Chapitre 14

Échantillonage

Sommaire

14.1 Activité

Reprenons le contexte de l'activité d'introduction du chapitre 12.

Épreuve de Bernoulli : On lance un dé à 6 faces équilibré et on appelle succès « obtenir un 6 ».

Schéma de Bernoulli : On répète l'épreuve précédente de manière identique et indépendante n fois.

Variable aléatoire : On appelle X la variable aléatoire correspondant au nombre de succès.

Alors X suit la loi $\mathscr{B}\left(n ; \frac{1}{6}\right)$.
On suppose pour cette activité que $n = 30$. La calculatrice permet d'obtenir les probabilités ci-contre (arrondies au millième).
On cherche à partager l'intervalle $[0 ; n]$, où X prend ses valeurs, en trois intervalles :

- Un intervalle $[a ; b]$ (où a et b sont des entiers) tel que $p(a \leqslant X \leqslant b) \geqslant 0{,}95$;
- Et pour qu'il soit centré au maximum, on s'impose que $p(X < a) < 0{,}025$ et $p(X > b) < 0{,}025$.

k	$p(X = k)$	$p(X \leqslant k)$
0	0,004	
1	0,025	
2	0,073	
3	0,137	
4	0,185	
5	0,192	
6	0,160	
7	0,110	
8	0,063	
9	0,031	
10	0,013	
11	0,005	
12	0,001	
13	0,000	
...	...	
30	0	

1. Compléter la dernière colonne du tableau.

2. Déterminer à l'aide du tableau les valeurs de a et de b.

Pour tout échantillon de taille $n = 30$, c'est-à-dire pour toute série de 30 lancers, dans au moins 95 % des séries, le nombre de succès sera compris dans l'intervalle $[a\,;\,b]$ sur un grand nombre de telles séries.
Dit autrement : la fréquence d'apparition du succès sera, dans au moins 95 % des séries, dans l'intervalle $\left[\frac{a}{30}\,;\,\frac{b}{30}\right]$. Cet intervalle est appelé intervalle de fluctuation au seuil de 95 %.
Les statisticiens, physiciens [1], économistes, médecins, etc. utilisent cet intervalle de la manière arbitraire suivante : tant que la fréquence d'apparition du succès est dans l'intervalle de fluctuation, on ne peut en tirer aucune conclusion ; dès lors que la fréquence d'apparition n'est plus dans cet intervalle on peut considérer que la fréquence obtenue n'est pas due au hasard.
Par exemple, si un dé est inconnu et qu'on ne sait pas s'il est équilibré, 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 30 fois l'expérience et on note la fréquence f d'apparition du 6.
Si f n'est pas dans l'intervalle $\left[\frac{a}{30}\,;\,\frac{b}{30}\right]$, alors on rejette l'hypothèse.
Dans quels cas rejettera-t-on alors cette hypothèse ?

14.2 Bilan et compléments

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

Définition 14.1. 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 contienne exactement 95 % des fréquences, aussi cherche-t-on un intervalle centré sur p qui contient au moins 95 % des fréquences.

14.2.2 Approximation de l'intervalle de fluctuation vue en Seconde

La propriété suivante a été énoncée en Seconde :

Propriété. *Dans le cas où $n \geqslant 30$, $n \times p \geqslant 5$ et $n \times (1-p) \geqslant 5$, 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. On trouve parfois les conditions suivantes : $n > 25$ et $0{,}2 < p < 0{,}8$. Dans tous les cas l'idée est que n doit être suffisamment grand et que p ne doit être ni trop petite ni trop grande.

1. Les physiciens, dans certaines circonstances, utilisent un intervalle plus grand contenant 99,7 % des données, voire plus.

14.2.3 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 14.1. *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.

14.2.4 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, c'est-à-dire le risque de se tromper, 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.

14.3 Exercices

Exercice 14.1.
Cet exercice nécessite de disposer d'une calculatrice TI ou d'une calculatrice CASIO récente.
On dispose d'une partie de programme :

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.

TI	Casio
: PROMPT N	"N"? → N ↩
: PROMPT P	"P"? → P ↩
: 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 tels que définis à la propriété 14.1 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 ci-contre 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?

EXERCICE 14.2.
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	…	59	60	61	62	…
$p(X \leqslant k)$	0,0106	0,0177	0,0286	0,0444	…	0,9338	0,9561	0,9719	0,9827	…

   (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 14.3.
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 14.4.
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 14.5.
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 14.6.
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 14.7.
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 ?