ANALYSIS OF VARIANCE CALCULATIONS FOR IRREGULAR EXPERIMENTS[1]

Jonathan W. Wright[2]

ABSTRACT.--Irregular experiments may be more useful than much smaller regular experiments and can be analyzed statistically without undue expenditure of time. For a few missing plots, standard methods of calculating missing-plot values can be used. For more missing plots (up to 10 percent), seedlot means or randomly chosen plot means of the same seedlot can be substituted for missing plots, provided between-block differences are small. Whatever the number of missing plots (provided there is more than one plot per seedlot) or the size of the between-block differences, seedlot means and sums of squares can be estimated in terms of deviations from block means. The procedures for calculation of analysis of variance in terms of deviations are described. The procedures are also applicable to regular experiments.

Ideally, a forest genetic experiment should follow a regular design in which each seedlot is represented in an equal number of blocks in every plantation. However, perfect regularity is rarely possible. There are usually a few missing plots as the result of mortality. Also, many times planting stock is more limited for some seedlots than for others. When that happens, one can plan a regular experiment by reducing the numbers of families and replications, or one can go ahead and distribute each family to each replication as long as the planting stock lasts. If the first alternative had been followed, most NC-99[3] experiments would have been reduced in size by 90 percent. Most NC-99 members probably agree that the second alternative was more desirable even though it resulted in experiments with great irregularities.

---

[1] The work done here was supported in part by regional research funds from the U.S.D.A. under regional project NC-99 entitled "Improvement of Forest Trees Through Selection and Breeding."

[2] Professor of Forestry, Michigan State University, East Lansing, Michigan 48823.

[3] NC-99 is an organization of University and federal forest tree geneticists. The group has conducted cooperative provenance and progeny tests in northcentral United States for almost 20 years.

This paper has two objectives. The first is to present relatively simple methods to analyze data from experiments with various degrees of irregularity. The second is to allay frequently encountered fears of irregular data when planning experiments that might lose much information if made perfectly regular.

In using these methods, one must remember that the resulting analyses are only as strong as the data base. One can devise calculation methods to allow for missing data, but the inferences can never be as strong as if the data were not missing.

STANDARD PROCEDURE FOR CALCULATING MISSING PLOT VALUES

With a randomized complete block experiment in which seedlot A is missing from block 1, the standard procedure for calculating a substitute value is as follows (from Sokol and Rohlf 1969):

(1) Calculate all seedlot sums, block sums, and the overall sum, omitting the missing plot in each case.

(2) Calculate a missing plot mean as follows:

$$\text{Missing plot mean} = \frac{Nb(\text{Sum seedlot A}) + Ns(\text{Sum block 1}) - \text{Overall sum}}{(Nb-1)(Ns-1)}$$

where Nb = number of blocks and Ns = number of seedlots.

(3) Using this substitute value, calculate the analysis of variance in the normal manner.

(4) Reduce the degrees of freedom for error by 1 for each missing plot.

If there is more than one missing plot, the first value must be calculated by using reasonable estimates for the other values, using the first value and reasonable estimates for all other values to calculate the second, etc. When there are many missing values, the preliminary estimates and calculated values may differ appreciably, in which case it may be necessary to redo some of the calculations.

This is the best way to calculate a missing-plot value, but is laborious if the number of missing plots is large.

USE OF SEEDLOT MEANS OR RANDOMLY CHOSEN PLOT MEANS

If the number of missing plots is small and there are minor differences among blocks, either of two simple procedures can be used. For each missing plot, one can substitute either (1) the mean for that seedlot, or (2) the mean of a randomly selected plot of the seedlot. In either case, reduce the degrees of freedom by 1 for each missing plot.

The first method changes seedlot means the least but the second is easiest to use with a computer.

The validity of these procedures can be checked by analyzing data from a completely regular experiment, eliminating some plots arbitrarily, substituting values for them, and re-calculating the analysis of variance. This has been done with several sets of data. Generally, there has been less than a 1 percent change in any mean square or F value as the result of substituting for up to 10 percent of the plot means.

Use of one of these simple substitution methods will normally cause less than a 2 percent change in the mean for an individual seedlot if only one or two plots per seedlot are missing and block means differ by less than 20 percent. Another calculation method may be preferable if there are more missing plots or large differences among blocks.

DISCARDING UNDER-REPRESENTED SEEDLOTS OR BLOCKS

If the missing plots are concentrated in a few seedlots or blocks, those seedlots or blocks can be eliminated to make the remaining data perfectly regular. This procedure is often satisfactory unless the incompletely represented seedlots are among the best.

ANALYSIS BASED UPON DEVIATIONS RATHER THAN TOTALS

Most students learn to do an analysis of variance by squaring the plot means, the seedlot sums, the block sums,etc. The analysis can be done, however, by working with the deviations of the plot means from the block means or seedlot means. By working with such deviations, one can avoid most of the problems encountered in irregular experiments.

Use of the deviation method requires transcription of a data array presented in terms of plot means to one presented in terms of deviations from either seedlot means or block means. This constitutes an extra step. However, subsequent computations are greatly simplified if the deviations are rounded off to the nearest whole number. Such rounding off is rarely of consequence. I often use deviations in preference to the standard method even when working with regular data arrays.

Adequacy of Deviations when Calculating Sums of Squares

Assume that one wishes to calculate the sum of squares and standard deviation of the series of 10 numbers included under column I in table 1. This can be done in either of two ways:

(1) Square each number as in column II. Find the total of the squares (280). Calculate the total for column I, square it and divide by 10 to obtain the correction factor ($50^2/10 = 250$). The "sum of squares" (280 - 250 = 30) is the total of the squares minus the correction factor.

(2) Determine the deviation of each number from the mean of 5.0, enter the deviations in column III, square them (column IV) and sum them to obtain a "sum of squares" of 30.

Notice that the two methods give identical results.

Table 1.--Example showing two ways of calculating the sum of squares of a series of 10 numbers

I Actual	II Square of actual number	III Deviation of actual number from mean	IV Square of deviation
2	4	-3	9
3	9	-2	4
4	16	-1	1
4	16	-1	1
5	25	0	0
5	25	0	0
6	36	+1	1
6	36	+1	1
7	49	+2	4
8	64	+3	9
Total 50	280	0	30
Mean 5.0		0.0	
Correction factor	$50^2/10 = 250$		$0^2/10 = 0$
Sum of squares	30		30

Slight Effects of Small Arithmetic Mistakes or Rounding Off Errors

Assume that one makes a mistake of 1 when adding column I in table 1, obtaining a total of 49 instead of 50. The correction factor would then be calculated as $49^2/10 = 2,401/10 = 240.1$. The sum of squares would be calculated as 280 - 240.1 = 39.9 instead of the true value of 30.

Assume also that a person made the same mistake when working with deviations, obtaining a total of +1 for column III. The correction factor would be calculated as $1^2/10 = .1$. The sum of squares would be calculated as 29.9 instead of 30.

Or, assume that the numbers are changed slightly to produce a total of 49 for column I in table 1. Rounding off the mean of 4.9 to 5.0 and the use of whole-number deviations would result in an error of 0.1 in the sum of squares.

Thus, even with the small deviations shown in table 1, rounding off errors are of little consequence. They are even less important with larger deviations. Also, when using deviations, small arithmetic mistakes may have negligible consequences.

Estimating Seedlot Means by Using Deviations from Block Means

Table 2 is an hypothetical data array showing how to calculate seedlot means by using deviations from block means. It was formulated by assuming a regular experiment with progressively greater growth from block 1 to 5 and from seedlot 1 to 9.

Table 2.--Example showing estimates of seedlot means when calculated from actual values and as deviations from block means for an experiment having many missing plots. (The example was formulated in such a way as to have theoretical means of 0, 7, 8....14 for seedlots 1 to 9, respectively and of 6, 8, 10, 12 and 14 for blocks 1 to 5, respectively, if there were no missing plots.)

Seedlot no.	Actual values: Height in block 1	2	3	4	5	Sum	Mean	Deviations from block means: Height in block 1	2	3	4	5	Sum	Mean	Estimated means: Theory	Act. val.	Deviations
1	2	4	6	--	--	12	4	-4	-4	-4	--	--	-12	-4	6	4	6.1
2	--	--	7	9	11	27	9	--	--	-3	-3	-3	-9	-3	7	9	7.1
3	4	6	8	10	12	40	8	-2	-2	-2	-2	-2	-10	-2	8	8	8.1
4	--	--	9	11	--	20	10	--	--	-1	-1	--	-2	-1	9	10	9.1
5	6	8	10	12	14	50	10	0	0	0	0	0	0	0	10	10	10.1
6	--	--	11	13	--	24	12	--	--	+1	+1	--	+2	+1	11	12	11.1
7	8	10	12	14	16	60	12	+2	+2	+2	+2	+2	+10	+2	12	12	12.1
8	--	--	13	15	17	45	15	--	--	+3	+3	+3	+9	+3	13	15	13.1
9	10	12	14	--	--	36	12	+4	+4	+4	--	--	+12	+4	14	12	14.1
n	5	5	9	7	5	31		5	5	9	7	5	31				
Sum	30	40	90	84	70	314		0	0	0	0	0	0				
Mean	6	8	10	12	14		10.1	0	0	0	0	0		0	10.0	10.1	10.1

The procedure is as follows:

(1) Calculate block totals and means as under "actual values".

(2) Calculate "deviations from block means" as in the central portion of the table.

(3) For each seedlot, add the average "deviation from block mean" to the plantation mean to obtain an estimated true mean (right-hand column).

While there is greater agreement between theory and practice in this hypothetical example than in actual practice, use of the deviations-from-block-means method always gives better estimates of seedlot means than can be obtained by averaging the actual plot means. The larger the number of missing plots and the greater the between-block differences, the greater the advantage of the deviation method.

Analysis of Variance Calculations Using Deviations from Block Means

To compute an analysis of variance, proceed as outlined below and in table 3.

Table 3.--Methods used to calculate means and analysis of variance for an irregular experiment, using deviations from block means (Sums of Squares and Degrees of Freedom are symbolized by SSQ and DF, respectively.)

Seedlot no.	True means for block 1	2	3	4	5	Sum	Ave.	Deviations from block means 1	2	3	4	5	Sum	Freq.	Ave.	Probable true mean
1	A	A	A	-	A			G	G	G	-	G	J	n_s	K	K+E = L
2	A	-	A	A	-			G	-	G	G	-	J	n_s	K	L
3	-	A	A	A	-			-	G	G	G	-	J	n_s	K	L
4	A	A	-	A	A			G	G	-	G	G	J	n_s	K	L
5	A	-	-	A	A			G	-	-	G	G	J	n_s	K	L
6	-	A	A	-	A			-	G	G	-	G	J	n_s	K	L
S	B	B	B	B	B	D		H	H	H	H	H	M		Zero	D
Freq.	n_b	n_b	n_b	n_b	n_b	n_t								n_t		
Mean	C	C	C	C	C		E								Zero	E

G = A - C

$SSQ_{seedlot} = \Sigma(J^2/n_s) - M^2/n_t = \Sigma(J^2/n_s)$, approximately

$SSQ_{block} = \Sigma(B^2/n_b) - D^2/n_t$, approximately

$SSQ_{error} = \Sigma G^2 - \Sigma(n^2/n_b) = \Sigma G^2$, approximately

$DF_{seedlot}$ = number of seedlots - 1

DF_{block} = number of blocks - 1

DF_{total} = total number of plots - 1

Arithmetic checks: each H and M should be less than $\pm n_b/2$ and $n_t/2$, respectively.

(1) For each living plot, enter a true mean (A) as in the left-hand portion of table 3. Calculate the sum, number of plots, and mean for each block (B, n_b and C, respectively) and for the entire plantation (D, n_t and E, respectively).

(2) Prepare a table similar to the right-hand portion of table 3. For each living plot enter a deviation-from-block-mean (G = A - C). Calculate the sum of deviations, number of plots and average deviation for each seedlot (J, n_s and K, respectively) and for the entire plantation (M, n_t and close to zero, respectively). Also, calculate the sum of the deviations' (H) for each block.

(3) Check the arithmetic. If there are no mistakes, $\Sigma H = \Sigma J = M$ = less than $n_t/2$; $\Sigma n_b = \Sigma n_s = n_t$; and ΣK = nearly zero. Also, each H and M should be less than $\pm n_b/2$ and $n_t/2$, respectively.

(4) For each seedlot calculate a probable true mean (L = K + E).

(5) Calculate the sums of squares (SSQ) and degrees of freedom (DF) as indicated at the bottom of table 3.

(6) Calculate mean squares (MSQ = SSQ/DF) and F values in the normal manner.

Analysis of Variance Calculations Using Deviations from Seedlot Means

If differences among blocks are small, the analysis of variance can be calculated in terms of deviations from seedlot means, using the methods outlined below:

(1) For each living plot, enter a true mean (A) as in the left-hand portion of table 4. Calculate the sum, number of plots (and mean) for each seedlot (F, n_s and L, respectively) and for the entire plantation (D, n_t and E, respectively).

(2) Prepare a table similar to the right-hand portion of table 4. For each living plot, enter a deviation-from-seedlot-mean. As an arithmetic check, calculate the sum (Q = less than $n_s/2$) for each seedlot.

(3) Calculate the sum of the deviations (P) for each block.

(4) Calculate the sums of squares (SSQ) and degrees of freedom (DF) as indicated at the bottom of table 4.

(5) Calculate mean squares and F values in the normal manner.

When differences among blocks are so small as not to be obvious in the field, they can be ignored in the analysis of variance calculations, often saving one-third in computation time. In this case, the signs of

the deviations can be ignored and the P's need not be calculated. When the F ratio for block is less than 3.0, separation of the block from the error sum of squares usually has a negligible effect on the error term.

Table 4.--Methods used to calculate means and analysis of variance for an irregular experiment, using deviations from seedlot means. (Sums of Squares and Degrees of Freedom are symbolized by SSQ and DF, respectively. Except where otherwise noted, symbolism is the same as for table 3.)

Seedlot no.	True means for block								Deviations from seedlot means						Deviation of true sum from expected sum
	1	2	3	4	5	Sum	Freq.	Ave.	1	2	3	4	5	Sum	
1	A	A	A	-	A	F	n_s	L	N	N	N	-	N	Q	$J = F - n_sE$
2	A	-	A	A	-	F	n_s	L	N	-	N	N	-	Q	J
3	-	A	A	A	-	F	n_s	L	-	N	N	N	-	Q	J
4	A	A	-	A	A	F	n_s	L	N	N	-	N	N	Q	J
5	A	-	-	A	A	F	n_s	L	N	-	-	N	N	Q	J
6	-	A	A	-	A	F	n_s	L	-	N	N	-	N	Q	J
Sum						D			P	P	P	P	P	M	
Freq.							n_t		n_b	n_b	n_b	n_b	n_b	n_t	
Mean								E							Zero

$N = A - L$

$SSQ_{seedlot} = \Sigma(J^2/n_s)$, approximately

$SSQ_{block} = \Sigma(P^2/n_b)$, approximately

$SSQ_{error + block} = \Sigma N^2$, approximately

$SSQ_{error} = SSQ_{error + block} - SSQ_{block}$

$DF_{seedlot}$, DF_{block}, DF_{total} = number of seedlots - 1, number of blocks - 1, and total number of plots - 1, respectively.

Arithmetic checks: each Q and M should be less than $\pm n_s/2$ and $n_t/2$, respectively.

Analysis of Variance, Using Data from Several Plantations

The procedures when using data from several plantations are described below and in table 5:

(1) Start with single-plantation analyses as outlined in tables 3 or 4.

It is not necessary to use the same calculation method for all plantations.

Table 5.--Methods used to calculate means and analysis of variance for an irregular experiment involving several plantations, using deviations from plantation means as calculated by use of tables 3 or 4 (Symbolism is a continuation of that used in those tables.)

Seedlot no.	Sum of deviations: Plantation 1	2	3	4	Sum	Number of plots: Plantation 1	2	3	4	Sum	Average deviation	Probable true mean
1	J	J	-	J	R	n_s	n_s	-	n_s	n_{sp}	$S = R/n_{sp}$	$U = S + T$
2	J	-	J	J	R	n_s	-	n_s	n_s	n_{sp}	S	U
3	J	-	-	J	R	n_s	-	-	n_s	n_{sp}	S	U
4	J	J	J	-	R	n_s	n_s	n_s	-	n_{sp}	S	U
5	J	J	J	J	R	n_s	n_s	n_s	n_s	n_{sp}	S	U
6	J	J	-	-	R	n_s	n_s	-	-	n_{sp}	S	U
Sum	Nearly Zero					n_t	n_t	n_t	n_t	n_{gt}		
Mean											Zero	T
					ACTUAL SUMS AND MEANS							
Sum	D	D	D	D	V							
Mean	E	E	E	E								$T = V/n_{gt}$

$SSQ_{seedlot} = \Sigma(R^2/n_{sp})$

$SSQ_{seedlot + seedlot\ X\ plantation} = \Sigma(J^2/n_s)$

$SSQ_{seedlot\ X\ plantation}$ is obtained by subtraction.

$SSQ_{plantation} = \Sigma(D^2/n_t) - V^2/n_{gt}$

$SSQ_{block\text{-}within\text{-}plantation}$ and SSQ_{error} are obtained by adding the SSQ_{block} and SSQ_{error}, respectively for the individual plantations.

$DF_{seedlot}$ and $DF_{plantation}$ are the number of seedlots - 1, and the number of plantations - 1, respectively.

$DF_{seedlot + seedlot\ X\ plantation}$ = the sum of the $DF_{seedlot}$ for the individual plantations; obtain $DF_{seedlot\ X\ plantation}$ by subtraction.

DF_{block} and DF_{error} = the sums of these for individual plantations.

If the measurements and analysis have been done by different individuals and in different units, a simple conversion is possible. If, for example, some plantations were measured in inches and others in centimeters, and all are to be converted to centimeters, all means and deviations based on inches should be multiplied by 2.54 and all sums of squares or mean squares based on inches should be multipled by $2.54^2 = 6.45$.

(2) Prepare a table similar to the left-hand portion of table 5 and insert a plantation-sum-of-deviations (J) and the number of plots (n_s) for each seedlot represented in any one plantation. For each plantation, insert the actual sum (D), actual mean (E) and total number of plots (n_t).

(3) For each seedlot, calculate the total sum-of-deviations (R) and average deviation ($S = R/n_{sp}$). For all plantations combined, calculate the actual sum (V), number of plots (n_{gt}) and mean (T).

(4) Calculate the probable true mean ($U = S + T$) for each seedlot.

(5) Calculate the sums of squares as indicated at the bottom of table 5. The sums of squares for error, block-within-plantation and "seedlot + seedlot X plantation" have already been calculated for individual plantations and may be summed to obtain the values for all plantations combined.

(6) Calculate degrees of freedom as indicated.

(7) Compute mean squares and F values in the normal manner.

LEAST SQUARES TECHNIQUE FOR USE WITH COMPUTERS

A least squares technique is available for those with access to a computer. This technique can be used with data such as are considered here. The analysis procedures are such as to stretch the capacity of even the largest computers. The routines are therefore effectively limited to experiments including a few score seedlots and a few plantations.

LIMITATIONS AND ADVANTAGES OF THE DEVIATIONS TECHNIQUES

With perfectly regular experiments, the deviations-from-means techniques yield exactly the same results as would be obtained with normal calculation procedures.

The standard method for calculating missing-plot values is laborious and therefore effectively limited to experiments with less than a few dozen missing plots. The simple substitution methods (seedlot mean or randomly chosen plot mean) are limited to experiments with small between-block differences and less than 10 percent missing plots. There are no such limitations to the use of the deviations-from-means techniques.

They can yield satisfactory results even if 50 percent of the plots are missing and there are large differences among blocks.

Whatever the calculation method, seedlots represented by a single plot in one plantation or in one plantation only in the case of a multi-plantation experiment should be excluded from an analysis of variance. Otherwise, the sum of squares due to seedlot will be inflated.

One situation is not amenable to the techniques described here: that in which certain blocks or plantations are composed primarily of seedlots considerably above or below average in the trait being measured. This might happen if seedlots of supposedly rapid growth are assigned to one series of plantations and seedlots of supposedly slow growth are assigned to other plantations. In such a case, there is a confounding of between-seedlot and between-block or between-plantation differences such that none of them can be estimated accurately.

Although the unequal sample sizes found in an irregular experiment do not pose serious problems in the calculation of analysis of variance or probable true means, they do in estimating the significance of a difference between any two means. Suppose that some seedlots are represented in 10 blocks and others in only 5 blocks. "Least Significant Differences" or multiple-range tests applicable to the former are not applicable to the latter. In such a case, means must be compared individually.

The deviations-from-mean techniques saves the most time in large experiments containing 100+ seedlots and several plantations, none of them complete. In such a case, the only reasonable alternative may be the exclusion of several seedlots or plantations from the analysis.

Even if few plots are missing, working with deviations can save time because of the lesser need for absolute accuracy in the computation work. As explained in connection with table 1, the smallest arithmetic mistake may have large consequences when working with true means and their sums of squares. That being the case, it is necessary to check every arithmetic operation until the slightest discrepancies are eliminated. Frequently, that requires much time. If working with deviations and their sums of squares, many slight arithmetic mistakes can be forgiven, so that much less cross checking is necessary.

LITERATURE CITED

Sokol, Robert R. and F. James Rohlf. 1969. Biometry. Freeman, San Francisco. 776 pp.