Fluctuation et estimation

Table des matières

I Identification de la situation

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

Intervalles de fluctuation	Intervalles de confiance
Dans l'urne U_1, on **connaît** la proportion p 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 change à chaque série des n tirages. Celle-ci appartient à un « intervalle de fluctuation » de centre p, dont l'amplitude diminue avec n. On est ici dans le domaine de **l'échantillonnage** et de **l'intervalle de fluctuation**.	Dans l'urne U_2, on **ignore** la proportion de boules rouges. On effectue des tirages avec remise de n boules ; on essaye alors d'estimer la proportion p de boules rouges dans l'urne, proportion dont on n'a aucune idée *a priori*. Cette estimation se fait au moyen d'un « intervalle de confiance ». Cet intervalle dépend d'un coefficient, le « niveau de confiance » que l'on attribue à l'estimation. On est ici dans le domaine de **l'estimation** et de **l'intervalle de confiance**

II Échantillonnage, intervalle de fluctuation asymptotique

II.1 Intervalle de fluctuation pour une loi binomiale (vu en Première)

Définition

Soit X une variable aléatoire qui suit la loi binomiale $\mathscr{B}(n\ ;\ p)$. On note $F = \dfrac{X}{n}$ la variable aléatoire qui correspond à la fréquence f de succès.
Soit $\alpha \in]0\ ;\ 1[$.
Un intervalle de fluctuation au seuil $1-\alpha$ est le plus petit intervalle de la forme $\left[\dfrac{a}{n}\ ;\ \dfrac{b}{n}\right]$ où a et b sont des entiers entre 0et n, tels que signifie que $P(a \leqslant X \leqslant b) \geqslant 1-\alpha$, c'est-à-dire $P\left(\dfrac{a}{n} \leqslant F \leqslant \dfrac{b}{n}\right) \geqslant 1-\alpha$.
On choisit donc a comme le plus grand entier tel que $P(X < a) \leqslant \dfrac{1-\alpha}{2}$ et b comme le plus petit entier tel que $P(X > b) \leqslant \dfrac{1-\alpha}{2}$.

Remarque : on prend en général $1-\alpha = 95\,\% = 0,95$ donc $\alpha = 0,25$.
a est donc le plus grand entier tel que tel que $P(X < a) \leqslant 0,025$ et b le plus petit entier tel que $P(X > b) < 0,025$.
On détermine alors les valeurs de a et b par les valeurs de $P(X \leqslant k)$ (obtenues avec la fonction BinomialFRép sur une TI).

Exemple 1 :

Dans une production de fruits, il y a 30 %de fruits abîmés. On prélève un échantillon aléatoire de 80 fruits (la population est suffisamment grande pour qu'on assimile ce tirage à un tirage avec remise).

Cherchons la valeur de k tel que la probabilité que l'échantillon compte au moins k fruits abîmés dépasse 0,025.
On note X la variable aléatoire qui compte le nombre de fruits abîmés. $X \hookrightarrow \mathscr{B}(80\ ;\ 0,3)$.

On cherche le plus petit entier k tel que $P(X \leqslant k) \leqslant 0,025$.
On trouve $k = 16$.

Exemple 2 :

Dans une urne contenant 3 boules rouges et 7 boules blanches, on effectue 100 tirages d'une boule au hasard, avec remise.
On cherche l'intervalle de fluctuation au seuil de 95 % de la fréquence de boules rouges.
On note X la variable aléatoire donnant le nombre de boules rouges obtenues lors des 100 tirages.
X suit la loi binomiale $\mathscr{B}(100\ ;\ 0,3)$. La fréquence est $F = \dfrac{X}{100}$.
À la calculatrice (ou avec un tableur), on obtient $a = 21$ et $b = 39$.

L'intervalle de fluctuation est donc $\boxed{I = [0,21\ ;\ 0,39]}$

II.2 IIntervalle de fluctuation asymptotique

Dans ce paragraphe, on suppose que la proportion **p** du caractère étudié est **connue**.

Définition

Pour tout α dans $]0\,;\,1[$, un intervalle de fluctuation asymptotique e la vavriable aléatoire F au seuil $1-\alpha$ est un intervalle déterminé à partir de p et de$n$ qui contient F avec une probabilité d'autant plus proche de $1-\alpha$ que n est grand.

Remarques : Cette définition met en lumière le fait qu'il n'existe pas un unique intervalle de fluctuation asymptotique à un seuil donné.

Quand on parlera de l'intervalle de fluctuation asymptotique à un seuil donné, i s'agira de celui étudié en Terminale.

Propriété

Pour tout nombre réel α de $]0\,;\,1[$, il existe un unique réel positif u_α tel que la probabilité que la variable aléatoire fréquence F_n prenne ses valeurs dans l'intervalle $I_n = \left[p - u_\alpha\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\,;\,p + u_\alpha\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ se rapproche de plus en plus de $1-\alpha$ quand la taille de l'échantillon n vient grande, ce que l'on note :

$$\lim_{n\to+\infty} P\left(F_n \in I_n\right) = 1-\alpha$$

où X_n est une variable aléatoire qui suit la loi binomiale $\mathscr{B}(n\,;\,p)$ et $F_n = \dfrac{X_n}{n}$.

Démonstration :

On pose $Z_n = \dfrac{X_n - np}{\sqrt{np(1-p)}}$.

D'après le théorème de De Moivre-Laplace, pour tous nombres a et b avec $a \leqslant b$, on a :

$$\lim_{n\to+\infty} P\left(a \leqslant Z_n \leqslant b\right) = \int_a^b \frac{1}{\sqrt{2\pi}\mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x}.$$

Or $Z_n = \dfrac{n\left(\dfrac{X_n}{n} - p\right)}{\sqrt{n} \times \sqrt{p(1-p)}} = \dfrac{F_n - p}{\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}}$.

$a \leqslant Z_n \leqslant b \Leftrightarrow a \leqslant \dfrac{F_n - p}{\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}} \leqslant b$

$\Leftrightarrow a \times \dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant F_n - p \leqslant b \times \dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \Leftrightarrow p + a\dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant F_n \leqslant p + b\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}$. Par conséquent : $\lim\limits_{n\to+\infty} P\left(p + a\dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant F_n \leqslant p + b\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right) = \displaystyle\int_a^b \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}}$ (E).

De plus, pour tout réel $\alpha \in]0\,;\,1[$, il existe un unique réel u_α tel que $\displaystyle\int_{u_\alpha}^{u_\alpha} \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-\frac{x^2}{2}} = 1-\alpha$.

En prenant $a = -u_\alpha$ et $b = u_\alpha$ dans l'égalité (E), la propriété s'ensuit.

Par propriété de la loi normale centrée réduite, on sait que pour $\alpha = 0,05$ et donc $1-\alpha = 0,95$, on a $u_{0,05} \approx 1,96$.

Définition

L'intervalle asymptotique au suis 0,95 de la variable aléatoire fréquence F est défini par :

$$p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\ ;\ p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}.$$

Exemple : On a une urne contenant des boules bleues et rouges ; la proportion de boules rouges est $p = 0,4$. On tire 50 boules ; on souhaite déterminer l'intervalle de fluctuation au seuil $\alpha = 0,1$.
À la calculatrice, on trouve $u_{0,1} \approx 1,645$ (on cherche β tel que $F(Z \leqslant \beta) = 1 - \dfrac{\alpha}{2} = 0,95$) où F est la fonction de répartition de la loi normale centrée réduite.
L'intervalle de fluctuation est donc $I_{50} = \left[0,4 - u_{0,1}\dfrac{\sqrt{0,4\times 0,6}}{\sqrt{50}}\ ;\ 0,4 + u_{0,1}\dfrac{\sqrt{0,4\times 0,6}}{\sqrt{50}}\right] = [0,286\ ;\ 0,514]$.
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 de 0,9.

Remarque : Pour 500 tirages, toujours avec une probabilité de 0,9, on obtient comme intervalle de fluctuation $I_{500} = [0,364\ ;\ 0,436]$; l'amplitude de l'intervalle a été divisée par plus de trois.

Cas particulier : On sait que, pour $\alpha = 0,05$, on a $\boxed{u_\alpha \approx 1,96}$.

On en déduit :

Définition

L'intervale de fluctuation au seuil de 0,95 de la variable aléatoire fréquence F est défini par :

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

Remarque : cet intervalle est inclus dans l'intervalle vu en seconde : $\left[p - \dfrac{1}{\sqrt{n}}\ ;\ p + \dfrac{1}{\sqrt{n}}\right]$

Démonstration : Soit $f : p \mapsto p(1-p))$.
$f(p) = p - p^2$; $f'(p) = 1 - 2p$; $f'(p) = 0 \Leftrightarrow p = \dfrac{1}{2}$; $f'(p) \geqslant 0 \Leftrightarrow p \leqslant \dfrac{1}{2}$.
On en déduit que f est croissante sur $\left[0\ ;\ \dfrac{1}{2}\right]$ puis décroissante sur $\left[\dfrac{1}{2}\ ;\ 1\right]$ et ce maximum est $\dfrac{1}{4}$.
On peut donc majorer 1,96 par 2 et $\sqrt{p(1-p)}$ par $\sqrt{\dfrac{1}{4}} = \dfrac{1}{2}$ donc $1,96\sqrt{p(1-p)}$ par $2\times\dfrac{1}{2} = 1$.
L'intervalle $\left[p - 1,96\times\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\ ;\ p + 1,96\times\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ est donc bien inclus dans $\left[p - \dfrac{1}{n}\ ;\ p = \dfrac{1}{n}\right]$.

II.3 Prendre une décision à partir d'une intervalle de décision

Remarque : pour tester un hypothèse à l'aide d'un intervalle de fluctuation asymptotique, on doit vérifier que les conditions d'utilisation suivantes sont vérifiées :

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

Si ces conditions sont vérifiées, on calcule l'intervalle de fluctuation asymptotique I_n au seuil $1-\alpha$ (en général, 0,95).

Si la fréquence observée n'appartient pas à l'intervalle I_n, on rejette j'hypothèse.

Exemple

Dans un casino, il a été décidé que les « machines à sous » doivent être réglées sur une fréquence de gain du jouer de $p = 0,06$.

Une fréquence inférieure est supposée faire fuir le client et une fréquence supérieure ruiner le casino.

Trois contrôleurs différents vérifient une même machine.

Le premier a joué 50 fois et gagné 2 fois, le deuxième a joué 120 fois et gagné 14 fois et le troisième a joué 400 lis et ganté 30 fois.

En utilisant à chaque fois l'intervalle de fluctuation asymptotique au seuil 95 %, quelle décision a pu rendre le contrôleur ?

Solutions :

- **Premier contrôleur** : $n = 50 \geqslant 30$ mais $np = 50 \times 0,06 = 3 < 5$ donc les conditions ne sont pas réunies.
- **Deuxième contrôleur** : $n = 120 > 30$; $np = 7,2 > 5$; $n(1-p) = 112,8 > 5$: les conditions sont réunies.

  $$I_{120} = \left[0,06 - 1,96 \times \frac{\sqrt{0,06 \times 0,94}}{\sqrt{120}} \,;\, 0,06 + 1,96 \times \frac{\sqrt{0,06 \times 0,94}}{\sqrt{120}}\right] \approx [0{,}0175 \,;\, 0{,}1025].$$

  La fréquence observée est $f_2 = \frac{14}{120} \approx 0{,}1167 \notin I_{120}$. Le contrôleur est conduit à rejeter l'hypothèse que $p = 0,06$ (il a trop souvent gagné).
- **Troisième contrôleur** : $n = 400 > 30$; $np = 24 > 5$; $n(1-p) = 376 > 5$: les conditions sont réunies.

  $$I_{400} = \left[0,06 - 1,96 \times \frac{\sqrt{0,06 \times 0,94}}{\sqrt{400}} \,;\, 0,06 + 1,9\S \times \frac{\sqrt{0,06 \times 0,94}}{\sqrt{400}}\right] \approx [0{,}0367 \,;\, 0{,}0833].$$

  La fréquence observée est $f_3 = \frac{30}{400} = 0{,}075 \in I_{400}$.. Le contrôleur est conduit à accepter l'hypothèse que $p = 0,06$

III Intervalle de confiance

Dans ce paragraphe, la proportion p du caractère étudié dans la population est inconnue ; on essaye de l'estimer à partit d'un échantillon pris dans la population.

Définition

Pour tout $\alpha \in]0\,;\,1[$, un **intervalle de confiance** pour une proportion p au niveau de confiance $1-\alpha$ est la réalissation, à partir d'un échantillon, d'un **intervalle aléatoire** contenant la proportion p avec une probabilité supérieure ou égale à $1-\alpha$.

Il y a plusieurs intervalles de confiance possibles.

En Terminale, on considère celui-là :

Propriété

L'intervalle de confiance au niveau de confiance de 95 % est l'intervale $I = \left[f - \frac{1}{\sqrt{n}}\ ;\ f + \frac{1}{\sqrt{n}}\right]$ où f ets la fréquence observée du caractère sur un échantillon de taille n.

Exemple 1 :

Lors d'un scrutin électoral, on souhaite connaître la proportion p de français votant pour le candidat « A ». L'institut SOFOS mène une enquête auprès de 1 000 personnes tirées au hasard et avec remise (c'est-à-dire qu'une même personne peut éventuellement être choisie plusieurs fois). Le résultat indique qu'une proportion $f = 49\,\%$ d'entre elles voteront pour le candidat « A ».

1. Quelle est la loi suivie par le nombre de personnes votant pour « A » dans une enquête avec remise effectuée auprès de 1 000 personnes ?
2. Étant donné le résultat de l'enquête SOFOS. donner un intervalle de confiance à 95 % pour l'estimation de p.
3. Peut-on affirmer d'après l'enquête que le candidat « A » n'aura pas la majorité des votes, c'est- à-dire que p < 0,5 ?

Solution

1. ¿ chaque étape de l'enquête, il y a une probabilité p de tirer une personne votant pour « A », car on procède à des tirages avec remise. De plus, les 1 000 tirages sont effectués de façon indépendante. L'enquête est donc un schéma de Bernoulli dont chaque succès correspond à tirer une personne votant pour « A », Le nombre de personnes votant pour « A » dans l'enquête suit donc une loi binomiale $\mathscr{B}(\ ;\ )$ où p est inconnu.
2. Notons F la proportion de personnes votant pour « A » parmi 1 000.
   D'après le cours, si $n \geqslant 30$, $np > 5$,$n(1-p) > 5$, un intervalle de confiance au niveau 95 % pour p est $\left[F - \frac{1}{\sqrt{1\,000}}\ ;\ F + \frac{1}{\sqrt{1\,000}}\right]$.
   Ici, $n = 1\,000 > 30$, $np \geqslant 5 \Leftrightarrow p \geqslant 0,005$ et $n(1-p) \geqslant 5 \Leftrightarrow p \leqslant 0,995$.
   On ne connaît pas p, mais au vu de l'enquête, $f = 0,49$, donc on peut raisonnablement penser que p les conditions sont vérifiées.
   L'intervalle de confiance est $\left[0,49 - \frac{1}{\sqrt{1\,000}}\ ;\ 0,49 + \frac{1}{\sqrt{1\,000}}\right] = [0,458\ ;\ 0,522]$.
3. L'intervalle de confiance précédent montre qu'i est tut à fait possible que $p \geqslant 0,5$.
   Donc il n'est pas raisonnable d'affirmer suite à l'enquête que le candidat « A » n'obtiendra pas majorité des votes.

Exemple 2 :

Dans une urne contenant des boules rouges et bleues en proportions inconnues, on effectue des tirages au hasard avec remise.

1. Après avoir effectué 100 tirages, on compte 52 boules rouges et 48 boules bleues.
   Donner un intervalle de confiance à 95 % de la proportion p de boules rouges dans l'urne.
2. Combien faudrait-il, au minimum, effectuer de tirages pour obtenir un intervalle de confiance à 95 % de longueur inférieure ou égale à 2×10^{-2} (c'est-à-dire une précision d'au moins 0,02) ?

Solutions

1. Avec $n=100$ et $f=0,52$, on a $f-\dfrac{1}{\sqrt{n}}=0,42$ et $f+\dfrac{1}{\sqrt{n}}=0,62$. Un intervalle de confiance à 95 % de la proportion p de boules rouges dans l'urne est : $[0.42\,;\,0,62]$.
2. On cherche le plus petit entier naturel n tel que $\dfrac{1}{\sqrt{n}}\leqslant 10^{-2}$. On trouve $n\geqslant 10^4$.

   En prélevant au moins 10 000 boules, on obtient un intervalle de confiance à 95 % à la précision 0,02.

Exemple 3 :

Une usine fabrique des pièces métalliques, qui sont censée résister à certaines contraintes mécaniques. Le responsable de fabrication souhaite estimer le taux de pièces défectueuses concernant la résistance mécanique dans la population.

Pour cela, il utilise la méthode par intervalle de confiance au niveau 95 %, en extrayant au hasard n pièces en fin de production qui sont soumises à contrainte mécanique jusqu'à la rupture. En fonction du niveau de contrainte à la rupture, on décide de la nature défectueuse ou pas de la pièce.

1. Chaque pièce testée étant détruite, le responsable souhaite minorer la taille de l'échantillon testé, tout en ayant un intervalle de confiance de de longueur inférieure à 0,1.

   Quelle taille d'échantillon lui conseiller ?
2. Il est décide de mener l'étude sur 100 pièces ; on trouve 40 pièces défectueuses. Quel intervalle de confiance au niveau de confiance 95 % obtient-on ?
3. L'année précédente, à l'issue d'un problème grave de rupture d'une pièce, une large étude avait débouché sur 130 pièces défectueuses sur un échantillon de 1 000.

   Peut-on supposer que la mise en place de nouvelles procédures de fabrication a vraiment diminué a proportion de pièces défectueuses ?

Solution

1. La longueur d'un intervalle de confiance est $\dfrac{2}{\sqrt{n}}$.

   On doit avoir $\dfrac{2}{\sqrt{n}}\leqslant 0,1$ d'où $\boxed{n\geqslant 400}$. Il faut donc tirer au moins 400 pièces.
2. La fréquebce est $f=\dfrac{40}{500}=0,08$.

   L'intervalle de confiance est est alors $I:\left[0,08-\dfrac{1}{\sqrt{500}}\,;\,0,08+\dfrac{1}{\sqrt{500}}\right]\approx\boxed{[0,0353\,;\,0,1247]}$.
3. La fréquence observée est $f'=\dfrac{130}{1\,000}=0,13$.

   L'intervalle de confiance au seuil 95 % est alors $I'=\left[0,13-\dfrac{1}{\sqrt{1\,000}}\,;\,0,13+\dfrac{1}{\sqrt{1\,000}}\right]\approx[0,0984\,;\,0,1616]$.

   Les deux intervalles de confiance obtenus ne sont pas disjoints ; il est possible que la proportion de pièces défectueuses n'ait pas changé.

IV Exercices (livre Math'x)

Pour décider la construction d'un grand stade, une municipalité veut sonder la population pour estimer si plus de 50 % des électeurs y sont favorables.

1. La municipalité réalise un sondage aléatoire de taille 100 et obtient 54 avis favorables.
   a) Quelle est la fréquence f d'avis favorables sur ce sondage ?
   b) La municipalité peut-elle décider la construction du stade en prétextant que plus de 50 % de la population est favorable ?
2. On suppose que la municipalité réalise un sondage de taille n et que la fréquence f des votes favorables reste égale à 0,54.
   Si p est la proportion (inconnue) d'avis favorables dans la population, donner l'expression d'un intervalle de confiance de p au niveau de 95 %.
3. La municipalité ne construira le stade que si la proportion p d'avis favorables dépase 50 %.
   a) Montrer que pour avoir $p > 0,5$, au seuil de confiance de 95 %, il suffit d'avoir : $0,54 - \dfrac{1}{\sqrt{n}} > 0,5$.
   b) Déterminer le plus petit entier n_0 à partir duquel l'inéquation est vérifiée.
   c) Si le sondage, avec la même fréquence observée portait sur 650 personnes, pensez-vous que le stade serait construit ?

Solutions

1. a) $f = 0,54$
   b) Au niveau de confiance de 95 %, l'intervalle de confiance est $\left[0,54 - \dfrac{1}{\sqrt{100}}\ ;\ 0,54 + \dfrac{1}{\sqrt{100}}\right] = [0,44\ ;\ 0,64]$.
   On ne peut pas affirmer, au niveau de confiance de 95 %, $p > 0,5$. La réponse est non.
2. $I_n = \left[0,54 - \dfrac{1}{\sqrt{n}}\ ;\ p + \dfrac{1}{\sqrt{n}}\right]$.
3. a) Pour avoir $p > 0,5$, il suffit, au niveau de confiance de 95 %, d'avoir $0,54 - \dfrac{1}{\sqrt{n}} > 0,5$.
   b) Cela équivaut à $n > 625$ donc $n_0 = 626$.
   c) Pour $n = 650$, on aurait $p > 0,5$ avec un niveau de confiance de 95 %. Le stade serait construit.