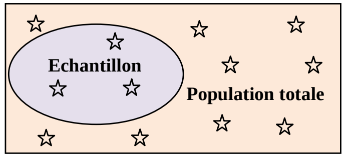
ECHANTILLONNAGE

Echantillonnage

Echantillon

Lorsqu'on étudie un caractère d'une population, la connaissance de la population entière n'est pas toujours envisageable.
On doit alors se contenter de la connaissance d'une partie – convenablement choisie - de cette population qu'on appelle alors échantillon.



Définition

À partir d'une population statistique, on appelle **échantillon** un sous-ensemble de cette population obtenu par prélèvement aléatoire. Le nombre d'individus de l'échantillon est appelé la **taille** de l'échantillon.

Exemple : une urne contient des boules vertes, rouges et bleues, la proportion des vertes étant $p = 0{,}4$.
On tire au hasard une boule de l'urne, on note sa couleur puis on la remet, et on répète ceci 4 fois de suite.
On obtient alors un échantillon de taille 4 ; si on a obtenu : V-R-B-V, la fréquence de boules vertes est $f = 0{,}5$.

Remarque : lorsqu'une partie est prélevée sans remise mais a une taille faible par rapport à la taille de la population entière, on peut la considérer comme un échantillon.

Exercice 1 :

Produire un échantillon de taille 30 dans le jeu du Pile ou Face (la population étant ici infinie) puis calculer la fréquence du Pile.

Fluctuation d'échantillonnage

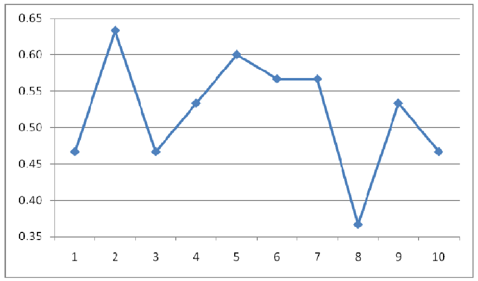
Dans l'exercice précédent, il est fort probable que les fréquences du Pile varient d'un élève à l'autre.
Pour une population donnée, des échantillons aléatoires produits suivants le même protocole peuvent avoir des compositions différentes : on dit qu'il y a fluctuation d'échantillonnage.

Définition

Si on réalise plusieurs échantillons de même taille, la distribution des fréquences varie d'un échantillon à l'autre . Ce phénomène est appelé **fluctuation d'échantillonnage**.

Exemple : Voici la simulation d'un pile ou face avec 10 échantillons de taille 30 :

Echantillon	1	2	3	4	5	6	7	8	9	10
Fréquence	0.47	0.63	0.47	0.53	0.60	0.57	0.57	0.37	0.53	0.47



Intervalle de fluctuation

Dans une population, la proportion d'un caractère est p.

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

Résultat

Si $0{,}2 \le p \le 0{,}8$ et $n \ge 25$, alors dans 95% des cas au moins, $f \in \left[p - \frac{1}{\sqrt{n}} \,;\, p + \frac{1}{\sqrt{n}} \right]$.

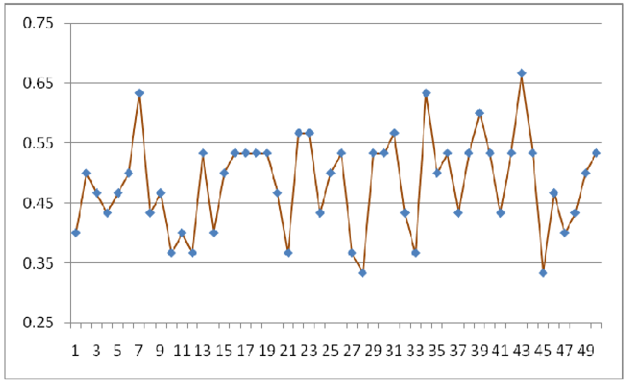
Cet intervalle est appelé **intervalle de fluctuation au seuil de 95%**.

Remarques

1. Ce résultat n'est valable que si $n \ge 25$, autrement dit l'échantillon est de taille « suffisante ».
2. Ce résultat signifie qu'il est « rare » (5% des cas) de trouver un échantillon dans lequel la fréquence f n'est pas dans l'intervalle indiqué. On pourra alors considérer qu'un tel échantillon n'est pas « normal », avec un risque de se tromper assez faible (5% des cas).
3. On peut remarquer que plus la taille des échantillons est grande, plus l'intervalle de fluctuation est petit.
   → Pour $n = 100$, on a $\frac{1}{\sqrt{n}} = 0{,}1$; → Pour $n = 1000$, on a $\frac{1}{\sqrt{n}} \simeq 0{,}03$.
4. Il existe bien sûr des résultats concernant des intervalles de fluctuation à d'autres seuils que 95%.

Exercice 2 :

Voici les fréquences du pile obtenues dans 50 échantillons de taille 50 du jeu Pile ou Face



1. Déterminer l'intervalle de fluctuation au seuil de 95% ; Faire apparaître cet intervalle sur le graphique.
2. Déterminer le pourcentage des fréquences f comprises dans cet intervalle.

Utilisation de l'échantillonnage

Prise de décision à partir d'un échantillon d'une population statistiquement connue

Exemple : Dans une population de truites de rivière, le *sex ratio* (proportion de mâles et de femelles) est de 0,5.
Certaines pollutions par des produits pharmaceutiques modifient ce *sex ratio* en augmentant la proportion de femelles.
Sur un prélèvement de 100 truites de rivière, on a relevé une fréquence de femelles égale à 0,64.

- Question : Cela est-il dû au seul hasard ou est-on en droit de suspecter une pollution ?
- Réponse :

→ On suppose que l'effectif de la population est très grand, le prélèvement ainsi effectué peut alors être assimilé à un échantillon de taille 100 de la population.

→ $p = 0{,}5$ donc $0{,}2 \leq p \leq 0{,}8$ et $n = 100$ donc $n \geq 25$;

→ L'intervalle de fluctuation au seuil 95% est : $\left[0.5-\frac{1}{\sqrt{100}}\,;0.5+\frac{1}{\sqrt{100}}\right]=\left[0.4\,;0.6\right]$;

→ Comme $0.64 \notin \left[0.4\,;0.6\right]$, on considère que cette fréquence n'est pas due à la fluctuation d'échantillonnage et donc on est en droit de suspecter une pollution !

Exercice 3 :

La proportion des yeux bleus en France est d'environ 0,31.

1. On a prélevé un échantillon de 50 individus dont 15 aux yeux bleus.
   Cet échantillon est-il représentatif de la population pour ce caractère ?
2. La Seconde … du lycée Grignard est-elle un échantillon représentatif de la population française pour ce caractère ?

Exercice 4 :

On a tiré 100 fois avec remise une boule dans une urne contenant la même quantité de boules rouges et de boules noires.
On a obtenu 40 boules rouges.
Si on effectue un nouveau tirage de 100 boules dans les mêmes conditions, à quel nombre de boules rouges peut-on s'attendre :

- moins de 40
- 40
- entre 40 et 60
- on ne peut pas savoir.

Exercice 5 :

On suppose qu'en moyenne 51% des nouveaux nés sont de sexe masculin.

1. Déterminer l'intervalle de fluctuation au seuil de 95% de la fréquence des garçons nouveau x nés dans des échantillons de taille 132 pris au hasard.
2.
   a. Dans une petite commune, il est né entre 1999 et 2003, 132 enfants dont 56 garçons.
   Peut-on considérer, avec 5% de risque, que cette situation est le seul fruit du hasard ?

   b. Au Canada, dans une réserve indienne, il est né entre 1999 et 2003, 132 enfants dont 46 garçons.
   Peut-on considérer, avec 5% de risque, que cette situation est le seul fruit du hasard ?

Exercice 6 :

Avant une campagne publicitaire, une société disposait de 40% de part de marché sur l'un de ses produits.
Après la campagne, elle effectue un test sur un échantillon de taille 200 et obtient 48% de part de marché.
Cette augmentation est-elle due à la publicité ?

Exercice 7 :

Dans une usine automobile, on contrôle les défauts de peinture de type « grains ponctuels sur le capot », invisibles par le client.

Lorsque le processus de fabrication est standard, on a 20 % de véhicules présentant ce type de défauts.

1. Lors du contrôle aléatoire de 50 véhicules, on en observe 26 % présentant ce type de défauts. Faut-il s'inquiéter ?
2. Sur 50 véhicules contrôlés, à partir de combien ayant ce type de défauts, devrait-on s'inquiéter ?

Estimation d'une proportion dans une population statistiquement inconnue, à partir d'un échantillon

Exemple : On s'intéresse à la proportion p d'un caractère dans une population.
Cette proportion p est supposée inconnue mais on sait que dans un échantillon de taille 100, la fréquence mesurée du caractère est $f = 0.58$.

- Question : Quelle est une estimation de p ?
- Réponse :

→ On peut penser que p est du même ordre de grandeur que f ;

→ Comme $0{,}2 \leq p \leq 0{,}8$ et $n = 100$ donc $n \geq 25$, dans 95% des cas au moins , $f \in \left[p-\frac{1}{\sqrt{100}}\,;p+\frac{1}{\sqrt{100}}\right]$ donc en supposant que 0,58 fait partie de ces cas, il vient :

$$0{,}58 \in \left[p-\frac{1}{\sqrt{100}}\,;p+\frac{1}{\sqrt{100}}\right] \Leftrightarrow p-\frac{1}{\sqrt{100}} \leq 0{,}58 \leq p+\frac{1}{\sqrt{100}}$$

d'où :

$$p-\frac{1}{\sqrt{100}} \leq 0{,}58 \Leftrightarrow p \leq 0{,}58+\frac{1}{\sqrt{100}} \quad \text{et} \quad 0{,}58 \leq p+\frac{1}{\sqrt{100}} \Leftrightarrow p \geq 0{,}58-\frac{1}{\sqrt{100}}\ ;$$

Ainsi, $p \in \left[0{,}58-\frac{1}{\sqrt{100}}\,;0{,}58+\frac{1}{\sqrt{100}}\right] \approx \left[0{,}48\,;0{,}68\right]$;

→ Il n'est pas sûr que l'intervalle $\left[0{,}48\,;0{,}68\right]$ contienne p mais, dans 95% des cas, le procédé fournit un intervalle qui contient p.

Résultat

Si $0{,}2 \leq p \leq 0{,}8$ et $n \geq 25$, alors dans 95% des cas au moins, $p \in \left[f-\frac{1}{\sqrt{n}}\,;f+\frac{1}{\sqrt{n}}\right]$.

Cet intervalle est appelé **intervalle de confiance au seuil 95%**.

Exercice 8 :

Lors d'un sondage effectué auprès de 900 personnes, 51% d'entre elles déclarent vouloir voter pour le candidat A.
En supposant que les personnes sondées ont répondu sincèrement et qu'elles ne changent pas d'avis le jour du vote, le candidat A peut-il raisonnablement penser qu'il sera élu au 1er tour ?

Exercice 9 :

Le taux d'audience du journal de 20h d'une chaîne de télévision a été mesuré à partir de 1000 appareils installés au hasard chez des téléspectateurs. On a relevé une audience de 31%.

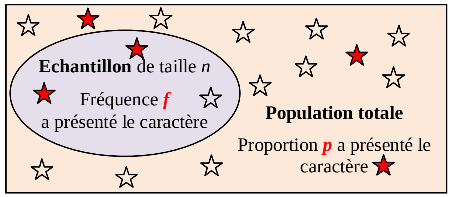
1. Déterminer la fourchette de sondage au seuil de 95% de l'audience de cette chaîne à cette heure là.
2. Une autre chaîne de télévision prétend avoir une meilleure audience à 20h en indiquant qu'un institut de mesure a relevé 40% d'audience sur 144 appareils installés au hasard chez des téléspectateurs.
   Que pensez-vous de cette affirmation ?

Exercice 10 :

Un sondage effectué au hasard sur 400 personnes en âge de voter, en donne 144 favorables au rétablissement de la peine de mort.

1. Donner la fourchette de sondage au seuil de 95% de la proportion p des personnes favorables au rétablissement de la peine de mort dans la population.
2. Un deuxième sondage effectué au hasard sur 625 personnes en âge de voter, en donne 175 favorables au rétablissement de la peine de mort. Ce résultat est-il compatible avec le précédent ?
3. Si ces deux sondages font partie de ceux dont la fourchette de sondage contient la proportion p, donner un encadrement de cette proportion.

En résumé



Intervalle de fluctuation au seuil de 95%	**Intervalle de confiance** au seuil de 95%
on connaît : la proportion p des individus ayant un caractère donné au sein d'une population.	**on connaît** : la proportion f des individus ayant un caractère donné au sein d'un échantillon de taille n.
on cherche : à savoir si un échantillon de taille n dans lequel la fréquence f de ce caractère, est « normal »	**on cherche** : à estimer la proportion p des individus ayant ce caractère donné au sein de la population totale.
Probablement si $f \in \left[p-\frac{1}{\sqrt{n}}\,;p+\frac{1}{\sqrt{n}}\right]$.	Probablement que $p \in \left[f-\frac{1}{\sqrt{n}}\,;f+\frac{1}{\sqrt{n}}\right]$.