PUBVET
Publicações em Medicina Veterinária e Zootecnia
www.pubvet.com.br ISSN: 1982-1263 DOI: 10.22256/pubvet.v10n9.643-649

Simulação do tamanho ótimo de amostra em duas fases para um custo fixo de amostragem

Juracy Mendes Moreira[1]*, Aurélio Ferreira Melo[2], Jose Marcelo de Oliveira[3], Daniela Silva Ataides[4], Marcelo Carlos Ribeiro[5], Juliano Bortolini[6]

[1]Prof. Ms. Faculdade Almeida Rodrigues (FAR), Rio Verde – Goiás, Brasil. E-mail: juracimendesmoreira@yahoo.com.br
[2]Prof. Doutorando. Faculdade Almeida Rodrigues (FAR), Rio Verde – Goiás, Brasil. E-mail: aurelioferreiramelo1@hotmail.com
[3]Prof. Ms. Faculdade Almeida Rodrigues (FAR), Rio Verde – Goiás, Brasil. E-mail: oliveirademarcelo@hotmail.com
[4]Profa. Esp. Faculdade Almeida Rodrigues (FAR), Rio Verde – Goiás, Brasil. E-mail: danisila@uol.com.br
[5]Doutorando em Estatística Aplicada e Biometria (UFV,. Viçosa – Minas Gerais, Brasil. E-mail marcelocarlosribeiro@hotmail.com
[6]Prof. Dr. Departamento de Estatística, Universidade Federal de Mato Grosso, Boa Esperança, Cuiabá – Brasil. E-mail: julianobortolini@ufmt.br
**Autor para correspondência*

RESUMO. As pesquisas genéticas na cultura do café, têm tido grande expansão e, na maioria dos casos, as amostras são constituídas da coleta de folhas ou de frutos em diferentes plantas constituindo amostragem em mais de um estágio. Na amostragem em dois estágios ou sub amostragem, designada amostragem hierárquica, a população é constituída por N_1 unidades primárias e cada unidade primária por N_2 indivíduos. São selecionadas n_1 unidades primárias e, de cada uma delas, selecionados n_2 indivíduos. Para a determinação do tamanho ótimo da amostra biológica é necessário que se tenha dados obtidos de experimentos bem conduzidos e que expressem fielmente a variabilidade entre plantas de café e entre frutos, nas plantas, para condições que possam variar de acordo com os genes pesquisados. Em geral, o tamanho da amostra biológica utilizado pode estar sendo subestimado em função principalmente da relação entre as variâncias e da relação de custos.

Palavras chave: Amostragem, unidades primarias, variabilidade

Sample size great simulation two phases for a cost fixed sampling

ABSTRACT. Genetic research on the coffee culture, have had great expansion and, in most cases, the samples consist of collecting leaves or fruits in different plants constituting sampling in more than one stage. In two-stage sampling or sub-sampling, designated hierarchical sampling, the population is N_1 primary units and each primary unit by N_2 individuals. n_1 primary units are selected, and each selected n_2 individuals. To determine the optimal size of the biological sample is necessary to have data obtained from experiments conducted well and faithfully express the variability between coffee plants and fruits, from plants for conditions which may vary according to the studied genes. In general, the size of the biological sample used may be underestimated due mainly to the relationship between the variances and cost ratio.

Keywords: Sampling, primary units and variability

Introdução

O processo de amostragem probabilística indicado para uma amostragem em duas fases é aquele em que na primeira fase procede-se a uma amostragem das unidades da população em estudo e na segunda fase, seleciona-se indivíduos das unidades selecionadas.

Assim o processo é caracterizado por:

a) seleção das unidades primárias, atribuindo a cada uma delas igual probabilidade.

b) seleção de indivíduos entre os que pertencem às unidades primárias selecionadas.

Como exemplos de amostragem em duas fases. Viana et al. (2002) no trabalho sobre reprodução humana no distrito de São Paulo, inicialmente

selecionou residências na qual havia uma ou mais mulheres da população objetivo e na segunda fase selecionou mulheres dessas residências para a realização de seu estudo.

Paranaíba (2014) utilizou a amostragem em duas fases para um estudo de simulação computacional investigando diversas estratégias de alocação de amostras, com o objetivo de selecionar a mais eficiente, visando aumentar a precisão das estimativas de interesse, na primeira fase selecionou assentamentos rurais de acordo com um plano de amostragem aleatória simples sem reposição e na segunda fase selecionou famílias nesses assentamentos sorteados, segundo o mesmo plano de amostragem.

Uma das características desse método é que a variância do estimador da média amostral é definida pela soma das variâncias entre e variância dentro. Quando a determinação das unidades da primeira fase e os indivíduos dentro de cada uma das fases é aleatória os parâmetros do modelo, exceto μ têm efeitos aleatórios.

Admitindo-se que os efeitos nas diferentes fases sejam aditivos, o modelo matemático para amostragem em duas fases é dado por:

$$Y_{ij} = \mu + \alpha_i + \beta_{ij},$$

em que: Y_{ij}representa a observação no indivíduo j da segunda fase, na unidade i da primeira fase; μ é uma constante; α_i representa o efeito da unidade i da primeira fase; β_{ij} representa o efeito do indivíduo j da segunda fase dentro da unidade i com i= 1, 2,..., n_1 e j = 1, 2,..., n_2 para cada i

Início do aparte

Dado o modelo $Y_{ij} = \mu + \alpha_i + \beta_{ij}$

Sendo

$i = 1,2,3, \dots, n_i$

$j = 1,2,3, \dots,$ para todo i

Determine $\bar{Y}$

$$\bar{Y} = \frac{\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} Y_{ij}}{n_1 n_2}$$

$$\bar{Y} = \frac{\sum_i \sum_j (\mu + \alpha_i + \beta_{ij})}{n_1 n_2}$$

$$\bar{Y} = \frac{\sum_{i=1}\sum_{j=1}(\mu) + \sum_i \sum_j (\alpha_i) + \sum_i \sum_j (\beta_{ij})}{n_1 n_2}$$

$$\bar{Y} = \frac{n_1 n_2 \mu}{n_1 n_2} + \frac{n_2 \sum_i \alpha_i}{n_1 n_2} + \frac{\sum_i \sum_j (\beta_{ij})}{n_1 n_2}$$

$$\bar{Y} = \mu + \frac{1}{n_1}\sum_{i=1}^{n_1} \alpha_i + \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \beta_{ij}$$

Fim do aparte

A média amostral ($\bar{Y}$) para o modelo de amostragem em duas fases é:

$$\bar{Y} = \mu + \frac{1}{n_1}\sum_{i=1}^{n_1} \alpha_i + \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} \beta_{ij},$$

considerando que a seleção das unidades na primeira fase e dos indivíduos na segunda fase é feita aleatoriamente e admitindo-se que os parâmetros do modelo tenham médias e covariâncias nulas e variâncias iguais a σ_α^2 e σ_β^2 a variância da média amostral para uma amostragem em duas fases é dada por:

Início do aparte

Dado

$$\bar{Y} = \mu + \frac{1}{n_1}\sum_{i=1}^{n_1}\alpha_i + \frac{1}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\beta_{ij}$$

e o modelo $Y_{ij} = \mu + \alpha_i + \beta_{ij}$ e admitindo que os parâmetros tenham medias e covariâncias iguais $\sigma_\alpha^2\ e\ \sigma_\beta^2$,

Determine $VAR(\bar{Y})$

$VAR(\bar{Y})$ = A soma das variâncias mais duas vezes a covariância combinadas duas a duas.

$$VAR(\bar{Y}) = VAR\left(\mu + \frac{1}{I}\sum_{i=1}^{I}\sigma_i + \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\beta_{ij}\right)$$

$$VAR(\bar{Y}) = VAR(\mu) + VAR\left(\frac{1}{I}\sum_{i=1}^{I}\sigma_i\right) + VAR\left(\frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\beta_{ij}\right) + 2$$

$$\left[COV\left(\mu, \frac{1}{I}\sum_{i=1}^{I}\sigma_i\right) + COV\left(\mu, \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\beta_{ij}\right) + COV\left(\frac{1}{I}\sum_{i=1}^{I}\sigma_i, \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\beta_{ij}\right)\right]$$

$$VAR(\bar{Y}) = VAR\left(\frac{1}{I}\sum_{i=1}^{I}\sigma_i\right) + VAR\left(\frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\beta_{ij}\right)$$

$$VAR(\bar{Y}) = \frac{1}{I^2}\sum_{i=1}^{I}(\sigma_i) + \frac{1}{IJ^2}\sum_{i=1}^{I}\sum_{j=1}^{J}\beta_{ij}$$

$$VAR(\bar{Y}) = \frac{1}{I^2}\sigma_i^2 + \frac{1}{(IJ^2)}IJ\sigma_\beta^2$$

$$VAR(\bar{Y}) = \frac{\sigma_\alpha^2}{I} + \frac{\sigma_\beta^2}{IJ}$$

Pelas propriedades de variância temos:

$VAR\ (a)\ = 0$

$VAR\ (a, x)\ = 0$

$VAR(cx) = c^2 VAR(x)$

$Considerando I = n_1 e J = n_2; Temos$:

$$VAR(\bar{Y}) = \frac{1}{n_1}\sigma_\alpha^2 + \frac{1}{n_1 n_2}\sigma_\beta^2,$$

Fim do aparte

$$VAR(\bar{Y}) = \frac{1}{n_1}\sigma_\alpha^2 + \frac{1}{n_1 n_2}\sigma_\beta^2,$$

O modelo de análise de variância para a amostragem em duas fases com as esperanças matemáticas dos quadrados médios considerando o modelo aleatório é apresentado na Tabela 1.

Tabela 1. Modelo de análise de variância para uma amostragem em duas fases – modelo aleatório.

Fontes de Variação	GL	QM	E(QM)
Entre unidades	$n_1 - 1$	QM_1	$\sigma_\beta^2 + n_2\sigma_\alpha^2$
Entre indivíduos: unidades	$n_1(n_2 - 1)$	QM_2	σ_β^2

Coeficiente de correlação intraclasse.

O coeficiente de correlação intraclasse é uma estimativa da variabilidade total de medidas devido a variações entre os indivíduos e pode ser obtido dividindo o valor da variação entre os indivíduos (V_e) pela variação total (V_t) que inclui a variação entre indivíduos e a variação não pretendida (V_{erro}). É definido pela seguinte expressão:

$$\rho = \frac{\sigma_E^2}{\sigma_E^2 + \sigma_D^2},$$

em que:σ_E^2é a variância entre classes eσ_D^2é a variância dentro de classes.

Para uma amostragem em duas fases, o coeficiente de correlação intraclasse é dado pela seguinte expressão:

$$\rho = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\beta^2},$$

Sendo que: ρ é o coeficiente de correlação intraclasse; σ_α^2 é a variância entre indivíduos e σ_β^2 é a variância dentro de indivíduos.

Função de custos para amostragem em duas fases

Os projetos de pesquisa envolvem custos, em maior ou menor escala associado também com a amostragem. Em se tratando de amostras em duas fases, esses custos podem ser ainda mais elevados. Assumindo que o custo é proporcional ao número de unidades amostradas nas duas fases e que o custo por unidade amostrada é conhecido, ou seja, custo de deslocamento, material utilizado, técnicos empregados e equipamentos usados nas análises das amostras, o custo total da amostra pode ser considerado como uma função linear do número de unidades amostradas. ConsiderandoC_1e C_2 como o custo de cada uma das n_1 unidade e cada um dos n_2 indivíduos amostrados respectivamente, o custo total da amostra C_t para uma amostragem em duas fases é:

$$C_t = C_1 n_1 + C_2 n_1 n_2,$$

O procedimento estatístico para o tamanho ótimo da amostra para a amostragem em dois estágios consiste na solução do sistema de equações constituído pela variância da média amostral e do custo total de amostragem como funções de n_1 e n_2:

$$VAR(\bar{Y}) = \frac{1}{n_1}\sigma_\alpha^2 + \frac{1}{n_1 n_2}\sigma_\beta^2,$$

$$c_t = c_1 n_1 + c_2 n_1 n_2.$$

A amostra ótima é constituída por $n_1' n_2'$ indivíduos sendo n_1' unidades primárias, com n_2' indivíduos selecionados em cada uma das unidades primárias. As soluções ótimas são obtidas fixando-se a variância da média amostral ou o custo total da amostra. Marcuse (1949) apresenta a alocação ótima para n estágios em amostragem hierárquica,

As soluções ótimas para a amostragem hierárquica em dois estágios são apresentadas a seguir. Nas expressões, a variância entre unidades primárias (σ_α^2) é designada por σ_1^2 e a variância entre indivíduos dentro das unidades, (σ_β^2) representada por σ_2^2. O número de unidades primárias, fixada a variância da média (v_0) e minimizando o custo total da amostra é:

$$n_1' = \sqrt{\frac{\sigma_1^2}{c_1}}\left(\frac{\sum_{i=1}^{2}\sqrt{c_i\sigma_i^2}}{v_0}\right)$$

Se o interesse for fixar o custo total da amostra (c_t) e minimizar a variância da média

amostral, o número ótimo de unidades primárias é:

$$n_1' = \sqrt{\frac{\sigma_1^2}{c_1}}\left(\frac{c_t}{\sum_{i=1}^{2}\sqrt{c_i\sigma_i^2}}\right).$$

O número ótimo de indivíduos em cada unidade primária não depende de que seja fixada a variância da média ou o custo total da amostra e é dado por:

$$n_2' = \sqrt{\frac{c_1\sigma_2^2}{c_2\sigma_1^2}}.$$

As expressões para o número ótimo de unidades primárias, em função den_2', fixada a variância da média amostral ou o custo total da amostra são, respectivamente:

$$n_1' = \frac{1}{v_0}\left(\sigma_1^2 + \frac{\sigma_2^2}{n_2'}\right),$$

e

$$n_1' = \frac{c_t}{c_1 + c_2 n_2'}.$$

A amostragem de variância mínima para um custo fixo é equivalente a um custo mínimo para uma variância fixa, é dada por n'_2 indivíduos por unidade. Zanon & Storck (2000), por:

$$n'_2 = \sqrt{\frac{C_1}{C_2}\left(\frac{1-\rho}{\rho}\right)},$$

esta fórmula demonstran'_2 como uma função crescente de ρ, ou seja, para valores altos de ρ são encontrados valores mínimos de n'_2 e também que o resultado não depende dos custos C_1 e C_2 mas do quociente entre eles, por isso é preferível minimizar essa relação e não o custo total.

A técnica de simulação de variáveis constitui uma importante ferramenta de trabalho, especialmente em estudos de características referentes à populações biológicas que tenham distribuições de probabilidade conhecidas. Um método muito utilizado em estatística para gerar dados de uma densidade e o método de simulação Monte Carlo. Esse método tem sido empregado como forma de se obter aproximações numéricas de funções complexas, envolvendo tipicamente a geração de observações com alguma distribuição de probabilidade. A principal vantagem da simulação de dados é minimizar os custos e tornar mais rápido a obtenção de dados bem como dar suporte a decisões em estudos futuros para programas de melhoramento tanto em espécies vegetais como animais.

Metodologia

Os dados utilizados neste trabalho foram obtidos através de simulação computacional pelo método de Monte Carlo, sendo utilizado no processo uma constante μ= 100 e as variâncias σ_α^2 = 0,310807 eσ_β^2 = 0,566276.

Através da simulação computacional de dados por meio do Software R foi possível determinar diversos tamanhos de parcelas (entre plantas e entre fruto dentro de plantas), a fim de determinar qual seria a mais eficiente, no sentido de aumentar a precisão das estimativas de interesse (alocar mais plantas e menos fruto ou mais fruto e menos plantas). Caracterizando assim uma assim uma amostragem em duas fases (plantas e frutos).

Em seguida esses valores foram agrupados dentro de cada conjunto de dados formando diversos tamanhos de parcelas e para cada tamanho de parcela foi determinado o valor da estimativa do coeficiente de variação da média; mediana e moda, como a média sofre influencia de valores extremos demos preferência por utilizar a moda do coeficiente de variação neste estudo.

Resultado e discussão

Os valores dos coeficientes de variação para diferentes tamanhos de parcelas encontram-se na tabela 2. Como se pode perceber, o valor do coeficiente de variação diminui com o aumento do tamanho da parcela, independente da estimativa analisada.

Com resultado desse agrupamento e com o uso Software Team (2013) foram estimados os valores de todos os parâmetros dos modelos citados anteriormente. Com isso pode ser determinado o tamanho de parcela para cada método.

Tabela 2. Dimensão e tamanho de parcelas e suas estimativas agrupadas

Tamanho de Parcela	Media CV	Mediana CV	Moda CV
1	9,3411	9,3351	9,3081
2	6,5792	6,5733	6,5738
3	6,2018	6,1936	6,1894
6	4,9574	4,9521	4,9389
9	3,5318	3,5187	3,4956
10	2,8673	2,8402	2,6735
18	2,8118	2,7885	2,7500
27	2,2738	2,2493	2,1233
30	2,1462	2,1214	1,9614
45	1,2122	1,1814	1,1315
90	0,6986	0,6584	0,5715
135	0,8315	0,7028	0,4819

Obtidos os valores das estimativas de $S_{\alpha}^2 = 0{,}4514$ e$S_{\beta}^2 = 0{,}3381$ e considerando o custo de amostragem de planta $C_1 = 0{,}70$, o custo de amostragem de frutos dentro de plantas $C_2 = 0{,}15$ e o custo total fixo $C_t = 25{,}00$. O numero de plantas e o numero de frutos dentro de planta podera ser estimado atraves das segunites expressões:

$$n'_{planta} = \sqrt{\frac{S_{Planta}^2}{C_{Planta}}}\left[\frac{C_t}{\sqrt{C_{Planta} * S_{Planta}^2 + \sqrt{C_{Fruto} * S_{Fruto}^2}}}\right]$$

$$n'_{Planta} = \sqrt{\frac{0{,}4514}{0{,}70}}\left[\frac{25{,}00}{\sqrt{0{,}70 * 0{,}4514} + \sqrt{0{,}15 * 0{,}3381}}\right]$$

$$n'_{planta} \cong 25$$

$$n'_{Fruto} = \sqrt{\frac{C_{Planta} * S_{Fruto}^2}{C_{Fruto} * S_{Planta}^2}}$$

$$n'_{Fruto} = \sqrt{\frac{0{,}70 * 0{,}3381}{0{,}15 * 0{,}4514}}$$

$$n'_{Fruto} \cong 2$$

Conclusão

Para um custo fixo por amostragem de R$ 25,00, deve-se selecionar aproximadamente 25 plantas e aproximadamente dois frutos por planta.

Para a determinação do tamanho ótimo da amostra biológica é necessário que se tenham dados obtidos de experimentos bem conduzidos e que expressem fielmente a variabilidade entre

plantas de café e entre frutos nas plantas para condições que possam variar de acordo com os interesses pesquisados.

Em geral, o tamanho da amostra biológica utilizado pode estar sendo subestimado em função principalmente da relação entre as variâncias e da relação de custos, para uma relação de custo alta, o tamanho ótimo de amostra se torna ainda maior. Com isso, a obtenção de amostras muito pequena poderá afetar a precisão do experimento diminuindo a confiabilidade, o número de frutos na amostra ótima aumenta à medida que o custo de amostrar uma planta aumenta em relação ao custo de amostrar um fruto, para uma mesma proporção de variâncias.

Referências bibliográficas

Marcuse, S. (1949). Optimum allocation and variance components in nested sampling with an application to chemical analysis. *Biometrics*, 5, 189-206.

Paranaíba, P. F. (2014). Proposição e avaliação de métodos para estimar o tamanho ótimo de parcelas experimentais, Universidade Federal de Lavras, Lavras.

Team, R. C. (2013). R: A language and environment for statistical computing. p 1-14.

Viana, A. E. S., T. Sediyama, P. R. Cecon, S. C. Lopes, and M. A. N. Sediyama. (2002). Estimativas de tamanho de parcela em experimentos com mandioca. Horticultura Brasileira 20: 58-63.

Zanon, M. L. B., and L. Storck. (2000). Tamanho ótimo de parcelas experimentais para Eucalyptus saligna Smith em dois estádios de desenvolvimento. Cerne 6: 104-111.

Article History
Received 13 May 2016
Accepted 31 May 2016
Available on line 9 July 2016