

Échantillonnage

1. Deux cas : échantillonnage et estimation

Exemple :

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

1) Deux cas se présentent

a) Dans U_1 , on connaît la proportion p de boules rouges.

- On procède à des tirages successifs 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 et dont l'amplitude diminue quand n augmente.
- On est ici dans le domaine de l'**échantillonnage**.

b) Dans U_2 , on ignore la proportion p de boules rouges.

- En procédant à des tirages avec remise de n boules, on va essayer d'estimer la proportion p de boules rouges dans l'urne. Cette estimation se fait à l'aide d'un **intervalle de confiance**.
- Celui-ci dépend d'un coefficient appelé **niveau de confiance** que l'on attribue à l'estimation.
- On est ici dans le cas de l'**estimation**.

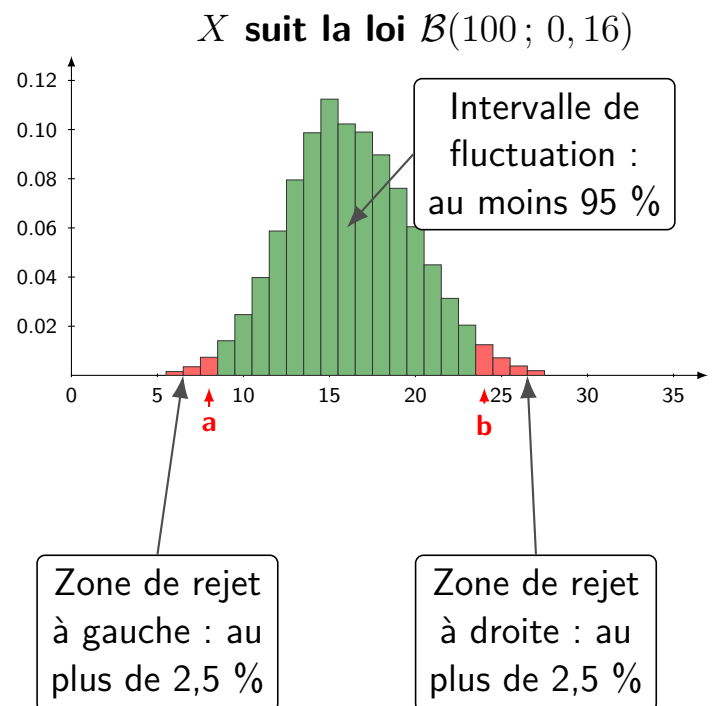
2) Quel intervalle choisir ?

a) Échantillonnage

On utilise un **intervalle de fluctuation** quand on connaît la proportion p de présence du caractère dans la population ou quand on peut faire une hypothèse sur la valeur de cette proportion.

Exemple :

Sous quelles conditions peut-on décider qu'une pièce de monnaie est ou non « équilibrée » ? On fait l'hypothèse que la fréquence d'apparition de « pile » est 0,5 et on « teste » cette hypothèse : c'est une situation d'échantillonnage qui mène à une prise de décision.



b) Estimation

On utilise un **intervalle de confiance** quand on ne connaît pas la valeur de la proportion p de présence du caractère dans la population.

Exemple :

Une usine fabrique des feux d'artifice. Sur 100 fusées choisies au hasard, on en trouve 12 défectueuses. Comment se faire une idée de la proportion de fusées défectueuses ? : c'est une situation d'estimation.

Remarque : Cette estimation est nécessaire, car on ne peut pas tester toutes les fusées !

2. Premier cas : Échantillonnage

■ PROPRIÉTÉ :

Si la taille n de l'échantillon est supérieure ou égale à 25 et si p appartient à l'intervalle $[0,2; 0,8]$, alors l'intervalle $\left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$ constitue un **intervalle de fluctuation au seuil de 95 %**.

Exemple :

Dans un casino, sur 1 000 lancers de dés, 465 ont donné un nombre pair.

1. Quelle est la proportion p théorique de résultats « pairs » ?
2. On réalise donc ici 1000 lancers, ainsi les conditions de la propriété précédente sont vérifiées. Déterminer alors l'intervalle de fluctuation au seuil de 95%.
3. Faut-il prévoir une enquête pour des soupçons d'utilisation de dés truqués ?

■ PROPRIÉTÉ :

Si le nombre X d'individus de l'échantillon possédant un caractère donné suit une loi binomiale de paramètres n et p , et si a et b sont des entiers tels que :

- a est le plus grand entier vérifiant

$$P(X \leq a) \leq 0,025$$

- b est le plus petit entier vérifiant

$$P(X \leq b) \geq 0,975$$

alors $P(a \leq X \leq b) \geq 0,95$ et l'intervalle $\left[\frac{a}{n}; \frac{b}{n}\right]$ est appelé **intervalle de fluctuation au seuil de 95 %**.

Remarques :

- $P(a \leq X \leq b) \geq 0,95$
 $\iff P\left(\frac{a}{n} \leq \frac{X}{n} \leq \frac{b}{n}\right) \geq 0,95$
- L'intervalle $\left[\frac{a}{n}; \frac{b}{n}\right]$ est « presque » centré sur p dès que n est « assez grand »
- Pour n « assez grand » et p dans $[0,2; 0,8]$, l'intervalle $\left[\frac{a}{n}; \frac{b}{n}\right]$ est « presque » le même que $\left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$

Exemple :

Un médecin de santé publique veut savoir si, dans sa région, le pourcentage d'habitants atteints d'une certaine maladie est égal à la valeur de 16 % récemment publiée pour des populations semblables.

En notant p la proportion de malades dans la population de sa région, le médecin formule l'hypothèse $p = 0,16$.

Pour vérifier cette hypothèse, le médecin constituera un échantillon de $n = 100$ habitants de la région ; il déterminera la fréquence f de malades (l'échantillon est prélevé au hasard et la population est suffisamment importante pour considérer qu'il s'agit de tirages avec remise).

Lorsque la proportion dans la population vaut $p = 0,16$, la variable aléatoire X correspondant au nombre de malades observé dans un échantillon aléatoire de taille $n = 100$, suit la loi binomiale de paramètres $n = 100$ et $p = 0,16$.

On cherche à partager l'intervalle $[0; 100]$, où X prend ses valeurs, en trois intervalles $[0; a]$, $[a; b]$ et $[b; 100]$ de sorte que la variable aléatoire X prenne ses valeurs dans chacun des intervalles extrêmes avec une probabilité proche de 0,025, sans dépasser cette valeur.

On recherche donc :

- le plus grand entier a tel que $P(X \leq a) \leq 0,025$
- le plus petit entier b vérifiant $P(X \leq b) \geq 0,975$

La calculatrice fournit $a = 9$ et $b = 23$.

La **règle de décision** pour le médecin sera la suivante : si la fréquence observée f sur l'échantillon étudié appartient à l'intervalle de fluctuation $[0,09; 0,23]$, on considère que l'hypothèse selon laquelle la proportion de malades dans la population est $p = 0,16$ est correcte et on l'accepte.

3. Intervalle de fluctuation asymptotique

Le cadre dans lequel on se place pour cette partie est le suivant : dans une population, on étudie un caractère dont on connaît la proportion p .

On considère un échantillon de taille n prélevé dans cette population et on note X_n le nombre d'individus de l'échantillon possédant ce caractère.

Exemple :

On sait que la proportion de femmes dans la population française est $p = 51\%$.

On choisit au hasard 100 personnes dans la population française.

X_{100} sera le nombre de femmes dans l'échantillon de 100 individus.

■ PROPRIÉTÉ :

X_n est une variable aléatoire qui suit la loi binomiale $\mathcal{B}(n; p)$

Imaginer qu'on constitue l'échantillon en choisissant au hasard à la suite 100 individus dans la population et en observant à chaque fois si l'individu est ou non une femme : c'est un schéma de Bernoulli.

On note enfin F_n la fréquence du caractère étudié dans l'échantillon, donc $F_n = \frac{X_n}{n}$.

D'après le théorème de Moivre-Laplace

$$\lim_{n \rightarrow \infty} P\left(-u_\alpha \leq \frac{X_n - np}{\sqrt{np(1-p)}} \leq u_\alpha\right) = P(-u_\alpha \leq Z \leq u_\alpha) = 1 - \alpha$$

où $Z \hookrightarrow \mathcal{N}(0; 1)$. Or

$$\begin{aligned} -u_\alpha &\leq \frac{X_n - np}{\sqrt{np(1-p)}} \leq u_\alpha \\ \iff -u_\alpha \sqrt{np(1-p)} + np &\leq X_n \leq u_\alpha \sqrt{np(1-p)} + np \\ \iff p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} &\leq \frac{X_n}{n} \leq p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \\ \iff p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} &\leq F_n \leq p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \end{aligned}$$

Ainsi, l'intervalle

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

contient la fréquence F_n avec une probabilité qui tend vers $1 - \alpha$ lorsque n tend vers $+\infty$.

■ THÉORÈME :

Avec les notations précédentes :

$$\lim_{n \rightarrow \infty} P\left(p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq F_n \leq p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) = 1 - \alpha$$

■ DÉFINITION :

L'intervalle

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

est appelé **intervalle de fluctuation asymptotique de la fréquence F_n au seuil de confiance $1 - \alpha$** .

Remarque : En pratique, on utilise surtout l'intervalle de fluctuation asymptotique de F_n au seuil de confiance 95 % et comme $u_{0,05} \approx 1,96$ on obtient avec des conditions d'approximations admises (différentes de précédemment) :

■ THÉORÈME :

Si $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$ alors

$$P\left(F_n \in \left[p - 1,96 \frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 1,96 \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]\right) \approx 0,95$$

■ DÉFINITION :

L'intervalle

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

est l'**intervalle de fluctuation asymptotique de la fréquence F_n au seuil de confiance de 95 %**.

Exemple :

Revenons à l'exemple du nombre de femmes dans un échantillon de 100 personnes : on a

$$1,96 \frac{\sqrt{p(1-p)}}{\sqrt{n}} = 1,96 \frac{\sqrt{0,51 \times 0,49}}{\sqrt{100}} \approx 0,09798$$

et donc $I_{100} \approx [0,51 - 0,09798 ; 0,51 + 0,09798]$
donc $I_{100} \approx [0,412 ; 0,608]$

Ce résultat signifie que dans un échantillon de 100 personnes prises au hasard dans la population, la fréquence du caractère « être une femme » varie avec une probabilité de 95 % entre 0,412 et 0,608. On peut encore dire que, dans un échantillon de 100 personnes prises au hasard dans la population, avec une probabilité de 95 %, il y aura entre 42 et 60 femmes.

4. Prise de décision à partir d'un échantillon

■ PROPRIÉTÉ : Règle de décision

On suppose que dans une population la proportion d'un certain caractère est p .

On observe la fréquence f (ou f_{obs}) de ce caractère sur un échantillon de taille n sous les conditions : $n \geq 30$; $np \geq 5$; $n(1-p) \geq 5$.

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

l'intervalle de fluctuation de la fréquence à 95 % dans les échantillons de taille n .

- Si $f \in I$, alors on accepte l'hypothèse que la proportion est p .
- Si $f \notin I$, alors on rejette cette hypothèse au seuil (rejet à tort) de 5 %.

Exemple :

1. Une publicité affirme que 90 % des consommateurs d'une marque M sont satisfaits. Déterminer l'intervalle de fluctuation I à 95 % de la fréquence de consommateurs satisfaits sur un échantillon de taille 900.
2. Un sondage fait auprès de 900 consommateurs d'une marque M montre que 774 en sont satisfaits.
 - (a) Vérifier que les trois conditions d'approximations sont vérifiées.
 - (b) Peut-on affirmer qu'au seuil de 5 % les résultats du sondage sont en accord avec l'annonce publicitaire ?

5. Deuxième cas : Estimation d'une proportion

Le problème de l'estimation est le problème inverse de celui de l'échantillonnage : à partir de la fréquence observée sur un échantillon, on souhaite estimer la proportion p correspondante sur la population toute entière.

Exemple :

Ce problème se pose par exemple dans les sondages.

Le cadre dans lequel on se place pour cette partie est le suivant : on considère une population dans laquelle il y a une proportion p inconnue d'individus présentant un certain caractère.

Dans la population d'une grande ville, quelle est la proportion p de citoyens soutenant le maire en place ?

On sélectionne un échantillon de taille n dans cette population et on calcule la fréquence f d'individus ayant le caractère étudié dans cet échantillon.

51 % des individus d'un échantillon de taille 100 de la population de la grande ville soutiennent le maire.

On souhaite, à l'aide de cette fréquence f , évaluer la véritable proportion p .

Soit Z une variable aléatoire telle que $Z \sim \mathcal{N}(0; 1)$: on sait que

$$P(-2 \leq Z \leq 2) \approx 0,95449 > 0,954$$

On considère alors un réel $\varepsilon > 0$ tel que :

$$0 < \varepsilon < 0,004$$

alors en notant $\ell = P(-2 \leq Z \leq 2)$, on a

$$\ell - \varepsilon > 0,95$$

D'après le théorème de Moivre-Laplace :

$$\lim_{n \rightarrow \infty} P\left(-2 \leq \frac{X_n - np}{\sqrt{np(1-p)}} \leq 2\right) = P(-2 \leq Z \leq 2) = \ell$$

donc il existe un entier n_0 tel que si $n \geq n_0$ alors

$$P\left(-2 \leq \frac{X_n - np}{\sqrt{np(1-p)}} \leq 2\right) \in]\ell - \varepsilon ; \ell + \varepsilon[$$

Ainsi, pour $n \geq n_0$:

$$P\left(-2 \leq \frac{X_n - np}{\sqrt{np(1-p)}} \leq 2\right) \geq 0,95$$

et donc

$$P\left(p - 2\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq F_n \leq p + 2\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) \geq 0,95$$

Rappelons que la valeur p est inconnue et qu'on souhaite l'estimer. On peut observer que la fonction $p \mapsto p(1-p)$ atteint son maximum en $p = \frac{1}{2}$ et que ce maximum est $\frac{1}{4}$. Ainsi, pour tout $p \in]0; 1[$:

$$\sqrt{p(1-p)} \leq \frac{1}{2} \quad \text{et donc} \quad 2\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq \frac{1}{\sqrt{n}}$$

Ainsi :

$$\left[p - 2\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 2\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right] \subset \left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$$

et donc pour $n \geq n_0$:

$$P\left(p - \frac{1}{\sqrt{n}} \leq F_n \leq p + \frac{1}{\sqrt{n}}\right) \geq 0,95$$

Il reste à observer que :

$$p - \frac{1}{\sqrt{n}} \leq F_n \leq p + \frac{1}{\sqrt{n}} \Leftrightarrow F_n - \frac{1}{\sqrt{n}} \leq p \leq F_n + \frac{1}{\sqrt{n}}$$

et finalement, pour $n \geq n_0$:

$$P\left(F_n - \frac{1}{\sqrt{n}} \leq p \leq F_n + \frac{1}{\sqrt{n}}\right) \geq 0,95$$

■ THÉORÈME :

Si la variable aléatoire X_n suit la loi $\mathcal{B}(n; p)$ et si $F_n = \frac{X_n}{n}$, alors pour tout réel p de l'intervalle $]0; 1[$, il existe un entier n_0 tel que, si $n \geq n_0$, alors

$$P\left(F_n - \frac{1}{\sqrt{n}} \leq p \leq F_n + \frac{1}{\sqrt{n}}\right) \geq 0,95$$

Remarque : Ce théorème fournit donc une estimation de la proportion inconnue p , connaissant la taille n de l'échantillon et la fréquence f du caractère dans l'échantillon, avec une probabilité d'au moins 95 %.

L'intervalle $\left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}\right]$ est l'intervalle de confiance pour p au seuil de 95 %.

Exemple :

Revenons au sondage dans la grande ville : le théorème affirme qu'avec une probabilité de 95 %, la vraie proportion de la population soutenant le maire dans la ville se trouve dans l'intervalle

$$\left[0,51 - \frac{1}{\sqrt{100}}; 0,51 + \frac{1}{\sqrt{100}}\right] = [0,41; 0,61]$$

On remarquera que ce sondage ne permet pas de prévoir un soutien majoritaire au maire !

■ DÉFINITION :

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

L'intervalle $\left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}\right]$ 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 : Conditions d'utilisations de l'intervalle de confiance :

Comme p est inconnue, on approche la proportion p inconnue par la fréquence f observée sur l'échantillon considéré, puis on vérifie si les conditions suivantes sont satisfaites :

$$n \geq 30, \quad nf \geq 5 \quad \text{et} \quad n(1-f) \geq 5$$

Exemple :

Dans une urne contenant un très grand nombre de boules rouges et bleues, on réalise un tirage de 100 boules ; on obtient 59 rouges et 41 bleues : la fréquence de sortie du rouge est donc 0,59.

$$\left[0,59 - \frac{1}{\sqrt{100}}; 0,59 + \frac{1}{\sqrt{100}}\right] = [0,49; 0,69]$$

est un intervalle de confiance de la proportion de boules rouges dans l'urne au niveau de confiance 95 %.

Remarque : L'amplitude de l'intervalle

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

est $\frac{2}{\sqrt{n}}$ donc, plus la taille de l'échantillon est grande, plus les intervalles de confiance sont précis.