Training workshop for consultants of TB prevalence surveys

Sampling design and sample size calculation; theory and concepts

25 February 2011
Phnom Penh, Cambodia

Charalampous (Babis) Sismanidis
Sian Floyd
From red to lime book

2007, 1st ed.

Assessing TB prevalence through population-based surveys

2011, 2nd ed.

Tuberculosis PREVALENCE SURVEYS: a handbook
second edition
The recommended 9 steps to sampling design

1. A prior guess of the true population prevalence

2. The relative precision around the estimate of prevalence

3. A prior guess of the magnitude of the "design effect"

4. Apply the recommended equation for the sample size calculation of a TB prevalence survey

5. A prior guess of the participation rate
The 9 steps to sampling design

6. Stratification to ensure a representative and precise overall estimate of prevalence

7. Cluster selection

8. Selection of individuals within cluster

9. Eligible survey population
1. A prior guess of the true population prevalence

- First step is to make a prior guess for the true population prevalence of TB

- Use national surveillance data (see WHO Global TB Report) AND other available research data

- Use previous prevalence survey results (if surveys comparable) and guess by how much TB decreased

- Must be done in close collaboration between a team including a statistician and local TB experts
2. The relative precision

• This (relative) precision refers to "how far away" we are allowing the survey's estimate of prevalence to be from the true national prevalence

• In statistical terms this translates into the width of the 95% confidence interval around the TB prevalence estimate we expect the survey to give us

• The higher the precision the larger the sample size

• Recommended precision is between 20% and 25%

  e.g. If prevalence 200, then 95% CI (160, 240)
3. A prior guess of the "design effect"

- We sample groups of people (clustered-random design), and not individuals (simple-random design)
- Clustered-sampled (CS) surveys have more uncertain results compared to simple-random sampled ones (for given assumptions)
- Therefore, we need an increased sample size for CS (multiply sample size for SRS by a factor called the "design effect")
- We estimate it from previous surveys OR the likely variation of prevalence between clusters
- Design effect gets bigger the bigger the:
  - difference in prevalence between clusters,
  - cluster size (recommend 400-1000 individuals),
  - expected TB prevalence is
Figure 5.3 For $k=0.4$, different combinations of TB prevalence $p$ (per 100 000 population aged ≥15 years) and cluster size.
4. Sample size calculation based on precision

Sample size for clustered sampled survey =
(sample size for simple random sampling) x (design effect)

\[ N = \left[ 1.96^2 \frac{\left(1 - \pi_g\right)}{d^2 \pi_g} \right] \times \left[ 1 + (m - 1) \frac{k^2 \pi_g}{\left(1 - \pi_g\right)} \right] \]

<table>
<thead>
<tr>
<th>( N )</th>
<th>Number of people included in the survey</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \pi_g )</td>
<td>“Prior guess” of the true population prevalence of pulmonary TB (expressed as a proportion)</td>
</tr>
<tr>
<td>( d )</td>
<td>Relative precision (expressed as a proportion). Recommended 0.20 or 0.25</td>
</tr>
<tr>
<td>( m )</td>
<td>Cluster size (=number of targeted individuals), assumed to be constant across clusters</td>
</tr>
<tr>
<td>( k )</td>
<td>Coefficient of between-cluster variation. Recommended to assume ( k ) is in the range 0.4 – 0.6</td>
</tr>
</tbody>
</table>
### 4. Sample size calculation; examples (cont.)

<table>
<thead>
<tr>
<th>TB prevalence</th>
<th>Cluster size</th>
<th>precision=0.2</th>
<th>precision=0.25</th>
</tr>
</thead>
<tbody>
<tr>
<td>100 per 100,000</td>
<td>500</td>
<td>103612</td>
<td>66312</td>
</tr>
<tr>
<td></td>
<td>750</td>
<td>107453</td>
<td>68770</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>111295</td>
<td>71229</td>
</tr>
<tr>
<td>150 per 100,000</td>
<td>500</td>
<td>71598</td>
<td>45823</td>
</tr>
<tr>
<td></td>
<td>750</td>
<td>75440</td>
<td>48282</td>
</tr>
<tr>
<td></td>
<td>1000</td>
<td>79282</td>
<td>50740</td>
</tr>
</tbody>
</table>

*Difference in prevalence between clusters increases.*
4. Sample size calculation; cluster size (cont.)

- This is recommended to be between 400-1000. It is a difficult choice we must make considering opposing parameters.
- The smaller the cluster size, the larger the number of clusters required (larger number of field teams) and the more difficult standardization becomes.
- The larger the cluster size, the larger the design effect (larger total sample size) AND the longer cluster operations become (recommend to complete a single cluster's operations within a week).
- Population size of the smallest available geopolitical unit (i.e. census enumeration area) is usually a good starting point to guide target cluster size.
- Final cluster size is a difficult balancing act of all these considerations.
5. Adjust sample size for non-participants

- In a field survey, some people will either not attend the initial screening, or will drop out during the survey.

- Calculated sample size should be increased to allow for non-participation in the survey.

- Recommended value is 85–90%:

  e.g. \((\text{New sample size}) = \frac{\text{(old sample size)}}{0.85}\)
6. Stratification

• TB prevalence will typically vary across different geographical regions of a country (AKA strata)
• A stratified design should be used to increase the precision and representativeness of the overall country estimate of TB prevalence
• This means number of clusters from each stratum is proportional to population size in the stratum
• By design the approach of stratification allows the estimation of stratum-specific estimates of TB, but their precision is smaller compared to the overall nationwide estimate

If Java-Bali/Sumatera/Eastern Indonesia population split is 50%/40%/10%, and we need to randomly select 100 clusters then 50 should come from Java-Bali, 40 from Sumatera and 10 from Eastern Indonesia.
7. Cluster selection

- Once cluster size (m) is chosen and sample size (N) calculated; number of clusters = N/m
- Examples of clusters are villages, household blocks, census enumeration areas
- Cluster selection is a multi-stage process starting from larger to smaller sampling units, to promote geographical representation
- Selection of clusters in every stage is proportional to population size (PPS)
- Selection at final stage (where sampling units are roughly equally sized) is random
  e.g. Nigeria: Zones, states, LGAs, EAs
7. Cluster selection
8. Selection of individuals within cluster

- Once a cluster is selected, the target sample size of eligible survey individuals is identified and invited.

- Cluster is split into household groups (roughly equally sized) using e.g. paths, roads, natural boundaries.

- Household groups are selected at random.

- Number of participants in each cluster should be as similar as possible across clusters enrolled in the survey (this follows from the use of PPS to select clusters AND simplifies the analysis; not essential to apply weights).
9. Eligible survey population

- Eligible survey individuals should be representative of the target population.

- Eligibility is based only on:
  - age (aged 15 years or older)
  - residency status in the household (e.g. people living in the household for the past 4 weeks between pre-census and census visits)

- All eligible individuals should be classified as:
  (i) participants, (ii) absentees and (iii) non-consenters

- Enumerating those under 15 years is also important for adjustments to demographic changes.
Sample size calculation based on repeat surveys

\[
N = \frac{(1.65 + z_\beta)^2 \times \{\pi(1 - \pi) + (m_2 - 1)\pi^2 k_2^2\} \times N_1}{(\pi - p_1)^2 \times N_1 - (1.65 + z_\beta)^2 \times \{p_1(1 - p_1) + (m_1 - 1)p_1^2 k_1^2\}}
\]

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>(N)</td>
<td>Number of people included in the second survey</td>
</tr>
<tr>
<td>(\pi)</td>
<td>“Prior guess” of the true population prevalence of pulmonary TB for the second survey, expressed as reduction from the first (p_1)</td>
</tr>
<tr>
<td>(z_\beta)</td>
<td>The z-value of the test power = 1 - (\beta) (0.84 or 1.24 for power of 80% or 90%)</td>
</tr>
<tr>
<td>(m_1, m_2)</td>
<td>Cluster size, for first and second survey respectively (assumed to be constant across clusters within each survey)</td>
</tr>
<tr>
<td>(k_1, k_2)</td>
<td>Coefficient of between-cluster variation, for first and second surveys respectively. Recommended to assume (k) is in the range 0.4 – 0.6</td>
</tr>
</tbody>
</table>
Exercise; Part I

Example of sample size calculation

Ethiopia 2010-2011
1. A prior guess of the true population prevalence

- Use national surveillance data (see WHO Global TB Control Report) AND other available research data

- WHO estimate for prevalence of smear-positive pulmonary TB in total population (including children) was 286 per 100,000 in 2008

  Conservative assumption: assume currently 200 per 100,000

- Justified on basis of extensive DOTS service expansion in the country, recent involvement of Health Extension Workers in TB prevention and control activities in the community, and relatively low HIV prevalence

- 55% of total population is adult 15+ years old →
  Prevalence in adults 15+ years = 200 / 0.55 = 364 per 100,000 adults 15+ years old

- So: best guess = 364 per 100,000

- Estimate of plausible range: 300 to 400 per 100,000
2. The relative precision

• In statistical terms this is the width of the 95% confidence interval around the TB prevalence estimate that will come from the survey

• Recommended relative precision is between 20% and 25%

• Ethiopia: Aim for a relative precision of 20%

• But also did sample size calculation for a less stringent 25%, for comparison
3. A prior guess of the "design effect"

- We sample groups of people ("clusters"), and not individuals
- Design effect gets bigger, the bigger the:
  - variation in prevalence among clusters (k)
    Ethiopia best guess k=0.5; plausible range 0.4 to 0.6
  - cluster size (m)
    Ethiopia chose 550: can be completed in 1 week, and is within the recommended range of 400-1000
  - expected TB prevalence (\(\pi_g\))
    Ethiopia best guess = 364 per 100,000; plausible range 300 to 400; i.e. \(\pi_g = 0.00364\), range 0.003 to 0.004

- Design effect formula:

\[
DEFF = \left[1 + (m-1) \frac{k^2 \pi_g}{(1-\pi_g)} \right]
\]
3. A prior guess of the "design effect"

- So, example design effect calculation:
  \[ k=0.5, m=550, \pi_g = 0.00364 \]

  \[ DEFF = \left[ 1 + (550 - 1) \times \frac{0.5^2 \times 0.00364}{1 - 0.00364} \right] \]

  \[ DEFF = 1.501 \]
4. Final equation for the sample size calculation (not allowing yet for non-participation)

Ethiopia 2010

- $\pi_g = \text{prevalence best estimate } 0.00364$
- $m = \text{cluster size planned as } 550$
- “best estimate” of design effect 1.501 (with $k=0.5$)
- $d$ chosen to be 0.2

\[
N = \left[ 1.96^2 \frac{(1 - \pi_g)}{d^2 \pi_g} \right] \times \left[ 1 + (m - 1) \frac{k^2 \pi_g}{(1 - \pi_g)} \right]
\]

\[
N = \left[ 1.96^2 \times \frac{(1 - 0.00364)}{0.2^2 \times 0.00364} \right] \times \left[ 1 + (550 - 1) \frac{0.5^2 \times 0.00364}{(1 - 0.00364)} \right]
\]

- Sample size 39,470
5. A prior guess of the participation rate

- In a field survey, some people will either not attend the initial screening, or will drop out during the survey.

- Sample size should be adjusted to allow for non-participation in the survey.

- Ethiopia 2010: Assumed 85% participation
  
  So: (New sample size) = (old sample size) / 0.85
  
  = 39,470 / 0.85
  
  = 46,435

- Total number of clusters = 46,435 / 550 = 84.2, round up to 85
5. Sample size required for alternative scenarios of true TB prevalence, K and required precision (keeping cluster size fixed at 550 and participation rate fixed at 85%)

<table>
<thead>
<tr>
<th>TB prevalence</th>
<th>Relative precision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>precision=0.2</td>
</tr>
<tr>
<td></td>
<td>Design effect increases</td>
</tr>
<tr>
<td></td>
<td>$k=0.4$</td>
</tr>
<tr>
<td>300</td>
<td>47475</td>
</tr>
<tr>
<td>364</td>
<td>40853</td>
</tr>
<tr>
<td>400</td>
<td>38059</td>
</tr>
</tbody>
</table>

Sample size of 46,435 sufficient for all $k$, and all TB prevalence 300-400, with lower relative precision of $d = 0.25$
6. Stratification

- TB prevalence will typically vary across different geographical regions of a country (AKA *strata*)

- A **stratified** design should be used

- This means number of clusters from each stratum proportional to population in the stratum

**Ethiopia 2010:**

- Urban, rural, pastoralist

- 16% population urban – allocated $0.16 \times 85$ clusters = 14

- 77% population rural – allocated $0.77 \times 85$ clusters = 65

- 7% population pastoralist – allocated $0.07 \times 85$ clusters = 6
7. Cluster selection

- Once cluster size \(m\) is chosen and sample size \(N\) calculated; \(N\) clusters = \(N/m\)
- Examples of clusters are villages, household blocks, census enumeration areas
- Cluster selection is a multi-stage process starting from larger and moving to smaller sampling units

Ethiopia 2010:

- Administratively, has a total of 9 States and 2 City administrative areas, 94 Zones, 810 Woreda (comparable to a district), and 15022 Kebele
- Woreda = primary sampling unit, Kebele = secondary sampling unit
- Average population of a Kebele is 5000 individuals
- Cluster is a Kebele, and need to sub-sample from it to get target cluster size of 550
7. Cluster selection
8. Selection of individuals within cluster

- Once a cluster is selected, the target cluster number of eligible survey individuals is identified and invited.
- Cluster is split into household groups using e.g. paths, roads, natural boundaries.
- Household groups are selected at random.

**Ethiopia 2010:**

- Divide each selected Kebele into blocks of households, choose one block at random as a starting point, and then proceed in a clockwise manner from the starting point until the target sample size is achieved.
9. Eligible survey population

- Eligible survey individuals should be representative of the target population

- Eligibility is based only on:
  - age (aged 15 years or older)
  - residency status in the household (e.g. people living in the household for the past 4 weeks between pre-census and census visits)

Ethiopia 2010:
- Eligible individuals defined as those who were a member of a household in the Kebele at the time of the survey team’s second visit to the Kebele (the second visit is the one at which the community in the Kebele is informed about the survey and is done 2-3 weeks before the survey)
Estimating $K$ - Key step = think about plausible values of 2.5% and 97.5% centiles of distribution of cluster-level TB prevalence (based on knowledge of the country)

1) $k = \frac{SD}{true\ mean\ prevalence}$, by definition
2) $SD = \frac{50\%\ centile - 2.5\%\ centile}{1.96}$, if “normal” distribution
   $= \frac{(364-10)}{1.96} = 181$
3) $\rightarrow$ estimate of $k = 181 / 364 = 0.50$ (rounded to 2 decimal places)
Exercise; Part II

Example of cluster selection

Cambodia 2010-2011
Overview of cluster selection; how many?

- We need to sample 62 clusters of 640 individuals each (approximately 40,000 of total sample size)

- Stratified sampling design is used. Strata are defined as: urban/rural/"2002 survey excluded areas"

- Number of clusters to be sampled from each stratum using probability proportional to size:
  - Stratum population split: 21%/76%/3%
  - Number of cluster split: 13/47/2 (=62)

- Cluster sampling is done independently within each stratum
Overview of cluster selection; multi-stage (PPS and random selection)

- Four administrative geopolitical levels exist in Cambodia; provinces, districts, communes and villages (=clusters)

- **Stage 1:** Districts have been selected as the **primary sampling unit.** Firstly, from a complete list of all districts, we identify the desired number of districts using PPS

- **Stage 2:** Communes are the **secondary sampling unit.** Secondly, from a complete list of all communes **within each of the identified districts**, we identify the desired number of communes using PPS

- **Stage 3** (**last stage**): Villages, fairly similar in terms of their population size, are the **final sampling unit - AKA "cluster".** Lastly, from a complete list of all villages **within each of the identified communes**, we randomly sample (**all villages have equal weight of 1**) villages

- This process is repeated independently for each stratum
Stage 1; identifying 13 urban districts using PPS

Sampling interval (SI) = total cumulative population / number of districts we want to identify = 1 909 749/13 = 146 903

There are 48 districts in total with a combined cumulative population of 1 909 749

<table>
<thead>
<tr>
<th>District</th>
<th>Column A</th>
<th>Column B</th>
<th>Column C</th>
<th>Column D</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>District population aged ≥ 15 years</td>
<td>Cumulative population</td>
<td>Cumulative population point selected for sampling</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>15598</td>
<td>15598</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5621</td>
<td>21219</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>61372</td>
<td>82591</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>43075</td>
<td>125666</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>86189</td>
<td>211855</td>
<td>127785</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>15139</td>
<td>226994</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>27057</td>
<td>254051</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>16230</td>
<td>270281</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>9326</td>
<td>279607</td>
<td>274688</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>12933</td>
<td>292540</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>34628</td>
<td>327168</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>10356</td>
<td>337524</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>30630</td>
<td>368154</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>5504</td>
<td>373658</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>32595</td>
<td>406253</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>22550</td>
<td>428803</td>
<td>421591</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>6513</td>
<td>435316</td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>22197</td>
<td>457513</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>6074</td>
<td>463587</td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>28240</td>
<td>491827</td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>44411</td>
<td>536238</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>9073</td>
<td>545311</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Random start (RS) = random number between 1 and 146 903 (=SI)

1st identified district

2nd identified district

3rd identified district

= RS + SI = 127785 + 146903

= RS + (2*SI)
Stage 2; identifying communes within districts

- At the end of stage 1 we will have identified 13 urban districts

- Performing PPS at stage 2 is exactly the same as described for stage 1, using a list of randomly ordered communes (along with their population estimates) within each of the identified districts from stage 1. The number of units within each district we need to identify at this stage is typically small (1-2)
Stage 3; identifying villages within communes

- At the end of stage 2 we will have identified 13 urban communes

- During this last stage, using a list of all available (broadly equally sized) villages within each of the identified communes from stage 2, we randomly select (different approach to PPS) the desired number of villages (typically 1)

- We repeat this process for rural (47 clusters) and "2002 survey excluded areas " (2 clusters) strata
Part III

Perform sample size calculation

Uganda 201x
Practical exercise: Sample size calculation for Uganda

Uganda 2010

- Prevalence best estimate 300 per 100,000 in total population
- Estimate 53% of population is 15+ years old
- \( m = \) cluster size, planned as 910 (estimate this will take 2 weeks)
- \( k: \) Try different values, at least 0.4 and 0.6
- \( d: \) Try both 0.2 and 0.25
- Participation rate: Use 85% (more than this could be too optimistic)

Q1: What is the design effect, for each of \( k=0.4 \) and \( k=0.6 \)?

Q2: What is the calculated total sample size in each of the 4 scenarios (\( k=0.4 \) or \( k=0.6 \), \( d=0.2 \) or \( d=0.25 \))?