98

UNIVERSIDADE DE SÃO PAULO
INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

MODELO DE RESPOSTA ALEATORIZADA

Dissertação apresentada ao
IME da USP, para obtenção
do grau de Mestre em Ciências

JOSÉ FERREIRA DE CARVALHO (*)

Orientador:

Dr. FLÁVIO WAGNER RODRIGUES ()**

(*) Do Instituto de Ciências Matemáticas de São Carlos - USP

(**) Do Instituto de Matemática e Estatística - USP

R E S U M O

Este trabalho apresenta os modelos de entrevista baseados em respostas aleatorizadas. É dado um tratamento unificado ao problema, todos os modelos sendo formulados como casos particulares de um modelo linear geral. Para os casos mais importantes, são discutidos problemas relacionados ao planejamento de amostragens. Um outro método, não enquadrado no modelo geral, é apresentado para estimar proporções de dois atributos.

S U M M A R Y

This dissertation gives an account of several randomized response models. An unified approach is used, by showing that all models are particular cases of a linear model. Suggestions for the design of surveys are discussed for the most significative cases. Another method, that is not a particular case of the linear model presented, is derived to estimate the three proportions in a two-way classification.

ÍNDICE

0 - Introdução	1
1 - O Modelo de Resposta Aleatorizada	4
1.1 - O modelo de Warner ou modelo I para proporções	4
1.2 - O modelo geral de resposta aleatorizada	10
2 - Casos Particulares do Modelo de Resposta Aleatorizada	18
2.1 - Extensão do modelo I para estimar mais de uma proporção	18
2.2 - O modelo de resposta aleatorizada com questão não-relacionada ou modelo II para proporções	23
2.3 - Caso quantitativo	36
3 - Outro Modelo Para Estimar as Proporções de Dois Atributos	41
3.1 - Formulação do modelo	41
3.2 - Estimação de Π_A , Π_B e Π_{AB}	43
3.3 - Planejamento de amostragem	49
Apêndice	50
Referências	52

0 - INTRODUÇÃO

Pesquisas em populações humanas, cujos dados obtem-se de entrevistas, enfrentam duas fontes de erros:

- 1) A falta de respostas pela absoluta negativa de alguns entrevistados a responder;
- 2) A presença de respostas falsas, dadas por entrevistados que, eventualmente, não desejem revelar certas condições.

Hã problemas para os quais o método de entrevista usualmente empregado (direto, que consiste simplesmente em perguntar ao indivíduo seu "status") é inviável. Por exemplo:

- a estimação do número médio de abortos provocados em certa população;
- a estimação da proporção de indivíduos que já tiveram contato com substâncias psicotrópicas.

Os dois exemplos citados referem-se a situações ilegais. Uma mulher que induziu aborto ou um viciado em psicotrópicos estão sujeitos a punições, de modo que eles não responderão corretamente a uma questão que os revele como tais. Mas qualquer questão de caráter muito íntimo sofrerá as consequências das mesmas duas fontes de erro.

Algumas tentativas de formular esquemas de entrevista diferentes, de modo a eliminar os erros citados, foram tentadas, mas sem êxito. Em geral baseiam-se em assegurar ao entrevistado o anonimato. Por exemplo, o entrevistado responde a um questionário, a sós, simplesmente marcando alternativas correspondentes a seu caso particular. Depois deposita sua resposta numa urna lacrada.

É claro que o entrevistado imagina que a sua resposta pode ser a única da urna e que a urna pode ser aberta após sua saída. Assim, não necessariamente se sentirá seguro para responder.

Em 1965, Warner [6] idealizou um método engenhoso para resolver o problema de estimar a proporção de um atributo cuja revelação fosse problemática.

Consiste o processo no seguinte: suponha-se que deseja-

mos estimar a proporção Π de certo atributo. Seja A o conjunto de indivíduos portadores do atributo e $\bar{A}$ seu complementar. Obtenha-se uma amostra aleatória da população em estudo e entreviste-se cada elemento da amostra, com duas questões:

$V. \in A?$

$V. \in \bar{A}?$

das quais o entrevistado selecionará aleatoriamente uma para responder. Sua resposta será "sim" ou "não" e não revelará a qual pergunta se refere. Supomos conhecidas, apenas, as probabilidades p e $1-p$, respectivamente de seleção da primeira e da segunda pergunta, que serão as mesmas em todas as entrevistas.

Se $p \neq 1/2$, tal método fornece um estimador para Π , desde que, é claro, as respostas possam ser consideradas corretas. Agora, se $p \neq 0$ e $p \neq 1$, não temos condições de classificar - com certeza absoluta, o entrevistado em A ou $\bar{A}$. Se o entrevistado foi suficientemente esclarecido dessa situação, cessa o motivo para não colaborar. Portanto, escolhas convenientes de p podem levar os entrevistados a colaborar, respondendo corretamente.

A partir do trabalho de Warner, várias extensões foram feitas. Abul-Ela et al. [1] estenderam o método para estimar as proporções de m conjuntos mutuamente exclusivos e exaustivos. Greenberg et al. [4] desenvolveram um método em que a segunda pergunta ($V. \in \bar{A}$) é substituída por uma questão não-relacionada à primeira, buscando maior colaboração do entrevistado. Greenberg et al. [5] estenderam a aplicação do método à estimação da média de uma variável aleatória qualquer, cuja revelação fosse crítica.

Todos os distintos modelos apresentados na literatura são passíveis de formulação em termos de um único modelo do tipo linear.

Neste trabalho apresentamos uma formulação do modelo geral, de onde são deduzidos os casos particulares citados.

O capítulo 1 apresenta primeiro o modelo de Warner tal como tratado por aquele autor, para familiarizar o leitor com o processo de entrevista. Depois, é formulado o modelo geral. O estimador do modelo de Warner torna a ser deduzido como um exemplo de aplicação do modelo geral.

No Capítulo 2 são apresentados cada um dos restantes modelos mencionados, como casos particulares do modelo geral. Entretanto, são discutidas orientações para o planejamento de amostras em cada caso específico.

O Capítulo 3 apresenta uma solução para estimar proporções de dois atributos, que não se enquadra na formulação do modelo geral, porque usa duas questões aleatorizadas por entrevista.

... O ...

Ao completar este trabalho, lembramos que muitas pessoas, direta ou indiretamente, colaboraram em nosso estudo. Gostaríamos de agradecer a todas essas pessoas, especialmente a:

Dr. Flávio W. Rodrigues, orientador deste trabalho, cuja boa vontade superou as dificuldades impostas pela distância de São Carlos a São Paulo;

Dr. Carlos Alberto Barbosa Dantas (Caio), a cujo esforço pessoal devemos a existência do programa de pós-graduação em Estatística na U.S.P., que nós cursamos;

Júlio Ferreira de Carvalho, Ruy Duque Estrada, José da Cunha Oliveira, Arthur Bernardes Weiss e Dr. Odelar Leite Linhares: amigos que muito nos ensinaram e/ou professores que nos brindaram com sua amizade (o que dá no mesmo);

e a Gaby, esposa, amiga e companheira de trabalho, que colabora e incentiva.

... O ...

a impressão deste trabalho recebeu ajuda parcial da
CAPES e do CNPq

1 - O MODELO DE RESPOSTA ALEATORIZADA

1.1 - O modelo de Warner ou modelo I para proporções

1.1.1 - O Método de Entrevista

Seja uma população humana C e um atributo que particione essa população nos conjuntos A e $\bar{A}$, respectivamente dos elementos de C que possuem o atributo e dos que não possuem. Represente-se por Π a proporção dos elementos de C que pertencem a A ; a proporção dos elementos de C que pertencem a $\bar{A}$ é, assim, $1-\Pi$.

Imaginemos o atributo tal que a pertinência a A seja ilegal ou socialmente desaprovada. Se um elemento de A for indagado:

V. $\in$ A ?

responderá, provavelmente, não, buscando manter oculta sua situação.

Uma pesquisa estatística, para estimar Π , que selecione uma amostra aleatória de C e, através de entrevistas, pergunte: V. $\in$ A ? será tendenciosa, pelas respostas falsas que resultarão.

Vários métodos de entrevista podem ser tentados para obter informações, mantendo o sigilo do entrevistado. Por exemplo, solicitar ao entrevistado que, a sós, preencha um questionário e o ponha em uma urna lacrada. Claro está que há nesse processo muitas possibilidades de embuste e, por isso, o entrevistado poderá mentir, com o fim de manter-se oculto.

Uma solução para o problema de estimar Π consiste em idealizar um processo de entrevista no qual o entrevistado forneça informação sem, todavia, haver possibilidade de revelar-se.

O método inventado por Warner [6] é baseado nessa idéia e nele o entrevistado fornece informação sobre seu status sem dar oportunidade de ser revelado. A informação (parcial) que dá possibilidade a estimação de Π .

Selecionada uma amostra aleatória simples de tamanho n de C , entrevista-se cada elemento da amostra do seguinte modo: o

entrevistado recebe duas perguntas:

1 - $V. \in A?$

2 - $V. \in \bar{A}?$

e um dispositivo que, acionado, seleciona a questão 1, com probabilidade p , ou a questão 2, com probabilidade $1-p$. Solicita-se ao entrevistado que acione o dispositivo várias vezes para que esteja certo de que ambas as questões podem ser selecionadas. Longe da vista do entrevistador, o entrevistado aciona o dispositivo e responde à questão indicada. Dito "sim" ou "não", conforme o caso, torna a acionar o dispositivo, para manter o sigilo, e o devolve ao entrevistador.

O dispositivo para seleção das perguntas pode ser uma caixa com bolas brancas e azuis, nas proporções p e $1-p$, dotada de uma janela que deixe ver uma só bola. Sacode-se a caixa e verifica-se a cor da bola aparente na janela; se for branca, a questão 1 deve ser respondida; se for azul, a questão a ser respondida é a 2.

É claro que a resposta dada pode referir-se à questão 1 ou à questão 2. Dessa forma, não temos meios para classificarmos, com segurança absoluta, o entrevistado, que pode, então, responder corretamente.

Supondo que os entrevistados respondam corretamente, é possível estimar Π .

1.1.2 - Estimação de Π .

Definam-se as variáveis aleatórias

$$X_i = \begin{cases} 1 & \text{se a resposta da } i\text{-ésima entrevista for "sim"} \\ 0 & \text{se a resposta da } i\text{-ésima entrevista for "não"} \end{cases}$$

($i=1, \dots, n$) (1.1.2-1)

Nessas condições X_i é uma coleção de variáveis aleatórias de Bernoulli independentes e com mesma distribuição, de parâmetro:

$$\lambda = P\{X_i=1\} = \Pi p + (1-\Pi)(1-p) \quad i=1, \dots, n \quad (1.1.2-2)$$

A variável aleatória

$$M = \sum_{i=1}^n X_i, \quad (1.1.2-3)$$

que corresponde ao "número de respostas 'sim' na amostra", tem distribuição binomial, com parâmetros n e λ .

A função de verossimilhança dessa amostra é dada por

$$L(X_1, \dots, X_n; \Pi) = \lambda^M (1-\lambda)^{n-M} \quad (1.1.2-4)$$

O estimador de máxima-verossimilhança é dado pela solução de

$$\frac{\partial \log L}{\partial \Pi} = \frac{M(2p-1)}{\Pi p + (1-\Pi)(1-p)} + \frac{(n-M)(1-2p)}{\Pi(1-p) + p(1-\Pi)} = 0 \quad (1.1.2-5)$$

Temos, então, o estimador

$$\hat{\Pi} = \frac{1}{2p-1} \left\{ \frac{M}{n} + p-1 \right\} \quad (p \neq \frac{1}{2}) \quad (1.1.2-6)$$

O estimador obtido é não-tendencioso para Π . Com efeito

$$E(\hat{\Pi}) = E\left\{ \frac{1}{2p-1} \left(\frac{M}{n} + p-1 \right) \right\} = \frac{1}{2p-1} \left\{ \frac{E(M)}{n} + p-1 \right\}$$

Mas $E(M) = n\lambda$,

e assim

$$E(\hat{\Pi}) = \frac{1}{2p-1} \{ \Pi p + (1-\Pi)(1-p) + p-1 \} = \Pi \quad (1.1.2-7)$$

A variância do estimador $\hat{\Pi}$ é dada por

$$V(\hat{\Pi}) = \frac{\Pi(1-\Pi)}{n} + \frac{p(1-p)}{n(2p-1)^2} \quad (1.1.2-8)$$

A prova de (1.1.2-8) é fácil:

$$V(\hat{\Pi}) = V\left\{\frac{1}{2p-1}\left(\frac{M}{n} + p-1\right)\right\} = \frac{V(M)}{n^2(2p-1)^2} \quad (1.1.2-9)$$

Como M tem distribuição binomial (n, λ) , segue-se que

$$V(M) = n\lambda(1-\lambda) = n[(2p-1)^2 \Pi(1-\Pi) + p(1-p)] \quad (1.1.2-10)$$

Substituindo-se (1.1.2-10) em (1.1.2-9), obtemos a expressão (1.1.2-8) para $V(\hat{\Pi})$.

Observando (1.1.2-8), verificamos que $V(\hat{\Pi})$ é composta de duas partes:

$$\frac{\Pi(1-\Pi)}{n} \quad \text{e} \quad \frac{p(1-p)}{n(2p-1)^2}$$

A primeira parte é a variância mínima dada pela desigualdade de Cramér-Rao e atingida pelo estimador do esquema comum de entrevista (pergunta direta), quando as respostas são verídicas. A segunda parte é o acréscimo causado pela aleatorização da resposta. Parece claro, então, que o método de Warner só deve ser aplicado a questões verdadeiramente críticas, quando o processo usual leva o entrevistado a mentir.

1.1.3 - Eficiência de $\hat{\Pi}$

O estimador obtido pelo método de pergunta direta é ótimo (não-tendencioso e de variância mínima), quando as respostas são verídicas e não há recusa a responder. Quando há respostas falsas e/ou recusas a responder, tal estimador torna-se tendencioso e sua eficiência (em termos do erro médio quadrático comparado com a cota mínima de Cramér-Rao) diminui sensivelmente.

O método de entrevista de Warner pode favorecer a colaboração dos entrevistados, no sentido de diminuir a probabilidade de resposta falsa. Se consideramos crítica a pertinência a A, mas não a pertinência a $\bar{A}$, está claro que os membros de $\bar{A}$ não terão razões para mentir.

No caso em que, sob o método de Warner, a probabilidade de resposta correta para um elemento de A é menor do que 1, $\hat{\Pi}$ é também tendencioso. Se T_w representa a probabilidade de resposta correta - para um elemento de A, então a tendenciosidade de $\hat{\Pi}$ é dada por

$$B(\hat{\Pi}) = \Pi(T_w - 1) \quad (1.1.3-1)$$

O erro médio quadrático de $\hat{\Pi}$ é definido por

$$EMQ(\hat{\Pi}) = V(\hat{\Pi}) + B^2(\hat{\Pi})$$

As expressões correspondentes para o estimador dado pelo método usual, de entrevistas diretas, estão no apêndice 1.

Definimos a eficiência relativa entre o estimador de Warner, $\hat{\Pi}$ (quando a probabilidade de resposta correta é, para um elemento de A, T_w) e o estimador usual $\tilde{\Pi}$ (quando a probabilidade de resposta correta, para um elemento de A, é T_u) por

$$ef(\hat{\Pi}, \tilde{\Pi}) = EMQ(\tilde{\Pi}) / EMQ(\hat{\Pi})$$

A tabela a seguir, ilustra alguns casos de $ef(\hat{\Pi}, \tilde{\Pi})$. Dizemos que $\hat{\Pi}$ é mais eficiente do que $\tilde{\Pi}$, quando $ef(\hat{\Pi}, \tilde{\Pi}) > 1$. Se $ef(\hat{\Pi}, \tilde{\Pi}) < 1$, então $\tilde{\Pi}$ é dito mais eficiente do que $\hat{\Pi}$.

Da tabela (1.1-1) podemos concluir que o estimador $\hat{\Pi}$ é mais eficiente quando, efetivamente, $T_w > T_u$, isto é, quando o método de entrevista do modelo de Warner induz o entrevistado a colaborar. O quanto T_w deve exceder a T_u para que $ef(\hat{\Pi}, \tilde{\Pi}) > 1$ depende de Π e do valor de p .

Os números na tabela (1.1-1) são suficientes para mos-

TABELA 1.1-1

EFICIÊNCIA RELATIVA ENTRE O ESTIMADOR DE
WARNER E O USUAL SOB RESPOSTAS FALSAS

T _U	T _W = 0,3			T _W = 0,8			T _W = 1		
	p=0,1	p=0,2	p=0,4	p=0,1	p=0,2	p=0,4	p=0,1	p=0,2	p=0,4
0,1	1,60	1,51	0,74	13,20	8,84	1,25	35,16	15,17	1,33
0,2	1,27	1,19	0,59	10,45	6,99	0,99	27,83	12,01	1,05
0,3	0,97	0,92	0,45	8,02	5,37	0,76	21,37	9,22	0,81
0,4	0,72	0,68	0,33	5,92	3,96	0,56	15,78	6,81	0,59
0,9	0,04	0,03	0,02	0,30	0,20	0,03	0,79	0,34	0,03
1,0	0,02	0,02	0,01	0,15	0,10	0,01	0,39	0,17	0,01

π = 0,1

n = 1000

trar que o método de resposta aleatorizada pode ser preferível ao de pergunta direta.

1.1.4 - Planejamento de amostragem com o modelo de resposta aleatorizada

De (1.1.2-8) vemos que $V(\hat{\Pi})$ diverge quando $p \rightarrow 1/2$. O gráfico de $V(\hat{\Pi})$, como função de p , é simétrico em relação a $p=1/2$, com aspecto duplamente hiperbólico, decrescendo de forma monótona com $|p-1/2|$ crescente. Para $p=0$ ou $p=1$, $V(\hat{\Pi})=\Pi(1-\Pi)/n$, que é o mínimo dado pela desigualdade de Cramér-Rao, atingido pelo estimador do processo usual quando as respostas são corretas e não há recusas a responder. Em nosso caso, este mínimo não pode ser atingido, visto que $p=0$ e $p=1$ não são admissíveis.

Decorre, então, a seguinte regra: p deve ser tomado de modo a maximizar $|p-1/2|$, sem, contudo, diminuir a colaboração dos entrevistados.

O estimador do modelo I para proporções tem distribuição assintótica normal, de média Π e variância dada por (1.1.2-8). Isto segue diretamente do fato de $\hat{\Pi}$ ser uma combinação linear da variável aleatória M , cuja distribuição é binomial e depende do tamanho da amostra, n .

Assim, é possível, para grandes valores de n , construir intervalos de confiança aproximados para Π . Uma vez fixado p , também é possível calcular o menor tamanho da amostra para determinada precisão na estimação de Π .

1.2 - O modelo geral de resposta aleatorizada

A partir do modelo de Warner, foi desenvolvido por Greenberg et al. [4] um modelo alternativo para estimar a proporção de um atributo estigmatizante. Consiste o modelo em substituir a questão alternativa " $V. \in \bar{A}?$ " por: " $V. \in B?$ ", onde a pertinência a B é considerada não-estigmatizante e não relacionada com a pertinência a A . Por exemplo, as duas questões poderiam ser:

1- você, alguma vez, tomou drogas psicotrópicas ou fez uso de maconha?

2- você nasceu em Abril?

Este modelo foi chamado por seus autores de "modelo de resposta aleatorizada com questão não-relacionada". Aqui será referido simplesmente como "modelo II para proporções".

Este mesmo modelo foi estendido para a estimação da média de uma variável aleatória qualquer, cuja revelação fosse crítica - (ex: renda em determinado ano, número de abortos induzidos). Tal generalização está no trabalho de Greenberg et al. [5], onde foi denominada "modelo quantitativo".

Todos os modelos de resposta aleatorizada: o de Warner e o de questão alternativa não-relacionada (inclusive o caso quantitativo) são casos particulares de um modelo geral. O modelo geral pode ser formulado como um modelo linear e, assim, todas as propriedades dos modelos lineares são imediatamente aplicáveis.

Apresentaremos a seguir o modelo geral, que chamaremos "modelo de resposta aleatorizada".

1.2.1 - Formulação do modelo de resposta aleatorizada

Considere-se uma população humana, com indivíduos a cada um dos quais está associada uma determinação da variável aleatória m-dimensional $(X_1, \dots, X_m)$.

Sejam:

$f_k(x)$: a função de densidade (ou função de probabilidade) marginal de X_k

$$\mu_k = E(X_k)$$

$$\sigma_k^2 = V(X_k) = E(X_k - \mu_k)^2$$

O problema que se coloca é estimar as médias $\{\mu_k\}$ ($k=1, \dots, m$), impondo-se que algumas das variáveis aleatórias $\{X_k\}$ ($k=1, \dots, m$) sejam de tal natureza que certos indivíduos da população evitem revelar as correspondentes determinações.

Neste caso, adotamos o procedimento que segue:

1- selecionamos da população $r \geq 1$ amostras aleatórias simples de tamanhos $n_1, \dots, n_r$, tais que $n_1 + \dots + n_r = n$;

2- ao j -ésimo elemento da i -ésima amostra ($i=1, \dots, r$; $j=1, \dots, n_i$) é oferecido um dispositivo que, acionado, indica ao acaso uma das variáveis aleatórias X_k ; seja p_{ik} a probabilidade de a variável aleatória X_k ser selecionada mediante esse dispositivo. Tem-se

$$\sum_{k=1}^m p_{ik} = 1 \quad i = 1, \dots, r$$

3- O entrevistado aciona o dispositivo longe da vista do entrevistador e responde o valor da correspondente determinação da variável aleatória selecionada; torna a acionar o dispositivo, para manter o sigilo, e o devolve ao entrevistador.

A escolha das variáveis aleatórias $\{X_k\}$ ($k=1, \dots, m$) deve ser tal que nenhuma possível resposta revele qual a variável selecionada pelo entrevistado. Devem todas ser expressas na mesma unidade e suas distribuições tão assemelhadas quanto possível.

Definamos R_{ij} como sendo a variável aleatória "resposta do j -ésimo elemento da i -ésima amostra" ($i=1, \dots, r$; $j=1, \dots, n_i$).

A distribuição de R_{ij} é uma mistura das distribuições de $X_1, \dots, X_k$. Dentro da mesma amostra, isto é, para i fixado, os R_{ij} tem a mesma distribuição. Seja $\phi_i(x)$ a função de densidade (ou a função de probabilidade) de R_{ij} . Temos:

$$\phi_i(x) = \sum_{k=1}^m p_{ik} f_k(x) \quad i=1, \dots, r \quad (1.2.1-1)$$

A partir de (1.2.1-1) podemos obter as expressões para $E(R_{ij})$ e $V(R_{ij})$:

$$E(R_{ij}) = \sum_{k=1}^m p_{ik} E_{f_k}(R_{ij}) = \sum_{k=1}^m p_{ik} \mu_k \quad (1.2.1-2)$$

$$V(R_{ij}) = E(R_{ij}^2) - E^2(R_{ij})$$

$$E(R_{ij}^2) = \sum_{k=1}^m p_{ik} E_{f_k}(R_{ij}^2) = \sum_{k=1}^m p_{ik} (\sigma_k^2 + \mu_k^2)$$

Por substituição, obtemos

$$V(R_{ij}) = \sum_{k=1}^m p_{ik} (\sigma_k^2 + \mu_k^2) - \left[\sum_{k=1}^m p_{ik} \mu_k \right]^2$$

Finalmente,

$$V(R_{ij}) = \sum_{k=1}^m p_{ik} [\sigma_k^2 + (1-p_{ik})\mu_k^2] - \sum_{k=1}^m \sum_{\ell \neq k} p_{ik} p_{i\ell} \mu_k \mu_\ell \quad (1.2.1-3)$$

Portanto, podemos escrever o seguinte modelo linear:

$$R_{ij} = \sum_{k=1}^m p_{ik} \mu_k + \epsilon_{ij} \quad i=1, \dots, r; j=1, \dots, n_i \quad (1.2.1-4)$$

onde as variáveis aleatórias ϵ_{ij} são tais que

$$E(\epsilon_{ij}) = 0 \quad (1.2.1-5)$$

$$V(\epsilon_{ij}) = V(R_{ij})$$

$$\text{Cov}(\epsilon_{i_1 j_1}; \epsilon_{i_2 j_2}) = 0 \quad (i_1 \neq i_2 \vee j_1 \neq j_2)$$

O modelo (1.2.1-4) pode ser escrito, por conveniência, em forma matricial. Para isto, definam-se:

$$P = \begin{bmatrix} p_{11} & \dots & p_{1m} \\ \vdots & & \vdots \\ p_{11} & \dots & p_{1m} \\ \vdots & & \vdots \\ p_{21} & \dots & p_{2m} \\ \vdots & & \vdots \\ p_{21} & \dots & p_{2m} \\ \vdots & & \vdots \\ p_{r1} & \dots & p_{rm} \\ \vdots & & \vdots \\ p_{r1} & \dots & p_{rm} \end{bmatrix} \left. \begin{array}{l} \left. \begin{array}{l} \left. \begin{array}{l} p_{11} \\ \vdots \\ p_{11} \\ \vdots \\ p_{21} \\ \vdots \\ p_{21} \\ \vdots \\ p_{r1} \\ \vdots \\ p_{r1} \end{array} \right\} \right. \\ \left. \begin{array}{l} \dots \\ \dots \\ \dots \\ \dots \\ \dots \\ \dots \\ \dots \\ \dots \end{array} \right\} \\ \left. \begin{array}{l} p_{1m} \\ \vdots \\ p_{1m} \\ \vdots \\ p_{2m} \\ \vdots \\ p_{2m} \\ \vdots \\ p_{rm} \\ \vdots \\ p_{rm} \end{array} \right\} \end{array} \right\} \begin{array}{l} n_1 \text{ linhas} \\ n_2 \text{ linhas} \\ \vdots \\ n_r \text{ linhas} \end{array}$$

$$R^T = (R_{11}, \dots, R_{1n_1}; R_{21}, \dots, R_{2n_2}; R_{r1}, \dots, R_{rn_r})$$

$$\epsilon^T = (\epsilon_{11}, \dots, \epsilon_{1n_1}; \epsilon_{21}, \dots, \epsilon_{2n_2}; \epsilon_{r1}, \dots, \epsilon_{rn_r})$$

$$\mu^T = (\mu_1, \dots, \mu_m)$$

$$V = \text{diag}[V(\epsilon_{11}) \dots V(\epsilon_{1n_1}) \quad V(\epsilon_{21}) \dots V(\epsilon_{2n_2}) \dots V(\epsilon_{r1}) \dots V(\epsilon_{rn_r})]$$

Com esta notação, o modelo (1.2.1-4) pode ser escrito:

$$R = P\mu + \epsilon \quad (1.2.1-6)$$

Suporemos que V é positiva definida, i.e. :

$$V(\epsilon_{ij}) > 0 \quad i=1, \dots, r; j=1, \dots, n_i$$

Se houver (i,j) tal que $V(\epsilon_{ij})=0$, isto significa que a resposta do j -ésimo indivíduo da i -ésima amostra é uma variável aleatória quase certamente igual a uma constante. Neste caso, ela independe da seleção que o entrevistado faz e da determinação da variável aleatória selecionada. Em outras palavras, tal entrevistado não colabora.

Se supusermos que uma conveniente escolha de matriz P favorece a colaboração do entrevistado, podemos excluir a hipótese de $V(\epsilon_{ij})=0$.

É lícito, pois, supor que V é positiva definida.

As equações normais associadas ao modelo (1.2.1-5) são:

$$(P^T V^{-1} P) \mu = P^T V^{-1} R \quad (1.2.1-7)$$

A característica de $P^T V^{-1} P$ é a mesma de P . Suporemos, no que segue, que as linhas distintas de P sejam linearmente independentes, de modo que a característica de P seja sempre igual ao número de amostras obtidas, r .

Há então r combinações lineares dos μ_i linearmente independentes estimáveis. Em particular, se as m médias não são linearmente relacionadas e se se desejam estimar todas, tem-se que tomar m amostras, de tamanhos $n_1, \dots, n_m$, escolhendo a matriz P de tal modo que tenha característica m .

1.2.2 - Exemplo: o modelo I para proporções

Mostraremos aqui que o modelo descrito em 1.1 é um caso particular do modelo de resposta aleatorizada.

A cada elemento da população C , tal como definida em 1.1, está associada uma determinação da variável aleatória bi-dimensional (X_1, X_2) , onde X_1 tem distribuição de Bernoulli, de parâmetro

Π , e $X_2 = 1 - X_1$. X_1 representa a pertinência a A e X_2 , portanto, a pertinência a $\bar{A}$.

A matriz P é dada por:

$$P = \begin{bmatrix} p & 1-p \\ p & 1-p \\ \vdots & \vdots \\ p & 1-p \end{bmatrix}$$

e o vetor μ por:

$$\mu^T = (\Pi, 1-\Pi)$$

Seja R o vetor de respostas:

$$R^T = (R_1, \dots, R_n)$$

Cada variável aleatória R_i ($i=1, \dots, n$) tem distribuição de Bernoulli, com parâmetro

$$\lambda = p\Pi + (1-p)(1-\Pi) = \Pi(2p-1) + 1-p \quad (1.2.2-1)$$

Logo, V é a matriz escalar de ordem n

$$V = \text{diag} [\lambda(1-\lambda), \dots, \lambda(1-\lambda)]$$

Está formulado, então, o modelo linear

$$R = P\mu + \epsilon$$

onde ϵ é um vetor de dimensão n que atende às condições (1.2.1-5).

Como a característica de P é 1, apenas uma função linear dos parâmetros é linearmente independente estimável. Quaisquer

duas funções estimáveis são linearmente dependentes.

Em particular λ é estimável, pois é o produto de uma linha de P por μ (cf. Graybill [3], pg.230).

Lembrando que V é uma matriz escalar, obtida pelo produto do escalar $\lambda(1-\lambda)$ pela matriz identidade de ordem n , podemos escrever as equações normais (1.2.1-7), para este caso particular, como:

$$\frac{n}{\lambda(1-\lambda)} \begin{bmatrix} p^2 & p(1-p) \\ p(1-p) & (1-p)^2 \end{bmatrix} = \frac{1}{\lambda(1-\lambda)} \begin{bmatrix} p & \sum_{i=1}^n R_i \\ (1-p) & \sum_{i=1}^n R_i \end{bmatrix} \quad (1.2.2-2)$$

e o estimador de λ é, então:

$$\hat{\lambda} = \frac{\sum_{i=1}^n R_i}{n} = \frac{\sum_{i=1}^n R_i}{n} \quad (1.2.2-3)$$

Por ser Π uma função linear de λ , Π é também estimável e seu estimador é dado por (cf. Graybill [3], pp.230 e 116).

$$\hat{\Pi} = \frac{1}{2p-1} \left(\frac{\sum_{i=1}^n R_i}{n} + p-1 \right), \quad (1.2.2-4)$$

obtido pela solução da (1.2.2-3).

Verificamos então que o melhor estimador linear não-tendencioso de Π (1.2.2-4) coincide com o estimador de máxima-verossimilhança do mesmo parâmetro (1.1.2-6).

2 - CASOS PARTICULARES DE MODELO DE RESPOSTA ALEATORIZADA

O modelo apresentado em (1.2) contém, como casos particulares, todos os que foram apresentados na literatura: o modelo de Warner (ou modelo I para proporções); uma extensão deste modelo para estimar mais de uma proporção, devida a Abul-Ela et al. [1]; o modelo de resposta aleatorizada com questão não relacionada, publicado por Greenberg et al. [4]; uma extensão do último modelo ao caso quantitativo (Greenberg et al. [5]).

Com exceção do trabalho de Abul-Ela, todos os demais são especializações do modelo de resposta aleatorizada para $m=2$, seja com um modelo de característica incompleta (Warner), seja com um modelo de característica máxima (os dois trabalhos de Greenberg)

Teoricamente, portanto, todas as propriedades desses casos são obtíveis da análise do modelo de resposta aleatorizada. Todavia, a especialização para $m=2$ permite ir mais além do que se obteria analisando o modelo (1.2.1-4), no seu caso mais geral, porque as expressões simplificam-se, permitindo conclusões de relevância nas aplicações. Além disso, o caso em que $m=2$ ocorre com frequência na prática.

Este capítulo apresentará os modelos citados, com exceção do modelo I para proporções, que foi discutido em (1.1).

2.1 - Extensão do modelo I para estimar mais de uma proporção

Considere-se a população C particionada em m partes C_k ($k=1, \dots, m$). Suponhamos que a pertinência a C_{k_j} seja considerada vexatória ou socialmente reprovada, para no mínimo um e no máximo $m-1$ índices k_j escolhidos entre os inteiros de 1 a m . Em outras palavras, existe pelo menos um conjunto C_{k_j} da partição de C , tal que a pertinência a C_{k_j} não seja vexatória.

Por ser $\{C_k\}$ ($k=1, \dots, m$) uma partição de C , tem-se:

$$\bigcup_{k=1}^m C_k = C$$

$$C_{k_1} \cap C_{k_2} = \emptyset \quad k_1 \neq k_2$$

Chamemos Π_k a proporção de elementos de C que estão em C_k ($k=1, \dots, m$). Decorre, então, que

$$\sum_{k=1}^m \Pi_k = 1 \quad (2.1-1)$$

Associemos a cada elemento de C o vetor aleatório $(X_1, \dots, X_m)$ onde, se tal elemento pertence a C_k , então $X_k=1$ e os demais $X_l=0$ ($l \neq k$).

A distribuição marginal de cada variável aleatória X_k é Bernoulli, de parâmetro Π_k .

De acordo com a notação estabelecida em (1.2), seja $r=m-1$ o número de amostras aleatórias simples, de tamanhos $n_1, \dots, n_r$ obtidas de C . A matriz P é definida como em (1.2), exigindo-se que sua característica seja r .

Então o modelo (1.2.1-4) é imediatamente aplicável:

$$R = P\mu + \epsilon \quad (2.1-2)$$

onde R é o vetor de respostas, constituído de zeros e uns (não ou sim),

$$\mu^T = (\Pi_1, \dots, \Pi_m) \quad e$$

ϵ é o vetor de erro, definido em (1.2).

Na aplicação deste modelo, cada entrevistado recebe um conjunto de perguntas:

$$V. \in C_1?$$

$$V. \in C_2?$$

$$\vdots$$

$$V. \in C_m?$$

O entrevistado procede como descrito em (1.2), selecionando uma questão a que, de acordo com seu caso, responderá "sim" ou "não".

A variável aleatória R_{ij} , resposta do j -ésimo entrevistado na i -ésima amostra ($i=1, \dots, r$; $j=1, \dots, n_i$), tem distribuição de Bernoulli, com parâmetro

$$\lambda_i = \sum_{k=1}^r p_{ik} \Pi_k \quad i=1, \dots, r \quad (2.1-3)$$

Assim, a matriz V é uma matriz diagonal tendo, nos primeiros n_1 elementos da diagonal, $\lambda_1(1-\lambda_1)$, nos n_2 seguintes $\lambda_2(1-\lambda_2)$ e assim por diante.

As equações normais

$$(P^T V^{-1} P) \mu = P^T V^{-1} R$$

tem característica r , de modo que há r funções lineares dos parâmetros linearmente independentes estimáveis.

Note-se que λ_i ($i=1, \dots, r$) é um conjunto de funções estimáveis e são, pela hipótese sobre a característica de P , linearmente independentes. Mais ainda, podemos escrever:

$$\lambda_i = \sum_{k=1}^{n_i} (p_{ik} - p_{im}) \Pi_k + p_{im} \quad (2.1-4)$$

As variáveis aleatórias R_{ij} , por (2.1-3), são tais que

$$E \frac{\sum_{j=1}^{n_i} R_{ij}}{n_i} = \lambda_i$$

de onde decorre que

$$\frac{\sum_{j=1}^{n_i} R_{ij}}{n_i} - p_{im} = \sum_{k=1}^r (p_{ik} - p_{im}) \Pi_k + \eta_i \quad (i=1, \dots, r)$$

(2.1-5)

sendo $\{\eta_i\}$ variáveis aleatórias de médias nulas e

$$v(\eta_i) = \frac{\lambda_i(1-\lambda_i)}{n_i}$$

$$\text{Cov}(\eta_{i_1}; \eta_{i_2}) = 0 \quad i_1 \neq i_2$$

Sejam, então:

$$P^* = (p_{ik} - p_{im}) \quad i, k=1, \dots, r$$

$$Y^T = \left(\frac{\sum_{j=1}^{n_1} R_{1j}}{n_1}, \frac{\sum_{j=1}^{n_2} R_{2j}}{n_2}, \dots, \frac{\sum_{j=1}^{n_r} R_{rj}}{n_r} \right)$$

$$\eta^T = (\eta_1, \dots, \eta_r)$$

$$\mu^* = (\Pi_1, \dots, \Pi_r)$$

O modelo

$$Y = P^* \mu^* + \eta$$

(2.1-6)

constitui uma reparametrização de (2.1-2) e tem característica máxima, r . A matriz P^* é não-singular, de modo que a solução independe da matriz de variância-covariância de η .

Com efeito, seja

$$W = [E(\eta_i \cdot \eta_j)] \quad i=1, \dots, r; \quad j=1, \dots, r$$

O estimador de μ^* é dado por

$$\hat{\mu}^* = (P^{*T} W^{-1} P^*)^{-1} P^{*T} W^{-1} Y$$

Mas, como existe $(P^*)^{-1}$, segue-se que

$$\hat{\mu}^* = P^{*-1} W (P^{*T})^{-1} P^{*T} W^{-1} Y$$

e finalmente,

$$\hat{\mu}^* = (P^*)^{-1} Y \quad (2.1-7)$$

Agora, de (2.1-1) temos

$$\Pi_m = 1 - \sum_{k=1}^r \Pi_k$$

e, assim, o estimador de Π_m é dado por

$$\hat{\Pi}_m = 1 - \sum_{k=1}^r \hat{\Pi}_k \quad (2.1-8)$$

Está completo o vetor $\hat{\mu}$

$$\hat{\mu} = \begin{pmatrix} \hat{\mu}^* \\ \dots \\ \hat{\Pi}_m \end{pmatrix} \quad (2.1-9)$$

que é o melhor estimador linear não-tendencioso de Π .

O estimador (2.1-9) coincide, também, com o estimador de máxima-verossimilhança de μ , deduzido por Abul-Ela et al. [1]. Aqueles autores estudam com algum detalhe o caso em que $m=3$ e concluíram que este método oferece vantagem potencial sobre o método de questão direta - mormente quando a pertinência a alguns dos C_k for ilegal ou socialmente reprovada.

Ao se projetar um plano de amostragem empregando este modelo, deve-se cuidar que a matriz P^* seja não-singular. Isto equivale a fazer P tal que sua característica seja $r=m-1$.

A escolha da matriz P não é tão fácil quanto a escolha do parâmetro p no caso do modelo I para proporções. Por isso, deixamos de fazer qualquer recomendação adicional a este respeito.

2.2 - O modelo de resposta aleatorizada com questão não-relacionada ou modelo II para proporções

Conforme adiantamos em (1.2) este modelo é uma variação do modelo I para proporções em que a questão alternativa é substituída por uma questão que não guarda relação com a primeira. Do mesmo modo que o modelo I para proporções, o modelo II para proporções é um caso particular do modelo de resposta aleatorizada. Diferentemente do I, este tem característica máxima.

O primeiro estudo deste caso do modelo de resposta aleatorizada foi realizado por Greenberg et al [4].

2.2.1 - Formulação do modelo

Seja uma população C , cujos elementos podem ser classificados, segundo um critério, em A ou em $\bar{A}$; segundo outro critério, em B ou em $\bar{B}$. Suponhamos a pertinência a A vexatória, por ser ilegal ou por ser socialmente condenada. A pertinência a B ou a $\bar{B}$ não é vexatória.

Represente-se por:

Π_A = proporção de elementos de C que estão em A.

Π_B = proporção dos elementos de C que pertencem a B.

Os atributos A e B não são relacionados, no sentido de que vale a igualdade

$$\Pi_{AB} = \Pi_A \Pi_B$$

onde Π_{AB} representa a proporção dos elementos de C que estão em A e em B.

Como um exemplo, A pode ser o conjunto das pessoas de uma população que tiveram contato com drogas psicotrópicas ou maconha e B pode ser o conjunto das pessoas dessa população que nasceram em abril.

Associemos a cada elemento de C o vetor aleatório (X_1, X_2) onde X_1 e X_2 tem distribuições de Bernoulli e são tais que

$$X_1 = \begin{cases} 1 & \text{se o elemento considerado pertence a A.} \\ 0 & \text{se o elemento considerado pertence a } \bar{A}. \end{cases}$$

$$X_2 = \begin{cases} 1 & \text{se o elemento considerado pertence a B.} \\ 0 & \text{se o elemento considerado pertence a } \bar{B}. \end{cases}$$

Então o parâmetro da distribuição de X_1 é Π_A e o da distribuição de X_2 é Π_B .

Tomem-se duas amostras aleatórias simples de tamanhos n_1 e n_2 , com $n = n_1 + n_2$, de C e aplique-se o esquema de entrevista descrito em 1.2.

Os elementos definidores do modelo (1.2.1-4), neste caso são:

$$P = \left[\begin{array}{cc} p_1 & 1-p_1 \\ \vdots & \vdots \\ p_1 & 1-p_1 \\ p_2 & 1-p_2 \\ \vdots & \vdots \\ p_2 & 1-p_2 \end{array} \right] \left. \begin{array}{l} \\ \\ \\ \end{array} \right\} \begin{array}{l} n_1 \text{ linhas} \\ \\ n_2 \text{ linhas} \end{array}$$

$$\mu^T = (\Pi_A, \Pi_B)$$

$$R^T = (R_{11}, \dots, R_{1n_1}, R_{21}, \dots, R_{2n_2})$$

onde R_{ij} são variáveis aleatórias de Bernoulli de parâmetros

$$P \{R_{1j} = 1\} = \lambda_1 = p_1 \Pi_A + (1-p_1) \Pi_B \quad j=1, \dots, n_1$$

$$P \{R_{2j} = 1\} = \lambda_2 = p_2 \Pi_A + (1-p_2) \Pi_B \quad j=1, \dots, n_2$$

P é suposta de característica máxima, isto é:

$$p_1 \neq p_2 \tag{2.2.1-1}$$

e ϵ é um vetor aleatório, de média nula, de componentes não-correlacionadas e com variâncias:

$$V(\epsilon_{1j}) = V(R_{1j}) = \lambda_1(1-\lambda_1) = v_1 \quad j=1, \dots, n_1 \tag{2.2.1-2}$$

$$V(\epsilon_{2j}) = V(R_{2j}) = \lambda_2(1-\lambda_2) = v_2 \quad j=1, \dots, n_2$$

Seja V a matriz de covariância de ϵ .

O melhor estimador linear não-tendencioso de μ é dado por

$$\hat{\mu} = (P^T V^{-1} P)^{-1} P^T V^{-1} R$$

A solução explícita pode ser calculada. O sistema de equações normais é

$$\begin{bmatrix} \frac{n_1 p_1^2}{v_1} + \frac{n_2 p_2^2}{v_2} & \frac{n_1 p_1 (1-p_1)}{v_1} + \frac{n_2 p_2 (1-p_2)}{v_2} \\ \frac{n_1 p_1 (1-p_1)}{v_1} + \frac{n_2 p_2 (1-p_2)}{v_2} & \frac{n_1 (1-p_1)^2}{v_1} + \frac{n_2 (1-p_2)^2}{v_2} \end{bmatrix} \begin{bmatrix} \Pi_A \\ \Pi_B \end{bmatrix} = \begin{bmatrix} \frac{p_1}{v_1} \sum_{j=1}^{n_1} R_{1j} + \frac{p_2}{v_2} \sum_{j=1}^{n_2} R_{2j} \\ \frac{1-p_1}{v_1} \sum_{j=1}^{n_1} R_{1j} + \frac{1-p_2}{v_2} \sum_{j=1}^{n_2} R_{2j} \end{bmatrix} \quad (2.2.1-3)$$

A solução de (2.2.1-3) pode ser obtida pelo método de Gramer.

Com efeito, o determinante da matriz de coeficientes é

$$\Delta = \frac{n_1 n_2 (p_1 - p_2)^2}{v_1 v_2}$$

e os determinantes relativos a Π_A e Π_B , calculados após a troca de correspondentes colunas pelo vetor de constantes, são

$$\Delta \Pi_A = \frac{n_2}{v_1 v_2} (p_1 - p_2) (1-p_2) \sum_{j=1}^{n_1} R_{1j} - \frac{n_1}{v_1 v_2} (p_1 - p_2) (1-p_1) \sum_{j=1}^{n_2} R_{2j}$$

$$\Delta \Pi_B = \frac{n_1}{v_1 v_2} p_1 (p_1 - p_2) \sum_{j=1}^{n_2} R_{2j} - \frac{n_2}{v_1 v_2} p_2 (p_1 - p_2) \sum_{j=1}^{n_1} R_{1j}$$

E assim, finalmente, chegamos ao melhor estimador linear não-tendencioso de μ , em forma explícita:

$$\hat{\Pi}_A = \frac{(1-p_2) \frac{\sum_{j=1}^{n_1} R_{1j}}{n_1} - (1-p_1) \frac{\sum_{j=1}^{n_2} R_{2j}}{n_2}}{p_1 - p_2} \quad (2.2.1-4)$$

$$\hat{\Pi}_B = \frac{p_1 \frac{\sum_{j=1}^{n_2} R_{2j}}{n_2} - p_2 \frac{\sum_{j=1}^{n_1} R_{1j}}{n_1}}{p_1 - p_2} \quad (2.2.1-5)$$

Estes estimadores coincidem também com os estimadores dados pelo método de máxima-verossimilhança, conforme se verifica comparando (2.2.1-4) e (2.2.1-5) com as expressões obtidas por Greenberg et al. [4].

As variâncias de $\hat{\Pi}_A$ e de $\hat{\Pi}_B$ são dadas por:

$$V(\hat{\Pi}_A) = \frac{1}{(p_1 - p_2)^2} \left[(1-p_2)^2 \frac{\lambda_1 (1-\lambda_1)}{n_1} + (1-p_1)^2 \frac{\lambda_2 (1-\lambda_2)}{n_2} \right] \quad (2.2.1-6)$$

$$V(\hat{\Pi}_B) = \frac{1}{(p_1 - p_2)^2} \left[p_2^2 \frac{\lambda_1 (1-\lambda_1)}{n_1} + p_1^2 \frac{\lambda_2 (1-\lambda_2)}{n_2} \right] \quad (2.2.1-7)$$

Evidentemente, todo o interesse que temos concentra-se na estimação de Π_A . Interessa, portanto, apenas o comportamento de (2.2.1-6):

2.2.2 - Planejamento de Amostragem

Os parâmetros de que dispomos para otimizar a estimação de Π_A são p_1 , p_2 , n_1 e n_2 , estes sujeitos, eventualmente, à restrição $n = n_1 + n_2$.

Sob a condição de que as respostas sejam corretas, o estimador $\hat{\Pi}_A$ é, por construção, não-tendencioso. Portanto, o que temos a fazer para otimizar a estimação de Π_A é, apenas, procurar minimizar a variância.

O parâmetro p_1 deve ser escolhido de modo a minimizar $|p_1 - 1/2|$. Verificamos empiricamente, através da observação de vários gráficos de $V(\hat{\Pi}_A)$, para valores selecionados dos demais parâmetros, que as menores variâncias ocorrem quando p_1 tende a 0 ou a 1. É claro que valores muito próximos de 0 ou de 1 induzirão o entrevistado a mentir. Vale aqui, então, a mesma regra do modelo I para proporções: p_1 deve ser escolhido de modo a maximizar $|p_1 - 1/2|$, sem diminuir a colaboração dos entrevistados.

Fixado p_1 , a escolha ótima de p_2 pode ser deduzida da observação dos gráficos de $V(\hat{\Pi}_A)$, considerada como função de p_2 somente. Os gráficos tem aspecto duplamente hiperbólico, com uma assíntota em $p_1 = p_2$, onde $V(\hat{\Pi}_A)$ diverge. Para $|p_1 - p_2|$ crescente, $V(\hat{\Pi}_A)$ decresce de forma monótona. É claro que p_2 não pode ser escolhido mais próximo de 0 ou de 1 do que p_1 , pela própria escolha de p_1 . Portanto, uma eleição razoável para p_2 é

$$p_2 = 1 - p_1$$

(2.2.2-1)

A amostra pode ser repartida em suas duas partes, de tamanhos n_1 e n_2 , de modo a minimizar $V(\hat{\Pi}_A)$. Suponhamos então, que $n = n_1 + n_2$ é fixado a priori e calcularemos n_1 e n_2 de modo a atender a exigência de $V(\hat{\Pi}_A)$ mínima.

Da observação de (2.2.1-6), vemos que $V(\hat{\Pi}_A)$ é dada pela soma de duas parcelas:

$$V(\hat{\Pi}_A) = K \left[\frac{\lambda_1(1-\lambda_1)(1-p_2)^2}{n_1} + \frac{\lambda_2(1-\lambda_2)(1-p_1)^2}{n_2} \right] \quad (2.2.2-2)$$

onde K independe de n_1 e de n_2 .

Achar o ponto (n_1, n_2) que minimiza (2.2.2-2), sujeito à restrição,

$$n = n_1 + n_2 \quad (2.2.2-3)$$

é resolvido facilmente empregando multiplicadores de Lagrange.

A solução pode ser encontrada desde que

$$\lambda_1(1-\lambda_1)(1-p_2)^2 \neq \lambda_2(1-\lambda_2)(1-p_1)^2 \quad (2.2.2-4)$$

e é dada por

$$n_1 = \frac{n \frac{\lambda_1(1-\lambda_1)(1-p_2)^2}{\lambda_2(1-\lambda_2)(1-p_1)^2}}{1 + \frac{\lambda_1(1-\lambda_1)(1-p_2)^2}{\lambda_2(1-\lambda_2)(1-p_1)^2}} \quad (2.2.2-5)$$

$$n_2 = n - n_1$$

Se a desigualdade (2.2.2-4) não se verifica, qualquer solução de (2.2.2-3) serve para nosso propósito.

Para a aplicação de (2.2.2-5) é necessário ter-se uma idéia da localização de Π_A e de Π_B , que são supostos desconhecidos. Qualquer informação de que dispomos deve ser empregada para repartir a amostra, posto que o ganho em eficiência pode ser

apreciável, comparado com a escolha de $n_1 = n_2$.

A escolha da questão não-relacionada pode também, eventualmente, ser feita de modo a minimizar $V(\hat{\Pi}_A)$. Verificamos que, se $\Pi_A < 0,5$, $V(\hat{\Pi}_A)$, como função de Π_B , decresce quando Π_B decresce; se $\Pi_A > 0,5$, $V(\hat{\Pi}_A)$ decresce com Π_B crescente. Deste modo, deve-se escolher Π_B mínimo quando $\Pi_A < 0,5$ e máximo quando $\Pi_A > 0,5$. Todavia, segundo a observação de Greenberg et al. [4], esta proporção não deve ser muito próxima dos extremos 0 ou 1, por que isto pode revelar as respostas dos entrevistados, criando resistência por sua parte.

Na escolha dos parâmetros p_1 , n_1 , n_2 e possivelmente, Π_B , devemos ter o maior cuidado em obter a colaboração do entrevistado e só secundariamente, dentro desta condição, minimizar $V(\hat{\Pi}_A)$.

2.2.3 - Caso em que Π_B é conhecida

Quando Π_B é conhecida, basta uma única amostra (i.e., $r=1$) para estimar Π_A . Um dos modos de assegurar o conhecimento de Π_B é ter resultados de um censo; porém, o mais seguro é o processo citado em Greenberg et al. [5], consistindo numa alteração da caixa para selecionar as questões (vd. exemplo em 1.1.1) que conteria três tipos de bolas, azuis, brancas e amarelas, em proporções conhecidas; se a bola sorteada for branca, o entrevistado deve responder "sim" ou "não" sobre a pertinência a A; se for azul deve responder "não" e se for amarela, "sim". A proporção Π_B equivale à razão entre o número de bolas amarelas e o total de bolas não brancas.

Com apenas uma amostra e Π_B conhecida, temos um caso formalmente semelhante ao modelo 1 para proporções, com

$$P = \begin{bmatrix} p_1 & 1-p_1 \\ \vdots & \vdots \\ p_1 & 1-p_1 \end{bmatrix} \quad \mu = \begin{bmatrix} \Pi_A \\ \Pi_B \end{bmatrix}$$

$$R^T = (R_1, \dots, R_n)$$

onde $\{R_i\}$ são variáveis aleatórias de Bernoulli, de parâmetro

$$\lambda = \Pi_A p_1 + \Pi_B (1-p_1)$$

Π_A é estimável e seu estimador é dado por

$$(\hat{\Pi}_A / \Pi_B) = \frac{\frac{\sum_{i=1}^n R_i}{n} - \Pi_B (1-p_1)}{p_1} \quad (2.2.3-1)$$

com variância

$$V(\hat{\Pi}_A / \Pi_B) = \frac{\lambda(1-\lambda)}{np_1^2} \quad (2.2.3-2)$$

Há um ganho considerável em eficiência quando Π_B é conhecida, conforme mostraremos na seção seguinte. Quando há erro no valor de Π_B , $(\hat{\Pi}_A / \Pi_B)$ torna-se tendencioso (Greenberg et al. [4]), mas a eficiência deste estimador não é muito sensível para pequenos erros em Π_B .

2.2.4 - Eficiência do modelo II para proporções

As tabelas a seguir mostram que o modelo II para proporções pode ser mais eficiente do que o modelo em (1.1), se a probabilidade de resposta correta é aumentada ou, pelo menos, não é diminuída. Greenberg et al. [4] argumentam, a este respeito, que o modelo II favorece psicologicamente a colaboração dos entrevistados por lhes dar a oportunidade de responder a uma questão não-relacionada com o atributo A. Cabe lembrar que, realmente, as duas questões do modelo I contam, no fundo, a mesma coisa. Neste caso as

respostas "sim" e "não" são equivalentes. Mas no modelo II, como observam Greenberg et al. [4], uma resposta "sim" pode ser embaraçosa, ao passo que uma resposta "não" nunca é embaraçosa. Aachamos, portanto, que só a experiência pode mostrar, em cada caso, qual o modelo que mais favorece a colaboração do entrevistado.

2.2.4-1 - EFICIÊNCIA RELATIVA ENTRE OS ESTIMADORES(*)
DO MODELO II E DO MODELO I

P(1)	$\Pi_A = 0,2$ (3)				$\Pi_A = 0,4$ (3)			
	$\Pi_B = 0,1$	$\Pi_B = 0,3$	$\Pi_B = 0,7$	$\Pi_B = 0,9$	$\Pi_B = 0,1$	$\Pi_B = 0,3$	$\Pi_B = 0,7$	$\Pi_B = 0,9$
0,1	1,30	1,13	1,00	1,00	1,11	1,03	1,00	1,00
0,2	1,56	1,22	1,00	1,00	1,20	1,06	1,00	1,00
0,3	1,77	1,28	1,00	1,00	1,27	1,08	1,00	1,00
0,4	1,91	1,32	1,00	1,00	1,31	1,09	1,00	1,00

(1) $p_1 = p, \quad p_2 = 1-p_1$

(2) Supondo respostas corretas em ambos os modelos e repartição ótima da amostra no modelo II.

(3) A tabela é simétrica para $\Pi_A - 1/2$ e $p - 1/2$

(4) Π_B suposta desconhecida

A tabela 2.2.4-1 mostra que o modelo II é, sempre, pelo menos tão eficiente quanto o modelo I, supondo que as respostas são corretas sob ambos os modelos e que a repartição das amostras no modelo II é ótima; em suma, quando os dois modelos estão no máximo de suas possibilidades.

(*) eficiência em termos do quociente entre os erros médios quadráticos, como em 1.1.3.

Quando isso não ocorre, a situação pode inverter-se. Com efeito, se $n_1=n_2$, para $\Pi_A = 0,2, p=0,1$ e $\Pi_B = 0,1$, a eficiência relativa cai de 1,30 para 0,77.

Cabe analisar, ainda, a eficiência relativa entre o estimador do modelo II com o do modelo I, quando as probabilidades de respostas corretas, respectivamente T_I e T_{II} , variam nos dois modelos.

2.2.4-2 - EFICIÊNCIA RELATIVA ENTRE OS ESTIMADORES DO MODELO II E DO MODELO I SOB RESPOSTAS FALSAS

$$\Pi_A = 0,3 \quad p=p_1 = 1-p_2=0,2 \quad \Pi_B=0,1$$

$$n = 1000; \quad n_1=n_2=500$$

$T_{II} \backslash T_I$	1,00	0,90	0,80	0,50	0,40
1,00	1,39	3,28	9,09	49,02	69,98
0,90	0,48	1,13	3,12	17,06	24,36
0,80	0,16	0,38	1,04	5,72	8,17
0,50	0,03	0,07	0,18	1,01	1,43
0,40	0,02	0,05	0,13	0,70	1,01

T_I = probabilidade de resposta correta no modelo I.

T_{II} = probabilidade de resposta correta no modelo II

Da tabela 2.2.4-2 vemos que o modelo II para proporções é mais eficiente do que o modelo I ainda que $T_I=T_{II}$.

Os cálculos da tabela 2.2.4-1 foram feitos sem repartição ótima da amostra e supondo Π_B desconhecida. Evidentemente, para T_I e T_{II} fixados, a adoção de Π_B conhecido e repartição ótima pode

apenas aumentar a eficiência relativa entre o estimador do modelo II e o do modelo I.

Para termos uma idéia do ganho em eficiência quando empregamos um Π_B conhecido, calculamos diversos valores da eficiência relativa entre os estimadores do modelo II para proporções, com Π_B conhecida e com Π_B desconhecida.

2.2.4-3 - EFICIÊNCIA RELATIVA ENTRE O ESTIMADOR DO MODELO II COM Π_B CONHECIDA E O ESTIMADOR DO MODELO II COM Π_B DESCONHECIDA

Π_B	$\Pi_A = 0,1$					$\Pi_A = 0,3$				
	p=0,1	p=0,3	p=0,6	p=0,7	p=0,8	p=0,1	p=0,3	p=0,6	p=0,7	p=0,8
0,1	0,03	0,65	9,36	3,55	2,42	0,05	0,85	8,96	3,41	2,36
0,3	0,01	0,51	9,83	3,75	2,50	0,03	0,65	9,36	3,55	2,42
0,5	0,01	0,50	9,84	3,76	2,51	0,02	0,62	9,46	3,59	2,43
0,7	0,02	0,55	9,67	3,68	2,48	0,03	0,65	9,36	3,55	2,42
0,9	0,03	0,65	9,36	3,55	2,42	0,04	0,78	9,08	3,45	2,38

Da tabela 2.2.4-3 aprendemos que é fundamental usarmos p do lado contrário, em relação a 1/2, de Π_A . Isto é, se $\Pi_A < 1/2$, p deve ser superior a 1/2 para este caso do modelo II (Π_B conhecida) ser mais eficiente do que o caso de Π_B desconhecida; se $\Pi_A > 1/2$, p deve ser tomado menor do que 1/2.

2.3 - Caso quantitativo

2.3.1 - O modelo

Consideremos o problema de estimar a renda média de certa população (sujeita a imposto sobre a renda). Aqui os indivíduos, se

perguntados, poderão responder um número falso, seja por vaidade ou modéstia, seja por ter feito declaração fiscal falsa. Ainda mais grave seria este efeito se o problema fosse a estimação do número médio de abortos induzidos.

O modelo de resposta aleatorizada é imediatamente aplicável a casos como os acima descritos, mais nada precisando acrescentar. Contudo, especializaremos o modelo para a estimação da média de uma variável aleatória, X_1 , cuja revelação seja crítica.

Suponhamos que seja possível associar a C uma outra variável aleatória, expressa na mesma unidade de X_1 e com distribuição de probabilidade assemelhada. Chamemos de X_2 tal variável aleatória e suponhamos que a revelação de X_2 não seja crítica.

Obtenhamos, então, $r=2$ amostras aleatórias, de tamanhos n_1 e n_2 , de C e apliquemos entrevistas segundo o modelo de resposta aleatorizada, com probabilidades: p_1 e $1-p_1$; p_2 e $1-p_2$, respectivamente na primeira e na segunda amostras.

Um exemplo de como se obtém X_2 é ilustrado na pesquisa levada a cabo por Greenberg et al. [5], sobre a renda na Carolina do Norte (EUA), em que as perguntas foram:

- 1 - Aproximadamente quanto dinheiro, em dólares, ganhou o chefe desta família o ano passado?
- 2 - Quanto você acha que um chefe de uma família como a sua precisa ganhar por ano?

Temos, então, um caso absolutamente idêntico a 2.2, com a diferença única que X_1 e X_2 não são mais variáveis aleatórias de Bernoulli, e o vetor de médias é geral, não representando proporções:

$$\mu^T = (\mu_1, \mu_2)$$

Claro está que o vetor R , de respostas, é também geral e estará expresso na mesma unidade de X_1 (e de X_2).

As equações normais 2.2.1-3 valem aqui e a solução é, portanto, idêntica, mudando apenas o vetor μ . Temos assim que o me-

lhör estimador linear não-tendencioso de μ é dado por:

$$\hat{\mu}_1 = \frac{(1-p_2) \frac{\sum_{j=1}^{n_1} R_{1j}}{n_1} - (1-p_1) \frac{\sum_{j=1}^{n_2} R_{2j}}{n_2}}{p_1 - p_2} \quad (2.3.1-1)$$

$$\mu_2 = \frac{p_1 \frac{\sum_{j=1}^{n_2} R_{2j}}{n_2} - p_2 \frac{\sum_{j=1}^{n_1} R_{1j}}{n_1}}{p_1 - p_2} \quad (2.3.1-2)$$

Ainda de acordo com a notação de 1.2, se

$$\sigma_1^2 = V(X_1) \quad \text{e} \quad \sigma_2^2 = V(X_2),$$

empregando (1.2.1-3), temos

$$V(R_{ij}) = \sigma_2^2 + p_i(\sigma_1^2 - \sigma_2^2) + p_i(1-p_i)(\mu_1 - \mu_2)^2 \quad (2.3.1-3)$$

$i=1,2$

Cabe lembrar que, na solução de (2.2.1-3), as variâncias foram levadas em conta, de forma geral, e simplificaram-se. Assim, o fato de as variâncias não coincidirem com as do modelo II para proporções não impossibilita o emprego da solução obtida para (2.2.1-3), pois esta é independente das variâncias.

Seja

$$\bar{R}_i = \frac{\sum_{j=1}^{n_i} R_{ij}}{n_i}$$

Então $V(\bar{R}_i) = V(R_{ij})/n_i$ V_j

As variâncias dos estimadores 2.3.1 e 2.3.2 são dadas por

$$V(\hat{\mu}_1) = \frac{1}{(p_1 - p_2)^2} \{ (1-p_2)^2 V(\bar{R}_1) + (1-p_1)^2 V(\bar{R}_2) \} \quad (2.3.1-4)$$

$$V(\hat{\mu}_2) = \frac{1}{(p_1 - p_2)^2} \{ p_2^2 V(\bar{R}_1) + p_1^2 V(\bar{R}_2) \} \quad (2.3.1-5)$$

2.3.2 - Nota sobre o planejamento

Todas as consequências estudadas em 2.2 são válidas aqui. Todavia, (2.3.1-3) nos permite ver que $V(\hat{\mu}_1)$ cresce com σ_2^2 e com $(\mu_1 - \mu_2)^2$. Assim, devemos escolher X_2 de tal modo que sua média seja tão próxima quanto possível de μ_1 . Já quanto a σ_2^2 não deve ser inferior a σ_1^2 para que os entrevistados não tenham receio de responder valores extremos de X_1 .

Procedimento análogo ao de 2.2 nos leva a repartir a amostra em n_1 e n_2 , tais que

$$\frac{n_1}{n_2} = \sqrt{\frac{(1-p_2)^2 V(\bar{R}_1)}{(1-p_1)^2 V(\bar{R}_2)}}$$

com $n = n_1 + n_2$. Na aplicação desta fórmula, precisamos saber aproximadamente os valores de μ_1 , μ_2 , σ_1^2 e σ_2^2 , o que, em geral, é bem mais difícil do que no modelo II para proporções. Todavia, se a variável aleatória X_2 foi bem escolhida, no sentido de que $\sigma_1^2 \approx \sigma_2^2$, então a seguinte expressão pode ser usada:

$$\frac{n_1}{n_2} = \frac{p_1}{p_2} \quad (\text{com } p_1 + p_2 = 1)$$

$$n = n_1 + n_2$$

segundo Greenberg et al. [5].

As dificuldades na seleção da variável aleatória X_2 são consideravelmente maiores do que no modelo II para proporções. Então, a aplicação do modelo de resposta aleatorizada, em cada caso, deve ser criteriosamente examinada, a fim de obter estimativas por meio de estimadores eficientes.

3 - OUTRO MODELO PARA ESTIMAR AS PROPORÇÕES DE DOIS ATRIBUTOS

3.1 - Formulação do modelo

Considere-se uma população C particionada, segundo dois critérios, em A e $\bar{A}$ e em B e $\bar{B}$. Suponhamos a pertinência a A e/ou a B vexatória, ilegal ou socialmente condenada. Por exemplo, A pode ser o conjunto de mulheres que provocaram aborto no último ano e B o conjunto de mães solteiras.

Trataremos aqui de um outro modelo para estimar:

Π_A = proporção de elementos de C que estão em A

Π_B = proporção de elementos de C que estão em B

Π_{AB} = proporção de elementos de C que estão em A e em B

Tal modelo não se enquadra na formulação de 1.2, porque se baseia em duas perguntas por entrevista, cada uma das quais aleatorizada conforme descrito em 1.1.

O modelo de resposta aleatorizada aplica-se também neste caso. Consideremos os conjuntos mutuamente exclusivos:

$A-B$ = de elementos de C que estão em A , mas não em B

$B-A$ = de elementos de C que estão em B , mas não em A

AB = de elementos de C que estão em A e em B

$\overline{A \cup B}$ = de elementos de C que não estão em A nem em B

e as respectivas proporções Π_{A-B} , Π_{B-A} , Π_{AB} , $\Pi_{\overline{A \cup B}}$, as quais somam 1.

Podemos, então, aplicar o modelo de resposta aleatorizada e estimar Π_A , Π_B e Π_{AB} por

$$\hat{\Pi}_A = \hat{\Pi}_{A-B} + \hat{\Pi}_{AB}$$

$$\hat{\Pi}_B = \hat{\Pi}_{B-A} + \hat{\Pi}_{AB} ,$$

onde as proporções à direita são estimadas por 2.1-9.

Do ponto de vista prático, a aplicação do modelo de resposta aleatorizada exige a formulação de quatro perguntas a serem selecionadas aleatoriamente. Tais perguntas referir-se-ão à inclusão (ou exclusão) a dois conjuntos, prejudicando, talvez, o entendimento do entrevistado.

Por exemplo, se A e B são os conjuntos descritos acima, as questões seriam:

- 1- você induziu aborto no último ano e não é mãe solteira?
- 2- você não induziu aborto no último ano e é mãe solteira?
- 3- você induziu aborto no último ano e é mãe solteira?
- 4- você não induziu aborto no último ano e não é mãe solteira?

Admitimos que tais questões poderiam ser melhor formuladas, mas isto só prova que a formulação (e que dizer do entendimento!) não é trivial.

Para contornar tal problema, separamos as questões em duas aleatorizadas:

- 1- v. induziu aborto no último ano?
v. não induziu aborto no último ano?
- 2- v. é mãe solteira?
v. não é mãe solteira?

cada uma das quais é respondida como em 1.1, isto é, selecionadas ao acaso, sendo p_1 a probabilidade de selecionar a pergunta afirmativa na questão 1 e p_2 a probabilidade de selecionar a pergunta afirmativa na questão 2.

3.2 - Estimação de Π_A , Π_B e Π_{AB}

Tome-se uma amostra aleatória simples de tamanho n de C e a cada elemento ofereça-se o par de questões aleatorizadas:

1- $V. \in A ?$

$V. \notin A ?$

2- $V. \in B ?$

$V. \notin B ?$

As respostas são constituídas de pares escolhidos de {sim; não}. As possíveis respostas são:

{SIM; SIM}, {SIM,NÃO}, {NÃO,SIM}, {NÃO,NÃO}

Podemos calcular as probabilidades de ocorrência de cada uma das possíveis respostas, em função dos parâmetros populacionais e amostrais:

$$\begin{aligned}
 \lambda_{11} &= P\{SIM;SIM\} = p_1 p_2 \Pi_{AB} + p_1 (1-p_2) (\Pi_A - \Pi_B) + \\
 &\quad + (1-p_1) p_2 (\Pi_B - \Pi_{AB}) + (1-p_1) (1-p_2) (1-\Pi_A - \Pi_B + \Pi_{AB}) \\
 \lambda_{10} &= P\{SIM;NÃO\} = p_1 p_2 (\Pi_A - \Pi_{AB}) + p_1 (1-p_2) \Pi_{AB} + \\
 &\quad + (1-p_1) p_2 (1-\Pi_A - \Pi_B + \Pi_{AB}) + (1-p_1) (1-p_2) (\Pi_B - \Pi_{AB}) \\
 \lambda_{01} &= P\{NÃO;SIM\} = p_1 p_2 (\Pi_B - \Pi_{AB}) + p_1 (1-p_2) (1-\Pi_A - \Pi_B + \Pi_{AB}) + \\
 &\quad + (1-p_1) p_2 \Pi_{AB} + (1-p_1) (1-p_2) (\Pi_A - \Pi_{AB}) \\
 \lambda_{00} &= P\{NÃO;NÃO\} = p_1 p_2 (1-\Pi_A - \Pi_B + \Pi_{AB}) + p_1 (1-p_2) (\Pi_B - \Pi_{AB}) + \\
 &\quad + (1-p_1) p_2 (\Pi_A - \Pi_{AB}) + (1-p_1) (1-p_2) \Pi_{AB}
 \end{aligned} \tag{3.2-1}$$

Agrupando os coeficientes de Π_A , Π_B e Π_{AB} , (3.2-1) pode ser escrito

$$\left\{ \begin{array}{l} (1-p_2)(2p_1-1)\Pi_A + (1-p_1)(2p_2-1)\Pi_B + (2p_1-1)(2p_2-1)\Pi_{AB} = \lambda_{11} - (1-p_1)(1-p_2) \\ p_2(2p_1-1)\Pi_A + (1-p_1)(1-2p_2)\Pi_B + (1-2p_1)(2p_2-1)\Pi_{AB} = \lambda_{10} - (1-p_1)p_2 \\ (1-p_2)(1-2p_1)\Pi_A + p_1(2p_2-1)\Pi_B + (1-2p_1)(2p_2-1)\Pi_{AB} = \lambda_{01} - p_1(1-p_2) \\ p_2(1-2p_1)\Pi_A + p_1(1-2p_2)\Pi_B + (2p_1-1)(2p_2-1)\Pi_{AB} = \lambda_{00} - p_1p_2 \end{array} \right. \quad (3.2-2)$$

Este sistema é compatível, de modo que pode ser reduzido ao sistema equivalente

$$\left\{ \begin{array}{l} (2p_1-1)\Pi_A = \lambda_{11} + \lambda_{10} + p_1 - 1 \\ (2p_2-1)\Pi_B = \lambda_{11} + \lambda_{01} + p_2 - 1 \\ p_2(2p_1-1)\Pi_A + p_1(2p_2-1)\Pi_B + (2p_1-1)(2p_2-1)\Pi_{AB} = \lambda_{00} - p_1p_2 \end{array} \right. \quad (3.2-3)$$

Sejam:

n_{11} = número de respostas {SIM;SIM} na amostra

n_{10} = número de respostas {SIM;NÃO} na amostra

n_{01} = número de respostas {NÃO;SIM} na amostra

n_{00} = número de respostas {NÃO;NÃO} na amostra

e adotemos a convenção de representar a soma de variáveis multi-indexadas pela substituição do índice de somação por um ponto. Assim, por exemplo:

$$\lambda_{11} + \lambda_{10} = \lambda_{1.}$$

$$n_{11} + n_{01} = n_{.1}$$

Definamos os vetores

$$\Pi^T = (\Pi_A, \Pi_B, \Pi_{AB})$$

$$Y = \begin{bmatrix} \frac{n_{1.}}{n} + p_1^{-1} \\ \frac{n_{.1}}{n} + p_2^{-1} \\ \frac{n_{00}}{n} + p_1 p_2 \end{bmatrix}$$

e seja P a matriz de coeficientes do sistema (3.2-3). Tal sistema pode, então, ser escrito:

$$E(Y) = P\Pi$$

o que define um modelo linear, de característica máxima. O melhor estimador linear não-tendencioso de Π , neste caso particular, é dado por

$$\hat{\Pi} = P^{-1} Y \quad (3.2-4)$$

desde que a matriz de covariância de Y seja positiva definida. Lembremos que, se P é não-singular, (3.2-3) é válida qualquer que seja $V(Y)$, suposta positiva definida.

Os componentes de $\hat{\Pi}$ podem ser explicitados:

$$\hat{\Pi}_A = \frac{1}{2p_1^{-1}} \left\{ p_1^{-1} + \frac{n_{1.}}{n} \right\} \quad (3.2-5)$$

$$\hat{\Pi}_B = \frac{1}{2p_2-1} \{ p_2^{-1} + \frac{n \cdot 1}{n} \} \quad (3.2-6)$$

$$\hat{\Pi}_{AB} = \frac{1}{(2p_1-1)(2p_2-1)} \{ \frac{n_{00}}{n} + p_2 \left[p_1^{-1} + \frac{n \cdot 1}{n} \right] + p_1 \left[p_2^{-1} + \frac{n \cdot 1}{n} \right] - p_1 p_2 \} \quad (3.2-7)$$

Notemos que $\hat{\Pi}_A$ e $\hat{\Pi}_B$ coincidem com o estimador do modelo I para proporções, cada questão considerada separadamente.

Podemos mostrar que (3.2-3) é também o estimador de máxima-verossimilhança de Π .

Com efeito, Y é o estimador de máxima-verossimilhança de $P\Pi$, como se vê facilmente. Como P é não-singular decorre que o estimador de máxima-verossimilhança de Π é (3.2-3).

Matriz de Covariância dos Estimadores

Dois momentos de segunda ordem de $\hat{\Pi}$ já são conhecidos: do fato de que $\hat{\Pi}_A$ e $\hat{\Pi}_B$ são formalmente idênticos ao estimador do modelo I para proporções decorre que:

$$V(\hat{\Pi}_A) = \frac{\Pi_A(1-\Pi_A)}{n} + \frac{p_1(1-p_1)}{n(2p_1-1)^2} \quad (3.2-8)$$

$$V(\hat{\Pi}_B) = \frac{\Pi_B(1-\Pi_B)}{n} + \frac{p_2(1-p_2)}{n(2p_2-1)^2} \quad (3.2-9)$$

Os restantes momentos de segunda ordem, a saber: $V(\hat{\Pi}_{AB})$ e as tres covariâncias, tem de ser calculados.

Consideremos as probabilidades λ_{11} , λ_{10} , λ_{01} , λ_{00} e também $\lambda_{1.}$ e $\lambda_{.1}$, bem como os estimadores respectivos, representados pelos mesmos símbolos, acrescidos de um "hat".

Calcularemos a matriz de covariância de $\hat{\lambda}_{1.}$, $\hat{\lambda}_{.1}$, $\hat{\lambda}_{00}$.

A partir daí, é fácil obter a matriz de covariância de $\hat{\Pi}$.

As variâncias de $\hat{\lambda}_{1.}$, $\hat{\lambda}_{.1}$ e $\hat{\lambda}_{00}$ são imediatas:

$$v(\hat{\lambda}_{1.}) = \frac{\lambda_{1.}(1-\lambda_{1.})}{n} \quad (3.2-10)$$

$$v(\hat{\lambda}_{.1}) = \frac{\lambda_{.1}(1-\lambda_{.1})}{n} \quad (3.2-11)$$

$$v(\hat{\lambda}_{00}) = \frac{\lambda_{00}(1-\lambda_{00})}{n} \quad (3.2-12)$$

Cálculo de Cov ($\hat{\lambda}_{1.}$, $\hat{\lambda}_{00}$)

Temos a relação

$$\hat{\lambda}_{00} + \hat{\lambda}_{1.} = \hat{\lambda}_{00} + \hat{\lambda}_{10} + \hat{\lambda}_{11} = 1 - \hat{\lambda}_{01}$$

Assim,

$$v(\hat{\lambda}_{00} + \hat{\lambda}_{1.}) = v(\hat{\lambda}_{00}) + v(\hat{\lambda}_{1.}) + 2\text{Cov}(\hat{\lambda}_{1.}, \hat{\lambda}_{00}) = v(\hat{\lambda}_{01})$$

ou seja

$$\text{Cov}(\hat{\lambda}_{1.}, \hat{\lambda}_{00}) = \frac{v(\hat{\lambda}_{01}) - v(\hat{\lambda}_{00}) - v(\hat{\lambda}_{1.})}{2}$$

Finalmente:

$$\text{Cov}(\hat{\lambda}_{1.}, \hat{\lambda}_{00}) = \frac{\lambda_{00}(\lambda_{0.}-1)}{2} \quad (3.2-13)$$

Cálculo de Cov. ($\hat{\lambda}_{00}$, $\hat{\lambda}_{.1}$)

Procedemos como acima, observando que

$$\hat{\lambda}_{.1} + \hat{\lambda}_{00} = \hat{\lambda}_{01} + \hat{\lambda}_{11} + \hat{\lambda}_{00} = 1 - \hat{\lambda}_{10}$$

Obtemos:

$$\text{Cov}(\hat{\lambda}_{00}, \hat{\lambda}_{.1}) = \frac{\lambda_{00}(\lambda_{.0}-1)}{2} \quad (3.2-14)$$

$$\text{Cálculo de Cov}(\hat{\lambda}_{1.}, \hat{\lambda}_{.1})$$

Este momento é facilmente calculável a partir da própria definição. Basta notar que

$$E(\hat{\lambda}_{1.}, \hat{\lambda}_{.1}) = \frac{\lambda_{11}}{n}$$

de onde obtemos diretamente

$$\text{Cov}(\hat{\lambda}_{1.}, \hat{\lambda}_{.1}) = \frac{\lambda_{11} - \lambda_{.1} \lambda_{1.}}{n} \quad (3.2-15)$$

A matriz de covariância de $\hat{\Pi}$ pode então ser completada, em função das expressões de (3.2-9) a (3.2-14):

$$\text{Cov}(\hat{\Pi}_A, \hat{\Pi}_B) = \frac{\text{Cov}(\hat{\lambda}_{1.}, \hat{\lambda}_{.1})}{(2p_1-1)(2p_2-1)} \quad (3.2-16)$$

$$\text{Cov}(\hat{\Pi}_A, \hat{\Pi}_{AB}) = \frac{\text{Cov}(\hat{\lambda}_{1.}, \hat{\lambda}_{00}) + p_2 V(\hat{\lambda}_{1.}) + p_1 \text{Cov}(\hat{\lambda}_{1.}, \hat{\lambda}_{.1})}{(2p_1-1)^2 (2p_2-1)} \quad (3.2-17)$$

$$\text{Cov}(\hat{\Pi}_B, \hat{\Pi}_{AB}) = \frac{\text{Cov}(\hat{\lambda}_{.1}, \hat{\lambda}_{00}) + p_1 V(\hat{\lambda}_{.1}) + p_2 \text{Cov}(\hat{\lambda}_{.1}, \hat{\lambda}_{1.})}{(2p_1-1)(2p_2-1)^2} \quad (3.2-18)$$

$$\begin{aligned} V(\hat{\Pi}_{AB}) = & \frac{1}{(2p_1-1)^2 (2p_2-1)^2} V(\hat{\lambda}_{00}) + p_2^2 V(\hat{\lambda}_{1.}) + p_1^2 V(\hat{\lambda}_{.1}) + \\ & + 2 p_2 \text{Cov}(\hat{\lambda}_{00}, \hat{\lambda}_{1.}) + p_1 \text{Cov}(\hat{\lambda}_{00}, \hat{\lambda}_{.1}) + p_1 p_2 \text{Cov}(\hat{\lambda}_{1.}, \hat{\lambda}_{.1}) \end{aligned} \quad (3.2-19)$$

(3.2-15) pode ainda ser escrita:

$$\text{Cov}(\hat{\Pi}_A, \hat{\Pi}_B) = \frac{\Pi_{AB} - \Pi_A \cdot \Pi_B}{n} \quad (3.2-16')$$

que, verificamos, não depende de p_1 nem de p_2 .

É possível chegar a expressões explícitas para estes momentos, em termos de Π , p_1 e p_2 . Todavia, isto é extremamente trabalhoso.

As fórmulas (3.2-15) a (3.2-18), juntamente com (3.2-7) e (3.2-8) já permitem o cálculo da matriz de covariância de $\hat{\Pi}$, bem como permite estimá-la (substituindo os " λ " pelos correspondentes estimadores).

3.3 - Planejamento de amostragem

Procedemos conforme descrito em 1.1, para a seleção de p_1 e de p_2 . Com isto, minimizamos $V(\hat{\Pi}_A)$ e $V(\hat{\Pi}_B)$.

Note-se que $V(\hat{\Pi}_{AB})$ não é definida para $p_1=1/2$ ou $p_2=1/2$, divergindo quando qualquer destes dois parâmetros tende a $1/2$. É possível, então, que a regra dada em 1.1 seja adequada para a estimação de Π_{AB} . Todavia, isto será matéria para estudo posterior.

...ooo...

A P Ê N D I C E

Ineficiência do estimador do método usual quando há respostas falsas

O esquema usual de entrevista consiste em perguntar a cada elemento da amostra

V. $\in$ A ?

Se a pertinência a A é de revelação crítica, então um elemento de A, entrevistado segundo o método usual, dirá a verdade com probabilidade T, geralmente menor do que 1.

O estimador de Π , proporção de elementos de A na população é

$$\tilde{\Pi} = \frac{\sum_i X_i}{n},$$

onde X_i é a variável aleatória de Bernoulli, afixo do evento "resposta sim na i-ésima entrevista".

Supomos que os elementos de $\bar{A}$, não tem razões para ocultar sua situação.

Se $T < 1$, $\tilde{\Pi}$ é um estimador tendencioso para Π , com tendenciosidade (Cochran [2]).

$$B(\tilde{\Pi}/T) = \Pi(T-1)$$

O erro médio quadrático de $\tilde{\Pi}$ é dado por

$$EMQ(\tilde{\Pi}/T) = \frac{\Pi(1-\Pi)}{n} + B^2(\tilde{\Pi}/T)$$

Note-se que a variância mínima para a estimação de Π , dada pela desigualdade de Cramér-Rao é

$$V_{\min} = \frac{\Pi(1-\Pi)}{n}$$

A variância de $\tilde{\Pi}$ atinge esse mínimo quando $T=1$, isto é, quando não há respostas falsas.

A eficiência de $\tilde{\Pi}$

$$ef(\tilde{\Pi}/T) = V_{\min} / [V_{\min} + B^2(\tilde{\Pi}/T)]$$

pode ser escrita:

$$ef(\tilde{\Pi}/T) = \frac{1}{1 + \frac{B^2(\tilde{\Pi}/T)}{V_{\min}}}$$

Portanto $ef(\tilde{\Pi}/T)$ decresce com T decrescente. A tabela seguinte ilustra essa situação

EFICIÊNCIA DO ESTIMADOR USUAL, SOB RESPOSTAS FALSAS

Π	0,1		0,2		0,3		0,4		0,5	
	TEND	EFIC	TEND	EFIC	TEND	EFIC	TEND	EFIC	TEND	EFIC
1,0	-	1,000	-	1,000	-	1,000	-	1,000	-	1,000
0,5	-0,05	0,035	-0,10	0,015	-0,15	0,009	-0,20	0,005	-0,25	0,003
0,4	-0,06	0,024	-0,12	0,011	-0,18	0,006	-0,24	0,004	-0,30	0,002
0,3	-0,07	0,018	-0,14	0,008	-0,21	0,004	-0,28	0,003	-0,35	0,002
0,2	-0,08	0,014	-0,16	0,006	-0,24	0,003	-0,32	0,002	-0,40	0,001
0,1	-0,09	0,011	-0,18	0,004	-0,27	0,002	-0,36	0,001	-0,45	0,001

TEND - TENDENCIOSIDADE

EFIC - EFICIÊNCIA

REFERÊNCIAS

- [1] Abul-Ela, A.A.-L., Greenberg, B.G., Horwitz, D.G. (1967): "A multi-proportions randomized response model". Journal of American Statistical Assoc. (62), p. 990-1008.
- [2] Cochran, W.G. (1965): "Sampling Techniques". J. Wiley and Sons
- [3] Graybill, F.A. (1961): "An Introduction to Linear Statistical Models", vol. I. Mc-Graw Hill
- [4] Greenberg, B.G., Abul-Ela, A.A.-L., Simmons, W.R., Horwitz, D.G. (1969). "The unrelated question randomized response model: theoretical framework". Journal of American Statistical Assoc. (64), p. 520-539.
- [5] Greenberg, B.G., Kuebler, R.R., Abernathy, J.R., Horwitz, D.G. (1971): "Application of the randomized response model in obtaining quantitative data". Journal of American Statistical Assoc. (66), p. 243-250
- [6] Warner, S.L. (1965): "Randomized response: a survey technique for eliminating evasive answer bias". Journal of American Statistical Assoc. (60), p. 63-69.

E R R A T A

pag./linha	texto	correção
5/8	"determinarmos"	"determinar"
10/23	"tamanho do"	"tamanho de"
14	(matriz) P	suprimir os pontos entre as linhas " $p_{11} \dots p_{1m}$ " e " $p_{21} \dots p_{2m}$ "
14	(vetor) R^T	acrescentar "...;" entre " R_{2n_2} ;" e " R_{r1} ".
14	(vetor) ϵ^T	acrescentar "...;" entre " ϵ_{2n_2} ;" e " ϵ_{r1} ".
17	expressão 1.2.2-2	acrescentar " μ ", imediatamente antes do sinal "=".
28/12	"minimizar"	"maximizar"
31/9,10,11	"A solução ... e é dada por:"	"A solução é dada por:"
31/14,15	"Se a desigualdade... ... propósito."	suprimir
31	Expressão (2.2.2-5)	$n_1 = \frac{n \sqrt{\frac{\lambda_1 (1-\lambda_1) (1-p_2)^2}{\lambda_2 (1-\lambda_2) (1-p_1)^2}}}{1 + \sqrt{\frac{\lambda_1 (1-\lambda_1) (1-p_2)^2}{\lambda_2 (1-\lambda_2) (1-p_1)^2}}}$
41/13	"enquada"	"enquadra"
45/17 46/6 e 46/10	"(3.2-3)"	"(3.2-4)"
48/10	"(3.2-9) a (3.2-14)"	"(3.2-10) a (3.2-15)"
49/7-7	"(3.2-15) a (3.2-18), juntamente com (3.2.7) e (3.2.8)"	"(3.2-16) a (3.2-19), juntamente com (3.2-8) e (3.2-9)"