
QUELQUES INTERROGATIONS DU PROFESSEUR DE LYCEE AUTOUR DES INTERVALLES DE FLUCTUATION

Véronique CERCLÉ
Irem de Montpellier

Résumé : *Un nouvel objet a fait son apparition dans les programmes de lycée : l'intervalle de fluctuation. Comment le définir ? Comment l'obtenir ? Comment l'utiliser ? On s'interrogera au passage sur la notion d'échantillon, de proportion/probabilité, et on analysera les présentations qui en sont données en Seconde, Première et Terminale pour expliquer, comparer et articuler les trois définitions proposées.*

Un nouvel objet a fait son apparition dans les programmes de lycée : l'intervalle de fluctuation. Comment un professeur en lycée comme moi, qui ai peu de connaissances universitaires en probabilités ou en statistiques, peut-il s'approprier cette notion pour pouvoir l'enseigner ? Quelles représentations peut-on forger sans le formalisme fourni par la théorie ?

Je propose ici de partager quelques-unes de mes interrogations, et quelques-unes des réponses que j'ai pu y apporter au fil de mes notes, en particulier :

- Les programmes donnent trois présentations successives : en quoi sont-elles différentes ?
- Quelle définition donner en Seconde ?
- Quel sens donner à cette notion d'intervalle de fluctuation ?

- Quel lien avec les simulations ? Qu'est-ce qu'un échantillon ?

Cet article montre les difficultés et les questions auxquelles j'ai été confrontée. Il s'attache à comprendre l'articulation de la Seconde à la Terminale des différentes présentations des intervalles de fluctuation (partie 1) et de leurs utilisations (partie 4). Il attire aussi l'attention des enseignants sur quelques précautions à prendre lors de cet enseignement, en particulier autour de la définition à donner en Seconde (partie 2) et des notions d'échantillon, de proportion et de probabilité (partie 3 et annexe).

La lecture de cet article peut compléter l'article intitulé « La prise de décision de la Seconde à la Première » publié dans le numéro 85 de la revue *Repères Irem* d'octobre 2011.

De quoi s’agit-il ?

On observe qu’en lançant une pièce 100 fois, on obtient des fréquences de pile comprises *essentiellement* entre 0,4 et 0,6. Inversement, le fait d’obtenir une fréquence en dehors de cet intervalle est *plutôt rare*. Il s’agit de donner corps à cette observation.

Au lycée, on va donc s’intéresser à la *fluctuation des fréquences* obtenues lors de la répétition d’une expérience aléatoire. Cette fluctuation est formalisée par la notion d’intervalle de fluctuation des fréquences¹, dont la définition trouvée dans wikipédia [W] m’a paru très claire :

« C’est un *intervalle dans lequel on s’attend à trouver la fréquence, avec une forte probabilité*. Il permet de détecter un écart important entre la fréquence observée sur un échantillon et la fréquence théorique (proportion ou probabilité). Le fait d’obtenir une valeur en dehors de cet intervalle s’interprète alors en mettant en cause la représentativité de l’échantillon ou la fréquence théorique. À l’inverse, le fait que la valeur observée soit comprise dans l’intervalle n’est pas une garantie de la validité de l’échantillon ou du modèle. »

Les programmes du lycée privilégient le seuil² de 95%, la « forte probabilité » avec laquelle

la fréquence observée est censée se trouver dans l’intervalle considéré est fixée arbitrairement à 95%. [P.O.2, p 8, ci-dessus]

On cherche ainsi un intervalle IF qui a la propriété de contenir la fréquence F dans un échantillon avec une probabilité d’au moins 95% :
 $P(F \in IF) \geq 0,95$.

Partie 1

Les programmes du lycée introduisent trois modes de détermination. Ont-ils un lien entre eux ?

Une lecture transversale des programmes de la Seconde à la Terminale montre trois présentations différentes :

Ces présentations successives amènent à s’interroger sur la diversité de ces définitions, et sur les relations possibles entre ces définitions.

1. Plusieurs intervalles de fluctuation sont possibles

On comprend assez rapidement que plusieurs choix sont possibles, et qu’il ne faut donc pas parler de l’intervalle de fluctuation à 95% mais de un intervalle de fluctuation à 95% (cadre ci-dessous). On peut déjà remarquer une première différence :

En Seconde	En Première	En Terminale
$IF2 = [p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}}]$	$IF1 = [\frac{a}{n} ; \frac{b}{n}]$ déterminé à l’aide de la loi binomiale (tableur ou algorithme) de sorte que $P(X < a) \leq 0,025$ et $P(X > b) \leq 0,025$	$IFT = [p - 1,96 \sqrt{\frac{p(1-p)}{n}} ; p + 1,96 \sqrt{\frac{p(1-p)}{n}}]$

¹ On peut aussi s’intéresser à la fluctuation de la moyenne observée sur un échantillon par rapport à la moyenne dans la population.

² Doit-on dire au seuil de 95% (on considère les fréquences majoritaires) ou au seuil de 5% (qui mesure le risque pris en rejetant les fréquences extérieures) ?

l'intervalle de Première (noté par la suite IF1) est bilatéral, mais non nécessairement centré, les intervalles de Seconde (noté par la suite IF2) et celui de Terminale (noté par la suite IFT) sont centrés sur p , mais pas nécessairement bilatéraux.

2. De la Seconde à la Terminale, trois modes de détermination de l'IF

On prélève un échantillon de taille n dans une population comportant une proportion p d'individus présentant un caractère C .

La variable aléatoire X représentant le nombre d'individus de l'échantillon présentant le caractère C suit la loi binomiale $B(n;p)$, on lui associe la variable aléatoire $F = X/n$ représentant la fréquence d'individus de l'échantillon présentant le caractère C . On cherche un intervalle de fluctuation pour cette variable F .

a. Détermination "à la main" en Première pour IF1

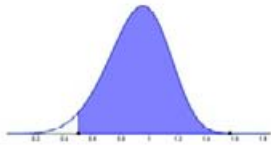
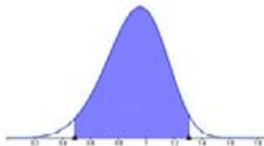
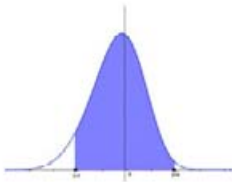
On détermine en Première un *intervalle de fluctuation à partir de la distribution des fréquences obtenue par la loi binomiale*, pour laquelle on sait calculer exactement la distribution des fréquences

Exemple : $B(20 ; 0,25)$

- intervalle unilatéral à droite: $IF = [0,1;1]$
- intervalle unilatéral à gauche: $IF = [0;0,4]$
- intervalle bilatéral : $IF = [0,1;0,45]$
- intervalle centré sur p : $IF = [0,05; 0,45]$

b. Formulation explicite pour n grand en Terminale pour IFT

Sous certaines conditions³, cette loi bino-

<p>Intervalle unilatéral à droite : $IF = [A ; 1]$ avec A tel que $P(F \geq A) \geq 0,95$ ou $P(F < A) \leq 0,05$</p> 	<p>Intervalle bilatéral (on symétrise les risques): $IF = [A;B]$ avec A tel que $P(F < A) \leq 0,025$ et B tel que $P(F > B) \leq 0,025$</p> 	<p>Intervalle centré sur p : $IF = [p-e;p+e]$ avec $P(p-e < F < p+e) \geq 0,95$</p> 
<p>Remarques : on peut définir de même un intervalle unilatéral à gauche : $IF = [0;B]$ avec B tel que $P(F \leq B) \geq 0,95$ ou $P(F > B) \leq 0,05$</p>	<p>Lorsque la loi est symétrique, l'intervalle bilatéral et l'intervalle centré sont identiques</p>	

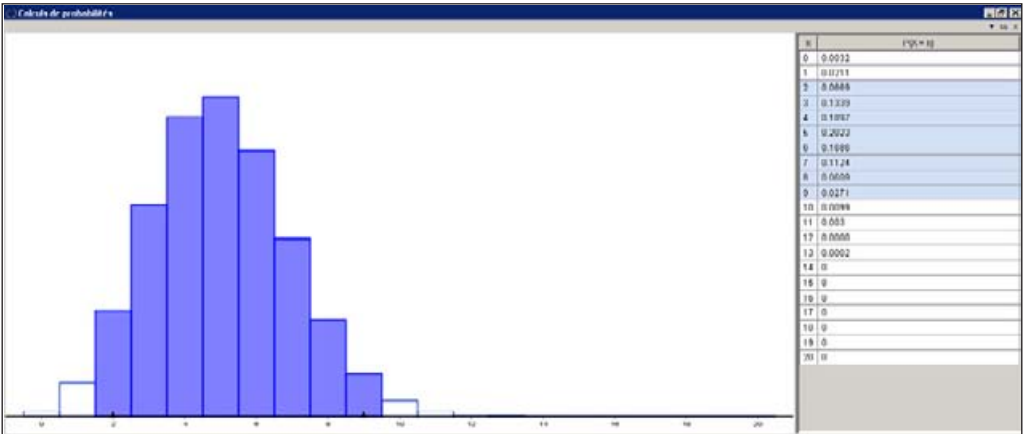
³ Les conditions d'approximation d'une loi binomiale par une loi normale varient selon les auteurs et les sources. En

général elles quantifient l'idée « n assez grand et p pas trop voisin de 0 ou de 1 » ; le programme 2011 précise : « $n > 30, np > 5$ et $nq > 5$ »).

... AUTOUR DES INTERVALLES DE FLUCTUATION

P.O.2, p 8

* 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 . Cet intervalle peut être obtenu, de façon approchée, par simulation.



Le programme de Seconde impose un IF centré sur p (voir P.O.2 p8 ci-dessus). Le document ressources de Première [R.1 page 38] privilégie un IF bilatéral avec la méthode illustrée ci-dessus.

miale peut être approchée par la loi normale. Pour travailler avec la loi normale, on dispose de la fonction de répartition qui permet de déterminer un IF. Pour un IF centré sur p , de la forme $IF = [p - ks'; p + ks']$ où s' est l'écart type, les tables statistiques donnent l'équivalence $P(F \in IF) = 0,95$ si, et seulement si, $k = 1,96^4$. On en déduit ainsi une formulation explicite de l'intervalle de fluctuation à 95% centré sur p , valable pour une loi normale :

$$IFT = \left[p - 1,96 \frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 1,96 \frac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$$

Cet intervalle de fluctuation, valable pour une loi normale, est un intervalle de fluctuation asymptotique (i.e. valable quand n tend vers l'infini ; en pratique pour n grand⁵) pour une loi binomiale.

c. Formulation simplifiée pour n grand en Seconde pour IF2

On peut approcher l'intervalle de Terminale IFT par un intervalle à la formulation plus simple :

$$IF2 = \left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}} \right]$$

L'intervalle IF2 est plus large que l'intervalle IFT puisque obtenu par deux majorations : celle de 1,96 par 2, et celle de $\sqrt{p(1-p)}$ par 0,5. La première majoration est une bonne approximation, tandis que la seconde majoration l'est d'autant moins que p s'éloigne de 0,5. On peut considérer qu'elle est satisfaisante lorsque p est compris entre 0,2 ou 0,8, car alors $0,4 \leq \sqrt{p(1-p)} \leq 0,5$.

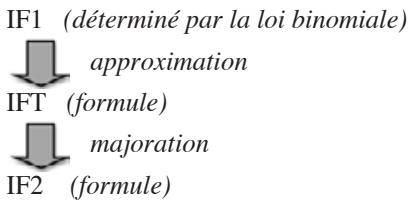
4 En fait cette valeur est déjà un arrondi.

5 Conséquence du théorème de Moivre-Laplace.

L'IF de Seconde est donc le résultat de deux opérations : d'abord l'approximation de la loi binomiale par une loi normale, puis la majoration simplifiant la formulation des bornes.

d. Conclusion

Les trois intervalles sont reliés entre eux d'après le schéma suivant :

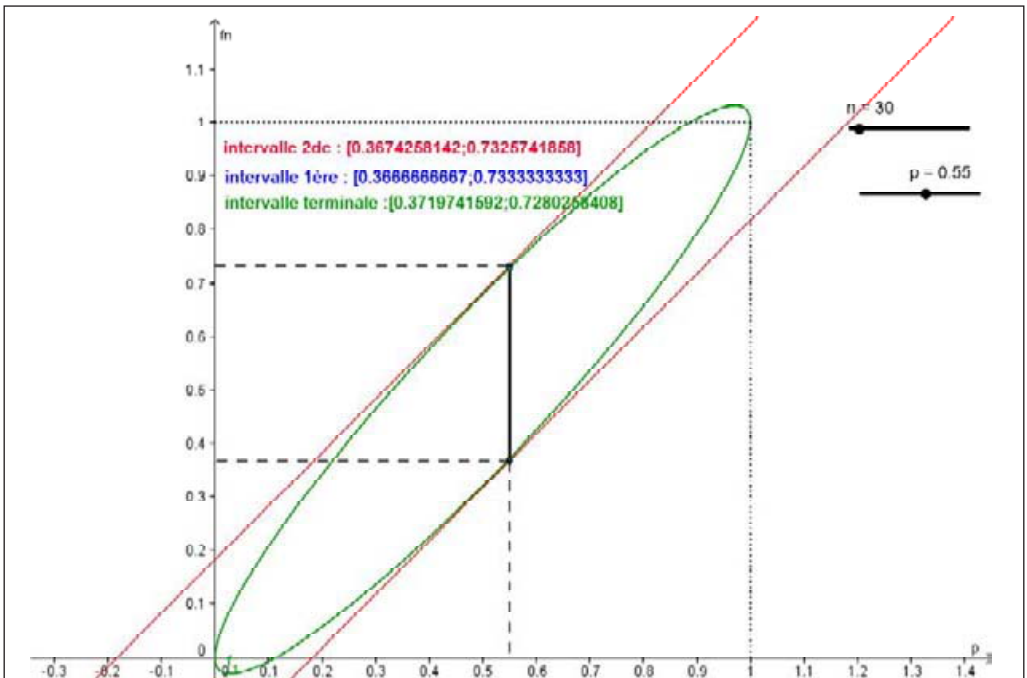


Partie 2 :

**Obtient-on trois intervalles différents ?
 Que se passe-t-il lorsqu'ils sont distincts ?**

Du fait des majorations, IFT est toujours inclus dans IF2 : l'IF de Seconde est plus large donc moins précis que celui de Terminale. En revanche la position de celui de Première par rapport à IF2 est variable. Or lorsqu'un IF ne contient pas IF1, cela signifie qu'il peut contenir moins de 95% des fréquences des échantillons prélevés, et donc ne pas satisfaire exactement la définition.

Cela est-il possible ? C'est ce que nous allons regarder :



Lecture : en abscisses, la probabilité p ; en ordonnées, les bornes des IF correspondants :
 la courbe ovale est l'ensemble des bornes inférieure et supérieure de IFT,
 les deux droites parallèles représentent les bornes inférieure et supérieure de IF2

1. Ces intervalles sont-ils distincts ?

a. Ecriture de ces intervalles

Par construction, les bornes de IF1 sont des fréquences possibles : le document ressource de Première demande d'ailleurs de l'écrire sous la forme $[a/n ; b/n]$, où a et b sont des entiers. En revanche ce n'est pas le cas pour IFT et IF2, mais il peut être éclairant d'écrire les trois IF sous cette forme :

* pour $n = 50$ et $p = 0,5$, on trouve $IF2 = [0,3585... ; 0,6414...] = [17,92/50 ; 32,07/50]$ ce qui donne un IF2 réel de :

$$IF2 = [18/50 ; 32/50] = [0,36 ; 0,64]$$

* pour $n = 60$ et $p = 0,5$, on trouve $IF2 = [0,3709... ; 0,6920...] = [22,25/60 ; 37,75/60]$ ce qui donne un IF2 réel de :

$$IF2 = [23/60 ; 37/60]$$

b. Des cas où les trois intervalles sont égaux

Comme on l'a vu, le calcul des bornes peut faire croire qu'on a toujours des IF différents, mais lorsqu'on rectifie en tenant compte du caractère discret des fréquences, on s'aperçoit du contraire :

* ils peuvent être tous trois égaux c'est-à-dire $IF2 = IF1 = IFT$: par exemple, pour $n = 50$ et $p = 0,43$; les calculs donnent

$$IF2 = [0,289 ; 0,571] , IF1 = [0,3 ; 0,56] \text{ et } IFT = [0,293 ; 0,567],$$

mais il s'agit en fait du même

$$IF = [15/50 ; 28/50] = [0,3 ; 0,56] ;$$

* on a aussi souvent simplement l'égalité $IF1 = IFT$ ⁶.

Dans ces situations, nos trois intervalles contiennent bien la fréquence observée pour au moins 95% des échantillons prélevés.

2. Cas de l'intervalle de Terminale

Néanmoins, dans quelques cas, et du fait de l'approximation effectuée, il arrive que IFT contienne la fréquence pour moins de 95% des échantillons prélevés.

* Cela se produit lorsqu'on a une stricte inclusion : IFT inclus dans IF1, et IF1 inclus dans IF2. (figure ci-dessous)

Par exemple pour $n = 50$ et $p = 0,24$:

$$IF2 = [0,10 ; 0,38]$$

$$IF1 = [0,12 ; 0,36]$$

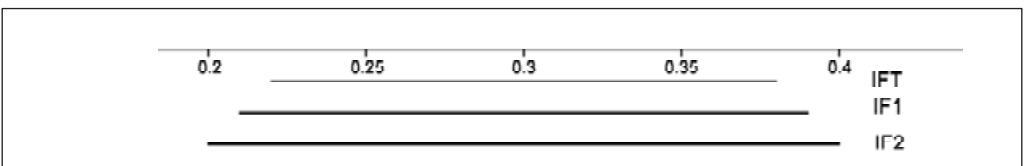
$$IFT = [0,14 ; 0,34]$$

La probabilité que F appartienne à IFT n'est que de 0,934, donc inférieure à 95%.

* Cela se produit aussi lorsque $IF1 = IF2$ mais avec un IFT distinct, par exemple avec $n = 50$ et $p = 0,44$: $IF1 = IF2 = [15/50 ; 29/50]$ mais $IFT = [16/50 ; 28/50]$ avec une probabilité réduite à 93,7%.

On retiendra que si la probabilité que la fréquence soit dans IF1 est, par construction, au

Trois intervalles possibles...



⁶ Pour $n = 50$ et $p = 0,36$ par exemple.

moins égale à 95%, en revanche ce n'est plus forcément vrai pour IFT, qui est une approximation de IF1. Néanmoins, dans tous les exemples présentés jusqu'ici, lorsqu'on majore IFT, l'intervalle IF2 obtenu contient IF1, ce qui lui garantit une probabilité d'au moins 95%.

On pourrait donc conjecturer que c'est toujours vrai, mais la suite va montrer le contraire.

3. Cas de l'intervalle de Seconde

a. IF2 contient-il 95% des fréquences ?

La loi normale étant une approximation de la binomiale, on comprend que IFT est une approximation de IF1, donc peut contenir la fréquence pour moins de 95% des échantillons observés. Mais par le procédé de majoration amenant celui de Seconde, on peut penser qu'on récupère plus de 95%. Le programme de Seconde [PO2, page 8] précise en effet (encadré 1 ci-dessous) :

Encadré 1

Le professeur peut indiquer aux élèves le résultat suivant, utilisable dans la pratique pour des échantillons de taille $n \geq 25$ et des proportions p du caractère comprises entre 0,2 et 0,8 : si f désigne la fréquence du caractère dans l'échantillon, f appartient à l'intervalle $\left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}} \right]$ avec une probabilité d'au moins 0,95. Le professeur peut faire percevoir expérimentalement la validité de cette propriété mais elle n'est pas exigible.

Encadré 2

Note
Ce théorème est vrai pour $n \geq 25$ et pour $p \in [0,2; 0,8]$.

b) Une propriété générale

Les résultats observés dans l'exemple précédent sont en accord avec la propriété générale suivante, démontrée en théorie des probabilités et des statistiques.

THÉORÈME

La fréquence du caractère dans un échantillon de taille n appartient à l'intervalle $\left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}} \right]$, avec une probabilité d'au moins 0,95.

Cette indication est reprise dans les manuels sous forme de « propriété admise », et même de théorème [AA, page 254 : encadré 2 ci-dessous].

Or *il admet des contre-exemples* : le contre-exemple cité le plus souvent⁷ vient pour $n = 30$ et $p = 0,55$, la probabilité n'est que de 93,5%, cet exemple est détaillé ci-après. Mais on peut en trouver d'autres : pour $n = 99$ et $p = 0,53$, elle est 94,3%... D'ailleurs ce problème apparaît dans le document ressources proposé en accompagnement de la partie probabilités-statistiques du programme de Terminale [RT, pages 24-25], sans que le problème soit signalé (voir encadré 3 de la page suivante).

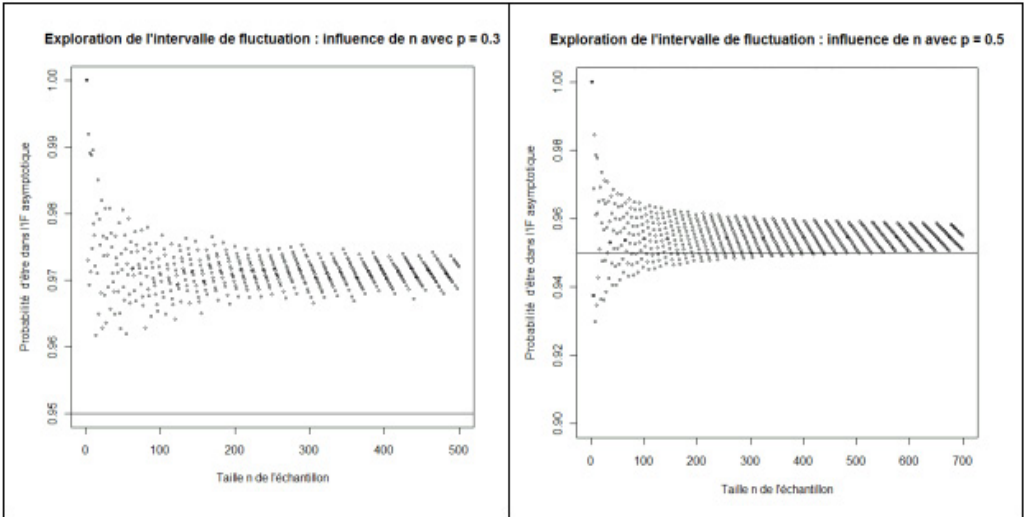
b. Analyse des cas où IF2 contient moins de 95% des fréquences

Essayons de comprendre ce qui se passe :

- la formule de Seconde donne $IF2 = [0,36743 ; 0,73257] = [11,02/30 ; 21,98/30]$

⁷ Plusieurs articles ont déjà évoqué ce problème.

Encadré 3



- cet intervalle est un intervalle de fluctuation des fréquences, mais on voit mieux avec les effectifs X , qui sont des nombres entiers ; comme $F = X/30$, on aura :

$$F \in \text{IF2} \text{ssi } 0,36743 \leq F \leq 0,73257$$

$$\text{ssi } 11,02 \leq X \leq 21,98.$$

X ne prenant que des valeurs entières, ceci équivaut à $12 \leq X \leq 21$, et un élève de Première pourra facilement vérifier que $P(12 \leq X \leq 21) = 0,935$; ce qui signifie que $P(F \in \text{IF2}) < 95\%$.

L'erreur vient du fait que les valeurs prises par les fréquences, comme celles prises par les effectifs, sont discrètes (elles ne prennent que $n + 1$ valeurs $0/n, 1/n, \dots$). Or ceci n'apparaît pas quand on utilise l'écriture décimale induite par l'expression des bornes de IF2. En revanche le problème se comprend mieux en regardant l'intervalle de fluctuation des effectifs, pour lequel la formule de Seconde donne $\text{IFe} = [np - \sqrt{n} ; np + \sqrt{n}]$.

Ces bornes peuvent être entières, auquel cas elles le sont simultanément. Si ces bornes ne sont pas entières, on obtient⁸ en réalité :

$$\text{IFe} = [E(np - \sqrt{n}) + 1 ; E(np + \sqrt{n})]$$

où $E(x)$ désigne la partie entière du réel x .

On voit ainsi que le problème risque de se produire lorsque les bornes de l'intervalle de fluctuation des effectifs obtenues avec la formule sont « presque entières », donnant un intervalle de la forme $[N + \varepsilon ; N' - \varepsilon']$ avec N et N' entiers, et ε et ε' proches de 0. En effet, dans ces cas, l'intervalle réellement pertinent $[N + 1 ; N' - 1]$ est « beaucoup » plus petit que l'intervalle de la formule.

c. Un « théorème statistiquement vrai »

Ainsi, avec les contraintes $n > 25$ et p compris entre 0,2 et 0,8, les probabilités associées

⁸ Evidemment une personne raisonnable confrontée à ces bornes arrondira en élargissant son IF2 de façon à être sûr de contenir 95% : on prendrait $\text{IF} = [11/30 ; 22/30]$.

aux IF2 se situent en gros entre 0,93 et 0,99. Elles ne sont donc pas toujours supérieures à 0,95, même si c'est le cas pour la grande majorité.

La non-universalité du résultat proposé n'est en pratique pas si grave. Pour le statisticien, elle ne pose pas de problème car les contre-exemples sont rares, et qu'il sait adapter ses outils. Pour le lycéen, ce qui est important est qu'un intervalle de fluctuation sert à indiquer la fourchette dans laquelle se trouve la « grande majorité » des fréquences des échantillons observés, le seuil de 95% étant fixé arbitrairement. Néanmoins, pour le professeur, il faudra réfléchir à une formulation satisfaisante pour la Seconde, ce qui va être l'objet de la suite.

Partie 3

Quelques difficultés liées au vocabulaire en jeu : quelle formulation donner en Seconde ?

Nous allons maintenant revenir sur la présentation proposée par le programme de Seconde [PO.2 : encadré ci-dessous], et aux difficultés conceptuelles sous-jacentes.

1. Deux interprétations de la « fréquence théorique » p

Dans le modèle de prélèvement d'un échantillon, on considère⁹ implicitement le fait que la proportion p qui est un paramètre de la population Ω se traduit sur chaque individu en termes de probabilité : la *proportion p dans la population Ω* d'individus présentant un caractère C devient la *probabilité p qu'un individu de la*

population Ω tiré au hasard présente ce caractère C . C'est ce principe de modélisation qui permet de se placer dans la théorie des probabilités.

Les exemples présentés en annexe montrent que cette correspondance n'est pas toujours aussi simple. De plus, elle est rarement explicitement formulée par le professeur. En revanche elle est largement utilisée dans la vie courante, dès qu'une proportion collective est identifiée à une probabilité individuelle. Par exemple, en santé publique, lorsqu'on compare les risques d'avoir telle maladie, corrélée à tel facteur, « trois fois plus de risque d'avoir un infarctus chez les fumeurs », on compare les proportions de malades dans les deux populations (fumeurs / non fumeurs), qu'on infère en termes de probabilité d'être malade pour chaque individu des deux populations. De même, une publicité qui annonçait « deux chances sur trois de vendre votre bien immobilier dans les deux mois », traduisant la proportion constatée en probabilité.

Il est donc facile d'interpréter la probabilité initiale p . Mais quel sens donner à la probabilité $p' = 0,95$ associée à l'IF ? Nous allons pour cela devoir nous intéresser à la population des individus concernés par cette probabilité, c'est-à-dire aux échantillons.

2. Questions autour du mot « échantillon »

a. Combien d'échantillons ?

En Seconde, puis en Première, on donne du sens aux probabilités par l'intermédiaire des simulations. On mène donc l'expérience aléatoire « concrètement », et on obtient des échantillons,

* 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 . Cet intervalle peut être obtenu, de façon approchée, par simulation.

⁹ C'est un choix de modèle.

« produits par l'expérience ». Ainsi dans une classe de 35 élèves où chacun lance une pièce dix fois, l'usage dira qu'on obtient « 35 échantillons de taille 10 » pour dire qu'il s'agit de 35 prélèvements d'échantillons.

Cette approche est d'ailleurs signalée dans la présentation du programme qui demande de faire des expériences réelles ou des simulations pour « obtenir IF, de manière approchée, par des simulations ». On dira donc en classe qu'« on simule N échantillons » (ou qu'on « prélève N échantillons par simulation »). Cette phrase sous-entend qu'on peut produire autant d'« échantillons » de taille n que l'on veut, par exemple produire par simulation une série de 1000 « échantillons ».

Cette conception initiale paraît pourtant remise en cause en Première, puisqu'on va déterminer une loi de probabilités sur ces échantillons, et pour cela compter le nombre d'échantillons théoriques :

Le programme de Seconde 2009 ne retient que ce type d'échantillons : "Un échantillon est constitué des résultats de n répétitions indépendantes de la même expérience".

On constitue (avec remise) des échantillons de taille 3.

On peut ainsi constituer $4^3=64$ échantillons.

Ainsi lorsqu'on s'intéresse au prélèvement d'un échantillon de taille n dans une population de taille N , cette définition dénombre N^n prélèvements d'échantillons possibles. Dans le cadre du schéma de Bernoulli, on obtient 2^n observations d'échantillons de taille n possibles. Ces échantillons *observables* sont les issues élémentaires au sens de la théorie des probabilités (c'est-à-dire des listes de longueur

n), ils sont en nombre fini fixé. Il me semblait alors contradictoire, qu'en Seconde on puisse produire « autant d'échantillons de taille n que l'on veut », alors qu'en Première ils étaient en nombre fini fixé égal à 2^n . Cette contradiction apparente m'a amenée à m'interroger sur la notion d'échantillon.

b. *qu'est-ce qu'un échantillon ?*

Soit p la proportion d'individus d'une population Ω présentant un caractère C , p peut s'interpréter aussi la probabilité pour chaque individu de la population Ω de présenter le caractère C . Le choix d'un individu est modélisé par une expérience de Bernoulli. Cette expérience aléatoire est traduite par une variable aléatoire $X : \Omega \rightarrow \{0 ; 1\}$. La valeur prise par X est 0 ou 1 suivant l'individu ω tiré au hasard dans la population Ω .

Un échantillon est alors obtenu par la répétition de l'expérience aléatoire consistant à tirer au hasard un individu et relever s'il présente le caractère C . Lorsque la population est assez importante par rapport à n , on peut assimiler le prélèvement (en principe un tirage sans remise) d'un échantillon de n individus à un tirage avec remise, et le modéliser par un schéma de Bernoulli.

Un échantillon de taille n est donc une liste (E_n) constituée de n « répétitions » de la variable $X : E_n = (X_1, \dots, X_n)$. Un échantillon de taille n est une variable aléatoire :

$$E_n : \Omega^n \rightarrow \{0 ; 1\}^n.$$

Au lycée, un échantillon est assimilé à une liste des issues des n tirages : par exemple un échantillon de trois lancers de pièce peut donner (P, P, F) associé à (1;1;0). On voit déjà que ce qu'on appelle échantillon au lycée est une réalisation de l'échantillon du supérieur, un

résultat des n réalisations de l'expérience aléatoire, autrement dit une image $E_n(\omega)$ où $\omega = (\omega_1, \dots, \omega_n) \in \Omega^n$. Ainsi, le sens du mot échantillon est différent entre le supérieur (l'application E_n) et le lycée (une image de cette application).

c. des échantillons d'échantillons ?

Un échantillon de taille n est donc au lycée une liste de n résultats.

- Lorsqu'on « réalise une série de n lancers », on réalise un échantillon de taille n , constitué de la liste des résultats de ces lancers. Avec deux résultats possibles (P associé à 1 ; F associé à 0), je peux faire des listes de n résultats.
- Lorsqu'on « réalise une série de n' échantillons de taille n », on réalise un échantillon de taille n' , constitué de la liste de n' échantillons de taille n . Avec 2^n échantillons élémentaires possibles, je peux faire des listes de n' échantillons.

Autrement dit, lorsqu'on Seconde on réalise des simulations pour obtenir IF de façon

approchée (voir programme), on doit constituer des échantillons d'échantillons...

3. l'échantillon et l'intervalle de fluctuation, interfaces entre les probabilités et les statistiques

En fait l'échantillon, et plus généralement la problématique de l'intervalle de fluctuation, se trouvent au carrefour de plusieurs problématiques, d'une part entre les probabilités et les statistiques, d'autre part entre l'expérience (le lancer d'une pièce) et sa répétition (l'échantillon)

a. Deux niveaux de lecture

Un échantillon de taille n est une liste de n résultats. Cette liste peut être regardée :

- comme une population, constituée de n individus qui sont les résultats ; sur cette population, j'opère un traitement statistique en calculant une fréquence,
- comme un individu d'une population ; j'opère un traitement probabiliste en étudiant le nombre d'issues élémentaires résultant du choix d'un individu dans cette population.

	Traitement probabiliste : Probabilité pour un individu de présenter le caractère	Traitement statistiques : fréquence, dans la population, d'individus présentant le caractère
Un individu = un résultat de lancer d'une pièce caractère = le résultat est pile, associé à 1	Probabilité pour un individu (résultat) de présenter le caractère (pile) : p avec $p=0,5$	Fréquence du caractère (« pile ») dans une population (échantillon) constituée de n individus : f_n avec, si n est grand, $f_n \approx 0,5$
Un individu = un échantillon de taille n caractère = la fréquence de pile dans cet échantillon est dans IF	Probabilité pour un individu (échantillon) de présenter le caractère (avoir une fréquence dans IF) : p' avec $p'=0,95$	Fréquence du caractère (« avoir une fréquence dans IF ») dans une population (série d'échantillon) constituée de n individus (échantillons) : t_n' avec, si n' est grand, $t_n' \approx 0,95$

* 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 . Cet intervalle peut être obtenu, de façon approchée, par simulation.

Chacun des deux niveaux fait un lien entre une probabilité et une fréquence mesurée :

- la probabilité $p = 0,5$ sur chaque lancer de faire « pile » se traduit par une fréquence $f \approx 0,5$ mesurée sur des échantillons de lancers.
- la probabilité $p' = 0,95$ sur chaque échantillon d'être dans IF se traduit par une fréquence $t' \approx 0,95$ mesurée sur des séries d'échantillons de lancers.

b. Fluctuation et loi des grands nombres

La proximité de f et p peut être formalisée par la loi des grands nombres, qui se retrouve

donc elle aussi aux deux niveaux (voir le tableau ci-dessous).

c. Deux regards sur l'échantillon

L'échantillon de taille n , constitué des résultats de n expériences identiques, se retrouve donc dans les deux niveaux. Mais on va changer le regard qu'on porte sur lui, de la même façon qu'un triplet (1;1;0) peut être considéré comme une liste de trois nombres, mais aussi par exemple comme un vecteur (et lui appliquer un traitement vectoriel). Ainsi, l'échantillon de Seconde, qui appartient au contexte statistique, « échantillon physique, dans la réa-

	Traitement probabiliste : Probabilité pour un individu de présenter le caractère	Interface : Loi des grands nombres	Traitement statistiques : fréquence, dans la population, d'individus présentant le caractère						
Individu = un lancer d'une pièce	Univers $\Omega =$ ensemble des lancers variable aléatoire $X : \Omega \rightarrow \{0 ; 1\}$ Pour une pièce équilibrée, la loi de probabilité de X est : <table border="1" style="margin-left: 20px;"> <tr> <td>k</td> <td>0</td> <td>1</td> </tr> <tr> <td>P(X=k)</td> <td>0,5</td> <td>0,5</td> </tr> </table> <p>$p=0,5$</p>	k	0	1	P(X=k)	0,5	0,5	Échantillon (X_1, \dots, X_n) $F_n = \frac{X_1 + \dots + X_n}{n}$ $F_n : \Omega^n \rightarrow [0;1]$ fréquence de pile dans l'échantillon $\lim_n P(F_n - p > \epsilon) = 0$	Une série de n lancers = un échantillon empirique $(\omega_1, \dots, \omega_n)$ de taille n , où chaque élément est le résultat d'un lancer f_n la fréquence de pile dans l'échantillon si n est grand, $f_n \approx 0,5$
k	0	1							
P(X=k)	0,5	0,5							
Individu = un échantillon de taille n	variable aléatoire $I : \Omega^n \rightarrow \{0 ; 1\}$ qui vaut 1 si $F_n(\omega) \in IF$, 0 sinon ; loi de probabilité de I : <table border="1" style="margin-left: 20px;"> <tr> <td>k</td> <td>0</td> <td>1</td> </tr> <tr> <td>P(I=k)</td> <td>0,05</td> <td>0,95</td> </tr> </table> <p>$p'=0,95$</p>	k	0	1	P(I=k)	0,05	0,95	Échantillon $(I_1, \dots, I_{n'})$ $T_{n'} = \frac{I_1 + \dots + I_{n'}}{n'}$ $T_{n'} : (\Omega^n)^{n'} \rightarrow [0;1]$ fréquence d'échantillons dans IF $\lim_n P(T_{n'} - p' > \epsilon) = 0$	Une série de n' échantillons de taille n = un échantillon empirique de taille n' , où chaque élément est un échantillon de taille n . $t_{n'}$ la fréquence d'échantillons ayant une fréquence dans IF si n' est grand, $t_{n'} \approx 0,95$
k	0	1							
P(I=k)	0,05	0,95							

lité » va être regardé en Première dans le contexte des probabilités, d'« issue élémentaire produite par la répétition de l'expérience de Bernoulli ». La réponse à la question « combien d'échantillons » ne dépend pas de l'objet en jeu (l'échantillon) mais du regard porté sur lui. Et c'est ce nouveau regard, porté sur l'échantillon à l'aide de la théorie des probabilités, qui va permettre de justifier la proximité de f et de p constatée sur les échantillons statistiques.

En Troisième, on porte un double regard sur les lancers : le regard des probabilités (on modélise par $p = 0,5$) et le regard des statistiques (on observe que $f \approx 0,5$). Cette proximité de p et f est étayée par l'intervalle de fluctuation puisque lorsque n est grand, l'amplitude de la fluctuation diminue du fait du resserrage de l'intervalle de fluctuation des fréquences, qui contient l'essentiel des fréquences (95%), autour de p . Or pour parler d'IF, il faut regarder les échantillons. En Seconde, on porte un regard statistique sur les échantillons : on observe que $t' \approx 0,95$. Mais pour expliquer ce constat, on va porter en Première un regard probabiliste sur les échantillons, pour obtenir $p' = 0,95$. Le problème étant que la proximité de t' et p' , pour être étayée, nécessiterait un nouvel IF¹⁰ : on est face à une belle mise en abyme !

4. Fluctuations de p et de p'

Ainsi lorsqu'on étudie la notion d'intervalle de fluctuation, on ne s'intéresse plus à la probabilité p qu'un individu ait le caractère (C) mais à une autre probabilité p' qu'un échantillon prélevé présente le caractère (C'). Or le lien entre la probabilité $p = 0,5$ et les fréquences observées sur les séries d'individus fait appel à un premier IF associé à $p = 0,5$. Le lien entre la probabilité $p' = 0,95$ et les fréquences constatées sur les séries d'échantillons va donc faire, lui, appel à un second IF associé à $p' = 0,95$.

a. Exemple du pile ou face : $p = 0,5$.

Je simule 100 lancers ; cette série de 100 lancers peut être assimilée au prélèvement d'un échantillon de taille $n = 100$. Je regarde la fréquence f de pile : f est proche de $p = 0,5$.

La fluctuation de f autour de $p = 0,5$ est encadrée par l'intervalle de fluctuation $IF = [0,4 ; 0,6]$ avec $P(F \in IF) = p'$. Nous prendrons $p' = 0,95$, même si ici en réalité $p' = 0,965$. Ce qui signifie que j'ai de grandes chances (95%) pour que le nombre de pile observé sur cet échantillon soit compris entre 40 et 60.

Mais comment interpréter ce p' ?

Je simule 100 prélèvements d'« échantillons », et je regarde la fréquence t' de ces échantillons dont la fréquence observée f est dans l'intervalle $[0,4 ; 0,6]$: t' est proche de $p' = 0,95$. La fluctuation de t' autour de $p' = 0,95$ est encadrée par un nouvel intervalle de fluctuation $IF'^{10} = [0,91; 0,99]$ avec $P(T' \in IF') = 0,95$ (même remarque qu'au dessus).

Autrement dit j'ai de grandes chances (95%) pour que le nombre d'échantillons simulés ayant une fréquence située dans l'intervalle de fluctuation $IF = [0,4; 0,6]$ soit compris entre 91 et 99.

Remarque : en réalité dans cet exemple avec $p = 0,5$, pour des échantillons de taille 100 on aura $IF = [0,4 ; 0,6]$, avec $P(F \in IF) = 0,965$. Lorsqu'on réalise 100 séries de tels échantillons, pour lesquels $p' = 0,965$, on obtient $IF' = [0,93 ; 1]$ avec $P(T' \in IF') = 0,976$. On peut donc s'attendre, de manière quasiment certaine, à ce que les séries présentent entre 93% et 100% d'échantillons ayant une fréquence

10 Calculé avec la loi binomiale.

dans IF. Et c'est effectivement ce qu'on constate en pratique lorsqu'on met en œuvre des simulations avec les élèves.

b. *A quelle probabilité donne-t-on du sens ?*

On a donc un enchaînement. Si p est la probabilité pour un individu d'avoir le caractère (C), on voit apparaître plusieurs phénomènes de fluctuation :

- une fluctuation autour de p : les fréquences constatées sur les échantillons prélevés seront proches de p (car situées dans IF), avec une probabilité $p' = 0,95$ (environ) d'être dans IF
- une fluctuation autour de $p' = 0,95$: les fréquences des échantillons dont la fréquence observée seront proches de p' avec une nouvelle probabilité $p'' = 0,95$, probabilité pour une série d'échantillon d'avoir le caractère (C') = il y en a 95% dans IF'.

Finalemment en voulant donner du sens à la probabilité p par la loi des grands nombres et sa quantification grâce l'intervalle de fluctuation, on s'appuie sur une autre probabilité $p' = 0,95$...

5. En conclusion : une formulation pour la Seconde

Compte tenu des éléments évoqués, on ne peut pas écrire en Seconde que « la fréquence F appartient à IF2 avec une probabilité d'au moins 95% » car c'est faux. De plus, même si dans la pratique elle ne semble pas présenter de difficulté aux élèves de Seconde, il convient d'être conscient de la complexité cachée de cette définition qui renvoie une probabilité à une autre, s'appuie sur le lien entre probabilité et fréquence établi en Troisième tout en utilisant le mot échantillon sans précaution.

Cette complexité me semble opacifier l'intérêt de l'IF, alors que l'idée importante à ce niveau est qu'il s'agit d'un « intervalle dans lequel la fréquence F de l'échantillon qu'on va prélever est censée se trouver avec une forte probabilité (ici de 95 %) ».

Il me semble donc qu'une formulation suffisante en Seconde pourrait être : « environ 95% des échantillons ont une fréquence qui appartient à IF », « intervalle dans lequel se trouvent l'essentiel (soit environ 95%) des fréquences constatées sur les échantillons ». Une telle définition est correcte quel que soit le sens donné à chaque terme, elle suffit à faire comprendre la notion d'intervalle de fluctuation à ce niveau, tout en permettant son utilisation pratique.

Partie 4

Trouver une cohérence au programme du lycée

Dans le programme de Seconde précédent, on se contentait de constater la fluctuation d'échantillonnage, tout en faisant remarquer que cette fluctuation diminuait lorsque la taille de l'échantillon augmentait : la fréquence observée était d'autant plus proche de p que l'échantillon sur lequel elle était mesurée était grand.

Maintenant nous disposons d'un outil qui permet de mesurer cette fluctuation en l'encadrant dans un intervalle. *On peut ainsi prévoir dans quelle fourchette va se trouver la fréquence obtenue. Cette problématique fournit ainsi un fil conducteur pour le thème probabilités-statistique au lycée.*

Chaque classe va alimenter la problématique par son regard sur les différents aspects de la question :

Seconde (noté IF2)	Première (noté IF1)	Terminale (noté IFT)
Centré sur p	Bilatéral (2,5% maximum de chaque côté)	Centré sur p
Formules pour les bornes	Pas de formules pour les bornes	Formules pour les bornes
Pas de théorie en Seconde : sensibilisation, justification par simulations	Théorie avec la loi binomiale	Théorie avec la loi normale et avec le théorème de Moivre-Laplace
$P(F \in IF2) \approx 0,95$ c'est-à-dire IF2 contient environ 95% des fréquences F des échantillons prélevés (du fait de l'approximation de la binomiale par la normale)	$P(F \in IF1) \geq 0,95$ c'est-à-dire IF1 contient par construction au moins 95% des fréquences F des échantillons prélevés	$P(F \in IFT) \approx 0,95$ c'est-à-dire IFT contient environ 95% des fréquences F des échantillons prélevés (du fait de l'approximation de la loi binomiale par une loi normale)

1. Déterminer l'IF

Ainsi en Seconde on observe IF sur des prélèvements d'échantillons par simulation, en Première on obtient IF en modélisant ce prélèvement par la loi binomiale. Le programme de Terminale vient faire le lien, en répondant aux deux problèmes soulevés en Seconde (pas de théorie) et en Première (pas de formule), mais au prix d'une approximation : c'est ce troisième intervalle qui est l'outil réel des statisticiens, et qui à ce titre est enseigné dans le supérieur.

2. Expliquer la proximité de f et p

On répète n fois l'expérience aléatoire consistant à choisir au hasard un individu dans une population, c'est-à-dire que l'on prélève un

échantillon de taille n ; on garde les notations précédentes :

p = probabilité théorique (proportion dans la population), f = fréquence observée (proportion dans l'échantillon)

p' = probabilité pour un échantillon d'avoir f dans IF ; t' = fréquence observée d'échantillon ayant f dans IF.

(Voir le tableau récapitulatif ci-dessous)

3. Mesurer la fluctuation d'échantillonnage

Il s'agit non seulement d'observer, mais aussi de quantifier la fluctuation des fréquences constatées sur les échantillons (voir encadré de la page suivante).

Troisième	observation de la proximité entre f et p , d'autant plus proches que n est grand
Seconde	On fixe n, on calcule IF (défini à partir de n et p), et on observe par simulations que $t' \approx 0,95$
Première	On fixe n, on pose $p'=0,95$ et on détermine IF
Terminale	On calcule IF (défini à partir de n et p), et on prouve par TML que $t' \approx 0,95$ lorsque n est grand ; ainsi IFT fournit une explication de la propriété $t' \approx 0,95$ observée en Seconde, et une approximation de l'intervalle cherché en Première.

Troisième	Constat expérimental (lancers de pièces et de dés) de la différence entre fréquence constatée et probabilité théorique, et de leur proximité lorsque l'expérience est répétée un grand nombre de fois
Seconde	Constat expérimental (lancer de pièces et de dés) de la fluctuation . Première quantification de ce phénomène : quantification de l'ampleur de la fluctuation par l'IF ; vérification expérimentale de l'IF par simulations : environ 95% des fréquences constatées sur les échantillons prélevés sont dans IF
Première	Apport d'un cadre théorique à ce prélèvement d'échantillons Schéma de Bernoulli et loi binomiale permettent un calcul explicite des probabilités et donc de l'IF : on peut calculer, pour un échantillon de taille donnée, la probabilité de chaque fréquence possible. La construction possible de l' histogramme de répartition permet la visualisation de la fluctuation théorique .
Terminale	Passage du discret au continu : modélisation histogramme par courbe de densité (donc calcul de la probabilité par une intégrale) qui donne une idée visuelle simple de l'allure de la fluctuation.

4. Prendre une décision sur la conformité d'un échantillon

L'intervalle de fluctuation est surtout un outil important de la prise de décision. Rappelons [W] qu'« il permet de détecter un écart important entre la fréquence observée sur un échantillon et la fréquence théorique (proportion ou probabilité). Le fait d'obtenir une valeur en dehors de cet intervalle s'interprète alors en mettant en cause la représentativité de l'échantillon ou la fréquence théorique ». C'est d'ailleurs cette utilisation qui est indiquée dans le programme de Terminale : « utiliser l'intervalle de fluctuation pour rejeter ou non une hypothèse sur une proportion ».

On utilisera donc un IF lorsqu'on connaît p (ou qu'on émet une hypothèse sur sa valeur) et n (la taille de l'échantillon)¹¹, puisqu'on peut alors déterminer un IF qui fournit une fourchette pour les fréquences possibles. Dans la modélisation de l'échantillonnage, p est le *paramètre de la loi binomiale utilisée*. En pratique, ce paramètre

p est une probabilité assimilée à la proportion dans la population-mère, ce qui explique la demande du programme. Mais les exemples donnés en annexe montrent qu'identifier la « proportion » en jeu n'est pas toujours si simple.

Lorsque la fréquence observée f n'est pas dans IF, on dira que cette fréquence n'est pas conforme à ce qu'on attend. On rejette donc l'hypothèse selon laquelle « notre échantillon a été prélevé au hasard dans une population où un certain caractère (C) a une probabilité p ». Les raisons de cette non-conformité de l'échantillon sont :

- soit que, dans la population où est fait le prélèvement, le caractère (C) a une probabilité qui prend une autre valeur que celle présumée (voir exemples) ;
- soit que l'échantillon n'a pas été prélevé au hasard.

On peut remettre alors en cause soit p , soit le procédé de prélèvement « au hasard » de l'échantillon.

Cette partie devrait bien évidemment être la plus importante de l'article : elle montrerait l'intérêt de l'introduction de ce nouvel objet au

¹¹ Remarquons que la taille de la population globale n'intervient pas ; il faut juste qu'elle soit assez grande pour assimiler le prélèvement de l'échantillon à un tirage avec remise.

lycée, son utilité, son importance. Un approfondissement de l'utilisation de l'IF dans la prise de décision a été donné dans [YD]. Mais surtout, la problématique de la prise de décision va être l'objet des cours de statistiques du supérieur, dans de nombreux domaines (contrôle-qualité, médecine, ...).

Conclusion

Un tout petit point de programme, mais dont la nouveauté a amené de nombreuses interrogations et échanges, pour tenter d'éclaircir

quelques points délicats : échantillon, proportion/probabilité. Nous avons aussi dû confronter les définitions proposées de façon à en comprendre l'articulation, ce qui nous a amenés à modifier notre présentation en Seconde.

Mais l'enjeu me paraît important, parce que je pense que cette partie de probabilités-statistiques nouvellement introduite au lycée est à la fois indispensable et très riche, les problèmes en jeu sont intéressants et posent des questions sur lesquelles il convient de réfléchir avec nos élèves.

Sources et bibliographie

Cet article est issu avant tout d'échanges menés avec mes collègues du lycée et les membres du Gralym¹². Je remercie en particulier Bruno Chambon et Pascale Boulais qui ont eu la patience de s'intéresser à mes questionnements.

Je remercie aussi Yves Ducel pour la gentillesse avec laquelle il m'a éclairée sur des points théoriques et amenée à reformuler, et donc approfondir.

Citons enfin Brigitte Chapat et Pierre Grihon dont les exposés dans le cadre de l'IREM ou de l'APMEP m'ont permis de répondre à quelques-unes de mes questions.

[YD] Ducel Yves et Sausseur Bruno, « La prise de décision de la Seconde à la Première », *Repères IREM*, 85, octobre 2011, pages 31-49

[PO.2] Programme de Mathématiques, classe de Seconde, BO n°30 du 23 juillet 2009

[PO.T] Programme de Mathématiques, classe de Terminales, BO spécial n°8, du 13 octobre 2011

[R.1] Ressources pour la classe de Première générale et technologique, statistiques et probabilités, page 38

[R.T] Ressources pour la classe de Terminale générale et technologique, statistiques et probabilités

[BC] Chapat, Brigitte : [Intervalles de fluctuation - APMEP Régionale de Toulouse](#), diaporama consultable à partir du site de l'APMEP

[AA] Antibi, André (sous la direction de) : *MATH 2de, travailler en confiance*, manuel scolaire, Nathan, 2010

[W] Wikipédia, article « Intervalle de fluctuation »

12 Gralym : Groupe académique lycée de l'académie de Montpellier (groupe animé par l'équipe des IPR de mathématiques)

ANNEXE

Trois situations d'utilisation de l'IF :
quelle « proportion » rejette-t-on ?

Exemple 1 : Les bonbons¹³

Situation : Un distributeur de bonbons est censé contenir $p = 25\%$ de bonbons à la menthe. Je prélève un échantillon de 20 bonbons, pour lequel j'observe une certaine fréquence f . Il apparaît que f n'appartient pas à IF.

Analyse :

p = la proportion de bonbons à la menthe dans le distributeur (population-mère) se traduit en : p est la probabilité pour chaque bonbon prélevé d'être à la menthe.

Si f n'appartient pas à IF, on rejette la probabilité $p = 0,25$, donc on rejette l'hypothèse d'une proportion $p = 25\%$ dans le distributeur.

Ici on rejette l'hypothèse sur la proportion en soupçonnant que l'échantillon est prélevé dans la population-mère initiale qui n'a pas une proportion p , mais une autre proportion.

Exemple 2 : L'affaire Castaneda¹⁴

Situation : La population du Texas compte une proportion $p = 79\%$ de personnes d'origine mexicaine. Or sur un échantillon 830 jurés tirés au sort dans cette population, on observe une certaine proportion f . Il apparaît que f n'appartient pas à IF.

Analyse :

La proportion de mexicains dans la population texane (population-mère) se traduit en : p est la probabilité pour un Texan d'être Mexicain.

Si f n'appartient pas à IF, on rejette ..., mais que rejette-t-on ? On ne rejette pas la proportion de mexicains au Texas ! Donc on ne rejette pas l'hypothèse d'une proportion $p = 79\%$ dans la population-mère initiale. On rejette

- * soit la population mère choisie : je n'avais pas choisi la bonne population-mère, puisque les jurés sont tirés au sort dans une sous-population (on doit être inscrit sur les listes électorales, ou être anglophone, ou ...). Ce qui revient bien à rejeter l'hypothèse « la proportion dans la population-mère constituée des Texans éligibles est $p = 0,79$ »
- * soit on rejette la conformité de l'échantillon à la proportion $p = 0,79$ en émettant le soupçon qu'il n'a pas été prélevé au hasard.

De même, l'an dernier j'avais une classe de Seconde avec 80% de filles. En prenant comme hypothèse $p = 0,5$ (proportion de filles sur l'ensemble des Seconde), on constate que f n'est pas dans IF. On peut penser :

13 .Source : Brigitte Chaput

14 Source : Document ressource

- * soit que la classe n'est pas constituée au hasard (pourquoi ?)
- * soit que mon échantillon « classe » a été constitué dans une population-mère constituée des élèves dans laquelle la proportion n'est pas de 0,5 (et c'était le cas, classe constituée d'élèves ayant choisi l'option Littérature).

Cet exemple 2 est donc plus subtil : *on rejette une hypothèse sur la proportion en soupçonnant que l'échantillon est prélevé non pas dans la population-mère initiale (de proportion p), mais dans une autre population.*

Exemple 3 : Les enfants de Xicun

Situation : Dans le village de Xicun, on a enregistré 20 naissances dont 16 garçons. Cette fréquence observée $f = 0,8$ n'est pas dans l'intervalle de fluctuation de $p = 0,5$ pour $n = 20$. Il apparaît que f n'appartient pas à IF.

Analyse :

On connaît la proportion de garçons parmi les enfants chinois déjà nés, à partir de laquelle on peut inférer la probabilité, pour un enfant existant choisi au hasard, d'être un garçon. Ici c'est plutôt la proportion de garçons parmi les naissances ayant déjà eu lieu (population passée) qu'on connaît, et à partir de laquelle on infère la probabilité, pour une naissance à venir (population future), de donner un garçon.

On ne rejette pas la valeur de p sur une population existante mais sur une population potentielle, une population à venir. Ainsi *on ne rejette pas une hypothèse sur une proportion, calculée sur une population existante* (les bonbons du distributeur, les élèves option Littérature, les Texans éligibles comme jurés), *mais une hypothèse sur une probabilité, inférée sur une population à venir* : les enfants à naître à Xicun.