Sample size for the assessment of soybean inbred populations

José Manoel Colombari Filho† and Isaias Olívio Geraldi‡

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 (F₂ or F₃) or in advanced generations of inbreeding (from F₆ 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). 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 F₂ bulk method was used from F₂₃ to F₂₇, beginning with 20...
Table 1. Analysis of variance of the eight experiments, overall mean (X), and coefficients of experimental variation (CV %) for soybean grain yield (g m⁻²)

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

* and **: significant (p ≤ 0.05) and significant (p ≤ 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 ($\hat{\sigma}^2_1$) 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 ($\hat{h}^2_X$). 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 ($\hat{\sigma}^2_1$) 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 ($\hat{\sigma}^2_1$) and corresponding confidence intervals, for sample sizes of 24, 48, 72, 96, 120, 144, 168, and 192 lines for soybean grain yield (g m$^{-2}$).](image)

<table>
<thead>
<tr>
<th>SV</th>
<th>24</th>
<th>48</th>
<th>72</th>
<th>96</th>
<th>120</th>
<th>144</th>
<th>168</th>
<th>192</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lines/Exp.</td>
<td>37,222.0 **</td>
<td>36,895.6 **</td>
<td>38,550.8 **</td>
<td>37,592.0 **</td>
<td>38,800.1 **</td>
<td>38,412.6 **</td>
<td>41,552.8 **</td>
<td>40,666.2 **</td>
</tr>
<tr>
<td>Error/Exp.</td>
<td>4,357.2</td>
<td>3,951.3</td>
<td>3,876.3</td>
<td>4,047.5</td>
<td>3,783.4</td>
<td>3,756.7</td>
<td>3,852.8</td>
<td>3,798.0</td>
</tr>
<tr>
<td>$\hat{\sigma}^2_1$</td>
<td>6,573.0</td>
<td>6,588.8</td>
<td>6,934.9</td>
<td>6,708.9</td>
<td>7,003.3</td>
<td>6,931.2</td>
<td>7,540.0</td>
<td>7,373.6</td>
</tr>
<tr>
<td>$\hat{h}^2_X$</td>
<td>0.883</td>
<td>0.891</td>
<td>0.898</td>
<td>0.892</td>
<td>0.902</td>
<td>0.902</td>
<td>0.907</td>
<td>0.907</td>
</tr>
</tbody>
</table>

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

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


Sample size for the assessment of soybean inbred populations


