Theor. Appl. Genet. 51, 29-33 (1977)

Selection Studies in Sugarcane (*Saccharum* sp. hybrids)

III. A Method to Determine Sample Size for the Estimation of Population Variance *

K.K. Wu, D.J Heinz, H.K. Meyer and S.L. Ladd
Department of Genetics and Pathology, Hawaiian Sugar Planters' Association Experiment Station, Aiea, Hawaii (USA)

Summary. An approximate method to determine sample size for the estimation of population variance, σ^2, is given. The estimate of σ^2 is denoted as s^2. Based on the assumption of a normal distribution for $(s^2/\sigma^2 - 1)$, the sample size is approximately equal to $20{,}000\, z_p^2/k^2$; where z is a standard normal deviate, p is the probability that $\Delta s^2 (\equiv 100|s^2 - \sigma^2|/\sigma^2)$ is less than, or equal to, a critical value k, and k (measured as Δs^2) is the desired precision of s^2.

The expected value of Δs^2, with respect to sample size, and the expected cumulative frequencies of Δs^2 over sample size for various k values are given. Their goodness of fit to the observed results was satisfactory except for populations that were different from normal. The observed values were taken from a study on four yield components in five sugarcane polycross progenies, grown in two contrasting environments over 2 years in three selection stages.

The expected Δs^2 was found to be independent of the population coefficient of variance.

Key words: Sugarcane - Sample Size - Population - Variance - Polycross

Introduction

Variance is an important population parameter, especially for studies in quantitative genetics. Sprague (1966) mentioned that a sample size of 250, with two replications, is commonly used in North Carolina experiments to provide a reliable estimate of variance components in corn. Skinner (1971) suggested that 75 seedlings from a cross in sugarcane would be satisfactory in determining the value of a cross. Wu et al. (In preparation) suggested that a sample size should never be less than 40 to estimate mean and acceptable variance in a hybrid population of surgarcane. An empirical sampling method was used by Wu et al. (In preparation), but no mention of methods was made by Skinner (1971) and Sprague (1966).

The objective in this study was to find an approximate method which could be used in determining the sample size for the estimation of population variance.

Methods and Results

The variance of a quantitative character, x, in a population is usually denoted as σ^2, and its estimate from a random sample of the population, as s^2. The value of Δs^2, defined as $100|s^2 - \sigma^2|/\sigma^2$, is used here as a measurement for the precision of the estimate: the closer Δs^2 is to zero, the more precise the estimate.

Two conditions are to be considered in obtaining an estimate for a given sample size: (1) that Δs^2 should be less than or equal to critical value, k, of certain precision of s^2, and (2) that $\Delta s^2 \leqslant$ should occur frequently. Therefore, information is obtained on the precision and confidence of the estimate of s^2 for a particular sample size.

Based on the assumption that $(s^2/\sigma^2 - 1)$ is normally distributed, the expected Δs^2 (or $E\Delta s^2$) is approximately equal to $112.8/\sqrt{n-1}$, and its standard deviation (SD) is approximately $85.2/\sqrt{n-1}$ (Appendix 1). The relationship between $E\Delta s^2$ and $E\Delta s^2$ + SD with respect to sample size, n, is shown in Fig. 1. Based on data in Fig. 1, a sample size of 30 may have an expected $\Delta s^2 = 20$, with a $E\Delta s^2$ + SD of about 40. Again assuming $(s^2/\sigma^2 - 1)$ is normally distributed, the probability that Δs^2 is less than or equal to k can be obtained by $P(|z| \leqslant k\sqrt{n-1}/100\sqrt{2})$, where z is a standard normal deviate and n the sample size

* Research suppoted in part by USDA, ARS, grant #12-14-5001-34. Published with the approval of the Director as Paper No. 412 in the Journal Series of the Experiment Station, Hawaiian Sugar Planters' Association.

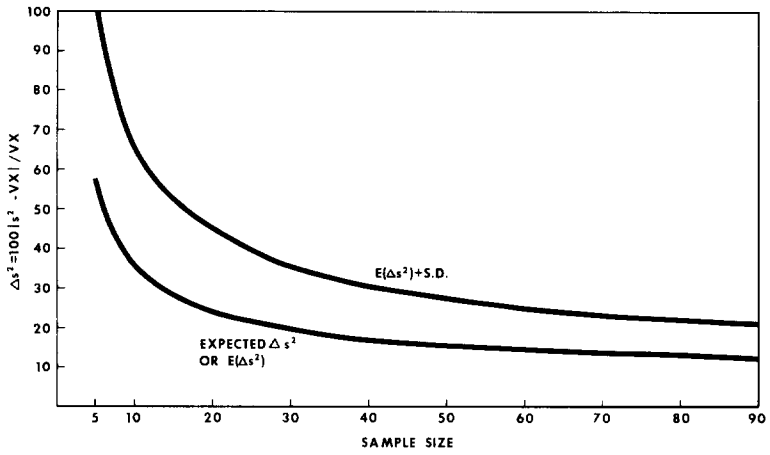

Fig. 1. Expected Δs^2 for different sample sizes

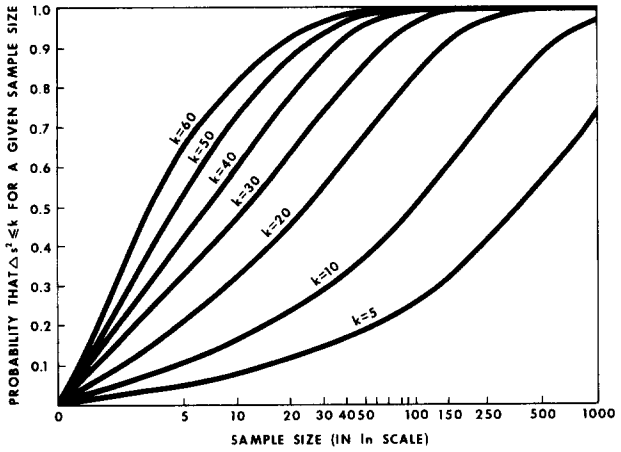

Fig. 2. Cumulative frequencies of Δs^2 over sample sizes for various k values

(Appendix 2). Shown in Fig. 2 are the expected cumulative frequencies of Δs^2, $[P(\Delta s^2 \leqslant k)]$ for seven different critical values of k and various sample sizes (in natural logarithmic scale). For a given sample size of 30, and a critical value of k = 40, the probability that the estimated s^2 with Δs^2 is less than or equal to 40 is about 0.87.

The observed results were obtained from repeated random samples drawn from 120 populations. The populations were derived from data on four yield components in five sugarcane polycross progenies, grown in two contrasting environments over 2 years in three selection stages (Wu et al., In preparation). Each population consisted of 250 data points. Δs^2 and $\bar{\Delta} s^2$ were calculated for each sample size. Results of chi-square test (Appendix 3) for goodness of fit between the observed and the expected Δs^2 for four different characters are listed in Table 1. Of the 24 tests, 17 (or 70%) were not significantly different (P = 0.01) from the expected Δs^2.

The observed frequencies for eight classes of Δs^2 were obtained for each sample size. The classes were 0 to 4.9, 5 to 9.9, 10 to 19.9, 20 to 29.9, 30 to 39.9, 40 to 49.9, 50 to 59.9, and more than 60 (Fig. 2). For each sample size, 150 values of Δs^2

Table 1. Chi-square tests for goodness of fit between observed $\bar{\Delta} s^2$ and expected $\Delta(s^2)$ for each character and selection stage in windward (W) and leeward (L) locations

Selection	Chi-square							
	Stalk number		Stalk diameter		Stalk length		Refractometer solids	
Stage	W	L	W	L	W	L	W	L
FT1	8.1	151.3**	10.9	4.4	5.7	9.5	17.1*	15.6*
FT2	7.4	68.1**	5.7	7.4	3.6	30.4**	16.0*	15.7*
FT4	5.0	44.8*	7.4	9.5	29.3**	21.2**	11.4	29.7**

*, ** Significant at 5% and 1% level, respectively

$\chi^2_{(s)} = \sum^{s} 1.7518\, m\, D^2$; where $D = (\bar{\Delta} s^2 / E\Delta s^2) - 1$; $E\Delta s^2 \tilde{=} 112.83/\sqrt{n-1}$;

s = 8, the eight different sample sizes as shown in Fig. 1;
m = 50, the 50 Δs^2 calculated from each of 50 random samples obtained from 5 populations, 10 random samples each

Table 2. Chi-square tests for goodness of fit between the observed and expected frequency distributions of Δs^2 for each sample size in windward (W) and leeward (L) locations

Sample size	Degree of freedom	Chi-squares							
		Stalk number		Stalk diameter		Stalk length		Refractometer solids	
n		W	L	W	L	W	L	W	L
5	7	4.04	17.45*	4.17	3.47	4.04	9.86	14.67*	6.12
10	7	7.96	16.75*	15.91*	6.34	8.03	2.07	6.46	13.95
20	7	5.07	16.09*	8.48	1.40	9.61	8.21	11.67	2.17
30	6	8.44	7.39	4.61	7.38	8.51	9.75	8.77	7.93
40	6	19.24**	37.75**	4.14	17.14**	8.63	6.62	5.77	7.04
50	5	3.54	77.95**	2.52	1.96	5.93	2.62	10.85	7.96
60	5	12.04*	111.05**	5.09	8.04	2.07	11.05	3.66	4.89
70	4	6.75	47.07**	10.49*	1.70	5.23	3.04	9.66*	11.28*

*, ** Significant at 5% and 1% level, respectively
$\chi^2 = \sum (\text{Expected} - \text{Observed})^2/\text{Expected}$
For each sample size, the total observed frequencies was 150 Δs^2 obtained by 10 random samples from each of 15 populations (5 progenies × 3 selection stages).

Table 3. Frequency distributions with significant skewness and kurtosis, in percent of total number (120) of populations tested

Parameter	Characters			
	Stalk number	Stalk diameter	Stalk length	Refractometer solids
Skewness	87	20	23	40
Kurtosis	57	33	27	40

were obtained by taking 10 random samples from each of 15 populations (5 progenies in 3 selection stages) within each location and yield component. Observed frequency distributions of Δs^2 were then compared with their expected frequency distributions which were obtained by conversion from the cumulative frequency distributions as shown in Fig.2. Chi-square tests for goodness of fit between the observed and expected frequency distributions are given in Table 2. Fifty-eight out of 64 chi-square tests (or 90%) were not significantly different (P = 0.01).

The Kolmogorov-Smirnov test for goodness of fit (Sokal and Rohlf 1969) was used to check the normality of each population. Of the 120 populations, 116 were not significantly different (P = 0.05) from the null hypothesis of normality. The four populations which were significantly different from normal were all different for number of stalks: one polycross progeny at the leeward location in all 3 selection stages and one polycross progeny at the windward location in the first selection stage.

Snedecor and Cochran (1969) pointed out that populations may be noticeably skewed, although the chi-square test does not reject the null hypothesis of normality. Skewness and kurtosis were determined on all 120 populations. The data for stalk number had the highest number of frequency distributions with skewness and kurtosis (Table 3). The four non-normal populations had the highest magnitude of skewness (g_1) and kurtosis (g_2): 2.96(g_1) and 11.80(g_2), 2.45 and 10.60, 1.64 and 5.20, and 1.03 and 1.30.

Discussion

Large chi-square values were mainly obtained for stalk number in the leeward location for both the expected Δs^2 and its frequency distributions (Tables 1 and 2). They were probably caused by the three non-normal populations of one progeny having extremely large values of skewness and kurtosis for stalk number. It appears that the adequacy of the assumption that $(s^2/\sigma^2 - 1)$ is normally distributed depends on the distribution of a variable in the population. The percentage of non-significant chi-square tests between the expected and observed Δs^2 (Appendix 3) was 78% (P = 0.01) or 55% (P = 0.05), and that between the expected and the observed frequency distributions of Δs^2 was 97% (P = 0.01), or 87% (P = 0.05), if one ignores those tests for number of stalks in Tables 1 and 2. The agreement between expected and observed results for Δs^2 was lower than that for the frequency distributions of Δs^2. The lower percentage of agreement between the observed and expected Δs^2 is probably due to the method used here (Appendix 3).

Since the percentage of agreement between the observed and the expected frequency distributions of Δs^2 was satisfactory, the latter can be used in estimating the sample size for population variance. The expected Δs^2 and its standard deviation as shown in Fig.2 could provide information on the precision and confidence of s^2 for a given sample size.

A working formula to estimate the sample size, n, for the variance of a population is as follows:

$$n \tilde{=} 20{,}000\ z_p^2/k^2, \text{ (Appendix 2)}$$

where z is a standard normal deviate, k is the critical value with certain precision of s^2, and p is the probability that $\Delta s^2 \leqslant k$. Values of z can be found in a table of the standard normal distribution. For example, z equals 1.64 when p equals 0.9. If the critical value of k is equal to 10, the sample size will be 537. This sample size can also be approximated from Fig.2.

For a given sample size, k represents the precision and p the confidence of the estimate. The choice of k and p is often a matter of subjective judgment. Statistically, for instance, the p value could be more than or equal to 0.9, while the k value could be less than or equal to 10 in order to obtain a satisfactory estimate. This will require a sample size of about 500, as calculated from the formula. This was the sample size suggested by Sprague (1966). However, the population variance may not be as useful as the population mean, especially when estimating the relative importance of a cross. One may reduce the precision and confidence of the variance estimated and examine fewer individuals per cross to be able to test more crosses in the same land area (Skinner 1971; Wu et al., In preparation).

Wu et al. (In preparation) in their empirical sampling study, found a similar pattern of Δs^2 with respect to sample size for the four characters studied although the population CV's were different for different characters. From the expected Δs^2, it was shown that CV was not related to $E\Delta s^2$. Therefore, regardless of the CV for a variable in a population, its $E\Delta s^2$ will remain the same as shown in Fig.1.

The probability that Δs^2 is less than or equal to k can also be obtained from a chi-square distribution: $P(\Delta s^2 \leqslant k) = P[(n-1)(1-k/100) \leqslant \chi^2_{(n-1)} \leqslant (n-1)(1+k/100)]$. This would be considered a more precise method. The probability obtained from this method, for example, is equal to 0.11 for n = 5, k = 10; 0.43 for n = 5, k = 40; 0.29 for n = 30, k = 10; 0.88 for n = 30, k = 40; 0.47 for n = 80, k = 10; 0.98 for n = 80, k = 40. These values of probability are very close to the results approximated from Fig.2 for the same values of n and k. We did not use the chi-square method because no simple formula for the sample size can be derived from it.

Appendix 1

Expected Δs^2 and its standard deviation:

Let $y = (s^2/\sigma^2) - 1$, and hence $E(Y) = 0$, $V(Y) \tilde{=} 2/(n-1)$, where n is the sample size. Assuming that y is approximately normally distributed with mean zero and variance $2/(n-1)$, then $E\Delta s^2 = E[100|s^2 - \sigma^2|/\sigma^2] = 100\ E|Y| \tilde{=} 100 \sqrt{\frac{2}{\pi}} \sqrt{\frac{2}{n-1}} = 112.8/\sqrt{n-1}$ and $V\Delta s^2 = 100^2\ V|Y| \tilde{=} 100^2\ (1 - 2/\pi) \times [2/(n-1)]$ (Kendall 1952). The standard deviation of Δs^2 is approximately $85.2/\sqrt{n-1}$.

Appendix 2

Probability that $\Delta s^2 \leqslant k$ for a given sample size of n:

Under the same assumption as in Appendix 1,

$$P(\Delta s^2 \leqslant k) = P(100|Y| \leqslant k) \tilde{=} P(|z| \leqslant k\sqrt{n-1}/100\sqrt{2})$$

where z is a standard normal deviate. Let $P(\Delta s^2 \leqslant k) = p$, then $z_p^2 \leqslant k^2 (n-1)/20{,}000$, and $n \tilde{=} 20{,}000\ z_p^2/k^2$.

Appendix 3

Chi-square test for goodness of fit between expected Δs^2 and observed $\bar{\Delta} s^2$, the average of Δs^2, of the same sample size $n \left(\text{or } \bar{\Delta} s^2 = \sum_{i=1}^{m} \Delta s_i^2/m \right)$:

Under the same assumption in Appendix 1 and considering

$$\bar{\Delta} s^2/E\Delta s^2 = \frac{100 \sum_{}^{m} \left| \frac{s^2 - \sigma^2}{\sigma^2} \right|}{m} \Bigg/ \left(100 \sqrt{\frac{2}{\pi}} \sqrt{\frac{2}{n-1}} \right) =$$

$$\frac{1}{m\sqrt{\frac{2}{\pi}}} \frac{\sum_{}^{m} \left| \frac{s^2}{\sigma^2} - 1 \right|}{\sqrt{\frac{2}{n-1}}} \tilde{=} \frac{1}{m\sqrt{\frac{2}{\pi}}} \sum_{}^{m} |z| \text{; where } z \sim N(0, 1).$$

Since $E\left(\sum_{}^{m} |z|\right) = mE|z| = m\sqrt{\frac{2}{\pi}}$ and $V\left(\sum_{}^{m} |z|\right) = mV|z| \tilde{=} m\left(1 - \frac{2}{\pi}\right)$, then $\sum_{}^{m} |z| \tilde{=} \left(\frac{\bar{\Delta}s^2}{E\Delta s^2}\right)\sqrt{\frac{2m^2}{\pi}}$ is approximately distributed as $N\left(\sqrt{\frac{2m^2}{\pi}}, m\left(1 - \frac{2}{\pi}\right)\right)$. (1)

From (1), $\frac{\left(\frac{\bar{\Delta}s^2}{E\Delta s^2}\right)\sqrt{\frac{2m^2}{\pi}} - \sqrt{\frac{2m^2}{\pi}}}{\sqrt{m\left(1 - \frac{2}{\pi}\right)}}$ is approximately distributed as $N(0, 1)$. It implies

$$\left[\frac{\sqrt{\frac{2m^2}{\pi}}\left(\frac{\bar{\Delta}s^2}{E\Delta s^2} - 1\right)}{\sqrt{m\left(\frac{\pi - 2}{\pi}\right)}}\right]^2 \sim \chi^2_{(1)} \quad \text{that is}$$

$$\frac{m^2\left(\frac{2}{\pi}\right)\left(\frac{\bar{\Delta}s^2}{E\Delta s^2} - 1\right)^2}{m\left(1 - \frac{2}{\pi}\right)} = 1.7518\, m\left(\frac{\bar{\Delta}s^2}{E\Delta s^2} - 1\right)^2 \sim \chi^2_{(1)},$$

for a given sample size of n.

Literature

Kendall, M.G.: The advanced theory of statistics. Vol. I. 5th ed., New York: Hafner Publ. 1952

Skinner, J.G.: Selection in sugarcane: A review. Proc. ISSCT 14, 149-162 (1971)

Snedecor, G.W.; Cochran, W.G.: Statistical methods. 6th ed., Ames, Iowa: Iowa State Univ. Press 1969

Sokal, R.R.; Rohlf, F.J.: Biometry. San Francisco: W.H. Freeman 1969

Sprague, G.F.: Quantitative genetics in plant improvement. In: Plant breeding (ed. Frey, K.J.), chapter 8. Ames, Iowa: Iowa State Univ. Press 1966

Wu, K.K.; Heinz, D.J; Meyer, H.K.; Ladd, S.L.: Selection studies in sugarcane (*Saccharum* sp. hybrids) II. Minimum sample size for estimating progeny mean and variance. (In preparation)

K.K. Wu
D.J Heinz
H.K. Meyer
S.L. Ladd
Department of Genetics and Pathology
Hawaiian Sugar Planters' Association
Experiment Station
Aiea, Hawaii 96701 (U.S.A.)

Received April 29, 1977
Accepted by H.F. Linskens