Crop Breeding and Applied Biotechnology 14: 71-75 2014
Brazilian Society of Plant Breeding. Printed in Brazil

ARTICLE
http://dx.doi.org/10.1590/1984-70332014v14n2a13

Sample size for the assessment of soybean inbred populations

José Manoel Colombari Filho[1] and Isaias Olívio Geraldi[2*]

Received 11 December 2013

Accepted 23 January 2014

Abstract – *In plant breeding programs, the knowledge about the appropriate sample size for the evaluation of populations is very important. A small sample reduces the chance of selecting superior genotypes, whereas a very large sample may lead to unnecessary increases in cost and labor. A population consisting of 192 soybean lines was divided in groups of 24 lines, which were assessed for grain yield in eight randomized complete block experiments. Analyses of variance were performed for each experiment as well as for groups of experiments, resulting in analyses of variance consisting of 24, 48, 72, 96, 120, 144, 168, and 192 lines. As the sample size increased, the width of confidence intervals of parameter estimates decreased, stabilizing with samples of 144 lines. Therefore, an appropriate sample size for the evaluation of soybean inbred populations should contain about 150 lines.*

Key words: *Genetic variance, heritability, parameter estimation accuracy, confidence interval.*

INTRODUCTION

In plant breeding programs of autogamous species, selection can be initiated in early generations of inbreeding (F_2 or F_3) or in advanced generations of inbreeding (from F_6 onwards), when the population reaches homozygosis and consists of a sample of inbred lines. In any case, knowing the best-suited sample size becomes important, i.e., a sample size that represents the genetic variability of the population. An undersized sample can reduce the chances of selecting superior genotypes that occur at low frequencies (transgressive types) and even promote the fixation of undesirable alleles, whereas an oversized sample may lead to unnecessary increases in cost and labor (Falconer and Mackay 1996, Pinto et al. 2000). Knowing the appropriate sample size is also relevant for an accurate estimation of parameters (Marquez-Sanchez and Hallauer 1970). However, little research has been conducted to determine the appropriate number of genotypes (plants, progenies or lines) in soybean breeding programs. Most studies addressed the size and shape of the experimental plot.

One way to determine the appropriate sample size is through the accuracy of the genetic parameters estimates such as genetic variances and heritability coefficients, which can be evaluated by their confidence intervals. This process was used by Pinto et al. (2000) for maize and by Badan et al. (1998) for rice. According to this method, the appropriate sample size is one in which the amplitude of the confidence interval is stabilized.

When determining the appropriate sample size, the effective population size (Ne) should also be considered, i.e. the number of genetically different plants that compose a sample and effectively contribute to form the following generation (Falconer and Mackay 1996). Different types of progenies require different sample sizes, since each type of progeny has a different Ne; the lower the Ne, the greater will be the number of progenies required to represent the population (Souza Júnior 2001, Vencovsky and Crossa 2003).

The objective of this study was the determination of the appropriate sample size for the evaluation of soybean populations in advanced generations of selfing and, therefore, consisted of a sample of inbred lines.

MATERIAL AND METHODS

The population used in this study was derived from the cross between the parents 'Gaúcha' and 'BR-80-8858' and consisted of a sample of 192 inbred lines. This population was chosen for its wide genetic variability for the trait grain yield. For the development of this population the within F_2 bulk method was used from $F_{2:3}$ to $F_{2:7}$, beginning with 20

[1] Embrapa Rice and Beans, CP 179, 75.375-000, Santo Antônio de Goiás, GO, Brazil
[2] Department of Genetics, "Luiz de Queiroz" College of Agriculture (ESALQ), University of São Paulo (USP), CP 83, 13.418-900, Piracicaba, SP, Brazil.
*E-mail: iogerald@usp.br

$F_{2:3}$ progenies. Ten inbred lines were randomly taken from each bulk in the $F_{2:7}$ generation, giving rise to 200 F_8 inbred lines. Due to some losses, 192 lines were left, representing the plant material used in this study.

This population with 192 lines was evaluated in eight experiments with 24 treatments (lines) each, corresponding to a random sample of the original population. We used a randomized complete block design with five replications and plots consisting of one 2-m row, spaced 0.50 m apart, i.e., 1 m^2 plots, containing 35 plants after thinning. The trait grain yield was recorded (in g m^{-2}).

The analysis of variance for each experiment were performed according to the random model $y_{ij} = \mu + t_i + r_j + e_{ij}$, where y_{ij} is the observation related to line i in replication j; μ is the overall mean; t_i, with i = 1, 2,... I is the random effect of treatments (lines); r_j, with j = 1, 2,... R, is the random effect of replications; and e_{ij} is the experimental error (Steel and Torrie 1980).

The analyses of variance were then repeated by sequential grouping of the lines. In the grouped analysis, lines were grouped from 1 to 24 (Experiment 1), 1 to 48 (Experiments 1 and 2), 1 to 72 (Experiments 1 through 3), 1 to 96 (Experiments 1 through 4), 1 to 120 (Experiments 1 through 5), 1 to 144 (Experiments 1 through 6), 1 to 168 (Experiments 1 through 7), and 1 to 192 (Experiments 1 through 8), in a total of eight sampling groups. In each case, the grouping was performed by pooling each source of variation, i.e., by summing the sums of squares and degrees of freedom.

For the eight sample sizes, the variance components ($\hat{\sigma}^2_1$, $\hat{\sigma}^2_F$ and $\hat{\sigma}^2$) and the heritability coefficient based on line means ($\hat{h}^2_{\bar{X}}$) were estimated by the expressions (Vencovsky and Barriga 1992): $\hat{\sigma}^2_1 = (MS_L - MS_E)/ R$, $\hat{\sigma}^2_F = (MS_L/R)$, $\hat{\sigma}^2 = MS_E$ and $\hat{h}^2_{\bar{X}} = \hat{\sigma}^2_1 / \hat{\sigma}^2_F$, where MS_L; MS_E; R; $\hat{\sigma}^2_1$; $\hat{\sigma}^2_F$ and $\hat{\sigma}^2$ represent, respectively, mean square of lines; mean square of error; number of replications; genetic variance among lines; phenotypic variance based on line means; and variance of the error.

The confidence intervals (CI) (95% probability) of the genetic variances among lines and heritability coefficient estimates were calculated using the following expressions (Knapp et al. 1985, 1987, Barbin 1993): IC ($\hat{\sigma}^2_1$): $P[(nt\sigma^2_1 / \chi^2_{nt;0.975}) \leq \hat{\sigma}^2_1 \leq (nt\sigma^2_1 / \chi^2_{nt;0.025})] = 0.95$, and IC($\hat{h}^2_{\bar{X}}$): $P\{1-[(MS_L/MS_E)F_{0.975;f_2;f_1}]^{-1}\} \leq \hat{h}^2_{\bar{X}} \leq \{1-[(MS_L/MS_E)F_{0.025;f_2;f_1}]^{-1}\} = 0.95$, where χ^2, f_1, f_2, nt and F correspond, respectively, to tabulated chi-square values at the 0.025 and 0.975 levels, degrees of freedom of error mean square, degrees of freedom of lines mean square, degrees of freedom of genetic variance among lines estimate, and tabulated F values at 0.025 and 0.975 levels. The value of nt was computed as proposed by Satterthwaite: $nt = (MS_L - MS_E)^2/[(MS_L^2/f_2) + (MS_E^2/f_1)]$ (Barbin 1993). The genetic variance among lines and heritability coefficients obtained for the eight sample sizes, along with the confidence intervals, were plotted on graphs for ease of comparison and interpretation.

RESULTS AND DISCUSSION

The individual analyses of variance (Table 1) showed significant differences among lines for all sample sizes by the F test, which is an indicator of the genetic variability in the population. It was also observed that the error mean squares were very similar, which evidently allows a combined analysis with the different sample sizes. The coefficients of experimental variation ranged from 25.8 to 31.1 %. These, although apparently high, were similar to those previously reported for this type of plot in soybean (Barona et al. 2012). Moreover, one also has to consider that the population means in different experiments were not high, i.e., in the order of 2 t ha^{-1}, mainly due to low rainfall which, of course, contributed to raise the coefficient variation. In this situation, the experimental precision can be considered satisfactory. The population mean ranged from 191.1 g m^{-2} (Exp. 5) to 228.1 g m^{-2} (Exp. 2), i.e., the estimates were very close. This was expected, since the treatments of each experiment corresponded to different samples of lines of the same population.

Table 1. Analysis of variance of the eight experiments, overall mean ($\bar{X}$), and coefficients of experimental variation (CV %) for soybean grain yield (g m^{-2})

SV	df	Experiments							
		1	2	3	4	5	6	7	8
Replications	4	1,434.1	11,060.8**	5,365.1	10,372.5	3,056.9	12,198.2 **	20,866.2 **	3,205.0
Lines	23	37,222.0**	36,569.0**	41,861.0 **	34,716.0 **	43,633.0 **	36,475.0 **	60,394.0 **	34,460.0 **
Error	92	4,357.2	3,470.8	3,741.8	4,544.4	2,762.9	3,609.7	4,439.0	3,420.8
$\bar{X}$	-	212.4	228.1	217.6	219.4	191.1	203.7	215.5	222.0
CV%	-	31.1	25.8	28.1	30.7**	27.5	29.5	30.9	26.3

* and **: significant ($p \leq 0.05$) and significant ($p \leq 0.01$) by the *F* test, respectively.

The combined analyses of variance (Table 2) showed very similar mean squares for lines and error in the different analyses, and significance between lines was detected for all analyses (sample sizes) by the F test. The estimates of genetic variance among lines ($\hat{\sigma}_1^2$) were also all very close, varying from 6,573.0 (g m^{-2})^2 in a sample with 24 lines to 7,540.0 (g m^{-2})^2 in a sample with 168 lines.

A similar fact occurred for the heritability coefficient estimates ($\hat{h}_{\bar{X}}^2$). Besides, these coefficients became practically stable at a sample size of 120 or more lines (Table 2). It also appears that the heritability coefficient estimates were high (around 90%), which may be surprising for a quantitative trait such as grain yield. However, one has to take into consideration that: i) the population was chosen due to its high genetic variability, ii) the heritability coefficients were calculated based on line means, where the environmental component of variation is divided by the number of replications (five in this case), which increases this coefficient, iii) the treatments correspond to a sample of lines in the F_8 generation, without previous selection. It is well known that in situations as this, the additive genetic variance among lines ($\hat{\sigma}_1^2$) is twice as high as that of the F_2 generation, while the dominant genetic variance is reduced to zero, contributing substantially to increase the heritability coefficient (Mather and Jinks 1982).

However, as already mentioned, the accuracy of an estimate is not determined by its value, but by its confidence interval. Thus, narrower confidence intervals indicate higher accuracy of the estimates, i.e., the estimate represents the population parameter with reasonable accuracy (Pinto al. 2000). In other words, the parameter can assume any value within the confidence interval, and therefore, very wide confidence intervals indicate low precision or low reliability of the estimates.

The confidence intervals of the estimates of genetic variance (Figure 1) illustrate this fact clearly. It was observed that as the number of lines increased, the width of the confidence interval decreased. When using 144 lines, a stabilization of the amplitude of the confidence interval was noted and from that point onwards, the degree of accuracy of the estimates was similar. A similar fact occurs for the heritability coefficient estimates (Figure 2). Although the magnitudes of these were practically constant for different sample sizes (Table 2), ranging from 89.1 (sample of 48 lines) to 90.7 % (samples of 168 and 192 lines), the same does not occur with the corresponding confidence intervals. Clearly, there was an almost linear reduction of the confidence intervals as the sample size increased, since the confidence interval was highest for the sample with 24 lines and smallest for that with 192 lines. However, the confidence interval was stabilized at sample sizes between 120 and 144 lines. Therefore, the estimates obtained with samples of 144 lines were satisfactorily accurate, requiring no larger samples.

Confidence intervals are calculated based on the parameter estimates and on the degrees of freedom of the sources of variation of analyses of variance. Once the estimates

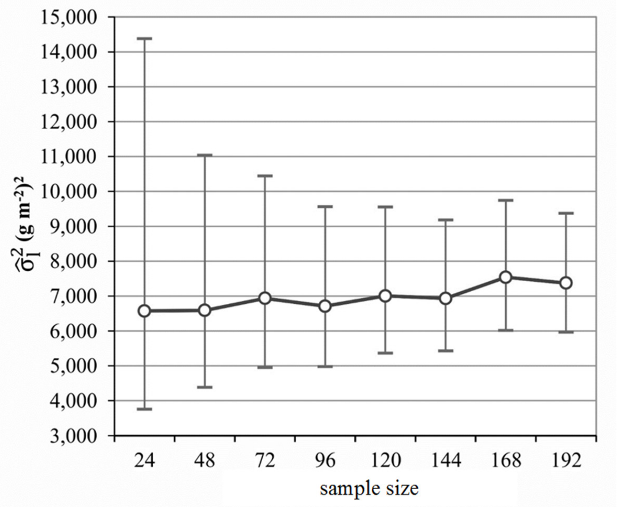

Figure 1. Estimates of genetic variance among lines ($\hat{\sigma}_1^2$) and corresponding confidence intervals, for sample sizes of 24, 48, 72, 96, 120, 144, 168, and 192 lines for soybean grain yield (g m^{-2}).

Table 2. Combined analysis of variance, estimates of genetic variance among lines ($\hat{\sigma}_1^2$) and heritability coefficients based on line means ($\hat{h}_{\bar{X}}^2$), for sample sizes of 24, 48, 72, 96, 120, 144, 168, and 192 lines for soybean grain yield (g m^{-2})

SV	Sample sizes (number of lines)							
	24	48	72	96	120	144	168	192
Lines/Exp.	37,222.0 **	36,895.6 **	38,550.8 **	37,592.0 **	38,800.1 **	38,412.6 **	41,552.8 **	40,666.2 **
Error/Exp.	4,357.2	3,951.3	3,876.3	4,047.5	3,783.4	3,756.7	3,852.8	3,798.0
$\hat{\sigma}_1^2$	6,573.0	6,588.8	6,934.9	6,708.9	7,003.3	6,931.2	7,540.0	7,373.6
$\hat{h}_{\bar{X}}^2$	0.883	0.891	0.898	0.892	0.902	0.902	0.907	0.907

* and **: significant ($p \leq 0.05$) and significant ($p \leq 0.01$) by the *F* test, respectively.

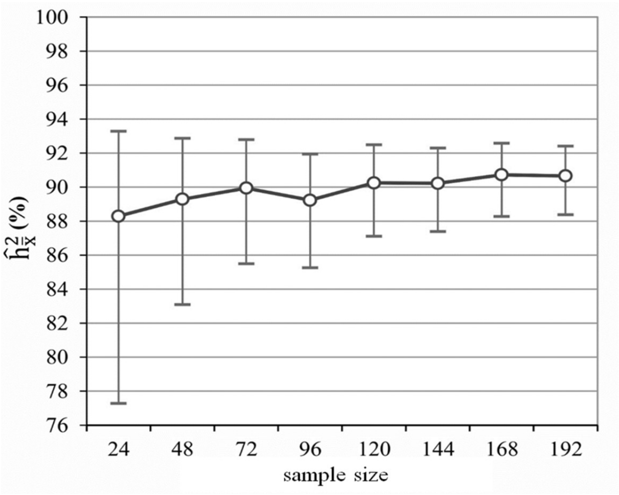

Figure 2. Estimates of heritability coefficients based on line means ($\hat{h}^2_{\bar{x}}$) and corresponding confidence intervals, for sample sizes of 24, 48, 72, 96, 120, 144, 168, and 192 lines, for soybean grain yield (g m^{-2}).

of genetic variance were relatively stable, the degrees of freedom had the greatest influence on the amplitude of the confidence interval (Knapp et al. 1987), i.e., estimates obtained with a higher number of degrees of freedom are more accurate. Therefore, the greater the number of treatments, the greater the number of degrees of freedom and the narrower the confidence interval of the variance estimates, resulting in more accurate estimates. Of course, the same reasoning applies to the heritability estimates.

Knowledge about the appropriate sample size is very important in plant breeding programs, since inadequate samples can lead to misleading conclusions about the properties of populations for breeding purposes. In addition, small samples can lead to a distortion of population properties, genetic drift and loss of transgressive genotypes (Falconer and Mackay 1996). On the contrary, very large samples involve high cost and labor, and can even make the maintenance of germplasm collections very difficult (Marquez-Sanchez 1972, Vencovsky and Crossa 2003), apart from the problems they cause in breeding programs.

In breeding programs for most species the most important trait is grain yield, since it is probably the most complex trait, for being strongly influenced by the environment. Thus, the appropriate sample size should be determined primarily for this trait, because if the number of treatments is suitable for this, it will certainly be for the other traits with less environmental influence. In this study, we concluded that for soybean grain yield, about 150 treatments would be an appropriate minimum number for estimating parameters, when the base population is a random sample of inbred lines.

Reports on this subject are scarce. In the literature, sample sizes from 50 to 1,000 lines are reported, which is frequently an arbitrarily set number. Marquez-Sanchez and Hallauer (1970) found that a sample of 200 maize plants is sufficient for an accurate estimation of genetic parameters. Similar results were reported by Omolo and Russel (1971), also in maize.

In a study on two maize populations, similar to ours, Pinto et al. (2000) concluded that the appropriate (or minimum) sample size to estimate parameters for grain yield is 200, when using S_1 progenies. They emphasized, however, that this number varies with the type of progeny, since different types of progenies correspond to different effective population sizes (Ne). Studies with different types of maize progenies suggested a minimum effective size of 200. Since each progeny of half-sibs, full sibs and of selfing (S_1) has effective sizes of 4, 2 and 1, respectively, a minimum of 50, 100 and 200 progenies of each type is required to adequately sample the population (Souza Júnior 2001). Therefore, studies on maize in the literature often report the use of at least 50 half-sib progenies. For carrot, 42 and 52 half-sib progenies were found to be sufficient to represent the traits xylem and phloem color, respectively (Silva and Vieira 2010). Accordingly, in this study it was concluded that to test inbred soybean populations, the number of lines should be around 150.

Tamanho da Amostra para a Avaliação de Populações Endogâmicas de Soja

Resumo – *No melhoramento genético de plantas é muito importante o conhecimento do tamanho da amostra para avaliar as populações. Uma amostra pequena reduz a chance de seleção de genótipos superiores, enquanto que uma amostra grande pode acarretar aumentos desnecessários de custo e trabalho. Uma população composta de 192 linhagens de soja foi dividida aleatoriamente em grupos de 24 linhagens, que foram avaliadas para a produção de grãos em oito experimentos em blocos ao acaso. Foram realizadas análises de variância para os oito experimentos e para o agrupamento destes, obtendo-se análises de variância com 24, 48, 72, 96, 120, 144, 168, and 192 linhagens. Observou-se que conforme o tamanho da amostra aumentou, os intervalos de confiança das estimativas de parâmetros diminuíram, estabilizando com amostras de 144 linhagens. Portanto, uma amostra apropriada para a avaliação de uma população endogâmica de soja deve conter aproximadamente 150 linhagens.*

Palavras-chave: *Variância genética, herdabilidade, precisão de estimativas de parâmetros, intervalo de confiança.*

REFERENCES

Badan AC, Geraldi IO and Guimarães EP (1998) **Tamanho da amostra para representar as populações de arroz**. CNA-IRAT e PCT-4. CIAT, Colômbia, 4p.

Barbin D (1993) **Componentes de variância: teoria e aplicações**. FEALQ, Piracicaba, 120p.

Barona MAA, Colombari-Filho JM, Santos VS and Geraldi IO (2012) Epistatic effects on grain yield of soybean [*Glycine max* (L.) Merrill]. **Crop Breeding and Applied Biotechnology 12:** 231-236.

Falconer DS and Mackay TFC (1996) **Introduction to quantitative genetics**. Longman Group Limited, Edinburgh, 464p.

Knapp SJ, Ross WM and Stroup WW (1987) Precision of genetic variance and heritability estimates from sorghum populations. **Crop Science 27**: 265-268.

Knapp SJ, Stroup WW and Ross WM (1985) Exact confidence intervals for heritability on a progeny mean basis. **Crop Science 25**: 192-194.

Marquez-Sanchez F (1972) Tamaño de muestra para representar poblaciones de maiz. **Agrociencia 8**: 163-177.

Marquez-Sanchez F and Hallauer AR (1970) Influence of sample size on the estimation of genetic variances in a synthetic variety of maize. I. Grain yield. **Crop Science 10**: 357-361.

Mather K and Jinks JL (1982) **Biometrical genetics**. Cambridge University Press, Cambridge, 396p.

Omolo E and Russel WA (1971) Genetic effects of population size in the reproduction of two heterogeneous maize populations. **Iowa State Journal Science 45**: 499-512.

Pinto RMC, Lima-Neto FP and Souza Jr. CL (2000) Estimativa do número apropriado de progênies S_1 para a seleção recorrente em milho. **Pesquisa Agropecuária Brasileira 35**: 63-73.

Silva GO and Vieira JV (2010) Seleção e número mínimo de famílias para a avaliação de parâmetros de cor em uma população meio-irmã de cenoura. **Revista Ceres 57**: 66-72.

Steel RGD and Torrie JH (1980) **Principles and procedures of statistics**. McGraw Hill, New York, 633p.

Souza Júnior CL (2001) Melhoramento de espécies alógamas. In Nass A, Valois ACC, Melo IS and Valadares-Inglis MC (Eds.) **Recursos genéticos e melhoramento: plantas**. Editora Fundação MT, Rondonópolis, p. 159-199.

Vencovsky R and Barriga P (1992) **Genética biométrica no fitomelhoramento**. Sociedade Brasileira de Genética, Ribeirão Preto, 486p.

Vencovsky R and Crossa J (2003) Measurements of representativeness used in resources conservation and plant breeding. **Crop Science 43**: 1912-1921.