# Échantillonnage et estimation, cours, terminale S

F.Gaudon

31 mai 2018

## Table des matières

# 1 Distinction entre échantillonnage et estimation

**Définition :**

On considère une population d'individus.

- Lorsque l'on connaît ou lorsque l'on fait une hypothèse sur la proportion $p$ d'individus ayant une caractéristique donnée dans une population et que l'on effectue un nombre $n$ de tirages avec remise dans cette population, la fréquence observée appartient avec une certaine probabilité à un intervalle appelé *intervalle de fluctuation* de centre $p$ et de longueur qui diminue lorsque $n$ augmente. On parle alors de situation *d'échantillonnage*.
- Lorsque l'on ne connaît pas la proportion d'individus ayant une caractéristique donnée, en procédant à un nombre $n$ de tirages avec remise on peut estimer à l'aide de la fréquence $f$ obtenue la proportion $p$ d'individus ayant cette caractéristique. Cette *estimation* se fait à l'aide d'un *intervalle de confiance* dont l'amplitude diminue lorsque le nombre $n$ de tirages augmente.

# 2 Échantillonnage

**Propriété et définition :**

Soit $X_n$ une variable aléatoire suivant une loi binomiale $\mathcal{B}(n;p)$ et $\alpha$ un réel tel que $0 < \alpha < 1$. Si $X$ est une variable aléatoire suivant la loi normale centrée réduite $\mathcal{N}(0;1)$, on appelle $u_\alpha$ l'unique réel tel que $P(-u_\alpha \leq X \leq u_\alpha) = 1-\alpha$.
Alors

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

c'est à dire que l'intervalle $I_n = \left[p - u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ contient la fréquence $\frac{X_n}{n}$ avec une probabilité qui se rapproche de $1-\alpha$ quand $n$ augmente. On dit que c'est un *intervalle de fluctuation asymptotique* au seuil $1-\alpha$.
En particulier, l'intervalle de fluctuation asymptotique au seuil de 95% est :

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

**Preuve :**

Puisque $X_n \sim \mathcal{B}(n;p)$, on a $E(X_n) = np$ et $\sigma(X_n) = np(1-p)$. On pose donc $Z_n = \frac{X_n - np}{\sqrt{np(1-p)}}$. On sait d'après le théorème de De Moivre-Laplace que

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

c'est à dire que

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

ou encore

$$\lim_{n\to+\infty} P(-u_\alpha\sqrt{np(1-p)} \leq X_n - np \leq u_\alpha\sqrt{np(1-p)}) = 1-\alpha$$

donc

$$\lim_{n\to+\infty} P(np - u_\alpha\sqrt{np(1-p)} \leq X_n \leq np + u_\alpha\sqrt{np(1-p)}) = 1-\alpha$$

d'où

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

Le cas particulier est obtenu en se rappelant que $u_{0,05} = 1,96$.

**Remarque :**
En pratique, on utilise cette propriété dès que les conditions $n \geq 30$, $np \geq 5$ et $n(1-p) \geq 5$ sont vérifiées.

**Test d'hypothèse :**
On considère une population dans laquelle on suppose que la proportion d'un caractère est $p$.On fait l'hypothèse « La proportion dans la population est $p$ ». On observe la fréquence $f$ d'apparition de ce caractère sur un échantillon de taille $n$ et on calcule l'intervalle de fluctuation asymptotique $I$ au seuil de 95%.

- Si $f \in I$, au risque de 5% d'erreur (ou au seuil de confiance de 95%), on accepte l'hypothèse que la proportion dans la population est $f$.
- Si $f \in I$, au risque de 5% d'erreur on rejette l'hypothèse que la proportion dans la population est $p$.

**Exemple :**
Un fournisseur d'accès à l'internet affirme que, sur sa hotline, seuls 20% des clients attendent plus de 5 minutes pour obtenir un interlocuteur. Une association de consommateurs interroge au hasard 200 personnes ayant eu à s'adresser à cette hotline. Parmi ces personnes, 53 ont dû attendre plus de 5 minutes. Peut-on mettre en doute l'affirmation du fournisseur d'accès ?

L'hypothèse à tester est « 20% des clients attendent plus de 5 minutes ».
$\frac{53}{200} = 0,265$ et $I = [0,2 - 1,96 \times \frac{\sqrt{0,2\times0,8}}{\sqrt{200}}; 0,2 + 1,96 \times \frac{\sqrt{0,2\times0,8}}{\sqrt{200}}] = [0,144; 0,255]$.
Or $0,265 \notin I$ donc au seuil de confiance de 95%, on rejette l'affirmation du fournisseur d'accès.

# 3 Estimation

**Propriété et définition :**

Soit $X_n$ une variable aléatoire suivant la loi binomiale $\mathcal{B}(n;p)$ et $F_n = \frac{X_n}{n}$ où $p$ est la proportion inconnue d'apparition d'un caractère.

- Alors, pour $n$ assez grand, $P(p \in [F_n - \frac{1}{\sqrt{n}}; F_n + \frac{1}{\sqrt{n}}]) \geq 0,95$.
- Si on appelle $f$ la fréquence d'apparition du caractère sur un échantillon de taille $n$, Alors l'intervalle $[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}]$ est appelé *intervalle de confiance de la proportion p inconnue au niveau de confiance 0,95*.

**Idée de la preuve :**

On a vu que

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

On montre d'abord que $[p-u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}};p+u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}}]$ est inclus dans $[p-\frac{1}{\sqrt{n}};p+\frac{1}{\sqrt{n}}]$.

Pour cela, on pose $f(p)=1,96\sqrt{p(1-p)}$ pour tout $p\in[0;1]$. $f$ est dérivable sur $]0;1[$ et $f'(p)=1,96\frac{-2p+1}{2\sqrt{p(1-p}}$ qui est du signe de $-2p+1$. D'où

| $p$ | 0 | | $\frac{1}{2}$ | | 1 |
|---|---|---|---|---|---|
| $f'(p)$ | | + | 0 | - | |
| $f(p)$ | 0 | ↗ | 0,98 | ↘ | 0 |

D'où pour tout $p\in[0;1]$, $0\leq f(p)<1$ ce qui, lorsque $p-u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq F_n \leq p+u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}}$ implique :

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

D'où $p>F_n-\frac{1}{\sqrt{n}}$ et $p<F_n+\frac{1}{\sqrt{n}}$ donc :

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

Par une étude sur les limites on obtient alors l'inégalité voulue.

**Remarque :**

Un intervalle de confiance au niveau de 95% a une amplitude de $\frac{2}{\sqrt{n}}$. L'amplitude diminue lorsque la taille $n$ de l'échantillon augmente.

**Exemple :**

Un candidat à une élection municipale fait effectuer un sondage. Sur 100 personnes de la ville interrogées, 63 déclarent vouloir voter pour lui.

$I=[0,63-\frac{1}{\sqrt{100}};0,63+\frac{1}{\sqrt{100}}]=[0,53;0,73]$

On peut donc estimer que la proportion de personnes dans la ville voulant voter pour lui est comprise dans l'intervalle I=[0,53 ;073].