Designing Minimum-Cost Multi-Stage Sample Designs

Matthias Ganninger*

Abstract

In cross–national sample surveys, a huge variety of sample designs is often applied in participating countries. In order to achieve estimates of comparable precision, the samples drawn according to these different sampling schemes must have the same *effective sample size*, $n_{\text{eff}} = \frac{n}{\textit{deff}}$, where n is the net sample size and *deff* is the design effect. As *deff*, among another parameter, depends on the average cluster size, $\bar{b}$, increasing the number of sampled clusters, ceteris paribus, decreases the design effect and hence increases n_{eff}. At a given cost structure (costs per interview and costs per sampled cluster), there exists an optimal number of clusters to sample so that a pre–defined effective sample size is exactly achieved — at minimum total costs.

Key Words: Design Effects, multi-stage sampling, optimal

1. Introduction

Comparative sample surveys like the European Social Survey (ESS) aim at providing high quality data that yield estimates of comparable precision at minimum bias and high precision. As far as precision is concerned, a necessary condition to assure comparability between samples of different countries is to achieve the same effective sample size $n_{\text{eff}} = \frac{n}{\textit{deff}}$ in the samples of participating countries where n is the net sample size and *deff* is the *design effect*. The design effect is a measure for the inflation of variance of an appropriate estimator for a population parameter θ under a complex sample design compared to the variance of an appropriate estimator for the same parameter under simple random sampling with replacement (Kish, 1965). In the case of two-stage sampling with unequal inclusion probabilities, the design effect can be decomposed into 1) the design effect due to unequal inclusion probabilities ($\textit{deff}_p$) and 2) the design effect due to clustering ($\textit{deff}_c$), *deff* being the product of the two (Gabler et al., 1999; Ganninger, 2010):

$$\textit{deff} = \textit{deff}_p \cdot \textit{deff}_c \tag{1}$$

with

$$\textit{deff}_p = n \cdot \frac{\sum_{i=1}^{n} w_i^2}{\left(\sum_{i=1}^{n} w_i\right)^2} \tag{2}$$

and

$$\textit{deff}_c = 1 + \left(\bar{b} - 1\right) \cdot \rho \qquad . \tag{3}$$

In the above equations w_i is the usual design weight associated with the ith element, $\bar{b} = \frac{n}{m}$ is the average cluster size, m is the number of clusters in the sample and ρ is the intra-class correlation coefficient. Hence, the effective sample size can be written as

$$n_{\text{eff}} = \frac{n}{\textit{deff}_p \cdot \textit{deff}_c} = \frac{n}{\left(n \cdot \frac{\sum_{i=1}^{n} w_i^2}{\left[\sum_{i=1}^{n} w_i\right]^2} \right) \cdot \left(1 + \left[\bar{b} - 1\right] \cdot \rho\right)} \tag{4}$$

*GESIS – Leibniz-Institute for the Social Sciences, B2 1, 68159 Mannheim, Germany

or, taking $deff_p$ as given

$$n_{\text{eff}} = \frac{n}{deff_p \cdot \left(1 + \left[\bar{b} - 1\right] \cdot \rho\right)} \quad . \tag{5}$$

With a given definition of primary sampling units in the population (e.g. municipalities), the magnitude of ρ is determined by the distribution of the values of teh study variable within and between the clusters (see Kish, 1965, 139). Any unbiased estimator $\hat{\rho}$ for ρ based on sample data from a sample design in which primary sampling units are drawn at the first stage (e.g. one-stage or two-stage sample design) will, on average, reproduce ρ with some sampling distribution. The only means to influence the magnitude of $deff_c$ in the planning stage of a survey is by changing the number of sampled clusters m and hence the average cluster size, $\bar{b}$. Substituting $\frac{n}{m}$ for $\bar{b}$ in (5) and solving for m gives

$$m_{\text{opt}} = \frac{n_{\text{eff}} \cdot n \cdot deff_p \cdot \rho}{n - n_{\text{eff}} \cdot deff_p + n_{\text{eff}} \cdot deff_p \cdot \rho} \tag{6}$$

which is the optimal number of clusters to sample in order to reach a specified effective sample size.

For example, let $n = 2000$, assume $\rho = 0.04$ and let $deff_p$ be 1.2 as usual for a sample of household where the only variation in weights comes from different inclusion probabilities within the households. Finally, assume the required effective sample size as 1500, as for example in the ESS (ESS, 2005). Given these constraints

$$m_{\text{opt}} = \frac{1500 \cdot 2000 \cdot 1.2 \cdot 0.04}{2000 - 1500 \cdot 1.2 + 1500 \cdot 1.2 \cdot 0.04} = 529.42 \approx 530,$$

i.e. at least 530 clusters of average size $\bar{b} = \dfrac{2000}{530} = 3.78$ have to be sampled to reach n_{eff}.

2. Minimum cost sample design

Usually (Kish, 1965, pp. 99), total fieldwork costs are assumed to arise from two sources:

c_I interview-related costs, i.e. costs that arise for a conducted interview

c_T travel-related costs, i.e. costs that arise for an interviewer to go to a certain psu

Hence, following a simple linear cost model (Kish, 1965, 268) total fieldwork costs are defined as

$$c = c_I \cdot n + c_T \cdot m. \tag{7}$$

At fixed interview and travel costs and at a fixed net sample size c increases with m. Assuming that c_T is independent of the number (and hence also of the average size) of sampled clusters, think of m_{opt} as a function of n, namely $m_{\text{opt}}^{(n)}$ and substitute it for m in (7). Then

$$c_{m_{\text{opt}}}^{(n)} = c_I \cdot n + c_T \cdot m_{\text{opt}}^{(n)} \tag{8}$$

is a U-shaped function of n as demonstrated in the following figure. Figure 1 illustrates the behavior of the total costs as a function of n assuming values of all other parameters as above and $c_I = 80$ and $c_T = 240$. The dashed line indicates the minimum total costs. Associated with that net sample size is an optimal number of sampled clusters to ensure $n_{\text{eff}} = 1500$ of $m_{\text{opt}}^{(n)} \approx 276$ with an average size of $\bar{b} \approx 8.5$.

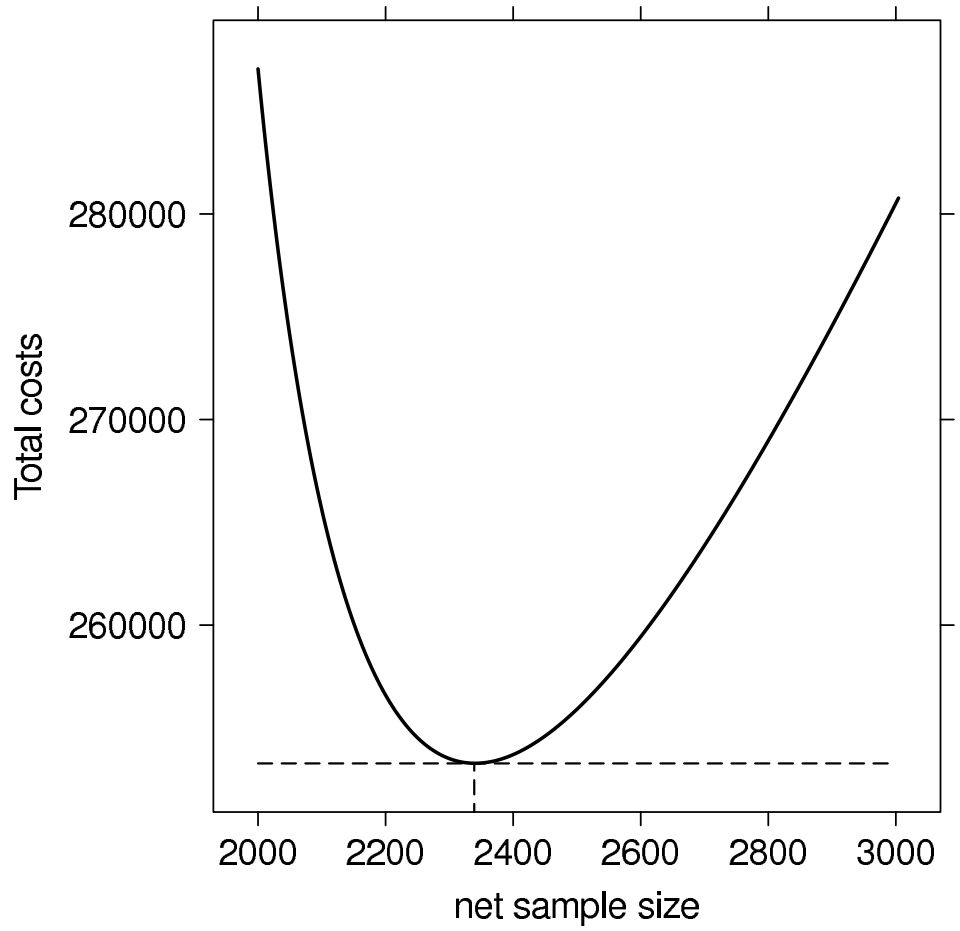

Figure 1: Total costs as a function of net sample size

Generally, the minimum of (8) can be found by setting the first derivative to zero and checking that the the second derivative is larger than zero. Then, we substitute (6) for m^n_{opt} in (8) and get

$$c(n) = \frac{(c_T + c_I) \cdot \mathit{deff}_p \cdot n \cdot n_{\text{eff}} \cdot \rho - c_I \cdot \mathit{deff}_p \cdot n \cdot n_{\text{eff}} + c_I \cdot n^2}{\mathit{deff}_p \cdot n_{\text{eff}} \cdot \rho - \mathit{deff}_p \cdot n_{\text{eff}} + n}. \tag{9}$$

Setting the first derivative to zero

$$\begin{aligned} c'(n) &= c_I + \frac{n_{\text{eff}} \cdot c_T \cdot \mathit{deff}_p \cdot \rho}{n - n_{\text{eff}} \cdot \mathit{deff}_p + n_{\text{eff}} \cdot \mathit{deff}_p \cdot \rho} - \\ &\quad \frac{n_{\text{eff}} \cdot c_T \cdot n \cdot \mathit{deff}_p \cdot \rho}{(n - n_{\text{eff}} \cdot \mathit{deff}_p + n_{\text{eff}} \cdot \mathit{deff}_p \cdot \rho)^2} = 0 \end{aligned}$$

and solving for n gives two solutions

$$\begin{aligned} n_1 &= \frac{c_I \cdot n_{\text{eff}} \cdot \mathit{deff}_p - c_I \cdot n_{\text{eff}} \cdot \rho \cdot \mathit{deff}_p}{c_I} - \\ &\quad \frac{\sqrt{c_I \cdot n_{\text{eff}}^2 \cdot c_T \cdot \mathit{deff}_p^2 \cdot \rho - c_I \cdot n_{\text{eff}}^2 \cdot c_T \cdot \mathit{deff}_p^2 \cdot \rho^2}}{c_I} \\ n_2 &= \frac{c_I \cdot n_{\text{eff}} \cdot \mathit{deff}_p - c_I \cdot n_{\text{eff}} \cdot \rho \cdot \mathit{deff}_p}{c_I} + \\ &\quad \frac{\sqrt{c_I \cdot n_{\text{eff}}^2 \cdot c_T \cdot \mathit{deff}_p^2 \cdot \rho - c_I \cdot n_{\text{eff}}^2 \cdot c_T \cdot \mathit{deff}_p^2 \cdot \rho^2}}{c_I} \end{aligned}$$

of which the second always yields positive values in the second derivative of (9) and can thus be interpreted as the net sample size with minimum total costs. Finally, if we substitute the right hand side of n_2 into

$$b_{\text{opt}} = \frac{n_2}{m_{\text{opt}}} = \frac{n_2}{\frac{n_{\text{eff}} \cdot n \cdot \mathit{deff}_p \cdot \rho}{n - n_{\text{eff}} \cdot \mathit{deff}_p + n_{\text{eff}} \cdot \mathit{deff}_p \cdot \rho}}$$

after some lines of algebra gives

$$b_{\text{opt}} = \sqrt{\frac{c_T\,(1-\rho)}{c_I \cdot \rho}}$$

as (8.3.7) in Kish (1965, 269).

Substituting the values from above into the formula for n_2 gives

$$\begin{aligned} n_{\text{min cost}} &= \frac{80 \cdot 1500 \cdot 1.2 - 80 \cdot 1500 \cdot 0.04 \cdot 1.2}{80} + \\ &\quad \frac{\sqrt{80 \cdot 1500^2 \cdot 240 \cdot 1.2^2 \cdot 0.04 - 80 \cdot 1500^2 \cdot 240 \cdot 1.2^2 \cdot 0.04^2}}{80} \\ &= 2338.94 \approx 2339 \end{aligned}$$

as already indicated graphically by Figure 1 with $b_{\text{opt}} = \sqrt{\frac{240\,(1-0.04)}{80 \cdot 0.04}} \approx 8.5$.

3. Fixed costs

So far, we have taken a quality–based perspective as we regarded total costs as subject to variation. In some situations, however, one is faced with a restricted budget. That leads us to another perspective, namely the cost–optimal view on survey planning. Now, assume c in (7) as given and substitute $m_{\text{opt}}^{(n)}$ for m. Then we have

$$c = c_I \cdot n + c_T \cdot \frac{n_{\text{eff}} \cdot n \cdot \mathit{deff}_p \cdot \rho}{n - n_{\text{eff}} \cdot \mathit{deff}_p + n_{\text{eff}} \cdot \mathit{deff}_p \cdot \rho} \tag{10}$$

Solving this for n gives[1]

$$\begin{aligned} n &= -\frac{1}{2c_I}\Big(-c - c_I \cdot n_{\text{eff}} \cdot d_p + c_I \cdot n_{\text{eff}} \cdot d_p \cdot \rho + n_{\text{eff}} \cdot d_p \cdot \rho \cdot c_T \pm \\ &\quad \sqrt{\left(c + c_I \cdot n_{\text{eff}} \cdot d_p - c_i \cdot n_{\text{eff}} \cdot d_p \cdot \rho - n_{\text{eff}} \cdot d_p \cdot \rho \cdot c_T\right)^2 4c_I\,(c \cdot n_{\text{eff}} \cdot d_p \cdot \rho - c \cdot n_{\text{eff}} \cdot d_p)}\Big) \end{aligned}$$

Substituting the same values as above and assuming 270,000 $ as an upper cost limit of the survey gives $n_1 = 2815.9$ and $n_2 = 2071.1$. With the first solution, there is associated an optimal number of clusters of $m_{\text{opt}}^{(2815.9)} = 186.36 \approx 187$ and hence an average cluster size of 15.1. If we go for the second solution, the optimal number of sampled clusters is $m_{\text{opt}}^{(2071.1)} = 434.6 \approx 435$ with an average cluster size of $\bar{b} \approx 11.1$.

That is how far you can go with a certain amount of additional money exceeding the cost minimum. This may be valuable information as the number of sampled clusters associated with the minimum cost net sample size may not be in every case a practical solution, e.g. the fieldwork institute may be unable to sample exactly that number of clusters, but only a couple more or less.

[1]Note that for typesetting reasons deff_p was substituted by d_p.

4. Discussion

The cost model underlying the current analysis assumes interview and travel costs to be independent of the number of sampled clusters. This may be an unrealistic assumption as travel costs may increase with the number of different locations an interviewer has to visit increases. On the other hand, sampling more clusters from the same frame will, most likely, result in a sample of clusters spatially evenly spread and hence not cause interviewers to go to locations far away from where they would have gone with less but larger clusters.

References

ESS (2005). European social survey, round 3: Specification for participating countries. Specification, European Social Survey.

Gabler, S., Häder, S., and Lahiri, P. (1999). A model based justification of kish's formula for design effects for weighting and clustering. *Survey Methodology*, 25(1):105–106.

Ganninger, M. (2010). *Design Effects: Design-based versus Model-based Approach*. GESIS Series. GESIS, Bonn.

Kish, L. (1965). *Survey Sampling*. John Wiley & Sons.