Echantillonnage, prise de décision et estimation.

1) Le contexte :

Dans ce chapitre, on s'intéresse à un **caractère de proportion** p dans une **population** donnée.

Ce caractère doit être objectivement quantifiable pour être étudié mathématiquement.
Que penser de la couleur des cheveux par exemple ?

Cette proportion est parfois connue (**échantillonnage**), parfois supposée connue (**prise de décision**) et d'autres fois inconnue (**estimation**).

Pour des raisons pratiques, on étudie ce caractère non pas sur la population entière, mais sur des **échantillons de taille** n .

Pour obtenir ces échantillons, on peut **prélever au hasard des individus représentatifs** de cette population un par un **avec remise**. On parle d'**échantillons aléatoires non exhaustifs**.

Dans des situations telles que les sondages, un tel prélèvement est impensable : on pourrait interroger la même personne plusieurs fois.

On effectue alors un **prélèvement sans remise** de n individus dans cette population.

Si la taille de cet échantillon est suffisamment faible devant la taille de la population étudiée, de l'ordre de 10% maximum, alors ce prélèvement ne modifie pas sensiblement la proportion du caractère dans la population au fur et à mesure des prélèvements successifs.

L'échantillon ainsi construit est assimilé à un échantillon aléatoire non exhaustif.

Tels sont les échantillons considérés dans ce chapitre.

2) La fluctuation des fréquences d'échantillonnage :

Considérons un caractère de proportion connue p dans une population.
On prélève un échantillon de taille n dans cette population.
On note X la variable aléatoire égale au nombre d'individus qui possèdent ce caractère dans cet échantillon. Nous savons que la variable X suit la loi binomiale $B(n;p)$.

La **variable aléatoire fréquence** associée à cet échantillon est la variable $f=\frac{X}{n}$.

Nous savons également que cette fréquence d'échantillonnage **fluctue** dans un intervalle centré en p, ce qui peut se vérifier expérimentalement ou en utilisant un algorithme.

La probabilité que cette fréquence appartienne à un intervalle donné I dépend alors de cet intervalle, qui dépendra lui-même de la proportion p, de la taille de l'échantillon n, et de la probabilité minimale attendue, ce que nous allons voir maintenant sur un exemple.

Exemple :

On lance 120 fois un dé bien équilibré.
On appelle N la variable aléatoire égale au nombre de 6 obtenus.
On souhaite déterminer la probabilité que $N\in[12;28]$ en utilisant un échantillon de taille 100 de cette expérience aléatoire.

Pour cela, le programme suivant effectue 100 fois ces 120 lancers.
Puis il affiche le nombre de fois où la variable N est dans l'intervalle $[12\,;\,28]$.

```
# intervalle de fluctuation
from random import randint
total = 0
for i in range(100):
    succes = 0
    for j in range(120):
        de = randint(1,6)
        if de == 6:
            succes = succes + 1
    if succes >= 12 and succes <= 28:
        total = total + 1
print("nombre d'échantillons qui conviennent", total)
```

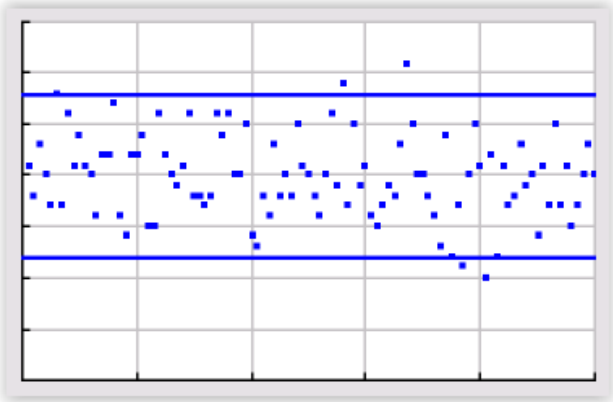
On obtient 96 % de réussite sur cet essai :

```
*** Console de processus distant Réinitialisée ***
nombre d'échantillons qui conviennent 96
>>>
```

L'intervalle $[12\,;\,28]$ est alors nommé :

« intervalle de fluctuation de N au seuil de 96% ».

Voici la répartition des 100 résultats entre les bornes 12 et 28 :



3) Intervalle de fluctuation :

Dans ce paragraphe, la proportion p du caractère étudié est connue.

Définition 1 : Soit X_n une variable aléatoire qui suit la loi binomiale $B(n,p)$.
Soit a et b deux réels de $[0;n]$, et un réel $\alpha \in]0;1[$.

On dit que $[a;b]$ est un **intervalle de fluctuation** de X_n **au seuil** de $1-\alpha$ si et seulement si : $P(X \in [a;b]) = P(a \leqslant X_n \leqslant b) \geqslant 1-\alpha$.

Exemple : Considérons la variable aléatoire X suivant la loi binomiale $B(110;0,6)$.
On a alors $P(46 \leqslant X \leqslant 71) = P(X \leqslant 71) - P(X \leqslant 45)$.
En utilisant la calculatrice, on obtient $P(46 \leqslant X \leqslant 71) = 0,858\ldots$
Donc l'intervalle $[46;71]$ est un intervalle de fluctuation de X au seuil de 85 %.
Cet intervalle correspond à la valeur $\alpha = 0,15$.

Définition 2 : Considérons trois réels a, b et α de l'intervalle $]0;1[$.
Soit $F_n = \dfrac{X_n}{n}$ la variable aléatoire fréquence associée à la variable X_n.

L'intervalle $[a;b]$ est un **intervalle de fluctuation asymptotique** de F_n au seuil de $1-\alpha$ si l'on a $\lim\limits_{n \to +\infty} P(F_n \in I) = 1-\alpha$.

Remarques :

- Il existe plusieurs intervalles de fluctuation asymptotiques à un seuil donné.
- La probabilité que $F_n \in I$ n'est pas égale à $1-\alpha$, mais elle s'en rapproche lorsque n augmente, c'est le sens du mot asymptotique.
- La variable aléatoire fréquence F_n ne suit pas une loi binomiale, car elle ne prend pas que des valeurs entières.
- Comme vu en classe de première, on peut toujours déterminer un intervalle de fluctuation, qui dépend implicitement des paramètres n et p.

Exercice n°1 : Soit X suivant la loi $B(n=110; p=0,6)$ et $F=\frac{X}{n}$ la variable aléatoire fréquence correspondante.
Déterminer l'intervalle I de fluctuation de F au seuil de 95 %.

Cet intervalle est de la forme $I=\left[\frac{a}{n};\frac{b}{n}\right]$ où a et b sont définis par :

- a est le plus petit entier k tel que $P(X \leqslant k) \geqslant 0,025$
- b est le plus petit entier k tel que $P(X \leqslant k) > 0,975$

En utilisant un tableur on obtient $a=56$ et $b=76$, donc $I=\left[\frac{56}{110};\frac{76}{110}\right] \approx [0,509; 0,690]$.

Conclusion : il y a 95 % de chance pour que $F \in I$.

Propriété 1 : Intervalle de fluctuation asymptotique.

Si la variable aléatoire X_n suit la loi binomiale $B(n,p)$ alors :
pour tout réel $\alpha \in]0\,;1[$, il existe un unique réel u_α tel que
l'intervalle $I_n=\left[p-u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}}\,;\,p+u_\alpha\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right]$
soit **un intervalle de fluctuation asymptotique** de $\frac{X_n}{n}$ **au seuil de** $1-\alpha$,
c'est-à-dire vérifiant : $\lim\limits_{n\to+\infty} P\left(\frac{X_n}{n} \in I_n\right)=1-\alpha$.

On appelle **variable fréquence**, la variable aléatoire $F_n=\frac{X_n}{n}$ qui à tout échantillon de taille n associe la fréquence f obtenue.

preuve :

Dans les conditions d'approximation d'une loi binomiale par une loi normale, c'est-à-dire lorsque $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$, le théorème de Moivre-Laplace nous assure que la variable aléatoire $Z_n = \dfrac{X_n - np}{\sqrt{np(1-p)}}$ se rapproche d'une loi $N(0;1)$ lorsque n augmente.

D'après les propriétés de la loi normale centrée réduite, on sait que pour tout $\alpha \in]0;1[$, il existe un unique réel u_α tel que $P(-u_\alpha \leqslant Z_n \leqslant u_\alpha) = 1-\alpha$.

Or $-u_\alpha \leqslant Z_n \leqslant u_\alpha$ équivaut à $-u_\alpha \leqslant \dfrac{X_n - np}{\sqrt{np(1-p)}} \leqslant u_\alpha$

équivaut à $-u_\alpha \sqrt{np(1-p)} \leqslant X_n - np \leqslant u_\alpha \sqrt{np(1-p)}$

équivaut à $np - u_\alpha \sqrt{np(1-p)} \leqslant X_n \leqslant np + u_\alpha \sqrt{np(1-p)}$

équivaut à $p - u_\alpha \dfrac{\sqrt{np(1-p)}}{n} \leqslant \dfrac{X_n}{n} \leqslant p + u_\alpha \dfrac{\sqrt{np(1-p)}}{n}$

équivaut à $p - u_\alpha \dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \leqslant \dfrac{X_n}{n} \leqslant p + u_\alpha \dfrac{\sqrt{p(1-p)}}{\sqrt{n}}$ car $\dfrac{\sqrt{n}}{n} = \dfrac{1}{\sqrt{n}}$

équivaut à $\dfrac{X_n}{n} \in I_n$

On obtient donc que $\lim\limits_{n \to +\infty} P\left(\dfrac{X_n}{n} \in I_n\right) = 1 - \alpha$.

L'intervalle de fluctuation asymptotique à connaître qui est utilisé en classe de terminale correspond à $\alpha = 0{,}05$ et donc $1-\alpha = 0{,}95$.
On sait que $u_{0{,}05} = 1{,}96$, ce qui permet d'écrire la propriété suivante :

Propriété 2 : Intervalle de fluctuation asymptotique au seuil de 0,95 .

Soit X_n une variable aléatoire suivant la loi $B(n;p)$

et $F_n = \dfrac{X_n}{n}$ la variable aléatoire fréquence associée.

Dans les conditions d'approximation $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$,
l'intervalle $I_n = \left[p - 1{,}96 \dfrac{\sqrt{p(1-p)}}{\sqrt{n}} ; p + 1{,}96 \dfrac{\sqrt{p(1-p)}}{\sqrt{n}} \right]$ peut être considéré comme un intervalle de fluctuation asymptotique de F_n au seuil de 95% .

On a donc $\lim\limits_{n \to +\infty} P(F_n \in I_n) = 0{,}95$.

Exemple : Reprenons l'exemple des 120 lancers de dé à jouer avec N comme variable aléatoire.

On a $p-1{,}96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\approx 0{,}100$ et $p+1{,}96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\approx 0{,}233$ avec $n=120$ et $p=\dfrac{1}{6}$
donc $I_n=[0{,}100\,;0{,}233]$ est l'intervalle de fluctuation asymptotique au seuil de 95% de la variable aléatoire fréquence $N/120$.

En multipliant par 120, on revient à la variable N et l'intervalle est alors $I_n=[12;28]$, ce qui confirme le résultat affiché par l'algorithme, 96%.

Remarque : En classe de seconde, l'intervalle de fluctuation au seuil de 0,95 utilisé est l'intervalle $J_n=\left[p-\dfrac{1}{\sqrt{n}};p+\dfrac{1}{\sqrt{n}}\right]$.
Montrons que cet intervalle contient l'intervalle I_n précédent.

En effet, la fonction $f(x)=x(1-x)=x-x^2$ est une fonction du second degré qui s'annule en 0 et 1, qui admet donc un maximum (coefficient de x^2 négatif) en 0,5, avec $f(0{,}5)=0{,}25$.
Elle est positive entre 0 et 1, donc on peut écrire :

$0\leqslant p(1-p)\leqslant 0{,}25$ qui donne $0\leqslant\sqrt{p(1-p)}\leqslant\sqrt{0{,}25}=0{,}5$.
On en déduit que $0\leqslant 1{,}96\sqrt{p(1-p)}\leqslant 1$ et donc $0\leqslant 1{,}96\dfrac{\sqrt{p(1-p)}}{\sqrt{n}}\leqslant\dfrac{1}{\sqrt{n}}$.
On a donc bien $I_n\subset J_n$, et donc dans la plupart des cas : $P(F_n\in J_n)\geqslant 0{,}95$.

Exercice n°2 :

Déterminer l'intervalle de fluctuation asymptotique au seuil de 0,95 pour $n=40$ et $p=0{,}5$.

4) Prise de décision :

Dans ce paragraphe, la proportion p du caractère étudié prend une valeur supposée.

Propriété 3 : Prise de décision.

Soit f_{obs} la fréquence observée d'un caractère sur un échantillon de taille n issu d'une population donnée. On suppose que les conditions de l'approximation normale de la loi binomiale sont remplies : $n \geqslant 30$, $np \geqslant 5$ et $n(1-p) \geqslant 5$.

On fait une hypothèse sur la valeur de la proportion p du caractère étudié dans la population toute entière.

On détermine ensuite l'intervalle de fluctuation asymptotique correspondant I_n .

- Si $f_{obs} \in I_n$, on ne peut pas rejeter l'hypothèse faite sur p .
- Si $f_{obs} \notin I_n$, on rejète l'hypothèse faite sur p .

Exercice n°3 :

Pour créer ses propres colliers, on peut acheter un kit contenant des perles de cinq couleurs différentes (marron, jaunes, rouges, vertes et bleues), dans des proportions affichées sur le paquet. Ainsi les perles marron et les perles jaunes sont annoncées comme représentant chacune 20% de l'ensemble des perles tandis que les perles rouges sont annoncées à 10%.
Sur un échantillon aléatoire de 690 perles, on a dénombré 140 perles marron.

1) Déterminer l'intervalle de fluctuation asymptotique au seuil de 95% pour la proportion de perles marron.
2) Calculer la proportion de perles marron dans l'échantillon. Que peut-on en conclure ?
3) Dans le même échantillon, il y avait 152 perles jaunes et 125 perles rouges. Que peut-on conclure de ces résultats ?

5) Estimation :

Dans ce paragraphe, la proportion p du caractère est inconnue.

On cherche alors à estimer p à partir d'un échantillon de taille n. On calcule alors la fréquence f_{obs} des individus de cet échantillon ayant ce caractère.

On estime ensuite la proportion p en utilisant un **intervalle de confiance** que l'on détermine à partir de la fréquence f_{obs} et de la taille n de l'échantillon.

Remarque :

La fréquence f_{obs} calculée varie d'un échantillon à l'autre du fait de la fluctuation d'échantillonnage. Il est donc nécessaire d'apprécier l'incertitude en donnant une estimation par un intervalle.

On suppose les trois conditions d'approximation remplies :

$$n \geqslant 30 \qquad n f_{obs} \geqslant 5 \qquad n\left(1 - f_{obs}\right) \geqslant 5$$

Propriété 4 :

Soit F_n la variable aléatoire qui à chacun des échantillons de taille n associe la fréquence du caractère dans cet échantillon.

La proportion inconnue p est telle que : $P\left(F_n - \frac{1}{\sqrt{n}} \leqslant p \leqslant F_n + \frac{1}{\sqrt{n}}\right) \geqslant 0,95$.

preuve : On a vu que l'intervalle de fluctuation au seuil de 95% pouvait être simplifié par :

$$J_n = \left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}}\right].$$

On a donc :

$$p - \frac{1}{\sqrt{n}} \leqslant F_n \leqslant p + \frac{1}{\sqrt{n}}$$

$$-\frac{1}{\sqrt{n}} \leqslant F_n - p \leqslant \frac{1}{\sqrt{n}}$$

$$-F_n - \frac{1}{\sqrt{n}} \leqslant -p \leqslant -F_n + \frac{1}{\sqrt{n}}$$

$F_n - \frac{1}{\sqrt{n}} \leqslant p \leqslant F_n + \frac{1}{\sqrt{n}}$ et ainsi $P\left(F_n - \frac{1}{\sqrt{n}} \leqslant p \leqslant F_n + \frac{1}{\sqrt{n}}\right) \geqslant 0,95$.

Définition 3 : On observe la fréquence f_{obs} sur un échantillon de taille n, et p désigne la proportion inconnue d'apparition du caractère dans la population entière.

On appelle **intervalle de confiance de p au niveau asymptotique de 95 %** l'intervalle :

$$I_c = \left[f_{obs} - \frac{1}{\sqrt{n}} ; f_{obs} + \frac{1}{\sqrt{n}} \right]$$

Cet intervalle de confiance a pour **amplitude** $\frac{2}{\sqrt{n}}$.

Ainsi, si l'on souhaite encadrer p dans un intervalle de longueur a, on doit avoir :

$$\frac{2}{\sqrt{n}} \leqslant a \text{ et donc } n \geqslant \frac{4}{a^2}.$$

Exemple :

Pour obtenir une précision sur p à 10^{-1} près, il faut choisir $n \geqslant \frac{4}{\left(10^{-1}\right)^2} = 400$,

donc un échantillon de taille 400.

Remarque :

On pourrait utiliser l'intervalle de confiance asymptotique suivant, mais il est impossible de le justifier en terminale :

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

Exercice n°4 : Un sondage pour l'élection présidentielle du 21 avril 2002.

Voici les résultats d'un sondage IPSOS réalisé avant l'élection présidentielle de 2002 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 18 ans et plus et inscrite sur les listes électorales.
On suppose cet échantillon constitué de manière aléatoire (même si en pratique cela n'est pas le cas). Les intentions de vote au premier tour pour les principaux candidats sont les suivantes :

Jacques Chirac : 20 % Lionel Jospin : 18 % Jean-Marie Le Pen : 14 %.

Les médias se préparent pour un second tour entre Jacques Chirac et Lionel Jospin.

a) Déterminer pour chaque candidat, l'intervalle de confiance au niveau de confiance de 0,95 de la proportion inconnue d'électeurs ayant l'intention de voter pour lui.

b) Le 21 avril, les résultats du premier tour des élections sont les suivantes : Jacques Chirac : 19,88 %, Lionel Jospin : 16,18 %, Jean-Marie Le Pen : 16,86 %. Les pourcentages de voix recueillies par chaque candidat sont-ils bien dans les intervalles de confiance précédents ?

c) Pouvait-on, au vu de ce sondage, écarter avec un niveau de confiance de 0,95, l'un de ces trois candidats second tour ?