CHAPITRE IV

CHOIX DE L'ECHANTILLON

Ce chapitre s'adresse aux coordinateurs et au personnel technique de l'enquête.
Il vous permettra de :

✓ Comprendre les concepts de base du sondage.
✓ Calculer la taille de l'échantillon nécessaire à des estimations nationales et sous-nationales.
✓ Déterminer le nombre de grappes à utiliser.
✓ Choisir un plan de sondage.

COMPRENDRE LES CONCEPTS DE BASE DU SONDAGE

Dans le contexte des enquêtes à indicateurs multiples, le sondage est une opération consistant à sélectionner, dans la population, des personnes à enquêter. Dans notre cas, les enquêtés seront généralement les mères, ou les personnes qui s'occupent des enfants de chaque ménage visité[1], et elles répondront à toutes les questions des modules concernant la Santé des Enfants. Les modules sur l'Eau et les Sanitaires et sur l'Iodation du Sel portent sur l'ensemble du ménage. N'importe quel membre adulte du ménage peut répondre à ces questions, qui sont posées même dans les ménages qui ne comptent aucun enfant de moins de 15 ans.

En principe, notre enquête pourrait couvrir *tous* les ménages de la population. Si toutes les mères interrogées pouvaient fournir des réponses parfaites, on pourrait mesurer tous les indicateurs de façon exacte. Cependant, interroger toutes les mères prendrait beaucoup de temps, serait coûteux et peu économique. Il est donc nécessaire d'interroger un *échantillon* de ces femmes pour obtenir des *estimations* des indicateurs réels.

La différence entre l'estimation et l'indicateur réel est appelé *erreur de sondage*. Les erreurs de sondage sont causées par le fait qu'un échantillon –et non la population dans son entier– est enquêté.

L'erreur de sondage peut être minimisée en prenant certaines précautions :

- Choisir l'échantillon de personnes à enquêter de façon non biaisée.
- Sélectionner un échantillon de taille suffisante pour que les estimations soient précises.

Choisissez votre échantillon à partir de tous les ménages. Evitez de choisir des échantillons dont pourrait résulter des estimations *biaisées*.

---

[1]Un *ménage* est un groupe de personnes qui vivent et mangent ensemble.

✎ **EXEMPLE :**

Ne choisissez pas des échantillons exclusivement à partir de groupes de population particuliers, comme les enfants qui se rendent dans les cliniques.

Ne demandez pas aux mères d'amener leurs enfants à un lieu central de la communauté, parce que certaines d'entre elles ne viendront pas; vous ne pourrez pas déterminer combien de mères ne sont pas venues et en quoi elles sont différentes de celles qui sont venues.

N'utilisez pas des échantillons choisis au gré de l'enquêteur, du superviseur ou du directeur des activités de terrain.

Ne limitez pas votre échantillon aux familles vivant dans des ménages facilement accessibles, comme ceux proches de la route principale ou près du centre du village; il se peut que les familles vivant dans des zones difficilement accessibles soient les plus pauvres et en moins bonne santé que les autres.

N'omettez pas les ménages dans lesquels personne n'est présent au moment de la première visite. Déterminez si le logement est inhabité et faites une seconde visite ultérieurement.

Si un échantillon ne représente pas exactement l'ensemble de la population étudiée, l'estimation sera *biaisée*. Cela signifie que l'estimation sera déplacée dans une direction ou une autre par rapport à la valeur réelle.

Un tirage probabiliste garantit que tous les individus de la population-cible (par exemple, toutes les mères de jeunes enfants) ont une chance connue d'être interviewés.

Pour éviter les biais, vous devez procéder à un tirage probabiliste pour sélectionner votre échantillon de personnes à enquêter. Les biais dépendent de la procédure de sélection et non de la taille de l'échantillon. Les échantillons de grande taille tendent à être plus précis mais ils ne sont pas nécessairement moins biaisés.

Pour pouvoir procéder à un tirage probabiliste, vous devez disposer d'une *base de sondage*. Il s'agit d'une liste de toutes les unités (états ou provinces, districts, communautés, ménages, individus, etc...) de la population dans laquelle vous allez choisir votre échantillon. Pratiquement aucun pays ne dispose de listes mises à jour de tous les individus et seulement quelques pays disposent de listes des ménages. De ce fait, votre base de sondage comprendra généralement des listes d'unités de population plus larges, telles que des états ou provinces, des districts, des villes, des villages ou des sections d'énumération. Vous devez également disposer d'une estimation de la population ou du nombre de ménages de chacune de ces unités; en outre, vous devez vous assurer que *toutes* les unités sont incluses dans la base de sondage.

Etant donné que dans les enquêtes à indicateurs multiples, on cherche généralement à obtenir des indicateurs au niveau national, il est nécessaire d'utiliser un *sondage à plusieurs degrés*. Par exemple, vous pourriez commencer par sélectionner un certain nombre de provinces (premier degré), puis un certain nombre de districts dans chacune des provinces sélectionnées (deuxième degré), un certain nombre de communautés dans chaque district choisi (troisième degré) et, finalement, un certain nombre de ménages dans chaque communauté retenue (quatrième degré). Chacun de ces tirages peut se faire en utilisant un des différents procédés de sondage présentés ci-dessous. Nous

présenterons, plus loin dans ce chapitre, quelques exemples concrets de la façon de combiner différents procédés de sondage aux différents degrés.

Echantillonnage aléatoire simple

L'échantillonnage aléatoire simple est la forme la plus simple de sondage probabiliste. Des nombres aléatoires sont choisis au moyen d'une calculatrice, d'un programme informatique ou d'une table de nombres aléatoires. Un autre procédé consiste à écrire sur des morceaux de papiers les noms ou numéros d'identification de toutes les communautés, ménages ou individus; l'échantillon souhaité est alors sélectionné en tirant le nombre nécessaire de morceaux de papier. Chaque communauté, personne ou ménage correspondant aux nombres choisis est alors inclus dans l'échantillon. Dans l'échantillonnage aléatoire simple, la sélection de chaque individu est *indépendante* de la sélection d'un autre individu.

Echantillonnage systématique

L'échantillonnage systématique est une modification de l'échantillonnage aléatoire simple. Il consiste à tirer chaque *Nième* ($5^{ème}$ ou $10^{ème}$ ou $50^{ème}$, et ainsi de suite) ménage d'une liste complète des ménages. En utilisant l'échantillonnage systématique, vous devez vous assurer que la liste que vous utilisez n'est ordonnée d'aucune manière régulière, ce qui biaiserait votre échantillon.

✎ **EXEMPLE :**

Si la liste est ordonnée de la façon suivante : chef de village, sous-chef, autres ménages, puis un autre chef de village, sous-chef, autres ménages, et ainsi de suite, un échantillonnage systématique peut biaiser vos résultats. Par exemple, si vous commencez par le cinquième ménage dans chaque village, vous n'inclurez jamais aucun chef ou sous-chef de village.

Echantillonnage par grappes

Comme cela a été noté précédemment, dans les pays où des enquêtes sont nécessaires pour mesurer les indicateurs des objectifs, les listes de ménages ou d'individus ne sont habituellement pas disponibles. Il est généralement trop onéreux et trop long d'établir de telles listes et de localiser les ménages. Le choix logique est alors d'utiliser le sondage par grappes —c'est-à-dire de sélectionner des groupes de ménages qui sont, géographiquement, proches les uns des autres.

Le sondage par grappes implique généralement deux degrés de tirage.

✎**EXEMPLE :**

Un certain nombre de communautés[2] peuvent être sélectionnées, au départ, à partir d'une base de sondage (liste des communautés); on choisit ensuite, aléatoirement, une grappe à l'intérieur

---

[2]Une *communauté* est définie comme un groupement à l'intérieur d'une population : il s'agit, par exemple, d'une ville, d'un village ou d'une section d'énumération. Dans ce manuel, on utilise le terme *grappe* pour désigner un échantillon sélectionné aléatoirement à partir de la communauté. Il s'agit du terme tel qu'il est utilisé par le PEV; dans la littérature statistique, la grappe *est* la communauté, et d'autres publications peuvent faire référence à ce terme de cette façon.

de chacune de ces communautés. Tous les ménages de ces grappes sont alors enquêtés.

Les ménages sont les *unités de base d'échantillonnage* (c'est-à-dire, les plus petites unités à sélectionner). On peut aussi utiliser les enfants de moins de cinq ans comme unité de base d'échantillonnage, mais les ménages sont beaucoup plus faciles à identifier et à lister que les enfants de moins de cinq ans. Le choix final des ménages est effectué dans la communauté. Les différentes procédures de sélection des ménages dans la communauté sont présentées au chapitre 6.

Le sondage par grappes est toujours utilisé dans les enquêtes à indicateurs multiples.

Le sondage par grappes présente de nombreux avantages. Il réduit le temps et les coûts de déplacement et simplifie le travail sur le terrain. Il simplifie aussi la supervision et l'administration de l'enquête. Cela est important puisqu'une meilleure supervision des enquêteurs aura pour résultat une meilleure qualité des données. Cependant, le sondage par grappes implique que chaque enquêté n'est pas choisi indépendamment des autres enquêtés —ce sont des voisins— et cela peut accroître l'erreur de sondage.

Echantillonnage avec probabilité proportionnelle à la taille

Les communautés peuvent varier considérablement en taille de population. Si on utilise l'échantillonnage aléatoire simple ou l'échantillonnage systématique, les grandes et les petites communautés auront la même probabilité d'être incluses, ce qui n'est pas correct. Une façon de compenser ces différences consiste à choisir des grappes, à partir de la base de sondage, avec une probabilité proportionnelle à la taille (PPT).

Les communautés avec des populations importantes ont, proportionnellement, plus de chances d'inclure une grappe sélectionnée que les communautés plus petites.

Si elle est utilisée correctement, un avantage de la PPT réside dans le fait que chaque ménage de l'échantillon a une chance égale d'être sélectionné. Dans ce cas, l'échantillon est dit *auto-pondéré*, ce qui facilitera l'analyse.

La procédure de l'échantillonnage avec PPT est présentée dans la section suivante. On peut utiliser un tirage à PPT pour la plupart des degrés du plan de sondage. Par exemple, vous pouvez utiliser l'échantillonnage avec PPT pour sélectionner les états ou les provinces dans un pays, les municipalités à l'intérieur d'un état, des communautés ou des villages à l'intérieur d'un district. L'exemple ci-dessous traite de la sélection de communautés à l'intérieur d'un district, mais la procédure est exactement la même quelles que soient les unités de sondage choisies.

Procédure d'échantillonnage avec probabilité proportionnelle à la taille

Pour sélectionner les communautés à utiliser dans l'enquête, vous aurez besoin :

- d'une liste de toutes les communautés du district à enquêter (dans cet exemple, chaque

ménage dans le district appartient à une —et seulement une— communauté)

- d'une estimation de la taille de la population (ou du nombre de ménages) dans chaque communauté[3]
- du nombre de grappes que vous voulez avoir dans le district (expliqué plus loin)

Pour sélectionner les communautés avec PPT, vous devez exécuter les sept étapes suivantes :

1. **Faites un tableau à 3 colonnes, comme celui présenté ci-dessous en tant qu'exemple.**
Colonne (1) : attribuer un numéro à chaque communauté. Lister les communautés.
Colonne (2) : Lister la population de chaque communauté.
Colonne (3) : Faire le cumul de la population de chaque communauté —c'est-à-dire, la somme de la population de cette communauté et de celles de toutes les communautés listées au-dessus dans le tableau.

☞ Quand vous faites la liste des communautés, essayez d'utiliser un ordre logique, comme un ordre géographique. Reportez-vous à la *stratification implicite* présentée ci-dessous.

2. **Calculez l'*intervalle de tirage* en utilisant la formule suivante :**
Intervalle de tirage = population totale cumulée ÷ nombre de grappes nécessaires
En utilisant l'exemple ci-dessous et en supposant que l'on ait besoin de sélectionner trois grappes, ceci donnerait :

Intervalle de tirage = 6 700 ÷ 3 = 2 233

3. **Sélectionnez un nombre aléatoire qui est égal ou plus petit que l'intervalle de tirage.** Pour l'exemple ci-dessus, il faut choisir un nombre aléatoire compris entre 1 et 2 233. Vous pouvez sélectionner ce nombre aléatoire en utilisant une calculatrice, un ordinateur, une table de nombres aléatoires provenant d'un manuel statistique, ou même en demandant à quelqu'un (qui ne connaît pas l'enquête) de choisir un nombre compris dans l'intervalle. Supposons que le nombre choisi est 1 814.

4. **Revenez au tableau**. Localisez la première grappe en trouvant la communauté dont la population cumulée dépasse ce nombre aléatoire. Dans cet exemple, la première grappe sera localisée dans le village 4, dont la population cumulée est de 1 900.

5. **Ajoutez l'intervalle de tirage au nombre aléatoire**. Dans l'exemple, nous avons, 2 233 + 1 814 = 4 047.

---

[3]Si la taille moyenne des ménages ne varie pas considérablement d'une communauté à l'autre, vous pouvez utiliser la formule suivante pour estimer la population de chaque communauté : *taille de la population = nombre de ménages x taille moyenne d'un ménage.*

6. **Choisissez la communauté dont la population cumulée dépasse juste ce nombre**. La seconde grappe sera localisée dans cette communauté. Dans cet exemple, la seconde grappe sera localisée dans le village 6.

TABLEAU ILLUSTRATIF
Cumul de la population des communautés

Village	Taille de la population	Cumul
1	1 000	1 000
2	400	1 400
3	200	1 600
4	300	1 900
5	1 200	3 100
6	1 000	4 100
7	1 600	5 700
8	200	5 900
9	350	6 250
10	450	6 700

7. **Identifiez chaque grappe suivante** en ajoutant l'intervalle de tirage au nombre qui a permis d'identifier la grappe précédente. Arrêtez-vous quand vous aurez identifié autant de grappes que nécessaire.

Si une communauté a une population plus importante que l'intervalle de tirage, elle peut être sélectionnée deux fois. Deux grappes indépendantes doivent alors être sélectionnées dans la même communauté. Cela est parfaitement correct.

Une fois que vous avez procédé à votre échantillonnage avec PPT et que vous vous êtes assuré qu'il n'y a pas d'erreur, ne cédez pas à la tentation de le refaire dans le cas où votre premier échantillon ne vous plairait pas.

Sondage stratifié (Stratification)

Vous pouvez souhaiter de mesurer des indicateurs pour des sous-groupes pris à l'intérieur de la population nationale. Comme certains de ces sous-groupes peuvent compter moins d'individus que les autres, vous voudrez vous assurer que tous ces groupes sont correctement représentés dans votre échantillon. L'utilisation de PPT peut avoir pour conséquence de petits échantillons pour des groupes représentant de faibles proportions de la population dans son ensemble. La stratification résout ce

problème. Vous constituerez une strate pour chaque sous-groupe et vous tirerez un échantillon dans chaque strate séparément.

La stratification est utilisée couramment pour fournir des données à un niveau sous-national –par exemple, au niveau de l'état ou de la province. Si vous ne procédez pas à une stratification, les grands états ou provinces auront des échantillons beaucoup plus importants que les petits états ou provinces. Pour obtenir une précision identique pour tous les états (voir explication ci-dessous), vous pouvez sélectionner le même nombre de ménages dans chaque état et, ultérieurement, pondérer les résultats en fonction de la vrai proportion de la population nationale qui correspond à chaque état (voir Chapitre 7).

En plus de fournir des indicateurs à un niveau sous-national, la stratification peut être utilisée pour fournir des données sur des sous-groupes de population vivant dans des conditions particulières.

✎**EXEMPLE :**

Dans certains pays, les habitants des bidonvilles du milieu urbain peuvent constituer seulement une petite fraction de la population, mais peuvent également présenter des besoins particuliers. Aussi, au début de la procédure de sondage, on peut lister les bidonvilles urbains séparément des autres zones et s'assurer qu'un nombre suffisant de logements de ces quartiers sont inclus dans l'échantillon, en leur affectant une probabilité de tirage supérieure à celle des autres logements. *Cet échantillon comprendra alors deux strates : les habitants des bidonvilles et les autres.*

Quand on utilise la stratification, la taille de l'échantillon de chaque strate n'est généralement pas proportionnelle à sa population réelle. Ce déséquilibre doit être corrigé au moment de l'analyse des données. Par exemple, si les habitants des bidonvilles représentent 5 pour cent de la population d'un pays, mais que votre procédure de stratification conduit à un échantillon dans lequel 10 pour cent des habitants appartiennent à cette catégorie, les données de l'échantillon doivent être *pondérées* pour produire les résultats au niveau national[4].

Stratification implicite

La stratification implicite est une forme particulière de stratification. Elle consiste à lister les unités à sélectionner (par exemple, des grappes) selon un ordre logique, puis à faire un tirage systématique ou avec PPT à partir de cette liste.

✎ **EXEMPLE :**

Vous pouvez suivre un ordre géographique du nord au sud, ou de l'ouest à l'est, ou vous pouvez partir de la capitale de district et vous éloigner de ce point.

---

[4]Dans cet exemple, à chaque enfant de bidonville, on attribuera un poids de 0,5 (puisque 5 pour cent est la moitié de 10 pour cent); de même, on attribuera un poids de 0,95/0,90 aux habitants des autres zones car ils représentent 95 pour cent de la population, mais seulement 90 pour cent de l'échantillon. La formule générale pour la pondération est égale à : (proportion dans la population)/(proportion dans l'échantillon). Voir Chapitre 6.

Grâce à cette procédure, les unités seront réparties régulièrement sur la liste et cela évite que, du fait du hasard, un type de communauté finisse par être sous-représenté. Cet ordre logique peut être géographique, mais peut aussi suivre la distribution d'une variable importante, comme la proximité des services de santé ou les indicateurs socio-économiques.

En plus de faciliter la production de résultats fiables au niveau sous-national, la stratification est une procédure très utile pour améliorer la précision d'ensemble des estimations nationales.

Les différentes procédures d'échantillonnage discutées ci-dessus —aléatoire simple, systématique, par grappes, PPT et stratification—sont nullement exclusives. Comme illustré par les exemples ci-dessous, ces procédures peuvent être combinées aux différents degrés de sondage pour produire des estimations nationales.

COMMENT CALCULER LA TAILLE DE L'ECHANTILLON

L'estimation de la taille de l'échantillon est une procédure simple, même si la plupart des gens sont déconcertés par sa complexité apparente. Cette section vous guidera à travers les différentes étapes nécessaires à son calcul. A la fin de la section, on présente un programme de feuille d'analyse qui permet ces calculs. La compréhension des principes de base de l'estimation de la taille de l'échantillon, présentés ci-dessous, est essentielle à l'utilisation correcte de la feuille de calcul.

L'enquête d'un échantillon d'enfants vous donnera une valeur estimée de l'indicateur vrai, comme la proportion des ménages utilisant du sel iodé. Du fait de l'erreur de sondage, cette valeur estimée n'est pas une mesure exacte. Cependant, l'utilisation du sondage probabiliste permet de calculer un intervalle dans lequel il est raisonnable de penser que l'indicateur réel se situe. Les limites supérieure et inférieure de cet intervalle sont connues sous le terme de *marge d'erreur*[5].

✎ **EXEMPLE :**

Les enquêtes PEV classiques fournissent une estimation de la couverture vaccinale avec une marge d'erreur de plus ou moins 0,10 ou 10 points de pourcentage. Cela signifie que, si l'enquête a trouvé une couverture du BCG de 75 pour cent, on peut, raisonnablement, être sûr que la valeur vraie dans la population se situe entre 65 et 85 pour cent.

Par "raisonnablement être sûr", on veut dire qu'il y a seulement 5 pour cent de chances pour que la couverture vaccinale réelle de cette population se situe en dehors de la marge d'erreur déterminée par l'enquête. En d'autres termes, on peut affirmer que, dans 95 pour cent des cas, le vrai taux de couverture de la population se situe à l'intérieur de l'intervalle.

Pour estimer la taille de l'échantillon, il faut obtenir ou estimer certaines informations de base. Suivez les étapes suivantes pour calculer l'échantillon nécessaire à chaque indicateur. On utilisera la couverture vaccinale du BCG comme exemple. Souvenez-vous que, *dans les calculs, vous devez*

---

[5]Le terme "marge d'erreur", utilisé dans ce document, est parfois appelé "niveau de précision" ou "intervalle de confiance" dans d'autres publications.

toujours utiliser les pourcentages sous la forme de nombre décimaux (par exemple, utilisez 0,6 pour 60 pour cent).

1. **Supposez/prévoyez la *proportion* que vous allez mesurer**. En utilisant n'importe quelles informations disponibles, faites une estimation grossière de la proportion d'enfants qui sont vaccinés. On appelle cela une *proportion prévue*.

✎ **EXEMPLE :**

Vous pouvez supposer que 80 pour cent des enfants ont été vaccinés. Par conséquent, votre proportion prévue sera 0,8.

☞ Vous ne disposez sûrement pas de bonnes estimations de la couverture actuelle. (C'est, en fait, la raison pour laquelle vous faites une enquête!) Cependant, même une information approximative est adéquate. Dans le doute, approchez-vous de 0,5 car les estimations proches de 0,5 nécessitent, pour une marge d'erreur donnée, des échantillons plus importants que des taux de couvertures plus forts ou plus faibles.

✎ **EXEMPLE :**

Si vous pensez que la couverture se situe entre 60 et 70 pour cent, choisissez 0,6.

2. **Décidez d'une *marge d'erreur* acceptable pour votre estimation.** C'est-à-dire, décidez de l'écart entre le vrai taux et votre estimation. Pour l'évaluation des objectifs au niveau national, plus ou moins 5 points de pourcentage[6] est une marge d'erreur raisonnable mais, pour des estimations sous-nationales, on peut se satisfaire de 10 points de pourcentage.

✎ **EXEMPLE :**

Vous pouvez souhaiter estimer la couverture du BCG avec une marge de plus ou moins 5 points de pourcentage. Par conséquent, si votre enquête trouve une couverture de 82 pour cent, le vrai taux de couverture dans la population se situe probablement entre 77 et 87 pour cent.

☞ Plus la marge d'erreur acceptable est faible, plus grand doit être l'échantillon utilisé dans l'enquête.

Vous devez vous assurer que votre échantillon est suffisamment grand pour atteindre la marge d'erreur absolue que vous recherchez. Pour l'évaluation des objectifs au niveau national, des marges d'erreur de, disons, 15 points de pourcentage ou plus sont certainement trop importantes pour être d'une utilité quelconque. Mais vous devez également équilibrer les besoins en précision avec les coûts

[6]Dans ce texte, les marges d'erreur sont exprimées en termes de nombre absolu de points de pourcentage. Par exemple, une couverture estimée à 60 pour cent avec une marge d'erreur de plus ou moins 5 points de pourcentage signifie que la vraie couverture devrait se situer entre 55 et 65 pour cent.

entraînés par une enquête très vaste.

✎ **EXEMPLE :**

Si vous voulez estimer la couverture du BCG avec une marge d'erreur de plus ou moins 1 point de pourcentage, vous aurez besoin de plus de 10 000 enfants de 12-23 mois. Cela implique que, dans un pays en développement typique, vous visitiez 70 000 ménages, une opération dans laquelle la plupart des pays ne se risqueront pas.

3. **Choisissez un *effet du plan de sondage* approprié.** Lorsqu'on utilise un sondage par grappes, les enquêtés ne sont pas choisis complètement indépendamment des autres enquêtés du même quartier.

✎ **EXEMPLE :**

Si les enfants d'une grappe sont semblables du point de vue de leur état vaccinal, on dira qu'ils sont "homogènes". Cela signifie que, si un enfant donné a été vacciné, il est probable que ses voisins aient aussi été vaccinés. L'homogénéité s'applique aussi aux enfants non vaccinés : il est probable que les autres enfants de la même grappe ne soient pas vaccinés.

Si les enfants sont plus homogènes (ou semblables) à l'intérieur d'une grappe que dans la population entière, il en résultera une marge d'erreur plus importante que si les mêmes enfants avaient été choisis par le biais d'un tirage aléatoire strict –c'est-à-dire, s'ils avaient été sélectionnés de façon complètement indépendante les uns des autres.

✎ **EXEMPLE :**

Un échantillon de 1 000 enfants obtenu par un sondage par grappes fournira une estimation de la couverture moins précise qu'un autre échantillon de 1 000 enfants choisis au moyen d'un échantillonnage aléatoire simple.

Certains indicateurs seront, plus que d'autres, affectés par le sondage par grappes. La couverture vaccinale peut être affectée du fait de l'éloignement des centres de santé et à cause des campagnes locales de vaccination. L'homogénéité est particulièrement importante pour les indicateurs d'incidence des maladies infectieuses comme la rougeole (qui est propagée d'un enfant à l'autre) ainsi que pour les variables en rapport avec l'eau et les sanitaires (en effet, les voisins partagent souvent les mêmes sources d'eau et les mêmes installations sanitaires).

L'expression mathématique de cet effet de grappe est appelé *effet du plan de sondage*. Il dépend, à la fois, du degré de similarité des enfants d'une même grappe et de la taille des grappes. Toutes choses étant égales par ailleurs, les grandes grappes provoqueront un effet de plan de sondage plus important que les petites.

Comme les calculs de la taille de l'échantillon sont généralement basés sur des échantillons aléatoires simples, la taille de l'échantillon doit être augmentée pour compenser un effet de plan de sondage important. En se basant sur l'expérience des autres enquêtes par grappes, généralement, on envisage un effet de plan de sondage de 2,0 pour la plupart des variables, ce qui nécessite de doubler

la taille de l'échantillon par rapport à un échantillon aléatoire simple. Pour les variables concernant l'eau et les sanitaires, on a utilisé un effet de plan de sondage de 10.

☞ Il est très difficile de prévoir ce que sera l'effet du plan de sondage avant de réaliser l'enquête. A moins que vous ayez de bonnes raisons pour ne pas le faire, utilisez les valeurs suggérées ci-dessus. Après la fin de l'enquête, on peut calculer l'effet du plan de sondage (voir Chapitre 7) et l'utiliser pour calculer la marge d'erreur réelle.

4. **Calcul de la taille de l'échantillon en utilisant la formule :**

$$Taille\,de\,l'échantillon = \frac{4 \times proportion \times (1-proportion) \times effet\,du\,plan\,de\,sondage}{marge\,d'erreur \times marge\,d'erreur}$$

✎ **EXEMPLE :**
En utilisant cette formule et les valeurs données dans les exemples ci-dessus pour la couverture du BCG, où :

proportion = 8,0 (c'est-à-dire 80 pour cent)
effet du plan de sondage = 2
marge d'erreur = 0,05 (± 5 points de pourcentage)

on obtient :

$$Taille\,de\,l'échantillon = \frac{4 \times 0{,}8 \times (1-0{,}8) \times 2}{0{,}05 \times 0{,}05} = 512$$

Ainsi l'échantillon devrait compter 512 enfants de 12-23 mois. Cela signifie que, pour obtenir une estimation avec une marge d'erreur de plus ou moins 5 points de pourcentage (95 pour cent de chances que le vrai taux se situe entre 75 et 85 pour cent), vous devrez enquêter les mères de 512 enfants âgés de 1 an.

Ces calculs doivent être répétés pour chacun des indicateurs à mesurer dans l'enquête. Ceci peut être réalisé à l'aide d'un programme de feuille de calcul (voir Encadré 4.1, qui commence à la page 4.14).

5. **Estimation du nombre de ménages que vous devez visiter.** La formule ci-dessus et la feuille de calcul fournissent le nombre de personnes de la population-cible nécessaire pour mesurer l'indicateur. Remarquez que la population-cible varie selon l'indicateur —par exemple, les enfants de 12-23 mois pour la couverture vaccinale, les enfants ayant eu récemment la diarrhée pour le taux d'utilisation de la thérapie de réhydratation par voie orale (TRO), les ménages pour l'eau et les

sanitaires, et ainsi de suite— si bien qu'un dénominateur commun est nécessaire.

☞ En calculant le nombre de ménages à visiter pour trouver l'effectif requis de population-cible, on doit comparer les exigences de tous les indicateurs du point de vue de la taille d'échantillon. Le nombre de ménages nécessaire le plus élevé sera alors retenu. Grâce à cela, tous les autres indicateurs seront mesurés avec une marge d'erreur identique ou plus faible.

Pour estimer le nombre de ménages nécessaires pour trouver un nombre donné d'enfants, vous devez connaître la taille moyenne des ménages et la proportion d'enfants dans la population. Remarquez que plus les ménages sont grands et plus le taux de natalité est élevé, moins vous aurez à visiter de ménages pour trouver un nombre suffisant d'enfants d'un groupe d'âges quelconque.

✎ **EXEMPLE :**

Si la taille moyenne des ménages est de six personnes, et si les enfants de 12-23 mois représentent 3 pour cent de la population, pour trouver 512 enfants de ce groupe d'âges, vous devrez visiter $512 \div (6 \times 0{,}03) = 2\,844$ ménages.

☞ Utilisez le dernier recensement disponible pour trouver la taille moyenne des ménages et la proportion d'enfants de moins de 5 ans dans la population. Si vous ne disposez pas de telles données, les valeurs "classiques" sont de six personnes par ménage et de 3 pour cent pour les enfants de chaque groupe d'âges d'une année (par exemple, 0-11 mois, 12-23 mois, etc.) : dans ce cas, la population comprend 15 pour cent d'enfants de moins de cinq ans. Vous pouvez également utiliser 25 pour cent comme fréquence de la diarrhée au cours des deux semaines précédentes. Le Tableau 4.1 donne des conseils sur la façon d'estimer le nombre de ménages nécessaire pour mesurer les indicateurs de la mi-décennie.

6. **Tenez compte des déperditions possibles.** Dans toutes les enquêtes, certaines des interviews prévues n'ont pas lieu parce que des gens sont absents de leur logement et à cause des refus. Les moyens à mettre en oeuvre pour minimiser ces déperditions sont présentés au Chapitre 6. Cependant, lorsqu'on calcule la taille de l'échantillon, il est prudent d'ajouter à l'échantillon la proportion de déperditions attendue. Les déperditions peuvent varier de façon importante de pays à pays, et du milieu urbain au milieu rural, en particulier là où il est courant que les mères travaillent à l'extérieur de la maison. Vous devez estimer cette proportion à partir des enquêtes précédentes réalisées dans le pays.

Ajoutez au moins 10 pour cent à votre taille d'échantillon estimée pour tenir compte des déperditions.

✎ **EXEMPLE :**

Si vous avez estimé votre échantillon à 2 844 ménages, ajoutez environ 10 pour cent et essayez d'atteindre 3 100 ménages.

Tableau 4.1. Nombres habituels de ménages à visiter pour trouver, en moyenne, un individu appartenant à la population-cible de chaque indicateur visant à mesurer les objectifs de la mi-décennie

Indicateur	Population-cible	Pourcentage de la population totale[a]	Nombre de ménages à visiter pour trouver un individu[b]
Couverture vaccination antitétanique2	Mères d'enfants de 0-11 mois	3%	5,56
Autres couvertures vaccinales	Enfants de 12-23 mois	3%	5,56
Couverture de la Vitamine A	Enfants de 0-23 mois	6%	2,78
Consommation de sel iodé	Ménages	-	1,0
Pourcentage avec un faible poids ou taille pour âge	Enfants de 0-59 mois	15%	1,11
Utilisation de la TRO en cas de diarrhée	Enfants de 0-59 mois avec diarrhée au cours des 2 semaines précédentes	4% (25% de 15%)	4,17
Inscription scolaire	Enfants de 5-9 ans	15%	1,11
Eau et sanitaires	Population entière	100%	0,17

[a] Estimation habituelle pour les pays en développement.
[b] En supposant 6 personnes par ménage.

La présentation de l'estimation de la taille de l'échantillon est suivie, ci-dessous, par trois exemples de tailles d'échantillon : pour estimer des tendances (Encadré 4.2), pour des analyses à un niveau sous-national (Encadré 4.3) et pour des indicateurs pour des sous-groupes –par exemple, par sexe– (Encadré 4.4).

RESUME DES CALCULS DE LA TAILLE DE L'ECHANTILLON

D'après un certain nombre de simulations basées sur les données de différents pays, la mesure des Objectifs de la mi-décennie avec une précision de 5 points de pourcentage nécessite un échantillon d'environ 4 000-5 000 ménages.

La taille de la population totale n'est pas importante pour définir la taille de l'échantillon.

Encadré 4.1
ESTIMATIONS DE LA TAILLE DE L'ECHANTILLON POUR MESURER LES OBJECTIFS DE LA MI-DECENNIE

Instructions pour l'utilisation de la Feuille de Calcul SAMPGOAL.WK1

Entrée des données

Veuillez suivre ces différentes étapes pour entrer les données :

1. Faites tourner Lotus ou Quattro-Pro (ou n'importe quel programme courant de Feuille de Calcul) et ouvrez le fichier SAMPGOAL.WK1.
2. Entrez seulement des données dans les cellules grisées. Notez que tous les pourcentages et proportions doivent être entrés en valeurs décimales (e.g., 15% doit être entré comme 0,15).
3. Entrez les données propres au pays en ce qui concerne les hypothèses de base suivantes :
   Cellules C4 et D4. Il s'agit ici de l'estimation des effets du plan de sondage dus à l'utilisation d'un sondage par grappes. Par défaut, le programme utilise 2 pour la plupart des variables et 10 pour les indicateurs eau/sanitaires (dans la mesure où ces derniers indicateurs sont probablement plus affectés par le sondage par grappes). A moins que vous n'ayez des informations locales à ce sujet, veuillez laisser ces valeurs telles qu'elles sont.
   Cellule C5. Nombre moyen de personnes par ménage, obtenu à partir d'une enquête ou recensement local.
   Cellule C6. Pourcentage de la population de moins de 5 ans, obtenu également à partir d'une enquête ou recensement local.
   Cellule C7. Prévalence de la diarrhée au cours des 15 jours précédant parmi les enfants de moins de 5 ans. Trouvez ce nombre à partir d'études précédentes réalisées dans le pays. Si vous ne disposez d'aucune information, utilisez la valeur par défaut de 0,25.
4. Entrez les données sur les estimations du niveau de couverture pour chacun des indicateurs dans les cellules D12 à D24. Vous pouvez changer les indicateurs (colonnes A et B), mais assurez-vous qu'ils se rapportent toujours à la même population-cible (colonne C), ou vous obtiendrez certainement des résultats erronés.
5. Entrez la marge d'erreur que vous êtes prêt à accepter dans les cellules E12 à E24. Ces valeurs doivent être exprimées en points de pourcentage.

Résultats

La Feuille de Calcul calcule l'échantillon-cible (à la colonne F) —par exemple, le nombre d'enfants de 12-23 mois nécessaire pour calculer chaque indicateur. Les informations fournies en haut de la Feuille sont également utilisées pour calculer combien de ménages vous devrez visiter (colonne G) pour trouver le nombre requis de personnes de l'échantillon-cible. Enfin, parmi ces nombres, le plus est sélectionné comme taille requise de l'échantillon (cellule G26).

L'application la plus utile de cette Feuille de Calcul est, peut-être, de réaliser des simulations répétées en changeant les paramètres, de façon à mesurer leur effet sur la taille de l'échantillon. En particulier, de petits changements dans la marge d'erreur peuvent avoir des effets importants sur l'échantillon nécessaire.

Le tableau suivant fournit un exemple de Feuille de Calcul.

Encadré 4.1 (suite)
Estimations de la taille de l'échantillon pour mesurer les objectifs de la mi-décennie

Hypothèses de base	Faible	Haute
Effet du plan de sondage	2	10
Personnes par ménage	6	
% de la population <5 ans	0.15	
Prévalence diarrhée/15 jours	0.25	

Objectif	Indicateur	Population cible	Prévalence estimée	Marge d'erreur	Echantillon cible requis	Nombre requis de ménages
1.1	Couverture DTCoq3	12-23 m.	0,7	0,05	672	3 733
1.2	Couverture rougeole	12-23 m.	0,4	0,05	768	4 267
1.3	Couverture VPO3	12-23 m.	0,5	0,05	800	4 444
1.4	Couverture BCG	12-23 m.	0,8	0,05	512	2 844
1.6	Couverture antitétanique3 (grossesse)	0-11 m.	0,3	0,05	672	3 733
5.1	Couverture Vitamine A	0-23 m.	0,3	0,05	672	1 867
6.1	Consommation sel iodé	Ménages	0,2	0,05	512	512
7.1	Utilisation de TRO (1) en cas de diarrhée	Diarrhée <5 ans	0,4	0,05	768	3 413
7.2	Utilisation de TRO (2) en cas de diarrhée	Diarrhée <5 ans	0,5	0,05	800	3 556
11.1	% faible poids/âge	Tous < 5 ans	0,4	0,05	768	853
12.4	Scolarisation	5-9 ans	0,7	0,05	672	747
13.1	Eau potable	Population	0,6	0,05	3 840	640
13.2	Sanitaires	Population	0,2	0,05	2 560	427
	Nombre de ménages requis					**4 444**

Encadré 4.2
CALCUL DE LA TAILLE DE L'ECHANTILLON POUR MESURER DES TENDANCES

La présentation qui précède se réfère aux calculs de taille d'échantillon nécessaire pour estimer une proportion avec une précision donnée. De tels calculs conviennent à la plupart des objectifs de la Mi-décennie et de l'an 2000, comme, par exemple, atteindre un taux de vaccination et un taux d'utilisation de la TRO de 80 pour cent. Cependant, certains des objectifs sont exprimés en termes de baisses attendues, comme, par exemple, une diminution de 20 pour cent de la prévalence de la malnutrition entre 1990 et 1995, ou une réduction d'un tiers du manque d'éducation au cours de la même période de cinq ans. Cette évaluation suppose de disposer d'au moins deux enquêtes réalisées à des époques différentes. Comme la plupart des pays ne disposent pas de données fiables pour 1990, ils ne pourront pas mesurer ces objectifs.

Les calculs de la taille de l'échantillon pour mesurer ces *tendances* sont assez complexes et ne sont pas présentées en détail. Il y a, cependant, deux méthodes empiriques utiles :

- Si vos deux échantillons sont absolument sans relation (par exemple, différentes grappes sont utilisées dans les deux enquêtes) et si votre estimation a une marge d'erreur de plus ou moins cinq points de pourcentage, le plus faible changement de tendance que vous serez capable de détecter sera de sept points de pourcentage, ou 1,4 fois votre marge d'erreur de départ.

- Si vous utilisez les mêmes grappes dans les deux enquêtes, votre précision s'améliore et vous serez capable de mesurer une modification de même amplitude (par exemple, points de pourcentage) que votre marge d'erreur*.

L'utilisation de mêmes grappes au cours d'enquêtes répétées a, à la fois, des avantages et des inconvénients. En plus de réduire la taille de l'échantillon, cette procédure simplifie le travail sur le terrain puisque les zones d'enquêtes ont déjà été cartographiées lors de la première enquête. Par contre, les enquêtes peuvent propager des messages éducatifs ou provoquer une prise de conscience des problèmes de santé par la communauté et, par conséquent, elles peuvent entraîner des changements de comportement par rapport aux questions de santé (effet connu sous le nom "d'effet de Hawthorne"). Les enquêtes suivantes dans les mêmes zones fourniront alors des informations erronées dans la mesure où ces communautés ne sont plus représentatives du pays. Lorsque des enquêtes répétées sont envisagées, chaque pays doit évaluer ces avantages et inconvénients.

* Note technique : Ceci suppose une corrélation de 0,5 entre les estimations de chaque grappe spécifique entre la première et la seconde enquête.

Encadré 4.3
TAILLES D'ECHANTILLON POUR DES ESTIMATIONS SOUS-NATIONALES

Jusqu'à présent, on s'est intéressé aux tailles d'échantillon nécessaire pour fournir des estimations, au niveau national, des indicateurs pour les objectifs de la Mi-décennie. Plusieurs de ces indicateurs sont basés sur des sous-groupes spécifiques (et souvent, réduits) de population, comme les enfants de 12-23 mois, ou les enfants qui ont eu récemment la diarrhée. Cependant, la plupart des pays souhaite également utiliser les enquêtes pour fournir des estimations à un niveau sous-national —par exemple, au niveau des régions, des états, des provinces ou des districts. Ces informations peuvent être utilisées pour identifier des zones dans lesquelles des efforts particuliers doivent être fournis, ainsi que pour des besoins de programmation et d'évaluation.

Le fait que la vaccination soit basée uniquement sur les enfants de 12-23 mois et que l'utilisation de la TRO s'applique aux enfants ayant eu la diarrhée au cours des deux semaines précédentes, sont les facteurs les plus limitatifs qui affectent la taille de l'échantillon national.

Une solution envisageable consiste à présenter, au niveau sous-national, des indicateurs pour *tous les enfants de moins de cinq ans* (ou pour les enfants de 12-59 mois, en ce qui concerne la vaccination*), ainsi qu'à accepter une marge d'erreur de plus ou moins 10 points de pourcentage. Ceci limiterait à, environ, 300 ménages par district/région les tailles d'échantillon requises au niveau sous-national. Une version à part de la Feuille de Calcul, appelée DISTRICT.WK1, est fournie pour ces calculs au niveau du district.

Avec 300 ménages, l'utilisation de la TRO ne sera toujours pas mesurée correctement. Deux options sont à envisager :

1. Etendre la période de référence de 15 jours à un mois, ce qui aura pour effet de faire passer le pourcentage d'enfants de moins de cinq ans avec la diarrhée d'environ 25 pour cent à, environ, 40 pour cent dans un district "typique". Ceci peut entraîner certains biais dus au fait que les enquêtés se souviendront plus facilement des épisodes de diarrhée les plus sévères et qui sont généralement traités de façon plus intensive; mais cela ne devrait pas affecter le classement des districts. Même avec une période de référence d'un mois, vous aurez encore besoin de 500-600 ménages par district pour une marge d'erreur de 10 points de pourcentage. Vous aurez également besoin de modifier la première question du module sur la diarrhée pour faire référence à une période d'un mois. Une question supplémentaire sera nécessaire pour séparer les épisodes de diarrhée qui se sont produits au cours des 15 jours avant l'enquête (pour les estimations des Objectifs de la mi-décennie) des autres épisodes.

2. Une seconde option consiste à demander aux responsables locaux de la planification sanitaire de regrouper les districts qui, d'après eux, ont un taux identique d'utilisation de la TRO, et d'analyser les données au niveau de plusieurs districts.

En augmentant la taille des populations-cible pour les indicateurs des objectifs de la mi-décennie, vous pourrez produire, à faible coût, des informations utiles pour le pays à des fins de programmation et, en même temps, vous pourrez fournir, au niveau national, les informations concernant les sous-groupes demandés.

* Augmenter la taille des groupes d'âges peut entraîner certaine perte de validité pour certains indicateurs, du fait de mauvaises déclarations; cependant, ce biais possible sera certainement minime du point de vue des planificateurs de la santé. En outre, on peut s'attendre à ce que ce biais agisse uniformément sur tous les districts, ce qui n'affectera pas le classement des zones du point de vue de la priorité des interventions.

Encadré 4.4
TAILLES D'ECHANTILLON POUR DES ANALYSES DE SOUS-GROUPES

Les analyses de sous-groupes peuvent comprendre des décompositions des indicateurs par sexe, par groupe socio-économique, etc.. Les indicateurs qui portent sur des sous-groupes seront moins précis que ceux calculés pour l'échantillon dans son ensemble.

Plus le sous-groupe est petit, moins l'estimation est précise.

Les exemples ci-dessous montre de combien augmente la marge d'erreur en fonction de la diminution de la taille des sous-groupes. Sur la base d'un échantillon global avec une marge d'erreur de plus ou moins cinq points de pourcentage pour un indicateur donné*, on devrait avoir :

- Une marge d'erreur d'environ ± 6,3 points de pourcentage pour des indicateurs par sexe spécifique (avec 50 pour cent de garçons et 50 pour cent de filles dans l'échantillon).
- Une marge d'erreur d'environ ± 8,6 points de pourcentage pour un sous-groupe correspondant à 20 pour cent de l'échantillon (par exemple, une catégorie socio-économique donnée).

Ces résultats montrent que si la marge d'erreur globale est d'environ cinq points de pourcentage, des résultats assez précis seront également obtenus pour des indicateurs par sexe spécifique ainsi que pour d'autres sous-groupes correspondant à 20 pour cent ou plus de l'échantillon total.

* **Note technique** : Basé sur une estimation de la couverture de 50 pour cent et un effet de plan de sondage de 2,0. On suppose également que la couverture de chaque sous-groupe est proche de 50 pour cent.

Décisions concernant le nombre total de grappes

Un élément de base à prendre en compte pour décider de la taille de vos grappes est l'estimation du nombre de ménages qui pourront être enquêtés en un jour par votre équipe d'enquêteurs. Vous ne souhaitez pas que vos agents de terrain passent plus d'une journée dans une grappe donnée et vous ne souhaitez pas, non plus, qu'ils terminent leurs interviews à, disons, 3 heures de l'après-midi et qu'ils n'aient rien d'autre à faire pour le reste de la journée.

✎ **EXEMPLE :**

Supposons que votre interview prenne 10 minutes pour chaque enfant de moins de 5 ans et 3 minutes pour les enfants de 5-9 ans (module sur l'éducation seulement), plus 5 minutes pour les questions sur le ménage. Un ménage type comprendra un enfant de chacun de ces groupes (voir Tableau 4.1) et l'enquête prendra donc 18 minutes. Ajoutez 5 minutes pour se déplacer de maison à maison. En envisageant une journée de travail de 6 heures (pour permettre le transport, vers et en provenance des zones sélectionnées et une pause à mi-journée), un enquêteur couvrira environ 15 ménages par jour [(6 heures x 60 minutes) ÷ 23 minutes]. Une équipe de deux enquêteurs couvrira 30 ménages par jour, ce qui devient alors la taille de la grappe.

☞ Vous devrez faire un pré-test pour estimer la durée de chaque interview. La formation contribuant à réduire de façon importante la durée de l'interview, les agents de terrain doivent bien connaître le questionnaire avant le pré-test . A cause des distances à parcourir, le travail de terrain en zone rurale est toujours beaucoup plus lent qu'en zone urbaine.

Les tailles type de grappe varient de 20 à 40 ménages. N'utilisez pas de grappes comportant plus de 100 ménages. Il est toujours préférable d'utiliser *davantage de petites grappes* que moins de grappes plus grandes.

Le nombre de grappes est donné par la formule :

$$Nombre\ de\ grappes = \frac{Nombre\ total\ de\ ménages\ dans\ l'échantillon}{taille\ de\ la\ grappe}$$

✎ **EXEMPLE :**
En supposant que l'équipe d'enquêteurs peut s'occuper de 30 ménages par jour et en visant un échantillon national de 4 500 ménages, vous aurez besoin de 4 500/30 = 150 grappes.

La plupart des échantillons nationaux incluront 100 grappes ou plus. pour des estimations à un niveau sous-national, essayez d'avoir 15 à 20 grappes dans chaque province, état ou district. Dans le doute, augmentez le nombre de grappes à sélectionner.

Utilisez un ordinogramme pour choisir votre plan de sondage

Quand vous choisissez un plan de sondage, il est préférable de rechercher l'échantillon le plus précis et le moins biaisé. Cependant, ce choix dépend également des données sur la population disponibles pour élaborer le plan de sondage. Enfin, les facteurs temps, faisabilité et coût joueront également un rôle majeur.

A moins que l'enquête puisse être accomplie en quelques mois et à un coût raisonnable, elle ne peut pas être utilisée pour le suivi des objectifs et la planification à un niveau sous-national.

Un grand nombre de combinaisons de différents types de sondage peuvent être utilisées pour produire un échantillon à plusieurs degrés pour une enquête à indicateurs multiples. L'ordinogramme présenté au Graphique 4.4 (voir page 4.21) montre les décisions qui doivent être prises. Cet exemple se réfère à un échantillon à trois degrés dans lequel de grandes zones sont d'abord sélectionnées; le deuxième degré consiste en la sélection de petites zones à l'intérieur des grandes zones; et le troisième degré consiste en la sélection de ménages à l'intérieur de chaque petite zone. Les alternatives finales,

numérotées A à D et qui figurent dans la partie inférieure droite de l'ordinogramme, sont classées par ordre de préférence.

☞ Pour concevoir le plan de sondage, vous aurez besoin de l'aide de spécialistes qui connaissent les différents types de données sur la population disponibles dans le pays (comme des cadres ou des démographes du bureau du recensement), ainsi que d'experts en sondage. Si des experts en sondage ne sont pas disponibles localement, une assistance internationale devra être prévue.

La vaste expérience acquise au cours des enquêtes PEV dans les années 1980 a montré qu'une enquête n'a pas à être parfaite pour fournir des données utilisables. Les méthodes d'enquête utilisées doivent être suffisamment satisfaisantes pour répondre aux objectifs pour lesquels les résultats seront utilisés.

Comment sélectionner les grandes zones

La première question posée dans l'ordinogramme est de savoir si des données récentes sur la population des grandes zones sont disponibles. Il peut s'agir de divisions administratives comme des états ou des provinces, des districts, des municipalités, etc... Généralement les grandes zones ont des populations qui s'échelonnent de plusieurs milliers à plusieurs millions d'habitants. Pour utiliser un sondage probabiliste, il est nécessaire de disposer de données appropriées sur la population de ces grandes zones –c'est-à-dire, une base de sondage. Plusieurs options sont possibles (par ordre de précision) :

1. Vous pouvez vous référer à un recensement récent (disons, effectué dans les 5-10 ans).

2. Une Enquête Démographique et de Santé (EDS) récente ou une autre enquête ménage récente au niveau national constitue une autre option. Bien que ces enquêtes ne fournissent généralement pas de données sur la population en soi, les bases de sondage au niveau national doivent être mises à jour avant que ces enquêtes ne soient réalisées.

3. Si aucun recensement ou base de sondage récent n'est disponible, mais que, depuis le dernier recensement, la croissance de la population ait été relativement uniforme à travers tous le pays, les données sur la population du recensement peuvent être utilisées.

4. Si aucun recensement ou autre base de sondage récent n'est disponible et si la croissance de la population a été variable (du fait de migrations ou de catastrophes, par exemple), vous devrez essayer d'estimer les changements intervenus dans la population régionale en consultant les experts locaux, y compris les démographes, les planificateurs ou les employés des bureaux de statistiques.

Bien que vous deviez chercher à obtenir la meilleure base de sondage possible, elle ne doit pas, nécessairement, être parfaitement exacte. Les degrés suivants de sondage peuvent permettre de corriger certaines des inexactitudes de votre base.

Les Options 1 à 3 présentées ci-dessus fournissent les données permettant d'utiliser la population des grandes zones pour mettre en oeuvre le premier degré de la procédure d'échantillonnage. Si elle est réussie, l'Option 4 permet également cette mise en oeuvre.

Une fois que la base de sondage est disponible, vous pouvez réaliser le tirage de l'échantillon au premier degré en utilisant la stratification, les probabilités proportionnelles à la taille (PPT) ou une combinaison des deux procédures. Ce degré peut également être subdivisé en deux ou plus sous-degrés.

✎ **EXEMPLE** (1) :

Supposons que vous vouliez des estimations séparées pour chaque province du pays. Vous procéderez à la stratification du pays en provinces et prendrez des échantillons de taille égale dans chacune d'entre elles pour obtenir les estimations au niveau provincial. Pour estimer les indicateurs au niveau national, il sera nécessaire de pondérer les résultats des provinces en fonction de la population de chaque province.

✎ **EXEMPLE (2) :**

Supposons que vous ne puissiez pas vous permettre d'obtenir des estimations séparées pour chaque province, mais que votre pays soit divisé en trois régions approximativement identiques, la capitale (avec 30 pour cent de la population), les provinces de la côte (avec 50 pour cent) et les provinces de l'intérieur (avec 20 pour cent). Vous pouvez alors sélectionner des échantillons de même taille dans chaque province pour obtenir des estimations régionales. En pondérant ultérieurement les résultats en fonction de la population, vous produirez les estimations nationales.

Dans quelques pays, en particulier ceux en situation de crise, aucune des options présentées ci-dessus pour obtenir une base de sondage peut être applicable (voir l'ordinogramme). On devra alors travailler avec des personnes bien informées de la situation locale pour stratifier le pays en vastes zones. Cette approche ne fournit pas un échantillon probabiliste au sens strict du terme du fait du manque d'information sur la population. Cependant, si aucune base de sondage convenable n'est disponible, cette option peut, au moins, fournir des informations approximatives sur le niveau des indicateurs.

✎ **EXEMPLE :**

En conservant l'exemple ci-dessus, le pays peut être divisé en trois régions, la capitale, la région côtière et la région de l'intérieur, et un échantillon de taille similaire peut être tiré dans chaque région. Contrairement au dernier exemple, il ne sera pas possible de pondérer les résultats puisque vous ne connaîtrez pas la population de chaque strate.

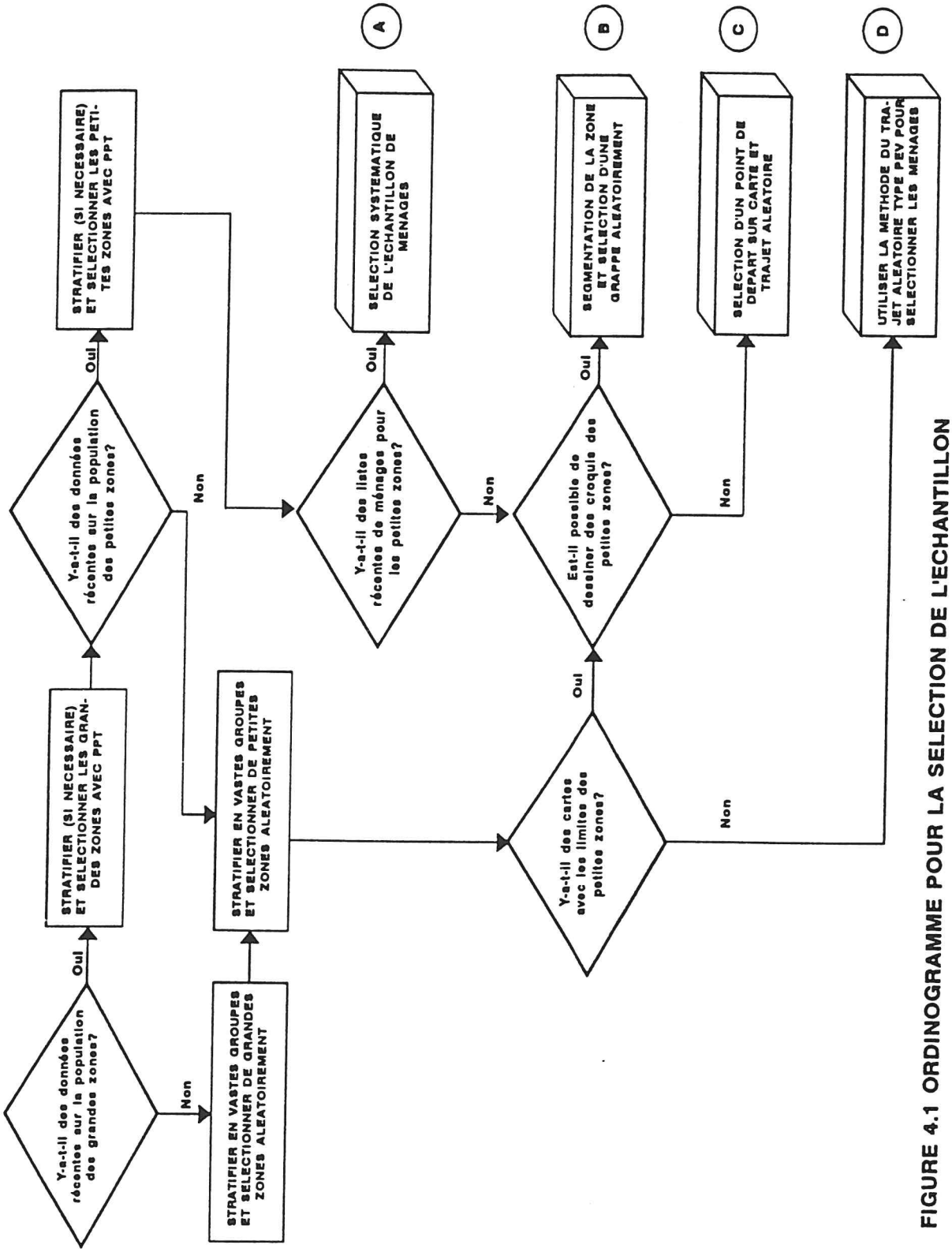

FIGURE 4.1 ORDINOGRAMME POUR LA SELECTION DE L'ECHANTILLON

Une fois l'échantillon des grandes zones choisi, l'étape suivante consiste à sélectionner les petites zones.

Comment sélectionner les petites zones

La seconde question posée dans l'ordinogramme porte sur les petites zones. Ces zones sont composées des sections d'énumération ou des petites unités administratives, comme les communes, villages ou unités administratives qui se situent au-dessous des municipalités. Généralement, les petites zones sont composées de moins de 1 000 ménages. Des données sur la population et des cartes délimitant les limites de ces zones sont nécessaires si l'on doit utiliser un sondage probabiliste à ce niveau. L'ordinogramme suppose que si des données sur la population sont disponibles, des cartes le seront également.

Si un recensement a été réalisé récemment, les données sur les sections d'énumération devraient être disponibles. Cependant, ces données peuvent ne pas être facilement accessibles aux chercheurs. Si elles sont disponibles, les données sur les sections d'énumération ont l'avantage de (1) découper la population en sections mutuellement exclusives de façon que chaque ménage appartient à une section et seulement à une section, et (2) être généralement de taille uniforme puisqu'elles sont le plus souvent définies comme des sections qu'un seul agent recenseur peut couvrir.

Une alternative à l'utilisation des sections d'énumération est l'utilisation des petites unités administratives locales, pourvu que ces unités incluent toute la population et qu'il n'y ait pas de chevauchement entre différentes unités.

Une fois que les informations sur la population des petites zones ont été obtenues, ces zones peuvent être tirées avec PPT. Le tableau illustrant le tirage avec PPT dans la section ci-dessus montre comment les villages sont tirés d'un district.

Si on utilise le village comme unité de sondage, on doit s'assurer que chaque ménage rural isolé appartient bien au village donné.

Si les sections d'énumération sont utilisées et qu'elles sont approximativement de même taille, vous pouvez utiliser un échantillonnage aléatoire simple plutôt qu'un PPT puisque chaque ménage aura approximativement la même probabilité d'être sélectionné.

Si, en suivant l'ordinogramme, vous trouvez que les données sur les petites zones ne sont pas disponibles, vous devrez travailler avec des personnes bien informées de la situation locale pour diviser les grandes zones en vastes groupes de petites zones, puis faire un tirage aléatoire à l'intérieur de chaque groupe.

✎**EXEMPLE :**

Supposons que les grandes zones (district) comprennent une grande ville ainsi que des petites villes, des villages et des ménages isolés. Supposons également qu'il y ait des populations de montagne et de plaine. Des strates devront être créées en fonction, si possible, de ces caractéristiques et en utilisant des estimations grossières de la population. Les personnes bien informées de la situation locale devront alors essayer

de dresser une liste complète des noms de toutes les zones (par exemple, villes, villages ou quartiers ruraux) de chaque strate. Un certain nombre de petites zones -basé sur les estimations grossières de la population- seront alors sélectionnées aléatoirement dans chaque strate.

Comme mentionné précédemment, le recours à des personnes bien informées ne fournit pas un échantillon probabiliste au sens strict du terme mais, dans certaines circonstances, cela peut constituer la seule option possible.

Une fois les petites zones sélectionnées, on doit alors choisir les grappes comprenant les ménages à enquêter.

Comment sélectionner les grappes

La plupart des petites zones comprendront un nombre de ménages plusieurs fois supérieur à la taille choisie de nos grappes. Il est alors nécessaire de sélectionner les ménages à l'intérieur de ces zones. L'ordinogramme montre qu'il y a quatre façon élémentaires de procéder à cette sélection : le dénombrement (option A dans l'ordinogramme), la segmentation (option B), ou la méthode du trajet aléatoire du PEV (subdivisée plus loin en options C et D). Les principaux avantages et inconvénients de ces méthodes sont synthétisés au Tableau 4.2.

En choisissant une méthode, vous devez tenir compte de ses avantages et inconvénients, ainsi que des contraintes locales, y compris la disponibilité de cartes et d'informations sur la population, les facteurs temps et coût, les caractéristiques géographiques du pays et le genre de personnel de terrain disponible.

Enumération. L'énumération consiste à cartographier et à lister tous les ménages des petites zones, puis de tirer un échantillon, généralement de façon systématique. C'est la méthode à choisir dans les pays où des listes et des cartes sont disponibles (par exemple, dans les établissements sanitaires), Cependant, dans le contexte des enquêtes conçues pour fournir des indicateurs multiples, en raison du temps nécessaire pour réaliser l'étape initiale d'énumération et de cartographie des ménages, on ne recommande pas cette méthode à moins que des listes récentes de ménages ne soient disponibles (voir Tableau 4.2). Une liste "récente" de ménages ne doit pas dater de plus de deux ans.

Segmentation. L'ordinogramme montre que l'option suivante est la segmentation. Notez que, si des données sur la population ne sont pas disponibles, mais si des cartes où figurent les limites de zones le sont, la segmentation peut encore être utilisée

La segmentation nécessite de circuler ou de marcher à travers toute la petite zone et de dresser une carte grossière avec la localisation des ménages. Cette carte ne doit pas nécessairement être parfaitement exacte de même qu'il n'est pas nécessaire d'avoir une liste complète de tous les ménages. Cependant, la carte doit couvrir la petite zone en entier et les segments doivent être exclusifs l'un par rapport à l'autre..

Tableau 4.2 Avantages et inconvénients des principales méthodes de sélection des ménages à l'intérieur des petites zones

Méthode de sélection	Avantages	Inconvénients
Enumération	• Echantillon strictement probabiliste. • S'adapte aux changements de la population. • N'utilise pas des ménages contiguës et est donc plus proche d'un échantillon aléatoire simple.	• Nécessite de lister et cartographier chaque ménage. • Les ménages ne sont pas contigus, le travail de terrain dure donc plus longtemps. • Les logements doivent être identifiés avec précision.
Segmentation	• Echantillon strictement probabiliste. • S'adapte aux changements de la population. • Economise du temps en utilisant des ménages contiguës.	• Nécessite du temps supplémentaire pour faire des croquis. • L'utilisation de ménages contigus peut affecter l'homogénéité de l'échantillon. • Les superviseurs ont besoin d'une formation approfondie pour la cartographie et la segmentation des zones.
Trajet aléatoire	• Economise du temps en utilisant des ménages contiguës. • Ne nécessite pas de cartographie. • Très connu et utilisé dans de nombreux pays par des non-experts. • Convient particulièrement bien aux populations dispersées et aux vastes zones. • Peut produire un échantillon non-biaisé.	• Echantillon non strictement probabiliste. • Ne tient pas compte des changements de la population. • L'utilisation de ménages contigus peut affecter l'homogénéité de l'échantillon. • Nécessite une formation approfondie des enquêteurs pour identifier les logements et une étroite supervision.

Après avoir dressé le croquis, la zone est divisée en segments de taille approximativement égale. La taille de chaque segment est donnée par la formule ci-dessous :

$$\textit{Taille du segment} = \frac{\textit{Nombre réel de ménages dans la zone} \times \textit{taille souhaitée de la grappe}}{\textit{nombre estimé de ménages dans la zone}}$$

✎ EXEMPLE :

D'après le dernier recensement, vous estimez qu'il devrait y avoir 510 ménages dans la zone. Comme vous voulez des grappes de 40 ménages, en moyenne, la zone devra être divisée en 12,75 segments (ou 13, en arrondissant). Votre croquis montre que, depuis le recensement, le nombre de ménages est passé à 600. Divisez la zone en 13 segments de taille égale de 46 ménages (arrondi de 46,15). Sélectionnez un ces

segments de façon aléatoire et intégrez tous les ménages à votre échantillon.

Comme elle permet une correction automatique des inexactitudes de la base de sondage et aussi parce qu'elle permet d'obtenir un échantillon strictement probabiliste, la segmentation doit être utilisée chaque fois qu'il est possible de visiter la zone sélectionnée dans sa totalité, de dresser des croquis et de segmenter la zone avant la collecte des données. Les facteurs à prendre en compte pour déterminer la faisabilité comprennent le temps, le coût, la taille géographique et la disponibilité de personnel de supervision expérimenté -par exemple, personnel du département des statistiques ou du bureau du recensement. Le Chapitre 6 décrit en détails les différentes étapes à suivre pour procéder à la segmentation.

Un avantage important des méthodes d'énumération et de segmentation consiste dans le fait qu'elles permettent de corriger les inexactitudes des bases de sondages pour les petites zones.

Méthodes du trajet aléatoire. Les méthodes du trajet aléatoire sont basées sur la méthode des enquêtes PEV en grappes. Ces méthodes ne fournissent pas d'échantillons strictement probabilistes mais, dans de nombreux cas, elles peuvent fournir des échantillons de ménages pratiquement non biaisés. Elles sont aussi très faciles et rapides à mettre en oeuvre et, dans certains pays, elles peuvent constituer la seule alternative possible.

Même lorsque des cartes des petites zones sont disponibles, il peut ne pas être possible de faire des croquis comportant les ménages. Ceci peut se produire lorsque les unités de sondage à ce degré (par exemple, les sections d'énumération ou les villages) couvrent de vastes zones géographiques et/ou lorsque les ménages sont dispersés dans toute la zone. Dans de tels cas, la cartographie pourrait prendre beaucoup de temps. Vous choisirez alors l'option C, en choisissant aléatoirement un des différents points de départ sur la carte où figurent les limites de zone et vous utiliserez ce point de départ pour démarrer le trajet aléatoire. Cette méthode est décrite plus loin au Chapitre 6.

Si vous ne disposez pas de carte où figurent les limites de la zone, un certain nombre d'adaptations de la méthode du trajet aléatoire du PEV peuvent être utilisées (option D). La méthode d'origine et ses limites sont présentées dans l'Encadré 4.5 et le Chapitre 6 présente une version modifiée de la méthode qui résout certains de ces problèmes.

Par rapport à l'option D, l'option C présente l'avantage que vous pouvez utiliser une carte de limites existante pour positionner les points de départ à l'intérieur de la petite zone. Avec l'option D, les cartes ne sont pas disponibles, aussi vous devez avoir recours à d'autres moyens pour trouver un point de départ, ce qui est expliqué dans l'Encadré 4.5 ainsi qu'au Chapitre 6.

EXEMPLES DE PAYS

L'expérience de trois pays –Bangladesh, Brésil et Jordanie– dans la réalisation d'enquêtes à indicateurs multiples est décrite aux sections suivantes.

Bangladesh

Des enquêtes à indicateurs multiples ont été réalisées au Bangladesh, à l'échelon national, en 1992, 1993 et 1994. Ces enquêtes visaient à fournir des données sur les indicateurs visant à mesurer les objectifs de la mi-décennie, au niveau national, ainsi que des informations à but programmatique, au niveau du district. A partir de la base de sondage du Recensement de 1991, le pays a été divisé en 75 strates (ou grandes zones) : les zones rurales de 64 districts et 11 domaines urbains, comprenant les quatre grandes zones métropolitaines –deux d'entre elles ont été, ultérieurement, décomposées en strates de bidonville et de non bidonville– et des échantillons des autres zones urbaines.

Dans chaque grande zone, toutes les unions (unités administratives de 3 000-4 000 ménages) ont été listés selon un ordre géographique, et 13 d'entre elles ont été sélectionnées avec PPT. Pour chaque union sélectionnée, tous les *mouzas* (en milieu rural) ou *mahallahs* (en milieu urbain) ont été listés. Les *mouzas* et les *mahallahs* sont des subdivisions administratives comprenant, chacune, environ 250-300 ménages; elles correspondent aux "petites zones" dans cette description de l'enquête du Bangladesh.

Après que les listes aient été dressées, une petite zone a été sélectionnée dans chaque union avec, encore une fois, PPT. Par conséquent, 13 petites zones ont été choisies dans chaque district ou dans chaque domaine urbain. Dans chacune de ces 13 zones, la méthode de la segmentation a été utilisée. La plupart des zones avaient déjà des cartes à jour, positionnant tous les ménages, qui avaient été établies par des agents de santé, si bien que des croquis n'ont été nécessaires que dans quelques cas. Chaque segment contenait, en moyenne, 40 ménages. Un segment a été sélectionné aléatoirement dans chaque petite zone et tous ses ménages ont été intégrés à l'échantillon. Chaque district comprenait donc, en moyenne, 520 ménages (13 x 40). Cet effectif était suffisant pour estimer tous les indicateurs des objectifs de la mi-décennie avec une marge d'erreur de 10 points de pourcentage, ou moins.

Les données des différents échantillons ruraux et urbains ont été pondérés pour produire les estimations des indicateurs au niveau national. Comme l'échantillon national comportait 39 000 ménages (520 x 75), sa marge d'erreur était beaucoup plus petite que les cinq points de pourcentage utilisés dans les exemples ci-dessus. Ce grand échantillon résultait du grand nombre de districts dans le pays et du besoin en données pour chaque district. Comme cela a été mentionné dans la section sur la taille de l'échantillon, la plupart des pays ont besoin d'échantillons beaucoup plus petits.

Une autre particularité des enquêtes du Bangladesh réside dans l'utilisation de 13 grappes par district alors que, ci-dessus, au moins 20 grappes ont été recommandées. Ce choix était basé sur des

considérations pratiques résultant du grand nombre de districts et de strates urbaines, ainsi que sur le fait que ces zones étaient géographiquement de petite taille et que les caractéristiques de la population à l'intérieur de ces zones semblaient être relativement uniformes. L'échantillon national comprenait 975 grappes (13 x 75).

L'exemple du Bangladesh montre qu'il est possible de combiner la procédure permettant d'obtenir des estimations nationales pour suivre les objectifs et la production de données, au niveau du district, dans un but programmatique.

Brésil

Des enquêtes à indicateurs multiples ont été réalisées dans plusieurs états du Nord-Est du Brésil au début des années 1990. Ces enquêtes visaient à mesurer des indicateurs de santé et de nutrition des enfants dans un but de suivi et pour des besoins programmatiques. L'exemple qui suit est basé sur l'enquête de 1991 dans l'Etat de Pernambuco qui a une population d'environ sept millions d'habitants.

Un échantillon de 1 000 enfants de moins de cinq ans était nécessaire pour estimer les indicateurs avec une marge d'erreur de plus ou moins trois points de pourcentage, en utilisant un effet de plan de sondage de 2,0. A partir d'informations sur la taille des ménages et sur la proportion de la population de moins de cinq ans, 1 920 ménages ont été visités. Comme le questionnaire était long -il portait sur les enfants de moins de cinq ans ainsi que sur les femmes de 15-49 ans- et que les ménages étaient souvent dispersés, la taille de la grappe avait été fixée à 12 ménages.

Au premier degré du sondage, toutes les municipalités ("grandes zones") ont été listées et numérotées en partant de la capitale et en suivant un ordre géographique. En utilisant un tirage avec PPT, 20 numéros ont été sélectionnés. Comme la capitale comprenait 15 pour cent de la population, elle a été sélectionnée trois fois (0,15 x 20).

Au second degré, huit sections d'énumération ("petites zones") ont été choisies par simple tirage aléatoire dans chaque municipalité sélectionnée: au total, 160 grappes (20 x 8) ont été tirées dans l'état. Dans la capitale, qui avait été sélectionnée trois fois, 24 sections d'énumération (3 x 8) ont été tirées. Chaque section d'énumération comprenant 200-300 ménages et leur taille étant uniforme, à ce degré de sondage, il n'a pas été nécessaire de procéder à un tirage avec PPT. Ces sections couvrent toute la municipalité, y compris les ménages ruraux et urbains. Quatre à six points de départ étaient positionnés sur les cartes de chaque section d'énumération et un de ces points étaient tirés aléatoirement. Douze ménages proches de ce point et appartenant à la section d'énumération étaient alors inclus dans l'échantillon.

Remarquez que cet échantillon était auto-pondéré, si bien qu'il n'a pas été nécessaire de pondérer les estimations.

Jordanie

Une enquête nationale a été réalisée en Jordanie, en 1991, en suivant la méthodologie de l'enquête PEV par grappes, dans le but d'évaluer la couverture vaccinale, le suivi de la diarrhée et la mortalité infantile. Un échantillon de 12 600 ménages était nécessaire pour l'estimation de la mortalité infantile par cause. Remarquez que cet échantillon est considérablement plus grand que celui généralement nécessaire pour estimer la couverture vaccinale ou celle de la TRO, mais les études de la mortalité nécessitent souvent de visiter de nombreux ménages pour identifier un nombre suffisant de décès.

Encadré 4.5
LA METHODE STANDARD DE L'ENQUETE PEV PAR GRAPPES

Dans la méthode standard PEV, 30 villages étaient sélectionnés dans un district avec PPT. Un point de départ était sélectionné à une position centrale, par exemple un établissement sanitaire. On faisait tourner une bouteille à ce point central et l'enquêteur comptait alors le nombre de ménages jusqu'à la limite du village et sélectionnait un ménage au hasard. Il s'agissait du premier ménage de l'échantillon. Les ménages suivants étaient sélectionnés en se rendant dans le logement dont la porte d'entrée était la plus proche de la maison qui venait d'être visitée, jusqu'à ce que sept enfants de 12-23 mois (la tranche d'âges intéressant pour la couverture vaccinale) soient trouvés. Cette méthode est connue sous le terme de "trajet aléatoire".

Un certain nombre de problèmes ont été soulevés concernant cette méthode, dont :

1. Dans la plupart des pays, les populations rurales ne sont pas bien organisées en villages, avec des frontières ou limites bien définies, mais vivent dans des logements plus ou mois éparpillés.

2. Le plan de sondage peut suréchantillonner les ménages proches du centre du village qui peuvent ne pas être représentatifs du village dans son ensemble. En se déplaçant d'un ménage au ménage suivant le plus proche, il est probable que les ménages ruraux dispersés soient sous-représentés.

3. En s'arrêtant lorsqu'on atteint un nombre fixe d'enfants (non de ménages), la méthode suréchantillonne les zones avec peu d'enfants (où beaucoup de ménages doivent être visités pour trouver sept enfants) et sous-représente les zones avec beaucoup d'enfants (où sept enfants sont trouvés dans peu de ménages).

4. Il est très difficile de vérifier que les agents de terrain ont bien suivi les instructions à propos du "trajet aléatoire".

5. La méthode de sélection ne garantit pas que chaque ménage a une probabilité connue d'être sélectionné. Par conséquent, il ne s'agit pas d'un échantillon strictement probabiliste. Il s'agit ici d'un point assez technique, mais il a inquiété de nombreux statisticiens parce que, à strictement parler, on ne peut pas calculer les marges d'erreur des indicateurs.

Ces problèmes ne diminuent pas l'utilité des enquêtes type-PEV pour estimer la couverture vaccinale à des fins programmatiques, pour lesquelles une large marge d'erreur peut être tolérée. Ces problèmes montrent, cependant, que la méthode peut être améliorée.

Au premier degré du sondage, le Royaume a été divisé en sept gouvernorats. Comme des estimations séparées étaient recherchées pour chacun des gouvernorats, l'échantillon PEV standard de 30 grappes par gouvernorat a été adopté. Chaque grappe comprenait 60 ménages, grâce à quoi l'échantillon national atteignait la taille souhaitée (30 x 60 x 7 = 12 600). L'utilisation d'échantillons fixes dans tous les gouvernorats, indépendamment de leur population, a nécessité, ultérieurement, une pondération, au moment de l'analyse.

A l'intérieur de chaque gouvernorat, un sondage à plusieurs degré a été utilisé. Tout d'abord, les localités ont été stratifiées en neuf groupes, selon leur taille en population, et un certain nombre de grappes ont été attribuées à chaque strate sur la base de sa population totale. Le nombre nécessaire de localités était alors sélectionné à l'intérieur de chaque strate avec PPT. Ensuite, des pâtés de maisons ont été choisis à l'intérieur de chaque localité tirée et, enfin, des ménages ont été sélectionnés à l'intérieur de chacun de ces pâtés de maisons.