Chapitre 9

Échantillonage

Sommaire

9.1 Activité

Reprenons le contexte suivant :
On lance un dé équilibré. On gagne si on fait un 6. On note S, comme « succès », cette issue et $\overline{S}$ l'autre issue.
On répète sept fois l'expérience de façon indépendante et on note X la variable aléatoire qui, à chaque série de lancers, associe le nombre de succès obtenus au cours de ces sept répétitions.
Alors X suit la loi $\mathscr{B}\left(7\,;\frac{1}{6}\right)$.
La calculatrice permet d'obetnir le tableau suivant (les valeurs sont arrondies au millième). Compléter la dernière ligne.

k	0	1	2	3	4	5	6	7
$p(X=k)$	0,279	0,391	0,234	0,078	0,016	0,002	0	0
$p(X \leqslant k)$								

On cherche à partager l'intervalle $[0\,;7]$, où X prend ses valeurs, en trois intervalles :

- Un intervalle $[a\,;b]$ (où a et b sont des entiers) tel que $p(a \leqslant Z \leqslant b) \geqslant 0,95$;
- Et pour qu'il soit centré au maximum, on s'impose que $p(Z < a) < 0,025$ et $p(Z > b) < 0,025$.
  Déterminer à l'aide du tableau précédant les valeurs de a et de b.
- Sur un dé inconnu, pour tester l'hypothèse « le 6 a une probabilité de sortir égale à $\frac{1}{6}$ » au seuil de 95 %, connaissant l'intervalle $[a\,;b]$ correspondant à cette probabilité, on applique la règle suivante : on répète avec ce dé 7 fois l'expérience et on note la fréquence f d'apparition du 6.
  Si f n'est pas dans l'intervalle $\left[\frac{a}{7}\,;\frac{b}{7}\right]$, alors on rejette l'hypothèse.
  Dans quels cas rejettera-t-on alors cette hypothèse ?

9.2 Bilan et compléments

9.2.1 Définition de l'intervalle de fluctuation au seuil de 95 %

Définition 9.1. L'intervalle de fluctuation au seuil de 95 %, relatif aux échantillons de taille n, est l'intervalle centré autour de p, proportion du caractère dans la population, où se situe, avec une probabilité égale à 0,95, la fréquence observée dans un échantillon de taille n.

Dans la pratique, il est impossible de trouver un intervalle qui contient exactement 95 % des fréquences, aussi cherche-t-on un intervalle centré sur p qui contient au moins 95 % des fréquences.

9.2.2 Approximation de l'intervalle de fluctuation vue en Seconde

La propriété suivante a été énoncée en Seconde :

Propriété. *Dans le cas où $n > 25$ et $0,2 < p < 0,8$, l'intervalle I suivant est centré sur p et il contient l'intervalle de fluctuation, c'est-à-dire que la probabilité qu'il contienne la fréquence observée est* au moins *égale à 95 % :*

$$I = \left[p - \frac{1}{\sqrt{n}}\,;\, p + \frac{1}{\sqrt{n}}\right]$$

Remarque. Les conditions d'application de cette propriété sont un peu arbitraires. Si telle qu'elle est énoncée ($n > 25$ et $0,2 < p < 0,8$) elle est rigoureusement conforme au texte du programme de Seconde, en Terminale le programme indique que les conditions d'application d'une propriété similaire sont les suivantes : elle s'applique à condition que $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$, aussi les professeurs de mathématiques de Seconde choisissent-ils parfois d'indiquer qu'elle s'applique dans les conditions « de Terminale »[1].

9.2.3 Intervalle de fluctuation avec la loi binomiale

La loi binomiale nous permet de calculer très exactement les probabilités des différentes fréquences observables dans un échantillon de taille n, à savoir les valeurs $\frac{k}{n}$, avec $0 \leqslant k \leqslant n$, quels que soient n et p. La règle est alors la suivante :

Propriété 9.1. *L'intervalle de fluctuation au seuil de 95 % associé à une variable aléatoire X suivant la loi binomiale $\mathscr{B}(n\,;\,p)$, est l'intervalle $\left[\frac{a}{n}\,;\,\frac{b}{n}\right]$, où a et b sont les deux entiers naturels définis par :*

- *a est le plus petit des entiers k vérifiant $p(X \leqslant k) > 0,025$;*
- *b est le plus petit des entiers k vérifiant $p(X \leqslant k) \geqslant 0,975$.*

Remarques.

- Lorsque n est assez grand, il est quasiment centré sur p.
- Cet intervalle s'obtient grâce aux possibilités des calculatrices (ou des logiciels) par la lecture des probabilités cumulées croissantes.
- Cet intervalle est le plus petit intervalle centré contenant au moins 95 % des fréquences ; en particulier, si l'approximation de cet intervalle vu en Seconde en est assez proche lorsque $n > 25$ et $0,2 < p < 0,8$, celui-ci est de la plus petite amplitude possible (celui de Seconde contient parfois largement plus de 95 % des fréquences) : on ne peut faire mieux.

1. Elles ont l'avantage de pouvoir utiliser l'intervalle de fluctuation pour l'étude de la sortie d'un dé équilibré à 6 faces ($p = \frac{1}{6}$) pour peu que le nombre de lancers soit suffisant ($n \geqslant 30$)

9.2.4 Prise de décision avec la loi binomiale

On considère une population dans laquelle *on suppose* que la proportion d'un certain caractère est p, ou bien une expérience aléatoire dont *on suppose* que la probabilité d'un évènement particulier est p.
Pour juger de cette *hypothèse*, on prélève dans la population, au hasard et avec remise [2], un échantillon de taille n sur lequel on observe une fréquence f du caractère ou bien on repète l'expérience aléatoire de manière identique et indépendante n fois et on observe la fréquence f d'apparition de l'évènement particulier.
On rejette l'hypothèse selon laquelle la proportion dans la population est p ou selon laquelle la probabilité de l'évènement particulier est p, lorsque la fréquence f observée est trop éloignée de p, dans un sens ou dans l'autre.
On choisit de fixer le seuil à 95 % de sorte que la probabilité de rejeter l'hypothèse, alors qu'elle est vraie, soit inférieure à 5 %.
La règle de décision adoptée est alors la suivante :

Si la fréquence observée f n'appartient à l'intervalle de fluctuation au seuil de 95 % $\left[\frac{a}{n};\frac{b}{n}\right]$, on considère que l'hypothèse selon laquelle la proportion est p dans la population ou selon laquelle la probabilité de l'évènement particulier est p est rejetée.

9.3 Exercices

Exercice 9.1.
Cet exercice nécessite de disposer d'une calculatrice TI ou d'une calculatrice CASIO récente.
On dispose d'une partie de programme :

TI	Casio
: PROMPT N	"N"? → N ↩
: PROMPT P	"P"? → N ↩
: 0 → I	0 → I ↩
: While binomFRép(N,P,I) ⩽ 0,025	While BinomCD(I,N,P) ⩽ 0,025 ↩
: I+1 → I	I+1 → I ↩
: End	WhileEnd ↩
: I → A	I → A ↩

Remarque. binomFRép(n,p,k) ou BinomCD(I,N,P) calculent $p(X \leqslant k)$, où X est une variable aléatoire suivant la loi $\mathscr{B}(n\,;\,p)$.

1. (a) À quoi correspondent N et P demandés en début de programme ?
   (b) À quoi correspond A à la fin du programme ?
2. Comment modifier ce programme pour qu'il obtienne A **et** B à la fin du programme ?
3. Comment modifier ce programme pour qu'il calcule les deux bornes de l'intervalle de fluctuation au seuil de 95 % de la loi binomiale de paramètres n et p ?
4. On a exécuté ce programme avec $p = 0,4$ et on a obtenu les résultats suivants pour la borne inférieure :

n	20	50	200	1 000	5 000
Borne	0,2	0,26	0,335	0,37	0,3864

   Comparer ce résultat avec la borne inférieure de l'intervalle de fluctuation introduit en Seconde. Qu'observe-t-on ?

2. Lorsque la taille de l'échantillon est petit par rapport à la taille de la population, on peut considérer qu'un tirage sans remise est quasiment équivalent à un tirage avec remise.

EXERCICE 9.2.
Monsieur Z, chef du gouvernement d'un pays lointain, affirme que 52 % des électeurs lui font confiance. On interroge 100 électeurs au hasard (la population est suffisamment grande pour considérer qu'il s'agit de tirages avec remise) et on souhaite savoir à partir de quelles fréquences, au seuil de 95 %, on peut mettre en doute le pourcentage annoncé par Monsieur Z, dans un sens, ou dans l'autre.

1. On fait l'hypothèse que Monsieur Z dit vrai et que la proportion des électeurs qui lui font confiance dans la population est 0,52. Montrer que la variable aléatoire X, correspondant au nombre d'électeurs lui faisant confiance dans un échantillon de 100 électeurs, suit la loi binomiale de paramètres $n = 100$ et $p = 0,52$.
2. On donne ci-dessous un extrait de la table des probabilités cumulées $p(X \leqslant k)$ où X suit la loi binomiale de paramètres $n = 100$ et $p = 0,52$.

k	40	41	42	43	...	60	61	...
$p(X \leqslant k)$	0,0106	0,0177	0,0286	0,0444	...	0,9561	0,9719	...

   (a) Déterminer a et b tels que :
   - a est le plus petit entier tel que $p(X \leqslant a) > 0,025$;
   - b est le plus petit entier tel que $p(X \leqslant b) \geqslant 0,975$.

   (b) Comparer l'intervalle de fluctuation au seuil de 95 %, $\left[\frac{a}{n} ; \frac{b}{n}\right]$, ainsi obtenu grâce à la loi binomiale, avec l'intervalle de Seconde.
3. Énoncer la règle de décision permettant de rejetter ou non l'hypothèse que la proportion des électeurs qui font confiance à Monsieur Z dans la population est 0,52, selon la valeur de la fréquence f des électeurs favorables à Monsieur Z obtenue sur l'échantillon.
4. Sur les 100 électeurs interrogés au hasard, 43 déclarent avoir confiance en Monsieur Z. Peut-on considérer, au seuil de 95 %, l'affirmation de Monsieur Z comme exacte ?

EXERCICE 9.3.
Un médecin de santé publique veut savoir si, dans sa région, le pourcentage d'habitants atteints d'hypertension artérielle est égal à la valeur de 16 % récemment publiée pour des populations semblables. En notant p la proportion d'hypertendus dans la population de sa région, le médecin formule l'hypothèse $p = 0,16$. Pour vérifier cette hypothèse, le médecin constitue un échantillon de $n = 100$ habitants de la région; il détermine la fréquence f d'hypertendus (l'échantillon est prélevé au hasard et la population est suffisamment importante pour considérer qu'il s'agit de tirages avec remise).
Pour quelles valeurs de f, la médecin rejettera-t-il cette hypothèse ?

EXERCICE 9.4.
En novembre 1976 dans un comté du sud du Texas, Rodrigo Partida est condamné à huit ans de prison. Il attaque ce jugement au motif que la désignation des jurés de ce comté est, selon lui, discriminante à l'égard des Américains d'origine mexicaine. Alors que 80 % de la population du comté est d'origine mexicaine, sur les 870 personnes convoquées pour être jurés lors des années précédentes, il n'y a eu que 339 personnes d'origine mexicaine.
Devant la Cour Suprême, un expert statisticien produit des arguments pour convaincre du bien fondé de le requête de l'accusé. En vous situant dans le rôle de cet expert, pouvez-vous décider si les Américains d'origine mexicaine sont sous-représentés dans les jurys de ce comté ?

EXERCICE 9.5.
Un groupe de citoyens demande à la municipalité d'une ville la modification d'un carrefour en affirmant que 40 % des automobilistes tournent un utilisant une mauvaise file.
Un officier de police constate que sur 500 voitures prises au hasard, 190 prennent une mauvaise file.

1. Déterminer, en utilisant la loi binomiale sous l'hypothèse $p = 0,4$, l'intervalle de fluctuation au seuil de 95 %.
2. D'après l'échantillon, peut-on considérer, au seuil de 95 %, comme exacte l'affirmation du groupe de citoyens ?

EXERCICE 9.6.
Dans le monde, la proportion de gauchers est 12 %. Soit n le nombre d'élèves dans votre classe.

1. Déterminer, à l'aide de la loi binomiale, l'intervalle de fluctuation au seuil de 95 % de la fréquence des gauchers sur un échantillon aléatoire de taille n.
2. Votre classe est-elle « représentative » de la proportion de gauchers dans le monde ?

EXERCICE 9.7.
Deux entreprises recrutent leur personnel dans un vivier comportant autant d'hommes que de femmes. Voici la répartition entre hommes et femmes dans ces deux entreprises :

	Hommes	Femmes	Total
Entreprise A	57	43	100
Entreprise B	1350	1150	2500

Peut-on suspecter l'une des deux de ne pas respecter la parité hommes-femmes à l'embauche ?