Introduction à la théorie des sondages - Cours 6

Thomas Merly-Alpa
thomas.merly-alpa@insee.fr

INSEE, département des méthodes statistiques

20 avril 2017

Organisation

- 6 cours (avec des TD en fil rouge)
- 2 intervenants :
  - Thomas Merly-Alpa - `thomas.merly-alpa@insee.fr`
  - Martin Chevalier - `martin.chevalier@insee.fr`
- Les slides et TD du cours sont à l'adresse `http://nc233.com/teaching` et sur moodle
- Notation : un devoir maison à rendre

Sommaire

Chapitre 1

Méthodes de redressement

Où en sommes-nous ?

La méthodologie d'Horvitz-Thompson permet d'obtenir un **estimateur sans biais** avec une **variance calculable**.

La **stratification** améliore l'estimation obtenue par sondage aléatoire simple en exploitant l'information auxiliaire de la base de sondage et **diminue la variance d'estimation**.

La **correction de la non-réponse** (imputation ou repondération) permet (dans la plupart des cas) de **neutraliser le biais** introduit par la non-réponse de certaines unités échantillonnées.

Ce qu'il reste à faire Améliorer encore l'estimateur en utilisant l'information auxiliaire **au moment de l'estimation**.

Objectifs des méthodes de redressement

1. Exploiter l'information auxiliaire qui n'a pas pu l'être au moment du tirage pour **améliorer la précision de l'estimateur**.
2. **Assurer la cohérence** entre les estimations produites par l'enquête et une ou plusieurs sources de référence.

En pratique Ajustement de l'estimateur d'Horvitz-Thompson. . .

- . . . pour garantir une **estimation parfaite** de certaines variables. . .
- . . . et ainsi **diminuer sa variance**. . .
- . . . tout en gardant le caractère **sans biais**.

Remarque Dans l'ensemble de cette partie, le plan de sondage est un sondage aléatoire simple et il n'y a pas de non-réponse.

Application : Enquête sur la fréquentation des cinémas

Le distributeur d'un film souhaite connaître le **nombre d'entrées réalisées une semaine donnée**.

Habituellement des remontées sont effectuées tous les mois, mais il souhaite avoir une **information plus rapidement** pour ajuster sa campagne promotionnelle.

Pour ce faire, il interroge un **échantillon de 100 cinémas** (parmi les 2 020 exploitants en activité) tiré par sondage aléatoire simple.

La variable d'intérêt est le **nombre d'entrées réalisées par le film** pour la semaine du 20 au 27 février 2017.

L'estimateur d'Horvitz-Thompson obtenu est de **464 923** avec un **intervalle de confiance à 95 % de [243 061 ; 686 785]**.

Application : Enquête sur la fréquentation des cinémas

Le distributeur n'est **pas très satisfait** de cette fourchette extrêmement large.
Il envisage d'exploiter une information disponible quelques jours après l'enquête, le **nombre de projections du film** :

- sur l'ensemble de la France, le film a été projeté 5 061 fois ;
- à partir de l'échantillon, ce nombre est estimé à 3 333 fois.

Intuition

- Nombre de projections et nombre d'entrées étant **corrélées**, le distributeur pourrait être tenté de **redresser** l'estimateur du nombre d'entrées en le multipliant par $\frac{5061}{3333} = 1,52$.
- L'utilisation du nombre de projections comme information auxiliaire pourrait venir « **stabiliser** » l'estimateur.

Partie 1

Redressement par le ratio

Définition

L'estimateur par le ratio est utilisé quand la variable auxiliaire X est **quantitative**.

Sachant que le total de la variable auxiliaire $T(X)$ est connu, on définit l'**estimateur par le ratio du total de la variable** Y par :

$$\hat{T}_{ratio}(Y) = \hat{T}_{HT}(Y) \times \frac{T(X)}{\hat{T}_{HT}(X)}$$

Intuition Si $T(X) > \hat{T}_{HT}(X)$, l'estimateur par le ratio de Y est supérieur à l'estimateur d'Horvitz-Thompson.

Propriétés

1. Asymptotiquement sans biais :

$$B(\hat{T}_{ratio}(Y)) \xrightarrow[n\to+\infty]{} 0$$

2. Variance d'autant plus faible que Y est corrélée à X :

$$V(\hat{T}_{ratio}(Y)) \approx N^2 \left(1 - \frac{n}{N}\right) \frac{S_Y^2 + R^2 S_X^2 - 2RS_{X,Y}}{n}$$

avec $R = \dfrac{T(Y)}{T(X)}$

3. **Propriété de calage** T(X) est estimé parfaitement :

$$\hat{T}_{ratio}(X) = \hat{T}_{HT}(X) \times \frac{T(X)}{\hat{T}_{HT}(X)} = T(X)$$

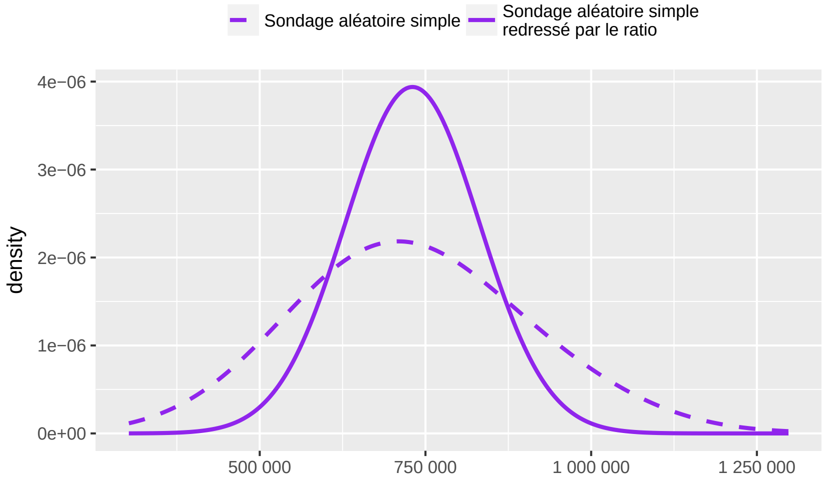
Application : Enquête sur la fréquentation des cinémas

Une fois les informations complètes sur le film remontées, le distributeur **évalue la pertinence d'un redressement par le ratio** en utilisant le nombre de projections comme variable auxiliaire.
Il tire 1 000 échantillons de taille 100 et calcule pour chacun la valeur de l'estimateur d'Horvitz-Thompson et celle de l'estimateur redressé par le ratio.

Exemple L'estimateur par le ratio associé au premier échantillon tiré donne :

$$\hat{T}_{ratio}(Y) = 464923 \times \frac{5061}{3333} = 705963$$

Application : Enquête sur la fréquentation des cinémas



Application : Enquête sur la fréquentation des cinémas

Valeur dans la population 731 892 entrées

Estimateur d'Horvitz-Thompson (1 000 simulations)

- moyenne empirique : 730 942
- écart-type empirique : 153 835

Estimateur redressé par le ratio (1 000 simulations)

- moyenne empirique : 730 299
- écart-type empirique : 16 039

Redressement par le ratio et repondération

L'estimation par le ratio peut être vue comme une repondération. En notant $d_k = \frac{1}{\pi_k}$ le poids de sondage de l'unité k, l'estimateur d'Horvitz-Thompson s'écrit en effet :

$$\hat{T}_{HT}(Y) = \sum_{k \in s} d_k y_k$$

Dès lors, on peut réécrire l'estimation par le ratio :

$$\hat{T}_{ratio}(Y) = \sum_{k \in s} d_k y_k \times \frac{T(X)}{\hat{T}_{HT}(X)} = \sum_{k \in s} \left(d_k \times \frac{T(X)}{\hat{T}_{HT}(X)} \right) \times y_k = \sum_{k \in s} w_k y_k$$

avec $\forall k \in s \quad w_k = d_k \times \dfrac{T(X)}{\hat{T}_{HT}(X)}$

Redressement par le ratio et repondération

En pratique Les redressements sont effectués **une fois pour toutes** au moment de la production d'une enquête. Un vecteur de **poids redressés** est ainsi produit et a vocation à être utilisé à la place des poids de sondage.

Partie 2

Redressement par post-stratification

Définition

L'estimateur post-stratifié est utilisé quand la variable auxiliaire X est **qualitative** (ou recodée en tranches).

On peut alors définir H groupes d'unités (les **post-strates**) selon les modalités de cette variables et calculer l'estimateur post-stratifié :

$$\hat{T}_{post}(Y) = \sum_{h=1}^{H} \hat{T}_{h,HT}(Y) \frac{N_h}{\hat{N}_{h,HT}}$$

où N_h est le nombre d'unités de la population dans la post-strate h et $\hat{N}_{h,HT}$ son estimateur à partir de l'échantillon.

Remarque Quand le plan de sondage est stratifié selon X, $\hat{N}_{h,HT} = N_h$ et donc $\hat{T}_{post}(Y) = \hat{T}_{HT}(Y)$.

Propriétés

1. Sans biais si tous les N_h **sont entiers**.
2. Variance supérieure à celle d'un SAS stratifié avec allocation proportionnelle ;

$$V(\hat{T}_{post}(Y)) \approx \underbrace{N^2 \frac{1-f}{n} \sum_{h=1}^{H} \frac{N_h}{N} S_h^2}_{\text{Variance d'un SAS stratifié avec alloc. proportionnelle}} + \underbrace{N^2 \frac{1-f}{n^2} \sum_{h=1}^{H} \frac{N-N_h}{N} S_h^2}_{\text{Variance supplémentaire due à la post-stratification}}$$

3. **Propriété de calage** La taille des H post-strates est estimée parfaitement :

$$\forall h = 1, \ldots, H \quad \hat{N}_{h,post} = N_h$$

Application : Enquête sur la fréquentation des cinémas

Le distributeur envisage également d'utiliser comme variable auxiliaire le fait que la zone dans laquelle sont situés les cinémas a été en **vacances scolaires** du 20 au 27 février.

Il constitue donc **deux post-strates** et les utilise pour redresser l'estimateur d'Horvitz-Thompson. À nouveau l'évaluation de la performance de ce redressement est effectuée sur 1 000 simulations.

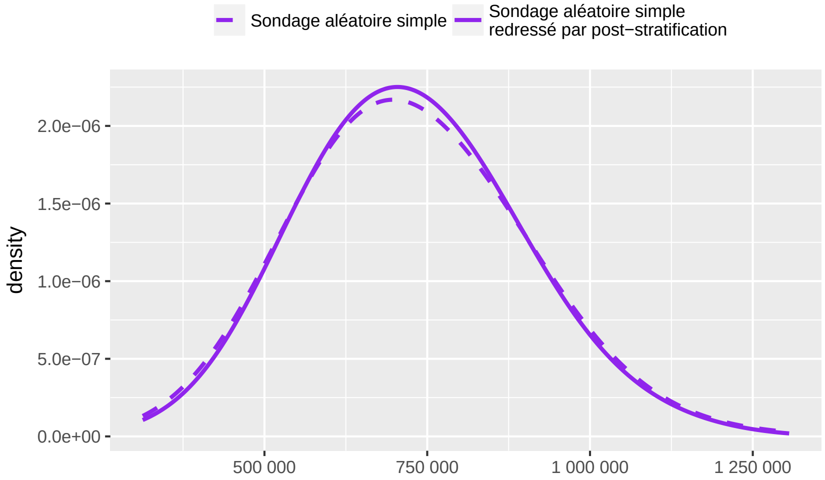
Application : Enquête sur la fréquentation des cinémas

Exemple À partir du tout premier échantillon, on estime :

- le nombre de cinéma dans une zone en vacances scolaires à 1 192 (contre 1 244 dans la population) ;
- le nombre de cinéma dans une zone non en vacances scolaires à 828 (contre 776 dans la population).

$$\hat{T}_{post}(Y) = 443915 \times \frac{1244}{1192} + 21008 \times \frac{776}{828} = 483042$$

Application : Enquête sur la fréquentation des cinémas



Application : Enquête sur la fréquentation des cinémas

Valeur dans la population 731 892 entrées

Estimateur d'Horvitz-Thompson (1 000 simulations)

- moyenne empirique : 725 513
- écart-type empirique : 153 752

Estimateur redressé par le ratio (1 000 simulations)

- moyenne empirique : 725 812
- écart-type empirique : 145 332

Post-stratification et repondération

Comme pour l'estimation par le ratio, il est possible de réécrire l'estimateur post-stratifié sous la forme d'une répondération :

$$\hat{T}_{post}(Y) = \sum_{h=1}^{H} \sum_{k \in s_h} d_k y_k \frac{N_h}{\hat{N}_{h,HT}} = \sum_{k \in s} \left(d_k \times \underbrace{\frac{N_h}{\hat{N}_{h,HT}}}_{h | k \in s_h} \right) y_k = \sum_{k \in s} w_k y_k$$

En pratique À nouveau, cette propriété permet de simplifier la mise en œuvre des redressements en calculant au moment de la production de l'enquête un **vecteur de poids redressés** à utiliser à la place des poids de sondage.

Application

TD 4 - Exercice 2

Partie 3

Généralisation : Calage sur marges

Redresser sur plusieurs variables simultanément

Le redressement par le ratio ou la post-stratification sont des méthodes simples et classiques pour utiliser de l'information auxiliaire au moment de l'estimation.

Néanmoins, elles présentent l'une et l'autre une limite principale : **elles ne peuvent intégrer l'information auxiliaire que d'une seule variable**.

Exemple On ne peut pas utiliser conjointement dans les redressements l'information sur le nombre de projections et les vacances scolaires.

Remarque Dans le cas de la post-stratification, une possibilité consiste à croiser les modalités de toutes les variables (qualitatives) que l'on souhaite utiliser, mais cela suppose d'avoir une **information auxiliaire sur leur distribution jointe**.

Calage sur marges : intuition et principe

Au moment de l'estimation on dispose des éléments suivants :

- pour chaque unité k de l'échantillon, un poids de sondage d_k ;
- p **variables de calage** formant la matrice $X = (x_1 \ x_2 \ \ldots \ x_p)$ et renseignées pour chaque unité k de l'échantillon ;
- la valeur du total **dans la population** des p variables de calage : $T(X) = (T(x_1) \ T(x_2) \ \ldots \ T(x_p))$

Intuition

- Utiliser les poids de sondage d_k **garantit une estimation sans biais**. . .
- . . . mais les modifier de façon à obtenir une estimation parfaite des marges de calage **améliore la précision des estimateurs**.

Principe du calage sur marges Trouver le **vecteur de poids calés** w_k qui conduise à **estimer parfaitement les marges de calage** et qui soit **le plus proche possible de** d_k.

Calage sur marges : formulation du problème

D'un point de vue mathématique, ce problème se formule de la façon suivante :

$$\begin{cases} \min_{w_k} \sum_{k \in s} d_k G\left(\frac{w_k}{d_k}\right) \\ \text{sous la contrainte } \sum_{k \in s} w_k X_k = T(X) \end{cases}$$

où G est une certaine **fonction de distance** entre les poids initiaux d_k et les poids finaux w_k :

- $G(1) = 0$;
- $G\left(\frac{w_k}{d_k}\right)$ est d'autant plus grand que $\frac{w_k}{d_k}$ est différent de 1.

Calage sur marges : fonction de distance et résolution

La résolution de ce problème fait intervenir la **fonction réciproque de la dérivée** de la fonction G, notée en général F.

La forme de la fonction F identifie la **méthode de calage** mise en œuvre, dont les propriétés diffèrent :

- méthode linéaire : $F(x) = 1 + x$
- méthode exponentielle (ou *raking ratio*) : $F(x) = exp(x)$
- méthode logistique : $F(x) = \dfrac{L(U-1) + U(L-1)exp(Au)}{U - 1 + (1-L)exp(Au)}$

  avec L et U des bornes pour $\dfrac{w_k}{d_k}$ et A une constante
- méthode linéaire tronquée : $F(x) = 1 + x$ pour $x \in [L; U]$

Dans tous les cas, la résolution s'appuie sur un **algorithme itératif**.

Illustration : *Raking ratio* sur deux variables dichotomiques

Identifiant	Sexe	Île-de-France	Poids de sondage d_k
A	H	Oui	10
B	H	Non	10
C	H	Non	10
D	F	Oui	10
E	F	Oui	10
F	F	Non	10

Marges dans la population

- $T(\text{Sexe} = \text{H}) = 20$
- $T(\text{Sexe} = \text{F}) = 40$
- $T(\text{Île-de-France} = \text{Oui}) = 40$
- $T(\text{Île-de-France} = \text{Non}) = 20$

Illustration : *Raking ratio* sur deux variables dichotomiques

Étape 1

Sexe \ IdF	Oui	Non	Marge	
H	10	20	30 (20)	$\times\ 20/30$
F	20	10	30 (40)	$\times\ 40/30$
Marge	30 (40)	30 (20)	60 (60)	

Étape 2

Sexe \ IdF	Oui	Non	Marge
H	6,67	13,33	20 (20)
F	26,67	13,33	40 (40)
Marge	33,34 (40)	26,67 (20)	60 (60)
	$\times\ 40/33{,}34$	$\times\ 20/26{,}67$	

Illustration : *Raking ratio* sur deux variables dichotomiques

Étape 3

Sexe \ IdF	Oui	Non	Marge	
H	8	10	18 (20)	$\times$ 20/18
F	32	10	42 (40)	$\times$ 40/42
Marge	40 (40)	20 (20)	60 (60)	

Étape 4

Sexe \ IdF	Oui	Non	Marge
H	8,88	11,11	20 (20)
F	30,48	9,52	40 (40)
Marge	39,36 (40)	20,63 (20)	60 (60)
	$\times$ 40/39,36	$\times$ 20/20,63	

Illustration : *Raking ratio* sur deux variables dichotomiques

Étape 5

9,03	10,77	19,80 (20)
30,97	9,23	40,20 (40)
40 (40)	20 (20)	60 (60)

Étape 6

9,12	10,88	20 (20)
30,81	9,19	40 (40)
39,93 (40)	20,07 (20)	60 (60)

Étape 7

9,14	10,84	19,98 (20)
30,86	9,16	40,02 (40)
40 (40)	20 (20)	60 (60)

Étape 8

9,15	10,85	20 (20)
30,85	9,15	40 (40)
40 (40)	20 (20)	60 (60)

Poids finaux

Tableau final

9,15	10,85	20 (20)
30,85	9,15	40 (40)
40 (40)	20 (20)	60 (60)

Détermination des poids finaux Multiplication du poids initial d_k par le rapport entre les totaux de chaque cellule après/avant l'algorithme de calage.

Exemple $d_B = 10$, $sexe_B =$ H et $idf_B =$ Non

- total final/initial de la cellule : 10,85/20
- poids final $w_b = 10 \times 10{,}85/20 = 5{,}425$

Illustration : *Raking ratio* sur deux variables dichotomiques

Identifiant	Sexe	Île-de-France	d_k	Poids calé w_k
A	H	Oui	10	9,150
B	H	Non	10	5,425
C	H	Non	10	5,425
D	F	Oui	10	15,425
E	F	Oui	10	15,425
F	F	Non	10	9,150

Vérification des contraintes de calage

- $\hat{T}(\text{Sexe} = \text{H}) = 9{,}150 + 5{,}425 + 5{,}425 = 20 = T(\text{Sexe} = \text{H})$
- $\hat{T}(\text{Sexe} = \text{F}) = 15{,}425 + 15{,}425 + 9{,}150 = 40 = T(\text{Sexe} = \text{F})$
- $\hat{T}(\text{Idf} = \text{Oui}) = 9{,}150 + 15{,}425 + 15{,}425 = 40 = T(\text{Idf} = \text{Oui})$
- $\hat{T}(\text{Idf} = \text{Non}) = 5{,}425 + 5{,}425 + 9{,}150 = 20 = T(\text{Idf} = \text{Non})$

Propriétés de l'estimateur obtenu par calage

1. Quelle que soit la méthode, asymptotiquement sans biais :

$$B(\hat{T}_{calage}(Y)) \xrightarrow[n \to +\infty]{} 0$$

2. Quelle que soit la méthode, variance **approximativement égale** et qui s'exprime en fonction d'un résidu :

$$V(\hat{T}_{calage}(y)) \approx V(\hat{T}_{calage}(\varepsilon))$$

où ε est le **résidu de la régression (linéaire) de Y sur les variables de calage**.

Moralité **Plus les variables de calage X sont corrélées à Y, plus le résidu de la régression de Y sur X est faible et plus la variance de l'estimateur du total de Y est elle-même faible.**

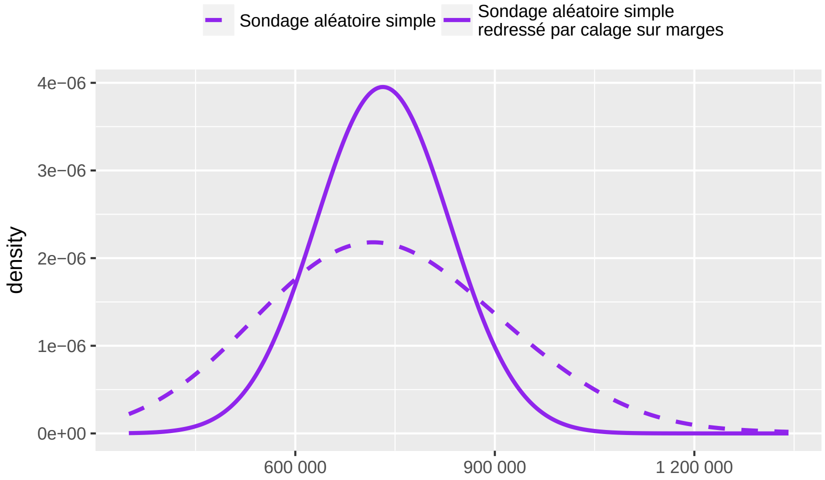
Application : Enquête sur la fréquentation des cinémas

Le distributeur souhaite **exploiter conjointement** l'information auxiliaire sur le nombre de projections et les périodes de vacances scolaires.

Pour ce faire, il introduit ces deux variables dans un calage sur marges par la **méthode exponentielle** (ou méthode du *raking ratio*).

À nouveau, il évalue les propriétés de l'estimateur en **répliquant 1 000 fois** l'ensemble des opérations (tirage puis redressement) et en représentant la **distribution des estimations ainsi obtenues**.

Application : Enquête sur la fréquentation des cinémas



Application : Enquête sur la fréquentation des cinémas

Biais Moyenne empirique sur 1 000 simulations

- Valeur dans la population : 731 892 entrées
- Estimateur d'Horvitz-Thompson : 733 832
- Estimateur par le ratio : 730 299
- Estimateur par post-stratification : 725 812
- Estimateur par calage sur marges : 731 625

Précision Écart-type empirique sur 1 000 simulations

- Estimateur d'Horvitz-Thompson : 153 932
- Estimateur par le ratio : 16 039
- Estimateur par post-stratification : 145 332
- Estimateur par calage sur marges : 13 607

Le calage sur marges en pratique

La plupart des enquêtes par sondage font l'objet d'un calage sur marges sur les **grandes structures de la population**.

En effet, une telle opération **ne peut qu'améliorer la précision** et garantit la **cohérence avec des sources extérieures à l'enquête**.

Est ainsi diffusé dans le fichier de l'enquête non pas le poids de sondage mais le **poids calé** sur de nombreuses marges.

En pratique, le calage sur marges est implémenté dans de nombreux logiciels :

- SAS : macro **%calmar** ;
- **R** : *packages* `sampling` et `icarus`.

En guise de conclusion

Les méthodes de redressement cherchent à **exploiter l'information auxiliaire disponible** au moment de l'estimation pour **améliorer la précision**.

Les estimateurs par le **ratio** et **post-stratifié** présentent une **variance plus faible** que l'estimateur d'Horvitz-Thompson pour autant que la variable d'intérêt soit **bien corrélée** à la variable explicative utilisée.

La méthode du **calage sur marges** généralise ce principe et permet de tirer parti de plusieurs variables auxiliaires simultanément.

Chapitre 2

Panorama de questions en sondages

Partie 1

Gestion de la collecte : suivi, multimode

Mode de collecte

Une fois qu'un échantillon a été tiré, et avant de disposer des données, il faut que la collecte se déroule. Plusieurs modes de collecte :

- CAPI : enquêteur en face à face
- CATI : enquêteur par téléphone
- CAWI : collecte par Internet
- Papier : questionnaire auto-administré

Ces modes peuvent être proposés concurentiellement ou séquentiellement.

Multimode

Lorsque plusieurs modes sont proposés :

- Effet de mode :
  - On ne répond pas pareil au téléphone et sur papier
  - Biais de désirabilité
  - Exemple classique : consommation de drogues des adolescents
- Effet de sélection :
  - Ce ne sont pas les mêmes personnes qui répondent sur Internet sur papier
  - Personnes plus jeunes et plus diplômées, ayant accès à Internet
  - Grandes disparités dans les réponses au recensement

Suivi de collecte

Principe :

- Au cours de la collecte, on sait qui a répondu ou pas
- Les profils peuvent être équilibrés ou non
- On peut concentrer les relances, par exemple téléphoniques
- On peut même débloquer des échantillons dits de réserve

Dispositif mis en oeuvre par les prestataires de collecte, et certains instituts comme Statistiques Canada.

Partie 2

Bases imparfaites et partage de poids

Base imparfaite

Retour sur la base de sondage parfaite :

- permet d'identifier les individus de façon non ambiguë
- est exhaustive (on parle sinon de défaut de couverture)
- est sans double compte
- contient de l'information auxiliaire (voir cours suivants)

Pas toujours le cas. Exemple : si l'on souhaite faire une enquête auprès des sans domicile ?

Enquête Sans Domicile

Enquête menée par l'INSEE en 2001 et en 2012. Comment atteindre la population ?

- on liste les services d'aide aux SDF (hébergement, repas) ;
- on en échantillonne certains ;
- on se rend dans le centre ;
- on interroge une personne sur p.

Quel est le poids de sondage d'un individu ?

Partage des poids

Supposons qu'un individu A est allé dans un centre d'hébergement une seule nuit. Un individu B lui par contre y dort tous les soirs. A t-on autant de chance de sélectionner A ou B ?

Idée : prendre en compte le nombre de nuits passées dans des centres ; c'est le nombre de **liens**. On utilise ensuite ce qu'on appelle le **partage des poids** : on divise le poids par le nombre de liens.

Partie 3

Tirage à probabilités inégales

Probabilités de sélection

Rappel : le plan de sondage permet de déterminer des probabilités d'inclusion pour chaque unité de la population.

Probabilité d'inclusion simple $\pi_k = \sum_{s \in \mathcal{S}} \delta_k p(s)$

Probabilité d'inclusion double $\pi_{k,l} = \sum_{s \in \mathcal{S}} \delta_k \delta_l p(s)$

Ces probabilités peuvent être très différentes entre les individus.

Tirage selon une variable X

Une méthode classique : le sondage à probabilités inégales selon une variable X disponible dans la base de sondage. On veut ici que π_k et X_k soient proportionnels : plus X est fort, plus il y a de chances qu'on sélectionne un individu. On a :

$$\pi_k = \frac{X_k}{\sum_{k \in U} X_k}$$

Propriété de ce plan de sondage : on estime parfaitement le total de la variable X :

$$\hat{X}_{HT} = \sum_{k \in S} \frac{X_k}{\pi_k} = \sum_{k \in U} X_k$$

Sondage auto-pondéré

On souhaite interroger des salariés d'entreprises, et on veut qu'ils aient tous la même chance d'être sélectionnés :

1. On échantillonne des entreprises à probabilités inégales selon X, le nombre de salariés ;
2. On sélectionne 5 salariés dans l'entreprise.

Pourquoi ils ont la même chance ?

- Une entreprise de 10 salariés a une chance sur 20 d'être sélectionnée. Ensuite, chaque salarié a une chance sur 2 : au total, une chance sur 40.
- Une entreprise de 100 salariés a une chance sur 2 d'être sélectionnée. Ensuite, chaque salarié a une chance sur 20 : au total, une chance sur 40.

Partie 4

Tirage équilibré

Retour sur l'échantillon "représentatif"

Lorsque l'on réalise un sondage aléatoire simple, on ne connaît pas la structure de l'échantillon obtenu : ratio homme/femme, etc.

Pour pallier ce problème, on peut stratifier.

Mais comment faire si l'on souhaite une structure précise pour :

- Sexe ;
- Âge ;
- Région. . .

Tirage équilibré

Cela demanderait trop de strates : à chaque fois qu'on rajoute un critère, il faut le croiser avec tous les autres, ce qui augmente très rapidement le nombre de strates.

Une autre méthode est possible : l'échantillonnage équilibré

- On choisit des variables X pour la structure : qualitatives ou quantitatives
- On sélectionne un échantillon s qui est correct sur ces variables X, c'est à dire que :

$$\hat{X}_{HT} = T(X)$$

- Si ce n'est pas possible, on cherche à être le plus proche possible.

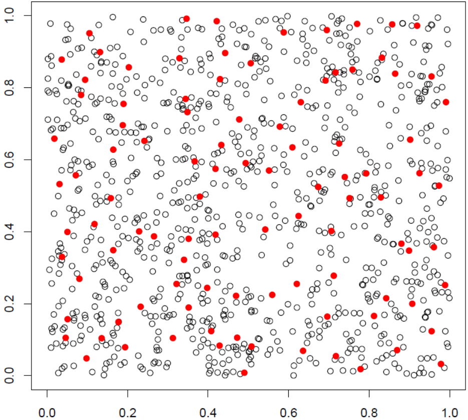
En pratique

Comment utiliser le tirage équilibré en pratique ?

- Méthode réjective : tirer des échantillons jusqu'à obtenir un échantillon qui convienne. Problème : quelles sont les vraies probabilités de sélection ?
- Méthode du Cube, qui respecte les π_i. Méthode assez complexe, qui est implémentée :
    - en SAS, via la macro Cube : `https://www.insee.fr/fr/information/2021904`
    - en R, par exemple dans les packages *sampling* et *BalancedSampling*.

Equilibrage spatial

On peut même chercher à répartir dans l'espace les points choisis, par exemple pour des sondages forestiers ou agricoles.



Partie 5

Unités influentes

Valeurs abberantes

On obtient parfois des réponses étonnantes :

- Erreur de compréhension de la question ;
- Erreur d'unité (€ contre k€)
- Mensonge

On peut corriger ces erreurs par des contrôles ou en recontactant la personne.

Unités influentes

Mais ces réponses peuvent être valides !

- Erreur de classification dans la base ;
- Évolution de l'unité (*strata jumper*) ;
- Manque d'information auxiliaire.

Problème : si quelqu'un a répondu 1 000 000 au milieu de répondants à 1 ou 2, son impact sur l'estimation est énorme.

Compromis biais-variance

On peut être tenté de le retirer de l'échantillon, en se disant qu'il est trop atypique. Mais cela crée du biais !

Une autre idée est de ne le pas le retirer, mais de réduire son poids de sondage. En effet, l'impact sur l'estimation HT dépend de la valeur de $w_i y_i$; réduire le poids diminue l'influence.

Quand on fait cela, on introduit du biais mais on réduit la variance. Idée : choisir le seuil de $w_i y_i$ à partir duquel on réduit le poids pour limiter au maximum la variance sans introduire trop de biais. La méthode usuelle : la **winsorisation**.

Partie 6

Estimation sur des petits domaines

Domaines

Rappel : un domaine d est un sous-ensemble de la population U.
Par exemple :

- Les hommes ou les femmes ;
- Les entreprises d'un secteur particulier.

Lorsque l'on veut estimer le total de Y sur le domaine d :

- La taille totale de l'échantillon n ne joue pas.
- C'est le nombre d'unités de l'échantillon qui sont dans le domaine d.
- Il est aléatoire, sauf si le domaine d est une strate.

Petits domaines

On souhaite parfois obtenir des résultats sur des domaines d qui sont de très petite taille. Par exemple :

- Les habitants d'un département ou d'une ville spécifique ;
- Les spectateurs d'une chaîne télé spécialiste ;
- Les salons de coiffure.

Quand la taille de l'intersection entre l'échantillon et le domaine n_d est très petite, il est difficile d'estimer avec précision. La solution optimale est souvent d'augmenter cette taille en amont du sondage, mais ce n'est pas toujours possible.

Estimation sur petits domaines

Il existe néanmoins des méthodes pour obtenir des estimations avec une précision correcte sur ces petits domaines. L'idée est la suivante :

1. On détermine un lien entre des variables X et les Y observés sur la population entière, par exemple par régression linéaire ;
2. On calcule via le modèle sur les X (connus) du domaine d une valeur de Y_d.

Rien ne garantit que le modèle calculé à l'étape 2 soit valide pour le domaine sur lequel nous travaillons ; il convient donc de traiter les résultats obtenus avec précaution.