Quick Analysis Methods for Random Balance Screening Experiments*

F. J. ANSCOMBE

Princeton University

A short expository account of random balance is given, in which some different types of sampling are distinguished. As a quick significance test of effectiveness of single factors, a simple analysis of variance method is recommended. For the sake of sensitivity, it is suggested that the number of levels of quantitative factors should preferably be less than five. The degree of unbalance of a random balance design is studied, largely through an example, and a desirable upper bound is suggested for the number of levels of any factor, namely one eighth of the total number of observations.

1. RANDOM BALANCE DESIGNS

Suppose it is desired to test f factors in a factorial experiment, each factor at some stated number (two or more) of levels, not necessarily the same for every factor. The total number N of treatment combinations in a complete replication is equal to the product of the numbers of levels of all the factors, and will be exceedingly large if f is large. Satterthwaite [3, 4] has suggested that a useful experiment with some small number n of experimental units or tests can be obtained if one test is made at each of a set of n treatment combinations drawn at random from the population of all N possible treatment combinations. The more orthodox procedure would be to choose for the n treatment combinations a highly systematic sample from the population of all possible combinations, namely a "fractional replicate". One might compare this with using a Monte Carlo computing procedure instead of orthodox non-stochastic computation.

There is more than one way in which a random sample of n combinations can be chosen. First, sampling may be either with or without replacement. If n is much smaller than N ($n \ll N$), replacement makes almost no difference; replacement will be assumed below. Secondly, sampling may be conditional on each factor's being represented in the sample a prearranged number of times at each level. Thus, if the first factor is to be tested at three levels and if n is divisible by 3, we may impose the condition that all three levels should be tested equally often, with similar conditions for all the other factors. I call this conditional sampling. Alternatively, no such condition may be imposed, and we then have simple unconditional sampling.

* Prepared in connection with research sponsored by the Office of Naval Research. A draft of this paper was presented at the Annual General Meeting of the Institute of Mathematical Statistics, Cambridge, Mass., August 27, 1958.

The design may be written out as an array of n rows (one for each experimental unit or test) and f columns (one for each factor); the entries are the levels of the factors. This is called the design matrix. To obtain *conditional sampling with replacement*, prearranged numbers of each of the level-symbols for any one factor are distributed at random in the corresponding column (by shuffling cards, for example), independently of the allocation of levels in the other columns. To obtain *unconditional sampling with replacement*, the entries in each column are a random sample (independent of every other column) from some chance distribution of levels. In either case, the entries in different columns are independent, and the factors are said to have random balance.*

TABLE 1

Design matrix and yields: random balance design, conditional sampling with replacement, $f = 8$, $N = 4320$, $n = 12$.

Experimental unit number	Levels of factors								Yields
	A	*B*	*C*	*D*	*E*	*F*	*G*	*H*	
1	*0*	*2*	*0*	*1*	*3*	*1*	*0*	*1*	99
2	*1*	*1*	*1*	*2*	*0*	*0*	*1*	*1*	38
3	*0*	*2*	*2*	*2*	*1*	*1*	*0*	*0*	90
4	*2*	*0*	*0*	*2*	*0*	*0*	*1*	*1*	65
5	*0*	*1*	*0*	*0*	*1*	*0*	*0*	*1*	45
6	*0*	*1*	*2*	*0*	*2*	*0*	*1*	*0*	42
7	*2*	*0*	*1*	*1*	*3*	*0*	*0*	*1*	64
8	*1*	*0*	*0*	*3*	*2*	*1*	*1*	*0*	76
9	*2*	*0*	*1*	*1*	*3*	*1*	*1*	*0*	100
10	*1*	*2*	*2*	*0*	*4*	*0*	*0*	*0*	59
11	*2*	*2*	*1*	*3*	*4*	*1*	*0*	*0*	101
12	*1*	*1*	*2*	*3*	*1*	*1*	*1*	*1*	93

There are eight factors, denoted by *A*, *B*, *C*, *D*, *E*, *F*, *G*, *H*, having various numbers of levels, from two to five. Factor *A* has three levels, denoted conventionally by *0*, *1*, *2*, each appearing four times, so that the entries in the first column of the design matrix were obtained by randomly distributing a stock of twelve symbols consisting of four *0*'s, four *1*'s, and four *2*'s. Factors *B* and *C* also have three levels each, distributed similarly (but independently). Factor *D* has four levels, denoted by *0*, *1*, *2*, *3*, each appearing three times. Factor *E* has five levels, denoted by *0*, *1*, *2*, *3*, *4*, of which *1* and *3* have been chosen to occur three times each and *0*, *2*, *4* twice each. Factors *F*, *G*, *H* have two levels each, denoted by *0* and *1*, each level of each factor occurring six times.

On the right of the design matrix is shown a column of fictitious yields. Those for which factor *F* is at level *0* are a random sample from a normal population having mean 50 and standard deviation 10; the rest (factor *F* at level *1*) are from a normal population having mean 90 and standard deviation 10.

* If sampling is *without* replacement, the entries in different columns are not completely independent. The same will be true, presumably, if conditions are imposed on the correlations between columns in the design matrix. One might suppose that conditional sampling with replacement, of the simple kind described above and illustrated in Table 1, would appeal most to users, but other types of sampling have been considered. It is desirable that in published work on random balance experiments the type of sampling, or method of writing down the design, should be specified clearly.

An example of a random balance design (conditional sampling with replacement) is shown in Table 1. For ease of illustration, n and f have been chosen to be a good deal smaller than they typically would be in practice. Although in some respects this is an unfavorable example, it does illustrate several general properties, and will be discussed in detail below. Note that with random balance designs there is no need for any particular arithmetic relation between the numbers n, f, N, as there would be for orthodox fractional replication.

Such an experiment is easy to design. It would appear most likely to be satisfactory if it was a factor-screening experiment, in which most of the factors were expected to be irrelevant, and the primary object was to identify one or more factors that had some appreciable effect. In that case, the following simple type of analysis suggested by Satterthwaite might suffice. Suppose only one observation is made on each experimental unit, say a yield. Plot the yields against the levels of each factor in turn, obtaining f scatter diagrams. Because of the random balance (independent assignment of levels for each factor), it is legitimate to forget about all the other factors when looking at any one scatter diagram, i.e. one may validly test the significance of the regression of yield on level of any one factor without making allowance for the levels of the other factors. If in fact none of the other factors has any appreciable effect this test will be not only legitimate but efficient. Satterthwaite suggests that on the diagram showing the largest regression, if there is one, a regression curve should be drawn, deviations of the yields from it should be found and plotted against the levels of all the other factors, and the process then be repeated. (For a non-quantitative factor, the term "regression curve" is not quite appropriate. Deviations would be found from the mean yield for each level separately.) In this way "significant" factors are identified one by one.

Thus random balance permits of a simple type of analysis of the observations, in which factors are considered singly. This is not necessarily the most sensitive possible method of analysis, but presumably it will be the more effective, the fewer the factors that influence yield, and fully effective if there is only one such factor.

The first stage of such a graphical analysis of the data of Table 1 is shown in Fig. 1. The most pronounced regression is clearly on factor F. The next stage would be to subtract the mean yield for the corresponding level of F, roughly 52 for level *0* and 93 for level *1*, from the given yields, and plot these residuals against the levels of each of the remaining seven factors.

2. Quick significance tests

It may be helpful on occasion to supplement the visual inspection of scatter diagrams as described above by an objective significance test, while still considering the factors only one at a time.

With most types of experimentation it is no doubt fair to say that significance tests are inappropriate. Factors are included only when they are thought likely, even certain, to influence yield, and the purpose of the experiment is to measure responses, rather than primarily to detect the possible presence of responses. But in a screening experiment, the experimenter may fully expect that most

of the factors included (all of them, if he is out of luck) will have no appreciable effect, and he may reasonably adopt the line that he will consider any factor to be irrelevant unless there is clear evidence to the contrary. The weight of such evidence is measured by a significance test.* The tests discussed here relate to each factor separately. If all f factors are without effect on yield, and if a significance test is made on each, the expected number of results "significant" at the 5% level will be $f/20$, and so on.

Suppose the factor under consideration has a small number k of levels ($2 \leq k \ll n$). If the levels are not quantitative, the obvious test criterion is derived from analysis of variance of the yields between and within levels. Even if the levels are quantitative, this is still a reasonable criterion, since one cannot be

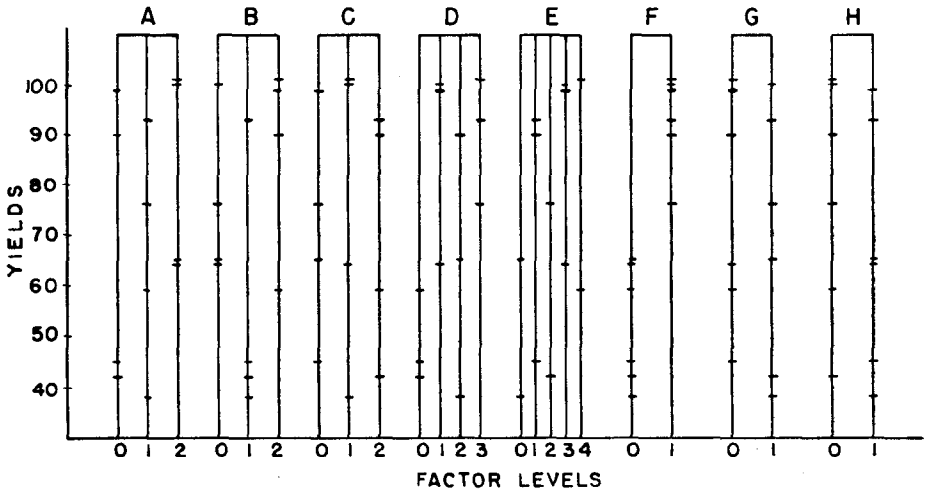

FIGURE 1—Scatter diagrams derived from Table 1.

sure what sort of regression curve is correct. One might make a normal-theory test based on the F-distribution, or (what usually comes to nearly the same thing) one might make a randomization test, following Welch [6], of the non-parametric hypothesis that the given set of yields has been associated at random with the given set of levels. Since the same set of yields will be considered repeatedly in relation to each factor in turn, the randomization test is attractive.

For the given set of yields (or possibly for the set of deviations of the yields from a previously fitted regression curve on another factor), let T denote the total sum of squares about the mean and let R denote the residual sum of squares of deviations of the yields from level means. One might take as test criterion the ratio of residual mean square to total mean square. If this is called $1 - U$, we have

* It is not *necessary*, however, to approach this problem through significance tests. Beale and Mallows [1] have considered simultaneous least-squares estimation of all responses in the light of a prior probability assumption that most responses will be small.

$$U = 1 - \frac{n-1}{n-k}\frac{R}{T}. \tag{1}$$

The upper tail of the distribution for U corresponds to a significant association between yields and levels. From Welch's results ([6], p. 152) one finds, for random association of levels with yields,

$$\mathcal{E}(U) = 0, \quad \text{var}(U) = \frac{2(k-1)}{(n+1)(n-k)}\left\{1 - \frac{K_4}{nK_2^2}\right\}, \tag{2}$$

where K_2 and K_4 are respectively the second and fourth sample cumulants or "k-statistics" of the n yields, in Fisher's terminology. (It is supposed here that each of the k levels appears equally often; otherwise a further term given by Welch needs to be added to the expression for var(U).) If the set of yields does not look markedly unlike a sample from a normal population, the factor in curly brackets can be ignored. If $n \gg k$, $(n-k)U + (k-1)$ has approximately a χ^2 distribution with $(k-1)$ degrees of freedom. The usual normal-theory test is nearly the same as this. The ratio of the mean square "between levels" to the residual mean square is

$$\frac{(n-k)U + (k-1)}{(k-1)(1-U)},$$

and this has the F-distribution with $(k-1)$ and $(n-k)$ degrees of freedom.

In Table 2 values of U are shown for all eight factors in the imaginary experiment of Table 1. The standard deviations of U have been calculated from equation (2). Only one U is clearly significant, that for factor F. Residuals from the level means for factor F have therefore been obtained, and U's recalculated for all the other seven factors. Only one U (for factor A) now exceeds, barely, twice its standard error, and since the sampling distribution for U in that case has presumably positive skewness like a χ^2 with 2 degrees of freedom the deviation cannot be considered remarkable. One will conclude that only factor F has given clear evidence of affecting yield. This conclusion happily agrees with the way the fictitious yields were composed.

A much quicker test than U is available when $k = 2$, namely Tukey's compact two-sample test [5]. A count is made of the number of yields at one of the levels that are above all the yields for the other level, plus the number of yields for the latter level below all the yields for the first—provided both these numbers are positive; otherwise no count is made. Tukey has found that percentage points of the chance distribution of the total count are nearly independent of n, under random permutation of the level-symbols, assuming approximately $\frac{1}{2}n$ observations at each level. For $n = 12$ he gives 7 as the two-sided 5% point and 9 as the two-sided 1% point. From Fig. 1 we obtain the following counts: for F, 12; for G, 3; for H, 3. The first is clearly significant, the others not. After subtracting the level means for F, we obtain: for G, no count; for H, 3; neither significant.

For a *quantitative* factor whose levels can be adjusted finely at will, it is not necessary that the number of levels k should be small. The levels tested in the

TABLE 2

Analysis of yields in Table 1: significance tests corresponding to the graphical analysis of Fig. 1.

Factor	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
				Analysis of original yields				
U	−0.11	0.22	−0.21	0.40	−0.04	0.80	−0.07	−0.04
s.d.(*U*)	0.20	0.20	0.20	0.26	0.31	0.13	0.13	0.13
				Analysis of residuals after fitting *F*				
U	0.41	0.18	−0.16	−0.05	0.41	—	0.05	−0.07
s.d.(*U*)	0.19	0.19	0.19	0.25	0.31	—	0.13	0.13

experiment could even be all different ($k = n$). In that case, a test criterion that suggests itself as being easy to calculate and independent of any assumption concerning the nature of the regression curve is one based on squared successive differences. Let the yields be arranged in ascending order of the levels of the factor, and let S denote the sum of squares of the $(n - 1)$ successive differences of the yields in that order. Then a test criterion corresponding to U above is

$$V = 1 - \frac{S}{2T}, \tag{3}$$

with the upper tail corresponding to significant regression. Young ([7], pp. 294–5) has found the first four moments of the distribution for V, under random reordering of the set of yields. In particular,

$$\mathcal{E}(V) = 0, \quad \text{var}(V) = \frac{2n - 3 - (m_4/m_2^2)}{2n(n - 1)}, \tag{4}$$

where m_2 and m_4 denote respectively the second and fourth sample moments of the yields (defined as sums of second or fourth powers of deviations from the mean, divided by n). If the set of yields does not look markedly unlike a sample from a normal population, we shall have approximately

$$\text{var}\,(V) = \frac{1}{n + 2},$$

and the distribution of V is nearly normal.

Another quick test is to group the levels together into a small number of groups and use U. But in that case the number of levels would have been better small to start with. How many levels are advisable, if the U-test is to be used? The variance of U increases with k. If we could be sure that the response to the factor (if any) would be nearly linear, there should be only two levels, spaced wide apart. Often, however, response curves have a maximum, or a noticeable curvature, and three or four, or even five, levels might be thought safer and more interesting.

Now if in fact there is a pronounced regression of yield on level, of the sort

just mentioned, and if the levels of the factor are distributed with equal spacing over a certain fixed interval, the criteria U (with k small) and V (with $k = n$) have roughly equal expectations, for they are approximately the same function of the ratio of residual variance of yields about the regression curve to gross variance. But on the null hypothesis V has a much greater variance than U. (For example, if $n = 50$ and U is based on as many as 5 factor levels, i.e. $k = 5$, the variance of V is roughly five times that of U.) It follows that V gives a much less powerful test than U.

The following conclusions may be drawn concerning quantitative factors. Suppose we are willing to assume that if any of the factors has an appreciable effect on yield the effect can be represented by a low-degree polynomial regression curve with not more than one maximum. Then

(i) no use should be made of the squared successive difference criterion V (which effectively measures serial correlation), and

(ii) analysis will be easier and nothing lost if the number of levels for each factor is small, not more than 5, preferably 3 or 4.

3. How much balance?

If two or more of the factors tested in a random balance experiment do in fact have a substantial effect and the experimenter becomes aware of this, he will wish to estimate the responses to those factors. At this point it is of interest to inquire how far the effects of the factors can be disentangled, that is, how close the design comes to being orthogonal. The fact that the degree of nonorthogonality or unbalance is random can be made the basis for an objection to the whole notion of random balance designs. Such designs may work well on the average, but should I trust to one now on this occasion? A similar objection can be raised against Monte Carlo methods of computation. And a similar objection again can be raised against randomizing the layout of an experiment having a conventional systematic design; though there the randomization seems less drastic than in a random balance experiment.

Let us consider further the imaginary experiment of Table 1. The last three columns of the design matrix each contain six *0*'s and six *1*'s. It could have happened, by an accident of randomization, that two of these three columns were identical, or identical except for a consistent interchange of *0*'s and *1*'s. In that case the two corresponding factors would have been completely confounded; if a pronounced response was associated with them, we could not tell at all (without further experimentation or background knowledge) which of the two factors was responsible. Similarly, it could have happened that two of the first three columns (which each contain four *0*'s, four *1*'s and four *2*'s) were identical, or identical except for a consistent interchange of levels; again the two corresponding factors would have been completely confounded. It is easy to calculate that any such perfect confounding is improbable, and in fact it has not happened in the actual layout of Table 1.

At the other extreme, two columns of the design matrix may have perfect orthogonality, in the sense that, associated with any chosen level of one of the factors, all the levels of the other factor appear in the same relative proportion

as they do in the experiment as a whole. In Table 1, the first and sixth columns (for factors A and F) have this perfect orthogonality, for opposite the four *0*'s in the A-column there appear equal numbers of *0*'s and *1*'s in the F-column; and similarly for the four *1*'s and for the four *2*'s in the A-column. In fact, six of the twenty-eight possible pairs of columns in the design matrix of Table 1 are orthogonal, namely A and F, A and H, C and F, C and G, F and G, G and H. If two columns are orthogonal, the corresponding factors are completely unconfounded; if one of them has an effect, it cannot cause the other to appear to have an effect. If all columns were mutually orthogonal, the response to each factor could be estimated by least squares independently of the responses to all the other factors; and this is what happens in orthodox full factorial experiments.

In Table 1, no accident of randomization could possibly have made all eight columns mutually orthogonal,* but with luck every pair of columns may be not far from orthogonal. How can one measure the lack of orthogonality of two columns? If both the factors are at two levels only, a natural measure is the square of the ordinary correlation coefficient between the column entries. It is easy to see that the result will be the same whatever numbers have been written to represent the levels. In Table 1 the levels of the two-level factors are denoted by *0* and *1*, but any other pair of unequal numbers could have been used for either factor. This squared correlation coefficient, ρ^2 say, is equal to the loss of efficiency, due to nonorthogonality, in estimating the response to each factor when both responses are simultaneously estimated by least squares (all other factors being ignored). That is, the error variance for each response is equal to what it would have been if the two factors had been orthogonal, divided by $1 - \rho^2$. We can also interpret ρ^2 as a "coefficient of influence", as follows. If one of the factors, Y say, has a real effect, while the other, X, does not, ρ^2 is the proportion of the real effect of Y (squared response) that reappears as an apparent effect of X, if X is considered separately, with no allowance made for the nonorthogonality of X and Y.

A similar squared correlation coefficient can be calculated between any pair of columns of the design matrix, even if the factors concerned have more than two levels. But now the result will depend on the numerical scoring of the levels, and will relate to particular single components of the responses to the factors. If the numerical scoring shown in Table 1 is used, the correlations will relate to the linear components of the responses, on the assumption that each factor is quantitative and its levels are a linear function of the level-symbols. All the coefficients obtained in this way from the design matrix of Table 1 are shown in Table 3.

The following proposition† can easily be proved: If the levels of two factors are represented in a design matrix by any numerical symbols (not all equal),

* Two coefficients are needed to specify the response to factor A, and two more for each of B and C, three for D, four for E, and one each for F, G and H; sixteen coefficients in all. Among the twelve observations only eleven independent comparisons can be made. All sixteen coefficients cannot therefore be estimated independently.

† Dr. C. L. Mallows informed me of this result, for the case of factors at 2 levels.

TABLE 3
Squared correlation coefficients between pairs of columns of the design matrix in Table 1.

	B	*C*	*D*	*E*	*F*	*G*	*H*
A	0.250	0.016	0.133	0.051	0.000	0.042	0.000
B		0.141	0.008	0.091	0.042	0.375	0.042
C			0.008	0.023	0.000	0.000	0.167
D				0.048	0.356	0.089	0.000
E					0.061	0.242	0.242
F						0.000	0.111
G							0.000

the average value of the squared correlation coefficient between the two columns, under random permutation of the entries in either column, is equal to $1/(n-1)$, n being the number of rows. The various levels need not appear equally frequently.

Thus the expected value of all entries in Table 3 is 1/11, and the average value of the 28 entries is in fact almost exactly that, namely 0.0906. The individual values range from 0 to 0.375.

However, if a factor is at more than two levels, we shall most likely be interested in more than just one particular component of its effect. If the levels are quantitative, the linear component of response may be the most important, but curvature is interesting too. If the factor is qualitative, probably no single contrast is known in advance to be more interesting than all others. We might ask, concerning two factors, between what limits ρ^2 would lie, if all possible components of the effects of the factors were considered. This is a question of canonical correlations. It is convenient to think in terms of coefficients of influence of one factor on another, as defined roughly above and more explicitly in the Appendix. The influence of Y on X depends on the nature of the response to Y, if Y is at more than two levels; there is usually a range of possible values for the influence coefficient. If X and Y are at an unequal number of levels, the range of values for the influence of Y on X is not necessarily the same as that for X on Y. In Table 4 are shown the greatest and least values of the influence coefficients for the design matrix of Table 1. Also shown is an average value for each coefficient, appropriate as an expectation when all patterns of response to the influencing factor are judged equally likely *a priori*. Where only one entry is shown, it is the only possible value, and so at once the lower limit, upper limit and average.

As an example to illustrate the meaning of Table 4, consider the influence on factor B by factor C. The coefficient may come anywhere from 0 to $\frac{1}{4}$. If the true response to C is proportional to 1 for level *0*, -1 for level *1*, and 0 for level *2*, while the true response to B is zero, it is easy to see that no apparent response to B is induced by C, and the influence coefficient is 0. If, on the other hand, the true response to C is proportional to 1 for levels *0* and *1* and -2 for level *2*, some of this response reappears as an apparent response to B, namely 1 for level *0* and $-\frac{1}{2}$ for levels *1* and *2*. The sum of squares for the apparent response to B is one-quarter that for the true response to C, and so the influence coefficient

TABLE 4

Lower and upper limits, and average values, of coefficients of "influence" of one factor on another, calculated from the design matrix in Table 1.

On factor	By factor *A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*
A		0.033–0.467 (av. 0.250)	0.033–0.467 (av. 0.250)	0.000–0.250 (av. 0.167)	0.000–0.402 (av. 0.146)	0.000	0.167	0.000
B	0.033–0.467 (av. 0.250)		0.000–0.250 (av. 0.125)	0.000–0.333 (av. 0.111)	0.000–0.500 (av. 0.208)	0.167	0.500	0.167
C	0.033–0.467 (av. 0.250)	0.000–0.250 (av. 0.125)		0.000–0.333 (av. 0.111)	0.000–0.458 (av. 0.146)	0.000	0.000	0.167
D	0.250	0.000–0.333 (av. 0.167)	0.000–0.333 (av. 0.167)		0.000–1.000 (av. 0.417)	0.556	0.111	0.111
E	0.181–0.402 (av. 0.292)	0.333–0.500 (av. 0.417)	0.125–0.458 (av. 0.292)	0.000–1.000 (av. 0.556)		0.222	0.556	0.556
F	0.000	0.000–0.167 (av. 0.083)	0.000	0.000–0.556 (av. 0.185)	0.000–0.222 (av. 0.056)		0.000	0.111
G	0.000–0.167 (av. 0.083)	0.000–0.500 (av. 0.250)	0.000	0.000–0.111 (av. 0.037)	0.000–0.556 (av. 0.139)	0.000		0.000
H	0.000	0.000–0.167 (av. 0.083)	0.000–0.167 (av. 0.083)	0.000–0.111 (av. 0.037)	0.000–0.556 (av. 0.139)	0.111	0.000	

is $\frac{1}{4}$. For other true responses to C, the influence coefficient will lie between 0 and $\frac{1}{4}$; the average is $\frac{1}{8}$.

As another example, consider the influence on factor E by factor D. For three possible (mutually orthogonal) true responses to D, the induced apparent responses to E, and the ratios of sums of squares of E-responses to D-responses, are as follows.

True response to D				Apparent response to E					Ratio of sums of squares (influence coefficient)
0	*1*	*2*	*3*	*0*	*1*	*2*	*3*	*4*	
1	0	0	−1	0	0	0	0	0	0
−1	0	2	−1	2	0	−1	0	−1	$\frac{2}{3}$
−1	3	−1	−1	−1	−1	−1	3	−1	1

The average of the three canonical influence coefficients on the right is $\frac{5}{9}$ or 0.556, and this is quoted in Table 4 as the average. The influence on factor D by factor E can be studied similarly. It is possible to find four orthogonal possible true responses for E, such that the influence coefficients on D are 0, 0, $\frac{2}{3}$, 1, and the average of these is $\frac{5}{12}$ or 0.417, given in the table.

The following proposition, analogous to the one already quoted, can be proved, as indicated in the Appendix: The influence coefficient of a factor Y on a factor X, averaged over permutations of entries in the columns of the design matrix, is equal to $(k - 1)/(n - 1)$, where k is the number of levels of factor X. The k levels of X need not appear equally frequently; the number of levels of Y is immaterial.

Thus the theoretical expected value for the entries in the first three rows of Table 4 is $\frac{2}{11}$ or 0.182, and the average of the 21 "average" coefficients shown is 0.157. For the next row (factor D) the expected value is $\frac{3}{11}$ or 0.273, and the actual average is 0.254. For the next row (factor E) the expected value is $\frac{4}{11}$ or 0.364, and the actual average is 0.413. For the last three rows the expected value is $\frac{1}{11}$ or 0.091 and the actual average is 0.067.

It is clear from Table 4 that the actual influence coefficients range much higher than the theoretical mean of $(k - 1)/(n - 1)$. The following approximate result is obtained in the Appendix: If ρ^2 measures the influence exerted by a factor Y (which has a certain real effect on the yields) on a factor X having no real effect, the chance distribution of $(n - 1)\rho^2$ under random permutation of the columns of the design matrix is approximately χ^2 with $(k - 1)$ degrees of freedom, k being the number of levels of X, provided $n \gg k$.

Accepting this χ^2 approximation, we can make calculations of the following kind. Suppose we agree that only rarely, say in not more than 5% of cases, should an influence coefficient exceed some sizable value, say $\frac{1}{4}$. Then $(n - 1)$ should be not less than four times the upper 5% point of the χ^2 distribution with $(k - 1)$ degrees of freedom. We find:

if all factors are at 2 levels, n should be not less than 16,
if some factors . . 3 25,
. . . 4 32,
. . . 5 . . . . 39.

These recommendations can be briefly summarized: take $n \geq 8k$, where k is the greatest number of levels of any factor. Naturally, a more or less cautious person will wish to increase or reduce (respectively) the factor 8; but it is clear at any rate that if k is large, n should be much larger.

Thus, in considering the confusion that nonorthogonality causes in the estimation of responses, we are led to much the same conclusion as before when considering significance tests, namely, that caution is advisable in assigning many levels to the factors. The fact that a random balance design can just as easily accommodate a factor at a dozen levels as at two should not encourage the thought that one can *just as well* have a dozen levels.

4. General remarks

The above discussion has concentrated on a simple approach to analysis, likely to be adequate if very few of the factors tested have an appreciable effect. Graphical and other quick methods have an essential place in statistics, if only as a guard against blunders in a more elaborate analysis. But when the carrying out of an experiment is expensive, it would usually be foolish to stop at a quick analysis, if there is the possibility that the latter may have failed to extract all the useful information available. Dempster [2] has made an interesting study of the relation between the observations in a random balance experiment (where sampling is unconditional and without replacement) and the totality of observations that would have been obtained in a complete replication (N observations), and has proposed certain analysis methods designed to bring out this relation. Beale and Mallows [1] have studied least-squares estimation of constants, when the number of the latter is of the same order of size as the number n of observations available; they show that good estimation is made possible by a small amount of prior information to the effect that most of the constants are small in magnitude, and they have discussed the basis for such prior information. Their method does not relate specifically to random balance experiments, but might be expected to be valuable for them. Satterthwaite has explored several approaches to the analysis of screening and other complex factorial experiments; see [4].*

Studies such as these are needed before we can begin to assess fairly the value of random balance designs in comparison with other types. It may well be that a systematic design can always be found which, correctly analysed, will be technically more efficient than a random balance design. But at least two things can be said to the credit of random balance:—

1. Random balance designs are very easy to write down, and the whole conception is strikingly simple. They may therefore appeal to experimenters with little statistical knowledge, for whom the only practical alternatives are simple factorial designs with few factors, or else "one at a time" methods.

2. Random balance has provided a challenge and stimulus to the purveyors of orthodox systematic designs, in two distinct directions—(a) greater flexibility

* A draft of [4] came to hand only as the present paper was being completed. It is therefore not discussed here.

in accommodating factors at various numbers of levels, (b) the feasibility of screening many factors with few observations.

5. Acknowledgment

I am much indebted to John Tukey, Colin Mallows, and the referees for helpful suggestions and improvements.

References

[1] Beale, E. M. L. and Mallows, C. L. On the analysis of screening experiments. *Annals of Mathematical Statistics* (to appear).

[2] Dempster, A. P. Random allocation designs, I: On general classes of estimation methods. *Annals of Mathematical Statististics* (to appear).

[3] Satterthwaite, F. E. Random balance experimental designs. *1957 Middle Atlantic Conference Transactions, American Society for Quality Control*, pp. 61–2.

[4] Satterthwaite, F. E. Random balance experimentation. *Technometrics*, 1 (1959).

[5] Tukey, J. W. A quick compact two-sample test to Duckworth's specifications. *Technometrics*, 1 (1959), 31–48.

[6] Welch, B. L. On tests for homogeneity. *Biometrika*, 30 (1938), 149–58.

[7] Young, L. C. On randomness in ordered sequences. *Annals of Mathematical Statistics*, 12 (1941), 293–300.

Appendix

Suppose that factor X has k levels and factor Y has l levels. Let n_{ij} denote the number of experimental units (rows of the design matrix) for which X is at level i and Y is at level j, where i ranges over $0, 1, 2, \cdots, k-1$ and j over $0, 1, 2, \cdots, l-1$. Let

$$n_{i\cdot} = \Sigma_j n_{ij} , \qquad n_{\cdot j} = \Sigma_i n_{ij} , \qquad n = \Sigma_i \Sigma_j n_{ij} .$$

For a fixed design matrix, let us regard the yields as realizations of independent random variables. Suppose that Y is the only factor having a real effect, and that at level j of Y the expected yield is $\mu_j = \bar{\mu} + \alpha_j$, where $\Sigma_j n_{\cdot j}\alpha_j = 0$. Then the apparent mean yield at level i of X has expected value

$$\bar{\mu} + \frac{\Sigma_j n_{ij}\alpha_j}{n_{i\cdot}}.$$

We define the influence coefficient of Y on X to be the ratio of the sum of squares of deviations from $\bar{\mu}$ of the expected mean yields of X (with weights $n_{i\cdot}$) to the corresponding sum of squares for the expected mean yields of Y, that is

$$\rho^2 = \frac{\Sigma_i(\Sigma_j n_{ij}\alpha_j)^2/n_{i\cdot}}{\Sigma_j n_{\cdot j}\alpha_j^2}. \tag{5}$$

The symbol ρ^2 is used, because the coefficient can be regarded as a squared multiple correlation coefficient, in the following sense. In the Y-column of the design matrix, let the level-symbols (j) be replaced by (α_j). In the X-column, let the level symbols (i) be replaced by arbitrary numbers (β_i). Then ρ^2 as defined above is the greatest value of the squared correlation coefficient between the

X and Y columns, for variation of (β_i). Thus ρ^2 is the squared correlation between a *particular* response to Y and that component of the X-effect which has greatest correlation with the response to Y. For different possible responses to Y, i.e. different (α_i), we have in general different values of ρ^2.

Stationary values of ρ^2, for variation of the vector (α_j), are easily seen to be roots λ of the $l \times l$ matrix Q whose (j, j')th term is

$$q_{jj'} = \frac{1}{n_{\cdot j}} \Sigma_i \left(\frac{n_{ij} n_{ij'}}{n_{i\cdot}} \right).$$

The corresponding (α_j) are the eigenvectors, thus:

$$\Sigma_{j'} q_{jj'} \alpha_{j'} = \lambda \alpha_j \ .$$

One root of Q is 1, with unit eigenvector. The eigenvectors for roots not equal to 1 all satisfy $\Sigma_j n_{\cdot j} \alpha_j = 0$, and so represent possible true responses to factor Y. Thus the stationary values of the influence coefficient of Y on X are all the roots of Q remaining after a 1 has been deleted. (If we now consider the influence coefficient of X on Y, the non-zero roots will be the same as for Y on X, but the number of zero roots will be different if $k \neq l$.)

To find an expected value for the influence coefficient, averaging over all possible effects (α_j) of Y, it is convenient to assign the following probability distribution to (μ_j): μ_j independently normally distributed with common mean and with variance proportional (and let us say equal) to $1/n_{\cdot j}$. In the case when all the $n_{\cdot j}$ are equal, this implies a spherically symmetric distribution for (α_j); and for unequal $n_{\cdot j}$ a simple distortion of that. It is now straightforward to show that the denominator of the right-hand side of equation (5) is distributed like χ^2 with $l - 1$ degrees of freedom, independently of ρ^2 itself. Hence $\mathcal{E}(\rho^2)$ is the ratio of expectations of numerator and denominator, and is found to be

$$\begin{aligned} \mathcal{E}(\rho^2) &= \left\{ \Sigma_i \Sigma_j \frac{n_{ij}^2}{n_{i\cdot} n_{\cdot j}} - 1 \right\} / (l - 1) \\ &= \{\text{trace } (Q) - 1\}/(l - 1) \ \{\text{trace } (Q) - 1\}/(l - 1) \\ &= \text{the average of the } l - 1 \text{ stationary values of } \rho^2. \end{aligned}$$

All this relates to a fixed design matrix. Suppose now that the entries in the columns of the design matrix are permuted at random. Then each n_{ij} has a hypergeometric distribution, well known in the theory of contingency tables; and the expected value of the above is found to be

$$\mathcal{E}(\rho^2) = \frac{k - 1}{n - 1} \ ,$$

the ratio of the degrees of freedom for X to the total degrees of freedom.

This result can be established directly as the mean value for ρ^2 under permutation of the columns of the design matrix, for a fixed effect of Y, without averaging over a probability distribution for the effect of Y. It is essentially Welch's result quoted at (2) above, that $\mathcal{E}(U) = 0$. For suppose we calculate the statistic U

corresponding to factor X, using not the actual yields but the expected yields (μ_i), which depend on the level of Y. We find, from (1) and (5), that

$$U = 1 - \frac{n-1}{n-k}(1-\rho^2),$$

or

$$(n-1)\rho^2 = (n-k)U + (k-1).$$

From (2) we can immediately obtain an expression for $\text{var}(\rho^2)$, which is nearly but not quite independent of the response to Y if Y is at more than two levels. If $n \gg k$, the χ^2 approximation to the distribution for U can be re-expressed: $(n-1)\rho^2$ has approximately the χ^2 distribution with $(k-1)$ degrees of freedom.