Vidéo 18 sur 21 : La pondération
Correction de la non-réponse

l'échantillonnage

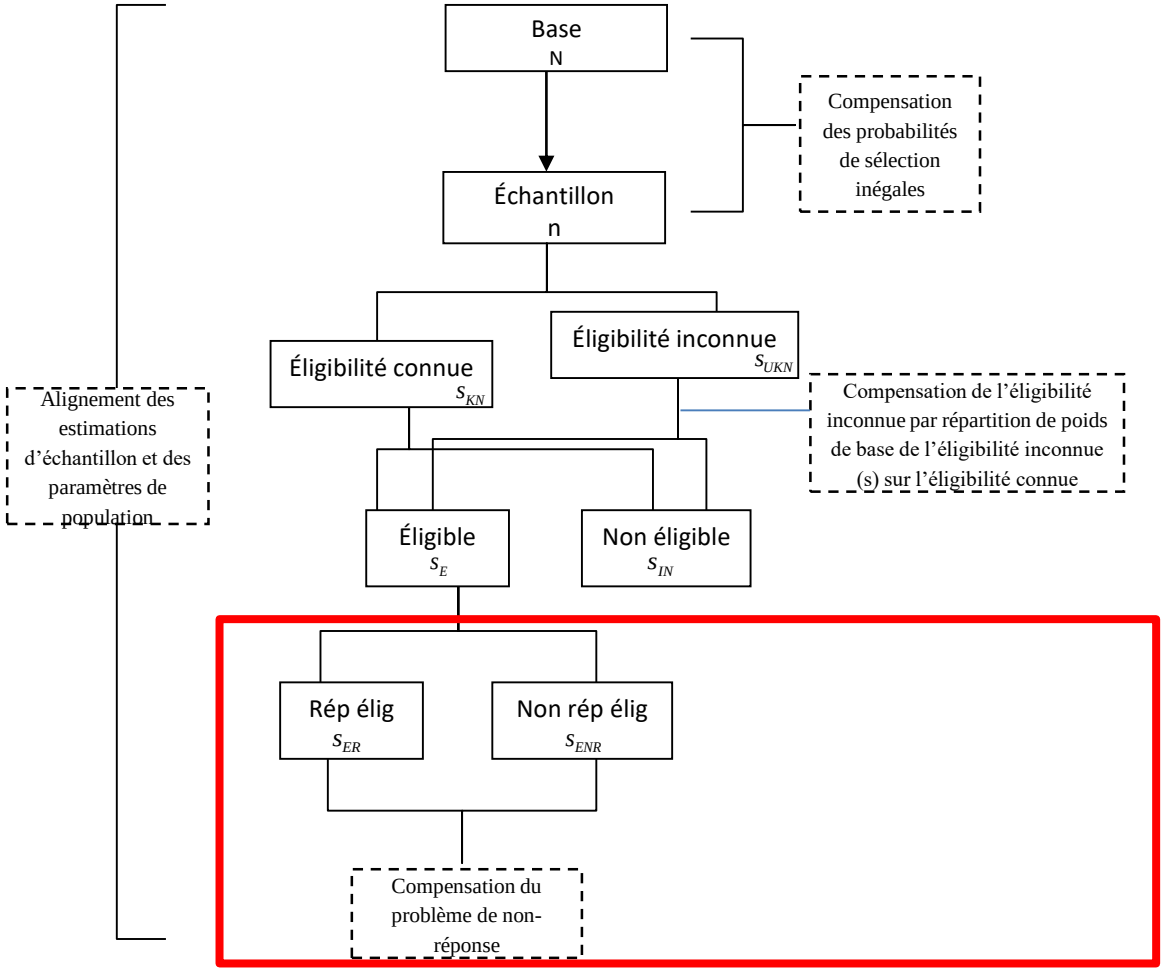
THE WORLD BANK

MANNHEIM BUSINESS SCHOOL

Pondération de la non-réponse (I)

- Dans une enquête idéale, toutes les unités de la population sont incluses dans la base d'échantillonnage, et toutes les personnes de l'échantillon participent à l'enquête
  - Mais en pratique, aucune de ces conditions ne se réalise
- Certaines unités ne sont pas incluses dans la base (sous-couverture), et certaines ne répondent pas (non-réponse)
- Une pondération de non-réponse est généralement appliquée pour réduire le biais lorsque certaines unités de l'échantillon ne répondent pas (non-réponse de l'unité)
- L'ensemble de répondants *r* peut être considéré comme un sous-ensemble de l'échantillon *s*

Grandes étapes de la pondération : Pondération de la non-réponse



Pondération de la non-réponse (II)

- Définissons les indicateurs d'échantillonnage et de réponse suivants :

$$I_i = \begin{cases} 1, \text{ si l'unité } i \textit{ est séléctionnée dans l'échantillon} \\ 0, \text{dans } \textit{le cas contraire} \end{cases}$$

$$R_i = \begin{cases} 1, \textit{ si l'unité répond} \\ 0, \text{dans le cas } \textit{contraire} \end{cases}$$

- La probabilité de faire partie de l'échantillon *s* est $P(I_i = 1) = \pi_i$
- La probabilité de répondre (faire partie de l'ensemble de répondants *r*) dans la mesure où l'unité *i* appartient à l'échantillon *s* est $P(R_i = 1 | I_i = 1) = \phi_i$

Pondération de la non-réponse (III)

- On utilise l'inverse de la probabilité de sélection, π_i^{-1}, pour ajuster les différentes probabilités de sélection
- On utilise l'inverse de la probabilité de réponse, ϕ_i^{-1}, pour les corrections de la non-réponse
  - La pondération finale serait alors $(\pi_i \times \phi_i)^{-1}$
- Il n'est toutefois pas possible d'employer la probabilité de réponse réelle ϕ_i.
  - On utilise à la place l'estimation $\hat{\phi}_i$ de la probabilité de réponse
- La méthode employée pour estimer ces propensions à répondre dépendra du nombre et du type de variables auxiliaires disponibles pour les *répondants* *ET* les non-répondants, et des hypothèses sur les données manquantes
- Pour la terminologie sur les données manquantes, voir Little & Rubin (2019)

Pondération de la non-réponse (IV)

- En l'absence de toute variable auxiliaire, l'approche la plus simple consiste à se baser sur un mécanisme MCAR (Missing Completely at Random) dans lequel il est possible d'estimer la propension à répondre à l'aide du taux global de réponse, $\hat{\phi}_i = RR$

Pondération de la non-réponse (V) Ajustement par classe

- L'ajustement par classe part du principe qu'il est possible de créer des classes dont toutes les unités présentent la même probabilité de réponse ou une valeur « *y* » quasiment identique -- mécanisme MAR (*Missing at Random*)
  - Il est possible de modéliser la probabilité de réponse estimée $\hat{\phi}_i$ si on dispose d'un ensemble de variables auxiliaires $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ pour chaque unité d'échantillon, qu'il y ait réponse ou non
- Les variables auxiliaires peuvent être :
  - Des variables démographiques, comme le sexe, l'âge, l'origine ethnique et le niveau d'éducation,
  - Des variables de conception ou de base, comme la région
  - Des variables observées au cours de la collecte de données (paradonnées)
- Dans l'idéal, ces covariables devraient être liées à la *propension à répondre* ET *à la valeur* « y » mesurée

Pondération de la non-réponse (VI) Ajustement par classe

- On peut calculer le facteur d'ajustement de la non-réponse pour les unités de la classe *c* en prenant l'inverse du taux de réponse non pondéré :

$$a_{4,c} = \frac{n_{c,E}}{n_{c,ER}}$$

- Ou l'inverse du taux de réponse pondéré :

$$a_{4,c} = \frac{\sum_{i \epsilon s_{c,E}} d_{3i}}{\sum_{i \epsilon s_{c,ER}} d_{3i}}$$

- Le poids de non-réponse est alors le produit du poids de sondage et du facteur d'ajustement de non-réponse :

$$d_{4i} = d_{3i} \times a_{4i}$$

Pondération de la non-réponse (VII)
Exemple d'ajustement par classe

- Considérons les régions comme des classes pour l'ajustement par classe de la non-réponse, et observons la répartition suivante des réponses :

Région	$n_{c,ER}$	$n_{c,ENR}$	$n_{c,E} = n_{c,ER} + n_{c,ENR}$	$a_{4,c} = n_{c,E}/n_{c,ER}$
Ouest	60	38	98	98/60 = 1,6333
Midwest	73	34	107	107/73 = 1,4658
Nord-est	98	63	161	161/98 = 1,6429
Sud	57	42	99	99/57 = 1,7368

Pondération de la non-réponse (VIII) Ajustement du score de propension

- Modélisez la non-réponse à l'aide de la régression logistique, de modèles probit ou à logarithme double complémentaire
- L'indicateur de réponse R_i fonctionne comme une variable dépendante, et les variables auxiliaires disponibles (pour les répondants et les non répondants) fonctionnent comme des variables indépendantes
- On peut écrire la probabilité de réponse estimée de la manière suivante

$$\hat{\phi}(\boldsymbol{x}_i) = \frac{exp\left(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}\right)}{1 + exp\left(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}\right)}$$

- où $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})$ est le vecteur de variables indépendantes et $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)$ représentent les coefficients estimés de régression logistique

Pondération de la non-réponse (IX)
Ajustement du score de propension

- Après estimation de la probabilité de réponse, on peut choisir entre :
  - la pondération par la propension : en prenant directement l'inverse de $\hat{\phi}_i$ pour l'ajustement de la pondération
    - Cette méthode ajoute de la dépendance au modèle de propension à répondre
  - la stratification par la propension : en prenant $\hat{\phi}_i$ pour créer des classes d'ajustement (Little, 1986) :
    - Après estimation de la propension à répondre $\hat{\phi}(\boldsymbol{x}_i)$, on trie le fichier par ordre croissant en fonction de $\hat{\phi}(\boldsymbol{x}_i)$
    - On forme ensuite des classes avec environ le même nombre d'unités d'échantillon (répondants et non répondants) initial dans chaque classe
    - Répartir la variable $\hat{\phi}(\boldsymbol{x}_i)$ de propension à répondre par quintiles ou déciles peut être une bonne technique de regroupement

Pondération de la non-réponse (X) Ajustement du score de propension

- la stratification par la propension : en prenant $\hat{\phi}_i$ pour créer des classes d'ajustement (cont.) :
  - Après la création des classes, il y a plusieurs possibilités pour calculer un ajustement unique dans chaque classe *c* :
    - la propension moyenne estimée non pondérée : $\hat{\phi}_i = \sum_{i \epsilon s_c} \hat{\phi}(\boldsymbol{x}_i)/n_c$; où n_c représente le nombre non pondéré de cas dans la classe *c*
    - la propension moyenne estimée pondérée : $\hat{\phi}_i = \sum_{i \epsilon s_c} d_i \hat{\phi}(\boldsymbol{x}_i)/\sum_{i \epsilon s_c} d_i$; où d_i est le poids d'entrée de l'étape NR et $\sum_{i \epsilon s_c} d_i = \hat{N}_c$, le nombre estimé d'unités de population dans la classe *c*
    - le taux de réponse non pondéré *:* $\hat{\phi}_i = n_{cR}/n_c$; où n_{cR} représente le nombre non pondéré de répondants de la classe *c*
    - le taux de réponse estimé pondéré : $\hat{\phi}_i = \sum_{i \epsilon s_{cR}} d_i / \sum_{i \epsilon s_c} d_i$
    - la propension médiane estimée non pondérée : $\hat{\phi}_i = median\left[\hat{\phi}(\boldsymbol{x}_i)\right]_{i \in s_c}$

FIN DE LA VIDÉO 18