Intervalles et échantillons

De nombreuses questions se posent aux élèves (et parfois à leurs professeurs) à propos des intervalles de confiance et des intervalles de fluctuation asymptotique.

Cette chronique a pour objectif de vulgariser un peu ces notions pour mieux les faire passer auprès des élèves.

J'ai déjà traité du sujet « intervalle de fluctuation asymptotique » en mai 2017 ; le texte de l'époque est disponible ici. Excusez les répétitions !

1 Fluctuation d'échantillonnage

La notion primordiale est la notion de « fluctuation d'échantillonnage ».

Quand on connaît une proportion dans une population de grande taille et qu'on extrait des échantillons de cette population, ils ne sont pas tous constitués de la même façon.

- Par exemple, si on a une urne qui contient 100 000 boules dont 30 000 boules blanches, et qu'on extrait de cette urne 1 000 boules, il n'y en aura pas « forcément » 300 blanches.
  Dans cette urne, la proportion de boules blanches est de 30 % donc la probabilité d'extraire une boule blanche de l'urne est 0,3.
  On n'oubliera pas de parler aux élèves d'équiprobabilité.
  Si on constitue des échantillons « avec remise » (ce qui correspond à un tirage « non exhaustif ») – c'est-à-dire que l'on extrait une boule, on note si elle est blanche ou non, puis on la remet dans l'urne – la variable aléatoire X qui donne le nombre de boules blanches dans l'échantillon de taille 1 000 suit la loi binomiale de paramètres $n = 1\,000$ et $p = 0{,}3$.
  La probabilité d'avoir exactement 300 boules blanches sur les 1 000 est donnée par la formule $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ donc $P(X = 300) = \binom{1\,000}{300} \times 0{,}3^{300} \times 0{,}7^{700} \approx 0{,}027\,5$.
  Il n'y a donc qu'environ 2,75 % de chances que l'échantillon extrait contienne 300 boules blanches.

  > Dans la pratique, on ne remet pas dans l'urne les boules déjà tirées, et la variable aléatoire X qui donne le nombre de boules blanches suit alors une loi hypergéométrique de paramètres $N = 100\,000$, $n = 1\,000$ et $p = 0{,}3$. Mais si N tend vers l'infini, la loi hypergéométrique tend vers la loi binomiale de paramètres n et p.
  > Pour éviter toute ambigüité, on trouve souvent dans les sujets des phrases du type : « On estimera que la population est de taille suffisamment grande pour que l'on puisse assimiler ce tirage à un tirage avec remise. »
  > On trouve d'ailleurs dans un document d'accompagnement de la classe de seconde : « ... un tirage sans remise d'un échantillon dans une population suffisamment nombreuse est assimilable à la répétition d'un tirage avec remise dans une urne. »
  > Ainsi on travaille directement avec une loi binomiale.
- Comme autre exemple simple, on peut faire lancer 20 fois une pièce et compter le nombre de fois où elle tombe du côté « pile ». Là encore, on serait en droit d'espérer que la pièce tombera 10 fois sur « pile », ce qui n'a qu'une probabilité d'arriver égale à :
  $\binom{20}{10} \times 0{,}5^{10} \times (1 - 0{,}5)^{10-5} \approx 0{,}176\,2$, soit dans environ 17,62 % des cas.
- Dernier exemple : on peut faire jeter un dé 60 fois et faire compter le nombre de fois que le 6 apparaît. La probabilité qu'il sorte 10 fois est $\binom{60}{10} \times \left(\frac{1}{6}\right)^{10} \times \left(1 - \frac{1}{6}\right)^{60-10} \approx 0{,}137\,0$.
  Ce qui veut dire qu'un échantillon de taille 60 contiendra exactement 10 fois le 6 dans environ 13,70 % des cas.

2 Intervalle de confiance

On utilise un « intervalle de confiance » quand, à partir d'un échantillon donné, on veut estimer une proportion dans une population ; c'est typiquement ce que l'on fait dans des sondages d'opinion.

D'après ce qu'on a vu dans le paragraphe précédent, il faut être prudent quand on manipule des échantillons.

> Voyons un exemple.
> Une personne lance une pièce 1 000 fois et elle tombe 490 fois sur « pile » ce qui peut arriver dans environ 2,07 % des cas. Est-ce que cette personne va dire que la probabilité d'obtenir « pile » est de 0,49 ?
> Attention de ne pas être distrait par le 2,07 % qui peut paraître faible : l'obtention de 500 « pile » n'arrive que dans environ 2,57 % des cas !
> Une autre personne, sur 1 000 lancers, obtient 505 fois « pile » ; va-t-elle en déduire que la probabilité d'apparition de « pile » est de 0,505 ?
> Si on se contente d'une estimation ponctuelle, les deux probabilités sont envisageables.
> Mais est-ce bien raisonnable ?

On va donc faire une estimation par intervalle de confiance.

On cherche à estimer une proportion p d'un caractère dans une population donnée.
On extrait un échantillon de taille n dans cette population, et on calcule la fréquence f_n du caractère dans cet échantillon.
On choisit un seuil de confiance, souvent 95 %.

> Déterminer un intervalle de confiance au niveau de confiance 95 %, c'est déterminer un intervalle I tel que la probabilité que p appartiennent à I soit supérieure ou égale à 0,95.

Après des calculs un peu compliqués (et des théorèmes importants qu'un élève de terminale S doit connaître), on peut dire qu'un intervalle de confiance au niveau de confiance 95 % est

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

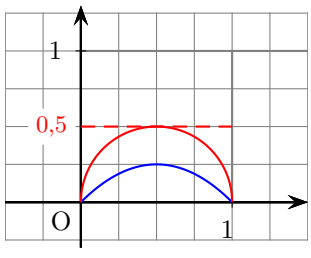
Le tout est assorti de quelques conditions : $n \geqslant 30$, $nf_n \geqslant 5$ et $n(1-f_n) \geqslant 5$.

Le nombre f_n est une fréquence, donc est compris entre 0 et 1.
Pour $0 \leqslant x \leqslant 1$, on a $0 \leqslant x(1-x) \leqslant 0{,}25$, donc
$0 \leqslant \sqrt{x(1-x)} \leqslant 0{,}5$.
On en déduit que

$$1{,}96\sqrt{\frac{f_n(1-f_n)}{n}} = \frac{1{,}96\sqrt{f_n(1-f_n)}}{\sqrt{n}} \leqslant \frac{1{,}96 \times 0{,}5}{\sqrt{n}} < \frac{1}{\sqrt{n}}$$

et donc que $f_n - \dfrac{1}{\sqrt{n}} < f_n - 1{,}96\sqrt{\dfrac{f_n(1-f_n)}{n}}$ et que

$f_n + 1{,}96\sqrt{\dfrac{f_n(1-f_n)}{n}} < f_n + \dfrac{1}{\sqrt{n}}$.



Fonction $x \longmapsto x(1-x)$
Fonction $x \longmapsto \sqrt{x(1-x)}$

Donc $\left[f_n - 1{,}96\sqrt{\dfrac{f_n(1-f_n)}{n}} \ ; \ f_n + 1{,}96\sqrt{\dfrac{f_n(1-f_n)}{n}} \right] \subset \left[f_n - \dfrac{1}{\sqrt{n}} \ ; \ f_n + \dfrac{1}{\sqrt{n}} \right]$.

C'est souvent l'intervalle $\left[f_n - \dfrac{1}{\sqrt{n}} \ ; \ f_n + \dfrac{1}{\sqrt{n}} \right]$ que l'on prend comme intervalle de confiance au niveau de confiance 95 % ; son amplitude est de $\dfrac{2}{\sqrt{n}}$.

> Petite fiction
>
> Lors d'une élection, on a le choix entre deux candidats A et B. On effectue un sondage auprès de 1 000 personnes et il y en a 510 qui déclarent vouloir voter pour le candidat A. Peut-on dire que, parmi tous les électeurs, il y en a 51 % qui sont prêts à voter pour A ?
>
> La fréquence dans l'échantillon de taille 1 000 est $f_{1\,000} = 0{,}51$.
> On a $n = 1\,000 \geqslant 30$, $nf_n = 510 \geqslant 5$ et $n(1-f_n) = 490 \geqslant 5$, donc les conditions sont réalisées pour que l'on établisse un intervalle de confiance au niveau de confiance 95 % :
> $\left[f_n - \dfrac{1}{\sqrt{n}} \,;\, f_n + \dfrac{1}{\sqrt{n}}\right] = \left[0{,}51 - \dfrac{1}{\sqrt{1000}} \,;\, 0{,}51 + \dfrac{1}{\sqrt{1000}}\right] \approx \left[0{,}478 \,;\, 0{,}542\right]$.
> On peut donc dire que la probabilité que le pourcentage d'électeurs prêts à voter pour le candidat A soit compris entre 47,8 % et 54,2 %, est supérieure à 0,95.

Que penser alors de phrases que – l'on entend régulièrement – du style « La cote de popularité de M. X a baissé d'un point ce mois-ci. » ?

Si le sondage se fait auprès de 1 000 personnes, on aura $\dfrac{1}{\sqrt{1\,000}} \approx 3{,}16$, donc l'intervalle de confiance aura une amplitude de 6,3 points. Est-ce bien raisonnable de parler d'une baisse de 1 point quand l'amplitude de l'intervalle de confiance est 6 fois plus grande ?

3 Intervalle de fluctuation asymptotique

La problématique liée à l'intervalle de fluctuation asymptotique est très différente de celle liée à l'intervalle de confiance.

Il s'agit ici de « tester une hypothèse », c'est-à-dire que l'on fait une hypothèse sur une proportion dans une population, et on va essayer de dire si cette hypothèse est justifiée en testant la fréquence du caractère étudié sur un échantillon extrait de cette population.

> Exemple : on a entendu dire que 54 % des français aimaient le jazz ; on veut vérifier la pertinence de cette affirmation.

Plus généralement, on fait l'hypothèse que, dans une population, un caractère étudié se trouve dans la proportion p. On veut vérifier cette hypothèse en calculant la fréquence de ce caractère dans un échantillon de taille n ; il s'agit d'un « test d'hypothèse ».

Il faut d'abord vérifier que l'on a un échantillon de taille suffisante et que la proportion n'est ni trop proche de 0 ni trop proche de 1 : $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$.
Cela exclut cette méthode lorsqu'on travaille avec des échantillons de faibles effectifs, notamment dans le milieu médical.

On prend donc un échantillon aléatoire de taille n, et on aimerait que la fréquence f_n du caractère dans cette échantillon soit « assez proche » de la proportion p, c'est-à-dire en fait dans un intervalle centré sur p et dont l'amplitude dépend du niveau de risque (ou du seuil de confiance).

Dans la pratique, on prend pour intervalle de fluctuation asymptotique, l'intervalle

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

quand le niveau de confiance est de 95 %.

Là encore, il y a des théorèmes à mettre en place et des démonstrations que des élèves de TS doivent connaître.

Ensuite on établit le diagnostic : si la fréquence calculée f_n appartient à l'intervalle I, il n'y a pas de raison de rejeter l'hypothèse selon laquelle la proportion dans la population globale est p. Sinon, on peut rejeter cette hypothèse. Le tout étant attaché à un risque de 5 %.

Comme l'intervalle est centré sur p, on a affaire à un test « bilatéral », le seul en vigueur dans les programmes du secondaire. Cela sous-entend qu'il existe un type de test « unilatéral » qui conduit à un intervalle de type $[p\,;\,p+a]$ ou $[p-a\,;\,p]$; mais c'est une autre histoire.

4 Remarque sur le 1,96

On demande souvent dans un intervalle de fluctuation asymptotique de donner des valeurs approchées des bornes au millième. Personnellement, j'approche par défaut la borne gauche et par excès la borne droite.

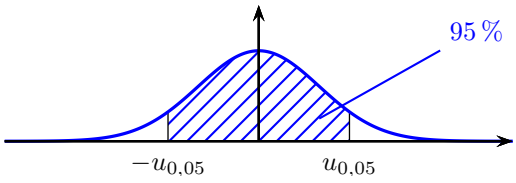
Mais comment peut-on avoir une valeur approchée au millième correcte alors qu'on utilise le nombre 1,96 qui est arrondi au centième ?

Rappelons comment on trouve cette valeur de 1,96.

Dans le cas général, on cherche un nombre u_α tel que, si la variable aléatoire X suit la loi normale centrée réduite, on ait $P(-u_\alpha < X < u_\alpha) = 1-\alpha$.

Pour un seuil de 5 %, on a $\alpha = 0{,}05$ et on cherche donc $u_{0{,}05}$ tel que $P(-u_{0{,}05} < X < u_{0{,}05}) = 0{,}95$.

C'est en effet dans cette valeur de u_α qu'apparaît le niveau de confiance de 95 % (c'est utile de le faire comprendre aux élèves).



D'après la symétrie de la courbe autour de l'axe des ordonnées, on a $P(X \geqslant u_{0{,}05}) = \dfrac{1-0{,}95}{2}$ c'est-à-dire $P(X \geqslant u_{0{,}05}) = 0{,}025$; ce qui signifie que $P(X < u_{0{,}05}) = 0{,}975$.

On trouve à la calculatrice que $u_{0{,}05}$ vaut environ 1,959 964. Ouf !

Pourquoi ouf ? Parce que la valeur arrondie au millième de $u_{0{,}05}$ est égale à 1,960, et que la valeur arrondie au dix-millième de $u_{0{,}05}$ est 1,960 0.

Il n'est donc pas totalement incohérent d'arrondir les bornes d'un intervalle de confiance ou de fluctuation asymptotique au millième, voire au dix-millième.

Personnellement, je préfèrerais que l'on donne 1,960 comme valeur approchée de $u_{0{,}05}$ au lieu de 1,96 pour bien signifier que c'est la valeur arrondie au millième.

5 Pour être clair

Si on veut estimer le nombre de personnes aimant le jazz dans une population, on utilisera un intervalle de confiance.

Si on veut tester l'hypothèse selon laquelle 54 % des gens aiment le jazz dans une population, on utilisera un intervalle de fluctuation asymptotique.