Calculating the sample size for surveys for the prevalence of TB
Why are we doing this?

- Compare to past values
- Estimate current prevalence
- Measure future trends
Step 1: Base Sample-size Calculation

\[
\frac{n' = t^2 \times p(1-p)}{m^2}
\]

- **Description:**
  - \( n \) = required sample size
  - \( t \) = confidence level at 95% (standard value of 1.96)
  - \( p \) = estimated prevalence of TB (SS+) in the project area
  - \( m \) = relative precision
Precision

• **Absolute precision:**
  – Refers to the actual uncertainty in a quantity. For example, prevalence of TB is 20% ± 10%, the absolute uncertainty is 10%.

• **Relative precision:**
  – Expresses the uncertainty as a fraction of the quantity of interest. For our example of a prevalence of 20% ± 10%, the relative uncertainty is 10% of 20% which is equal to 2%.
Sample size

The design effect

Random

Cluster
Step 2: Design effect

The design effect, $D$, gives the increase in the variance arising from the cluster design and hence the amount by which we have to increase the sample size.

\[ D = 1 + m\alpha^2\mu \]

$m$ is the cluster size
$\alpha$ is the between cluster variation (std. dev./mean)
$\mu$ is the mean prevalence of TB

$D$ is big if:
- **There are a few large clusters (big $m$);**
- The clusters are very different (big $\alpha$);
- The prevalence is high (big $\mu$).
Step 2: Design effect (2)

\[ n'' = n' \times d \]

• **Description:**
  – \( n'' \) = required sample size correcting for design effect
  – \( n' \) = previously calculated sample size
  – \( d \) = design effect

• In most prevalence TB disease surveys a design effect of 1.5 to 2 seems reasonable
Step 3: Participation rate

$$n''' = n'' \times (100 + (1 - pr))$$

- **Description:**
  - $n'''$ = required sample size correcting for participation rate
  - $n''$ = previously calculated sample size
  - $pr$ = participation rate

- In most prevalence TB disease surveys a participation rate of 85% seems reasonable
If we want to estimate prevalence by gender, age, geography or anything else we need to multiply the sample size by the number of strata.
Calculating the Cost Effectiveness

\[ C \propto \frac{1.96}{\mu \varepsilon^2} \left( 1 + \alpha^2 m \mu \right) + l \rho = \frac{1.96}{\mu \varepsilon^2} \left( 1 + \alpha^2 m \mu \right) \left( 1 + \frac{\rho}{m} \right) \]

Cost, C, is a minimum when

\[ \tilde{m} = \sqrt{\frac{\rho}{\alpha^2 \mu}} \]

\[ \tilde{l} = \frac{n}{\tilde{m}} \]
In Cambodia the marginal cost of adding a new cluster was 750 times the cost of adding one person to an existing cluster.
1. Do a rough calculation of sample size
2. Make sure that you have the right question
3. Make an estimate of the design effect
4. Design your sampling strategy
5. Think about the costs and logistics and make sure that you have a sensible cluster size