INTERVENTION EXPERIMENTS, RANDOMIZATION AND INFERENCE

Oscar Kempthorne, Department of Statistics, Iowa
State University, Ames, Iowa

Abstract

This essay gives a discussion of processes of design and analysis of a study of the effect of two or more interventions or treatments on a set of experimental material (e.g., an agricultural area, or a set of mice, or a human). The problems of design, which includes, critically, the plan by which treatments are conjoined to experimental units, and of analysis are discussed. The author suggests that everything be based on randomization, both design and analysis by randomization tests and inversion thereof. The problem that usual conventional randomization gives bad plans is discussed and suggestion made to overcome it. Parametric models are not used, so defects in conventional parametric inference do not arise. Discussion is given on subjectivity and objectivity.

Introduction

The term *experiment* is commonly interpreted to mean a variety of activities. It can mean nothing more than observation of a piece of space-time; e.g., observing the moon by sending a moon shot. It can mean making a piece of material and measuring attributes of this piece. It can mean doing a study to attempt to determine the effects of a treatment protocol on a disease in humans. It is not entirely unusual to refer to a study estimating an attribute of a defined population such as the human population of the United States as an experiment, though most statisticians would say that such a study is a survey. Then we have the writings of theoretical statisticians that an experiment is a triple $(X, A, P(\theta))$ where X is a sample space, A is an algebra of subsets of X and $P(\theta)$ is a set of probability measures indexed by a parameter θ.

I have taken the position that there is a case for distinguishing three types of *experiment* with associated types of inference that I named sampling, observation and experimental (Kempthorne, 1979).

In the sampling problem, there is a real existent population, say, the totality of human beings of the United States. Each individual has unambiguously defined attributes, such as age, height, weight, amount of education and so on. The problem is very simple to state and to understand; namely, what is the frequency distribution of an attribute in this real population?

Preprint Number 88-30, Statistical Laboratory, ISU, November 1988, Department of Statistics, Iowa State University, Ames, Iowa 50011. Journal paper no. J-13287 of the Iowa Agriculture and Home Economics Experimental Station, Ames, Iowa, Project 890.

It is easy to imagine having a huge army of enumerators – measurers, so that every human is located, enumerated and measured. The inference problem in this case is also obvious: as a simple example, there is a population of ages, and this population has a mean. An inference problem is then to obtain data and then to make useful statements about the unknown mean.

In the observation problem, we observe a whole population, but we hope and wish that this population that we observe is *representative of* a much larger population. Our explorers on the moon observed a portion of the surface of the moon over a very brief period (hours, I imagine), but the hope is that the observations are more or less typical of what would be observed over an extensive time. Similarly, we hope that our observations of planet Earth relate to its status over a significant period, e.g., years, decades, or centuries, etc. We are currently concerned about the ozone layer and wonder what its status will be in, say, 20 ot 50 years. Obviously, to speculate about this, we must have observations at a few times and validated dynamic model of how the status changes. So, then, in the observation problem, we must have a model that represents what we hypothesize about the unobserved world, unobserved because it is in the past or in the future, or at present and not looked at.

In the present essay, I wish to address solely the third class of problem, which is easily exemplified. Let me give some examples. Atherosclerosis of the heart is a common enough problem: rather worrying, I am sure, and I know. How should this be treated? There are treatments by drugs, by diet, etc., and there is one treatment that is rather *heavy* – heart bypass surgery. There is then an obvious question. Is it a good idea to treat the sick person with bypass surgery? Other *heavy* questions arise with the disease of cancer in humans. What treatments are effective, which treatments are better than other treatments? The nature of situations of this sort is that we have a problem developing under its own dynamic, and the question is of what intervention will help.

The Intervention Experiment

A rather generally accepted, and, I imagine, not to be challenged, partial model is that we have materiel and a set of interventions. The partial design of the experiment is to partition the experimental materiel into *pieces* and then place one of the interventions on each piece of materiel. The branch of statistics called the design of experiments was started by R. A. Fisher at the Rothamsted Agricultural Experiment Station. The materiel was agricultural land, planted with certain crops such as wheat, or mangolds, or grass, etc., which was partitioned into pieces called plots, and the treatments were various agricultural interventions such as nutritional supplements. An example that seems superficially quite different is a psychological experiment in which the materiel is part of the life of a human subject for example, the 6 days of a week, and the pieces are human-days. The treatments could be various drug regimes. The aim of the experiment might be to palliate depression, for instance.

The performance of the experiment consists of the following steps:

(i) defining the problem, which will consist of specifying the experimental material and specifying the interventions (treatments) that are to be compared;

(ii) *dividing* the experimental material into *plots*, each of which is to receive a treatment;

(iii) deciding how to conjoin the set of *plots* and the set of treatments, taking into account the totally obvious fact that a *plot* can *receive* only one of the treatments;

(iv) letting the experiment proceed to the prechosen termination point; e.g., the point of harvest of an agricultural crop, or recovery or judged failure of a medical treatment;

(v) taking measurements that are thought to be relevant to the problem;

(vi) *analyzing* the resultant data: I put the word analyzing in quotation marks because this is by no means a well-defined operation; and the *drawing of conclusions*, with the same obscurity;

(vii) discussing usefully how the conclusions can be extended to what is often called the *target population.*

The "Design" of the Experiment

It is commonplace among statisticians who actually work with real investigators (not individuals who only write about the design of experiments) to consider *all three* of steps (i), (ii) and (iii) as critical components of the design of the comparative intervention experiment. Both adjectives comparative and intervention are essential.

It is useful, I think, to mention for comparison, the type of study in which the outcome is *thought* or *modelled* to be a realization of a random variable, X say, which is distributed according to a distribution determined by some control variables, say z, and indexed by some parameter θ, where z and θ may be vectors. Such a study is purely mathematical.

It is rather obvious, at least by hindsight, that a natural field for thinking about the comparative intervention experiment is farm agriculture or garden agriculture. Suburbia consists mostly of houses on individual lots with associated grassed areas – commonly called lawns. Almost all suburbanites experience problems with their lawns. The grass is thin, is dying or has died. What should be done to obtain a lawn that is *good looking*? What interventions should be made? In trying to teach the design of experiments I have often used

this problem as an example. It is not at all surprising that the formulation of a set of procedures for the experiment was done at the Rothamsted Agricultural Experiment Station. The beginning of *experimental agriculture* was made by Lawes and Gilbert in, say, 1843. The most famous Rothamsted experiment is, surely, the Broadbalk field experiment on wheat which was started in 1852 and has continued to present time. The field, Broadbalk, was divided into 13 plots for different nutritional treatments. The yields of wheat were analyzed in a certain way by Fisher (1921). Later Fisher (1924) gave a data analysis of the yields (or years 1852 to 1918) attempting to determine the influence of rainfall on yield.

The use of intervention studies obviously goes back for centuries or millennia – humans found that eating certain plants was harmful or even fatal. It was only in this century that a partial logic was developed.

That the design and analysis of intervention experiments did not originate in connection with human nutrition or human medical problems is not surprising, perhaps, because the comparative intervention experiment requires conjoining one of several treatments to each experimental unit, e.g., human. There were obviously no ethical problems in treating a plot of land with one of several treatments.

There was the recognition that there was variability between experimental units that received the same treatment, and it was obvious that this variability was not the result of measurement error. The existence of such variability was exhibited completely by the various uniformity trials that were conducted, after agricultural scientists recognized that there were problems of design and of analysis.

The Field Plot Experiment

Suppose that our initial problem is that of Lawes and Gilbert in 1843. We wish to determine the effectiveness of several nutritional treatments for wheat. We realize that the yield of wheat grown under the same regime varies over England. Obviously, the yields at Rothamsted will not be the same as the yields in Cornwall or even on a farm 5 miles from Rothamsted. We are able to perform the experiment at Rothamsted and have the field Broadbalk to use. Then, obviously, we can hope only to determine somewhat the effectiveness of the treatments on Broadbalk field of Rothamsted. We realize that we can only, at best, determine the differences among treatments as measured on Broadbalk field in year, say, 1852. Suppose that we can determine these differences exactly. Then to apply the results to what will happen elsewhere and in different years (e.g., 1990), the only process we can use is to assume that the treatment differences will be the same or that the differences are related to some variables that are known for the other circumstances.

This thinking leads me to a view of the fundamental problem of what we might (but should not necessarily) call experimental inference. I state this in very simple form:

> We have a collection, a set, of experimental material. We have a set of interventions or treatments. Our task is to form judgments on the effects of the treatments on this collection of material.

The extension of conclusions to some larger set of material is a problem I shall not address. I merely make the comment that making the assumption that the material used in the experiment that is performed is a random sample from some large population of material is unjustifiable, though perhaps the only way to make even a guess.

I shall discuss agronomic field experiments later, but I first wish to consider what I call experimentation on a line.

Experimentation "On A Line"

Suppose we have an oil processing plant with an inflowing pipeline of feed stock. We wish to examine the differential effects of some treatment processes; e.g., the use of different catalysts. Then our procedure will be to take time slugs of the input and treat each slug with one or other of the treatments. We shall use time slugs that are separated by intervals necessary to make the alterations in the processing and to allow the processing to reach equilibrium status under each given treatment.

As a result of such considerations we shall have experiment time slugs that can be indexed by 1, 2,..., the integers. Suppose now that we have 4 treatments, say A, B, C and D, and we have decided to use 20 successive time slugs. Then the question must be faced of how we are to assign A, B, C, D to the slugs. An obvious suggestion is to use the sequence $ABCDABCD...$ but only a fool would do this. Why do I say this? There will be undoubtedly a *time trend* in the nature of the feed stock and one would expect there to be variation around the time trend. I put the words *time trend* in bold because I find it difficult to find another term. One would expect that if one made a uniformity trial, thereby using only one treatment – say A, that the difference squared between observations on different time slugs would depend on the *distance* between the time slugs. In the particular example I am using, the uniformity trial will have been given by preexperiment records.

There would be no computational difficulty with any treatment assignment in using a linear model,

$$y_2 = \tau_{(i)} + e_i,$$

where $\tau_{(i)}$ is the effect of treatment in slug i and e_i is the error, and then to *assume* that the set $\{e_i\}$ is a realization of 20 independent random Gaussian variables that have mean 0 and variance σ^2 (unknown). From even an elementary first course in statistics one can set this up as a Gauss Markov Normal Linear (GMNL) model, do the ANOVA, make the usual tests of significance, set up the usual *confidence* intervals, etc.

The experimental scientist with even minuscule understanding of variability should object to the plan – the treatment assignment above and the ensuing analysis as given by the usual elementary procedures in the attempt statement of precision of estimation of the differences between treatments – for the simple reason that treatments A and B are contiguous, treatments A and C occur at *points* that are apart by 2 units, and A and D are contiguous half the time and apart by 3 units the other half. So one would expect the difference between treatments A and B to have lower variance than that between A and C.

What then should be done? It is a standard cliche of *the design of experiments* that one has to contemplate analysis to evaluate designs. It is less standard (and even not accepted by some) that the proper analysis (*if there is one*, and this is by no means sure) is determined to a considerable extent by the design.

Suppose that one has used the treatment assignment stated above; i.e., *ABCDABCD...ABCD*. At the end of the experiment, one has observations y_1, $y_2, \ldots, y_{20}$. How should one "analyze" the data? I imagine that 10 statisticians would produce perhaps 5 different analyses. There is the *obvious* one mentioned above. A second one would be to note that the whole sequence is made up of 5 *blocks* each containing the 4 treatments A, B, C and D. Then to compound the naiveté, the statistician could say that he is doing a randomized block analysis, though this can reasonably be characterized only as a *block analysis*. But why do this? Such an analysis ignores almost completely that the units are *on a line*.

Why not consider the model

$$y_i = \beta_0 + \beta_1 i + \tau_{(i)} + e_i,$$

or

$$y_i = f_i + \tau_{(i)} + e_i,$$

where f_i is some function of i (e.g., a quadratic or higher degree polynomial) and e_i is a term that is called error?

The range of possible models with regard to the systematic part – the non-*error* part of the model is huge. In our little case, it is just the number of functions definable on the set of 20 values of i. It is perhaps of interest to mention that I remember with vividness being given a set of data of an experiment like this and the task of *analyzing* the data when I had completed a bachelor degree in mathematics at Cambridge. I was scared stiff – petrified, then. After many decades of being comfortable with the standard programs of *statistical methods*, I find I am again scared, except when randomization is used.

An aspect of standard statistical methods that should cause questioning, but seems not to, is the nature of *error*. What is this error that statisticians talk and write about? One *part* of error is error of measurement, and this is very easy to understand. We have a process of measurement and often, or always in our imagination, we can measure without affecting the object or entity being observed. We assume without questioning, it seems, that individual unknown

errors of measurements are independent realization of a scalar random variable. With this mode of thinking, it is natural to think of a large number of measurements of the entity being measured, and that the error of a particular measurement is the deviation of the result from the average. Curiously then, this error is conceptualized by means of what would be observed with repetition, with what might have happened – a notion objectionable, it seems, to *Bayesians.*

In a real experiment with the usual nature of experimental units, there are, in fact, differences between the units, and there will be differences between units in the absence of measurement error, with the same treatment, as we would observe in a uniformity trial. These are called *plot errors* or *experimental unit errors.* Is it proper to use the term *error* for such variability?

Suppose for definiteness that I wish to quantify the result of applying a treatment to 2 units: I do the *experiment* and I obtain 2 numbers y_1 and y_2. Is the difference between y_1 and y_2 an indication of error in this little study? We learned in our elementary statistics the role and importance of replication. I suggest, however, that we, including our founding fathers, have not thought out and told us what replication is. It seems easy and unquestionable that replication consists of repetition under constant circumstances. But we never have constant circumstances. Perhaps nearly so in a chemical or physical laboratory but not in, say, interventional research on humans. Fisher (1937, Sections 25 and 26) gives an interesting and relevant but not totally convincing discussion. In the case of the agronomic field experiment, he says that the problem of the impossibility of testing two or more treatments in the same year and on identically the same land can be overcome by testing the treatments on random samples of the same experimental area. Perhaps this will make my doubts seem reasonable. In the case of a field experimental area that is divided into parts, 2 plots are *the same* only if we agree to say this, and if we look at them sufficiently carefully, they will be found to be different. So it seems that we never have what may be called *real replication* in any sort of intervention experiment. This seems almost an absurd line of thought. We can have replication only in the sense of repeating a set of operations (e.g., of baking a cake).

A *natural* model to characterize the variability of the observations is to assume that the errors, e_i, $i = 1(1)20$, are a realization of a short section of a time series; e.g., a moving average process or an autoregressive process. However, one can surmise that the choice of a parametric class of models and subsequent fitting will be difficult. Finally, the assessment of uncertainty in treatment effects will be difficult. It is curious that methods based on such ideas have not been well developed and used.

The Fundamental Problem of the Intervention Experiment

I take the basic common structure of the intervention experiment to be that we have units of material that we index by i, and we have interventions that we index by j. If we conjoin unit i and intervention j, we obtain an observation

y_{ij}. The fundamental problem is that we cannot determine how the observation y_{ij} is caused. We cannot conjoin more than one treatment with unit i. If, for instance, we could observe y_{11} and y_{12}, we could conclude that the effect of treatment 2 minus the effect of treatment 1 on *unit* 1 is $y_{12} - y_{11}$. We shall observe, say, y_{11} and y_{22}. Then the difference $y_{22} - y_{11}$ can be attributed equally well (and equally badly) to this difference being the effect of treatment 2 minus the effect of treatment 1 or the effect of unit 2 minus that of unit 1. It is obvious that we have to deal with a set of units, some of which receive treatment 1 and some treatment 2. Suppose then we observe in a small experiment

$$y_{11} = 10,\ y_{22} = 15,\ y_{31} = 13,\ y_{42} = 20.$$

We are inclined to view that the effect of treatment 2 minus the effect of treatment 1 is

$$\tfrac{1}{2}(15 + 20) - \tfrac{1}{2}(10 + 13) = 6$$

But we can equally well conclude that this difference should be attributed to

$$\text{(unit 2 + unit 4) minus (unit 1 + unit 3)}$$

In fact, the size of the experiment is irrelevant to the difficulty. If we have treatment 1 on 1,000 units and treatment 2 on a different set of 1,000 units, whatever mean difference we observed can be equally well attributed to difference of effects of treatment 1 and treatment 2 or the difference between the 2 sets of units.

This leads to the absurd conclusion that we cannot determine whether any intervention produces some effect. Obviously the conclusion is false. What has often enabled the conclusion that an intervention is, e.g., successful, is a sort of empirical Bayesian reasoning. If, for instance, in the past all humans who have contracted a disease subsequently died, and one individual who contracted the disease and received an intervention survived, then one concludes that the intervention was successful. It may be, of course, that there is something unique about the individual and, thus, the intervention has not produced the successful outcome. One guesses that most so-called quack remedies have come about by this route.

This procedure is, of course, the method of historical controls, which has been very successful in many contexts. The method has been successful when the result produced after intervention is hugely different from the historical record.

The first act that must be considered in thinking about an intervention is to ask what the historical record is without intervention and with intervention. Such questioning is usual, of course, in the case of treatment for illness, especially when the intervention is not reversible or removable; e.g., in a partial gastrectomy. In many situations with a new intervention, there is no historical record of the outcome from it. In many cases, the outcome without intervention and

with intervention is very variable. Insofar as there is a historical record, it is imprecise and exhibits variability. It would then be very difficult to determine a historical control.

Even though the idea of a historical control is very appealing, there is a very difficult problem of deciding whether a proposed historical control is appropriate. What indeed makes a historical record relevant to evaluation of proposed intervention? In raising this question, I am thinking about interventional studies in connection with human illness and disease. We are told frequently that an attempt to determine if an intervention helps must incorporate its own controls. An exemplar case in which controls must be included in the experiment is that of agricultural research; for example, evaluation of a nutritional treatment on farm animals or farm crops.

Holland (1986) has written very informatively on the general problem I am discussing.

Design and Analysis

These are surely interrelated. The quality of a design can be determined only by means of the method of analysis and the quality of the conclusions. So the first step in considering design must revolve around the method of analysis.

The first step in *standard* theory of data analysis is to assume that the data D are a realization of a random variable X that has a distribution function F_X, which depends on a parameter θ. The next step is to determine if the data are in agreement with a particular value θ_0.

This step in Neyman-Pearson-Wald theory is to construct a rule for rejecting the hypothesis that $\theta = \theta_0$. This rule is to have the property that the probability under the model that it rejects $\theta = \theta_0$ when θ is in fact θ_0 is some pre-chosen α. Then, with this done for every θ_0, the values of θ that are not rejected by this rule are said to constitute a $(1 - \alpha)$ confidence set for the unknown θ.

Related to this process, but different from it, is the use of significance levels, often called P values. Inversion of the whole family of related significance tests of $\theta = \theta_0$ for a set of values of θ_0 gives a region of values of θ that *agree with the data* to a designated extent.

My preference is to regard the regions so obtained as consonance regions, regions that specify values of θ that are consonant with the data at chosen levels.

These procedures, however characterized by particular words, do not give probabilities of hypotheses such as probability that θ belongs to any chosen region of the parameter space.

If, then, the aim of the whole exercise, design, performance and analysis of the experiment is the obtaining of such probabilities, the procedures are totally unsuccessful.

The group of statisticians known as Bayesians take the position that the aim of all investigation must be the obtaining of such probabilities. Then it is obvious that one can reach the result with the introduction of a prior distribution. Unfortunately there is no logic that forces choice of a prior. It is

the conclusion of this line of development that the probability outcome is a belief probability that depends critically, obviously, on the prior belief probability.

My opinion is that the processes of science and technology do not require belief probabilities. The processes of science and technology require the obtaining of data under circumstances chosen by the investigator, and analysis of the data, which consists of making judgment of whether the data are consonant with particular models suggested by previous investigations or of determining new models from the data that are obtained. The idea that one has a realization from the *holy trinity* (to use a phrase of Basu) is simply ludicrous, so ludicrous that I can only suggest that those who base their ideas of learning about the real world, its present position and its dynamics have no experience of the nature of the processes one must use. One never knows the model! Did Newton know of the inverse square gravitation law? I say, "Obviously not". He and other scientists knew that motion of the planets was elliptic – they knew this by observation and data analysis. The *Bayesians* write as though the past workers *knew* that the law of force was d^{γ}, where d is the distance and γ is a parameter, and that they also had a belief distribution or a prior distribution on γ.

Another example that comes to my mind, though I have no depth of understanding, is the nature of the universe. It is expanding it seems, but will it continue to do so, or will it stop expanding or stay as it is, or start contracting and reach the size of a golf ball, or something even smaller? The idea that analysis of astronomical data should use a parametric model determined by some θ with a prior belief distribution on θ seems to me to be an antithesis of scientific method.

I therefore take the view that the Bayesian prescription, which is being heavily touted as the prescription by which all the uncertainty about this world in which we have to live can be handled, is not worth considering. The prescription is very beautiful in its simplicity and its power. There are many nice theorems in its theory. But it is based on assumptions and ideas that cannot be validated. It is true, of course, that any *reasonable* prior will be overcome by data eventually if the data come from an unvarying stochastic process. This, however, is essentially useless in that (a) any individual has a finite life and (b) the models that are consonant with past data change with new data. A critical process of science is the determination of a model that is consonant with all data accumulated in the past and then challenging that model, which is done only by new *experiments* and determining if predictions from the *old* model are realized in the *new* experiment. The lesson of science of the past century is surely that the models of yesteryear, while having predictive value for circumstances under which they were developed, are found to fail. It follows then that evaluations of goodness of fit (e.g., of the question of whether a prediction and the actual realization agree) is an essential element of science. It is, of course, an essential element of decision making. Where does the particular $(\mathfrak{S}, A, P, \theta)$ come from? The very neat presentations start off with the assumption that this is known. How silly this is! I think I have said enough.

Randomization "Inference"

I have tried to communicate my opinion that the usual frequentist theory and Bayesian theory, which purport to address the problems of inference and decision making, are failures. The failure of frequentist theory is not as deep, because it does recognize, though not at all adequately, that a stochastic model for a particular situation is a *pure invention*, which must be discovered, checked out and validated by means of real world data.

It is useful, perhaps, to discuss the matter of subjectivity and objectivity, which seems to require discussion forever (see, for example, Berger and Berry, 1988). The background seems to be the perception that Neyman-Pearson-Wald theory claims to be objective in contradistinction to Bayesian theory, which is subjective. The described polarity is partly *fake*. The real story is that both theories qua theories are theories, and neither is subjective or objective, just as a theory is not heavy or light in the sense of weight avoirdupois.

The only question is whether the practical use of either of the two *rival theories* is subjective or objective. My answer to this is that the NPW theory is partly objective in that the statistical models it uses must be confronted by the associated data, even though theory books say nothing about this. Practitioners of Bayesian theory (*if there really are any*) seem to *pull* their models and their prior distributions *out of thin air* but obviously do not. They do, however, make beliefs an absolutely essential component of their procedures, and any reasonable use of language must characterize the introduction of beliefs as subjective.

In the Bayesian framework, the conditional distribution of the supposed random variable given the parameter value is checkable. If for instance $X|\theta$ is $N(\mu, \sigma^2)$, we can check this by looking at a normal plot. If, however, we wish to adjoin to this the assumption that θ is $N(\nu, \phi^2)$, how are we to check the appropriateness of this assumption? Someone else could declare that he would like to assume that θ is distributed Cauchy (or whatever). Even more simply, where do ν and ϕ^2 come from? The fact that the values seem not to matter (but, of course, they do!) gives me no comfort, and I think I am not alone.

The obvious conclusions that should be drawn from the objective-subjective polarity that seems to be necessary are twofold:

(a) use of NPW theory requires data confrontation, which is *not* discussed in any theory book but uses use portions of general distribution theory and, obviously, significance tests to make such confrontation;

and

(b) use of Bayesian theory requires data confrontation, but this is not discussed in any exposition of the theory – I include exposition by *any* of the purported founding fathers. I shall not give references. Let any reader of this essay pull his (her) favorite exposition and examine it with respect to what I am discussing; in fact, I think, Bayesian users use the

distribution theory and significance tests that NPW users use; finally, Bayesian theory is subjective in that a prior is *plucked out of thin air* or quasi-derived by theory which itself is not validated for use even though based on axioms that seem (but are not) unchallengeable.

The story really is that NPW theory is the half-clothed emperor while Bayesian theory is the emperor without any clothes.

My discussion does not include empirical Bayes procedures which depend on data analysis in the choice of constituent distributions and face the same difficulties as NPW theory.

Where Do I Come Out?

I have given my views about the general mix of NPW decision theory and Bayesian theory. It seems to me that there are huge lacunae or gaps between the currently available theory and needed applications.

I now turn to the intervention experiment problem. My perception of the history is that our founding father, Fisher, recognized almost all the problems that I have mentioned, but was not as explicit as he could have been.

I am of the opinion that the assumption in a comparative intervention experiment that the outcome is a random variable from a probability distribution of a family of distributions indexed by some parameter of interest is not supportable.

So the question then is: Can anything be done? An answer is that something can be done; namely, use randomization in the conjoining of units and treatments and then use tests of significance (= tests of consonance) that are based on the frame of reference induced by the randomization process used.

Obviously, I am of the opinion that tests of significance are useful. If one regards them as useless, one is, it seems, in the position of being unable to determine objectively that a data set is not consonant with a particular model.

The value of randomization and the randomization test of significance in the randomized intervention experiment is that the probabilities that arise in the justification are not *belief probabilities* but are *frequency-in-repetition* probabilities determined by the randomization process used.

I think that most of the criticism of use of P values comes from a literal interpretation of Neyman-Pearson theory with its accept-reject rules and its type I error. Such a test of $\theta = 0$, say, carries with it the idea that θ may really be 0. In the significance testing outlook the achieved significance level is a measure of strength of evidence against the hypothesis $\theta = 0$. Use of the significance test of $\theta = 0$ carries no implication that θ may be exactly zero. Also, no one should have a strongly different outlook if P were 0.049 rather than 0.051 as Neyman-Pearson theory suggests.

The determination of confidence intervals or regions or, as I prefer, consonance intervals or regions for a parameter θ requires a formulation of how results resulting from θ_2 will differ from results from θ_1. In the case of

intervention experiments, the idea is used that if a unit with intervention j gives a result of y then with that same unit intervention j' would give the result $y + (\tau_{j'} - \tau_j)$.

What Randomization Process to Use?

This is, I judge, the basic question to be addressed. I think it was not addressed properly in past years. In experimentation on a line with say 8 units and 2 treatments denoted by A and B, the plan

$$AAAABBBB$$

is obviously a bad one.

What makes a plan *bad*? It is obvious that, with n units and t treatments, there are n^t possible treatment assignments, so in the case of 8 units and 2 treatments, there are 64 possible assignments, 2 of which are completely useless. The first plan is bad because the units are on a line. The 4 units that receive B occur later than the 4 units that receive A. In the second plan, B occurs after A in step. The third one that I give is a *sandwich* plan, which was discussed by Yates (1939). A plan is bad if the treatment assignment *favors* or *seems* to *favor* the treatments unequally. If one *knows nothing* about the units and they are labelled 1, 2 to 8, the first plan is not "bad". What one wants is that the plan be *balanced* with respect to the variability among the units that *one thinks may be present.* Choice of randomization process is then a matter of informal Bayesian thinking. A plan is *bad* is the investigator thinks so.

With 8 units on a line, *one may have the opinion* that the position of the units on the line tells one nothing about the variability among the units. One may think that most of the variability is expressed by a difference between the first 4 units and the second 4 units. One would then use this partition as a block partition. But, obviously, doing this is only part of the problem. One would still have to decide how to place A and B within each block. The sandwich plan seems not unreasonable. Another plan would be to partition the 8 units segmentally in *blocks of 2.* Then one would have to decide how to place the treatments within the resultant blocks. It is *reasonable* to surmise that

$$|\ AB\ |\ AB\ |\ AB\ |\ AB\ |$$

is a *bad* plan.

Suppose we wish to compare 2 treatments on a piece of land. We could partition the land into 2 pieces, one of which would receive A and the other B. This would be an appallingly bad choice. Why? What informative model can one use? How could one obtain an idea of error of conclusions? We could divide into 4 pieces of land, into 8 pieces, into 16 pieces, and then decide on a partitioning of the pieces into blocks. We could partition the land into a 2×2 array and assign the treatments according to a 2×2 Latin square. We could

partition the land into a 4×4 array and then use a plan in which A and B each occur twice in each row and in each column. There are undoubtedly many other possibilities.

How should one choose among all the possibilities? Why does the problem of choice arise? It arises because we have to decide how to partition the experimental material into *pieces* such that all subpieces of a piece receive the same treatment and then, of course, assign the treatments to the pieces. In the case of the agronomical field plot trial, the *pieces* are called *plots*, and the choice of plots is a matter that is discussed under the rubric *Field plot technique.* I shall not discuss this.

It is obvious intuitively that the *pieces*, the *plots* or the *experimental units* should be partitioned into subsets that are as *alike* as possible, with a subset for each treatment. But one can only guess about the *alikeness* of the units. One's guesses about *alikeness* may prove to be very poor. The actual experiment must be such that one can form a judgment about the *alikeness* of the units and then apply that judgment to form objective judgment about the *alikeness* of units receiving different treatments.

I am saying nothing new in these remarks. The ideas are all in Fisher's *The Design of Experiments.* Fisher discussed only two designs, the randomized block design and the Latin square design in that book. Various other designs are discussed by Cochran and Cox (1957). Later, Yates initiated the ideas of incomplete block designs and designs for two-way elimination of heterogeneity.

The ideas used for making analysis of the resultant data were those of linear models and analysis of variance. Fisher proved (insofar as Fisher proved anything!) that, if one used the customary randomization of the randomized block design and of the randomized Latin square design, then treatment comparisons were unbiased (meaning that the comparisons estimated by the use of the ordinary linear models and the method of least squares were unbiased for what one would observe if one could assign every treatment to every unit). Also the variance over randomizations of estimated treatment comparisons could be estimated by analysis of variance, if unit-treatment additivity holds, though Fisher was not aware of this requirement. Later, Yates gave the idea that the design should be unbiased in the sense that the expectation of the treatment mean square should equal the expectation of the residual (error) mean square in the absence of treatment effects.

The properties indicated in the previous paragraph hold in the case of the randomized design only if each block comprises a completely randomized design. The requirement for the Latin square design is unclear, except that the properties are realized if one chooses a Latin square plan from the totality of Latin squares of the given size.

It was realized, for example, by Grundy and Healy (1950), that the use of such randomization gave some realizations that were *bad.* For example, on a piece of land with 3 blocks of 4 plots, the blocks being aligned, one might obtain the plan:

BCAD
BCAD
BCAD

This is obviously a *bad* plan. An 8×8 Latin square design involving several factors each at 2 levels could result in the levels of one of the 2 level factors occurring in the 4 quarters of the square. Grundy and Healy made a suggestion of a restricted randomization plan. Youden (1956) discussed the problem, as did Sutter, Zyskind and Kempthorne (1963). There has been extensive work in recent years by Bailey and others.

The whole line of development with regard to restricted randomization appears to have been dominated by analysis of variance unbiasedness.

It is worthwhile to note that the randomized block design is a restriction of the completely randomized design and that the Latin square design is a restriction of a particular randomized block design, so the idea of restricted randomization goes back to the beginnings of the subject of design.

In recent years, I (Kempthorne, 1986a, b) have reached the opinion that the whole matter of randomization, and associated estimation and tests of significance, needs to be rethought in what is, conceptually, a very simple way. We realize, or should do so, that use of the *classical* designs is based on a sort of informal Bayesian process, in which one guesses or judges, or suspects or surmises (*but does not believe*) that the pattern of variability among the experimental units is such and such; for example, units within blocks are very much alike, while the units in different blocks differ appreciably.

The suggested procedure is that the experimenter specifies a set of plans, which he *surmises* will give *fair* comparisons among the treatments. He (she) then uses this set as a randomization frame for choice of plan that is used and for the randomization test of the null hypothesis of no treatment differences and for the randomization test of any shift alternative by adjusting the data to the null hypothesis.

In the case of *experimentation on a line*, the only attribute of a unit that is known is x, equal to its position. One can then pick out of the totality of plans, those for which Σx is nearly the same for the various treatments and Σx^2 is nearly the same: Any one plan in which this occurs can be regarded as a *systematic design*, of course. Indeed, any plan produced by randomizations looks to be systematic if one looks at it long enough.

I use the case of experimentation on a line because the implications are obvious. The extension of the basic idea to experimentation on a plane or a set of units in R^k is intuitively clear but not easy to implement.

My discussion brings to mind the argumentation in the '30s and '40s about the value of systematic designs. Fisher (1937) gives a discussion on this that is useful but not forcing. He assumed, without even mentioning so, that the *proper* way to analyze a systematic design was by means of the same linear model as that he used for a randomized design. He then gave compelling reasons in that framework for his view that the systematic Latin square designs were

variance biased in the sense that the expectation under the null hypothesis of the treatment mean square would be less than the expectation of the *error* mean square. I say that Fisher's discussion is not forcing because it is not at all clear that the analysis of the data set resulting from any plan should be based on the obvious Gauss-Markov-Normal-Linear Model (GMNLN) theory. Considerations of expectations, variances and covariances under randomization does suggest that GMNLN theory can be used as approximating randomization distribution theory if the classical randomization procedures are followed.

The Work of R. A. Bailey

Bailey (1983, 1985) has written very informatively on restricted randomization versus blocking and cites much literature that is strongly relevant. I suggest that these papers be read. She made (Bailey, 1983, p. 17) critical remarks about blocking that are very similar to those I have made in this essay and in Kempthorne (1986b, c), where I failed badly in not knowing and recognizing her work.

It appears that, if the plots (or units) lie be in a regular configuration with *nice dimensions* (e.g., a 2×4 or 8×8 array), one can bring ideas of permutation groups to bear.

I have three comments on this line of work. First, it seems that it is only in very special cases that the conditions demanded can be met. What is a good thing to do, for instance, with 12 units on a line and 3 treatments? Second, the requirement is imposed that the design has to be valid in the sense that the analysis of variance based on a linear model gives a treatment mean square and error mean square that have equal expectations under the randomization in the absence of treatment effects. Third, along with the use of analysis of variance, which I have just questioned, there is the problem of how to make tests of significance and how to make interval statements about treatment effects. This is where Kempthorne *came in* some decades ago. In his book (Kempthorne, 1952), he took the viewpoint that GMNLN theory can be used as an approximation to randomization theory, with respect to estimation of effects, estimation of error and statistical tests (and, hence, intervals on parameters). It is rather obvious, I think, that with restricted randomization this will not happen. It does not happen with small *classical* restricted randomized experiments; e.g., the 3×3 Latin square design (Kempthorne, 1952, pp. 193-195). It is on the basis of such thinking that I advocate the construction of a list of acceptable plans and using this list for design and statistical testing. The point is that an estimate and standard error of estimate are useless except for the construction of a pivotal, and a pivotal with distribution over 2 or 6 points is rather useless.

Some Closing Remarks on D. Basu

I have found myself in an anomalous position with respect to the writings of Basu (to whom I have referred at times as *my beloved enemy*). It is

obvious that Basu is highly expert in *mathematical statistics* at an advanced measure theory level. I can only admire this aspect. It is also obvious that Basu is deeply interested in inference. I have found that I agree rather strongly with some of his criticisms of NPW theory. However, I judge that Basu is *a sort of Bayesian*, and it is clear from the present essay, I imagine, that I am strongly averse to Bayesian writing that I have seen.

I am particularly averse to the introduction of formal Bayesian processes in the design and analysis of the comparative intervention experiment. I would like to read an account by a dedicated Bayesian of a real experimental situation, with the real outcome and with the statement of conclusions. In the absence of such, I suggest that Bayesian writings be ignored.

I am not at all clear on whether Basu has written on the problems I discuss. I hope that I have not committed any injustices.

I attempted (Kempthorne, 1980) to give my reactions to Basu's writing on the Fisher randomization (Basu, 1980) and decided that repetition of this would serve no useful purpose. The aspect that I did not emphasize then is the matter of design. The obviously Bayesian nature of design surely needs consideration. The discussion or argumentation of 1980 had little, if any, relevance to the problems of experimental method.

References

Bailey, R. A. (1983): Restricted randomization, *Biometrika* 70, 183-198.

Bailey, R. A. (1985): Restricted randomization and blocking, *International Statistical Review* 53, 171-182.

Bailey, R. A. (1986): Randomization constrained, *Encyclopedia of Statistical Sciences* 7, 519-524.

Basu, D. (1980). Randomization analysis of experimental data: The Fisher randomization test, *J.A.S.A.* 75, 575-582.

Berger, J. O. and D. A. Berry (1988): Statistical analysis and the illusion of objectivity, *American Scientist* 76, 159-165.

Cochran, W. G. and G. Cox (1957): *Experimental Designs*, Wiley, New York.

Finch, P. D. (1986): Randomization I, *Encyclopedia of Statistical Sciences* 7, 516-519.

Fisher, R. A. (1921): Studies in crop variation. I. An examination of the yield of dressed grain from Broadbalk, *J. Agric. Sci.* 11, 107-135.

Fisher, R. A. (1924): The influence of rainfall on the yield of wheat at Rothamsted, *Phil. Trans. Roy. Soc.* (B) 213, 89-142.

Fisher, R. A. (1935): *The Design of Experiments*, 2nd ed., 1937, Oliver and Boyd, Edinburgh.

Grundy, P. M. and M. J. R. Healy (1950): Restricted randomization and quasi-Latin squares, *J.R.S.S.* (B) 12, 286-2981.

Holland, P. W. (1986): Statistics and causal inference, *J.A.S.A.* 81, 945-970.

Kempthorne, O. (1952): *The Design and Analysis of Experiments*, Wiley, New York (reprinted Krieger).

Kempthorne, O. (1955): The randomization theory of experimental inference, *J.A.S.A.* 50, 946-967.

Kempthorne, O. (1975): Inference from experiments and randomization, in: *A Survey of Statistical Design and Linear Models*, ed. J. N. Srivastava, North Holland Publishing Co., Amsterdam, pp. 303-331.

Kempthorne, O. (1977): Why randomize?, *J. Statistical Planning & Inference* 1, 1-25.

Kempthorne, O. (1979): Sampling inference, experimental inference and observation inference, *Sankhya* 40, 115-145.

Kempthorne, O. (1984): Statistical methods and science, in: W. G. Cochran's *Impact on Statistics*, ed. P. S. R. S. Rao, Wiley, New York, pp. 287-308.

Kempthorne, O. (1986a): Comment on paper by D. Basu, *J.A.S.A.* 75, 584-587.

Kempthorne, O. (1986b): Randomization II, *Encyclopedia of Statistical Sciences* 7, 519-524.

Kempthorne, O. (1986c): Comparative experiments and randomization, in: *Statistical Design: Theory and Practice*, eds. C. E. McCulloch, S. J. Schwager, G. Casella, S. R. Searle, Cornell University, Ithaca, New York, pp. 43-48.

Kempthorne, O. and L. Folks (1971): *Probability, Statistics and Data Analysis*, Iowa State University Press, Ames.

Sutter, G. J., G. Zyskind, and O. Kempthorne (1963): Some aspects of constrained randomization, ARL 63-18 of Aeronautical Research Laboratories.

Yates, F. (1933): The formation of Latin squares for use in agricultural experiments, *Emp. J. Exp. Agric.* 1, 235-244.

Yates, F. (1939): The comparative advantages of systematic and randomized arrangements in the design of agricultural and biological experiments, *Biometrika* 30, 440-469.

Yates, F. (1970): A fresh look at the basic principles of the design and analysis of experiments, in: *Experimental Designs: Selected Papers of Frank Yates*, C. B. E., F. R. S., Hafner, Darien, Connecticut.

Youden, W. J. (1956): Randomization and experimentation, *Ann. Math. Stat.* 27, 1185-1186.