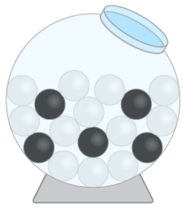
FLUCTUATION ET ESTIMATION

Le mathématicien d'origine russe *Jerzy Neyman* (1894 ; 1981), ci-contre, pose les fondements d'une approche nouvelle des statistiques. Avec l'anglais *Egon Pearson*, il développe la théorie de l'estimation et de la prise de décision sur un échantillon.
Ses travaux trouveront rapidement des applications dans de nombreux domaines concrets, tels la médecine, l'astronomie ou la météorologie.

Dans ce chapitre, on va étudier deux domaines des statistiques qu'il faut savoir distinguer :

Echantillonnage – Prise de décision	**Estimation**
- Une urne contient un très grand nombre de boules blanches et de boules noires dont **on connaît la proportion** *p* de boules blanches. On tire avec remise *n* boules (échantillon) et on observe la fréquence d'apparition des boules blanches. Cette fréquence observée appartient à un intervalle, appelé **intervalle de fluctuation** de centre *p*. - Dans le cas où on ne connaît pas la proportion *p* mais on est capable de faire une hypothèse sur sa valeur, on parle de **prise de décision**. On veut par exemple savoir si un dé est bien équilibré. On peut faire l'hypothèse que l'apparition de chaque face est égale à 1/6 et on va tester cette hypothèse à l'aide d'une expérience. Le résultat de l'expérience va nous permettre d'accepter ou rejeter l'hypothèse de départ.	Une urne contient un très grand nombre de boules blanches et de boules noires dont **on ignore la proportion** *p* de boules blanches. On tire avec remise *n* boules dans le but d'estimer la proportion *p* de boules blanches. On obtient ainsi une fréquence d'apparition qui va nous permettre d'estimer la proportion p à l'aide d'un **intervalle de confiance**. 

Conditions sur les paramètres : Dans tout le chapitre, sauf mention contraire, la taille de l'échantillon n et la proportion p du caractère étudié dans la population vérifient :

$$n \geq 30\,,\ n \times p \geq 5 \text{ et } n \times (1-p) \geq 5.$$

I. Echantillonnage

1) Intervalle de fluctuation asymptotique

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

Exemple :
On dispose d'une urne contenant un grand nombre de boules blanches et noires. La proportion de boules blanches contenues dans l'urne est *p* = 0,3.
On tire successivement avec remise $n = 50$ boules.
Soit X_{50} la variable aléatoire dénombrant le nombre de boules blanches tirées.

X_{50} suit la loi binomiale $B(50\,;0{,}3)$.

En effectuant 50 tirages dans cette urne, on va prouver dans ce chapitre que la fréquence d'apparition d'une boule blanche est comprise dans l'intervalle [0,173 ; 0,427] avec une probabilité de 0,95.

Cette intervalle s'appelle l'intervalle de fluctuation asymptotique au seuil 0,95 (ou 95%).

Définition : X_n est une variable aléatoire qui suit une loi binomiale $B(n;p)$.

La variable aléatoire $F_n = \frac{X_n}{n}$ s'appelle la variable aléatoire fréquence de succès pour un schéma de Bernoulli de paramètres n et p.

Propriété : Soit $\alpha \in \left]0;1\right[$ et X_n une variable aléatoire qui suit une loi binomiale $B(n;p)$.

La probabilité que la fréquence F_n prenne ses valeurs dans l'intervalle $I_n = \left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\,;\, p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ se rapproche de $1-\alpha$ quand la taille de l'échantillon n devient grande. On note : $\lim\limits_{n\to+\infty} P(F_n \in I_n) = 1-\alpha$.

Définition : I_n est appelé intervalle de fluctuation asymptotique de la fréquence F_n au seuil $1-\alpha$.

Démonstration (exigible BAC) :

X_n suit la loi binomiale $B(n;p)$ donc la suite de variables aléatoires $Z_n = \frac{X_n - E(X_n)}{\sigma(X_n)}$ suit une loi normale centrée réduite $N(0;1)$ et d'après le théorème de Moivre-Laplace, on a :

$\lim\limits_{n\to+\infty} P(a \le Z_n \le b) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,dx$, pout tous réels a et b avec $a < b$.

Or $Z_n = \frac{X_n - E(X_n)}{\sigma(X_n)} = \frac{X_n - np}{\sqrt{np(1-p)}} = \frac{n\left(\frac{X_n}{n} - p\right)}{n\frac{\sqrt{p(1-p)}}{\sqrt{n}}} = \frac{F_n - p}{\frac{\sqrt{p(1-p)}}{\sqrt{n}}}$.

Donc $\lim\limits_{n\to+\infty} P\left(p + a\frac{\sqrt{p(1-p)}}{\sqrt{n}} \le F_n \le p + b\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right) = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\,dx$

Comme, pour tout réel $\alpha \in \left]0;1\right[$, il existe un unique réel positif u_α tel que $P\left(-u_\alpha \le X \le u_\alpha\right) = 1-\alpha$ où *X* suit une loi normale centrée réduite $N\left(0;1\right)$, on a :

$$\int_{-u_\alpha}^{u_\alpha} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}\, dx = 1-\alpha .$$

En prenant $a = -u_\alpha$ et $b = u_\alpha$, on a :

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

Remarque :
La probabilité définie dans la propriété se rapproche de $1-\alpha$ sans être nécessairement égale d'où l'emploi du terme "asymptotique".

Exemple :
Vidéo https://youtu.be/k_Q2FN07jQ0

Démontrons le résultat donné dans l'exemple en début de paragraphe :

$$I_{50} = \left[0,3 - 1,96 \times \frac{\sqrt{0,3\times 0,7}}{\sqrt{50}} \,;\, 0,3 + 1,96 \times \frac{\sqrt{0,3\times 0,7}}{\sqrt{50}} \right]$$ car $u_{0,05} = 1,96$.

Soit $I_{50} = \left[0,173\,;\,0,427\right]$.

Pour 500 tirages, on obtient :

$$I_{500} = \left[0,3 - 1,96 \times \frac{\sqrt{0,3\times 0,7}}{\sqrt{500}} ; 0,3 + 1,96 \times \frac{\sqrt{0,3\times 0,7}}{\sqrt{500}} \right] = \left[0,26\,;\,0,34\right]$$

On constate que l'intervalle, pour un même seuil, se resserre fortement lorsqu'on augmente le nombre de tirages.

Définition : On appelle intervalle de fluctuation au seuil 0,95 de la variable aléatoire fréquence l'intervalle : $\left[p - 1,96 \times \frac{\sqrt{p\left(1-p\right)}}{\sqrt{n}} \,;\, p + 1,96 \times \frac{\sqrt{p\left(1-p\right)}}{\sqrt{n}} \right]$.

2) Prise de décision

Dans ce paragraphe, la proportion du caractère étudié n'est pas connue mais est supposée être égale à *p*.
La prise de décision consiste à valider ou invalider l'hypothèse faite sur la proportion *p*.

Propriété (Règle de décision) : Soit f la fréquence du caractère étudié d'un échantillon de taille n.
Soit l'hypothèse : "La proportion de ce caractère dans la population est *p*."
Soit *I* l'intervalle de fluctuation asymptotique au seuil 0,95.
- Si $f \in I$, alors on accepte l'hypothèse faite sur la proportion *p*.
- Si $f \notin I$, alors on rejette l'hypothèse faite sur la proportion *p*.

Remarque :
On peut interpréter cette propriété par le fait que la probabilité qu'on rejette à tort l'hypothèse sur *p* sachant qu'elle est vraie est approximativement égale à 5%.

Méthode : Prendre une décision à l'aide d'un intervalle de fluctuation

Vidéo https://youtu.be/QZ0YFthGI0Y

Un fabricant d'alarme commande auprès de son fournisseur deux types de composants électroniques : RS017 et P412. Il demande 900 composants de chaque sorte.
Au moment de la livraison, le service de contrôle retire 50 composants et constate que 19 sont des modèles RS017.
Peut-on affirmer qua la commande est respectée par le fournisseur ?

- Le fabricant a commandé autant de composants de chaque sorte. On peut donc supposer que la proportion de composants RS017 est égale à 0,5.
La taille de l'échantillon est $n = 50$.
La fréquence observée est donc $f = \frac{19}{50} = 0{,}38$.
- Vérifions si les paramètres n est *p* répondent aux conditions imposées :
$n = 50 \geq 30$, $n \times p = 50 \times 0{,}5 = 25 \geq 5$ et $n \times (1 - p) = 50 \times 0{,}5 = 25 \geq 5$
- L'intervalle de fluctuation asymptotique au seuil 0,95 est :
$$\left[0{,}5 - 1{,}96 \times \frac{\sqrt{0{,}5 \times 0{,}5}}{\sqrt{50}} ; 0{,}5 + 1{,}96 \times \frac{\sqrt{0{,}5 \times 0{,}5}}{\sqrt{50}}\right] \approx \left[0{,}361 ; 0{,}639\right].$$
La fréquence observée $f = 0{,}38$ appartient à l'intervalle de fluctuation asymptotique au seuil 0,95, d'après la règle de décision, l'hypothèse faite est acceptable.

II. Estimation

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

C'est le problème inverse de celui de l'échantillonnage. A partir de la fréquence observée sur un échantillon, on va estimer la proportion *p* d'un caractère dans la population tout entière.

Propriété : X_n est une variable aléatoire qui suit une loi binomiale $B(n;p)$.

$F_n = \frac{X_n}{n}$ est la fréquence associée à X_n.

Pour n suffisamment grand, p appartient à l'intervalle $J_n = \left[F_n - \frac{1}{\sqrt{n}} ; F_n + \frac{1}{\sqrt{n}}\right]$ avec une probabilité supérieure ou égale à 0,95.

- Admis -

Définition : Soit f une fréquence observée du caractère étudié sur un échantillon de taille n.
On appelle intervalle de confiance de la proportion p au niveau de confiance 0,95, l'intervalle $\left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}}\right]$.

Remarques :
- Un niveau de confiance 0,95 signifie que dans 95 cas sur 100, on affirme à juste titre que p appartient à l'intervalle de confiance.
- Il n'est pas vrai d'affirmer que p est égal au centre de l'intervalle de confiance. Il n'est pas possible d'évaluer la position de p dans l'intervalle de confiance.
- p étant inconnu, il n'est pas possible de vérifier si les conditions énoncées sur n et p en introduction de chapitre sont vérifiées.
Cependant, il faudra les vérifier sur la fréquence observée f :

$n \geq 30$, $n \times f \geq 5$ et $n \times (1-f) \geq 5$.

Exemple :
On dispose d'une urne contenant un grand nombre de boules blanches et noires. La proportion de boules blanches contenues dans l'urne n'est pas connue.
On réalise un tirage de 100 boules et on obtient 54 boules blanches.
La fréquence observée est donc $f = 0{,}54$.
L'intervalle de confiance de la proportion de boule blanche dans l'urne au niveau de confiance 95% est $\left[0{,}54 - \frac{1}{\sqrt{100}} ; 0{,}54 + \frac{1}{\sqrt{100}}\right] = [0{,}44 ; 0{,}64]$.

Méthode : Estimer une proportion inconnue par un intervalle de confiance

Vidéo https://youtu.be/cU5cJICVAM8

Un institut de sondage interroge 1052 personnes entre les deux tours de l'élection présidentielle sur leur intention de vote.
614 déclarent avoir l'intention de voter pour Martine Phinon.
En supposant que les votes seront conformes aux intentions, la candidate a-t-elle raison de croire qu'elle sera élue ?

- La proportion p des électeurs de Martine Phinon est inconnue.
La taille de l'échantillon est $n = 1052$.

La fréquence observée est $f = \frac{614}{1052} \approx 0,5837$.
- Vérifions si les paramètres n est f répondent aux conditions imposées :
$n = 1052 \geq 30$, $n \times f \approx 1052 \times 0,5837 \approx 614 \geq 5$ et
$n \times (1-p) \approx 1052 \times (1-0,5837) \approx 438 \geq 5$.
- L'intervalle de confiance de la proportion *p* au niveau de confiance 0,95 est :
$\left[0,5837 - \frac{1}{\sqrt{1052}} \ ; \ 0,5837 + \frac{1}{\sqrt{1052}}\right] \approx [0,553 \ ; \ 0,615]$.
Pour être élue, la proportion *p* doit être strictement supérieure à 0,5. Selon ce sondage, il est envisageable que Martine Phinon soit élue.

Intervalle de FLUCTUATION V.S. Intervalle de CONFIANCE :

Vidéo https://youtu.be/97vzxWsyie8

Méthode : Déterminer une taille d'échantillon suffisante pour obtenir une estimation d'une proportion

Vidéo https://youtu.be/ogmMVpkBVgs

Un constructeur automobile fait appel à un institut de sondage afin de mesurer le degré de satisfaction du service après-vente.
L'institut souhaite estimer la proportion de clients satisfaits au niveau de confiance 0,95 avec une amplitude d'au plus 5 centièmes.
Combien de personnes au minimum faut-il interroger ?

On appelle *p* la proportion de clients satisfaits. Cette proportion est inconnue.
Une estimation de cette proportion peut être obtenue à l'aide de l'intervalle de confiance au niveau de confiance 0,95 : $\left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}}\right]$, où f est la fréquence observée.

Cet intervalle a pour longueur $\frac{2}{\sqrt{n}}$.

Donc $\frac{2}{\sqrt{n}} \leq 0,05$ soit $n \geq \frac{4}{0,05^2} = 1600$.
L'institut de sondage devra donc interroger au moins 1600 personnes.