Échantillonnage

La presse présente très régulièrement des sondages accompagnés de pourcentages et de commentaires. Ces sondages sont-ils fiables ? Quelles notions sous-tendent-ils ?

Qu'est-ce qu'un intervalle de confiance, quel lien avec la fluctuation ?

Prenons le cas d'une population dont on veut connaître les intentions de vote, avant une élection. Il est de fait malaisé d'interroger l'ensemble des personnes concernées. On constitue alors un échantillon représentatif (le mot « représentatif » signifie que l'on va respecter les répartitions définies dans la population, comme, par exemple, le pourcentage d'hommes et de femmes, les tranches d'âge, etc.). On va ensuite étendre les résultats obtenus à partir de l'échantillon à toute la population.

L'expérience montre que, lorsque l'on choisit un autre échantillon représentatif, on obtient des résultats assez proches mais pas exactement les mêmes. Aussi, pour avoir une meilleure approximation du résultat, va-t-on donner un intervalle plutôt qu'un nombre. Si on reprend l'exemple de l'élection, supposons qu'à partir du sondage réalisé sur l'échantillon, un candidat obtienne 45 % des intentions de vote. À partir de ce résultat, dans quel intervalle se situent les intentions de vote de la population ?

Cet intervalle s'appelle « intervalle de confiance », afin de limiter les effets de la fluctuation d'échantillonnage.

Que signifie le terme « au seuil de 95 % de la fréquence » ?

Le pourcentage de 95 % détermine la marge d'erreur. Ici, le risque est de 5 %. La phrase « au seuil de 95 % en fréquence » signifie donc « avec une marge d'erreur inférieure à 5 % ». Le seuil de 5 % est le plus utilisé, mais on peut très bien définir un autre seuil.

Intervalle de fluctuation asymptotique au seuil de 95 % de la fréquence

Soit X une variable aléatoire qui suit la loi binomiale $B(n\ ;\ p)$ avec $0 < p < 1$, $n \geqslant 30$, $np > 5$ et $n(1-p) > 5$.

On appelle **intervalle de fluctuation asymptotique au seuil de 95 % de la fréquence** l'intervalle :

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

Contrairement à la fréquence f de l'intervalle de confiance, la proportion p est ici déjà connue.

On utilise la **loi binomiale** $B(n\ ;\ p)$ car on renouvelle n fois de manière indépendante une épreuve de Bernoulli de paramètre p.

Intervalle de confiance

Il s'agit de savoir comment estimer la proportion p d'individus d'une population ayant une propriété, à partir de la fréquence f observée sur un échantillon : on utilise un intervalle de confiance.

Définition : en utilisant les notations du point précédent, on appelle **intervalle de confiance de la proportion p avec un niveau de confiance de 95 %**, l'intervalle $\left[f - \frac{1}{\sqrt{n}}\ ;\ f + \frac{1}{\sqrt{n}}\right]$ où n est la taille de l'échantillon.

Méthode : on considère une population et un échantillon de taille n de cette population. À partir de l'échantillon, on calcule la fréquence f des individus ayant une propriété. La proportion p des individus de la population ayant la propriété appartient à l'intervalle de confiance, avec un niveau de confiance de 95 %, qui est : $\left[f - \frac{1}{\sqrt{n}}\ ;\ f + \frac{1}{\sqrt{n}}\right]$.

UN ARTICLE DU *MONDE* À CONSULTER

• **Avec la méthode française, la marge d'erreur ne peut pas être calculée mathématiquement** p. 58
(Pierre Le Hir, *Le Monde* daté du 17.03.2002)

MOTS CLÉS

ÉCHANTILLON
En statistique, la population est l'ensemble sur lequel on étudie une série statistique. Un échantillon est une partie (un sous-ensemble) de la population.

FRÉQUENCE
En statistique, la fréquence d'une valeur est le quotient :

$$\frac{\text{effectif de la valeur}}{\text{taille de la population}}.$$

On l'exprime sous la forme d'un pourcentage ou d'un nombre décimal.

SIMULATION
• Simuler une expérience aléatoire consiste à produire une liste de n résultats (à l'aide de la touche RANDOM de la calculatrice par exemple) que l'on peut assimiler (ou faire correspondre) à n résultats de l'expérience. On a ainsi produit un échantillon de taille n de l'expérience.

• Entre deux simulations, ou entre deux échantillons, les distributions de fréquences varient, c'est ce que l'on appelle la fluctuation d'échantillonnage.

INTERVALLE DE FLUCTUATION
Pour une variable aléatoire X qui suit la loi binomiale $B(n\ ;\ p)$ avec $0 < p < 1$, $n \geqslant 30$, $np > 5$ et $n(1-p) > 5$, on appelle intervalle de fluctuation asymptotique au seuil de 95 % de la fréquence l'intervalle :

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

INTERVALLE DE CONFIANCE
• Si f est la fréquence obtenue avec un échantillon de taille n, un intervalle de confiance à un niveau de confiance de 0,95 est

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

• Pour un échantillon de taille n, l'amplitude de cet intervalle de confiance est $\frac{2}{\sqrt{n}}$.

Amérique du Nord (mai 2013)

Une boulangerie industrielle utilise une machine pour fabriquer des pains de campagne pesant en moyenne 400 grammes.
Pour être vendus aux clients, ces pains doivent peser au moins 385 grammes.
Un pain dont la masse est strictement inférieure à 385 grammes est un pain non commercialisable ; un pain dont la masse est supérieure ou égale à 385 grammes est commercialisable.
La masse d'un pain fabriqué par la machine peut être modélisée par une variable aléatoire X suivant la loi normale d'espérance $\mu = 400$ et d'écart type $\sigma = 11$.
Les probabilités seront arrondies au millième le plus proche.
(Les parties A et B peuvent être traitées indépendamment les unes des autres.)

Partie A

On pourra utiliser le tableau suivant dans lequel les valeurs sont arrondies au millième le plus proche.

x	380	385	390	395	400
$P(X \leqslant x)$	0,035	0,086	0,182	0,325	0,5

x	405	410	415	420
$P(X \leqslant x)$	0,675	0,818	0,914	0,965

1. Calculer $P(390 \leqslant X \leqslant 410)$.
2. Calculer la probabilité p qu'un pain choisi au hasard dans la production soit commercialisable.
3. Le fabricant trouve cette probabilité p trop faible. Il décide de modifier ses méthodes de production afin de faire varier la valeur de σ sans modifier celle de μ.
Pour quelle valeur de σ la probabilité qu'un pain soit commercialisable est-elle égale à 96 % ? (On arrondira le résultat au dixième.)
On pourra utiliser le résultat suivant : lorsque Z est une variable aléatoire qui suit la loi normale d'espérance 0 et d'écart type 1, on a $P(Z \leqslant -1{,}751) \approx 0{,}040$.

Partie B

Les méthodes de production ont été modifiées dans le but d'obtenir 96 % de pains commercialisables.
Afin d'évaluer l'efficacité de ces modifications, on effectue un contrôle qualité sur un échantillon de 300 pains fabriqués.
1. Déterminer l'intervalle de fluctuation asymptotique au seuil de 95 % de la proportion de pains commercialisables dans un échantillon de taille 300.
2. Parmi les 300 pains de l'échantillon, 283 sont commercialisables. Au regard de l'intervalle de fluctuation obtenu à la question **1.**, peut-on décider que l'objectif a été atteint ?

La bonne méthode

Partie A

1. Utiliser le tableau et le fait que si X est une variable aléatoire suivant une loi continue :
$P(a \leqslant X \leqslant b) = P(X \leqslant b) - P(X \leqslant a)$.
2. Traduire à l'aide d'une variable aléatoire et d'une probabilité le fait qu'un pain choisi au hasard dans la production soit commercialisable.
3. Traduire l'énoncé à l'aide d'une variable aléatoire et d'une probabilité, puis centrer et réduire. Utiliser la valeur donnée dans l'énoncé.

Partie B

1. Utiliser les données de l'énoncé pour déterminer les bornes de l'intervalle de fluctuation.
2. Calculer la fréquence observable de l'échantillon et vérifier si elle appartient ou non à l'intervalle de fluctuation précédemment déterminé.