Fluctuation d'échantillonnage

1 Expériences aléatoires

Définition : Une **expérience aléatoire** est un action dont le résultat ne peut être prévu avec certitude car le hasard intervient dans son déroulement.

Exemples :

1. Jeter un dé à 6 faces et observer le résultat apparaissant sur la face supérieure ;
2. Mesurer l'intervalle de temps entre deux rames de métro à la station Dupleix ;
3. Pratiquer une prise de sang sur une personne et mesure son taux de glycémie ;
4. Interroger un français choisi au hasard sur son intention de vote entre deux candidats X et Y au deuxième tour d'une élection présidentielle.

Bien que le hasard intervienne lors de l'exécution d'une expérience aléatoire, il est possible de déterminer l'ensemble des résultats possibles sur lesquels cette expérience aléatoire peut déboucher.

Définition : L'ensemble des résultats possibles lors de l'exécution d'une expérience aléatoire est appelé **univers**. Les éléments qui le composent sont les **issues** ou **éventualités**.

Exercice 1. Associer un univers Ω à chacune des expériences alétoires suivantes :

1. On jette un dé à 6 faces dont 2 bleues, 1 face est rouge et les 3 autres jaunes. On s'intéresse à la couleur de la face supérieure.
2. Sachant que l'intervalle de temps maximal entre deux rames de métro est de 10 min, déterminer l'ensemble des résultats possibles lorsque l'on mesure la durée s'écoulant entre le passage de deux rames à la station Dupleix.

Imaginons que l'on effectue maintenant 10 fois de suites l'expérience aléatoire suivante : on jette un dé à 6 faces.
Il est tout à fait raisonnable de supposer que le dé ne se modifie par lors de ces 10 lancers et que le résultat d'un lancer n'a aucune influence sur le résultat des 9 autres. On dit qu'on a effectué 10 expériences identiques et indépendantes.

Définition : Soit n un entier naturel non nul. Réaliser n expériences aléatoires **identiques et indépendantes** signifie que :

1. les n expériences aléatoires sont les mêmes et sont réalisées dans les mêmes conditions.
2. le résultat obtenu à l'une d'entre elles n'influe pas sur les résultats obtenus aux autres.

Simulation d'expériences aléatoires identiques et indépendantes : les calculatrices possède des fonctions qui permettent de simuler le hasard. Par exemple, l'appel de la fonction rand affiche un nombre au hasard de l'intervalle $]0;1[$.
La fonction randint(a,b) de la TI-82 STATS permet d'afficher au hasard un entier k tel que $a \leqslant k \leqslant b$ (a et b sont entiers. Si de plus on tape randint(a,b,N), où N est un entier, on obtient une liste de N nombres entiers appartenant à l'intervalle $[a;b]$.
Les appels successifs à ces fonctions sont indépendants.

Exercice 2. Simuler à la calculatrice 10 jets d'un dé équilibré à 6 faces et calculer la fréquence de l'occurence "6" à l'issue de ces 10 lancers.

Exercice 3. Soit $N \in \mathbb{N}^*$. Écrire un algorithme qui permette de simuler N jets d'un dé équilibré à 6 faces et qui indique à la fin la fréquence de l'occurence "6".
Programmer cet algorithme sur la calculatrice et le faire fonctionner pour différentes valeurs de N.
Quelles remarques cela-vous inspire-t-il ?

Exercice 4. Les biologistes ont constaté que dans l'espèce humaine, il naît 52 garçons pour 48 filles. Écrire un algorithme qui permette de simuler la naissance de N bébés qui indique à la fin la fréquence des garçons.
Programmer cet algorithme sur la calculatrice et le faire fonctionner pour différentes valeurs de N.

2 Qu'est-ce qu'un échantillon ?

On considère une population d'effectif $N \in \mathbb{N}^*$ sur laquelle on souhaite étudier un caractère statistique X.
Soit $n \in \mathbb{N}$, $n \leqslant N$.

Définition : Constituer un **échantillon aléatoire** de la population consiste à procéder de l'une ou l'autre des façons suivantes :

1. **[Échantillonnage "avec remise"]** On itère n fois les actions suivantes :
   on choisit **au hasard** un individu de la population ; on enregistre la valeur du caractère X pour cet individu ; on le remet dans la population.
   On obtient une liste de n résultats, que l'on appelle **échantillon de taille** n du caractère X.
2. **[Échantillonnage "sans remise"]** On choisit **au hasard et simultanément** n individus dans la population. On enregistre alors la valeur de X pour les n individus sélectionnés.
   On obtient une liste de n résultats, que l'on appelle **échantillon exhaustif de taille** n du caractère X.

Exercice 5. Parmi les deux types d'échantillonnage décrits ci-dessus, lequel correspond à la répétition de n expériences aléatoires identiques et indépendantes ?

Remarque – Les formules mathématiques que nous allons voir dans ce chapitre ne sont valables pour les échantillons non exhaustifs (ceux du type 1). Mais en pratique, lorsque la population étudiée est d'un effectif supérieur à 30, le statisticien ne fait pas de différence entre un échantillonnage "avec remise" et un échantillonnage "sans remise" : il choisit aléatoirement, et simultanément, dans la population, les n individus qui lui permettront de constituer son échantillon.
Attention ! Dans le langage courant, l'échantillon désigne aussi la partie de la population sur laquelle l'étude a porté.

Quel est l'intérêt de travailler sur un échantillon ? On souhaite effectuer une étude statistique sur une population d'effectif N. Une telle étude peut s'avérer impossible sur l'intégralité de la population pour deux raisons.

1. L'étude détruit l'individu : c'est le cas lorsque l'on s'intéresse par exemple, à la durée de vie des ampoules dans un lot ; si l'on teste toutes les ampoules pour obtenir la durée de vie moyenne de toutes les ampoules du lot, on n'aura plus d'ampoules à vendre ...
2. N est trop grand pour que l'étude statistique soit réalisable en un temps et avec un coût raisonnables : c'est le cas lorsque l'on s'intéresse à l'issue d'une élection présidentielle ; on ne peut en général pas interroger l'intégralité des votants sur leurs intentions ...

Le statisticien décide donc de faire porter son étude sur une **partie** de la population : il n'est pas nécessaire de manger tout le boeuf pour se rendre compte que la viande n'était pas tendre...

3 Fluctuation d'échantillonnage

Exemple – On lance un même dé dont les faces sont numérotées de 1 à 6, bien équilibré, et on repère le nombre qui apparaît sur la face supérieure. On répète ce lancer deux fois 100 fois.
On obtient ainsi deux échantillons A et B de taille 100 : on a consigné les fréquences d'apparition de chaque face dans un tableau :

Nombre	1	2	3	4	5	6
Fréquence A	$0,14$	$0,17$	$0,19$	$0,18$	$0,17$	$0,15$
Fréquence B	$0,15$	$0,16$	$0,16$	$0,18$	$0,17$	$0,18$

Sur ces deux échantillons de taille 100, les fréquences d'apparition des différents nombres ne sont pas les mêmes, bien qu'on ait lancé le même dé équilibré, dans des conditions identiques et indépendantes. Ce phénomène est appelé **fluctuation d'échantillonnage**.

Remarque – Désormais, on ne considérera que des expériences aléatoires à deux issues. De telles expériences sont appelées **épreuves de Bernoulli**.

Exercice 6. Simuler trois séries de 50 naissances et donner la fréquence f du nombre de garçons.
Mise en commun des résultats – Sur la classe, nous obtenons donc $3 \times 28 = 84$ fréquences. Quelle est la plus petite fréquence observée ? la plus grande ?

3.1 Intervalle de fluctuation d'une fréquence

Théorème [admis] et définition : Soit un caractère statistique dont la fréquence dans une population est p et n un entier naturel.
Si $n \geqslant 30$ et si $0,2 \leqslant p \leqslant 0,8$, alors dans 95% des échantillons de tailles n que l'on peut constituer aléatoirement à partir de la population, la fréquence observée f du caractère statistique étudié se trouve dans l'intervalle $\left[p - \dfrac{1}{\sqrt{n}}; p + \dfrac{1}{\sqrt{n}}\right]$.
Cet intervalle est appelé **intervalle de fluctuation** au seuil de 95% des fréquences observées.

Exercice 7. 26% des Français se déclarent allergiques aux pollens. On étudie la fréquence f des personnes allergiques dans un échantillon de taille 400.

1. Au seuil de 95%, dans quel intervalle f fluctue-t-elle ?
2. Sur un échantillon de taille 400, on relève 120 personnes allergiques : peut-on dire que le taux d'allergiques dans cet échantillon est anormalement élevé ?

Application : une règle de décision
Lorsque l'on effectue une échantillonnage aléatoire sur une population, il y a donc 95% de "chances" pour que la fréquence observée f d'un certain caractère dont la proportion est p dans la population soit dans l'intervalle de fluctuation défini ci-dessus. Mais la fluctuation d'échantillonnage peut expliquer que dans 5% des cas, cette fréquence observée "sorte" de cet intervalle. On met alors en place la règle de décision suivante :

- si la proportion f observée sort de l'intervalle de fluctuation prévu, on ne sera pas en mesure de valider l'hypothèse : "l'échantillon a été construit de façon aléatoire" et on la rejettera, en gardant à l'esprit que dans 5% des cas, c'est le hasard (et non un quelconque trucage) qui pourrait expliquer cette sortie de l'intervalle de fluctuation.
- si la proportion f observée est dans l'intervalle de fluctuation prévu, on ne sera pas en mesure de rejeter l'hypothèse : "l'échantillon a été construit de façon aléatoire" et on la validera donc, en gardant à l'esprit que même pour un échantillon "truqué", on pourrait observer une fréquence dans l'intervalle de fluctuation prévu quand on manipule des échantillons aléatoires.

3.2 Intervalle de confiance d'une proportion

Exercice 8. Soit f, p deux réels compris entre 0 et 1, et $n \in \mathbb{N}^*$. Prouver que l'équivalence suivante est vraie :

$$f \in \left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right] \iff p \in \left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}\right]$$

Application et définition : À partir d'une fréquence observée f dans un échantillon de taille n, on peut estimer la valeur d'une proportion inconnues p à l'aide de l'intervalle $\left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}\right]$, appelé **intervalle de confiance** au seuil de confiance 95% de la proportion p, ou encore au risque d'erreur de 5%.

Exercice 9. Lors d'un élection, un sondage portant sur un échantillon aléatoire de 1000 personnes donne 400 votants en faveur d'un candidat A. Au risque d'erreur de 5%, quelle information peut-on obtenir sur la proportion réel d'électeurs envisageant de voter pour A ?.