Chapitre VI
Échantillonages et simulations

Commentaires : Récursivement, les commentaires ne sont pas à l'attention des élèves.

1. Fluctuation d'échantillonnage

Définition 1 :

> *En statistiques, un* ***échantillon de taille n*** *est la liste des n résultats obtenus par n répétitions indépendantes de la même expérience.*

Exemple :

- On effectue deux séries de 100 lancés d'une pièce équilibre et on note la fréquence d'apparition du côté Pile. On obtient alors deux échantillons A et B de taille 100.
- Dans un groupe de 1000 personnes, on selectionne un individu, on note la couleur de ses cheveux et on le replace dans le groupe. En répetant 50 fois cette experience, on obtient un échantillon de taille 50.

Si l'on constitue deux échantillons de taille n pour une même experience, on constate que les distributions des fréquences des deux échantillons ne sont pas les mêmes : c'est ce qu'on appelle la **fluctuation d'échantillonnage**.

2. Intervalle de fluctuation

Commentaires :
Deux introductions possibles pour poser la problématique :

- À partir d'une expérience aléatoire, on réalise des échantillons de taille n. Pour chaque échantillon, on note la fréquence d'un caractère dont la probabilité est p. Les fréquences fluctuent autour de la probabilité. Cependant, sous certaines conditions, on peut encadrer cette fluctuation pour une majorité des échantillons.
- Soit une experience aléatoire pour laquelle la probabilité p d'un caractère est connue. On souhaite réaliser un échantillon de taille n de cette expérience. En raison de la fluctuation d'échantillonnage, la fréquence qui sera observée pour le caractère ne sera pas égale à la probabilité attendue. Cependant, sous certaines conditions et en admettant un certain risque, l'écart entre ces valeurs peut être "encadrer".

Ainsi la fréquence f calculée sur un échantillon de taille n diffère très souvent de la probabilité p d'apparition du caractère étudié. Sous certaines conditions, on peut encadrer cette fluctuation.

Théorème 1 :

Soit un échantillon de taille $n \geqslant 25$ sur lequel on observe un caractère de probabilité d'apparition p tel que $0,2 \leqslant p \leqslant 0,8$.
Si f désigne la fréquence du caractère observé dans l'échantillon, alors f appartient à l'intervalle $\left[p - \frac{1}{\sqrt{n}} \ ; \ p + \frac{1}{\sqrt{n}} \right]$ avec une probabilité d'au moins $0,95$.
Cet intervalle est appelé ***intervalle de fluctuation de taille n au seuil de 95%****.*

Application 1 : Prévision

Autrement dit, dans au moins 95% des cas, l'écart entre la fréquence observée et la probabilité est inférieur à $\frac{1}{\sqrt{n}}$. Ainsi, si je réalise un échantillon, je peux affirmer en prenant un risque de 5%, que la fréquence que j'obtiendrais sera dans l'intervalle de fluctuation.
La valeur $\frac{1}{\sqrt{n}}$ est appelé la ***marge d'erreur au seuil de 95%****.*

Commentaires : Attention, plus mon échantillon est grand, plus l'intervalle est précis mais le risque d'erreur reste à 5%.

Exemple : Roger prévoit de lancer 2500 fois une pièce de monnaie équilibrée. Puisque $n = 2500 \geqslant 25$ et $p = 0,5$ est compris entre $0,2$ et $0,8$, je peux parier **avant de réaliser l'expérience, et en acceptant un risque de 5%** que Roger aura entre une fréquence dans l'intervalle $\left[0.5 - \frac{1}{\sqrt{2500}}; 0.5 + \frac{1}{\sqrt{2500}}\right] = [0.48; 0.52]$.
Après l'expérience, Roger connait la valeur de f et peut savoir si ma prévision était correcte.

Application 2 : Prise de décision

Dans une population donnée, on étudie un caractère et on émet l'hypothèse :

La proportion du caractère est p

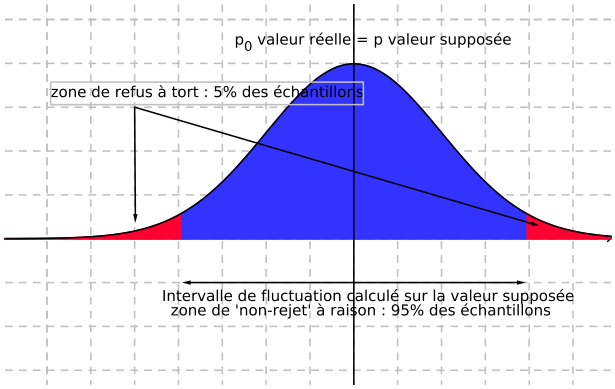
Pour juger de cette hyptothèse, on prélève un échantillon de taille n, et on calcule la fréquence, notée f, du caractère au sein de l'échantillon.

- *Si $f = p$, on ne rejette pas l'hyptothèse mais ce cas est peu fréquent en raison de la fluctuation d'échantillonnage.*
- *Si $f \neq p$, on regarde l'écart entre f et p :*
  - *Si cet écart (la marge d'erreur) est inférieur à $\frac{1}{\sqrt{n}}$ (i.e f est dans l'intervalle de fluctuation), on ne rejette pas l'hypothèse au motif que le hasard produirait un tel écart dans 95% des échantillons envisageables.*
  - *Si cet écart est supérieur à $\frac{1}{\sqrt{n}}$ (i.e f est hors l'intervalle de fluctuation), on rejette l'hyptothèse puisqu'un tel écart n'apparait que dans 5% des échantillons envisageables. Soit il s'agit vraiment un échantillon exceptionnel (et on a eu tort de rejeter l'hyptothèse) soit l'intervalle de fluctuation n'est pas correct, ce qui signifie que la proportion choisie n'était pas adaptée.*

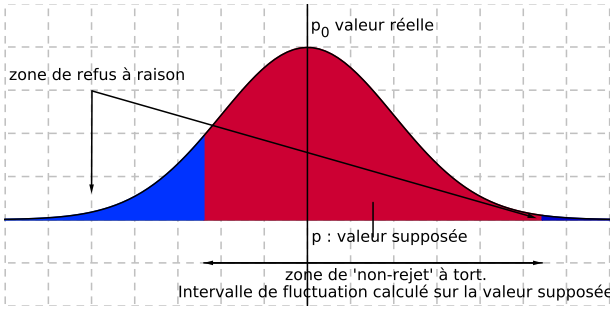
Exemple : Voir correction de l'exo 6.3 du livret

Commentaires :

- Attention, il y a une différence entre **accepter** une hypothèse et **ne pas la rejetter**. En fait, on conduit une sorte de raisonnement par l'absurde sur la valeur de la proportion p réelle. Si f est hors de l'intervalle de fluctuation, on a une "contradiction" puisque que cela devrait arriver que rarement (5% des cas). Mais si f est dans l'intervalle de fluctuation, on n'a aucune 'contradiction' ...ce qui ne prouve rien sur l'hypothèse émise. En effet, on peut très bien avoir le cas où la valeur supposée pour p est proche de celle la valeur réelle et ainsi l'intervalle de fluctuation supposé chevauche l'intervalle de fluctuation réel, créant ainsi une zone de non rejet à tort (voir commentaire suivant).
  En bref, si f est dans l'intervalle de flucutation, on rejette l'hypothèse sinon on ne la rejette pas car « on n'est pas sur qu'elle n'est pas fausse ».
- Pour la prise de décision, il y a deux erreurs possibles :
  - ◇ rejeter une hypothèse juste. La probabilité de cette erreur est de 5%. On parle d'**erreur de première espèce**.
    En effet, si l'hypothèse est correcte alors 95% des échantillons sont dans l'intervalle de fluctation et donc 5% des échantillons sont en dehors. Ce sont ceux là qui nous font rejeter l'hypothèse d'où une erreur estimée à 5%.



  - ◇ Ne pas rejeter une hypothèse fausse. La probabilité de cette erreur n'est pas connue. On parle d'**erreur de deuxième espèce**.
    En effet, si la probabilité supposée est fausse, alors l'intervalle de fluctuation est centré sur cette valeur (et non pas sur la vraie valeur). La probabilité de ne pas rejeter (à tort) l'hyptothèse est égale à l'aire du domaine rouge ... qui est non calculable de façon générale.



- Ci-dessous la correction de l'exos.
  Ayant lancé une pièce de monnaie, on dispose de deux échantillons : l'un de taille 100 (43 fois pile, soit 43% des lancers) et l'autre échantillon de taille 2500 (1150 fois pile soit 46% des lancers). Ces échantillons représentent-ils un modèle de lancer de pièce équilibrée ? (autrement dit, une probabilité de 0.5 a-t-elle pu simuler ces échantillons ?)
  Une première (fausse) idée est de dire que comme 46% est plus proche de 43%, l'échan-

tillon 2 représente plus le lancer d'une pièce équilibrée que l'échantillon 1.

En effet, supposons que la pièce soit équilibrée. Alors la probabilité d'apparition de la face Pile est 0.5 et l'intervalle de fluctuation pour le premier échantillon (de taille 100) est

$$\left[0.5 - \frac{1}{\sqrt{100}} \ ; \ 0.5 + \frac{1}{\sqrt{100}}\right] = [0.4; 0.6]$$

et le deuxième échantillon (de taille 2500) est

$$\left[0.5 - \frac{1}{\sqrt{2500}} \ ; \ 0.5 + \frac{1}{\sqrt{2500}}\right] = [0.48; 0.52]$$

La fréquence du premier échantillon est dans l'intervalle de fluctuation. Cet échantillon fait partie de ceux observables dans 95% des cas et donc on ne rejette pas qu'il puisse provenir du lancer d'une pièce équilibrée.
La fréquence du second échantillon n'est pas dans l'intervalle de fluctuation. Soit il fait donc partie de ceux observables dans 5% des cas (cas rare), soit l'intervalle de fluctuation (et donc la probabilité choisie pour le calculer) n'est pas correct. On rejette l'hypothèse que le second échantillon représente un modèle de pièce équilibrée.

3. Intervalle de confiance

L'objectif de cette section est de voir comment on peut estimer une proportion inconnue dans une population à partir d'un échantillon.

Sous les conditions du théorème, on a que si f est la fréquence d'un échantillon de taille n alors dans 95% des cas, on a

$$\begin{aligned} p - \frac{1}{\sqrt{n}} &\leqslant f \leqslant p + \frac{1}{\sqrt{n}} \\ -f - \frac{1}{\sqrt{n}} &\leqslant -p \leqslant -f + \frac{1}{\sqrt{n}} \\ f - \frac{1}{\sqrt{n}} &\leqslant p \leqslant f + \frac{1}{\sqrt{n}} \end{aligned}$$

On en déduit que, pour plus de 95% des échantillons aléatoires de taille n, la probabilité inconnue p appartient à l'intervalle $\left[f - \frac{1}{\sqrt{n}} \ ; \ f + \frac{1}{\sqrt{n}} \right]$, calculé à partir de l'échantillon.

Application 3 : Estimation

On adopte la procédure d'estimation suivante :
Pour estimer la probabilité d'un caractère au sein d'une population, on prélève UN échantillon aléatoire de taille n pour lequel on obtient UNE fréquence f. La probabilité p, inconnue, de la population mère est dans l'intervalle $\left[f - \frac{1}{\sqrt{n}} \ ; \ f + \frac{1}{\sqrt{n}} \right]$ *dans 95% des cas. Cet intervalle est appelé* ***intervalle de confiance de p au niveau de confiance 95%****.*

Remarques :

- L'intervalle de confiance ne correspond pas à une probabilité. Ce n'est pas par hasard que l'on parle de « confiance » et non de « probabilité ». On a un seul intervalle centré sur f obtenu après l'expérience aléatoire du tirage dans l'urne. Il n'y a plus de hasard : p est, ou non, dans l'intervalle de confiance (et on n'en sait rien).
  La probabilité est dans la fiabilité de la procédure (95% des cas) et non sur le seul intervalle que l'on connaisse. Donc ne pas dire que « p a 95% de chances d'être dans un intervalle de confiance donné » mais plutôt p est dans 95% des intervalles de confiance.
- Ne pas confondre (malgré la symétrie dangereuse de la formule) intervalle de fluctuation et intervalle de confiance. Il y a autant d'intervalles de confiance que d'échantillons. Ils sont déterminés à partir de la fréquence f de l'échantillon. Il y a un seul intervalle de fluctuation déterminé à partir de la probabilité.

Bilan :

Il y a deux démarches très différentes :

- *Soit on suppose une probabilité p pour le caractère étudié et dans ce cas on prélève un échantillon puis on utilise l'intervalle de fluctuation supposé au seuil de* 95% *pour juger l'hypothèse. Le nombre* $0,95$ *correspond dans un cas à une probabilité.*
- *Soit on ne connait pas p et dans ce cas, plutot que de baser sur une estimation ponctuelle, on calcule l'intervalle de confiance au seuil de* 95%*. On ne peut pas affecter une probabilité au fait que p appartienne à l'intervalle de confiance. C'est soit vrai, soit faux. Le nombre 0,95 n'a aucune interprétation en terme de probabilité. On parle de « niveau de confiance ».*