Australian Journal of Crop Science AJCS

AJCS 9(11):1064-1068 (2015) ISSN:1835-2707

A multivariate approach to determine sample size for morphological characterization of pepper fruits

Anderson Rodrigo da Silva[1*], Reginaldo Francisco Hilário[2], Elizanilda Ramalho do Rêgo[3], Naysa Flávia Ferreira do Nascimento[3], Carlos Tadeu dos Santos Dias[2], Renato Paiva de Lima[4]

[1]Department of Agronomy, Goiano Federal Institute, Geraldo Silva Nascimento Road, Km 2.5 – 75790-000 – Urutaí, GO – Brazil
[2]Department of Exact Sciences, University of São Paulo – ESALQ/USP – Av. Pádua Dias, 11 – 13418-900 – Piracicaba, SP – Brazil
[3]Laboratory of Vegetal Biotechnology, Federal University of Paraíba – 58397-000 – Areia, PB – Brazil
[4]Department of Soil Sciences, University of São Paulo – ESALQ/USP, Brazil

***Corresponding author: anderson.silva@ifgoiano.edu.br**

Abstract

In chilli pepper, the calculation of the effective or minimum sample size can minimize costs with characterization. In order to determine the effective sample size, a general multivariate statistical method consisting of resampling subsamples from a reference sample is presented. Data from a field experiment involving eight accessions of *Capsicum* pepper are used to illustrate the method. Six response variables relating to morphological characterization of fruits were analyzed: mean weight, peduncle length, fruit length, largest diameter, lowest diameter, pericarp thickness. The reference sample consisted of the vector of scores of the first principal component, thus representing 30 observations on the 6 morphological variables. Through the percentile bootstrap method, a 99% confidence interval was created for two parameters: mean and standard deviation of the reference sample, which was then resampled with replacement, creating 500 subsamples of sizes ranging from 2 to 29. Afterwards, we estimated both mean and standard deviation for each subsample of each size. The proportion of estimates outside their respective confidence interval was computed. We also compared the results of the multivariate approach with its univariate form. The multivariate approach has taken into account the correlations among the response variables and was more efficient than the univariate form. A sample containing 22 fruits is considered suitable for estimating the mean of pepper fruit traits, whereas 24 fruits should be enough to estimate the standard deviation.

Keywords: *Capsicum spp*.; Principal component analysis; Resampling; Pepper; Bootstrap.
Abbreviations: MW_fruit mean weight; PL_peduncle length; FL_fruit length; LD_fruit largest diameter; LowD_fruit lowest diameter; PT_pericarp thickness; CI_confidence interval.

Introduction

Pepper (Capscicum spp.) is an important spice and vegetable crop in Brazil, where several types and forms of fruits of this crop are grown (Rêgo et al., 2012). The species and varieties are differentiated by botanical traits, mainly relating to flowers and fruits (Nascimento et al., 2013). The morphological characterization of pepper fruits has been essential for understanding the enormous diversity of the *Capsicum* species, which has fostered several breeding programs (Rêgo et al., 2003; Nascimento et al., 2013). It generates subsidies that have facilitated the decisions of the breeders as well as the identification of duplicate genotypes, so that they can properly plan their experiments, knowing the genetic diversity available (Pickersgill, 1997). In this sense, Silva et al. (2011) accentuated the importance of characterizing the fruits based on an appropriate sample size. The authors stated that an alternative way to obtain efficient sample sizes for estimating population parameters is the technique of resampling subsamples with replacements from a reference sample. According to Leite et al. (2009), this technique allows one to make efficient comparisons of the sample size effects on the estimation of genetic and phenotypic parameters. The method, in most cases, indicates a relatively smaller sample size, which is able to provide estimates with similar accuracy as those from the reference sample, thus decreasing costs of characterization while keeping reliable estimates. The technique is currently implemented in free software programs, such as Genes (Cruz, 2006) and R (R Core Team, 2015) through the package *biotools* (Silva, 2015). Using this methodology to determine sample sizes from a reference sample of 30 *Capsicum* pepper fruits, Silva et al. (2011) obtained reductions of around 50% of the reference sample size, depending on the morphometric trait. The authors then observed that the recommendation found in the descriptors for *Capsicum* (IPGRI 1995), of 10 mature fruits at the second harvest, were not enough to represent the reference sample. Using resampling with replacement techniques for estimating genetic and phenotypic parameters in sugarcane, Leite et al. (2009) stated that sample size estimates varied according the evaluated parameter and trait. The technique was also used to estimate the plastochron in pigeonpea (Cargnelutti Filho et al., 2013), to estimate the means of jack beans and velvet beans traits (Cargnelutti Filho et al., 2012) and to estimate the Pearson correlation coefficient among maize traits (Cargnelutti Filho et al., 2010). Herrmann et al. (2010) used random subsamples to determine sample size in diversity studies on alfalfa. In morphological characterization, not only of peppers but of many other crops, the calculation of appropriate sample sizes based on objective methods is still underexplored. Moreover, when calculations are made, the correlations among the response variables are usually devalued. Another point is that many methods currently used to determine sample size based on power analysis, via t-test, F-test etc., admit Gaussian distribution, and most of them provide results whose target parameter is only the population mean. Nonetheless, the technique of resampling subsamples is wider, since it does not assume any distribution, and is more flexible in terms of the target parameter, whatever its complexity. The goal of this study was twofold: (1) to present a statistical method to determine the effective sample size in a multivariate way, through the technique of resampling subsamples from a reference sample; (2) to make decisions regarding the appropriate sample size for performing morphological characterization of chilli pepper fruits.

Results and Discussion

Importance of traits on sample size calculation

The first principal component retained 52.4% of the total residual variation. Its coefficients are:

$$Z_1 = 0.58MW + 0.40PL + 0.34FL + 0.54LD + 0.31LowD - 0.02PT \qquad (1)$$

Note that except for the pericarp thickness, the other variables contributed together and with similar weight to the fruit variability. Moreover, because PT presented a low coefficient in Z_1 (-0.02), it is expected that PT had less influence on the calculation of the effective sample size than the other fruit traits.

Multivariate sample size

The bootstrapped 99% confidence interval for the population mean of the first principal component corresponds to the limiting values: -0.29 and 0.28 (Fig 1A). A sample size of 22 fruits reached the proportion of 0.008 (0.8%) points outside the $CI_{99\%}$. The following subsamples showed a decreasing proportion. These findings are similar to those found by Silva et al. (2011), who also based their study on a reference sample of 30 fruits and found subsample sizes ranging from 16 to 19, with the same accuracy of the reference sample. Nevertheless, note that the authors used α = 0.05, thus the calculated sample sizes are expected to be smaller than those found here. Michereff et al. (2011), calculating sample size for quantifying cercospora leaf spots in sweet pepper, stated that the number of plants to be sampled was reduced when the degree of acceptable error was increased. Monitoring the impact of *Bt* maize on butterflies, Lang (2004) concluded that the number of field margins that must be sampled in order to achieve a higher statistical power must be increased when monitoring a single butterfly species. Using α = 0.05, Lúcio et al. (2003) found reductions of around 25% on the population size (reference sample of size 72) of sweet pepper plants cultivated in greenhouses. These authors used the method based on the t-Student random variable, as proposed by Cochran (1977). The 99% confidence interval for the population standard deviation of the first principal component corresponds to the limiting values: 0.33 and 0.73 (Fig 1B). It can be observed that the behaviour of the estimated standard deviations is similar to the behaviour of the estimated means. For the standard deviation, a subsample size of 24 seems to be appropriate, at which 0,1% points are located outside the $CI_{99\%}$. The following subsamples showed a decreasing proportion.

Multivariate versus univariate approach

Considering the univariate form of the same technique, i.e., by calculating the effective sample size for estimating the population mean of each fruit trait, we observe the following values: 28 for fruit weight, 26 for peduncle length, 28 for fruit length, 28 for largest diameter, 27 for lowest diameter and 28 (or > 28) for pericarp thickness (Fig 2). Comparing these results with those obtained using the multivariate approach, it can be seen that, besides the differences among the calculated sample sizes, the effective sample size based on the multivariate approach was smaller (22, for estimating the mean). The most probable cause to this is the presence of correlations among the variables, which is another source of variability in the data which is not being considered by the univariate method. Furthermore, it is noteworthy to recap the low influence of PT over Z_1 (Eqn. 5), probably due to the lack of correlation (~ 0.05, data not presented) with the other variables. In fact, the univariate calculation of the sample size based on the estimated mean of PT is the most contradictory, not exactly presenting a decreasing behaviour (Fig 2). Even though there are studies (Leite et al., 2009; Silva et al., 2011; Cargnelutti Filho et al., 2010, 2012) in which the univariate approach has been used to determine the effective sample size, we could not find any published paper, involving any species, that used the multivariate form; therefore, we could not make any comparisons of results obtained with this technique.

Materials and Methods

Plant materials

This study was based on a field experiment involving four species of chilli pepper, genus *Capsicum*, presenting fruits morphologically distinct. We studied the species *Capsicum chinense* (accessions 2, 12, 13, 15 and 74)*, C. annuum* (accession 14)*, C. baccattum* (accession 72) and *C. frutescens* (accession 4), all from the Germoplasm Bank of the Federal University of Paraíba (UFPB-CCA). Plants were sown in polystyrene trays with 128 cells containing commercial substrate. After presenting six true leaves, plants were transplanted to the field.

Experimental design and data

The experiment was carried out under a generalized randomized block design with two replications (blocks) and thirty within-plot replications (fruits), and eight accessions. The experimental area was located at: 06°57′ S, 35°41′ W, 618 m a.s.l.. Six response variables related to the morphological characterization of fruits were evaluated: mean fruit weight (MW), peduncle length (PL), fruit length (FL), largest diameter (LD), lowest diameter (LowD) and pericarp thickness (PT). For each experimental plot, 30 fruits were collected.

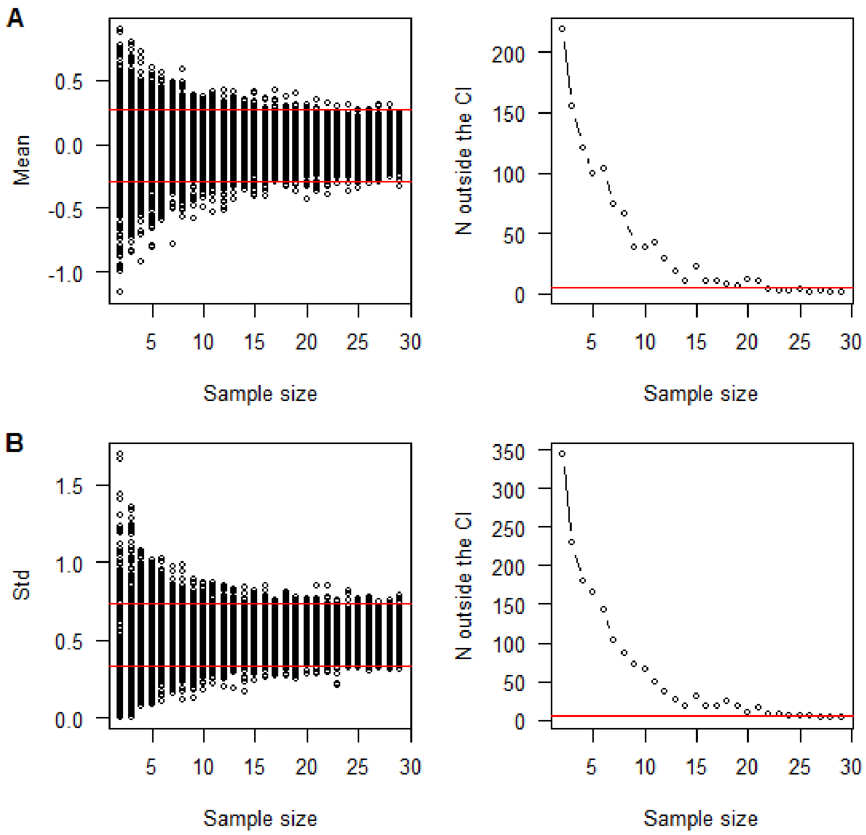

Fig 1. Dispersion of the estimates for each subsample (left), and number of estimates out of the bootstrapped 99% (percentile) confidence interval (right) for (A) the mean and (B) the standard deviation of the first principal component.

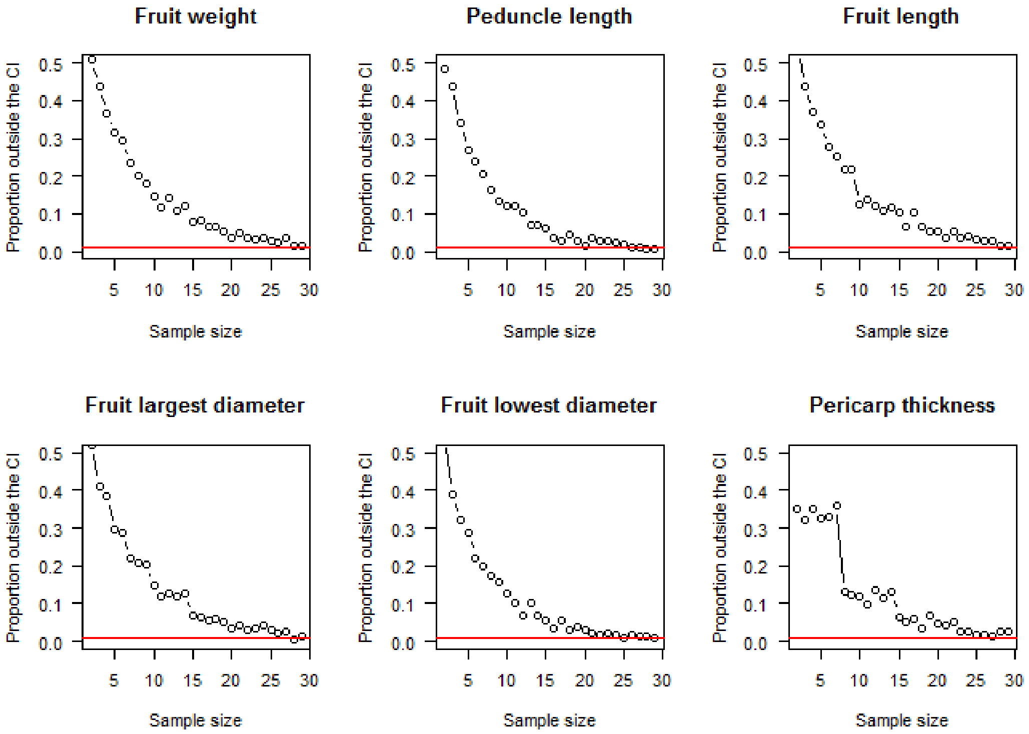

Fig 2. Proportion of estimates out of the bootstrapped 99% (percentile) confidence interval for the mean of each fruit trait.

Statistical analysis

A multivariate analysis of variance was performed according the following model:

$$Y = 1\mu^{T} + X\alpha + Z\beta + W\gamma + \varepsilon \qquad (2)$$

where Y corresponds to a (n×p)-dimensional response matrix, 1 is an n-dimensional vector containing the value 1 only, μ is a p-dimensional vector of population means,ά is a (7×p)-dimensional matrix of accession effects, β is a (1×p)-dimensional matrix of block effects, γ is a (7×p)-dimensional matrix of interactions between accessions and blocks, X, Z and W are incidence (model) matrices, and **ε** is a (n×p)-dimensional matrix of residuals. In this experiment, n = 540 and p = 6.

The residual matrix was used to perform a principal component analysis based on the correlation matrix (R), represented by the spectral decomposition: $R=VAV^{T}$, in which V is the matrix of eigenvectors and A is the diagonal matrix of eigenvalues. The means of the first principal component scores were calculated for each fruit (within-plot replication), $\overline{Z}_1 = \varepsilon V_1$. For a quick reading on multivariate analysis of variance and principal components analysis, see Yeater et al. (2015). For a more detailed approach, see Krzanowski (2000).

Sample size determination (the multivariate approach)

The reference sample used to determine the sample size was the vector of scores of the first principal component, representing 30 observations of the 6 morphological variables. Through the percentile bootstrap method, a 99% confidence interval was built for the following parameters of the reference sample (scores): mean (Eqn. 2) and standard deviation (Eqn. 3).

$$\hat{\mu}^{*}_{\alpha/2} \le \mu \le \hat{\mu}^{*}_{1-\alpha/2} \qquad (3)$$

$$\hat{\sigma}^{*}_{\alpha/2} \le \sigma \le \hat{\sigma}^{*}_{1-\alpha/2} \qquad (4)$$

where $\hat{\mu}^{*}_{\alpha/2}$ and $\hat{\sigma}^{*}_{\alpha/2}$ are the quantile α/2 of the bootstrap estimates for the population mean and standard deviation, respectively. We considered α = 0.01, along with 200 bootstrap estimates.

Subsamples, with size ranging from 2 to 29, were resampled with replacement to estimate both statistics. For each smaller sample size, 500 subsamples were taken in order to compute the proportion of estimates outside their respective confidence interval.

The procedure can be described by the following algorithm:

1) Consider *n* as the size of the reference sample and *s* < *n* the size of a subsample, $s = 2, 3, ..., n-1$.
2) Take with replacement a length-*k* sequence of independent subsamples of size *s*, say $x_1^{(s)}, x_2^{(s)}, ..., x_k^{(s)}$.
3) For each resampled vector, compute the statistic of interest: $f\left(x_1^{(s)}\right), f\left(x_2^{(s)}\right), ..., f\left(x_k^{(s)}\right)$.
4) Calculate the proportion of estimates (Eqn. 4) outside the (1 – α)100% confidence interval (*CI*) for that statistic (*f*), based on reference sample.

$$p(s) = \frac{1}{k}\sum_{j=1}^{k} I\left[f\left(x_j^{(s)}\right) \notin CI\right] \qquad (5)$$

where *I[.]* is an *indicator variable*.

5) Consider *s* as an appropriate sample size if $p(s) \le \alpha$. Otherwise, repeat the previous steps, considering $s = s+1$.

We validate the results by comparing the effective sample size obtained with the multivariate approach and the effective sample size obtained using the univariate version of the method, for each fruit trait.

Computing

All statistical procedures were performed using the software R. The sample size algorithm described is available from the package *biotools* with the function sample.size(). To initialize the resampling process, a seed equal to ‘123’ was used through the function set.seed().

Conclusions

The multivariate approach has taken into account the correlations among the response variables and was more efficient than the univariate form on determining the effective sample size of *Capsicum* fruits. A sample containing 22 fruits is considered suitable for estimating the mean of pepper fruit traits, whereas 24 fruits should be enough to estimate the standard deviation.

Acknowledgments

Authors thank the Goiano Federal Institute (IFGoiano, Brazil) for the financial support.

References

Cargnelutti Filho A, Toebe M, Burin C, Silveira TR, Casarotto G (2010) Sample size for estimating the pearson correlation coefficient among corn characters. Pesqui Agropecu Bras. 45:1363–1371.

Cargnelutti Filho A, Toebe M, Burin C, Fick AL, Alves BM, Facco G (2012) Sample size for estimating the average length, diameter and weight of seeds of jack bean and velvet bean. Cienc Rural. 42:1541–1544.

Cargnelutti Filho A, Toebe M, Facco G, Santos GO, Alves BM, Bolzan A (2013) Sample size to estimate the plastochron in pigeonpea. Eur J Agron. 48:12–18.

Cochran WG (1977) The estimation of sample size. In: Sampling Techniques, 3rd edn. John Wiley & Sons, New York.

Cruz CD (2006) Programa Genes - análise multivariada e simulação. Editora UFV, Viçosa.

Herrmann D, Flajoulot S, Julier B (2010) Sample size for diversity studies in tetraploid alfalfa (*Medicago sativa*) based on codominantly coded SSR markers. Euphytica. 171:441–446.

IPGRI - International Plant Genetic Resources Institute (1995) Descriptors for Capsicum (Capsicum spp.). Rome: The Asian Vegetable Research and Development Center, Taipei, Taiwan, and the Centro Agronómico Tropical de Investigación y Enseñanza, Turrialba, Costa Rica, p 51.

Krzanowski W (2000) Principles of multivariate analysis: a user’s perspective, 2nd edn. Oxford University Press, New York.

Lang A (2004) Monitoring the impact of Bt maize on butterflies in the field: estimation of required sample sizes. Environ Biosafety Res. 3:55–66.

Leite MSO, Peternelli LA, Barbosa MHP, Cecon PR, Cruz CD (2009) Sample size for full-sib family evaluation in sugarcane. Pesqui Agropecu Bras. 44:1562–1574.

Lúcio AD, Souza MF, Heldwein AB, Lieberknecht D, Carpes RH, Carvalho MP (2003) Sample size and sampling method for sweet pepper evaluations in greenhouse. Hortic Bras. 21:180–184.

Michereff SJ, Martins RB, Noronha MA, Machado LP (2011) Sample size for quantification of cercospora leaf spot in sweet pepper. J Plant Pathol. 93:183–186.

Nascimento NFF, Nascimento MF, Santos RMS, Bruckner CH, Finger FL, Rêgo ER, Rêgo MM (2013) Flower color variability in double and three-way hybrids of ornamental peppers. Acta Hortic. 1000:457–464.

Pickersgill B (1997) Genetic resources and breeding of capsicum spp. Euphytica. 96:129–133.

R Core Team (2015) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. Available from: http://www.R-project.org/. Accessed on 09 July, 2015.

Rêgo ER, Rêgo MM, Cruz CD, Finger FL, Amaral DSSL (2003) Genetic diversity analysis of peppers: a comparison of discarding variables methods. Crop Breed Appl Biotechnol. 3:19–26.

Rêgo ER, Santos RMC, Rego MM, Nascimento NFF, Nascimento MF, Bairral MAA (2012) Quantitative and multicategoric descriptors for phenotypic variability in a segregating generation of ornamental peppers. Acta Hortic. 1:289–296.

Silva AR (2015) biotools: tools for biometry and applied statistics in agricultural science. R package version 2.1. Available from: http://cran.r-project.org/package=biotools. Accessed on 09 July, 2015.

Silva AR, Rêgo ER, Cecon PR (2011) Sample size for morphological characterization of pepper fruits. Hortic Bras. 29:125–129.

Yeater KM, Duke SE, Riedell WE (2015) Multivariate analysis: greater insights into complex systems. Agron J. 107:779–810.