrt{n}} \,;\, F_n + \dfrac{1}{\sqrt{n}}\right]$ est un intervalle de confiance de la proportion p de niveau (supérieur ou égal à) $1-\alpha = 0.95$.

Pour une observation f de F, on obtient l'intervalle $ic'_p = \left[f_n - \dfrac{1}{\sqrt{n}} \,;\, f_n + \dfrac{1}{\sqrt{n}}\right]$.

2. Énoncés

✯ EXERCICE 1

Une boîte contient 10 boules. Sur chacune d'elles on a inscrit un nombre suivant le tableau ci-contre :

Nombre inscrit	5	6	10	11	12	13	14
Nombre de boules	1	2	1	3	1	1	1

Un joueur mise 10 euros, tire une boule au hasard et reçoit la somme (en euros) inscrite sur la boule. Toutes les boules ont la même probabilité d'être tirées.

1. Le joueur joue une fois. On appelle p_1 la probabilité qu'il perde de l'argent (c'est-à-dire qu'il reçoive moins de 10 euros à l'issue du tirage) et p_2 la probabilité qu'il reçoive plus de 10 euros. Donner p_1 et p_2.
2. Soit X la variable aléatoire qui à chaque tirage fait correspondre le « gain » du joueur (une perte est un « gain » négatif). Par exemple, s'il tire le nombre 12, son « gain » est +2 ; s'il tire le 6, son « gain » est −4.
   a. À l'aide d'un tableur, proposer une simulation de 500 parties de ce jeu (une partie correspondant à un tirage) ; calculer les fréquences de chaque valeur du « gain », puis la moyenne du gain pour les 500 parties. Qu'observe-t-on ?
   b. Quelles sont les valeurs prises par la variable aléatoire X ? Donner la loi de probabilité de X et recopiant et complétant le tableau suivant :

Valeurs de X : x_i	-5	-4	0	1	2	3	4
$p_i = P(X = x_i)$							

   c. Calculer l'espérance mathématique $E(X)$. Que représente $E(X)$ pour le joueur ?
   d. Calculer la variance et la valeur approchée à 10^{-2} près de l'écart-type de X.
3. Il s'agit maintenant, en changeant le nombre sur une boule, de rendre ce jeu équitable, c'est-à-dire de rendre l'espérance mathématique nulle. Proposer une solution.

✯ EXERCICE 2

Un grossiste en fourniture de bureau revend des rouleaux de ruban adhésif transparent et affirme que seulement 0,8 % des rouleaux présente un défaut de jaunissement du papier. Un client achète 500 rouleaux et constate que 6 rouleaux, soit 1,2% des rouleaux, jaunissent le papier. Le client peut-il faire une réclamation auprès du grossiste ?

✯ EXERCICE 3

Sur 200 plantes examinées, on en compte 134 d'un phénotype « A » et 66 d'un phénotype « a ». Peut-on admettre la loi de Mendel, prévoyant les proportions respectives 3/4 et 1/4 des deux phénotypes ?

✯ EXERCICE 4

Avril 2002

Voici un extrait d'article, publié dans le journal « Le Monde » par le statisticien Michel Lejeune, après le premier tour de l'élection présidentielle de 2002. « Pour les rares scientifiques qui savent comment sont produites les estimations, il était clair que l'écart des intentions de vote entre les candidats Le Pen et Jospin rendait tout à fait plausible le scénario qui s'est réalisé. En effet, certains des derniers sondages indiquaient 18 % pour Jospin et 14 % pour Le Pen. Si l'on se réfère à un sondage qui serait effectué dans des conditions idéales [...], on obtient sur de tels pourcentages une incertitude de plus ou moins 3 % étant donné la taille de l'échantillon [...]. »

1. Si l'on tient compte de l'incertitude liée au sondage, entre quels pourcentages pourraient se situer réellement (à 95 % de confiance) les deux candidats si le sondage donne 18 % pour l'un et 14 % pour l'autre ?
2. Représenter sur un même graphique les deux « fourchettes » calculées à la question précédente. Peut-on prévoir l'ordre des candidats ?
3. Au premier tour de l'élection présidentielle de 2002, L. Jospin a obtenu 16,18 % des voix et J.-M. Le Pen 16,86 %. Expliquer la phrase « l'écart des intentions de vote entre les candidats Le Pen et Jospin rendait tout à fait plausible le scénario qui s'est réalisé ».

3. Éléments de correction

✯ Correction de l'exercice 1

1. $p_1 = \dfrac{3}{10}$ et $p_2 = \dfrac{6}{10} = \dfrac{3}{5}$

2. a. Simulation de 500 parties : voir le fichier joint (**Fichier 1**) (Simulation faite pour 500, 1000 et 5000 : Observations liées au point suivant dans les commentaires des programmes « À l'aide de simulations et d'une approche heuristique de la loi des grands nombres, on fait le lien avec la moyenne et la variance d'une série de données ».

   b. Valeurs prises par la variable aléatoire X : $-5, -4, 0, 1, 2, 3$ et 4.

   c.

Valeurs de X : x_i	-5	-4	0	1	2	3	4
$p_i = P(X = x_i)$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{3}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{10}$

   d. $E(X) = \cdots = -\dfrac{1}{10}$: cette espérance représente **le gain moyen obtenu sur un grand nombre de parties** (conformément au libellé du programme).

   e. En utilisant $V(X) = E(X^2) - (E(X))^2$, on obtient $V \approx 8,89$, puis $\sigma(X) = \sqrt{V}$ soit $\sigma(X) \approx 2,98$.

3. Il suffit par exemple de remplacer la boule portant le n°5 par une boule portant le n°4.

✯ Correction de l'exercice 2

« corrigé » avec les données telles que … mais on peut remarquer qu'on n'est pas dans les bonnes conditions pour utiliser les intervalles de fluctuation et de confiance $\left[p - \dfrac{1}{\sqrt{n}}; p - \dfrac{1}{\sqrt{n}}\right]$,… ; d'ailleurs on a une borne de l'intervalle qui est négative.

Le garde-t-on ainsi ? ? ? Les réponses ci-dessous ne seront pas satisfaisantes, mais ça peut aussi montrer l'intérêt de garder un esprit critique face aux exercices rencontrés dans les ouvrages de seconde.

Avec l'intervalle de fluctuation :

$I_f = \left[0,008 - \dfrac{1}{\sqrt{500}}; 0,008 + \dfrac{1}{\sqrt{500}}\right] \approx [-0,037; 0,053]$.

Cet intervalle contient 1,2 % donc, à priori, le client ne peut pas se plaindre : le lot acheté se comporte comme 95 % des échantillons de 500 rouleaux.

Si on choisit le point de vue de l'intervalle de confiance , le fabricant ne sera pas en défaut non plus puisque

$\left[0,012 - \dfrac{1}{\sqrt{500}}; 0,012 + \dfrac{1}{\sqrt{500}}\right] \approx [-0,033; 0,057]$.

. Or on est sûr à 95 % que la moyenne théorique est dans cet intervalle, donc la moyenne proposée par le fabricant ne peut pas être rejetée.

✯ Correction de l'exercice 3

200 plantes examinées, parmi lesquelles 134 d'un phénotype « A » et 66 d'un phénotype « a ».

$\dfrac{134}{200} = 0,67$ donc l'intervalle de confiance correspondant à cet échantillon est $\left[0,67 - \dfrac{1}{\sqrt{200}}; 0,67 + \dfrac{1}{\sqrt{200}}\right] \approx [0,599; 0,741]$.

Il ne contient pas 0,75.

On peut interpréter ce résultat comme signifiant que cet échantillon fait partie des 5 % ne contenant pas la moyenne théorique, la loi de Mendel étant considérée comme fiable….

Si on raisonne à partir de l'intervalle de fluctuation en partant des proportions théoriques de la loi de Mendel : cet intervalle est $\left[0,75 - \dfrac{1}{\sqrt{200}}; 0,75 + \dfrac{1}{\sqrt{200}}\right] \approx [0,679; 0,921]$ et donc ne contient pas 0,67.

On peut interpréter le résultat en disant que l'hypothèse est rejetée avec une marge d'erreur de 5 %, c'est-à-dire d'être tombé sur un échantillon qui ne convient pas. Si on avait eu une fréquence de 0,685, l' hypothèse $p = \dfrac{3}{4}$ n'aurait pas été rejetée…

✯ Correction de l'exercice 4

1. Pour celui qui a 18% des intentions de vote, son résultat se situera entre 15% et 21% (au seuil de confiance de 95%).
   Pour celui qui a 14% des intentions de vote, son résultat se situera entre 11% et 17% (au seuil de confiance de 95%).

2. Schéma…. Compte-tenu de l'intersection non vide des deux intervalles, on ne peut pas vraiment prévoir l'ordre des candidats.

3. L. Jospin a obtenu en réalité 16,18% des voix et J.M. Le Pen 16,86% ce qui est conforme aux intervalles donnés ci-dessus et explique la phrase « l'écart des intentions de vote entre les candidats Le Pen et Jospin rendait tout à fait plausible le scénario qui s'est réalisé ».