Fluctuation et Échantillonnage

1) *Introduction*

On considère une urne contenant des boules rouges et des boules noires. Une partie consiste en 50 tirages successifs avec remise d'une boule de cette urne (cette succession de 50 épreuves indépendantes et identiques s'appelle ***un échantillon de taille 50***) et à l'issue d'une partie, on s'intéresse à **la *fréquence d'apparition des boules rouges*.**
Pour étudier le comportement de cette fréquence, on a simulé 100 parties consécutives et on a observé que la distribution des fréquences varie avec l'échantillon prélevé : on parle alors de ***fluctuation d'échantillonnage*.**

• Si ***on connait la proportion p*** de boules rouges dans l'urne, ou bien si on fait une hypothèse sur la valeur de cette proportion, on peut alors se demander si les fréquences d'apparition observées d'une boule rouge appartiennent à un intervalle donné, dit ***intervalle de fluctuation.***

• Si ***on ne connait pas la proportion p*** de boules rouges dans l'urne, on peut chercher à estimer cette proportion, à partir des fréquences observées d'apparition d'une boule rouge : on est alors dans le domaine de ***l'estimation*.**

2) *Intervalle de fluctuation.*

Dans cette partie, la proportion d'un certain caractère dans une population est ***connue***.

Soit X_n une variable aléatoire qui suit une loi binomiale $B(n;p)$, $n \in \mathbb{N}^*$

On définit la Variable Aléatoire F_n par : $F_n = \dfrac{X_n}{n}$ nommée Variable Aléatoire Fréquence

X_n prend les valeurs : 0 ; 1 ; 2 ; ..... k ..... ; n

F_n prend les valeurs : $\dfrac{0}{n}; \dfrac{1}{n}; \dfrac{2}{n}; .....; \dfrac{k}{n}; ..... \dfrac{n}{n}$.

F_n indique donc la fréquence d'apparition de succès dans le schéma de Bernoulli associé à la Variable Aléatoire X_n (nombre de fois que l'issue « S » apparait, divisé par le nombre n d'épreuves du schéma de Bernoulli).

Rappel :

Pour tout entier k tel que $0 \le k \le n$: $P\left(F_n = \dfrac{k}{n}\right) = P\left(\dfrac{X_n}{n} = \dfrac{k}{n}\right) = P(X_n = k) = \binom{n}{k} p^k (1-p)^{n-k}$

On connait donc la loi de probabilité de la Variable Aléatoire F_n (qui n'est pas une loi binomiale !!).

L'espérance de F_n est: $E(F_n) = E\left(\dfrac{X_n}{n}\right) = \dfrac{1}{n}E(X_n) = \dfrac{1}{n} \times np = p$

L'écart-type de F_n est: $\sigma(F_n) = \sigma\left(\dfrac{X_n}{n}\right) = \left|\dfrac{1}{n}\right|\sigma(X_n) = \dfrac{1}{n} \times \sqrt{np(1-p)} = \sqrt{\dfrac{np(1-p)}{n^2}} = \dfrac{\sqrt{p(1-p)}}{\sqrt{n}}$

a) *Rappel des classes précédentes*

➢ ***En seconde***

La taille n de l'échantillon doit être supérieure à 25, p doit appartenir à l'intervalle $[0,2;0,8]$
Alors, dans ces conditions, dans plus de 95 % des cas, la fréquence f_{obs} du caractère appartient à l'intervalle $I_F = \left[p - \dfrac{1}{\sqrt{n}}; p + \dfrac{1}{\sqrt{n}}\right]$

Exercice :

La fréquence des yeux bleus en France est d'environ 0,31. On a prélevé un échantillon de 50 individus dont 15 ont les yeux bleus.
Quel est l'intervalle de fluctuation au seuil de 95 % ?

Solution :

On doit vérifier que les conditions sont vérifiées. Ici $n = 50$, et $p = 0{,}31$. On peut donc appliquer la formule $I_F = \left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$ Soit donc $I_F = \left[0{,}31 - \frac{1}{\sqrt{50}}; 0{,}31 + \frac{1}{\sqrt{50}}\right]$

$I_F \approx [0{,}168; 0{,}452]$ ici, on a : $f_{obs} = \frac{15}{50} = 0{,}3$. (voir prise de décision)

- ***En première***

On a rappelé en préambule que X suit une loi Binomiale de paramètre n et p, notée $B(n;p)$
On partage alors $[0;n]$ en trois intervalles : $[0;a-1]$ $[a;b]$et $[b+1;n]$.
a et b sont choisis de telle sorte que X prenne ses valeurs dans les deux intervalles extrêmes avec une probabilité proche de 2,5% sans jamais la dépasser. (Intervalles de rejet)

Définition :

L'intervalle de fluctuation à 95% d'une fréquence correspondant à la réalisation, sur un échantillon de taille n, d'une variable aléatoire X de loi binomiale $B(n;p)$ est l'intervalle $\left[\frac{a}{n};\frac{b}{n}\right]$défini par :
a est le plus petit entier tel que $P(X \leq a) > 0{,}025$
b est le plus petit entier tel que $P(X \leq b) \geq 0{,}975$

Remarque :

- On peut alors montrer que $P(a \leq X \leq b) \geq 0{,}95$
- Dans au moins 95% des cas, la fréquence observée f_{obs} appartient à l'intervalle $[\frac{a}{n};\frac{b}{n}]$
- Pour $n \geq 30$, $np \geq 5$et $n(1-p) \geq 5$ l'intervalle de fluctuation à 95% est sensiblement le même que celui donné en classe de seconde et défini par : $I_F = \left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$

Exercice :

Le taux de réussite du bac S en France en 2021 a été de 88,2 %.
Déterminer un intervalle de fluctuation asymptotique du taux de réussite dans une classe de TS qui est composée de 35 élèves.

Solution :

Soit X la variable aléatoire qui compte le nombre de réussite au BAC parmi les 35 ;
La variable aléatoire X suit la loi binomiale de paramètres $n = 35$ et $p = 0{,}882$.
On lit sur le tableur que le plus petit entier a tel que $P(X \leq a) > 0{,}025$ est $a = 27$ et le plus petit entier b tel que $P(X \leq b) \geq 0{,}975$ est $b = 34$.
En divisant par n, on obtient $\left[\frac{a}{n};\frac{b}{n}\right] = \left[\frac{27}{35};\frac{34}{35}\right] = [0{,}771; 0{,}972]$. L'intervalle de fluctuation est donc $[0{,}771; 0{,}972]$.

Attention :

On arrondit toujours la bonne inférieure par défaut et la borne supérieure par excès. On a toujours cherché un intervalle de Fluctuation à 95%.

b) Intervalle de fluctuation asymptotique au seuil de $1-\alpha$

Théorème et définition :

• Soit X_n une Variable Aléatoire qui suit une loi binomiale $B(n;p)$, $n \in \mathbb{N}^*$ et $\alpha \in]0;1[$
Si X une Variable Aléatoire qui suit la loi normale centrée réduite $N(0;1)$, on note u_α l'unique réel strictement positif tel que $P(-u_\alpha \leq X \leq u_\alpha) = 1-\alpha$.

On note I_n l'intervalle $I_n = \left[p - u_\alpha \dfrac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + u_\alpha \dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$.

Alors $\lim\limits_{n\to+\infty} P\left(\dfrac{X_n}{n} \in I_n \right) = 1-\alpha$ ou encore $\lim\limits_{n\to+\infty} P(F_n \in I_n) = 1-\alpha$ avec $F_n = \dfrac{X_n}{n}$.

• Lorsque n tend vers $+\infty$, la fréquence F_n appartient à l'intervalle I_n avec une probabilité qui se rapproche de $1-\alpha$. I_n est appelé ***l'intervalle de fluctuation asymptotique de la fréquence*** F_n ***au seuil de*** $1-\alpha$

Démonstration

$$F_n \in I_n \Leftrightarrow \frac{X_n}{n} \in I_n \Leftrightarrow p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq \frac{X_n}{n} \leq p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

$$\Leftrightarrow np - u_\alpha \frac{n\sqrt{p(1-p)}}{\sqrt{n}} \leq X_n \leq np + u_\alpha \frac{n\sqrt{p(1-p)}}{\sqrt{n}}$$

$$\Leftrightarrow np - u_\alpha \sqrt{np(1-p)} \leq X_n \leq np + u_\alpha \sqrt{np(1-p)}$$

$$\Leftrightarrow -u_\alpha \sqrt{np(1-p)} \leq X_n - np \leq u_\alpha \sqrt{np(1-p)}$$

$\Leftrightarrow -u_\alpha \leq \dfrac{X_n - np}{\sqrt{np(1-p)}} \leq u_\alpha$. En posant, $Z_n = \dfrac{X_n - np}{\sqrt{np(1-p)}}$ on obtient $\Leftrightarrow -u_\alpha \leq Z_n \leq u_\alpha$

D'après le théorème de Moivre-Laplace, si X_n est une variable aléatoire qui suit la loi $B(n;p)$, alors $\lim\limits_{n\to+\infty} P\left(-u_\alpha \leq \dfrac{X_n - np}{\sqrt{np(1-p)}} \leq u_\alpha \right) = P(-u_\alpha \leq Z_n \leq u_\alpha) = 1-\alpha$ où Z_n suit une $N(0;1)$

Donc $\lim\limits_{n\to+\infty} P(F_n \in I_n) = 1-\alpha$.

Remarque :

L'intervalle de fluctuation vu en seconde est une approximation plus large que celui de TS.

c) Intervalle de fluctuation asymptotique au seuil de 95%

Dans le chapitre précédent, nous avons établi que :

$P(-1{,}96 \leq X \leq 1{,}96) \simeq 0{,}95 \Leftrightarrow P(-1{,}96 \leq X \leq 1{,}96) \simeq 1 - 0{,}05$

Autrement dit $u_{0{,}05} \simeq 1{,}96$. On en déduit le résultat suivant :

Propriété :

Pour une variable aléatoire X_n suivant une loi binomiale $B(n;p)$, ***l'intervalle de fluctuation asymptotique au seuil de 95%*** de la fréquence de succès $F_n = \dfrac{X_n}{n}$ est l'intervalle :

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

d) ***Règle de prise de décision***

Dans une population d'effectif *N*, on suppose que la proportion d'un certain caractère est *p*. On souhaite juger cette hypothèse (valider ou rejeter). Pour cela, on prélève dans la population, au hasard et avec remise, un échantillon de taille *n* sur lequel on observe la fréquence f_{obs} de ce caractère.

Si X_n désigne la Variable Aléatoire égale au nombre d'individus de l'échantillon présentant le caractère étudié, X_n suit la loi binomiale $B(n;p)$. f_{obs} est une valeur de la Variable Aléatoire $F_n = \dfrac{X_n}{n}$ donnant la fréquence de succès.

Si les conditions d'approximation $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$ ***sont remplies***, alors dans environ 95% des cas, la fréquence F_n est dans l'intervalle de fluctuation asymptotique :

$I_n = \left[p - 1{,}96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 1{,}96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$ On adopte la règle de prise de décision suivante :

Règle de décision:

- Si $f_{obs} \in I_n$, on accepte l'hypothèse que la proportion est *p* au seuil de 95%.
- Si $f_{obs} \notin I_n$, on rejette cette hypothèse au seuil de risque de 5%.

Exemple :

L'institut national de prévention et d'éducation pour la santé (INPES) a étudié la proportion de fumeurs quotidiens de tabac en France en 2010. Chez les 15-19 ans, 26% des garçons et 20% des filles sont des fumeurs quotidiens.
Sur un échantillon de 1000 lillois de cette tranche d'âge, dont 450 sont des filles, on a dénombré 178 fumeurs quotidiens chez les garçons et 98 chez les filles.
On fait l'hypothèse que dans la ville de Lille, la proportion de fumeurs quotidiens chez les 15-19 ans est de 26% pour les garçons et de 20% pour les filles.
Répondre aux questions suivantes pour les garçons puis pour les filles.

1. Déterminer l'intervalle de fluctuation asymptotique au seuil de 95% de la fréquence de fumeurs quotidiens chez les 15-19 ans (on n'oubliera pas de vérifier les conditions d'application de la règle de prise de décision).
2. Pouvez-vous considérer, au seuil de risque de 5%, que la fréquence observée de jeunes fumeurs quotidiens lillois dans cet échantillon est en accord avec la proportion de jeunes fumeurs quotidiens de la population française ?

Solution :

Pour les garçons :
On vérifie d'abord que les conditions d'application sont vérifiées :
$n = 550$, $np = 143$ et $n(1-p) = 407$
On peut ainsi définir l'intervalle de fluctuation asymptotique au seuil 0,95 dans un échantillon de taille $n = 550$:

$$\left[0{,}26 - 1{,}96 \times \frac{\sqrt{0{,}26 \times 0{,}74}}{\sqrt{550}} ; 0{,}26 + 1{,}96 \times \frac{\sqrt{0{,}26 \times 0{,}74}}{\sqrt{550}} \right] \approx [0{,}2233 ; 0{,}2967]$$

La fréquence observée, fréquence des garçons fumeurs dans l'échantillon considéré, est égale à : $f_{obs} = \dfrac{178}{550} = 0{,}3236$
Comme cette fréquence n'appartient pas à l'intervalle de fluctuation asymptotique au seuil 0,95, l'hypothèse selon laquelle 26 % des garçons sont des fumeurs est rejetée.

Pour les filles :
On vérifie d'abord que les conditions d'application sont vérifiées :
$n=450$, $np=90$ et $n(1-p)=360$
On peut ainsi définir l'intervalle de fluctuation asymptotique au seuil 0,95 dans un échantillon de taille $n=450$:

$$\left[0,2-1,96\times\frac{\sqrt{0,2\times0,8}}{\sqrt{450}};0,2+1,96\times\frac{\sqrt{0,2\times0,8}}{\sqrt{450}}\right]\approx[0,1630;0,2370]$$

La fréquence observée, fréquence des filles fumeuses dans l'échantillon considéré, est égale à : $f_{obs}=\frac{98}{450}=0,2177$
Comme cette fréquence appartient à l'intervalle de fluctuation asymptotique au seuil 0,95, l'hypothèse selon laquelle 20 % des filles sont des fumeurs est acceptée.

3) Estimation par intervalle de confiance

Dans cette partie, la proportion d'un certain caractère dans une population est ***inconnue***.
On cherche alors à ***estimer*** cette proportion p, $p\in[0;1]$ à partir d'un échantillon de taille n (n tirages successifs au hasard et avec remise), avec un certain ***niveau de confiance***.
Si X_n désigne la Variable Aléatoire égale au nombre d'individus de l'échantillon présentant le caractère étudié, X_n suit la loi binomiale $B(n;p)$ et $F_n=\frac{X_n}{n}$ donne la fréquence de succès.
Si les conditions d'approximation $n\geq30$, $np\geq5$ **et** $n(1-p)\geq5$ ***sont remplies***, alors dans environ 95% des cas, la fréquence F_n est dans l'intervalle de fluctuation asymptotique:

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

Comparons les intervalles $I_n=\left[p-1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}};p+1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ et $J_n=\left[p-\frac{1}{\sqrt{n}};p+\frac{1}{\sqrt{n}}\right]$, tous deux centrés en p.
La fonction polynôme du second degré $p\mapsto p(1-p)=-p^2+p$ admet 0,25 pour maximum, atteint en 0,5. On a donc $0\leq p(1-p)\leq\frac{1}{4}$ et par croissance de la fonction racine carrée :
$0\leq\sqrt{p(1-p)}\leq\frac{1}{2}$, et par suite $0\leq1,96\sqrt{p(1-p)}\leq1$, $\forall p\in[0;1]$.
D'où $0\leq\frac{1,96\sqrt{p(1-p)}}{\sqrt{n}}\leq\frac{1}{\sqrt{n}}$ et donc l'intervalle I_n est contenu dans l'intervalle J_n.
$I_n\subset J_n$ donc $P(F_n\in I_n)\leq P(F_n\in J_n)$. Si les conditions d'approximation ci-dessus sont bien remplies, $P(F_n\in I_n)\approx0,95$ donc $P(F_n\in J_n)\geq0,95$ dans la plupart des cas. D'où le résultat suivant :

Propriété :
Soit X_n une variable aléatoire qui suit une loi binomiale $B(n;p)$ où p est la proportion ***inconnue*** d'apparition d'un caractère et $F_n=\frac{X_n}{n}$ la fréquence associée à X_n. Alors, pour n suffisamment grand, $p\in\left[F_n-\frac{1}{\sqrt{n}};F_n+\frac{1}{\sqrt{n}}\right]$ avec une probabilité supérieure ou égale à 0,95.

Démonstration :

Il suffit d'inverser les doubles inégalités:

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

soit alors $\Leftrightarrow F_n - \frac{1}{\sqrt{n}} \le p \le F_n + \frac{1}{\sqrt{n}} \Leftrightarrow p \in \left[F_n - \frac{1}{\sqrt{n}} ; F_n + \frac{1}{\sqrt{n}}\right]$

Deux évènements équivalents ont la même probabilité donc $P\left(p \in I_c\right) \ge 0{,}95$ lorsque n est assez grand, ce qui montre que l'intervalle a un niveau de confiance de 95% pour l'estimation de p.

Définition :

On note p la proportion inconnue d'un caractère dans une population et f_{obs} la fréquence observée d'apparition de ce caractère sur un échantillon de taille n.

- On appelle ***Intervalle de confiance de p au niveau de confiance 0, 95*** (ou 95%), l'intervalle $I_c = \left[f_{obs} - \frac{1}{\sqrt{n}} ; f_{obs} + \frac{1}{\sqrt{n}}\right]$.
- L'amplitude de cet intervalle est $\frac{2}{\sqrt{n}}$

Remarques :

- Le niveau de confiance 0,95 nous indique simplement que dans plus de 95 cas sur 100, on énonce à juste titre que la proportion inconnue p appartient à l'intervalle de confiance considéré. Par contre, on ne peut faire aucun pronostic sur une localisation possible de cette proportion p dans l'intervalle. En particulier, ***la proportion inconnue n'est pas nécessairement le centre de l'intervalle***.
- A chaque tirage d'un échantillon, on obtient un I_c différent.
- L'intervalle de confiance au niveau de confiance 0,95 est centré en la fréquence observée f_{obs}. Cette condition n'étant pas imposée par la définition générale, ce n'est donc pas nécessairement le cas de tous les intervalles de confiance.
- La proportion p étant inconnue, on ne peut pas vérifier si les conditions exigées sur n et p en début de chapitre sont vérifiées afin d'utiliser l'intervalle de confiance au niveau de confiance 0,95. Pour remédier à ce problème, ***on approche la proportion inconnue p par la fréquence observée*** f_{obs} sur l'échantillon considéré, puis on vérifie si les conditions suivantes sont satisfaites : $n \ge 30, n \times f \ge 5$ **et** $n \times (1-f) \ge 5$

Exercice :

Après avoir examiné 100 poissons pêchés dans un lac, on a constaté que 20% d'entre eux étaient malades. Par quel intervalle de confiance au niveau de confiance de 95% peut-on estimer la proportion de poissons malades dans le lac ?
On estime la population de poissons dans le lac suffisamment importante pour assimiler le prélèvement des 100 poissons à des tirages successifs avec remise.

Solution :

On a donc : $n = 100$ et $f_{obs} = 0{,}2$; On vérifie que les conditions sont vérifiées :

$n \ge 30$ et $n \times f = 20$ et $n \times (1-f) = 80$. On a donc : $I_c = \left[0{,}2 - \frac{1}{\sqrt{100}} ; 0{,}2 + \frac{1}{\sqrt{100}}\right] = \left[0{,}1 ; 0{,}3\right]$

I_c est un intervalle de confiance de la proportion des poissons malades du lac au niveau de confiance de 95%.

Exercice :

Une urne contient des boules rouges et des boules noires. On cherche à estimer la proportion p (inconnue) de boules rouges de l'urne à l'aide de tirages successifs avec remise. Combien de tirages suffit-il d'effectuer pour déterminer un encadrement de p d'amplitude 0,1 au niveau de confiance de 95% ?

Solution :

On rappelle l'expression de l'intervalle de confiance : $I_c = \left[f_{obs} - \frac{1}{\sqrt{n}} ; f_{obs} + \frac{1}{\sqrt{n}} \right]$

L'amplitude de l'intervalle étant $\frac{2}{\sqrt{n}}$, je cherche n afin que : $\frac{2}{\sqrt{n}} \leq 0{,}1$ soit alors $n \geq 400$.

Ainsi, à partir de 400 tirages, on aura un encadrement de la proportion des boules rouges de l'urne avec une précision de 0,1 avec un niveau de confiance de 0,95.