

PROBABILITÉS

Échantillonnage – Estimation

I) 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 on **peut faire une hypothèse** sur la valeur de cette proportion

Exemple : Comment décider si 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 !

II) Premier cas : Échantillonnage

P1 (Rappel de Seconde) :

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 %.

E1 : Dans un casino, sur 1 000 lancers de dé, 465 ont donné un nombre pair.

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

P2 (Rappel de Première) : 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 $p(X \leq a) \leq 0,025$, et 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 %**.

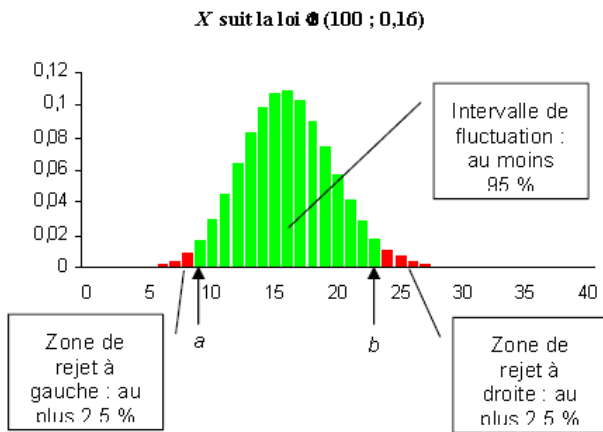
Remarques :

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

Exemple d'utilisation

- 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 $p(X \leq a) \leq 0,025$ et 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.

III) Intervalle de fluctuation asymptotique

P3 (Rappel) : Si la variable aléatoire X_n suit la loi $B(n, p)$, d'après le théorème de Moivre-Laplace, on peut, sous certaines conditions, approcher la loi de la variable aléatoire $Z_n = \frac{X_n - E(X_n)}{\sigma(X_n)} = \frac{X_n - np}{\sqrt{np(1-p)}}$ par la loi normale $\mathcal{N}(0 ; 1)$.

P4 : Les conditions communément admises permettant cette approximation sont :

$$n \geq 30 ; np \geq 5 ; n(1-p) \geq 5.$$

On note $F_n = \frac{X_n}{n}$, F_n est la **fréquence de succès lors des n épreuves**. Et Z_n s'écrit aussi :

$$Z_n = \frac{X_n - np}{\sqrt{np(1-p)}} = \frac{\frac{X_n}{n} - p}{\sqrt{\frac{np(1-p)}{n}}} = \frac{\frac{X_n}{n} - p}{\sqrt{p(1-p)}} = \frac{F_n - E(F_n)}{\sigma(F_n)}$$

(d'après les propriétés de l'espérance et de l'écart-type)

P5 : Soit n un entier naturel et p un réel dans $]0 ; 1[$.

Si la variable aléatoire X_n suit la loi $B(n, p)$, alors pour tout réel α dans l'intervalle $]0 ; 1[$, si u_α désigne l'unique réel tel que $P(-u_\alpha \leq Z \leq u_\alpha) = 1 - \alpha$ où Z suit la loi normale $\mathcal{N}(0 ; 1)$ et si I_n est l'intervalle

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

alors on a : $\lim_{n \rightarrow +\infty} P\left(\frac{X_n}{n} \in I_n\right) = 1 - \alpha$,

c'est-à-dire $\lim_{n \rightarrow +\infty} P(F_n \in I_n) = 1 - \alpha$

Application : Quand on sait qu'une suite converge vers une limite L , on peut considérer que pour n assez grand le terme de rang n constitue une approximation de L .

Ici, on inverse les rôles. On connaît la limite, mais pas les valeurs des termes de la suite.

On admet donc que, on peut approcher le terme de rang n

de la suite $P\left(\frac{X_n}{n} \in I_n\right)$ par sa limite $1 - \alpha$.

Rappelons que les conditions communément admises sont : $n \geq 30 ; np \geq 5 ; n(1-p) \geq 5$.

D1 : Si la variable aléatoire X_n suit la loi $B(n, p)$,

$$\text{l'intervalle } \left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$$

est un **intervalle de fluctuation asymptotique** (ou approché) de la variable fréquence $\frac{X_n}{n}$ au seuil $1 - \alpha$.

P6 (Rappel) : Si Z suit la loi $\mathcal{N}(0 ; 1)$ alors $u_{0,05} \approx 1,96$.

Cas particulier important :

Intervalle de fluctuation au seuil de 95 % :

Si la variable aléatoire X_n suit la loi $B(n, p)$, un

intervalle de fluctuation de $F_n = \frac{X_n}{n}$ au seuil de 95 %

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

E2 : (Lien avec l'intervalle vu en Seconde)

a) Déterminer les variations sur l'intervalle $[0 ; 1]$ de la fonction $x \mapsto x(1-x)$.

b) En déduire que pour tout p de $]0 ; 1[$, $\sqrt{p(1-p)} \leq \frac{1}{2}$.

c) En notant I_n l'intervalle de fluctuation asymptotique au seuil de 95%, en déduire que

$$I_n \subset \left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}} \right], \text{ intervalle noté } J_n. \text{ Pourquoi}$$

peut-on dire que J_n est un intervalle de fluctuation asymptotique à un seuil au moins égal à celui de I_n ?

E3 : La proportion de droitiers dans la population est d'environ 85 %.

Déterminer un intervalle de fluctuation asymptotique au seuil de 95 % de la fréquence des droitiers dans les échantillons de taille 400.

On arrondira les bornes à 10^{-3} près.

IV) Prise de décision à partir d'un échantillon

P7 (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 %.

E4 : 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 ?

E5 : On considère qu'une machine à former des pilules fonctionne de manière satisfaisante si le pourcentage de pilules défectueuses est de 1 pour 1000.

Sur un échantillon de 10 000 pilules, on a trouvé 15 pilules défectueuses.

On veut savoir si la machine est bien réglée.

- a) Déterminer un intervalle de fluctuation asymptotique au niveau de confiance 0,95 de la fréquence de pilules défectueuses dans les échantillons de taille 10 000, prélevés au hasard et avec remise.
(On arrondira les bornes de l'intervalle à 10^{-4} près.)
- b) Énoncer la règle de décision permettant d'accepter ou non l'hypothèse $p = 0,001$ au seuil de 5%. Conclure.

V) 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. Ce problème se pose par exemple dans les sondages.

P8 (admise) : Si la variable aléatoire X_n suit la loi

$B(n, p)$ 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

$$n \geq n_0 \text{ alors } P\left(p \in \left[F_n - \frac{1}{\sqrt{n}} ; F_n + \frac{1}{\sqrt{n}}\right]\right) \geq 0,95.$$

D2 : On réalise l'expérience aléatoire de n tirages au hasard, et on appelle f la fréquence observée d'apparition du caractère. L'intervalle

$$\left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}} \right] \text{ est appelé } \mathbf{intervalle de}$$

confiance de p au niveau de confiance 0,95,

où p est la proportion (inconnue) d'apparition du caractère dans la population.

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 : $n \geq 30$, $n \times f \geq 5$ et $n \times (1-f) \geq 5$.

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.

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

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

Remarque : L'amplitude de l'intervalle

$$\left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}} \right] \text{ est } \frac{2}{\sqrt{n}} \text{ donc plus la taille de}$$

l'échantillon est grande, plus les intervalles de confiance sont précis.

E6 : Dans une population donnée, on s'intéresse à la proportion de blonds. Dans un échantillon de taille 100 extrait de cette population, on a trouvé 22 blonds. Déterminer un intervalle de confiance de la proportion des blonds au niveau de confiance de 95%.

E7 : À l'occasion d'une élection, on réalise un sondage sur un échantillon de n personnes afin de connaître le pourcentage d'électeurs qui souhaitent voter pour un candidat donné. On suppose la population suffisamment importante pour que ce sondage soit assimilé à un tirage avec remise. Quelle doit être la taille minimale de l'échantillon afin que l'intervalle de confiance de cette proportion au niveau de confiance 0,95 nous donne celle-ci à 0,01 près ?