ÉCHANTILLONNAGE ET ESTIMATION

1) INTRODUCTION

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

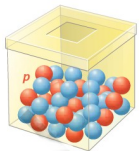
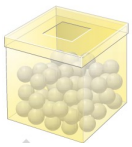
- la méthode exhaustive (ou recensement) : on examine chacun des éléments de la population. En général, cette méthode est jugée trop longue.
- la méthode des sondages : on n'examine qu'une partie de la population pour essayer d'en déduire des informations sur la totalité de la population. Cette méthode 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.

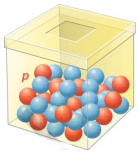
B) QUEL INTERVALLE UTILISER ?

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 ASYMPTOTIQUE



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

Si on tire successivement avec remise, n boules dans l'urne $(n\in\mathbb{N}^*)$, et si on appelle X_n la variable aléatoire dénombrant les boules rouges tirées, alors X_n suit une loi binomiale $B(n\,;\,p)$

Théorème et définition :

Soit X_n une variable aléatoire suivant une loi binomiale $B(n\,;\,p)$, et α un réel tel que $0<\alpha<1$.
Si X est une variable aléatoire suivant la loi normale centrée réduite $N(0\,;\,1)$, on appelle u_α l'unique réel tel que :

$$P(-u_\alpha\leqslant X\leqslant u_\alpha)=1-\alpha$$

On appelle I_n l'intervalle :

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

Alors :

L'intervalle I_n contient la fréquence $F_n=\dfrac{X_n}{n}$ avec une probabilité qui se rapproche de $1-\alpha$ lorsque n augmente.

On dit que c'est un intervalle de fluctuation asymptotique de F_n au seuil $1-\alpha$.

Preuve : *exigible*

On pose $Z_n=\dfrac{X_n-np}{\sqrt{np(1-p)}}$ et on applique le théorème de Moivre-Laplace.

Si X suit la loi normale $N(0\,;\,1)$: $\lim\limits_{n\to+\infty} P(Z_n\in[-u_\alpha\,;\,u_\alpha])=P(X\in[-u_\alpha\,;\,u_\alpha])=1-\alpha$

Or $Z_n\in[-u_\alpha\,;\,u_\alpha]\Leftrightarrow -u_\alpha\leqslant\dfrac{X_n-np}{\sqrt{np(1-p)}}\leqslant u_\alpha$

Donc $Z_n\in[-u_\alpha\,;\,u_\alpha]\Leftrightarrow np-u_\alpha\sqrt{np(1-p)}\leqslant X_n\leqslant np+u_\alpha\sqrt{np(1-p)}\Leftrightarrow p-u_\alpha\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\leqslant\dfrac{X_n}{n}\leqslant p+u_\alpha\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}$

D'où $Z_n\in[-u_\alpha\,;\,u_\alpha]\Leftrightarrow\dfrac{X_n}{n}\in I_n$

Exemple :

On tire 50 boules de l'urne décrite ci-dessus, et on souhaite déterminer un intervalle de fluctuation au seuil 0,9 (c'est à dire avec $\alpha=0{,}1$)
A l'aide de la calculatrice, on trouve $u_{0{,}1}\approx 1{,}645$ (à 10^{-3} près)
On obtient pour intervalle de fluctuation :

$$I_{50}=\left[0{,}4-u_{0{,}1}\frac{\sqrt{0{,}4\times 0{,}6}}{\sqrt{50}}\,;0{,}4+u_{0{,}1}\frac{\sqrt{0{,}4\times 0{,}6}}{\sqrt{50}}\right]\approx[0{,}286\,;\,0{,}514]$$

Ainsi en effectuant 50 tirages dans cette urne, la fréquence d'apparition d'une boule rouge est comprise entre 0,286 et 0,514 avec une probabilité d'environ 0,9.
Pour 500 tirages, au même seuil 0,9, on obtient $I_{500}\approx[0{,}364\,;\,0{,}436]$.
L'amplitude de l'intervalle a, pour un même seuil, été divisée par plus de 3 en passant de 50 à 500 tirages …

Remarque :
On a déjà vu que pour la loi normale, $u_{0{,}05}=1{,}96$. On en déduit la propriété ci-dessous :

Propriété :

Pour une variable aléatoire X_n suivant une loi binomiale $B(n\,;\,p)$, l'intervalle de fluctuation asymptotique au seuil de 95% est :

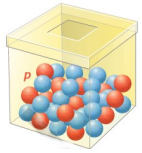
$$I_n=\left[p-1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\,;\,p+1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$$

Remarques :

- Un intervalle de fluctuation asymptotique au seuil $1-\alpha$ correspond à une approximation . On ne connaît pas les termes de la suite (p_n) où $p_n=P\left(\dfrac{X_n}{n}\in I_n\right)$. On sait toutefois que (p_n) converge vers $1-\alpha$, et on considère que la limite $1-\alpha$ est une valeur approchée de p_n avec les conditions $n\geqslant 30$, $np\geqslant 5$ et $n(1-p)\geqslant 5$.
- En majorant $1{,}96\sqrt{p(1-p)}$, on retrouve l'intervalle de fluctuation présenté en classe de seconde. *(cf : démonstration suivante)*

3) ESTIMATION

A) 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.

Théorème :

Soit X_n une variable aléatoire suivant une loi binomiale $B(n\,;\,p)$, où p est la proportion inconnue d'apparition d'un caractère, et $F_n=\dfrac{X_n}{n}$ la fréquence associée à X_n.

Alors, pour n suffisamment grand, $p\in\left[F_n-\dfrac{1}{\sqrt{n}}\,;F_n+\dfrac{1}{\sqrt{n}}\right]$ avec une probabilité supérieure ou égale à 0,95.

Preuve : *exigible*

Soit la variable aléatoire $Z_n=\dfrac{X_n-np}{\sqrt{np(1-p)}}$, et la suite a définie pour tout entier $n\geqslant 1$, par $a_n=P(-2\leqslant Z_n\leqslant 2)$.

D'après le théorème de Moivre-Laplace, la suite a converge vers l avec :
$l=P(-2\leqslant Z\leqslant 2)$ où Z suit la loi normale $N(0\,;1)$.

Or, on a : $l\geqslant 0{,}9544$ (avec la calculatrice $l\approx 0{,}9545$)
Soit un réel ε tel que : $0<\varepsilon<0{,}004$ (ainsi $l-\varepsilon\geqslant 0{,}95$).
Par définition de la convergence vers l, il existe un entier naturel n_0, tel que : si $n\in\mathbb{N}$ et $n\geqslant n_0$, alors $a_n\in]l-\varepsilon\,;l+\varepsilon[$

Ainsi pour $n\geqslant n_o$: $a_n\geqslant 0{,}95$. Comme dans la démonstration précédente, on a :

$$P(-2\leqslant Z_n\leqslant 2)\geqslant 0{,}95\Leftrightarrow P\left(p-2\frac{\sqrt{p(1-p)}}{\sqrt{n}}\leqslant F_n\leqslant p+2\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)\geqslant 0{,}95$$

L'étude de la fonction $p\longmapsto p(1-p)$ sur l'intervalle $]0\,;1[$ permet de majorer $p(1-p)$ par son maximum $\dfrac{1}{4}$ sur $]0\,;1[$.

On en déduit que :

$$\sqrt{p(1-p)}\leqslant\frac{1}{2}\text{ et }2\frac{\sqrt{p(1-p)}}{\sqrt{n}}\leqslant\frac{1}{\sqrt{n}}\ (1)$$

Cette majoration a pour effet “d'agrandir” l'intervalle, donc d'augmenter sa probabilité . On obtient finalement :

$$\text{Pour } n\geqslant n_0\ ,\quad P\left(p-\frac{1}{\sqrt{n}}\leqslant F_n\leqslant p+\frac{1}{\sqrt{n}}\right)\geqslant P\left(p-2\frac{\sqrt{p(1-p)}}{\sqrt{n}}\leqslant F_n\leqslant p+2\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)\geqslant 0{,}95$$

$$\text{On a }\left(p-\frac{1}{\sqrt{n}}\leqslant F_n\leqslant p+\frac{1}{\sqrt{n}}\right)\Leftrightarrow\left(F_n-\frac{1}{\sqrt{n}}\leqslant p\leqslant F_n+\frac{1}{\sqrt{n}}\right)$$

Grâce à la majoration (1), on peut ecrire que pour tout entier $n\geqslant n_0$, $P\left(p-\dfrac{1}{\sqrt{n}}\leqslant F_n\leqslant p+\dfrac{1}{\sqrt{n}}\right)\geqslant 0{,}95$.

On en déduit que pour tout entier $n\geqslant n_0$, $P\left(F_n-\dfrac{1}{\sqrt{n}}\leqslant p\leqslant F_n+\dfrac{1}{\sqrt{n}}\right)\geqslant 0{,}95$.

Définition :

On réalise l'expérience aléatoire de n tirages au hasard, et on appelle f la fréquence observée d'apparition du caractère.

L'intervalle $\left[f-\dfrac{1}{\sqrt{n}}\,;\,f+\dfrac{1}{\sqrt{n}}\right]$ est appelé intervalle de confiance de p au niveau de confiance 0,95, où p est la proportion (inconnue) d'apparition du caractère dans la population.

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 $\left[0{,}59-\dfrac{1}{\sqrt{100}}\,;0{,}59+\dfrac{1}{\sqrt{100}}\right]=[0{,}49\,;0{,}69]$ est un intervalle de confiance de la proportion de boules rouges dans l'urne au niveau de confiance 95%.

B) PRÉCISION D'UNE ESTIMATION ET TAILLE DE L'ÉCHANTILLON

On a vu ci-dessus qu'en tirant 100 boules de l'urne, l'intervalle de confiance obtenu est de longueur 0,2 ; on peut trouver cet intervalle trop grand.
En procédant à un tirage de 400 boules, si f est la fréquence observée de sortie du rouge, on obtient l'intervalle de confiance au niveau 95% égal à :

$$\left[f-\frac{1}{\sqrt{400}}\,;\ f+\frac{1}{\sqrt{400}}\right]=\left[f-0{,}05\,;\ f+0{,}05\right]$$

Son amplitude, deux fois moindre que la précédente, est de 0,1.

On retient :

Propriété :

Un intervalle de confiance au niveau 95 % est d'amplitude $\frac{2}{\sqrt{n}}$.
Plus la taille de l'échantillon est grande, plus les intervalles de confiance obtenus sont précis.

Exemple :

Pour obtenir un intervalle de confiance d'amplitude inférieur à 0,01 de la proportion de boules rouges dans l'urne, il faut procéder à des tirages de n boules, avec :

$$\frac{2}{\sqrt{n}}\leqslant 0{,}01\Leftrightarrow\frac{4}{n}\leqslant 10^{-4}\Leftrightarrow n\geqslant 4\times 10^{4}$$

Ainsi, pour obtenir des intervalles de confiance au niveau 0,95 d'amplitude inférieur à 0,01, il faut procéder à au moins 40000 tirages ...