PLAN DE SONDAGE DES ENQUÊTES GÉNÉRATION

Utilisation d'un calage pour suréchantillonner les extensions

Journées de Méthodologie Statistique de l'Insee

13/06/2018

Christophe Barret – Mady Cissé – Christophe Dzikowski

Les enquêtes « Génération »

- Enquête de la Statistique Publique sur l'insertion professionnelle des jeunes
- Génération de sortants du système éducatif (et non une génération d'âge!)
  - interrogés 3 ans après leur sortie
  - et une fois sur deux réinterrogés à 5 et 7 ans après la sortie (voire même 10 ans)
- Le Céreq constitue la base de sondage des présumés sortants du système éducatif en collectant des bases d'élèves auprès des ministères (lorsqu'il existe des bases centralisées) ou directement auprès des établissements scolaires.
- Cette base de sondage présente
  - des défauts de couverture (non-réponse de certains établissements)
  - des défauts de sur-couverture (nombreux individus hors champ de l'enquête)
- Plusieurs acteurs publics partenaires demandent des extensions d'échantillon
  - Sur certains niveaux ou spécialités de formation ou sur certains territoires
  - Ces extensions se recoupant entre elles.

↳ **Définition d'un plan de sondage adapté à ces contraintes.**

Les différentes phases de l'échantillonnage

- Phase A : le calcul des probabilités individuelles de tirage
  - Étape A1 : le taux de couverture
  - Étape A2 : détermination des probabilités de tirage en l'absence d'extension
  - Étape A3 : prise en compte des extensions dans le calcul des probabilités de tirage

- Phase B : le tirage équilibré de l'échantillon
  - Étape B1 : Tirage de l'échantillon global (principal + réserve)
  - Étape B2 : Tirage de la réserve et de l'échantillon principal

Calcul des probabilités individuelles de tirage

- Chaque individu *i* de la BS se voit attribuer un poids de couverture individuel ($pcouv_i$)
- On modélise sur la base d'une enquête passée (ou plusieurs) ,
  - la probabilité de réponse anticipée (pr_i)
  - la probabilité de réponse anticipée dans le champ (pr_i^C)
- On calcule la probabilité de tirage $\pi_{i1} = \frac{pcouv_i}{pr_i} * coeff1$
- On calcule coeff1 pour obtenir le nombre de répondants dans le champ (nr^C) satisfaisant les besoins du Céreq (hors extension)
- On fait **comme si** on faisait un premier tirage selon les probabilités $\frac{pcouv_i}{pr_i}$ puis on simulait le fait d'être répondant dans le champ pour obtenir un premier nombre de répondants dans le champ. Cela permet de calculer le coefficient unique $coeff1$ pour obtenir le nombre souhaité nr^C.
  Mais en fait on ne le fait pas, on obtient tout par calcul direct.

$$coeff1 = 8500 \Big/ \sum_{i \in U} \left(pr_i^C * \frac{pcouv_i}{pr_i} \right)$$

Prise en compte des extensions

- Des cibles d'extension détaillées essentiellement sur des formations ciblées (et parfois sur des territoires)

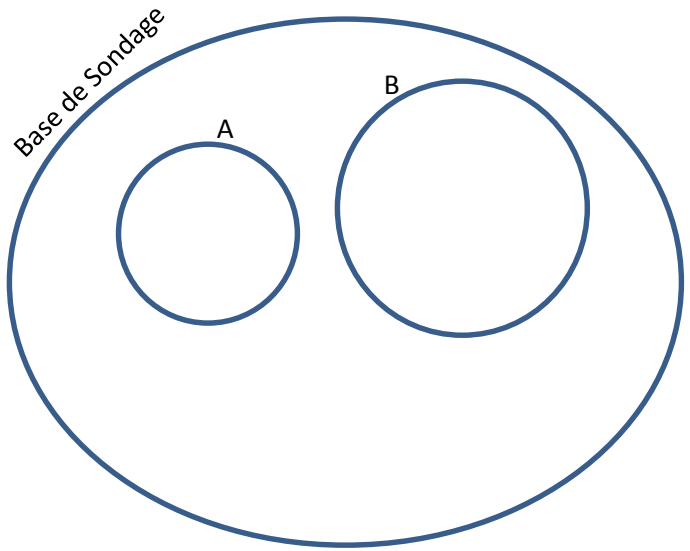
Sous population d'extension	codage des sous populations	nombre de questionnaires à partir de l'échantillon socle	cible (effectif extension +réserve)
Collège	c1	404	903
Baccalauréat général et technologique	c2	570	1 068
Baccalauréat professionnel	c3	1 374	2 391
CAP	c4	860	1 373
…	…	…	…
Ensemble	c38	8 490	24 700

- Des extensions qui se recoupent :
  - exemple : formation environnementale, de niveau bac pro, résidant en QPV, (dans une région donnée)

- Prenons un exemple pour comprendre la difficulté rencontrée

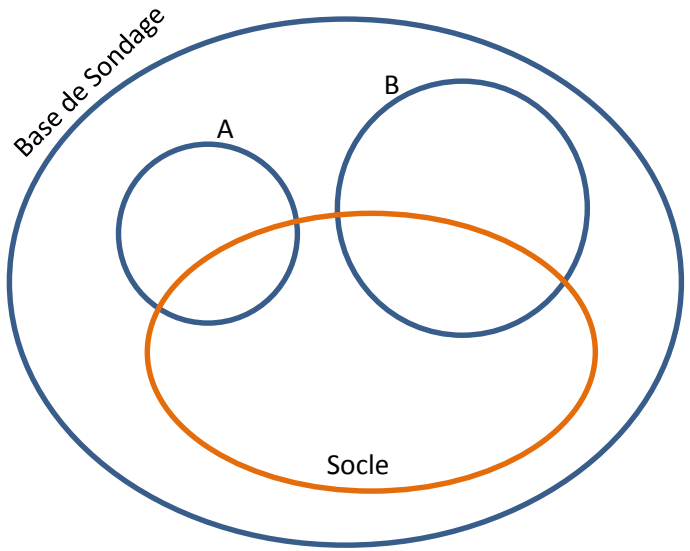
Les extensions : un exemple avec intersection vide

- Supposons deux sous-populations d'intérêt A et B. Les cibles de questionnaires fixées sont respectivement C_A et C_B.
- Sur l'échantillon socle (visant 8500 questionnaires), on estime le nombre de questionnaires de A, noté $nsoc_A$, et de B, noté $nsoc_B$.
- <u>1er cas : supposons $A \cap B = \emptyset$</u>



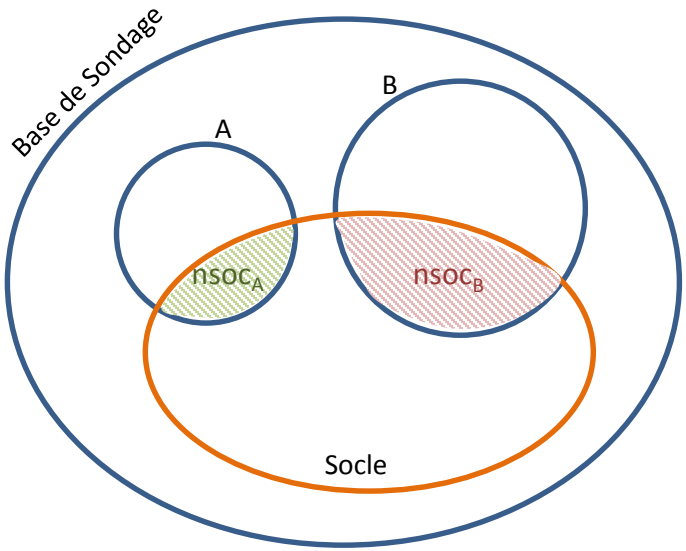
Les extensions : un exemple avec intersection vide

- Supposons deux sous-populations d'intérêt A et B. Les cibles de questionnaires fixées sont respectivement C_A et C_B.
- Sur l'échantillon socle (visant 8500 questionnaires), on estime le nombre de questionnaires de A, noté $nsoc_A$, et de B, noté $nsoc_B$.
- 1er cas : supposons $A \cap B = \emptyset$



Les extensions : un exemple avec intersection vide

- Supposons deux sous-populations d'intérêt A et B. Les cibles de questionnaires fixées sont respectivement C_A et C_B.
- Sur l'échantillon socle (visant 8500 questionnaires), on estime le nombre de questionnaires de A, noté $nsoc_A$, et de B, noté $nsoc_B$.
- <u>1er cas : supposons $A \cap B = \emptyset$</u>



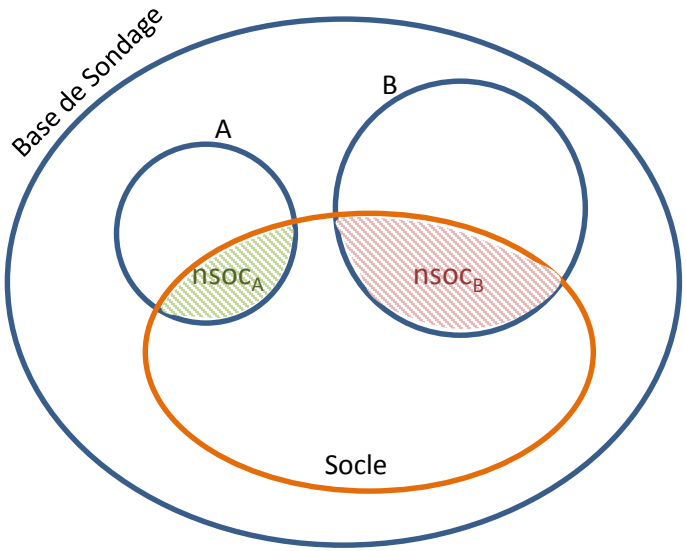
Les extensions : un exemple avec intersection vide

- Supposons deux sous-populations d'intérêt A et B. Les cibles de questionnaires fixées sont respectivement C_A et C_B.

- Sur l'échantillon socle (visant 8500 questionnaires), on estime le nombre de questionnaires de A, noté $nsoc_A$, et de B, noté $nsoc_B$.

- 1er cas : supposons A∩B=∅

$$\mathrm{dil}_A = \frac{C_A}{nsoc_A} \text{ et } \mathrm{dil}_B = \frac{C_B}{nsoc_B}$$



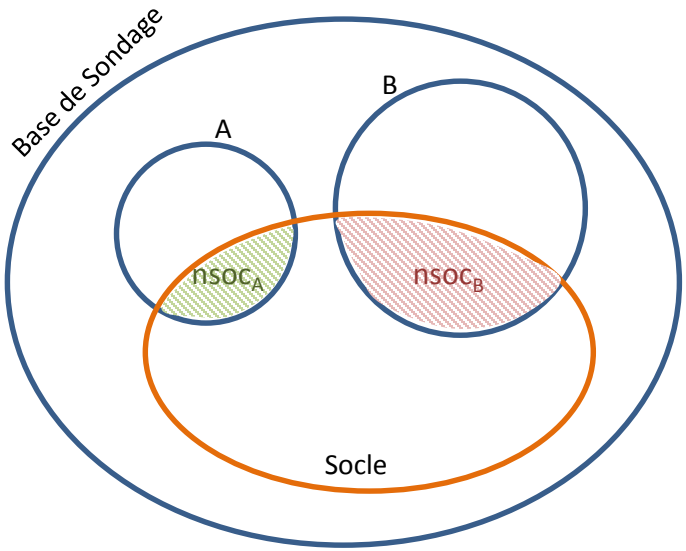
Les extensions : un exemple avec intersection vide

- Supposons deux sous-populations d'intérêt A et B. Les cibles de questionnaires fixées sont respectivement C_A et C_B.
- Sur l'échantillon socle (visant 8500 questionnaires), on estime le nombre de questionnaires de A, noté $nsoc_A$, et de B, noté $nsoc_B$.
- 1er cas : supposons A∩B=Ø

$$\text{dil}_A = \frac{C_A}{nsoc_A} \text{ et } \text{dil}_B = \frac{C_B}{nsoc_B}$$

$$\text{pour i}\epsilon\text{A}, \quad \pi_{i2} = \pi_{i1} \times \text{dil}_A = \pi_{i1} \times \frac{C_A}{nsoc_A}$$

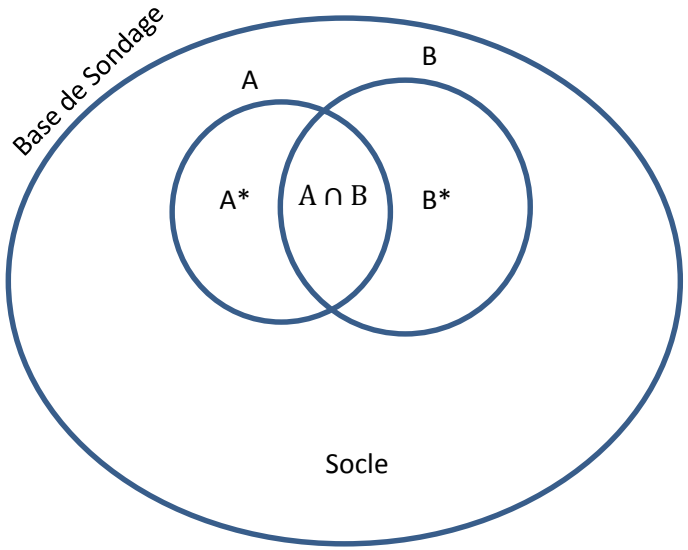
$$\text{pour i}\epsilon\text{B}, \quad \pi_{i2} = \pi_{i1} \times \text{dil}_B = \pi_{i1} \times \frac{C_B}{nsoc_B}$$



Les extensions : un exemple avec intersection **non** vide

- 2e cas : supposons $A \cap B \neq \emptyset$

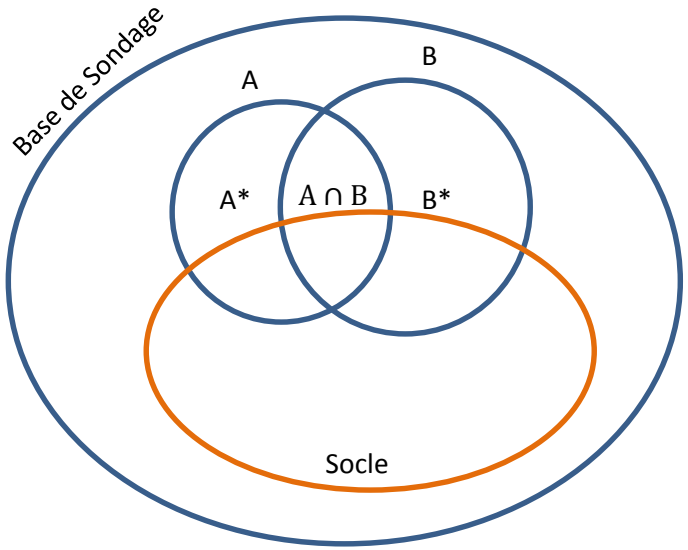
$$A^* = A \backslash (A \cap B)$$
$$B^* = B \backslash (A \cap B)$$
$$Z = SOCLE \backslash (A \cup B)$$



Les extensions : un exemple avec intersection **non** vide

- 2e cas : supposons $A \cap B \neq \emptyset$

$$A^* = A \backslash (A \cap B)$$
$$B^* = B \backslash (A \cap B)$$
$$Z = SOCLE \backslash (A \cup B)$$

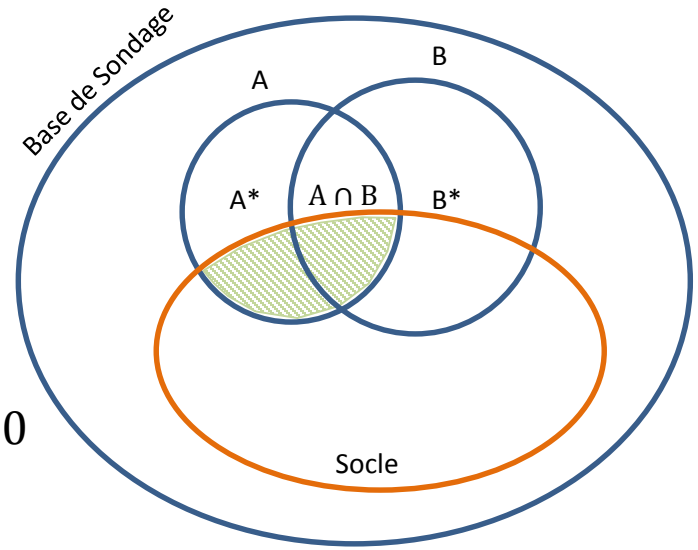


Les extensions : un exemple avec intersection **non** vide

- 2e cas : supposons A∩B≠∅

$$A^* = A \backslash (A \cap B)$$
$$B^* = B \backslash (A \cap B)$$
$$Z = SOCLE \backslash (A \cup B)$$

- Prenons :
  - $nsoc_A = 100$, $nsoc_B = 200$, $nsoc_{A \cap B} = 50$
  - $C_A = 1000$ et $C_B = 1300$

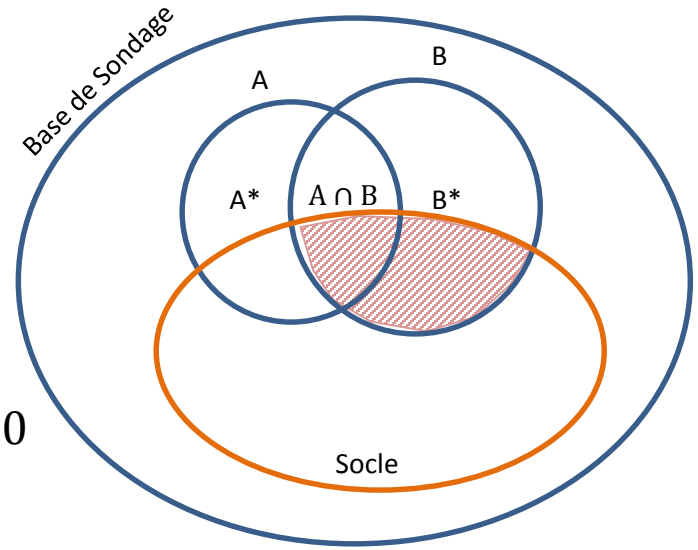


Les extensions : un exemple avec intersection **non** vide

- 2e cas : supposons A∩B≠∅

$$A^* = A \backslash (A \cap B)$$
$$B^* = B \backslash (A \cap B)$$
$$Z = SOCLE \backslash (A \cup B)$$

- Prenons :
  - $nsoc_A = 100$, $nsoc_B = 200$, $nsoc_{A \cap B} = 50$
  - $C_A = 1000$ et $C_B = 1300$

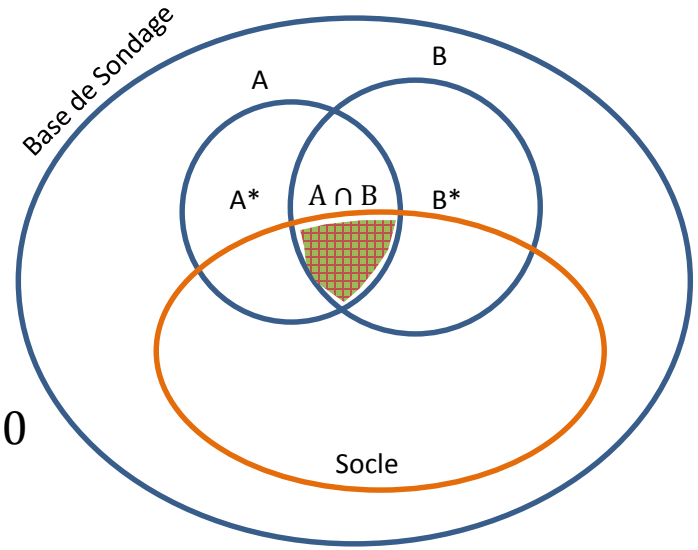


Les extensions : un exemple avec intersection **non** vide

- 2e cas : supposons A∩B≠∅

$$A^* = A \backslash (A \cap B)$$
$$B^* = B \backslash (A \cap B)$$
$$Z = SOCLE \backslash (A \cup B)$$

- Prenons :
  - $nsoc_A = 100$, $nsoc_B = 200$, $nsoc_{A \cap B} = 50$
  - $C_A = 1000$ et $C_B = 1300$

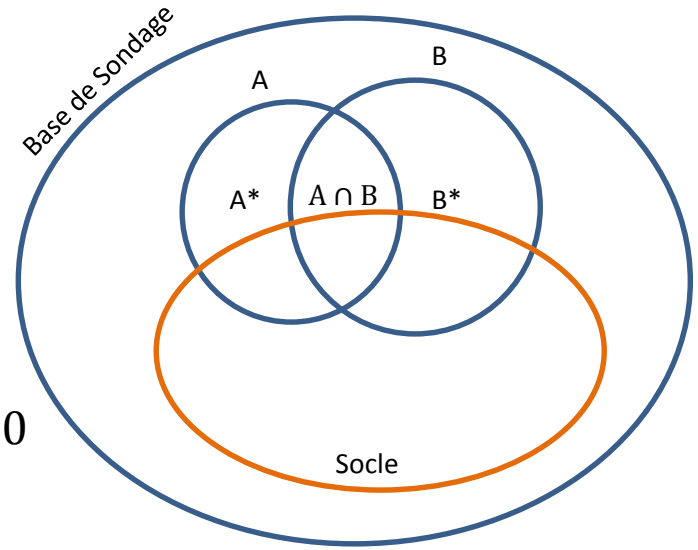


Les extensions : un exemple avec intersection **non** vide

- 2e cas : supposons A∩B≠∅

$$A^* = A \backslash (A \cap B)$$
$$B^* = B \backslash (A \cap B)$$
$$Z = SOCLE \backslash (A \cup B)$$

- Prenons :
  - $nsoc_A = 100$, $nsoc_B = 200$, $nsoc_{A \cap B} = 50$
  - $C_A = 1000$ et $C_B = 1300$



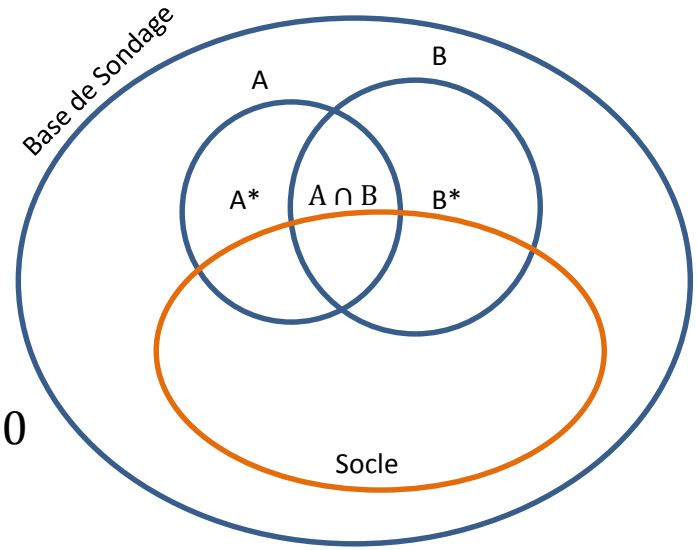
		METHODE 1 : d'abord A puis B	
		d'abord A (on applique dilA=1000/100 à la population A)	puis B (on applique dilB=1300/650 à la population B)
	nombre de répondants estimé dans l'échantillon socle	nombre de répondants estimé dans l'échantillon intermédiaire	nombre de répondants estimé dans l'échantillon final méthode 1 (F1)
A*	50	500	500
B*	150	150	300
A∩B	50	500	1000
Z	8250	8250	8250
Total	8500	9400	10050
A	100	1000	1500
B	200	650	1300

Les extensions : un exemple avec intersection **non** vide

- 2e cas : supposons A∩B≠∅

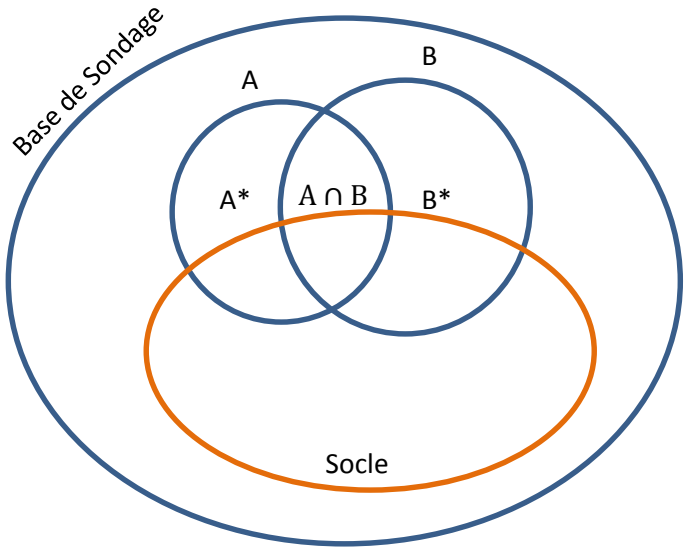
$$A^* = A \backslash (A \cap B)$$
$$B^* = B \backslash (A \cap B)$$
$$Z = SOCLE \backslash (A \cup B)$$

- Prenons :
  - $nsoc_A = 100$, $nsoc_B = 200$, $nsoc_{A \cap B} = 50$
  - $C_A = 1000$ et $C_B = 1300$



		METHODE 1 : d'abord A puis B		METHODE 1bis : d'abord B puis A	
		d'abord A (on applique dilA=1000/100 à la population A)	puis B (on applique dilB=1300/650 à la population B)	d'abord B (on applique dilB=1300/200 à la population A)	puis A (on applique dilA=1000/375 à la population B)
	nombre de répondants estimé dans l'échantillon socle	nombre de répondants estimé dans l'échantillon intermédiaire	nombre de répondants estimé dans l'échantillon final méthode 1 (F1)	nombre de répondants estimé dans l'échantillon intermédiaire	nombre de répondants estimé dans l'échantillon final méthode 1 (F1bis)
A*	50	500	500	50	133
B*	150	150	300	975	975
A∩B	50	500	1000	325	867
Z	8250	8250	8250	8250	8250
Total	8500	9400	10050	9600	10225
A	100	1000	1500	375	1000
B	200	650	1300	1300	1842

Les extensions : un exemple avec intersection **non** vide



- Pour pallier le problème de dépendance à l'ordre de traitement des extensions.

On peut définir :

$\mathrm{dil}_A = 1 + \alpha_A$ et $\mathrm{dil}_B = 1 + \alpha_B$

- Puis on dilate les probabilités comme suit :

pour $i \epsilon A^*$, $\pi_{i2} = \pi_{i1} \times (1 + \alpha_A)$

pour $i \epsilon B^*$, $\pi_{i2} = \pi_{i1} \times (1 + \alpha_B)$

pour $i \epsilon A \cap B$, $\pi_{i2} = \pi_{i1} \times (1 + \alpha_A + \alpha_B)$

	METHODE 2 : A et B simultanément		
	nombre de répondants estimé dans l'échantillon socle	coefficient de dilatation	nombre de répondants estimé dans l'échantillon final méthode 2 (F2)
A*	50	10	500
B*	150	6,5	975
A∩B	50	15,5	775
Z	8250	1	8250
total	8500		10500
A	100		1275
B	200		1750

Les extensions : un exemple avec intersection **non** vide

- Les inconvénients de l'approche :
  - Dépendance à l'ordre (pour la méthode F1)
  - Dépassement d'une cible (F1) voire des deux cibles (F2)
  - Extensions surreprésentées dans l'échantillon final

	socle	F1	F1bis	F2
poids des répondants de A∩B parmi ceux de A	50%	67%	87%	61%
poids des répondants de A∩B parmi ceux de B	25%	77%	47%	44%

- D'où l'idée « naturelle » d'itérer la méthode F1 jusqu'à convergence sur les effectifs cibles de A et B.

- Ce qui revient à effectuer un calage sur marges sur un tableau du type

Partition	A	B	effectif
P1 (A*)	1	0	50
P2 (B*)	0	1	150
P3 (A∩B)	1	1	50

- Quelques réglages en aval : on s'assure que les proba π_{i2} soient entre π_{i1} et 1

Le tirage équilibré de l'échantillon

- Tirage <u>en une seule étape</u> selon les probabilités inégales de tirage π_{i2}

- Tirage équilibré sur la variable de probabilité de tirage pour chaque sous-population *J*

$$\sum_{i \in S} \frac{1_{i \in J} * \pi_{i2}}{\pi_{i2}} = \sum_{i \in U} 1_{i \in J} * \pi_{i2}$$

$$\sum_{i \in S} 1_{i \in J} = \sum_{i \in U} 1_{i \in J} * \pi_{i2}$$

$$n_{j=} \sum_{i \in U} 1_{i \in J} * \pi_{i2}$$

  - cela permet d'obtenir un tirage de taille fixe (pour *J*)
  - Qui permet d'obtenir le nombre d'individus échantillonnés adapté à l'obtention de la cible sur *J*

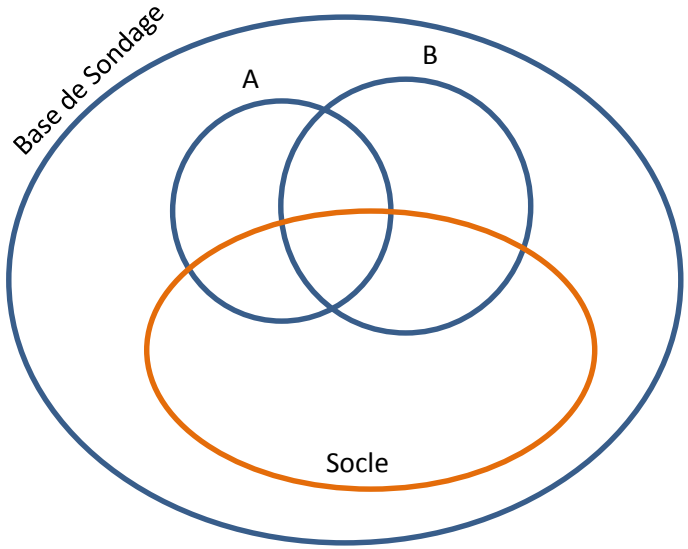
- Après cette étape de tirage de l'échantillon global, tirage de l'échantillon principal (pour constituer une réserve… non utilisée…)

Une méthode alternative

- La méthode précédente consistait :
  - à effectuer un calcul de probabilité individuel qui prennent directement en compte les surplus d'échantillonnage liés aux extensions
  - puis d'effectuer un tirage en une seule étape.

- Une méthode alternative plus classique consisterait :
  - à faire un tirage en deux étapes :
    - Tirage du socle
    - Puis tirage de chaque extension dans le complémentaire
  - Et d'effectuer un partage de poids pour gérer les surreprésentations des intersections des extensions.

Une méthode alternative



- Tirage de l'échantillon Socle

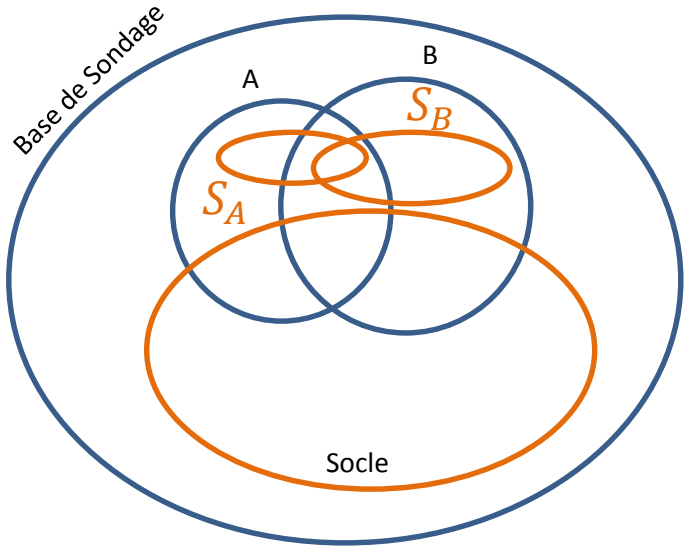
- Tirages indépendants des suppléments d'extension :
  - Dans A\Socle : Echantillon S_A de taille $C_A - nsoc_A$ avec probabilité $\pi_{i,A}$
  - Dans B\Socle : Echantillon S_B de taille $C_B - nsoc_B$ avec probabilité $\pi_{i,B}$

- Partage des poids pour compter les compter les individus présents dans plusieurs extensions : (les individus de A inter B ont deux chances d'être tirés)

$$poids_i = \frac{1}{\mathbb{I}_{i\in A}+\mathbb{I}_{i\in B}} \cdot \left(\frac{\mathbb{I}_{i\in S_A}}{\pi_{i,A}} + \frac{\mathbb{I}_{i\in S_B}}{\pi_{i,B}}\right)$$

- Permet de donner le bon poids aux individus des intersections mais n'empêche pas a priori le suréchantillonnage de l'intersection (en termes de nombre d'individus tirés)

Une méthode alternative



- Tirage de l'échantillon Socle

- Tirages indépendants des suppléments d'extension :
  - Dans A\Socle : Echantillon S_A de taille $C_A - nsoc_A$ avec probabilité $\pi_{i,A}$
  - Dans B\Socle : Echantillon S_B de taille $C_B - nsoc_B$ avec probabilité $\pi_{i,B}$

- Partage des poids pour compter les compter les individus présents dans plusieurs extensions : (les individus de A inter B ont deux chances d'être tirés)

$$poids_i = \frac{1}{\mathbb{I}_{i\in A}+\mathbb{I}_{i\in B}} \cdot \left(\frac{\mathbb{I}_{i\in S_A}}{\pi_{i,A}} + \frac{\mathbb{I}_{i\in S_B}}{\pi_{i,B}}\right)$$

- Permet de donner le bon poids aux individus des intersections mais n'empêche pas a priori le suréchantillonnage de l'intersection (en termes de nombre d'individus tirés)

Conclusion : Avantages et inconvénients des deux méthodes

Avantages de la méthode en une étape (avec calcul des proba par calage)

- Le nombre souhaité exactement pour chaque extension
- Limiter le nombre d'individus dans les intersections
- Un seul tirage (la méthode en deux étapes nécessite autant de tirages que d'extensions, c'est plus long lorsque celles-ci sont nombreuses)
- Estimations en population entière incluant facilement les extensions

Avantages de la méthode en deux étapes (avec partage des poids) :

- Plus intuitive ?
- Permet de dissocier les individus sur-échantillonnés pour les extensions de ceux du socle
  - ↳ si une extension n'est pas reconduite pour la réinterrogation, il suffit de supprimer les individus suréchantillonnés de cette extension.

DIAPOS COMPLEMENTAIRES

Tableau 1 : extrait de la table PARTITION

Parties	c1	c2	c3	c4	c6	c7	c8	c9	c10	c11	c12	c13	c14	c15	c17	c20	c21	c22	c23	c24	c25	c26	c27	c28	c29	c30	c31	c32	c33	c34	c35	c37	Effectif agrégé
P1	1	2	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2,33
P2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	401,51
P3	2	1	2	2	2	2	2	1	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	16,07
…																																	…
P93	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	1	2	154,14
P94	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	2	203,66

$$min_{somme\ calée_p} \sum_{p \in PARTITION} eff_p * G\left(\frac{eff\ calé_p}{eff_p}\right)$$

Sous la contrainte :

$$\sum_{p \in PARTITION} eff\ calé_p * \begin{pmatrix} 1_{pj1} \\ 1_{pj2} \end{pmatrix} = t_{Xj}$$

$$\text{où } t_{Xj} = \begin{pmatrix} Mar1_j \\ Mar2_j \end{pmatrix}$$

Tableau 1 : écart entre échantillon global et cibles

Sous population d'extension	Effectif répondant à partir de l'échantillon global	cible (effectif extension +réserve)	différence entre échantillon global et les cibles
Collège	895	903	-8
Baccalauréat général et technologique	1 059	1 068	-9
Baccalauréat professionnel	2 403	2 391	12
CAP	1 372	1 373	-1
CGDD	3 372	3 349	23
CGDD niveau 1	831	842	-11
CGDD niveau 2	826	817	9
CGDD niveau 3	564	567	-3
CGDD niveau 4	744	748	-4
CGDD niveau 5	407	416	-9
DGESIP	7 350	7 026	324
DGESIP autre	2 044	1 959	85
DGESIP BTS	974	972	2
DGESIP Grandes écoles	802	771	31
DGESIP IUT	348	350	-2
DGESIP Licence	1 666	1 464	202
DGESIP Master	1 516	1 509	7
Quartier priorité de la ville	2 145	2 287	-142
Santé social	5 971	5 981	-10
Santé social aide-soignante	1 171	1 180	-9
Santé social assistant social	395	349	46
Santé social auxiliaire puéricultrice	510	454	56
Santé social conseiller économie familiale	207	208	-1
Santé social éducateur jeunes	323	324	-1
Santé social éducateur spécialisé	531	474	57
Santé social ergothérapeute	62	62	0
Santé social infirmier	1 190	1 191	-1
Santé social masseur kinésithérapeute	425	427	-2
Santé social moniteur éducateur	521	461	60
Santé social orthophoniste	77	78	-1
Santé social orthoptiste	31	31	0
Santé social podologue pédicure	140	140	0
Santé social psychomotricien	84	84	0
Santé social sage femmes	304	306	-2
Sport	2 605	1 678	927
Thèse	2 486	2 446	40
Docteur en santé	110	112	-2
Ensemble	26 734	24 700	2 034

Tableau 1 : écart entre échantillon principal et cible

Sous population d'extension	Effectif répondant à partir de l'échantillon principal	Cibles échantillon principal	Différence entre échantillon principal et les cibles
Collège	895	903	-8
Baccalauréat général et technologique	1 059	1 068	-9
Baccalauréat professionnel	2 403	2 391	12
CAP	1 372	1 373	-1
CGDD	3 163	3 080	83
CGDD niveau 1	753	765	-12
CGDD niveau 2	750	742	8
CGDD niveau 3	509	515	-6
CGDD niveau 4	744	680	64
CGDD niveau 5	407	378	29
DGESIP	6 665	6 387	278
DGESIP autre	1 840	1 781	59
DGESIP BTS	887	884	3
DGESIP Grandes écoles	730	701	29
DGESIP IUT	318	318	-1
DGESIP Licence	1 518	1 331	186
DGESIP Master	1 373	1 372	1
Quartier priorité de la ville	1 980	2 079	-99
Santé social	5 595	5 438	158
Santé social aide-soignante	1 071	1 073	-1
Santé social assistant social	359	318	41
Santé social auxiliaire puéricultrice	461	413	48
Santé social conseiller économie familiale	207	208	-1
Santé social éducateur jeunes	323	324	-1
Santé social éducateur spécialisé	488	431	57
Santé social ergothérapeute	62	62	0
Santé social infirmier	1 085	1 083	2
Santé social masseur kinésithérapeute	425	427	-2
Santé social moniteur éducateur	477	419	58
Santé social orthophoniste	77	78	0
Santé social orthoptiste	31	31	0
Santé social podologue pédicure	140	140	0
Santé social psychomotricien	84	84	0
Santé social sage femmes	304	306	-2
Sport	2 360	1 525	835
Thèse	1 863	1 500	363
Docteur en santé	100	100	0
Ensemble	24 570	23 000	1 570

Notons n^j la cible pour l'extension E^j et n^j_{socle} le nombre d'individus de l'extension E^j dans l'échantillon socle.

Si l'échantillon socle est équilibré sur les effectifs des extensions, n^j_{socle} est fixe et est connu à l'avance : $n^j_{socle} = \sum_{i \in E^j} \pi_{i1}$

$n^j_{supp} = n^j - n^j_{socle}$.

pour chaque extension d'échantillon E^j,

$$\pi_i^j = \frac{\pi_{i1} \cdot n^j_{supp}}{\sum_{k \epsilon E^j / S_{socle}} \pi_{k1}}$$

$$W_{i,ext} = \left(\sum_{k=1}^{p} \frac{L_i^{j_k}}{\pi_i^{j_k}} \right) \cdot \frac{\sum_{k=1}^{p} L_i^{j_k}}{p}$$

Avec $L_i^{j_k}$ l'indicatrice que l'individu i est tiré pour l'extension E^{j_k}