Optimum sampling fraction among the nonrespondents with nonlinear cost function

Oleksandr Chernyak

Kyiv National Taras Shevchenko University, Ukraine
e-mail: chernyak@univ.kiev.ua

Abstract

Developed the method of definition optimum sampling fraction among the nonrespondents with nonlinear cost function whicn minimizes cost for a fixed value of variance of estimate or minimizes variance for a fixed cost.

1 Introduction

We use the method of double sampling for determinate optimum sampling fraction among the nonrespondents when nonlinear cost function is chosen. By this method after first attempt of sampling random subsample are obtained from nonresponse units. This technique was first developed for survey in which the initial attempt was made by mail, a subsample of persons who did not return the completed questionnaire being approached by the more expensive method of a personal interview. This method is an application of the technique of double sampling for stratification (see Cochran (1977), section 12, Rao (1973)). In this paper which is a continuation of research by Chernyak (1997, 2000, 2001) we give results for nonlinear cost functions.

2 Results

We take a simple random sample of n' units on the first step or phase. Let n_1' be the number of units in the sample who responded, n_2' the number in the nonresponse group; $n' = n_1' + n_2'$. The random subsample $n_2 = \nu_2 n_2'$ of the n_2' are obtained on the second step or phase, $0 < \nu_2 \le 1$. Hancen and Hurwitz (1946) use the notation $\nu_2 = \frac{1}{k}$, so that $n_2 = \frac{n_2'}{k}$.

In the framework of double sampling for stratification there are two strata. Stratum 1 consist of those who would respond to a first attempt, with a measured sample of size $n_1 = n_1'$, so that $\nu_1 = 1$. Stratum 2 consist of those who would respond to the second attempt with $n_2 = \nu_2 n_2'$.

$w_1 = \frac{n_1'}{n'}, w_2 = \frac{n_2'}{n'}$ are the sample proportions in the two strata of the first sample of size n'.

Let W_1 and W_2 are the population proportions or weights in the two strata; $W_1 + W_2 = 1$; μ_k is the true mean in stratum k, $k = 1,2.$ w_k is unbiased estimate of W_k, $Ew_k = W_k, k = 1,2.$

As an estimate the population mean $\mu = \sum_{k=1}^{2} W_k \mu_k$ we take

$$\bar{y}_d = w_1 \bar{y}_1 + w_2 \bar{y}_2 = \frac{(n_1' \bar{y}_1 + n_2' \bar{y}_2)}{n'},$$

where $\bar{y}_1, \bar{y}_2$ are the means of the samples of sizes $n_1 = n_1'$ i $n_2 = \nu_2 n_2'$.

The estimate $\bar{y}_d$ will be unbiased. The variance of this estimate is

$$D\bar{y}_d = S^2\left(\frac{1}{n'} - \frac{1}{N}\right) + \frac{W_2 S_2^2}{n'}\left(\frac{1}{\nu_2} - 1\right),$$

where S^2 - is the population variance, $S_2^{\,2}$ – is the variance in stratum 2:

$S_2^2 = \frac{1}{N_2 - 1}\sum_{i=1}^{N_2}(Y_{2i} - \mu_2)^2$ and $S^2(N-1) = \sum_{k=1}^{2}(N_k - 1)S_k^{\,2} + \sum_{k=1}^{2} N_k(\mu_k - \mu)^2$, N is population size and N_k is stratum size, $k = 1,2$ (see Cochran (1977), section 13.6, Rao (1973)).

The costs of taking sample are: c_0 is a cost of making the first attempt per unit; c_1 is the cost of processing the results from the first attempt per unit; c_2 is the cost of getting and processing the data in the second stratum. Let the cost function is following $C = c_0(n')^{\alpha} + c_1 n_1' + c_2 n_2 = c_0(n')^{\alpha} + c_1 w_1 n' + c_2 w_2 \nu_2 n', \alpha > 0.$

Then the expected cost is $C^* = EC = c_0(n')^{\alpha} + c_1 W_1 n' + c_2 W_2 \nu_2 n'.$

We used the following properties: $n_k' = w_k n'$, $n_k = \nu_k w_k n'$ and $En_k = \nu_k n' \cdot W_k$, $\nu_1 = 1$.

Theorem 1. The variance of the estimated mean $\bar{y}_d$ is a minimum for a specified expected cost C^* and the expected cost is a minimum for a specified variance V when

$$\nu_2 = \frac{S_2}{S_u\sqrt{c_2}} \cdot \sqrt{\alpha c_0 (n')^{\alpha-1} + c_1 W_1}, \tag{1}$$

where $S_u^2 = S^2 - W_2 S_2^2 > 0.$

For the first problem n' is unique positive root of the following equation

$$c_0(n')^{\alpha} + c_1 W_1 n' + \sqrt{c_2} \cdot W_2 n' \frac{S_2}{S_u} \cdot \sqrt{\alpha c_0 (n')^{\alpha-1} + c_1 W_1} - C^* = 0. \tag{2}$$

For the second problem n' is unique positive root of the following equation

$$\frac{S_u^2}{n'} + \frac{\sqrt{c_2} \cdot W_2 \cdot S_2 \cdot S_u}{n'\sqrt{\alpha c_0 (n')^{\alpha-1} + c_1 W_1}} - V_1 = 0, \text{ where } V_1 = V + \frac{S^2}{N}. \tag{3}$$

The Lagrange multipliers method are used to proof the theorem.

Remark 1. If $\alpha = 1$ then the results of theorem 1 coincide with the results of Cochran (1977), section 13.6, p.372.

Theorem 2. If the expected cost function is of the logarithmic form $C^* = c' \ln(n') + c_1 W_1 n' + c_2 W_2 \nu_2 n'$ i $S_u^2 = S^2 - W_2 S_2^2 > 0$, then the variance of the estimated mean $\bar{y}_d$ is a minimum for a specified cost C^* and the expected cost is a minimum for a specified variance V when

$$\nu_2 = \frac{S_2}{S_u\sqrt{c_2}} \cdot \sqrt{\frac{c_0}{n'} + c_1 W_1}. \tag{4}$$

For the first problem n' is unique positive root of the following equation

$$c_0 \ln(n') + c_1 W_1 n' + \sqrt{c_2} \cdot W_2 n' \frac{S_2}{S_u} \cdot \sqrt{\frac{c_0}{n'} + c_1 W_1} - C^* = 0; \tag{5}$$

For the second problem n' is unique positive root of the following equation

$$\frac{S_u^2}{n'} + \frac{\sqrt{c_2} \cdot W_2 \cdot S_2 \cdot S_u}{n'\sqrt{\frac{c_0}{n'} + c_1 W_1}} - V_1 = 0, \text{ where } V_1 = V + \frac{S^2}{N}. \tag{6}$$

The solutions require a knowledge of W_1, W_2, S^2 and S_2^2. These variables can often be estimated from previous experience.

3 Example

This example is condensed from the paper by Hansen and Hurwitz (1946) and the classical book by Cochran (1977). The first sample is taken by mail and the response rate W_1 is expected to be 50 %. The precision desired is that which would be given by a simple random sample of size 1000 if there were no nonresponse. Let the cost of mailing a questionnaire is $c_0 = \$0{,}45$ per unit (cost of fist attempt), the cost of processing the completed questionnaire is $c_1 = \$1$ per unit (cost of processing data for a respondent). To carry out a personal interview cost is $c_2 = \$10$ per unit (cost of obtaining and processing data for a nonrespondent). How many questionnaires should be sent out and what percentage of the nonrespondents should be interviewed? We will also consider the case when the expected cost $C^* = \$1000$ is given for research.

If the variances S^2 and S_2^2 are assumed equal and N is assumed to be large then $S_u^2 = (1-W_2)S^2 = 0{,}5S^2$. Note that we have put $V = S^2/1000$, since this is the variance that the sample mean would have if a sample of 1000 were taken and complete response were obtained.

3.1. Let the expected cost function is of the linear form $C^* = c_0(n') + c_1W_1n' + c_2W_2\nu_2n' = 0{,}95n' + 5\nu_2n';\ W_1 = W_2 = 0{,}5.$

For a specified expected cost $C^* = \$1000$ from (1), $\nu_2 = \dfrac{1}{\sqrt{1-W_2}\sqrt{c_2}} \cdot \sqrt{c_0 + c_1W_1} = 0{,}436$,

and from (2) $n' = \dfrac{\sqrt{1-W_2}}{\sqrt{c_0 + c_1W_1}} \cdot \dfrac{C^*}{(\sqrt{1-W_2}\sqrt{c_0 + c_1W_1} + \cdot\sqrt{c_2} \cdot W_2)} \approx 320$.

Consequently, 320 questionnaires should be mailed, of the 160 that are not returned, we interview a random subsample of $160 \cdot 0{,}436 = 69$. The minimal variance is $V_{\min} = 0{,}0051 \cdot S^2$.

For a specified variance V from (3)

$$n' = \frac{1}{\sqrt{c_0 + c_1W_1}} \cdot \frac{(\sqrt{1-W_2}\sqrt{c_0 + c_1W_1} + \sqrt{c_2} \cdot W_2)}{(\dfrac{1}{1000} + \dfrac{1}{N})} \approx 2328,$$

and $\nu_2 = 0{,}436$.

Consequently, 2328 questionnaires should be mailed, of the 1164 that are not returned, we interview a random subsample of $1164 \cdot 0{,}436 = 508$. The minimal expected cost is \$7291,60.

3.2. Let the expected cost function is of the form

$$C^* = c_0\sqrt{n'} + c_1W_1n' + c_2W_2\nu_2n' = 0{,}45\sqrt{n'} + 0{,}5n' + 5\nu_2n'.$$

For a specified expected cost $C^* = \$1000$ from (2) we have the equation for n'

$$c_0\sqrt{n'} + c_1W_1n' + \sqrt{c_2} \cdot W_2n' \frac{1}{\sqrt{1-W_2}} \cdot \sqrt{\frac{c_0}{2\sqrt{n'}} + c_1W_1} - C^* = 0,$$

than

$$9(n')^2 + 4003{,}69n' - 1{,}8n'\sqrt{n'} + 3600\sqrt{n'} - 4000000 = 0,$$

and unique positive root is $n' = 476$. From (1)

$$\nu_2 = \frac{1}{\sqrt{1-W_2}\sqrt{c_2}} \cdot \sqrt{\frac{c_0}{2\sqrt{n'}} + c_1W_1} = 0{,}317.$$

Consequently, 476 questionnaires should be mailed, of the 238 that are not returned, we interview a random subsample of $238 \cdot 0{,}317 = 76$. The minimal variance is $V_{\min} = 0{,}0044 \cdot S^2$.

For a specified variance n' is unique positive root of the following equation (see (3))

$$\frac{1-W_2}{n'}+\frac{\sqrt{c_2}\cdot W_2\cdot\sqrt{1-W_2}}{n'\sqrt{\frac{c_0}{2\sqrt{n'}}+c_1W_1}}-\frac{1}{1000}=0\text{, or }\frac{0{,}5}{n'}+\frac{\sqrt{5}\cdot 0{,}5}{n'\sqrt{\frac{0{,}45}{2\sqrt{n'}}+0{,}5}}-\frac{1}{1000}=0,$$

and unique positive root of this equation is $n'=2074$, $\nu_2=0{,}317$.

Consequently, 2074 questionnaires should be mailed, of the 1037 that are not returned, we interview a random subsample of $1037\cdot 0{,}317=329$. The minimal expected cost is \$4347,49.

3.3. Let the expected cost function is of the form

$$C^*=c_0\ln(n')+c_1W_1n'+c_2W_2\nu_2n'=0{,}45\ln(n')+0{,}5n'+5\nu_2n'.$$

For a specified expected cost $C^*=\$1000$ from (5) we have the equation for n'

$$c_0\ln(n')+c_1W_1n'+\sqrt{c_2}\cdot W_2n'\frac{1}{\sqrt{1-W_2}}\cdot\sqrt{\frac{c_0}{n'}+c_1W_1}-C^*=0\text{, or}$$

$$0{,}9\cdot\ln(n')+n'+n'\cdot\sqrt{\frac{9}{n'}+10}-2000=0.$$

and unique positive root of this equation is $n'=479$ .From (4)

$$\nu_2=\frac{1}{\sqrt{1-W_2}\sqrt{c_2}}\cdot\sqrt{\frac{c_0}{n'}+c_1W_1}=0{,}316$$

The 479 questionnaires should be mailed, of the 240 that are not returned, we interview a random subsample of $240\cdot 0{,}316=76$. The minimal variance is $V_{\min}=0{,}0043\cdot S^2$.

For a specified variance n' is unique positive root of the following equation (see (6))

$$\frac{1-W_2}{n'}+\frac{\sqrt{c_2}\cdot W_2\cdot\sqrt{1-W_2}}{n'\sqrt{\frac{c_0}{n'}+c_1W_1}}-\frac{1}{1000}=0\text{,or }\frac{0{,}5}{n'}+\frac{\sqrt{5}\cdot 0{,}5}{n'\sqrt{\frac{0{,}45}{n'}+0{,}5}}-\frac{1}{1000}=0,$$

then $(n')^3+999{,}1(n')^2-2250900n'+225000=0$ and unique positive root of this equation is $n'=2080$, $\nu_2=0{,}316$.

The 2080 questionnaires should be mailed, of the 1040 that are not returned, we interview a random subsample of $1040\cdot 0{,}316=329$. The minimal expected cost is \$4329,83.

I. $C=\$1000$					II. $V=0{,}001\cdot S^2$				
Cases	n'	ν_2	n_2	$V_{\min}/S^2$	Cases	n'	ν_2	n_2	$C_{\min}$ (\$)
1	320	0,436	69	0,0051	1	2328	0,436	508	7291,60
2	476	0,317	76	0,0044	2	2074	0,317	329	4347,49
3	479	0,316	76	0,0043	3	2080	0,316	329	4329,83

References

Chernyak O.I. (1997) Allocation problem in two-stage stratified sampling with a nonlinear cost function. *Theory of Stochastic Processes,* **3(19)**, **1-2**, 147-153.

Chernyak O.I., Chornous G.O. (2000) Optimal allocation in stratified sampling with a nonlinear cost function. *Theory of Stochastic Processes,* **6(22)**, **3-4,** 6-17.

Chernyak O.I. (2001) Optimal allocation in stratified and double sampling with a nonlinear cost function . *Journal of Mathematical Sciences*, **103**, **4**, 525-528.

Cochran , W.G. (1977). *Sampling Techniques.* Third edition. John Wiley and Sons, New York.

Hansen M.M., Hurwitz W.N. (1946) The problem of nonresponse in sample surveys. *Journal of American Statistical Association,* **41**, 517-529.

Rao , J.N.K. (1973) On double sampling for stratification and analytical surveys. *Biometrika*, **60, 1**, 125-133.