5.1 Sample size calculation for a simple random sample survey, and definition and estimation of the design effect

This annex is designed for readers who would like to have a deeper understanding of the mathematical principles that underpin calculation of the sample size for a simple random sample survey, and the size of the design effect.

1. Sample size calculation assuming a simple random sample survey is to be done

If the number of people included in the survey is denoted by $N$ and the number of TB cases found in the survey is $t$ then:

The point estimate of the true prevalence of TB will be estimated from the survey as:

$$ p = \frac{t}{N} \quad (1) $$

and the variance of the survey estimate $p$ is given by the usual binomial expression

$$ \sigma_p^2 = \frac{\pi(1-\pi)}{N} \quad (2) $$

and the standard error by

$$ \sigma_p = \sqrt{\frac{\pi(1-\pi)}{N}} \quad (3) $$

The 95% confidence interval for the prevalence of TB is calculated as:

$$ p \pm 1.96 \sigma_p \quad (4) $$

It is thus clear that, the greater is the required precision (i.e. the narrower is the required width of the 95% confidence interval), the bigger must be the sample size $N$.

The next step is thus to define the relative precision that is required. We can denote this as $d_\pi$, where $d$ is a proportion and $\pi$ is the population prevalence of TB. For example, if the required relative precision is 0.2 (20%) then $d_\pi = 0.2\pi$; if the required relative precision is 0.25 (25%) then $d_\pi = 0.25\pi$.

We are then requiring that:

$$ 1.96 \sqrt{\frac{\pi_g(1-\pi_g)}{N}} = d_\pi \pi_g \quad (5) $$

where $\pi_g$ is our "prior guess" of the true TB prevalence.

If we re-arrange this equation, we obtain

$$ N = \frac{1.96^2 \pi_g (1-\pi_g)}{d_\pi^2 \pi_g} \quad (6) $$

as the required total sample size, depending on the required $d$ and prior guess of TB prevalence $\pi_g$.

2. Design effect correction to sample size calculation, due to cluster sample survey design

In a cluster sample survey, the total population is partitioned into “clusters” of individuals, for example into towns and villages. As explained in Section 5.3, “clusters” are selected at random from the total population of clusters, usually in a multi-stage process, and then a random sample of individuals from each selected cluster is included in the TB prevalence survey.

The extra uncertainty about the true prevalence of TB due to a cluster sample survey design, compared with a simple random sample survey design, is called the “design effect”.

It is assumed in the following discussion of the design effect that the number of eligible individuals included in the survey is the same in each cluster, though in practice some clusters may include a higher number of individuals than the target. A constant cluster size (in terms of eligible individuals included in the survey) follows from the use of PPS sampling, and is the preferred approach from a logistical (and analytical) point of view.

To estimate the design effect, an estimate of either:

(i) The between-cluster variation in true TB prevalence, $\sigma_{\pi_b}^2$

or

(ii) The correlation between individuals in the same cluster for whether or not they have TB, i.e. the intra-cluster
correlation coefficient, \( \rho \) is required. Technically, they are equivalent but one (\( \sigma^2_B \)) is an absolute measure and the other (\( \rho \)) is relative and constrained to be between 0 and 1. If individuals in the same cluster are no more alike to each other than they are to individuals in a different cluster, then \( \rho \) is 0; at the other extreme, if in the same cluster each individual has the same value for TB (yes or no), and this occurs for all clusters, then \( \rho \) is 1.

From the perspective of (i), we are taking into account that the true prevalence of pulmonary TB varies among clusters, and thus that observed variation in the prevalence of pulmonary TB between clusters is not just due to sampling variation. We can represent this by assuming that the true prevalence in cluster \( i \) is \( \pi_i \), that the true prevalence averaged over all clusters in the country’s population is \( \pi \), and that the variation of the \( \pi_i \) about the overall prevalence \( \pi \) is \( \sigma^2_B \). Such variation in the true prevalence of TB between clusters is termed “over-dispersion”.

In the literature on how over-dispersion can be modelled, the simplest way in which this can be done is to assume that

\[
\sigma^2_B = \Phi \pi (1 - \pi) \tag{7}
\]

where \( \Phi \) is an unknown scale parameter that takes a value \( \geq 0 \) and \( \pi \) is the true overall prevalence, averaging across all clusters in the population[Collett, 2003] . With this framework, and with all clusters having the same number of eligible individuals included in the survey, it can be shown that the variance of the survey estimate \( \hat{p} \) is equal to:

\[
\frac{\pi (1 - \pi)}{N} \left[ 1 + (m - 1) \Phi \right] \tag{8}
\]

where \( m \) is the number of eligible individuals included in the survey from each cluster (cluster size) and \( N \) is the total sample size. Comparing this with equation (2), we can see that the variance of the survey estimate \( \hat{p} \) is increased by a factor of \( 1 + (m - 1) \Phi \), and it is this factor that is termed the “design effect”.

From the perspective of (ii), the intra-cluster correlation, the variance of the survey estimate \( \hat{p} \) can be shown [Collett, 2003] to be:

\[
\frac{\pi (1 - \pi)}{N} \left[ 1 + (m - 1) \rho \right] \tag{9}
\]

Thus \( \rho \) can be equated to \( \Phi \), and so we can also represent the between-cluster variation as:

\[
\sigma^2_B = \rho \pi (1 - \pi) \tag{10}
\]

or, equivalently,

\[
\rho = \frac{\sigma^2_B}{\pi (1 - \pi)} \tag{11}
\]

and the design effect as:

\[
DEFF = \left[ 1 + (m - 1) \rho \right] \tag{12}
\]

3. Estimation of the design effect in terms of \( k \), the coefficient of variation

From section 2. above, an estimate of \( \rho \) or \( \sigma^2_B \) is required in order to make a “prior guess” about the value of the design effect. With a binary outcome, such as TB yes or no, it is easiest to think in terms of \( \sigma^2_B \), the between-cluster variation. It also turns out to be easiest to make a prior guess as to the value of \( \sigma^2_B \) in terms of the coefficient of variation \( k \), because \( k \) (like \( \rho \)) is a relative measure (although it is not constrained to be less than 1).

The coefficient of variation, \( k \), of the cluster-specific TB prevalences is defined as standard deviation (SD) / true overall population value, and thus in the case of a TB prevalence survey,

\[
k = \frac{\sigma_B}{\pi} \tag{13}
\]

Thus, for example, if \( \sigma_B \) is estimated as 0.4\( \pi \), then we have that \( k = 0.4\pi / \pi = 0.4 \).

Because \( k = \sigma^2_B / \pi^2 \), we can substitute \( k \) into equation (11) to obtain

\[
\rho = \frac{k \pi * (k \pi)}{\pi (1 - \pi)} \tag{14}
\]

which simplifies to

\[
\rho = \frac{k^2 \pi}{(1 - \pi)} \tag{15}
\]

We can then express the design effect in terms of cluster size, \( k \), and \( \pi \):

\[
DEFF = \left[ 1 + (m - 1) \frac{k^2 \pi}{(1 - \pi)} \right] \tag{16}
\]
We thus have 2 ways to estimate the design effect, using (16) or using

\[ DEFF = \left[1 + (m - 1)\rho\right] \] (17)

If we use equation (16) to predict the value of the design effect, then for \( \pi \) we substitute the value of our “prior guess” \( \pi_g \). Thus our equation for the sample size, corrected for the design effect and using equation (16) for the design effect, becomes:

\[ N = \left[1.96^2 \left(1 - \pi_g \right) / d^2 \pi_g \right] \left[1 + (m - 1) k^2 \pi_g / (1 - \pi_g) \right] \] (18)

Reference