## Chapitre 3 : Sondage Stratifié

1. Principe :

Le principe de la **stratification** est de découper la population en sous-ensembles appelés **strates** et réaliser un sondage dans chacun de ces ensembles.

L’intérêt de cette méthode est d’améliorer la précision des estimateurs. Ceci nécessite l’utilisation d’une information auxiliaire connue pour l’ensemble de la population.

Par **exemple**, si nous souhaitons estimer l’âge moyen d’une population composée d’étudiants et d’enseignants, la variable âge ne se comporte pas de la même manière dans ces deux classes. On peut penser que la population enseignant est plus âgée que la population étudiante. Cette information (auxiliaire) sera donc utile et sera prise en compte dans le plan de sondage.

L’information auxiliaire peut être utilisée à deux moments :

* A L’étape de la conception du plan de sondage.
* A l’étape de l’estimation des paramètres.

Nous allons considérer, dans le cas de ce cours, l’utilisation de cette information pour bâtir le plan de sondage.

1. Plan de sondage stratifié :

Soit une population de individus et la valeur du caractère mesurée sur le individu, . On cherche à estimer la moyenne de la population .

(Nous pouvons étendre les concepts à l’estimation d’un total ou d’une proportion comme dans le chapitre 2).

On suppose que la population est partagée en sous-ensembles ou strates notées . i.e., est une partition de .

On définit :

* La taille de la strate h : ;
* La moyenne de la strate : ;
* La variance de la strate : ;
* La variance corrigée de la strate : ;
  + - Proposition 1 :

On peut écrire la moyenne et la variance de la population en fonction des moyennes et des variances des strates, i.e.

Et

En effet,

Et

Le premier terme représente la moyenne des variances des strates et le second terme est dû aux différences entre strates.

* + - Définition

Un plan de sondage est dit **stratifié** si dans chaque strate on sélectionne un échantillon aléatoire de taille ﬁxe , tel que et que les sélections sont réalisées **indépendamment** d’une strate à une autre. Dans ce cas, est un échantillon de individus de la population .

On supposera pour la suite qu’au sein de chaque strate les plans sont simples et sans remise (PESR).

Dans ce contexte, il y a échantillons possibles.

**Remarque**: Ce sont les plans de sondage de type ST qui sont classiquement utiliséspour les enquêtes auprès des entreprises.

**Exemple :**

Considérons la stratification de la population de 20 individus d’une université suivant : étudiant () et enseignant (), dont 12 étudiants (strate 1, ) et 8 enseignants (strate 2, ).

Le nombre total d’individus est

On effectue un plan de sondage stratifié et on sélectionne un échantillon aléatoire de taille de la manière suivante : et (voir la figure Fig. 1 ci-dessous).

Fig. 1: Echantillonnage stratifié : à gauche la population, à droite l’échantillon.

**Notations utilisées**

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Population | Echantillon |
|  | Taille |  |  |
|  | Moyenne |  |  |
| Total | Variance |  |  |
|  | Variance corrigée |  |  |
| Votre texte ici |  |  |  |
|  | Taille |  |  |
|  | Moyenne |  |  |
| Strate | Variance |  |  |
|  | Variance corrigée |  |  |
|  |  |  |  |

Tab 1 : Notations pour le plan stratifié.

1. Estimateur de la moyenne :
   * + Définition

Considérons un plan de sondage stratifié à strates et soit la moyenne calculée sur l’échantillon issu de la strate , .

Alors, l’estimateur de la moyenne de la population est défini par

,

et est l’échantillon de la strate .

**Exemple** :

Soit pour , l’âge du individu présent dans l’échantillon .

On note où contient les étudiants de l’échantillon et contient les enseignants de l’échantillon.

On calcule l’âge moyen des individus de l’échantillon dans les strates 1 et 2, respectivement.

* Âge moyen des individus de la strate 1 ;
* Âge moyen des individus de la strate 2 ;

est un estimateur de l’âge total de la population des étudiants et est un estimateur de l’âge total de la population des enseignants.

D’où, est un estimateur de l’âge total de toute population de l’université et

est un estimateur de l’âge moyen de la population .

**Application numérique :**

Les résultats du sondage sont présentés dans le tableau suivant :

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Strate | 1 | 1 | 1 | 2 | 1 | 2 | 2 | 1 | 1 | 2 |
| Age | 20 | 25 | 23 | 50 | 26 | 42 | 45 | 21 | 22 | 48 |

Tab 1 : Age des individus sondés

On calcule la moyenne des âges des individus de l’échantillon par strate :

et

Une estimation de µ est donc :

* + - Propriétés de l’estimateur

L’estimateur de la moyenne , pour un plan stratifié, vérifie les propriétés suivantes :

* ;
* (3.1)

Un estimateur aléatoire sans biais de est le réel

1. Le choix de

Il s’agit de s’intéresser à la question suivante :

Combien de personnes doit-on sonder par strate pour que l’estimateur soit le plus précis possible ?

Pour répondre à cette question on applique l’un des plans suivants.

* + 1. Plan avec allocation proportionnelle (STP)
* Définition

Dans un plan stratifié avec allocation proportionnelle (STP), on choisit les de telle sorte que la proportion d’individus provenant de la strate dans l’échantillon soit la même que dans la population, c’est-à-dire :

D’où,

Dans ce cas, les strates ont dans l’échantillon des poids égaux à leurs poids dans la population. Ceci revient à appliquer le même taux de sondage dans toutes les strates, i.e.,

**Attention :** Cette procédure ne donne généralement pas de résultat entier. Il faut alors recourir à une procédure d’arrondi (et vérifier que l’on a toujours ).

En pratique, pour tout on prend le plus petit entier tel que

Si on a , on ajuste en ajoutant ou en enlevant une unité pour les échantillons les plus nombreux.

**Exemple** :

Dans l’exemple de l’âge moyen de la population des étudiants et enseignants de l’université, un tel plan signifie que les proportions de chaque strate dans la population sont les mêmes que dans l’échantillon. On considère la répartition suivante :

|  |  |
| --- | --- |
| Strate |  |
| Etudiants | 6000 |
| Enseignants | 4000 |

Alors, un plan stratifié avec allocation proportionnelle de taille consistera à sonder

* étudiants,
* enseignants.
* Proposition



24

Soit l’estimateur construit pour un plan avec allocation proportionnelle (STP). On a alors

(3.2)

**Remarque :**

La stratification avec allocation proportionnelle donne presque toujours de meilleurs résultats qu’un plan de simple puisque l’on supprime la variance inter-strate dans l’expression de la variance de l’estimateur : Si les tailles de chaque strate sont grandes, on a d’où

Avec . Les résultats seront d’autant plus satisfaisants lorsque la variance inter-strate est grande. Celle-ci est grande quand la variable de stratification est fortement liée à la variable d’intérêt. C’est pourquoi il faut toujours stratifier avec une variable très dépendante de la variable d’intérêt.

i.e.

Si et sont suffisamment grands, on peut montrer que

**Exemple :**

Considérons le tableau suivant dans lequel on donne pour chaque individu de l’université

* Son âge
* Sa catégorie : 1 si étudiant, 2 si enseignant
* Sa couleur de cheveux : a si brun, b si blond, c si châtain

On considère une population de 16 individus

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Âge | 24 | 52 | 19 | 26 | 45 | 23 | 39 | 24 | 22 | 48 | 24 | 26 | 46 | 23 | 39 | 18 |
| Catg | 1 | 2 | 1 | 1 | 2 | 1 | 2 | 1 | 1 | 2 | 1 | 1 | 2 | 1 | 2 | 1 |
| Coulchev | c | a | c | b | c | a | a | b | c | a | a | b | b | c | a | c |

1. On souhaite estimer la moyenne à l’aide d’un plan simple. La variance de l’estimateur est
2. On désire stratifier suivant la catégorie. La population est divisée selon la figure suivante

26

23

24

22

24

26

23

18

24

52

45

39

48

46

39

Les moyennes et variances corrigées par strates sont

* et
* et

D’où la variance de l’estimateur

1. On choisit maintenant de stratifier suivant la couleur des cheveux. La population est divisée selon la figure suivante

24

23

24

26

45

46

18

52

48

23

22

26

39

24

19

39

En raisonnant de la même façon que la question précédente, on obtient que la variance de l’estimateur, pour ce plan de sondage, vaut  3.6077.

En comparant les variances de l’estimateur

|  |  |
| --- | --- |
| Plan |  |
| Simple  Strate Catg  Strate Coulchev | 4.68  1.78  3.6077 |

On voit que les deux plans stratifiés possèdent des variances inférieures au plan simple. Le gain de la stratification par la catégorie est significatif comparé à celui de la couleur des cheveux. Ceci vient du fait que la variable d’intérêt (âge) dépend plus de la catégorie que de la couleur de cheveux. Il sera donc beaucoup plus pertinent de stratifier par rapport à la catégorie que par rapport à la couleur de cheveux.

D’où, le plan avec allocation proportionnelle est plus précis que le plan simple, en terme de variance de l’estimateur.

* + 1. Plan avec allocation optimale (STO)
* Définition

Dans un plan stratifié avec allocation optimale, on choisit les tailles d’échantillons telles que et telles que la variance de l’estimateur soit minimale.

La solution de ce problème est

En pratique, pour tout on prend le plus petit entier tel que

Il dépend ainsi de la taille de strate et de la dispersion des valeurs de Y dans la strate

* Si , alors on prend et on calcule les autres tailles sans prendre en compte l’échantillon
* Si on ajuste en enlevant une unité pour les échantillons les plus nombreux.

Par définition, parmi tous les plans stratifiés, l’estimateur construit avec un plan d’allocation optimale (STO) possède la plus petite variance possible.

i.e.

Pour construire un tel estimateur, il nous faut connaître la variance corrigée du caractère dans chaque strate de la population.

La variance de l’estimateur associé au plan (STO) est calculée par (3.1). La formule (3.2) est valable uniquement pour un plan avec allocation proportionnelle (STP).

**Remarque :**

Si on dispose d’une information permettant la stratification de la population, on a tout intérêt à l’utiliser pour améliorer l’estimation de . Le plan de sondage aléatoire de type STO donne les meilleurs résultats.

* Intervalle de confiance

Pour les estimateurs construits par plans stratifiés, on peut calculer des intervalles de confiance comme pour les plans simples. Un intervalle de confiance de niveau 1 − α est donné par

Où est le quantile d’ordre de la loi normale centrée réduite.

**Exemple**

On s’intéresse à l’estimation de l’âge moyen d’un personnel d’une entreprise contenant 10000 personnes. Des études antérieures ont montré que la variable que l’on cherche à analyser est très contrastée selon les trois grandes catégories qui vont former les strates. Considérons le tableau d’informations suivant,

|  |  |  |
| --- | --- | --- |
| Catégories | Effectifs | Ecart-type des âges |
| 1 | 2000 | 18 |
| 2 | 3000 | 12 |
| 3 | 5000 | 3.6 |
| Ensemble | 10000 | 16 |

On désire estimer l’âge moyen du personnel à partir d’un échantillon de taille .

Si désigne l’âge de la personne interrogée alors l’estimateur de est

de variance , où

D’où,

On envisage d’appliquer un plan de sondage stratifié avec allocation proportionnelle, alors si est le nombre de personnes interrogés dans chaque strate alors l’estimateur de la moyenne

Où est l’âge moyen des personnes interrogées dans la strate . Les effectifs sont choisis tels que Par conséquent,

et

Les variances corrigées par strate sont aussi calculées, avec

et .

Pour un plan avec allocation optimale, les effectifs sont obtenus tels que , avec , on obtient

et

On arrondit :

et

en vérifiant que la somme fait 100. D’où la variance de la formule (3.1),

* Choix de la taille de l’échantillon

Soit un échantillon prélevé lors d’une étude préliminaire. La taille d’échantillon à choisir pour avoir une incertitude relative de au niveau 100(1−α)%, α ∈]0,1[, inférieure ou égale à (100×)% est le plus petit tel que ≤ . En particulier,

* Pour un plan de sondage aléatoire de type STP
* Pour un plan de sondage aléatoire de type STO