Chapitre 19. Echantillonnage. Estimation

• Ce chapitre ne peut en aucun cas être étudié si on n'a pas d'abord étudié le chapitre 16 sur la loi binomiale et le chapitre 18 sur la loi normale.
• Les différentes expressions mathématiques écrites dans ce chapitre ont un aspect très rebutant. On ne doit cependant pas s'effrayer du contenu. Au bout du compte, comme le signale le programme officiel, « les attendus de ce chapitre sont modestes » et ce que doit savoir faire un élève est très simple.

I. Intervalles de fluctuation associés à une loi binomiale

1) La variable aléatoire fréquence associée à une loi binomiale

On étudie une certaine population où chaque individu de la population possède un certain caractère C avec une probabilité p ou ne possède pas le caractère C avec une probabilité $1-p$. La probabilité p est donc la proportion de la population qui est constituée de personnes possédant le caractère C.

Par exemple, dans un pays, chaque individu vote ou ne vote pas pour un candidat donné, dans une urne contenant un grand nombre de boules blanches et noires, chaque boule est blanche ou pas, dans une population de vaches, chaque vache a ou n'a pas une certaine maladie, dans une population de télévisions, chaque appareil peut tomber en panne ou pas ... Le mot « individu » a donc un sens très général.

Par la suite, on supposera que $p \in]0,1[$ car si $p=1$ (ou $p=0$), tout le monde possède la caractère C (ou personne ne le possède) et il n'y a pas besoin de faire des calculs de probabilités.

On prélève maintenant un **échantillon** de cette population de taille n et on compte dans cet échantillon le nombre d'individus possédant le caractère C. On fait donc de l'**échantillonnage**.

On note X_n la variable aléatoire qui, à chaque échantillon de taille n associe le nombre de personnes possédant le caractère C. Si la taille n d'un échantillon est petit devant la taille de la population, on peut assimiler un échantillon (qui est un tirage successif sans remise de n individus dans la population) à un tirage successif avec remise de n individus dans la population. En effet, si n est petit devant la population, après avoir tiré un individu sans le remettre dans la population, la probabilité que l'individu suivant possède le caractère C n'a quasiment pas varié.

On sait que la variable aléatoire X_n suit la loi binomiale $\mathscr{B}(n,p)$ de paramètres n et p. On définit alors la **variable aléatoire fréquence** associée à la variable X_n :

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

Le caractère C peut apparaître, 0 fois, une fois, deux fois, ..., n fois dans l'échantillon de taille n et ceci avec une fréquence égale à $\frac{0}{n}=0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{k}{n}, \ldots, \frac{n-1}{n}, \frac{n}{n}=1$ suivant le cas. Les valeurs prises par la variable F_n sont ces fréquences et chaque fréquence est une réalisation de la variable F_n et a une certaine probabilité d'être obtenue.

La variable aléatoire F_n ne prend pas des valeurs entières contrairement à X_n. Par contre, les probabilités sont les mêmes :

$$\text{pour tout entier naturel } k \text{ tel que } 0 \leqslant k \leqslant n,\ P\left(F_n = \frac{k}{n}\right) = P(X_n = k) = \binom{n}{k} \times p^k \times (1-p)^{n-k}.$$

On note que l'espérance de F_n est

$$E(F_n) = E\left(\frac{X_n}{n}\right) = \frac{1}{n}E(X_n) = \frac{1}{n} \times np = p$$

et l'écart-type de F_n est

$$\sigma(F_n) = \sigma\left(\frac{X_n}{n}\right) = \frac{1}{n}\sigma(X_n) = \frac{1}{n} \times \sqrt{np(1-p)} = \sqrt{\frac{np(1-p)}{n^2}} = \sqrt{\frac{p(1-p)}{n}}.$$

Par exemple, en France, la probabilité qu'un individu donné regarde une certaine émission de télé est $p = \frac{1}{8}$.
On demande à chaque personne d'un échantillon de 500 personnes si cette personne a ou n'a pas regardé l'émission. Si on note F_{500} la variable aléatoire égale au nombre de personnes ayant effectivement regardé l'émission, la loi de probabilité de F_{500} est

$$\text{pour tout entier naturel } k \text{ tel que } 0 \leqslant k \leqslant 500,\ p\left(F_{500} = \frac{k}{500}\right) = \binom{500}{k}\left(\frac{1}{8}\right)^k\left(\frac{7}{8}\right)^{500-k}.$$

Ainsi, la probabilité qu'exactement 100 personnes de l'échantillon aient regardé l'émission est environ 6×10^{-7}, la probabilité que moins de 60 personnes de l'échantillon aient regardé l'émission est environ $0,4$ (ces valeurs ont été obtenues avec Excel).

Enfin, dire que l'espérance de F_{500} est égale à $\frac{1}{8} = 0,125$ revient à dire que, en répétant un grand nombre de fois l'expérience qui consiste à choisir un échantillon de taille 500, en moyenne, sur l'ensemble des échantillons de taille 500, une personne sur 8 d'un échantillon regarde l'émission.

2) Intervalles de fluctuation associés à une loi binomiale

a) Intervalles de fluctuation. Intervalles de fluctuation asymptotique

Notations. Soit X_n une variable aléatoire suivant une loi binomiale de paramètres $n \in \mathbb{N}^*$ et $p \in]0,1[$.
On note $F_n = \dfrac{X_n}{n}$ la variable fréquence associée, $Z_n = \dfrac{X_n - np}{\sqrt{np(1-p)}}$ la loi centrée réduite associée à X_n et Z une variable aléatoire suivant la loi normale centrée réduite. ❑

Dans ce paragraphe, **on suppose connue la proportion** p**.** On choisit un échantillon de taille n dans la population et on s'intéresse aux probabilités du type $P(a \leqslant F_n \leqslant b)$ ou encore $P(F_n \in [a,b])$.

L'intervalle $[a,b]$ est un intervalle dans lequel **fluctue** la variable fréquence F_n avec une certaine probabilité calculée à partir de la fonction de répartition de la loi binomiale. De manière générale, on peut définir la notion d'intervalle de fluctuation à un certain seuil :

Définition 1. Soit α un réel de $]0,1[$.
On dit que l'intervalle $[a,b]$ est un intervalle de fluctuation de $F_n = \dfrac{X_n}{n}$ au seuil $1-\alpha$ si et seulement si
$$P(F_n \in [a,b]) \geqslant 1-\alpha.$$

On va maintenant s'intéresser à un intervalle qui est une approximation d'un intervalle de fluctuation. C'est la notion d'**intervalle de fluctuation asymptotique**.

On revient sur le théorème de Moivre-Laplace (théorème 1 du chapitre 18). Celui-ci a la conséquence suivante :

Théorème 1. Soit $\alpha \in]0,1[$. Soit $I_n = \left[p - u_\alpha \dfrac{\sqrt{p(1-p)}}{\sqrt{n}}, p + u_\alpha \dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$. Alors,
$$\lim_{n\to+\infty} P(F_n \in I_n) = 1-\alpha.$$

Commentaire. Le fractile u_α a été défini dans le théorème 4 du chapitre 18 : u_α est l'unique réel strictement positif tel que $P(-u_\alpha \leqslant Z \leqslant u_\alpha) = 1-\alpha$. ❑

Démonstration. Soit α un réel élément de $]0,1[$.

$$\begin{aligned}
F_n \in I_n &\Leftrightarrow \frac{X_n}{n} \in I_n \Leftrightarrow p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant \frac{X_n}{n} \leqslant p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}} \\
&\Leftrightarrow np - u_\alpha \frac{n\sqrt{p(1-p)}}{\sqrt{n}} \leqslant X_n \leqslant np + u_\alpha \frac{n\sqrt{p(1-p)}}{\sqrt{n}} \\
&\Leftrightarrow np - u_\alpha \sqrt{np(1-p)} \leqslant X_n \leqslant np + u_\alpha \sqrt{np(1-p)} \text{ (car } \frac{n}{\sqrt{n}} = \frac{\sqrt{n}\times\sqrt{n}}{\sqrt{n}} = \sqrt{n}) \\
&\Leftrightarrow -u_\alpha \sqrt{np(1-p)} \leqslant X_n - np \leqslant u_\alpha \sqrt{np(1-p)} \\
&\Leftrightarrow -u_\alpha \leqslant \frac{X_n - np}{\sqrt{np(1-p)}} \leqslant u_\alpha \\
&\Leftrightarrow -u_\alpha \leqslant Z_n \leqslant u_\alpha.
\end{aligned}$$

Puisque les événements $F_n \in I_n$ et $-u_\alpha \leqslant Z_n \leqslant u_\alpha$ se produisent simultanément, on en déduit qu'ils ont la même probabilité :

$$P(F_n \in I_n) = P(-u_\alpha \leqslant Z_n \leqslant u_\alpha).$$

• Le théorème de Moivre-Laplace affirme que $\lim\limits_{n\to+\infty} P(-u_\alpha \leqslant Z_n \leqslant u_\alpha) = P(-u_\alpha \leqslant Z \leqslant u_\alpha)$ (où Z suit une loi normale centrée réduite).
• D'autre part, $P(-u_\alpha \leqslant Z \leqslant u_\alpha) = 1-\alpha$ par définition du fractile u_α.

Donc $\lim\limits_{n\to+\infty} P(F_n \in I_n) = 1-\alpha$.

L'intervalle $I_n = \left[p - u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}, p + u_\alpha \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ a donc la propriété suivante : quand n est grand, la probabilité que F_n appartienne à I_n vaut environ $1-\alpha$. Ce n'est pas un intervalle de fluctuation au seuil $1-\alpha$ au sens de la définition 1 car la probabilité que F_n appartienne à I_n peut être strictement plus petite que $1-\alpha$. Néanmoins, l'intervalle I_n « approche » un intervalle de fluctuation au seuil $1-\alpha$. On dit que I_n est un **intervalle de fluctuation asymptotique** au seuil $1-\alpha$.

Le programme officiel de Terminale S donne la définition suivante d'un intervalle de fluctuation asymptotique :

Définition 2. Un intervalle de fluctuation asymptotique de la variable aléatoire F_n au seuil $1-\alpha$ est un intervalle déterminé à partir de p et de n et qui contient F_n avec une probabilité d'autant plus proche de $1-\alpha$ que n est grand.

On analyse maintenant un cas particulier d'intervalle de fluctuation asymptotique. Dans le chapitre précédent, on a donné la valeur de u_α quand $\alpha = 0,05$ de sorte que $1-\alpha = 0,95$:

$$u_{0,05} = 1,959\ldots$$

En particulier, $u_{0,05} < 1,96$. Par suite, l'intervalle $\left[p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ contient l'intervalle $I_n = \left[p - u_{0,05}\frac{\sqrt{p(1-p)}}{\sqrt{n}}, p + u_{0,05}\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$. On en déduit que

$$P\left(\frac{X_n}{n} \in \left[p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]\right) \geqslant P\left(\frac{X_n}{n} \in \left[p - u_{0,05}\frac{\sqrt{p(1-p)}}{\sqrt{n}}, p + u_{0,05}\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]\right).$$

Comme $\lim\limits_{n\to+\infty} P\left(\frac{X_n}{n} \in \left[p - u_{0,05}\frac{\sqrt{p(1-p)}}{\sqrt{n}}, p + u_{0,05}\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]\right) = 0,95$, on a donc

$$\lim_{n\to+\infty} P\left(\frac{X_n}{n} \in \left[p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]\right) \geqslant 0,95.$$

On peut même être plus précis. La calculatrice donne

$$\lim_{n\to+\infty} P\left(\frac{X_n}{n} \in \left[p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]\right) = P(-1,96 \leqslant Z \leqslant 1,96) = 0,950004\ldots$$

Donc, $\lim\limits_{n\to+\infty} P\left(\frac{X_n}{n} \in \left[p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]\right) > 0,95$. Par suite, en appliquant la définition de la limite d'une suite, à partir d'un certain rang, on a $P\left(\frac{X_n}{n} \in \left[p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]\right) \geqslant 0,95$ et on peut donc énoncer :

Théorème 2. Pour tout réel $p \in]0,1[$, il existe un entier naturel non nul n_0 tel que, pour tout entier naturel n supérieur ou égal à n_0,

$$P\left(F_n \in \left[p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]\right) \geqslant 0,95.$$

Commentaire. L'entier n_0 dépend du réel p. Par exemple, on peut montrer que si p est proche de 0 ou proche de 1, l'entier n_0 est bien plus grand que si p est proche de $\frac{1}{2}$. ❑

Ainsi, l'intervalle $\left[p - 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}; p + 1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$ est une approximation d'un intervalle de fluctuation au seuil $0,95$ quand n est grand ce qui motive une fois de plus l'appellation « intervalle de fluctuation asymptotique ».

Dans la pratique, on a l'habitude d'utiliser cet intervalle quand $n \geqslant 30$, $np \geqslant 5$, $n(1-p) \geqslant 5$ (ces conditions sont établies grâce à des calculs dépassant largement le niveau d'une classe de terminale).

Résumons tout ce qui précède. Ce que l'on doit retenir et qui est l'essentiel du cours est :

1) L'intervalle de fluctuation asymptotique au seuil 95% est

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

2) si $n\geqslant 30$, $np\geqslant 5$ et $n(1-p)\geqslant 5$,

$$P\left(F_n\in\left[p-1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}};p+1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]\right) \text{ est environ égale à } 0,95.$$

3) pour tout réel $p\in]0,1[$, il existe un entier n_0 dépendant de p tel que si $n\geqslant n_0$

$$P\left(F_n\in\left[p-1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}};p+1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]\right)\geqslant 0,95.$$

Commentaire. Attention, l'entier 30 du 2) n'est pas l'entier n_0 du 3) car même si on est dans les conditions de validité $n\geqslant 30$, $np\geqslant 5$ et $n(1-p)\geqslant 5$, il est possible que la probabilité $P(F_n\in I_n)$ soit strictement plus petite que $0,95$ bien que proche de $0,95$. ❑

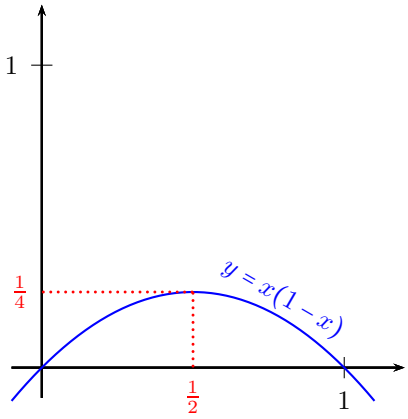
b) L'intervalle de fluctuation de la classe de seconde

On va voir que l'« intervalle de la classe de seconde » $J_n=\left[p-\frac{1}{\sqrt{n}},p+\frac{1}{\sqrt{n}}\right]$ contient l'« intervalle de Terminale » $I_n=\left[p-1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}};p+1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$.

Il s'agit pour cela de vérifier que $p-\frac{1}{\sqrt{n}}\leqslant p-1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\leqslant p+1,96\frac{\sqrt{p(1-p)}}{\sqrt{n}}\leqslant p+\frac{1}{\sqrt{n}}$ ce qui revient à vérifier que $1,96\sqrt{p(1-p)}\leqslant 1$.

La fonction $f:x\mapsto x(1-x)=-x^2+x$ est un polynôme du second degré s'annulant en 0 et 1. Puisque le coefficient de x^2 est strictement négatif, la fonction f admet un maximum en $\frac{1}{2}$ qui est le milieu de $[0,1]$. Ce maximum est égal à

$$f\left(\frac{1}{2}\right)=\frac{1}{2}\left(1-\frac{1}{2}\right)=\frac{1}{4}.$$



Par suite, pour tout réel p de $]0,1[$, $p(1-p)\leqslant\frac{1}{4}$ puis

$$1,96\sqrt{p(1-p)}\leqslant 2\times\sqrt{\frac{1}{4}}=1.$$

On peut donc énoncer :

Théorème 3. **1)** Pour tout réel $p\in]0,1[$ et tout entier naturel non nul n,

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

2) Il existe un entier naturel non nul n_0 tel que pour $n\geqslant n_0$, $P\left(F_n\in\left[p-\frac{1}{\sqrt{n}},p+\frac{1}{\sqrt{n}}\right]\right)\geqslant 0,95$.

3) Pour $n\geqslant 30$, $np\geqslant 5$, $n(1-p)\geqslant 5$, $P\left(F_n\in\left[p-\frac{1}{\sqrt{n}},p+\frac{1}{\sqrt{n}}\right]\right)$ est environ égale à $0,95$.

3) Prise de décision au seuil de 5%

Dans ce paragraphe, p **n'est pas connue** mais on fait une hypothèse sur p : on suppose que p est égale à une certaine valeur précise p_0. On se demande alors si notre hypothèse est cohérente à partir de la connaissance d'un échantillon de taille n.

On adopte la démarche suivante :

- On calcule la fréquence f d'appartion du caractère C dans l'échantillon.
- On calcule l'intervalle de fluctuation asymptotique au seuil $0,95$:

$$I = \left[p_0 - 1,96\frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}, p_0 + 1,96\frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}\right].$$

- On applique la **règle de décision** suivante :

Si f n'appartient pas à l'intervalle de fluctuation I, on rejette l'hypothèse que $p = p_0$ au risque de se tromper de 5% et si f appartient à l'intervalle I, on peut accepter l'hypothèse faite sur p mais on ne connait pas le rsique de se tromper.

Exercice 1. Dans un site sur Internet, on trouve le tableau ci-dessous qui donne la répartition des groupes sanguins en France :

Rhésus \ Groupe sanguin	O	A	B	AB	Total
Rhésus positif	37%	39%	7%	2%	85%
Rhésus négatif	6%	6%	2%	1%	15%
Total	43%	45%	9%	3%	100%

Nous allons tester si ce tableau peut être considéré comme correct à partir d'un sondage effectué dans une classe de Terminale S constituée de 36 élèves.

1) a) Quelle hypothèse peut-on faire sur la proportion de personnes ayant un rhésus négatif en France ?
b) Déterminer l'intervalle de fluctuation asymptotique au seuil 95% correspondant.
c) Enoncer la règle de décision que l'on va appliquer.
d) Dans la classe, 6 élèves ont un rhésus négatif. Que peut-on en conclure ?

2) a) Quelle hypothèse peut-on faire sur la proportion de personnes étant du groupe O en France ?
b) Déterminer l'intervalle de fluctuation asymptotique au seuil 95% correspondant.
c) Dans la classe, 9 élèves sont du groupe O. Que peut-on en conclure ?

3) Quelle conséquence peut-on en tirer sur les données du tableau ?

Solution.

1) a) On peut supposer qu'en France, la proportion des personnes ayant un Rhésus négatif est $p_0 = 0,15$.

b) Ici, $n = 36$ et $p = 0,15$. On note que $n \geqslant 30$ puis $np = 5,4$ et donc $np \geqslant 5$ et enfin $n(1-p) = 30,6$ et donc $n(1-p) \geqslant 5$.

L'intervalle de fluctuation asymptotique au seuil 95% correspondant est

$$\left[0,15 - 1,96\frac{\sqrt{0,15\times 0,85}}{\sqrt{36}}; 0,15 + 1,96\frac{\sqrt{0,15\times 0,85}}{\sqrt{36}}\right].$$

En arrondissant les bornes de cet intervalle de manière à l'élargir un peu, on trouve l'intervalle $[0,033; 0,267]$.

c) La règle de décision est : « si la fréquence observée dans la classe n'appartient pas à cet intervalle, on rejette l'hypothèse que $p_0 = 0,15$ avec un risque de 5% de se tromper et si la fréquence observée appartient à cet intervalle, on accepte l'hypothèse que $p_0 = 0,15$ mais on ne connait pas le risque de se tromper dans ce cas ».

d) La fréquence des élèves ayant un rhésus négatif dans la classe est $\dfrac{6}{36} = 0,166\ldots$ Cette fréquence appartient à l'intervalle de fluctuation et on peut donc accepter l'hypothèse que 15% des personnes ont un rhésus négatif mais on ne connaît pas le risque de se tromper.

2) a) On peut supposer qu'en France, la proportion des personnes étant du groupe O est $p_0 = 0,43$.

b) Donc $n = 36$ et $p = 0,43$. Donc, $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$.
L'intervalle de fluctuation asymptotique au seuil 95% correspondant est

$$\left[0,43 - 1,96\frac{\sqrt{0,43 \times 0,57}}{\sqrt{36}}; 0,43 + 1,96\frac{\sqrt{0,43 \times 0,57}}{\sqrt{36}}\right].$$

En arrondissant les bornes de cet intervalle de manière à l'élargir un peu, on trouve l'intervalle $[0,268; 0,592]$.

c) La fréquence des élèves étant du groupe O dans la classe est $\frac{9}{36} = 0,25\ldots$ Cette fréquence n'appartient pas à l'intervalle de fluctuation et on peut donc rejeter l'hypothèse que 43% des personnes sont du groupe O avec un risque de se tromper de 5%.

3) Si le pourcentage des rhésus négatifs est (peut-être) correct et le pourcentage des groupes O est (peut-être) incorrect, il est possible que le pourcentage des personnes du groupe O^+ donnée dans le tableau, à savoir 37%, soit un peu faux (peut-être).

II. Estimation

Dans le paragraphe précédent, on connaissait la proportion p ou on faisait une hypothèse sur cette proportion.

Dans ce paragraphe, nous allons faire l'inverse. **On connaît la fréquence** f d'apparition du caractère C dans un certain échantillon de taille n et **on ne connaît pas la proportion** p de la population possédant le caractère C. On veut alors évaluer la proportion p. On note que la fréquence f est une réalisation de la variable aléatoire F_n.

Cette situation est par exemple celle des sondages d'opinion avant une élection. On connaît, dans un « échantillon représentatif » de la population, le pourcentage de gens votant pour tel candidat ou encore la fréquence f de personnes de l'échantillon votant pour ce candidat, et on cherche à évaluer la proportion p de personnes votant pour ce candidat dans la population toute entière. La réponse apportée ne sera pas pas la valeur exacte de p bien sûr, mais une fourchette dans laquelle se situe p avec un certain niveau de confiance. Cette fourchette est un intervalle associé à un certain niveau de confiance $1-\alpha$, intervalle appelé **intervalle de confiance au niveau** $1-\alpha$. Par la suite, nous n'envisagerons que le niveau de confiance 0,95.

Il s'agit d'inverser les rôles de f et p à partir de l'intervalle de fluctuation asymptotique au seuil 0,95 de la classe de seconde. Ceci se fait de la façon suivante :

$$\begin{aligned} p - \frac{1}{\sqrt{n}} \leqslant F_n \leqslant p + \frac{1}{\sqrt{n}} &\Leftrightarrow p - \frac{1}{\sqrt{n}} \leqslant F_n \text{ et } F_n \leqslant p + \frac{1}{\sqrt{n}} \\ &\Leftrightarrow p \leqslant F_n + \frac{1}{\sqrt{n}} \text{ et } F_n - \frac{1}{\sqrt{n}} \leqslant p \\ &\Leftrightarrow F_n - \frac{1}{\sqrt{n}} \leqslant p \leqslant F_n + \frac{1}{\sqrt{n}} \end{aligned}$$

En particulier, $P\left(p - \frac{1}{\sqrt{n}} \leqslant F_n \leqslant p + \frac{1}{\sqrt{n}}\right) = P\left(F_n - \frac{1}{\sqrt{n}} \leqslant p \leqslant F_n + \frac{1}{\sqrt{n}}\right)$. Le théorème 3 peut alors se réénoncer de la façon suivante :

Théorème 4. **1)** Il existe un entier naturel non nul n_0 tel que pour $n \geqslant n_0$, l'intervalle aléatoire $\left[F_n - \frac{1}{\sqrt{n}}, F_n + \frac{1}{\sqrt{n}}\right]$ contient p avec une probabilité supérieure ou égale à 0,95.

2) Pour $n \geqslant 30$, $nf \geqslant 5$, $n(1-f) \geqslant 5$, l'intervalle aléatoire $\left[F_n - \frac{1}{\sqrt{n}}, F_n + \frac{1}{\sqrt{n}}\right]$ contient p avec une probabilité environ égale à 0,95.

Revenons à notre problème : on se place dans la situation où on connaît la fréquence f d'apparition du caractère C dans un échantillon de taille n.
L'intervalle $\left[f - \frac{1}{\sqrt{n}}, f + \frac{1}{\sqrt{n}}\right]$ est alors une réalisation de l'intervalle aléatoire $\left[F_n - \frac{1}{\sqrt{n}}, F_n + \frac{1}{\sqrt{n}}\right]$.
Si n est suffisament grand, on peut affirmer avec un certain niveau de confiance que p appartient à cet intervalle. Plus précisément,

Définition 3. Si $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$, l'intervalle $\left[f - \frac{1}{\sqrt{n}}, f + \frac{1}{\sqrt{n}}\right]$ est appelé **intervalle de confiance de la proportion** p **au niveau de confiance** 95%.

Commentaire 1. Chaque choix d'un échantillon fournit une fréquence f et donc chaque choix d'un échantillon fournit un intervalle de confiance. Deux échantillons différents fourniront éventuellement deux intervalles de confiance différents. ❑

Commentaire 2. L'intervalle $\left[f-\frac{1}{\sqrt{n}}, f+\frac{1}{\sqrt{n}}\right]$ contient p avec un niveau de confiance environ égal à 95% si $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$ et avec un niveau de confiance supérieur ou égal à 95% si n dépasse un certain entier n_0. ❑

Exercice 2. Une élection est organisée. Il y a deux candidats C_1 et C_2.

1) Un sondage est effectué auprès d'un échantillon représentatif de 600 personnes.
342 personnes de l'échantillon ont l'intention de voter pour le candidat C_1.
Déterminer une fourchette dans laquelle se trouvera le pourcentage de personnes votant pour le candidat C_1 lors de l'élection avec un niveau de confiance de 95%.

2) Déterminer la taille minimale que doit avoir l'échantillon pour que l'amplitude de l'intervalle de confiance soit au maximum de 2%.

Solution.

1) La fréquence de personnes de l'échantillon votant pour le candidat C_1 est

$$f = \frac{342}{600} = 0,57.$$

L'intervalle de confiance au niveau de confiance 95% est

$$I = \left[0,57-\frac{1}{\sqrt{600}}, 0,57+\frac{1}{\sqrt{600}}\right] = [0,529\ldots, 0,610\ldots].$$

En arrondissant les bornes de manière à élargir légèrement cet intervalle, on obtient l'intervalle $[0,529; 0,611]$. Donc, au niveau de confiance 95%, on peut affirmer que le candidat C_1 aura entre $52,9\%$ et 61% des suffrages lors de l'élection.

2) Soit n $(n \in \mathbb{N}^*)$ la taille de l'échantillon et f la fréquence des personnes qui vote pour le candidat C_1.
L'intervalle de confiance au niveau de confiance 95% est

$$I = \left[f-\frac{1}{\sqrt{n}}, f+\frac{1}{\sqrt{n}}\right] = [0,529\ldots, 0,610\ldots].$$

L'amplitude de cet intervalle est $\frac{2}{\sqrt{n}}$. On veut donc que $\frac{2}{\sqrt{n}} \leqslant 0,02$. Or

$$\frac{2}{\sqrt{n}} \leqslant 0,02 \Leftrightarrow \frac{1}{\sqrt{n}} \leqslant 0,01 \Leftrightarrow \sqrt{n} \geqslant 100 \Leftrightarrow n \geqslant 10000.$$

Un échantillon de 10000 personnes donnera une fourchette d'amplitude 2% au niveau de confiance 95%.

Commentaire. Comme on l'a dit au début du chapitre, le cours est compliqué mais ce qu'il faut savoir faire est simple. ❑