Chapitre 8 FLUCTUATION ET ESTIMATION TES

« [...] Dans le petit nombre de choses que nous pouvons savoir avec certitude [...], les principaux moyens de parvenir à la vérité [...] se fondent sur les probabilités. »

P.S. de Laplace

Dans ce chapitre, on s'intéresse à un caractère dans une population donnée dont la proportion est p. Cette **proportion** sera dans quelques cas **connue** (intervalle de fluctuation), dans certains cas **supposée connue** (prise de décision) et dans d'autres cas **inconnue** (estimation).

I Variable aléatoire fréquence

Pour des raisons généralement économiques, on étudie le caractère, non pas sur la population entière, mais sur des échantillons de taille n extraits de cette population. Pour ce faire, on peut prélever au hasard des individus de cette population un par un avec remise.
On note X_n la variable aléatoire égale au nombre de personne possédant le caractère étudié parmi les personnes de l'échantillon de taille n.
X_n suit donc la loi binomiale de paramètres n et p.

Définition La variable aléatoire égale à la fréquence du caractère étudiée dans un échantillon de taille n se note F_n et s'appelle la variable aléatoire fréquence. On a la relation :

$$F_n = \frac{X_n}{n}$$

Remarque

X_n suit précisément la loi binomiale uniquement si l'on effectue des tirages avec remise. Or dans certains cas, par exemple pour un sondage, on ne souhaite pas interroger deux fois la même personne. X_n ne devrait donc plus suivre la loi binomiale. Toutefois si la taille de l'échantillon n'excède pas 10% de la population totale, on pourra dire que la loi de X_n est très proche de la loi binomiale. On supposera dans tout ce cours que cette hypothèse est vérifiée.

On supposera donc : $n \geq 30$ $\quad n \times p \geq 5$ $\quad n \times (1-p) \geq 5$

II Intervalle de fluctuation et prise de décision

2.1 Intervalle de fluctuation asymptotique au seuil de 95 %

Dans ce paragraphe, on suppose que la proportion p du caractère étudié est **connue**.

Propriété Soit $p \in]0;1[$ et $n \in \mathbb{N}^*$ vérifiant les hypothèses indiquées dans la partie I.
La probabilité que la variable aléatoire fréquence F_n appartienne à l'intervalle
$\left[p - 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ se rapproche de 0,95 lorsque n devient très grand.
On note $\lim\limits_{n\to+\infty} P\left(p - 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq F_n \leq p + 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) = 0{,}95$

Définition L'intervalle $[p - 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}]$ s'appelle **l'intervalle de fluctuation asymptotique au seuil 0,95 (ou 95 %)** de la variable aléatoire fréquence.

Remarque

En classe de seconde, l'intervalle de fluctuation étudié était l'intervalle $\left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}\right]$. Le nouvel intervalle que nous venons de définir est « de meilleur qualité ».

En effet : Soit f la fonction définie sur $]0;1[$ par $f(p) = p(1-p)$. f y est dérivable et $f'(p) = 1 - 2p$. f admet un maximum égal à ¼ pour p=1/2.
Donc, pour tout $p \in]0;1[, 0 < p(1-p) \leq \frac{1}{4}$ et donc $\sqrt{p(1-p)} \leq \frac{1}{2}$, d'où :
$1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq 1{,}96 \times \frac{0{,}5}{\sqrt{n}} \leq \frac{1}{\sqrt{n}}$ (car 1,96*0,5=0,98).
D'où $[p - 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}]$ est bien inclus dans $[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}}]$.

Exercice

Les acariens sont responsables de 70% des asthmes allergiques de l'enfant. On choisit au hasard un groupe de 300 enfants asthmatiques.

1. Déterminer l'intervalle de fluctuation asymptotique au seuil de 95% de la fréquence des asthmes dus aux acariens chez l'enfant.
2. Déterminer le nombre d'enfants présentant un asthme allergique dû aux acariens que l'on peut obtenir avec une probabilité proche de $0,95$.

Solution

1. *On a ici $p = 0,7$ et $n = 300$. L'intervalle de fluctuation asymptotique au seuil de 95% est donc :*

$$I = \left[p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right] \approx [0,648; 0,752]$$

2. *L'intervalle obtenu dans la question précédent nous permet de dire que $P(0,648 \leqslant F_n \leqslant 0,752) \approx 0,95$.*
   Dans cette question on cherche le nombre d'enfant et non plus la fréquence. Il faut donc multiplier la fréquence par le nombre total : 300.
   Ainsi on peut dire avec une probabilité proche de 0,95, qu'entre 194 et 226 enfants présentent un asthme allergique dû aux acariens.

2.2 Prise de décision

Dans ce paragraphe, la proportion du caractère **est supposée** être égale à une valeur p donnée dans l'énoncé du problème.

La prise de décision consiste, à partir d'un échantillon de taille n dans lequel on peut compter le nombre d'individu présentant le caractère étudié, à valider ou rejeter l'hypothèse faite sur la valeur de la proportion p.

Méthode

- On calcule la fréquence observée f du caractère étudié dans l'échantillon de taille n.
- Si les conditions sur les paramètres n et p sont vérifiées ($n \geqslant 30$, $n \times p \geqslant 5$ et $n \times (1-p) \geqslant 5$) on calcule l'intervalle de fluctuation asymptotique au seuil de $0,95$ de la fréquence à l'aide de la formule donnée dans le paragraphe précédent.
- On applique la **règle de décision** suivante :

Propriété

• Si la fréquence observée f appartient à l'intervalle de fluctuation asymptotique au seuil de $0,95$, on accepte l'hypothèse faite sur la proportion p.

• Si la fréquence observée f n'appartient pas à l'intervalle de fluctuation asymptotique au seuil de $0,95$, on rejette l'hypothèse faite sur la proportion p avec un risque de 5% de se tromper.

Remarque

Lorsqu'on accepte l'hypothèse faite sur p on peut pas donner la probabilité de se tromper.

Exemple

A la réception d'un nombre important de sacs de fèves de cacao (chaque sac pèse soit 25 kg soit 40 kg), un chocolatier demande de peser 50 sacs choisis au hasard : 35 ont un poids de 40 kg. La commande effectuée spécifie qu'il doit y avoir 60% de sacs de 40 kg. Ainsi d'après la commande la proportion de sacs de 40 kg est supposée être égale à 0,6.

Le chocolatier doit-il accepter la livraison ?

- *La fréquence observée du nombre de sac de 40 kg dans l'échantillon pesé est égale à* $f = \dfrac{35}{50} = 0,7$.
- *On a ici* $n = 50$ *et* $p = 0,6$. *On a donc bien :*

$$n \geqslant 30 \qquad np = 30 \geqslant 5 \qquad n(1-p) = 20 \geqslant 5$$

L'intervalle de fluctuation asymptotique au seuil de 0,95 de la fréquence des sacs de 40 kg est donc pour cette taille d'échantillon :

$$I = \left[p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right] \approx [0,46; 0,74]$$

- *Comme* $f = 0,7 \in I$ *la règle de décision nous dit que l'on peut accepter l'hypothèse faite sur la valeur de p. Le chocolatier peut donc accepter ce lot. La règle de décision ne nous dit pas quelle est sa probabilité de se tromper.*

Remarque Cette règle de décision peut aussi permettre, lorsqu'on connaît la valeur de p de dire si un échantillon donné est représentatif ou non de la population.

III Estimation et intervalle de confiance

Dans cette partie on suppose que la proportion du caractère étudiée est **inconnue**.

Définition Une **estimation ponctuelle** de la proportion p est la valeur de la fréquence observée du caractère étudié dans un échantillon donné.

Définition Un **intervalle de confiance pour p au niveau de confiance 0,95 (ou 95 %)** est un intervalle I tel que $P(p \in I) \geq 0{,}95$.

On dispose d'un échantillon de taille n connue dans lequel on peut calculer la fréquence d'apparition du caractère étudiée que l'on note f.

Propriété
Si les conditions : $n \geqslant 30$, $n \times f \geqslant 5$ et $n \times (1-f) \geqslant 5$ sont vérifiée alors un intervalle de confiance pour la proportion p au niveau de confiance 0,95 est l'intervalle

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

Remarque
On peut vous demander deux choses différentes :
- Si on connait n et f on peut vous demander de donner l'intervalle de confiance.
- On peut aussi vous demander quelle taille d'échantillon il faut prendre pour obtenir avec une précision donnée une estimation de la proportion cherchée.

Voici deux exemples illustrant ces deux questions.

Exemple 1
Un grossiste maraîcher vient de recevoir deux tonnes et demi de pommes de terre supposées être de calibre 35-55 en sac de 25 kg. Il décide de choisir au hasard dans chaque sac une pomme de terre et d'en vérifier le calibre : sur les 100 pommes de terre testées, 17 ne sont pas au bon calibre.

On s'intéresse à la proportion p de pommes de terre qui ne sont pas au bon calibre.

Donner un intervalle de confiance au niveau de confiance 0,95 de cette proportion.

On a ici $n = 100$ *et* $f = \dfrac{17}{100} = 0{,}17$.

D'après le cours, comme $n = 100 \geqslant 30$, $nf = 17 \geqslant 5$ *et* $n(1-f) = 83 \geqslant 5$, *un intervalle de confiance au niveau de confiance 0,95 pour* p *est*

$$I = \left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}\right] \approx [0{,}07; 0{,}27]$$

Ainsi grâce à son échantillon de 100 pommes de terre le grossiste est sûr à 95 % que la proportion de pommes de terre mal calibrées est comprise entre $0{,}07$ *et* $0{,}27$.

Exemple 2
Dans l'exemple précédent, quelle quantité de pommes de terre faudrait-il tester pour avoir une estimation de p à 0,1 près au niveau de confiance 0,95 ?

Ce que l'énoncé impose ici c'est en fait l'amplitude de l'intervalle de confiance de l'on souhaite obtenir.

Or l'amplitude de l'intervalle de confiance du cours est égale à :

$$f + \frac{1}{\sqrt{n}} - \left(f - \frac{1}{\sqrt{n}}\right) = f + \frac{1}{\sqrt{n}} - f + \frac{1}{\sqrt{n}} = \frac{2}{\sqrt{n}}$$

On cherche donc la valeur de n telle que :

$$\frac{2}{\sqrt{n}} \leqslant 0{,}1 \Leftrightarrow 2 \leqslant 0{,}1 \times \sqrt{n} \Leftrightarrow 20 \leqslant \sqrt{n} \Rightarrow 400 \leqslant n$$

Il faut donc tester au moins 400 pommes de terres pour avoir ensuite un intervalle de confiance pour p *au niveau de confiance 0,95 d'amplitude inférieure à 0,1.*