



What is the best way to define sample size for cauliflower seedlings?

Karina Chertok Bittencourt¹  Rafael Rodrigues de Souza²  Stella Bonorino Pazetto³ 
Marcos Toebe^{1*}  Iris Cristina Datsch Toebe⁴ 

¹Departamento de Ciências Agronômicas e Ambientais, Universidade Federal de Santa Maria (UFSM), 98400-000, Frederico Westphalen, RS, Brasil. E-mail: m.toebe@gmail.com. *Corresponding author.

²Departamento de Fitotecnia, Universidade Federal de Santa Maria (UFSM), Santa Maria, RS, Brasil.

³Universidade Federal do Pampa (UNIPAMPA), Itaqui, RS, Brasil.

⁴Programa de Pós-graduação em Informática na Educação, Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, RS, Brasil.

ABSTRACT: *This study compared four methods for defining the ideal sample size per experimental unit to estimate the overall experimental mean for the total length, shoot length, root length, and the number of leaves of cauliflower seedlings. An experiment was carried out where the number of leaves, shoot, root, and total length were measured, and the general, perpendicular distance, linear response plateau, and spline methods were tested. While the general method may under or overestimate sample size and the sampling of 10 seedlings suggested by the spline method is still too far from the stabilization point of the curve, the perpendicular distance and linear plateau response methods are recommended to obtain results corresponding to narrower confidence interval widths. Therefore, according to the perpendicular distance method, at least 15 seedlings per experimental unit are required to estimate the overall experimental mean of cauliflower seedlings reliably for the traits here measured.*

Key words: *Brassica oleracea, horticulture, experimental planning, maximum curvature point.*

Qual é a melhor maneira para definir o tamanho amostral para mudas de couve-flor?

RESUMO: *Este estudo comparou quatro métodos para definir o tamanho amostral ideal por unidade experimental para estimar a média experimental geral para o comprimento total, comprimento de parte aérea, comprimento de raiz e número de folhas de mudas de couve-flor. Um experimento foi conduzido no qual foram medidos o número de folhas, o comprimento total, de parte aérea e de radícula, e foram testados os métodos: geral, de distâncias perpendiculares, resposta linear platô e spline. Enquanto o método geral pode sub ou superestimar o tamanho amostral e a amostragem de 10 mudas sugerida pelo método spline ainda se encontra muito distante do ponto de estabilização da curva, os métodos de distâncias perpendiculares e resposta linear platô são recomendados por obterem resultados que corresponderam a amplitudes de intervalos de confiança menores. Portanto, de acordo com o método de distâncias perpendiculares, pelo menos 15 mudas por unidade experimental são necessárias para estimar a média experimental geral de mudas de couve-flor confiavelmente para as variáveis aqui mensuradas.*

Palavras-chave: *Brassica oleracea, horticultura, planejamento experimental, ponto de máxima curvatura.*

Different methodologies have been proposed to define sample size based on the maximum curvature point (FEDERER, 1955), such as the general, perpendicular distance, linear plateau response, and spline methods (SILVA & LIMA, 2017). However, CARGNELUTTI FILHO et al. (2021) showed that different methods obtained quite different results when defining optimal plot size for several crops, making the selection of the method the first crucial step in sample size definition since, if not appropriate, it may lead to unrepresentative numbers. Also, little attention has been given to the definition of sample

size per experimental unit, that is, considering experimental restrictions, which are present in experiments that use experimental designs, being this the case of most experiments performed with horticultural crops.

Cauliflower (*Brassica oleracea* var. *botrytis* L.) is an example of a widely studied horticultural crop that has been the object of several experiments thorough the years in which different sample sizes have been chosen empirically, once the lack of standardization for this number can be easily visualized in the literature. While THOMSON et al. (2013) assessed 20 cauliflower

plants per plot, TEMPESTA et al. (2019) used a sample of 5 plants from each experimental unit, and COSTA et al. (2020) collected only 1 plant per cultivar. Thus, a recommendation of the number of plants to be collected per experimental unit based on the comparison of methods may be extremely useful to researchers that perform experiments with cauliflower seedlings, facilitating experimental planning and the obtention of more reliable results. Therefore, this study compared four methods for defining the ideal sample size per experimental unit to estimate the overall experimental mean for the total length, shoot length, root length, and the number of leaves of cauliflower seedlings.

The experiment was conducted in a greenhouse at the Federal University of Pampa (UNIPAMPA), Itaqui, Rio Grande do Sul, Brazil. Cultivar Teresópolis Gigante was sown using three substrates (50% Mecplant® + 50% Carolina Padrão®, 75% Mecplant® + 25% rice husk, and 75% Carolina Padrão® + 25% rice husk), in 72 and 128 cell-trays (3x2 two-factor scheme) with four replications, in a completely randomized design. Thirty days later, twenty seedlings were randomly collected from each experimental unit, considering higher sample numbers are rarely used in cauliflower studies (THOMSON et al., 2013; TEMPESTA et al., 2019; COSTA et al., 2020). Then, the following traits were measured: a) Number of Leaves (NL) in units, b) Shoot Length (SL), from neck to leaflet insertion, in cm; c) Root Length (RL), from neck to root apex, in cm; and d) Total Length (TL), as the sum of SL and RL, in cm. Other experiments with 1, 2, ..., 100 seedlings per experimental unit were simulated using bootstrap resampling, with 10,000 resamples with repositioning (EFRON, 1979).

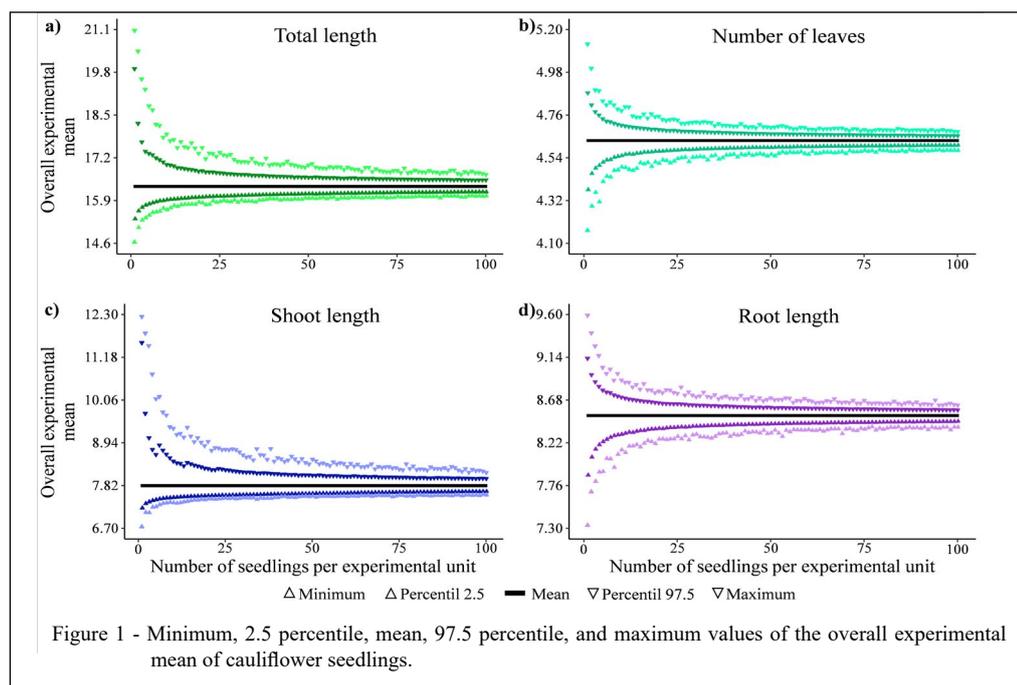
The statistical analyses were performed using R software (R DEVELOPMENT CORE TEAM, 2021) in several of its functions, and R package soilphysics (SILVA & LIMA, 2015) according to the applications carried out to determine the sample size by SILVA & LIMA (2017). After subdividing the database per experimental unit, these sample sizes were subjected to analysis of variance, performed according to the following model: $Y_{ijk} = m + T_i + S_j + (TS)_{ij} + \varepsilon_{ijk}$, where Y_{ijk} is the value observed in the response variable in plot ijk , m is the overall mean, T_i is the fixed effect of level i ($i = 1$ and 2) of the tray-cell-size

factor, S_j is the fixed effect of level j ($j = 1, 2, 3$) of the substrate factor, $(TS)_{ij}$ is the interaction fixed effect of level i of the tray-cell-size factor with level j of the substrate factor and ε_{ijk} is the experimental error effect (STORCK et al., 2016). Thereafter, m effect was extracted from this model in each resampling, using specific routines with *sample()* and *aov()* functions.

Re-samplings for each planned sample scenario were subjected to descriptive analysis defining minimum, percentiles of 2.5, mean, percentiles of 97.5, and maximum values. Ninety five percent confidence interval width ($CI_{95\%}$) was estimated from the difference between percentiles of 97.5 and percentiles of 2.5. Next, $CI_{95\%}$ estimates were fitted using *nls()* function through the following power model: $CI_{95\%} = \alpha \times n^\beta + \varepsilon$, where α is the coefficient of interception, n , sample size, β , exponential rate of decay, and ε , random effect error. Posteriorly, four methods for determining the maximum curvature point were used: the general, perpendicular distances, linear plateau response, and spline methods, according to SILVA & LIMA (2017), using *maxcurv()* function from soilphysics package (SILVA & LIMA, 2015). This point was considered the representative sample size.

In the reference experiment, the effects of the substrate, tray-cell size, and substrate \times tray-cell size interaction factors were significant. As expected, for all traits, $CI_{95\%}$ decreased exponentially as sample size increased up to a stabilization point (Figure 1), that is, the sampling of 1 seedling corresponds to a much wider $CI_{95\%}$ compared with the sampling of 100 seedlings per experimental unit. This reflects the higher the number of seedlings collected, the more representative the sample (SIEGEL, 2016), once too small sample sizes may subject results to over or underestimation (CARGNELUTTI FILHO et al., 2018). Nevertheless, the mean property of m was constant for all traits (4.62 units for NL, 7.82 cm for SL, 8.51 cm for RL, and 16.34 cm for TL), which was also observed by TOEBE et al. (2018), who reported this statistic as a non-biased estimator. Moreover, power models presented satisfactory fitting-quality (MOINESTER & GOTTFRIED, 2014), verified through the coefficient of determination (R^2), root mean square error (RMSE), and d index (Table 1).

Although, the optimal sample size varied slightly between traits, the four methods led to quite different results (Figure 2). Whilst the

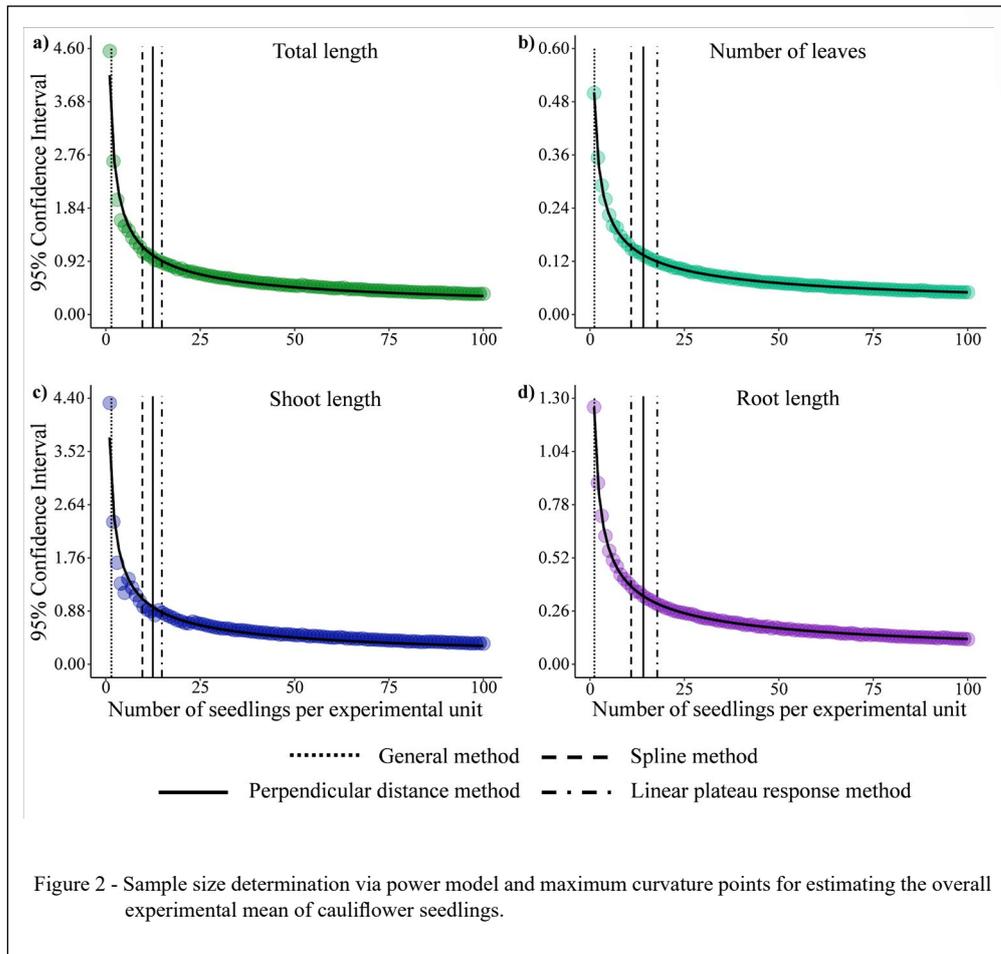


general method considered two seedlings were enough, the others required either 10 (spline), 15 (perpendicular distance), or even 19 seedlings per experimental unit (linear plateau response)

to estimate m reliably. However, considering the $CI_{95\%}$ observed when sampling only two seedlings, such a low number would most likely lead to unreliable estimates. Also, the sampling of

Table 1 - Coefficient of determination (R^2), root mean square error (RMSE) and d index of the power models, and maximum curvature points and sample sizes for the overall experimental mean of the number of leaves (NL), shoot length (SL), root length (RL) and total length (TL) of cauliflower seedlings.

Trait	Power model	R^2	RMSE	d index
NL	$CI_{95\%} = 0.5019 \times n^{-0.4985}$	0.9995	0.0015	0.9999
SL	$CI_{95\%} = 3.7533 \times n^{-0.5466}$	0.9588	0.1009	0.9892
RL	$CI_{95\%} = 1.2576 \times n^{-0.5017}$	0.9998	0.0022	0.9998
TL	$CI_{95\%} = 4.1439 \times n^{-0.5534}$	0.9837	0.0703	0.9957
Trait	Maximum curvature method	Maximum Curvature	Maximum $CI_{95\%}$	Sample size
NL	General method	1.0000	0.5019	1
NL	Spline method	9.9940	0.1593	10
NL	Perpendicular distance method	14.4865	0.1324	15
NL	Linear plateau response method	18.9407	0.1158	19
SL	General method	1.4951	3.0126	2
SL	Spline method	9.4000	1.1028	10
SL	Perpendicular distance method	13.9518	0.8887	14
SL	Linear plateau response method	16.9607	0.7987	17
RL	General method	1.0000	1.2576	1
RL	Spline method	9.9542	0.3971	10
RL	Perpendicular distance method	14.4469	0.3294	15
RL	Linear plateau response method	18.8022	0.2886	19
TL	General method	1.5941	3.2014	2
TL	Spline method	9.3205	1.2048	10
TL	Perpendicular distance method	13.8726	0.9668	14
TL	Linear plateau response method	16.7023	0.8724	17



10 seedlings per experimental unit is still too far from the stabilization point of the curve, meaning the general and spline methods may not be ideal choices for the conditions under study.

Moreover; although both perpendicular distance and linear plateau response methods presented representative sample numbers, the precision gain obtained by the linear plateau response in relation to the perpendicular distance method is too little to justify selecting the first over the latter. Thus; even though both can be used reliably, the results obtained using the perpendicular distance method may come up as more efficient under the practical perspective of researchers for being closer to the minimum sufficient number of seedlings to be sampled per experimental unit enough to reach high precision. This avoids collecting greater samples, as recommended by the linear plateau response, which often requires more resources and manpower. Therefore, we highly encourage the sampling of at

least 15 cauliflower seedlings per experimental unit in order to estimate the overall experimental mean reliably for the traits here measured.

In this sense, even if the sampling recommendations here proposed might be highly practical and efficient for optimizing experimental planning of experiments with cauliflower seedlings, considering the majority of them use experimental restrictions, these should not be applied to other horticultural crops without performing preliminary studies. The method comparison here presented should only serve as a basis to researchers whose aim is to define sample size for other species from the Brassicaceae family.

ACKNOWLEDGEMENTS

We thank the Universidade Federal do Pampa (UNIPAMPA) for infrastructure and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brasil - Finance code 001, for financial support.

DECLARATION OF CONFLICT OF INTERESTS

We have no conflict of interest to declare.

AUTHORS' CONTRIBUTIONS

Conceptualization: KCB and RRS. Data acquisition: KCB and SBP. Design of methodology and data analysis: RRS and MT. Supervision and coordination: MT and ICDT. KCB and RRS prepared the draft of the manuscript. All authors critically revised the manuscript and approved of the final version.

REFERENCES

- CARGNELUTTI FILHO, A. et al. Number of leaves for modelling the leaf area of velvet bean according to leaf dimensions. *Revista de Ciências Agroveterinárias*, v.17, p.571–578, 2018. Available from: <<http://dx.doi.org/10.5965/223811711732018571>>. Accessed: Sept. 21, 2021. doi: 10.5965/223811711732018571.
- CARGNELUTTI FILHO, A. et al. Comparison of methods for estimating the optimum plot size for pearl millet, slender leaf rattlebox, and showy rattlebox. *Revista Caatinga*, v.34, p.249–256, 2021. Available from: <<http://dx.doi.org/10.1590/1983-21252021v34n201rc>>. Accessed: Sept. 21, 2021. doi: 10.1590/1983-21252021v34n201rc.
- COSTA, L. F. et al. Cauliflower growth and yield in a hydroponic system with brackish water. *Revista Caatinga*, v.33, p.1060–1070, 2020. Available from: <<http://dx.doi.org/10.1590/1983-21252020v33n421rc>>. Accessed: Sept. 23, 2020. doi: 10.1590/1983-21252020v33n421rc.
- EFRON, B. Bootstrap methods: another look at the jackknife. *Annals of Statistic*, v.7, p.1–26, 1979. Available from: <<https://doi.org/10.1214/aos/1176344552>>. Accessed: Sept. 18, 2021. doi: 10.1214/aos/1176344552.
- FEDERER, W. J. *Experimental design: theory and application*. New York : Oxford & IBH Publishing, 1955.
- MOINESTER, M.; GOTTFRIED, R. Sample size estimation for correlations with pre-specified confidence interval. *Quantitative Methods for Psychology*, v.10, p.124–130, 2014. Available from: <<https://doi.org/10.20982/tqmp.10.2.p0124>>. Accessed: Sept. 25, 2021. doi: 10.20982/tqmp.10.2.p0124.
- R DEVELOPMENT CORE TEAM. 2021. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria.
- IEGEL, A.F. *Practical business statistics*. London : Academic Press, 2016.
- SILVA, A. R. da; LIMA, R. P. soilphysics: an R package to determine soil preconsolidation pressure. *Computers and Geosciences*, v.84, p.54–60, 2015.
- SILVA, A. R. da; LIMA, R. P. Determination of maximum curvature point with the R package soilphysics. *International Journal of Current Research*, v.9, p.45241–45245, 2017.
- STORCK, L. et al. *Plant Experimentation*. Santa Maria : UFSM, 2016.
- TEMPESTA, M. et al. Optimization of nitrogen nutrition of cauliflower intercropped with clover and in rotation with lettuce. *Scientia Horticulturae*, v.246, p.734–740, 2019. Available from: <<https://doi.org/10.1016/j.scienta.2018.11.020>>. Accessed: Sept. 25, 2021. doi: 10.1016/j.scienta.2018.11.020.
- THOMSON, G. et al. Effects of elevated carbon dioxide and soil nitrogen on growth of two leafy Brassica vegetables. *New Zealand Journal of Crop and Horticultural Science*, v.41, p.69–77, 2013. Available from: <<https://doi.org/10.1080/01140671.2013.772905>>. Accessed: Sept. 23, 2021. doi: 10.1080/01140671.2013.772905.
- TOEBE, M. et al. Sample size for estimating mean and coefficient of variation in species of crotalarias. *Anais da Academia Brasileira de Ciências*, v.90, p.1705–1715, 2018. Available from: <<https://doi.org/10.1590/0001-3765201820170813>>. Accessed: Sept. 23, 2021. doi: 10.1590/0001-3765201820170813.