

Regression type strategy for randomized response

Strategia tipo regressione per le risposte casualizzate

Giancarlo Diana ¹

Dipartimento di Scienze Statistiche,
Università di Padova, e-mail: giancarlo.diana@unipd.it

Pier Francesco Perri ²

Dipartimento di Economia e Statistica,
Università della Calabria, e-mail: pierfrancesco.perri@unical.it

Riassunto: I metodi basati sulle *risposte casualizzate* vengono impiegati in indagini su argomenti particolarmente sensibili ed imbarazzanti per ridurre i tassi di non risposta e le distorsioni nelle risposte. Nel lavoro viene proposto un nuovo stimatore per la proporzione di individui che presentano una caratteristica sensibile. Lo stimatore è ottenuto modificando quello originariamente proposto da Warner (1965) ed è utilizzabile se è possibile rilevare una variabile ausiliaria non sensibile, di media nota.

Keywords: Proportion, sensitive question, regression estimator

1. Introduction

In surveys concerning sensitive questions, such as gambling, alcoholism, sexual behavior, drug taking, tax evasion, illegal income, and many others, direct techniques for collecting information may induce interviewed people to refuse to answer or give untruthful or misleading responses. To reduce rates of nonrespondents and biased response arising from sensitive, embarrassing, threatening, or even incriminating questions, some special statistical techniques, which ensure interviewee's anonymity or, at least, a high degree of confidentiality to the respondents, may be employed. Different ingenious techniques have been devised for dealing with sensitive items. Such techniques are known as *Randomized Response Methods*. In general, these methods involve the use of a randomization device, such as dice or a deck of cards, rather than a direct response to collect reliable information on sensitive issues. Depending on the result produced by the randomized device, the interviewee gives an answer concerning his/her true status. Because the interviewer is unaware of the result of the randomized device, the use of these methods ensures that an individual respondent cannot be identified on the basis of his/her answer.

Since the pioneering work of Warner (1965), many authors have dealt with the problem of estimating the proportion (π) of people with a sensitive attribute and various methods have been suggested for making the technique more and more efficient. Recently, Zaizai (2006) improved the Warner's estimator making use of an auxiliary variable through the ratio estimator. Motivated by this work, we suggest a new estimator of π which is always more efficient than the estimators suggested by Warner (1965) and Zaizai (2006).

¹Address of correspondence: Via Cesare Battisti, 241, 35121 Padova

²Address of correspondence: Via P. Bucci, Cubo 0C, 87036 Arcavacata di Rende (CS)

2. The new estimator

Let $P = \{1, 2, \dots, N\}$ be a finite population and Y a variable taking value $Y_i = 1$ if the i -th person possesses the sensitive characteristic, $Y_i = 0$ otherwise, $i = 1, 2, \dots, N$. Let X be a not sensitive auxiliary variable, easy to collect, seemingly neutral but statistically connected with Y . Suppose that its mean, $\bar{X} = N^{-1} \sum_{i=1}^N X_i$, is known. For the sake of simplicity, we assume that X is also an indicator variable taking value 1 if the i -th person possesses the auxiliary characteristic, 0 otherwise.

To estimate the unknown proportion of people possessing the sensitive characteristic, $\pi = N^{-1} \sum_{i=1}^N Y_i$, a sample of size n is selected from P according to the simple random sampling with replacement. Aimed at enhancing people cooperation, we consider a randomized device consisting in a deck of cards. On each card is written one of the following two pairs of statements: (a) “I possess the value $Y_i = 1$; I possess the value $X_i = 1$ ”; (b) “I possess the values $Y_i = 0$; I possess the value $X_i = 1$ ”. The two types of cards occur in the proportion of p and $1 - p$. Each respondent is asked to select at random a card from the deck and report a “yes” (“no”) response if his/her status, with respect to the sensitive and the auxiliary characteristic, matches (does not match) with the statements on the card.

Let $Z_i = D_i Y_i + (1 - D_i)(1 - Y_i)$ denotes the randomized response about Y for the i -th population unit, where D_i takes value 1 if a card with the statements (a) is selected, 0 otherwise. To estimate π , Zaizai (2006) suggested the estimator $\hat{\pi}_Z = [\hat{\pi}_r - (1 - p)] / (2p - 1)$, $\hat{\pi}_r = \bar{z}\bar{X}/\bar{x}$. Here, we propose

$$\hat{\pi}_{rr} = \frac{\tilde{\pi}_d - (1 - p)}{2p - 1}$$

where $\tilde{\pi}_d = \bar{z} - b(\bar{x} - \bar{X})$ is the *difference estimator*. For b known, the estimator is unbiased and the variance is given by:

$$Var(\hat{\pi}_{rr}) = \frac{\pi(1 - \pi)}{n} + \frac{p(1 - p)}{n(2p - 1)^2} - \frac{b}{n(2p - 1)^2} [2(2p - 1)\sigma_{xy} - b\sigma_x^2]$$

where σ_x^2 is the variance of X and σ_{xy} the covariance between X and Y .

The variance is minimized for $b_o = (2p - 1)\sigma_{xy}/\sigma_x^2$. It follows that:

$$\min Var(\hat{\pi}_{rr}) = \frac{1}{n} \left[\pi(1 - \pi) + \frac{p(1 - p)}{(2p - 1)^2} - \frac{\sigma_{xy}^2}{\sigma_x^2} \right]. \quad (1)$$

By (1) we can immediately state that $\hat{\pi}_{rr}$, with $b = b_o$, is always more efficient than the Warner's estimator. Moreover, it is also more efficient than $\hat{\pi}_Z$.

To conclude we observe that in many situations the optimum b_o is generally unknown and may be replaced with a consistent estimate. In this case, using Taylor series expansion, and considering expectation term-by-term up to include terms of order n^{-1} , we can derive a first order mean square error which is equivalent to (1).

References

- Warner S.L. (1965) Randomized response: a survey technique for eliminating evasive bias, *Journal of the American Journal of Statistics*, 60, 63–69.
- Zaizai Y. (2006) Ratio method of estimation of population proportion using randomized response technique, *Model Assisted Statistics and Applications*, 1, 125–130.