Synthèse de cours PanaMaths
→ Fluctuation et estimation

Introduction

Bien souvent, des situations réelles sont modélisées grâce à la loi binomiale : obtention de PILE (ou FACE !) lors du jet d'une pièce (bien équilibrée ou non), obtention de telle ou telle face lors du jet d'un dé (bien équilibré ou non), dysfonctionnement d'un appareil ou d'un système à l'issue du processus de fabrication ou en cours d'utilisation, nombre d'appels à un service client durant une période donnée, nombre de votants pour un candidat donné lors d'une élection, …

Dans la pratique, le premier paramètre de la loi, *n*, est connu (on lance la pièce ou le dé un certain nombre de fois, on teste un certain nombre de produits/systèmes, on interroge un certain nombre de votants,…) et il correspond à la taille de l'échantillon sur lequel on travaille.

Si on note X_n la variable aléatoire suivant la loi binomiale considérée, on s'intéresse souvent à la loi $\overline{X}_n = \dfrac{X_n}{n}$ correspondant à la fréquence empirique des succès.
Par exemple, si on lance 1000 fois une pièce bien équilibrée et que l'on s'intéresse au nombre de PILE obtenus, la fréquence empirique sera donnée par la variable aléatoire $\overline{X}_{1000} = \dfrac{X_{1000}}{1000}$.
Ainsi, si on a obtenu 435 PILE (i.e. la variable aléatoire X_{1000} prend la valeur 435), la variable aléatoire $\overline{X}_{1000}$ prendra la valeur $\dfrac{435}{1000} = 0,435 = 43,5\%$.

Dans certains cas, l'espérance de la loi est connue.
Par exemple, dans la situation ci-dessus, si la pièce est bien équilibrée, la variable aléatoire X_{1000} suivra la loi binomiale $\mathscr{B}(1000\,;0,5)$. On s'intéresse alors légitimement à la loi de la variable aléatoire $\overline{X}_{1000}$. On niveau de la classe de terminale ES, on peut intuitivement admettre que l'espérance de cette loi soit égale à 0,5. La fréquence empirique étant aléatoire, on parle de « fluctuation d'échantillonnage ». Lorsque *n* est grand, on peut donner des résultats précis sur la variable $\overline{X}_n$. C'est l'un d'eux qui apparaît dans le programme de la classe de terminale ES : on précise en fait l'intervalle de fluctuation vu en classe de seconde.

Dans d'autres situations, on ne connaît pas le second paramètre *p* de la loi.
Par exemple, si l'on effectue un sondage à la sortie de bureaux de vote dans le cas d'une élection donnée, on sait qu'il y a une probabilité *p* qu'un votant ait porté son choix sur un candidat donné. Il n'est absolument pas nécessaire d'interroger tous les votants pour se faire une idée précise de la valeur de *p* ! On va estimer *p* à partir d'un échantillon réduit.
Une telle estimation peut être rendue incontournable pour des raisons économiques (coût de la démarche qui consisterait à interroger tous les votants, à tester certains produits/systèmes …) ou de simple bon sens (si on teste une production de pétards, on ne va pas tous les faire sauter pour vérifier la qualité de la production … si ? ☺).

Fluctuation

Théorème-définition

Soit X_n une variable aléatoire réelle suivant la loi binomiale $\mathscr{B}(n\,;\,p)$.

Soit $\bar{X}_n = \dfrac{X_n}{n}$ la variable aléatoire réelle associée à X_n et correspondant à la fréquence empirique des succès.

On appelle « intervalle de fluctuation asymptotique au seuil de 95% » l'intervalle :

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

On a :

$$\lim_{n\to+\infty} p\left(\bar{X}_n \in IF_n\right) = 0{,}95$$

En d'autres termes, pour *n* suffisamment grand (voir ci-après), la probabilité que la fréquence empirique se trouve dans l'intervalle IF_n est approximativement égale à 0,95.

Remarques :

- La notation « IF_n » n'est pas standard.
- Rappelons que l'écart type σ_{X_n} de la variable aléatoire X_n suivant la loi binomiale $\mathscr{B}(n\,;\,p)$ est égal à $\sqrt{np(1-p)}$.
- Dans la pratique, on pourra utiliser cet intervalle de fluctuation lorsque les conditions suivantes seront vérifiées :

$n \geq 30$	n doit être "grand"
$np \geq 5$	p ne doit pas être trop "proche" de 0
$n(1-p) \geq 5$	p ne doit pas être trop "proche" de 1

- L'intervalle IF_n est centré en *p* et sa longueur vaut $3{,}92\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}$. Ainsi, cet intervalle est d'autant moins grand que *n* est élevé. On notera cependant que cette diminution est en « $\dfrac{1}{\sqrt{n}}$ ». En « jouant » sur la taille de l'échantillon, *n*, on pourra donc, par exemple, diviser la taille de l'intervalle de fluctuation par 2 à condition de multiplier celle de l'échantillon par 4 …

- (Hors programme) Ce résultat découle d'un théorème célèbre appelé « théorème central limite » ou « théorème de la limite centrée » qui permet d'affirmer ici que la loi de la variable aléatoire $\overline{X}_n$ peut être approchée, pour *n* grand, par la loi normale :

$$\mathscr{N}\left(p\,;\frac{p(1-p)}{n}\right) \text{ où } p=\mathrm{E}(X_n)=\mathrm{E}(\overline{X}_n) \text{ et } \frac{p(1-p)}{n}=V(\overline{X}_n)=\frac{V(X_n)}{n^2}=\left(\frac{\sigma_{X_n}}{n}\right)^2$$

Il s'agit donc de la loi normale ayant même espérance et même écart type que la loi de la variable aléatoire $\overline{X}_n$.

On ne s'étonnera donc pas de voir apparaître le coefficient 1,96 alors que l'on considère une probabilité de 0,95 …

Un exemple

Supposons que l'on dispose d'un dé bien équilibré.
On lance ce dé et on considère que l'on gagne la partie (événement SUCCES) si le 5 ou le 6 sort.

La probabilité de gagner est donc de $p=\frac{2}{6}=\frac{1}{3}$.

On joue 100 parties.

La variable aléatoire X_{100} qui comptabilise le nombre de parties gagnées suit la loi binomiale $\mathscr{B}\left(100\,;\frac{1}{3}\right)$.

On a ici $n=100>30$, $np=100\times\frac{1}{3}>5$ et $n(1-p)=100\times\frac{2}{3}>5$.

On a alors :

$$p-1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}=\frac{1}{3}-1{,}96\frac{\sqrt{\frac{1}{3}\times\frac{2}{3}}}{\sqrt{100}}=\frac{1}{3}-\frac{1{,}96\sqrt{2}}{30}=\frac{10-1{,}96\sqrt{2}}{30}\simeq 0{,}241$$

$$p+1{,}96\frac{\sqrt{p(1-p)}}{\sqrt{n}}=\frac{10+1{,}96\sqrt{2}}{30}\simeq 0{,}426$$

L'intervalle de fluctuation asymptotique au seuil de 95% de la variable aléatoire $\overline{X}_{100}=\frac{X_{100}}{100}$ est donc : $[0{,}241\,;0{,}426]$.

Cela signifie que la probabilité que la fréquence empirique $\overline{X}_{100}$ se trouve dans l'intervalle $[0{,}241\,;0{,}426]$ est de 0,95. Dit autrement, si on effectue plusieurs séries de 100 lancers avec ce dé et que l'on calcule à chaque fois la fréquence empirique, environ 95% des valeurs ainsi calculées se trouveront dans l'intervalle $[0{,}241\,;0{,}426]$.

L'intervalle de fluctuation donné en seconde

Dans les mêmes conditions que celle du théorème-définition ci-dessus, on a donné en seconde l'intervalle de fluctuation suivant : $\left[p-\frac{1}{\sqrt{n}} ; p+\frac{1}{\sqrt{n}} \right]$.

Sur l'intervalle $[0;1]$, on montre facilement que la fonction $p \mapsto p(1-p)$, qui prend des valeurs positives, admet $\frac{1}{4}$ pour valeur maximale. On en déduit alors que la fonction $p \mapsto \sqrt{p(1-p)}$ admet, elle, $\frac{1}{2}$ pour valeur maximale.

On a donc : $\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq \frac{\frac{1}{2}}{\sqrt{n}}$ puis $1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \leq \frac{1,96\times\frac{1}{2}}{\sqrt{n}} \leq \frac{1}{\sqrt{n}}$.

On en déduit :

$$\forall p \in [0;1], \left[p-1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p+1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}} \right] \subset \left[p-\frac{1}{\sqrt{n}} ; p+\frac{1}{\sqrt{n}} \right]$$

L'intervalle de fluctuation asymptotique donné en seconde, qui a le mérite de la simplicité au niveau de l'expression de ses bornes, est donc un intervalle de fluctuation asymptotique à un seuil <u>au moins égal</u> à 95%.

Remarque : le calcul ci-dessus n'est valable qu'avec une loi binomiale ! Le cadre d'application du théorème central limite étant plus large, on doit donc s'efforcer de bien garder présent à l'esprit le fait que l'intervalle fourni en seconde est directement lié à la loi binomiale.

Estimation

Introduction et mise en garde

Ce qui suit ressemble beaucoup à … ce qui précède ! Bien qu'il y ait des éléments théoriques sous-jacents communs, la problématique n'est pas la même.

Ici, on va considérer un échantillon aléatoire à partir duquel on va calculer une fréquence empirique qui sera une réalisation de la variable aléatoire $\overline{X}_n = \frac{X_n}{n}$, la variable aléatoire X_n suivant la loi binomiale $\mathscr{B}(n;p)$. Mais si n est connu (taille de l'échantillon), le paramètre p ne l'est pas ! La problématique consiste ici à estimer ce paramètre. Plus précisément, dans le cadre du programme de terminale ES, on va fournir un intervalle dans lequel ce paramètre aura une certaine probabilité (élevée) de se trouver.

Théorème-définition

Soit X_n une variable aléatoire réelle suivant la loi binomiale $\mathscr{B}(n\,;\,p)$, le paramètre p n'étant pas connu.
Soit $\overline{X}_n = \dfrac{X_n}{n}$ la variable aléatoire réelle associée à X_n et correspondant à la fréquence empirique des succès.

Soit f une réalisation de la variable aléatoire $\overline{X}_n$ (calculée à partir d'un échantillon de taille n).

On appelle « intervalle de confiance de p au seuil (ou « au niveau ») de confiance de 95% » l'intervalle :

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

On parle d' « estimation de p par intervalle de confiance ».

On a :

$$\lim_{n\to+\infty} p\left(p \in IC_n\right) \geq 0{,}95$$

En d'autres termes, pour n suffisamment grand (voir ci-après), la probabilité que le paramètre p se trouve dans l'intervalle IC_n est supérieure ou égale à 0,95.

Remarques :

- La notation « IC_n » n'est pas standard.
- L'intervalle de confiance IC_n est centré sur la fréquence empirique f et a pour longueur $\dfrac{2}{\sqrt{n}}$.
- La fréquence empirique f étant une réalisation de la variable aléatoire $\overline{X}_n$, l'intervalle $\left[f - \dfrac{1}{\sqrt{n}}\,;\, f + \dfrac{1}{\sqrt{n}} \right]$ est une réalisation de l'intervalle aléatoire $\left[\overline{X}_n - \dfrac{1}{\sqrt{n}}\,;\, \overline{X}_n + \dfrac{1}{\sqrt{n}} \right]$.
- Dans la pratique, on pourra accepter cet intervalle de confiance lorsque les conditions suivantes seront vérifiées (pour la 2ème et la 3ème, la vérification se fait A POSTERIORI, i.e. après obtention de l'intervalle de confiance) :

$n \geq 30$	n doit être "grand"
$n\left(f - \dfrac{1}{\sqrt{n}} \right) \geq 5$	$f - \dfrac{1}{\sqrt{n}}$ ne doit pas être trop "proche" de 0
$n\left(1 - \left(f + \dfrac{1}{\sqrt{n}} \right) \right) \geq 5$	$f + \dfrac{1}{\sqrt{n}}$ ne doit pas être trop "proche" de 1

- (Hors programme) L'écart type de la variable aléatoire X_n (et donc de la variable aléatoire $\overline{X}_n$) s'exprimant en fonction du paramètre *p*, une expression plus précise de l'intervalle de confiance de *p* au seuil de confiance de 95% est :

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

Un exemple

Un responsable de production cherche à estimer la proportion d'appareils défectueux produits quotidiennement sur son site.
Des tests sont effectués sur 400 appareils produits et on constate que 32 d'entre eux sont défectueux.

On a donc ici $f = \frac{32}{400} = 0{,}08$.

D'où $f - \frac{1}{\sqrt{400}} = 0{,}08 - \frac{1}{20} = 0{,}08 - 0{,}05 = 0{,}03$ et $f + \frac{1}{\sqrt{400}} = 0{,}08 + 0{,}05 = 0{,}13$.

On a alors :

- $n = 400 \geq 30$
- $n\left(f - \frac{1}{\sqrt{n}} \right) = 400 \times 0{,}03 = 12 \geq 5$
- $n\left(1 - \left(f + \frac{1}{\sqrt{n}} \right) \right) = 400 \times (1 - 0{,}13) = 348 \geq 5$

L'intervalle de confiance au seuil de 95% est donc $[0{,}03 ; 0{,}13] = [3\% ; 13\%]$.

Ainsi, la probabilité que la proportion d'appareils défectueux soit comprise entre 3% et 13% est supérieure ou égale à 95%.

Remarque : l'intervalle obtenu peut sembler bien … grand et, de fait, l'estimation de la proportion d'appareils défectueux assez imprécise. On doit souligner que :

- L'intervalle $\left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}} \right]$ est un intervalle volontairement « conservateur ». Ce que l'on gagne en « certitude » en statistique (0,95 est une probabilité élevée !) est classiquement perdu en précision.
- La longueur de l'intervalle utilisée étant égale à $\frac{2}{\sqrt{n}}$, celui-ci sera d'autant plus petit que la taille de l'échantillon sera grande. Si ce constat est général, la solution consistant à augmenter significativement la taille de l'échantillon se heurte souvent dans la pratique à des considérations de faisabilité et/ou économiques.