MODELOS DE RESPOSTA ALEATORIZADA PARA VARIÁVEIS QUANTITATIVAS: MODELO DE GREENBERG

Olga Lidia Solano DÁVILA[1]

- RESUMO: Neste trabalho estudamos alguns modelos de resposta aleatorizada (MRA), comumente usados em pesquisas nas quais estão envolvidas perguntas delicadas. O Modelo de Resposta Aleatorizada para variáveis quantitativas – Modelo de Greenberg – foi aplicado em uma pesquisa realizada na Faculdade de Ciências Matemáticas da Universidade Nacional Maior de São Marcos no Peru, para investigar o gasto médio em bebidas alcoólicas (pergunta delicada) e o gasto médio em materiais de estudo (pergunta não relacionada). A população em estudo compreende os alunos matriculados no semestre 96 – II.

- PALAVRAS CHAVES: Modelos de resposta aleatorizada, não-resposta, mecanismo aleatório, modelo de Greenberg, pergunta delicada.

1 Introdução

Em pesquisas realizadas no passado, nas quais estão envolvidas perguntas altamente delicadas ou pessoais, detecta-se que muitas vezes as pessoas não estão inclinadas a responder com honestidade ou simplesmente assinalam uma negativa por resposta. Isto é, a não-resposta, respostas evasivas e respostas falsas são tão comuns na prática que dificilmente podem ser medidas e controladas. Para solucionar esse problema é possível utilizar procedimentos alternativos, se nossa intenção é obter dados confiáveis nessa classe de investigação, em vez de realizar pesquisas pela modalidade de entrevista direta.

A primeira técnica de Resposta Aleatorizada foi introduzida por Warner (1965), com o objetivo de obter dados confiáveis; para isto projetam-se habilmente as perguntas de tal maneira que não se revela a identidade do entrevistado no decorrer da pesquisa. O entrevistador deve utilizar um mecanismo aleatório (urna com bolas, flecha giratória, uma moeda etc.) para selecionar uma de duas perguntas (Pertenço ao grupo A ou não pertenço ao grupo A), cada uma das quais requer uma resposta "Sim" ou "Não" por parte do entrevistado, sem revelar ao entrevistador sua posição pessoal com respeito à pergunta.

O modelo estima a proporção de elementos da população que pertence ao grupo A a partir de uma amostra aleatória com reposição.

Greenberg et al. (1969) modificou o modelo original de Warner com a finalidade de que o entrevistado tenha maior disposição a colaborar, substituindo o segundo enunciado "Não Pertenço ao grupo A" por uma pergunta referente a uma característica B, não relacionada com A.

[1] Departamento de Estadística, Universidad Nacional Mayor de San Marcos, Av. Venezuela s/n, Lima, Perú. E-mail: *solano_olga@yahoo.com.br*.

O entrevistado deve selecionar por meio de um mecanismo aleatório uma das seguintes proposições a que se supõe responde corretamente.

1. Pertenço ao grupo A.
2. Pertenço ao grupo B.

Os parâmetros que se deseja estimar são:

π_A - proporção de elementos da população que pertence a A .

π_B - proporção de elementos da população que pertence a B .

Esses dois parâmetros são estimados com base em duas amostras aleatórias independentes de tamanho n_1 e n_2 respectivamente.

Greenberg et al. (1971) sugerem uma extensão do modelo da "pergunta não relacionada" para características quantitativas.

Nesse modelo utilizam-se duas variáveis quantitativas; a primeira variável é *X*, cujo valor resulta comprometedor ao entrevistado (gasto em bebidas alcóolicas, número de abortos, salários-familia etc.); a segunda variável é *Y* não relacionada a *X*; $g(.)$ é função de densidade de *X* e $h(.)$ função de densidade de *Y*.

O entrevistado utiliza um mecanismo aleatório para selecionar a pergunta delicada com probabilidade *P* e a pergunta "não relacionada" com probabilidade (1-*P*).

Os parâmetros que se deseja estimar são:

μ_X - média populacional da variável *X*, associada à pergunta delicada.

μ_Y - média populacional da variável *Y*, associada à pergunta "não relacionada".

Esses dois parâmetros são estimados com base em duas amostras aleatórias independentes de tamanho n_1 e n_2, e o estimador está relacionado à variável aleatória associada à resposta, chamada *Z*, que é a mistura de *X* e *Y* segundo *P* e (1-*P*).

O Modelo de Resposta Aleatorizada para variáveis quantitativas – Modelo de Greenberg – foi aplicado em uma pesquisa realizada na Faculdade de Ciências Matemáticas da Universidade Nacional Maior de São Marcos no Peru, para investigar o gasto médio em bebidas alcoólicas (pergunta delicada) e o gasto médio em materiais de estudo (pergunta não relacionada). A população em estudo compreende os alunos matriculados no semestre 96 – II.

Modelo de resposta aleatorizada para variáveis quantitativas

Greenberg et al. (1971) sugere uma extensão do modelo da "pergunta não relacionada" para características quantitativas.

Seja *X* a variável quantitativa cujo valor resulta comprometedor ao entrevistado (gasto em bebidas alcóolicas, número de abortos, salários familiares, etc.), e $g(.)$ é função de densidade de *X*.

Seja *Y* a variável quantitativa não relacionada a *X*; e $h(.)$ a função de densidade de *Y*.

O entrevistado utiliza um mecanismo aleatório para selecionar a pergunta delicada com probabilidade *P* e a pergunta "não relacionada" com probabilidade (1-*P*).

Estimação da média populacional μ_X , com μ_Y conhecida

Suponhamos que selecionamos uma amostra simples com reposição, de tamanho n. As probabilidades de selecionar a pergunta aleatorizada na amostra são:

P: Probabilidade de que a pergunta delicada associada a X seja selecionada pelo entrevistado.
Q =1- P: Probabilidade de que a pergunta "não relacionada" associada a Y seja selecionada pelo entrevistado.

Então a resposta aleatorizada Z, tem função de densidade:

$$f(z) = Pg(z) + Qh(z) \tag{2.1}$$

Por ser $g(.)$ função de densidade de X, utilizando a definição de esperança matemática e variância, obtemos:

$$\mu_Z = P\mu_X + Q_Y \tag{2.2}$$

$$\sigma_Z^2 = P\sigma_X^2 + Q\sigma_Y^2 + PQ(\mu_X - \mu_Y)^2 \tag{2.3}$$

A média e variância amostral são:

$$\overline{Z} = \frac{\sum_{i=1}^{n} Z_i}{n} \qquad e \qquad S^2 = \frac{\sum_{i=1}^{n} (Z_i - \overline{Z})^2}{n-1},$$

Estimador não-viciado de μ_X

$$\hat{\mu}_X = \frac{(\overline{Z} - Q\mu_Y)}{P} \tag{2.4}$$

Estimador não-viciado da variância de μ_X

$$\hat{Var}(\hat{\mu}_X) = \frac{S^2}{nP^2} \tag{2.5}$$

Estimação da média populacional μ_X , com μ_Y desconhecida

Suponhamos que sejam selecionadas duas amostras independentes simples com reposição de tamanhos n_1 e n_2. As probabilidades de selecionar a pergunta aleatorizada em cada amostra são:

P_1: Probabilidade de que a pergunta delicada associada a X seja selecionada pelo entrevistado na amostra 1.
P_2: Probabilidade de que a pergunta delicada associada a X seja selecionada pelo entrevistado na amostra 2.

Q_1: Probabilidade de que a pergunta "não relacionada" associada a Y seja selecionada pelo entrevistado na amostra 1.

Q_2: Probabilidade de que a pergunta "não relacionada" associada a Y seja selecionada pelo entrevistado na amostra 2.

E as Respostas Aleatorizadas Z_1 e Z_2 na amostra 1 e 2, respectivamente, têm função de densidade:

$$f_1(z_1) = P_1 g(z_1) + Q_1 h(z_1) \quad e \quad f_2(z_2) = P_2 g(z_2) + Q_2 h(z_2) \tag{2.6}$$

Por ser g(.) função de densidade de X, utilizando a definição de esperança matemática e variância, obtemos:

$$\mu_{Z_1} = P_1 \mu_X + Q_1 \mu_Y \,,\ \mu_{Z_2} = P_2 \mu_X + Q_2 \mu_Y \quad \text{e}$$

$$V(z_1) = P_1 \sigma_X^2 + Q_1 \sigma_Y^2 + P_1 Q_1 (\mu_X - \mu_Y)^2 \,, \tag{2.7}$$

$$V(z_2) = P_2 \sigma_X^2 + Q_2 \sigma_Y^2 + P_2 Q_2 (\mu_X - \mu_Y)^2$$

As médias e variâncias amostrais são:

$$\overline{Z}_1 = \frac{\sum_{i=1}^{n_1} Z_i}{n_1} \,,\quad \overline{Z}_2 = \frac{\sum_{i=1}^{n_2} Z_i}{n_2} \quad \text{e} \quad S_1^2 = \frac{\sum_{i=1}^{n_1} (Z_i - \overline{Z}_1)^2}{n_1 - 1} \,,\quad S_2^2 = \frac{\sum_{i=1}^{n_2} (Z_i - \overline{Z}_2)^2}{n_2 - 1} \tag{2.8}$$

onde

$\overline{Z}_1$: média amostral de Z_1 na amostra 1,

$\overline{Z}_2$: média amostral de Z_2 na amostra 2,

S_1^2: Variância amostral de Z_1 na amostra 1 e

S_2^2: Variância amostral de Z_2 na amostra 2.

Estimador não-viciado de μ_X e μ_Y

$$\hat{\mu}_X = \frac{\left[(1-P_2)\overline{Z}_1 - (1-P_1)\overline{Z}_2\right]}{P_1 - P_2} \text{ e } \hat{\mu}_Y = \frac{\left[P_2 \overline{Z}_1 - P_1 \overline{Z}_2\right]}{P_2 - P_1}, \text{ sempre que } P_1 \neq P_2 \tag{2.9}$$

Estimador não viciado da variância de μ_X e μ_Y

$$\hat{Var}(\hat{\mu}_X) = \frac{\left[\frac{(1-P_2)^2 S_1^2}{n_1} + \frac{(1-P_1)^2 S_2^2}{n_2}\right]}{(P_1 - P_2)^2} \quad \text{e} \quad \hat{Var}(\mu_Y) = \frac{\left[\frac{P_2^2 S_1^2}{n_1} + \frac{P_1^2 S_2^2}{n_2}\right]}{(P_2 - P_1)^2} \,, \text{ sempre} \tag{2.10}$$

que $P_1 \neq P_2$

Escolha de P_1 , P_1, n_1 e n_2

No trabalho de Greenberg et al. (1971) são apresentadas as seguintes considerações:

1. Na escolha de P_1 e P_2 deve-se levar em consideração a diminuição da variância e também a disposição da colaboração por parte do entrevistado. Para isso, propõem P_1+P_2=1 e elege-se P_1 tão longe de 0,5 quanto seja factível. Em experiências anteriores obtiveram-se resultados satisfatórios com P_1 entre 0,70 e 0,80 ou o seu complemento.
2. Escolha da pergunta "não relacionada" associada a *Y*. Uma regra fundamental é que a pergunta não relacionada deve estar na mesma unidade de medida que a pergunta delicada, por exemplo, reais, cm, ou número de vezes da ocorrência de um evento.
   Em geral recomenda-se que:
   σ_Y^2 tão perto quanto possível de σ_X^2
   μ_Y perto de μ_X
   σ_Y^2 tão pequeno quanto possível.
3. Determinação ótima de n_1 e n_2, a fim de minimizar $V\hat{a}r(\hat{\mu}_X)$, sujeita à condição $n_1 + n_2 = n$. Essa determinação ótima é dada por:

$$\frac{n_1}{n_2} = \frac{Q_2\hat{\sigma}_1}{Q_1\hat{\sigma}_2}, \tag{2.11}$$

onde $\hat{\sigma}_1^2$: variância de *Z* na amostra 1 e $\hat{\sigma}_2^2$: variância de *Z* na amostra 2.

Tamanho da amostra *n*

Para achar o tamanho da amostra e determinação ótima da mesma, usaremos a condição $n_1 + n_2 = n$, e a determinação ótima mencionada anteriormente, $V\hat{a}r(\hat{\mu}_X)$, a normalidade assintótica de $\hat{\mu}_x$, e, após manipulações algébricas temos o tamanho da amostra calculada por:

$$n = \left(\frac{Z_\alpha}{E}\right)^2 \frac{[\sigma_2 + P_1(\hat{\sigma}_1 - \hat{\sigma}_2)]^2}{(2P_1 - 1)^2}, \tag{2.12}$$

para um erro máximo de estimação prefixado *E* e um nível de confiabilidade (1- α).

Aplicação do Modelo de Greenberg

Pesquisa de estudantes sobre o gasto médio em bebidas alcoólicas

O modelo de Resposta Aleatorizada de Greenberg, da pergunta não relacionada para variáveis quantitativas, foi utilizado em uma pesquisa sobre o gasto em bebidas alcoólicas. Sua aplicação foi feita na Faculdade de Ciências Matemáticas da Universidade

Nacional Maior de São Marcos (UNMSM), considerando duas amostras aleatórias independentes selecionadas da listagem de alunos matriculados no semestre 96-II.

O propósito desta investigação foi mostrar a utilidade da técnica de Resposta Aleatorizada para obter uma estimativa do gasto por bebidas alcoólicas, reduzindo principalmente o vício de resposta (por respostas deliberadamente falsas) e obter taxas de resposta maiores.

Escolha das variáveis

As variáveis escolhidas para este estudo foram as seguintes:

1. Gasto em bebidas alcoólicas (variável associada à pergunta delicada)
2. Gasto em materiais de estudo (variável associada à pergunta não relacionada)

Escolha do mecanismo aleatório e da probabilidade de seleção das perguntas

O mecanismo aleatório que serviu para selecionar um de dois enunciados, cada um dos quais requer uma resposta expressa em moeda nacional (novos soles), por parte do entrevistado, sem revelar ao entrevistador sua posição pessoal com respeito à pergunta, foi uma caixa de plástico transparente e selada, de aproximadamente 38 cm de comprimento, 25,5 cm de largura e 10,5 cm de altura, que continha bolas vermelhas e brancas.

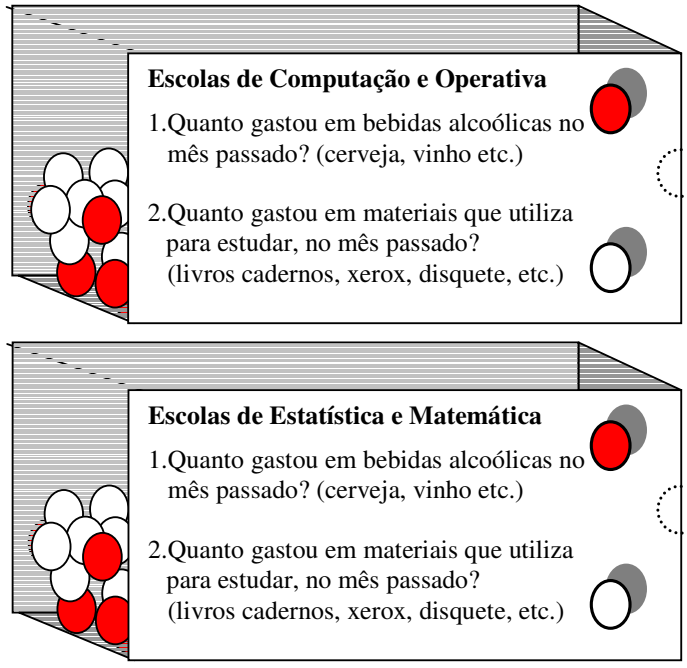

FIGURA 1 – Mecanismo aleatório usado na pesquisa.

A caixa continha dois enunciados localizados na tampa e identificados com uma cor de bola cada um (ver Figura 1).

Os enunciados foram:

1. Quanto gastou em bebidas alcoólicas no mês passado? (cerveja, vinho, rum, etc.)
2. Quanto gastou em materiais que utiliza para estudar no mês passado? (livros, cadernos, xerox, disquete, impressões etc.).

A primeira pergunta estava identificada com uma bola vermelha. Na parte esquerda desta estava desenhado um círculo de cor vermelha.

A segunda pergunta estava identificada com uma bola branca. Na parte esquerda desta estava desenhado um círculo de cor branca.

Ao entrevistado foi solicitado que remexesse a caixa e removesse as bolas com o propósito de que uma das bolas ficasse na janela da caixa, que era visível ao entrevistado. A cor da bola que aparecia na janela determinava qual das duas perguntas teria de responder.

Se uma bola vermelha ficasse na janela, o entrevistado respondia à primeira pergunta (gasto em bebidas alcoólicas); se uma bola branca ficasse na janela, respondia à segunda pergunta (gasto em materiais de estudo).

O entrevistador encontrava-se de costas para o entrevistado, de tal maneira que não se informava que pergunta havia sido selecionada pelo entrevistado. Para evitar o conhecimento de qual pergunta havia respondido o entrevistado, a caixa era removida novamente de tal maneira que nenhuma bola ficasse perto da janela da caixa.

O total de bolas e a probabilidade de seleção por tipo de variável a ser pesquisada é mostrado a seguir:

Para realizar a entrevista utilizaram-se duas caixas.

A caixa utilizada para a amostra 1 (Escola de Computação e Operativa) continha 51 bolas vermelhas e 9 bolas brancas. Logo, a probabilidade (P_1) de selecionar a pergunta 1 é de 0,85 e a probabilidade (1- P_1) de selecionar a pergunta 2 é de 0,15.

A caixa utilizada para a amostra 2 (Escola de Estatística e Matemática) continha 34 bolas brancas e 6 bolas vermelhas. Logo, a probabilidade (P_2) de selecionar a pergunta 1 é de 0,15 e a probabilidade (1-P_2) de selecionar a pergunta 2 é de 0,85.

Tamanho da amostra

O tamanho da amostra foi calculado com base em uma amostra piloto de tamanho 30, para o qual se formaram dois estratos. No primeiro estrato estava a Escola de Computação e Operativa e no segundo, a Escola de Matemática e Estatística. Da amostra piloto obtiveram-se os seguintes resultados:

$\hat{\sigma}_1 = 39{,}77$, $\hat{\sigma}_2 = 43{,}47$, $\hat{\mu} = 37{,}45$,

onde $\hat{\mu}$: Gasto médio em bebidas alcoólicas de um aluno na Faculdade de Ciências Matemáticas.

Calcularam-se vários tamanhos de amostra para diferentes erros de estimação (E) e para diferentes valores de P_1.

Exemplo:

$$Z_{5\%} = 1{,}96, \qquad E = 15\%\mu, \qquad E = 15\% \times 37{,}45 = 5{,}6175,$$

$$\left(\frac{z_{5\%}}{E}\right)^2 = \left(\frac{1{,}96}{5{,}6715}\right)^2 = 0{,}12.$$

Substituindo em 2.12, obtemos $n = 404$.

Tabela 1 - Tamanho de amostra

Erro de estimação E	Probabilidade de selecionar a pergunta 1 (P_1)					Tamanho de amostra η no método convencional
	0,70	0,75	0,80	0,85	0,90	
10%	2.861	1.814	1.248	909	690	479
15%	1.272	806	555	404	306	213
20%	715	454	312	227	172	120

Para a aplicação do trabalho utilizou-se uma amostra de tamanho 227, com um erro de estimação de 20% e 95% de confiabilidade.

Tamanho da amostra para cada estrato

Usando a equação (2.11) e substituindo os resultados obtidos no teste piloto,

$$\frac{n_1}{n_2} = \frac{Q_2\hat{\sigma}_1}{Q_1\hat{\sigma}_2}, \qquad \frac{n_1}{n_2} = \frac{0{,}85}{0{,}15} \times \frac{39{,}77}{43{,}47} = 5{,}18 \rightarrow n_1 = 5{,}18\ n_2.$$

$n_1 + n_2 = n = 227$

Além disso, sabe-se que:

Utilizando os resultados obtidos anteriormente obtemos que:

$n_1 = 190 \qquad n_2 = 37.$

Então, selecionamos um tamanho de amostra de 190 nas Escolas de Computação e Operativa (estrato 1) e 37 nas Escolas de Estatística e Matemática (estrato 2).

Aplicação do Método convencional

De um total de 1.667 alunos matriculados no semestre 96-II, com idades entre 16 e 57 anos, selecionamos uma amostra aleatoria sim reposição de 227 alunos. A entrevista realizou-se de 27 a 30 de novembro do ano 1996. Os resultados estão na Tabela 2.

Tabela 2 - Resultados da pesquisa segundo o método

Método	Não-Resposta		Taxa de Resposta	Gasto médio em bebidas alcoólicas	Error de Estimação E
	Freq.	Taxa			
MRA Greenberg	0	0,00%	$n = 227$ 100,00%	20,50	6,12
Entrevista Direta	13	5,73%	$n = 227$ 94,27%	17,81	4,98

Observa-se que a taxa de resposta no modelo de resposta aleatorizada foi de 100%, enquanto o método de entrevista direta implicou uma taxa de resposta de 94,27% e uma taxa de não resposta de 5,73%.

O gasto médio em bebidas alcoólicas de um aluno da Faculdade de Ciências Matemáticas foi de 20,50 novos soles, enquanto no método de entrevista direta foi de 17,81 novos soles, no mês de outubro de 1996, o que implica uma provável redução do vício de resposta ao aplicar o modelo de Greenberg.

Intervalos de confiança de 95%, para $\hat{\mu}_X$ e $\hat{\mu}_Y$

$\hat{\mu}_X \in < 14{,}39;\ 26{,}61 >$ $\quad$ $\hat{\mu}_Y \in < 12{,}71;\ 36{,}27 >$

Conclusões

A aplicação do MRA – Modelo de Greenberg – permitiu pôr em prática uma técnica de amostragem que é eficiente em pesquisas com perguntas delicadas.

O MRA é mais eficiente quando é utilizado em pesquisas nas quais o problema é muito delicado, para o qual se requer um tamanho de amostra maior que o convencional.

A desvantagem da técnica é que o gasto na capacitação dos entrevistadores (o treinamento da técnica) e o tempo gasto na entrevista para explicar a técnica ao entrevistado é maior que o método tradicional.

A técnica de resposta aleatorizada é mais eficiente quando utilizada em pesquisas em que o problema é altamente sensível e para o qual se requer um tamanho de amostra maior que o método tradicional (ver Tabela 1).

DÁVILA, O. L. Randomized response models for quantitative characters: Greenberg model. *Rev. Mat. Estat.*, São Paulo, v.22, n.3, p.47-56, 2004.

▪ *ABSTRACT: In this study we described some randomized response models (RRM) for quantitative characters. RRM for quantitative characters – Greenberg Model – were applied in a survey conducted at the Faculty of Mathematical Science of Universidad Nacional Mayor de San Marcos in Peru in order to estimate the average monthly expenditure on alcoholic drinks (sensitive question) and the average monthly expenditure on textbooks (unrelated question). The population under study consisted of students registered in the 96-II school term.*

▪ *KEYWORDS: Randomized response models; no response; randomization device; Greenberg model; sensitive question.*

Referências

CHAUDHRI, A.; MUKERJEE, R. *Randomized response*: theory and techniques. New York: Marcel Dekker, 1983. 161p.

COCHRAN, G. W. *Técnicas de muestreo.* México: Continental, 1977. 513p.

DES, R. *Teoría del muestreo*. México: Fondo de Cultura Económica, 1980. 305p.

GREENBERG, B. G. et al. The unrelated question randomized response model: theoretical frame work. *J. Am. Stat. Assoc.,* Alexandria, v.64, p.520-539, 1969.

GREENBERG, B. G. et al. Aplication of the randomized response technique in obtaining quantitative data, *J. Am. Stat. Assoc.,* Alexandria, v.66, n.334, p.243-250, 1971.

LANKE, S. On the choice of the unrelated question in Simmons version of randomized response model. *J. Stat. Assoc.*, New York, v.70, p.80-83, 1975.

LIU, P. T.; CHOW, L. P. A new discrete quantitative randomized response model. *J. Am. Stat. Assoc*., Alexandria, v.71, n.353, p.72-73, 1976.

MAHMOOD, M.; SINGH, S.; HORN, S. On the confidentiality guaranteed under randomized response sampling: A comparison with several new techniques. *Biom. J.*, Berlin, v.40, n.2, p.237-242, 1998.

MOORS, J. J. A Optimization of the unrelated question randomized response model. *J. Stat. Assoc.*, New York, v.66, n.361, p.627-629, 1971.

SHIMIZU, I. M.; BONHAM, G. S.Randomized response technique in a national survey. *J. Am. Stat. Assoc.,* Alexandria, v.73, n.361, p.35-39, 1978.

SOLANO, O. L. *Modelos de respuesta aleatorizada para variables cuantitativas*. 1997. 98f. Tese (Licenciada en Estadística) – Facultad de Ciencias Matemáticas, Universidad Nacional Mayor de San Marcos, Lima, 1981.

WARNER, S. L. Randomized response: A survey technique for elimination evasive answer bias. *J. Stat. Assoc.*, New York, v.60, p.63-69, 1965.

Recebido em 09.02.2004.

Aprovado após revisão em 22.12.2004.