1 Échantillonnage

On s'intéresse à l'étude d'un caractère (quantitatif ou qualitatif) des N individus d'une population. Pour chacun des individus de la population, le caractère peut à priori prendre des valeurs aléatoirement différentes.

Lorsqu'on n'a pas accès à l'ensemble de la population, **on procède à un échantillonnage**, i.e. au choix de n individus dans la population, sur lesquels on observe la valeur du caractère.

Exemple. Lorsqu'on lance un dé un certain nombre de fois ou lorsqu'on interroge des électeurs sur le nom du candidats pour lequel ils comptent voter. On dit qu'on dispose d'un échantillon de données.

Par exemple, l'expérience qui consiste à lancer 100 fois un dé peut conduire à l'échantillon :

Numéro	1	2	3	4	5	6
Effectif	15	21	17	13	16	18

Le résultat d'un lancer n'influe pas sur le suivant, on dit que les lancers sont indépendants.

Définition 1 :

Un **échantillon de taille n** est la collection des n résultats obtenus après n répétitions indépendantes d'une même expérience aléatoire.

Exemple. « FONDAMENTAL »
L'épreuve de Bernoulli consiste en une expérience aléatoire n'ayant que deux issues.

Ainsi, effectuer un sondage dans une population amenée à choisir lors d'une élection entre deux candidats A et B revient à obtenir un échantillon du type : A - B - A - B - B - ...

2 Intervalle de fluctuation

Remarque. « préliminaire » Revenons sur l'exemple du lancer de 100 dés, en réitérant l'expérience, on obtient un nouvel échantillon qui n'a aucune raison de fournir les mêmes résultats. Il en va de même pour un second sondage effectué sur une même population. Ce phénomène est appelé **la fluctuation d'échantillonnage**. On peut cependant dans le second cas, avoir une idée de cette fluctuation.

Théorème 1 : (admis)

Considérons une expérience de Bernoulli où l'échantillon est de taille $n \geqslant 25$.
Supposons que l'une des issues a pour probabilité p avec $0{,}2 \leqslant p \leqslant 0{,}8$
Lorsqu'on dispose d'un grand nombre d'échantillons (de même taille), pour au moins 95% d'entre eux, les fréquences observées seront comprises dans l'intervalle :

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

Ce dernier est appelé **intervalle de fluctuation de la fréquence de f au seuil de 95%** .
Ainsi lorsqu'on dispose d'un échantillon, on a 95% de chances que la fréquence observée appartienne à cet intervalle. L'échantillon est représentatif (ou non biaisé) si et seulement si sa fréquence d'apparition de la caractéristique est dans cet intervalle.

Remarque. La longueur de l'intervalle diminue lorsque la taille de l'échantillon devient grand. Ainsi, la fréquence d'observation se rapproche de p : il s'agit d'une illustration de la loi des grands nombres.

On dira que les fluctuations d'échantillonnage de f autour de p sont d'autant plus faibles que n est grand. Quand la taille de l'échantillon, n, tend vers l'infini, la fréquence observée f tend vers p.

Exemple. On sait que dans une population donnée, il y a 60% de fumeurs, soit une proportion $p = 0{,}6$ de fumeurs. Sur 400 malades atteints d'un cancer des bronches, on trouve 333 fumeurs, soit une fréquence $f = 0{,}833$ de fumeurs. Un tel résultat ne suffit pas à prouver que le tabagisme augmente les chances de cancer des bronches. Il faut savoir si la différence entre 0,6 et 0,83 est significative d'un comportement différent des malades atteint du cancer des bronches. Si les malades se comportent comme le reste de la population, on devrait encore avoir une proportion $p = 0{,}6$ de fumeurs chez les malades.

Pour un échantillon de $n = 400$ personnes issues de la population, l'intervalle de fluctuation de f serait $\left[p - \frac{1}{\sqrt{n}} \ ; \ p + \frac{1}{\sqrt{n}}\right] = \left[0{,}6 - \frac{1}{\sqrt{400}} \ ; \ 0{,}6 + \frac{1}{\sqrt{400}}\right] = [0{,}55 \ ; \ 0{,}65]$.

Comme $f = 0{,}83$ n'appartient pas à cet intervalle, on peut considérer que la proportion $p = 0{,}6$ n'est pas compatible avec l'observation f, et ici qu'il y a plus de fumeurs chez les malades atteints d'un cancer des bronches. Comme toujours, on a considéré que l'échantillon observé faisait partie des 95% d'échantillons donnant une fréquence dans l'intervalle de fluctuation. Le risque d'erreur est donc de 5%.

3 Applications

Dans la suite, on dispose d'un échantillon pour laquelle on observe une fréquence f .

3 1 On suppose p connu et on teste l'hypothèse $f = p$

C'est par exemple le cas si l'on cherche à savoir si une pièce est truquée à partir d'un échantillon de lancers de pièces. Ainsi, si l'on obtient 2 050 fois piles en lançant 4 000 fois une pièce et que l'on veuille tester l'hypothèse $f = p = 0{,}5$, cela revient à tester si la pièce est truquée. Or on sait que si $n = 4\,000$ alors pour au moins 95% des expériences (qui consistent à lancer 4 000 pièces), les fréquences appartiendront à l'intervalle : $\left[0{,}5 - \frac{1}{200} \ ; \ 0{,}5 + \frac{1}{200}\right] = [0{,}495 \ ; \ 0{,}505]$. Ici, on a : $f = \frac{2050}{4000} = 0{,}512\,5 \notin [0{,}495 \ ; \ 0{,}505]$. Ce qui signifie qu'on a 95% de chances de ne pas se tromper en supposant la pièce truquée mais aussi 5% de faire erreur.

Exemples.

a. Dans la réserve indienne d'Aamjiwnaag, située au Canada il est né entre 1999 et 2003, 132 enfants dont 46 garçons. Est-ce le fruit du hasard ?

46 garçons sur 132 naissances alors que la fréquence théorique de garçons est $p = 0{,}5$.

$0{,}5 - \frac{1}{\sqrt{132}} \simeq 0{,}413$ et $0{,}5 + \frac{1}{\sqrt{132}} \simeq 0{,}587$

Donc l'intervalle de fluctuation au seuil de 95% est $[0{,}413 \ ; \ 0{,}587]$. Il n'y a donc que 5% de chances d'obtenir une valeur en dehors de cet intervalle.

La valeur $\frac{46}{132} = 0{,}348$ est nettement en dehors de cet intervalle. Il y a lieu de se poser des questions et de chercher quelle peut-en être la raison.

b. Les entreprises sont sensées ne pas faire de discrimination quant au sexe des personnes employées. Deux entreprises A et B ont respectivement 41 femmes pour 100 employés et 4 850 femmes sur 10 000 employés. Pour chacune des entreprises, la sélection est-elle équitable ?

Pour l'entreprise A l'intervalle de fluctuation de la fréquence au seuil de 95% est :

$\left[0{,}5 - \frac{1}{100} \ ; \ 0{,}5 + \frac{1}{100}\right] = [0{,}4 \ ; \ 0{,}6]$ or $f = \frac{41}{100} = 0{,}41$ et on a $0{,}41 \in [0{,}4 \ ; \ 0{,}6]$ donc l'échantillon est représentatif d'une situation de parité.

Pour l'entreprise B, l'intervalle de fluctuation est : $\left[0{,}5 - \frac{1}{10000} \ ; \ 0{,}5 + \frac{1}{10000}\right] = [0{,}49 \ ; \ 0{,}51]$

or ici $f = \frac{4850}{10000} = 0{,}485$ et 0,485 n'appartient pas à l'intervalle, donc l'échantillon n'est pas représentatif d'une situation de parité.

3 2 On suppose p inconnu et on teste l'hypothèse $p = f$

Le parti d'un candidat commande un sondage réalisé à partir de 1 600 personnes à l'issue duquel il est donné gagnant avec 52% des voix. A-t-il des raisons d'être confiant ?

On peut répondre à la question si on montre $p > 0{,}5$ avec une grande probabilité. Le problème est que p est inconnu ...

On remarque alors que :

$$p - \frac{1}{40} \leqslant f \leqslant p + \frac{1}{40} \iff f - \frac{1}{40} \leqslant p \leqslant f + \frac{1}{40} \quad \textbf{intervalle de confiance au risque de 5\%}$$

$$\iff 0{,}52 - 0{,}025 \leqslant p \leqslant 0{,}52 + 0{,}025$$

$$\iff 0{,}495 \leqslant p \leqslant 0{,}545$$

On ne peut donc conclure. De toute façon, même si l'on avait obtenu $p > 0{,}5$, on aurait eu 5% de chances de se tromper en pensant que l'élection était gagnée.

Exemples.

a. **Lors d'un référendum**, un sondage aléatoire simple pratiqué sur 1 000 personnes a donné 55% pour le "Oui" et 45% pour le "Non". Peut-on prévoir le résultat du référendum ?

L' **intervalle de confiance** est $\left[f - \frac{1}{\sqrt{n}} \ ; \ f + \frac{1}{\sqrt{n}}\right] = \left[0{,}55 - \frac{1}{\sqrt{1000}} \ ; \ 0{,}55 + \frac{1}{\sqrt{1000}}\right] = [0{,}518 \ ; \ 0{,}582]$

Avec un risque d'erreur de 5%, on peut dire le "Oui" va l'emporter.

b. **Si, pour un référendum**, on sait que "oui" se situe autour de 50%, combien de personnes faudrait-il interroger pour que la proportion de "Oui" soit connue à 1% près ? (en plus ou en moins)

L' **intervalle de confiance** est $\left[f - \frac{1}{\sqrt{n}} \ ; \ f + \frac{1}{\sqrt{n}}\right]$. On veut $\frac{1}{\sqrt{n}} \leqslant 0{,}01$ soit $n \geqslant \left(\frac{1}{0{,}01}\right)^2 = 10\,000$

c. **Pourcentage de garçons à la naissance.**

Dans un pays, sur 429 440 naissances, on a dénombré 221 023 garçons. Ce résultat est-il conforme à l'hypothèse selon laquelle il y a 50% de naissances masculines (et donc 50% de naissances féminines) ?

Intervalle de confiance de niveau 0,95 : $\left[\frac{221\,023}{429\,440} - \frac{1}{\sqrt{429\,440}} \ ; \ \frac{221\,023}{429\,440} + \frac{1}{\sqrt{429\,440}}\right] = [0{,}5132 \ ; \ 0{,}5162]$

soit entre 51.32% et 51.62% de naissances masculines donc non conformité avec l'hypothèse.

d. Le dernier sondage de 2002 ne prévoyait pas la présence de Jean-Marie Le Pen au second tour. Pouvait-on croire au sondage ?

21 Avril 2002 second tour de l'élection présidentielle en France.

Les sondages (1000 p) prévoient :	M Chirac : 19 %	M Jospin : 18 %	M Le Pen : 14 %
Les Résultats sont : Surprenant !	M Chirac : 19,88 %	M Jospin : 16,18 %	M Le Pen : 16,88 %
Les intervalles de confiance à 95 %	[16 % ; 22 %]	[15 % ; 21 %]	[11 % ; 17 %]

Il n'y a pas de surprise, seulement que M Jospin est dans la partie basse et M Le Pen dans la partie haute.

Remarque. Il y a autant d'intervalles de confiance que d'échantillons. Ils sont centrés sur la fréquence f de l'échantillon.

3 3 Qu'est-ce qu'un sondage ?

Un maire voudrait connaître le pourcentage de personnes de sa commune favorables à un projet d'urbanisme, et ceci à partir d'une enquête portant sur un nombre restreint d'individus.

Il demande à quatre collaborateurs comment procéder.

- le 1er propose d'ouvrir à la mairie un registre pour recueillir l'avis des personnes désirant s'exprimer sur le sujet ;
- le 2e d'interroger les 1 350 habitants de son quartier ;
- le 3e d'interroger les 100 premières personnes rencontrées dans la rue, le mardi suivant à partir de 10 h.
- le 4e de sélectionner, de façon totalement aléatoire, 100 individus à interroger, à partir de la liste des habitants de la commune.

Tous les quatre pensent ensuite utiliser la formule bien connue $\left[f - \frac{1}{\sqrt{n}} \ ; \ f + \frac{1}{\sqrt{n}}\right]$. Ils affirment avoir une probabilité 0,05 de se tromper en disant que la proportion cherchée est dans cet intervalle.

À la place du maire, et indépendamment de toute considération de coût ou de difficulté de réalisation pratique (et en supposant que toutes les personnes interrogées répondent), quelle méthode choisiriez-vous ?

Celle du collaborateur :

1 : *Non, car seules les personnes intéressées feront la démarche d'aller jusqu'à la mairie. Celles qui n'en ont pas le temps, ou qui ne se sentent pas assez concernées, ne seront pas consultées. L'échantillonnage ainsi réalisé ne sera pas représentatif de toute la population.*

2 : *Non, la taille de l'échantillon est grande, ce qui permettrait une bonne précision si on avait un échantillon vraiment aléatoire, mais il ne s'agit pas d'un tirage au hasard sur l'ensemble de la population, puisqu'on exclut du sondage tous les habitants des autres quartiers.*

3 : *Non, car on exclut du sondage toutes les personnes qui travaillent le mardi matin. On risque de n'interroger que des femmes au foyer, des retraités ou des chômeurs. L'échantillon ne serait pas représentatif de l'ensemble des habitants de la ville.*

4 : **Oui**, c'est la seule démarche qui permette de justifier le recours à la formule donnant l'intervalle de confiance. Il est nécessaire d'avoir un échantillon aléatoire simple : tous les habitants ont la même chance d'être choisis, et de façon indépendante. Personne n'est exclu du sondage.

Le maire décide donc de choisir, à partir d'une liste de plusieurs milliers de noms, 100 personnes, "totalement au hasard". Mais comment faire pour être sûr d'agir "en toute objectivité" ? Une solution est l'utilisation de tables de nombres au hasard, ou de procédés informatiques. À partir d'une liste numérotée de N noms, choisir les numéros de n personnes, de façon à ce que chacun ait la même probabilité d'être choisi, et de façon indépendante.

D'autre part, si le maire pense que son projet risque d'être ressenti différemment par les hommes et les femmes (implantation d'un stade de foot-ball par exemple), ou selon les tranches d'âge, et que la liste d'habitants dont il dispose mentionne le sexe et l'âge, que faire ?

Il peut améliorer la précision de son estimation en choisissant au hasard un certain nombre d'hommes, un certain nombre de femmes, un certain nombre d'individus par tranche d'âge. C'est ce que l'on appelle un sondage **stratifié**. De même, il peut être logique de procéder dans certains cas à des sondages à probabilités inégales : par exemple si les individus sont des entreprises, il peut être utile de les choisir avec des probabilités proportionnelles à leur chiffre d'affaire, ou au nombre de leurs salariés.

Remarque. Une méthode de sondage consiste à définir la façon dont on va prélever les individus dans la population afin de constituer un échantillon. Lorsque tous les individus ont la même probabilité d'appartenir à l'échantillon sélectionné, on parle de **sondage à probabilités égales**. Un sondage aléatoire est dit simple si tous les échantillons de taille n fixée sont réalisables avec la même probabilité. Il existe également des **sondages stratifiés** (s'appuyant sur des sous-populations appelées strates constituées à partir des données portant sur l'ensemble de la population), des **sondages par la méthode des quotas** (analogue aux sondages stratifiés mais avec probabilité inégales d'appartenir à l'échantillon sélectionné), ...