

# Échantillonnage

## I Problématiques

On va donner deux exemples de problématiques pour nous guider dans cette séquence :

### Exemple 1 : *Un dé truqué?*

Nous savons que sur un dé à quatre faces équilibré, la probabilité d'obtenir un « 4 » est  $\frac{1}{4}$  soit 25%.

Tom lance plusieurs fois un dé à quatre faces. Il obtient 50% de « 4 ».

Une question se pose : « le dé est-il truqué ou est-ce un tirage exceptionnel? »

Cela dépend de plusieurs paramètres et en particulier, du **nombre de lancers** :

En effet, si Tom n'a lancé que 2 fois le dé et obtenu un « 4 », on ne peut pas conclure, le nombre d'essais étant trop faible.

En revanche, si Tom a lancé 1000 fois le dé et obtenu 500 « 4 », on peut se dire que soit le dé est truqué, soit il s'agit d'une série de tirages vraiment exceptionnelle. Mais quelle certitude a-t-on?

### Exemple 2 : *Qui va gagner?*

Avant une élection, il est impossible de connaître avec certitude le nombre de personnes qui vont voter pour un candidat.

Afin de fournir des estimations, on réalise des sondages.

Imaginons : En France, au second tour de la présidentielle, il reste deux candidats, M.CHEMISE et Mme.ROBE.

Une enquête réalisée sur un groupe d'individus montre que 54% des personnes de cet échantillon voteront pour Mme.ROBE.

Plusieurs questions se posent : « Quel pourrait être le résultat de l'élection? », « Quelle certitude a-t-on? »

Encore une fois, cela dépend de plusieurs paramètres dont le **nombre de personnes interrogées** :

En effet, il sera plus difficile d'estimer le résultat de l'élection si seulement 50 personnes ont été interrogées.

## II Échantillon, simulation et fluctuation

### Définition : *Échantillon*

Un **échantillon** de taille  $n$  est constitué des résultats de  $n$  répétitions indépendantes de la même expérience.

### Exemples 3 : *Exemples d'échantillons*

- ☞ on lance une pièce 50 fois et on regarde si on obtient pile;
- ☞ on tire 20 fois une carte d'un jeu de 32 cartes en la remettant et on regarde si c'est un cœur;
- ☞ on interroge 1 000 personnes et on leur demande si elles voteront.

### Définition : *Fluctuation d'échantillonnage*

Deux échantillons de même taille issus de la même expérience aléatoire ne sont généralement pas identiques.

On appelle **fluctuation d'échantillonnage** les variations des fréquences des valeurs relevées.

## Remarques :

- ☞  $n$  est le nombre d'éléments de l'échantillon. C'est l'**effectif** ou la **taille de l'échantillon**.  
On dit que l'échantillon est de taille  $n$ .
- ☞  $f_o$  est la **fréquence** du caractère observé dans l'échantillon.
- ☞  $p$  est la **proportion effective** du caractère observé dans la population.
- ☞ Plus la taille de l'échantillon augmente, plus les fréquences observées se rapprochent de  $p$ .

## III Prise de décision : intervalle de fluctuation ( $p$ est connu)

### Protocole :

Soit une population pour laquelle on étudie la proportion d'un caractère.

On émet une hypothèse sur la proportion  $p$  du caractère étudié dans la population. On considère donc  $p$  comme connu car il a une valeur conjecturée.

Un échantillon de taille  $n$  de cette population est prélevé et on observe une fréquence  $f_o$  du caractère étudié.

### La question :

Peut-on, à partir de l'observation de  $f_o$ , valider la conjecture faite sur  $p$ ?

La fréquence observée,  $f_o$ , est-elle proche ou éloignée de la probabilité ou proportion théorique,  $p$  ?

### Définition : **Intervalle de fluctuation**

L'**intervalle de fluctuation au seuil de 95%**, relatif aux échantillons de taille  $n$ , est l'intervalle centré autour de  $p$  qui contient la fréquence observée  $f_o$  dans un échantillon de taille  $n$  avec une probabilité égale à 0,95.

### Propriété :

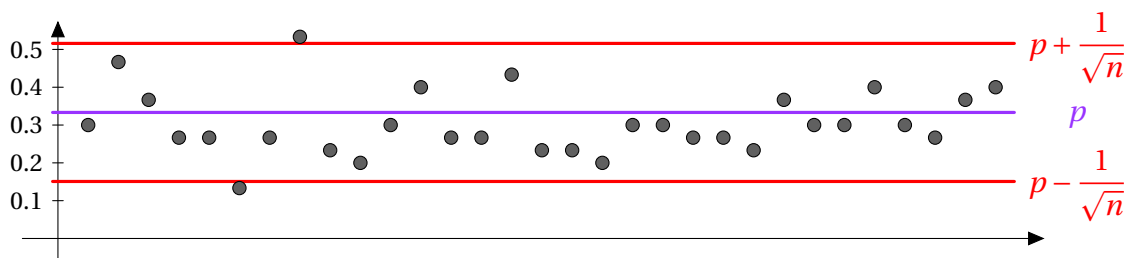
Soit  $p$  la proportion effective d'un caractère d'une population comprise entre 0,2 et 0,8 et  $f_o$  la fréquence du caractère dans un échantillon de taille  $n$  supérieure ou égale à 25.

$f_o$  appartient à l'intervalle  $\left[ p - \frac{1}{\sqrt{n}}, p + \frac{1}{\sqrt{n}} \right]$  avec une probabilité d'environ 0,95.

### Exemple 4 : **Lancers de dés**

Vous avez effectué chacun 30 lancers d'un dé à six faces en notant la fréquence d'apparition de 5 ou de 6. La probabilité  $p$  d'obtenir un 5 ou un 6 est de  $\frac{1}{3} \approx 0,33$ .

Voici les résultats en fréquences de vos 31 séries de 30 lancers :



On observe que 29 séries sur 31 sont dans l'intervalle :  $\left[ p - \frac{1}{\sqrt{30}}, p + \frac{1}{\sqrt{30}} \right]$ .

Or, 95% de 31 correspond à 29,45.

## Remarques :

La taille de l'intervalle de fluctuation  $\left(\frac{2}{\sqrt{n}}\right)$  diminue si  $n$  augmente.

Dans les conditions de la définition et de la propriété :

- ☞ On émet une hypothèse sur la proportion du caractère de la population  $p$ .
- ☞ On détermine l'intervalle de fluctuation au seuil de 95% de la proportion  $p$  dans des échantillons de taille  $n$ .
  - Si  $f_o$  n'appartient pas à cet intervalle, on rejette l'hypothèse sur  $p$  **avec un risque d'erreur de 5%**.
  - Si  $f_o$  appartient à cet intervalle, on ne rejette pas l'hypothèse faites sur  $p$ .

## Exemple 5 : Prendre une décision

Dans la réserve indienne d'Aamjiwnaag, située au Canada, à proximité d'industries chimiques, il est né entre 1999 et 2003, 132 enfants dont 46 garçons. Est-ce normal?

### Correction :

On fait ici l'hypothèse  $P$  suivante : « le sexe d'un enfant qui naît dans cette réserve est un garçon avec une probabilité de 0,5 ».

La taille de l'échantillon est  $n = 132$  ( $n \geq 25$ ) et la fréquence observée est  $f_o = \frac{46}{132} \approx 0,34$ .

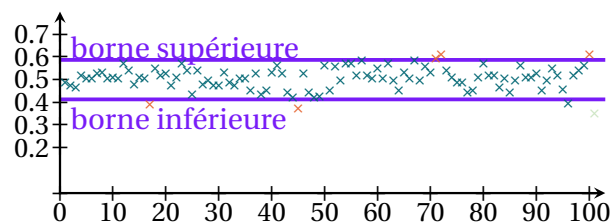
L'intervalle de fluctuation au seuil de 95% est :

$$IF = \left[ 0,5 - \frac{1}{\sqrt{132}}; 0,5 + \frac{1}{\sqrt{132}} \right] \approx [0,41; 0,58]$$

$f_o \notin IF$  et on rejette l'hypothèse  $P$ . La probabilité qu'un garçon naisse dans cette réserve n'est pas de 0,5.

Les 95% sont illustrés avec le graphique qui suit :

On simule 100 fois le comptage de garçons sur 132 naissances. Dans 94 simulations, la proportion des garçons nés se trouve dans l'intervalle de fluctuation.



Tom a lancé cent fois un dé à quatre faces et il a obtenu 50 « 6 ».

Que peut-on conclure?

## IV Estimation : Intervalle de confiance ( $p$ est inconnu)

L'intervalle de fluctuation permet d'avoir un intervalle où se situe la proportion inconnue  $p$  avec une probabilité de 0,95%.

### Propriété :

On considère un échantillon de taille  $n$  ( $n \geq 25$ ) tel que  $f_o \in [0,2; 0,8]$ .

Alors  $p$  appartient à l'intervalle  $\left[ f_o - \frac{1}{\sqrt{n}}; f_o + \frac{1}{\sqrt{n}} \right]$  avec une probabilité de 0,95.

## Définition : Intervalle de confiance

Un **intervalle de confiance** au seuil de 95%, relatif aux échantillons de taille  $n$ , est un intervalle centré autour de  $f_0$  où se situe la proportion  $p$  du caractère dans la population avec une probabilité égale à 95%. L'intervalle  $\left[ f_0 - \frac{1}{\sqrt{n}}; f_0 + \frac{1}{\sqrt{n}} \right]$  est donc appelé intervalle de confiance au seuil de 95%.

### Remarque :

- ☞ On réalise un échantillon de taille  $n$  et on y obtient une fréquence observée  $f_0$ .
- ☞ On construit l'intervalle de confiance à partir de  $n$  et  $f_0$ .

La proportion réelle dans la population se situe dans cet intervalle **avec une probabilité d'environ 0,95**.

### Exemple 6 : Estimer la proportion d'un caractère

Le 4 mai 2007 soit deux jours avant le second tour des élections présidentielles, on publie le sondage suivant réalisé auprès de 992 personnes :

S. Royal : 45%
N. Sarkozy : 55%

Interpréter ce sondage.

### Correction :

On calcule l'intervalle de confiance pour N. Sarkozy :

$$I = \left[ f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}} \right] = \left[ 0,55 - \frac{1}{\sqrt{992}}; 0,55 + \frac{1}{\sqrt{992}} \right] \approx [0,518; 0,582]$$

La proportion des votants en faveur de N. Sarkozy se trouvant dans  $[0,518; 0,582]$  avec 95% de chance, on peut en déduire qu'il avait de grandes chances d'être élu.

### Remarque :

Les sondages sont souvent réalisés auprès d'environ 1000 personnes car cela permet de connaître la proportion d'un candidat à 3% près.



En France, au second tour de la présidentielle, il reste deux candidats, M.CHEMISE et Mme.ROBE. Une enquête réalisée sur un groupe de 900 individus montre que 54% des personnes de cet échantillon voteront pour Mme.ROBE.

Que peut-on conclure?