On Internet survey sampling

Janusz Wywiał

University of Economics, Department of Statistics, Bogucicka 14, 40-226 Katowice, Poland, email: wywial@ae.katowice.pl

1. Models of an Internet survey

A population: $U = (1, 2, ..., N)$ of Internet respondent.

The sample: $s = [s_1\ s_2 ... s_N]$.

If a $k-th$ population element is in the sample, $s_k = 1$ (a respondent return a questionnaire by email).

If a $k-th$ popul. element is not in the sample, $s_k = 0$ (a respondent do not return a questionnaire by email).

The elements of the random sample S are independent, (respondents return the questionnaire independently)

Response probabilities for $k = 1, ..., N$:

$$P(S_k = 1) = \pi, \qquad P(S_k = 0) = 1 - \pi$$

The Bernoulli sampling design:

$$P(s) = \prod_{k=1}^{N} \pi^{s_k}(1-\pi)^{1-s_k} \quad \text{for} \quad s \in \boldsymbol{S}.$$

The sample size $n(s)$ is not fixed, $0 \leq n(s) \leq N$.

The Poisson sampling design (more realistic):

$$P(s) = \prod_{k=1}^{N} \pi_k^{s_k}(1-\pi_k)^{1-s_k} \qquad \text{for} \qquad s \in \boldsymbol{S}.$$

Response probabilities:

where: $P(S_k = 1) = \pi_k,$

$$P(S_k = 0) = 1 - \pi_k, \qquad k = 1, ..., N.$$

2. Estimation under the Bernoulli model

The purpose is estimation of the total or mean value:

$$y = \sum_{k=1}^{N} y_k, \qquad \bar{y} = \frac{1}{N} \sum_{k=1}^{N} y_k.$$

2.1. Basic results

The unbiased estimators of the mean:

$$\bar{y}_{HTS} = \frac{1}{N\pi} \sum_{k=1}^{N} y_k S_k = \frac{1}{N\pi} \sum_{k \in S} y_k. \tag{1}$$

$$D^2 \left(\bar{y}_{HTS} \right) = \frac{1-\pi}{N^2 \pi} \sum_{k=1}^{N} y_k^2. \tag{2}$$

The unbiased est. of the variance (see Tillé (2006)):

$$D_S^2 \left(\bar{y}_{HTS} \right) = \frac{1-\pi}{N^2 \pi^2} \sum_{k=1}^{N} y_k^2 S_k. \tag{3}$$

If the probability π is estimated by the sample fraction $\pi_S = \frac{n(S)}{N}$, the approximately unbiased estimator is:

$$\bar{y}_S = \frac{1}{n(S)} \sum_{k=1}^{N} y_k S_k = \frac{1}{n(S)} \sum_{k \in S} y_k. \tag{4}$$

$$D^2 \left(\bar{y}_S \right) \approx \frac{1-\pi}{N\pi} v_{yy} \tag{5}$$

where $v_{yy} = \frac{1}{N} \sum_{k=1}^{N} (y_k - \bar{y})^2$.

The estimator of the variance is:

$$D_S^2 \left(\bar{y}_S \right) = \frac{N - n(S)}{N n(S)} v_{yyS} \tag{6}$$

where $v_{yyS} = \frac{1}{n(S)-1} \sum_{k \in S} (y_k - \bar{y}_S)^2$.

If an auxiliary variable's values x are observed in the population U, the ratio and regression estimators can usful:

$$\bar{y}_{ratio,S} = \bar{y}_S \frac{\bar{x}}{\bar{x}_S}$$

$$\bar{y}_{regS} = \bar{y}_S + B_S \left(\bar{x} - \bar{x}_S \right), \quad B_S = \frac{\sum_{k \in S}^{N} (x_k - \bar{x}_S) y_k}{\sum_{k=1}^{N} (x_k - \bar{x}_S)^2}$$

2.2. Stratified sample

Response Homogeneity Group Model.

N_h is the size of a stratum: U_h and $h = 1, ..., H$.

The fraction $w_h = N_h/N$.

π_h is the response probability for an $h - th$ stratum.

Z_h is the Bernoulli sample of size $n(Z_h)$, $Z_h \subseteq U_h$.

The sample: $S = \bigcup_{h=1}^{H} Z_h$.

The sample size: $n(S) = \sum_{h=1}^{H} n(Z_h)$.

The approximately unbiased estimator of a popul. mean:

$$\tilde{y}_S = \sum_{h=1}^{H} w_h \bar{y}_{Z_h} \tag{7}$$

where

$$\bar{y}_{Z_h} = \frac{1}{n(Z_h)} \sum_{k \in Z_h} y_k, \tag{8}$$

$$D^2\left(\tilde{y}_S\right) = \sum_{h=1}^{H} w_h^2 D^2\left(\bar{y}_{Z_h}\right) \tag{9}$$

where

$$D^2\left(\bar{y}_{Z_h}\right) \approx \frac{1-\pi_h}{N_h \pi_h} v_{hyy} \tag{10}$$

$$v_{hyy} = \frac{1}{N_h} \sum_{k=1}^{N_h} (y_k - \bar{y}_h)^2.$$

The estimator of the variance :

$$D_S^2\left(\tilde{y}_S\right) = \sum_{h=1}^{H} w_h^2 D_{Z_h}^2\left(\bar{y}_{Z_h}\right) \tag{11}$$

$$D_{Z_h}^2\left(\bar{y}_{Z_h}\right) = \frac{N - n(Z_h)}{N n(Z_h)} v_{yyZ_h} \tag{12}$$

$$v_{yyZ_h} = \frac{1}{n(Z_h) - 1} \sum_{k \in Z_h} (y_k - \bar{y}_{Z_h})^2.$$

In the case of the stratification of the sample after its selection, the above results are valid, too.

3. Estimation under the Poisson model

3.1. The case of known response probabilities

The unbiased estimator of $\bar{y}$:

$$\bar{y}_{HTPS} = \frac{1}{N}\sum_{k=1}^{N}\frac{y_k S_k}{\pi_k} = \frac{1}{N}\sum_{k\in S}\frac{y_k}{\pi_k}. \qquad (13)$$

$$D^2\left(\bar{y}_{HTPS}\right) = \frac{1}{N^2}\sum_{k=1}^{N}\frac{y_k^2(1-\pi_k)}{\pi_k}. \qquad (14)$$

The unbiased est. of the variance (see Tillé (2006)):

$$D_S^2\left(\bar{y}_{HTPS}\right) = \frac{1}{N^2}\sum_{k=1}^{N}\frac{y_k^2(1-\pi_k)}{\pi_k^2}S_k. \qquad (15)$$

The response probabilities should be estimated.

3.2. Logit approx. of response probabilities

$\mathbf{x}_k$ is the row vector of m-auxiliary variables values attached to a k-the population element, $k = 1, ..., N$.

The logit model of the probabilities:

$$\pi_k \approx \frac{1}{1 + exp\{-q_k\}}, \qquad k = 1, ..., N \tag{16}$$

where

$$q_k = \mathbf{x}_k \boldsymbol{\beta}. \tag{17}$$

and $\boldsymbol{\beta}$ is the column vector of parameters.

The estimator (the maximum likelihood method):

$$\hat{\pi}_{Sk} = \frac{1}{1 + exp\{-\hat{q}_{Sk}\}}, \qquad q_{Sk} = \mathbf{x}_k \boldsymbol{\beta}_S. \tag{18}$$

The estimator of the $\bar{y}$:

$$\bar{y}_{logS} = \frac{1}{N} \sum_{k \in S} \frac{y_k}{\hat{\pi}_{Sk}}. \tag{19}$$

The approximate variance:

$$D^2\left(\bar{y}_{logS}\right) \approx D^2\left(\bar{y}_{HTPS}\right) + \\ + \frac{3}{N^2}\sum_{t=1}^{N} y_t(1-\pi_k)\mathbf{x}_t\mathbf{H}_{log}^{-1}\sum_{k=1}^{N} y_k(1-\pi_k)\mathbf{x}_k^T$$

where $D^2\left(\bar{y}_{HTPS}\right)$ is given above and

$$\mathbf{H}_{log} = \sum_{k=1}^{N}\pi_k(1-\pi_k)\mathbf{x}_k^T\mathbf{x}_k,$$

The estimator of the above variance:

$$D_S^2\left(\bar{y}_{logS}\right) = \frac{1}{N^2}\Bigg(\sum_{k=1}^{N}\frac{y_k^2}{\hat{\pi}_{Sk}^2}(1-\hat{\pi}_{Sk})S_k + \\ +3\sum_{t=1}^{N}\frac{y_t(1-\hat{\pi}_{St})S_t}{\hat{\pi}_{St}}\mathbf{x}_t\mathbf{H}_{log,S}^{-1}\sum_{k=1}^{N}\frac{y_k(1-\hat{\pi}_{Sk})S_k}{\hat{\pi}_{Sk}}\mathbf{x}_k^T\Bigg)$$

$$\mathbf{H}_{log,S} = \sum_{k=1}^{N}\hat{\pi}_{Sk}(1-\hat{\pi}_{Sk})\mathbf{x}_k^T\mathbf{x}_k,$$

3.3. Probit approx. of response probabilities

$F(q)$ is the distribution function of the standard normal random variable.

The probit model of the response probabilities and theirs estimators (the maximum likelihood method):

$$\pi_k = F(q_k), \qquad \tilde{q}_k = \mathbf{x}_k \boldsymbol{\beta}, \qquad k = 1, ..., N.$$

$$\tilde{\pi}_{Sk} = F(\tilde{q}_{Sk}), \qquad \tilde{q}_{Sk} = \mathbf{x}_k \tilde{\boldsymbol{\beta}}_S.$$

The estimator of the mean $\bar{y}$:

$$\bar{y}_{probS} = \frac{1}{N} \sum_{k \in S} \frac{y_k}{\tilde{\pi}_{S_k}}. \tag{20}$$

$$D^2\left(\bar{y}_{probS}\right) \approx D^2\left(\bar{y}_{HTPS}\right) +$$

$$+ \frac{3}{N^2} \sum_{k=1}^{N} \frac{y_k f_q(q_k)}{\pi_k} \mathbf{x}_k \mathbf{H}_{prob}^{-1} \sum_{k=1}^{N} \frac{y_k f_q(q_k)}{\pi_k} \mathbf{x}_k^T$$

$$\mathbf{H}_{prob} = \sum_{k=1}^{N} \frac{f^2(q_k)}{F(q_k)(1 - F(q_k))} \mathbf{x}_k^T \mathbf{x}_k.$$

The estimator of the variance:

$$D_S^2\left(\bar{y}_{probS}\right) = \frac{1}{N^2}\Bigg(\sum_{k=1}^{N}\frac{y_k^2}{\tilde{\pi}_{Sk}^2}(1-\tilde{\pi}_k)S_k+$$

$$+3\sum_{k=1}^{N}\frac{y_k f_q(\tilde{q}_k)S_k}{\tilde{\pi}_k^2}\mathbf{x}_k\mathbf{H}_{prob}^{-1}\sum_{k=1}^{N}\frac{y_k f_q(\tilde{q}_k)S_k}{\tilde{\pi}_k^2}\mathbf{x}_k^T\Bigg)$$

$$\mathbf{H}_{prob,S} = \sum_{k=1}^{N}\frac{f^2(\tilde{q}_{Sk})}{F(\tilde{q}_{Sk})(1-F(\tilde{q}_{Sk}))}\mathbf{x}_k^T\mathbf{x}_k.$$

4. Example

Some expenditure is a variable under study y.

The age of respondents is an auxiliary variable x.

A population of size $N = 30$.

The sample size 26.

The estimated probit model: $\tilde{\pi}_k = F(-0,11x_k+3,43)$.

The estimated logit model: $\hat{\pi}_k = (1+e^{0,21x_k-6,15})^{-1}$.

The population divided into 4 strata according to x.

$\tilde{y}_s = 441,90,$

$\bar{y}_{probs} = 442,11$

$\bar{y}_{logs} = 442,08.$

Table 1:

x	N_h	$n(S_h)$	π_{Z_k}	$\hat{\pi}_{Sk}$	$\tilde{\pi}_{Sk}$
19	10	9	0,900	0,896	0,897
20	8	7	0,875	0,875	0,875
21	6	5	0,833	0,851	0,850
22	6	5	0,833	0,822	0,822

$D_s^2\left(\tilde{y}_S\right) = 205,67.$

$D_s^2\left(\tilde{y}_{logS}\right) = 1734,22$

$D_s^2\left(\tilde{y}_{probS}\right) = 1728,55.$

Bibliography

Bethlehem J.G. (1988). Reduction of nonresponse bias through regression estimation. *J. Official Stat.*, vol. 4, no. 3, pp. 251-260.

Chow (1983) G. C. *Econometrics*. Mc Graw Hill Book Company, New York.

Ekholm A., Laaksonen S. (1991). Weighting via response modelling in the Finish Household Budget Survey. *J. Official Stat.*, vol. 7, no. 3, pp. 325-338.

Horvitz, D. G., Thompson, D. J. (1952). A generalization of the sampling without replacement from finite universe. *JASA*, vol. 47, pp. 663-685.

Kelley C.T. (2003). *Solving Nonlinear Equations with Newton's no 1 in Fundamentals of Algorithms*. SIAM, Philadelphia.

Särndal C. E., B. Swensson, J. Wretman (1992): *Model Assisted Survey Sampling*. Springer Verlag, New York.

Tillé Y. (2006): *Sampling Algorithms*. Springer, New York.

WywiałJ. L. (2009). On Application of non-response model in internet survey sampling. Studia Ekonomiczne-Zeszyty Naukowe, 53, 19-34.