Jour. Ind. Soc. Ag. Statistics
57 (Special Volume), 2004 : 103-115

Sample Size for Collecting Plant Diversity

R.L. Sapra
Indian Agricultural Research Institute, New Delhi

SUMMARY

The present paper, re-examines the general model by Sapra *et al.* [13] which was based on the mathematical deduction and empirical verifications. It develops a complete mathematical proof of the general expression determining the minimum sample size for collecting plant germplasm diversity for genetic conservation with an overall objective of retaining at least a copy of each allele with some preassigned probability of conservation. It considers sampling from a large heterogeneous diploid or auto-polyploid population under a broad range of mating systems.

Key words : Allele, Auto-polyploid, Conservation, Diploid, Germplasm, Polysomic model, Sample.

1. Introduction

Plant explorers and conservationists are generally concerned with the ultimate sampling problem of collecting genetic material (vegetative or seeds) from large populations with a view to conserve the germplasm diversity with a certain degree of assurance. The number of plants needed to conserve the germplasm diversity has been discussed in a number of papers (Allard [1], Bennett [2], Qualset [12], Marshall and Brown [10], Bogyo *et al.* [3], Chapman [5], Yonezawa [15], Crossa [6], Yonezawa and Ichihashi [16], Crossa *et al.* [7], Lawrence *et al.* ([8], [9]) and Brown and Marshall [4] using probability models mainly for diploid species. Sapra *et al.* [13] proposed .a polysomic model for collecting plant variability from a population under broad range of mating system. However, the said model was based on the empirical verification of the results. The present paper, re-examines the said model and develops the general expression with complete algebraic treatment.

2. Diploid Model

Di-allelic and Single Locus

In considering the problem of sample size, it is convenient to begin by considering the simplest case of conserving pair of alleles at a single locus in a sample drawn at random from a population of individual plants. The analysis of the problem is then extended to the situation where there are more than two alleles per locus at several of the independent loci. Suppose we wish to conserve a gene, which occurs in a population in just two allelic forms, A_1 and A_2. The genetical variation at this locus will have been completely conserved when a sample drawn from the population contains at least one copy of each of these alleles. We assume that the species in question is a diploid one. Let the frequencies of the three genotypes at this locus be G_1 of A_1A_1, G_2 of A_1A_2 and G_3 of A_2A_2, where $G_1 + G_2 + G_3 = 1$. Consider, now, the composition of a sample of size, m, drawn at random from this population. Our objective of conserving the variation of this locus will be achieved if the sample contains one heterozygote, A_1A_2 or one each of the homozygotes, of A_1A_1 and A_2A_2. There is a risk, however, that a sample will contain either A_1A_1 or A_2A_2 individuals only, a possibility which we wish to avoid. The probability that a sample size m contains A_1A_1 individuals only is, by the normal probability rules, $(G_1)^m$. Similarly the probability that all m individuals in the sample are A_2A_2 is $(G_3)^m$. Hence the probability that the sample does not consist of either A_1A_1 or A_2A_2 individuals only is $1 - (G_1)^m - (G_3)^m$. But any sample which is not one or other of these extreme kinds must contain at least one A_1 and at least one A_2 allele. Thus, the probability, P that a randomly drawn sample of size m contains at least one copy of each allele is

$$P[A_1, A_2] = 1 - (G_1)^m - (G_3)^m$$

In order to define the relationship between the genotype frequencies, G_1, G_2 and G_3; and the allele frequencies of A_1 and A_2, p_1 and p_2, respectively, where $p_1+p_2=1$. When individuals set a proportion of their seeds, s, by self-fertilization and the remainder, (1-s) by random mating, the relationship between genotype frequencies and allele frequencies is

Table 1. Diploid with 2 alleles

Genotype	Designation	Frequency
A_1A_1	G_1	$p_1^2(1 - F_1) + p_1F_1$
A_1A_2	G_2	$2p_1p_2(1 - F_1)$
A_2A_2	G_3	$p_2^2(1 - F_1) + p_2F_1$

where $F_1 = s/(2 - s)$ is the coefficient of inbreeding. Let the allele A_2 is rare having a frequency of p_0 (say $p_0 \le 0.1$). Then the probability P is given by

$$P = 1 - (1 - p_0)^m - (p_0)^m$$

or $$\alpha = (1 - p_0)^m + (p_0)^m \tag{1}$$

where, $\alpha = 1 - P$

Multi-allelic and Single Locus

3 Alleles

Let us consider a population with 3 alleles A_1, A_2, and A_3 having allele frequencies as p_1, p_2 and p_3 at a locus. The five genotypes along with the frequencies are as follow

Table 2. Diploid with 3 alleles

Genotype	Frequency
A_1A_1	$p_1^2\,(1 - F_1) + p_1F_1$
A_2A_2	$p_2^2\,(1 - F_1) + p_2F_1$
A_3A_3	$p_3^2\,(1 - F_1) + p_3F_1$
A_1A_2	$2p_1p_2(1 - F_1)$
A_2A_3	$2p_2p_3(1 - F_1)$
A_3A_1	$2p_3p_1(1 - F_1)$

The expression for evaluating sample size, m can be formulated as

$$P\,[A_1, A_2, A_3\,] = (1 - \alpha) = 1 - \sum_{i=1}^{3} P(A_i)^c + \sum_{1=i<j\leq 3}^{3} P(A_iA_j)^c$$

where P $[A_1, A_2, A_3]$ is the probability of including all the 3 alleles at least once in a sample of size m, $P(A_i)^c$ is the probability of missing allele, A_i, and $P(A_i\,A_j)^c$ is the probability of missing both the alleles, A_i and A_j. Let us assume that out of these three alleles one is abundant and the other two are rare having identical frequencies (say p_0). After calculating various probabilities and simplifying the expression we get

$$P[A_1, A_2, A_3\,] = (1 - \alpha)$$

where

$$\begin{aligned}\alpha = (2p_0)^m\,\{2p_0(1 - F_1) + F_1\}^m - 2p_0^{\,m}\,\{p_0\,(1 - F_1) + F_1\}^m \\ + 2(1 - p_0)^m\{(1 - p_0)(1 - F_1) + F_1\}^m \\ - (1 - 2p_0)^m\{(1 - 2p_0)(1 - F_1) + F_1\}^m \qquad (2)\end{aligned}$$

4 Alleles

Let us consider a population with 4 alleles A_1, A_2, A_3 and A_4 having allele frequencies as p_1, p_2, p_3 and p_4. The ten genotypes along with the frequencies are as follow

Table 3. Diploid with 4 alleles

Genotype	Frequency
A_1A_1	$p_1^2(1 - F_1) + p_1F_1$
A_2A_2	$p_2^2(1 - F_1) + p_2F_1$
A_3A_3	$p_3^2(1 - F_1) + p_3F_1$
A_4A_4	$P_4^2(1 - F_1) + p_4F_1$
A_1A_2	$2p_1p_2(1 - F_1)$
A_1A_3	$2p_1p_3(1 - F_1)$
A_1A_4	$2p_1p_4(1 - F_1)$
A_2A_3	$2p_2p_3(1 - F_1)$
A_2A_4	$2p_2p_4(1 - F_1)$
A_3A_4	$2p_3p_4(1 - F_1)$

The probability expression for evaluating m for including at least a copy of each allele is

$$P[A_1, A_2, A_3, A_4]$$

$$= 1 - \sum_{i=1}^{4} P(A_i)^c + \sum_{1=i<j\le 4}^{4} P(A_iA_j)^c - \sum_{1=i<j<k\le 4}^{4} P(A_iA_jA_k)^c$$

where $P(A_i)^c$ is the probability of missing A_i , $P(A_i\ A_j\ )^c$ is the probability of missing both A_i and A_j; and $P(A_i\ A_j\ A_k)^c$ is the probability of missing 3 alleles (A_i, A_j, A_k) at a time. Let us assume again that A_1 is abundant and A_2, A_3 and A_4 are rare and each having an identical frequency of p_0. After calculating various probabilities and simplifying the expression we get

$$P\,[\,A_1, A_2, A_3, A_4\,] = 1 - \alpha$$

where

$$\begin{aligned}\alpha = {} & (3p_0)^m\{3p_0(1 - F_1) + F_1\}^m - 3(2p_0)^m\{2p_0(1 - F_1) + F_1\}^m \\ & + 3p_0^m\{p_0(1 - F_1) + F_1\}^m + 3(1 - p_0)^m\{(1 - p_0)(1 - F_1) + F_1\}^m \\ & - 3(1 - 2p_0)^m\{(1 - 2p_0)(1 - F_1) + F_1\}^m \\ & + 3(1 - 3p_0)^m\{(1 - 3p_0)(1 - F_1) + F_1\}^m \qquad (3)\end{aligned}$$

On critically examining (1), (2) and (3), we can generalize the case for ‘a’ alleles, under the assumption that there are ‘a’ alleles at each locus and out of ‘a’ alleles, (a-1) are rare having identical frequency of p_0 and the a^th allele having a frequency of $[1 - (a - 1)p_0]$. Thus, for ‘a’ alleles we get the following probability expression for evaluating the sample size

$$\begin{aligned}\alpha = {} & \sum_{r=1}^{a-1}(-1)^{r-1}\binom{a-1}{r-1}\{(a-r)p_0\}^m\{(a-r)p_0(1-F_1)+F_1\}^m \\ & + \sum_{r=1}^{a-1}(-1)^{r-1}\binom{a-1}{r}\{(1-rp_0\}^m\{(1-rp_0)(1-F_1)+F_1\}^m \qquad (4)\end{aligned}$$

or $\alpha = \alpha_1 + \alpha_2$

where α_1 is the first summation and α_2 is the second summation. Now we will evaluate α_1 and α_2 as follow

The rth term of summation α_1, contains the term $\{(a - r)p_0(1 - F_1) + F_1\}^m$ which is so small that it is almost negligible. Therefore

$$\alpha_1 \approx 0$$

(because $(a - r)p_0 < 1$

$\therefore \quad (a - r)\, p_0 (1 - F_1) < (1 - F_1)$

$\Rightarrow \quad (a - r)\, p_0 (1 - F_1) + F_1 < 1$)

Let us evaluate α_2. We will consider the first term of α_2 and it will be shown for this case, that the other terms of α_2 are negligible and can be dropped. After dropping these terms we get

$$\alpha_2 = (a - 1)(1 - p_0)^m\{(1 - p_0)(1 - F_1) + F_1\}^m$$

Since $\alpha_1 \approx 0$, therefore, $\alpha \approx \alpha_2$. Taking log on both sides and solving for m we get

$$m > \frac{\log\alpha - \log(a-1)}{\log[(1-p_0)^2(1-F_1)+F_1]} \tag{5}$$

Proof for ignoring all the terms of α_1 and α_2 (except first) are negligible by calculating the limit of rth term of both the summations. The rth term of α_1 is

$$[(a-r)p_0]^{m'}[(a-r)p_0(1-F_1)+F_1]^{m'} = \mathrm{Exp}(Y_1)$$

where

$$Y_1 = \frac{\{\log\alpha - \log(a-1)\}}{\log\{(1-p_0)^2(1-F_1)+F_1\}}\log[(a-r)p_0][(a-r)p_0(1-F_1)+F_1]$$

where $m' = m$ as given by (5) and Y_1 is equal to the terms inside the exponential. Now it can be shown that Y_1 is negative, therefore, Exp (Y_1) is maximum when Y_1 tends to zero. In that case $\mathrm{Exp}(Y_1) = 1$. Therefore minimum value of p_0, a and r that make Y_1 approaching to 0 are required. Now we will obtain the limiting value of Exp (Y_1) by calculating the individual limits

$$\operatorname*{lt}_{p_0 \to 0}[\log(1-p_0) + \log\{(1-p_0)(1-F_1)+F_1\}]$$

$$= \operatorname*{lt}_{p_0 \to 0}\log(1-p_0) + \operatorname*{lt}_{p_0 \to 0}\log\{(1-p_0)(1-F_1)+F_1\}$$

$$= -p_0(2-F_1)$$

$$\operatorname*{lt}_{p_0 \to 0}\log(a-r)p_0 + \operatorname*{lt}_{p_0 \to 0}\log[(a-r)p_0(1-F_1)+F_1] = -\{1-(a-r)p_0\}\{2-F_1\}$$

Putting the values of these limits, we get the limiting value of the rth term of α_1 as follows

$$\text{Exp}(Y) = \left[\frac{\alpha}{a-1}\right]\{1-(a-r)p_0\}/p_0$$

The maximum value of this limit is attained when r = 1 and a = 2

= 1.91E–25 for $\alpha = 0.05$ and $p_0 = 0.05$

= 1.0E–9 for $\alpha = 0.05$ and $p_0 = 0.10$

Now we will prove that the terms of summation α_2 (other than the first i.e. $r \geq 2$) are negligible. The rth term of α_2 is

$$[1-rp_0]^{m'}[(1-rp_0)(1-F_1)+F_1]^{m'} = \text{Exp}(Y_2)$$

where

$$Y_2 = \frac{\{\log\alpha - \log(a-1)\}}{\log\{(1-p_0)^2(1-F_1)+F_1\}}\log[(1-rp_0)][(1-rp_0)(1-F_1)+F_1]$$

where m′ = m as given by (5) and Y_2 is equal to the terms inside the exponential. Now it can be shown that Y_2 is negative, therefore, Exp (Y_2) is maximum when Y_2 tends to zero. In that case Exp (Y_2) = 1. Therefore, minimum value of p_0, a and r that make Y_2 approaching to 0 are required. Now we will obtain the limiting value of Exp (Y_2) by calculating the individual limits.

$$\underset{p_0\to 0}{\text{lt}}\log[(1-rp_0)\{(1-rp_0)(1-F_1)+F_1\}]$$

$$= \underset{p_0\to 0}{\text{lt}}\log(1-rp_0) + \underset{p_0\to 0}{\text{lt}}\log\{1-rp_0(1-F_1)\}$$

$$= -rp_0 - rp_0(1-F_1) = -rp_0(2-F_1)$$

Putting the values of individual limits, we get the limiting value of the rth term of α_2 as follows

$$\text{Exp}(Y) = \left[\frac{\alpha}{a-1}\right]^r$$

$$= \alpha^{(1-p_0)/p_0}$$

Exp(Y_2) is maximum when r = 2 and a = 3. Therefore, the maximum value of the limit is at $(\alpha/2)^2$; and for α = 0.05 and 0.1, values are 0.000625 and 0.0025. Thus, these are the maximum possible values of the second term of summation α_2.

Multi-allelic and Multi-locus

Let there is λ independent loci and at each locus the first (a-1) alleles occur at an identical low frequency of p_0 and that the ath allele occurs at a frequency of $[1-(a-1)p_0]$. Then the probability for capturing alleles from all the loci is the product of the probabilities for each locus. The probability for a single locus as from expression (5) is

$$P = 1 - \alpha \approx 1 - (a - 1)\ \{\beta_l\}^m \qquad (6)$$

where $\beta_l = \log\{(1 - p_0)^2(1 - F_l) + (1 - p_0) F_l\}$

Now the probability that at least a copy of each allele from all the loci can be obtained as follow

$$P = 1 - \alpha \approx [1 - (a - 1)\ \{\beta_l\}^m]^\lambda \qquad (7)$$

Taking log on both sides and after simplification, (7) reduces to

$$m > \frac{\log\{1 - (1 - \alpha)^{1/\lambda}\} - \log(a - 1)}{\log \beta_l} \qquad (8)$$

From expression (8), we can derive following two important expressions; one for random mating (s = 0) and the other for exclusive selfing (s = 1).

Random Mating (s = 0)

$$m > \frac{\log\{1 - (1 - \alpha)^{1/\lambda}\} - \log(a - 1)}{2 \log(1 - p_0)} \qquad (8a)$$

Selfing (s = 0)

$$m > \frac{\log\{1 - (1 - \alpha)^{1/\lambda}\} - \log(a - 1)}{\log(1 - p_0)} \qquad (8b)$$

The expression (8a) obtained here for inbred population (s = 1) is the same as of Crossa *et al.* [7] obtained for a population of infinite size subdivided into many highly homozygous lines with a class of alleles.

3. Polysomic Model

Diallelic and Single Locus

Let us consider a large 2k-auto-polyploid population with 2 alleles A_1 and A_2 at a single locus having frequencies p_1 and p_2 respectively, reproducing by constant proportions of selfing (s) and random mating (1 – s), with no double reduction or selection. Such a population at equilibrium can be denoted as

$Z \equiv (A_1^{2k}, A_1^{2k-1} A_2, A_1^{2k-i} A_2^{i}, \ldots\ldots\ldots\ldots\ldots\ldots\ldots, A_2^{2k})$

$p_1^{2k}(1\text{-}F_k) + p_1 F_k,\ {}^{2k}C_1\, p_1^{2k-1} p_2 (1\text{-}F_k),\ {}^{2k}C_i\, p_1^{2k-i} p_2^{i} (1\text{-}F_k)\ldots,\ p_2^{2k}(1\text{-}F_k)$
$+ p_2 F_k$

The sum of the allelic frequencies and the sum of the genotypic frequencies as given above are one. F_k is the theoretical population inbreeding coefficient at equilibrium for a 2k-ploid organism and is related to the proportion of selfing (s) by the following formula (McConnell and Fyfe [11])

$F_k = s / \{2k\text{-}(2k\text{-}1)s\}$

When k = 1, the population becomes a diploid. Our objective of conservation that a randomly drawn sample from a 2k-ploid population captures at least one copy of each of these alleles can be achieved if the sample contains either one of the heterozygotes or one each of the homozygotes A_1^{2k} and A_2^{2k}. The probability of capturing at least one copy of each allele excluding the probability of selecting only one of the homozygotes in a sample of size m, as suggested above by Lawrence *et al.* [8] for diploid models. Thus, following the same logic, the probability that a randomly drawn sample of size m contains at least a single copy of each allele at the said locus is

$$P[A_1, A_2] = 1 - \alpha = 1 - \{p_1^{2k}(1 - F_k) + p_1 F_k\}^m - \{p_2^{2k}(1 - F_k) + p_2 F_k\}^m \qquad (9)$$

We can further simplify the above expression by assuming that the allele A_2 is rare in nature and occurs with a frequency of p_0. Then we can rewrite (9) as

$$\alpha = \left[(1 - p_0)^{2k}(1 - F_k) + (1 - p_0)F_k\right]^m + \left[p_0^{2k}(1 - F_k) + p_0 F_k\right]^m \qquad (10)$$

Tetraploid with 3 Alleles

The expression for a tetraploid population with 3 alleles A_1, A_2, and A_3; and with 15 genotypes is formulated as

$$P[A_1, A_2, A_3] = (1\text{-}\alpha) = 1 - \sum_{i=1}^{3} P(A_i)^c + \sum_{1=i<j\le 3}^{3} P(A_i A_j)^c \qquad (11)$$

After calculating the various probabilities as mentioned earlier and assuming that alleles A_2 and A_3 are rare, having a low frequency of p_0 and A_1 is abundant with a frequency of $(1\text{-}2p_0)$, we get the simplified expression as below

$$\begin{aligned}\alpha = (2p_0)^m\{(2p_0)^3(1 - F_1) + F_1\}^m - 2p_0^m\{p_0^3(1 - F_1) + F_1\}^m \\ + 2(1 - p_0)^m\{(1 - p_0)^3(1 - F_1) + F_1\}^m \\ - (1 - 2p_0)^m\{(1 - 2p_0)^3(1 - F_1) + F_1\}^m \end{aligned} \qquad (12)$$

Now looking at (1), (2), (3), (10) and (12), we can very easily generalize expression for auto-ployploids (2k-ploid) for 'a' alleles as follows

$$\begin{aligned}\alpha = \sum_{r=1}^{a-1}(-1)^{r-1}\binom{a-1}{r-1}\{(a-r)p_0\}^m\left[\{(a-r)p_0\}^{2k-1}(1 - F_k) + F_k\right]^m \\ + \sum_{r=1}^{a-1}(-1)^{r-1}\binom{a-1}{r}\{(1 - rp_0\}^m\{(1 - rp_0)^{2k-1}(1 - F_k) + F_k\}^m \end{aligned} \qquad (13)$$

or $\alpha = \alpha_1 + \alpha_2$

where α_1 and α_2 are the first and second summation of α. Let us again evaluate α_1 and α_2 for auto-polyploids. Here, it can be seen that

$$\alpha_1 \leq \sum_{r=1}^{a-1} (-1)^{r-1} \binom{a-1}{r-1} \{(a-r)p_0\}^m$$

Now the term $[(a\text{-}r)p_0]^m$ is so small that it is almost negligible. Therefore, $\alpha_1 \approx 0$. We will only consider the first term of the summation, α_2. Other terms of α_2 are almost negligible. Thus, after dropping the negligible terms of α_2, we get

$$\alpha \approx (a\text{-}1)(1\text{-}p_0)^m[(1\text{-}p_0)^{2k\text{-}1}(1\text{-}F_k)+F_k]^m$$

The expression for evaluating sample size, m can be written as

$$m > \frac{\log\alpha - \log(a-1)}{\log[\beta_k]} \qquad (14)$$

where

$$\beta_k = \log\{(1-p_0)^{2k}(1-F_k) + (1-p_0)F_k\}$$

Proof for Ignoring Terms of α_1 and α_2

Now we will prove that all the terms of α_1 and α_2 (except first term of α_2) are negligible by calculating the limit of r^th^ term of both the summations. The r^th^ term of α_1 here in this case is

$$[(a-r)p_0]^{m'}[\{(a-r)p_0\}^{2k-1}(1-F_k)+F_k]^{m'} = \text{Exp}(Y_1)$$

where

$$Y_1 = \frac{\log\alpha - \log(a-1)}{\log\{(1-p_0)^2(1-F_k)+F_k\}} \log[(a-r)p_0][\{(a-r)p_0\}^{2k-1}(1-F_k)+F_k]$$

where $m' = m$ as given above and Y_1 is equal to the terms inside the exponential. Now it can be shown that Y_1 is negative, therefore, Exp (Y_1) is maximum when Y tends to zero. In that case Exp $(Y_1) = 1$. Therefore, minimum value of p_0, a, k and r that make Y_1 approaching to 0 are required. Now we will obtain the limiting value of Exp (Y_1) by calculating the individual limits. The limiting value of the r^th^ term is

$$\text{Exp}(Y_1) = \left[\frac{\alpha}{a-1}\right]^L$$

where

$$L = \frac{\{1-(a-r)p_0\} + (1-F_k)(1-(a-r){p_0}^{2k-1})}{p_0 + (1-F_k)\{1-(1-p_0)^{2k-1}\}}$$

Now for maximizing $\text{Exp}(Y_1)$, L has to be minimized. L is minimum when $k = 1$, $r = 1$ and $a = 2$. In that case the limit becomes

$$\text{Exp}(Y_1) = (\alpha)^{(1-p_0)/p_0}$$

$= 1.91\text{E-}25$ for $\alpha = 0.05$ and $p_0 = 0.05$

$= 1.0\text{E-}9$ for $\alpha = 0.05$ and $p_0 = 0.10$

Similarly we can evaluate the rth term of the summation, α_2. The limiting value is given by

$$\mathrm{Exp}(Y_2) = \left[\frac{\alpha}{a-1}\right]^L$$

where

$$L = \frac{rp_0 + (1-F_k)\{1-(1-rp_0)^{2k-1}\}}{p_0 + (1-F_k)\{1-(1-p_0')^{2k-1}\}}$$

$\mathrm{Exp}(Y_2)$ is maximum when $a = 2$, $F_k = 0$, $r = 2$ and $k = 2$. Thus, the maximum value of 2nd term of the summation, α_2 is 8.21168E-4 for $p_0 = 0.05$ and 3.12E-3 for $p_0 = 0.1$.

Multi-allelic and Multi-locus

From expression (8) and (14), we can easily deduce the final expression determining minimum plant sample size for 2k-polyploids, for capturing multi-allelic and multi-locus diversity

$$m > \frac{\log\{1-(1-\alpha)^{1/\lambda}\} - \log(a-1)}{\log(\beta_k)} \tag{15}$$

Here, it can be seen that when $k = 1$, i.e., when the population is diploid the expression (15) reduces to (8). Expression (15) can be further re-written to isolate the effect of polysomic inheritance on the sample size as

$$m > \frac{A}{B+C} \tag{16}$$

where

$A = \log\ \{(1 - (1 - \alpha)^{1/\lambda}\} - \log\ (a - 1)$

$B = \log\ (1 - p_0)$

$C = \log\ \{(1 - p_0)^{2k-1}(1 - F_k) + F_k\}$

When the individuals mate at random i.e., $s = 0$, (16) reduces to

$$m > \frac{A}{2B} \tag{16a}$$

and when there is a complete selfing i.e., $s = 1$, we get

$$m > \frac{A}{B} \tag{16b}$$

4. Discussion

Our expression (16), $m > A/(B+C)$, determines minimal sample size, whereas, Crossa *et al.* [7] obtained $m > A/B$, which is possible when $C = 0$ or the population in consideration is completely inbred ($s = 1$). The term C involves the polyploidy parameter (k) and the corresponding inbreeding coefficient (F_k); hence, it accounts for a reduction in sample size owing to deviations from selfing and diploidy. For a given ploidy level, reduction in sample size continues until the state of complete random mating, where sample size reaches a minimum value of A/2kB. Thus, for any given population, minimum sample size lies between A/2kB and A/B. The upper bound (A/B) is attained under the condition of no random mating (s = 1) and is unaffected by the ploidy level. As we deviate from complete inbreeding, the role of ploidy in reducing minimal sample size increases until the minimum value (A/2kB) is reached. This occurs under the condition of no selfing (s = 0). Thus, the sample sizes under this condition for diploid, tetraploid, hexaploid and octaploid populations are almost $m_u / 2$, $m_u / 4$, $m_u / 6$, and $m_u / 8$, respectively. Here m_u is the upper þound of the sample size i.e. A/B.

We made an attempt to determine a theoretical minimum number of vegetative samples for capturing all the alleles from a population with a given probability of conservation. We developed a general model by considering a 2k-auto polyploid population under a broad range of mating systems by extending statistical treatment to the models of Lawrence *et al.* [8]. The required minimum sample size under our model is A/(B + C) which lies between the bounds A/2kB and A/B, attained under the extreme conditions of no selfing and no random mating. Crossa *et al.* [7] reported a similar conclusion, but for a diploid model. They indicated that if there are no associations between genes within individuals at any loci, then the required sample size is exactly half the sample size of that under perfect association. If the degree of association is unknown, then the required sample size is between m/2 and m. Our general model yields the same results as given by the said authors when k = 1 and s = 0 or 1. Minimum sample sizes for given probabilities of conservation, rare allele frequencies, and numbers of alleles and loci, under our set of assumptions, have a lower bound, A/B for all inbred populations irrespective of ploidy level. Sample sizes are smallest under random mating equilibrium. Sample sizes in this state reduce further with increasing ploidy levels. The behavior of our model confirms the conservative characteristics of genetic variability related to polysomic inheritance.

Notably, our treatment here only provides a model for the required minimum sample size for collecting the plant materials. With most species, however, it will usually be possible and, indeed, considerably more practicable, to collect seed from the individuals of a population. When seed is collected, sampling is done from the next generation, because the plants raised from this seed are the offspring of the plants from which the collections have been made. When the individuals of a population always set their seed by self-fertilization, all of the progeny raised from seed taken from a single individual are expected to be, mutation apart, identically the same as their parent. Hence, when s = 1, nothing can be gained by taking more than one seed per plant. When individuals mate at random, the offspring raised from seed taken from a single individual are no longer expected to be genetically identical either among themselves or to their maternal parent. Potentially, questions of how many seeds per plant should be sampled or whether we can achieve greater efficiency by collecting more seeds from a smaller number of plants have been investigated by Yonezawa and Ichihashi [16] using probability models, Lawrence *et al.* [9] based on the analytical procedures of quantitative genetics and Sapra *et al.* [14].

ACKNOWLEDGEMENTS

The author wishes to thank with regards to Dr. Prem Narain for his valuable guidance and whole hearted support in carrying out this research activity in the area of plant germplasm sampling.

REFERENCES

[1] Allard, R.W. (1970). *Population Structure and Sampling Methods. Genetic Resources in Plants - Their Exploration and Conservation.* Blackwell publications, Oxford and Edinburgh, UK, 97-107.

[2] Bennett, E. (1970). *Adaptation in Wild and Cultivated Plant Populations. Genetic Resources in Plants - Their Exploration and Conservation.* Blackwell publications, Oxford and Edinburgh, UK, 115-129.

[3] Bogyo, T. P., Poreceddu, E. and Perrino, P. (1980). Analysis of sampling strategies for collecting genetic material. *Econ. Bot.*, **34**, 160-174.

[4] Brown, A. H. D. and Marshall, D. R. (1995). A basic sampling strategy: Theory and practice. *Collecting Plant Genetic Diversity*, CAB International, UK, 75-91.

[5] Chapman, C. G. D. (1984). *On the Size of Genebank and the Genetic Variation it Contains. Crop Genetic Resources: Conservation and Evaluation.* Allen and Unwin publications, UK, 102-108.

[6] Crossa, J. (1989). Methodologies for estimating the sample size required for conservation of outbreeding crops. *Theor. Appl. Genet.*, **77**, 153-161.

[7] Crossa, J., Harnandez, C. M., Bretting, P., Eberhart, S. A. and Taba, S. (1993). Statistical genetic considerations for maintaining germplasm collections. *Theor. Appl. Genet.*, **86**, 673-678.

[8] Lawrence, M. J., Marshall, D. F. and Davies, P. (1995). Genetics of genetic conservation .I. Sample size when collecting germplasm. *Euphytica*, **84**, 89-99.

[9] Lawrence, M. J., Marshall, D. F. and Davies, P. (1995). Genetics of genetic conservation. II. Sample size when collecting seed of cross pollinating species and the information that can be obtained from the evaluation of material held in gene banks. *Euphytica*, **84**, 101-107.

[10] Marshall, D.R. and Brown, A. H. D. (1975). *Optimum Sampling Strategies in Genetic Conservation. Crop Genetic Resources Today and Tomorrow.* Cambridge University Press, UK, 53-80.

[11] McConnell, Gillian and Fyfe, J. L. (1975). Mixed selfing and random mating with polysomic inheritance. *Heredity,* **34**, 271-272.

[12] Qualset, C. O. (1975). *Sampling Germplasm in a Center of Diversity: An Example of Disease Resistance. Crop Genetic Resources Today and Tomorrow.* Cambridge University Press, UK, 81-96.

[13] Sapra, R. L., Narain, Prem and Chauhan, S.V.S. (1998). A general model for sample size determination for collecting germplasm. *J. Biosci.,* **23**, 647-652.

[14] Sapra, R. L., Narain, Prem, Chauhan, S.V.S., Lal, S.K. and Singh, B.B. (2003). Sample size for collecting germplasms - A polyploid model with mixed mating system. *J. Biosci.,* **28**, 155-161.

[15] Yonezawa, K. (1985). A definition of the optimal allocation of effort in conservation of plant genetic resources with application to sample size determination for field collection. *Euphytica,* **34**, 345-354.

[16] Yonezawa, K. and Ichihashi, H. (1989). Sample size for collecting germplasm from natural plant population in view of genotypic multiplicity of seed embryos borne on a single plant. *Euphytica,* **41**, 91-97.