Chapitre 20 : Échantillonnage

I Rappels des années précédentes

I.1 Notion d'intervalle de fluctuation d'une fréquence

On nomme p la proportion d'un caractère dans une population.
Soit n un entier strictement positif.
Soit X_n la variable aléatoire qui à chaque échantillon de taille n prélevé dans cette population associe le nombre d'individus possédant le caractère étudié.
La variable aléatoire X_n suit la loi binomiale $\mathscr{B}(n;p)$.
La variable aléatoire fréquence associée est $F_n = \dfrac{X_n}{n}$.

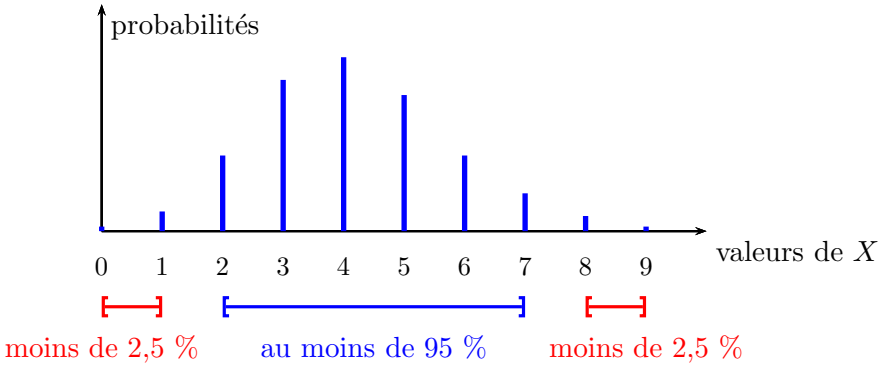
Définition
Soit α un réel de $]0;1[$.
Tout intervalle I tel que $P(F_n \in I) \geqslant 1-\alpha$ est un **intervalle de fluctuation de la fréquence F_n au seuil de $1-\alpha$.**

Cas $\alpha = 0,05$:
Dire qu'un intervalle I est un intervalle de fluctuation de la fréquence F_n au seuil de _____ % signifie que :
. . . . . .

I.2 Intervalle de fluctuation vu en seconde

Propriété (intervalle de fluctuation d'une fréquence)
Soit un caractère dont la proportion dans une population donnée est p.
Lorsque $n \geqslant 25$ et $0,2 \leqslant p \leqslant 0,8$, au moins 95 % des échantillons de taille n issus de cette population sont tels que la fréquence f du caractère dans l'échantillon appartient à l'intervalle $\left[p - \dfrac{1}{\sqrt{n}}; p + \dfrac{1}{\sqrt{n}}\right]$.
L'intervalle $I = \left[p - \dfrac{1}{\sqrt{n}}; p + \dfrac{1}{\sqrt{n}}\right]$ est appelé intervalle de fluctuation au seuil de 95 %.

I.3 Intervalle de fluctuation associé à la loi binomiale, 1ère S



Définition
Soit X une variable aléatoire suivant la loi binomiale $\mathcal{B}(n;p)$.
On note f la fréquence associée à un échantillon aléatoire de taille n de X.
L'intervalle de fluctuation au seuil de 95 % de la fréquence f est l'intervalle $\left[\dfrac{a}{n}; \dfrac{b}{n}\right]$, défini par :
- a est le plus petit entier tel que $P(X \leqslant a) > 0,025$,
- b est le plus petit entier tel que $P(X \leqslant b) \geqslant 0,975$.

Remarque

1. On a donc $P(a \leqslant X \leqslant b) \geqslant 0,95$.
2. Cet intervalle de fluctuation s'applique à des échantillons de variables aléatoires suivant une loi binomiale, et ce quelles que soient les valeurs de n et p, contrairement à l'intervalle de fluctuation vu en seconde.
3. L'intervalle de fluctuation $\left[\frac{a}{n}; \frac{b}{n}\right]$ n'est pas nécessairement centré sur p.

   Il n'y a pas de formule donnant directement les bornes $\frac{a}{n}$ et $\frac{b}{n}$.

Exercice 1

Monsieur Z, chef du gouvernement d'un pays lointain, affirme que 52 % des électeurs lui font confiance. On interroge 100 électeurs au hasard (la population est suffisamment grande pour considérer qu'il s'agit de tirages avec remise) et on souhaite savoir à partir de quelles fréquences, au seuil de 95 %, on peut mettre en doute le pourcentage annoncé par Monsieur Z, dans un sens, ou dans l'autre.

1. On fait l'hypothèse que Monsieur Z dit vrai et que la proportion des électeurs qui lui font confiance dans la population est $0,52$. Montrer que la variable aléatoire X, correspondant au nombre d'électeurs lui faisant confiance dans un échantillon de 100 électeurs, suit la loi binomiale de paramètres $n = 100$ et $p = 0,52$.
2. On donne ci-contre un extrait de la table des probabilités cumulées $P(X \leqslant k)$ où X suit la loi binomiale de paramètres $n = 100$ et $p = 0,52$.

k	$P(X \leqslant k) \approx$
40	0,0106
41	0,0177
42	0,0286
43	0,0444
...	...
61	0,9719
62	0,9827
63	0,9897
64	0,9897

   Déterminer a et b tels que :
   a est le plus petit entier tel que $P(X \leqslant a) > 0,025$;
   b est le plus petit entier tel que $P(X \leqslant b) \geqslant 0,975$.
3. En déduire l'intervalle de fluctuation $\left[\frac{a}{n}; \frac{b}{n}\right]$ au seuil de 95 % pour une fréquence f de personnes lui faisant confiance.
4. Comparer cet intervalle à l'intervalle de fluctuation $\left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$ vu en seconde.
5. Sur les 100 électeurs interrogés au hasard, 43 déclarent avoir confiance en Monsieur Z. Peut-on considérer, au seuil de 95 %, l'affirmation de Monsieur Z comme exacte ?

Exercice 2 (sécurité au carrefour)

Un groupe de citoyens demande à la municipalité d'une ville la modification d'un carrefour en affirmant que 40 % des automobilistes tournent en utilisant une mauvaise file.
Un officier de police constate que sur 500 voitures prises au hasard, 190 prennent une mauvaise file.

1. Déterminer, en utilisant la loi binomiale sous l'hypothèse $p = 0,4$, l'intervalle de fluctuation au seuil de 95 %.
2. D'après l'échantillon, peut-on considérer, au seuil de 95 %, comme exacte l'affirmation du groupe de citoyens ?

II Théorème de Moivre-Laplace

Rappel :
Si X suit la loi binomiale $\mathscr{B}(n;p)$, alors $E(X) = np$, et $\sigma(X) = \sqrt{np(1-p)}$.

Théorème (Théorème de Moivre-Laplace)
Soit $p \in]0;1[$. On suppose que pour tout entier non nul n, la variable X_n suit la loi binomiale de paramètres n et p.
Soit Z_n la variable aléatoire $Z_n = \dfrac{X_n - np}{\sqrt{np(1-p)}}$.
Alors, pour tous réels a et b tels que $a < b$,

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

III Intervalle de fluctuation asymptotique

Rappel (relatif à la loi normal centrée réduite) :
Si X suit la loi normale centrée réduite $\mathscr{N}(0;1)$, alors pour tout réel $\alpha \in]0;1[$, il existe un unique réel strictement positif u_α tel que $P(-u_\alpha \leqslant X \leqslant u_\alpha) = 1-\alpha$.
Soit X_n une variable aléatoire suivant la loi binomiale $\mathscr{B}(n;p)$. On note F_n la variable aléatoire $F_n = \dfrac{X_n}{n}$ correspondant à la fréquence de succés dans la répétition indépendante de n épreuves de Bernoulli de paramètre p.

Propriété
Si la variable aléatoire X_n suit la loi binomiale $\mathscr{B}(n;p)$, alors, pour tout $\alpha \in]0;1[$, $\lim\limits_{n\to+\infty} P\left(\dfrac{X_n}{n} \in I_n\right) = 1-\alpha$ où I_n est l'intervalle $\left[p - u_\alpha\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}; p + u_\alpha\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$

Démonstration (à connaître)

$$\begin{aligned}
\frac{X_n}{n} \in I_n &\Leftrightarrow p - u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant \frac{X_n}{n} \leqslant p + u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}} \\
&\Leftrightarrow np - u_\alpha\frac{n\sqrt{p(1-p)}}{\sqrt{n}} \leqslant X_n \leqslant np + u_\alpha\frac{n\sqrt{p(1-p)}}{\sqrt{n}} \\
&\Leftrightarrow np - u_\alpha\sqrt{np(1-p)} \leqslant X_n \leqslant np + u_\alpha\sqrt{np(1-p)} \\
&\Leftrightarrow -u_\alpha \leqslant Z_n \leqslant u_\alpha
\end{aligned}$$

en posant $Z_n = \dfrac{X_n - np}{\sqrt{np(1-p)}}$.
D'après le théorème de Moivre-Laplace, $P(-u_\alpha \leqslant Z_n \leqslant u_\alpha)$ tend vers $\displaystyle\int_{-u_\alpha}^{u_\alpha} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x$ lorsque n tend vers $+\infty$.
Donc $\lim\limits_{n\to+\infty} P\left(\dfrac{X_n}{n} \in I_n\right) = \lim\limits_{n\to+\infty} P(-u_\alpha \leqslant Z_n \leqslant u_\alpha) = \displaystyle\int_{-u_\alpha}^{u_\alpha} \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-\frac{x^2}{2}}\,\mathrm{d}x$.
Or, cette dernière intégrale est égale à $P(-u_\alpha \leqslant Y \leqslant u_\alpha)$ où Y est une variable aléatoire suivant la loi normale centrée réduite $\mathscr{N}(0;1)$.
Comme $P(-u_\alpha \leqslant Y \leqslant u_\alpha) = 1-\alpha$ (par définition de u_α), on a montré que

$$\lim_{n\to+\infty} P\left(\frac{X_n}{n} \in I_n\right) = 1-\alpha. \qquad \square$$

Définition
On dit que l'intervalle $I_n = \left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ est un intervalle de fluctuation asymptotique au seuil de confiance $1-\alpha$ de la variable aléatoire $F_n = \frac{X_n}{n}$.

Remarque

1. Cet intervalle contient les fréquences $F_n = \frac{X_n}{n}$ avec une probabilité qui tend vers $1-\alpha$ lorsque n tends vers l'infini.
2. Il est toujours centré sur p.
3. On admet que pour $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$, on peut approcher $P\left(\frac{X_n}{n} \in I_n\right)$ par $1-\alpha$.

   On convient donc d'utiliser l'intervalle de fluctuation asymptotique lorsque les conditions suivantes sont remplies :
   - $n \geqslant 30$,
   - $np \geqslant 5$,
   - $n(1-p) \geqslant 5$.

Dans le cas où $\alpha = 0,05$, $1-\alpha = 0,95$ et on a vu que $u_\alpha \approx 1,96$. On en déduit un intervalle de fluctuation asymptotique au seuil de 95 %.

Propriété
L'intervalle de fluctuation asymptotique au seuil de 95 % de la fréquence F_n d'un caractère dans un échantillon de taille n est :

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

III.1 Détermination pratique de l'intervalle de fluctuation au seuil de 95 %

— Si $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$, alors on utilise l'intervalle de fluctuation asymptotique

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

— Sinon, on utilise l'intervalle $\left[\frac{a}{n}; \frac{b}{n}\right]$ vu en première, où :
- a est le plus petit entier tel que $P(X \leqslant a) > 0,025$,
- b est le plus petit entier tel que $P(X \leqslant b) \geqslant 0,975$.

III.2 Autres seuils possibles

1. Intervalle de fluctuation au seuil de 99 % (avec un risque de 1%).
   $u_{0,01} \approx 2,58$, ce qui donne l'intervalle

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

2. Cas général : intervalle de fluctuation au seuil $1-\alpha$ (risque α) :

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

   où u_α est tel que $P(-u_\alpha \leqslant X \leqslant u_\alpha) = 1-\alpha$ avec X suivant $\mathcal{N}(0;1)$.

IV Prise de décision à partir d'un échantillon

Théorème (prise de décision)
On fait une hypothèse sur la valeur de p (proportion d'un caractère dans la population).
On calcule la fréquence observée f du caractère étudié dans un échantillon de taille n.
Après s'être assuré des conditions d'approximation liées à l'intervalle de fluctuation asymptotique, on détermine celui-ci.

- Si la fréquence observée f n'appartient pas à l'intervalle de fluctuation asymptotique au seuil de 95 %, alors on rejette l'hypothèse faite sur p, avec un risque de 5 % (de rejeter à tort une hypothèse vraie).
- si la fréquence observée f appartient à l'intervalle de fluctuation asymptotique, alors on ne peut pas rejeter l'hypothèse faite sur p (éviter de dire qu'on l'accepte).

Remarque

1. Dans le cas où $f \in I_n$, le risque d'erreur n'est pas quantifié.
2. le risque de 5% signifie que la probabilité que l'on rejette à tort l'hypothèse faite sur la proportion p alors qu'elle est vraie est **approximativement** égale à 5%. C'est une probabilité conditionnelle.

V Intervalle de fluctuation simplifié

Propriété
Soit $p \in]0;1[$.
Soit, pour tout entier $n > 0$, X_n une variable aléatoire suivant la loi binomiale $\mathscr{B}(n;p)$.
Alors, il existe un entier n_0 tel que pour tout $n \geqslant n_0$,

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

Démonstration
Soit p un nombre réel de $]0;1[$.
Soit, pour tout n entier strictement positif, une variable aléatoire X_n suivant une loi binomiale de paramètres n et p.
Posons, pour tout n de $\mathbb{N}^*$, $J_n = \left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$.
Soit, pour tout n de $\mathbb{N}^*$, Z_n la variable aléatoire centrée réduite associée à X_n.
Rappel : $Z_n = \dfrac{X_n - np}{\sqrt{np(1-p)}}$.
Posons, pour tout n de $\mathbb{N}^*$, $I_n = \left[p - 2\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 2\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$.

1. Soit (a_n) la suite définie par, pour tout n de $\mathbb{N}^*$, $a_n = P\left(\frac{X_n}{n} \in I_n\right)$.
   (a) Démontrer que, pour tout n de $\mathbb{N}^*$, $a_n = P(-2 \leqslant Z_n \leqslant 2)$.
   (b) Démontrer que la suite (a_n) est convergente et que sa limite ℓ vérifie : $\ell \approx 0,954$ à 10^{-3} près par défaut.
   (c) Justifier qu'il existe un entier strictement positif n_0 tel que, si l'entier n vérifie $n \geqslant n_0$, alors $a_n \geqslant 0,95$.
2. (a) Déterminer le maximum de la fonction g définie sur $[0;1]$ par $g(x) = \sqrt{x(1-x)}$.
   (b) Montrer que, pour tout entier strictement positif n, $I_n \subset J_n$.
3. Conclure.

□

VI Estimation d'une proportion

La proportion p de caractère dans la population est inconnue.
On essaie d'**estimer** p à partir de la fréquence d'un échantillon.

Propriété
Soit p un nombre réel de $]0;1[$.
Soit, pour tout n entier strictement positif, une variable aléatoire X_n suivant la loi binomiale $\mathscr{B}(n;p)$ et $F_n = \dfrac{X_n}{n}$.
Il existe un entier strictement positif n_0 tel que, pour tout $n \geqslant n_0$,

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

Démonstration
$F_n = \dfrac{X_n}{n}$ et d'après la propriété précédente, pour n assez grand,

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

Or,

$$\begin{aligned} F_n \in \left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right] &\Leftrightarrow p - \frac{1}{\sqrt{n}} \leqslant F_n \leqslant p + \frac{1}{\sqrt{n}} \\ &\Leftrightarrow F_n - \frac{1}{\sqrt{n}} \leqslant p \leqslant F_n + \frac{1}{\sqrt{n}} \end{aligned}$$

Donc, pour n assez grand, $P\left(F_n - \dfrac{1}{\sqrt{n}} \leqslant p \leqslant F_n + \dfrac{1}{\sqrt{n}}\right) \geqslant 0,95$. □

Définition
Soit f la fréquence d'un caractère observée sur un échantillon de taille n d'une population dans laquelle la proportion du caractère est p.
Alors l'intervalle $\left[f - \dfrac{1}{\sqrt{n}}; f + \dfrac{1}{\sqrt{n}}\right]$ est un intervalle de confiance de la proportion p au seuil de 95 %.

Remarque
On utilise cet intervalle dès que $n \geqslant 30$, $nf \geqslant 5$ et $n(1-f) \geqslant 5$.