1. Intervalle de fluctuation

Contexte : dans une population donnée, la proportion d'individus présentant le caractère C est p. Que peut-on dire de la fréquence f de C sur un échantillon aléatoire de taille n ?

Définition : Soit X une variable aléatoire, définie sur un intervalle contenant $[a;b]$. Soit α un réel appartenant à $]0;1[$. Dire que $[a;b]$ est un intervalle de fluctuation de X au seuil $1-\alpha$ signifie que $P(a \le X \le b) \ge 1-\alpha$

Remarque : cette définition est très générale, dans ce chapitre nous supposerons que X suit systématiquement la loi binomiale $\mathscr{B}(n\,;p)$.

Propriété : si X_n suit la loi binomiale $\mathscr{B}(n\,;p)$, avec $p \in]0;1[$, alors pour tout réel α de $]0;1[$, on a :

$$\lim_{n\to+\infty} P\left(\frac{X_n}{n} \in I_n\right) = 1-\alpha \text{ , où } I_n = \left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$$

et u_α désigne le nombre réel tel que $P(-u_\alpha \le Z \le u_\alpha) = 1-\alpha$ lorsque Z suit la loi normale $\mathscr{N}(0\,;1)$.

C'est-à-dire que $F_n = \frac{X_n}{n}$ appartient à I_n avec une probabilité approximativement égale à $1-\alpha$ (remarque : la suite $P\left(\frac{X_n}{n} \in I_n\right)$ n'étant pas monotone, on ne peut pas savoir si la probabilité est inférieure ou supérieure à $1-\alpha$)

Rappel (?) : $F_n = \frac{X_n}{n}$ est la variable aléatoire « fréquence », qui, comme son nom l'indique, mesure la fréquence des résultats obtenus.

Démonstration : on pose $Z_n = \frac{X_n - np}{\sqrt{np(1-p)}}$.

Rappel : np est la moyenne, et $\sqrt{np(1-p)}$ est l'écart-type de la loi binomiale X_n.

Alors d'après le théorème de De Moivre-Laplace $\lim_{n\to+\infty} P(-u_\alpha \le Z_n \le u_\alpha) = P(-u_\alpha \le Z \le u_\alpha) = 1-\alpha$, où Z suit la loi normale $\mathscr{N}(0\,;\,1)$ (on se souvient du cours sur les lois : « $P(a \le Z_n \le b)$ tend vers $\int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$ lorsque n tend vers $+\infty$ », et $\exists!\, u_\alpha$ tel que $P(-u_\alpha \le X \le u_\alpha) = 1-\alpha$.)

Or $P(-u_\alpha \le Z_n \le u_\alpha) = P\left(-u_\alpha \le \frac{X_n - np}{n} \times \frac{\sqrt{n}}{\sqrt{p(1-p)}} \le u_\alpha\right)$ on va isoler X_n :

$$= P\left(np - u_\alpha\sqrt{np(1-p)} \le X_n \le np + u_\alpha\sqrt{np(1-p)}\right) \text{ on va diviser par } n :$$

$$= P\left(p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \le \frac{X_n}{n} \le p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)$$

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

donc $\lim_{n\to+\infty} P\left(\frac{X_n}{n} \in I_n\right) = P(-u_\alpha \le Z \le u_\alpha) = 1-\alpha$.

Remarque pratique : on admet que pour les conditions suivantes :

- $n \geq 30$
- $np \geq 5$
- $n(1-p) \geq 5$ alors on peut approcher $P\left(\frac{X_n}{n} \in I_n\right)$ par $1-\alpha$.

Exemple : pour $\alpha = 0{,}05$, on a vu que $u_{0,05} \simeq 1{,}96$. On prendra donc comme intervalle de fluctuation au seuil de 95% de F_n l'intervalle $I_n = \left[p - 1{,}96 \cdot \frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 1{,}96 \cdot \frac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$.

De même, au seuil de 99%, $I_n = \left[p - 2{,}58 \cdot \frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 2{,}58 \cdot \frac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$.

Prise de décision au seuil de 5%

On cherche à savoir, au seuil de décision (ou au risque) de 5%, si la proportion p du caractère C dans la population vaut $p = p_0$ ou non, à partir d'un échantillon de taille n, dans lequel la proportion observée de C vaut f.

On suppose que les conditions pratiques sont réunies ($n \geq 30$, $np \geq 5$, $n(1-p) \geq 5$).

On raisonne ainsi :

- Calcul de $I = \left[p - 1{,}96 \cdot \frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}} ; p + 1{,}96 \cdot \frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}} \right]$
- La fréquence f appartient-elle à I ?
- On applique la règle de décision au seuil de risque de 5% :
  - Si $f \notin I$, on rejette l'hypothèse $p = p_0$ avec un risque de 5 % de se tromper.
  - Si $f \in I$, on accepte l'hypothèse (on ne connaît pas dans ce cas le risque de se tromper).

Remarque :

On n'a pas démontré « il y a 95 % de chances que la proportion soit p_0 ». La proportion étant fixée, soit elle est de p_0, soit non (elle est de p_0 totalement, à 100 % ou pas du tout, à 100 % aussi !). De même, dans aucun cas on n'a démontré que la proportion est de p_0 ou non, on prend juste une décision de la considérer comme telle. On peut donc toujours se tromper. Cette décision n'est plus du ressort du mathématicien, mais de celui du scientifique qui demande au mathématicien de faire le test.

Par ailleurs, le risque est de refuser à tort le modèle. Plus le risque est petit et plus on aura tendance à accepter le modèle. Ceci peut être gênant. Donc, ce n'est pas parce que le risque est plus petit que c'est « mieux »… Par exemple, il vaut peut-être mieux choisir de ne pas dire qu'un médicament est efficace (au seuil de risque de 99 %) si ses effets secondaires peuvent être très graves. Et à l'inverse, accepter comme efficace un médicament (au seuil de risque de 95 %) sans effets secondaires.

Le risque de se tromper de 5 % est la probabilité conditionnelle que l'on ait déclaré $p \neq p_0$ « sachant » que $p = p_0$ (« sachant » entre guillemets car on ne le sait pas… puisqu'on se trompe !). De plus, cette probabilité vaut approximativement 5 % et non exactement 5 % : le théorème ci-dessus utilise une limite.

Exemple : M. et Mme Gluckenstimmelinsdörf attendent un enfant. Ils pensent avoir une chance sur deux d'avoir un garçon, et une chance sur deux d'avoir une fille. En se rendant sur le site de l'INED, ils apprennent qu'en 2010, il y a eu, en France Métropolitaine, 410140 garçons sur 802224 naissances (soit une fréquence F égale à 0,51). On souhaite savoir si cette value observée de la variable aléatoire F_n permet de remettre en cause l'hypothèse d'équiprobabilité des sexes à la naissance formulée par le couple Gluckenstimmelinsdörf.

- On fait l'hypothèse que le couple Gluckenstimmelinsdörf a raison, et qu'une naissance est une expérience aléatoire de probabilité 0,5 d'avoir une fille. Soit X_n la variable aléatoire comptant le

nombre de naissances de filles dans un échantillon aléatoire de n naissances. Quelle est la loi suivie par X_n ?

- X_n suit $\mathscr{B}(802224 ; 0{,}5)$.

- Déterminer l'intervalle de fluctuation asymptotique au seuil de 95 % correspondant à la variable aléatoire $F_n = \frac{X_n}{n}$. Les conditions sont-elles remplies pour pouvoir utiliser l'approximation $P(F_n \in I_n) \simeq 0{,}95$?
  - $I_{802224} = \left[0{,}5 - 1{,}96 \cdot \frac{\sqrt{0{,}5^2}}{\sqrt{802224}} ; 0{,}5 + 1{,}96 \cdot \frac{\sqrt{0{,}5^2}}{\sqrt{802224}}\right] = [0{,}499 ; 0{,}501]$ et conditions remplies.
- Conclure.
  - $F \notin I_{802224}$ donc on peut rejeter l'hypothèse « on a une chance sur deux que l'enfant soit une fille/un garçon, en France Métropolitaine, en 2010 ».
  -

2. Estimation

Contexte : dans une population, on prélève un échantillon de taille n. La fréquence d'individus présentant le caractère C dans l'échantillon est f. Que peut-on dire de la proportion p de C dans l'ensemble de la population ?

Théorème : soit X_n une variable aléatoire suivant la loi binomiale $\mathscr{B}(n ; p)$, avec $0 < p < 1$. Soit $F_n = \frac{X_n}{n}$.

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

Formulation équivalente : il existe $n_0 \in \mathbb{N}$ tel que si $n \geq n_0$ alors $P\left(p \in \left[F_n - \frac{1}{\sqrt{n}} ; F_n + \frac{1}{\sqrt{n}}\right]\right) \geq 0{,}95$.

Démonstration :

On a vu au 1°) que l'intervalle de fluctuation à 95% de F_n est $I_n = \left[p - 1{,}96 \cdot \frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 1{,}96 \cdot \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$.

Or, comme $0 \leq p \leq 1$, on montre facilement que $1{,}96 \cdot \sqrt{p(1-p)}$ est majoré par 1 (en étudiant la fonction $x \mapsto 1{,}96 \cdot \sqrt{p(1-p)}$ sur $[0;1]$) :

On en déduit que $\left[p - 1{,}96 \cdot \frac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 1{,}96 \cdot \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right] \subset \left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}}\right]$ (intervalle que vous avez peut-être vu en seconde).

Comme $F_n = \frac{X_n}{n}$ appartient à I_n avec une probabilité supérieure ou égale à 0,95, et que l'intervalle $\left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}}\right]$ est plus grand, alors la probabilité que F_n appartienne à ce dernier intervalle est aussi supérieure ou égale à 0,95 (l'intuition nous dit que cette probabilité est encore plus grande, mais méfiance ici : ce n'est pas valable pour toutes les valeurs de n –il faut n « grand »-).

On a donc le lemme suivant :

Lemme : soit X_n une variable aléatoire suivant la loi binomiale $\mathscr{B}(n ; p)$, avec $0 < p < 1$. Soit $F_n = \frac{X_n}{n}$.

Il existe $n_0 \in \mathbb{N}$ tel que si $n \geq n_0$ alors $P\left(p - \frac{1}{\sqrt{n}} \leq F_n \leq p + \frac{1}{\sqrt{n}}\right) \geq 0{,}95$.

Il suffit alors pour obtenir le théorème de constater que $p - \frac{1}{\sqrt{n}} \leq F_n \leq p + \frac{1}{\sqrt{n}} \Leftrightarrow F_n - \frac{1}{\sqrt{n}} \leq p \leq F_n + \frac{1}{\sqrt{n}}$

Définition : Soit *f* la fréquence d'un caractère C sur un échantillon aléatoire de taille *n*, prélevé dans une population donnée.

L'intervalle $\left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}\right]$ est un intervalle de confiance à 95% de la proportion inconnue *p* du caractère C dans l'ensemble de la population.

Remarques :

- Les conditions pratiques d'utilisation sont les mêmes que pour l'intervalle de fluctuation :
  - $n \geq 30$
  - $np \geq 5$
  - $n(1-p) \geq 5$
- L'intervalle de confiance dépend donc de la taille de l'échantillon utilisé, mais pas de la taille de la population. En effet, si on « sonde » toute la population, alors on connaît précisément *p* !
- La précision de l'intervalle est donnée par sa longueur de $\frac{2}{\sqrt{n}}$.
- « Pour de vrai », on utilise plutôt $\left[f - 1{,}96 \cdot \frac{\sqrt{p(1-p)}}{\sqrt{n}}; f + 1{,}96 \cdot \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$, qu'on ne peut pas justifier en terminale, qui est un peu plus précis que $\left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}}\right]$.

Exemple :

Voici les résultats d'un sondage IPSOS effectué pour Le Figaro et Europe 1, les 17 et 18 avril 2002 auprès de 989 personnes, constituant un échantillon national représentatif de la population française âgée de plus de 18 ans et inscrite sur les listes électorales.

On suppose cet échantillon réalisé de manière aléatoire. En pratique ce n'est pas du tout le cas, les techniques de sondage ne relèvent pas des mathématiques mais d'un artisanat, certes très travaillé, mais non scientifique (ce n'est pas un jugement de valeur).

Les intentions de vote au premier tour pour les principaux candidats à l'élection présidentielle étaient les suivants : 20 % pour J. Chirac, 18 % pour L. Jospin, et 14 % pour J. M. Le Pen.

- Déterminer pour chaque candidat l'intervalle de confiance au niveau de 0,95, de la proportion inconnue d'électeurs ayant l'intention de voter pour lui.
- Le 21 avril 2002, les résultats du premier tour de l'élection présidentielle sont les suivants : 19,88 % pour J. Chirac, 16,86 % pour J. M. Le Pen et 16,18 % pour L. Jospin. Ces résultats sont-ils cohérents avec les intervalles de confiance ? Pouvait-on écarter un des trois candidats pour le second tour, au vu desdits intervalles ?
  - $\left[f - \frac{1}{\sqrt{989}}; f + \frac{1}{\sqrt{989}}\right]$ donne respectivement $[0{,}168\,;0{,}232]$ pour Chichi, $[0{,}148\,;0{,}212]$ pour Jojo, et $[0{,}108\,;0{,}172]$ pour l'Affreux. Soit une marge d'erreur de ± 3% (si le sondage avait été fait suivant la méthode mathématique).
  - Les résultats obtenus sont bien dans les intervalles de confiance. Les trois intervalles se recouvrent, il était donc impossible d'établir un classement fiable.

3. Résumé.

Règle générale :

- On utilise un intervalle de fluctuation lorsque la proportion *p* dans la population est connue ou si l'on fait une hypothèse sur sa valeur.
- On utilise un intervalle de confiance lorsque l'on veut estimer une proportion inconnue dans une population.

Exemples :

- Test de conformité d'une proportion : on veut déterminer si la proportion observée dans un échantillon est conforme à une valeur de référence connue dans la population. Sous l'hypothèse que l'échantillon est issu d'un tirage aléatoire correspondant à un schéma de Bernoulli (tirage avec remise ou s'y apparentant), la variable fréquence appartient à un intervalle de fluctuation avec une probabilité déterminée.
  En fonction de l'appartenance ou non de la fréquence observée à cet intervalle, on peut prendre une *décision* concernant la conformité de l'échantillon.
  Si les conditions d'utilisation sont réunies, on détermine l'intervalle de fluctuation asymptotique, sinon on a recours à un intervalle de fluctuation calculé avec la loi binomiale.
- Estimation d'une proportion inconnue p grâce à un échantillon aléatoire
  On se place dans le cas où l'échantillon comporte au moins 30 éléments afin de pouvoir utiliser l'intervalle de confiance au programme.
  Si la fréquence observée f est telle que $n \cdot f \geq 5$ et $n \cdot (1-f) \geq 5$, on considère qu'on peut conclure qu'un intervalle de confiance de p au niveau de confiance 0,95 est $\left[f - \frac{1}{\sqrt{n}} ; f + \frac{1}{\sqrt{n}} \right]$.

4. Limites des statistiques

   Aux USA, les cantons dans lesquels il y a le moins de cancers du rein sont ruraux, peu densément peuplés, et votent majoritairement républicains. Conclusion ?
   Aux USA, les cantons dans lesquels il y a le plus de cancers du rein sont ruraux, peu densément peuplés, et votent majoritairement républicains. Conclusion ?

   ➔ les deux statistiques précédentes sont vraies.
   ➔ On peut tout rationnaliser, c'est une des caractéristiques de notre fonctionnement intellectuel
   ➔ Les échantillons sont trop petits pour donner une conclusion fiable.