How to Choose the Number of Call Attempts in a Telephone Survey – in the Presence of Nonresponse and Measurement Errors

Annica Isaksson[1] , Peter Lundquist [2] and Daniel Thorburn [3]

[1] Linköping University, Sweden

[2] Statistics Sweden

[3] Stockholm University, Sweden

Abstract

In telephone surveys, several call attempts are often needed to establish contact with selected individuals. We consider the problem of choosing the maximum number of call attempts to be made to each selected individual before he or she is classified as unavailable. This is an important question since a large number of call attempts makes the data collection costly and time-consuming; a small number of attempts decreases the response set from which conclusions are drawn. We start from the assumptions that the survey goal is to estimate a population total, that the sample can be divided into response homogeneity groups, and that the direct weighting estimator is to be used. We suggest a strategy for choosing the number of call attempts that takes the survey design, non-response, measurement errors and costs of data collection into account. The strategy relies on models for non-response and measurement errors. In our models, the impact of the error sources on the estimator may differ for different numbers of call attempts. According to our strategy, the number of call attempts is decided through a comparison of the estimator's standard error for different numbers of call attempts but the same cost.

Next we study the case when the response homogeneity groups do not work properly but there remains a non-response bias. Our solution relies on models for many different survey errors (the sampling error, measurement errors and errors due to nonresponse) as well as the costs of data collection. We apply our strategy to the Swedish Labour Force Survey. By use of process data, we estimate the bias and variance of a survey estimator as well as the cost for different numbers of call attempts. The estimates refer to the variable annual salary, which is available from a tax register for all sampled individuals. Through comparisons of estimates of average salary based on the full sample (i.e. both respondents and non-respondents) and corresponding estimates based solely on the respondents, the non-response bias is estimated. We further investigate optimal numbers of call attempts for different subsets of the sample.

1 Background

Consider the simple case of a single occasion survey with direct sampling from a frame with good population coverage. We assume that the population consists of H groups with Nh units in each group. The intended sample size is nh.We will use the standard estimator i.e. "the direct weighting estimator"

$$\hat{t}_y^{(A)} = \sum_{r^{(A)}} \frac{y_k}{\pi_k \hat{\theta}_{k|s}^{(A)}}$$

where A is the maximum number of call attempts. The sum is over all responding units, Yk is the answer from person k, π_k is his selection probability and $\hat{}\Theta^{(A)}_{k|s}$ is the observed response proportion in the group containing person k.

We consider three sources of variation: Sample selection design, Response mechanism; contact will be obtained at each call attempt with a certain probability and the result may be an interview or a refusal and Measurement error may, which be different for few or many call attempts. A parameter, gamma, is used to describe whether the measurement variance increases, stays constant or decreases with the number of call attempts. The expressions for the variance and the bias are complicated expressions and will not be given here. Interested readers are referred to Isaksson & al (2008).

2 Cost model

A cost model for a survey may be formulated in many different ways depending on its purpose. For an example of a detailed cost model for a centralized telephone facility, see Groves (1989, section 11.6). Here, we confine ourselves with a much simpler model, which focuses on the number of call attempts. Our model resembles those in Lindström (1991, chapter 6) and Thorburn (2004). Let the total cost of the survey be denoted K. The total cost is composed of a fixed cost C0 , which neither is affected by the sample size nor by the maximal number of call attempts, and a cost which increases with the sample size and with the number of call attempts. The marginal cost may be different in different RHG-groups. In group h the cost for each person consists of

$C_{start,h}$ The average basic cost e.g. for locating the person and finding his phone number. This cost applies to all persons in the sample.

$C_{contact,a,h}$ The average extra cost for trying to contact a person for the a'th time (i.e. after a-1 failures). It contains costs for one more phone call without an answer, talking to persons who are not the correct respondent or speaking to the correct person and booking an interview for another time and similar things. (We will in some places assume that this does not depend on a and write $C_{contact,h}$). The probability that a sampled person has not answered (or refused) after a-1 attempts is denoted by $(1-\omega^{(a-1)})$. The expected value of total costs for all call attempts to a person will then be $\Sigma_1^A (1-\omega^{(a-1)})\, C_{contact,a,h}$, if the maximal number of attempts is A.

$C_{interview,h}$ The average cost for an interview with a person who has already been approached and given his consent to an interview. This cost also contains other costs, which applies only to persons who answer, like editing. The probability of getting an interview at all is $\theta^{(A)}$. The expected cost is thus $\theta^{(A)} C_{interview,h}$, if A is the maximal number of attempts.

We will denote the total expected cost for a person in group h by $C_h^{(A)}$. It is thus

$$C_h^{(A)} = C_{start,h} + \Sigma_1^A (1-\omega^{(a-1)})\, C_{contact,h,a} + \theta^{(A)}\, C_{interview,h}$$

The total expected cost of the survey is thus

$$C_0 + \Sigma_h n_h C_h^{(A)} = n\, \Sigma_h \rho_h C_h^{(A)}$$

where ρ_h is the proportion of the sample taken in group h.

3 Optimisation

3.1 Without a bias

Our next problem is to find the optimal value of A. Assume that the bias is 0 and that we want to get the smallest variance. We first assume that the same A should be used for all groups. If one has allocated a fixed amount K to the study one can see that one can afford to take a sample of size $n(A) = (K\text{-}C0) / \Sigma_h \rho_h C_h^{(A)}$. This value can be inserted into the variance formula

$$Var(\hat{}t) = Var(\Sigma_h \hat{}t_h) = (1/n)\, \Sigma_h N_h^2 V_h^{(A)} / \rho_h$$

(The expression $V_h^{(A)}$ is with our set-up a complicated expression, but can easily be found from Isaksson al (2008). The best choice of A is thus the value that minimizes
$\Sigma_h \rho_h C_h^{(A)} * \Sigma_h N_h^2 V_h^{(A)} / \rho_h$.

One may also try to find different optimal values A_h for the number of call attempts in different groups. In that case one wants to minimize the variance

$Var(\hat{}t) = Var(\Sigma_h \hat{}t_h) = \Sigma_h N_h^2 V_h^{(A_h)} / n_h$ subject to the constraint $K\text{-}C_0 = \Sigma_h n_h C_h^{(A_h)}$

In the same way as above, we find that the optimal A_h is the value which minimizes $(V_h^{(A_h)} C_h^{(A_h)})^{½}$

But in this case one may also see that different sample sizes should be chosen in the groups. A use of Lagrange multiplicators gives that n_h should be proportional to $N_h (V_h^{(A_h)} / C_h^{(A_h)})^{½}$. Inserting this into the constraint gives that the optimal n_h should be

$$N_h (V_h^{(A_h)} / C_h^{(A_h)})^{½} (K\text{-}C_0) / \Sigma_h N_h (C_h^{(A_h)} V_h^{(A_h)})^{½}$$

In other words: A smaller number of call attempts should in some strata be compensated by larger sample sizes.

3.2 Optimising MSE with a bias present

The MSE depends on two parts: a bias component, which does not decrease with sample size and a variance component, which decreases. The bias term will be more and more important as the sample size increases. This means that the choice of A must depend more and more on the influence of A on the bias and its influence on the variance is less important.

One great problem, though, is how to define the bias. For a periodic survey like the Swedish Labour Force Survey (LFS) the comparison between successive months is probably much more important than the absolute level. Thus one should probably be more concerned about the bias in the estimated differences between two months than about the absolute bias at one occasion.

A study like the LFS has many study-variables. The difference in bias behaviour between them is much larger than the behaviour in variance. Thus a choice of A that is optimal for the bias of one study-variable may be far from optimal for other variables. Optimising the number of call attempts for the variance of one variable is probably not so dangerous for other variables.

Another problem is that the bias may not decrease with the number of call attempts. According to our empirical study below there were groups such that the bias changed sign at some A-value (c.f. figure 1). This may have been due the fact that the simplest persons to contact were unoccupied persons who were at home during daytime (e.g. house-wives and sick-pensioners). As the sampling continued more and more people with ordinary works were found and answered to the survey. The last part, with persons who never were reached, was dominated by people with no permanent address, students not living at home, backpackers and so on. Thus a look only at the bias suggested that the sampling should end after only two or three call attempts.

In this section we assume only one stratum and write

$$MSE(A) = N^2\, Bias^2(A) + N^2\, V^{(A)} / n$$

If one has allocated a fixed amount K to the study we can afford to take a sample of size

$$n(A) = (K - C0) / C^{(A)}.$$

If this is inserted into the MSE we get

$$MSE(A) = N^2 \, Bias^2(A) + N^2 \, V^{(A)} \, C^{(A)} / (K\text{-}C0)$$

Thus we should choose the A-value that minimizes

$$Bias^2(A) + V^{(A)} \, C^{(A)} / (K\text{-}C0)$$

Or a form that is more close to the previous one

$$V^{(A)} \, C^{(A)} + Bias^2(A)(K\text{-}C0).$$

It is easy to generalise this to several strata assuming that the cost allocation of the strata is fixed. However, if we also want to reallocate between strata the answer will depend on variance, cost and bias functions in a complicated manner.

4. A case study. The Swedish Labour Force Survey

The target population of the Swedish labour force survey (LFS) consists of all Swedish residents, 15-74 years old. A monthly panel of roughly 21 500 is selected every month according to a rotating scheme so that each selected person is included in altogether eight panels three months apart (The design may here be taken as SRS). The data are collected by telephone.

Our study consists of the response behaviour of all persons selected for an interview in March – Dec 2007, supplemented with their annual salary 2006 according to the Swedish Tax Register (our y). We have also access to data from the process monitoring system WinDati. For each person we know the number of times the interviewer has handled this person. However, all recorded events are not true call attempts. Only about one third are real call attempts. We will call them WD-events. The exact number of call attempts is not possible to obtain.

Figure 1 Relative bias, Monthly averages

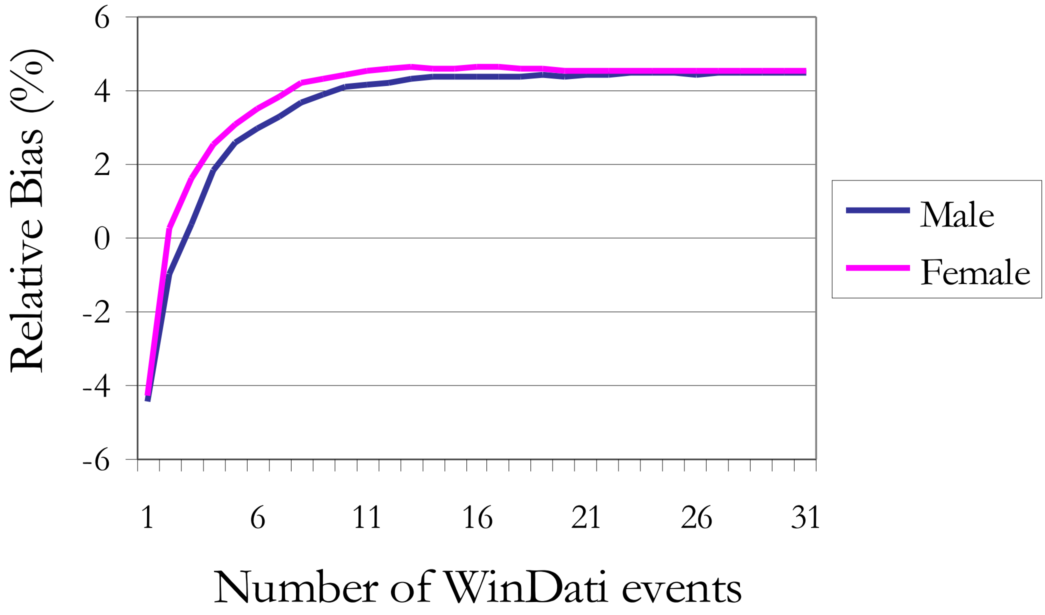


Figure 2 Function determining the optimal number of call attempts

No bias, lowest value is optimum

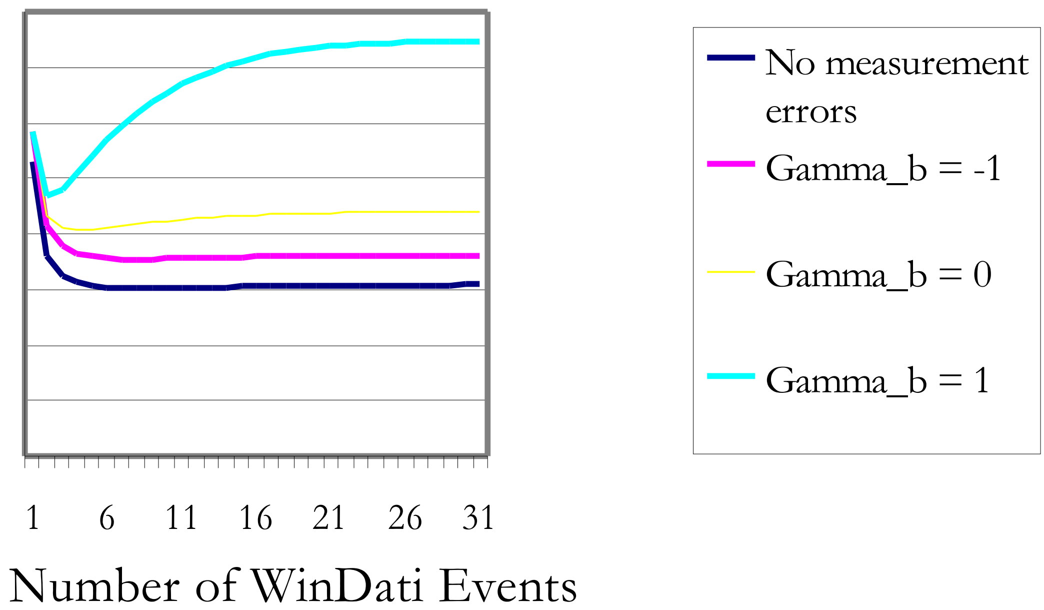
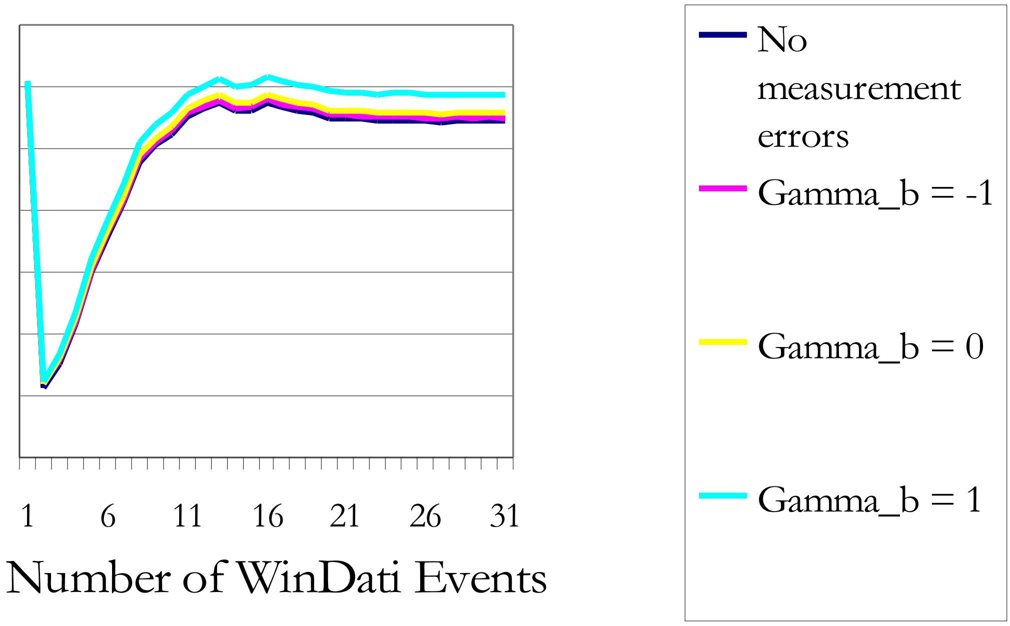


Figure 3 Function determining the optimal number of call attempts

Including raw bias, lowest value is optimum



References

Groves, R. M. (1989). *Survey Errors and Survey Costs.* New York: Wiley.

Isaksson, A, Lundquist, P, Thorburn, D. (2008), *Use of Process data to determine the number of call attempts in a telephone survey*. Research Report, Statistics Sweden.

Lindström, H. L. (1991). Interacting Nonresponse and Response Errors. *R & D Report 1991:3, Statistics Sweden.*

Thorburn, D. (2004). Officiell statistik. *Kurskompendium, Department of Statistics, Stockholm University.*