ÉCHANTILLONNAGE ET ESTIMATION

1) INTRODUCTION

Pour étudier une population statistique, on a recours à deux méthodes :

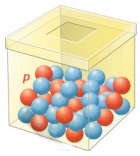
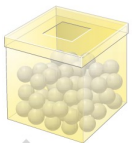
- la méthode exhaustive (ou recensement) : on examine chacun des éléments de la population. En général, cette méthode est jugée trop longue.
- la méthode des sondages : on n'examine qu'une partie de la population pour essayer d'en déduire des informations sur la totalité de la population. Cette méthode comprend deux parties :
  - l'échantillonnage qui permet de passer de la population totale à une partie seulement de cette population (l'échantillon).
  - l'estimation permet d'induire, à partir des résultats observés sur l'échantillon, des informations sur la population totale.

Remarque :
Ces domaines appartiennent au champ des statistiques « inférentielles ».

Ce que dit « wikipedia » :
L'inférence statistique consiste à **induire** *les caractéristiques inconnues d'une* **population** *à partir d'un* **échantillon** *issu de cette population. Les caractéristiques de l'échantillon, une fois connues, reflètent avec une certaine* **marge d'erreur** *possible celles de la population . Strictement, l'inférence s'applique à l'ensemble des membres (pris comme un tout) de la population représentée par l'échantillon, et non pas à tel ou tel membre particulier de cette population .*
Par exemple, les intentions de vote indiquées par l'échantillon, ne peuvent révéler l'intention de vote qu'a tel ou tel membre particulier de la population des électeurs de la circonscription électorale.
L'inférence statistique est donc un ensemble de méthodes permettant de tirer des conclusions fiables à partir de données d'échantillons statistiques .
L'interprétation de données statistiques est, pour une large part, le point clé de l'inférence statistique. Elle est guidée par plusieurs principes et axiomes.

A) IDENTIFICATION DE LA SITUATION

On considère deux urnes U_1 et U_2 contenant chacune un très grand nombre de boules rouges ou bleues.

Domaine de l'échantillonnage	Domaine de l'estimation
Dans l'urne U_1, on connaît la proportion p de boules rouges. 	Dans l'urne U_2, on ignore la proportion de boules rouges. 
On procède à des tirages avec remise de n boules, et on observe la fréquence d'apparition d'une boule rouge. Cette fréquence observée appartient « en général » à **un intervalle de fluctuation** de centre p, dont l'amplitude diminue avec n.	En procédant à des tirages avec remise de n boules, on va essayer d'estimer la proportion p de boules rouges dans l'urne, proportion inconnue à priori. Cette estimation se fait à l'aide d'**un intervalle de confiance**. Cet intervalle dépend d'un coefficient, le niveau de confiance, que l'on attribue à l'estimation.

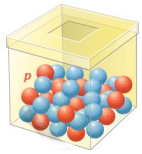
B) QUEL INTERVALLE UTILISER ?

On s'intéresse à une population dont on étudie un caractère particulier.

Échantillonnage - intervalle de fluctuation	Estimation - intervalle de confiance
• On connaît la proportion p de présence du caractère dans la population. • On fait une hypothèse sur la valeur de cette proportion (on est alors dans la « prise de décision »)	On ignore la valeur de la proportion p de présence du caractère dans la population, et on ne formule pas d'hypothèse sur cette valeur.
Exemple : On dispose d'une pièce de monnaie. Comment décider qu'elle est équilibrée ou pas ? On va faire ici l'hypothèse que la fréquence d'apparition de Pile, par exemple, est égale à 0,5, et on va tester cette hypothèse.	**Exemple :** Une usine fabrique des fusées de feux d'artifices . Sur 100 fusées choisies au hasard à l'issue du processus de fabrication et mises à feu, on trouve 12 fusées qui ne fonctionnent pas . Comment se faire une idée de la proportion des fusées défectueuses dans la production ? On n'a, au départ, aucune idée de la valeur de la proportion étudiée dans la production.

Remarque : *Quand l'estimation est obligatoire …*
On ne peut pas allumer toutes les fusées en sortie de production pour vérifier leur bon fonctionnement …
Il est donc impossible de faire une étude exhaustive sur toute la population.

2) ÉCHANTILLONNAGE - INTERVALLE DE FLUCTUATION



On dispose d'une urne contenant un très grand nombre N de boules rouges et bleues.
On sait que la proportion de boules rouges dans l'urne est égale à $p = 0{,}4$.

On tire successivement avec remise, n boules dans l'urne $(n \in \mathbb{N}^*)$. On obtient ainsi un échantillon de taille n.
On note f la proportion de boules rouges dans l'échantillon.

Propriété et définition :

Quand on prélève un échantillon de taille n dans une population qui contient une proportion p du caractère étudié, alors la proportion f du caractère dans l'échantillon appartient à l'intervalle $I_n = \left[p - 1{,}96 \dfrac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 1{,}96 \dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$ avec une probabilité de 95%.

Cet intervalle est appelé **intervalle de fluctuation** à 95% . (On dit aussi intervalle de fluctuation asymptotique)

Certaines conditions sont requises pour que l'intervalle de fluctuation soit valide : $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$.

Exemple :

On tire 50 boules de l'urne décrite ci-dessus, et on souhaite déterminer un intervalle de fluctuation à 95%.

On obtient pour intervalle de fluctuation :

$$I_{50} = \left[0{,}4 - 1{,}96 \frac{\sqrt{0{,}4 \times 0{,}6}}{\sqrt{50}} ; 0{,}4 + 1{,}96 \frac{\sqrt{0{,}4 \times 0{,}6}}{\sqrt{50}} \right] \approx [0{,}264 ; 0{,}536]$$

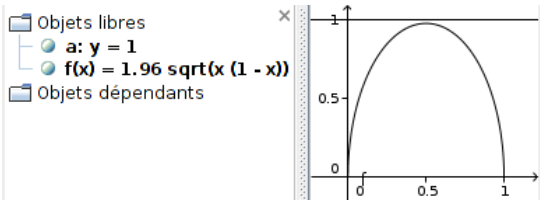
Ainsi en effectuant 50 tirages dans cette urne, la fréquence d'apparition d'une boule rouge est comprise entre 0,264 et 0,536 avec une probabilité d'environ 0,95.

Pour 500 tirages, au même seuil 95% , on obtient $I_{500} \approx [0{,}357 ; 0{,}443]$.

L'amplitude de l'intervalle a, pour un même seuil, été divisée par plus de 3 en passant de 50 à 500 tirages …

Remarques :

- Plus la taille de l'échantillon est importante, plus l'intervalle de fluctuation est resséré autour de p.
- L'intervalle ne dépend pas de la taille N de la population initiale.
- En majorant $1{,}96\sqrt{p(1-p)}$ par 1 , on retrouve l'intervalle de fluctuation présenté en classe de seconde $\left[p - \dfrac{1}{\sqrt{n}} ; p + \dfrac{1}{\sqrt{n}} \right]$.



La proportion p dans la population est connue et la fréquence f aussi . Ce qui est inconnue, c'est si l'échantillon provient bien de la population.
On fait l'hypothèse qu'il provient bien de la population.

Propriété :

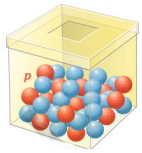
Si f n'appartient pas à l'intervalle $I_n = \left[p - 1{,}96 \dfrac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 1{,}96 \dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$, on rejette l'hypothèse au seuil 95%.
Dans le cas contraire, on l'accepte au seuil de 95%.

Exemple :

Dans l'urne ci-dessus, on réalise un tirage de 50 boules ; on obtient 23 boules rouges.

$f = \dfrac{23}{50} \approx 0{,}46$. On a $0{,}46 \in I_{50}$. .
On peut donc estimer au seuil de 95 % que l'échantillon provient bien de l'urne.

3) ESTIMATION - INTERVALLE DE CONFIANCE



On dispose d'une urne contenant un très grand nombre de boules rouges et bleues.
On ignore quelle est la proportion p de boules rouges dans l'urne et rien ne permet de faire une hypothèse sur la valeur de p.
L'estimation consiste à deviner, avec un certain niveau de confiance, quelle valeur peut prendre p, en s'appuyant sur les informations recueillies en procédant à des tirages au sort aléatoires.

Propriété :

Quand un échantillon de taille n contient une proportion f du caractère étudié, alors la proportion p du caractère dans la population appartient à l'intervalle $\left[f - 1{,}96\frac{\sqrt{f(1-f)}}{\sqrt{n}} ; f + 1{,}96\frac{\sqrt{f(1-f)}}{\sqrt{n}} \right]$ avec une probabilité de 95%.

Cet intervalle est appelé **intervalle de confiance** à 95%.

Certaines conditions sont requises pour que l'intervalle de confiance soit valide : $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$.

Remarque :
Le niveau de confiance 95%, signifie que si l'on effectuait un très grand nombre de tirages de 100 boules, on devrait obtenir moins de 5% d'intervalles de confiance ne contenant pas la proportion p de boules rouges.

Exemple :

Dans l'urne ci-dessus, on réalise un tirage de 100 boules ; on obtient 59 rouges et 41 bleues.
La fréquence observée de sortie du rouge est donc 0,59.

L'intervalle $\left[0{,}59 - 1{,}96\frac{\sqrt{0{,}59\times 0{,}41}}{\sqrt{100}} ; 0{,}59 + 1{,}96\frac{\sqrt{0{,}59\times 0{,}41}}{\sqrt{100}} \right] \approx [0{,}494 ; 0{,}686]$ est un intervalle de confiance de la proportion de boules rouges dans l'urne au niveau de confiance 95%.

Dans de nombreuses situations, il est intéressant de comparer des échantillons dans lesquels on a pu observer des proportions f_1 et f_2 d'un caractère pour savoir s'ils viennent de la même population . On fait alors l'hypothèse qu'ils viennent effectivement de la même population et on explicite les deux intervalles de confiance.

Propriété :

Si les deux intervalles de confiances à 95 % sont disjoints, alors on rejette l'hypothèse au seuil 95%.
Dans le cas contraire, on l'accepte au seuil de 95%.

Exemple :

Dans l'urne ci-dessus, on réalise un autre tirage de 100 boules ; on obtient 39 rouges et 61 bleues.
La fréquence observée de sortie du rouge est donc 0,39.

L'intervalle $\left[0{,}39 - 1{,}96\frac{\sqrt{0{,}39\times 0{,}61}}{\sqrt{100}} ; 0{,}39 + 1{,}96\frac{\sqrt{0{,}39\times 0{,}61}}{\sqrt{100}} \right] \approx [0{,}294 ; 0{,}486]$ est un intervalle de confiance de la proportion de boules rouges dans l'urne au niveau de confiance 95%.

Cet intervalle et l'intervalle de l'exemple précédent sont disjoints . On rejette donc l'hypothèse que les deux échantillons viennent de la même urne.