Échantillonnage et estimation

Terminale S
Lycée Charles PONCET

Mai 2013

Table des matières

Le symbole ✏ indique les exemples à traiter, des démonstrations à trouver.
Le symbole ☛ indique des points importants, des pièges possibles, des notations particulières, etc.

1 Introduction

Une urne contient 60 % de boules rouges. On prélève au hasard et avec remise 100 boules. Il y a équiprobabilité. On désigne par X le nombre de boules rouges prélevées.

1. a. Quelle est la loi suivie par X ? Quels sont ses paramètres.
   Calculer l'espérance mathématique et l'écart-type de X.
   b. Calculer, à 10^{-4} près, $P(X = 60)$, $P(X \leqslant 49)$, $P(X \leqslant 50)$, $P(X \geqslant 69)$ et $P(X \geqslant 70)$.
2. a. Exprimer en fonction de X la variable aléatoire F qui associe à chaque prélèvement de 100 boules la fréquence de boules rouges prélevées.
   b. On considère l'intervalle de fluctuation vu en seconde.
   Déterminer l'intervalle $I = \left[p - \dfrac{1}{\sqrt{n}} \,;\, p - \dfrac{1}{\sqrt{n}}\right]$ avec $n = 100$ et $p = 0{,}6$.
   Calculer, à 10^{-4} près, $P(F \in I)$.
3. a. On utilise l'intervalle de fluctuation vu en première.
   Déterminer le plus petit intervalle $J = \left[\dfrac{a}{n} \,;\, \dfrac{b}{n}\right]$ avec $n = 100$, où a et b sont des entiers naturels tels que $P\left(F < \dfrac{a}{n}\right) \leqslant 0{,}025$ et $P\left(F > \dfrac{b}{n}\right) \leqslant 0{,}025$.
   b. Que peut-on en déduire pour $P(F \in J)$?
4. Soit X_n la variable aléatoire qui suit la loi binomiale $\mathscr{B}(n\,;\,p)$.
   En utilisant le théorème de MOIVRE – LAPLACE, calculer à 10^{-4} près :

$$\lim_{n\to+\infty} P\left(-1{,}96 \leqslant \frac{X_n - np}{\sqrt{np(1-p)}} \leqslant 1{,}96\right).$$

2 Échantillonnage

2.1 Conséquence du théorème de MOIVRE – LAPLACE

Théorème 2.1.1
On considère un entier naturel n non nul et p un nombre réel de l'intervalle $]0\,;\,1[$.
Si la variable aléatoire X_n suit la loi binomiale $\mathscr{B}(n\,;\,p)$, alors, pour tout nombre réel α de l'intervalle $]0\,;\,1[$ on a :

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

le nombre réel u_α étant défini par $P(-u_\alpha \leqslant Y \leqslant u_\alpha) = 1 - \alpha$, où Y suit la loi normale centrée réduite.

☛ La variable aléatoire $F_n = \dfrac{X_n}{n}$ représente la *fréquence de succès* pour un schéma de BERNOULLI de paramètres n et p.

✏ Démontrer le théorème 2.1.1 en utilisant le théorème de MOIVRE – LAPLACE.

2.2 Intervalle de fluctuation

On considère un entier naturel n non nul et p un nombre réel de l'intervalle $]0\,;\,1[$.

Dans une certaine population, la proportion d'individus présentant le caractère C est p. Que peut-on dire de la fréquence f de C, sur un échantillon de taille n ?

2.2.1 Intervalle de fluctuation avec la loi binomiale (rappels)

Proposition 2.2.1 (et définition)
Si X_n est une variable aléatoire associant à chaque échantillon de taille n le nombre d'individus présentant le caractère C (de proportion p) et si $F_n = \dfrac{X_n}{n}$ est la variable aléatoire correspondant à la fréquence de C dans l'échantillon, alors :

- *X_n suit la loi binomiale $\mathscr{B}(n\,;\,p)$;*
- *on peut déterminer un intervalle $\left[\dfrac{a}{n}\,;\,\dfrac{b}{n}\right]$ avec a le plus grand entier naturel et b le plus petit entier naturel tels que $P\left(F_n < \dfrac{a}{n}\right) \leqslant 0{,}025$ (ou $P(X_n < a) \leqslant 0{,}025$) et $P\left(F_n > \dfrac{b}{n}\right) \leqslant 0{,}025$ (ou $P(X_n > b) \leqslant 0{,}025$).*

L'intervalle $\left[\dfrac{a}{n}\,;\,\dfrac{b}{n}\right]$ vérifie $P\left(F_n \in \left[\dfrac{a}{n}\,;\,\dfrac{b}{n}\right]\right) \geqslant 0{,}95$. Il est appelé intervalle de fluctuation *au seuil $0{,}95$ de F_n.*

2.2.2 Intervalle de fluctuation asymptotique

Définition 2.2.1
L'intervalle $I_n = \left[p - u_\alpha\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\,;\,p + u_\alpha\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ du théorème 2.1.1 est appelé intervalle de fluctuation asymptotique *au seuil de confiance $1-\alpha$ de la variable aléatoire $F_n = \dfrac{X_n}{n}$ qui, à tout échantillon de taille n, associe la fréquence obtenue.*

☛ L'intervalle de fluctuation asymptotique contient F_n avec une probabilité proche de $1-\alpha$, quand n est suffisamment grand.
On peut utiliser cette approximation lorsque $\boxed{n \geqslant 30,\ np \geqslant 5 \text{ et } n(1-p) \geqslant 5.}$

Proposition 2.2.2
Un intervalle de fluctuation asymptotique au seuil de confiance 95% de la fréquence F_n d'un caractère C dans un échantillon de taille n est :

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

où p est la proportion du caractère C dans la population.

☛ En effet, pour $\alpha = 0{,}05$ on a $1-\alpha = 0{,}95$ et $u_{0{,}05} = 1{,}96$.

Proposition 2.2.3
Un intervalle de fluctuation asymptotique au seuil de confiance 99% de la fréquence F_n d'un caractère C dans un échantillon de taille n est :

$$\left[p - 2{,}58\frac{\sqrt{p(1-p)}}{\sqrt{n}}\,;\,p + 2{,}58\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$$

où p est la proportion du caractère C dans la population.

☛ En effet, pour $\alpha = 0{,}01$ on a $1-\alpha = 0{,}99$ et $u_{0{,}01} = 2{,}58$.

Exemple

Dans une urne contenant 3 boules rouges et 7 boule bleues, on effectue 100 tirages avec remise. On suppose qu'il y a équiprobabilité.

On désigne par X le nombre de boules rouges obtenues et on pose $F = \dfrac{X}{100}$.

1. Donner la loi de probabilité de X.
2. Déterminer l'intervalle de fluctuation de la variable aléatoire F au seuil de 95 %.
3. Déterminer l'intervalle de fluctuation asymptotique de la variable aléatoire F au seuil de 95 %. Comparer avec le résultat obtenu à la question précédente.

3 Prise de décision à partir d'un échantillon

On cherche à savoir, au seuil de décision de 5 %, si la proportion p du caractère C dans la population vaut $p = p_0$ ou non, à partir d'un échantillon de taille n.

On suppose que $n \geqslant 30$, $np_0 \geqslant 5$ et $n(1-p_0) \geqslant 5$.

La procédure est la suivante :

- On détermine $I = \left[p_0 - 1{,}96\dfrac{\sqrt{p_0(1-p_0)}}{\sqrt{n}} \; ; \; p_0 + 1{,}96\dfrac{\sqrt{p_0(1-p_0)}}{\sqrt{n}} \right]$.
- On calcule la fréquence f du caractère C de l'échantillon de taille n.
- On applique la *règle de décision*, au seuil de 5 % :
  - si $f \notin I$, on rejette l'hypothèse $p = p_0$;
  - si $f \in I$, on accepte l'hypothèe $p = p_0$.

☛ D'après le théorème 2.1.1, la probabilité de rejeter à tort l'hypothèse $p = p_0$ est environ égale à 0,05. Le seuil de décision correspond à ce risque.

On peut utiliser d'autres seuils de décision, par exemple, 1 %, on utilise alors l'intervalle de la proposition 2.2.3.

Lorsque les critères $n \geqslant 30$, $np_0 \geqslant 5$ et $n(1-p_0) \geqslant 5$ ne sont pas vérifiés, on utilise l'intervalle de fluctuation définie avec la loi binomiale (cf. paragraphe 2.2.1).

Exemple

Selon la théorie de MENDEL, certaines cosses de petits pois devraient fournir des petits pois jaunes et verts dans les proportions de 75 % et 25 %. On souhaite tester l'hypothèse selon laquelle la proportion de petits pois jaunes est $p = 0{,}75$ en mettant en place une expérience permettant d'obtenir 224 petits pois considérés comme une échantillon aléatoire.

1. Sous l'hypothèse $p = 0{,}75$, déterminer l'intervalle de fluctuation asymptotique au seuil de 95 % de la variable aléatoire correspondant à la fréquence des petits pois jaunes sur un échantillon de taille 224 (arrondir les bornes de l'intervalle à 0,01).
2. Énoncer la règle de décision permettant de rejeter, ou non, l'hypothèse $p = 0{,}75$, au seuil de 5 % sur un échantillon aléatoire de taille 224.
3. L'expérience a permis d'obtenir 176 petits pois jaunes et 48 verts. Que peut-on en conclure ?
4. Un autre expérience donne 185 petits pois jaunes. Qu'en conclure cette fois ?

4 Estimation d'une population

4.1 Notion d'estimation

Théorème 4.1.1
Soit X_n *la variable aléatoire qui suit la loi binomiale* $\mathscr{B}(n\ ;\ p)$*, avec* n *un entier naturel non nul et* p *un nombre de l'intervalle* $]0\ ;\ 1[$*. On pose* $F_n = \dfrac{X_n}{n}$*.*
Il existe un entier naturel n_0 *non nul tel que si* $n \geqslant n_0$ *alors* $P\left(p - \dfrac{1}{\sqrt{n}} \leqslant F_n \leqslant p + \dfrac{1}{\sqrt{n}}\right) \geqslant 0{,}95$.

Corollaire 4.1.2
Si F_n *est la variable aléatoire qui associe à chaque échantillon de taille* n *la fréquence d'un caractère* C *extrait d'une population dont la proportion du caractère* C *est* p *alors, pour* n *assez grand :*

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

4.2 Intervalle de confiance

Définition 4.2.1
On considère n *un entier naturel non nul et* f *la fréquence d'un caractère* C *sur un échantillon de taille* n*.*
L'intervalle $\left[f - \dfrac{1}{\sqrt{n}}\ ;\ f + \dfrac{1}{\sqrt{n}}\right]$ *est un* intervalle de confiance à 95 % *de la proportion inconnue* p *dans la population.*

☛ L'intervalle de confiance dépend de l'échantillon utilisé, mais pas de la taille de la population.

La précision de l'intervalle de confiance, donnée par sa longueur, est $\dfrac{2}{\sqrt{n}}$.

On peut parfois utiliser un intervalle de confiance à 95 % plus précis :

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

Exemple

Dans une urne contenant des boules rouges et des boules 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 0,02 (c'est-à-dire une précision d'au moins 0,02) ?