CHAPITRE VII

TRAITEMENT DES DONNEES

Ce chapitre s'adresse au coordinateur de l'enquête et aux cadres techniques.
Il vous aidera à :

- ✔ Effectuer la saisie des données.
- ✔ Vérifier et éditer les données, et à créer un fichier de données apurées pour l'analyse.
- ✔ Evaluer la qualité des données de l'enquête.
- ✔ Calculer les estimations des indicateurs.

INTRODUCTION

Le traitement des données doit démarrer aussi rapidement que possible. Les premiers résultats de l'enquête, à savoir le rapport préliminaire, devraient paraître rapidement, quelques semaines après la fin des travaux sur le terrain. Un rapport de fin de mission complet de l'équipe d'enquête, passant en revue les premiers résultats, contribue à l'élaboration d'un rapport final de bonne qualité. L'enthousiasme et l'intérêt qu'a occasionné le travail sur le terrain disparaîtront rapidement à la fin des opérations de collecte, aussi assurez-vous de conserver le crédit que les résultats de l'enquête vous procurent en les publiant rapidement.

Les enquêteurs doivent vérifier les questionnaires lorsqu'ils ont fini une interview : cela fait partie des procédures de terrain pour réduire les erreurs. Les erreurs peuvent être corrigées sur le champ sans qu'il soit nécessaire de procéder à une réinterview. Les superviseurs de terrain doivent aussi vérifier tous les questionnaires dès qu'ils les récupèrent des enquêteurs. Sont-ils complets? Sont-ils clairs? Les instructions de "saut" ont-elles été suivies correctement? Le responsable de l'enquête doit aussi vérifier tous les questionnaires au début du travail sur le terrain et continuer à vérifier des échantillons de questionnaires pendant toute l'enquête.

Il est préférable de commencer la saisie des données pendant que les enquêteurs sont sur le terrain. Ceci permet de repérer et de corriger des erreurs que certains enquêteurs ou équipes peuvent commettre. Les problèmes sérieux qui peuvent échapper à l'attention du superviseur de terrain peuvent ainsi être détectés rapidement, suffisamment tôt pour reprendre la formation du personnel de terrain et corriger les erreurs importantes. Les superviseurs de terrain ont pour responsabilité de contrôler la complétude et la cohérence des questionnaires, et de classer toutes les réponses que l'enquêteur n'a pas pu coder. A part vérifier le nombre de questionnaires retournés du terrain et enregistrer les numéros d'identification sur des formulaires, le personnel de bureau ne devrait pas avoir besoin de procéder à l'édition ou à la codification des questionnaires qui arrivent au bureau.

Si des modifications doivent être apportées au questionnaire, les programmes de saisie et d'analyse devront être modifiés par un programmeur ayant une connaissance approfondie du logiciel EPI INFO.

Le système de saisie doit être bien conçu et le personnel de saisie doit être bien formé et motivé. Les programmes de saisie des données, de vérification et d'analyse doivent être mis au point et testés en utilisant les données de l'enquête pilote avant que ne commence la phase d'enquête principale. Tous les problèmes doivent être résolus et, si nécessaire, les programmes doivent être modifiés avant que les données de l'enquête ne commence à parvenir du terrain.

Le traitement informatique des données des questionnaires comporte quatre phases principales:

1. Saisie des données
2. Vérification des données et édition
3. Distribution des fréquences
4. Elaboration des variables et tabulations

Il existe différents logiciels permettant de réaliser ces opérations. Un programme complet de saisie des données, de vérification et d'analyse pour calculer les sept indicateurs pour les objectifs de la mi-décennie (c'est-à-dire ceux qui peuvent être calculés au moyen d'une enquête à indicateurs multiples) est fourni avec ce manuel pour être utilisé avec le logiciel EPI INFO (Version 6). Les programmes sont également listés en Annexe 3.

EPI INFO a été développé par le Centers for Disease Control and Prevention en collaboration avec l'OMS (disponible auprès de USD, Inc., 2075 A West Park Place, Stone Mountain, GA 30087, USA) et permet de mener à bien les quatre phases du traitement des données. Une copie de EPI INFO (Version 6) est incluse dans ce manuel.

Le responsable du traitement des données doit mettre en place, à l'avance, le système de saisie des données. Des conseils pour cette mise en place sont présentés également en Annexe 3.

SAISIE INFORMATIQUE DES DONNEES ET PRODUCTION DE FICHIERS DE DONNEES APUREES

Saisie des données

Aussitôt que les données d'une grappe arrivent au bureau central, la saisie des données doit commencer. La saisie doit être menée par petit lot (par exemple, une grappe à la fois).

Les données des questionnaires doivent être saisies pour chaque ménage suivant l'ordre dans lequel elles ont été collectées : les données sur le ménage de la feuille sur les membres du ménage, les données sur tous les enfants de la liste des enfants, les modules sur l'eau et les sanitaires et l'iodation du sel, les donnés sur l'éducation pour chaque enfant ayant atteint l'âge d'entrée à l'école,

le module sur la vaccination antitétanique pour chaque mère, le module sur le suivi des maladies respiratoires aiguës pour chaque mère et, en dernier, les modules sur la santé des enfants pour chaque enfant. Une fois que toutes les données sont saisies pour un ménage, l'agent de saisie continue en entrant les données du ménage suivant.

Remarquez que les informations non déterminées pendant le travail sur le terrain sont enregistrées par des 9 -par exemple, 9, 99, 999, etc. Les champs vides au moment de la saisie et les réponses inapplicables (par exemple, pour des réponses à des questions optionnelles ou à des questions sautées) sont enregistrées comme des champs vierges dans la base de données. Quand les données de chaque ménage sont saisies, le programme procède à un *contrôle d'étendue*. Ceci garantit que, pour une rubrique donnée, aucune valeur saisie n'est en dehors de l'étendue donnée des réponses pour cette rubrique.

✎ **EXEMPLE :**
La date de l'interview doit être enregistrée pendant la période du travail sur le terrain. Une date d'interview enregistrée comme 01 11 93 est en dehors de l'étendue des dates permises pour cette enquête. Si le programme de saisie a déterminé l'étendue de cette variable, l'ordinateur reconnaîtra que cette date saisie est invalide.

Il y a deux façons d'entrer les données. Les données peuvent être saisies deux fois, par deux agents de saisie différents. Les deux fichiers sont ensuite comparés pour incohérences et d'autres vérifications sont réalisées. Une autre approche consiste à saisir les données une seule fois, mais à procéder à un certain nombre de vérifications pour éviter les erreurs au moment de l'entrée des données. Une vérification interactive des données peut ralentir le processus de saisie, mais peut aussi prévenir de nombreuses erreurs de saisies commises par l'agent. Des vérifications de cohérence sont alors exécutées et le fichier corrigé. Pour ces enquêtes à indicateurs multiples, nous conseillons la dernière option, qui est l'approche qui a été utilisée dans les programmes de saisie fournis avec ce manuel.

Encadré 7.1
POUR ET CONTRE LA DOUBLE SAISIE

Pour

Les erreurs de saisie sont pratiquement éliminées si la double saisie est faite correctement. Cette option, suivie par des contrôles d'erreurs en série peut se révéler peu coûteuse et rapide quand on emploie des agents de saisie professionnels.

Contre

La double saisie double le coût de la saisie des données et elle complique la circulation des questionnaires dans le bureau, ce qui peut aussi conduire à des erreurs.

Un compromis possible est d'entrer quelques questionnaires deux fois en tant que contrôles à l'improviste de la qualité du travail des agents de saisie et pour identifier les erreurs systématiques qui sont commises.

Vérification des données

L'expérience a montré que les erreurs d'étendue sont presque toujours des erreurs de saisie et qu'il peut être très utile de les contrôler et de les corriger immédiatement au moment de la saisie. Toutefois, les erreurs de cohérence doivent être corrigées en retournant auprès du ménage ou en examinant avec attention le questionnaire. Il est donc préférable de mener le contrôle de cohérence lors d'une étape séparée, en imprimant les erreurs, ce qui permettra de les pointer au moment de la correction.

Le programme de saisie de EPI INFO fourni avec ce manuel procède automatiquement à un grand nombre de vérifications des données, soit en alertant l'agent de saisie quand il entre une réponse hors étendue, soit en signalant les incohérences lors du passage final du programme de vérification des données.

☞ Une fois que les données sont au bureau central, il n'est généralement plus possible de retourner auprès des ménages pour corriger les réponses incohérentes. Le coordinateur de l'enquête ou un superviseur expérimenté auront alors à faire la "meilleure hypothèse" possible, sur la base des informations existantes, pour corriger le questionnaire. Si ce n'est pas possible, la variable devra être traitée comme "manquant".

Après avoir saisi les données d'une grappe entière, le lot doit être contrôlé pour cohérence et édité. Le programme procède à des vérifications pour s'assurer que toutes les variables appropriées au cas sont dans le fichier (c'est-à-dire, la structure du fichier est correcte) et qu'elles se situent dans l'étendue correcte. Le programme vérifie la cohérence interne pour s'assurer que les données ont été enregistrées et saisies correctement. Les instructions de saisie qui se trouvent en Annexe 3 expliquent les procédures à suivre quand des erreurs se produisent. A ce niveau, pour trouver la solution aux problèmes rencontrés dans les données, il est souvent utile de se référer au questionnaire.

✎ **EXEMPLE:**

Les contrôles de cohérence vérifient que les conditions suivantes sont respectées:

— La date de naissance de chaque enfant doit précéder la date de l'interview.

— Les dates de naissances de tous les enfants doivent se situer dans les 15 dernières années.

— Si l'origine de l'eau est "robinet dans le logement," alors la distance doit être "sur place".

— Si un enfant allaité reçoit "SEULEMENT du lait maternel", alors aucun poste ne doit avoir une réponse affirmative

— La date du premier DTCoq doit précéder celle du second DTCoq, qui doit précéder celle du troisième.

— Les indices anthropométriques ne sont pas positionnés comme invalides du fait d'erreurs de saisie de l'âge ou du poids.

Il peut être utile d'examiner le nombre d'erreurs commises par chaque équipe ou enquêteur, de façon à donner un complément de formation au début des activités de terrain.

Une fois que les corrections concernant les incohérences ont été notées, les enregistrements correspondants sont alors mis à jour directement. Le programme de contrôle doit alors être exécuté à nouveau et les différentes procédures doivent être répétées jusqu'à ce que toutes les erreurs qui peuvent être corrigées au bureau le soient. Après avoir procédé à ces vérifications, les données sont "apurées" et les fichiers sont prêts pour l'analyse.

EVALUATION DE LA QUALITE DES DONNEES DE L'ENQUETE

L'étape suivante consiste à produire des *distributions de fréquences* pour toutes les variables de votre fichier de données et pour toutes les valeurs qu'elles peuvent prendre (voir Tableau 7.1). Cela vous donne l'occasion de vérifier, encore une fois, les enregistrements inappropriés et de noter le nombre de valeurs manquantes dans les données.

Tableau 7.1. Exemple de distribution des fréquences

Numéro de rubrique	Source d'approvisionnement en eau du ménage	Distribution
1	Robinet dans le logement	146
2	Robinet public	440
3	Borne fontaine	102
4	Puits protégé	95
5	Puits non protégé	660
6	Etang, rivière ou ruisseau	35
7	Camion citerne , vendeur	55
9	Autre	15
.	Manquant	6
Total		**1 554**

Il se peut que vous ayez à retourner aux questionnaires pour vérifier les données saisies. Vous aurez alors à corriger le fichier de données brutes et à recommencer cette étape.

☞ Selon l'exemple du Tableau 7.1 : Supposons que vous vouliez examiner les 15 questionnaires qui comportent des réponses "autres" pour la source d'approvisionnement en eau. Vous aurez besoin d'un programme pour sélectionner seulement les enregistrements qui ont la valeur "9" (autre) pour la variable "source de l'eau". Le programme doit vous permettre de lister le numéro d'identification des ménages avec ces valeurs. Vous pouvez alors utiliser ces numéros d'identification pour retourner aux questionnaires et les examiner.

Vous devez conserver un enregistrement du nombre de cas incomplets dans le fichier, du nombre de ménages qui étaient listés mais qui n'ont pas été contactés, et du type d'erreurs qui ont été détectées dans les données. Le pourcentage de ménages éligibles qui n'ont pas été contactés doit être examiné avec attention et reporté. Un taux de réponse inférieur à 90-95 pour cent signifie que l'enquête ne représente pas nécessairement la population dans son ensemble.

On peut vérifier la qualité des données en examinant la cohérence interne des données collectées. Les contrôles de cohérence qui sont exécutés au moment de la saisie des données peuvent être exécutés à nouveau sur l'ensemble des données pour déterminer le nombre d'enregistrements pour lesquels restent des incohérences. (Voir l'exemple ci-dessus sur les conditions d'incohérences).

Un grand nombre d'incohérences non résolues est un indicateur d'une mauvaise collecte des données (et parfois d'une falsification des données) et pourrait mettre en cause la validité des résultats.

Un second indicateur de la qualité des données est le pourcentage de réponses "Ne sait pas" ou "Autre". On peut connaître ces pourcentages en examinant la distribution des fréquences pour une variable dans le fichier de données. Un pourcentage élevé de "Ne sait pas" signifie que les enquêtés n'ont pas compris les questions ou que les informations recherchées sont trop difficiles à obtenir. Les résultats basés sur de telles variables sont douteux. Un pourcentage élevé de réponses "Autre" indique que les catégories utilisées dans le questionnaire ne correspondent pas aux réponses données le plus couramment.

Un autre indicateur de la qualité des données est leur cohérence avec d'autres données existantes. Par exemple, la répartition par âge des mères et des enfants et le ratio des garçons par rapport aux filles devraient être similaires à la répartition par âge et au rapport de masculinité trouvés au cours d'un recensement récent ou au cours d'une autre enquête récente portant sur la même population. Même en l'absence d'un recensement ou d'une enquête récente, le nombre de garçons et de filles dénombrés devraient être approximativement le même. Le nombre d'enfants de chaque année d'âge devrait également être le même, avec une légère diminution avec l'augmentation en âge.

CALCUL DES INDICATEURS

Les tableaux élémentaires nécessaires au calcul de chaque indicateur sont listés en Annexe 4 et les programmes pour produire les estimations sont présentés en Annexe 3. Toutes les estimations peuvent être calculées avec le programme qui se trouve sur la disquette fournie avec ce manuel.

Autrement, il est possible d'utiliser un logiciel standard, comme SPSS PC.

Vous pouvez obtenir certains des indicateurs —par exemple, le pourcentage de ménages ayant du sel iodé—directement à partir des réponses à une variable du questionnaire. Cependant, pour la plupart des indicateurs, vous devrez examiner les réponses à plusieurs questions pour déterminer l'information nécessaire pour l'indicateur. Par exemple, l'accès à l'eau salubre est basé, à la fois, sur la source d'approvisionnement de l'eau et sur la distance par rapport à la source. Dans ces cas, une variable doit être créée pour indiquer si le ménage, la mère *ou* l'enfant s'accordent avec la définition de l'indicateur.

☞ Le programme EPI INFO fourni avec ce manuel devra être adapté pour fournir les indicateurs qui varient de pays à pays—par exemple, l'eau et les sanitaires.

La plupart des logiciels, y compris SPSS PC et EPI INFO, produisent une table d'informations sur la variable créée pour l'indicateur. Pour rendre compte de l'indicateur, le résultat doit être tiré de cette table.

✎ **EXEMPLE:**

Une table concernant un faible poids-pour-âge peut être générée comme suit :

Nom de la valeur	Valeur	Fréquence	Pourcentage	Pourcentage valide	Pourcentage cumulé
≥ -2 ET	0	1 183	77,9	79,0	79,0
<-2 ET	1	314	20,7	20,9	100,0
	.	21	1,4	manquant	
Total =		**1 518**	**100,0**	**100,0**	
Cas valides : 1 497					*Cas manquants : 21*

ET = écart type

Dans ce cas, pour l'Indicateur 11.1, le nombre à publier est 20,9 pour cent, à savoir le pourcentage d'enfants dont le poids-pour-âge était faible, parmi les enfants ayant un poids-pour-âge valide dans le fichier de données. Dans le logiciel EPI INFO inclus dans ce manuel, les indicateurs sont imprimé directement en utilisant la fonction REPORT, de telle sorte que seule l'information nécessaire est montrée.

Les indices anthropométriques doivent être calculés pour chaque enfant à partir des mesures du poids et de la taille. EPI INFO utilise un sous-programme separé pour comparer l'âge et le poids (ou la taille) d'un enfant à la référence internationale, puis attribue un score à chaque enfant sur la base de son écart par rapport à la valeur médiane de référence (un score d'écart type). Le calcul peut être fait au moment de la saisie des données, comme c'est le cas avec le logiciel EPI INFO fourni ici,

ou peut être fait au cours d'un traitement ultérieur des données en utilisant un logiciel spécial d'anthropométrie comme ANTHRO (précédemment CASP) ou EPINUT. L'avantage d'EPI INFO réside dans le fait que les indices peuvent être vus à l'écran aussitôt que les données sur la taille et le poids sont entrées, de façon telle que le personnel de saisie peut immédiatement détecter les indices qui sont manifestement trop élevés ou trop faibles (généralement à cause d'erreurs de frappe au moment de la saisie des données). Les indices inférieurs à -6,0 scores d'écart type ou supérieurs à 6,0 sont généralement considérés comme aberrants.

✎ **EXEMPLE:**
Les résultats générés par le programme REPORT fourni avec ce manuel se présentent comme suit :

Proportion d'enfants avec un Score d'écart type faible pour le poids-pour-âge
(Indicateur 11.1,2)

Nombre total d'enfants	S d'ET pour PPA < -2	S. d'ET pour PPA < -3
1 497	21%	4%

☞ Dans certains cas, le logiciel d'analyse ne produira pas l'indicateur lui-même. Par exemple, une enquête ne donne pas les nombres absolus de la population. Pour convertir le pourcentage de la population ayant un accès facile à de l'eau salubre (valeur produite par le programme REPORT) en nombres absolus (Indicateur 13.1), il est nécessaire de multiplier ce pourcentage par la taille de la population totale.

Estimation de la marge d'erreur pour chaque indicateur

La plupart des logiciels d'analyse estiment l'erreur type des pourcentages et des moyennes, en supposant que l'échantillon est un "simple échantillon aléatoire"[1]. Cependant, ces erreurs standard sont généralement sous-estimées parce qu'elles ne prennent pas en compte le fait que les données proviennent d'enquêtes par sondage en grappes. Des logiciels d'analyse spécialement conçus pour travailler avec des enquêtes par sondage en grappes sont nécessaires pour

Le programme CSAMPLE d'EPI INFO produit automatiquement l'estimation avec sa marge d'erreur. Voir Annexe 3 pour les instructions d'utilisation de ce programme.

[1]La formule de calcul de l'*erreur standard* (e.s) pour une proportion, en supposant un échantillon aléatoire simple, est la suivante : $e.s. = \sqrt{[p \times (1 - p) \div n]}$

où p est la proportion étudiée et n est la taille de l'échantillon du groupe. La *marge d'erreur* pour cette proportion égale (2 × e.s.). L'*intervalle de confiance* pour la proportion est calculée comme $p \pm$ la marge d'erreur.

estimer correctement les erreurs standard. EPI INFO (Version 6) a un module, appelé CSAMPLE, pour calculer des erreurs standard correctes pour les pourcentages et les moyennes, bien que ce module ne soit pas inclus dans le module ANALYSIS. L'Annexe 3 contient les instructions pour calculer les erreurs standard et les intervalles de confiance pour chaque indicateur en utilisant ce module.

✎ **EXEMPLE:**
La marge d'erreur pour l'indicateur et son intervalle de confiance à 95 pour cent sont présentés. Par exemple, pour l'analyse du poids-pour-âge, le résultat est :

Le pourcentage avec un faible poids-pour-âge est de 21,0% avec un intervalle de confiance de (19,3; 22,7).

Calcul des facteurs de pondération pour les échantillons non auto-pondérés

Si des bases de sondage séparées ont été utilisées pour différentes régions au premier degré du sondage, l'échantillon national n'a pas été choisi avec probabilité proportionnelle à la taille (PPT). Ceci peut également se produire si vous avez procédé à une stratification selon certains autres facteurs (par exemple, urbain/rural ou quartiers pauvres/autres quartiers) et utilisé des taux de sondage (proportions) différents dans chaque strate. Ces échantillons ne sont pas "auto-pondérés" , et vous devez pondérer votre échantillon quand vous calculez les estimations nationales. Cela signifie que vous devez vous assurer que la part de chaque sous-échantillon séparé—par exemple, chaque région—contribue à l'échantillon national seulement pour la part qu'il aurait eu si l'échantillon de l'enquête, au niveau national, avait été tiré avec PPT. La procédure pour calculer les pondérations est illustrée au Tableau 7.2 en utilisant un exemple théorique.

☞ Créez une nouvelle variable pour représenter le facteur de pondération. Ce facteur doit être utilisé pour calculer les estimations *nationales* à partir des différentes enquêtes sous-nationales. Chaque cas sera alors pondéré par cette variable.

Par exemple, supposons que le pays a une population de 1 333 415 ménages et comprend sept régions, chacune d'entre elles ayant été enquêtée séparément. L'échantillon total des ménages était de 11 312.

Pour calculer le facteur de pondération approprié pour chaque échantillon enquêté, créez un tableau comme le Tableau 7.2.

☞ Pour calculer la couverture réelle pour l'échantillon entier, créez une variable représentant le facteur de pondération :

```
IF (REGION = 1) THEN WTFACTOR = 2.73
IF (REGION = 2) THEN WTFACTOR = 0.48
IF (REGION = 3) THEN WTFACTOR = 0.99
IF (REGION = 4) THEN WTFACTOR = 0.26
```

IF (REGION = 5) THEN WTFACTOR = 1.77
IF (REGION = 6) THEN WTFACTOR = 0.41
IF (REGION = 7) THEN WTFACTOR = 0.23

Les données analysées doivent alors être pondérées par la variable WTFACTOR. Les résultats vous donneront des indicateurs au niveau national correctement pondérés pour tenir compte du plan de sondage non auto-pondéré.

☞ Ces calculs peuvent aussi être faits manuellement. Multipliez chaque estimation régionale par son facteur de pondération, faites la somme des résultats des produits et divisez la somme par la somme des poids.

✎ **EXEMPLE:**

(1)	(2)	(3)	(4)
Région	Estimation	Facteur de pondération	Estimation x FACTEUR PND.
1	0,85	1,7	1,445
2	0,75	0,2	0,15
3	0,70	1,0	0,70
Total		Σ = 2,9	2,295

Estimation nationale moyenne pondérée : (somme de la colonne 4 ÷ 2,9) = 0,79

Estimations par âge et sexe

☞ Pour calculer chaque indicateur selon le sexe de l'enfant, vous devez sélectionner dans le fichier de données les enfants d'un sexe particulier et mener une analyse à part pour ce groupe. Faites de même pour les enfants de l'autre sexe.

Suivez la même procédure si vous voulez calculer des indicateurs pour de petits groupes d'âges ou pour des groupes sous-nationaux comme les régions, les districts ou la résidence urbain/rural. Sélectionnez chaque groupe séparément et exécutez le programme d'analyse pour obtenir des résultats pour chaque groupe. Répétez la procédure pour chaque groupe pour lequel vous voulez calculer des indicateurs séparés. Si vous n'utilisez pas le programme d'analyse de EPI INFO, il se peut que vous trouviez plus simple de procéder à des tabulations de la variable croisée avec le sexe de l'enfant, ou avec d'autres variables similaires qui définissent des sous-groupes de population (par exemple, au niveau régional, urbain/rural).

☞ Assurez-vous que les titres sont imprimés clairement (par exemple, garçons seulement) pour éviter toute confusion ultérieure.

Maintenant, vous êtes prêt pour rédiger votre rapport préliminaire et pour représenter ces indicateurs graphiquement.

Tableau 7.2. Tableau illustratif pour le calcul des facteurs de pondération

Strate	(1) Taille de la Population (ménages)	(2) Population de la strate en tant que proportion de la population nationale	(3) Taille finale de l'échantillon	(4) Echantillon de la strate en tant que proportion de l'échantillon total	(5) Facteur de pondération
		(col. 1 ÷ 1 333 415)	*n (ménages)*	(col. 3 ÷ 11 312)	(col. 2 ÷ col. 4)
1	550 088	0,412	1 707	0,1509	2,730
2	93 088	0,070	1 664	0,1471	0,476
3	192 580	0,144	1 648	0,1457	0,988
4	47 873	0,036	1 560	0,1379	0,261
5	330 312	0,248	1 587	0,1403	1,768
6	77 124	0,058	1 592	0,1407	0,412
7	42 350	0,032	1 544	0,1374	0,233
Total	1 333 415	1,000	11 312	1,000	