OPHI
OXFORD POVERTY & HUMAN DEVELOPMENT INITIATIVE
www.ophi.org.uk

UNIVERSITY OF
OXFORD

AF Measure Analysis Issues I

José Manuel Roche

Washington, 11 July 2013

Analysis Issues I

1. Metadata
2. Survey design and representativeness
3. Non response rate and other non sampling error
4. Missing values, inconsistencies, “don’t know”
5. Eligible population
6. Sample drop and bias analysis

Metadata

Metadata is “data about the data”. The metadata of a household surveys provides us information about the survey sample design, fieldwork activities, questionnaires, structure of the dataset, definitions and coding, etc.

How to use the sample weights?

Who are eligible?

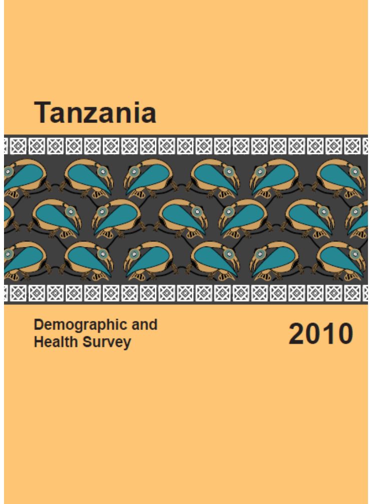
How to treat missing values?

How to interpret the coding?

DHS Country Report :

It includes comprehensive survey results and country survey specificities (sample representativeness, nonresponse rate, fieldwork obstacles)

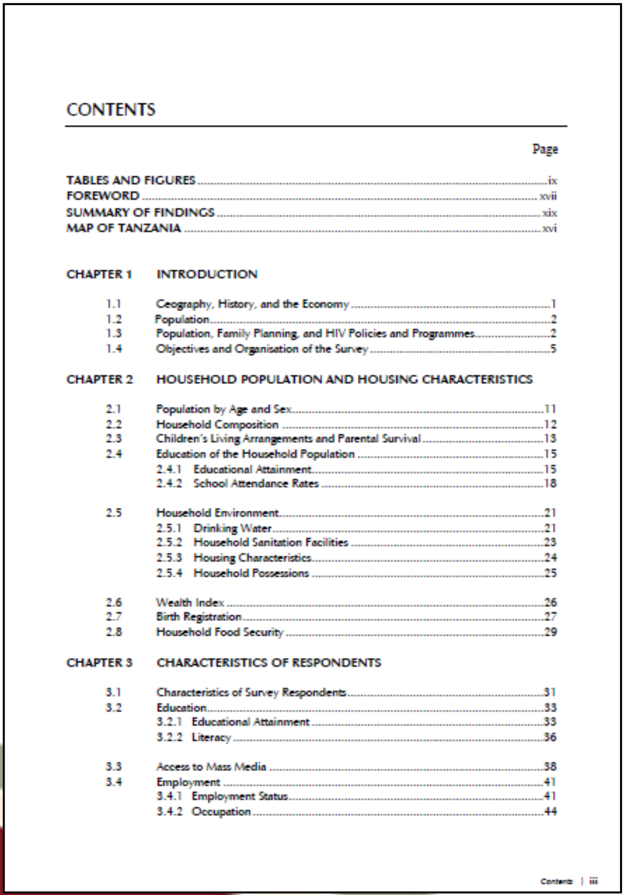
http://www.measuredhs.com/publications/publications-by-type.cfm



Tanzania

Demographic and Health Survey

2010



CONTENTS

Page

Contents | iii

Sampling design and representativeness

Tables and results

viii | Contents

Reports are useful but not enought!

Large Survey projects provide plenty of metadata

For example, DHS General Data Manuals: available online

http://www.measuredhs.com/data/Data-Tools-and-Manuals.cfm

Guide to DHS Statistics

The Guide to DHS Statistics is a reference to help users who work with DHS survey indicators and datasets to better understand indicator definitions and the calculations used to generate the survey results. The Online Guide to DHS Statistics is also available.

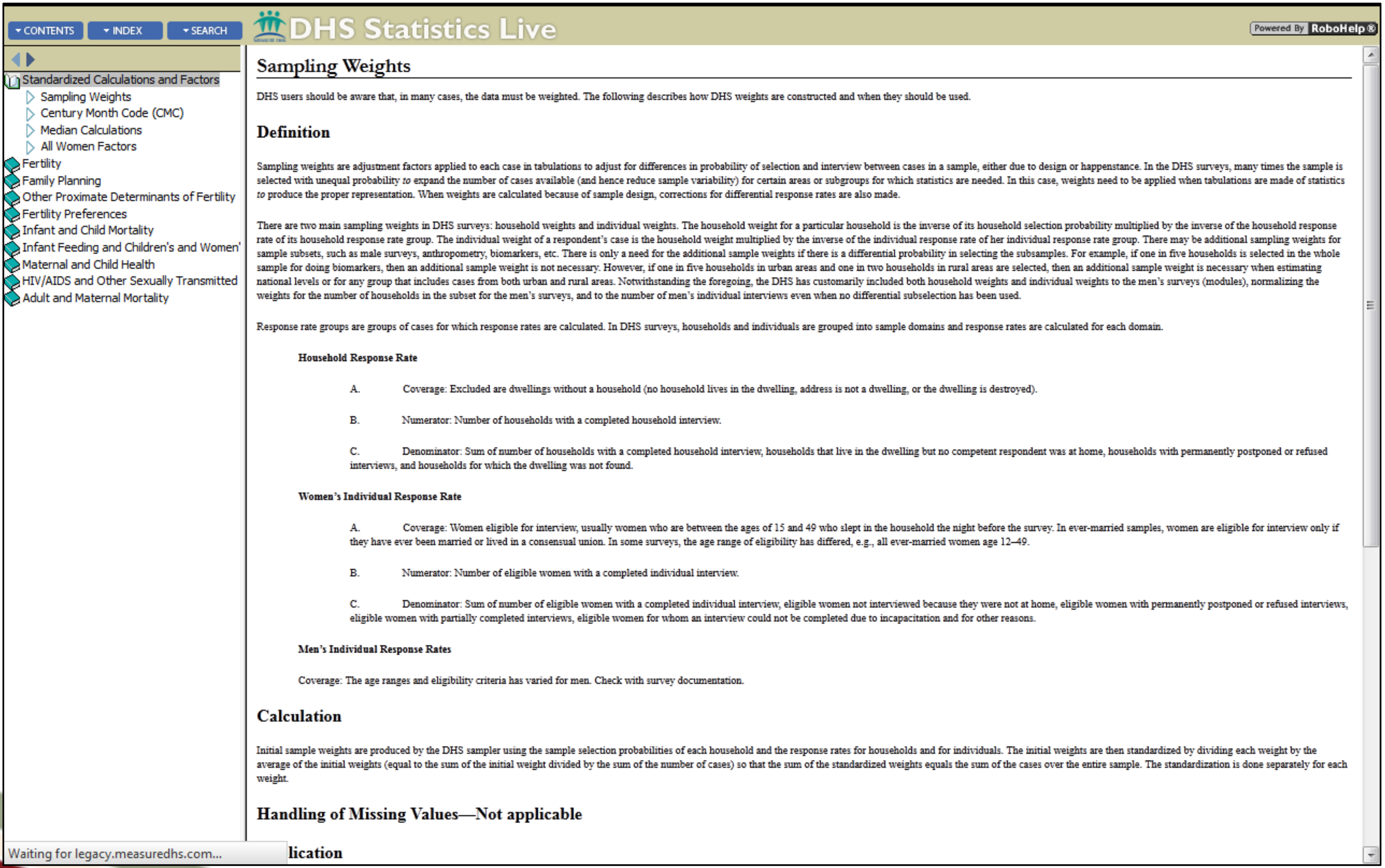
DHS Recode Manual

The Recode Manual provides the information necessary to understand these datasets. It describes each data file and the variables contained in them. Dataset users are strongly encouraged to download the DHS recode manual for use with all recode files.

DHS Data Editing and Imputation

This paper presents the methodology used by DHS for the production of edited data files. The paper focuses primarily on the editing of dates of events, and the imputation of incomplete dates. The paper discusses various approaches to the problems of partial and inconsistent data, and the need for procedures to handle these data.

Online Guide to DHS Statistics
– quite handy to search for detailed info!


CONTENTS
INDEX
SEARCH
DHS Statistics Live
Powered By RoboHelp®
Standardized Calculations and Factors
Sampling Weights
Century Month Code (CMC)
Median Calculations
All Women Factors
Fertility
Family Planning
Other Proximate Determinants of Fertility
Fertility Preferences
Infant and Child Mortality
Infant Feeding and Children's and Women'
Maternal and Child Health
HIV/AIDS and Other Sexually Transmitted
Adult and Maternal Mortality
Waiting for legacy.measuredhs.com...

Sampling Weights

DHS users should be aware that, in many cases, the data must be weighted. The following describes how DHS weights are constructed and when they should be used.

Definition

Sampling weights are adjustment factors applied to each case in tabulations to adjust for differences in probability of selection and interview between cases in a sample, either due to design or happenstance. In the DHS surveys, many times the sample is selected with unequal probability *to* expand the number of cases available (and hence reduce sample variability) for certain areas or subgroups for which statistics are needed. In this case, weights need to be applied when tabulations are made of statistics *to* produce the proper representation. When weights are calculated because of sample design, corrections for differential response rates are also made.

There are two main sampling weights in DHS surveys: household weights and individual weights. The household weight for a particular household is the inverse of its household selection probability multiplied by the inverse of the household response rate of its household response rate group. The individual weight of a respondent’s case is the household weight multiplied by the inverse of the individual response rate of her individual response rate group. There may be additional sampling weights for sample subsets, such as male surveys, anthropometry, biomarkers, etc. There is only a need for the additional sample weights if there is a differential probability in selecting the subsamples. For example, if one in five households is selected in the whole sample for doing biomarkers, then an additional sample weight is not necessary. However, if one in five households in urban areas and one in two households in rural areas are selected, then an additional sample weight is necessary when estimating national levels or for any group that includes cases from both urban and rural areas. Notwithstanding the foregoing, the DHS has customarily included both household weights and individual weights to the men’s surveys (modules), normalizing the weights for the number of households in the subset for the men’s surveys, and to the number of men’s individual interviews even when no differential subselection has been used.

Response rate groups are groups of cases for which response rates are calculated. In DHS surveys, households and individuals are grouped into sample domains and response rates are calculated for each domain.

Household Response Rate

A. Coverage: Excluded are dwellings without a household (no household lives in the dwelling, address is not a dwelling, or the dwelling is destroyed).

B. Numerator: Number of households with a completed household interview.

C. Denominator: Sum of number of households with a completed household interview, households that live in the dwelling but no competent respondent was at home, households with permanently postponed or refused interviews, and households for which the dwelling was not found.

Women’s Individual Response Rate

A. Coverage: Women eligible for interview, usually women who are between the ages of 15 and 49 who slept in the household the night before the survey. In ever-married samples, women are eligible for interview only if they have ever been married or lived in a consensual union. In some surveys, the age range of eligibility has differed, e.g., all ever-married women age 12–49.

B. Numerator: Number of eligible women with a completed individual interview.

C. Denominator: Sum of number of eligible women with a completed individual interview, eligible women not interviewed because they were not at home, eligible women with permanently postponed or refused interviews, eligible women with partially completed interviews, eligible women for whom an interview could not be completed due to incapacitation and for other reasons.

Men’s Individual Response Rates

Coverage: The age ranges and eligibility criteria has varied for men. Check with survey documentation.

Calculation

Initial sample weights are produced by the DHS sampler using the sample selection probabilities of each household and the response rates for households and for individuals. The initial weights are then standardized by dividing each weight by the average of the initial weights (equal to the sum of the initial weight divided by the sum of the number of cases) so that the sum of the standardized weights equals the sum of the cases over the entire sample. The standardization is done separately for each weight.

Handling of Missing Values—Not applicable

lication

MICS Metadata: available online

http://www.childinfo.org

Country reports

It includes comprehensive survey results and country survey specificities (sample representativeness, nonresponse rate, fieldwork obstacles)

Questionnaires

Including: Flow of questionnaire modules, Household questionnaire, Women's questionnaire, Children under-5 questionnaire, Additional situation specific modules, Optional modules

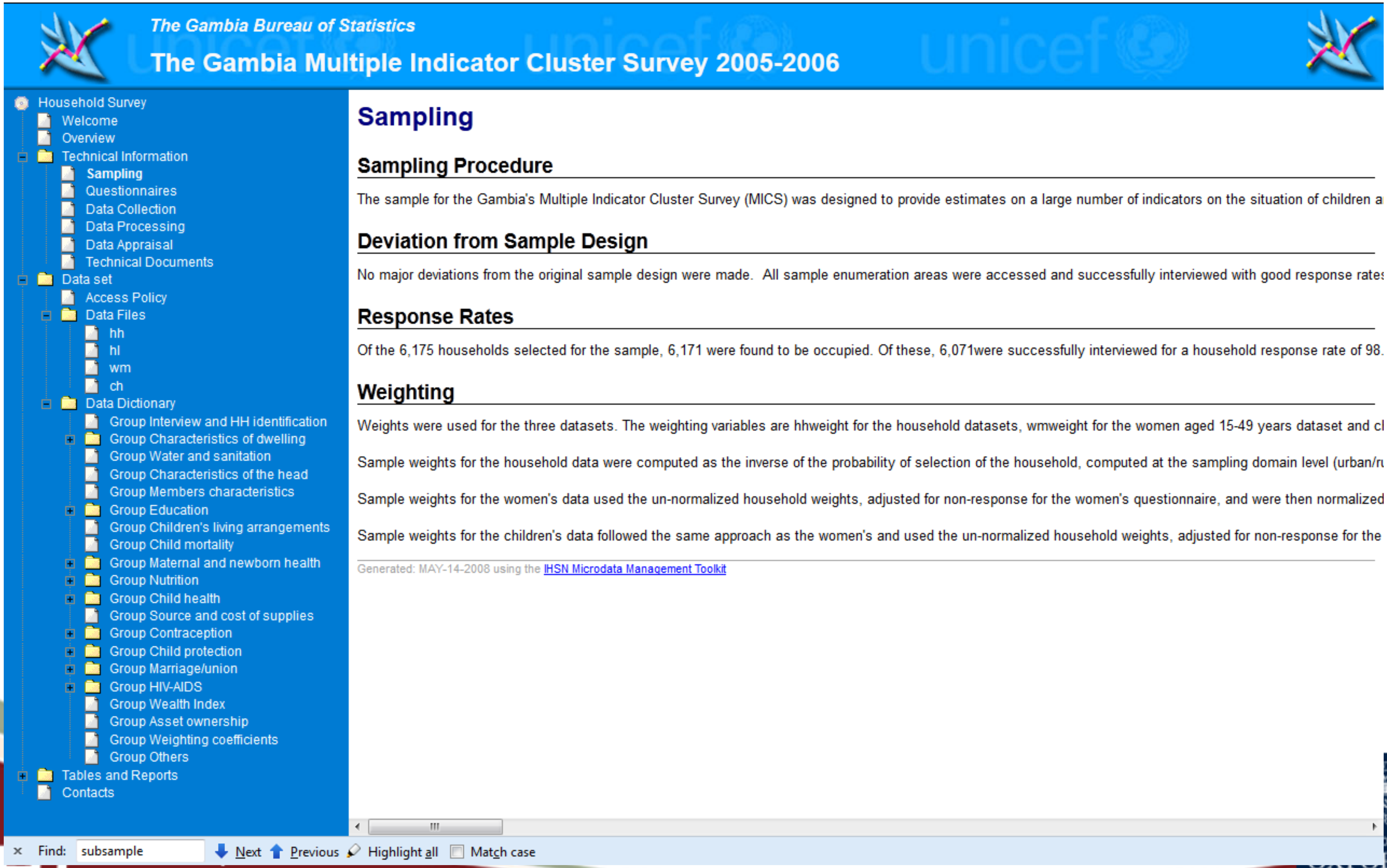
MICS Manual

Including: Flow of questionnaire modules, Household questionnaire, Women's questionnaire, Children under-5 questionnaire, Additional situation specific modules, Optional modules

Other various documents

Including: Standard memorandum of understanding, Sample Size (Households) Calculation Template, Pictorials for Water and Sanitation Facilities, One-page pictorial on cooking methods using solid fuels. Illustrations of cooking methods commonly used, Wealth Index, Sample weight calculation.

Online MICS Survey archive: It provides detailed information describing the content, methodology and implementation of the survey
Example: http://www.childinfo.org/mics/mics3/archives/gambia/survey0/index.html



The Gambia Bureau of Statistics

The Gambia Multiple Indicator Cluster Survey 2005-2006

Sampling

Sampling Procedure

The sample for the Gambia's Multiple Indicator Cluster Survey (MICS) was designed to provide estimates on a large number of indicators on the situation of children a

Deviation from Sample Design

No major deviations from the original sample design were made. All sample enumeration areas were accessed and successfully interviewed with good response rate

Response Rates

Of the 6,175 households selected for the sample, 6,171 were found to be occupied. Of these, 6,071were successfully interviewed for a household response rate of 98.

Weighting

Weights were used for the three datasets. The weighting variables are hhweight for the household datasets, wmweight for the women aged 15-49 years dataset and cl

Sample weights for the household data were computed as the inverse of the probability of selection of the household, computed at the sampling domain level (urban/ru

Sample weights for the women's data used the un-normalized household weights, adjusted for non-response for the women's questionnaire, and were then normalized

Sample weights for the children's data followed the same approach as the women's and used the un-normalized household weights, adjusted for non-response for the

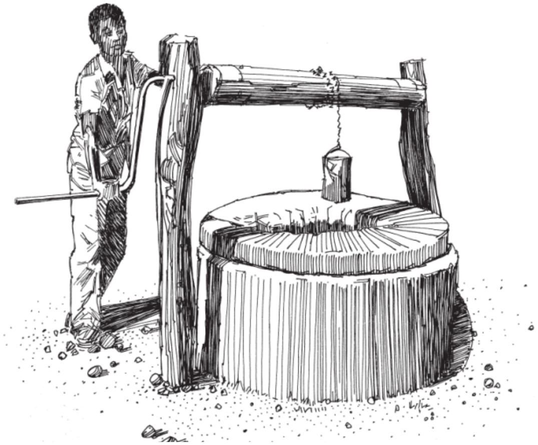
Generated: MAY-14-2008 using the IHSN Microdata Management Toolkit

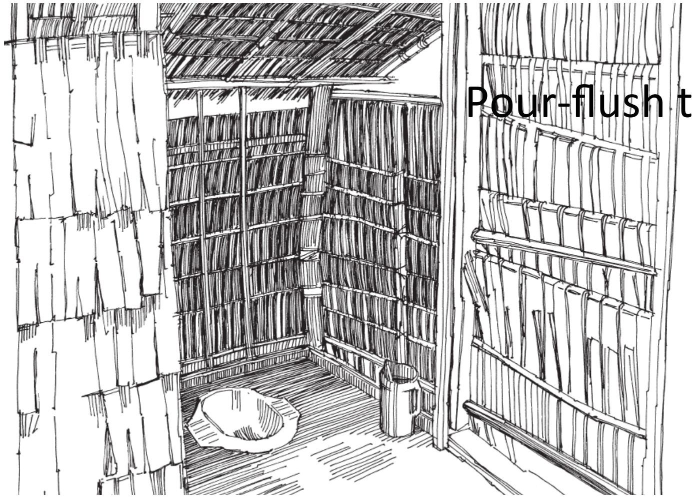
Find: subsample

Always search also among materials used during fieldwork

For example: Pictorial illustrations on access to water supply and sanitation facilities for use in national household surveys by JPM and UNICEF

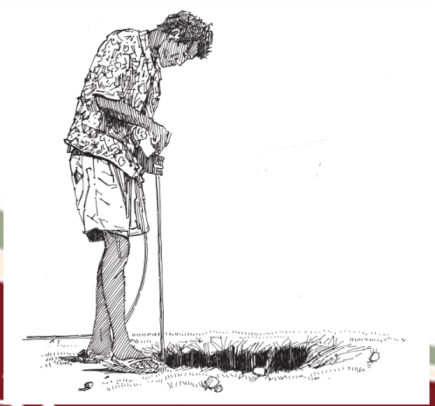
Protected dug well?

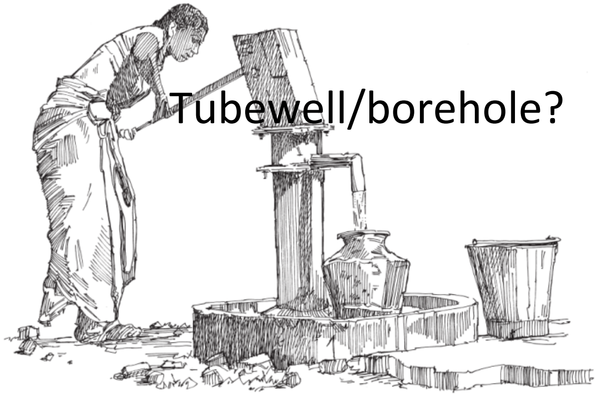




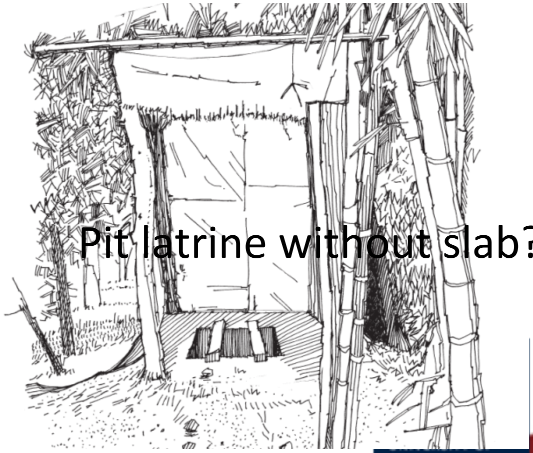
Pour-flush to pit?

Unprotected dug well?





Tubewell/borehole?



Pit latrine without slab?

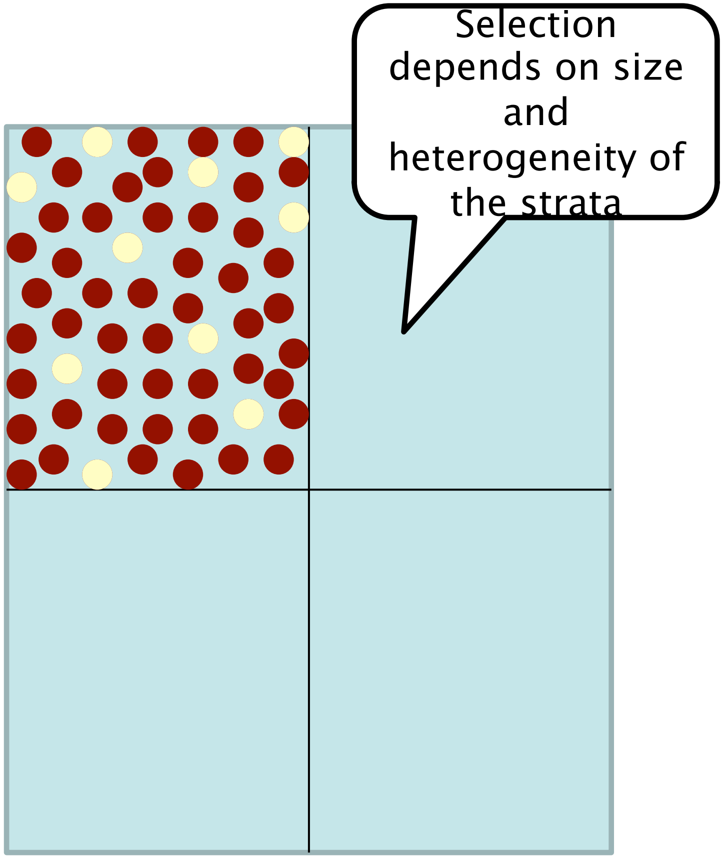
2. Survey Design and Representativeness

Usually household surveys follow a complex sampling design in two stages:

1. Clusters (e.g. PSU) are selected in a first stage from within each strata (e.g. Region+urban/rural)

2. Households are systematically selected from household listings within each cluster (listing are generated during census operations or at a later updates)

The results is a representative and yet efficient sample that reduces cost and increases quality of data collection



Selection depends on size and heterogeneity of the strata

What about probability of selection?

What are the sample weights?

The weights are computed as the inverse probability of selection: probability of selecting the cluster and probability of selecting the household within the cluster. They may also adjust by response rate and/or by the demographic structure of the population (see: Yansanhe 2005).

- When there are not different probabilities weights may not be calculated
- Use of sample weight is a statistical requirement rather than an econometric issue – it is advisable to use even in most regression analysis! (Deaton 1997)
- Ignoring the weights would produce significantly bias results

Samples and subsamples

Some data can be particularly more difficult and expensive to collect, either because it takes longer (e.g. revisits) or it requires enumerators with more expertise (hence supervision is more difficult).

Check the metadata for subsamples and how to undertake analysis with it!

“Changes over Time (regarding Children and Women's Nutritional Status)

In phases of the DHS survey before phase IV (DHS+), only children of interviewed women and who were under five years old (or the cutoff for the health section of the individual questionnaire) were weighed and measured. In many surveys, only a subsample of these children were selected for anthropometry. All comparisons between surveys, either over time or between countries, should take into account the possible differences in the defined population base”

What geographical level can you decompose?
Are all ethnic group represented well in the sample?

Survey representativeness depends on the sample design, and will limit how far one can undertake decomposition analysis.

For example in Tanzania:

"The 2010 TDHS sample was designed to provide estimates for the entire country, for urban and rural areas in the Mainland, and for Zanzibar.

For specific indicators such as contraceptive use the sample design allowed the estimation of indicators for each of the then 26 regions. To estimate geographic differentials for certain demographic indicators, the regions of mainland Tanzania were collapsed into seven geographic zones. Although these are not official administrative zones, this classification is used by the Reproductive and Child Health Section of the MoHSW. Zones were used in each geographic area in order to have a relatively large number of cases and a reduced sampling error."

3. Non response rate and other non sampling error

It is also a good practice to report the nonresponse rate and any other non sampling error (e.g. problems during fieldwork logistics)

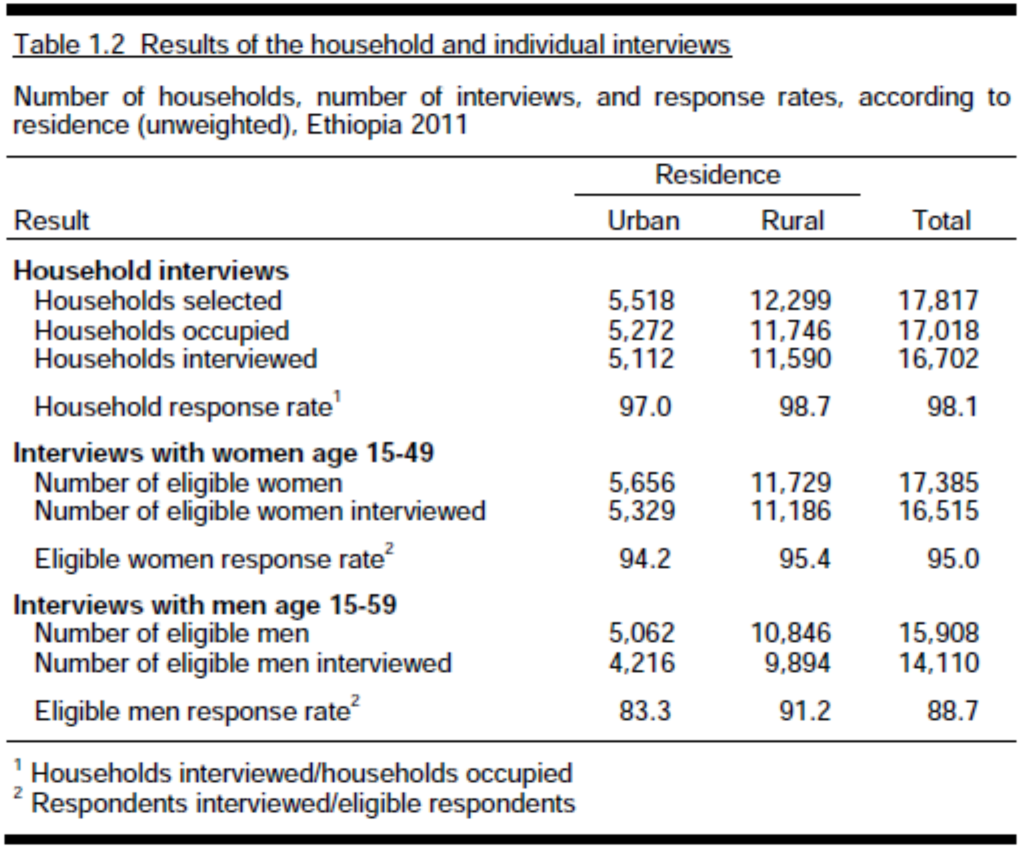


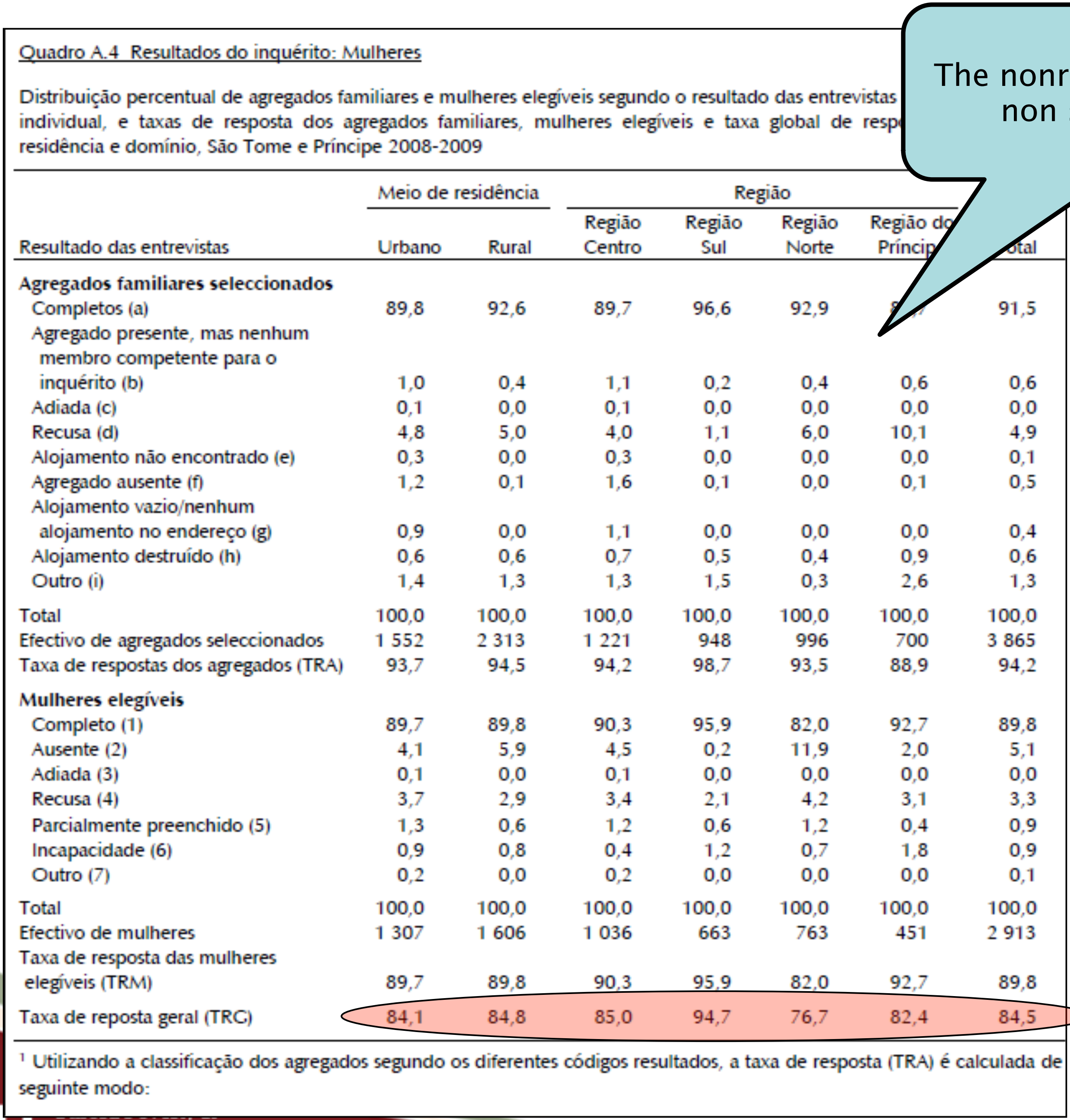
Table 1.2 Results of the household and individual interviews

Number of households, number of interviews, and response rates, according to residence (unweighted), Ethiopia 2011

	Residence		
Result	Urban	Rural	Total
Household interviews			
Households selected	5,518	12,299	17,817
Households occupied	5,272	11,746	17,018
Households interviewed	5,112	11,590	16,702
Household response rate[1]	97.0	98.7	98.1
Interviews with women age 15-49			
Number of eligible women	5,656	11,729	17,385
Number of eligible women interviewed	5,329	11,186	16,515
Eligible women response rate[2]	94.2	95.4	95.0
Interviews with men age 15-59			
Number of eligible men	5,062	10,846	15,908
Number of eligible men interviewed	4,216	9,894	14,110
Eligible men response rate[2]	83.3	91.2	88.7

[1] Households interviewed/households occupied
[2] Respondents interviewed/eligible respondents

The nonresponse introduces non statistical error



Quadro A.4 Resultados do inquérito: Mulheres

Distribuição percentual de agregados familiares e mulheres elegíveis segundo o resultado das entrevistas individual, e taxas de resposta dos agregados familiares, mulheres elegíveis e taxa global de resp residência e domínio, São Tome e Príncipe 2008-2009

Resultado das entrevistas	Meio de residência Urbano	 Rural	Região Região Centro	 Região Sul	 Região Norte	 Região do Princip	 otal
Agregados familiares seleccionados							
Completos (a)	89,8	92,6	89,7	96,6	92,9	[illegible]	91,5
Agregado presente, mas nenhum membro competente para o inquérito (b)	1,0	0,4	1,1	0,2	0,4	0,6	0,6
Adiada (c)	0,1	0,0	0,1	0,0	0,0	0,0	0,0
Recusa (d)	4,8	5,0	4,0	1,1	6,0	10,1	4,9
Alojamento não encontrado (e)	0,3	0,0	0,3	0,0	0,0	0,0	0,1
Agregado ausente (f)	1,2	0,1	1,6	0,1	0,0	0,1	0,5
Alojamento vazio/nenhum alojamento no endereço (g)	0,9	0,0	1,1	0,0	0,0	0,0	0,4
Alojamento destruído (h)	0,6	0,6	0,7	0,5	0,4	0,9	0,6
Outro (i)	1,4	1,3	1,3	1,5	0,3	2,6	1,3
Total	100,0	100,0	100,0	100,0	100,0	100,0	100,0
Efectivo de agregados seleccionados	1 552	2 313	1 221	948	996	700	3 865
Taxa de respostas dos agregados (TRA)	93,7	94,5	94,2	98,7	93,5	88,9	94,2
Mulheres elegíveis							
Completo (1)	89,7	89,8	90,3	95,9	82,0	92,7	89,8
Ausente (2)	4,1	5,9	4,5	0,2	11,9	2,0	5,1
Adiada (3)	0,1	0,0	0,1	0,0	0,0	0,0	0,0
Recusa (4)	3,7	2,9	3,4	2,1	4,2	3,1	3,3
Parcialmente preenchido (5)	1,3	0,6	1,2	0,6	1,2	0,4	0,9
Incapacidade (6)	0,9	0,8	0,4	1,2	0,7	1,8	0,9
Outro (7)	0,2	0,0	0,2	0,0	0,0	0,0	0,1
Total	100,0	100,0	100,0	100,0	100,0	100,0	100,0
Efectivo de mulheres	1 307	1 606	1 036	663	763	451	2 913
Taxa de resposta das mulheres elegíveis (TRM)	89,7	89,8	90,3	95,9	82,0	92,7	89,8
Taxa de reposta geral (TRG)	84,1	84,8	85,0	94,7	76,7	82,4	84,5

[1] Utilizando a classificação dos agregados segundo os diferentes códigos resultados, a taxa de resposta (TRA) é calculada de seguinte modo:

4. Missing values, inconsistencies, “don’t know”

Missing value: a variable that should have a response, but because of interview errors the question was not asked.

Inconsistent: This code is generally used by people in the secondary editing group, when a value or code is not plausible.

“Don’t know” responses: These codes are normally pre-coded in the questionnaires, but they are consistently used throughout the recode file.

How should we treat missing values in a composed indicator? (e.g. we know the source of water but do not know the distance)

What to do if we only have a few missing cases when constructing a aggregate household indicator? (e.g. Years of schooling or school attendance)

5. Eligible population

example from DHS

Eligible Women for interview

Women eligible for interview, usually women who are between the ages of 15 and 49 who slept in the household the night before the survey. In ever-married samples, women are eligible for interview only if they have ever been married or lived in a consensual union. In some surveys, the age range of eligibility has differed, e.g., all ever-married women age 12–49.

Eligible Male for interview

In some cases the Man's Questionnaire is administered to all men age 15-59 in each household but in some other cases only a subsample of man are interviewed (e.g. 50%)

Who is eligible for anthropometrics?

Usually, DHS measures the height and weight of all children under age 5 and of women and men age 15-49. Occasionally only a subsample is measured (this is now more frequent)

How should we handle the non eligible?
The may be a degree of under estimation if not all members are eligible

How should we treat a missing value when computing the Adjusted Headcount Ratio?

Suppose the following matrix, and a poverty line of k>=1:

$$\begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ . & 0 & 1 \\ 0 & 0 & 0 \\ . & . & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

How do we compute the average deprivation with missing information?

Is this individual poor?

In practice we reduce the sample to only cases with information in all indicators, having a consequent "sample drop" due to missing information

6. Sample drop and bias analysis

The sample drop may…

Affects the representativeness of the sample

- Need to check the proportion of missing values for each indicator and analyze the proportion of total sample drop

Affects the population share when regions are decomposed

- Need to check how the share of each region changes before and after sample drop – Is there a bias towards a particular region?

6. Sample drop and bias analysis

In practice this “bias analysis” can be undertaken with a series of hypothesis test for difference of means or proportions.

Often the cause of a large sample loss is only few indicators. So one could assess if the subsample after sample drop has a significant bias in any given indicator.

$$H_0\colon \mu_1 = \mu_2$$

$$H_1\colon \mu_1 \neq \mu_2$$

Where μ_1 represents the estimation for the full sample, and μ_2 represents the estimation for the subsample after sample drop

www.ophi.org.uk