Comparaison de deux méthodes de repondération : Cas de l'enquête santé de l'Observatoire de population de Ouagadougou

Yacouba Compaoré [1], Bruno Lankoandé [2], Abdramane Soura [3]

[1] *ISSP 03 BP 7118 Ouagadougou 03, Burkina Faso,* ***ycompaore@issp.bf***
[2] *ISSP 03 BP 7118 Ouagadougou 03, Burkina Faso,* ***blankoande@issp.bf***
[3] *ISSP 03 BP 7118 Ouagadougou 03, Burkina Faso,* ***asoura@issp.bf***

Résumé.
Cette communication compare deux méthodes de repondération qui visent à corriger le biais induit dans l'estimation par la non réponse totale. La première méthode de repondération se fonde sur une hypothèse d'uniformité du taux de réponse à l'enquête pour tous les individus de l'échantillon. Quant à la seconde méthode, elle se base sur une hypothèse d'un mécanisme de réponse homogène à l'intérieur de sous populations. Celles - ci ont été constituées à l'aide d'une régression logistique sur des variables auxiliaires (sexe, situation matrimoniale, groupe d'âge, type de zone) connues sur tous les individus de l'échantillon et ayant un lien statistiquement significatif avec le comportement de réponse.

Les données utilisées proviennent de l'enquête santé réalisée au sein de l'Observatoire de Population de Ouagadougou (Burkina Faso) en 2010. Les fortes variations des taux de réponse au sein des sous populations qui ont été constituées indiquent que l'hypothèse d'uniformité générale est irréaliste. Cependant ce travail devrait être complété par une analyse des variances des estimations.

Mots-clés. enquête santé, repondération, observatoire de population

Introduction

La non réponse est un phénomène toujours présente dans les enquêtes. Quelque soit les mesures préventives prises par le chercheur pendant la collecte, il ne peut pas garantir un taux de non réponse nul à la fin de la collecte. Ainsi, il doit toujours se préparer à y faire face afin d'assurer une bonne précision des indicateurs qu'il va calculer sur ces données. On distingue essentiellement deux types de non réponse dans les enquêtes : la non réponse partielle qui est une indisponibilité des réponses pour certaines variables, et la non réponse totale qui est une absence complète de l'information sur un ensemble d'individus de l'échantillon. L'ignorance de cette absence totale ou partielle de l'information pour certains individus, réduisent les chances d'obtenir une bonne estimation des indicateurs.

Par exemple, il est reconnu que dans la plupart des enquêtes qui concernent des individus issus de différents milieux de résidence, les urbains ont une forte propension de non

réponse plus élevée que les ruraux à cause de plusieurs facteurs comme l'intrusion dans la vie privée (Goyder (1987)) cité par Durand et Blais (2000) ou le manque de disponibilité plus élevés en milieu urbain qu'en milieu rural (Durand et Blais (2000)). Estimer un indicateur uniquement sur les répondants dans ces types d'enquêtes conduirait à mal estimer la contribution des individus vivant en milieu urbain. Cet exemple illustre clairement que l'ignorance des non réponses dans l'estimation des indicateurs peut conduire à des erreurs.

Généralement, en cas de non réponse totale, les estimations obtenues sur l'ensemble des répondants de l'enquête uniquement sont biaisées si les répondants et les non répondants ont des comportements différents par rapport à la thématique de l'étude (Caron (2005)). Dans la littérature, il existe une multitude de méthodes de traitement de la non réponse totale dans les enquêtes. L'une d'entre elles, qui est couramment utilisée est la repondération qui consiste à augmenter judicieusement le poids de sondage des individus répondants. Elle nécessite cependant d'émettre des hypothèses sur le comportement de réponse des individus.

L'objectif de cette communication est de comparer deux méthodes de repondération basées sur des hypothèses différentes du comportement des individus vis à vis de la non réponse. Elle applique ces deux méthodes sur les données d'une enquête santé réalisée dans l'observatoire de population de Ouagadougou en 2010 afin de déterminer la méthode la mieux adaptée à la correction de la non réponse totale dans cette enquête. Ainsi, nous allons d'abord faire une brève présentation de l'observatoire et de l'enquête, ensuite présenter les deux méthodes de repondération et enfin présenter les résultats des estimations suivant les deux méthodes.

1 Présentation de l'observatoire de population de Ouagadougou et de l'enquête santé

Dans cette section, il sera question de faire une brève présentation de l'Observatoire de Population de Ouagadougou et de l'enquête santé.

1.1 Présentation de l'Observatoire de Population de Ouagadougou

L'Observatoire de Population de Ouagadougou (OPO) est une plateforme de recherche et d'interventions en milieu urbain dont l'objectif est de fournir des fondements scientifiques aux politiques de santé en Afrique sub-saharienne. C'est un outil fondamental dans la recherche de la relation entre les changements démographiques et l'évolution de la santé de la population. Les données collectées au sein de l'observatoire portent essentiellement sur les grossesses, les avortements, les naissances, les décès et les causes de décès par le biais des autopsies verbales. La zone de l'observatoire est située à la périphérie nord de Ouagadougou et couvre une superficie de 15,32 km^2. Elle est composée de cinq quartiers

repartis dans deux types de zones : deux formelles (Kilwin et Tanghin)et trois informelles (Nonghin, Polesgho, et Niokho II). Ces deux zones présentent des caractéristiques très contrastées. Pendant que la zone formelle est administrativement reconnue, et dispose de routes bien tracées, de structures de santé (pharmacies et centres de santé) publiques et même privées, d'écoles, d'électricité et d'eau la zone non formelle a un faible accès aux services sociaux de base. Dans l'ensemble, la population suivie était évaluée à 86000 individus en 2012 dont plus de la moitié vivaient en zone non formelle (52,7%). L'observatoire ne constitue pas un échantillon urbain représentatif mais un laboratoire pour observer en profondeur les spécificités des populations urbaines les plus pauvres, afin d'appuyer la conception puis le test de programmes destinés à réduire les inégalités de santé en ville.

1.2 Présentation de l'enquête santé

Réalisée en 2010, l'enquête santé avait pour objectif d'approfondir des thématiques particulières sur la santé des enfants, des adultes et des personnes âgées. L'état de santé, l'accès aux soins, les maladies chroniques, les accidents et violences sont entre autres les thématiques qui ont été abordées. La base de sondage était constituée de l'ensemble des ménages du deuxième passage de la collecte de routine de l'observatoire. Même si l'objectif était d'enquêter 1000 enfants de moins de 5 ans, 2000 adultes de 15 à 49 ans et 1000 personnes âgées de 50 ans et plus, l'unité d'échantillonnage était le ménage. L'échantillon final a été constitué grâce à quatre tirages systématiques, dont les détails sont consignés dans le tableau ci dessous :

Tableau 1 – Nombre de ménage et pas de tirage à chaque étape

Tirages	**Pas de tirage**	**Nombre de ménage**
1	18	791
2	13	650
3	9	350
4	6	150
Total		**1941**

Dans cette communication on s'intéressera uniquement aux individus âgés de 15 ans et plus.

2 Méthodologie

Les techniques de repondération sont surtout utilisées pour la correction du biais induit par la non réponse totale. Ce biais est d'autant plus grand que les répondants et

les non répondants ont des comportements différents vis à vis de la thématique de l'étude. Plus précisément, la correction du biais consiste à augmenter judicieusement le poids de sondage des répondants pour compenser celui des non répondants (Caron (2005)).

Soit Y notre variable d'intérêt,

y_k la valeur prise par un individu k sur la variable Y

π_k la probabilité d'inclusion de l'individu k dans l'échantillon,

N la taille totale de la population supposée connue,

S_r l'ensemble des répondants de l'enquête,

p_k la probabilité de réponse d'un individu k.

On se propose ici d'estimer la moyenne de Y en présence de non réponse totale. Il est démontré que

$$\bar{y} = \frac{1}{N} \sum_{k \in S_r} \frac{y_k}{\pi_k p_k}$$

est un estimateur sans biais de la moyenne (Chauvet (2012)). Cependant, les p_k ne sont pas connues et doivent être estimées. Nous proposons ici, deux méthodes d'estimation des p_k. L'estimation sera d'autant plus bonne que l'hypothèse d'estimation est proche de la réalité qui est inconnue.

2.1 Méthodologie 1

On suppose un mécanisme de réponse globalement uniforme, c'est à dire que la probabilité de réponse est la même pour chaque individu. Cette probabilité est approchée par le taux de réponse de l'enquête.

Soit n la taille de l'échantillon et n_r le nombre de répondants à l'enquête. L'estimation de la moyenne de Y est ainsi donnée :

$$\bar{y_1} = \frac{n}{n_r N} \sum_{k \in S_r} \frac{y_k}{\pi_k}$$

2.2 Méthodologie 2

En ce qui concerne la seconde méthode, nous modéliserons le mécanisme de la non réponse à l'aide des caractéristiques socio-démographiques des répondants et des non répondants. Les étapes de cette modélisation sont les suivantes (Chauvet (2012)) :

- Identification des non-répondants ;
- Identification des facteurs explicatifs de la non réponse à l'aide d'une régression logistique ;
- Constitution des groupes homogènes de réponse par croisement des modalités ;
- Estimation des probabilités de réponse au sein de chaque groupe ;
- Calcul des poids corrigés.

Soit H le nombre de groupes homogènes,

S_{rh} l'ensemble des répondants du groupe h,
n_{rh} le nombre de répondants du groupe h,
n_h le nombre d'individus échantillonné du groupe h,
$\hat{p_h} = \dfrac{n_{rh}}{n_h}$ la probabilité de réponse estimée au sein du groupe h.
L'estimation de la moyenne de Y est donnée par :

$$\hat{\bar{y}_2} = \frac{1}{N} \sum_{h=1}^{H} \frac{n_h}{n_{rh}} \sum_{k \in S_{rh}} \frac{y_k}{\pi_k}$$

3 Résultats

Les résultats des estimations sont illustrés par les deux méthodes en considérant le fait d'être hypertendu ou non comme variable d'intérêt. En effet, l'enquête santé a collecté des informations sur la tension artérielle des enquêtés de 15 ans et plus. Celles - ci ont été utilisées pour construire une variable qui indique pour chaque individu si il est hypertendu ou pas. Ainsi, notre variable d'intérêt est définit de la manière suivante :

$y_k = 1$ si l'individu k est hypertendu et 0 sinon.

3.1 Résultats de la méthode 1

Le taux de réponse estimé au sein des 15 ans et plus est de 71,28%, soit 2355 enquêtés et 3304 individus échantillonnés. Ainsi, après repondération la prévalence estimée de l'hypertension dans les zones de l'OPO est de 13,68%.

3.2 Résultats de la méthode 2

A l'aide des analyses bivariées et de la régression logistique, nous avons identifié que les variables suivantes sont associées au mécanisme de non-réponse : le type de zone, le statut matrimonial, le sexe et le groupe d'age. Toutes ces variables sont significatives à un seuil de 1% dans la régression logistique. Le tableau 2 ci-dessous indique la proportion des répondants selon les modalités de chaque variable.

Tableau 2 – Proportion des répondants par catégorie

Variable	**Modalités**	**Taux de réponse**
Type de zone	*formelle*	68,53%
	non formelle	75,13%
Sexe	*homme*	66,10%
	femme	76,35%
Statut matrimonial	*marié*	75,75%
	non marié	64,67%
Groupe d'âge	*15-49 ans*	67,54%
	50 ans et plus	77,21%

Le taux de réponse est plus élevé en zone non formelle par rapport à la zone formelle. En effet, les zones non formelles sont des quartiers d'habitat spontané qui se sont développés à la périphérie de Ouagadougou (Burkina Faso). Celles-ci ne sont pas viabilisées et les habitants n'ont pas accès aux services sociaux de base (eau, électricité, assainissement). Par ailleurs, certains y habitent dans l'espoir d'être propriétaire de terrain au moment de la viabilisation. Ces individus ont tendance à confondre les enquêtes aux opérations de lotissement que la commune pourrait entreprendre. Ils sont donc très disponibles pour l'administration des questionnaires. Quant aux habitants de la zone formelle, en raison de leur activité, de leur niveau élevé d'instruction et de leur méfiance, ils sont des fois hésitants pour répondre aux questions.

Comme attendu, le taux de réponse au sein des femmes est plus élevé par rapport aux hommes. La plus grande occupation des hommes est sans doute le facteur explicatif de cette différence.

En ce qui concerne le groupe d'âge, le taux de réponse plus faible des 15-49 ans est probablement due à leur plus grande mobilité.

En raison du faible nombre de variables explicatives et de leur nature qualitative, les groupes homogènes ont été constitués en croisant toutes les modalités possibles. Les probabilités de réponse estimées par groupe homogène sont consignées dans le tableau 3. Les hommes mariés de 15 à 49 ans vivant en zone formelle sont ceux qui ont le taux de réponse le plus faible. Le taux de réponse le plus élevé est observé chez les femmes mariées de 50 ans et plus qui vivent en zone non formelle.

Tableau 3 – Probabilités de réponse par groupe homogène

Groupe	Type de zone	Statut marital	Sexe	Groupe d'âge	Taux de réponse
1	formelle	marié	homme	15-49 ans	50,57%
2	non formelle	non marié	homme	15-49 ans	57,06%
3	formelle	non marié	homme	15-49 ans	57,39%
4	formelle	non marié	homme	50 ans et plus	59,46%
5	formelle	non marié	femme	15-49 ans	59,80%
6	non formelle	non marié	femme	15-49 ans	65,93%
7	non formelle	marié	homme	15-49 ans	73,20%
8	formelle	marié	homme	50 ans et plus	74,15%
9	non formelle	non marié	homme	50 ans et plus	77,27%
10	non formelle	marié	homme	50 ans et plus	77,90%
11	non formelle	non marié	femme	50 ans et plus	78,34%
12	formelle	marié	femme	50 ans et plus	79,15%
13	formelle	non marié	femme	50 ans et plus	80,10%
14	formelle	marié	femme	15-49 ans	80,95%
15	non formelle	marié	femme	15-49 ans	83,19%
16	non formelle	marié	femme	50 ans et plus	83,33%

Conclusion

En constituant les groupes homogènes, la prévalence estimée de l'hypertension est de 14,63%. Cette valeur est plus élevée que celle obtenue en utilisant la première méthode. L'écart entre les deux estimations est de 1 point. Vu que les probabilités de réponse sont très différentes au regard des groupes que nous avons constitués, il est fort probable que cette méthode offre de meilleurs résultats. L'hypothèse d'une distribution uniforme de la non réponse n'est donc pas réaliste. Il aurait été intéressant de comparer les résultats obtenus par les deux méthodes avec d'autres sources d'estimation de l'hypertension. Malheureusement, l'estimation disponible date de 2003 (23%, Niakara et al., (2003)) et portait sur les adultes de 18 ans et plus.

Une comparaison de la qualité des estimations obtenues par les deux méthodes de repondération ne serait complète en l'absence de l'estimation des variances. Ce travail devrait être complété par une analyse des variances des estimateurs obtenus.

Bibliographie

[1] Allison, P.D, (2002), Missing data , Sage, Newbury Park.
[2] Caron, N (2005), La correction de la non réponse par repondération et par imputation, Document de travail numéro M0502, INSEE.
[3] Chauvet, G (2012), Données manquantes dans les enquêtes, Notes de cours ENSAI.
[4] Durand, C. et Blais A., (2000), A la recherche des déterminants culturels de la non-réponse, Communication présentée au deuxième colloque francophone sur les sondages, Bruxelles 22 - 23 juin 2000.
[5] Goyder,J (1987), The silent minority. Westview press, Boulder, Colorado. 232 pages.
[6] Niakara, A et al., (2003), Connaissance d'une enquête urbaine sur l'hypertension artérielle : enquête prospective menée à Ouagadougou, Burkina Faso *Revue des Organisateurs du congrès internationale de médecine francophone*, 96, 219–222.
[7] Tossou B.D, (2012), Analyse de la sensibilité de l'estimation des indicateurs de l'éducation et de l'emploi au traitement de la non-réponse, ENSAE, Dakar.