Il est parfois impossible ou trop coûteux de recueillir des données sur l'ensemble d'une population. On étudie alors un **échantillon** de cette population à l'aide d'un sondage.

La fréquence d'un caractère sur un échantillon est une estimation de sa fréquence sur la population.

## 1. Modélisation de la situation

**Définitions**

- Lorsqu'on étudie une partie de la population, on dit qu'on étudie un **échantillon**.
- Le nombre d'individus formant l'échantillon est appelé **taille de l'échantillon**.

Un critère peut être : « voter 'oui' lors d'un référendum », ou bien « avoir plus de 50 ans »...

▶ **Notation** : On note $p$ la proportion de la population vérifiant le critère étudié et $\hat{p}$ la proportion de l'échantillon vérifiant ce critère.

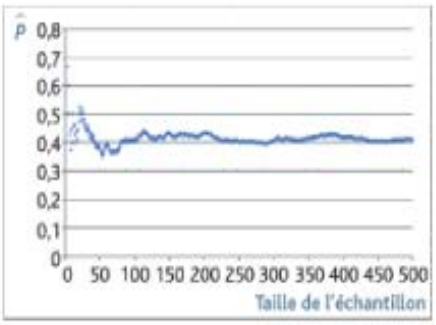
**Théorème de stabilisation des fréquences**

Plus la taille de l'échantillon est grande, plus $\hat{p}$ se rapproche de $p$.

La proportion de l'échantillon vérifiant un critère est $\hat{p} = \frac{n}{N}$, avec $n$ le nombre de fois où le critère est réalisé et $N$ la taille de l'échantillon.

**Exemple**

Pour un sondage, on sait que la proportion de personnes ayant répondu « oui » était de 0,4. On connaît donc $p$ : $p = 0,4$. Voici une représentation des valeurs de $\hat{p}$ en fonction de la taille de l'échantillon.




On constate que plus la taille de l'échantillon est grande, plus $\hat{p}$ se rapproche de $p$.

## 2. Intervalle de fluctuation, intervalle de confiance

**Théorème de l'intervalle de fluctuation**

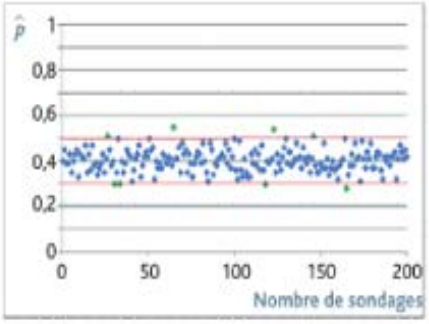
**Conditions d'application :**
- La taille $n$ de l'échantillon doit être supérieure ou égale à 25.
- $p$ doit appartenir à l'intervalle $[0,2 ; 0,8]$.

Dans ces conditions, dans plus de 95 % des cas, $\hat{p}$ appartient à l'intervalle $\left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}}\right]$.

On parle aussi d'intervalle de fluctuation au seuil 95 % ou 0,95.

**Exemple**

Pour le sondage précédent ($p = 0,4$), on a effectué 200 fois un sondage sur 100 personnes (donc $n = 100$). On a représenté ici les 200 valeurs de $\hat{p}$ obtenues. $p - \frac{1}{\sqrt{100}} = 0,3$ et $p + \frac{1}{\sqrt{100}} = 0,5$.




On voit que 8 valeurs (sur 200) sont à l'extérieur de l'intervalle $[0,3 ; 0,5]$, soit 4 %.

Si on connaît $p$ (et que les conditions sont respectées), on utilise le théorème de l'intervalle de fluctuation.
Si on ne connaît pas $p$, on utilise le théorème de l'intervalle de confiance.

➡ **Remarque** : $\hat{p} \in \left[p - \frac{1}{\sqrt{n}} ; p + \frac{1}{\sqrt{n}}\right]$ si, et seulement si, $p \in \left[\hat{p} - \frac{1}{\sqrt{n}} ; \hat{p} + \frac{1}{\sqrt{n}}\right]$. D'où :

**Théorème de l'intervalle de confiance**

**Conditions d'application :**
- La taille $n$ de l'échantillon doit être supérieure ou égale à 25.
- $\hat{p}$ doit appartenir à l'intervalle $[0,2 ; 0,8]$.

Dans ce cas, dans plus de 95 % des cas, $p$ appartient à l'intervalle $\left[\hat{p} - \frac{1}{\sqrt{n}} ; \hat{p} + \frac{1}{\sqrt{n}}\right]$.

On parle aussi d'intervalle de confiance au niveau 95 % ou 0,95.

*exemples corrigés*

**On souhaite savoir si une entreprise exerce une discrimination à l'embauche vis-à-vis du personnel féminin.**
**S'il n'y a pas de discrimination, la proportion de femmes dans cette entreprise devrait être représentative de la proportion de femmes dans la population active. On admet que la proportion de femmes dans la population active est 0,5.**

**1. En utilisant l'intervalle de fluctuation au seuil 0,95, déterminer si une entreprise contenant 1 183 femmes sur 2 540 salariés exerce une discrimination à l'égard des femmes.**
**2. Quel doit être le nombre minimal de femmes dans cette entreprise pour que la proportion $\hat{p}$ de femmes appartienne à l'intervalle de fluctuation $[0,48 ; 0,52]$ ?**

**Solution**

**1.** La taille de l'échantillon est $n = 2540$.
Dans cet exercice, on connaît la proportion théorique de femmes : $p = 0,5$.
Les conditions étant respectées, on peut donc utiliser le théorème.
La proportion $\hat{p}$ de femmes dans un échantillon de taille 2 540 appartient dans plus de 95 % des cas à l'intervalle $\left[0,5 - \frac{1}{\sqrt{2540}} ; 0,5 + \frac{1}{\sqrt{2540}}\right]$, soit l'intervalle $[0,48 ; 0,52]$.
Ici, $\hat{p} = \frac{1183}{2540} = 0,466$. Donc $\hat{p} \notin [0,48 ; 0,52]$.
On en conclut que cette entreprise exerce très probablement une discrimination à l'égard des femmes.

**On vérifie que les conditions suivantes sont respectées avant d'utiliser le théorème :**
**• La taille $n$ de l'échantillon doit être supérieure ou égale à 25.**
**• $p$ doit appartenir à l'intervalle $[0,2 ; 0,8]$.**

**2.** Pour que $\hat{p}$ appartiennent à l'intervalle de fluctuation, on doit avoir : $\hat{p} \geqslant 0,48$. On doit donc avoir un nombre de femmes supérieur à $0,48 \times 2540$.
Il faudrait donc au moins 1 220 femmes dans cette entreprise pour que $\hat{p}$ soit dans l'intervalle de confiance.

**Lors du second tour des élections présidentielles, un candidat souhaite connaître les intentions de vote des français en sa faveur.**
**Un premier sondage sur 250 personnes interrogées donne une intention de vote de 54 %.**
**Un second sondage sur 1 900 personnes interrogées donne une intention de vote de 53 %.**
**Quelle est le sondage qui est le plus favorable au candidat ?**

**Solution**

On ne connaît pas la proportion $p$ de Français qui vont voter pour le candidat ayant commandé le sondage.

**Le premier sondage** donne une proportion $\hat{p} = 0,54$ sur un échantillon de taille $n = 250$.
On peut donc déterminer l'intervalle de confiance : dans 95 % des cas, $p \in \left[0,54 - \frac{1}{\sqrt{250}} ; 0,54 + \frac{1}{\sqrt{250}}\right]$, soit :
$p \in [0,477 ; 0,604]$.
Remarque : pour être élu, il faut plus de 50 % des voix ; cet intervalle de confiance ne permet donc pas d'affirmer que $p \geqslant 0,5$.

**Le second sondage** donne une proportion $\hat{p} = 0,53$ sur un échantillon de taille $n = 1900$.
On peut donc déterminer l'intervalle de confiance : dans 95 % des cas, $p \in \left[0,53 - \frac{1}{\sqrt{1900}} ; 0,53 + \frac{1}{\sqrt{1900}}\right]$, soit :
$p \in [0,507 ; 0,553]$.
Cet intervalle de confiance permet d'affirmer que, dans plus de 95 % des cas, on devrait avoir $p \geqslant 0,5$ (ce qui signifie que le candidat serait élu).

C'est donc le second sondage qui est le plus favorable au candidat.

**Une société fabrique des écrans plasma.**
**En moyenne, 21 % des écrans sont défectueux.**
**Lors d'un contrôle d'un lot de 40 écrans, 14 sont défectueux.**
**1. Calculer la proportion d'écrans défectueux sur ce lot.**
**2. Doit-on s'inquiéter ?**
**3. Afin de tester une nouvelle machine, on contrôle 400 écrans au hasard. 28 % d'entre eux sont défectueux. Ce résultat semble satisfaisant. Est-ce correct ?**

## Solution

1. La proportion d'écrans défectueux sur ce lot est : $\hat{p} = \frac{14}{40} = 0{,}35$.

2. Il y a 21 % d'écrans défectueux en moyenne, donc on sait que $p = 0{,}21$.
D'après le théorème de l'intervalle de fluctuation (les conditions d'application sont respectées), dans 95 % des cas, la proportion $\hat{p}$ d'écrans défectueux sur un lot de 40 écrans est située dans l'intervalle $\left[0{,}21 - \frac{1}{\sqrt{40}} ; 0{,}21 + \frac{1}{\sqrt{40}}\right]$, soit $[0{,}051 ; 0{,}369]$.
La proportion 0,35 d'écrans défectueux trouvée appartient à cet intervalle : il n'y a pas lieu d'être inquiet.

3. Dans 95 % des cas, sur un lot de 400 écrans, la proportion d'écrans défectueux est située dans l'intervalle de fluctuation $\left[0{,}21 - \frac{1}{\sqrt{400}} ; 0{,}21 + \frac{1}{\sqrt{400}}\right]$, soit $[0{,}16 ; 0{,}26]$.
La proportion 0,28 d'écrans défectueux trouvée n'appartient pas à cet intervalle.
Contrairement à ce que l'on pourrait croire, ce pourcentage est anormalement élevé.

**Intervalles de confiance – Applications**

Voici les résultats d'un sondage IPSOS réalisé avant l'élection présidentielle de 2002 pour Le Figaro et Europe1, les 17 et 18 avril 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 que cet échantillon est constitué de manière aléatoire (même si en pratique cela n'est pas le cas).
Les intentions de vote au 1er tour pour les principaux candidats sont les suivantes :
20% pour J. Chirac, 18% pour L.Jospin et 14% pour J.-M. Le Pen.
Les médias se préparent donc pour un second tour entre J.Chirac et L.Jospin.

1. Déterminer pour chaque candidat, l'intervalle de confiance au seuil de 95% de la proportion d'électeurs ayant l'intention de voter pour lui.
2. Le 21 avril, les résultats du 1er tour sont les suivants : 19,88% pour J.Chirac, 16,18% pour L.Jospin et 16,86% pour J.-M. Le Pen.
   a. Les pourcentages des voix recueillies par chaque candidat sont-ils dans les intervalles de confiance précédents ?
   b. Pouvait-on, au vu de ce sondage, écarter avec un niveau de confiance de 95%, l'un des 3 candidats pour le second tour ?

**Correction : 1)** La condition n>25 est vérifiée. Par contre les fréquences f=0,18 et f=0,14 ne sont pas dans l'intervalle [0,2;0,8] condition attendue en seconde.
Dans ce cas, l'intervalle de confiance reste valable si $nf>5$ et $n(1-f)>5$ ce qui est vérifié ici.

Pour Chirac : $n=989$ , $f=0,2$
$p_{C\,hirac} \in \left[0,2-\frac{1}{\sqrt{989}};0,2+\frac{1}{\sqrt{989}}\right]$ au seuil de confiance de 0,95 soit [0,168;0,232]

Pour Jospin : $n=989$ , $f=0,18$

$p_{Jospin} \in \left[0,18-\frac{1}{\sqrt{989}};0,18+\frac{1}{\sqrt{989}}\right]$ au seuil de confiance de 0,95 soit [0,148;0,212]

Pour Le Pen: $n=989$ , $f=0,14$

$p_{LePen} \in \left[0,14-\frac{1}{\sqrt{989}};0,14+\frac{1}{\sqrt{989}}\right]$ au seuil de confiance de 0,95 soit [0,108;0,172]

2) a) oui car 0,1988∈[0,168;0,232], 0,1618∈[0,148;0,212] et 0,1686∈[0,108;0,172].

b) Non à cause du chevauchement des 3 intervalles de confiance.

# Échantillonnage

## I Fluctuation d'échantillonnage

**Définition 1**
*Un* ***échantillon*** *de taille $n$ est obtenu à partir d'une population en réitérant $n$ fois de suite l'opération suivante : on prélève au hasard un de ses éléments, on note la valeur du caractère prélevé et on remet l'élément prélevé dans la population.*

**Exemple 1**
On lance un dé numéroté de 1 à 6, bien équilibré, et on repère le chiffre qui apparaît sur la face supérieure
on répète ce lancer 100 fois pour obtenir un échantillon $A$ de taille 100 et encore 100 fois pour obtenir un échantillon $B$ de taille 100.
On a noté les fréquences d'apparition de chaque chiffre dans un tableau de distribution des fréquences :

| Chiffre | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Fréquence $A$ | 0,14 | 0,17 | 0,19 | 0,18 | 0,17 | 0,15 |
| Fréquence $B$ | 0,15 | 0,16 | 0,16 | 0,18 | 0,17 | 0,18 |

Dans l'exemple précédent, on constate que les distributions des fréquences des deux échantillons ne sont pas les mêmes : c'est ce qu'on appelle la **fluctuation d'échantillonnage**.
La moyenne de l'échantillon $A$ est de ..... et celle de $B$ est .....

## II Intervalle de fluctuation

On se place dans le cas où on réalise une expérience aléatoire dont le résultat est soit un succès (noté 1) avec la probabilité $p$, soit un échec (noté 0) avec la probabilité $1-p$ *(appelée expérience de Bernoulli).*

**Propriété 1**
Pour des probabilités $p$ comprises entre $0,2$ et $0,8$ et $n \geq 25$, on peut assurer que pour environ 95% des échantillons de taille $n$, la fréquence d'apparition du 1 appartient à l'intervalle $\left[ p - \frac{1}{\sqrt{n}} \ ; \ p + \frac{1}{\sqrt{n}} \right]$.
Cet intervalle s'appelle l'**intervalle de fluctuation au seuil de** 95%**.**

Cette propriété signifie que dans 95% des cas, la fréquence d'apparition d'un succès est située dans l'intervalle centré en $p$ d'amplitude $\frac{2}{\sqrt{n}}$.

**Exemple 2**
Un joueur tire une carte dans un jeu de cartes, puis il la remet dedans. Il gagne si l'on obtient un « coeur ». Il renouvelle cette expérience $n$ fois. La probabilité de gagner est donc de $p =$ ..... à chaque fois.

- Si $n = 100$, dans 95% des cas, la fréquence d'apparition d'un coeur fluctue dans l'intervalle [..... ; .....],
- Si $n = 10\,000$, dans 95% des cas, la fréquence d'apparition d'un coeur fluctue dans l'intervalle [..... ; .....].