

Échantillonnage en Seconde

1. Echantillons

Lorsqu'on travaille sur une population de grande taille, il est rarement possible d'avoir accès aux données relatives à l'ensemble de la population.

On utilise alors un **échantillon** de cette population.

Définition

Un **échantillon de taille n** est une sélection de n individus choisis « au hasard » dans une population.

Exemple

On étudie la répartition mâle/femelle d'une population de truites peuplant une rivière.

Il est pratiquement impossible de recenser toutes les truites de la rivière. On décidera donc de travailler sur un échantillon en prélevant, par exemple, 100 truites.

La taille de l'échantillon doit être suffisamment élevée pour fournir des résultats fiables (mais pas trop pour ne pas entrainer un surcroit de travail important !)

Remarque

Il existe deux manières d'effectuer un échantillonnage:

- **sans remise** : Dans l'exemple précédent, si l'on prélève les 100 truites simultanément, on obtient 100 individus différents

- **avec remise** : On prélève une truite au hasard, on note son sexe puis on la remet dans la rivière. Et on répète cette expérience 100 fois. Dans ce cas, il est possible de prélever plusieurs fois le même individu.

En pratique, si l'effectif global est nettement supérieur à la taille de l'échantillon (c'est à dire, ici, si la rivière abrite beaucoup plus de 100 truites) les deux méthodes donneront des résultats également satisfaisants.

2. Intervalle de fluctuation

Si l'on effectue plusieurs échantillonnage de même taille sur une même population, on obtiendra en général des fréquences légèrement différentes pour un caractère donné.

Voici, par exemple, les résultats que l'on pourrait obtenir en prélevant 5 échantillons de 100 truites :

Echantillons	n°1	n°2	n°3	n°4	n°5
Pourcentage de truites femelles	52\%	55\%	42\%	50\%	48\%

Ce phénomène s'appelle **fluctuation d'échantillonnage**.

Le résultat suivant précise cette notion :

Théorème et définition

On note p la proportion d'un caractère dans une population donnée.

On prélève un échantillon de taille n de cette population et on note f la fréquence du caractère dans l'échantillon.

Si $0,2 \leq p \leq 0,8$ et si $n \geq 25$ alors, dans au moins 95% des cas, f appartient à l'intervalle :

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

I est appelé **l'intervalle de fluctuation au seuil 95%**.

Remarques

- On applique le théorème ci-dessus si **on connaît la proportion p du caractère dans la population**.

On peut aussi utiliser ce théorème en **supposant** que le caractère est présent dans une proportion p . Suivant la (ou les) fréquence(s) observée(s) dans un (ou plusieurs) échantillon(s) on acceptera ou on rejettera l'hypothèse.

- Bien retenir la signification de chacune des variables :
 - p = proportion du caractère dans **l'ensemble de la population**
 - f = fréquence du caractère dans **l'échantillon**
 - n = taille de l'échantillon
- Au niveau Seconde, les intervalles de fluctuation seront toujours demandés au seuil de 95%.

Ce seuil a été choisi car :

- il conduit à une formule assez simple
- on peut considérer comme « *raisonnablement fiable* » un résultat validé dans 95% des cas

Exemple

Supposons que notre rivière contienne 50% de truites femelles (et donc 50% de mâles...).

Pour nos échantillons de taille 100, $n = 100 \geq 25$; par ailleurs $p = 0,5 \in [0,2; 0,8]$

Donc **l'intervalle de fluctuation au seuil de 95%** sera $I =$

$$\left[0,5 - \frac{1}{\sqrt{100}} ; 0,5 + \frac{1}{\sqrt{100}} \right] \text{ c'est à dire } I = [0,4 ; 0,6].$$

