Chapitre 4 : La constitution de l'échantillon

A. Introduction

Comment, à partir de l'observation d'un nombre restreint d'individus (l'échantillon), tirer des conclusions valables pour un groupe d'individus plus étendu (la population) ? Tel est le problème de l'échantillonnage, qui constitue un des moments importants dans nombre d'enquêtes pratiquées en sciences sociales.

On se pose un problème sur une **population**, c'est-à-dire un ensemble d'**individus**. L'individu, en statistiques, n'est pas forcément une personne, mais peut être un groupe de personnes. Par exemple, la population peut être l'ensemble des classes de CM1 en Côte d'Ivoire ; ou l'ensemble des élèves qui redoublent leur CP au Sénégal ; ou encore l'ensemble des ménages de la ville de Ouagadougou. Dans ces trois exemples, l'individu est défini respectivement comme : une classe, un élève ou un ménage.
On souhaite connaître certaines caractéristiques de cette population, en particulier une **moyenne** ou un **pourcentage**. Soit par exemple : le niveau scolaire moyen à un test standardisé et le pourcentage de femmes enseignant dans les classes de CM1 en Côte d'Ivoire ; le milieu socio-culturel des élèves redoublant leur CP au Sénégal ; et le niveau de vie des ménages de Ouagadougou.

On a deux bonnes raisons de ne pas chercher à interroger ou observer toute la population :
· Ce n'est **pas** toujours **faisable**, et c'est souvent très coûteux ;
· Cela risque de ne **pas** être **fiable**, car lors d'une collecte de données sur des dizaines de milliers d'individus, la qualité est difficile à contrôler et la période de recueil d'information risque d'être trop longue pour que toutes ces données soient comparables (en particulier, des tests de niveaux doivent être passés sur une période relativement courte si on veut que les élèves aient bénéficié des mêmes durées d'apprentissage).
Pour ces deux raisons, on va se limiter à un nombre restreint **d'individus : l'échantillon**.

A partir de là, les deux problèmes symétriques de l'échantillonnage sont :

1. Selon la façon dont a été constitué l'échantillon, avec quelle précision puis-je estimer les grandeurs qui m'intéressent concernant la population ?
2. Sachant mes contraintes (budgétaires notamment), quelle est la meilleure façon de constituer mon échantillon pour répondre le plus précisément possible aux questions que je me pose sur la population que j'étudie ?

Nous allons envisager successivement trois types d'échantillon, du plus simple au plus complexe, afin de parvenir aux pratiques d'échantillonnage utilisables dans le cadre de nos enquêtes d'évaluation des systèmes éducatifs :
· L'échantillon aléatoire simple ;
· L'échantillon stratifié ;
· L'échantillon en grappes.

B. L'échantillon aléatoire simple

1. Le principe :

Le principe de l'échantillon aléatoire simple tient en peu de mots : **les individus de l'échantillon sont tirés au sort directement à partir de l'ensemble de la population, en donnant à chaque individu la même chance d'être retenu.**

C'est le principe de la tombola : chacun des N individus de la population a un ticket et un seul pour le représenter ; on rassemble les N tickets dans une même urne ; on tire n tickets de cette urne ; les n individus correspondant constituent l'échantillon[1] .
Concrètement, on peut procéder de multiples façons. Supposons qu'on veuille tirer par tirage aléatoire simple 20 élèves dans une classe de 60. On peut :
- Mettre 20 papiers marqués d'une croix et 40 non marqués dans une urne. Chaque élève vient tirer un papier à son tour et le garde. Celui qui tire un papier marqué d'une croix est retenu, l'autre non.
- Ranger les élèves par ordre alphabétique, et prendre un élève sur 3 à partir du nombre 1, 2 ou 3 tiré au hasard (si on tire 2, on prendra le 2ème, le 5ème, le 8ème,… élève de la liste).

[1] Une convention de notation est utilisée au long de ce chapitre : les lettre majuscules indiquent les grandeurs qui concernent la population ; les lettres minuscules celles qui concernent l'échantillon.

- Prendre une liste ordonnée des élèves (que ce soit l'ordre alphabétique ou un autre ordre) et utiliser une série de nombre aléatoire (en la générant à partir d'un ordinateur ou en recourant à des tables données par des statisticiens) pour retenir certains d'entre eux.

L'important dans tous ces cas de figure est qu'on procède à un tirage au sort à un seul niveau et que tous les individus ont la même chance d'être sélectionnés.

2. Les propriétés d'un échantillon aléatoire simple :

En général, on ne tire qu'un seul échantillon. Mais pour décrire les propriétés de l'échantillonnage, on se demande ce qui se passerait si on tirait plusieurs échantillons selon le même principe :

- Ces échantillons, si on les répétait à l'infini, nous permettraient-ils de retrouver, en moyenne, les propriétés de la population complète ?
  Par exemple, si je tire un grand nombre d'échantillons de 20 élèves parmi les 60 de la classe, et que je calcule le résultat moyen des élèves sur ces échantillons, est-ce que je trouve une valeur qui s'approche de la moyenne des 60 élèves ?
  Cela nous dit si, en moyenne, nos échantillons visent la bonne cible. Si oui, ils sont dits non biaisés.

- Comme dans la plupart des cas on n'a qu'un seul échantillon et qu'il n'est pas possible, par conséquent, de faire la moyenne sur un ensemble d'échantillons, la seconde question est :
  Quelle est la distance moyenne entre l'estimation donnée par un échantillon (il peut s'agir d'une moyenne, d'un pourcentage) et la vraie valeur[1] (moyenne ou proportion) prise par la population complète ? Autrement dit, quelle est la précision probable d'un échantillon pris au hasard ? **Cela nous dit non seulement si les échantillons visent la bonne cible, mais s'ils sont en moyenne proches ou éloignés du but.**

[1] On appelle généralement «vraie valeur» la valeur qui concerne la population totale. L'objectif de la théorie statistique est alors de juger des propriétés des différents estimateurs possibles par rapport à cette vraie valeur : sont-ils centrés autour ? en sont-ils proches ?

L'échantillonnage aléatoire simple a les propriétés suivantes :

Propriété 1 : Des échantillons aléatoires simples répétés tournent bien autour de la vraie valeur (l'échantillon est non biaisé) ;

Propriété 2 : La **précision d'un échantillon** donné dépend :

1. Du nombre d'individus présents dans l'échantillon *(noté n)* : **plus il y a d'individus dans mon échantillon, moins les caractéristiques de l'échantillon risquent de différer de celles de la population** ;
2. De la variabilité des caractéristiques des individus au sein de la population *(la variance, notée S^2)* : **moins les individus de la population sont hétérogènes, moins les caractéristiques de l'échantillon risquent de différer de celles de la population** ;
3. De la proportion d'individus de la population qui a été représentée dans l'échantillon *(noté n/N)* : **plus la proportion d'individus que j'ai représentés dans mon échantillon est grande, moins les caractéristiques de l'échantillon risquent de différer de celles de la population.**

Supposons que notre échantillon serve à mesurer une moyenne ou une proportion au sein d'une population (par exemple, le score moyen ou la proportion de filles dans une population d'élèves). Notons Y cette «vraie valeur», non mesurée et qu'on veut estimer.
On tire un échantillon de n individus. On calcule dessus la même moyenne (ou proportion) et on la note y. Autant Y est fixe (mais inconnue), autant y varie selon l'échantillon que l'on tire. **La variabilité de y en fonction de l'échantillon tiré est la source même de l'erreur d'échantillonnage.** En effet, plus les tirages possibles nous donnent des estimations différentes (plus y, donc, peut prendre des valeurs différentes), plus y peut s'écarter de Y.
C'est pourquoi l'amplitude du risque d'erreur d'échantillonnage est mesurée par la variance de y, notée V(y). Plus la variance est élevée, moins l'échantillon permet de tirer des conclusions précises sur les caractéristiques de la population.
On montre que :

$$V(y) = (1 - \frac{n}{N}) \cdot \frac{S^2}{n}$$

Cette formule permet d'estimer V(y) à partir d'un unique échantillon, à condition de prendre pour estimation de S^2 (la variance au sein de la population) la variance au sein de l'échantillon (s^2)[1] .

Pour bien comprendre comment V(y) affecte la fiabilité de l'estimation de Y, on utilise souvent la notion d'**intervalle de confiance** pour la vraie valeur Y. Un intervalle de confiance est une fourchette qui encadre la vraie valeur Y, avec un degré donné de certitude (par exemple à 95%). Ces fourchettes seront d'autant plus larges que V(y) est grand.
Plus précisément, on montre que l'intervalle de confiance à 95% se calcule de la façon suivante :

$$\left[y - 1{,}96 \times \sqrt{V(y)} ; y + 1{,}96 \times \sqrt{V(y)} \right]$$

Cet intervalle de confiance peut se lire de la façon suivante : «la valeur estimée de Y la plus probable, au vu de notre échantillon, est y, et on peut affirmer avec seulement 5% de chances de se tromper que Y se trouve entre $\underline{y}$ et $\overline{y}$ », où $\underline{y}$ et $\overline{y}$ sont deux bornes qui encadrent y.

Remarque : Dans les formules qui précèdent, tous les paramètres sont connus au moment de la préparation de l'échantillon, sauf un : S^2. Il ne pourra être estimé qu'une fois l'échantillon tiré. Si on veut donc prévoir la précision de l'estimation par échantillonnage avant d'effectuer le tirage, il va falloir choisir des valeurs raisonnables pour S^2 (par exemple en s'inspirant d'autres études).

Un exemple permet d'appliquer les formules et de saisir leur intuition et leur portée :

On veut mesurer le niveau moyen des élèves de CM1 dans un pays par un test standardisé. On note Y ce niveau moyen. Y ne pouvant être mesuré en faisant passer à tous les élèves le même test, on choisit de faire passer ce test à 400

[1] $s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})}{n-1}$

élèves tirés individuellement au hasard parmi les 40 000 élèves de la population (les 40 000 élèves qui se trouvent au CM1 cette année-là).

On obtient un échantillon dont les caractéristiques sont les suivantes :

- La moyenne (notée y) vaut 54 (sur 100) ;
- Sur les 400 élèves, la variance observée s^2 est de 16 (ce qui signifie en simplifiant que les individus de l'échantillon se situent en moyenne à 16 points du score moyen de 54). Plus précisément, on a obtenu s^2 par le calcul suivant :

$$s^2 = \frac{\sum_{i=1}^{400} (y_i - \overline{y})}{399} = 16$$

Que peut-on dire alors du niveau moyen de l'ensemble des élèves de la population ? L'échantillon permet d'estimer ce niveau moyen à 54. Mais il permet aussi d'estimer la précision de cette estimation : dans la formule de V(y), on estime S^2 (la variance des scores sur l'ensemble de la population) par s^2 (la variance des scores sur l'échantillon) ; par ailleurs, n et N sont connus d'où :

$$\hat{V}(y) = (1 - \frac{n}{N}) \times \frac{s^2}{n} = (1 - \frac{400}{40000}) \times \frac{16}{400} = 0{,}0392$$

et l'intervalle de confiance estimé à 95% est :

$$\left[\underline{y}; \overline{y}\right] = \left[y - 1{,}96 \times \sqrt{\hat{V}(y)}; y + 1{,}96 \times \sqrt{\hat{V}(y)} \right] = \left[53{,}6; 54{,}4\right]$$

Autrement dit, on connaît le score de la population sur 100 à 0,4 points près avec un degré de certitude de 95%.

Commentaire :

➢ Tout d'abord, constatons que, lorsque N est grand (c'est-à-dire lorsque la taille de la population est élevée) le terme $\frac{n}{N}$ *a un impact mineur : si on le néglige,* $\hat{V}(y)$ *passe simplement de 0,0392 à 0,04 : la différence est indiscernable pour peu qu'on arrondisse les bornes de l'intervalle de confiance comme nous l'avons fait. Or c'est le seul endroit où N apparaît dans la formule.*

Autrement dit, à partir du moment où la population est suffisamment nombreuse, sa taille n'influe pas sur la précision de l'estimation donnée par l'échantillon. Ce résultat, peut-être pas intuitif, est au fondement de l'efficacité des techniques de sondage : que notre population d'élèves soit de 40 000 ou de 80 000 n'a quasiment aucun impact sur la précision de notre estimation par échantillonnage. L'échantillonnage est donc une méthode très efficace pour estimer des grandeurs sur des populations très nombreuses.

- Restent les deux autres paramètres : dans quelle mesure est-il intéressant d'augmenter la taille de l'échantillon n ? Que se passe-t-il si la population des élèves de CM1 est en fait plus hétérogène (s^2 plus élevé) ?

- Voyons d'abord comment évolue notre intervalle de confiance selon le paramètre s^2. Le tableau donne l'ampleur de la fourchette (c'est-à-dire le paramètre $\left[1{,}96 \times \sqrt{\hat{V}(y)}\right]$) selon différents degrés d'hétérogénéité de la population étudiée.

S^2	0	1	5	10	16	22	30	40
Intervalle pour une précision de 95%	0	±0,10	±0,22	±0,31	±0,39	±0,46	±0,53	±0,62

Prenons le cas extrême où l'hétérogénéité est nulle dans l'échantillon ($s^2=0$).*On va alors estimer que l'hétérogénéité est nulle dans la population* ($S^2=0$). *La conclusion s'impose : puisque tous les individus sont identiques, n'importe quel échantillon nous donnera la vraie valeur de la moyenne, qui est aussi le score de chaque élève. D'où l'intuition de la proposition suivante :*

Plus l'hétérogénéité de la population est grande, plus les échantillons vont différer entre eux et plus ils vont pouvoir s'écarter de la moyenne de la population totale. Les intervalles dans lesquels on pourra situer la vraie moyenne sans risque de se tromper seront alors d'autant plus vastes.

C'est ce qu'on constate en se déplaçant vers les colonnes de droite du tableau

➢ *Examinons ensuite les effets de la taille de l'échantillon sur la taille de la fourchette d'estimation :*

n	100	200	400	600	800	10 000	20 000	40 000
Intervalle pour une précision de 95%	±0,78	±0,55	±0,39	±0,32	±0,27	±0,07	±0,04	0

Considérons le cas extrême où l'échantillon est égal à la population totale (n=40 000). Là encore, la moyenne de la population est directement évaluée : aucun risque d'erreur d'échantillonnage. Mais à mesure que la taille de l'échantillon diminue, le risque d'erreur s'accroît. Pourtant, de très grands échantillons ne fournissent pas un gain majeur. Par exemple, on gagne plus à passer de 100 à 200 individus (100 individus de plus dans l'échantillon réduisent la marge d'erreur de 0,78 - 0,55 = 0,23 points) que de 10 000 à 20 000 (10 000 individus de plus réduisent la marge d'erreur de seulement 0,07 - 0,04 = 0,03 points).

Le gain en précision lié à l'ajout d'un individu supplémentaire décroît avec la taille de l'échantillon, au point de devenir négligeable lorsque l'échantillon est grand. Il y a donc une «juste taille» à trouver pour un échantillon : à partir d'une certaine taille, augmenter encore l'échantillon ne se justifie plus.

3. Les limites concrètes de l'échantillon aléatoire simple :

Malgré ses bonnes propriétés et sa simplicité théorique, l'échantillon aléatoire n'est que rarement utilisé, et ne pourra nous servir seul dans le cadre d'enquêtes sur une population scolaire.
Pour s'en convaincre, il suffit de réaliser qu'un échantillon aléatoire simple de 400 élèves parmi 40 000 supposerait d'abord d'avoir une liste complète des 40 000 élèves de la population ; et qu'il soit pratiquement possible d'aller ensuite tester individuellement 400 élèves où qu'ils se trouvent, c'est-à-

dire probablement dans près de 400 écoles différentes. On voit bien que le coût serait important, et qu'il ne serait pas rentable de ne tester qu'un seul élève dans une école une fois qu'on a fait le déplacement.

L'échantillon aléatoire simple sert donc surtout de référence. On va lui comparer d'autres types d'échantillonnages plus complexes et plus adaptés aux réalités du terrain. Mais il sert aussi à comprendre intuitivement les différents paramètres qui affectent la précision d'une estimation par échantillonnage, et qui sont toujours un peu les mêmes : l'hétérogénéité de la population considérée, la taille absolue de l'échantillon et, dans une moindre mesure, sa taille relative à la population.

Nous sommes donc armés pour comprendre des types d'échantillonnages plus complexes. Commençons par envisager la stratification.

C. La stratification

1. Le principe

Le principe de la stratification est lui aussi assez simple.

Supposons qu'il existe dans une population **différents groupes assez hétérogènes entre eux**, et que ces groupes soient connus. **L'échantillon aléatoire simple présente le risque, dans un tirage donné, de ne pas représenter l'un ou l'autre de ces groupes. Partant, la moyenne obtenue ne sera pas représentative de ces groupes dont aucun membre n'a été tiré dans l'échantillon**. On risque de grosses erreurs d'échantillonnage dues à la non représentation de certains groupes. Et plus les groupes sont différents les uns des autres, plus la non représentation de l'un ou l'autre conduira à une évaluation imprécise de la moyenne de la population complète.

La stratification consiste à diviser la population en groupes d'individus dont on pense qu'ils se ressemblent pour les caractéristiques qui nous intéressent. C'est une partition de la population : chaque individu doit appartenir à un groupe et un seul. On procède alors à un échantillonnage distinct dans chacun de ces groupes. Cela permet de **contrôler la représentation de chacun des groupes dans l'échantillon final**. Par construction, cette stratégie exclut toute

une partie des échantillons possibles avec un tirage aléatoire simple, c'est-à-dire ceux où un (ou plusieurs) des groupes n'est pas représenté. Comme ces échantillons que l'on a écartés sont source d'erreurs d'estimation, l'opération permet d'obtenir des échantillons plus fiables.

On peut faire une stratification selon plusieurs critères (qui constituent les différents niveaux de stratification) : un premier niveau peut distinguer les régions socio-économiques d'un pays ; au sein de chacune de ces régions, on peut distinguer milieu urbain et milieu rural (deuxième niveau de stratification), etc. **L'important est d'avoir de bonnes raisons de croire que ces critères de stratification répartissent les individus dans des sous-groupes plus homogènes que le groupe de départ.**

Avant de montrer les propriétés d'un échantillonnage stratifié, prenons un exemple simplifié pour bien mettre en évidence l'intuition qui se trouve derrière les propriétés statistiques :

2. Un exemple simplifié

Soit un pays comportant deux régions A et B avec chacune deux écoles avec une classe de CM1 chacune. On veut évaluer le niveau moyen des maîtres de CM1 de l'ensemble du pays. Notre population comporte donc seulement quatre individus (les maîtres). On a prévu un test pour ces maîtres, mais on n'a le courage de corriger les copies que de deux d'entre eux. Quelle procédure d'échantillonnage choisir ?

L'idée d'utiliser un échantillon stratifié vient lorsqu'on se souvient que la région A est pauvre, au climat rigoureux, en somme peu attrayante, et que les maîtres chevronnés n'ont qu'un souci : la quitter. Au contraire, la région B attire les maîtres les plus qualifiés. Comme les affectations se font suivant une logique de mérite, on pense que tous les maîtres qualifiés se trouvent dans la région B.
On choisit alors de procéder par strates : dans la région A, on tire un maître au sort sur deux. On fait de même dans la région B. L'opération est-elle bénéfique ?

Le tableau indique les niveaux des différents maîtres de la population et les tirages au sort possible selon qu'on utilise le tirage aléatoire simple ou l'échantillonnage stratifié :

Tirages possibles avec ou sans stratification:

Population totale					
	Région A		Région B		Niveau moyen
	Maître 1 : Score=20	Maître 2 : Score=25	Maître 3 : Score=70	Maître 4 : Score=65	**45**
Tirages aléatoires simples possibles					
Tirage 1	OUI	OUI	-	-	22,5
Tirage 2	OUI	-	OUI	-	45
Tirage 3	OUI	-	-	OUI	42,5
Tirage 4	-	OUI	OUI	-	47,5
Tirage 5	-	OUI	-	OUI	45
Tirage 6	-	-	OUI	OUI	67,5
Moyenne des estimations : **45**					
Dispersion des estimations (variance) : **170,8**					
Tirages stratifiés possibles					
Tirage 2	OUI	-	OUI	-	45
Tirage 3	OUI	-	-	OUI	42,5
Tirage 4	-	OUI	OUI	-	47,5
Tirage 5	-	OUI	-	OUI	45
Moyenne des estimations : **45**					
Dispersion des estimations (variance) : **3,1**					

Les deux types d'échantillonnage mènent en moyenne, s'ils sont répétés, à la vraie valeur, c'est-à-dire 45 points. Mais dans le cas de l'échantillonnage aléatoire simple, on a une chance sur trois de faire une grosse erreur d'échantillonnage : les tirages 1 et 6 sont tous deux à 22,5 points de la vraie valeur. A contrario, l'échantillonnage stratifié exclut ces deux tirages, et tous les tirages possibles (les tirages 2, 3, 4 et 5) donnent des estimations relativement proches de la vraie valeur. Cette différence de précision des deux types d'échantillon est traduite par la variance beaucoup plus élevée de l'estimateur fondé sur un tirage aléatoire simple.

Ce tableau montre clairement le bénéfice qu'il y a à choisir un échantillonnage stratifié : on évite les pires erreurs d'échantillonnage. On peut pressentir que le gain en fiabilité est d'autant plus élevé que les régions A et B diffèrent par les caractéristiques de leurs maîtres.

Il suffit pour le voir d'envisager le cas où les niveaux des maîtres seraient respectivement 10 et 15 pour la région A, et 80 et 75 pour la région B. Les deux types d'échantillonnage resteraient bien sûr centrés autour de la vraie