TRAITEMENT DE NON REPONSE TOTALE

Presenté par :
Crépin MEDEHOUIN

Formateur : Dr. KOUAME Darès

PLAN

INTRODUCTION

Introduction

On dit qu'il y a non-réponse vis-à-vis de la variable Y pour l'individu échantillonné i dès lors que l'on ne dispose pas de la valeur Y_i relative à cet individu.

On distingue deux types de non réponse :

1. les non réponses totales
2. les non réponses partielles

Introduction

La **non-réponse totale** est habituellement traitée par une méthode de **repondération**:

1. on supprime du fichier les non-répondants totaux,
2. on augmente les poids des répondants pour compenser de la non-réponse totale.

La non-réponse partielle est habituellement traitée par imputation
$\Rightarrow$ une valeur manquante est remplacée par une valeur plausible.

Introduction

L'objectif prioritaire est de **réduire** autant que possible **le biais de non-réponse** : cela passe par une recherche des facteurs explicatifs de la non-réponse

Non réponses totales

On a une non réponse totale lorsque l'on n'a aucune donnée sur l'unité d'observation.

Autrement dit, on a aucune réponse aux questions posées. Cela ne signifie pas que l'on ne dispose aucune information sur le non-répondant :

En général, on dispose tout de même de renseignements présents dans la base de sondage, ou collectés sur le terrain (Par exemple des renseignements obtenus auprès de tierces personnes)

Non réponses totales

On va traiter ce problème par **repondération** : on fait porter aux répondants le poids des non-répondants. Cette repondération se justifie sous une modélisation du mécanisme de non-réponse.

Cette modélisation permet d'estimer les probabilités de réponse à l'enquête, pour obtenir les poids corrigés de la non-réponse totale.

Quelques facteurs de non-réponse totale (Haziza, 2011)

- Mauvaise qualité de la base de sondage;
- Impossibilité de joindre l'individu;
- Type d'enquête (obligatoire ou volontaire);
- Fardeau de réponse;
- Méthode de collecte (interview, téléphone, courrier, ...);
- Durée de collecte;
- Suivi (et relance) des non-répondants;
- Formation des enquêteurs.

Hors champs

Un point important : la distinction entre individus hors-champ et individus non-répondants

On parle de hors champs quand une valeur manquante est dû au fait que l'enquêté n'est pas concerné par la question :

1. S'il concerne toute les variables, elle est total (HCT)
2. Ou partiel s'il concerne quelques variables (HCP)

La non réponses modifie les poids de sondage alors que le hors champs ne les modifie pas.

PRINCIPE

Les étapes du traitement de la non-réponse totale

1. Identification des non-répondants,
2. Modélisation du mécanisme de non-réponse (recherche des facteurs explicatifs),
3. Estimation des probabilités de réponse,
4. Calcul des poids corrigés de la non-réponse totale.

Modélisation du mécanisme de non-réponse

On note r_k la variable indicatrice de réponse pour l'individu k, valant 1 si l'individu a répondu à l'enquête et 0 sinon.

$$r_k = \begin{cases} 1 & \text{si l'individu } k \text{ a répondu,} \\ 0 & \text{sinon.} \end{cases}$$

On note $p_{k|S} \equiv p_k$ la probabilité de réponse pour l'unité k :

$$p_k = \Pr(k \in S_r \mid S) = \Pr(r_k = 1 \mid S).$$

Modélisation du mécanisme de non-réponse

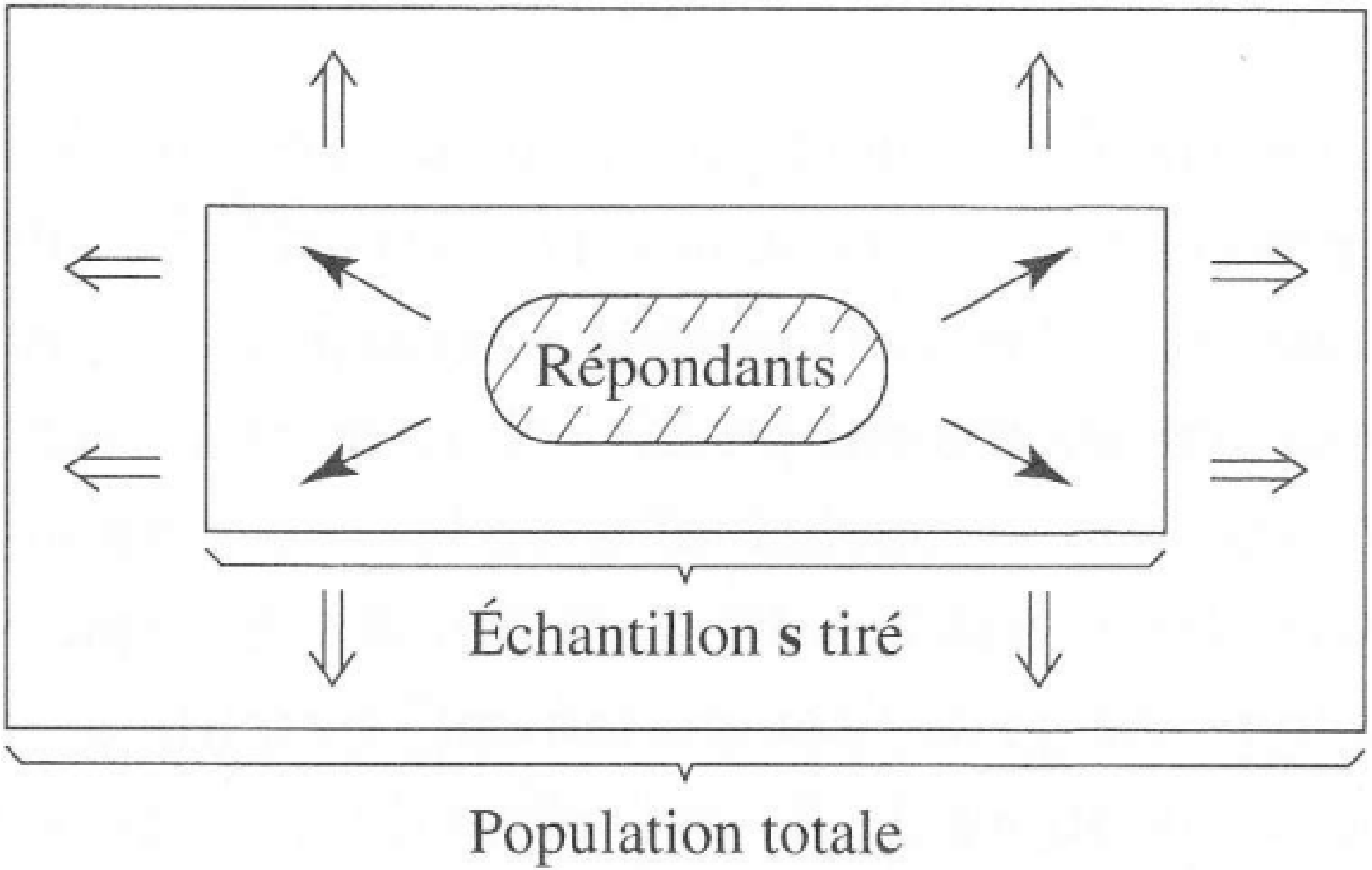
Si on veut estimer le total T d'une variable dans l'échantillon total, T est estimé sans biais par l'estimateur de **Horvitz-Thompson** :

$$\hat{T} = \sum_{k \in S} w_k y_k = \sum_{k \in S} \frac{y_k}{\pi_k}$$

avec $w_k = \frac{1}{\pi_k}$ le poids de sondage de l'unité k

En présence de non réponse totale, on obtient un estimateur sans biais du total T :

$$\hat{T}_r = \sum_{k \in S_r} \frac{y_k}{\pi_k \times p_k}$$


Répondants
Échantillon s tiré
Population totale

Hypothèses

On fait l'hypothèse que :

- toutes les probabilités de réponse vérifient $0 < p_k \leq 1$: pas de non-répondants irréductibles,
- les individus répondent indépendamment les uns des autres :

$$\Pr(k, l \in S_r \mid S) \equiv p_{kl} = p_k p_l.$$

Cette dernière hypothèse peut être affaiblie (Haziza et Rao, 2003 ; Skinner et D'Arrigo, 2011).

Types de mécanisme

On distingue schématiquement trois types de mécanisme de non-réponse :

1. uniforme (ou MCAR),
2. ignorable (ou MAR),
3. non-ignorable (ou NMAR).

Méthode de repondération uniforme

Mecanisme uniforme (ou MCAR)

Le mécanisme est dit uniforme (ou Missing Completely At Random) quand $p_k = p$, i.e. quand tous les individus ont la même probabilité de réponse.

C'est une hypothèse généralement peu réaliste.

Exemple : non-réponse provenant de la perte de questionnaires.

Exemple

Prenons l'exemple d'une population de **10 personnes**, dans laquelle on tire **5 individus** par sondage aléatoire simple.

Supposons (cas d'école) que y_k soit égal à 1 pour chacun des 10 individus, et que, **parmi les 5** personnes tirées, seulement **2 acceptent** de répondre. On a:

$N = 10, \quad n = 5 \quad$ et $\quad \pi_k = f = \frac{n}{N} = \frac{1}{2} \quad$ pour tout k

Le vrai total est T = 10. Si on ne corrige pas des non-réponses, on utilise naivement :

$$\hat{T} = \sum_{k \in S_r} \frac{y_k}{\pi_k} = \frac{1}{\frac{1}{2}} + \frac{1}{\frac{1}{2}} = 4$$

Exemple

Puisqu'on ne dispose des valeurs y_k que pour deux individus. T est manifestement un mauvais estimateur de T, fortement biaisé.

Pour limiter ce biais, on peut cependant raisonner ainsi :

sur les cinq personnes tirées, puisque deux répondent, on peut penser que chacune a une probabilité $r_k = \frac{2}{5}$ de répondre.

$$\hat{T}_r = \sum_{k \in S_r} \frac{y_k}{\pi_k \times p_k} = \frac{1}{\frac{1}{2} \times \frac{2}{5}} + \frac{1}{\frac{1}{2} \times \frac{2}{5}} = 10$$

L'idée qui vient d'être appliquée consiste à estimer la probatrilité de réponse supposée commune par un taux de réponse empirique.

Mecanisme uniforme (ou MCAR)

La méthode de repondération uniforme est facile à mettre en œuvre et il n'y a pas de calcul complexe pour la probabilité de réponse et le fait d'attribuer la même probabilité de réponse à chaque individu garantit un traitement équitable pour toutes les observations.

Cependant, cette méthode ne tient pas en compte l'hétérogénéité des données

Mécanisme de réponse homogène

Mécanisme de non-réponse ignorable (MAR)

On parle de mécanisme de non-réponse ignorable (ou Missing At Random) quand les probabilités de réponse peuvent être expliquées à l'aide de l'information auxiliaire disponible :

$$\Pr(r_k = 1 \mid y_k, z_k) = \Pr(r_k = 1 \mid z_k),$$

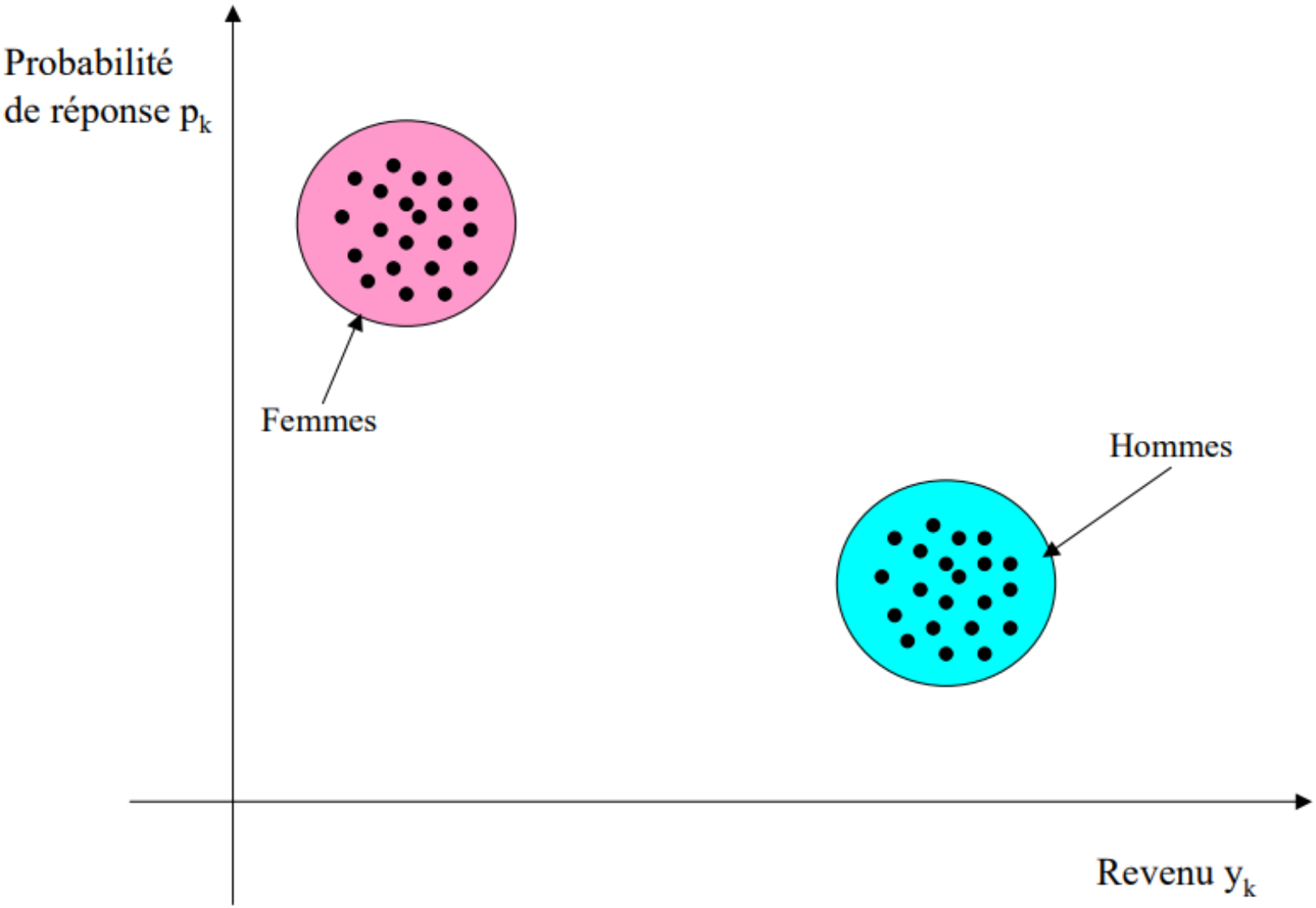
avec

- y_k la variable d'intérêt,
- z_k le vecteur des valeurs prises par un vecteur $\mathbf{z}$ de variables auxiliaires pour l'individu k de S.

Exemple : enquête sur le revenu + non-réponse expliquée par le sexe des individus.

Mécanisme de non-réponse ignorable (MAR)

Cette méthode consiste à repartir les unités de la population en sous-groupes disjoints appelés groupes de réponses homogènes. A l'intérieur de chaque groupe, les individus ont des comportements de réponse indépendants et une probabilité de réponse commune et non nulle. La probabilité de réponse est égale au rapport des unités répondants dans un groupe par le nombre d'unités dans le groupe.


Probabilité
de réponse p_k
Femmes
Hommes
Revenu y_k

Exemple

Dans une enquête sur le revenu, avec un échantillon initial de 10 000 personnes et un échantillon répondant de taille 8 000, il serait grossier d'estimer r_k par 80% pour chaque individu.

Une solution plus satisfaisante consisterait à considérer les catégories socioprofessionnelles (CSP) si on en dispose pour chaque individu tiré, répondant ou non (ou, mieux, les croisements CSP - âge si l'information existe), et à estimer un taux de réponse par catégorie.

En effet, il est très probable que les taux de réponse différeront fortement selon la catégorie, et on pouffa ainsi obtenir des estimations des r_k plus proches dela réalité.

Exemple

Par exemple, supposons que l'on obtienne :

	Échantillon tiré	Échantillon répondant	Taux de réponse estimé (%)
CSP à revenu élevé	2 000	1 000	50
CSP à faible revenu.....	8 000	7 000	87,5
TOTAL	10 000	8 000	80

Ainsi, si S_{r_1} désigne l'échantillon de répondants parmi les CSP à fort revenu (respectivement S_{r_2}, parmi les CSP à faible revenu), on utilisera :

$$\hat{T}_r = \sum_{k \in S_{r_1}} \frac{y_k}{\pi_k \times 0.5} + \sum_{k \in S_{r_2}} \frac{y_k}{\pi_k \times 0.875}$$

Le mécanisme de non-réponse ignorable (NMAR)

Mécanisme de non-réponse non ignorable (NMAR)

Un mécanisme de non-réponse qui n'est pas ignorable est dit non-ignorable (ou Non Missing At Random).

Cela signifie que la non-réponse dépend de la variable d'intérêt, même une fois que l'on a pris en compte les variables auxiliaires.

Il est très difficile de corriger de la non-réponse non ignorable, ou même de la détecter.

Exemple : enquête sur le revenu + non-réponse expliquée par le croisement sexe $\times$ revenu.

Mécanisme de non-réponse ignorable (NMAR)

En pratique, les probabilités de réponse p_k sont inconnues et doivent être estimées. On postule alors un modèle de réponse de la forme

$$p_k = f(z_k, \beta_0),$$

avec

- z_k un vecteur de variables auxiliaires connu sur S,
- $f(\cdot, \cdot)$ une fonction connue,
- β_0 un paramètre inconnu.

Mécanisme de non-réponse ignorable (NMAR)

La liaison en question peut prendre les formes les plus diverses. Une des plus ”célèbres” consiste à écrire :

$$r_k \simeq \frac{a.e^{bX_k}}{1 + a.e^{bX_k}} \Rightarrow Log\frac{r_k}{1 + r_k} \simeq Loga + bX_k$$

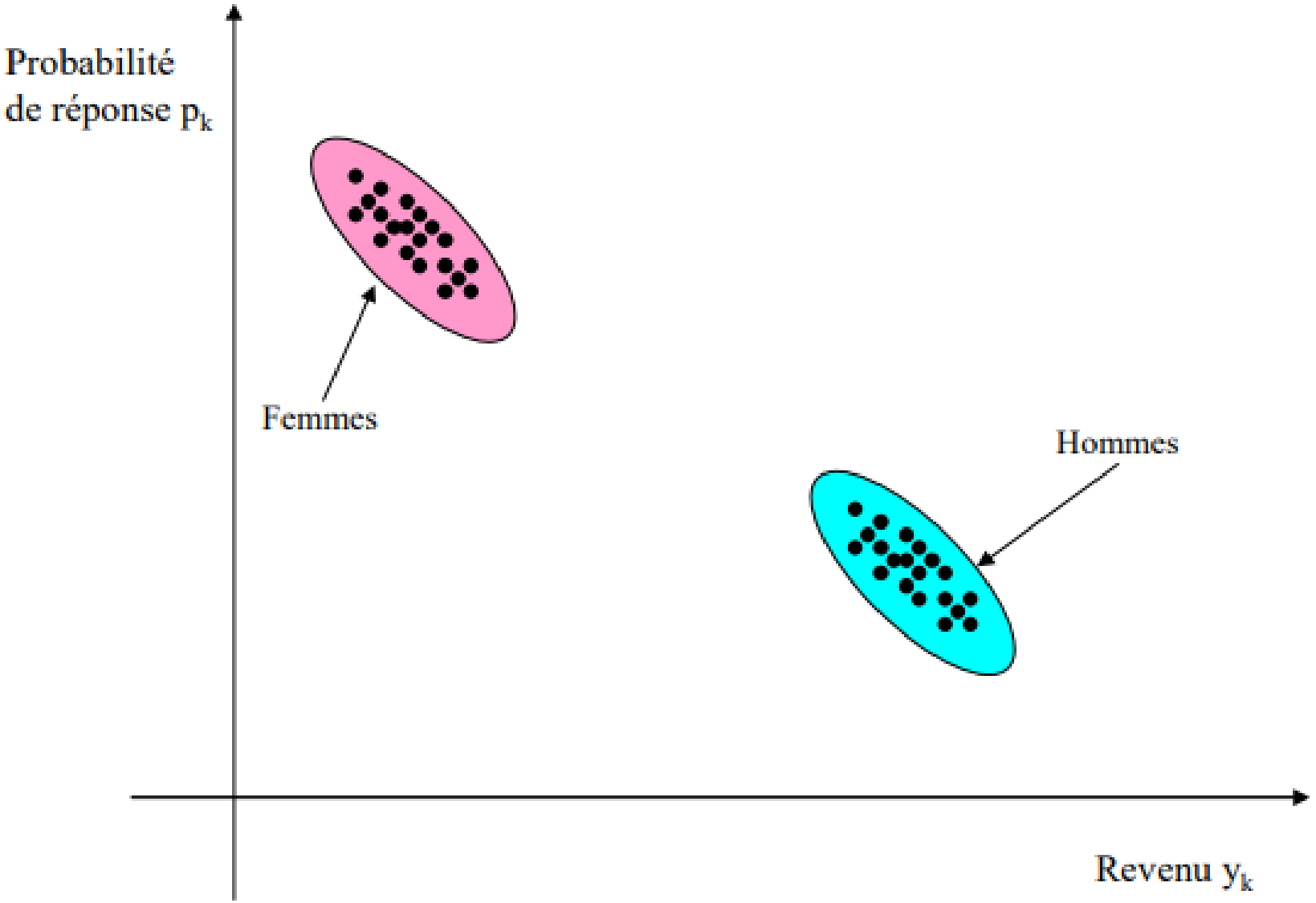
On parle alors de modèle **LOGIT** et on sait mettre en œuvre des procédures pour estimer les paramètres a et b selon certains critères optimaux

Mécanisme de non-réponse ignorable (MAR)

On peut aussi trouver d'autres fonctions de X, qui ont bonne allure et qui sont toujours comprises entre 0 et 1. Par exemple, on peut s'appuyer sur la fonction de répartition $\phi(x)$ d'une loi normale centrée réduite, soit

$$\phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \, dt$$

et ajuster le modèle $r_k \simeq \phi(a + bX_k)$ On parle cette fois de modèle **PROBIT**, et on sait également estimer a et b par des techniques appropriées.


Probabilité
de réponse p_k
Femmes
Hommes
Revenu y_k

CAS PRATIQUE

Conclusion

Conclusion

En conclusion, les méthodes de repondération face aux non-réponses sont cruciales pour garantir la fiabilité des résultats d'enquête. La repondération uniforme, le mécanisme de réponse homogène, et l'estimation des probabilités de réponse permettent de corriger les biais liés aux non-réponses.

Chacune présente des avantages et des limites selon le contexte. Leur choix doit être adapté aux spécificités de l'enquête et de la population cible.

Une approche bien choisie améliore la représentativité des données et la qualité des analyses.

Références

1. Ardilly, Pascal. 2006. Les Techniques de Sondage / Pascal Ardilly,... Éditions Technip. https://bibliotheque.univ-catholille.fr/Default/doc/SYRACUSE/340736/les-techniques-de-sondage-pascal-ardilly.

2. Chauvet, Guillaume. n.d. “Données Manquantes dans les Enquêtes.”

MERCI POUR VOTRE ATTENTION