ESTRATOS DE GRANDE VARIÂNCIA EM LEVANTAMENTOS POR AMOSTRAGEM[1]

Francisco Alberto Pino[2]
Vera Lúcia Ferraz dos Santos Francisco[3]

1 - INTRODUÇÃO

Quando se levantam dados estatísticos por amostragem estratificada, em particular para obter estimativas de safras agrícolas, é comum que se construam estratos de tal forma que alguns deles contenham os elementos mais importantes do ponto de vista das principais variáveis a serem levantadas. Eventualmente, também se constrói um estrato com os elementos menos importantes ou de menor expressão. A produção agrícola é exemplo desse caso em que pelo menos um estrato contém os elementos pouco importantes. Ocorre frequentemente que esse estrato, dos elementos menos expressivos economicamente, perfaça grande número de elementos na população, com alta variabilidade entre eles. Como se pretende que a amostra seja pequena em tal estrato, podem resultar estimativas que, embora não viesadas, estejam muito longe da realidade, devido ao seu grande erro de amostragem.

Um exemplo dessa situação ocorreu no levantamento para previsão e estimativa da safra de laranja do Estado de São Paulo, realizada pelo Instituto de Economia Agrícola (IEA) e pela Coordenadoria de Assistência Técnica Integral (CATI), conforme Carta Acordo firmada com a Companhia Nacional de Abastecimento (CONAB), desde o ano-safra 2010 (CAMARGO; FRANCISCO, 2011). Especificamente, é o estrato que corresponde à cauda inferior da curva de distribuição do tamanho do pomar laranjeiro. Como dentro desse estrato o tamanho da amostra é muito pequeno em relação ao tamanho da população, o erro de amostragem resultante é muito alto, podendo gerar tanto estimativas muito altas quanto muito baixas para a produtividade e, consequentemente, estimativas ruins para a produção do estrato. Embora a estimativa da produção total do Estado seja muito maior do que a estimativa para esse estrato, convém diminuir sua variabilidade para maior acurácia nas estimativas.

Na amostra utilizada no levantamento da safra agrícola 2007/08 (safra industrial 2008/09) esse problema não aconteceu, porque os estratos foram divididos de outra forma, resultando em estratos menores (PINO; FRANCISCO, 2011). Uma vez que os estratos são definidos *a priori*, nem sempre é possível perceber se haverá o problema aqui relatado antes de ir a campo levantar dados. Como não há consenso sobre exatamente o que seja um pomar doméstico nem a partir de que tamanho um pomar se torna comercial, qualquer divisão do estrato de pequenos produtores teria de ser feita estatisticamente. Porém, se o levantamento de campo já foi efetuado, é preciso lidar com o problema *a posteriori*. O presente artigo apresenta uma proposta para tratar dessa questão. Embora motivado pela questão da estimação de produção de laranja, o procedimento proposto poderá ser utilizado em outras situações semelhantes.

2 - MATERIAL E MÉTODOS

Descreve-se a seguir o procedimento proposto.

2.1 - Fonte dos Dados e Análise de Agrupamentos

A fonte dos dados foi o cadastro de produtores de laranja, gerado com base no censo agropecuário do Estado de São Paulo, conhecido por Projeto LUPA (TORRES et al., 2008), sobre o qual uma amostra foi calculada, estratificada e sorteada (CAMARGO; FRANCISCO, 2011). O estrato 1 desse levantamento amostral, objeto deste estudo, é constituído pelas unidades de

[1]Registrado no CCTC, IE-52/2001.

[2]Engenheiro Agrônomo, Doutor, Pesquisador Científico do Instituto de Economia Agrícola (e-mail: pino@iea.sp.gov.br).

[3]Estatística, Pesquisadora Científica do Instituto de Economia Agrícola (e-mail: veralfrancisco@iea.sp.gov.br).

produção agropecuária (UPAs) com pomares de laranja entre 0,1 e 6 hectares.

As UPAs (elementos) do estrato 1 foram classificadas em grupos, aplicando-se a análise de agrupamentos (*cluster analysis*), sendo os grupos sugeridos pelos próprios dados, com base na produtividade média de laranja (kg/ha). Dentre os métodos de partição de uma população para obter os *clusters*, utilizou-se uma variação do método das K-médias para análise não hierárquica, utilizando a distância euclidiana como coeficiente de parecença e a soma de quadrados residuais como critério de homogeneidade dentro do grupo e heterogeneidade entre grupos (SPATH, 1985, p. 62-63). Resumidamente, suponha-se uma partição dos *n* elementos em *k* grupos indicada por:

$$p(k) = \left(o_i(k) : 1 \le i \le n_k\right)$$

O centro do grupo *p(j)*, ou seja, o ponto formado pela média das coordenadas de seus membros, é representado por $\overline{o(j)}$. Desse modo, a soma de quadrados residuais dentro do *j-ésimo* grupo será:

$$SQ\,\mathrm{Re}\,s(j) = \sum_{i=1}^{n_j} d^2\left(o_i(j); \overline{o}(j)\right)$$

em que d^2 representa o quadrado da distância euclidiana do elemento *i*, do grupo *j*, ao seu centro.

Para a partição toda, a soma de quadrados residual será:

$$SQ\,\mathrm{Re}\,s = \sum_{i=1}^{k} SQ\,\mathrm{Re}\,s(j)$$

Quanto maior for esse valor, mais homogêneos serão os elementos dentro de cada grupo e melhor será a partição.

Dessa forma, o estrato original é subdividido, de tal forma que cada grupo (*cluster*) torna-se um novo estrato (ou sub-estrato), construído *a posteriori*.

Para a aplicabilidade da maioria das ferramentas da Estatística clássica e para simplificação da análise supõe-se normalidade dos dados. Para verificar a aderência dos dados à distribuição normal utilizou-se o teste de Kolmogorov-Smirnov (estatística *D*), para avaliar as hipóteses:

H_0: Os dados seguem uma distribuição normal;
H_1: Os dados não seguem uma distribuição normal.

2.2 - Cálculo das Estimativas

A pós-estratificação requer que as proporções de elementos em cada estrato da população sejam conhecidas e que cada elemento da amostra possa ser classificado nos novos estratos (KISH, 1965, p. 90-92), sendo ambos os requisitos satisfeitos no caso do problema da produção de laranja apresentado. Se dentro do estrato os elementos da amostra forem selecionados mediante amostragem casual simples (não estratificada), como é o caso citado, fórmulas para estimação usuais podem ser utilizadas (KISH, 1965).

Denota-se:

A_{hi}, a área informada na *i-ésima* UPA da amostra no grupo *h;*

P_{hi}, a produção informada na *i-ésima* UPA da amostra no grupo *h*;

N_h, o tamanho da população no grupo *h*;

$N = \sum_h N_h$, o tamanho da população no estrato;

n_h, o tamanho da amostra no grupo *h*;

$f_h = n_h / N_h$, a fração amostral no grupo *h*;

$W_h = N_h / N$, o peso do grupo *h*;

$\hat{A}$, a estimativa do total da área plantada no estrato;

$\hat{P}$, a estimativa do total da produção no estrato.

Então, segundo Kish (1965), as estimativas da área e da produção no estrato são obtidas por:

$$\hat{A} = N\sum_h W_h \frac{1}{n_h}\sum_i A_{hi} = \sum_h \frac{N_h}{n_h}\sum_i A_{hi}$$

e

$$\hat{P} = N\sum_h W_h \frac{1}{n_h}\sum_i P_{hi} = \sum_h \frac{N_h}{n_h}\sum_i P_{hi}$$

As respectivas variâncias das estimativas são dadas por:

$$\mathrm{var}\left(\hat{A}\right) = \sum_h W_h^2\left(1 - f_h\right)\frac{s_h^2(A)}{n_n}$$

e

$$\mathrm{var}\left(\hat{P}\right) = \sum_h W_h^2 \left(1 - f_h\right) \frac{s_h^2\left(P\right)}{n_n}$$

em que

$$s_h^2\left(A\right) = \frac{1}{n_h - 1}\left[\sum_i A_{hi}^2 - \left(\sum_i A_{hi}\right)^2 \Big/ n_h\right]$$

e

$$s_h^2\left(P\right) = \frac{1}{n_h - 1}\left[\sum_i P_{hi}^2 - \left(\sum_i P_{hi}\right)^2 \Big/ n_h\right]$$

Surgem problemas na estimação quando o tamanho da amostra no estrato é muito pequeno:

a) Se $n_h = 1$ para algum *h*, a variância não poderá ser calculada; zerá-la nesse grupo irá subestimar a variância da estimativa no estrato;
b) Se $n_h = 0$ para algum *h*, além do mesmo problema com a variância, também o total da variável não poderá ser calculado; zerá-lo nesse grupo irá subestimar a estimativa do total no estrato.

Os cálculos necessários para os procedimentos foram realizados com o *Statistical Analysis Software* (SAS INSTITUTE, 2011).

3 - RESULTADOS E DISCUSSÃO

O estrato correspondente à cauda inferior da distribuição do tamanho do pomar laranjeiro (em hectare) que contém unidades de produção agropecuária (UPAs) entre 0,1 e 6 hectares (aproximadamente até 3.000 pés de laranja plantados) perfaz 7.554 elementos. Embora representando 36% das UPAs com laranja, esse estrato representa somente 2,9% em número de pés e 2,8% em área plantada. De acordo com o delineamento amostral aplicado no último levantamento em março de 2011, foram alocados somente 4 elementos nesse estrato, sorteados aleatoriamente (amostra casual simples dentro do estrato).

Como resultado da análise de agrupamentos (*cluster analysis*), as UPAs da população foram classificadas em três grupos, quanto à produtividade (Tabela 1). O grupo 1 é o de maior produtividade, enquanto o grupo 3 é o de menor produtividade, sendo a primeira quase 2,5 vezes maior do que a última. Essa grande disparidade justifica o uso da produtividade para subdividir o estrato. Por outro lado, percebe-se que o grupo 3 é menos relevante que os outros dois para estimar a produção de laranja, uma vez que ele representa mais da metade das UPAs do estrato (54%), mas 30% da área plantada e 17% da produção do estrato. Por sua vez, a produção total do estrato, pouco acima de 445 mil toneladas (11 milhões de caixas de 40,8 kg) é ínfima (cerca de 3,4%) quando comparada à produção total da safra agrícola paulista, que em 2010 foi da ordem de 322 milhões de caixas de 40,8 kg, ou 13 milhões de toneladas (FRANCISCO; CAMARGO; CASER, 2011).

O fato de que médias e medianas, em cada grupo, não diferem muito entre si, indica certo grau de simetria na distribuição dentro de cada grupo. Isso é comprovado pelo teste de aderência à normalidade através da estatística *D* do teste de Kolmogorov-Smirnov, cuja não significância leva à conclusão de que não se rejeita a hipótese de normalidade (Tabela 1).

Considerando o procedimento usual, aplicado ao delineamento amostral original (com 4 elementos), obtém-se a estimativa pontual de 494 mil toneladas (12,1 milhões de caixas de 40,8 kg) para o domínio referente ao estrato de pomares até 6 hectares. Com base no procedimento de pós-estratificação foram alocados 2 elementos no grupo 1 e 2 elementos no grupo 2, portanto sem representantes para o grupo 3. Outra situação encontrada no atual levantamento foi a de que um dos elementos alocados no grupo 1, cuja característica é de melhor produtividade em relação aos outros, apresentou queda de produção devido ao aumento da incidência de clorose variegada dos citros (CVC) e morte de plantas. Assim, a estimativa calculada foi de 241 mil toneladas ou 5,9 milhões de caixas de 40,8 kg (Tabela 2).

Como os respectivos intervalos de confiança de dois desvios padrões se sobrepõem, verifica-se que não há diferença significativa entre os totais estimados, porém a magnitude do intervalo é bem menor na segunda opção, isto é, a estimativa com pós-estratificação é mais precisa e de menor variabilidade. Dessa forma, apesar do estrato em estudo não ser um dos principais estratos na composição da estimativa final, o procedimento proposto apresentou estimação pontual que não difere significativamente daquela obtida originalmente, porém com menor variabilidade e, consequentemente, mais alta precisão.

TABELA 1 - Análise de Agrupamento, Segundo a Produtividade de Laranja, em Pomares de até 6 hectares, Estado de São Paulo, 2007/08

Variável	Estatística	Grupo (*cluster*)			Estrato
		1	2	3	
UPAs	Número	550	2.930	4.074	7.554
	Percentual	7,28	38,79	53,93	100
Produtividade (kg/ha)	Média	31.049	24.439	12.882	21.884
	Mediana	30.000	25.000	14.100	
Número de pés	Média	1.733	1.484	572	
	Mediana	1.700	1.200	350	
Área (ha)	Média	4,83	4,02	1,48	
	Mediana	5	4	1	
	Total	2.654	11.788	6.047	20.489
	Percentual	12,95	57,54	29,51	100
Produção (kg)	Total	82.405.001	288.086.932	77.897.454	448.389.387
	Percentual	18,38	64,25	17,37	100
Teste Kolmogorov-Smirnov	*D*	0,151042	0,2041194	0,132245	
		n.s.	n.s.	n.s.	

N.s. = não significativo ao nível de significância de 1%.
Fonte: Elaborada pelos autores com base em Torres et al. (2011).

TABELA 2 - Estimativa da Produção de Laranja em Pomares de até 6 hectares (mil toneladas em caixas de 40,8 kg), Estado de São Paulo, 2011

Procedimento	Estimativa	Intervalo de dois desvios padrões	
		Inferior	Superior
Original	493,68	236,64	754,8
Pós-estratificação	240,72	228,48	252,96

Fonte: Dados da pesquisa.

4 - CONSIDERAÇÕES FINAIS

Sabe-se que é preferível prevenir eventuais problemas quando na etapa do delineamento amostral do que procurar soluções *a posteriori*. Mais especificamente, a amostragem estratificada usa informação *a priori* para dividir a população alvo em subgrupos internamente homogêneos tanto quanto possível. No entanto, em particular nos levantamentos para estimativa de safra, a definição de elementos que não são expressivos economicamente deve ser feita atentando para a variabilidade das estimativas que produzem. Na cultura da laranja ainda não há consenso nessa definição, todavia, verificou-se a necessidade de se estabelecer tal limite para uma amostragem mais eficiente e, consequentemente, maior acurácia na estimativa.

LITERATURA CITADA

CAMARGO, F. P.; FRANCISCO, V. L. F. S. Estimativa de safra de laranja no estado de São Paulo. **Informações Econômicas**, São Paulo, v. 41, n. 5, 2011.

FRANCISCO, V. L. F. S.; CAMARGO, F. P.; CASER, D. V. **Evolução da produção de laranja 2009-2011 no estado de São Paulo**. São Paulo; IEA: 2010. Disponível em: <http://www.iea.sp.gov.br/out/verTexto.php?codTexto=12052>. Acesso em: 11 abr. 2011.

KISH, L. **Survey sampling**. New York: Wiley, 1965. 643 p.

PINO, F. A.; FRANCISCO, V. L. F. S. Estimativa de safra de laranja em 2008: um suco amargo. 2011**. Informações Econômicas**, São Paulo, v. 41, n. 8, p. 41-58, ago. 2011.

SAS INSTITUTE INC., (2011) **SAS OnlineDoc version eight**. Cary: SAS Institute Inc., 2011. Disponível em: <http://v8doc.sas.com/sashtml>. Acesso em: abr. 2011.

SPATH, H. **Cluster dissection and analysis**. Chichester: Ellis Horwood, 1985.

TORRES, A. J. et al. (Org.). **Projeto LUPA 2007/08**: levantamento censitário de unidades de produção agrícola do Estado de São Paulo.São Paulo: IEA/CATI/SAA, 2009. Disponível em: <www.cati.sp.gov.br/projetolupa>. Acesso em: 11 abr. 2011.

ESTRATOS DE GRANDE VARIÂNCIA EM LEVANTAMENTOS POR AMOSTRAGEM

RESUMO: *Levantamentos por amostragem estratificada usualmente têm dois estratos extremos em relação a uma variável de interesse: um com os elementos mais importantes ou de maior valor e outro com os elementos menos importantes ou de menor valor. Se o estrato de menor valor tem variâncias relativamente altas, as estimativas do estrato podem ser imprecisas, afetando negativamente a precisão da estimativa geral. Esse é o caso dos pomares pequenos (0,1 a 6 hectares) em se tratando da estimação de safras de laranja no Estado de São Paulo. Apresenta-se e testa-se um procedimento para pós-estratificação do estrato de menor valor, baseado em análise de agrupamentos. Mostra-se que as estimativas resultantes não diferem significativamente das estimativas originais, mas são muito mais precisas.*

Palavras-chave: *pós-estratificação, levantamento por amostragem, agrupamento, estimação de safra.*

LARGE VARIANCE STRATA IN SAMPLE SURVEYS

ABSTRACT: *Stratified sample surveys usually have two extreme strata: one with more important or higher value elements, and the other with less important or lower value ones in relation to the variable of interest. If the lower valued stratum has relatively high variances, the stratum estimates may lack precision, negatively affecting the overall estimation accuracy. This is the case faced by small orchards (0.1 to 6 ha) in Sao Paulo state, Brazil, when estimating orange crops. A post-stratification procedure for the lower value stratum is presented and tested, based on a cluster analysis. The resulting estimates are shown to be not significantly different from the original estimates, but much more precise.*

Key-words: *post-stratification, sample survey, clustering, crop estimation.*

Recebido em 30/06/2011. Liberado para publicação em 08/08/2011.