Échantillonnage
Estimation

Christophe ROSSIGNOL*

Année scolaire 2019/2020

Table des matières

Table des figures

Activité : Activité 2 page 437 [1] [TransMath]

1 Échantillonnage – Intervalle de fluctuation

1.1 Définition – Utilisation

Rappel : On appelle échantillon de taille n la série statistique formée des résultats obtenus lorsqu'on répète n fois une expérience dans les mêmes conditions.

- Les distributions de fréquences varient d'un échantillon à l'autre pour la même expérience. C'est ce qu'on appelle la fluctuation d'échantillonnage.
- Même pour des échantillon de même taille, la distribution de fréquences peut varier.
- Lorsque la taille de l'échantillon augmente, les distributions de fréquences ont tendance à se stabiliser.

Remarque : Comme on répète dans les mêmes conditions une expérience n fois, on peut assimiler cet échantillon à une loi binomiale $\mathcal{B}(n\,;\,p)$, où p est la proportion du caractère étudié dans la population totale. La distribution de fréquence de cet échantillon peut alors être assimilée à la loi de fréquence F_n.

Définition : Soit X_n une variable aléatoire qui suit le loi binomiale $\mathcal{B}(n\,;\,p)$ et $F_n = \frac{X_n}{n}$ la fréquence de succès.
Soit α un nombre réel compris entre 0 et 1.
On dit que l'intervalle I_n est un intervalle de fluctuation de F_n au seuil de $1-\alpha$ si :

$$P(F_n \in I_n) \geq 1-\alpha$$

Remarques : 1. Les intervalles de fluctuation les plus utilisés sont les intervalles de fluctuation au seuil des 95 %, c'est-à-dire pour $\alpha = 0,05$.

2. On utilise donc les intervalles de fluctuation dans les deux cas suivants :
   - on connaît la proportion p de présence du caractère dans la population ;
   - on fait une hypothèse sur la valeur de cette proportion et on veut valider (ou invalider) cette hypothèse (on parle alors de prise de décision).

1.2 Intervalle de fluctuation simplifié

Ce résultat a été vu en classe de Seconde :

Propriété : Soit un caractère dont la proportion dans une population donnée est p. On considère un échantillon de taille n.
Si $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$, l'intervalle :

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

est un intervalle de fluctuation au seuil de 95 %.

1.3 Intervalle de fluctuation avec la loi binomiale

Ce résultat a été vu en classe de Première S :

Propriété : Soit un caractère dont la proportion dans une population donnée est p.
L'intervalle de fluctuation au seuil des 95 % , pour un échantillon de taille n, selon la loi binomiale de paramètres n et p est :

$$\left[\frac{a}{n}\,;\,\frac{b}{n}\right] \quad \text{avec} \quad \begin{cases} a : & \text{plus petit entier tel que } p(X_n \leq a) > 0,025 \\ b : & \text{plus petit entier tel que } p(X_n \leq b) \geq 0,975 \end{cases}$$

Remarque : On admettra que si $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$, cet intervalle de fluctuation est sensiblement le même que celui vu en Seconde.

1. Estimation par « Capture-recapture »

1.4 Intervalle de fluctuation asymptotique

Rappel : Soit Z une variable aléatoire suivant la loi $\mathcal{N}(0\,;1)$.
Soit α un nombre réel tel que $0<\alpha<1$.
Il existe un unique nombre strictement positif u_α tel que :

$$P(-u_\alpha < Z < u_\alpha) = 1-\alpha$$

Remarque : On a vu que $u_{0,05} \simeq 1,96$ et $u_{0,01} \simeq 2,58$.

Théorème : Soit X_n une variable aléatoire qui suit la loi binomiale $\mathcal{B}(n\,;p)$ et $F_n = \frac{X_n}{n}$ la loi de fréquence des succès.
Soit $\alpha \in]0\,;1[$. On a :

$$\lim_{n\to+\infty} P(F_n \in I_n) = 1-\alpha \qquad \text{où} \qquad I_n = \left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\,;\, p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$$

Démonstration :
On pose $Z_n = \frac{X_n - np}{\sqrt{np(1-p)}}$.
D'après le théorème de Moivre-Laplace, on a :

$$\lim_{n\to+\infty} P(-u_\alpha \leq Z_n \leq u_\alpha) = P(-u_\alpha \leq Z \leq u_\alpha) = 1-\alpha$$

$$\text{Or : } -u_\alpha \leq Z_n \leq u_\alpha \iff -u_\alpha \leq \frac{X_n-np}{\sqrt{np(1-p)}} \leq u_\alpha \iff -u_\alpha\sqrt{np(1-p)} \leq X_n - np \leq u_\alpha\sqrt{np(1-p)}$$

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

$$\iff p - u_\alpha\frac{\sqrt{n}\sqrt{p(1-p)}}{(\sqrt{n})^2} \leq F_n \leq p + u_\alpha\frac{\sqrt{n}\sqrt{p(1-p)}}{(\sqrt{n})^2} \iff p - u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq F_n \leq p + u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

On obtient donc :

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

Remarques : 1. 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 alors appelé intervalle de fluctuation asymptotique de F_n au seuil de $1-\alpha$.

2. Comme $u_{0,05} \simeq 1,96$, on a :

$$\lim_{n\to+\infty} P(F_n \in J_n) = 0,95 \qquad \text{où} \qquad J_n = \left[p - 1,96 \frac{\sqrt{p(1-p)}}{\sqrt{n}}\,;\, p + 1,96 \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$$

L'intervalle J_n est appelé intervalle de fluctuation asymptotique de F_n au seuil des 95 %.
Il est utilisé sous les conditions : $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$.
On peut montrer que, sous ces conditions, cet intervalle est plus précis que l'intervalle de fluctuation simplifié vu en Seconde.

Exercices : 14 page 448 et 30 page 454 [2] – 3 page 459 [3] [TransMath]

1.5 Règle de prise de décision

On considère une population dans laquelle on suppose que la proportion d'un caractère est p.
On observe la fréquence f_{obs} de ce caractère dans un échantillon de taille n et on considère l'hypothèse « la proportion de ce caractère dans la population est p ».

2. Contrôle de fluctuation.
3. R.O.C.

On considère que les conditions $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$ sont remplies et on note

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

l'intervalle de fluctuation asymptotique de F_n au seuil des 95 %.

On a alors la règle de décision suivante :
- Si $f_{\text{obs}} \in J_n$: on considère que l'hypothèse n'est pas remise en question et l'on accepte au seuil de risque de 5 % ;
- Si $f \notin J_n$: on rejette l'hypothèse au seuil de risque de 5 % (ce qui signifie que le risque d'erreur par rejet à tort de l'hypothèse est d'environ 5 %).

Exercices : 1, 2, 3 page 443 et 32, 33 page 455 [4] – 44, 45 page 459 [5] [TransMath]

2 Intervalle de confiance

On considère maintenant le cas où la proportion p du caractère dans la population totale est inconnue.
On veut estimer p à l'aide d'un échantillon de taille n, et on supposera que les conditions $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$ sont remplies.

Propriété : Soit X_n une variable aléatoire suivant la loi binomiale $\mathcal{B}(n\,;p)$ et $F_n = \frac{X_n}{n}$ la fréquence des succès.
Pour n suffisamment grand, on a :

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

Ce qui signifie que l'intervalle $\left[F_n - \frac{1}{\sqrt{n}} \,;\, F_n + \frac{1}{\sqrt{n}} \right]$ contient la proportion p avec une probabilité de plus de 0,95.

Démonstration :
On a vu que l'intervalle $\left[p - \frac{1}{\sqrt{n}} \,;\, p + \frac{1}{\sqrt{n}} \right]$ est un intervalle de fluctuation seuil des 95 % donc, pour n suffisamment grand :

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

Or : $p - \frac{1}{\sqrt{n}} \leq F_n \leq p + \frac{1}{\sqrt{n}} \iff -\frac{1}{\sqrt{n}} \leq F_n - p \leq \frac{1}{\sqrt{n}} \iff -F_n - \frac{1}{\sqrt{n}} \leq -p \leq -F_n + \frac{1}{\sqrt{n}}$
En multipliant ce dernier encadrement par -1 (ce qui inverse l'ordre), on obtient : $F_n - \frac{1}{\sqrt{n}} \leq p \leq F_n + \frac{1}{\sqrt{n}}$.
On obtient donc :

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

Définition : On observe une fréquence f_{obs} sur un échantillon de taille n.
On appelle intervalle de confiance de p au niveau de confiance de 95 % l'intervalle $\left[f_{\text{obs}} - \frac{1}{\sqrt{n}} \,;\, f_{\text{obs}} + \frac{1}{\sqrt{n}} \right]$.

Remarques :
1. Cela signifie que la proportion inconnue a plus de 95 % de chances de se trouver dans cet intervalle.
2. On admettra que l'intervalle $\left[f_{\text{obs}} - 1,96\frac{\sqrt{f_{\text{obs}}(1-f_{\text{obs}})}}{\sqrt{n}} \,;\, f_{\text{obs}} + 1,96\frac{\sqrt{f_{\text{obs}}(1-f_{\text{obs}})}}{\sqrt{n}} \right]$ est aussi intervalle de confiance de p au niveau de confiance de 95 %.

Exercices : 4, 5, 6 page 445 et 35, 36 page 456 [6] – 7, 9 page 446 [7] – 15 page 448 [8] – 46 page 459 [9] [TransMath]

4. Prise de décision.
5. Type BAC.
6. Estimer une proportion à partir d'un échantillon.
7. Taille de l'échantillon nécessaire pour une certaine précision.
8. Approximation de π.
9. Type BAC.

Références

[TransMath] transMATH Term S, programme 2012 (Nathan)

2, 3, 4