LES SONDAGES : PRATIQUE ET THÉORIE

Anne Ruiz-Gazen

TSE, Université Toulouse Capitole

13 Janvier 2024

Conférence Cercle Sofia Kovalevskaïa

Introduction

- Professeure de statistique à Toulouse School of Economics.
- Parcours et carrière universitaires à Toulouse :
  - DEUG en mathématiques appliquées aux sciences sociales à UT2,
  - Licence et maitrise de mathématiques pures, DEA et doctorat de mathématiques appliquées spécialité statistique à UT3,
  - Enseignante-chercheuse à UT1 (MCF/Professeure).
- Intérêt pour la statistique à partir de la maitrise. Domaine avec de nombreux débouchés qui permet développements théoriques et/ou appliqués.
- Métier passionnant avec propre équilibre à trouver entre enseignement, recherche et administration.
- Deux sujets de recherche principaux :
  - Statistique robuste et détection d'anomalies
  - Théorie des sondages

Introduction

1. Quelques concepts et intuitions sur les sondages
2. Introduction au formalisme mathématique de la théorie des sondages
3. Deux exemples de recherche en théorie des sondages

Où trouve-t-on des études par sondages ?

Un peu partout

- dans les offices nationaux de statistique comme l'INSEE en France pour compter :
  - le nombre de français,
  - le nombre de chômeurs en France au dernier trimestre,
  - le nombre de personnes qui ont été victimes d'un cambriolage dans une région donnée, ...
- dans les ministères et les grands instituts de recherche publique comme l'INSERM (santé), l'INED (démographie), Santé Publique France, ...
- dans les grandes entreprises comme La Poste, Médiamétrie,...
- dans les publications quotidiennes à visée du grand public (élection, avis, goûts,...).

Selon le cas, pas les mêmes moyens mis en œuvre et pas la même fiabilité attendue dans les résultats.

Population, paramètre d'intérêt, échantillon et estimation

Supposons que l'on souhaite connaître le nombre de chômeurs en France métropolitaine au dernier trimestre 2023.

On cherche à compter le nombre de personnes au chômage dans la population des personnes domiciliées en France métropolitaine.

Le paramètre d'intérêt est un total sur une population finie.

Parmi les difficultés rencontrées : on ne peut pas interroger tous les français tous les trimestres.

La solution proposée : on n'interroge que certaines personnes et à partir de leur réponse on propose une estimation du paramètre qui nous intéresse sur toute la population (inférence statistique).

On espère que l'estimation du paramètre que l'on fournit n'est pas très différente de la vraie valeur de ce paramètre (qu'en pratique on ne connaîtra pas).

Types d'erreur

La différence entre l'estimation du paramètre et sa vraie valeur est appelée erreur d'estimation.

Il existe de nombreuses sources d'erreur d'estimation.

- Erreurs de couverture.
- Erreur d'échantillonnage.
- Erreur de non-réponse.
- Erreur de mesure.

Mathématiquement, si on maitrise le tirage aléatoire de l'échantillon, on peut espérer maitriser l'erreur d'échantillonnage.

Les autres types d'erreur ne sont généralement pas maitrisées ou alors sous des hypothèses qui ne sont pas vérifiables en pratique.

Stratégie en sondage

Une stratégie en sondage consiste en

- un plan de sondage i.e. une méthode probabiliste de tirage d'échantillon,
- et une méthode d'estimation (ou estimateur) du paramêtre d'intérêt qui prend en compte les données de l'échantillon tiré.

Dans cet exposé, le plan de sondage est supposé fixé et maitrisé par le statisticien d'enquête.

Ce point est très important car la connaissance de la loi de probabilité sous-jacente au tirage de l'échantillon va permettre de dériver les propriétés mathématiques des estimateurs.

On ne connaît pas l'erreur d'échantillonnage pour une estimation donnée mais on va pouvoir calculer son espérance et aussi sa variance (variabilité de l'erreur).

L'objectif de ce petit exemple est de vous donner une intuition de

- ce qu'est l'erreur d'échantillonnage,
- ce qu'est une stratégie (simple),
- comment on peut modifier une stratégie.

- Imaginons une population de taille 5000.
- On veut connaître (paramètre d'intérêt) le nombre de personnes qui ont un compte Facebook actif dans cette population.

- Imaginons une population de taille 5000.
- On veut connaître (paramètre d'intérêt) le nombre de personnes qui ont un compte Facebook actif dans cette population mais on ne peut pas interroger tout le monde.
- On interroge 10 personnes parmi les 5000. On dit que l'on a tiré un échantillon de taille 10.

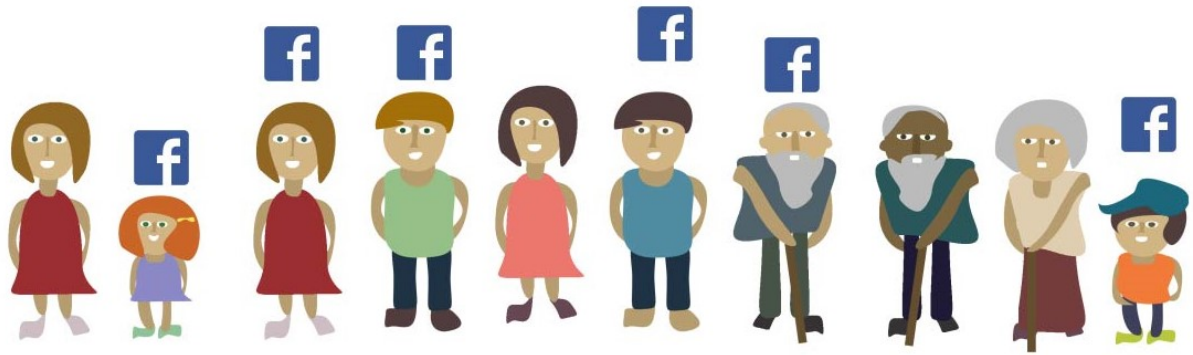
Important

On suppose que les personnes que l'on interroge sont choisies "au hasard" (de manière aléatoire), par exemple ici en donnant la même chance (probabilité) à chacun des échantillons de taille 10 d'être tiré.

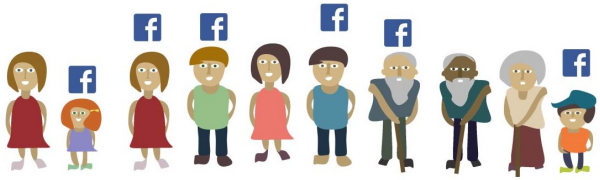
Cette méthode pour tirer un échantillon (basée sur l'équiprobabilité) est un plan de sondage particulier.

En pratique ce n'est pas si facile car cela suppose qu'on dispose d'une base de sondage (une liste de tous les individus de la population dans laquelle on va tirer l'échantillon).

- Imaginons une population de taille 5000 (dont 3000 personnes avec un compte Facebook mais on ne le sait pas).
- On veut connaître le nombre de personnes qui ont un compte Facebook actif dans cette population mais on ne peut pas interroger tout le monde.
- On interroge 10 personnes parmi les 5000. On dit que l'on a tiré un échantillon de taille 10.



- Dans cette population de taille 5000, 6 personnes parmi les 10 interrogées ont un compte Facebook.



- On va considérer que chacun des individus interrogés représente 500 personnes puisque

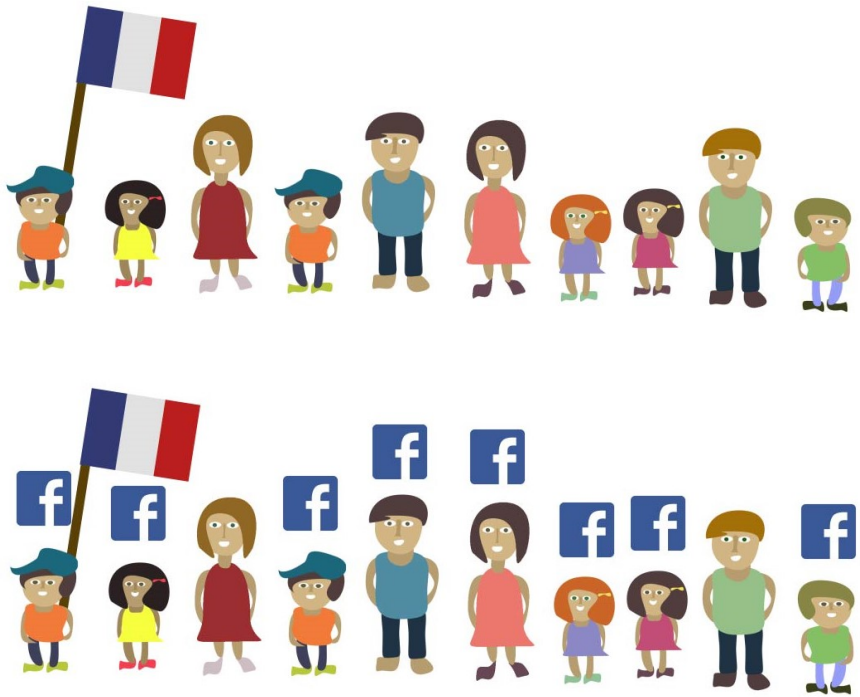
$$10 \times 500 = 5000$$

- On en déduit une estimation du nombre de personnes de la population qui ont un compte Facebook :

$$6 \times 500 = 3000$$

- On peut interpréter le calcul en disant que chaque individu interrogé "compte pour 500 personnes" ou a un poids de 500 dans le calcul de l'estimation. Il s'agit de l'estimateur de Hovitz-Thompson ou estimateur par les valeurs dilatées.

- Dans cette même population de taille 5000, on tire un autre échantillon :



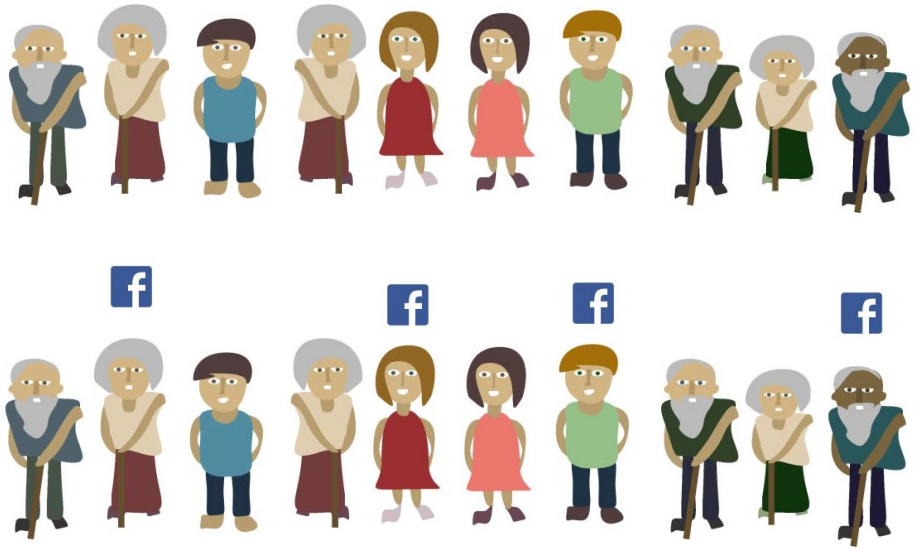
8 personnes parmi les 10 interrogées ont un compte Facebook.

- On en déduit une estimation du nombre de personnes de la population qui ont un compte Facebook :

$$8 \times 500 = 4000$$

- L'estimation dépend de l'échantillon tiré. Ici on a une surestimation.

- Dans cette même population de taille 5000, on tire un autre échantillon :



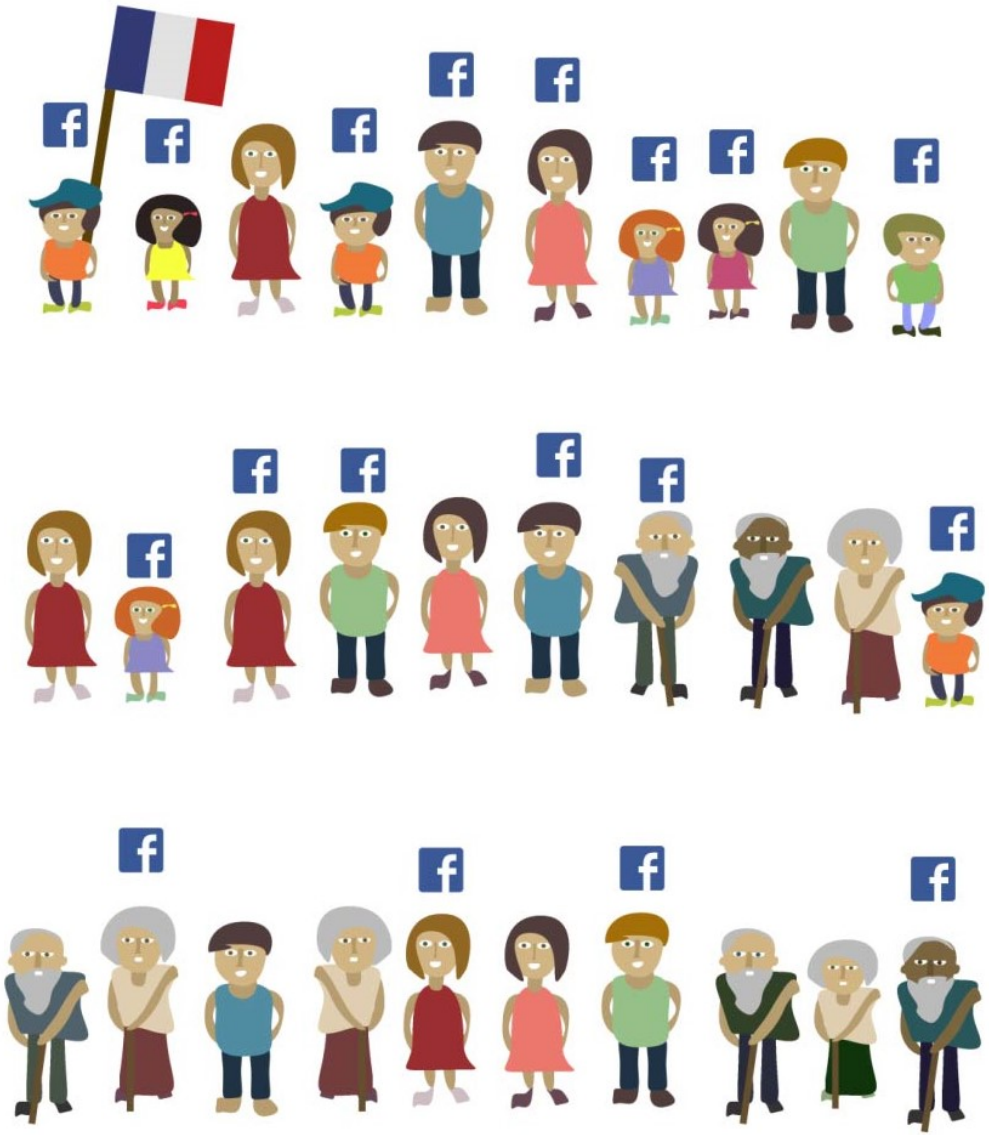
4 personnes parmi les 10 interrogées ont un compte Facebook.

- On en déduit une estimation du nombre de personnes de la population qui ont un compte Facebook :

$$4 \times 500 = 2000$$

- L'estimation dépend de l'échantillon tiré. Ici on a une sous-estimation.

Si on compare les 3 échantillons :



Changement de stratégie

On va voir que la stratégie précédente est sans biais (en espérance l'estimateur est égal au paramètre d'intérêt) mais l'estimateur est très variable.

- Une première idée pour éviter de trop grandes sous-estimations ou sur-estimations consiste à maitriser le tirage de l'échantillon en ne permettant pas de tirer un échantillon sans jeunes ou sans personnes agées.

  On parle de stratification ou d'échantillonnage stratifié.

- Une autre idée est de changer la méthode d'estimation. Dans les calculs d'estimation précédents, on a donné un poids de 500 à chaque personne de l'échantillon. Le poids est le même pour chaque personne de l'échantillon mais on peut donner des poids différents selon si l'individu est jeune ou agé par exemple.

  On parle de post-stratification ou d'estimation post-stratifié.

Trois petites remarques sur la partie qui précède :

- Dans la réalité, on ne connait pas la vraie valeur dans la population (on ne la connaîtra jamais !) et on ne dispose que d'un seul échantillon. Donc on ne sait pas si on sur-estime, sous-estime ou estime comme il faut la valeur que l'on cherche.
- On sait que l'estimation que l'on donne est "fausse" et donc on préfère en général donner un intervalle de valeurs probables. Par exemple, au lieu de dire que l'on estime le nombre de comptes Facebook actifs dans la population par 3000, on dira que le nombre de comptes Facebook est très probablement dans l'intervalle [2500 ;3500].
- Plus l'intervalle est petit, meilleure est l'estimation (plus précise). Et aussi plus la taille de l'échantillon est grande, plus l'intervalle est petit. Avec une taille d'échantillon de 10, comme dans l'exemple fictif, l'estimation est très peu précise.

La stratification consiste à partager la population en sous-populations appelées strates et à tirer un échantillon dans chaque strate en fixant la taille de l'échantillon dans chaque strate.

Dans l'exemple fictif, on partage la population en 3 strates (jeunes, adultes, personnes agées)

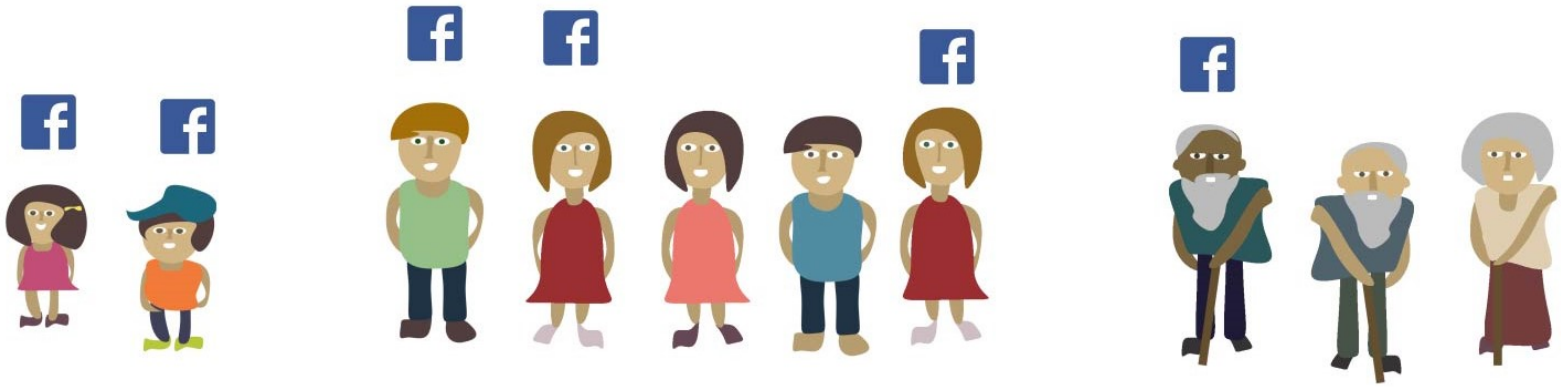
Dans l'exemple fictif, on partage la population en 3 strates (jeunes, adultes, personnes agées)

$$5000 = 1000 + 2500 + 1500$$

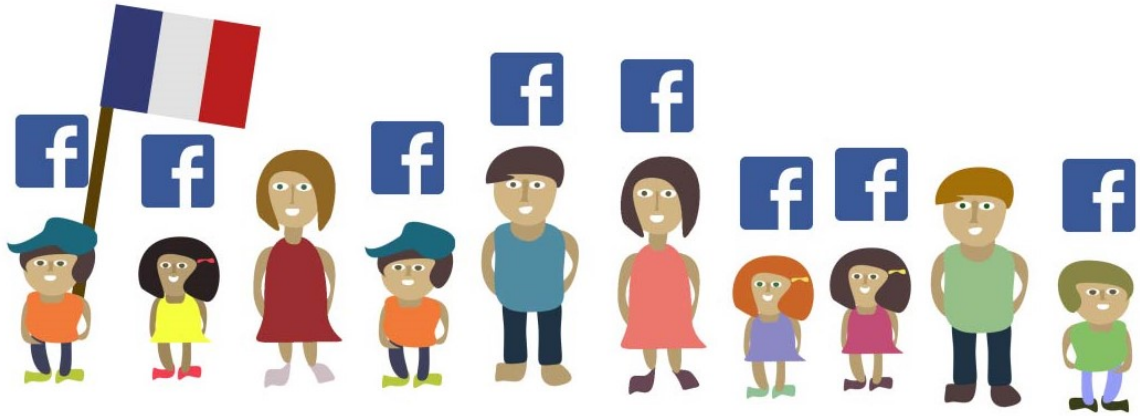
et on tire un échantillon de taille 10 avec une taille 2 dans la strate des jeunes, 5 dans la strate des adultes, 3 dans la strate des personnes agées.

On tire un échantillon de taille 10 avec une taille de 2 dans la strate des jeunes, de 5 dans la strate des adultes, de 3 dans la strate des personnes agées.

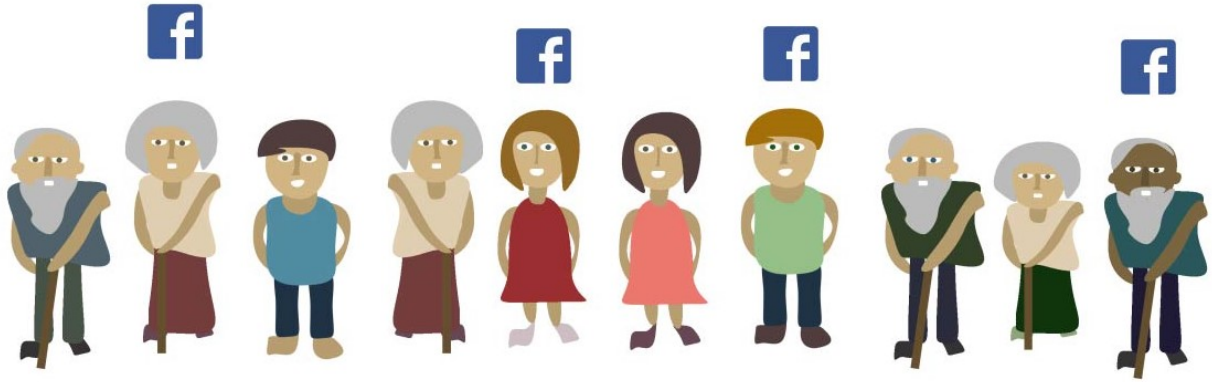
On obtient que 6 personnes sur les 10 ont un compte Facebook actif ce qui donne une estimation de 3000 personnes ayant un compte Facebook actif dans l'ensemble de la population.



La façon dont on tire un échantillon s'appelle le plan de sondage et on voit sur l'exemple fictif qu'avec un plan de sondage stratifié, on ne tirera jamais l'échantillon :

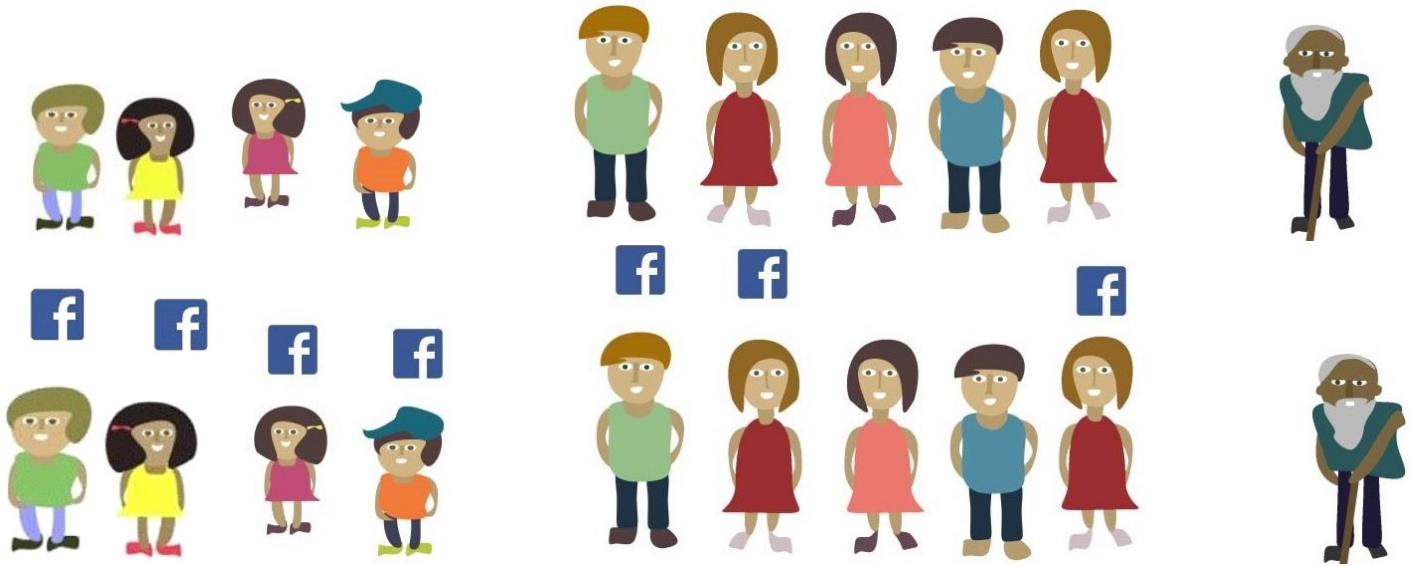


ni



Si on n'a pas utilisé un plan de sondage stratifié, on peut quand même améliorer l'estimation en changeant de méthode d'estimation. On peut utiliser notamment la post-stratification.

Si on a tiré l'échantillon :

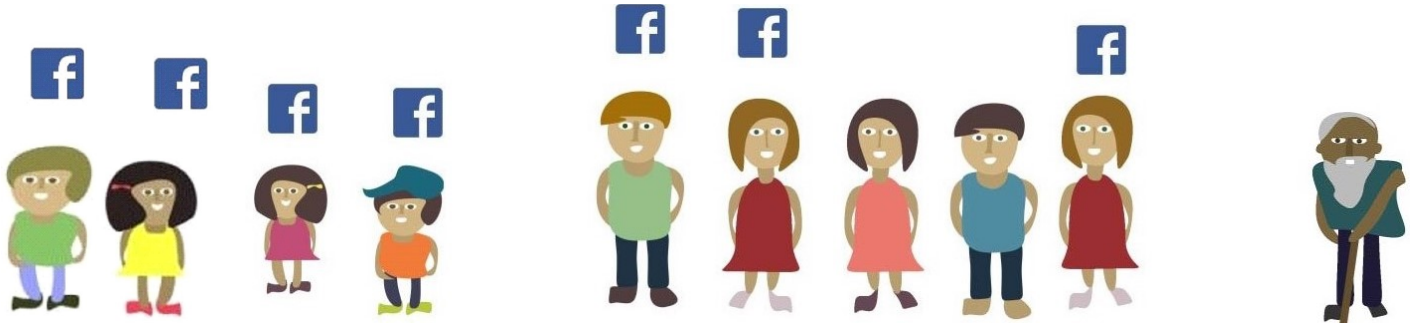


Si on utilise la méthode d'estimation précédente on obtient $7 \times 500 = 3500$.

Au lieu de donner le même poids de 500 à chacun des individus, on va donner un poids différent aux jeunes, aux adultes et à la personne agée.

On rappelle que :

On a tiré l'échantillon : $5000 = 1000 + 2500 + 1500$



On donne un poids de 250 aux jeunes, de 500 aux adultes, de 1500 à la personne agée. On obtient une estimation de :

$$4 \times 250 + 3 \times 500 = 2500.$$

Quelques remarques :

- Si on n'a pas utilisé un plan de sondage stratifié, on peut quand même améliorer l'estimation en changeant de méthode d'estimation. Si par exemple on constate que l'échantillon contient peu de personnes agées, on peut donner aux personnes agés présents dans l'échantillon un poids plus important qu'aux jeunes. On parle de méthode de redressement et dans l'exemple fictif de post-stratification.

- Pour stratifer ou post-stratifier, il faut savoir dans quelle strate se situe chacun des individus de l'échantillon. Il faut aussi connaître la taille des strates.

Certaines stratégies en sondages sont mises en place pour des raisons pratiques (faisabilité et coût) et pas pour amaliorer la qualité des estimations.

C'est le cas des plans à plusieurs degrés où on ne va pas directement échantillonner dans la population d'intérêt mais dans des bases de données agrégées. Par exemple pour l'enquête emploi en France, on sélectionne un certain nombre de "petites zones", appelées aires, ou grappes, et on interroge tous les logements des aires sélectionnées et tous les ménages de ces logements.

- L'intérêt de ce type de plan est qu'on a uniquement besoin de la base des logements pour les aires sélectionnées et qu'on peut limiter les coûts et diminuer les problèmes de non-réponse en plaçant des enquêteurs sur des-zones géographiques limitées.
- L'inconvénient est qu'on va interroger des ménages proches géographiquement et donc qui se ressemblent et les estimateurs seront moins précis.

1 Quelques concepts et intuitions sur les sondages

2 Introduction au formalisme mathématique de la théorie des sondages
- Eléments de contexte et notations
- Plan de sondage
- Estimateur de Horvitz-Thompson

3 Deux exemples de recherche en théorie des sondages

Eléments de contexte et notations

- Soit $U = \{1, \ldots, k, \ldots, N\}$ une population finie.
- Soit $Y = (y_1, \ldots, y_N)^T$ une variable d'étude.
- Soit $s \subset U$ un échantillon tiré aléatoirement selon un plan de sondage p.
- Un exemple de paramètre d'intérêt dans la population finie U est le total

$$t_Y = \sum_{k \in U} y_k$$

Remarque : comparée à d'autres domaines de la statistique, la théorie des sondages considère une population finie.

Plan de sondage

- L'objectif en théorie des sondages est de proposer une stratégie pour estimer un total si possible sans biais et donner une estimation de la précision de l'estimateur du total (intervalle de confiance).
- Inference statistique basée sur le plan : Y fixée.
- Plans de sondage en général sans remise.
- un échantillon tiré sans remis s : sous-ensemble de la population U (2^N possible).

Plan de sondage

Definition

Plan de sondage p sur U : distribution de probabilité sur l'ensemble de tous les sous-ensembles de U.

p défini par le statisticien avec des considérations sur la précision des résultats mais aussi sur la faisabilité et les coûts.

Un échantillon s est aussi un vecteur aléatoire d'indicatrices d'inclusion

$$(I_1, ..., I_k, ..., I_N)^\top \quad \text{où } I_k = \begin{cases} 1 & \text{if } k \in s \\ 0 & \text{otherwise.} \end{cases}$$

Deux exemples de plan de sondage

- Plan aléatoire simple sans remise SRSWOR(n) : taille d'échantillon fixée égale à n, et equiprobabilité sur tous les échantillons de taille n.

$$p(s) = \begin{cases} \dfrac{1}{\binom{N}{n}} & \text{if } s \text{ est de taille } n \\ 0 & \text{sinon.} \end{cases}$$

- Plan de Poisson $\mathcal{P}(\pi_1, \ldots, \pi_N)$: probabilités d'inclusion fixées $\pi_k \in]0; 1]$ avec $k \in U$, et I_k variables aléatoires independantes de loi de Bernoulli de paramètres π_k (taille d'échantillon aléatoire).

Probabilités d'inclusion

- Pour obtenir des estimateurs sans biais et calculer leur variance, on a uniquement besoin de connaître les probabilités d'inclusion d'ordre un et deux de p.

$$\pi_k = p(k \in s) = E(I_k) \ , \ k \in U : \text{probabilités d'inclusion d'ordre un,}$$

$$\pi_{kk'} = p(k \in s \text{ and } k' \in s) = E(I_k I_{k'}), \ k, k' \in U : \text{probabilités d'inclusion d'ordre deux.}$$

- A partir des probabilités d'inclusion d'ordre un et deux

$$Cov(I_k, I_{k'}) = \Delta_{kk'} = \pi_{kk'} - \pi_k \pi_{k'} \ k, k' \in U.$$

Remarque : les π_k sont fixées par le statisticien.

Estimateur de Horvitz-Thompson d'un total

- Considérons un plan de sondage p avec $\pi_k > 0$ et $\pi_{kk'} > 0$, $k, k' \in U$.

- Estimateur de Horvitz-Thompson (HT) de t_Y :

$$\hat{t}_Y = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in U} \frac{y_k}{\pi_k} I_k$$

$\hat{t}_Y$ est un estimateur sans biais de t_Y :

$$\mathrm{E}(\hat{t}_Y) = \sum_{k \in U} \frac{y_k}{\pi_k} \mathrm{E}(I_k) = \sum_{k \in U} \frac{y_k}{\pi_k} \pi_k = \sum_{k \in U} y_k = t_Y.$$

- Les $1/\pi_k$ sont appelés les poids de sondage.

Estimateur de Horvitz-Thompson d'un total

- Estimateur de Horvitz-Thompson (HT) : $\hat{t}_Y = \sum_{k \in s} \frac{y_k}{\pi_k} = \sum_{k \in U} \frac{y_k}{\pi_k} I_k.$

- Variance :

$$\text{Var}(\hat{t}_Y) = \sum_{k \in U} \sum_{k' \in U} y_k y_{k'} \, Cov\left(\frac{I_k}{\pi_k}, \frac{I_{k'}}{\pi_{k'}}\right) = \sum_{k \in U} \sum_{k' \in U} \Delta_{kk'} \frac{y_k}{\pi_k} \frac{y_{k'}}{\pi_{k'}}.$$

- Estimateur HT de la variance :

$$\hat{\text{Var}}(\hat{t}_Y) = \sum_{k \in s} \sum_{k' \in s} \frac{\Delta_{kk'}}{\pi_{kk'}} \frac{y_k}{\pi_k} \frac{y_{k'}}{\pi_{k'}}.$$

En résumé

Deux stratégies simples :

- Plan aléatoire simple sans remise et estimateur de Horvitz-Thompson.
- Plan de Poisson et estimateur de Horvitz-Thompson

De très nombreuses autres possibilités de stratégies existent en modifiant le plan de sondage (plan stratifié, à plusieurs degrés,...) ou en modifiant l'estimateur (post-stratifié, calé,...).

1. Quelques concepts et intuitions sur les sondages
2. Introduction au formalisme mathématique de la théorie des sondages
3. Deux exemples de recherche en théorie des sondages
   - Le projet Elfe et le plan produit
   - La Poste et le plan de sondage doublement indirect

- La recherche en sondages peut permettre de mieux comprendre certains résultats observés sur de grandes enquêtes nationales.
- Illustration à partir de deux projets de recherche récents :
  - **L'enquête logitudinale française depuis l'enfance (ELFE)** : plan de sondage produit (Cross Classified Sampling).
  - **L'enquête mensuelle du trafic postal de la poste française** : plan de sondage doublement indirect.

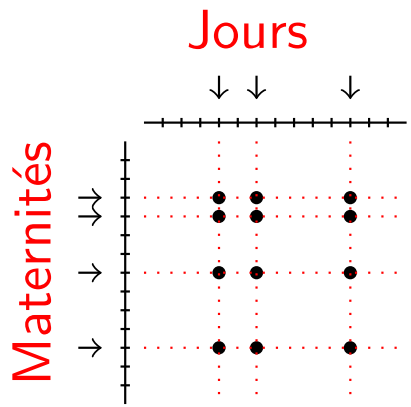
- Les propriétés théoriques des plans de sondage produit (CCS) et doublement indirect pas connues au moment de l'enquête.
- Ces plans ont été choisis pour des raisons pratiques (faisabilité, moindre coût).
- A première vue, les deux plans sont proches de plans plus classiques :
  - CCS ressemble à un plan **à deux degrés**.
  - le sondage doublement indirect ressemble au plan de sondage **indirect classique**.
- Nos recherches ont montré qu'en réalité ces plans de sondage peuvent conduire à des estimateurs peu précis comparés à des plans plus classiques.
- Seules des études théoriques avancées permettent de comprendre le comportement des estimateurs pour ce type de plans de sondage.

Elfe project

Etude Longitudinale Française depuis l'Enfance conduite par INED, INSERM and EFS.

- Plus de 18,000 enfants nés en 2011 sélectionnés sur la base de leur lieu et date de naissaince.
- Parents des bébés nés dans les maternités sélectionnées et les jours échantillonnés approchés par sage-femmes durant leur séjour à la maternité.
- Objectif : analyser la santé physique et mentale des enfants en fonction de conditions environnementales.
- Obtenir en particulier des intervalles de confiance pour l'estimation de paramêtres d'intérêt en prenant en compte le plan de sondage (nombre d'enfants nés par césarienne).
- Plan de sondage : produit (CCS).

Plan produit



sample of 320 maternités × sample of 25 jours

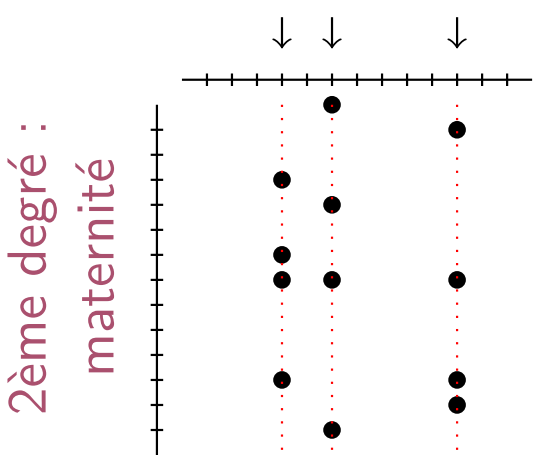
Un échantillon de 320 maternités tiré aléatoirement sur le territoire français métropolitain (parmi 544) et un échantillon de 25 jours répartis sur les 4 saisons et les différents jours de l'année 2011 (parmi 365).

Plan produit

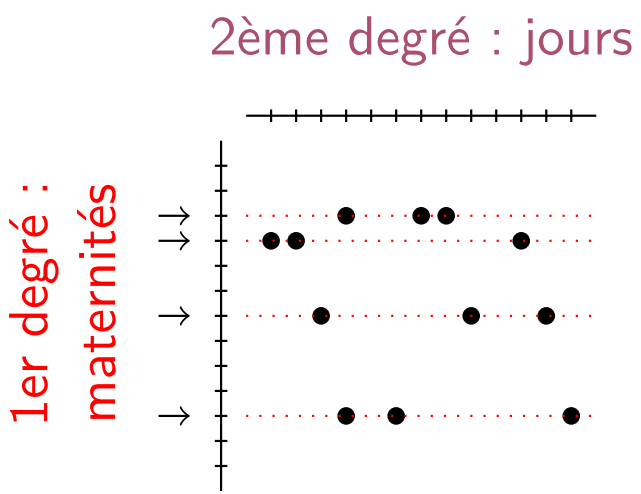
- De manière surprenante, la litérature sur le plan produit était très sporadique (Vos, 1964, Dalén and Ohlsson, 1995, Ohlsson, 1996, Skinner, 2015).
- La théorie générale n'avait jamais été développé pour ce plan.
- Dans Juillard et al. (2017), théorie générale développée et comparée avec les résultats du plan à deux degrés et proposition d'estimateurs de variance (positif).

Comparaison entre le plan produit et le plan à deux degrés

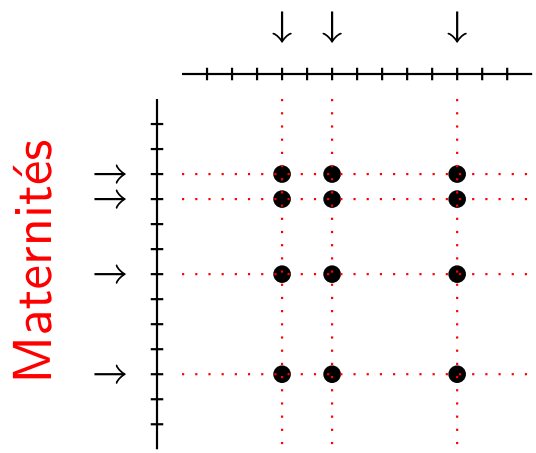
deux degrés DM



two-stage MD



CCS M $\times$ D



Comparaison entre le plan produit et le plan à deux degrés

Nous avons prouvé que :

- Pour des plans de taille fixe et des variables d'intérêt approximativement proportionnelles à la taille des maternités, CCS est moins précis que le deux degrés.
- Il n'est pas facile d'obtenir des estimateurs de variance qui sont toujours positifs pour le plan CCS.
- Le biais des estimateurs de variance pour le deux degrés peut être très grand si on utilise ces estimateurs avec un plan CCS.

Estimation de variance

Données Elfe

	Naissances	Né(e) par césarienne	Mère suivie par une sage-femme	Mère primipare
$\hat{t}_Y$	362924	33873	42337	162316
$\hat{\mathrm{V}}\left(\hat{t}_Y\right)$	7,6E+07	1,5E+07	3,9E+06	1,5E+07
$\mathrm{RD}\left(\hat{\mathrm{V}}_{\mathrm{SIMP1}}\right)$	-63,7 %	-95,5 %	-13,2 %	-46,5 %
$\mathrm{RD}\left(\hat{\mathrm{V}}_{\mathrm{SIMP2}}\right)$	-31,1 %	-1,9 %	-76,3 %	-41,4 %
$\mathrm{RD}\left(\hat{\mathrm{V}}_{\mathrm{SIMP3}}\right)$	5,2 %	2,6 %	10,5 %	12,2 %

où la différence relative RD entre $\hat{V}_{\mathrm{SIMP}}$ et l'estimateur sans biais $\hat{V}$ est :

$$RD = \frac{\hat{V}_{\mathrm{SIMP}i}\left(\hat{t}_Y\right) - \hat{V}\left(\hat{t}_Y\right)}{\hat{V}\left(\hat{t}_{Yr}\right)}.$$

La Poste

Pendant longtemps La Poste a estimé le traffic postal français par le biais d'une enquête portant sur la population des tournées de facteurs.

Tournées de Facteur

Population cible

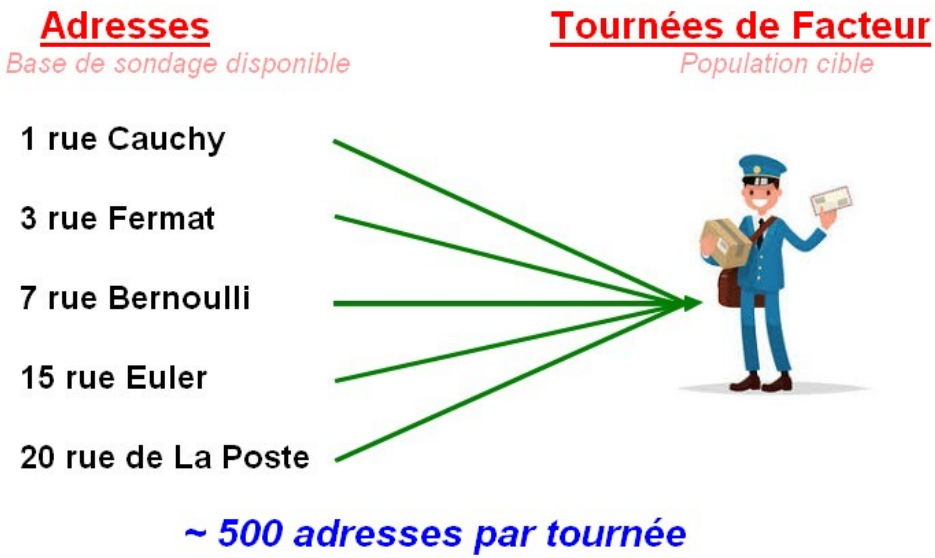
Exemple d'objectif : estimer le nombre de lettres postées en janvier 2022 en France métropolitaine.

La Poste

Depuis 2015, plus possible d'accéder à une base de sondage de la population des tournées de facteurs. Mais base d'adresses postales diponible.

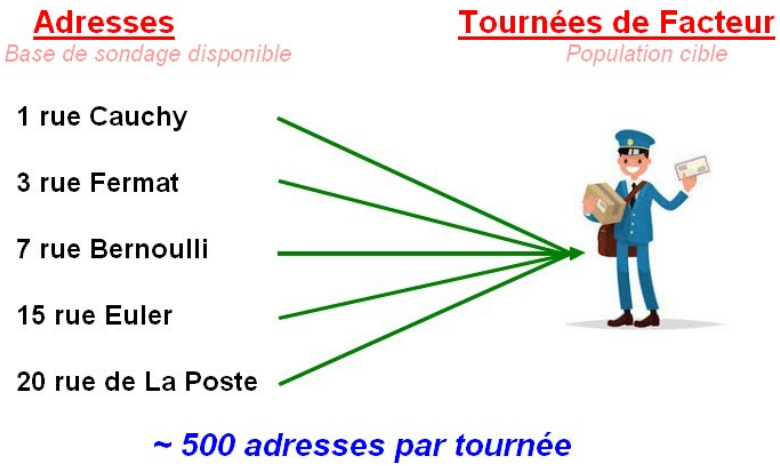
→ Tirage indirect d'un échantillon de tournées de facteurs via la population d'adresses et utilisation d'une méthode d'estimation dite de partage des poids.

→ trop coûteux en temps.

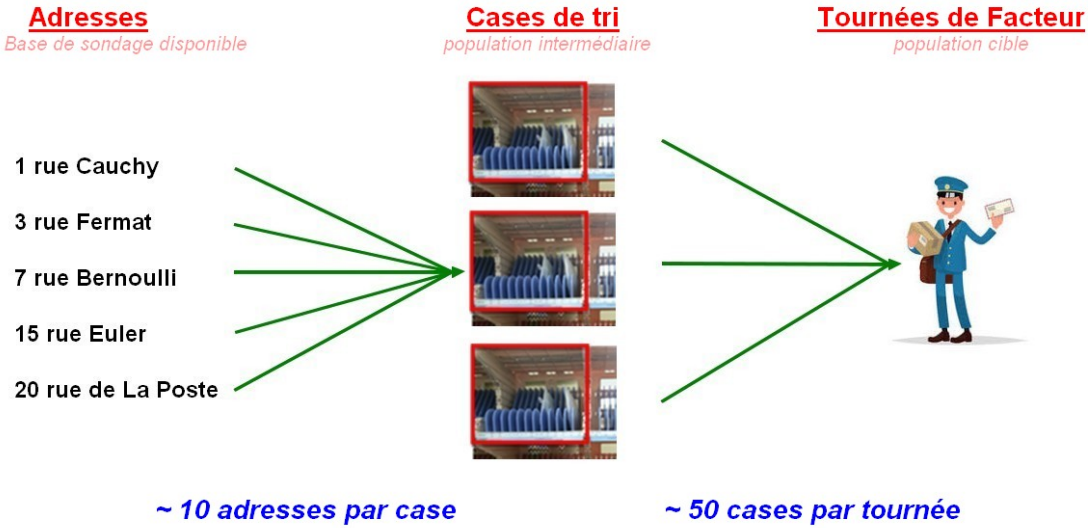


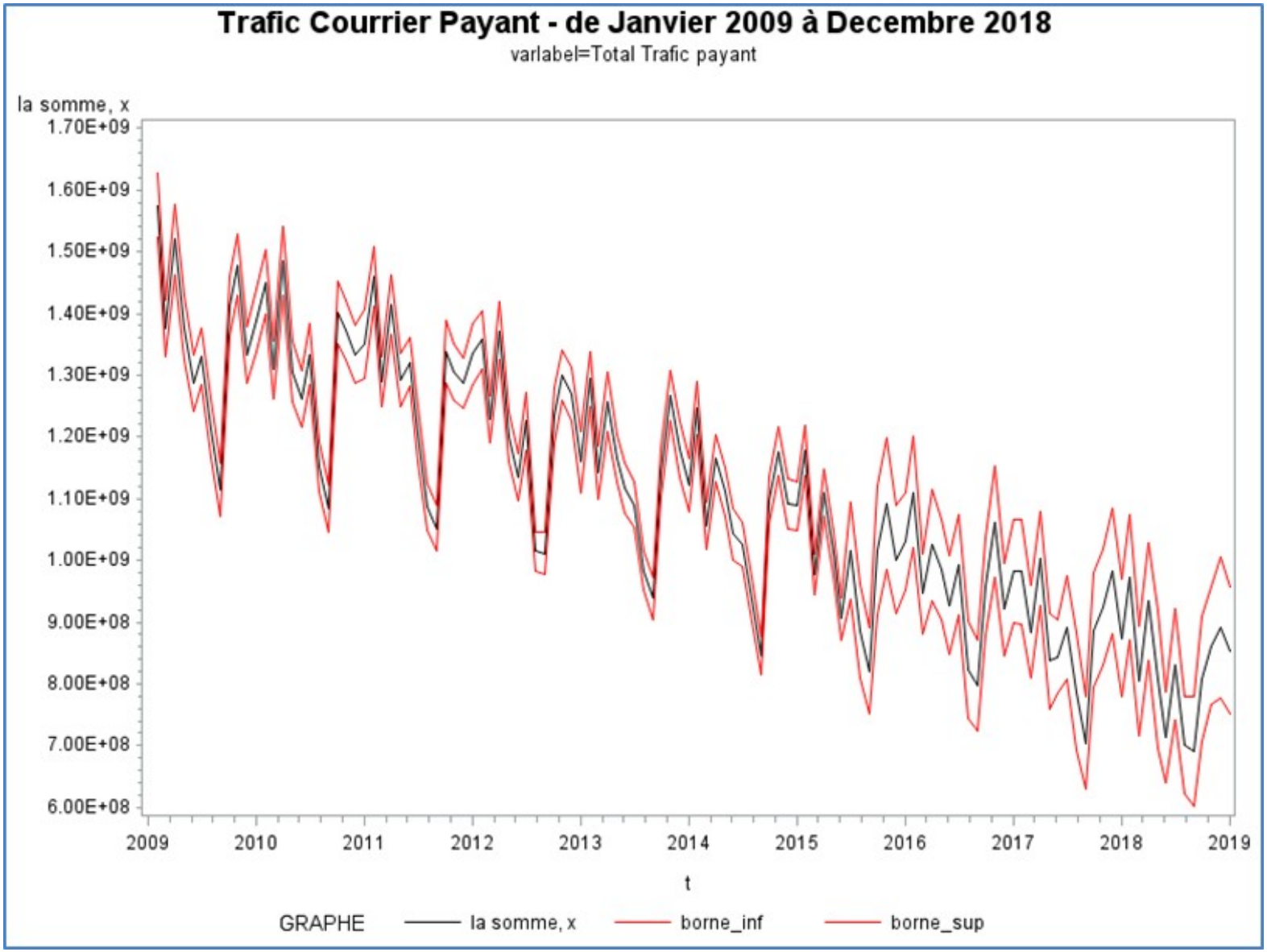
La Poste

Tirage indirect d'un échantillon de tournées de facteurs via la population d'adresses et utilisation d'une méthode d'estimation dite de partage des poids.
→ trop coûteux en temps.



→ La Poste utilise une deuxième population intermédiaire et un double sondage indirect.
Moins coûteux : 50 cases de tri par tournée et 10 adresses par case de tri.




Trafic Courrier Payant - de Janvier 2009 à Decembre 2018
varlabel=Total Trafic payant
la somme, x
1.70E+09
1.60E+09
1.50E+09
1.40E+09
1.30E+09
1.20E+09
1.10E+09
1.00E+09
9.00E+08
8.00E+08
7.00E+08
6.00E+08
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
t
GRAPHE
la somme, x
borne_inf
borne_sup

La Poste

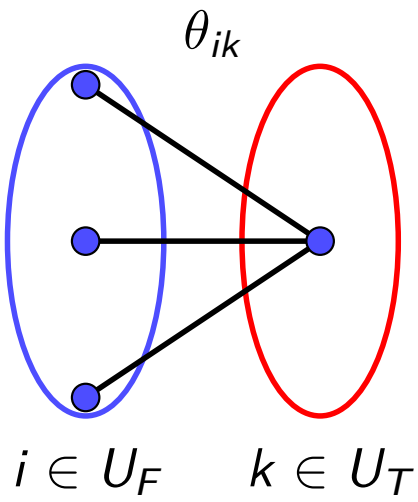
- Malheureusement perte importante de précision du sondage doublement indirect comparé au plan direct.

- Objectifs :
  - Comprendre cette perte de précision en comparant la variance de l'estimateur de Horvitz-Thompson pour estimer total entre sondage doublement indirect, sondage simplement indirect et sondage direct.
  - Améliorer la méthode d'estimation pour pallier cette perte de précision.

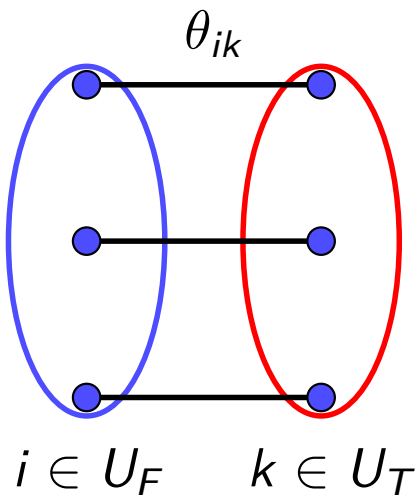
Liens entre populations

- Différents types de lien :

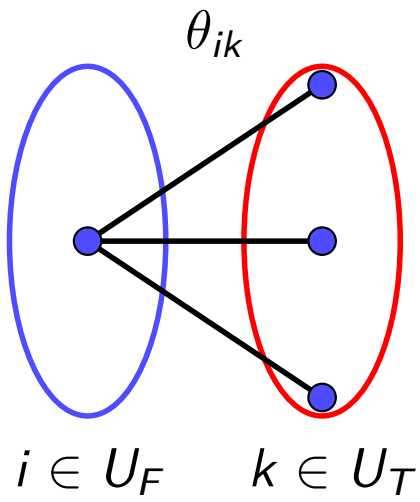
Tous pour Un



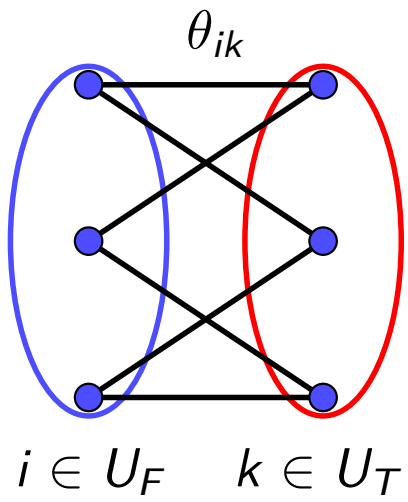
Un pour Un



Un pour Tous



Tous pour Tous



- A La Poste uniquement des liens Tous pour Un (MTO).
- Les liens sont pondérés par θ_{ik}, $i \in U_F$, $k \in U_T$.

Plan de sondage indirect simple

- θ_{ik} : poids de lien entre $i \in U_F$ et $k \in U_T$ tels que

$$\theta_{ik} = \begin{cases} > 0 & \text{si } i \text{ et } k \text{ sont liés,} \\ 0 & \text{sinon.} \end{cases}$$

- Poids de lien standardisés $\tilde{\theta}_{ik} = \dfrac{\theta_{ik}}{\sum_{i' \in U_F} \theta_{i'k}}$. Nous avons :

$$\sum_{i \in U_F} \tilde{\theta}_{ik} = 1 \text{ pour tout } k \in U_T$$

- Grâce à la standardisation des liens, le total t_Y se réécrit comme un total sur U_F :

$$t_Y = \sum_{k \in U_T} y_k = \sum_{k \in U_T} y_k \sum_{i \in U_F} \tilde{\theta}_{ik} = \sum_{i \in U_F} \tilde{y}_i \text{ avec } \tilde{y}_i = \sum_{k \in U_T} y_k \tilde{\theta}_{ik}.$$

Plan indirect simple

- Méthode de partage des poids (GWSM) pour estimer t_Y : estimateur de Horvitz-Thompson sur U_F (Deville and Lavallée, 2006 and Lavallée, 2009) :

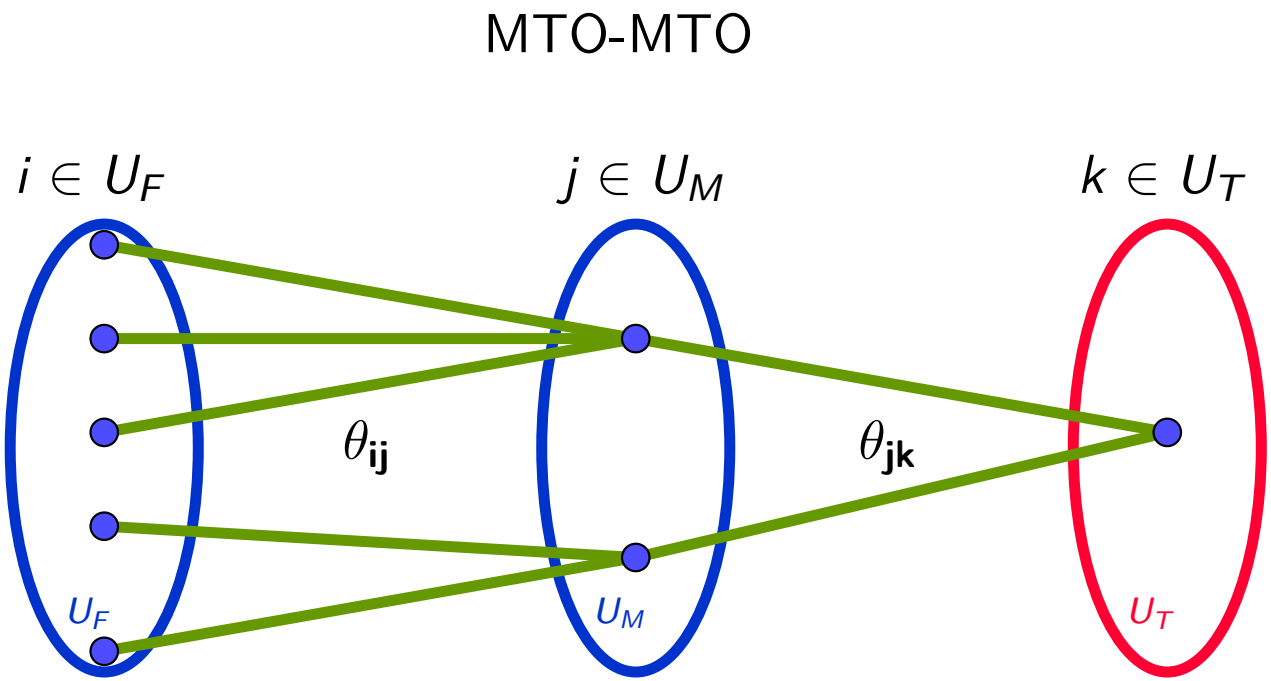
$$\hat{t}_{Y,GWSM} = \sum_{i \in s_F} \tilde{y}_i / \pi_i \quad \text{avec} \quad \tilde{y}_i = \sum_{k \in U_T} y_k \tilde{\theta}_{ik} \text{ et } \pi_i = \mathrm{P}(i \in s_F) > 0.$$

- Possibilité de choisir θ_{ik} en minimisant la variance de $\hat{t}_{Y,GWSM}$ sous contrainte de standardisation.

- Le choix le plus simple et le plus commun est $\theta_{ik} = 1$ (optimal pour le plan SRSWOR).

Plan indirect double

- Restriction au cas MTO-MTO case.
- Soit $\theta_{ij} > 0$ le poids de lien entre $i \in U_F$ et $j \in U_M$ et $\theta_{jk} > 0$ le poids de lien entre $j \in U_M$ et $k \in U_T$



Plan indirect double

- $\tilde{\theta}_{ij}$ and $\tilde{\theta}_{jk}$: poids de lien standardisés :

$$\sum_{i \in U_F} \tilde{\theta}_{ij} = 1 \text{ pour tout } j \in U_M \text{ et } \sum_{j \in U_M} \tilde{\theta}_{jk} = 1 \text{ pour tout } k \in U_T.$$

$$t_Y = \sum_{k \in U_T} y_k = \sum_{i \in U_F} \tilde{y}_i^D \text{ avec } \tilde{y}_i^D = \sum_{j \in U_M} \tilde{\theta}_{ij} \sum_{k \in U_T} y_k \tilde{\theta}_{jk}.$$

- La méthode de partage des poids double (dGWSM) pour estimer t_Y :

$$\hat{t}_{Y,dGWSM} = \sum_{i \in s_F} \frac{\tilde{y}_i^D}{\pi_i} \text{ where } \tilde{y}_i^D = \sum_{j \in U_M} \tilde{\theta}_{ij} \sum_{k \in U_T} y_k \tilde{\theta}_{jk}$$

Comparaison le plan indirect simple et double

MTO pour GWSM et MTO-MTO pour dGWSM.

- dGWSM nécessite moind d'information sur les liens que GWSM.

  La Poste : 500 adresses pour le simple et 60 pour le double.

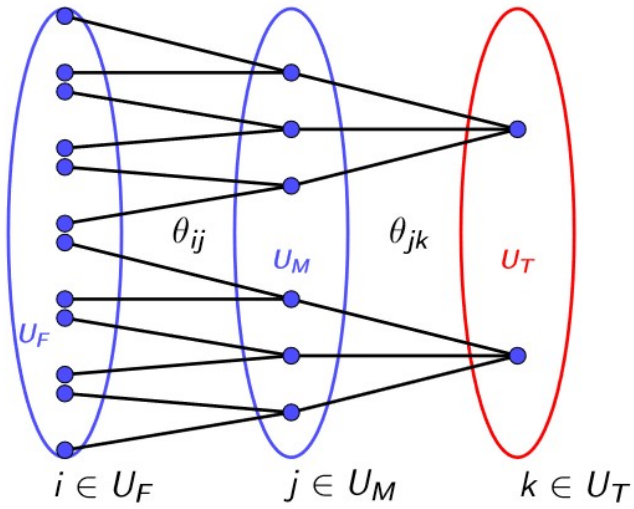
- Comparaison de variances entre GWSM et dGWSM :

  Pour SRSWOR, Poisson et plan stratifié, dans Medous et al. (2023), expressions explicite et simple de la difference entre la variance de dGWSM et la variance de GWSM poids de lien optimaux (différence toujours positive).

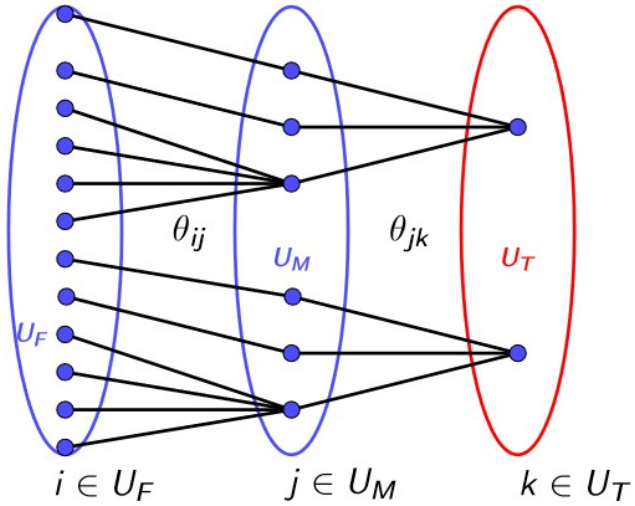
- Permet de comprendre la perte de précision qui dépend de la structure des liens entre populations U_F et U_M et U_M et U_T.

- Simulations à partir de données historiques de La Poste pour plan SRSWOR de taille n et poids de lien égaux à 1.

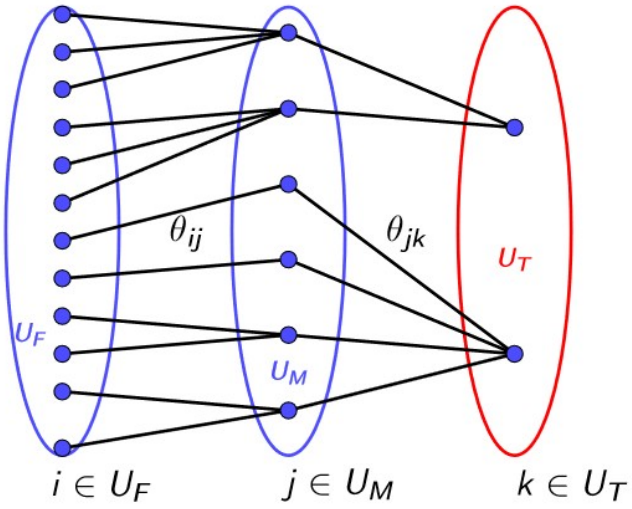
1. θ_{ij} & θ_{jk} uniform



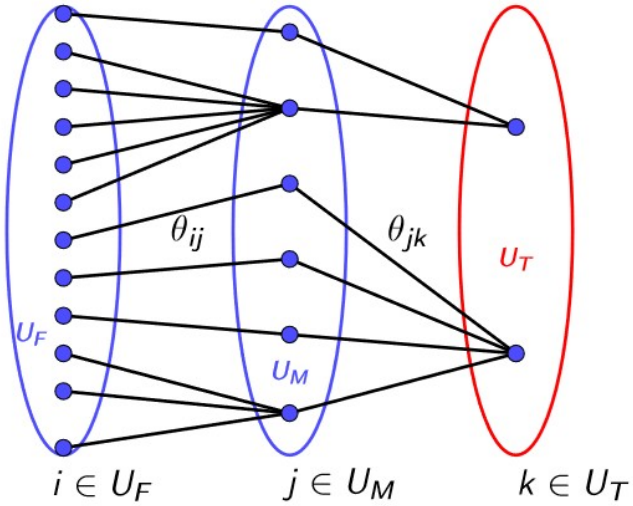
2. θ_{ij} not uniform & θ_{jk} uniform



3. θ_{ij} uniform & θ_{jk} not uniform



4. θ_{ij} & θ_{jk} not uniform



1. θ_{ij} & θ_{jk} uniformes

n	RRMSE
500	1.00
1000	1.00

60 liens à observer en moyenne par unité k dans U_T.

2. θ_{ij} non uniforme & θ_{jk} uniforme

n	RRMSE
500	7.19
1000	7.14

60 liens à observer en moyenne par unité k dans U_T.

$\sum_{j \in U^M} \tilde{\theta}_{ij}\tilde{\theta}_{jk} = \tilde{\theta}_{ik} \Rightarrow$ dGWSM=GWSM.

3. θ_{ij} uniforme & θ_{jk} non uniforme

n	RRMSE
500	1.00
1000	1.00

314 lens à observer en moyenne par unité k dans U_T.

4. θ_{ij} & θ_{jk} non uniformes

n	RRMSE
500	11.78
1000	11.55

314 liens à observer en moyenne par unité k dans U_T.

Pour GWSM, 500 liens à observer en moyenne par unité k dans U_T.

Conclusion

- Pour les deux exemples de recherche perte de précision non anticipée.
- Seule une étude théorique approfondie peut permettre compréhension claire du comportement des méthodes (ici plan de sondage) qui sont pertinentes d'un point de vue practique.
- Intérêt des allers retours entre pratique et théorie.
- Domaine de la théorie des sondages vaste avec nombreuses problématiques statistiques intéressantes.

Merci de votre attention !