International Statistical Review (2009), 77, 3, 415–429 doi:10.1111/j.1751-5823.2009.00084.x

Randomization in the Design of Experiments

D. R. Cox

Nuffield College, Oxford OX1 1NF, UK
E-mail: david.cox@nuffield.ox.ac.uk

Summary

A general review is given of the role of randomization in experimental design. Three objectives are distinguished, the avoidance of bias, the establishment of a secure base for the estimation of error in traditional designs, and the provision of formally exact tests of significance and confidence limits. The approximate randomization theory associated with analysis of covariance is outlined and conditionality considerations are used to explain the limited role of randomization in experiments with very small numbers of experimental units. The relation between the so-called design-based and model-based analyses is discussed. Corresponding results in sampling theory are mentioned briefly.

Key words: Analysis of covariance; bias elimination; conditionality; estimation of error; exact significance test; history of design; small experiments.

1 Introduction

Ideas about the design of investigations have a long history but a natural starting point for current discussion is the paper of Fisher (Fisher, 1926), leading to his book, still in print (Fisher, 1935). This was soon followed by a remarkable monograph on factorial experiments (Yates, 1937), being the culmination of a period of extremely rapid development. Subsequent major notions came with the emphasis on response surfaces, as contrasted with factorial effects (Box & Wilson, 1951), on randomized clinical trials and currently on the experimental investigation of computer models and on applications to issues of social policy. The theory of optimal design underpins some of these discussions at a more formalized level and provides a basis for exploring non-standard situations.

Fisher based his discussion on four ideas:

- The study of several potential effects simultaneously, the factorial principle.
- Local error control, based on comparing like with like.
- Replication.
- Randomization.

Of these, randomization raises perhaps the most acute conceptual issues. The present paper reviews some of the considerations involved, primarily in the context of experimental design, with a few remarks on randomization in sampling. By randomization is always meant the use of an impersonal procedure with specified probabilistic properties.

2 Purposes of Randomization

There are three broad purposes that may be used to justify randomization:

- to avoid selection and other biases and to do so in a publically convincing way,
- to provide a unified approach to the second-order analysis of many standard designs, and
- to provide a basis for exact tests of significance and related interval estimates.

The first of these is largely uncontroversial. The possibility of bias in treatment allocation, in implementation, and in measurement of outcomes is serious in many contexts in which personal choice is even marginally involved and randomization is often the most effective way of avoiding such biases. Moreover, it does so in a way that is convincing to those who are not directly involved in the investigation. In some contexts, failure to randomize appropriately may fatally compromise an investigation.

Thus, Student (1938), in a posthumous paper in which he argued strongly *against* the use of randomization in the design of agricultural variety trials, emphasized the importance of randomization when there was the possibility of observer's and other errors arising from personal differences or selection effects.

It is sometimes argued that, e.g. the allocation of patients to one of a number of possible treatment regimens may be determined by so many ill-specified considerations that observational data may be analyzed as if treatment allocation were essentially random; in particular, it is implied that actual randomization is not needed. While this may be the case in specific instances, such reliance on effective randomization arising from ill-specified procedures is likely to be hazardous and, particularly, a matter for retrospective concern if unexpected conclusions appear to have been obtained.

A general principle of design is that all parts of an investigation at which uncontrolled variation may enter must be considered. This often implies the need for several stages of randomization. For a detailed exploration of that aspect, see Brien & Bailey (2006). We shall not discuss situations in which a sequence of treatments is applied with later choices depending on earlier outcomes. Then, more subtle forms of bias are possible (Wermuth & Cox, 2008).

The discussion here is for the design of experiments: i.e. whether or not randomization is employed, the assignment of treatments is under the control of the investigator. The key feature is that no aspect of the experimental units should influence both the response variable under study and the allocation of treatments. Alternatives to randomization, namely allocation in systematic order, purposive choice to produce balance, and vaguely haphazard assignment, typically, depending on context, all have the potential to produce relatively subtle bias. Systematic allocation may have the attraction of leading to relatively easily and reliably implemented protocols and also may be enough to preclude some kinds of personal selection errors. The other two possibilities are best avoided.

The situation is quite different in comparable observational studies in which the aspects analogous to treatments are pre-determined in a way that is typically not open to direct investigation. The possibility of serious systematic distortion arising from unobserved features affecting both assignment and outcome can usually be discounted only from subject-matter knowledge of possible assignment processes and general judgement, often combined with a sensitivity analysis. For a thorough discussion, see Rosenbaum (2002).

We shall not discuss bias avoidance further, concentrating on the two more technical arguments for randomization outlined above. The relevance of the discussion for the somewhat parallel issues in sampling theory will be mentioned; the aspects about bias are very similar.

3 Second-Order Theory

3.1 Formulation

The usual explicit formulation, stemming from Neyman (1923), takes, as given, a set of n experimental units, $U_1, \ldots, U_n$ and t treatments $T_1, \ldots, T_t$. Just one treatment is applied to each unit, and it is supposed that if T_s is applied to U_j, the resulting observation is

$$\tau_s + \alpha_j, \tag{1}$$

independently of the allocation of treatments to the other units. This assumption is called *unit-treatment additivity*. We consider the generalizations in Section 5.

The α_j, in effect, encapsulate all aspects of the units as they are before the instant of randomization and that may affect the response to be observed.

The interpretation of a contrast of the τ's, in particular, of say $\tau_u - \tau_v$, is that it concerns the comparison of the outcome on a unit receiving treatment T_u as compared with the outcome that unit would have given if it had received T_v, other things remaining unchanged. Because it is supposed that each unit may be observed under only one treatment, the assumption that the treatment contrasts are constant across units is not directly testable. Some such test is, however, possible if an additional feature, z, say, is identified for each unit. We may then look for a treatment by z interaction.

A central point that will be discussed further below is that there is typically no assumption that the α_j are random variables. Any assumption that the units are, say, a random sample from a population of units or are generated by a stochastic process of a particular form is additional to the specification.

3.2 Outline Analysis

Suppose now that the treatments are allocated to units in accordance with a well-defined randomized procedure. Typically, this means that the units are arranged in a particular structure, e.g. spatially by blocks, rows, and columns, or in time, or by laboratories, sets of apparatus, etc. Then, the treatments are allocated to units at random, subject to balancing constraints. For a penetrating account, see Bailey (2008).

Estimated treatment effects and estimated standard errors are then defined essentially by reference to a linear model in which each constraint on the randomization is represented by a parameter. In the simpler cases, the treatment estimates are based directly on sample means and standard errors derived from a residual mean square, RMS, say.

We outline the details for a randomized block design. Thus, we specify the unit constants by α_{cj} corresponding to the j-th unit in block c. Consider for simplicity designs in which each of the t treatments occurs once in each of the b blocks so that the total number of units is $n = bt$. That is, $c = 1, \ldots, b; j = 1, \ldots, t$. We define the usual estimates and analysis of variance table associated with the randomized block design, e.g. derived from a linear model in which the constraints on the randomization are represented by unknown parameters, defining block effects.

This may appear conceptually as if the randomization analysis does not stand on its own but needs the model-based formulation in the background. This is, however, not really the case. A formal justification of the estimated treatment contrasts based on the sample means in the standard model-based formulation is that, as least-squares estimates, among estimates that are linear functions of the data and are unbiased, they have minimum variance. It is not difficult to show an exactly analogous randomization property: among all linear estimates that are

unbiased in the randomization distribution, the contrasts of means have minimum randomization variance.

Let E_R denote expectation over the randomization distribution. That is, the statistic in question is evaluated for every design in the randomization set and the result averaged.

Then, $\hat{\tau}_u - \hat{\tau}_v$, the difference between the corresponding treatment means, is such that the contribution of the α_{ck} sums to zero over the randomization, i.e.

$$E_R(\hat{\tau}_u - \hat{\tau}_v) = \tau_u - \tau_v. \tag{2}$$

Moreover, the variance of this contrast does not depend on the τ's and is a quadratic function of the α_{ck}, with a high degree of symmetry. In fact, because the randomization distribution is unaffected by reordering the units within a block and by permuting the blocks, we have

$$\begin{aligned} \mathrm{var}_R(\hat{\tau}_u - \hat{\tau}_v) &= A'\Sigma\alpha_{cj}^2 + B'\Sigma\alpha_{cj}\alpha_{dk} + C'\Sigma\alpha_{cj}\alpha_{ck} \\ &= A\Sigma\alpha_{cj}^2 + B\Sigma(\bar{\alpha}_{c.} - \bar{\alpha}_{..})^2 + C\Sigma(\alpha_{cj} - \bar{\alpha}_{c.})^2. \end{aligned} \tag{3}$$

Here, $\bar{\alpha}_{c.} = \Sigma_j \alpha_{cj}/t$, $\bar{\alpha}_{..} = \Sigma_{c,j}\alpha_{cj}/(bt)$, and A, B, and C are constants to be determined. Now, equation (3) is an identity holding for arbitrary $\{\alpha_{cj}\}$. The randomization variance is zero if either (i) $\alpha_{cj} = \bar{\alpha}_{..}$, a constant for all c, j or (ii) more generally $\alpha_{cj} = \bar{\alpha}_{c.}$ so that there is no variation within blocks. It follows from these choices that $A = B = 0$ so that

$$\mathrm{var}_R(\hat{\tau}_u - \hat{\tau}_v) = C\Sigma(\alpha_{cj} - \bar{\alpha}_{c.})^2. \tag{4}$$

An exactly analogous argument shows that for the RMS

$$E_R(\mathrm{RMS}) = D\Sigma(\alpha_{cj} - \bar{\alpha}_{c.})^2. \tag{5}$$

To complete the argument, we suppose that the α_{cj} are independent random variables of zero mean and unit variance and take expectations of the last two equations over their distribution to give, respectively, known infinite model variances and expectation. There results

$$2/b = (b-1)(t-1)C, \qquad 1 = (b-1)(t-1)D. \tag{6}$$

It is important that the supposition that the α_{cj} are random variables is a technical trick used to simplify a calculation, not a statistical assumption about the model. The special argument exploiting the relation between a property of a finite population and one of an infinite population is sometimes called inheritance.

An alternative approach, in some ways more direct but arguably less enlightening, is to introduce an indicator random variable for the treatment applied to U_j, writing $I_{js} = 1$ if T_s is applied and $I_{js} = 0$ otherwise. Then, the observation on U_j can be written as $\Sigma_s I_{js}\tau_s + \alpha_j$. Any specific scheme of randomization specifies the joint distribution of the $\{I_{js}\}$. In particular, the mean and the covariance matrix of the random variables determine the second-order properties of the design and, in the special case of randomized blocks, leads to equations (4) and (5) (Kempthorne, 1952, 1955; Hinkelmann & Kempthorne, 2008).

The essence is that the relation between estimated variance of a treatment contrast and the RMS is that holding in the standard analysis based on a linear model.

The immediate importance of this argument is that the standard analysis of a range of experimental designs can be regarded as a consequence of a single unifying assumption and does not require an *ad-hoc* specification of a new linear model for each design. Although the above argument does not depend on asymptotic considerations, large-sample considerations are required to justify in detail the standard distributional form for pivotal quantities such as the difference between estimated and population contrast divided by its estimated standard error.

It is shown below that approximate versions of the above argument apply to analysis of covariance, i.e. to one of the standard designs augmented by additional dependencies.

The implications of the arguments of this section for analysis other than by the least-squares-based study of linear models do not seem to have been much explored. It is plausible that the linear model implied by the randomization would also be a suitable starting point for other analyses, e.g. those of generalized linear form. For such models based on single-parameter distributions, such as the Poisson, binomial, or exponential, the role of the RMS would be replaced by the residual likelihood ratio (or deviance) statistic, regarded as a test of model adequacy, at least for those problems in which the individual observations are not too far from normality in distribution. Extended versions of such analyses allowing for overdispersion formalized by quasi-likelihood are more directly analogous to the discussion given earlier.

3.3 Analysis of Covariance

Now, consider one of the standard designs covered by the previous discussion and suppose available on each unit a vector $\boldsymbol{z}$ of concomitant variables. These may be of a number of types and serve different purposes. If they are intermediate variables, in the sense that they may be affected by the treatments, then their purpose may be as response variables in their own right or as suggesting pathways between the treatment and its effect on the ultimate response.

Here, we concentrate on variables that cannot have been affected by treatment, ideally because they were recorded before randomization. Their role is two-fold. One is to investigate possible interaction between treatments and $\boldsymbol{z}$; this is, in effect, a test of the basic assumption (1). The discussion here, largely following Cox (1982b), is concentrated on the second and more common role, that of precision improvement.

In the simplest standard linear model formulation and with each treatment randomized to r experimental units, we consider

$$\hat{\tau}_u - \hat{\tau}_v = \bar{y}_{u.} - \bar{y}_{v.} - \hat{\boldsymbol{\beta}}^T_{YZ.\mathrm{res}}(\bar{\boldsymbol{z}}_u - \bar{\boldsymbol{z}}_v). \tag{7}$$

Here, the vector of estimated regression coefficients of Y on $\boldsymbol{z}$, namely $\hat{\boldsymbol{\beta}}_{YZ.\mathrm{res}}$, is calculated residual to any effects such as block effects introduced into the design. It follows that in the standard linear model theory

$$\mathrm{var}(\hat{\tau}_u - \hat{\tau}_v) = \sigma^2\big\{2/r + (\bar{\boldsymbol{z}}_u - \bar{\boldsymbol{z}}_v)^T \boldsymbol{S}^{-1}_{zz.\mathrm{res}}(\bar{\boldsymbol{z}}_u - \bar{\boldsymbol{z}}_v)\big\}, \tag{8}$$

where $\boldsymbol{S}_{zz.\mathrm{res}}$ is the matrix of sums of squares and products of $\boldsymbol{z}$ residual to regression on the design parameters.

A more general form, assuming equal replication of treatments, is that

$$\mathrm{ave}_{u,v}\mathrm{var}(\hat{\tau}_u - \hat{\tau}_v) = 2\sigma^2\big\{1 + \mathrm{tr}\big(\boldsymbol{B}_{zz.\mathrm{res}}\boldsymbol{S}^{-1}_{zz.\mathrm{res}}\big)/(t-1)\big\}/r,$$

where $\boldsymbol{B}$ is the between treatments matrix of sums of squares and products of $\boldsymbol{z}$ analogous to $\boldsymbol{S}$.

Under randomization, the variables z are such that

$$E_R(\boldsymbol{S}_{zz.\mathrm{res}}/d_{\mathrm{res}}) = \boldsymbol{\Omega}_{zz.\mathrm{res}}, \tag{9}$$

$$E_R(\bar{\boldsymbol{z}}_u - \bar{\boldsymbol{z}}_v) = 0. \tag{10}$$

Here, $\boldsymbol{\Omega}_{zz.\mathrm{res}}$ is the finite population covariance matrix of $\boldsymbol{z}$ residual to regression on the design features. By a similar argument

$$E_R\{(\bar{\boldsymbol{z}}_u - \bar{\boldsymbol{z}}_v)(\bar{\boldsymbol{z}}_u - \bar{\boldsymbol{z}}_v)^T\} = 2\boldsymbol{\Omega}_{zz.\mathrm{res}}/r. \tag{11}$$

It follows from this that

$$\mathrm{var}(\hat{\tau}_u - \hat{\tau}_v) = \frac{2\sigma^2}{r}(1 + W_p/n), \tag{12}$$

where W_p has, under randomization, approximately a chi-squared distribution with p degrees of freedom. A rather naive interpretation of this is that before the randomization, W_p should be replaced by its expectation, namely p, so that the variance is inflated by a factor $1 + p/n$, corresponding for small p/n to the effective loss of one unit for each component of $\boldsymbol{z}$ fitted unnecessarily.

The situation is more complicated for two reasons. First, the choice of $\boldsymbol{z}$ to use in the final analysis may be to some extent data-dependent. Also, the variance actually achieved would depend on the configuration of $\boldsymbol{z}$ as realized so that the value of W_p might, if an unfortunate design were implemented, be in the upper tail of the chi-squared distribution. This suggests that a cautious estimate of the potential loss of information should use a reasonable upper percentage point of the chi-squared distribution rather than of the mean.

The discussion so far has used a standard model-based formulation for the primary analysis and randomization theory as a basis to study the impact of $\boldsymbol{z}$. An immediate extension of the previous results for a fully randomization-based analysis is not possible because the regression coefficient used in adjustment is not polynomial in the variables concerned. One can either make the approximation of ignoring variation in the regression coefficient or rely on the joint asymptotic normality of the functions involved.

The interchange between Student (1938) and Fisher about the merits of randomized as contrasted with systematic designs in agricultural variety trials centred essentially on the extent to which randomized designs could achieve balance with respect to plausible patterns of variability. Yates (1939), at the conclusion of the discussion, suggested that, while there might sometimes be small gains in precision to be achieved by systematic arrangements, the lack of security in the basis for error estimation in such designs distracted attention from key issues of interpreting the effects under study. More recent work in a similar vein, stemming from Bartlett (1978) and Wilkinson *et al.* (1983), has been based on explicit time series or spatial models of variability, often leading to the so-called neighbourhood balance designs. Again, however, the reality of any apparent gain in precision depends on the adequacy of the assumed model.

3.4 Conditionality and Its Implications

To be appropriate as a basis for precision improvement, as discussed in the previous subsection, the variable $\boldsymbol{z}$ must refer to the state of the experimental unit before the instant of randomization. In many cases, as is, in principle, desirable, the values of $\boldsymbol{z}$ are available at the time of randomization. Sometimes, however, $\boldsymbol{z}$ may become available only some time later. Because different configurations of $\boldsymbol{z}$ induce different precisions in the estimated treatment contrasts, conditionality considerations become relevant.

In the first case, there would rarely be justification in implementing a design notably unbalanced with respect to $\boldsymbol{z}$. At least, when the randomization set is rich, it is reasonable to condition the randomization and the associated distribution on close to balance, as assessed in the case of two treatments by the squared norm of the difference in the means of $\boldsymbol{z}$ and, in more complicated cases, by the corresponding normalized sum of squares. This could be done by re-randomizing, say, 20 times and choosing the design with most balance and using, in principle, the conditional randomization distribution. Note that if the initial design were completely randomized and the components of $\boldsymbol{z}$ were indicators of a set of blocks, repeated re-randomization in the way just described would lead exactly or nearly to a randomized block design and to its correct analysis.

One alternative procedure would be to aim to minimize to an acceptable level the maximum, over all components of $\boldsymbol{z}$ taken one at a time, of the standardized difference of that component between the treatment arms.

If $\boldsymbol{z}$ does not become available until after the implementation of the design, then any randomization analysis should be conditional on the measure of balance being reasonably close to that actually achieved. To the extent that the joint randomization distribution of relevant linear and quadratic functions of the data is approximately multivariate Gaussian, the standard procedures have a randomization-based justification, but this would collapse if the randomization set were to become very small.

In some situations, it is possible to avoid especially undesirable treatment configurations by the device of restricted randomization (Yates, 1951a; Grundy & Healy, 1951; Bailey & Rowley, 1987). In this, the second-order properties of a standard randomization, e.g. of a Latin square, are preserved by rejecting not only the objectionable squares, for instance, those with strong diagonal structure, but also suitable complementary squares. The mathematical concept underlying this is the double transitivity of the group of permutations underlying the randomization, this being sufficient to preserve the second-moment properties of the randomization.

There is a qualitative difficulty if $\boldsymbol{z}$ is high-dimensional or not well specified. Then, even if all apparently reasonable precautions have been taken to ensure essential balance with respect to important components of $\boldsymbol{z}$, there is the possibility that new features may be suggested as potential sources of bias not properly accommodated by the randomization. In practice, it is probably prudent to aim initially for effective balance with respect to a relatively small set of features chosen as plausibly relevant to the response of interest, but to be prepared to make additional *ad-hoc* checks if suggestions are made at a later stage. The randomized procedure of design and analysis, and indeed any other approach, is vulnerable to overenthusiasm for looking for anomalies.

For a corresponding discussion of the role of conditioning in sampling theory, see Durbin (1969) and Thompson (1997).

4 Exact Randomization Tests

The third and statistically most technical justification of randomization is that it provides a secure basis for an exact test of significance; here, of course, *exact* is a specialized term meaning that there is no approximation involved in finding the distribution of the test statistic. This aspect of randomization tests was emphasized by Fisher (1935), it seems, in part, as a reply to suggestions that some procedures connected with analysis of variance might be unduly sensitive to a normality assumption.

In its simplest form, the argument concerns the strong null hypothesis in the basic model (1) that all the τ_s are equal, implying that the observations on all units are completely unaffected by the treatment allocation received. It follows that the data that would have been obtained in any other arrangement of treatments are precisely known. For any test statistic, the exact distribution under the null hypothesis can thus be found by enumeration. In general, some appeal to either informal considerations or to optimality under a parametric model is needed to choose a test statistic.

Perhaps the simplest form of this argument leads for binary outcomes and two treatments to Fisher's exact test for a 2×2 contingency table (Barnard, 1947). Here, n individuals are randomly allocated, n_1 to receive treatment T_1, and the remaining $n_2 = n - n_1$ to receive treatment T_2. A binary outcome is observed, leading to r_1 and r_2 successes in the respective treatment groups. Under the strong null hypothesis, identical outcomes would be obtained for each individual

unit, whatever be the allocation of treatments, so that in all possible configurations, $r_1 + r_2$ is fixed. The only possible test statistic is r_1 or an equivalent and its null hypothesis distribution is hypergeometric. Note that the conditioning on the two marginal totals in this formulation is a consequence in the case of one margin of the design and in the case of the other margin of the null hypothesis. In other versions of the 2×2 table, the conditioning is primarily a technical device of frequentist theory to remove dependence on nuisance parameters.

More generally, for binary outcomes, the exact test for the equality of two treatments, adjusting for concomitant variables by logistic regression, can be regarded as a randomization test taken conditionally on the treatment means of the concomitant variables. This is implicit, but not explicit, in the discussion of logistic regression by Cox (1958a), who related the test to sampling a finite population randomly without replacement. If, as would often be the case, the conditional randomization is exactly or nearly degenerate, then exact conditioning has to be replaced by approximate conditioning. It is an open question as to precisely how that should be done.

Freedman (2008) has discussed the non-null randomization theory of logistic regression, emphasizing that the standard logistic regression coefficient does not have a direct interpretation in terms of the following extension of equation (1). For each experimental unit, there are two notional binary responses, one under T_2 and one under T_1. This defines an average finite population mean response for each treatment, an average probability, in fact; an associated logistic parameter is then the difference of the log odds of the two means. Freedman, having shown that this is not estimated by the standard logistic regression coefficient on treatment, gives an estimate obtained by first averaging estimated probabilities before calculating log odds.

The additive structure of equation (1) is not directly appropriate for discrete data except possibly in a large-frequency approximation. Copas (1973) obtained one-sided confidence limits by supposing that some unknown number Δ of the units that had outcome zero when tested under T_1 would have had outcome one had they been tested under T_2 and that otherwise treatments have no effect. In principle, consistency with any assigned non-negative integer value of Δ can be tested by reconstructing the data for alternative treatment allocations and hence a confidence interval can be found from all those values, consistent with the data at a pre-assigned significance level.

For randomization versions of standard normal-theory tests, substantial simplification of the test statistic is often possible and reasonably accurate approximations to its distribution are obtained. Thus, for the matched pair problem, let (y_{1j}, y_{2j}) be the observations on the j-th pair and let $d_j = y_{j2} - y_{j1}$. Then, under the null hypothesis of no treatment difference, the randomization distribution assigns equal probability 2^{-m} to the m sample values $\pm d_j$. Because the sum of squares of the notional values is constant for all possible samples, the standard Student t-statistic is a monotonic function of the sample total. The test statistic Σd_j has thus the distribution of

$$\Sigma I_j d_j, \tag{13}$$

where the I_j are independent and identically distributed, with the distribution putting equal probability $1/2$ on ± 1. Exact distributional calculations, an approximation based on the central limit theorem, or on a refinement thereof, are thus available.

This test is numerically identical to, but conceptually different from, the nonparametric permutation test of the hypothesis that the observations in any pair are independent and identically distributed, with an arbitrary distribution function possibly different from pair to pair.

A detailed investigation of the first four moments of standard statistics associated with randomized blocks and Latin squares was reported by Pitman (1937) and Welch (1937). They fitted Pearson's curves to the resulting moments, typically leading to adjusted forms of the standard t and F distributions.

For the standard linear model with possibly non-normal errors, the infinite model distribution of a test statistic is the superposition of randomization distributions, suggesting that the distinctions between the randomization and the infinite model distributional results are a small-sample effect.

Much recent discussion, following Yates's (1951b) criticism of Fisher's overemphasis on significance tests, has stressed estimation as opposed to testing. Here, the distinctions between the alternative distributional procedures are often less critical. From a computational point of view, the direct use of randomization distributions is now-a-days much easier than was previously, so interest in simple approximations is primarily for general enlightenment rather than for immediate use.

The analogous confidence interval argument for continuous data plausibly represented by equation (1) is direct. Any assigned values hypothesized for a contrast of the τ_u's can be regarded as a null hypothesis for revised data in which appropriate constants have been subtracted from the observations and consistency with that null hypothesis has been tested.

In some situations, there is a choice of the detailed scheme of randomization to be employed. For example, in the randomization of a Latin square, second-order properties are achieved by taking a fixed square and permuting both rows and columns at random. Further possibilities are to permute also the naming of treatments or to choose from the set of all Latin squares of the order in question. It is plausible but, so far as I know, unproved that the randomization distribution induced becomes closer to a standard continuous approximation the richer is the randomization set.

5 Generalizations of the Key Model

The key assumption of unit-treatment additivity, as formulated in equation (1), is, in principle, deterministic and as such both very strong and, without further information, essentially untestable. As a special case, the strong null hypothesis asserts the total irrelevance of the treatment allocation to the outcomes. Fisher's original discussion put much emphasis on such null hypotheses, but the more general formulation is needed to provide a basis for the estimation of contrasts.

The formulation (1) may be modified, either to deal with special complications or to make it more realistic. The possibilities include:

- A random term may be added representing sampling, measurement, or in general, technical error in determining the response. If these random terms are independent and identically distributed random variables, their inclusion has no direct effect on the above conclusions. If, however, the magnitude of the variance of the technical error can be estimated, either from external considerations or from internal replication, the information implied about the variance of the α_j may be useful.
- The possibility of a carry-over or residual effect from one experimental unit to another may be represented by a second set of treatment parameters. It is known that randomization theory cannot be used to justify the standard analysis of the balanced Latin squares commonly used in such experiments, e.g. in psychology, but the practical implications, if any, of this failure do not seem to have been studied in depth.
- The possibility of a hierarchy of experimental units may be represented, as is appropriate in, for instance, a split-unit design.
- If there is a vector of baseline variables $\boldsymbol{z}_j$ available for each unit, it may be helpful to decompose the α_j in the form $\alpha_j^* + \boldsymbol{\beta}^T \boldsymbol{z}_j$, leading to the analysis discussed in Section 3.3.

- In some ways, more interestingly, the treatment parameters applied to U_j may be modified to $\tau_s(1 + \boldsymbol{\gamma}^T \boldsymbol{z}_j)$. This represents a treatment by unit interaction of a simple form in which any treatment contrast is multiplied by the factor $1 + \boldsymbol{\gamma}^T \boldsymbol{z}_j$ when realized implicitly on U_j. An exact randomization test of the hypothesis $\boldsymbol{\gamma} = 0$ is possible.
- The representation (1) may be modified to deal with discrete, in particular, binary responses, as outlined in Section 4.

A formally fairly general specification is for the application of treatment s to unit j to introduce a correction term η_{js} to τ_s, producing a response

$$\tau_s + \alpha_j + \eta_{js} + \epsilon_{js}, \tag{14}$$

where the η sum to zero in a sense yet to be defined, and the ϵ_{js} are technical errors in the sense mentioned above. This formulation implies that, averaged over the technical errors, the "true" difference between the responses to treatments u and v, averaged over the units in the study, is

$$\tau_u - \tau_v + \text{ave}_j(\eta_{ju} - \eta_{jv}). \tag{15}$$

In some more recent studies, this is called the average causal effect.

Neyman (1935) introduced essentially this formulation with the η regarded as unknown constants defined to sum to zero over the units in the experiment. That is, $\tau_u - \tau_v$ is an average treatment effect over the units in question. This led to the suggestion that the estimate of error in the Latin square design is biased, a conclusion vigorously disputed by Fisher in the discussion of Neyman's paper. Wilk & Kempthorne (1957) confirmed Neyman's conclusion; Cox (1958b) suggested that, in particular, the null hypothesis of exact balance of treatment effects over a finite population of units is artificial and that a more reasonable formulation restores the unbiased character of the estimate of error variance. This property would follow directly if one assumed that the η's are uncorrelated random variables of zero mean and constant variance and averaged over their probability distribution.

6 Discussion

The choice between model-based and design-based approaches to inference is usually not critical numerically. Indeed that broadly the same numerical conclusions can be obtained by two routes is a source of strength. The main numerical difference is likely to be when the set over which the randomization distribution is defined is not very large. There is then, e.g. a limit on the level of significance that can be achieved, no matter how large is the numerical difference between, say, two groups. This, in effect, warns against overinterpretation of the security with which an apparently clear-cut effect can be established from small amounts of data, at least without external knowledge of the amount of random variability that is likely to be involved.

A conceptually more important point concerns the nature and generality of the conclusions reached. Consider for simplicity an experiment involving just two treatments. The target of inference on the basis of the formulation (1) is the difference between the mean response achieved if, under the conditions of the experiment, treatment T_2 were to be applied to all units and the corresponding mean achieved if treatment T_1 were to be applied to all units. Tests and confidence limits for this difference are obtained. If we took equation (1) literally, we would reach a conclusion for each unit individually but this would be to rely on untestable and often intrinsically implausible aspects of the model. Indeed, the conclusions derived from statistical comparisons of groups can be applied with confidence to specific individuals only when there is substantial external information.

The conclusion concerns the set of units employed in the study. As such, it is typically of limited interest; one wants to draw a conclusion that applies more generally. To do this in the current setting, we have to rely on a general uniformity principle; conclusions firmly established in the past will apply in future, unless there is a material change in relevant conditions. In some contexts, this may be all that needs to be and can be said; so far as I know, the early workers on the fruit fly *Drosophila* were not explicitly concerned with how broadly their conclusions might be applicable to other species. If the breadth of applicability is important, then the route to this is to establish the absence of interaction with important additional features and the inclusion of factors in a factorial experiment to do just this is part of standard recommendations for experimental design; see, e.g. Yates & Cochran (1938) and Cox (1958c, pp. 9–11). It can be argued that this, and preferably an understanding of underlying process, is the basis for a reasoned extrapolation of conclusions, and indeed of applying conclusions at some more individual level, not some notion of random sampling.

In a typical model-based formulation, the unit constants in equation (1) are regarded as realizations of random variables. In simple situations, these have a systematic structure representing block and other effects plus independent and identically distributed random components. These, in effect, regard the units as a random sample from a hypothetical infinite population or, in rather exceptional cases, from an existing target population. Thus, the model implies that the conclusions of a statistical analysis can be generalized. In general, however, they are not generalizable to a specific situation, such as, e.g. from the set of patients giving informed consent in a clinical trial to the target population of all relevant patients. That generalization would be possible only to the extent that the trial units are actually or effectively a random sample from the target population and this would typically be an implausible assumption. It does seem that, at least in many situations, the additional *specific* generalizability implied by the model-based approach is illusory. From this perspective, the conceptual argument in favour of the model-based approach is different. It is that the generalization is across different patterns of random variation that might plausibly arise. The dependence on the accidents of the particular random errors actually encountered is to some extent eliminated. This is useful but is more limited than generalization to a specific new situation.

Thus, Zheng & Zelen (2008) have argued that an analysis based essentially on equation (1) is appropriate for clinical trials, precisely because the patients taking part in such trials are not a random sample from the target population of patients. The same point could be made about agricultural field trials and indeed in many other contexts. The conditions at agricultural experimental stations are unlikely to be representative of standard farming practice. As noted above, this points to replication under a range of soil types, farming practices, and meteorological conditions (Yates & Cochran, 1938).

In investigations with a very long-term objective, it may be best to choose experimental units likely to show the phenomenon under study in as striking a form as possible so that it can be examined in depth. When the objective is more short term, e.g. about immediate issues of public policy, representativeness may be of more direct concern. Obviously, the relative time scales of study design, study execution and analysis, and decision making are important.

7 Randomization in Sequence

So far, it has been assumed that randomization takes place in one step or perhaps in a series of steps, one for each source of error or phase of investigation to be randomized. If, say, in a randomized block design, randomization is done one block at a time, no new considerations are involved unless the experience of early stages of the investigation leads to reconsideration of the

whole design. We exclude situations in which the outcomes of early stages of an investigation influence whether and how to continue.

When randomization is implemented in stages, a key issue is whether the allocation at previous stages is known and whether it might influence action at the current stage. For example, randomizing a unit at a time in a randomized block design would mean that at the end of each block, the next allocation would be known and the possibility of a selection bias would be thereby enhanced. A compromise is necessary between the requirements of balance to enhance precision and of bias avoidance. Atkinson (2002) has made a careful comparison of a number of procedures from this point of view.

Efron (1971) suggested achieving balance with high probability by the so-called biased coin design. In the simplest version of this for two treatments and no concomitant variables, the probability that the next unit received, say, the first treatment, is $1/2 \pm a$, where say $a = 1/4$, the choice of sign being made to tend to restore balance in the allocation. More generally, the choice is made on the basis of a Markov chain defined by the current state of treatment balance (Smith, 1984). Cox (1982a) considered some aspects of the randomization-based analysis of such a design. A linear model suitable for such a design is not entirely clear in that time trends in response are partly eliminated from error and therefore should be represented in any model to avoid overestimating the error variance.

8 Relation with Theory of Sampling

Broadly analogous issues arise in sampling contexts, i.e. the design and analysis of studies in which the target is some property, often a mean or total, of an explicit population of individuals, in principle, all of whom could be measured. Here, the initial objective is descriptive, and the issue is essentially equivalent to the estimation of the values associated with elements of the population that are not observed.

The requirements of freedom from systematic error and of the ability to estimate a relevant standard error are much as before; the issue of exact tests of significance will typically not arise, at least not in the descriptive as contrasted with the analytical parts of interpretation. Interestingly, there has been more recent explicit discussion in the sampling literature of the interplay and choice between model- and design-based approaches than in the context of experimental design. This is probably largely because sampling may involve unequal probabilities of selection and hence the need to use weighted estimates, such as the Horvitz–Thompson estimate. For a recent discussion and new developments, see Beaumont (2008).

In the model-based sampling theory, the individual outcomes are assumed to be random variables, i.e. in other terminology, to be samples from a superpopulation. The properties of the superpopulation are inferred by standard inferential methods and these are used to impute the missing values in the population of interest. The superpopulation serves as a technical device and is usually of no intrinsic interest. It might be argued that in the design of experiments, the concept corresponding to the superpopulation is potentially relevant for interpretation as representing the situations to which the conclusions can be generalized. This suggests that the design-based analysis may be relatively more appropriate for sampling than for the design of experiments.

9 Overall Assessment

As noted in Section 1, randomization plays a key role, sometimes indeed an absolutely essential role, in situations in which subjective biases may enter. Clinical trials are the most

obvious example, although there are many others, including the avoidance of bias in sampling of existing populations. Randomization provides a unifying thread to the least-squares analysis of a large family of balanced designs, avoiding the notion that the choice of an appropriate linear model is an *ad-hoc* matter in each case and establishing the important principle that, in a sense, the analysis of variance table, i.e. the structure of the design and the resulting data, comes first and the associated linear model comes second. In that sense, there are at least qualitative implications for analysis of such designs by methods other than the least-squares method. Randomization provides an elegant route to exact tests of significance, appealing perhaps particularly to those wishing a cautious approach to the analysis in question.

At least for the last two purposes, randomization is ineffective in very small investigations because the individual arrangements over which randomization takes place have recognizably relevant distinct features so that an implicit conditionality principle would be violated by an implicit or explicit reference to a randomization distribution. In larger studies, there is an implied appeal to a notion of what may be called searching critical abstinence.

Randomization seems to play little role in the theory of optimal design, presumably because such formal optimality arguments presuppose a strong specification, whereas randomization is concerned essentially with justifying such specifications. Early work on the personalistic Bayesian approach to statistics largely denied any role for randomization except perhaps for making a choice between equally appealing possible decisions. The personalistic approach might seem to point towards purposive selection rather than towards randomization, but the empirical evidence of the dangers of purposive selection is strong and wide-ranging. In many ways, a more fundamental point is that the pioneers in personalistic probability emphasized its intimate connection with *personal* decision making. Randomization plays an important role in game theory, but this is relevant to statistics only under the severe abstraction that analysis of data can be viewed as a game against nature (Wald, 1950).

Now, while the objective in designing and analyzing an experiment is likely, in part, to be to inform the personal views of the investigator, this is not the main purpose of scientific and technological reporting, which is largely to provide secure knowledge of the field in question, in principle, convincing to a broad group. For this, the personalistic Bayesian view of analysis is inappropriate; this is not at all the same as dismissing Bayesian numerical methods. Yet, the choice of a design is a decision problem for the investigators and so one would expect the general principles of Bayesian decision theory to apply, at least qualitatively. Fisher, in effect, resolved this dilemma by allowing personal choice, vague background knowledge, etc., to influence the design, e.g. through the choice of a randomized block in preference to a completely randomized design and by specifying the precise definition of the blocks. Randomization then ensured that even if the design was inappropriate, the validity of the conclusions was not compromised, although, of course, there would be loss of efficiency from poor choice.

In this sense, randomization is a cornerstone of the notion of the individually secure investigation, whether it is an experiment or a sample survey. Of course, in many fields, progress stems largely from the imaginative synthesis of information from different sources. In that context, while the conclusions from a single study are not often crucial, and the results of any one study are not decisive, nevertheless, a poorly designed and misleading study may be very disruptive and the security of individual studies remains desirable even in the broader context.

The role of randomization in different subject-matter fields depends on the relative importance of the aspects mentioned above. Thus, in clinical trials, randomization is typically highly desirable. In those laboratory-based studies in which relatively subtle effects close to the threshold of detectability may be under investigation, randomization may be needed to avoid systematic measurement errors. Provided that obviously biased arrangements are avoided in

agricultural field trials, randomization is often valuable but perhaps rarely crucial, and in response surface designs, in which the emphasis is strongly on complex treatment structure and on the use of small numbers of experimental runs in situations with relatively small error, randomization may, in general, be much less important or even undesirable.

Acknowledgements

This paper is a much revised version of a talk given at the Isaac Newton Research Centre, Cambridge, United Kingdom, in August 2008. I am grateful to Krista Gile and a referee for helpful comments.

References

Atkinson, A.C. (2002). The comparison of designs for sequential clinical trials with covariate information. *J. Roy. Stat. Soc. A*, **165**, 349–373.
Bailey, R.A. (2008). *Design of Comparative Experiments*. Cambridge: Cambridge University Press.
Bailey, R.A. & Rowley, C.A. (1987). Valid randomization. *Proc. Roy. Soc. London A*, **410**, 105–124.
Barnard, G.A. (1947). Significance tests for 2 × 2 tables. *Biometrika*, **34**, 122–138.
Bartlett, M.S. (1978). Nearest neighbour models in the analysis of field experiments (with discussion). *J. Roy. Stat. Soc. B*, **40**, 147–174.
Beaumont, J.-F. (2008). A new approach to weighting and inference in sample surveys. *Biometrika*, **95**, 539–553.
Box, G.E.P. & Wilson, K.B. (1951). On the experimental attainment of optimum conditions (with discussion). *J. Roy. Stat. Soc. B*, **13**, 1–45.
Brien, C.J. & Bailey, R.A. (2006). Multiple randomizations. *J. Roy. Stat. Soc. B*, **68**, 571–609.
Copas, J.B. (1973). Randomization models for the matched and unmatched 2 × 2 tables. *Biometrika*, **60**, 467–476.
Cox, D.R. (1958a). Regression analysis of binary sequences (with discussion). *J. Roy. Stat. Soc. B*, **20**, 215–232.
Cox, D.R. (1958b). The interpretation of the effects of non-additivity in the Latin square. *Biometrika*, **45**, 69–73.
Cox, D.R. (1958c). *Planning of Experiments*. New York: Wiley.
Cox, D.R. (1982a). A remark on randomization in clinical trials. *Util. Math.*, **21A**, 245–252.
Cox, D.R. (1982b). Randomization and concomitant variables in the design of experiments. In *Statistics and Probability: Essays in Honor of C.R. Rao,* Eds. G. Kallianpur, P.R. Krishnaiah & J.K. Ghosh, pp. 197–202. Amsterdam, The Netherlands: North Holland.
Durbin, J. (1969). Inferential aspects of the randomness of sample size in survey sampling. In *New Developments in Survey Sampling*, Eds. N.L. Johnson & H.V. Smith, pp. 629–651. New York: Wiley.
Efron, B. (1971). Forcing a sequential experiment to be balanced. *Biometrika*, **58**, 403–417.
Fisher, R.A. (1926). The arrangement of field experiments. *J. Minist. Agric.*, **33**, 503–513.
Fisher, R.A. (1935). *Design of Experiments*. Edinburgh, United Kingdom: Oliver and Boyd. Current 5th ed., Oxford University Press.
Freedman, D. (2008). Randomization does not justify logistic regression. *Statist. Sci.*, **23**, 237–249.
Grundy, P.M. & Healy, M.J.R. (1951). Restricted randomization and quasi-Latin squares. *J. Roy. Stat. Soc. B*, **12**, 286–291.
Hinkelmann, K. & Kempthorne, O. (2008). *Design and Analysis of Experiments*, Vol. 1. New York: Wiley.
Kempthorne, O. (1952). *Design and Analysis of Experiments*. New York: Wiley.
Kempthorne, O. (1955). The randomization theory of experimental inference. *J. Amer. Statist. Assoc.*, **50**, 946–967.
Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. English translation from Polish of Chapter 9 by D.M. Dobrowska and T.P. Speed (1990). *Statist. Sci.*, **9**, 465–480.
Neyman, J. (1935). Statistical problems in agricultural experimentation (with discussion). *Suppl. J. Roy. Stat. Soc.*, **2**, 107–180.
Pitman, E.J.G. (1937). Significance tests that may be applied to samples from any populations. III. The analysis of variance test. *Biometrika*, **29**, 322–335.
Rosenbaum, P.R. (2002). *Observational Studies*, 2nd ed. New York: Springer.
Smith, R.L. (1984). Properties of biased coin designs in sequential clinical trials. *Ann. Statist.*, **12**, 1018–1034.
Student (1938). Comparison between balanced and randomized arrangements of field plots. *Biometrika*, **29**, 363–379.
Thompson, M.E. (1997). *Theory of Sampling*. London: Chapman and Hall.
Wald, A. (1950). *Statistical Decision Functions*. New York: Wiley.

Welch, B.L. (1937). On the *z* test in randomized blocks and Latin squares. *Biometrika*, **29**, 21–52.
Wermuth, N. & Cox, D.R. (2008). Distortion of effects caused by indirect confounding. *Biometrika*, **95**, 17–33.
Wilk, M.B. & Kempthorne, O. (1957). Non-additivities in a Latin square design. *J. Amer. Statist. Assoc.*, **52**, 218–236.
Wilkinson, G.N., Eckert, S.R., Hancock, T.W. & Mayo, O. (1983). Nearest neighbour (NN) analysis of field experiments (with discussion). *J. Roy. Stat. Soc. B*, **45**, 157–211.
Yates, F. (1937). *Design and Analysis of Factorial Experiments*. Technical Communication 35. Harpenden, United Kingdom: Imperial Bureau of Soil Science.
Yates, F. (1939). The comparative advantages of systematic and randomized arrangements in the design of agricultural and biological experiments. *Biometrika*, **30**, 440–466.
Yates, F. (1951a). Quelques developpements modernes dans la planification des experiences. *Ann. Inst. H. Poincare*, **12**, 113–130.
Yates, F. (1951b). The influence of *Statistical Methods for Research Workers* on the development of the science of statistics. *J. Amer. Statist. Assoc.*, **46**, 19–34.
Yates, F. & Cochran, W.G. (1938). The analysis of groups of experiments. *J. Agric. Sci.*, **28**, 556–580.
Zheng, L. & Zelen, M. (2008). Multi-center clinical trials: Randomization and ancillary statistics. *Ann. Appl. Stat.*, **2**, 582–600.

Résumé

On passe en revue le rôle du traitement aléatoire dans la conception d'expériences. On distingue trois objectifs, la prévention de biais, la constitution d'une base solide pour l'estimation d'erreur dans les conceptions traditionnelles et la fourniture de tests formellement exacts de signification et de limites de confiance. La théorie du traitement aléatoire approximatif associé à l'analyse de covariance est présentée et des considérations de conditionnalité sont utilisées pour expliquer le rôle limité du traitement aléatoire dans les expériences avec de tròs petits nombres d'unités expérimentales. La relation entre les analyses dites à base de conception et à base de modòle est discutée. Les résultats correspondants dans la théorie des sondages sont briòvement mentionnés.

Mots clés: analyse de covariance, élimination de biais, conditionnalité, estimation d'erreur, test de signification exact, historique de la conception; petites expérimentations.

[*Received January 2009, accepted March 2009*]