

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 entraîner un surcroît 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. Simulation d'une expérience aléatoire

Simuler une expérience, c'est la remplacer par une autre qui permet d'obtenir des résultats similaires. En général, on simule à l'aide d'une table, d'une calculatrice ou d'un ordinateur.

a. En utilisant une table (quasiment abandonné)

Tables de chiffres aléatoires entre 1 et 10. On conviendrait par exemple, avec les truites, que les chiffres pairs correspondent à « femelle » et les chiffres impairs à « mâle ». Puis on lirait la table chiffre par chiffre .

Exemple : échantillon de taille 8 :

.....

Table de nombres au hasard

0	8	7	4	6
4	9	3	8	9
9	6	5	5	9
8	1	1	1	4
9	6	7	5	6

1	1	5	1	3
6	1	3	7	2
9	6	8	7	6
0	8	6	0	8
2	0	6	4	0

0	8	6	4	4
9	4	7	1	1
5	4	9	0	4
2	4	1	7	0
1	4	6	1	7

4	6	4	2	8
2	2	8	9	6
0	3	0	8	5
9	3	0	7	3
3	1	1	4	3

5	1	5	0	4
3	3	7	3	7
8	0	5	5	0
4	9	6	8	0
2	4	3	2	9

7	0	7	3	7
1	8	7	3	5
7	4	9	4	9
1	0	2	9	0
1	3	7	8	8

9	9	5	4	6
4	5	6	9	8
9	3	7	8	3
9	4	2	2	7
8	7	6	5	5

0	9	7	8	4
4	9	1	4	0
1	6	6	0	4
0	5	9	3	6
2	7	3	2	5

2	2	5	6	2
8	7	7	5	8
1	1	1	7	5
2	2	6	8	1
8	7	9	6	7

3	1	6	0	9
4	0	8	0	0
8	1	2	6	9
0	8	0	7	2
1	1	3	9	2

7	3	3	2	1
1	0	0	6	4
5	9	3	3	9
6	0	4	5	8
9	7	8	1	1

1	6	1	9	1
7	3	1	6	3
1	0	8	5	9
2	5	6	8	1
9	2	5	9	8

1	4	7	8	6
8	5	1	7	7
9	5	7	2	2
9	3	8	0	7
9	0	7	5	2

4	4	1	4	5
2	6	0	9	0
9	1	5	8	7
4	9	1	5	7
1	7	7	2	3

9	0	0	8	2
5	2	0	9	6
2	3	8	3	1
0	2	4	9	0
0	0	7	6	8

1	1	1	8	9
8	4	3	1	0
7	0	7	7	2
4	7	7	2	4
3	8	3	0	0

5	2	1	5	8
7	7	0	7	9
4	0	9	0	3
8	6	9	2	0
1	7	8	4	4

0	9	1	4	3
7	0	7	6	2
6	9	3	8	6
9	4	3	5	3
7	6	0	5	8

b. Avec la fonction random de la calculatrice

	Casio	T.I.
Nombre au hasard (RANDOM)	• Menu OPTN • Sélectionner PROB • Choisir RAN # • Taper sur EXE	• Menu MATH • Sélectionner PRB • Choisir 1 : NbrAléat • Taper sur ENTER
Partie entière (Ent)	• Menu OPTN • Sélectionner NUM • Commande Int	• Menu MATH • Menu NUM • Commande ent (ou int)

Cette touche donne un nombre aléatoire réel compris entre 0 et 1 (inclus).

Cette touche fournit une liste de 10 chiffres après la virgule ex : 9846634426.

Si moins de 10 chiffres s'affichent, les chiffres suivants sont des 0.

c. Avec la commande ALEA() d'un tableur

La formule =ENT(ALEA() + 0,5) donne soit 0, soit 1.

Dans l'exemple, on prendrait par exemple 0 pour « mâle » et 1 pour « femelle ».

3. Intervalle de fluctuation

Si l'on effectue plusieurs échantillonnages 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	échantillon n°1	échantillon n°2	échantillon n°3	échantillon n°4	échantillon 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 :

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

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

Remarque : L'intervalle est centré autour de p .

Exemple : Prendre une décision à partir d'un échantillon

Deux entreprises A et B recrutent dans un bassin d'emploi où il y a autant de femmes que d'hommes, avec la contrainte du respect de la parité.

Dans l'entreprise A, il y a 100 employés dont 43 femmes (soit 43 %).

Dans l'entreprise B, il y a 2500 employés dont 1150 femmes (soit 46 %).

Or, 46 % est plus proche de 50 % que 43 % : les chiffres parlent d'eux-mêmes !

Si on admet que la parité, c'est exactement 50 % de femmes, il est vrai que B en est plus proche que A.

Question 1 : Peut-on alors affirmer que l'entreprise B respecte mieux la parité que l'entreprise A ?

La proportion effective p est égale à 0,5 (50% de femmes) et $n > 25$. Les conditions sont respectées. Pour l'entreprise A, l'intervalle de fluctuation au seuil de 95% de $p = 0,5$ est :

$$\left[0,5 - \frac{1}{\sqrt{100}}; 0,5 + \frac{1}{\sqrt{100}} \right] = [0,4; 0,6].$$

Pour l'entreprise B, l'intervalle de fluctuation au seuil de 95% de $p = 0,5$ est :

$$\left[0,5 - \frac{1}{\sqrt{2500}}; 0,5 + \frac{1}{\sqrt{2500}} \right] = [0,48; 0,52].$$

La valeur 43% est donc dans l'intervalle de fluctuation de l'entreprise A alors que la valeur 46% n'est pas dans l'intervalle de fluctuation de l'entreprise B.

La proportion de 46% s'observe donc dans moins de 5% des échantillons de taille 2500.

On peut alors rejeter l'hypothèse que l'entreprise B respecte la parité.

Par contre, pour l'entreprise A, on peut accepter cette hypothèse.

Question 2 : Quel nombre d'employés devrait avoir l'entreprise B pour être dans les 95 % de chances de respecter la parité ?

On veut que la proportion de femmes appartienne à l'intervalle de fluctuation $[0,48; 0,52]$, c'est-à-dire $0,48 \leq f \leq 0,52$. Pour un nombre d'employés de 2500, le nombre de femmes doit donc être compris entre $0,48 \times 2500$ et $0,52 \times 2500$, donc entre 1200 et 1300.

4. Intervalle de confiance

Théorème et définition :

On cherche à estimer p (inconnue) 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 observée du caractère dans l'échantillon.

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

$$IC = \left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}} \right]$$

IC est appelé l'intervalle de confiance au seuil 95%.

Démonstration : On sait que, dans au moins 95% des cas, $f \in \left[p - \frac{1}{\sqrt{n}}; p + \frac{1}{\sqrt{n}} \right]$, ce qui est équivalent à $p \in \left[f - \frac{1}{\sqrt{n}}; f + \frac{1}{\sqrt{n}} \right]$.

Exemple : Estimer une proportion inconnue

1) Avant les élections, le candidat A commande un sondage effectué sur 250 personnes. 138 personnes interrogées déclarent avoir l'intention de voter pour le candidat A.

Le candidat A peut-il espérer être élu ?

2) Le candidat A commande un second sondage effectué sur 1000 personnes pour lequel 538 personnes déclarent avoir l'intention de voter pour lui.

Le candidat A peut-il espérer être élu ?

1) Soit p la proportion effective d'électeurs pour le candidat A.

La proportion observée (fréquence) est égale à $f = \frac{138}{250} = 0,552$

L'intervalle de confiance de p au seuil de 0,95 est :

$$\left[0,552 - \frac{1}{\sqrt{250}}; 0,552 + \frac{1}{\sqrt{250}} \right] \text{ soit de façon approchée } [0,49 ; 0,62].$$

On a donc : $0,49 < p < 0,62$. Il est donc possible que le candidat A ne soit pas élu.

2) La proportion observée (fréquence) est égale à $f = \frac{538}{1000} = 0,538$

L'intervalle de confiance de p au seuil de 0,95 est :

$$\left[0,538 - \frac{1}{\sqrt{1000}}; 0,538 + \frac{1}{\sqrt{1000}} \right] \text{ soit de façon approchée } [0,51 ; 0,57].$$

On a donc : $0,51 < p < 0,57$.

La proportion effective évaluée est supérieure à 50%. Le candidat A peut donc espérer être élu puisque 95% des échantillons possibles de taille 1000 serait compris dans cet intervalle.