PSC Working Paper Series

12-1-2009

Double Sample to Minimize Bias Due to Non-response in a Mail Survey

Herbert L. Smith
University of Pennsylvania, hsmith@pop.upenn.edu

Smith, Herbert L. 2009. "A Double Sample to Minimize Bias Due to Non-response in a Mail Survey." Philadelphia, PA: Population Studies Center, University of Pennsylvania. PSC Working Paper Series, No. 09-05. [English version of Smith 2008.]

Double Sample to Minimize Bias Due to Non-response in a Mail Survey

Abstract

A large study of nurses conducted in the U.S. states of California (CA) and Pennsylvania (PA) is based on two large samples: n^CA≈100,000 and n^PA≈65,000. The study was conducted by mail and had response rates of: p^CA=.27 and p^PA=.39 ;; the number of respondents is thus, respectively, : n_1^CA≈28,000 and n_1^PA≈25,000. Although there are many respondents, we must concern ourselves with the possibility of substantial bias due to non-response. In order to estimate and correct for this bias, a second random sample (n_01=1,300 in the two states combined) was drawn from among the non-respondents to the first survey. Thanks to financial incentives and, above all, a shorter questionnaire, we obtained a response rate above 90%. In each state, the two samples were combined to create a virtually unbiased double sample.

Keywords

Biases, California, Double sample, Efficiency, Errors, Mail surveys, Non-response, Nursing, Pennsylvania, Random sample, Sample design, Sample surveys, Sampling, Statistical methods, Statistics, Survey Data, Survey methodology, Surveys

Disciplines

Applied Statistics | Demography, Population, and Ecology | Quantitative, Qualitative, Comparative, and Historical Methodologies | Sociology | Statistics and Probability

Comments

Smith, Herbert L. 2009. "A Double Sample to Minimize Bias Due to Non-response in a Mail Survey." Philadelphia, PA: Population Studies Center, University of Pennsylvania. PSC Working Paper Series, No. 09-05. [English version of Smith 2008.]

A Double Sample to Minimize Bias Due to Non-response in a Mail Survey

Herbert L. Smith
Population Studies Center
University of Pennsylvania
Philadelphia, PA 19104-6298 USA
hsmith@pop.upenn.edu

This is the English-language translation of the original version, which appears as:

Smith, Herbert L. 2008. « Un échantillon double pour minimiser le biais dû à la non réponse dans un sondage par courrier ». Pp. 334-339 in *Méthodes de sondage : applications aux enquêtes longitudinales, à la santé, aux enquêtes électorales et aux enquêtes dans les pays en développement*, édité par Anne Ruiz-Gazen, Philippe Guilbert, David Haziza, and Yves Tillé. Paris: Dunod.

In several places in this version words or phrases are added to clarify points that may have been too clipped in the original, and an error in notation has been rectified.

This research was conducted with support from NINR R01-NR004513, "Outcomes of Nurse Practice" (Linda H. Aiken, PI). The second ("double") sample was carried out under a TRIO award from the NICHD R24-HD044964, "Population Research Center Grant," (Herbert L. Smith, PI). Thanks to Robyn Cheung and Jeannie Cimiotti for their organization of the follow-up study, to Laurent Toulemon for his comments on the paper's original (French) version, and to David Haziza for that paper's editing and review.

A Double Sample to Minimize Bias Due to Non-response in a Mail Survey

Abstract

A large study of nurses conducted in the U.S. states of California (CA) and Pennsylvania (PA) is based on two large samples: $n^{CA} \approx 100{,}000$ and $n^{PA} \approx 65{,}000$. The study was conducted by mail and had response rates of: $p^{CA} = .27$ and $p^{PA} = .39$;; the number of respondents is thus, respectively, : $n_1^{CA} \approx 28{,}000$ and $n_1^{PA} \approx 25{,}000$. Although there are many respondents, we must concern ourselves with the possibility of substantial bias due to non-response. In order to estimate and correct for this bias, a second random sample ($n_{01} = 1{,}300$ in the two states combined) was drawn from among the non-respondents to the first survey. Thanks to financial incentives and, above all, a shorter questionnaire, we obtained a response rate above 90%. In each state, the two samples were combined to create a virtually unbiased double sample.

A Double Sample to Minimize Bias Due to Non-response in a Mail Survey

1. Introduction

The idea of a double sample dates to Hansen and Hurwitz (1946), who showed that a second random sample, drawn from among the non-respondents to the initial sample, could be combined with the first sample in order to create unbiased estimators of means, even though the survey did not initially wind up with a 100% response rate. This method is used only rarely, for several reasons, one of which is that the same factors that lead to a non-response in the initial survey make it difficult to obtain a better response rate among the sample of individuals who already refused to respond. There is also a tendency to dedicate all available resources toward the largest possible sample size.

The nurses' study took place in a tense climate among nurses, hospitals, and politicians, especially in California, where there was a new law — much contested — that had fixed the number of nurses per patient. One objective of this study was to understand the organization of work in the various hospitals from the standpoint of the nurses working inside each one. But the "natural" sampling design, a two-stage survey, first of hospitals, then of the nurses in each of the hospitals selected, was abandoned out of fear that certain hospitals would refuse to participate, and that this self-selection of hospitals, connected to the phenomenon under study, would create more bias than a large single-stage survey of nurses, in which the nurses could report their workplace so that their responses could be aggregated to the hospital level. Thus the hospitals could not refuse to participate (through their nurses) in the study. On the other hand, since we could not know beforehand which nurse was working in which hospital, a very large initial sample was necessary, conducted by mail, with a low response rate as a result.

2. Relative Efficiency of a Double Sample

Unbiased estimators in a double sample for the mean and associated standard error of a variable Y are well known (Glynn, Laird, and Rubin 1993). They exist under the assumption that all of initial non-respondents who are drawn in the second sample respond on that occasion. If this assumption is not strictly true, but the response rate in the second sample is very high (such as ours, at 91%), then these are "nearly unbiased" estimators (Levy and Lemeshow 1999).

The mean $\overline{y}^C$ (C for "classic") is $\overline{y}^C = (n_1\overline{y}_1 + n_0\overline{y}_{01})/n$. The estimated standard error, under the assumption of an infinite population, is

$$SE(\overline{y}^C) = \sqrt{\frac{n_1}{n^2}s_1^2 + \frac{n_0^2}{n^2 n_{01}}s_{01}^2 + \frac{n_0 n_1}{n^2}(\overline{y}_{01} - \overline{y}_1)^2} \ ,$$

where s_1 and s_{01} are the standard deviations in, respectively, the first and second samples. When there is only a single sample (and one does not know the possible bias due to non-response), the estimators of the mean and its standard error are $\overline{y}_1$ and $SE(\overline{y}_1) = \sqrt{s_1^2/n_1}$,. The variance ratio ($K$) for a mean calculated with respect to a double sample as against that from a single sample is the following:

$$Var(\overline{y}^C)/Var(\overline{y}_1) = p^2 + p(1-p)f^{-1}\left(\frac{s_{01}^2}{s_1^2}\right) + p^2(1-p)\left(\frac{\overline{y}_{01} - \overline{y}_1}{s_1}\right)^2 = K$$

with $p = n_1/n$ the proportion of the first sample that responded and $f = n_{01}/n_0$ the proportion of non-respondents to the first sample selected for the second sample. Under the suppositions $s_{01}^2 \approx s_1^2$ (the variances in the two samples are more or less equal) and $s_1 \gg \overline{y}_{01} - \overline{y}_1$[1] (the within-sample variance is larger than the mean difference between the two samples), the variance ratio becomes $p^2 + p(1-p)f^{-1}$. The relative efficiency $K = 1$ occurs when $f = p\,(1-p)/(1-p^2)$.

Figure 1 shows, for several values of $\sqrt{K}$, the ratio of standard errors, the proportion of the original sample n required for various initial rates of response p. The ordinate of these curves is $g = f(1-p)$, the proportion of the original sample that will eventually need to be re-interviewed as the second sample. The abscissa is the initial response rate, p, along a scale that is reversed, going from 1 (on the left) to 0 (on the right). The solid, thickest curve corresponds to $K = 1$, where the "classic" estimator of a double-sample mean has the same variance (or standard error) as that which would be calculated with an initial collection of respondents of size $n_1 = np$. As the response rate declines, the number of respondents required in the second sample to maintain the same *relative* efficiency increases up to $p = .414$, at which point it *diminishes*; fewer and fewer respondents in the second stage are required to maintain the relative efficiency as the response rate declines. It's the same with the other curves,

[1] This inequality appears (incorrectly) as $s_1 \gg \overline{y}_{01} - \overline{y}_0$ in the French (original) publication.

corresponding to values of $K > 1$, except that the inflection point, $\max(p|K)$ drops, toward an asymptote at 1/3. It seems a bit odd to think that a worse response rate $p < \max(p|K)$ can require a smaller sample size in the second sample to maintain the same sampling error, but this is a relative equality. The absolute efficiency diminishes as a function of $p^{-0.5}$, as indicated by the dashed line in Figure 1. In effect, if we consider p, the initial response rate, to be fixed, and $g = n_{01}/n$ as well: We do a survey and get a rate of response. Persuading the non-respondents to participate in the second stage is not easy, especially when we have to obtain a response from *all* of the subjects drawn in the second sample. It's a matter of asking ourselves, "With a little more effort (in terms of time and money), how much would the efficiency of the estimator improve with a little larger second sample n_{01}?" Since costs tendency to be linear with respect to n_{01}, we benefit the most when outside the region $.41 > p > .33$, because it is there that K is maximally dispersed, hence increases in n_{01} least efficient.

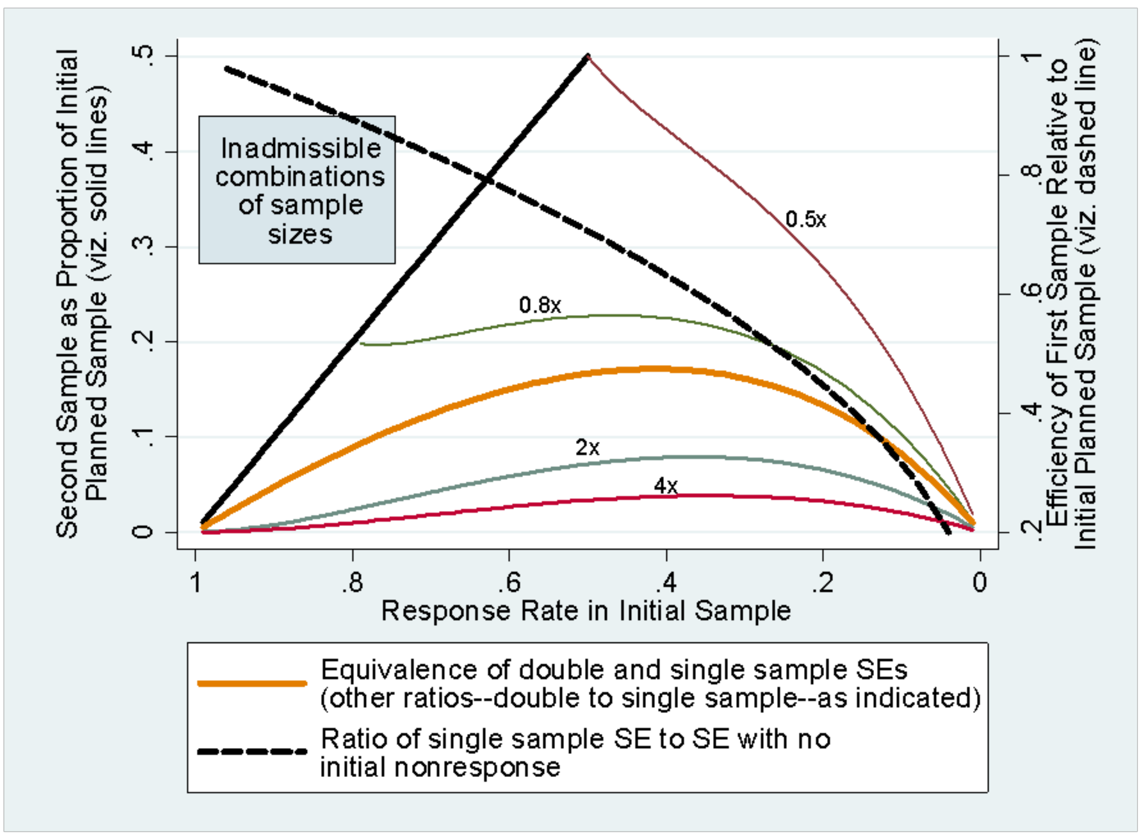

Figure 1. Relative Efficiency of a Double Sample

Even if n_{01}, hence g, are small, the double sample estimator is to be preferred, because sampling error is not the only source of survey error. We can take into account bias via

$MSE(y_c) = Var(y_c) + Bias^2$, the mean squared error of the estimator y_c. The "classic" estimator for the double sample $\overline{y}^C$ is unbiased and its mean squared error is a function of its variance: $MSE(\overline{y}^C) = Var(\overline{y}^C) = K \cdot Var(\overline{y}_1)$. But the estimator based only on the first sample is probably biased: $Bias(\overline{y}_1) = \overline{y}_1 - \overline{y}^C = \overline{y}_1 - [p\overline{y}_1 + (1-p)\overline{y}_{01}] = (1-p)(\overline{y}_1 - \overline{y}_{01})$. Thus its mean square error is $MSE(\overline{y}_1) = (1-p)^2(\overline{y}_1 - \overline{y}_{01})^2 + Var(\overline{y}_1)$. This means that the estimator for the double sample is a better estimator of the "true value" (Kish 1995, p. 9) of the mean, even in the case where its variance is greater than that of the first sample, because it does not suffer from bias due to non-response. And for $RMSE$, the root mean squared error, we find that

$$|\overline{y}_1 - \overline{y}_{01}| > \frac{\sqrt{K-1}}{1-p} \times \frac{s_1}{\sqrt{n_1}} \Rightarrow RMSE(\overline{y}^C) < RMSE(\overline{y}_1)$$

3. Application to the Survey of Nurses

Table 1 presents some results. The most interesting measure is in the last column, the ratio of the two root mean square errors: the total error of the double sample compared to that of the very large initial single sample, with its equally large non-response rate. When the ratio is less than unity (1), the double sample estimator is more efficient; the bias in the initial survey (unknown in the absence of the second survey) is large enough as not to be compensated for by its small sampling error.

On the other hand, when the ratio is larger than 1, the bias is sufficiently small that the estimator resulting from the first sample is preferable — after the fact — to the double sample estimator. With second samples of a size less than 1% that of the initial survey and response rates in the neighborhood of 25-40%, ratios of 4 and 5 show that the second sample is not always efficient. In particular, for the evaluations of the hospitals, the second sample of non-respondents does not improve the estimates, since it turns out that there was no difference in these evaluations between the nurses who responded in the first place and those who refused to respond initially. It does not matter that this second set of respondents is different with respect to demographic characteristics (sex, race, national origin) — the type of information found, for example, in a sampling frame, which is used to "inform" the weighting scheme connected to the method of post-stratification.

Of course, one does not know these things in the absence of the second survey of the non-respondents. But after having found them out, must one present the standard errors obtained via the estimator for the double sample? In the absence of bias, this is very costly: We would have preferred to know beforehand if this bias existed or not!

Table 1. Means and Root Mean Squared Error by Sample: First (1), Second (2), and Double Sample Estimator (Double)

Questionnaire Item	State	Mean			RMSE		
		1 ($\bar{y}_1$)	2 ($\bar{y}_{01}$)	Double ($\bar{y}^C$)	1	Double	Ratio
Demographic items (CA: $n_1 \approx 28{,}000$, $n_{01} \approx 525$; PA: $n_1 \approx 25{,}000$, $n_{01} \approx 580$)							
Gender (Proportion male)	CA	.060	.106	.094	.034	.010	0.29
	PA	.040	.077	.062	.024	.007	0.28
Race (Proportion white)	CA	.714	.650	.667	.047	.016	0.32
	PA	.945	.921	.930	.015	.007	0.46
National origin (Proportion Filipino)	CA	.100	.158	.142	.042	.142	0.27
	PA	.004	.010	.008	.004	.003	0.70
Education (Proportion with bachelor's degree or more)	CA	.599	.527	.546	.053	.016	0.30
	PA	.469	.465	.466	.004	.013	3.27
Work in patient care (Proportion staff nurse)	CA	.588	.631	.620	.032	.015	0.48
	PA	.560	.585	.575	.015	.012	0.81
Evaluation of job (CA: $n_1 \approx 23{,}000$, $n_{01} \approx 440$; PA: $n_1 \approx 19{,}600$, $n_{01} \approx 490$; $s \approx 0.80$)							
How satisfied are you with your job? (1=Very to 4)	CA	1.83	1.79	1.80	.028	.028	1.00
	PA	1.98	2.00	1.99	.012	.024	1.93
Feelings (CA: $n_1 \approx 22{,}000$, $n_{01} \approx 440$; PA: $n_1 \approx 18{,}500$, $n_{01} \approx 490$; $s \approx 1.75$)							
I feel used up at the end of the work day (1=Never to 7)	CA	4.19	4.04	4.08	.117	.062	0.53
	PA	4.33	4.37	4.36	.027	.050	1.84
I feel burned-out from my work (1=Never to 7)	CA	3.11	3.11	3.11	.012	.063	5.19
	PA	3.24	3.37	3.33	.082	.054	0.65
Desirable characteristics of hospital (CA: $n_1 = 15{,}914$, $n_{01} = 342$; PA: $n_1 = 13{,}430$, $n_{01} = 371$; $s \approx 0.60$)							
14-item scale (1=Strongly agree to 4)	CA	2.18	2.18	2.18	.006	.027	4.31
	PA	2.33	2.33	2.33	.005	.022	4.15

References

Glynn, R.J., N.M. Laird, and D.B. Rubin. 1993. "Multiple Imputation in Mixture Models for Nonignorable Nonresponse with Follow-ups." *Journal of the American Statistical Association* **88**:984-993.

Hansen, M.H., and W.N. Hurwitz. 1946. "The Problem of Non-Response in Sample Surveys.", *Journal of the American Statistical Association* **4**:517-529.

Kish, L. 1995 [1965]. *Survey Sampling*. New York: Wiley.

Levy, P.S., and S. Lemeshow. 1999. *Sampling of Populations: Methods and Applications*. New York: Wiley.