ASSESSMENT OF RISK ATTACHED TO RECOMMENDATIONS

Training Working Document No. 2

Prepared by

CIMMYT

TRAINING WORKING DOCUMENT

ASSESSMENT OF RISK ATTACHED TO RECOMMENDATIONS

Training Working Document No. 2

Prepared by
Roger Mead
Consultant
in collaboration
with CIMMYT staff

CIMMYT
Lisboa 27
Apdo. Postal 6-641,
06600 México, D.F., Mexico

PREFACE

This is one of a new series of publications from CIMMYT entitled *Training Working Documents*. The purpose of these publications is to distribute, in a timely fashion, training-related materials developed by CIMMYT staff and colleagues. Some Training Working Documents will present new ideas that have not yet had the benefit of extensive testing in the field while others will present information in a form that the authors have tested and found useful for teaching. Training Working Documents are intended for distribution to participants in courses sponsored by CIMMYT and to other interested scientists, trainers, and students. Users of these documents are encourage to provide feedback as to their usefulness and suggestions on how they might be improved. These documents may then be revised based on suggestions from readers and users and published in a more formal fashion.

CIMMYT is pleased to begin this new series of publications with a set of six documents developed by Professor Roger Mead of the Applied Statistics Department, University of Reading, United Kingdom, in cooperation with CIMMYT staff. The first five documents address various aspects of the use of statistics for on-farm research design and analysis, and the sixth addresses statistical analysis of intercropping experiments. The documents provide on-farm research practitioners with innovative information not yet available elsewhere. Thanks goes out to the following CIMMYT staff for providing valuable input into the development of this series: Mark Bell, Derek Byerlee, Jose Crossa, Gregory Edmeades, Carlos Gonzalez, Renee Lafitte, Robert Tripp, Jonathan Woolley.

Any comments on the content of the documents or suggestions as to how they might be improved should be sent to the following address:

CIMMYT Maize Training Coordinator
Apdo. Postal 6-641
06600 Mexico D.F., Mexico.

Document 2A

THE ASSESSMENT OF THE RISK ATTACHED TO RECOMMENDATIONS

THEORY

1. Preliminary Comments

It is assumed that we would like to recommend a change from treatment 0 to treatment 1. The evidence for this intention is a number of separate pieces of experimental information about the size of the advantage to be derived from this change. If the evidence includes estimates of the size of the advantage from a single set of trials which is accepted as totally representative of the situations in which the recommendation might be applied, then the methodology of the "Representation of Risk" report (Mead, 1990a) is relevant. Normally the requirements for such a set of trials would be that they be representative both of the population of possible locations and of the population of possible years.

The use of average yields for a range of locations, each averaged over a few years, might be regarded as an adequate basis for assessing the risk appropriate to the recommendation, in the sense of representing the variation across locations of average benefit over a number of years. That is, the risk would be for average performance over years for a single location rather than for performance in a single year at a single location.

The situation which we attempt to solve here is the assessment of risk for a single year at a single location when we do not have adequate direct information about the population variation for both years and locations. The information which is required is the two or three variance components for years, for locations and, if necessary, for year x location interaction. We could subsume the interaction component in the year variation if we redefine year variation as year-within-location variation, unless we can believe that interaction variation is negligible (year to year differences are consistent across locations within experimental error).

2. Variances and Variance Components.

Following Byerlee [following Binswanger and Barah, (1980)] define

$\sigma^2(s)$ as the variance between sites/locations (averaging out the differences between years);

$\sigma^2(y)$ as the variance between years (averaging out the differences between locations);

σ^2 as the interaction year x site/location variance, incorporating variation within a site/location.

If we are to think in terms of a combined σ^2 and $\sigma^2(y)$ we can further define
$\sigma^2(yws)$ as the variance between years within site/locations.

The total variance for a prediction for a single site/location for a single (future) year is either

$$\sigma^2 + \sigma^2(s) + \sigma^2(y)$$

or

$$\sigma^2(s) + \sigma^2(yws).$$

We assume that we will always have sufficient information about the variation between sites within (one or more) years. This is equivalent to

$$\sigma^2 + \sigma^2(s).$$

It is clear that we can obtain, directly from data on a set of trials, separable information about all three variance components if, and only if, the set includes at least several site/locations repeated over at least several years. This requirement appears to be rarely, if ever, satisfied and if the assessment of risk is to be accorded a high priority this situation may have to be reconsidered.

If we cannot estimate all three of $\sigma^2(s)$, $\sigma^2(y)$, and σ^2 directly we must find some way of predicting them from each other. We could, for example, demonstrate (or assume) that the primary causes of variation in the three forms were similar and examine the corresponding variance components for the causative variables. Thus, if we could derive a relation between yield difference and, say, rainfall (total for a period, or some combination of rainfall information then we could examine the variation of yield across sites (not necessarily the same as trial locations) and years.

3. A Practical Approach and Examples

Two quite substantial sets of data have been analysed as case histories. In addition a set of simulation data for ten sites for ten years, with rainfall data for the same 100 combinations has been constructed, and subsamples of data from the total data set have been analysed to provide further illustration of the possible procedures. I hope that by using all these examples I can demonstrate an overall philosophy for assessing risk and also assess the potential and implications of the implementation of that philosophy.

We consider four distinct situations:

3.1 Direct Estimation of Variance Components.

There is data on the yield difference between two alternative treatments for all, or most of, the combinations of a considerable number of years and a considerable number of sites. We can estimate the three variance components and their sum and thence calculate risk probabilities for a recommendation to change from one treatment to the other. We shall return to the consideration of how much data is required to obtain adequate information about the variance components. In this situation we do not have, nor would intend to use, information on ancillary variables.

This situation is rare (or possibly non-existent). The analysis of the complete simulated data set (Document 2D) demonstrates the estimation of the variance components.

3.2 Estimate from Total Sample Variance (No Components)

If, in (3.1) the evidence for consistent site or year differences is not compelling then we can treat a set of data for various site/year combinations as a representative sample from the population of possible site/year combinations. We can then estimate the variance of that population from the sample variance (ignoring site and year differences) and hence calculate the risk probabilities attached to the recommendation for change.

This would work well in situations where the domain for experimentation was homogeneous or where the pattern of variation of yield differences across years was completely different for different sites. The data from N x P trials in Poza Rica (Document 2B) appears to represent this situation well, though the possible identification of a consistent difference between site groups would lead to the recommendation being for a subarea of the district in which trials were performed.

This approach, of assuming that the complete sample of trial results is representative of the population for which the recommendation is required, is always a possible option. It is only convincing when the absence of systematic site and year variation (over and above the interaction variation) has been demonstrated. Nevertheless it will quite often offer an estimate of the population variance which is based on a good number of degrees of freedom and could be used in preference to other estimates based on more credible models but inadequate data.

3.3 Predict Year Variation from Regression.

If the data sample contains too few years to be able to deduce any useful estimate of the year variance component there may be information in the data sample about a relationship between yield difference and an explanatory variable. If this relationship is quite strong and there is additional information about the variation over years (and possibly also sites) of the explanatory variable then it may be feasible to construct estimates of the year (and other) variance components.

The procedure would involve extracting whatever information is available from the data about the variance components of field difference. Then the regression of yield difference on the explanatory variable is calculated. If the regression relationship is well estimated then we could assume that the variation of yield differences can be predicted from the variation of the explanatory variable. If the regression equation is

yield difference, $y = A + Bx$

and the variance of the estimated regression coefficient is

$\text{var}(B) = (\text{Residual Mean Square}) / (\text{Corrected SS of } x)$,

then the variance of y across years and sites is

$\text{var}(y) = B^2 (\text{variance of } x) + (\text{Mean of } x)^2 \text{ var}(B)$.

To split the variation of the predicted y values for different site - year combinations into variance components we assume that the proportional division of the variation of y into variance components is the same as the corresponding division of the variation of x. This is a dubious assumption but if the regression relationship is strong then it may be a reasonable approximation.

Rainfall is one potential explanatory variable which might produce a good regression and for which there will often be supplementary information over a range of years and sites. The rainfall measure could be an annual total or a period total or even a set of rainfall measures for several periods. Another possibility as an explanatory variable could be mean farm yield for the standard farm practice.

The data example from Ghana (Document 2C) provides an example where the possibility of using rainfall as an explanatory variable was developed by identifying, for each experimental site, the nearest point from which rainfall information was available. Analysis of the data, however, did not discover a useable regression relationship either with the rainfall measure or with the average yield of the two treatments. The only possibility for this data set was to revert to (3.1) and assume that the set of 27 experimental conditions could provide a valid estimate of the population variance for the recommendation prediction.

Three examples of typical structures of experimental information providing the basis for a recommendation are drawn from the simulated data (Document 2D) and show how the method could be applied and also illustrate the variability of the resulting estimate of the variance components and of the combined variance.

3.4 Using Computer Models

The other possibility is to use a computer modelling system to predict the variability in yield difference. If a suitable computer model is available and there is sufficient data to validate the model for the local Recommendation Domain, then the model predictions can be calculated for a range of site - year combinations for which the necessary input information for the model is available. No example of such a procedure is available.

4.Conclusions

Of the real data examples, I believe the risk probability assessments from the N x P trials at Poza Rica are probably reliable. They are based on a wide range of years and of sites and the lack of evidence for systematic site or year differences (omitting site group 7) is quite compelling. Although the two years of the Ghana data were very different, I would not be confident in using the risk probability assessments because of having only the two years. Nevertheless the risk assessments are better than nothing.

The simulation results make the difficulties in estimating variance components very clear. We can quantify this difficulty by considering the Chi-square distribution. For a variance, or variance component, estimated on v degrees of freedom, the divisors to construct 90% confidence limits for the true variance are calculated from the 5% and 95% points of the Chi-square distribution. Some values are shown.

Degrees of freedom of sample variance	Divisors (of sample variance) for Lower limit	Upper limit
4	0.18	2.37
6	0.27	2.10
9	0.37	1.88
12	0.44	1.75
15	0.48	1.67
20	0.54	1.57

Thus even for the estimates of $\sigma^2(s)$ and $\sigma^2(y)$ from the complete simulated data sample, based on 9 df, the uncertainty about the value of the variance component is between

Sample variance /1.88 and Sample variance /0.37.

I believe that in most cases the estimation of the predictive variance of the yield difference on which a recommendation is based will most effectively be based on the total sample of data values of the yield difference. Where the data sample allows the estimation of variance components to be attempted using the methods illustrated in documents 2B, 2C or 2D then this should be attempted. However for all estimates the precision of variance estimates should be assessed and the resulting calculated risk probabilities interpreted in the context of that precision.

The conclusions for the planning of verification experiments prior to making a recommendation are that the number and distribution of trials must be large.

Document 2B

DATA FROM 30 NxP TRIALS AT POZA RICA

1. Introduction

The trials are part of a larger set performed over two decades on farms in the Poza Rica region. Each trial has two replicates of twelve treatments arranged in two randomised complete blocks of twelve plots per block. The twelve treatments were all combinations of four levels of Nitrogen (0, 50, 100, 150 kgN/ha) with three levels of Phosphorus (0, 40, 80kg P_2O_5/ha).

The 30 trials are spread between 1976 and 1986, in two seasons (A and B) as shown:

	Season A	Season B
1976	1 trial	1 trial
1979	1 trial	2 trials
1980	3 trials	1 trial
1981		4 trials
1982	3 trials	2 trials
1983	3 trials	5 trials
1984	1 trial	2 trials
1986		1 trial

The trials are also spread geographically with some repetition of villages and some other possible clustering. Based on partial information seven groups of locations are defined as follows:

Group 1	Papatlarillo, Copal;
Group 2	El Palmar, Mahuapan, Mamey, Ind.Nat.II;
Group 3	Jiliapa, Zaragoza;
Group 4	La Reforma, Tihuatlan;
Group 5	Zapotalillo;
Group 6	Cardel, Zapotalillo (Calvo, Hernández);
Group 7	Tierra Blanca, Sabanillas.

A summary paper for 19 trials in 1973 to 1977 examining average responses for each season and using net benefit analysis suggests that the important treatments are N0P0, N1P0, N1P1 and N2P0.

For 26 of the trials, results from the preliminary analysis were available providing Error Mean Squares and F ratios for N and P main effects and for the interaction N x P. The general pattern of the evidence indicated that both main effects were important (11 and 8 trials with 5% significant F ratios for N and P respectively and almost all F ratios for P greater than 1). There were three significant interaction F ratios (two of them from experiments with very low error mean squares) but only nine of the 26 interaction F ratios were greater than 1. The overall evidence is, I believe, quite clear that a main effects model is adequate for these trials, not only overall but also for individual trials.

2. The Distribution of Treatment Differences

With a sample which includes a wide range of both years and locations, such as the present set of trials, it is reasonable to hope that the sample is sufficiently representative, simultaneously, of future years and locations. Therefore we first examine the distribution, over the 30 sample points, of treatment differences expecting that the mean and variance of the sample of treatment differences will be adequate for estimating risk from a recommendation based on the mean difference. We shall check the possibility of consistent year and location differences after this initial analysis.

If a main effects model is the correct summary for each trial then the best estimates of differences of yield between treatment combinations will be derived from the main effect estimates and not from comparisons of the particular combinations. This follows not from the results from individual experiments but from the overall pattern of results, and is an important consequence of being able to consider a substantial set of experiments.

2.1 Estimates Based on a Fitted Quadratic Response

Calculation of fitted response models can always be achieved through the use of a multiple regression program. The model to be fitted can be determined by comparing the residual mean squares for alternative possible models. A general model would be

Yield difference = $a + b1*N + c1*N^2 + b2*P + c2*P^2 + d*NP$

in which the $b1*N$, $c1*N^2$, $b2*P$ and $c2*P^2$ terms are components of main effects and the $d*NP$ term is part of the interaction.

Since we have decided that the interaction term is negligible we omit the d*PN term. The estimation of quadratic response terms for N and P can now be calculated separately for each factor. Instead of using multiple regression we can estimate the linear and quadratic terms directly from the main effect means for the two factors. This is quicker than using a regression program when we have only three or four levels of the factor.

For the response to P, for which there are three levels the quadratic will fit the mean yields for P0, P1 and P2 exactly. Our purpose in fitting quadratic responses is to estimate yield differences between levels of each factor. We do not therefore need to calculate the linear and quadratic terms from the mean yields and then calculate estimates of the mean yield differences between levels from the linear and quadratic terms, since we would simply get back to where we had started.

The main effect estimate of P1-P0 for any level of N will therefore be the difference between the mean yields for P1 and for P0. This difference will be based on 8 observations per mean and, for each experiment the estimate will have a variance

$$2\,\sigma^2/8$$

compared with the variance for N0P1-N0P0 which is

$$2\,\sigma^2/2$$

(where σ^2 is the random variance estimated by the experimental error mean square).

There are four levels of N and the quadratic response has to be estimated from the four mean yields for N0, N1, N2 and N3. This is done by finding appropriate contrasts of the mean yields. If the response model is written

$$\text{yield} = a + b(N - \text{Nmean}) + c(N - \text{Nmean})^2,$$

and we code the levels of N as 0,1,2 and 3 so that Nmean = 1.5, then the mean yields for N0, N1, N2 and N3 are

$$N0 = a - 1.5b + 2.25c$$
$$N1 = a - 0.5b + 0.25c$$
$$N2 = a + 0.5b + 0.25c$$
$$N3 = a + 1.5b + 2.25c.$$

By manipulation of these equations we can produce

$$N3 - N0 = 3b$$
$$N2 - N1 = b.$$

Hence, by simple regression, the estimate for b is

$$b = (3(N3-N0) + (N2-N1)) /10.$$

Similarly, from the equations for N0, N1, N2 and N3

$$N3 - N2 = b + 2c$$
$$N1 - N0 = b - 2c.$$

Hence, by subtraction, the estimate for c is

$$c = (N3 - N2 - N1 + N0) /4.$$

Finally $N1 + N2 = 2a + 0.5\,c$

so that the estimate for a is

$$a = (N2 + N1 - c/2) /2.$$

During the course of this derivation, which can also be developed from the general statistical theory of treatment contrasts, we have derived expressions for N1 - N0 and for N2 - N1. Hence the main effect estimates of N1-N0 and of N2-N1, based on the quadratic response are

Estimate of (N1-N0), Est(Q) = b - 2c

Estimate of (N2-N1), Est(Q) = b.

The estimates of b and c are linear combinations of mean yields for N0, N1, N2 and N3 and the variance of each mean yield is

$$\sigma^2/6.$$

Hence the variances of b and c are

Variance(b) = $\sigma^2/6\ (3^2 + 3^2 + 1^2 + 1^2)/10^2 = \sigma^2/30$

Variance(c) = $\sigma^2/6\ (1^2 + 1^2 + 1^2 + 1^2)/4^2 = \sigma^2/24$

and the variances of the estimates of N1-N0 and N2-N1 based on the quadratic response are

Variance (Est(Q) N1-N0) = $\sigma^2(1/30 + 4/24) = \sigma^2/5$

Variance (Est(Q) N2-N1) = $\sigma^2/30$

compared with the variances for N1P0-N0P0 and N2P0-N1P0 of

$2\sigma^2/2$.

Note the improved precision first from the use of main effects instead of individual treatment comparisons, and then from regression estimates of the response. The estimate of the linear term is particularly precise.

3. Comparisons of Effect Estimates

To illustrate the benefits of improved information from the use of estimates of treatment mean differences based on the main effects fitted quadratic response, we shall calculate estimates both from our fitted quadratic response and from direct comparison of mean yields for treatment comparisons.

3.1 Estimates of N1P0-N0P0.

The estimates of N1P0-N0P0 directly from the two experimental treatment means and from Est(Q) are calculated for each experiment (the quadratic response being also calculated for each experiment). The results are listed below and plotted in Figure 1.

Site/year	Direct	Est(Q)	Site/year	Direct	Est(Q)
1	+0.06	+0.47	2	-0.55	+0.06
3	+1.71	+1.16	4	+0.84	+0.74
5	+0.43	+0.44	6	-0.14	-0.14
7	+0.54	+0.22	8	+0.05	+0.50
9	+1.20	+0.96	10	-0.45	+0.35
11	+0.27	+0.18	12	-0.22	+0.29
13	+0.90	+0.83	14	+0.73	+0.52
15	-0.01	+0.83	16	-0.55	+0.14
17	+0.16	+0.24	18	-0.45	+0.10
19	-0.12	+0.23	20	-0.34	-0.07
21	-0.14	+0.48	22	-0.23	+0.04
23	-0.03	-0.42	24	+0.39	+0.33
25	+0.34	+0.31	26	-0.04	+0.15
27	+1.37	+1.18	28	+0.25	+0.08
29	+0.19	+0.47	30	+0.66	+0.23

Since the average experimental error mean square for the trials is about 0.35 the experimental standard errors for these estimates of these estimates are approximately

S.E. (direct) = $\sqrt{(2\ (0.35)/2)} = 0.59$

S.E. (Est(Q)) = $\sqrt{(0.35/5)} = 0.26$.

These express the average within-experiment precision with which each single value is estimated. The improved precision of the estimates based on the fitted quadratic response is reflected in the reduced variation of those values shown in figure 1.

The means and variances for the two samples of 30 values for N1P0-N0P0 are

	Direct	**Est(Q)**
Mean	+0.227	+0.363
Variance	0.324	0.136
Standard deviation	0.570	0.369
Experimental standard error	0.59	0.26

The variance and standard deviation summarize the variation of the effect estimates between the 30 experiments. Note that the standard deviation of the sample of 30 directly estimated values is almost the same as the approximate experimental standard error. This implies that using the imprecise direct estimates there is little evidence of variation of the treatment difference across site/year combinations. In contrast the standard deviation for the variation of Est(Q) across sites is clearly greater than the approximate experimental standard error.

3.2 Estimates of N2P0-N1P0

In exactly the same manner as in 3.1, the estimates of N2P0-N1P0 are calculated both directly from the two experimental treatment means and from Est(Q). The results are listed below and shown in Figure 2.

Site/year	Direct	Est(Q)	Site/year	Direct	Est(Q)
1	+0.39	+0.29	2	+1.10	+0.38
3	-0.41	+0.53	4	+0.20	+0.38
5	+0.12	+0.06	6	+0.62	+0.14
7	-0.43	+0.32	8	+1.49	+0.26
9	-0.04	+0.20	10	-0.35	+0.11
11	+0.25	+0.24	12	+0.65	+0.09
13	+0.06	+0.31	14	+0.20	+0.52
15	+0.58	+0.31	16	+0.86	+0.16
17	+0.22	+0.16	18	+0.07	+0.02
19	+0.67	+0.05	20	-0.08	+0.03
21	+0.81	+0.26	22	-0.11	+0.66
23	0.00	-0.12	24	+0.15	+0.13
25	+0.45	+0.25	26	+0.50	+0.07
27	+0.26	+0.54	28	+0.23	+0.16
29	+0.62	+0.19	30	+0.44	+0.03

The experimental standard errors for these values are approximately

$$S.E. (Direct) = \sqrt{(2(0.35)/2)} = 0.59$$

$$S.E. (Est(Q)) = \sqrt{(0.35/30)} = 0.11.$$

The means and variances for the two samples of 30 values for N2P0-N1P0 are

	Direct	Est(Q)
Mean	+0.317	+0.224
Variance	0.188	0.032
Standard Deviation	0.433	0.179
Experimental standard error	0.59	0.11

Note again that for the direct form of estimate the standard deviation across site/years is smaller than would be expected from the experimental standard error but that for the Est(Q) the standard deviation across site/years is larger than the experimental standard error.

A further aspect of the relative imprecision of the estimates obtained directly from the comparison of treatments is seen when we look at both the N1P0-N0P0 and N2P0-N1P0 estimates. The mean values of the direct estimates are +0.227 for N1P0-N0P0 and +0.317 for N2P0-N1P0. It would be most surprising if the second increase were really greater than the first. In contrast, the Est(Q) mean values are +0.362 for N1P0-N0P0 and +0.224 for N2P0-N1P0, showing the expected diminishing returns, as N is increased.

3.3 Estimates of N1P1-N0P1

The estimates of N1P1-N1P0 calculated directly from the experimental treatment means and from Est(Q) are listed below and plotted in Figure 3.

Site/year	Direct	Est(Q)	Site/year	Direct	Est(Q)
1	+1.01	+0.64	2	+0.75	+0.06
3	-0.49	+0.48	4	+1.08	+1.39
5	-0.51	+0.44	6	-0.46	+0.12
7	+0.56	+0.99	8	+1.12	+0.23
9	-0.44	-0.04	10	-0.05	+0.26
11	-0.22	+0.31	12	+0.70	+0.31
13	-0.25	-0.21	14	+0.21	+0.54
15	+0.68	+0.25	16	+0.55	+0.19
17	+0.05	+0.12	18	+0.70	+0.38
19	+0.33	+0.40	20	+0.45	+0.26
21	+0.48	+0.33	22	+0.51	+0.45
23	+1.02	+1.02	24	+0.27	+0.23
25	+1.07	+0.47	26	-0.13	-0.01
27	+0.77	+0.90	28	-0.19	+0.20
29	+1.57	+0.53	30	+0.63	+0.36

The experimental standard errors of these values are approximately

$$\text{S.E. (Direct)} = \sqrt{(2(0.35)/2)} = 0.59$$

$$\text{S.E. (Est(Q))} = \sqrt{(2(0.35)/8)} = 0.30.$$

The means and variances for the two samples of 30 values for N1P1-N1P0 are

	Direct	Est(Q)
Mean	+0.392	+0.387
Variance	0.313	0.105
Standard deviation	0.560	0.324
Experimental standard error	0.59	0.30

Once again we note that the variation of the direct estimate across site/years is slightly less than would be expected from the experimental standard errors and the Est(Q) varies slightly more than expected from its experimental standard error.

The large experimental standard errors, particularly for the less precise direct estimates, emphasize the difficulty of assessing variation of treatment difference across site/year combinations. Using the better Est(Q) form of estimates of the effects of changing treatments we can be more confident of the existence of variation across site/years and have a better estimate of the risk attached to the recommendation to change fertiliser level.

4. Differences between Sites and between Years

To investigate the possible consistency of differences between site groups and years we consider the N1P0-N0P0 effect estimated by Est(Q).

	Site Group						
Year/season	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1976 A	+0.47						
1976 B		+0.06					
1979 A						+0.50	
1979 B			+0.18		+0.96		
1980 A		+0.33		+0.31			
		+0.15					
1980 B							-0.42
1981 B	+0.04				+0.23	+0.48	-0.07
1982 A	+0.52	+0.83				+0.83	
1982 B					+0.29	+0.35	
1983 A			+0.47		+0.08		
			+0.23				
1983 B			+1.16	-0.14	+0.22		
			+0.74				
			+0.44				
1984 B		+1.18		+0.24			+0.14
1986 B						+0.10	

We could calculate a complete analysis (non-orthogonal) for site groups, seasons and years. Instead we look at each factor separately, initial, calculating separate analyses of variance, to assess the variation between and within (i) site groups, (ii) seasons and (iii) years.

		SS	**df**	**MS**
(i)	Between Site group	1.1812	6	0.197
	Within Site groups	2.7573	23	0.120
(ii)	Between Seasons	0.0823	1	0.082
	Within Seasons	3.8561	28	0.138
(iii)	Between Years	0.9254	7	0.132
	Within Years	3.0131	22	0.137

Only the effect of sites seems at all interesting.

If (using the sweeping method described in detail in Mead, 1990e) we subtract the site group means from the individual data values for that site group we can examine the season and year differences after allowing for site group differences.

	Site groups						
Year/season	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1976 A	+0.13						
1976 B		-0.45					
1979 A						+0.05	
1979 B			-0.36		+0.60		
1980 A		-0.18		+0.17			
		-0.36					
1980 B							-0.30
1981 B	-0.30				-0.13	+0.03	+0.05
1982 A	+0.18	+0.32				+0.38	
1982 B					-0.07	-0.10	
1983 A			-0.07		-0.28		
			-0.31				
1983 B			+0.62	-0.28	-0.14		
			+0.20				
			-0.10				
1984 B		+0.65		+0.10			+0.26
1986 B						-0.35	

Calculating means for seasons and for years shows virtually no difference between seasons but some differences between years. We therefore ignore seasons amalgamating the seasons results within years and continue to sweep out differences between years and between site groups until the final residuals shown are reached, as follows:-

	Site groups						
Years	**1**	**2**	**3**	**4**	**5**	**6**	**7**
1976	+0.26	-0.26					
1979			-0.46		+0.53	-0.06	
1980		-0.03		+0.37			-0.12
		-0.21					
1981	-0.26				-0.02	+0.10	+0.18
1982	0.00	+0.20			-0.18	+0.23	
						-0.25	
1983			-0.04	-0.18	-0.22		
			-0.28		-0.08		
			+0.65				
			+0.23				
			-0.07				
1984		+0.29		-0.21			-0.07
1986						0.00	

The sum of squares of these final residuals is $(0.26^2 + 0.26^2 + 0.46^2 + +0.07^2 + 0^2) = 1.9064$.

The summary analysis of variance is therefore

Source	SS	df	MS
Between site groups (ignoring years)	1.1812	6	0.197
Between years + seasons (adjusted for site differences) (calculated by subtraction)	0.8509	8	0.106
Residual	1.9064	15	0.127

As suggested by the initial analyses of variance, neither years nor seasons show any significance and this confirms that the only possible consistent pattern is that between site groups. Reverting to the initial analysis for site groups the original F ratio is 1.64 (on 6 and 23 df) some way below significance at 10%. The group means are

group 1	+0.34
group 2	+0.51
group 3	+0.54
group 4	+0.14
group 5	+0.36
group 6	+0.45
group 7	-0.14

These continue to stimulate some interest with the lowest group (7) being extreme to the North, group 4 being nearest geographically to group 7, and group 2 being the most extreme to the South. If we omit group 7 the analysis of variance is

	SS	df	MS
Between site groups	0.4059	5	0.0812
Within site groups	2.5972	21	0.1237

Any suggestion of between site group differences has gone and it might therefore be of interest considering the risk probabilities not only for the complete sample of 30 experiments, but also for the reduced sample of 27 omitting the three sites in group 7.

5. Assessment of Risk

5.1 Risk Based on the Complete Sample of 30 Experiments.

We shall consider each of the treatment comparisons, N1P0-N0P0, N2P0-N1P0 and N1P1-N1P0. Using the costs quoted in the previous summary paper the marginal changes in total costs that vary and the marginal net benefits can be calculated. We start with the table of mean yield changes and their standard deviations across site/year combinations (the methodology here is similar to that in Mead, 1990f.)

Change	N1P0-N0P0	N2P0-N1P0	N1P1-N1P0
Yield Difference			
Mean	+0.363	+0.224	+0.387
Standard deviation	0.369	0.179	0.324

Now these marginal increases in yield are converted to net benefits using a multiplier of $2160 for yields and the marginal costs from the previous calculations (using peso costs prevailing inn 1977).

Marginal TCV	$359	$359	$259
Marginal Net Benefits			
Mean	$425	$125	$577
Standard deviation	$797	$387	$700
Marginal Rate of Return			
Mean	1.18	0.35	2.23
Standard deviation	2.22	1.08	2.70

On the mean values both N1P0-N0P0 and N1P1-N1P0 give acceptable MRR over 100%. The N2P0-N1P0 does not.

The risk probabilities are calculated from the Normal probability distribution using the mean and standard deviation for the relevant variable.

Net Benefits

The probability of a net benefit greater than zero from the change N1P0-N0P0 is calculated from the normal deviate

$$z = (0 - 425)/797 = -0.53$$

giving

Prob(greater than zero) = 0.702
(from Normal distribution tables);

For the change N1P1-N1P0

$$z = (0 - 577)/700 = -0.82$$

giving

Prob(greater than zero) = 0.794.

Marginal Rate of Return

The probability of a MRR greater than 100% (1.0) is calculated in the same way.

For the change N1P0-N0P0

$$z = (1.0 - 1.18)/2.22 = 0.08$$

Prob(greater than 100%) = 0.532;

For the change N1P1-N1P0

$$z = (1.0 - 2.23)/2.70 = 0.45$$

Prob(greater than 100%) = 0.675.

Finally, note that because we have used estimates based on main effects, the assessment of the change N0P1-N0P0 will be exactly the same as that for N1P1-N1P0.

5.2 Risk Probabilities Omitting Sites in Group 7

It is interesting to examine the effect on the risk probability calculations of omitting the three experiments which were quite some distance North of the rest of the experiments. Although the justification for this omission is based on the analysis for the N1P0-N0P0 values only, we will examine the effect on the risk probabilities for both N1P0-N0P0 and N1P1-N1P0.

The revised figures are

Change	N1P0-N0P0	N1P1-N1P0
Yield Difference		
Mean	+0.416	+0.376
Standard Deviation	0.340	0.330
Marginal TCV	$359	$259
Marginal Net Benefits		
Mean	$540	$553
Standard Deviation	$734	$713
Marginal rate of Return		
Mean	1.50	2.14
Standard Deviation	2.04	2.75

Net Benefits

For the change N1P0-N0P0

$z = (0 - 540)/734 = -0.74$

Prob (greater than zero) = 0.770.

For the change N1P1-N1P0

$z = (0 -553)/713 = -0.78$

Prob (greater than zero) = 0-779.

Marginal Rate of Return

For the change N1P0-N0P0

$z = (1.0 - 1.50)/2.04 = 0.24$

Prob (greater than 100%) = 0.595.

For the change N1P1-N1P0

$$z = (1.0 - 2.14)/2.75 = 0.41$$

Prob (greater than 100%) = 0.659.

The restriction to 27 experiments, and to a smaller recommendation domain, has improved the risk probabilities for the N1P0-N0P0 change, which was the effect which suggested the omission. There is little affect on results for the N1P1-N1P0 change.

GHANA FERTILISER PLACEMENT TRIALS

This is a summary of an initial look at data provided with potential for a recommendation. The recommendation to be made in this instance was that a change of treatment would make no difference to yield, while reducing costs.

Data was available from 27 experiments, 11 in 1982 and 16 in 1983. From each experiment the following information was extracted:

- Site name
- Year
- Yield for treatment T1P1 (averaged over types of N)
- Yield for treatment T2P2 (averaged over types of N)
- Difference T2P2 - T1P1
- Mean of T1P1 and T2P2
- Rainfall at nearest source of information

The comparison between T1P1 and T2P2 was a comparison of previous recommendation with prospective recommendation. The calculated data is shown

Site	Year	T1P1	T2P2	Diff	Mean	Rainfall(dist)
Agogo	1982	6.24	7.05	+0.81	6.64	676 (30miles)
Mampong	1982	4.96	4.80	-0.16	4.88	676 (25miles)
Suromani	1982	5.84	5.35	-0.49	5.60	754 (30miles)
Wenchi	1982	3.93	4.13	+0.20	4.03	754 (35miles)
Abetifi	1982	4.78	6.75	+1.97	5.76	835 (0miles)
Akuase	1982	4.28	4.17	-0.11	4.22	1215 (60miles)
Enyan D.	1982	2.96	3.42	+0.46	3.19	915 (20miles)
Gomoa Adam	1982	2.65	3.33	+0.68	2.99	1075 (20miles)
Simbrofo	1982	3.38	2.71	-0.67	3.04	915 (30miles)
Kwami-Krom	1982	3.72	2.68	-1.04	3.20	935 (mean)
Dzolokpuita	1982	2.40	2.06	-0.34	2.23	739 (50miles)
Anwomaso	1983	0.60	0.56	-0.04	0.58	635 (0miles)
Ejura	1983	4.70	4.33	-0.37	4.52	606 (50miles)
Techiman	1983	0.63	0.82	+0.19	0.72	606 (20miles)
Yamfo	1983	0.16	0.21	+0.05	0.18	606 (10miles)
Pepease	1983	3.80	4.08	+0.28	3.94	677 (10miles)
Konko	1983	3.59	3.11	-0.48	3.35	388 (0miles)
Jukwa	1983	3.38	3.48	+0.10	3.43	387 (45miles)
Bawjiase	1983	1.73	1.78	+0.05	1.76	387 (15miles)
Logba	1983	5.15	4.58	-0.57	4.86	369 (20miles)
Okadzakrom	1983	5.12	5.06	-0.06	5.09	523 (80miles)
Dzalele	1983	2.24	2.64	+0.40	2.44	298 (30miles)
Matse	1983	4.63	4.21	-0.42	4.42	298 (30miles)
Damongo	1983	2.18	2.12	-0.06	2.15	750 (60miles)
Tampion	1983	0.88	1.01	+0.13	0.94	485 (15miles)
Salaga	1983	2.14	2.22	+0.08	2.18	750 (90miles)
Navrongo	1983	4.47	5.29	+0.82	4.88	485 (60miles)

Initial inspection of the data did not suggest any substantial difference T2P2 - T1P1, nor that the difference values were related to rainfall or, indeed, to site mean yield. Graphs of difference against rainfall, and of difference against mean site yield supported these impressions. A graph of mean site yield against rainfall confirmed the suspicion that these two variables were also unrelated.

A multiple regression of difference on rainfall and on mean yield confirmed the lack of relationships. The fitted model was

Difference = -0.338 +0.067(mean yield) +0.00025 (rainfall).

The analysis of variance was

Source	SS	df	MS
Regression	0.3945	2	0.1972
Residual	8.5694	24	0.3571
Total	8.9639	26	0.3448

In the absence of any relationship on which to base indirect assessment of year-to-year variability, the only possible summary of the variability must be derived from the sample of 27 year x site combinations.

Within the two sets of sites some pairs of sites were judged to be close enough to be thought of as similar. The difference values classified by site and year are shown

	1982	1983
Site type 1	+0.81	-0.04
Site type 2	-0.16	-0.37
Site type 3	+0.20	+0.19 +0.05
Site type 4	+1.97	+0.28
Site type 5	+0.68	+0.10
Site type 6	+0.46	+0.05
Site type 7	-0.34	-0.57 -0.06 -0.42

There does seem to be some consistency of result within a site type. The analysis of variance for sites and years has to be calculated for both orders of fitting and the results are

Source	SS	df	MS
Between years (ignoring types)	1.4376	1	1.4376
Between types (adjusted for years)	3.2147	6	0.5358
Error	1.2633	9	0.1404
Total	5.9156	16	
Between types (ignoring years)	3.4789	6	0.5798
Between years (adjusted for types)	1.1734	1	1.1734
Error	1.2633	9	0.1404

The analysis confirms the impression from the initial tabulation that there are consistent type differences, and also a year difference, which give mean squares larger by a factor between 3 and 4 than that for the interaction. However with only two years we cannot say anything useful about the variance component for years. Therefore, unless some relationship of the form searched for earlier is discovered, we must rely on the total sample of 27 values being representative of the population for which the recommendation is sought.

The data are therefore appropriately summarised by the mean and variance of the entire sample.

Mean difference = +0.052

Variance = 0.3446
(from the total sum of squares in the Regression ANOVA)

S.E.(mean) = 0.113

The best estimate of risk for a future site-year combination is obtained by assuming that the difference is normally distributed with mean +0.052 and standard deviation 0.59 (= square-root of 0.3446).

Thus the risk probability that the treatment difference for T2P2 - T1P1 is less than zero is calculated from the normal deviate

$$z = (0 - 0.052)/0.59 = -0.09$$

Prob (difference less than zero) = 0.46.

For the risk probability of a difference less than -0.5 (that is of a disadvantage from the recommendation of 0.5)

$$z = (-0.5 - 0.052)/0.59 = -0.93$$

Prob (difference less than -0.5) = 0.176.

For the risk probability of a difference less than -1.0

$$z = (-1.0 - 0.052)/0.59 = -1.78$$

prob (difference less than -1.0) = 0.04.

Document 2D

SIMULATED SITE x YEAR DATA AND ANALYSIS

1. Simulation Model

A simulation set of data for 100 site x year combinations was constructed to illustrate and test methods of assessing the risk attached to recommendations. The data set was constructed according to the following model.

$$\text{Treatment Yield Effect} = \text{Mean} + \text{Site Effect} + \text{Year Effect} + B\,(\text{Rain} - 7) + \text{Error}$$

The model is defined for a treatment yield effect rather than a treatment yield because we are considering the situation where a recommendation to change from one practice to a different one is intended and we therefore wish to assess the variability of the yield difference resulting from the change. The dependence on rainfall is included to represent some relation with an additional variable for which data would be available for other years and sites. The numerical values used in the simulation were as follows:

Overall mean = 30

Site effects ranging between -10 and +10

Year effects ranging between -10 and +10

Rainfall values ranging between 3 and 14 (with some patterns of consistency between sites and years)

Rainfall effect coefficient (B) = 2

Errors normally distributed with mean zero and standard deviation of 10.

The units were based, very loosely, on values in a large scale verification trial from Ghana (G.Edmeades) in which yield differences (measured in (bags/acre)*10) ranged between -10 and +100 (the range here is a bit less being between -14 and +69). The rainfall units envisaged are 100mm. Site and year variation is pure guesswork but is chosen to be large enough to make a substantial contribution to the overall variation in the sample.

Although these efforts have been made to clothe the simulation results in a degree of realism, it is important to recognize that the actual scale of values used is completely irrelevant. The question being asked is to what extent the known properties of the complete sample can be estimated from small subsamples with particular structures.

2. Simulated sample values

The resulting simulation values which are used in the rest of this paper are shown.

Rainfall

	Year									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Site										
1	4	6	3	9	10	6	7	11	8	5
2	3	9	3	6	12	5	8	13	8	3
3	5	12	6	8	9	4	4	9	11	7
4	3	8	7	7	7	5	3	10	10	4
5	5	11	4	9	11	7	7	12	13	3
6	8	13	11	8	12	9	11	13	11	9
7	9	12	10	9	11	11	8	12	14	8
8	7	7	7	7	10	9	7	10	9	7
9	7	10	7	10	13	8	12	14	12	11
10	9	9	8	10	8	10	11	11	9	7

Treatment Yield Effect

	Year									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Site										
1	30	33	2	23	33	12	38	39	23	7
2	11	34	18	9	32	23	42	63	19	-1
3	33	28	22	12	26	-11	22	34	8	17
4	28	28	-6	14	32	-14	16	23	5	-11
5	41	31	22	10	31	41	41	48	44	30
6	61	50	34	32	50	28	37	69	46	29
7	35	50	39	29	24	33	40	55	48	12
8	17	27	32	20	35	20	35	28	38	21
9	58	37	24	22	63	45	40	62	30	41
10	45	43	18	35	34	34	33	34	41	26

3. Analysis of the Complete Sample

The analysis of variance for the 10 sites x 10 years is shown.

Source	SS	df	MS
Sites	9394	9	1044
Years	7433	9	825
Error(Interaction)	7869	81	97
Total	24696	99	249

Variance components are estimated as follows:

$$\sigma^2 = 97$$

$$\sigma^2(s) = (1044 - 97)/\ 10 = 95$$

$$\sigma^2(y) = (825 - 97)/\ 10 = 73.$$

The variance of a predicted value for a future site and year combination is

$$\sigma^2 + \sigma^2(s) + \sigma^2(y) = 97 + 95 + 7'' = 265.$$

Note that for this balanced sample this would be well estimated by the variance for the total sample = 249.

4. A Sample of 7 Sites in 2 Years with 4 Common Sites

Seven sites were selected randomly from year 9 and a second random sample of seven from year 10, with the restriction that only four sites should appear in both samples (i.e. all 10 sites were included at least once). The resulting sample values for yield effect and rainfall are shown.

	Yield Effect			Rainfall	
	Year			Year	
	9	10		9	10
Site			**Site**		
1		7	**1**		5
2		19	**2**	8	.
3	8	17	**3**	11	7
4	5	-11	**4**	10	4
5		44	**5**	13	
6		29	**6**		9
7	48	12	**7**	14	8
8	38	21	**8**	9	7
9		41	**9**		11
10	41		**10**	9	

Examining the data we can see that we have two relatively extreme, and different years and a wide range of rainfall levels. We assume that additionally rainfall data are available from the common sites (3,4,7,8) for eight years (1,2,3,4,5,6,7,8) as shown.

	Year							
	1	2	3	4	5	6	7	8
Site								
3	5	12	6	8	9	5	3	10
4	3	8	7	7	7	5	3	10
7	9	12	10	9	11	11	8	12
8	7	7	7	7	10	9	7	10

We employ three forms of analysis. First analysis of variance of the observed data to estimate site, year and error variances. Second the regression of yield on rainfall. Finally the analysis of variance for the rainfall data to estimate site, year and error variances.

For the initial analysis of variance information about year and error variation is available only from the common sites.

	Year	
	9	10
Site		
3	8	17
4	5	-11
7	48	12
8	38	21

Source	SS	df	MS
Sites	1490	3	497
Years	450	1	450
Error	512	3	171

The analysis of variance for the whole sample (which could be obtained directly from a computer analysis) is:

Source	SS	df	MS
Sites	3150	9	350
Years	450	1	450
Error	512	3	171

The estimates of components of variance are

σ^2 = 171 (on 3 df)

$\sigma^2(s)$ = (350 - 171)/1.4 = 128 (on 9 df but using 3 df estimate of σ^2)

$\sigma^2(y)$ = (450 - 171)/4 = 70 (on 1 df)

The divisors used in calculating the variance components are (for sites) the average replication per site, and (for years) the replication of common sites, which are the only ones providing information about the year variation. Inevitably the estimates of variance components are very imprecise based on so few df and observations and we have been lucky in our sample in that these variance component estimates are not far from those calculated earlier for the complete sample (97,95 and 73).

The regression of yield effect on rainfall gives the following results:

Corrected SS for rainfall = 101
Corrected SS for yield = 4112
Corrected sum of products = 462

Regression line: y = 4.58 (rainfall) -18.11.

Analysis of variance:

Regression SS = 2113 on 1 df
Residual SS = 1999 on 12 df Residual MS = 167

Variance of regression coefficient = 167/101 =1.65.

If we propose to use the rainfall data for other year x site combinations to estimate year and site variation then we shall be assuming a relationship

$$y = A + B \text{ (rain)}$$

and the consequent expression for the variance for y

Variance (y) = B^2 variance(rain) + (mean rain)^2 variance(B).

We can split the variance(rain) into components for sites, years and error, but the same is not feasible for Variance(B).

The analysis of variance for rainfall is:-

Source	SS	df	MS
Sites	70.84	3	23.61
Years	82.97	7	11.85
Error	44.91	21	2.14
Total	198.72	31	6.41

The estimates of variance components are

σ^2 = 2.14 (on 21 df)

$\sigma^2(s)$ = 2.68 (on 3 df)

$\sigma^2(y)$ = 2.43 (on 7 df)

These variance components are for rainfall and to make them relevant to the variance of yield effects we multiply by the ratio of the total variances for yield effects (predicted from rainfall) and for rainfall

Variance (y) = $(4.58)^2$ 6.41 + $(7.91)^2$ 1.65 = 237

giving the ratio = 237/6.41 = 37.0

giving estimates of variance components for yield effect variation, based on the regression model, of

σ^2 = 2.14 x 37 = 79 (on 21 df)

$\sigma^2(s)$ = 2.68 x 37 = 99 (on 3 df)

$\sigma^2(y)$ = 2.43 x 37 = 90 (on 7 df).

These estimates are based on the assumption that the split of the variability of rainfall into site, year and error components is similar to that for yield effect variation. This assumption is dubious, to put it mildly, but seems to be the only way of utilizing the regression-based information on variation.

If we combine the two sets of variance component estimates, weighting them by their degrees of freedom we get the following:

Estimated from	**Data**		**Regression**	**Combined**
σ^2	171	(3df)	79 (21df)	90
$\sigma^2(s)$	128	(9df)	99 (3df)	121
$\sigma^2(y)$	70	(1df)	90 (7df)	87

The resulting predicted variance for a future site and year combination is

$$\sigma^2 + \sigma^2(s) + \sigma^2(y) = 90 + 121 + 87 = 298.$$

This is really very respectably close to the "correct" value of 265 from the complete sample. I suspect that this is a distinctly lucky sample.

5. A Sample of 8 Sites in 2 Years with 6 Common Sites

As in example 3 two samples of 8 sites from years 7 and 8 were selected randomly with the constraint that only six common sites were included. The additional rainfall data available is for years 1 to 6 of the common sites (1,4,6,8,9,10).

	Yield effects			Rainfall	
	Years			Years	
Sites	9	10	Sites	9	10
1	38	39	1	7	11
2	63		2		13
3	22		3	4	
4	16	23	4	3	10
5		48	5		12
6	37	69	6	11	13
7	40		7	8	
8	35	28	8	7	10
9	40	62	9	12	14
10	33	34	10	11	11

Additional rainfall data

	Year					
Site	1	2	3	4	5	6
1	4	6	3	9	10	6
4	3	8	7	7	7	5
6	8	13	11	8	12	9
8	7	7	7	7	10	9
9	7	10	7	10	13	8
10	9	9	8	10	8	10

The analysis of variance for the yield effect data is calculated in two stages, first for the common sites and then for the full set of 16 observations.

	Common sites		Full		
Source	SS	df	SS	df	MS
Years	261	1	261	1	261
Sites	1597	5	2559	9	284
Error	544	5	544	5	109
Total	2402	11	3364	15	224

The SS for Years and Error from the first analysis are transferred to the second analysis and the Sites SS in the second analysis is calculated by subtraction from the Total SS in that second analysis.

The regression analysis is calculated as shown:

Corrected SS for rainfall = 152
Corrected SS for yield = 3364
Corrected sum of products = 538

$$y = 3.54(\text{rain}) + 4.45$$

Regression SS = 1904 on 1 df
Residual SS = 1460 on 14 df — Residual Mean Square = 104

Variance of regression coefficient = 104/152 = 0.68

The analysis of variance for the additional rainfall data is shown

Source	SS	df	MS
Years	50.22	5	10.04
Sites	78.89	5	15.78
Error	66.45	25	2.66
Total	195.56	35	5.59

The variance component estimates for rainfall are

$\sigma^2 = 2.66$

$\sigma^2(s) = 2.19$

$\sigma^2(y) = 1.23$

Variance(y) = B^2 variance (rain) + (mean rain)^2 variance (B)

$= (3.54)^2\, 5.59 + (8.11)^2\, 0.68 = 115.$

Hence, ratio for scaling up variance component estimates is 115/5.59 = 20.5, and the estimated variance components for yield effects, based on the regression are

$\sigma^2 = 20.5 \times 2.66 = 55$ (on 25 df)

$\sigma^2(s) = 20.5 \times 2.19 = 45$ (on 5 df)

$\sigma^2(y) = 20.5 \times 1.23 = 25$ (on 5 df).

The weighted averages of the two sets of estimates of variance components are

Estimated from	Data	Regression	Combined
σ^2	109 (5 df)	55 (25 df)	64
$\sigma^2(s)$	109 (9 df)	45 (5 df)	85
$\sigma^2(y)$	25 (1 df)	25 (5 df)	25

The resulting predicted variance for a future site year observation is

$$\sigma^2 + \sigma^2(s) + \sigma^2(y) = 64 + 85 + 25 = 174.$$

Not so good as the first example, but still not too bad an estimate of the "correct" value of 265.

6. A Sample of 4,4 and 8 Sites in 3 Years

The sample data, randomly selected from the first three years, to include two sites common to all three years, is shown:

	Yield effects				**Rainfall**		
	Years				**Years**		
	1	**2**	**3**		**1**	**2**	**3**
Sites				Sites			
1	30	33	2	**1**	4	6	3
2		34	18	**2**		9	3
3			22	**3**			6
4	28		-6	**4**	3		7
5			22	**5**			4
6			34	**6**			11
7	35			**7**	9		
8	17	27	32	**8**	7	7	7
9			24	**9**			7
10		43		**10**	9		

The spread of rainfall values is not so wide and there seems to be a lot of inconsistency of site and year differences in the yield effects data. It is likely that this data set will prove more difficult.

The additional rainfall data is from sites 1,2,4 and 8 for years 4 to 10:

				Years			
	4	**5**	**6**	**7**	**8**	**9**	**10**
Site							
1	9	10	6	7	11	8	5
2	6	12	58	13	8	3	
4	7	7	5	3	10	10	4
8	7	10	9	7	10	9	7

The analysis of variance for the yield effects data has two forms because of the non-orthogonality of sites and years (even within those sites occurring in more than one year). The sums of squares given here are for fitting years (ignoring year differences) and for sites (adjusting for year differences).

Source	SS	df	MS
Years	726	2	363
Sites	846	9	94
Error	785	4	143
Total	2357	15	157

Immediately there is a problem in estimating the variance component since our estimate of the site variance would be negative. With small df (for Error) this can happen quite often. We shall assume that the estimate of the site variance component is zero and estimate the error variance from the combined sums of squares for Sites and Error:

σ^2 = 125 (13 df)

$\sigma^2(s)$ = 0

$\sigma^2(y)$ = (363 - 125) / 3.33 = 71 (2 df).

The regression calculations are

Corrected SS for yields = 2357
Corrected SS for rainfall = 90
Corrected sum of products = 204

$y = 2.27 \text{ (rain)} + 10.22$

Regression SS = 463 on 1 df
Residual SS = 1894 on 14 df   Residual MS = 135

Variance of Regression coefficient = 135/90 = 1.5.

The analysis of variance for the additional rainfall data is shown

Source	SS	df	MS
Years	117.21	6	19.54
Sites	13.43	3	4.48
Error	47.07	18	2.62
Total	177.71	27	6.58

The estimates of variance components for rainfall are

σ^2 = 2.62 (on 18 df)

$\sigma^2(s)$ = 0.27 (on 3 df)

$\sigma^2(y)$ = 4.23 (on 6 df).

The scaling-up factor is

$$((2.27)^2\ 6.58 + (7.71)^2\ 1.50\ )/\ 6.58 = 18.7.$$

The scaled-up estimates of variance components are

σ^2 = 49 (on 18 df)

$\sigma^2(s)$ = 5 (on 3 df)

$\sigma^2(y)$ = 79 (on 6 df).

The eighted averages of the variance component estimates are therefore

Estimated from	Data	Regression	Combined
σ^2	125 (13 df)	49 (18 df)	81
$\sigma^2(s)$	negative	5 (3 df)	5
$S^2(y)$	66 (2 df)	79 (6 df)	76

The resulting predicted variance for a future site year combination is

$$\sigma^2 + \sigma^2(s) + \sigma^2(y) = 81 + 5 + 76 = 162$$

Given the failure to get any useful estimate of between site variance this is not as bad as might have been expected but it is considerably less than the "correct" value.

7. Summary

The sets of estimated values of variance components are

	Full data	7 sites 2 years	8 sites 2 years	4,4 and 8 sites 3 years
σ^2	97	90	64	81
$\sigma^2(s)$	95	121	85	5
$\sigma^2(y)$	73	87	25	76
Combined	265	298	174	162

With two obvious exceptions the individual estimates of variance components are reasonably close to the "correct" values from the full data. Nevertheless these results do emphasize the well-known difficulty of estimating variance components accurately from small df.

To illustrate the problems in estimating variance components, the 100 values in the total sample were split into four sets of 5 sites by 5 years. An analysis of variance was calculated for each set and the three variance components calculated. The results were

Sites Years	(1,2,3,4,5) (1,2,3,4,5)	(1,2,3,4,5) (6,7,8,9,10)	(6,7,8,9,10) (3,4,5,6,7)	(6,7,8,9,10) (1,2,8,9,10)
σ^2	33	102	88	116
$\sigma^2(s)$	3	163	negative	64
$\sigma^2(y)$	85	167	14	53
Combined	121	432	102	233

These results make those from our three examples look quite good. The methods used to predict the variance for treatment yield differences, although they have rather dodgy aspects, seem to give reasonable estimates, given the amount of initial data. We probably cannot hope to get better estimates of the combined variance without considerably more information.

References

Mead, R., 1990a, Representation of Risk. CIMMYT Training Working Document 1B.

Binswanger, H.P., and B.C. Barah. 1980. Yield risk, risk aversion and genotype selection: Conceptual issues and approaches. ICRISAT Research Bulletin Series No. 2.

Mead, R., 1990e, Sweeping Methods for Analysis of Variance. CIMMYT Training Working Document 3C.

Mead, R., 1990f, Precision of Net Benefit Analysis. CIMMYT Training Working Document 1A.

Centro Internacional de Mejoramiento de Maíz y Trigo
International Maize and Wheat Improvement Center
Lisboa 27, Apartado Postal 6-641, 06600 México, D.F. México