

## FICHE CONSEIL NUMÉRO 5

### TECHNIQUES D'ÉCHANTILLONNAGE



L'échantillonnage permet aux statisticiens puis aux marketeurs de tirer des conclusions au sujet d'un tout, en n'en examinant qu'une partie. Les chercheurs ne s'intéressent pas à l'échantillon lui-même, mais à ce qu'il est possible d'apprendre à partir de l'enquête et à la façon dont on peut appliquer cette information à l'ensemble de la population. A la différence d'un recensement où tous les sujets de la population sont « examinés », dans l'échantillonnage, une partie des sujets de la population est étudiée.

→ Plusieurs échantillons peuvent être constitués

→ L'échantillon en lui-même n'est pas intéressant, ce sont les conclusions sur la population que l'on peut tirer de son observation qui en font l'intérêt : c'est l'**inférence**.

### SÉLECTION D'UN ÉCHANTILLON

#### Étapes pour sélectionner un échantillon

##### 1. Établir les objectifs de l'enquête

###### Définir la population cible

C'est la population totale pour laquelle on a besoin de l'information. Il faut définir les unités qui composent la population sous forme de caractéristiques l'identifiant (nature des données, emplacement géographique, dates ou encore critères sociodémographiques)

###### Déterminer les données à recueillir

(définition des termes, libellé des questions, définitions des méthodes de mesures, s'assurer que les exigences de l'enquête seront respectées sur le plan opérationnel)

##### 4. Fixer le degré de précision

Il y a un degré d'incertitude associé aux estimations établies à partir d'un échantillon qui dépend notamment de la méthode d'échantillonnage et de la taille de l'échantillon. **Quel degré peut-on accepter ?** Il faut établir un compromis entre le degré d'incertitude et le budget disponible pour l'enquête

##### ■ La taille de l'échantillon

Est souvent un compromis entre le degré de précision à atteindre et le budget de l'enquête mais aussi d'autres contraintes opérationnelles comme le temps disponible

Repose notamment sur :

La variabilité des caractéristiques que l'on mesure

La taille de la population (attention, ce n'est pas proportionnel)

La méthode d'échantillonnage

**Attention** : La population observée est différente de la population cible (la population cible est la population que nous **voulons observer**, tandis que la population observée est la population que nous **pouvons observer**) et les conclusions ne s'appliqueront qu'à la population réellement observée. L'utilisateur des résultats doit en être informé.



**Vocabulaire :**

La **base de sondage** est l'outil qu'on utilise pour avoir accès à la population. Le choix de la base de sondage aura des répercussions sur la sélection de la population observée. Par exemple, si on utilise une liste de numéros de téléphone pour sélectionner un échantillon de ménages, tous les ménages n'ayant pas le téléphone seront alors exclus de la population observée.

L'**unité d'échantillonnage** : fait partie de la base de sondage, peut être ou non sélectionnée

L'**unité déclarante** : fournit l'information qu'exige l'enquête.

L'**unité d'analyse ou de référence** : c'est l'unité au sujet de laquelle l'information est fournie  
Exemple : enquête sur les nouveau-nés

Unité d'échantillonnage : Ménage

Unité déclarante : L'un des deux parents ou le tuteur

Unité d'analyse : Le nouveau-né

Il existe deux types de méthodes d'échantillonnage : L'échantillonnage probabiliste et l'échantillonnage non probabiliste. La différence entre les deux tient au fait que dans le cas de l'échantillonnage probabiliste chaque unité a une « chance » d'être sélectionnée et que cette chance peut être quantifiée, ce qui n'est pas vrai pour l'échantillonnage non probabiliste; dans ce cas, chaque unité incluse à l'intérieur d'une population n'a pas une chance égale d'être sélectionnée.

**LES MÉTHODES ALÉATOIRES (PROBABILISTES)**

L'échantillonnage probabiliste entraîne la sélection d'un échantillon à partir d'une population, sélection qui repose sur le principe de la randomisation (la sélection au hasard ou aléatoire) ou la chance. Il est plus complexe, prend plus de temps et est habituellement plus coûteux que l'échantillonnage non probabiliste.

**L'échantillonnage aléatoire simple****■ Principe**

Il consiste à choisir des individus de telle sorte que chaque membre de la population a une chance égale de figurer dans l'échantillon. Ce choix peut se faire avec remise ou sans remise : avec remise, un individu peut être choisi plusieurs fois ; sans remise, un individu déjà choisi ne peut l'être de nouveau. C'est le cas habituel.

**■ Méthode**

Numéroter tous les individus de la liste correspondant aux individus de la population avec des nombres comportant un même nombre de chiffres. Puis utiliser une table de nombres aléatoires, une calculatrice ou un programme informatique, pour obtenir des nombres aléatoires comportant le nombre de chiffres désiré.

Enfin, sélectionner les nombres qui coïncident avec la liste. On rejette les nombres qui ne coïncident pas avec la liste ou qui se répètent, on s'arrête après avoir sélectionné n individus (n représentant le nombre d'individus souhaités dans l'échantillon).

**Avec Excel**

- Première colonne : identifie avec un nombre chaque individu de la liste de référence.
- Deuxième colonne : =alea()
- Recopier les deux colonnes en valeur à la même place.
- Trier les deux colonnes en fonction de l'ordre croissant (ou décroissant) de la deuxième colonne.
- Retenir les n premiers individus dans la colonne 1

**Combien peut-on réaliser d'échantillons ?**

Si l'on note  $n$  la taille de l'échantillon et  $N$  la taille de la population.

Avec remise :

$$N^n$$

Sans remise

$$C_N^n = \frac{N!}{n!(N-n)!}$$

■ **Avantage** de cette méthode : On peut espérer un échantillon «représentatif » puisque la méthode donne à chaque individu de la population une chance égale.

■ **Inconvénients**: la méthode n'est applicable que lorsqu'il existe une liste exhaustive de toute la population.

**L'échantillonnage systématique****■ Principe**

L'échantillonnage systématique est une méthode qui exige aussi l'existence d'une liste de la population où chaque individu est numéroté de 1 jusqu'à  $N$ .

Notons  $n$ , le nombre d'individus que doit comporter l'échantillon (la taille de l'échantillon). L'entier voisin de  $N/n$  sera noté «  $r$  » et appelé « raison » de sondage ou « pas » de sondage.

**■ Méthode**

Choisir au hasard un entier naturel  $d$  entre 1 et  $r$  (cet entier sera le point de départ). L'individu dont le numéro correspond à  $d$  est le premier individu, Pour sélectionner les autres, il suffit d'ajouter à  $d$  la raison de sondage : les individus choisis seront alors ceux dont les numéros correspondent à

$d + r$   
 $d + 2r$   
 $d + 3r$   
 etc.

■ **Avantages** : facile à sélectionner parce qu'un seul individu est choisi au hasard. On peut obtenir une bonne précision parce que la méthode permet de répartir l'échantillon dans l'ensemble de la liste.

■ **Inconvénients** : Les données peuvent être biaisées à cause de la périodicité.

**■ Exemple**

Étude des déplacements en autobus sur 365 jours en prenant un échantillon de taille 60. ( $N=365$  jours et  $n=60$ ).

**■ Remarques**

On a une population de 400 individus, on veut un échantillon de 100 individus

$$R = 4$$

On a donc que 4 échantillons possibles

1, 5, 9, .... 397  
 2, 6, 10, ... 398  
 3, 7, 11, ....399  
 4, 8, 12, ... 400





Si la population est distribuée au hasard dans la base de sondage, un échantillonnage systématique donnera des résultats similaires à ceux d'un échantillonnage aléatoire simple. Cette méthode est très utilisée dans les contrôles de qualité. L'échantillonnage avec une probabilité proportionnelle à la taille. Si la base de sondage renferme de l'information sur la taille de chaque unité (comme le nombre de médecins d'un hôpital) et si la taille de ces unités varie, on peut utiliser cette information pour accroître l'efficacité de l'échantillonnage. Plus la taille de l'unité est grande, plus sa chance d'être incluse dans l'échantillon est élevée.

### L'échantillonnage stratifié

#### ■ Principe

1. On subdivise la population en strates (groupes relativement homogènes) qui sont mutuellement exclusives
2. Proportionnellement à son importance dans la population, on calcule combien il faut d'individus au sein de l'échantillon pour représenter chaque strate.
3. Dans chacune des strates, on choisit au hasard le nombre nécessaire d'individus

*Les variables de stratification doivent être simples à utiliser, facile à observer et étroitement reliées au thème de l'enquête.*

#### ■ Avantages

Il est peu probable de choisir un échantillon absurde puisqu'on s'assure de la présence proportionnelle de tous les divers sous-groupes composant la population.

#### ■ Inconvénients

La méthode suppose l'existence d'une liste de la population. Il faut aussi connaître comment cette population se répartit selon certaines strates.

#### ■ Exemple

Choisir par échantillonnage stratifié 10 élèves dans un groupe de 60, en tenant compte du fait que 50% d'entre eux sont en CP, 30% en CE1 et 20% en CE2.

*La variance totale est la somme de la variance intrastrate et de la variance interstrate. On cherche à avoir la plus petite variance intrastrate et une grande variance interstrate*

#### ■ Estimation

Echantillonnage aléatoire simple intrastrate

Moyenne générale :	Précision
$\hat{\bar{Y}}_{ST} = \sum_{h=1}^H \frac{N_h}{N} * \bar{Y}_h$	$V(\hat{\bar{Y}}_{ST}) = \sum_{h=1}^H \left( \frac{N_h}{N} \right)^2 * (1 - f_h) * \frac{S_h^2}{n_h}$
H = Nombre de strates	fh = taux de sondage dans la strate h nh = taille de l'échantillon de la strate h S <sub>2h</sub> = dispersion vraie au sein de la strate h



■ **Application numérique**

Tranche de taille	N <sub>h</sub>	Y <sub>h</sub> (moyenne)	S <sub>h</sub> <sup>2</sup>	nh
0-9	500	5	1,5	130
10-19	300	12	4,0	80
20-49	150	30	8,0	60
50-499	100	150	100,0	25
500 et plus	10	600	2 500,0	5

On dispose de 1060 hôpitaux. On s'intéresse au nombre moyen Y de médecins par hôpital. La population est définie par 5 strates par tranches de taille en fonction du nombre de médecins. Cette information est obtenue à partir de

documents de l'APHP ne donnant pas le nombre exact de médecins mais seulement la tranche de taille. Réalisant un sondage aléatoire simple dans chaque strate h selon un budget permettant d'enquêter globalement 300 hôpitaux, on mesure y<sub>h</sub> et la dispersion S<sub>h</sub><sup>2</sup> de la variable nombre de médecins dans l'échantillon des hôpitaux tirés. Les allocations par strates sont données dans la dernière colonne du tableau.

**Quel est l'estimateur de Y, et quelle est sa précision ?**

Tranche de taille	N <sub>h</sub>	Y <sub>h</sub> (moyenne)	S <sub>h</sub> <sup>2</sup>	nh	Y <sub>h</sub> *nh	Terme de la variance de la moyenne
0-9	500	5	1,5	130	2 500	0,002
10-19	300	12	4,0	80	3 600	0,003
20-49	150	30	8,0	60	4 500	0,002
50-499	100	150	100,0	25	15 000	0,027
500 et plus	10	600	2 500,0	5	6 000	0,022
Total	1 060			300	31 600	0,055
				<b>Y =</b>	<b>29,8</b>	
				<b>Var Y =</b>	<b>0,055</b>	
				<b>ET Y</b>	<b>0,235</b>	
				<b>BS IC 95%</b>	<b>30,3</b>	
				<b>BI IC 95%</b>	<b>29,4</b>	

**Quelle serait l'allocation proportionnelle ?**

Tranche de taille	N <sub>h</sub>	nh	Allocation proportionnelle
0-9	500	130	142
10-19	300	80	85
20-49	150	60	42
50-499	100	25	28
500 et plus	10	5	3

Tranche de taille	N <sub>h</sub>	Y <sub>h</sub> (moyenne)	S <sub>h</sub> <sup>2</sup>	nh	Y <sub>h</sub> *nh	Terme de la variance de la moyenne
0-9	500	5	1,5	142	2 500	0,002
10-19	300	12	4,0	85	3 600	0,003
20-49	150	30	8,0	42	4 500	0,003
50-499	100	150	100,0	28	15 000	0,023
500 et plus	10	600	2 500,0	3	6 000	0,056
Total	1 060			300	31 600	0,086
				<b>Y =</b>	<b>29,8</b>	
				<b>Var Y =</b>	<b>0,086</b>	
				<b>ET Y</b>	<b>0,293</b>	
				<b>BS IC 95%</b>	<b>30,4</b>	
				<b>BI IC 95%</b>	<b>29,2</b>	





## L'échantillonnage par grappes

### ■ Principe

Dans les méthodes précédentes, l'unité statistique était choisie individuellement.

La technique de l'échantillonnage en grappes entraîne la division de la population en groupes ou grappes.

On sélectionne au hasard un certain nombre de grappes (unités primaires) pour représenter la population. Puis on sélectionne tous les individus des grappes choisies

### ■ Avantages

La méthode ne nécessite pas une liste globale de la population puisque seules les individus inclus dans les grappes comptent. Elle permet de limiter l'échantillon à des groupes compacts ce qui permet de réduire les coûts de déplacement, de suivi et de supervision.

### ■ Inconvénients

La méthode peut entraîner des résultats imprécis (moins précis que les méthodes précédentes) puisque les unités voisines ont tendance à se ressembler. Elle ne permet pas de contrôler la taille finale de l'échantillon.

## L'échantillonnage à plusieurs degrés

### ■ Principe

Ressemble à l'échantillonnage en grappes, sauf que dans ce cas on prélève un échantillon à l'intérieur de chaque grappe.

Ce qui implique au moins deux degrés (mais cela peut être plus) d'échantillonnage : on identifie au premier degré les grandes grappes (unités primaires), puis au second degré, à l'intérieur de chaque grappe, on sélectionne les unités (unités secondaires) qui vont faire partie de l'échantillon.

### ■ Avantage

L'échantillon est plus concentré ce qui réduit les coûts, pas besoin de disposer de la liste de toutes les unités. La méthode permet de contrôler la taille de l'échantillon notamment par stratification.

■ **Inconvénient** : précision des résultats, taille plus grande que dans le cas d'un échantillonnage aléatoire simple.

## L'échantillonnage à plusieurs phases

### ■ Principe

Les données de base sont collectées auprès d'un échantillon d'unité de grande taille, ensuite pour un sous-échantillon de ces unités, la collecte des données est plus détaillée.

Le plus couramment on utilise deux phases ou échantillonnage double. A première phase consiste donc à « filtrer » le premier échantillon par le biais d'un questionnaire par exemple.

L'échantillonnage à plusieurs phases est assez différent de l'échantillonnage à plusieurs degrés, malgré les similarités entre eux sur le plan de leur appellation.

L'échantillonnage à plusieurs phases est utile lorsqu'il manque à l'intérieur de la base de sondage des données auxiliaires qui pourraient servir à stratifier la population ou à rejeter à la sélection une partie de la population et lorsqu'on dispose d'un budget insuffisant pour recueillir des données auprès de l'échantillon entier (ou lorsque recueillir des données auprès de l'échantillon entier imposerait un fardeau excessif aux répondants).

## MÉTHODES NON ALÉATOIRES



On oppose aux méthodes aléatoires les méthodes non aléatoires ou empiriques.

Dans le cas de l'échantillonnage probabiliste, chaque unité a une chance d'être sélectionnée. Dans celui de l'échantillonnage non probabiliste, on suppose que la distribution des caractéristiques à l'intérieur de la population est égale. C'est ce qui fait que le chercheur croit que n'importe quel échantillon serait représentatif et que les résultats, par conséquent, seront exacts. Pour l'échantillonnage probabiliste, la randomisation est une caractéristique du processus de sélection, plutôt qu'une hypothèse au sujet de la structure de la population.

Dans le cas de l'échantillonnage non probabiliste, puisqu'on choisit arbitrairement des unités, il n'existe aucune façon d'estimer la probabilité pour une unité quelconque d'être incluse dans l'échantillon. Également, comme la méthode en question ne fournit aucunement l'assurance que chaque unité aura une chance d'être incluse dans l'échantillon, on ne peut estimer la variabilité de l'échantillonnage ni identifier le biais possible.

On ne peut mesurer la fiabilité d'un échantillonnage non probabiliste; la seule façon de mesurer la qualité des données en résultant consiste à comparer certains des résultats de l'enquête à l'information dont on dispose au sujet de la population. Encore une fois, rien ne fournit l'assurance que les estimations ne dépasseront pas un niveau acceptable d'erreur. Les statisticiens hésitent à utiliser les méthodes d'échantillonnage non probabiliste, parce qu'il n'existe aucun moyen de mesurer la précision des échantillons en découlant.

Elles sont souvent utilisées

- pour des études exploratoires;
- pour réduire les coûts;
- quand il est impossible ou non envisageable d'utiliser la méthode aléatoire.

### On distingue :

- l'échantillonnage à l'aveuglette ou de commodité : Ex.: déguster un échantillon de vin.
- l'échantillonnage de volontaires : Ex : expériences médicales ou psychologiques.
- l'échantillonnage au jugé : cette méthode implique la sélection d'individus en fonction de l'idée qu'on se fait de la composition de la population. On le fait pour des essais auprès des groupes cibles.
- l'échantillonnage par quotas : il est largement utilisé dans les enquêtes d'opinion et les études de marché notamment parce qu'il ne suppose pas de liste des individus de la population. On parle aussi d'échantillonnage dirigé ou par choix raisonné. On demande aux enquêteurs de faire un nombre d'entrevues dans divers groupes établis en fonction du secteur géographique, de l'âge, du sexe ou d'autres caractéristiques... L'enquêteur doit respecter son quota.

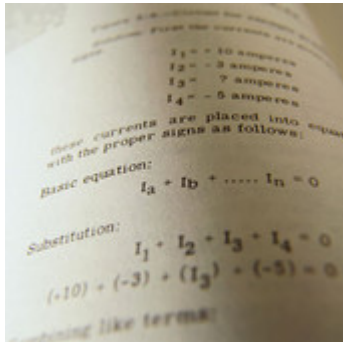
■ **Avantages** : Moins coûteuse et plus facile à réaliser.

■ **Inconvénients**: Beaucoup de non-réponses; difficulté de trancher lorsqu'il s'agit de sélectionner des individus d'un groupe d'âge ouvert (Ex : 65 ans et plus : faut-il prendre 66 ans, 70 ans ...).



## LES ERREURS

---



Les méthodes d'échantillonnage peuvent être sources d'erreurs. Un certain nombre d'erreurs pourront être éliminées, certaines pourront être réduites, mais d'autres persisteront.

### ■ Les erreurs dues aux instruments de mesure

Un instrument est fidèle s'il répond exactement de la même façon quand il est placé dans deux situations identiques. Exemple le thermomètre. Une question claire est dite fidèle quand tout le monde la comprend de la même façon.

Un instrument est valide lorsqu'il mesure vraiment ce qu'il est censé mesurer.

### ■ Les erreurs dues à l'organisation

Ce sont les erreurs qui se glissent lors de la collecte des données.

Est-ce que les consignes ont été respectées?

Les enquêteurs ont-ils agi de la même façon?

Pour éviter ces erreurs il faut utiliser les mêmes instruments, les mêmes conditions.

### ■ Les erreurs dues à la méthode d'échantillonnage

Il faut toujours vérifier, à la lumière des objectifs de l'étude statistique, que la méthode d'échantillonnage est adaptée, en particulier, éviter la sur-représentation de certaines parties de la population.

### ■ Les erreurs dues au phénomène de non-réponse

Même avec la meilleure méthode d'échantillonnage, il se présente toujours un certain nombre de non-répondants, ce qui peut entacher la représentativité de l'échantillon et amener des conclusions erronées.

### ■ L'erreur d'échantillonnage

Le fait d'étudier un échantillon plutôt qu'un autre engendre forcément une erreur. Cette erreur appelée erreur d'échantillonnage est inévitable.