ISSN 0258 - 7122
Bangladesh J. Agril. Res. 32(3) : 403-411, September 2007

ESTIMATION OF OPTIMUM SAMPLE SIZE AND NUMBER OF REPLICATIONS IN SPLIT-SPLIT PLOT DESIGN

MD. SAIFUL ISLAM[1]

Abstract

In field experiments, it is necessary to determine the optimum sample size as well as optimum number of replications if researchers have to use sampling techniques for collecting data from such experiments. Estimates of such optimum sample size and number of replications has been determined for split-split plot design minimizing the variance for a given cost of the experiment per treatment.

Key Words: Estimation, sample size, replications, split-split plot design

Introduction

Frequently it is not possible to measure yield and yield contributing characteristics on the whole of each experimental unit. It may be desirable to estimate the characteristics on a random sample basis. Kempthorne (1952) showed the optimum sample size and number of replications in randomized complete block design (RCBD) for equal sampling from each cell having multiple observations. Islam *et al.* (2000) also showed the optimum sample size and number of replications in RCBD with unequal observation per cell. Federer (1963) and Islam (2001) has cited some example (optimum sample size was found 5 & 9 in completely randomized design & split plot design). Usual four-way classification in Split-Split Plot Design with more than one observation per cell and equal variance are considered. Estimation of variance components are obtained from analysis of variance technique. None has yet used for split-split plot design to find out the optimum sample size and number of replications for measuring yield and yield contributing characters from experimental fields.

The main objective of this study is to determine optimum sample size and number of replications for measuring yield and yield contributing characters based on equal sampling from each sub-sub plot of split-split plot design.

[1]Senior Scientific Officer, Statistics Section, BARI, Joydebpur, Gazipur 1701, Bangladesh.

Method

Sample size depends on the variability associated with variable and the cost of reducing that variability. For such cases, it is necessary to choose optimum sample size and number of replications. Estimation of optimum sample size and number of replications are obtained by maximizing the information for a given cost.

Suppose that we have an experiment in r replications, p main plots, q sub-plots and s sub-sub plots and that the characteristic has been observed on a random sample of n equal sampling units from each plot (sub-sub plot). The observations may be denoted by Y_{ijklm}, where k denotes the replications (k=1,2,............,r); i denotes the number of main plots (i=1,2,..........,p); j denotes the sub-plots (j=1,2,.........q); l denotes the sub-sub plots (1,2,......,s) and m denotes the sampling unit (m= 1,2,..........,n) and assume the following model:

$$Y_{ijklm} = \mu + \rho_k + \alpha_i + \gamma_{ik} + \beta_j + (\alpha\beta)_{ij} + \gamma_{ijk} + \delta_l + (\alpha\delta)_{il} + (\beta\delta)_{il} + (\alpha\beta\delta)_{ijl} + \varepsilon_{ijkl} + \eta_{ijklm}$$

where μ = the general mean

ρ_k = the effect of kth replication

α_l = the main plot effect due to ith level of A

γ_{ik} = the main plot error

β_j = the sub plot effect due to jth level of B

$(\alpha\beta)_{ij}$ = the interaction effect due to ith level of A and jth level of B

γ_{ijk} = the sub plot error

δ_l = the sub-sub plot effect due to lth level of C

$(\alpha\delta)_{il}$ = the interaction effect due to ith level of A and lth level of C

$(\beta\delta)_{jl}$ = the interaction effect due to jth level of B and lth level of C

$(\alpha\beta\delta)_{ijl}$ = the interaction effect due to ith level of A, jth level of B and lth level of C

ε_{ijkl} = the experimental (sub-sub plot) error

η_{ijklm} = the sampling error of the mth observation.

We suppose for the present purpose that the η_{ijklm} 'S are normally and independently distributed with variance σ_n^2, the ε_{ijkl}' S are normally and independently distributed with variance σ_e^2, γ_{ik}' S are normally and independently distributed with variance σ_α^2 and γ_{ijk}' S are also normally and independently distributed with variance σ_b^2. The errors η_{ijklm}, ε_{ijkl}, γ_{ik} and γ_{ijk} are independent since the sampling is done at random.

The least square estimates of parameters and errors are obtained as follows:

$$\hat{\mu} = \bar{y}.....$$

$$\hat{\alpha}_i = \bar{y}_i.... - \bar{y}......$$

$$\hat{\beta}_j = \bar{y}._j... - \bar{y}......$$

$$\hat{\rho}_k = \bar{y}.._k.. - \bar{y}.......$$

$$\left(\hat{\alpha}\hat{\beta}\right)_{ij} = \bar{y}_{ij}... - \bar{y}_i.... - \bar{y}._j... + \bar{y}......$$

$$\hat{\gamma}_{ik} = \bar{y}_{i\cdot k}.. - \bar{y}_i.... - \bar{y}.._k.. + \bar{y}......$$

$$\hat{\gamma}_{ijk} = \bar{y}_{ijk}.. - \bar{y}_{ij}... - \bar{y}_{i\cdot k}.. + \bar{y}_i....$$

$$\hat{\delta}_l = \bar{y}..._l. - \bar{y}.......$$

$$\left(\hat{\alpha}\hat{\delta}\right)_{il} = \bar{y}_{i\cdot\cdot l}. - \bar{y}_i.... - y...._{l.} + y......$$

$$\left(\hat{\beta}\hat{\delta}\right)_{jl} = \bar{y}._j._l. - \bar{y}._j... - y..._l. + y......$$

$$\left(\hat{\alpha}\hat{\beta}\hat{\delta}\right)_{ijl} = \bar{y}_{ij\cdot l}. - \bar{y}_{ij}... - \bar{y}_{i\cdot\cdot l}. - \bar{y}._j._l. + \bar{y}_i.... + \bar{y}._j... + \bar{y}..._l. - \bar{y}......$$

$$\hat{\varepsilon}_{ijkl} = \bar{y}_{ijkl}. - \bar{y}_{ijk}.. - \bar{y}_{ij\cdot l}. + \bar{y}_{ij}...$$

$$\hat{\eta}_{ijklm} = y_{ijklm} - \bar{y}_{ijkl}.$$

Total SS can be partitioned into component SS as follows:

$$\text{Total SS} = \sum_i \sum_j \sum_k \sum_l \sum_m \left(y_{ijklm} - \bar{y}......\right)^2 = qrsn \sum_i \left(\bar{y}_i.... - \bar{y}.....\right)^2 +$$

$$pqsn \sum_i \left(\bar{y}.._k.. - \bar{y}......\right)^2 + qsn \sum_i \sum_k \left(\bar{y}_{i\cdot k}.. - \bar{y}_i.... - \bar{y}.._k.. + \bar{y}......\right)^2 + prsn$$

$$\sum_j \left(\bar{y}._j.. - \bar{y}......\right)^2 +$$

$$rsn \sum_i \sum_j \left(\bar{y}_{ij}... - \bar{y}_i.... - \bar{y}._j... + \bar{y}.....\right)^2 + sn \sum_i \sum_j \sum_k \left(\bar{y}_{ijk}.. - \bar{y}_{ij}.. - \bar{y}_{i\cdot k}.. + \bar{y}_i....\right)^2 +$$

$$pqrn \sum_l \left(\bar{y}..._l. - \bar{y}......\right)^2 + sn \sum_i \sum_j \sum_k \left(\bar{y}_{ijk}.. - \bar{y}_{ij}... - \bar{y}_{i\cdot k}.. + \bar{y}_i....\right)^2 +$$

$$prn\sum_{j}\sum_{l}\left(\bar{y}_{\cdot j\cdot l\cdot}-\bar{y}_{\cdot j\ldots}-\bar{y}_{\ldots l\cdot}+\bar{y}_{\ldots\ldots}\right)^{2}+rn\sum_{i}\sum_{j}$$

$$\sum_{l}\left(\bar{y}_{ij\cdot l\cdot}-\bar{y}_{ij\ldots}-\bar{y}_{i\cdot\cdot l\cdot}+\bar{y}_{i\ldots\cdot}+\bar{y}_{\cdot j\ldots}+\bar{y}_{\ldots l\cdot}-\bar{y}_{\ldots\cdot}\right)^{2}+$$

$$n\sum_{i}\sum_{j}\sum_{k}\sum_{l}\left(\bar{y}_{ijkl\cdot}-\bar{y}_{ijk\cdot\cdot}-\bar{y}_{ij\cdot l\cdot}+\bar{y}_{ij\ldots}\right)^{2}+\sum_{i}\sum_{j}\sum_{k}\sum_{l}\sum_{m}\left(y_{ijklm}-\bar{y}_{ijkl\cdot}\right)^{2}$$

= Main plot SS + Sub plot SS + Error$_1$ SS + Replication SS + Interaction (MP∞SP) SS + Error$_2$ SS + Sub-Sub plot SS + Interaction (MP × SSP) SS + Interaction (SP × SSP) SS + Interaction (MP × SP×SSP) SS + Experimental error SS + Sampling error SS

= SS (A) + SS (B) + SS (Error)$_1$ + SS (Rep) + SS(A×B) + SS(Error)$_2$ + SS(C) + SS(A×C) + SS (B×C) + SS(A×B×C) + SS (Experimental error) + SS (Sampling error).

The degrees of freedom for different sum of squares are:

Sum of Squares	degrees of freedom
Total	(pqrsn-1)
Replication	(r-l)
Main plot (A)	(p-1)
Error-1	(r-l) (p-l)
Sub plot (B)	(q-l)
Interaction (A×B)	(p-l) (q-l)
Error-2	p(q-l)(r-l)
Sub-sub plot (C)	(s-l)
Interaction (A×C)	(p-l)(s-l)
Interaction (B×C)	(q-l)(s-l)
Interaction (A×B×C)	(p-l)(q-l)(s-l)
Experimental error (Error-3)	pq(s-l)r-l)
Sampling error	pqrs(n-l)

The analysis of variance of table for the analysis on a sample basis is given below:

Table 1.

Source of variation	d. f	Sum of squares	Mean square	Expectation of mean square*
Replication	(r-l)	$pqsn\sum_{k}\left(\bar{y}_{\cdot\cdot k\cdot\cdot}-\bar{y}_{\cdot\cdot\cdot\cdot\cdot}\right)^2$	R	
Main plot (A)	(p-l)	$qrsn\sum_{i}\left(\bar{y}_{i\cdot\cdot\cdot\cdot}-\bar{y}_{\cdot\cdot\cdot\cdot\cdot}\right)^2$	MP	
Error-1	(r-l) (p-l)	$qsn\sum_{i}\sum_{k}\left(\bar{y}_{i\cdot k\cdot\cdot}-\bar{y}_{i\cdot\cdot\cdot}-\bar{y}_{\cdot\cdot k\cdot\cdot}+\bar{y}_{\cdot\cdot\cdot\cdot\cdot}\right)^2$	E_1	$qsn\sigma^2{}_a+sn\sigma^2{}_b+n\sigma^2{}_e+\sigma^2{}_n$
Sub plot (B)	(q-l)	$prsn\sum_{j}\left(\bar{y}_{\cdot j\cdot\cdot\cdot}-\bar{y}_{\cdot\cdot\cdot\cdot\cdot}\right)^2$	SP	
Interaction (A×B)	(p-l)(q-l)	$rsn\sum_{i}\sum_{j}\left(\bar{y}_{ij\cdot\cdot\cdot}-\bar{y}_{i\cdot\cdot\cdot\cdot}-\bar{y}_{\cdot j\cdot\cdot\cdot}+\bar{y}_{\cdot\cdot\cdot\cdot\cdot\cdot}\right)^2$	MP∞SP	
Error-2	p(q-l)(r-l)	$sn\sum_{i}\sum_{j}\sum_{k}\left(\bar{y}_{ijk\cdot\cdot}-\bar{y}_{ij\cdot\cdot\cdot}-\bar{y}_{i\cdot k\cdot\cdot}+\bar{y}_{i\cdot\cdot\cdot\cdot}\right)^2$	E_2	$sn\sigma^2{}_b+n\sigma^2{}_e+\sigma^2{}_n$
Sub-sub plot (C)	(s-l)	$pqrn\sum_{l}\left(\bar{y}_{\cdot\cdot\cdot l\cdot}-\bar{y}_{\cdot\cdot\cdot\cdot\cdot}\right)^2$	SSP	
Interaction (A×C)	(p-l)(s-)	$qrn\sum_{i}\sum_{l}\left(\bar{y}_{i\cdot\cdot l\cdot}-\bar{y}_{i\cdot\cdot\cdot\cdot}-\bar{y}_{\cdot\cdot\cdot l\cdot}+\bar{y}_{\cdot\cdot\cdot\cdot\cdot\cdot}\right)^2$	MP∞SSP	
Interaction (B×C)	(q-l) (s-l)	$prn\sum_{j}\sum_{l}\left(\bar{y}_{\cdot j\cdot l\cdot}-\bar{y}_{\cdot j\cdot\cdot\cdot}-\bar{y}_{\cdot\cdot\cdot l\cdot}+\bar{y}_{\cdot\cdot\cdot\cdot\cdot}\right)^2$	SP∞SSP	
Interaction (A×B×C)	(p-l) (q-l) (s-l)	$rn\sum_{i}\sum_{j}\sum_{l}(\bar{y}_{ij\cdot l\cdot}-\bar{y}_{ij\cdot\cdot\cdot}-\bar{y}_{i\cdot\cdot l}-\bar{y}_{\cdot j\cdot l\cdot}+$ $\bar{y}_{i\cdot\cdot\cdot\cdot}+\bar{y}_{\cdot j\cdot\cdot\cdot}+\bar{y}_{\cdot\cdot\cdot l\cdot}-\bar{y}_{\cdot\cdot\cdot\cdot\cdot})^2$	MP x SP x SSP	
Experimental error (Error-3)	pq(s-l) (r-l)	$n\sum_{i}\sum_{j}\sum_{k}\sum_{l}\left(\bar{y}_{ijkl\cdot}-\bar{y}_{ijk\cdot\cdot}-\bar{y}_{ij\cdot l\cdot}+\bar{y}_{ij\cdot\cdot\cdot}\right)^2$	E_3	$n\sigma^2{}_e+\sigma^2{}_n$
Sampling error	pqrs(n-l)	$\sum_{i}\sum_{j}\sum_{k}\sum_{l}\sum_{m}\left(y_{ijklm}-\bar{y}_{ijkl\cdot}\right)^2$	S	$\sigma^2{}_n$
Total	(pqrsn-l)	$\sum_{i}\sum_{j}\sum_{k}\sum_{l}\sum_{m}\left(y_{ijklm}-\bar{y}_{\cdot\cdot\cdot\cdot\cdot}\right)^2$	-	-

* The expression are given only for error mean square, sources Kempthorne (1952, pp213) and Islam *et al.*, (2000, pp 91)

R, MP, E_1 SP, MPxSP, E_2, SSP, MPxSSP, SPxSSP, MPxSPxSSP, E_3, & S are the mean squares of replication, main plot, error_1, sub plot, MPxSP interaction, error_2, sub-sub plot, MPxSSP interaction, SPxSSP interaction, MPxSPxSSP interaction, experimental error and sampling error, respectively.

Estimation

The component of variances : $\sigma^2{}_n, \sigma^2{}_e, \sigma^2{}_b$ *and* $\sigma^2{}_a$ are estimated as

$$\hat{\sigma}^2{}_n = S, E_3 = n\hat{\sigma}^2{}_e + \hat{\sigma}^2{}_e, E_2 = sn\hat{\sigma}^2{}_b + n\hat{\sigma}^2{}_e + \hat{\sigma}^2{}_n \text{ and } E_1$$
$$= qsn\hat{\sigma}^2{}_a + sn\hat{\sigma}^2{}_b + n\hat{\sigma}^2{}_e + \hat{\sigma}^2{}_n$$

or, $\hat{\sigma}^2{}_e = \frac{E_3 - S}{n}, \hat{\sigma}^2{}_b = \frac{E_2 - E_3}{ns}, \hat{\sigma}^2{}_a = \frac{E_1 - E_2}{snq}$

Now
$$\hat{\delta}_l = \bar{y}\ldots_l. - \bar{y}.....$$
$$\hat{\delta}_l' = \bar{y}\ldots_l'. - \bar{y}..... \qquad _l \neq l'$$

$$V\left(\hat{\delta}_l - \hat{\delta}_l'\right) = V\left(\bar{y}\ldots._l. - \bar{y}\ldots l'.\right) = V\left(\bar{y}\ldots_l.\right) + V\left(y\ldots_l'.\right)$$

$$V(\bar{y}\ldots_l.) = V(\bar{\gamma}..) + V(\bar{\gamma}\ldots) + V(\bar{e}\ldots_l) + V(\bar{\eta}\ldots_l.)$$

(Where $V(\bar{y}\ldots_l.)$= variance of treatment mean comes sub-sub plots (treatments), formula cited from Kempthorne (1952) & Islam *et al.* (2000))

$$= \frac{\sigma_a{}^2}{pr} + \frac{\sigma^2{}_b}{pqr} + \frac{\sigma_e{}^2}{pqr} + \frac{\sigma_n{}^2}{npqr} = \frac{n\left(q\sigma_a{}^2 + \sigma^2{}_b + \sigma_e{}^2\right) + \sigma_n{}^2}{npqr}$$

$$V\left(\hat{\delta}_l - \hat{\delta}l'\right) = 2\frac{n\left(q\sigma_a{}^2 + \sigma^2{}_b + \sigma_e{}^2\right) + \sigma_n{}^2}{npqr}$$

The estimated variacne of difference between two treatment (sub-sub plot) means on a per sample basis will be proportional to $\frac{n\left(q\hat{\sigma}_a{}^2 + \hat{\sigma}^2{}_b + \hat{\sigma}_e{}^2\right) + \hat{\sigma}_n{}^2}{npqr} = \frac{E_l - E_3 + sE_3}{npqrs}$ Similarly, the estimated variance of treatment comparison with r' replications, p' main plot, q' sub plot and n' samples per plot will be proportional to $\frac{\hat{\sigma}^2{}_n + n'\left(q'\hat{\sigma}^2{}_a + \hat{\sigma}^2{}_b + \hat{\sigma}^2{}_e\right)}{r'p'q'n'}$

$$= \frac{\hat{\sigma}^2{}_n}{r'p'q'n'} + \frac{\hat{\sigma}^2{}_a}{r'p'} + \frac{\hat{\sigma}^2{}_b}{p'q'r'} + \frac{\hat{\sigma}^2{}_e}{r'p'q'}$$

$$= \frac{S}{r'p'q'n'} + \frac{E_1 - E_2}{nsqr'p'} + \frac{E_2 - E_3}{snp'q'r'} + \frac{E_3 - S}{nr'p'q'} = \frac{E_1 - E_2}{snqp'r'} + \frac{E_2 - E_3}{snp'q'r'} + \frac{E_2}{np'q'r'} + \frac{S}{r'p'q'}\left[\frac{1}{n'} - \frac{1}{n}\right]$$

The relative information (RI) for r' replication, p' main plot, q' sub plot and n' samples per plot over r replication, p main plot, q sub plot and n samples per plot may then be estimated

$$RI = \frac{1}{\left[\frac{q'n'(E_1 - E_2) + qn'(E_2 - E_3) + sqn'E_3 + sqS(n - n')}{sqnr'p'q'n'}\right]} \Big/ \frac{1}{\left[\frac{E_1 - E_3 + sE_3}{spqrn}\right]}$$

$$= \frac{n'p'q'r'(E_1 - E_3 + sE_s)}{pr\left[n'q'(E_1 - E_2) + qn'(E_2 - E_3) + n'qsE_3 + (n - n')Sqs\right]}.$$

$$= \frac{\{n'p'q'r'(E_1 - E_3 + sE_3)\}/n'}{\left[pr\{n'q'(E_1 - E_2) + qn'(E_2 - E_3) + n'sqE_3 + (n - n')Sqs\}\right]/n'}$$

As $n' \to \propto$, RI = $\dfrac{p'q'r'(E_1 - E_3 + sE_3)}{pr\{q'(E_1 - E_2) + q(E_2 - E_3) + sq(E_3 - S)\}}$

and this is the maximum that can be achieved by increasing the number of sub sampling units. It will often be uneconomical to increase the number of sub sampling units i.e. n and it is necessary to consider the cost of the various operations to determine optimum n.

For the cost function, let us consider

C = pre-harvesting cost per plot

C_n = harvesting and post-harvesting cost per sample

With r replications p main plots, q sub plots and n sample size per plot, the cost of the experiment per treatment (sub-sub plot) is ($rpqC+rpqnC_n$)=$rpq(C+nC_n)$

The information on each treatment mean, assuming the variance components to be known is

$$\frac{pqrn}{\sigma^2{}_n + n\left(q\sigma^2{}_a + \sigma^2{}_e + \sigma^2{}_b\right)} \qquad (1)$$

We shall choose r, p, q and n to maximize the information in (1) for a given cost per treatment $C_0=rpq(C+nC_n)$ (2)

From (2), $r = \dfrac{C_0}{pq(C + nC_n)}$

Putting this value in equation (1), we get the information expressed in terms of costs as

$$\frac{nC_0}{\left(C+nC_n\right)\left\{\sigma^2{}_n+n\left(q\sigma^2{}_a+\sigma^2{}_b+\sigma^2{}_e\right)\right\}} \tag{3}$$

Then maximization of (1) is equivalent to maximization of (3)

The reciprocal of (3) is given by

$$\frac{1}{C_0}\left\{\frac{C\sigma^2 n}{n}+C_n\sigma^2{}_n+C\left(q\sigma^2{}_a+\sigma^2{}_b+\sigma^2{}_e\right)+nC_n\left(q\sigma^2{}_a+\sigma^2{}_b+\sigma^2{}_e\right)\right\} \tag{4}$$

Then the maximization of (3) is equivalent to the minimization (4)

Differentiating (4) with respect to n and equating to zero, we get

$$-\frac{C\sigma^2{}_n}{n^2}+C_n\left(q\sigma^2{}_a+\sigma^2{}_b+\sigma^2{}_e\right)=0,$$

or $$n^2=\frac{C\sigma_n^2}{C_n\left(q\sigma_a^2+\sigma_b^2+\sigma_e^2\right)},$$

which yields the optimum sample size as $n_{(opt)}=\sqrt{\dfrac{C\sigma^2{}_n}{C_n\left(q\sigma^2{}_a+\sigma^2{}_b+\sigma^2{}_e\right)}}$

The optimum number of replication is obtained substituting the optimum value $n_{(opt)}$ of n in equation (2) $r_{(opt)}=\dfrac{C_0}{pq\left[C+\sqrt{\dfrac{CC_n\sigma^2{}_n}{\left(q\sigma^2{}_a+\sigma^2{}_b+\sigma^2{}_e\right)}}\right]}$

Concluding remarks

The optimum sample size as well as optimum number of replications depend on unknown error variances $\sigma_n^2, \sigma_e^2, \sigma_b^2 \textit{ and } \sigma_a^2$. Unbiased estimators of the error variances are given from Table 1. Using such unbiased estimators of $\sigma_n^2, \sigma_e^2, \sigma_b^2 \textit{ and } \sigma_a^2$, estimates optimum sample size $(\hat{n})$and optimum number of replications $(\hat{r})$ can be obtained. The costs C and C_n can be obtained from similar experiments already conducted in real fields.

References

Kempthorne, O. 1952. Design and analysis of experiment, John Wiley and Sons, Inc, New York, USA.

Islam M. S., K. Sen and K. Rahim. 2000. Estimation of optimum sample size and number of replications in RCBD with unequal observation per cell. *Dhaka University Journal of Science* **48**(1): 89-94.

Islam, M. S. 2001. Statistical development of field plot technique for agronomic experiments. Unpublished Ph. D. thesis, Dhaka University, Dhaka, Bangladesh.

Federer W. T. 1963. Experimental design. Theory and application, Oxford and IBH Publishing Co. New Delhi, India.