Sample size, sample weights in household surveys
Outline

• Background
• Total quality in surveys
• Sampling Controversy
• Sample size, stratification and clustering effects
An overview of the quality dimensions in household surveys

The integration of sampling design, questionnaires and tools, field work and data management

with the goal of delivering analysts a reliable database on time.

Data loses their value if they don’t represent the reality of the day.

The end product
Background

• Did you know that cooks are excellent at sampling?
  – *When they add salt to a stew or a soup which they have just finished cooking, they mix it in, they take a spoonful, they taste it and decide whether there is enough salt or some more should be added. They do not have to eat all the soup they have prepared. They only taste a sample.*
  – Researchers and analysts are confronted with similar challenges
Because the number of individuals, households, programmes, etc. that must be studied is large, it would be impossible or very expensive to study all of them.

Hence, we study a relatively small sample, with the intention of inferring from the sample the situation of the entire population.

But, how confident can we be that the results observed in the sample properly represent the population?

Should it depend on how numerous the population is (the so called population size,) or on the number of items selected (the sample size,) and how diverse these items are in regards to what we wish to know?
Background cont’d

• In order to describe these concepts we use *sampling theory* (Formulae that frighten the lay and have acquired the reputation of being something that only experts can understand notwithstanding).
The Sampling controversy

• Shere Hite’s book “Women and Love: A Cultural Revolution in Progress (1987)” produced the following findings:
  – 84% of the women are not satisfied emotionally in their relationships (Page 804)
  – 95% of women report forms of emotional and psychological harassment from men with whom they are in love relationships (page 810)
III. The Sampling Controversy cont’d

• The book was widely criticised in the US and some referred to the conclusions as “dubious and doubtful”

• But what was the problem with the findings:
  • The research allowed women to discuss their experience to a great length in a way a multiple question questionnaire could not
  • He went ahead and generalised the results to all women and yet the sample was self selected (interviewed those willing to participate)
III. The Sampling Controversy cont’d

• Questionnaires were sent to professional women organisations, groups etc whose views may differ from the rest who are non members

• Some questions were “leading questions” while other were vague (according to Sharon Lohr)

• The questionnaire was so large that only those willing to fill it responded (120 pages)

• The response rate was extremely low (4.5% returned out of 100,000 questionnaires mailed)
III. The Sampling Controversy cont’d

• Does research that is not based on probability or random sample give one the right to generalise from the results of the study to the entire population?

• Answer: It depends on how large the sample is

• Questionnaires were sent to professional women organisations, groups etc whose views may differ from the rest who are non members

• Some questions were “leading questions” while other were vague

• This example is documented in the book “Sampling Design and Analysis by Sharon L. Lohr
Stratification considerations
Stratification considerations

• Geographic domains – Districts,
  *Urban/rural* – review definition, understand changes between censuses

Examples of other domains used in labour force surveys
Two-stage sampling

• Instead of taking a SRS

• We divide the territory into small areas, called Primary Sampling Units (PSUs).
  • In the first stage, we choose PSUs.
  • In the second stage, we select households in the chosen PSUs
Other Stratification considerations

- Number of households/population
- Employment groups
  - 1-4, 5-9, etc
- Gross output
  - Less than 1m, etc
- Type of Health facility
  - Health facility- hospitals, Health centre III, etc
  - Ownership (private/ Public)
Stratification (continued)

• Important to distinguish geographic domains and sampling strata
  • Domain requires minimum sample size to provide required level of precision
  • Sampling strata only require at least two PSUs per stratum
  • Possible to divide geographic domains into smaller more homogeneous sampling strata (for example, province, urban/rural)
Implicit stratification

- Implicit stratification - order the sampling frame of PSUs geographically in a serpentine manner (*rural urban within district, ascending order then descending order etc*)
- Ordering of PSUs in large city can be based on socioeconomic classification as well as geography
- Selection of PSUs systematically with PPS provides implicit stratification
- Implicit stratification ensures effective representation and proportional allocation to lower levels of geography
- Problem – some geographic codes are assigned alphabetically, not geographically
Allocation of sample to strata

• Proportional allocation
  • Effective for precision of estimates at the national level
• Equal allocation to each domain
  • Used when each domain requires same level of precision
• Optimum allocation – takes into account differential variance and costs by stratum
  • For example, variability may be higher in urban areas and enumeration costs may be higher in rural areas – in this case, higher sampling rate for urban areas
Subnational Estimates

• Sample size depends on number of different geographic domains for which separate, equally reliable estimates are required

• As a compromise, larger sampling errors accepted for subnational estimates
  • One proposal (by Dr. Vijay Verma) – increase national sample size by factor of $D^{0.65}$, where $D$ is the number of domains
  • Results in an average increase in the sampling errors for domain estimates by a factor of about 1.5
  • Minimum number of PSUs required for each domain – for example, 30 clusters

• Allocation of sample to domains
  • Equal allocation
  • Modified proportional allocation, with a minimum and maximum number of sample PSUs per domain
Stratification for special sub populations

• In some cases, it may be important to measure indicators for minority subpopulations
• Special strata identified for areas with concentration of minorities
• Higher sampling rates used for strata with concentration of minorities
• Example – Examples of minorities in Suriname?
Small area estimation

• Innovative methods linking Census to survey data are being applied to generate estimates using small area estimation techniques for smaller administrative units
  • Depends on availability of recent census together with survey results
• Different regression, and estimation models available
• Important to validate the results
SAMPLE SIZE DETERMINATION AND COMPLETION RATES

Sample Size
Completion rates
Major steps in designing a sample

- Define objectives
  - Key indicators
  - Desired level of precision
  - Sub-national domains of estimation

- Identify most appropriate sampling frame
  - Most recent census of population and housing
  - Sample for another survey conducted recently

- Determine sample size and allocation
  - Determine availability of previous results to provide measures of sampling parameters
Selection of key indicators for sample size determination

• Choose an important indicator that will yield the largest sample size

• Step 1: Select 2 or 3 target populations representing each a small percentage of the total population \((pb)\); typically
Selection cont’d

• **Step 2**: Review important indicators for these target groups but ignore indicators with very low or very high prevalence (less 10% or over 40%, respectively)

• Do not choose from the desirably low coverage indicators an indicator that is already acceptably low
Sample size

• **Requirements**
  • Margin of error/ precision requirements
    • A relative error, also known as *coefficient of variation* (*cv*) of 10 to 20 percent is, in fact, commonly specified as the precision needed for the key estimates of a survey
    • Statistically, the coefficient of variation is equal to the standard error of the survey estimate divided by the estimate.
    • For example, if you want to estimate each of its important items at the 95-percent level of confidence with, say, a relative error of 10 percent; for a 20-percent item this would translate into a standard error of 2 percentage points, while for a 40-percent item it would be 4 percentage points, and so forth
  • Reliability of the estimates for domains
  • Survey Budget and constraints
Sample size

\[ n = \frac{3.84f}{\sqrt{q}} \times \frac{q}{p} \]

\( n \) is the sample size we wish to calculate,
\( p \) is the anticipated proportion of facilities with the attribute of interest,
\( q \) is equal to \( 1 - p \)
\( f \) is the so-called design effect (shortened from deff),
    is the relative variance, (square of the relative error), and
\( \sqrt{3.84} \) is the square of the normal deviate (1.96) needed to provide an
    estimate at the 95 percent level of confident
sample calculation

<table>
<thead>
<tr>
<th>Illustration for sample calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.84 f</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>3.84</td>
</tr>
<tr>
<td>3.84</td>
</tr>
<tr>
<td>3.84</td>
</tr>
</tbody>
</table>
Determining sample size

• Important to examine tables of sampling errors and design effects from final report for previous surveys (e.g. Labourforce surveys, MICS, or DHS)

• In addition to using previous estimates for the indicators and design effects, possible to conduct simulation of sampling errors based on alternative sampling proposals, using results from previous survey
  • Example - Uzbekistan
Two-stage sampling

• Solves the problems of SRS
  • Reduces transportation costs
  • Reduces sample frame problems

• The sample can be made self-weighting if
  • We choose PSUs with Probability Proportional to Size (PPS), and then
    • We take a fixed number of households in each PSU

• The price to pay is cluster effect
Number of sample PSUs and cluster size

• Important to balance statistical efficiency and cost considerations
• Review DEFF for key indicators from previous survey reports to determine whether number of sample households per cluster should be changed
• Some household characteristics like safe water and improved sanitation have high intra class correlation, and thus high DEFFs, but these are less important
Completion rate

• The actual effective sample size depends on the completion rate, which is generally close to the response rate.

• The difference between the completion rate and the response rate depends on the number of selected households that are out-of-scope, such as selected vacant housing units.
Completion rate (continued)

• When a census list or older listing is used for selecting sample households, there will be more out-of-scope households selected, and the completion rate will be lower than the response rate.

• In this case the expected completion rate should be used instead of the response rate in the template for calculating the sample size.
CLUSTERING EFFECTS AND STAGES OF SAMPLE SELECTION
Cluster effect

Standard error grows if, instead of taking a Simple Random Sample of \( n \) households, we take a two-stage sample, with \( k \) PSUs and \( m \) households per PSU \((n=k\cdot m)\)

\[
e^2_{TSS} = e^2_{SRS} \left[ 1 + \rho (m - 1) \right]
\]
## Cluster Effect

For a total sample size of 12,000 households

<table>
<thead>
<tr>
<th>Number of PSUs</th>
<th>HHs per PSU</th>
<th>Intra-Cluster Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>3,000</td>
<td>4</td>
<td>1.03</td>
</tr>
<tr>
<td>2,000</td>
<td>6</td>
<td>1.05</td>
</tr>
<tr>
<td>1,500</td>
<td>8</td>
<td>1.07</td>
</tr>
<tr>
<td>1,000</td>
<td>12</td>
<td>1.11</td>
</tr>
<tr>
<td>800</td>
<td>15</td>
<td>1.14</td>
</tr>
<tr>
<td>600</td>
<td>20</td>
<td>1.19</td>
</tr>
<tr>
<td>400</td>
<td>30</td>
<td>1.29</td>
</tr>
<tr>
<td>300</td>
<td>40</td>
<td>1.39</td>
</tr>
<tr>
<td>200</td>
<td>60</td>
<td>1.59</td>
</tr>
<tr>
<td>150</td>
<td>80</td>
<td>1.79</td>
</tr>
<tr>
<td>100</td>
<td>120</td>
<td>2.19</td>
</tr>
</tbody>
</table>
Design effect

• In a two-stage sample
  Cluster effect = $e^2_{\text{TSS}} / e^2_{\text{SRS}}$

• In a more complex design
  (with two or more stages, stratification, etc.)
  Design effect = $Deff = e^2_{\text{Complex design}} / e^2_{\text{SRS}}$

• Can be interpreted as an apparent contraction of the sample size, as a result of clustering and stratification

• Can be estimated with special software
  (e.g., Stata’s svy commands)
Optimum cluster size

• Socioeconomic surveys – optimum cluster size in the range of 8 to 15 households

• Perhaps more research can be done on design effects and optimum cluster size, but general range of 10-15 households for income/expenditure surveys appear to be effective

• Discussion of experiences from the labour sampling experts

• Any other experiences?
Levels of clustering

• For three-stage design, with multiple clusters selected at second stage – two levels of clustering; reduces dispersion of sample, increases design effects

• When one cluster is selected in each PSU, treated as two-stage design, where the PSUs are considered intermediary stage for selecting clusters
First stage selection of PSUs

• Standard methodology for DHS and other household surveys – select Enumeration Area Blocks or clusters systematically with PPS

• Advantage – a constant number of sample households selected at second stage provides approximately self-weighting sample within stratum

• Provides implicit stratification
Sampling procedures for selecting PSUs with PPS

• Important to sort frame before selection, in order to ensure effective implicit stratification

• Traditional procedure – cumulate measures of size, determine sampling interval and random start, generate selection numbers

• Labourforce, MICS template – cumulate probabilities, formulas for identifying sample based on random start

• SPSS Complex Samples – option for selecting stratified sample systematically with PPS
  • Will not work if any PSU is larger than the sampling interval
Large sample PSUs in PPS sampling

• Sometimes a PSU may have a measure of size larger than the sampling interval
• PSU may be selected more than once in the systematic PPS selection
• Option 1 – if the PSU is selected two or more times, multiply the number of households to be selected by the number of “hits”
• Option 2 – separate the large PSUs and include in sample with a probability of 1
SECOND STAGE SELECTION OF SAMPLE HOUSEHOLDS
Household Listing

• Objectives:
  • Provide an updated list of all dwellings and households in each selected PSU to be used as sampling frame for second stage selection
  • Adjust for differences in PSU sizes during weighting: size used for PPS selection in the first stage differs from the observed size from the listing operation due to imperfections in the frame or demographic mobility
Household samples
Choosing the households

• The best sample frame is the full list of all households in the selected PSUs

• The household listing operation requires time and money. Relative to the project’s overall calendar and budget, these are
  • Marginal, if they are accounted for beforehand
  • Large enough to be a big headache, if they are not

• Information to be reported on the listing
  • Name and address, as a minimum
  • Additional data required for the selection (e.g., presence of pregnant women, or children any examples???)

• Households are generally selected from the listing by systematic equal probability sampling

Beware of imitations, such as random walks snowballing expert opinion

Do not ask additional information that is not essential
Household Listing cont’d

• Create the list of dwellings and households for the survey or borrow an existing list from a census or another survey

• Borrowing existing lists from a census or another survey:
  • Need to critically examine lists to ensure they are recent, complete and good maps are available
  • Information on the lists should allow the selected households to be located easily
Household Listing cont’d

• Borrowing existing lists from a census or another survey:
  • Lists that are more than 1 or 2 years old by the time of actual fieldwork can be outdated due to demographic mobility

• Household listing operation: a separate field operation before the survey starts or combined with household selection and interviewing into one single operation
Household Listing cont’d

• Household listing as a separate field operation:
  • More reliable as listing staff are less likely than interviewers to bias the sample by excluding households that are difficult to reach
  • Allows household selection to be done in a single central location using reliable and uniform procedures
  • More expensive but costs can be reduced by using segmentation
  • A separate household listing operation is usually recommended
Household Listing
Listing of households

• Common problems found in listing operations
  • Problem with quality of sketch maps – difficult to determine segment boundaries
  • Sometimes large differences found between number of households in frame (census) and number listed
Listing of households cont’d

• Importance of new listing to represent current population
• Problems with using previous listing (older than 1 year)
  • Does not represent newer households
  • Distribution of sample population by age group distorted, generally with higher median age
  • Difficulty of finding households in old list
Selection of sample households from listing

• Selection of households in the office following listing operation
  • Advantages – conducted by specialized staff, possible to avoid selection bias
  • Disadvantage – increased costs from having two field visits

• Selection of households in field
  • Advantage – cost savings of having one integrated field operation
  • Disadvantage - correct sampling may be difficult for field staff, selection may be biased
Household selection table

• One option for selection of households in field - household selection table

• When households selected in the field, it is best to avoid the use of random number tables and manual calculations, which can lead to mistakes and selection bias

• Excel spreadsheet used for generating systematic selection of fixed number of households based on number of households listed
Household selection table (continued)

• Providing household selection numbers to field staff provides more control
• Makes it difficult to cheat in selection of households, since selected households are not determined until the listing is completed and the total number of households is known
• Possible to verify later whether interviewer selected the correct households
Standard errors

- Effect of population size
- Effect of sample size
- Sampling error vs non sampling error
Effect of the population size

\[ e = \sqrt{1 - \frac{n}{N}} \sqrt{\frac{P(1 - P)}{n}} \]

Finite population correction

In practice this is almost always so close to 1 that we can safely ignore it.
Effect of the population size

In practice, the size of the population has very little influence on the sample size needed to achieve a given precision.
Effect of the sample size

To halve the standard error, the sample size needs to be quadrupled.
Sampling error vs non sampling error

Sample size vs Total error

Sampling error vs Non sampling error
Power calculations
• Power calculations permit making recommendations about the sample size needed for an impact evaluation
  • In the previous sessions, the key variable to recommend a sample size was the standard error, because the analytical technique was a point estimation with a confidence interval
  • In this session, the key variable is power (i.e. the probability of not getting a false negative), because the analytical technique is a statistical test
  • Both techniques are needed for an impact evaluation

• The key ingredient of a good power calculation is the standard error
Fundamentals of power calculations

- Power calculations permit making recommendations about the sample size needed for an impact evaluation
  - In the previous sessions, the key variable to recommend a sample size was the standard error, because the analytical technique was a *point estimation with a confidence interval*
  - In this session, the key variable is power (i.e. the probability of not getting a false negative), because the analytical technique is a *statistical test*
  - Both techniques are needed for an impact evaluation
- The key ingredient of a good power calculation is the *standard error*
Fundamental formula of power calculations

\[ MDE = (t_{1-\alpha/2} + t_{1-\beta})e \]

*\(MDE\): Minimum detectable effect
*\(\alpha/2\): Rate of Type I errors (false positives)
  (typically \(\alpha/2 = 2.5\%\))
*\(\beta\): Rate of Type II errors (false negatives)
  (typically \(\beta = 10\% - 20\% \iff \text{Power} = 90\% - 80\%\))
*\(e\): Standard error of the estimated effect
• Solve for $n$ the equation:

$$n = \left( \frac{t_{1-\alpha/2} + t_{1-\beta}}{MDE} \right)^2 2\sigma^2$$
### Table of normal deviates $t$

<table>
<thead>
<tr>
<th>Probability</th>
<th>Valor $t$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 tail</td>
<td>2 tails</td>
</tr>
<tr>
<td>$1-\gamma$</td>
<td>$t_{1-\gamma}$</td>
</tr>
<tr>
<td>80%</td>
<td>0.84</td>
</tr>
<tr>
<td>90%</td>
<td>1.28</td>
</tr>
<tr>
<td>95%</td>
<td>1.64</td>
</tr>
<tr>
<td>98%</td>
<td>2.05</td>
</tr>
<tr>
<td>99%</td>
<td>2.33</td>
</tr>
</tbody>
</table>

Value of $t$ for a 95% confidence level

\[ MDE = \left(1.96 + t_{1-\beta}\right)\sqrt{\frac{2\sigma^2}{n}} \]
Problem 1

• What do you think of this sampling design?
Answer 1

• Problems with the sampling design of this impact evaluation:
  1. It is using SRS for a household survey, which can be costly.
  2. The population of reference are only the young who enrolled in June 2013.
Problem 2

• Evaluate the power of this sampling design
  1. Reference power: 1-\(\beta\)=90%.
  2. Calculate the MDE:

\[
MDE = \left( t_{1-\alpha/2} + t_{1-\beta} \right) \sqrt{\frac{2\sigma^2}{n}}
\]
Table of normal deviates $t$

<table>
<thead>
<tr>
<th>Probability</th>
<th>Valor $t$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 tail</td>
</tr>
<tr>
<td>$1-\gamma$</td>
<td>$t_{1-\gamma}$</td>
</tr>
<tr>
<td>80%</td>
<td>0.84</td>
</tr>
<tr>
<td>90%</td>
<td>1.28</td>
</tr>
<tr>
<td>95%</td>
<td>1.64</td>
</tr>
<tr>
<td>98%</td>
<td>2.05</td>
</tr>
<tr>
<td>99%</td>
<td>2.33</td>
</tr>
</tbody>
</table>

Value of $t$ for a 95% confidence level

$$MDE = (1.96 + t_{1-\beta}) \sqrt{\frac{2\sigma^2}{n}}$$
### Answer 2

**Table of normal deviates t**

<table>
<thead>
<tr>
<th>Probability</th>
<th>Valor t</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-(\gamma)</td>
<td>(t_{1-\gamma}) (t_{1-\gamma/2})</td>
</tr>
<tr>
<td>80%</td>
<td>0.84</td>
</tr>
<tr>
<td>90%</td>
<td>1.28</td>
</tr>
<tr>
<td>95%</td>
<td>1.64</td>
</tr>
<tr>
<td>98%</td>
<td>2.05</td>
</tr>
<tr>
<td>99%</td>
<td>2.33</td>
</tr>
</tbody>
</table>

Value of \(t\) for a power of 90%

\[
MDE = (1.96 + 1.28) \sqrt{\frac{2\sigma^2}{n}}
\]
Answer 2

• The variance of the drop out rate \( p \) is equal to \( p(1-p) \).
• It is maximum for \( p=0.5 \)

\[
MDE = (1.96 + 1.28) \sqrt{\frac{2 \times 0.5(1 - 0.5)}{n}}
\]
Answer 2

• The sample size is 250.

\[ MDE = (1.96 + 1.28) \sqrt{\frac{2 \times 0.5(1 - 0.5)}{250}} \]
Answer 2

- The Minimum Detectable Effect is 14.5 percent points in the drop_out rate

\[ MDE = (1.96 + 1.28) \sqrt{\frac{2 \times 0.5(1 - 0.5)}{250}} = 0.145 \]
Sample Weights computation

• **Base weight calculations:**

  • Selection probabilities $p_1$ and $p_2$, where $p_1 = $ selection probability for PSUs, $p_2 = $ selection probability for households within PSU. These selection probabilities are available in an Excel format. The selection probability for the individual within each household $p_3$ is given by $1/$the number of eligible persons in the household ($hh2$). The number of eligible people in the household is obtained from the survey response data. The overall base weight ($wb$) is calculated as:
Base weight (wb)

\[ wb = \frac{1}{p_1 \times p_2 \times p_3} \]

• In addition, the base weight at the psu level \((wb_{psu})\) and at the household level \((wb_{hh})\) are calculated as:

\[ wb_{psu} = \frac{1}{p_1}, \quad wb_{hh} = \frac{1}{p_1 \times p_2} \]
Nonresponse adjustment:

• The non-response adjustment will be done at three levels: PSU level, household level and respondent level. The PSU level non-response adjustment is calculated by partitioning the 1079 PSUs into weighting classes defined by Region and residence, giving 37 * 2 = 74 adjustment cells. The PSU level non-response adjustment is:

\[ \text{psu}_\text{nr} = \frac{\sum \text{wb}_{\text{psu}}_{\text{eligible PSUs}}}{\sum \text{wb}_{\text{psu}}_{\text{non-missing PSUs}}} \]
Nonresponse adjustment: cnt’d

• The PSU non-response adjusted weight $wr_{psu}$ is the product of the base weight $wb$ and the PSU-level non-response adjustment.

<table>
<thead>
<tr>
<th>hhid</th>
<th>$\Sigma wb_{psu}$ eligible</th>
<th>$\Sigma wb_{psu}$ non-missing</th>
<th>psu_nr</th>
<th>$wr_{psu}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>167041</td>
<td>2870.18</td>
<td>2870.18</td>
<td>1.00</td>
<td>986.63</td>
</tr>
<tr>
<td>125081</td>
<td>5117.20</td>
<td>5117.20</td>
<td>1.00</td>
<td>4321.19</td>
</tr>
<tr>
<td>157361</td>
<td>19568.38</td>
<td>18827.02</td>
<td>1.04</td>
<td>1270.06</td>
</tr>
<tr>
<td>137371</td>
<td>22771.06</td>
<td>22070.41</td>
<td>1.03</td>
<td>6428.56</td>
</tr>
<tr>
<td>137741</td>
<td>22771.06</td>
<td>22070.41</td>
<td>1.03</td>
<td>4120.48</td>
</tr>
</tbody>
</table>
Nonresponse adjustment: cont’d

• The household non-response adjustment is calculated by PSU, so there are 1079 adjustment cells – one for each PSU. The household level non-response adjustment is calculated as:

\[ hh_{nr} = \frac{\sum wb_{hh}^{eligible\ household}}{\sum wb_{hh}^{completed\ rosters}} \]
Nonresponse adjustment: cont’d

- The household non-response adjusted weight $wr_{hh}$ is the product of the PSU non-response adjusted weight $wr_{psu}$ and the household non-response adjustment $hh_{nr}$. Due to large values of $hh_{nr}$ for 182 cases, the household non-response adjustment was trimmed at $hh_{nr} = 3$.

<table>
<thead>
<tr>
<th>hhid</th>
<th>$\Sigma wb_{hheligible}$</th>
<th>$\Sigma wb_{hhcompleted}$</th>
<th>$hh_{nr}$</th>
<th>$wr_{hh}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>167041</td>
<td>17759.25</td>
<td>15786.00</td>
<td>1.13</td>
<td>1109.95</td>
</tr>
<tr>
<td>125081</td>
<td>19445.36</td>
<td>19445.36</td>
<td>1.00</td>
<td>4321.19</td>
</tr>
<tr>
<td>157361</td>
<td>21995.03</td>
<td>19551.14</td>
<td>1.13</td>
<td>1427.56</td>
</tr>
<tr>
<td>137371</td>
<td>35307.65</td>
<td>29076.89</td>
<td>1.21</td>
<td>7806.11</td>
</tr>
<tr>
<td>137741</td>
<td>21299.70</td>
<td>13312.3</td>
<td>1.60</td>
<td>6592.76</td>
</tr>
</tbody>
</table>
Nonresponse adjustment: cont’d

• The person non-response adjustment is calculated by residence (urban/rural), gender, smoking status and age taken from the household roster. Therefore, there are 2*2*2*4=32 adjustment cells for the person non-response adjustment. The person-level non-response adjustment is:

\[ pp_{nr} = \frac{\sum wb_{\text{eligible households}}}{\sum wb_{\text{completed interviews}}} \]
Nonresponse adjustment: cont’d

• The final non-response adjusted weight is the product of the household non-response adjusted weight \((wr_{hh_{pp}})\) and the person non-response adjustment \((pp_{nr})\).

<table>
<thead>
<tr>
<th>hhid</th>
<th>(\Sigma wb_{eligible})</th>
<th>(\Sigma wb_{completed})</th>
<th>pp_nr</th>
<th>wr_{hh_pp}</th>
</tr>
</thead>
<tbody>
<tr>
<td>167041</td>
<td>233739.76</td>
<td>222133.55</td>
<td>1.05</td>
<td>1167.95</td>
</tr>
<tr>
<td>125081</td>
<td>3009504.18</td>
<td>2939829.79</td>
<td>1.02</td>
<td>4423.60</td>
</tr>
<tr>
<td>157361</td>
<td>4817285.25</td>
<td>4709716.88</td>
<td>1.02</td>
<td>1460.17</td>
</tr>
<tr>
<td>137371</td>
<td>4817285.25</td>
<td>4709716.88</td>
<td>1.02</td>
<td>7984.40</td>
</tr>
<tr>
<td>137741</td>
<td>1730377.38</td>
<td>1705108.19</td>
<td>1.01</td>
<td>6690.47</td>
</tr>
</tbody>
</table>
Post-stratification adjustment

- The post-stratification adjustment is calculated as:

\[ r = \frac{\text{pop}}{\sum \text{wb}_{hh pp}} \]
Weighting

- The final weight \((wf)\) is the product of the non-response adjusted weight \((wr_{hh\_pp})\) and the post-stratification adjustment \((r)\).

<table>
<thead>
<tr>
<th>hhid</th>
<th>State</th>
<th>Gender</th>
<th>Agegroup</th>
<th>pop</th>
<th>(\sum wr_{hh_pp})</th>
<th>r</th>
<th>wf</th>
</tr>
</thead>
<tbody>
<tr>
<td>167041</td>
<td>31</td>
<td>1</td>
<td>3</td>
<td>203029.0</td>
<td>217520.9</td>
<td>0.93</td>
<td>1090.13</td>
</tr>
<tr>
<td>125081</td>
<td>12</td>
<td>1</td>
<td>1</td>
<td>361006.0</td>
<td>102874.29</td>
<td>3.51</td>
<td>15523.29</td>
</tr>
<tr>
<td>157361</td>
<td>30</td>
<td>1</td>
<td>2</td>
<td>703585.0</td>
<td>550144.5</td>
<td>1.28</td>
<td>1869.07</td>
</tr>
<tr>
<td>137371</td>
<td>24</td>
<td>1</td>
<td>2</td>
<td>1656068.0</td>
<td>1435587.16</td>
<td>1.15</td>
<td>9210.67</td>
</tr>
<tr>
<td>137741</td>
<td>24</td>
<td>2</td>
<td>3</td>
<td>497528</td>
<td>565669.2</td>
<td>0.88</td>
<td>5884.52</td>
</tr>
</tbody>
</table>
Data Cleaning, editing
Data cleaning

• Data collection generates errors due to design of questionnaire as well as errors attributable to respondents and interviewers

• At Design level, efforts to minimize design errors is ensured. However, the flow of the questions may demand a more consistent system of editing

• Stages requiring data cleaning:
  • During data collection
  • During field supervision visits
  • Consistency checks and range checks during data entry/processing
  • Comparison with known demographics (sex ratio, cut-offs for nutrition data,
Data cleaning

• Treatment of screening questions for labourforce questions
• Treatment of extreme values
• Compare with data from other sources
• Review existing literature/ previous reports on similar or related subjects
• Conduct some re-interviews
• All these are possible BUT:
  • Requires staff who know the subject very well
  • Staff who are capable of detecting errors that may go beyond the machine edits
  • Need a small and knowledgeable team to edit and clean the data
Data cleaning, CAPI

• all possible range checks should be built into the application at the time of its development

• However, allow the application flexibility to continue to run even when an inconsistency is detected during an interview (this could be flagged for follow up after the interview)

• Plan for a small team of editors to validate the information as it comes in from the field. These could also perform a double role as editors/coders

• Develop do files or Cspro syntax to check population distribution, sex composition, sample distribution on a sample of data at regular intervals (every week or month,)
Coverage and content errors in Household surveys

• Coverage errors in Household surveys:
  • Ideally, a random sample is targeted but variation from the norm arises due to imperfect sampling frames
  • Sometimes, the sampling units are not identical to the unit of observations being studied
  • Non coverage error: failure to include some units of observation in the frame
  • Duplication of variables hence assigning more weight than is desired

• The main source of coverage error is the sampling frame (obsolete information which is not updated, and inappropriate Enumeration Area blocks etc)

• Ensure there is no overlap, no duplication, and assign unique identification in the frame,
Coverage and content errors in Household surveys cont’d

• Content errors in Household surveys:
  • Completion rates Vs response rates
  • Item non response rates Vs unit non response rates

• Refusals and failure to contact respondents partly responsible
• Respondents fatigue /burden increases content error

• Interviewer characteristics (training, level of education etc)
END

THANK YOU